SCIENTIFIC PROGRAMME

MULTIMEDIA TECHNOLOGIES

MPEG-21 DIGITAL ITEM PROCESSING ARCHITECTURE

Frederik De Keukelaere
Wesley De Neve
Peter Lambert
Boris Rogge
Rik Van de Walle
Multimedia Lab
Department of Electronics and Information Systems
Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent,
Belgium
E-mail: Frederik.DeKeukelaere@rug.ac.be

KEYWORDS
multimedia, MPEG-21, functional metadata, Digital Item Processing

ABSTRACT

Within the world of multimedia, a new standard is currently under development. The purpose of this new standard, MPEG-21, is to create an open framework for multimedia delivery and consumption. This open framework allows its users to produce and consume a large variety of content in an interoperable manner. The second part of the MPEG-21 standard, the Digital Item Declaration Language (DIDL), makes it possible to declare, in XML, a Digital Item composed of multiple multimedia resources. This paper describes how additional metadata can be included to add processing information to a Digital Item Declaration (DID). This extra processing information allows DIDs to be processed in an interoperable manner.

INTRODUCTION

Today, many elements exist to build an infrastructure for the delivery and consumption of multimedia content. There is, however, no “big picture” to describe how these elements relate to each other. Therefore a new standard ISO/IEC 21000, better known as MPEG-21 (Bormans and Hill 2002), is currently under development. The aim of MPEG-21 is to describe how these various elements fit together and to realize the “big picture”.

The result of MPEG-21 will be a new standard which provides an open framework for multimedia delivery and consumption. This generic multimedia framework will allow its users to create and consume multimedia on a large variety of terminals and networks.

Because of the generic character of MPEG-21, it will be suited for a large range of multimedia applications. This application independence implies that MPEG-21 needs to be a standard with great flexibility concerning the processing of the multimedia elements of the application.

In this paper we provide such flexibility by applying the ideas of Functional Metadata (Rogge et al. 2002) to MPEG-21 Digital Items. Digital Items were defined in Part 1 of MPEG-21 (Bormans and Rump 2001). Functional Metadata was designed to add functionality to multimedia presentations. Extending those ideas to MPEG-21 will allow the author of Digital Items to express his/her intentions concerning the processing of their Digital Items.

DIGITAL ITEM DECLARATION

The second part of MPEG-21, the Digital Item Declaration Language (DIDL) (Schwartz et al. 2002), makes it possible to declare a Digital Item composed of multiple multimedia resources. It allows *Users*, defined as both humans and machines in MPEG-21 (Bormans and Rump 2001) (i.e., any entity that interacts with the MPEG-21 environment at any point throughout the multimedia delivery and consumption chain), to describe in XML, the relationship between the different elements of a Digital Item.

As an example of how Digital Items can be used, suppose that a User would like to create a digital music album. This album could consist of two different music tracks, the album’s title and the titles of the two songs. An example of a Digital Item Declaration suited for this purpose can be found in figure 1. As can be seen from this figure, the Digital Item Declaration Language is a container structure allowing the author of a multimedia presentation to specify the multimedia elements that are part of the multimedia presentation.

The real power of the Digital Item Declaration Language resides in the possibility to include any type of media, going from text to future types of media that are currently undefined, and the possibility to configure the Digital Item conform the User’s preferences. The latter can be expressed using the choice mechanism included in the Digital Item Declaration Language.

As an example of use of the choice mechanism, suppose that the User who created the music album would like to include the possibility to choose between two different formats for the music tracks. Figure 2 shows how a more complex version of the music album could be used for this scenario. The Digital Item Declaration in that figure contains two different types of audio formats (i.e., wave and mp3 format) and a *Choice* element that allows the User to choose one of the audio formats.

```
<?xml version="1.0" encoding="UTF-8"?>
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-
NS">
  <Item>
    <Descriptor>
      <Statement mimeType="text/plain">
        Album title
      </Statement>
    </Descriptor>
    <Item id="track01">
      <Descriptor>
        <Statement mimeType="text/plain">
          Song title 01
        </Statement>
      </Descriptor>
      <Component>
        <Resource ref="track01.mp3"
          mimeType="audio/mpeg"/>
      </Component>
    </Item>
    <Item id="track02">
      <Descriptor>
        <Statement mimeType="text/plain">
          Song title 02
        </Statement>
      </Descriptor>
      <Component>
        <Resource ref="track02.mp3"
          mimeType="audio/mpeg"/>
      </Component>
    </Item>
  </Item>
</DIDL>
```

Figure 1: A Digital Item Declaration expressed in the Digital Item Declaration Language (DIDL)

THE NEED FOR DIGITAL ITEM PROCESSING

In the previous examples a small part of the expressivity of the Digital Item Declaration Language has been demonstrated. Both examples also illustrate that a Digital Item Declaration (DID) is a static XML declaration. Because of this a DID can be seen purely as a container combining metadata and resources while specifying the relationships between them. However, this means that the author of a Digital Item has no way to express how a Digital Item should be processed and what an MPEG-21 terminal should or could do with it.

Therefore, it is possible that different MPEG-21 terminals handle the same DID differently. From the end user's point of view, this is perceived as a lack of interoperability. Without a specification for processing Digital Items, authors cannot express how they want their Digital Items to be processed.

Now, reconsider the scenario of an author making a Digital Item containing a music album. In the new scenario the music album of figure 1 and figure 2 is extended to contain music clips, some advertisements and the necessary information to protect the intellectual property rights of the author. Suppose that the author had the following intentions:

- Present the different videos and songs to the User (for example a consumer) if the User has bought the rights to listen and watch them.

```
<?xml version="1.0" encoding="UTF-8"?>
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-
NS">
  <Item>
    <Descriptor>
      <Statement mimeType="text/plain">
        Album title
      </Statement>
    </Descriptor>
    <Choice>
      <Selection select_id="WAV"/>
      <Selection select_id="MP3"/>
    </Choice>
    <Item id="track01">
      <Descriptor>
        <Statement mimeType="text/plain">
          Song title 01
        </Statement>
      </Descriptor>
      <Component>
        <Condition require="WAV"/>
        <Resource ref="track01.wav"
          mimeType="audio/wav"/>
      </Component>
      <Component>
        <Condition require="MP3"/>
        <Resource ref="track01.mp3"
          mimeType="audio/mpeg"/>
      </Component>
    </Item>
    <Item id="track02">
      <Descriptor>
        <Statement mimeType="text/plain">
          Song title 02
        </Statement>
      </Descriptor>
      <Component>
        <Condition require="WAV"/>
        <Resource ref="track02.wav"
          mimeType="audio/wav"/>
      </Component>
      <Component>
        <Condition require="MP3"/>
        <Resource ref="track01.mp3"
          mimeType="audio/mpeg"/>
      </Component>
    </Item>
  </Item>
</DIDL>
```

Figure 2: Choices and different types of media in a Digital Item

- After the preview, present a form with the possibility to buy the rights to listen to certain or all parts of the music album.
- If the User buys certain parts of the music album construct a message in a predefined format and send it to distributor's sales server.
- During the whole process present advertisements about relevant products.

MPEG-21 as it is currently specified allows the inclusion of all the necessary data in a Digital Item to make processing as presented possible. However, currently there is no mechanism that allows the specification of the processing itself, i.e., there is currently no way for a User (for example an author) to include information on how the processing should be done.

```
<?xml version="1.0" encoding="UTF-8"?>
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-
NS">
  <Item>
    <Item>
      <!-- a list of DIP Methods -->
      <Component>
        <Resource mimeType="text/xml"
         ref="standardMusicAlbum#playTrack"/>
      </Component>
    </Item>
    <Item>
      <!-- mapping information -->
      <Component>
        <Descriptor>
          <Statement mimeType="text/plain">
            StandardMusicTrack
          </Statement>
        </Descriptor>
        <Resource mimeType="text/plain">
          track01
        </Resource>
        <Resource mimeType="text/plain">
          track02
        </Resource>
      </Component>
    </Item>
  </Item>
</DIDL>
```

Figure 3: A PDI containing the processing information for a CDI

```
<?xml version="1.0" encoding="UTF-8"?>
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-
NS">
  <Item>
    <Item id="playTrack">
      <!-- playTrack method -->
      <Item>
        <!-- arguments of the method -->
        <Component>
          <Resource mimeType="text/plain">
            StandardMusicTrack
          </Resource>
        </Component>
      </Item>
      <Item>
       <!-- implementation of the method -->
       <Component>
        <Resource mimeType="text/javascript">
    function playTrack(StandardMusicTrack)
    {
      //actual implementation
      play(StandardMusicTrack,"audio");
    }
        </Resource>
       </Component>
      </Item>
    </Item>
  </Item>
</DIDL>
```

Figure 4: A MDI containing the method's signature and implementation

DIGITAL ITEM PROCESSING

Processing information as metadata

Metadata can be generally defined as "structured data about data". What part of the data is metadata and what part of the data is the actual content depends on the application responsible for processing the data. Data that is metadata for one application can therefore be content for another application and vice versa.

In the context of this paper, processing information is actually data (containing the processing information) about the data contained within the DID and can hence be seen as metadata of the Digital Item.

Addressing the requirements of Digital Item Processing

In the previous sections it has been made clear why there is a need for Digital Item Processing. In order to respond to this need MPEG has specified the requirements for Digital Item Processing (MPEG 2002). This paper tries to address these requirements and proposes an architecture and implementation that satisfies the Digital Item Processing requirements.

Terminology

We introduce terminology that can be used in Digital Item Processing:

- A Digital Item Processing Method (DIP Method) is a method that can be applied to a Digital Item Declaration.
- A Digital Item Processing Engine (DIP Engine) is the part of the MPEG-21 terminal that will be responsible for the execution of the DIP Method.
- An Object in a Digital Item Declaration is every DIDL element that can be identified by a unique id or choice id, i.e. a Container, an Item, a Component, a Descriptor, an Anchor, a Choice or an Annotation.
- An Object Map is the association of the Objects in a Digital Item with an Object Type.
- An Object Type is a semantic type that can be assigned to an argument of a DIP Method.

DIFFERENT TYPES OF DIGITAL ITEM DECLARATIONS IN DIGITAL ITEM PROCESSING

For the use in Digital Item Processing, three different types of Digital Item Declarations are introduced:

- A Content Digital Item (CDI) is a Digital Item Declaration that contains the description of the relations between the different components of a multimedia presentation (figure 2).
- A Processing Digital Item (PDI) is a Digital Item Declaration containing the Object Map and the list of DIP Methods. It will be included in the CDI as metadata (figure 3).

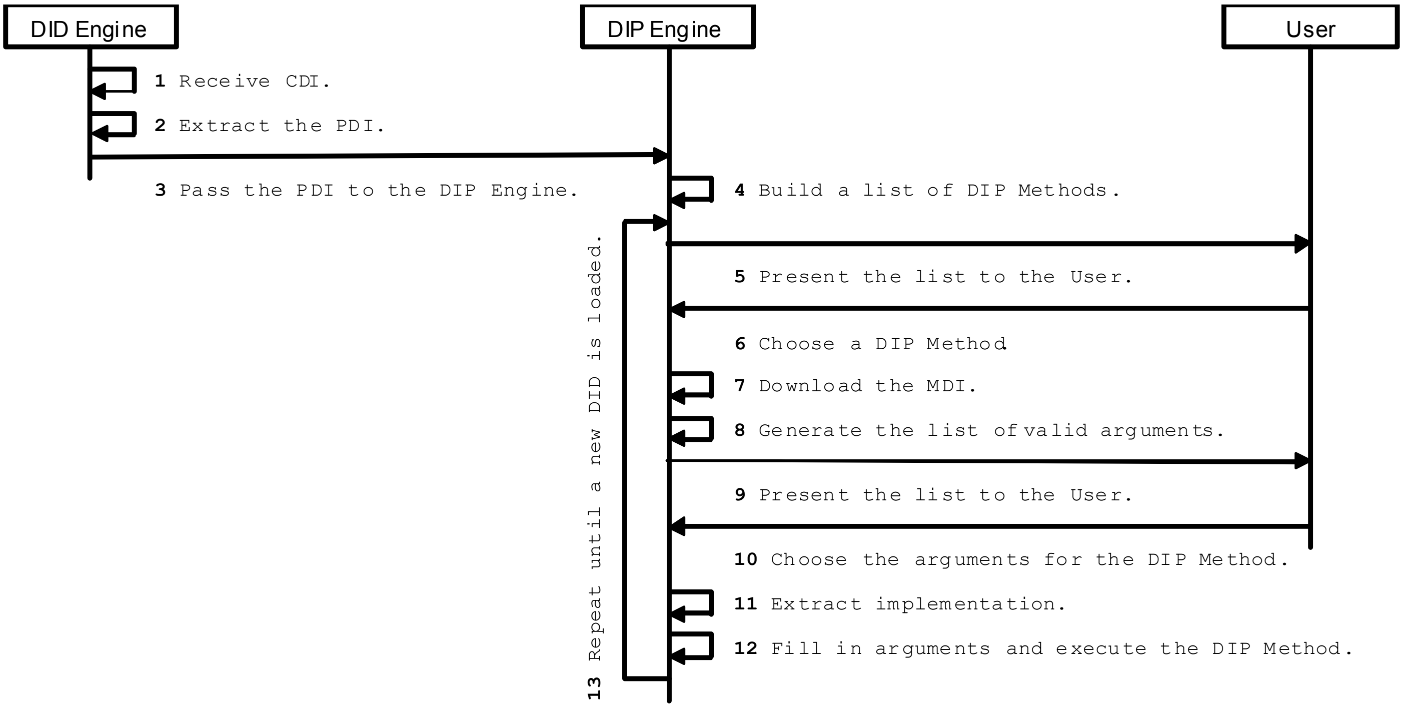

DID Engine
DIP Engine
User
1 Receive CDI.
2 Extract the PDI.
3 Pass the PDI to the DIP Engine.
4 Build a list of DIP Methods.
5 Present the list to the User.
6 Choose a DIP Method
7 Download the MDI.
8 Generate the list of valid arguments.
9 Present the list to the User.
10 Choose the arguments for the DIP Method.
11 Extract implementation.
12 Fill in arguments and execute the DIP Method.
13 Repeat until a new DID is loaded.

Figure 5: A high level algorithm for Digital Item Processing

- A Method Digital Item (MDI) is a Digital Item Declaration containing the prototype and the implementation of the DIP Methods a PDI refers to (figure 4).

Processing Digital Item

By defining how the data in a Processing Digital Item should be interpreted, it is possible to reuse DIDL syntax to include processing information for the CDI. Therefore the Processing Digital Item can be constructed according to the Digital Item Declaration Language and hence it will be a correct Digital Item.
The processing information included in a Processing Digital Item consists of a list of DIP Methods that can be applied to the Content Digital Item and the mapping of DID Objects in the Content Digital Item to a specific Object Type.
As an example, the Processing Digital Item in figure 3 contains the following information: a DIP Method ”play-Track”, for which the prototype and implementation can be found in the MDI standardMusicAlbum; and the Object Map. From the Object Map it can be derived that the DID Objects of the Content Digital Item, identified by the ids “track01” and “track02” are of the Object Type “Standard-MusicTrack”. This means that for every DIP Method that needs an argument of the Object Type “StandardMusicTrack” both “track01” and “track02” can be chosen as argument for such a DIP Method.

Method Digital Item

The same approach can be followed considering the Method Digital Item. In this case, we use a predefined structure to differentiate between the arguments of a method and the actual implementation of that method.
Figure 4 shows a Method Digital Item containing a “playTrack” DIP Method. This DIP Method has a single argument of the Object Type “StandardMusicTrack”. The Method Digital Item also includes how the “playTrack” DIP Method can be implemented in JavaScript, although implementations in other scripting/programming languages are possible too. This allows Digital Items to be processed on different platforms with different capabilities, i.e. with support for different scripting/programming languages.

A HIGH LEVEL ALGORITHM FOR PROCESSING DIP DATA

It is rather obvious that an MPEG-21 terminal needs to understand the structure and the information contained in the different types of DIDs. Therefore we introduce the concept of a DIP Engine that can be used to build a Digital Item Processing environment. This DIP Engine can be included in an MPEG-21 terminal and will be responsible for the processing of a Content Digital Item by using the information stored in a Processing Digital Item and a Method Digital Item. In this paper the information flow within a DIP enabled MPEG-21 terminal is presented by means of a “high level algorithm” for which a graphical representation can be found in figure 5. The algorithm in question can be described as follows:

1. In the first step, the DID Engine (MPEG 2002) (the part of an MPEG-21 terminal that is responsible for the parsing of the DIDs) receives the Content Digital Item.

2. The DID Engine extracts the Processing Digital Item, which was added as metadata to the Content Digital Item.
3. The DID Engine passes the Processing Digital Item to the DIP Engine.
4. The DIP Engine builds a list of methods, called DIP Methods, which can be applied to the Objects in the Content Digital Item.
5. The list of DIP Methods is presented to the User.
6. The User chooses which DIP Method that needs to be executed.
7. Based upon the information that resides in the Processing Digital Item, the DIP Engine downloads the Method Digital Item with the information about the DIP Method's prototype and implementation.
8. The DIP Engine uses the information of the prototype (extracted from the MDI) and of the Object Map (extracted from the PDI) to build a list of valid arguments for the DIP Method chosen by the User. By example, valid arguments for the "playTrack" DIP Method are DID Objects of the Object Type "StandardMusicTrack", therefore both "track01" and "track02" are valid.
9. This list is presented to the User.
10. The User chooses the arguments for the chosen DIP Method.
11. The DIP Engine extracts the implementation of the DIP Method from the Method Digital Item.
12. In the last step, the DIP Engine fills in the arguments of the DIP Method and executes the DIP Method.
13. Repeat steps 5 to 13 until the MPEG-21 terminal loads a new Digital Item.

CONCLUSIONS AND FUTURE WORK

This paper presented a solution to the problem of adding processing information to Digital Item Declarations. We have addressed the Digital Item Processing requirements by introducing two new types of Digital Items: the Processing Digital Item and the Method Digital item.

The Processing Digital Item is designed to be included in the Content Digital Item as metadata. It contains a list of DIP Methods that can be applied to the Content Digital Item. It also contains the Object Map. The latter provides a mechanism that allows the Objects in the Content Digital Item to be used as arguments for the DIP Methods in a generic way. The Method Digital Item contains the prototype and the implementation of the DIP Methods whereto the Processing Digital Item refers.

Besides introducing the Processing Digital Item, the Method Digital Item and the Digital Item Processing terminology, we have provided a high level algorithm that allows the implementation of the Digital Item Processing requirements using both the Processing Digital Item and the Method Digital Item. The proposed algorithm allows Digital Item Declarations to be processed in an interoperable and platform independent manner.

To allow interoperability at the level of the implementation of DIP Methods, we need to develop and standardize a high level language that is powerful enough to express the desired functionality in an easy way.

Finally, an online demo of the MPEG-21 Digital Item Processing Architecture can be found at the following address: http://multimedialab.elis.rug.ac.be/demo.asp.

ACKNOWLEDGMENTS

The research activities that have been described in this paper were funded by Ghent University, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the Belgian Federal Office for Scientific, Technical and Cultural Affairs (OSTC).

REFERENCES

Bormans, J.; and K. Hill. 2002. "MPEG-21 overview v 5.0." *ISO/IEC JTC1/SC29/WG11 N5231.*

Bormans, J.; and N. Rump. 2001. "ISO/IEC TR 21000-1: MPEG-21 Multimedia Framework Part 1: Vision,Technologies and Strategy."

Burnett I.; and R. Van deWalle. 2002. "Current Vision on MPEG-21 Digital Item Processing." *ISO/IEC JTC1/SC29/WG11 N5228.*

MPEG. 2002., "Draft requirements for digital item processing." *ISO/IEC JTC1/SC29/WG11 N4989.*

Rogge, B.; D. Van De Ville; I. Lemahieu; and R. Van de Walle. 2002. "Validating MPEG-21 encapsulated functional Metadata." in *Proceedings IEEE International Conference on Multimedia andExpo (ICME)*, Lausanne, 2002, 4 pages, published on CD-ROM.

Schwartz, T.; V. Iverson; Y.W. Song; R. Van de Walle; D. Chang; and E. Santos. 2002. "MPEG-21 digital item declaration FDIS." *ISO/IEC JTC1/SC29/WG11 N4813.*

AUTHOR BIOGRAPHY

FREDERIK DE KEUKELAERE was born in Ghent, Belgium April 23, 1980. He finished highschool in 1998. After completing basic military training for officers, he went studying at the Ghent University. He received a degree in Computer Science from the Ghent University in 2002. The subject of his graduation thesis was "Temporele synchronistatie van multimedia in MPEG-21-declaraties van digitale items" ("Temporal synchronization of media within MPEG-21 Digital Items"). Since his graduation, he has been working as a PhD researcher at Multimedia Lab - department of Electronics and Information Systems (ELIS) - at the Ghent University.

METADATA-BASED ACCESS TO COMPLEX DIGITAL OBJECTS IN MULTIMEDIA ARCHIVAL COLLECTIONS

Jeroen Bekaert
Robbie De Sutter
Rik Van de Walle
Department of Electronics and
Information Systems
Ghent University
Sint-Pietersnieuwstraat 41
B 9000, Ghent,
Belgium
{jeroen.bekaert, robbie.desutter,
rik.vandewalle}@rug.ac.be

Emiel De Kooning
Department of Architecture and
Urbanism
Ghent Univeristy
Jozef Plateaustraat 22
B 9000, Ghent,
Belgium
emiel.dekooning@rug.ac.be

KEYWORDS
Multimedia Archival Collections, Metadata, METS,
EAD, MPEG-7

ABSTRACT

A comprehensive approach to the access of digital objects necessitates the interplay of various types of metadata standards. Each of these standards fulfills its own part within the context of a 'metadata infrastructure'. In recent years, due to the growth in computational power and the increasing networking bandwidth and connectivity, electronic archival records have evolved from simple text-based files to complex digital objects that may contain embedded images (still and moving), drawings, sounds, hyperlinks, or complex multimedia data. The aim of this research is to investigate and implement a number of tools supporting time-dependent media within an archival context. Therefore, we have presented a metadata framework consisting of both the Encoded Archival Description (EAD) standard – supporting collection level descriptions – and the Metadata Encoding and Transmission Standard (METS) – supporting the digital item descriptions –. Then, an extension of this metadata framework has been proposed to time-based media content delivered via the MPEG-7 multimedia content description standard.

INTRODUCTION

In recent years, there has been a tremendous growth in computational power, and in networking bandwidth and connectivity. As a result, the number of organizations making digital information available has massively grown. This digitization process encouraged the development of standards that support the management, description, indexing, and long-term preservation of digital information.

Moreover, there has been a considerable increase of available digital multimedia data (digitized books, photo albums, audio of various nature, video fragments, etcetera) within the last few years. This enthusiasm has revealed the difficulties associated to managing such resources, mainly due to their joint complexity and a lack of appropriate indexing standards. A large number of recently started projects related to the resource discovery of different media types, including music, speech records, video, and images, indicate an acknowledgement of this problem and the importance of this field of research for both digital libraries and archives.

This research presents a metadata framework dealing with both time-dependent multimedia objects within the context of an archival collection. The paper starts with a brief overview of how archival collections can be described. Section III expresses briefly a mechanism for describing complex digital items. Section IV handles the descriptions of time-dependent data and how they can be incorporated in the proposed metadata infrastructure. Section V presents a prototype appplication and to conclude, we enumerate some further adjustments and draw some conclusions.

MECHANISMS FOR DESCRIBING ARCHIVAL COLLECTIONS

In an archival fund, the finding aid is an important tool for resource description. Finding aids differ from catalog records by being much longer, more narrative and explanatory, and highly structured in a hierarchical fashion. They generally start with a description of the gathering as a whole, indicating what types of materials it contains and why they are important. The finding aid describes the series into which the collection is organized, such as correspondence, business records, personal papers, and campaign speeches, and ends with an itemization of the contents of the physical boxes and folders comprising the collection.

Consequently, the description at various levels is a fundamental part of the archival descriptive practice. For resource discovery, the existence of collection-level descriptions supports the multi-level navigation of a large and heterogeneous archive collection. For example, a researcher may make use of collection-level descriptions to find the existence of a specific collection. Furthermore, he can target a selected collection by submitting collection-specific queries. Moreover, collection-level descriptions may be used to support controlled searching across multiple collections, and to assist users by reducing the number of individual hits returned in an initial response to a query (Waibel 2001).

```
<!ELEMENT  archdesc          (runner*, did, (%m.desc.full;)*)        >
<!ENTITY   % m.desc.full     '%m.desc.base; | dsc |
                             dao | daogrp | note'                    >
<!ENTITY   % m.desc.base  'admininfo | bioghist |
                             controlaccess | odd | … |               >
<!ATTLIST  archdesc
       %a.common;
       type                (inventory | register |
                            othertype)              #IMPLIED
      othertype            CDATA                    #IMPLIED
       %a.desc.all;
        encodinganalog     CDATA                    #IMPLIED
        relatedencoding    CDATA                    #IMPLIED        >
<!ENTITY   % a.common
        'id                ID                       #IMPLIED
        altrender          CDATA                    #IMPLIED

        audience           (external | internal)  #IMPLIED'         >
<!ENTITY   % a.desc.all
        'level             (%av.level;)             #REQUIRED
       %a.desc.base;'                                                >
<!ENTITY   % av.level      'series | collection | file |
                           fonds | item | otherlevel | … |           >
<!ENTITY   % a.desc.base
        'otherlevel     CDATA                     #IMPLIED
        langmaterial       CDATA                    #IMPLIED
       legalstatus         (public | private |
                           otherlegalstatus)        #IMPLIED
       otherlegalstatus    CDATA                    #IMPLIED'       >
```

Figure 1. EAD DTD for the <archdesc> element.

The archival community has already developed well-established national and international archival standards among which the General International Standard Archival Description (ISAD(G)) (International Council on Archives 2000) and the Encoded Archival Description (EAD) (The Library of Congress 1999) are most important. The EAD is capable of describing a collection as a hierarchy of potentially infinitely nesting levels down to the item level itself (Figure1). This standard was initiated by the University of California at Berkeley, and its development has been supported by the Society of American Archivists Standards Board. Because of its extensive facilities to link to digital objects, it is able to describe digital collections as well as their more traditional counterparts. At each level of description, the EAD supports varying degrees of specificity. At each level of description, the EAD supports varying degrees of specificity. It should also be noticed that the EAD can be successfully mapped to the ISAD(G) description standard. However, not all EAD-elements are conform to the ISAD(G) description rules, as the description parameters of the EAD standard extend much further.

MECHANISMS FOR DESCRIBING DIGITAL OBEJCTS

However, the EAD standard excels at item discovery, it has little provision for exploring the digital object itself. Therefore, the strategy for granular collections needs to be complemented with another schema capable of encapsulating single and complex multipart digital items. Yet another closely allied community, the library world, has developed a schema which adds the desired functionality to our information architecture. This schema is currently known as the Metadata Encoding and Transmission Standard (METS) (The Library of Congress 2002). This METS Standard is a rather newly devised metadata standard, which refines and extends the much earlier Making of America II metadata rules (Beaubien and Hurley 2001). It provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories (or between repositories and their users).

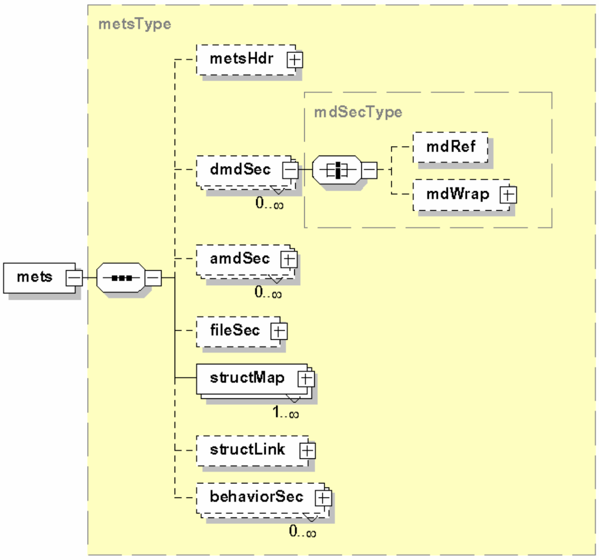


Figure 2: Metadata Encoding and Transmission Standard

The METS framework provides containers for descriptive and administrative metadata encoded according to standards external to METS itself. This offers the possibility to provide an interplay between both the EAD and the METS standard. Archives should be using EAD to provide metadata regarding an archival collection in its entirety, including describing the relationship of the various items in the collection to one another. Then, part of these items may include XLinks to a digital representation (i.e. the location and description of a Digital Archival Object or Record Group) of an archival object, which calls on the METS object (Figure2).

Likewise, the METS document will have to refer back to the specific position within the EAD collection as the descriptive home of the object. This way, we establish a bi-directional linking between the METS object and the corresponding EAD document.

An example of a more complex digital library object is a digitized photo album, which can be pointed to by an EAD (which can be successfully mapped to the ISAD(G) description standard) encoded finding aid. The diary may have 20 pages that were digitized as 600 dpi TIFF master files, 20 lower resolution JPEG viewing files, and 20 thumbnail files which were derived from the masters for each page. The digital diary object now includes the content (60 files of digitized page images), a reference to the EAD archival collection description metadata in which the digital item has to be situated, metadata describing the content.

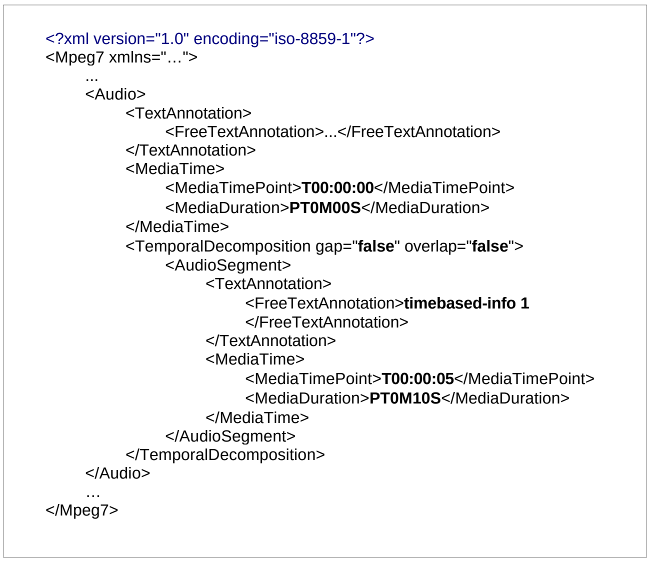
MECHANISMS FOR DESCRIBING MULTIMEDIA FEATURES

While this functionality easily accommodates items such as ‘photo-album’ objects, it still reveals its limits when the object in question is of a more time dependent media type. Clearly, the metadata framework needs to be complemented with other schemes and tools capable of managing and describing multimedia content and properties, such as the MPEG-7 Multimedia Content Description Standard (Martinez 1999).

Likewise the EAD standard, the METS framework can be extended with descriptive and administrative metadata responsible for the management of multimedia information. This way, we can attach catalog, semantic and information based on the MPEG-7 Multimedia Description Scheme to the metadata framework. This extension offers also the possiblity describing both temporal and spatial structures of the content (Figure3).

Hereby it is interesting to note that the METS framework offers the possibility for area linking, meaning the structMap section of the standard can reference a specific point or range in an audio/video file; i.e., a segment of a file (e.g., text screen, image line, audio/video clip), or a two-dimensional section of a file (e.g., subsection of an image, or a subsection of the video display of a video file). In addition to linking directly with only one file, multiple area elements may be used within parallel or sequential elements to link (regions in/parts of) multiple files to each other in a sequential or hierarchical order. This way we can construct within the Structural Map section of a METS document –temporal or spatial– segments composed of one or multiple components from one or multiple source objects respectively.

The MPEG-7 Segment Description Scheme can provide an extension scheme for describing the temporal and spatial segements and technical features related to these different segments. It describes the result of a spatial, temporal, or spatio-temporal partitioning of the audio-visual content. It can describe a recursive or hierarchical decomposition of the audio-visual content into segments that form a segment tree. The Segment DS forms the base abstract type of the different specialized segment types: audio segments, video segments, audio-visual segments, moving regions, and still
regions. As a result, a segment may have spatial and/or temporal properties. For example, the AudioSegment DS can describe a temporal audio segment corresponding to a temporal period of an audio sequence, the VideoSegment DS can describe a set of frames of a video sequence; the StillRegion DS can describe a spatial segment or region of an image or a frame in a video and so on.



```
<?xml version="1.0" encoding="iso-8859-1"?>
<Mpeg7 xmlns="…">
    …
    <Audio>
        <TextAnnotation>
            <FreeTextAnnotation>...</FreeTextAnnotation>
        </TextAnnotation>
        <MediaTime>
            <MediaTimePoint>T00:00:00</MediaTimePoint>
            <MediaDuration>PT0M00S</MediaDuration>
        </MediaTime>
        <TemporalDecomposition gap="false" overlap="false">
            <AudioSegment>
                <TextAnnotation>
                    <FreeTextAnnotation>timebased-info 1
                    </FreeTextAnnotation>
                </TextAnnotation>
                <MediaTime>
                    <MediaTimePoint>T00:00:05</MediaTimePoint>
                    <MediaDuration>PT0M10S</MediaDuration>
                </MediaTime>
            </AudioSegment>
        </TemporalDecomposition>
    </Audio>
    …
</Mpeg7>
```

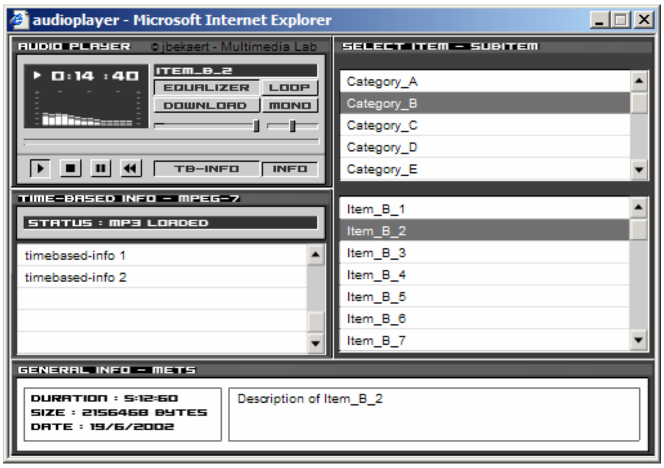
Figure 3: MPEG-7 Audio Segment Information (referenced by METS)

A PROTOTYPE APPLICATION

Based on the presented metadata framework, we developed an experimental client prototype application, capable of dealing with time-based content presentation. Therefore, a browser based front-end loads an external audio stream, and associated administrative and structural metadata, as well as the metadata, describing the temporal content related to the audio. The main issue hereby is the temporal synchronization of the audio stream and the associated temporal information. For instance, the digital object may consist of a speech record, consisting out of multiple segments. Each segment can have associated time-based information encoded in the MPEG-7 XML format. Note that the temporal synchronization should be maintained, even if the stream is interrupted. The actual playback of the application can be subdivided into the following sections:

- The server sends an EAD-document describing the hierarchical structure of a the multimedia collection (the collection used for the protoype is limited to 2 levels).
- The client requests the METS information corresponding to a specific digital object request of the user. For instance give me audio fragment no 5 in category no 2.
- The corresponding METS document is sent to the client.
- The application makes a connection with the external audio stream, based on the links contained within the METS fileSec section.
- The application receives the external (MPEG-7) metadata streams, based on the references contained within the METS dmdSec and amdSec sections. During playback of the audio stream, additional MPEG-7 time-dependent information is delivered.

The demo has been developed using the Shockwave scripting language. This is an object-oriented scripting language based heavily on the ECMAScript Language Specification (ECMA-262) (Figure4).



STATUS : MP3 LOADED

timebased-info 1
timebased-info 2

DURATION : 5:12:60
SIZE : 2156468 BYTES
DATE : 19/6/2002

Description of Item_B_2

Figure 4: screenshot of the prototype application

SHORTCOMMINGS AND CONCLUSIONS

A metadata architecture based on the standards outlined, so far delivers navigable and hierarchically structured objects of multimedia data. However, the framework we described has some shortcomings that we haven't addressed yet. So far, little attention has been paid to integrating the types of information that describe the technical characteristics of the digital items.

A more broad research should consider technical standards such as the Technical Metadata for Digital Still Images, developed by the National Inforamtion Standards rganization, the Audio Engineering Society (AES) and the Society of Motion Pictures and Television Engineers, both with links to the European Broadcasting Union. Furthermore, additional attention should be paid to the comparison of the proposed framework to the temporal synchronization of media within MPEG-21 Digital Item Declarations by using the Synchronized Multimedia Integration Language (SMIL) (De Keukelaere et al. 2001).

REFERENCES

Beaubien R. and Hurley B., University of California, Berkeley Library, *The Making of America II*, http://sunsite.berkeley.edu/MOA2, 2001.

Bekaert J., De Sutter R., Lerouge S., Rogge B., Van De Ville D., De Kooning E., and Van de Walle R., 2002, “Metadata-based access to multimedia architectural and historical archive collections”, *in Internet Multimedia Management Systems III*, John R. Smith, Thomas J. Watson, Sethuraman Panchanathan, and Tong Zhang, Eds., vol. 4862 of Proc. of SPIE. (July)

De Keukelaere F., Rogge B. and Van de Walle R., 2001, *Temporal synchronization of media within MPEG-21 Digital Items*, ISO/IEC JTC1/SC29/WG11 M7728, Pattaya.

International Council on Archives, 2000. *ISAD(G): General International Standard Archival Description*; Second edition, Ottawa

Martinez J. M., *Overview of the MPEG-7 Standard*, Technical Report, ISO/IEC JTC1/SC29/WG11/N3158, 1999.

The Library of Congress, 1999. *EAD Application Guidelines for Version 1.0*, http://www.loc.gov/ead/ag/agappb.html.

The Library of Congress , 2002, Metadata Encoding and Transmission Standard, http://www.loc.gov/ standards/mets (Feb)

Waibel G., 2001. “Granular Collections Access. An Information Architecture Informed by Standards”, *Electronic Imaging & The Visual Arts (EVA)*

ACKNOWLEDGEMENTS

This research was financially supported by the Fund for Scientific Research – Flanders (Belgium).

BIOGRAPHY

In 2001, Jeroen Bekaert received his engineer's Masters degree in Architecture and Urbanism at the Ghent University. He obtained a degree summa cum laude on his thesis which was entitled “IT and research into architectural history”. After his graduation, he has been working as a Research Assistant of the Fund for Scientific Research - Flanders, Belgium (FWO) at the Department of Electronics and Information Systems (ELIS) at the Ghent University, Belgium. Currently, his research activities are focusing on the description and managament of historical archive collections. From April 2003 unwards, Jeroen Bekaert will be working as a visiting research assistant at the Digital Library Research lab of the Los Alamos National Laboratories.

MY_ELIZA, A MULTIMODAL COMMUNICATION SYSTEM

Siska Fitrianie and Dr. Drs. L.J.M. Rothkrantz
Data and Knowledge Systems Group,
Faculty of Information Technology and Systems
Delft University of Technology
Mekelweg 4,
2600 GA Delft, The Netherlands
email: s.fitrianie@kbs.twi.tudelft.nl

KEYWORDS

Multimodal communication, Emotion recognition

ABSTRACT

My_Eliza is a computer model for a multimodal communication system, a combination of natural language processing and nonverbal communication. The development of this system is based on a famous question-answering system - QA system, Eliza (Weizenbaum 1966). A human user can communicate with the developed system using typed natural language. The system will reply with text-prompts and appropriate facial-expressions. In order to communicate using a nonverbal facial display, the system should be able to process natural language and emotional reasoning. A first prototype as a proof of concept has been developed that consists of a dialog box, an emotional recognizer based on stimulus response, and a facial display generator.

To implement the dialog box, the work of Wallace, A.L.I.C.E (Wallace 1995), has been used as a starting point. My_Eliza system has a rule engine that determines current system's affective state as reaction to the user's string input and conversation content.

INTRODUCTION

Recent advances in sensing, tracking, analyzing and animating human non-verbal communicative signals, have produced a surge of interest in affective computing by researchers of advanced HCI. This intriguing new field focuses on threefold: computational modeling of human perception of affective states, synthesis/animation of affective expressions, and design of affect-sensitive HCI. Indeed the first step towards an intelligent HCI having the abilities to sense and respond appropriately to the user's affective feedback is to detect and interpret affective states shown by the user in an automatic way. his paper focuses on a text based interface but is a first step to a multimodal interacting talking face.

The way QA systems represent and retrieve information is transparent by their memory structure. The memory structure functions as the systems' "brain" and is the foundation of the ability level of the system to "speak" in human natural language. The QA system retrieves the information from its memory and uses syntactic and semantic analysis to output a string as an answer to the user's string input.

Eliza worked by simple pattern-matching operation and substitution of keywords. First, the system identifies the "most important" keyword occurring in the user's input string. Next, it chooses an appropriate transformation rule and its mechanism. There are two transformation rules that are associated with certain keywords. Decomposition rule serves to decompose a data string according to certain criteria (pattern). Reassemble rule serves to reassemble a decomposed string according to certain assembly specifications (reply sentence). If Eliza finds a keyword, she will pattern-match the string input against each decomposition rule for that keyword. If it matches, she randomly selects one of the reassemble rules (for that decomposition rule).

```
keyword: your
 decomposition rule: * your *
  reassemble rule: Why are you concerned about
                   my (2) ?
  reassemble rule: What about your own (2) ?
  reassemble rule: Really, my (2) ?
 decomposition rule: …
  reassemble rule: …
. . .
Example fragment:
User : What is your name?
Eliza: What about your own name?
User : Only your name, please!
Eliza: Really, my name, please?
User : Just tell me your name!
Eliza: Why are you concerned about my name?
```

Figure 1: Example Unit of Eliza's Memory Structure

Finally, Eliza uses a selected reassemble rule to construct the reply. The keyword lists, and the list of decomposition rules and reassembly rules are constructed in a script, which controls all the behavior of Eliza. Figure 1 displays an example of one unit Eliza's memory structure (asterisk sign shows that it can contain any words or phrases).

The pattern matching operation of the original Eliza still has three major problems (Simmons 1970) (1) lack of anaphoric analysis, it cannot use previous question-answers to keep the continuity of the conversation content and to store information about the user, (2) lack of ability to restrict the conversation on its topic and (3) lack of ability to get the meaning beyond the sentence.

Another limitation of Eliza system is that users can only communicate with Eliza by exchanging text prompts. However beyond speech, human people can express their feelings or thoughts through the use of their body, facial expressions, and tone of voice. As indicated by Mehrebian (King 1997), it is proved that about 55 percent of the emotional meaning of a message is communicated through the nonverbal channel, which includes gestures, postures, and facial signals. Nonverbal communication is behavior other than spoken or written communication that creates or represents meaning. Human face-to-face conversation has provided an ideal model for designing a multimodal human-computer interaction (HCI) (Takeuchi and Nagao 1993). Characteristics of face-to-face conversation are the multiplicity and multi modality of the communication channel. Multimodal user interfaces are interfaces with multiple channels that act on multiple modalities. Conversation is supported by multiple coordinated activities of various cognitive levels. As a result communication becomes highly flexible and robust, so that failure of one channel is recovered by another channel and a message in one channel can be explained by another channel. This is the basic idea how a multimodal HCI should be developed to facilitate realistic human-machine interaction.

Nowadays, as computer acts as electronic secretaries or communication mediators, they become common entities in human society (Elliot 1993; Nakatsu 1999). The capability of communicating with humans using both verbal and nonverbal communication channels would be essential. This will surely make interaction between computers and humans more intimate and human-like (Lee 1999; Predinger and Ishuka 2001). Face to face communication is inherently natural and social for human-human interaction and substantial evidence suggest this may also be true for human-computer interaction. Using human-like faces as means to communicate have been found to provide natural and compelling computer interfaces.

Eliza has shocked AI community because it gave the impression of deep semantic linguistic processing but it was in fact based on shallow language processing. Many people become emotionally involved with the QA system. Automating the recognition of users' emotion would therefore be highly beneficial in order to give a proper user reply, both in the verbal channel and in the nonverbal channel. In recent advances of QA systems, facial expression recognition and adapting life-like agents open up the possibility of automatic emotion recognition from user interaction in conversation between human and computer. Emotions are an essential part of human lives; they influence how human think and behave and how human communicate with others, and facial displays are human primary means of communicating emotion (Schiano et.al. 2000). However, there are only a few researches involving research on human emotion recognition, because it is difficult to collect a large amount of utterances that contain emotion (Nakatsu et.al. 1999). Only a few of them work in recognizing emotion from text and none of them work in facilitating emotion recognition in a QA system. Moreover, the interpretation of emotion eliciting factors is strongly situation and culture dependent (Wierzbicka 1999)

As a first step in achieving automatic analysis of human behavior and face-to-face communication, automated emotion recognition in human conversation between the users and a QA system has been investigated. This paper discusses the results of the research, which ensued in the development of the my_Eliza – an advance version of the original Eliza. My_Eliza was aimed at the design and establishment of a QA system of a semi automated emotion recognition from human user written conversation. A user or client can communicate with the system using typed natural language. The system will reply by text-prompts and appropriate facial-expressions.

The problem of automating emotion recognition and generating appropriate nonverbal facial displays on a QA system as defined in this research comprises into three sub-discussions: (1) automatic generation of system's reply text prompts with ability of anaphoric analysis and ability to respond the conversation based on its topic (2) semi automatic emotion recognition of user's affective state and its intensity, and (3) automatic facial display selection from a facial expressions database based on emotion analysis.

In conversation, my_Eliza displays two kinds of emotional expressions: first, related to stimulus response when she hears the utterance and second, related to cognitive processing when she realizes the situation and the conversation content to convey her reply sentences.

NATURAL LANGUAGE PROCESSING

Nowadays, a QA system is also called chatterbot - a short for "chatter" and "bot" (Laven 1996), spreading in Internet. Bot is short for "robot". A.L.I.C.E is an example of this class of programs. Tackling the three limitations of Eliza above, Wallace proposed to expand memory structure using an extended-XML (Extensible Markup Language) script specification for programming the memory structure for a QA system, called AIML (Artificial Intelligence Markup

Language) (Wallace 1995). The most important AIML units are (Bush 2000):

- <aiml>, the tag that begins and ends an AIML document.
- <category>, the tag that marks a "unit of knowledge" in the system's memory structure.
- <pattern>, the tag that contains a simple input pattern rule that matches what a user may type.
- <topic>, the tag that contains current conversation topic pattern rule.
- <that>, the tag that refers to system's previous reply as a history pattern rule.
- <template>, the tag that contains the response to a user input.

In Eliza, <pattern> tag part is namely the decomposition rule and <template part> tag is the reassemble rule. Figure 2 displays an example of A.L.I.C.E's memory units in a topic about name.

```
<category> <that>*</that>
<pattern>WHAT IS YOUR NAME</pattern>
<template>My <set_topic>name</set> is
<bot name="name">.</template>
</category>

<topic name="NAME">
<category><that>MY NAME IS *</that>
<pattern>CAN I CALL YOU *</pattern>
<template><random>
<li> <star/>? Huh! Like I've told you my name
is <bot name="name">.</li>
 <li> You can call me whatever you like</li>
</random> </template>
</category>

<category><that>MY NAME IS *</that>
<pattern>I HATE *</pattern>
<template><random>
<li> I don't care, you can only call me,
<bot name="name">.</li>
  <li> Why? A sad memory perhaps?</li>
</random> </template>
</category>
</topic>
…
Example fragment:
User : What is your name?
Alice: My name is Alice
User : Can I call you Madonna?
Alice: Madonna? Huh! Like I've told you my name
is Alice.
User : I hate that name.
Alice: Why? A sad memory perhaps?
```

Figure 2: An Example of A.L.I.C.E's Memory Units

<set> and <get> tags are used to store information during conversation. A.L.I.C.E has much more possibilities of reply sentences based on their topic and history. Using AIML gives the possibility to create new content by a dialog analysis.

The matching operation is word-by-word, not category-by-category. The algorithm searches the best match pattern by ensuring that the most specific pattern matches first basically it finds the longest pattern matching an input. If there are two identical patterns but the later contains the same <that> tag, then it will take precedence over the other categories, if inside <that> tag matches the previous response. Any categories that are contained within a <topic> tag will be searched first if the current topic matches it. If neither of above is true, there is a default category with <pattern>*</pattern>. We used this AIML schema to build my_Eliza's memory structure.

NONVERBAL COMMUNICATION

This section deals with emotion reasoning and facial display generator. The main goal here is to explore the issues of design and implementation of a nonverbal QA system that could recognize the user's emotion and show a proper facial display accordingly. In general, three steps can be distinguished in tackling this issue: (1) define which and how many emotions can be recognized by the system, (2) define mechanisms for extracting emotion-eliciting factors in the observed text prompt, which devise the categorization mechanism and the emotion interpretation mechanism, and (3) define some set of categories of emotions that we want to use for facial displays classification and facial displays generation mechanism.

Currently, the interpretation of the emotion-eliciting factor is still semi automatic since we assume to use the memory structure approach of Weizenbaum's or Wallace's pattern matching operation. The memory structure of this approach does not store the semantic meaning of the text. It needs human intervention to interpret the affective semantic meaning.

EMOTION CLASSIFICATION

How many and what kind of emotional expressions are to be treated in a QA system are interesting but difficult issues. In this research we investigate three classification methods:

1. Reddy's (Reddy 2001) basic emotions: every emotion is either pleasant or unpleasant and every emotion has a varying intensity regarded as either shaping one's goals or reflecting one's goals.
2. Ekman and Friesen's (Ekman and Friesen 1975) seven universal emotions: neutrality, happiness, sadness, anger, fear, disgust, and surprise, in terms of facial expressions and mainly concentrated on primary or archetypal emotions, which are universally associated to distinct expressions.
3. Ortony, Clore and Collins theory's twenty-four emotions (Elliott 1993b) (OCC's theory 1988, see table 1). It is based on grouping human emotions by their eliciting conditions events, their consequences of their action, and their selections of computational implementation. They are resulted in three branches: (1) Attraction relates to emotions that are arising from aspects of the object, (2) Consequences of event relates to reaction of others' fortunes and (3) Attribution relates to approval of self or other. In addition, there is

a compound class that involves the emotions of gratification, remorse, gratitude and anger.

Table 1: Twenty-four OCC's Theory Emotion Types

Name and Emotion Type
Joy: pleased about an event **Distress:** displeased about an event
Happy-for: pleased about an event desirable for another **Gloating:** pleased about an event undesirable for another **Resentment:** displeased about an event desirable for another **Sorry-for:** displeased about an event undesirable for another
Hope: pleased about a prospective desirable event **Fear:** displeased about a prospective undesirable event
Satisfaction: pleased about a confirmed desirable event **Relief:** pleased about a disconfirmed undesirable event **Fears-confirmed;** displeased about a confirmed undesirable event **Disappointment:** displeased about a disconfirmed desirable event
Pride: approving of one's own act **Admiration:** approving of another's act **Shame:** disapproving of one's own act **Reproach:** disapproving of another's act
Liking: finding an object appealing **Disliking:** finding an object unappealing
Gratitude: admiration + joy **Anger:** reproach + distress **Gratification:** pride + joy **Remorse:** shame + distress
Love: admiration + liking **Hate:** reproach + disliking

Since classifications of some emotion eliciting factors are in a gray area, in this research, we add one emotion type: uncertainty.

Emotion Eliciting Factor Extraction

Most of developed systems that are able to devise emotion-eliciting factor information still need manual human intervention. Following three experiments dealing with representing and extracting emotions' information on the system's memory structure and how we map them in my_Eliza:

Emotive Lexicon Dictionary Look-up Parser.

This approach uses a list of lexicons associated to different type of emotions. Those lexicons, which are composed by words or phrases, are selected from the way human people expresses their feelings with its intensity. The system uses a shallow word matching parser to extract affective state from the context. Elliott (Elliot 1992; Elliot 1993) used this approach for his model of a multi-agent world where each agent is able to reason about emotion episodes that take place in one another's lives. He used an extended base lexicon of spoken phrases that includes 198 emotion words associated with twenty-four OCC's theory emotion types. Those words describe relationship, mood and emotional intensity. Each emotion type has a set of eliciting conditions. When the eliciting conditions are met, and various thresholds have been crossed, corresponding emotions result. The system applies minimal the detection of user's emotional inflection. Using this approach, it allows the user to teach the computer keywords in a new vocabulary relatively quickly and the system remains understandable no matter in which context the user is.

My_Eliza uses this approach to extract emotion-eliciting factor information in the text prompt in the conversation both of the user's string input and the system's reply sentence. Since the first prototype is dedicated as a "proof of concept", only six universal emotion types (Ekman's) will be used in emotive lexicons classification instead of twenty-four OCC's theory emotion types. We define six dictionaries containing lexicons in the following form: [<lexicon>: <intensity value>] with <intensity value> is an integer value [1..3].

We also define six affective counters C for each emotion type. The parser parses the sentence word-per-word against the dictionary. If it finds the same emotive lexicon in the dictionary, it will calculate the counters using following equations:

$$\forall \text{ Lexicon } l_i \in d_i \mid C_{i(t)} = C_{i(t-1)} + \boldsymbol{I_i} \;.\; \boldsymbol{s} \;; \qquad (1)$$
$$\forall\, j \neq i \mid C_{j(t)} = C_{j(t-1)} - \text{distance}[j,\, i]$$

where i = active emotion type, I = intensity level, s = summation factor and j = {happiness, sadness, anger, fear, disgust, surprise}

For the first prototype we use Hendrix and Ruttkay's distance values between expression emotions (Hendrix and Ruttaky 1998) shown in table 2 for the distance[j,i]. The result of this calculation is the candidate of affective state both for the user and the system, which is taken from the emotion type with the highest level of all counters.

Table 2: Distance Value Between Emotions

	Happiness	**Surprise**	**Anger**	**Disgust**	**Sadness**
Happiness	**0**	**3.195**	**2.637**	**1.926**	**2.554**
Surprise		**0**	**3.436**	**2.298**	**2.084**
Anger			**0**	**1.506**	**1.645**
Disgust				**0**	**1.040**
Sadness					**0**

Emotive Labeled Memory Structure Extraction

This approach labels each unit of memory structure with one or more of the emotions types. Most of the examples for systems using this approach are automatic story telling systems and automatic digitizer for cartoon movies. Each dialog sentence of each actor is labeled with an emotion type and decomposed in its phonological representation. Therefore, the system can show appropriate intonation and nonverbal display when it reads the dialog. (Pelachaud et.al. 1994) used it in their research by assuming the input as a file containing an utterance already decomposed and written in its phonological representation with its prosody in its bracketed elements. At each input, it specifies the desired affectual parameters and their intensity. Using this approach means to sidestep the issue of emotion recognition. The modeling of affect is not meaning based and it needs human manual work to label each memory unit.

We label my_Eliza's memory structure by adding two additional tags in AIML schema: <affect> tag that labels the user's affective situation and <concern> tag that labels the system's reaction situation. Inside those two new tags, based on Reddy's basic emotion we define four-possibility emotive situation type: (1) positive "+", (2) negative"-", (3) joking "#", and (4) normal/any "*". Figure 3 (below) displays an example of my_Eliza's memory structure units.

```
<category> <affect name=”*”>
<pattern>WHAT IS YOUR NAME</pattern>
<that>*</that>
<template><think><setconcern>+</setconcern>
  <setaffect>+</setaffect></think>My
  <set_topic>name</set> is <bot name=”name”>.
</template>
</affect>
</category>

<topic name="NAME">
<category> <affect name=”*”>
<that>MY NAME IS *</that>
<pattern>YOUR *</pattern>
<template><random>
 <li><think><setconcern>#</setconcern>
  </think>Your <star/> also!</li>
 <li><think><setconcern>+</setconcern>
  </think>Mind that, what is your name?</li>
</random> </template>
</affect></category>

<category> <affect name=”*”>
<that>MY NAME IS *</that>
<pattern>I HATE YOUR *</pattern>
<template><random>
 <li>><think><setconcern>-</setconcern>
  </think>Why? A bad memory perhaps?</li>
 <li>><think><setconcern>-</setconcern>
  </think>Why?</li>
</random>
<think><setaffect>-</setaffect></think>
</template>
</affect></category>

<category> <affect name=”-”>
<that>MY NAME IS *</that>
<pattern>YOUR *</pattern>
<template>
 <li><think><setconcern>+</setconcern>
  </think>Mind that, what is yours?</li>
 <li><think><setconcern>-</setconcern>
  </think>Don’t be a witty!</li>
</random> </template>
</affect></category>
</topic>
```

Figure 3: Example of My_Eliza's Memory Structure Units

Using the memory structure units in figure 3, we expect my_Eliza is able to converse with human user in positive situation (table 3) and in negative situation (table 4).

Table 3: Example of Fragment in Positive Situation

User:	What is your name?	🙂 My_Eliza
🙂 My_Eliza:	Your name is lovely.	
User:	My name is Eliza.	😀 My_Eliza
😀 My_Eliza:	Your name is lovely also!	

Table 4: Example of Fragment in Negative Situation

User:	What is your name?	🙂 My_Eliza
🙂 My_Eliza:	My name is Eliza.	
User:	I hate your name.	😮 My_Eliza
😖 My_Eliza:	Why? Bad memory perhaps?	
User:	Your name reminds me to my enemy.	🙁 My_Eliza
🙂 My_Eliza:	Mind that, what is yours?	

Using extra two tags, my_Eliza has the possibility to have a bigger memory structure than A.L.I.C.E. In matching-operation algorithm, the memory structure with the same situation type will be checked first. If none of them is matches, the system picks a category with a normal situation type (*).

Goal Based Emotion Reasoning.

This approach sets some goals, principles, preferences and moods in the system. Extraction emotion-eliciting factors leading to emotions falls into four major categories: those rooted in the effect of an event on the goal of the system, those rooted in the standard and principle invoked by an act of the user, those rooted in tastes and preferences with respect to an object (including the user treated as an object), and lastly a selected combination of the first three categories. Another way to view these categories is that they are rooted in the system's assessment of the desirability or undesirability of some event, the praiseworthiness or blameworthiness of some act, the attractiveness or unattractiveness of some object, or selected combinations of these assessment. (Elliot 1993b) also used this approach for his model to simulate social interactions between agents in incorporated models of individual affect and personality. Each agent interprets situations that are characterized in terms of the way they may or may not meet the eliciting conditions of emotions. Agents use a case based heuristic classification system to reason about the emotions of other agents' personalities that will help them to predict and explain future emotion episodes involving observed agents. Embodied in the simulation system, Elliott used a set of rules for the mapping from four categories emotion-eliciting factors above into twenty-four OCCs theory emotion types.

Mapping to my_Eliza, we define the system's goals, affective status and preferences (GSP) while she converses with the user. Several goals of my_Eliza are:

- *Answering questions* – if the user asks something, my_Eliza's goal is to answer it.
- *Persuasive agreement* – if the user persuades to do something or invites my_Eliza to do something, my_Eliza's goal is to show whether she agrees or not.
- *Topical focus* – to keep on conversing on the same topic and beware if it is changing.
- *Explanation statements* – to reply the user's statements that require specification and explanation.
- *Reflecting feeling* - to keep consistent with the user's current affective state.
- *Alignment* - to keep consistent with the system's current reaction affective state (system's status) and system's preferences.

The system recognizes the dialog using a dialog scheme adopted from (Carletta et.al. 1995). By distinguishing the by distinguishing a dialog state as a certain dialog act like a question, statement, acknowledgement, or pause, the system has to know which goal to pursue. Whether a certain goal is appealing or not appealing may influence the system's affective state. We also define my_Eliza's preference as the personal data about the system and can be used during conversation, for example: her name, birthday, the things she likes or hates, and so on. To be fair, using <set> tag the system also stores the user's personal data during conversation, for example the user's name, birthday, favorite stuff, personal data about family and so on. These data about system's GSP and user's personal data can be used for pragmatic analysis when the system constructs the reply sentence and defines its current affective status.

Emotion Recognition

For the activation of an emotion, (Elliot 1993a) proposed the use of threshold values by counting all associated elicitation factors, the excitatory (positive) and inhibitory (negative), from other emotions. They used an activation level range [0, max] where max is an integer value determined empirically. All emotions are always active, but their intensity must exceed a threshold level before they are expressed externally. The activation process is controlled by a knowledge-based system that synthesizes and generates cognitive-related emotions in the system.

We define six affective thermometers classified by six Ekman's universal emotion types. These thermometers observe the affective state of the system as reaction to the user's string input and the dialog content – the system's reaction affective state. If an emotion is active, the system calculates all of thermometers T, with the following equations:

$$T_{i(t)} = T_{i(t-1)} + \boldsymbol{I_i} \ . \ \boldsymbol{s} \qquad (2)$$
$$\forall \ j \neq i \ | \ T_{j(t)} = T_{j(t-1)} - \mathrm{distance}[j,\ i]$$

Where j = {happiness, sadness, anger, fear, disgust, surprise}, i is an active emotion type, s is a summation factor and I is the intensity. We again use Hendrix and Ruttkay's distance values to calculate those equations. The system takes the highest degree of all thermometers as the most dominant emotion.

To determine the system's affective state we formulate two knowledge based systems: (1) determines the system's reaction affective state as stimulus response to the user's input string and (2) determines the system's reaction affective state as the result of cognitive process of the conversation content to convey its reply sentence. We have defined a set of rules that specify the emotion recognition process of the system. We call these rule-sets preference rules, since they indicate preferences to exhibit system's "preference" reaction affective state rather than performing explicit actions, such as facial displays. Every rule in the set defines conditions of emotion eliciting factors and the affective thermometers to activate the rule and a preference that is expressed upon activation. The result from each knowledge-based system is one of twenty-four OCC's theory emotion types with addition of two emotion types: normal and uncertainty, for example:

1. Preference rule for stimulus response:
This rule will fire the preference first reaction joy if the following conditions are met:

- user is happy,
- user asks question,
- situation type of user is not negative,
- current maximum system's affective
- thermo is happy.

In this case my_Eliza will answer any questions from the user joyfully, because she enjoys the situation and she met the goal: making the user feel happy.

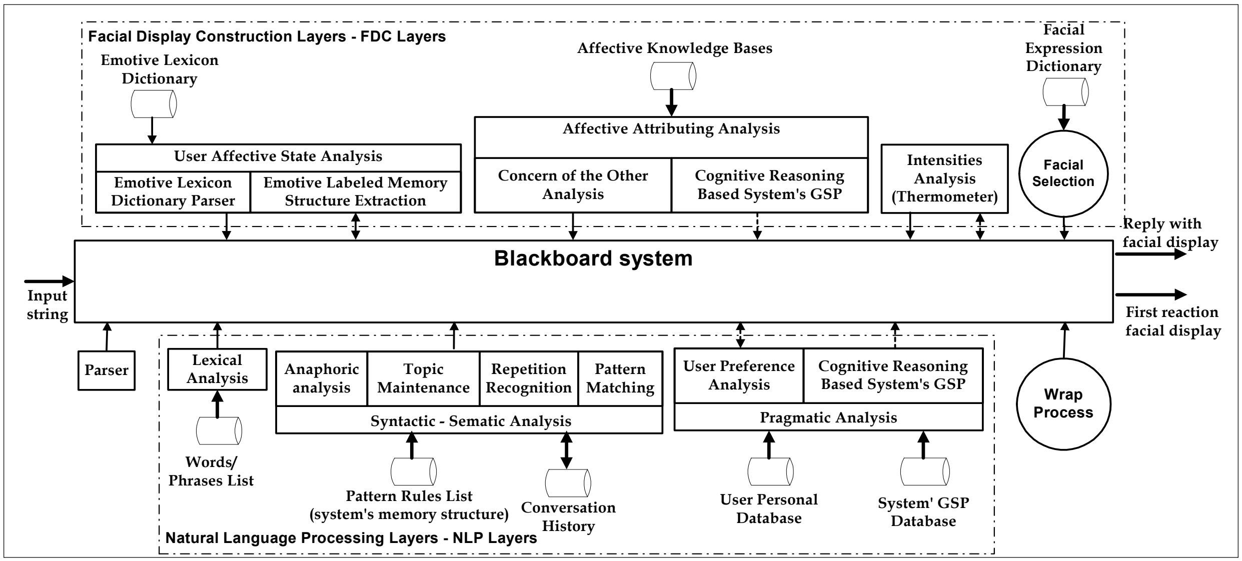

Facial Display Construction Layers - FDC Layers
Emotive Lexicon Dictionary
Affective Knowledge Bases
Facial Expression Dictionary
Affective Attributing Analysis
User Affective State Analysis
Emotive Lexicon Dictionary Parser
Emotive Labeled Memory Structure Extraction
Concern of the Other Analysis
Cognitive Reasoning Based System's GSP
Intensities Analysis (Thermometer)
Facial Selection
Reply with facial display
Blackboard system
Input string
First reaction facial display
Parser
Lexical Analysis
Anaphoric analysis
Topic Maintenance
Repetition Recognition
Pattern Matching
User Preference Analysis
Cognitive Reasoning Based System's GSP
Syntactic - Sematic Analysis
Pragmatic Analysis
Wrap Process
Words/ Phrases List
Pattern Rules List (system's memory structure)
Conversation History
User Personal Database
System' GSP Database
Natural Language Processing Layers - NLP Layers

Figure 4 My_Eliza's blackboard system

2. Preference rule for cognitive process:

This rule will fire the cognitive processed preference resentment if the following conditions are met:

- user is sad,
- system's reply is sad,
- situation type of user is joking,
- situation type of the system is negative,
- current maximum system's affective
- thermo is sad.

Here my_Eliza does not like the user makes a joke while she feels sad.

Facial Display Generator

In most of the works in facial display generation are used one to one corresponding facial display and emotions, distinguished by intensity (Elliot 1993, Prendinger and Ishiuka 2001). The other works used the correspondence between communication.

categorization and Ekman & Friesen's Facial Action Coding System (FACS) (Takeuchi and Nagao 1993) and between emotions with FACS (Pelachaud et.al. 1994). FACS is a notation to describe visible facial expression based on anatomical studies; how a feature is affected by specifying its new location and the intensity of changes. For the first prototype we use one to one corresponding facial display and emotion. We use twenty-two smiley nonverbal facial display classified by eight emotion types (neutrality, happiness, sadness, anger, surprise, fear, disgust and uncertainty) and three level of intensity (LOW, MEDIUM, HIGH) except for neutrality. Since the system's reaction affective state may be one of twenty-four OCC's theory emotion types we cluster every those emotion types into six Ekman's universal emotion types. We cluster normal into neutrality

DESIGN

The architecture of my_Eliza is illustrated in figure 4, which takes the idea of message passing on a blackboard system. The message flow and message process are always on the blackboard. If a new message comes, it will be analyzed, synthesized, and the result will always be put back on the blackboard. In my_Eliza, the message is the user's string input and the results are the reply sentences and facial displays. My_Eliza works by the following steps:

1. Generating a stimulus-response nonverbal signal

- User types a string input and puts it on the blackboard system.
- The Parser parses the input into words and puts it on the list on the blackboard system.
- The Lexical Analysis layer normalizes the string input by eliminating incorrect or incomplete words or phrases and checking relations between words or phrases. This layer puts the result on the blackboard system.
- The Affective State Analysis layer activates its two sub layers: Emotive Lexicon Dictionary Parser and Emotive Labeled Memory Structure Extraction. The Emotive Lexicon Dictionary Parser layer identifies the emotive lexicons from the user's input and the reply sentence (after the system has constructed the reply sentence). The Emotive Labeled Memory Structure Extraction layer extracts the label from the system's memory unit. Those results are put on the blackboard system.
- The Syntactic-Semantic Analysis layer performs a pattern matching operation based on the user's string input pattern to add the user's affective information on

the blackboard system. In this step, the system starts to work in parallel with the process to construct system's reply sentence. As a result, Syntactic-Semantic Analysis layer performs pattern-matching operation to generate system's reply sentence and puts its candidate on the blackboard system.

- The Affective Attributing Analysis layer activates its sub layer, the Concern of the other Analysis layer, to perform emotion-based reasoning to deliver current system's reaction affective state and put it on the blackboard system. This analysis based on stimulus response and directly related to the user's input string.
- The Intensity Analysis layer processes the message and calculates the current system's affective 'thermometers' and puts the calculation result on the blackboard system.
- The Facial Selection selects my_Eliza's facial display and puts the selection on the blackboard.
- The Wrap Process delivers and displays my_Eliza's stimulus-response facial display or the first facial display.

2. Constructing the candidate of reply sentence and generating a cognitive-processed facial display:

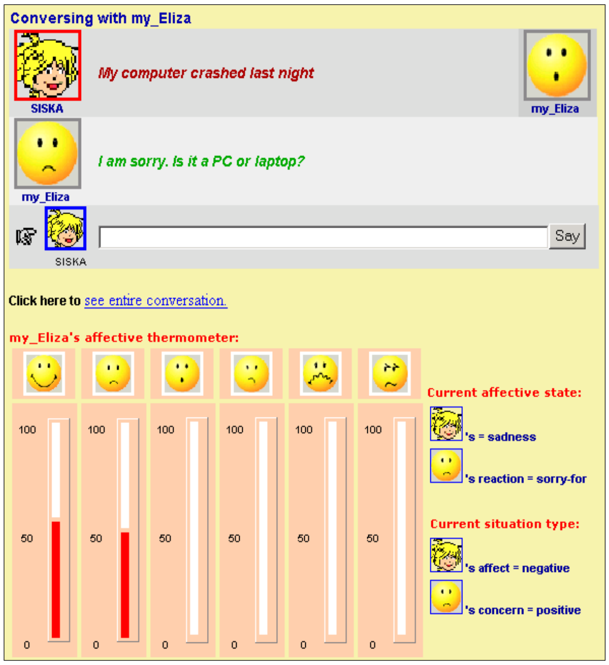
- The Pragmatic Analysis layer processes the candidate reply sentence and puts the result on the blackboard system. This layer reviews the candidate of reply sentence whether it violates user's preference and/or system's goals, status and preferences – system's GSP. If it does, the result will be sent back to the Syntactic-Semantic Analysis Layer to get new a reply sentence.
- The Affective Attributing Analysis layer activates its sub layer, the Cognitive Reasoning layer, to perform emotion-based reasoning to deliver current system's reaction affective state and put it on the blackboard system. This analysis is based on cognitive processing of system's GSP, system's reply sentence and user's affective state.
- Again, the Intensity Analysis layer processes the message and calculates the current system's affective 'thermometer' level. This layer also puts the result of calculation on the blackboard system.
- The Facial Selection selects my_Eliza's facial display and puts the selection on the blackboard.
- The Wrap Process delivers and displays my_Eliza's reply sentence and cognitive-processing result facial display.

IMPLEMENTATION

There are three incremental implementation layers: (1) create a dialog box that can engage in human conversation based on typed natural language and recognize the user's affective state and the system's reaction affective state, (2) build a stimulus-response of facial displays based on spontaneously spinal brain reasoning on user's string input, (3) build a cognitive processor of facial displays based on anaphora analysis, pragmatic analysis, dialog content and system's goals, status, and preferences.

Currently, we are in the second implementation layer and the result is called my_Eliza prototype-1. We use Program D A.L.I.C.E (Wallace 1995) as a starting point to build my_Eliza's dialog box. Program D A.L.I.C.E is written on Java Development Kit version 1.3 and XML, therefore we use a compiler and classes contained in the same languages for my_Eliza prototype-1. My_Eliza's dialog box contains many packages most of them derive from Program D A.L.I.C.E. This program has provided a robust client server and multi user communication. Therefore, My_Eliza is also controlled by a collection of autonomous client-server communication via TCP/IP. The user can communicate with my_Eliza's server through the HTTP server. My_Eliza's server provides the blackboard system. My_Eliza uses the Java expert system shell (Jess) for affective attributing knowledge based system shell (Friedman-Hill 2000).

Currently, my_Eliza prototype-1's emotive lexicon dictionary contains: 48 lexicons for happiness, 170 lexicons for sadness, 34 lexicons for surprise, 33 lexicons for fear, 93 lexicons for disgust, and 69 lexicons for anger. This prototype has 1953 categories in its list of pattern rules. Its affective knowledge bases contain 77 preference rules of stimulus response knowledge base and 151 preference rules of stimulus response knowledge base. We can add these databases and knowledge bases easily even while the server is still running. so we can build the system's knowledge base incrementally.. Figure 5 displays the main page of prototype-1 when she felt sorry-for the user's misfortune.



Conversing with my_Eliza

My computer crashed last night

I am sorry. Is it a PC or laptop?

Click here to see entire conversation.

my_Eliza's affective thermometer:

Current affective state:

's = sadness

's reaction = sorry-for

Current situation type:

's affect = negative

's concern = positive

Figure 5: My_Eliza Prototype-1 Main Page

CONCLUSION

Surveying the literature about the emotion recognition process leads to the conclusion that none of the research is particularly in the work of recognizing emotion in a QA system. However, from the work of researchers in the fields of multi agents system, emotion recognition from human speech intonation, automated character animation system and communicative facial display system, could give inspiration which allows us to fit in our approach with the fields. However from most of the work still need human manual intervention. A pragmatic advantage of using AIML form and preference rules to exhibit system's behavior is that the systems' memory structure and its behavior can be extended easily. One of future works in this research is to extend the rule-sets in system's affective knowledge bases. Current rule-sets that we have implemented in the system do not know the correlation emotion of current dialog and the emotion of entire conversation content. The rules-sets could also be made temporal. We can make specific rules for the opening of the conversation, during discussion or the end of the discussion. By adding more emotion eliciting-factors in the rule-sets, extra rules are needed for new eliciting factors.

Additionally the server interface should have extended functionality to show the thinking process of the system. Realistic virtual environments not only include believable appearance and simulation of the virtual world but also imply the natural representation of participants. That can be fulfilled by visualization of human character embodiment with animation. Moreover, using more possible facial displays, the system is able to convey many different kinds of emotion as different social situation arise. It needs to explore several ways so that real-time, animated, and virtual human characters can be given more intelligence and communication skills, therefore they can act, react, make decisions, and take initiatives. In a similar fashion, the system should be able to communicate with a broad range of conversation topics and it should be able to visually support these conversations with an equally broad range of emotion and expressions behaviors. The system also should have the ability to learn from conversation history. These additions to the system will be valuable assets to add new memory structure units and to add rule-sets of the system to generate its affective knowledge bases autonomous.

REFERENCES

Bazzan & Bordini, "A Framework for the Simulation of Agents with Emotions, Report on Experiments with the Iterated Prisoner's Dilemma", AGENT'01, *Communication of ACM*, p292-p299, Canada, 2001.

Carletta, J., et. al., "The Coding of Dialogue Structure in a Corpus, in Corpus-based Approaches to Dialogue Modeling", *9th Twente Workshop on Language Technology*, p25-p34, University of Twenty, 1995.

Cassell, J. et.al. "Animated Conversation: Rule Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversation Agents", in *SIGGRAPH'94*, 1994.

Ekman & Friesen, "Unmasking the Face", Prentice Hall, New Jersey, USA, 1975.

Elliott & Siegle, "Variables Influencing the Intensity of Simulated Affective States", In *AAAI Technical Report for Spring Symposium on Reasoning about Mental States: Formal Theories and Applications*, 58-67, American Association for Artificial Intelligence, 1993.

Elliott, "Using the Affective Reasoner to Support Social Simulations". In *Proceeding of the 13th International Joint Conference on Artificial Intelligence*, 194-200, Chambery, 1993.

Elliott, "Draft: Components of two-way emotion communication between humans and computers using a broad, rudimentary, model of affect and personality", http://condor.depaul.edu/~elliott/ar.html, 1995.

Friedman-Hill, Ernest, "Jess: Rule Engine of the Java Platform", Sandia National Laboratory, http://herzberg.ca.sandia.gov/jess/, USA, 2000.

Hendix & Ruttkay, "Exploring the Space of Emotional Faces of Subjects without Acting Experience", *ACM Computing Classification System*: H.5.2, I.5.3, J.4, ftp://ftp.cwi.nl/pub/CWIreports/INS/INS-R0013.ps.Z, 1998.

King, Donnell, A., "Nonverbal Communication", http://www2.pstcc.cc.tn.us/~dking, 1997.

Lee, Kwanyong, "Integration of Various Emotion Eliciting Factors for Life-Like Agents", ACM Multimedia'99, Part 2, p155-p158, 1999.

Nakatsu, Ryohei, et al. "Emotion Recognition and Its Application to Computer Agent with Spontaneous Interactive Capabilities", *Creativity and Cognition 99, Communication of ACM*, p135-p143, UK, 1999.

Pelachaud et.al., "Generating Facial Expression for Speech", Dept. of Computer and Information Science, University of Pennsylvania, 1994.

Prendinger & Ishizuka, "Social Role Awareness in Animated Agents", *AGENTS'01 ACM*, p270-p276, 2001.

Reddy, W.M., "The Navigation of Feeling, A Framework for the History of Emotion", Cambridge University Press, Edinburgh, 2001.

Schiano et.al., "Face to Interface: Facial Affect in (Hu)Man and Machine", *CHI Letters, Communications of the ACM*, Vol. II No. 1, p193-p200, 2000.

Simmons, R.F., "Computational Linguistic, Natural Language Question Answering System: 1969", *Communications of the ACM*, Vol. XIII No. 1, p15-p29, January 1970.

Takeuchi & Nagao, "Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation", *Agent Interfaces*, Section VII, p572-p579, 1993.

Wallace, Richard, "Alicebot", http://www.Alicebot.org, 1995.

Weizenbaum, J., "ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine*", Communication of the ACM* 9(1): p36-p45, 1966.

Wierzbicka, Anna, "Emotional Universal, Language Design 2", p23-p99, Australia, 1999.

Emotion Extraction Engine: Expressive Image generator

Xu Zhe, David John and Anthony C. Boucouvalas

Multimedia Communications Research Group,
School of Design, Engineering and Computing,
Bournemouth University,
Fern Barrow, Poole,
Dorset, BH12 5BB, UK.

{zxu, djohn, tboucouv}@bournemouth.ac.uk

Abstract:
In this paper, we present the latest development of an emotion extraction engine used for real-time internet text communication. Real time expressive communication is important as it provides aspects of the visual clues that are present in face-to-face interaction not available in ordinary text-based communications. In former papers, we demonstrated a text-to-emotion engine that can analyse the emotional content present in a real-time chat environment and can deliver the emotional parameters necessary to invoke an appropriate expressive image. In this paper, we present a quick and user-friendly method to generate the necessary expressive images. Images can be generated for six emotion categories from one original neutral image. In each category, three different emotion intensities can be achieved. Users only need to provide a single default image, define six control points and two control shapes to generate all the images. This paper also presents the preliminary findings from a series of experiments that have been carried out to test the efficiency of our expressive image generator.

Keywords: Emotion, Expression, Warping, Morphing.

1 Introduction

Emotion research can be traced back to 500BCE, Heraclitus (the Greek philosopher of the late 6th century BCE) who concluded that the emotional state is characterised by a mixture of expressive body parameters [1].

Communication of emotions through facial expression is an active research area. Darwin and other recent contemporary researchers such as Buck [2] and K. Dautenhahn [3] have shown great attention to this area.

With the rapid development of Computing and Internet, research on of emotions is associated with research on computer networks, for example the "Affective Communication" [16], "Computational Emotions" [17] projects, and our emotion extraction engine.

The emotion extraction engine, can analyse user input in the form of text sentences, has been developed and is presented in detail in [4]. Sentences are analysed in real-time for emotive content and the emotion is represented by an appropriate facial expression displayed automatically. In paper [5], a series of experiments that were carried out to test the performance and the effectiveness of the engine are described.

The experiment results in [5] show that most users prefer an online chatting interface that includes text with expressive images than text alone. The challenge is how to generate expressive images for all users in a fast and friendly way without individuals having to resort in taking many expressive photographs.

To solve this challenge, we report here the development of the expressive image generator discussed in detail in this paper.

Finally, we have carried out tests for assessing the performance of the expressive image generator**.** The first test assesses the correct recognition of the expressive images without any text information; the second test assesses the effective recognition of the expressive images including text information.

This paper is organised as follows. In section 2 the emotion extraction engine is reviewed. In section 3 the background knowledge of image generation is discussed. In section 4 the motion image generator is described in detail. In section 5 the test strategy for the expressive image generator is given. Section 6 illustrates possible applications using the engine. Finally in section 6 conclusions are given.

2 Emotion extraction engine

An emotion extraction engine that can analyse user input in the form of text sentences has been developed. When emotional content is detected the engine will send the parameters needed for selecting the expressive images across the network. When receiving side engine receives the parameters, the corresponding expressive images will be selected and displayed. In this way, the transmissions of images are avoided since the only data transmitted over the network is the text parameters. As a result the bandwidth requirement is extremely low.

The emotion extraction engine includes two sub-systems: the emotion analysis system and the expressive image generator. The emotion analysis system includes three parts: input text analysis, tagging system and parser. A general description is presented here. For detail information, please refer to [4] and [5].

The working flow of the emotion analysis system is shown in figure 1.

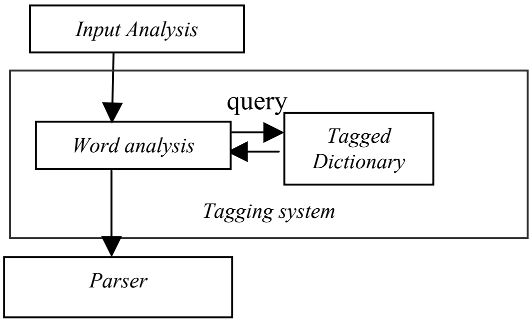


Figure 1: The working flow of the emotion analysis system

- ***Input analysis function***

Users' text input is sent from the user interface to *the input analysis function*. The engine analyses only one sentence at a time. The *input analysis function* will replace all punctuation with pre-defined characters and send the analysed sentences to the *tagging system*.

- ***The tagging system***

The w*ord analysis function* splits the sentences from *input analysis function* into words and searches *the tagged dictionary* to find the corresponding tag category. The outputs include the *word category* and the *emotional tag*.

Daily communications involve about two thousand words [6]. In order to identify the words, a special designed dictionary was set up. In this project, a database containing 16400 words was used. The database includes three fields: *word*, *category* and *emotional tag*. The *category* field contains the corresponding word category (noun, verb, adjective etc) and the *emotional tag* field describes whether that word belongs to one of the six emotion types.

Unlike tagging methods used in some existing systems, e.g., BNC [7] and Brown Corpus [8], the entire word is appropriately tagged in the dictionary in order to keep the response time to a minimum.

- ***Parser***

Receiving the output from the *tagging system*, the *parser* will try to identify the emotion content. The *parser*'s analysis is accomplished through the use of rewrite rules and tree representations [9]. According to pre-defined rules, the *parser* will search for the current emotional words, the person to whom the emotional words refer to and the intensity of the emotional words. The *parser*'s outputs are the emotional parameters and are sent across the network.

- **Expressive images**

The output from the parser will be sent out through the network to related users. When receiving the output, the corresponding expressive images are selected from a database and displayed.

3 Expressive image generation background

- **Universal Expressions**

Research in facial expression has concluded that there are six universal categories of facial expressions that are recognised across cultures [10]. The categories are *happiness, sadness, anger, fear, disgust* and *surprise*. Within each of the categories, a wide range of expression intensity and variation of detailed expressions exists. Our engine requires a set of expressive images representing the emotions.

- **Existing Expression image algorithms**

To generate expressive images a number of algorithms and theories have been developed. The most well known algorithms and software include the “Facial Action Coding System” [11], "FaceWorks"[12] and "Synthesising Realistic Facial Expressions" [13].

The major challenges for these algorithms and software are the need to reduce computation time and the development of a user-friendly interface. For example, some programs may require a database to hold different faces and compare the new face with the existing samples. Some programs may require users to create a complex model at the face that covers almost all the edges of the face. It is difficult for users to identify all the points required by the complex model accurately and it is not always possible to generate images to be used in real time communication systems.

Instead of using a complex model and a heavy burden computation algorithm, we developed a simple algorithm based on image warping and image morphing. Users only need to choose six control points and draw two control areas on the original neutral image. Then six images that correspondingly belong to the six expression categories will automatically be generated and stored for future use. In each emotion category, three different emotion intensities are achieved. In total eighteen expressive images are generated from a default image.

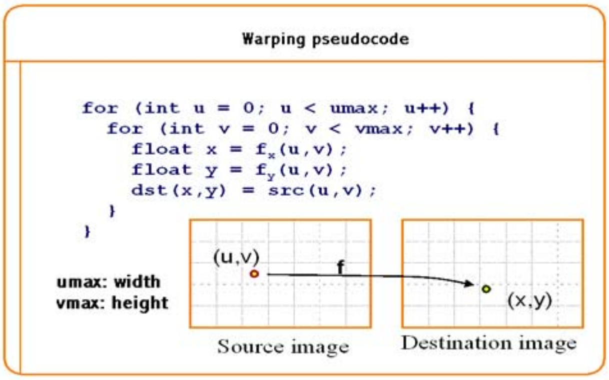
- **Image warping**

Image warping is the act of distorting a source image into a destination image according to a transformation between source space (u,v) and destination space (x,y) [14]. The transformation function f() describes the destination (x,y) for every location (u,v) in the source. The function is presented as follows.

$$\mathbf{x=f_x(u,v)}$$
$$\mathbf{y=f_y(u,v)}$$

The function f() will correspondingly stretch or compress an area defined by source space and destination space.

To apply warping to an image, we may apply the transformation function f() to each pixel. The pseudocode is shown in figure 2.



```
for (int u = 0; u < umax; u++) {
  for (int v = 0; v < vmax; v++) {
    float x = fx(u,v);
    float y = fy(u,v);
    dst(x,y) = src(u,v);
  }
}
```

Figure 2: the warping pseudocode

According to Ekman's Facial action Coding System [11] and PARKE's Computer Facial Animation [10], a limited number of muscles on the faces are responsible for the expression generations. For example only four muscles contribute to expressing sadness. It is not necessary to implement warping for facial expression generation across the whole image since most parts of the face remain unchanged.

- **Image morphing**

Image morphing is an image processing technique used for the metamorphosis from one image to another. The usual morphing technique is to generate a sequence of intermediate images. Those images put together with the original images would represent the change from first image to the last [15].

4 expressive image generator

To generate expressive images, users need to upload a neutral face image to the expressive image generator. The generator is based on the local area warping and morphing technology discussed above. The structure of the expressive image generator is shown in figure 3.

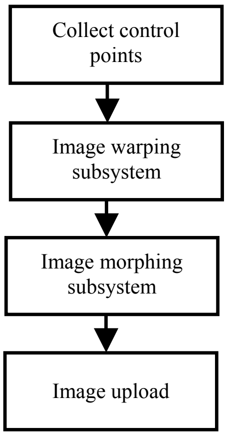


Figure 3: Expressive image generator structure.

- **Emotion categories**

In reference [10] and [11], 64 action units are defined that are responsible for the movement of the face. In this paper, we limit the investigation to the facial action units required to create expressive images and define the following rules to generate the expressive images.

Happiness: The eyebrows are relaxed, the mouth is wide with the corners pulled up toward the ears.
Sadness: The inner portions of the eyebrows are pulled up above the upper eyelid and the mouth is relaxed.
Anger: The eyebrows are pulled downward and together. The mouth is closed with the upper lip slightly compressed or squared off.
Disgust: The middle eyebrows are pulled upward and the mouth is slightly opened with the upper lip squared off.
Fear: Eyebrows are raised and pulled together. The eyebrows are bent upward. The mouth may be dropped slightly open.
Surprise: The eyebrows are raised up, the upper eyelids are opened and the mouth is dropped open [10].
For each expression category, three different emotion intensities are calculated. Based on the above rules, the movement of mouth, eyebrows and lips are quantified. From low intensity to high intensity, the movement is proportionately enlarged also.

- **Control Points**

To generate different expressions, start points and finish points are required. Users only need to select the start points. The finish points are calculated automatically based on the user's selection.

After uploading the neutral image to the system, users will be guided to select six start points and three control areas. The six start points include: left corner of the mouth (LM), right corner of the mouth (RM), outer edge of the left eyebrow (LOE), inner edge of the left eyebrow (LIE), outer edge of the right eyebrow (ROE) and inner edge of the right eyebrow (RIE).

The three control areas include the outer edge of the lips, and the inner edge of the eyelids for the left eye and right eye. These areas will be used in image morphing subsystem.

These parameters will be sent to the image warping subsystem to generate the expressive images.

- **Image Warping subsystem**

The image warping subsystem implements the model based warping. Instead of interactive manipulation, the model based warping subsystem has the advantages of accuracy and speed.

To generate the model, the facial action coding system and several different expressive images from different persons were analysed. We have named the kernel that generates the images the "expression model mask". The mask is constituted by two sets of points. The first set are the points selected by users, which is called the start points. The second set are the control points of the start points, which is called the finish points. The values of the finish points are calculated relative to the start points. By applying the masks to images of individual faces, corresponding expressions can be generated.

The finish points are calculated as the start points plus an integer value that depends on the expression being generated. The function to calculate finish points is shown below.

Finish.x = Start.x + a
Finish.y = Start.y + b

With user-selected points and the calculated finish points, The image warping subsystem can generate intermediate images. These images are sent to the image morphing subsystem.

- **Image morphing subsystem**

The weakness of the image warping subsystem is that it can not generate new pixels, for example to open the mouth and widen the eyes. The emotion *fear* and *surprise* require an opened mouth and widen eyes. The warping subsystem can not achieve this affect. To solve this problem, we implemented an image morphing subsystem.

First several different images of mouths and eyes were chosen from *surprise* and *fear* images. When the morphing subsystem receives the intermediate images from the warping subsystem, it will replace the mouth and eyes for the images belonging to the appropriate emotion categories. The mouth and eyes from pre-chosen images are pasted to the images. To remove the gaps between the pre-prepared images and the intermediate images, a gaussian blur operation is applied to the edges.

Examples of expressive images generated by our expressive image generator are shown in figure 4.

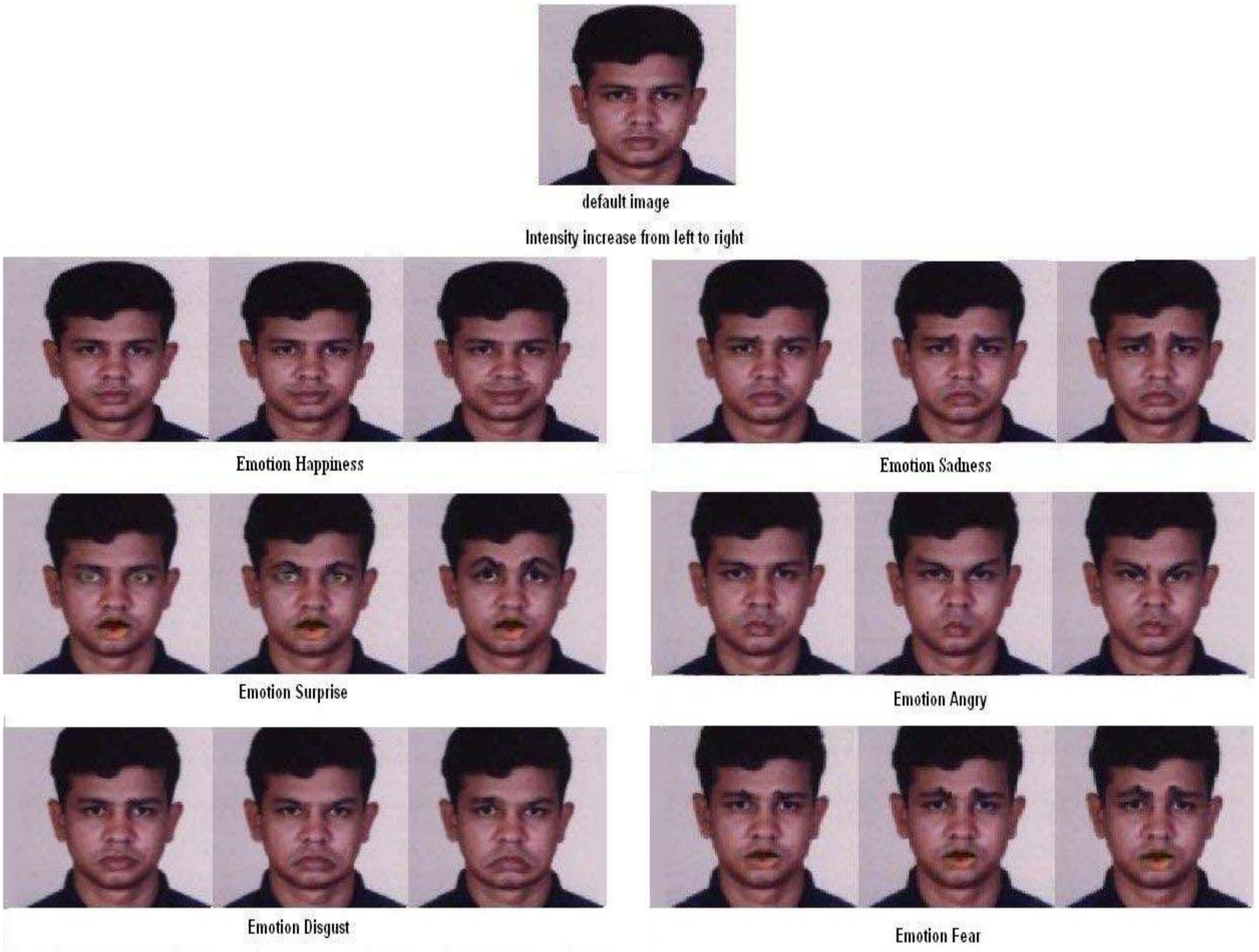


Figure 4: Generated expressive images

5 Experiment strategy

To test the correctness and effectiveness of the expressive image generator, two types of experiments were carried out. The first was an image only test, in which the test subjects view the generated images in isolation. The second type is the image plus text test, in which users view the images with contextual information.

Null hypothesis and chi-square test

The null hypothesis predicted that each of the six expression categories would be recognised only at the random level, at 17%. A recognition rate significantly above this would disprove the null hypothesis and prove its inverse, namely, that the images were recognised at a significantly greater rate than chance, and therefore that expressive image generator could be used successfully with the emotion extraction engine. The chi-square tests were carried out for both types to analyse the null hypothesis.

Image only experiment

The purpose of this experiment is to test the effectiveness of the expressive images themselves. Four people's facial images were upload to the image generator, and 18 images were generated for each person. The eighteen images belong to the six expression categories, each with three images of different expressive intensities were created.

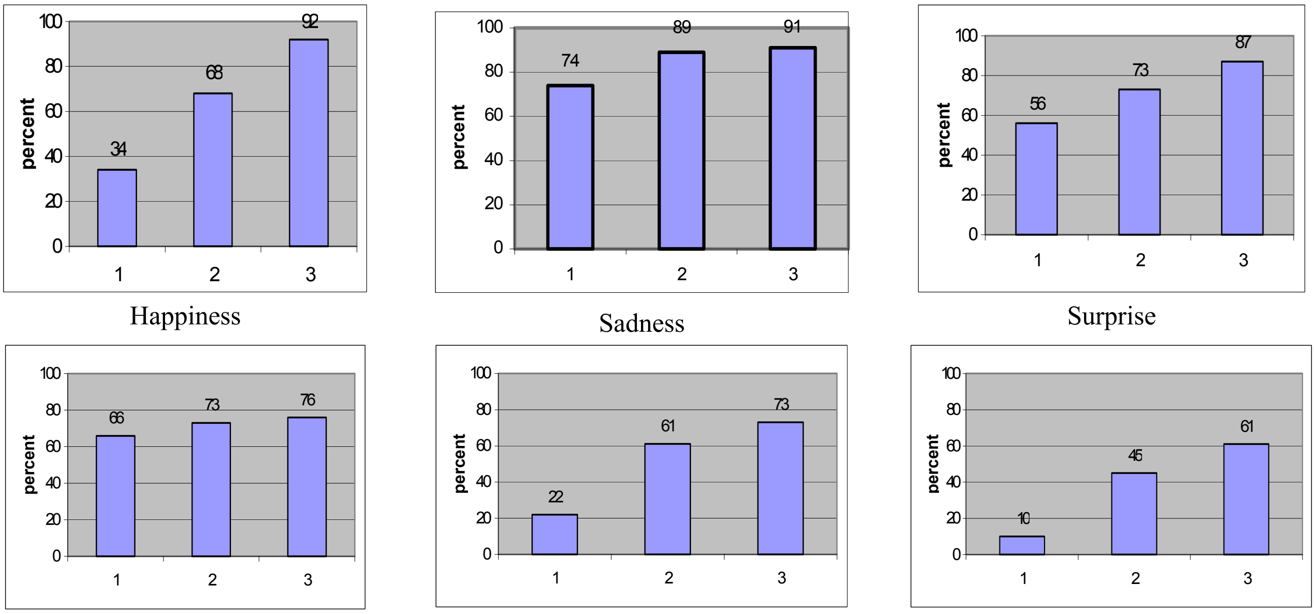
A total of 35 students and staff from Bournemouth University participated in this experiment. Each subject was told to view the images and identify which category of emotion each image expressed. The answers to be chosen include *happiness, sadness, anger, surprise, fear disgust* and *not sure*.

The results are shown in table 1. (1, 2, 3 in the x axis represents increasing corresponding intensity in that emotion category, 1 is the lowest and 3 is the highest. the y axis shows the percentage of subjects who identified the emotion correctly).

Result analysis

For each category of image, the chi-square test is carried out. The corresponding obtained values for the six emotion categories *Happiness, sadness, surprise, angry, fear and disgust* are 245.3, 456.1, 563.2, 123.4, 245.7 and 156.2. The values mean that they are extremely significant p < .01. The results show that users did not classify the images randomly.

The intensity increasing in the expressive image generator means that the characteristic of the corresponding emotion increases. We predicted that users should recognise more images correctly with the intensity increasing. The results proved our prediction. From table 1, we also find that the images in the emotion categories *disgust* and *fear* were not recognised as successfully as others.



Happiness

Sadness

Surprise

Anger

Fear

Disgust

Table 1 Results for different emotion categories

Image plus text test

A total of 35 students and staff from the Bournemouth University participated in this experiment. Each subject was told to view a set of 18 pre-selected expressive images, which were generated using our expressive image generator. Below each image a short sentence was written. For example, for a generated "*happy*" image with high intensity, the text is "I am extremely happy". The subjects answer whether the image is appropriate to the text or not.

The results are shown in table 2. (In these figures, 1, 2, 3 in the x-axis represents the increasing corresponding intensity in that emotion category; 1 is the lowest and 3 is the highest. The y-axis shows the percentage of the subjects who agree that the image is appropriate to the text.)

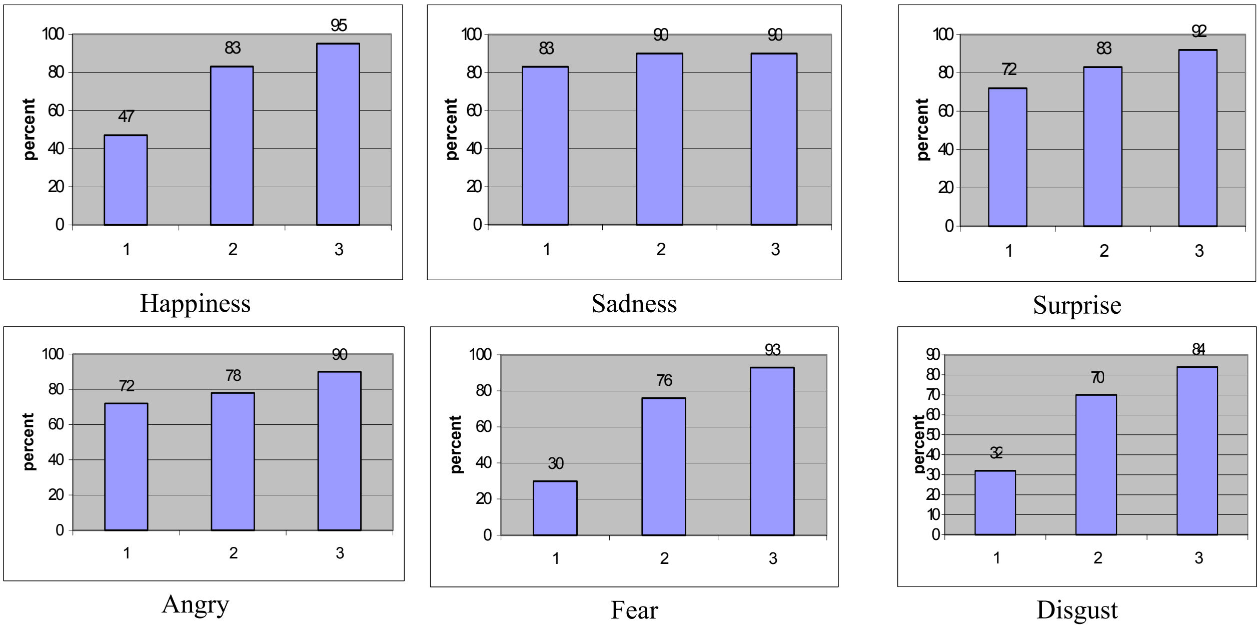
Result analysis

For each category of image, the chi-square test is carried out again. The corresponding obtained values for emotion categories *happiness, sadness, surprise* and *anger* are 8.257, 24.03, 9.91 and 9.94. Those are extremely significant at p < .01. For categories *fear* and *disgust*, the obtained values are 3.671 and 4.82, which are marginally significant at p<0.1. The results show that users did not choose their answers randomly.

This experiment also proved that by increasing expressive intensity, more subjects will correctly recognise the expressive images.

It is shown that with text context information, the subjects correctly recognised more expressive images than in the first test. For emotion categories *happiness, sadness, surprise, fear*, more than 70% images were correctly recognised. For emotion categories *fear* and *disgust*, on average more than 60% images were correctly recognised. For all images with medium and high intensity, more than 78% are recognised correctly.

In this case, the expressive image generator can be successfully used together with emotion extraction engine in a chatting environment.



Happiness

Sadness

Surprise

Angry

Fear

Disgust

Table 2: Results for image plus text test

6 Conclusions

A text based real-time Internet communication system using an emotion extraction engine was developed and operates successfully. Since only text parameters are transmitted over the network, the bandwidth requirements for real-time communications is extremely low with the text to emotion extraction engine.

From previous experiments in paper [5], we established that people prefer using expressive image plus text in the chatting environments. The challenge in this paper is how to generate the expressive images for each user in real time.

An expressive image generator has been developed successfully. The generator can receive an uploaded neutral facial image and generate eighteen different expressive facial images. The eighteen facial images belong to six universal categories: *happiness, sadness, disgust, anger, surprise* and *fear*. For each category, three images with different emotion intensities are archived.

Using the expressive image generator, expressive images for each user can be produced in real time and in a user-friendly way. The generator can be used with the emotion extraction engine in a number of environments that provide text context information e.g. chatting room, story reader and online games. The generator's effectiveness may be enhanced with improvements to the model: expression model mask. Possible refinements include adding more control points and more control areas.

The experiments show that by increasing the emotion intensity, more acceptable expressive images can be obtained.

The test results demonstrate that the generator can create acceptable expressive images for a chatting environment when accompanied by text labels. The expressive image generator solved the image problem described in paper [4] and [5].

7 References

[1] http://emotion.salk.edu/Emotion/History/Hgeneral.html

[2] Buck, R., 2001. *Emotional Expression, Communication, and Competence: Applications of a Developmental-Interactionist Theory of Biological, Social, Cognitive and Moral Emotions*. University of Connecticut. http://wattlab.coms.uconn.edu/ftp/users/rbuck/Germany01/

[3] Dautenhahn, K. & Werry, I., 2001. The AURORA Project: Using mobile Robots in Autism Therapy. *Learning Technology*, IEEE Computer Society Learning Technology Task Force, Vol. 3(1), January 2001, ISSN 1438-0625

[4] Zhe, X. & Boucouvalas, A.C., 2002. Text-to-Emotion Engine for Real Time Internet Communication**.** *International Symposium on Communication Systems, Networks and DSPs*, 15-17 July 2002, Staffordshire University, UK, pp 164-168

[5] Zhe, X, John,D & Boucouvalas, A.C., 2002. Text-to-Emotion Engine: tests of user preferences**.**ISCE 2002, 22-25 September 2002, Erfurt, Germany, pp B25-B30

[6]. Kuehn, S.A., 1993. Communication innovation on BBS, Interpersonal Computing & Technology, 1(2), ISSN: 1064-4326

[7] Leech, G. 1997. A Brief Users' Guide to the Grammatical Tagging of the British National Corpus. http://info.ox.ac.uk/bnc/what/gramtag.html

[8] Francis, W.N. & Kucera, H., 1979. *Brown Corpus Manual*, Department of Linguistics, Brown

[9] Russell, S. & Norvig, P., 1995. *Artificial Intelligence - a modern approach*, Prentice Hall, New Jersey

[10] Parke, F.I., & Waters, K., 1996. *Computer Facial Animation*, A.K. Peters, Wellesley, Massachusetts

[11] Paul Ekman and W.V. Friesen "facial action coding system"

[12] Cambridge digital research laboratory : faceworks. http://interface.digital.com/

[13] Frederic Pighin, JLOE Auslander, Dani Lischinski and David Salesin. Realistic Facial Animation Using Image-Based 3D Morphing. Technical report UW-CSE-97-01-03.

[14] "digital Image warping", George Wolberg ISBN: 0-8186-8944-7

[15].A. Goshtasby, "Piecewise Linear Mapping Functions for Image Registration", Pattern Recognition, Vol. 19, No. 6, pp. 459-466, 1986.

[16] Jocelyn Scheirer, R. Fernandez, J. Klein, and R. W. Picard (2002), "Frustrating the User on Purpose: A Step Toward Building an Affective Computer", Interacting with Computers, 14, 2 (2002), 93-118.

[17] Maes, P. Velfisquez, J.D. "Cathexis: A Computational Model of Emotions.

GENETIC PROGRAMMING OF VERTEX SHADERS

Jörn Loviscach
Fachbereich Elektrotechnik und Informatik
Hochschule Bremen
D-28199 Bremen, Germany
E-mail: jlovisca@informatik.hs-bremen.de

Jennis Meyer-Spradow
Fachbereich Informatik
Universität Hannover
D-30167 Hannover, Germany
E-mail: meyersp@gdv.uni-hannover.de

KEYWORDS
Genetic Programming, Evolutionary Art, Vertex Shader, Video Texture.

ABSTRACT

Modern consumer 3-D graphics chips can synthesize procedural textures at a speed comparable to or even better than typical CPUs. We propose genetic programming of vertex shader assembly code for the real-time display and interactive design of procedural video textures and for the approximation and artistic abstraction of given static textures by compact vertex shaders.

INTRODUCTION

Software such as Corel™ (Ex-MetaCreations) KPT Texture Explorer® has long been used to design 2-D textures by evolution of programs. The user sees a small set of genetic mutations and chooses from them the genotype for the next round of mutations. We extend this principle to employ the programmability of modern off-the-shelf graphics hardware: Static and animated procedural textures can efficiently be generated directly on the 3-D chip. Our software allows evolutionary design of *animated* textures based on aesthetic selection by the user and automatic evolution of *static* textures.

The major contributions of the ongoing work presented here are:

- Procedural textures are produced via vertex shaders subjected to genetic programming.
- Animated (not only static) textures are evolved under human supervision in real time (see Fig. 1).
- A given static texture is imitated by vertex shaders. This unsupervised process allows artistic abstraction as well as compact representation of an image (see Fig. 2).

RELATED WORK

Following the seminal work of Sims (Sims 1991), many researchers have applied genetic programming (Koza 1989) with human aesthetic selection to art and to graphic design, including such areas like 3-D morphogenesis and music generation (Benteley 1999). Abraham's Genshade (Ibrahim 1998) evolves RenderMan shaders either with or without human supervision. In unsupervised mode it tries to mimic given 2-D images like the Gentropy system by Wiens and Ross (Wiens 2002, Ross 2002). They mutate and recombine LISP expressions and employ a set of quality measures inspired by algorithms for query by example in image data bases. Saunders and Gero combine a genetic art software with a neural network judging the novelty of a generated picture (Saunders and Gero 2002). They study the sociology of groups of such agents with different behaviors.

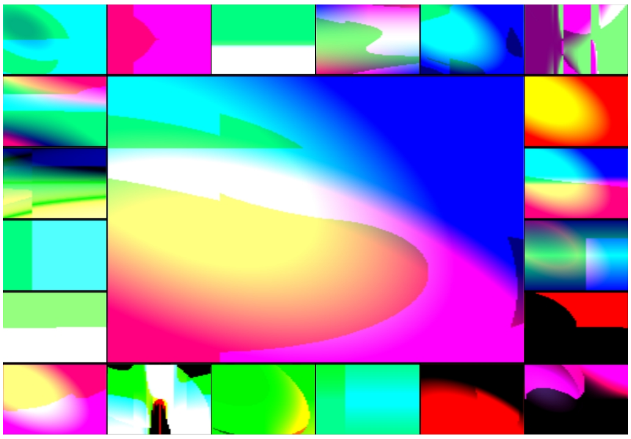


Figure 1: Interactive Evolution of Video Textures

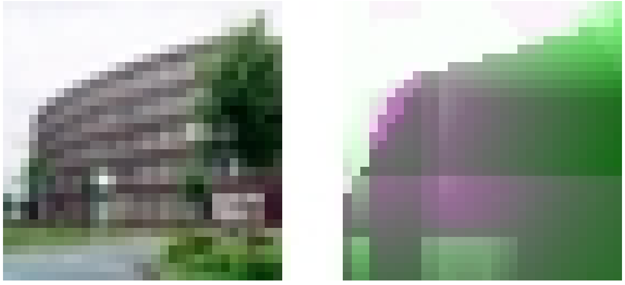


Figure 2: Artistic Abstraction of Static Textures

Thompson and co-workers analyze the performance of the graphics chip nVidia® GeForce4 Ti 4600™ in general-purpose applications such as matrix multiplication and 3-SAT (Thompson et al. 2002). They find that in many cases the graphics chip outperforms a Pentium® 4 class CPU.

TEXTURING BY VERTEX SHADERS

The programmabilty of modern graphics chips rests on vertex shaders (also known as vertex programs) and pixel shaders (also known as fragment programs). These are small programs run on the graphics chip for each vertex or for each pixel fragment, respectively, being processed.

Pixel shaders are the obvious choice for texture operations. However, the available graphics cards still offer only very restricted functionality for pixel shaders. Therefore, we chose to use vertex shaders to generate textures.

Our prototype supports the nVidia® GeForce3 Ti and GeForce4 Ti class ships. On them, a vertex shader can consist of up to 127 instructions from a set of 21 base types. Most of them are fully SIMD (single instruction multiple data), such as the component-wise addition of two vectors, each consisting of four 32 bit floating point values. A vertex shader has a workspace of 12 write- and readable registers. The instructions can read from 96 constant registers and from 16 vertex attribute registers. All of these registers store vectors of four floating point components.

A vertex shader operation can permute and/or negate the components of its inputs. Its effect can be limited to a subset of the four components of the output register. Indirect addressing is possible, but not used here due to its complexity in programming and results.

The evolved shaders only read the 2-D position from the vertex attributes. In our application, each of the four floating point components of the 96 constants are set to fixed pseudo-random floating point values between 0.0 and 1.0. For the evolution of animated textures the constants are filled with sine and cosine waves of integer frequencies so that the animation smoothly repeats every second.

To every evolved shader an instruction is appended to use the content of one of the registers as color of the vertex. The vertex shader assembler enforces that the vertex position be written. Thus, in total we are left with 125 instructions to evolve freely, see Fig. 3. (On the GeForce3 Ti and GeForce4 Ti there is no need to sacrifice instructions to initialize registers within a vertex shader; this is in contrast, for instance, to the ATI® Radeon™ 9700.)

In order to minimize the number of vertex shader runs, the software plots a display list of point drawing commands. The points address pixel per pixel of the buffer. Using polygons instead of points would drastically increase the number of vertices needed, although the vertex cache of the graphics card might lessen this problem. In the interactive version the display list is rendered into an offscreen buffer, which is painted as a texture to fill an arbitrarily large window with relatively little computation.

GENETIC VERTEX SHADERS

Genetic programming of machine code typically needs severe restrictions (Harvey et al. 1999) or to take precaution against syntactic errors and against operations that result in errors or compromise system stability (Kühling et al. 2002). However, for the vertex shaders of the class of 3-D chips employed here, loops as well as branching are not available, every operation can digest any input value, and no output can do any harm to the system. Therefore, such programs can easily be mutated and recombined as instruction lists. Their overall structure resembles Cartesian Genetic Programming (Miller and Thompson 2000).

```
!!VP1.1
MOV o[HPOS], v[OPOS];
ABS R4.xy, -c[30].yzxz;
DPH R4.w, -R11.yyzy, -c[75].wzwx;
LIT R5.z, c[43].xwzx;
ADD R9.xyw, v[OPOS].wyxx, c[36].yzwx;
EXP R11.xyz, -R4.x;
SGE R3.y, R9.wyww, -c[64].zzwx;
RCC R0.zw, -R9.w;
LOG R1.xw, -c[46].w;
RCC R1.yzw, v[OPOS].x;
MUL R7.xyw, R11.xyyw, -v[OPOS].xyzw;
DPH R11.xyz, -R9.zwxy, -c[75].wyzx;
MAX R0.xz, c[70].zyxw, R9.yzxy;
// 70 lines deleted from this listing
LIT R10.x, -R10.zxzz;
SUB R2.xyw, -c[66].wzyz, R10.xyzy;
DP3 R9.x, -R8.ywzw, R9.xwzw;
SUB R1.xyw, c[59].xwyy, R9.yxzz;
DP3 R1.xz, R3.xzxz, -R2.wwxx;
MAX R0.yzw, -R9.wwzy, -R1.zyxz;
MIN R0.xzw, c[32].wywz, -R9.yyyy;
MOV o[COL0], R0;
END#
```

Figure 3: Assembler Code of an Evolved Vertex Shader

Genetic Programming is characterized by the fast growth of functionally superfluous code such as introns (Soule and Heckedorn 2002). Such code bloat may possess advantages for genetic operations, especially by protecting other code against the harmful action of crossover. It is inefficient, however, for the evaluation of the evolved programs.

Therefore, for the rendering of the programs we have implemented an optimization so that all instructions are skipped which only write into register components not used at a later stage. Thus, typically about half of the instructions are discarded which nearly doubles the rate at which evolved vertex shaders are converted from assembler code and rendered.

In the interactive application for animated textures, the user clicks onto the videos to be varied on their own (mutated) or mixed with others (recombined). To accelerate the interactive search for complex graphical and temporal patterns, we found it helpful to restrict the number of registers used: Using only three of the twelve available registers leads to heavy use of former results in later instructions and thus to complexity.

For the unsupervised evolution of static textures guided by an input picture, the result is read back from the graphics buffer and compared to the input. As a measure for fitness we use the sum of the mean absolute difference (L_1) plus the maximum absolute difference (L_∞).

To use a PC cluster for accelerating the unsupervised evolution, we let each computer evolve its own population; additionally, the PCs exchange genotypes as files on a shared disk volume. This corresponds to the island model of multipopulation genetic programming (Fernández et al. 2003).

RESULTS

The implementation prototype has been written in C++ with OpenGL™. Our benchmarks for the GeForce4 Ti 4200 show a maximum speed of approximately 474,000 vertex instructions per second (distributed onto several execution units in the 3-D chip), which is equivalent to nearly 2 GFLOPS, roughly the processing power of the vector floating point instructions (so-called SSE) of Pentium® III class processors. Simultaneously displaying 20 small animated textures on 64 times 64 vertices and one large on 256 times 256 vertices, the interactive solution delivers (depending on the length of the shaders after optimization) 30 to 40 frames per second on a PC equipped with a Pentium® 4 running at 2.5 GHz.

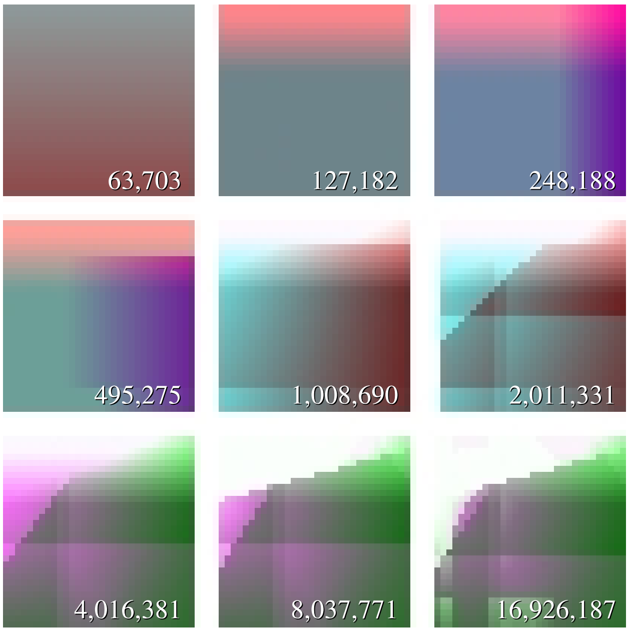


Figure 4: Best Fit after a Given Number of Evaluations

The automatic evolution suffers from the bottleneck that evolved programs cannot be sent as opcodes to the graphics driver but only as assembler code. The overhead for building assembler programs and letting the graphics driver parse them limits the performance for small pictures: At a size of 32 times 32 pixels, we achieve approximately 1200 evolved genotypes per second.

OUTLOOK

Instead of providing time as input to a shader evolved in real time, one could also use mouse position, motion, the bass level of a music signal and many other variables. A novelty detector (Saunders and Gero 2002) may be of help in the supervised mode. For automatic, unsupervised evolution, the measurement should mimic human perception like in (Wiens and Ross, 2002).

When the nVidia® GeForceFX generation of graphics chips becomes available, we will switch from vertex shaders to pixel shaders. They will then have similar or even more programmability features than the recent vertex shaders. Because pixel shaders employ a larger number of parallel execution units, we expect the throughput to double at least.

We hope to be able to formulate physiological measurements as such pixel shaders, not only to speed up the computation but also to skip reading back the generated textures to the main memory. Such measurement programs may be found by co-evolution of synthesis and analysis.

REFERENCES

Bentley, P. (Ed.). 1999. *Evolutionary Design by Computers.* Morgan Kaufmann, San Francisco, CA

Ibrahim, A.E. 1998. “Genshade: an Evolutionary Approach to Automatic and Interactive Procedural Texture Generation”. Doctoral Thesis, College of Architecture, A&M University, College Station, TX.

Fernández, F.; M. Tomassini; L. Vanneschi. 2003. “An Empirical Study of Multipopulation Genetic Programming”. *Genetic Programming and Evolvable Machines* 4, 21-51

Harvey, B.; J. Foster; and D. Frincke. 1999. “Towards Byte Code Genetic Programming”. *Proc. of the Genetic and Evolutionary Computation Conf.* (Orlando, FL, Jul. 13-17). Morgan Kaufmann, San Francisco, CA, 1234-1241.

Koza, J.R. 1989. “Hierarchical Genetic Algorithms Operating on Populations of Computer Programs”. In *Proc. of the 11th Int. Conf. on Genetic Algorithms* (San Mateo, CA, Aug. 20-25). Morgan Kaufmann, San Francisco, CA, 768-774.

Kühling, F.; K. Wolff; and P. Nordin. 2002. “Brute-Force Approach to Automatic Induction of Machine Code on CISC Architectures”. *Genetic Programming, Proc. of the 5th European Conf.* (Kinsale, Ireland, Apr. 3-5). Springer, Berlin, 288-297.

Miller, J.F. and P. Thomson. 2000. “Cartesian Genetic Programming”. *Genetic Programming, Proc. of EuroGP 2000* (Edinburgh, UK, Apr. 15-16). Springer, Berlin, 121-132.

Ross, B. and H. Zhu. 2002. “Procedural Texture Evolution Using Multiobjective Optimization”. Technical Report CS-02-18, Dept. of Computer Science, Brock University, St. Catharines, Ontario (Jul)

Saunders, R. and J.S. Gero. 2002. “How to Study Artificial Creativity”. *Proc. of the 4th Conf. on Creativity & Cognition* (Loughborough, UK, Oct. 14-16). ACM, 80-87

Sims, K. 1991. “Artificial Evolution for Computer Graphics”. *Computer Graphics* 25, No. 4, 319-328.

Soule, T. and R.B. Heckendorn. 2002. “An Analysis of the Causes of Code Growth in Genetic Programming”. *Genetic Programming and Evolvable Machines* 3, 283-309.

Thompson, J.; S. Hahn; and M. Oskin. 2002. “Using Modern Graphics Architectures for General-Purpose Computing: A Framework and Analysis”. *International Symposium on Microarchitecture* (Istanbul, Turkey, Nov. 18-22). IEEE, 306-320

Wiens, A.L. and B.J. Ross. 2002. “Gentropy: Evolving 2D Textures”. *Computers & Graphics* 26, 75-88.

LOW COST APPROACH TO ACOUSTIC FIELD ENHANCEMENT FOR STEREO HEADPHONES

Nobuyuki Iwanaga[†], *Wataru Kobayashi*[†,††], *Kazuhiko Furuya*[††], *Noriaki Sakamoto*[†], *Takao Onoye*[†], *Isao Shirakawa*[†]

[†] Dept. Information Systems Engineering, Osaka University
2-1 Yamada-Oka, Suita, Osaka, 565-0871 Japan
{iwanaga, kobachan, sakamoto, onoye, sirakawa}@ise.eng.osaka-u.ac.jp
[††] Arnis Sound Technologies, Co., Ltd.
2-7-9 Kita-Senzoku, Ota-ku, Tokyo, 145-0062 Japan
{kobachan, kzf}@arns.com

KEYWORDS

headphone stereo, 'out-of-head' acoustic field enhancement, 3-D Sound localization, frequency division

ABSTRACT

This paper proposes low cost implementation of acoustic field enhancement dedicatedly for stereo headphones. This aproach can reduce the computational costs for signal processing to widen an acoustic field generated through stereo headphones on the basis of the feature extraction of the head related transfer function (HRTF). Software implementation are attempted by using a DSP and an embedded RISC processor in order to demonstrate the practicability of the proposed approach. For both types of processors, acoustic field enhancement is achieved with dissipating 50mW. Subjective test results of the generated sound are also executed.

1. INTRODUCTION

In recent years, a number of approaches are extensively attempted to attain high fidelity acoustic environments. Especially, it is said that in the coming mobile era there will be strong demands to develop such excellent acoustic fields applicable for mobile computing.

In the audio listening with portable devices, listeners generally enjoy sounds through stereo headphones. Since stereo sounds from headphones directly reach to eardrums of listeners without spatial cues imparted, sound images are localized inside their heads, and thus listeners are always suffering from the so-called *listening fatigues*.

To settle this problem, several researches have been performed for the acoustic field enhancement, which portray the sound of a two-speaker system virtually to lessen the fatigues. Conventional approaches(Rubak 1991; Fujinami 1998; Lake Technology Ltd.), however, incur enormous computations with a huge size of memory, for which a single chip DSP can hardly execute the whole process.

Motivated by this tendency, the present paper describes a low-cost algorithm to enhance an acoustic field of stereo headphones to the extent of out-of-head, which extracts fundamental factors necessary for perceiving realistic sound effects through the human auditory sense on the basis of a 3D acoustic image localization algorithm run on an embedded DSP(Kobayashi et al. 2001; Sakamoto 2000). To demonstrate the practicability of the algorithm, two types of implementations using a TMS320C54x DSP and an ARM7 RISC processor are evaluated, for which subjective tests are carried out.

2. ACOUSTIC FIELD ENHANCEMENT

When the sounds from each speaker of a stereo system reach to eardrums of listeners, a set of acoustic factors influence the sounds. These factors are associated with the acoustics of a room, such as the attenuation by a sound path, the reflection/diffraction by the wall/floor, and the physical properties of the listeners' heads and outer ears. By receiving effected sounds the listeners can perceive the location of a sound source in terms of the direction and distance, as well as its pitch, tone, and level.

When one listens to sounds of a two-speaker stereo system, each ear of the listener receives the direct sound from each speaker (i.e. left speaker to left ear, right speaker to right ear), together with the cross-talk sound from the opposite side speaker. Therefore the listener can detect the direction and distance of the sound source by sensing several differences of the received sound, such as pressure and time differences.

Contrary to this, the stereo sound through headphones directly reaches to the listener's eardrums without any of spatial cues imparted, and thus left and right sounds are localized just beside the listener, while other sounds are inside the head, forming a straight line between ears. This in-head localization yields unnatural acoustic fields and the listener may feel uncomfortable, which causes listening fatigues.

To widen the acoustic field to the extent of out-of-head, a number of researches have been attempted(Rubak 1991; Fujinami 1998; Lake Technology Ltd.), which are based on virtual simulations of speaker systems. Generally, at first a listening room is supposed, where two or 5.1 channel speakers are placed. Then the characteristics of a sound propagation path, such as the arrival time from a virtual speaker to left/right ear, the head-related transfer function, the reflection by a wall of the listening room, etc., are extracted to simulate an acoustic field from multiple speakers in the virtual listening room.

In (Rubak 1991), eight 226 tap FIR filters are used for simulating wall-reflected sounds and cross-talk sounds diffracted by the listener's head. A key feature of wall-reflected sounds utilized in (Fujinami 1998) is that the impulse response in the high frequency band lasts for a shorter time than that of the low frequency band does. Specifically, in this approach, first the audible frequency band is divided into two subbands by using a filter bank, and then high tap FIR filters are used only for the lower frequency subband to reduce the total computational costs. A single universal HRTF setting is accomplished in (Lake Technology Ltd.), taking account of the direct sound and multiple reflections contributed by the room's surfaces and furnishings. Although two (stereo) to five (surround) speaker system in a virtual room can be simulated, two 7,000 tap FIR filters have to be employed in this algorithm to portray each virtual speaker.

These approaches entail an enormous cost of high-precision computations, for which multiple DSPs are used for 20- or 24-bit fixed-point or 32-bit floating-point calculations. Moreover, the required memory capacity is also large, and thus it can be hardly incorporated into an embedded system with practically low cost and low power consumption. In addition, the achieved acoustic field may fall into an unrealistic one due to the artificial reverberation and/or flanging effect, when an identical processing filter is applied to allover the audible frequency band.

3. NOVEL SCHEME FOR ACOUSTIC FIELD ENHANCEMENT

3.1. Algorithm

To cope with the rigid condition for embedded systems, a simple yet effective algorithm is discussed here, which extracts fundamental factors necessary for perceiving realistic sound effects through the human auditory sense. Specifically, this algorithm executes an acoustic field enhancement by using the following three sound paths illustrated in Fig. 1.

- *Direct*: Direct sound path from a speaker to the listener's corresponding ear.
- *Reflected*: Reflected sound path, which is reflected at a wall and reaches to the listener's corresponding ear.
- *Crosstalk*: Crosstalk sound path from a speaker to the listener's opposite ear, which is diffracted by the listener's head.

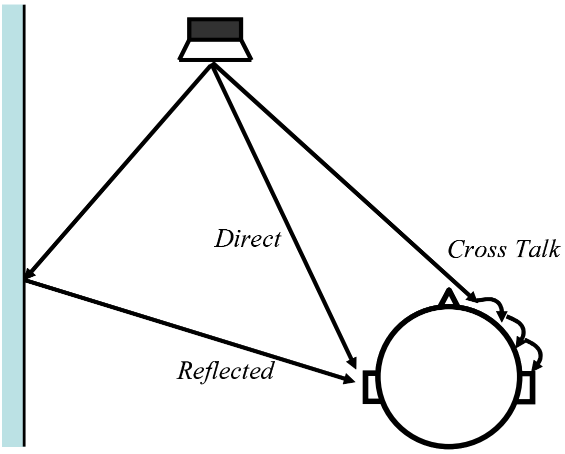


Figure 1: Direct, Reflected, and Crosstalk Sound to Human Ears

In order to reproduce those sounds through headphones, *Reflected*/*Crosstalk* sounds are obtainable by adding to *Direct* sounds the pressure and time differences between *Direct* and *Crosstalk* sounds.

As for *Crosstalk*, it should be noticed that the pressure and time differences vary with frequencies, since the sound path is diffracted by the human head and pinna. Thus, based on the scheme proposed in (Kobayashi et al. 2001; Sakamoto 2000), our algorithm divides the whole audible frequency band into three subbands, low, intermediate, and high subbands, such that the synthesis of *Crosstalk* sound in each subband can be efficiently achieved.

It is reasonable to assume that a human head has a spherical shape with a diameter of 150-200mm, although there are a little differences among individuals. If the half of a sound wavelength is longer than the head diameter, the effect of sound diffraction by the head is negligibly small. Thus frequency response in the low frequency band appears flat.

On the other hand, in the high frequency band, the sound diffraction by a human pinna dominantly effects on the frequency response. Suppose that the pinna is a cone with a base diameter of 35-55mm. If the half of a sound wavelength is shorter than the pinna diameter, then the diffracted sound issuperposed to the original sound, and thus the frequency response in this band looks like a comb.

Finally, the intermediate frequency band has a complicated frequency response, since there are a number of diffraction factors overlaid to the original sound.

Figs. 2 and 3 exemplify measured pressure differences (differences of amplitude spectrum) and measured time differences (differences of phase delay spectrum), respectively, between *Direct* and *Crosstalk* sounds. It can be seen from Fig. 2 that the frequency response of the pressure differences show complicated charactaristics. However, it is generally supposed that only the average level of the pressure differences is essential for perceiving horizontal directions of the sounds, while the spectrum variation in the high frequency subband is effective for perceiving vertical and cross directions. Furthermore, it is known that the time differences are also essential for perceiving horizontal directions, especially in the low frequency subband(see Fig. 3).

Since perceiving the horizontal direction of a sound is the most impotant for the acoustic field enhancement, the proposed algorithm is constructed by replicating mainly the average level of the pressure differences in all subbands and the time differences only in the low frequency subband.

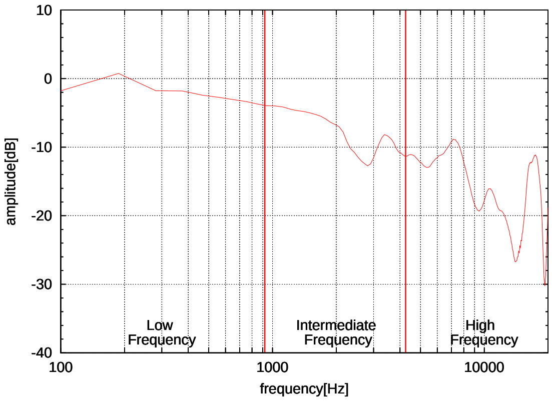


Figure 2: Example of Pressure Difference between *Direct* and *Crosstalk*

3.2. Signal Processing Flow

Fig. 4 outlines an invocation of the proposed algorithm. Input signals are used for *Direct* sounds without change, and *Reflected* sounds are controlled by delays depending on arrival times. As for *Crosstalk* sounds, subsequent to the frequency division, each of the low, intermediate, and high subbands is controlled by the phase-controller, so that different phase-controls can be applied to each subband. Finally, the sounds obtained by these processes are mixed with *Direct* sound so as to generate out-of-head stereo sounds.

To facilitate various types of implementations, the alternative architectures are prepared in each part of processing, the details of which are described bellow.

At first, in the frequency division part, low-pass, bandpass, and high-pass filters are used for dividing input sig-

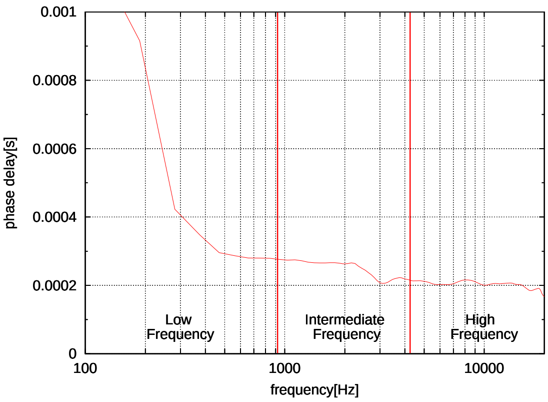


Figure 3: Example of Time Difference between *Direct* and *Crosstalk*

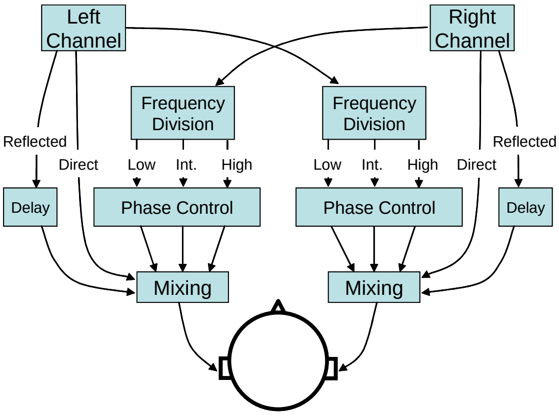
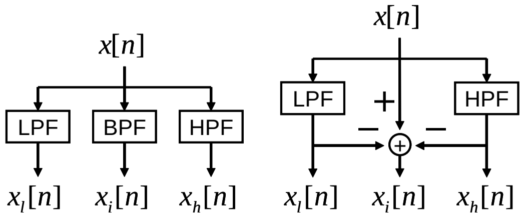


Figure 4: Overall Organization of Proposed Scheme

nals $x[n]$ into low subband signals $x_l[n]$, intermediate subband signals $x_i[n]$, and high subband signals $x_h[n]$, as illustrated in Fig. 5(a). Here, a 32nd order linear phase FIR filter is adopted for each of these filters, which can maintain 40dB/oct slope without phase adjustment.

The FIR filter processing requires a considerable amount of operations in the environment of low performance on the product-sum operations. In such an environment, the intermediate subband signals can be attained by subtracting low and high subband signals from input signals, as illustrated in Fig. 5(b).



(a)Architecture 1 (b)Architecture 2

Figure 5: Architecture of Frequency Division

In the phase control part, signal phases are controlled by using 2nd order IIR filters as PEQs (parametric equalizers) in conjunction with delay elements, as illustrated in Fig. 6. As described in 3.1, intermediate and high subband signals are not so important for the phase-control as compared with low subband signals, and thus PEQs for intermediate and high subbands can be removed in order to reduce the total computational costs, if necessary.

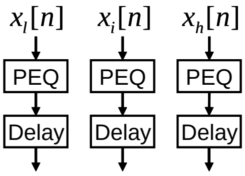
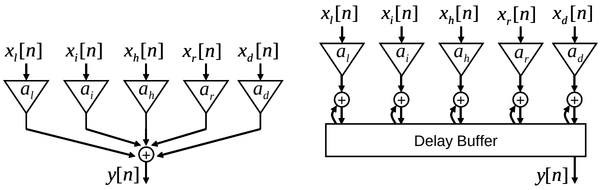


Figure 6: Architecture of Phase Control

Finally, in the mixing part, the gain-control is performed by multiplying *Direct* $x_d[n]$, *Reflected* $x_r[n]$ *Crosstalk* $x_l[n]$, $x_i[n]$, $x_h[n]$ by gains a_d, a_r, a_l, a_i, a_h, respectively, as illustrated in Fig. 7(a). This architecture offers low computational costs in the environment that supports the circular buffer operations. As another architecture as illustrated in Fig. 7(b), a delay buffer can be used repeatedly for mixing five intermediate signals to reduse the memory size.



(a)Architecture 1 (b)Architecture 2

Figure 7: Architecture of Mixing

4. IMPLEMENTATION RESULTS

4.1. Embedded DSP Implementation

The proposed scheme has been implemented on a Texas Instruments 16-bit fixed point DSP TMS320C54x. The input/output sampling rate is set to 44.1kHz, which is generally used in portable audio devices, such as CD and MD.

A block diagram of this implementation is shown in Fig. 9(a), where the frequency division of *Crosstalk* sounds is achieved by using three FIR filters.

In this implementation FIR filters for the frequency division are also used for controlling the phase of *Crosstalk* so that PEQs for intermediate and high subbands can be removed.

Input/output data of this implementation are in 16-bit precision integers, while intermediate data between the processes are in 32-bit fixed point representation of 24-bit integer and 8-bit decimal.

Fig. 8 shows an evaluation board of the DSP implementation. Computational costs of the acoustic field enhancement by the embedded DSP are summarized in Table 1, where totally 34.4MIPS is necessary for realtime processing. Requested sizes for ROM and RAM are 8k bytes and 4k bytes, respectively. Power consumption of this DSP implementation is evaluated as 56mW with 1.8V power supply. Consequently, it is confirmed that the out-of-head acoustic field enhancement scheme is compliant to the single chip DSP for mobile applications.

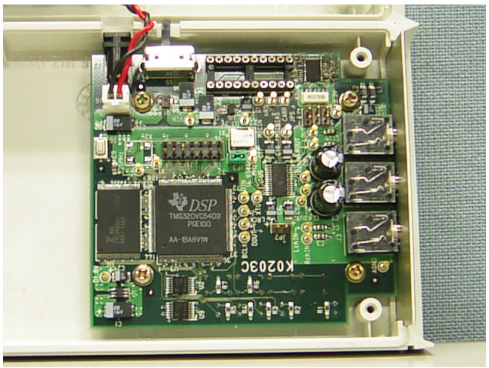


Figure 8: Picture of DSP Evaluation Board

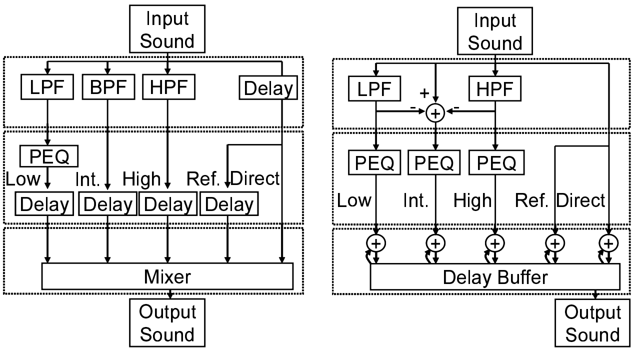
4.2. Embedded RISC Implementation

Considering the portability of our scheme to a number of embedded RISC processors, it should be practical to evaluate a software implementation in C, in which an arbitrary filter organization can be selected in accordance with its application.

Our software implementation on an ARM7TDMI processor, is evaluated with the use of ARMulator.

As shown in Fig. 9(b), the intermediate subband of *Crosstalk* sounds is achieved by subtracting, since the RISC processor has low performance on the product-sum operation. In contrast with DSP implementation, however, this frequency division scheme has a low capability of phase control. Thus PEQs are used for all subbands of *Crosstalk*. In addition, an RISC processor may suffer from limited on-chip memory size. To cope with this difficulty, in this implementation a delay buffer is repeatedly used.

As a result, 81MIPS can be attained for realtime processing, which can perform ARM7TDMI by using 0.13μm technology. Required sizes for ROM and RAM are 5.5k bytes and 1k bytes respectively. Power consumption is 18mW from a 1.8V supply.



(a) DSP implementation (b) RISC implementation

Figure 9: Block Diagram of Two Types of Implementation

Table 1: Impulementation Results

		DSP	RISC
Computational costs	[MIPS]	34	81
RAM	[k bytes]	4	1
ROM	[k bytes]	8	5.5
Power consumption	[mW]	56	18

5. SUBJECTIVE TESTS

The sound quality achieved in the proposed algorithm is evaluated by subjective tests based on Mean Opinion Score (MOS). In our subjective test, 20 participants have listened to the original and enhanced sounds through stereo headphones. Afterwards, each participant has evaluated the sound quality of the enhanced sounds in comparison with the original sounds. “Sense of distance", “direction", and “spatiality" are rated according to MOS with categories “excellent (5)", “good (4)", “fair (3)", “poor (2)", and “bad (1)". Moreover “feeling fatigue" is rated with “no fatigue (5)", “less fatigue (4)", “fair (3)", “fatigue (2)", and “hard fatigue (1)". The results are summarized in Table 2, which claims that the proposed algorithm can achieve the sound with good spatiality and less fatigue.

Table 2: Subjective Tests Results

distance	direction	spatiality	fatigue
3.6	3.2	4.8	3.9

6. CONCLUSIONS

This paper has devised an embedded implementation of an acoustic field enhancement for stereo headphones. Based on the analysis of the human auditory sense, this approach successfully performs the acoustic signal processing to widen an acoustic field generated through stereo headphones with low computational costs. The proposed scheme is experimentally implemented on a TMS320C54x DSP and an ARM7TDMI RISC processor. In both implementations, the acoustic field enhancement can be achieved with low computational costs and low memory size in comparison with the conventional approaches. Subjective test of the proposed scheme claims that good spatiality and less fatigue can be achieved.

Development is continuing further on the investigation of a more efficient out-of-head acoustic field enhancement scheme not only for stereo headphones but also for stereo speaker system.

REFERENCES

Rubak, P. 1991. “Headphone signal processing system for out-of-head localization," *Audio Eng. Soc. Preprint 3063* (Feb), 16.

Fujinami, Y. 1998. “Improvement of sound image localization for headphone listening by wavelet analysis," *IEEE Trans. Consumer Electronics*, vol. 44, no. 3 (Aug), 1183–1188.

Lake Technology Ltd., “Dolby Headphone DSP Implementation."

Kobayashi, W.; N. Sakamoto; T. Onoye; and I. Shirakawa. 2001. “3D acoustic image localization algorithm by embedded DSP," *IEICE Trans. Fundamentals*, vol. E84-A, no. 6 (June), 1423–1430.

Sakamoto, N.; W. Kobayashi; T. Onoye; and I. Shirakawa. 2000. “Low power DSP implementation of 3D sound localization," in *Proc. International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2000)* (July), 253–56.

A New Approach to Speech Synthesis Based on Fractal Dimension

S. Fekkai, M. Al-Akaidi
Faculty of Computing Sc. & Engineering,
De Montfort University, Leicester, LE1 9BH, UK.
Email: mma@dmu.ac.uk

Abstract

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms has been under development for several decades [1,2]. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem.

The goal of this work was to develop a new speech synthesis system, which is based mainly on the fractal dimension to create natural sounding speech. Our initial work in this area showed that by careful use of the fractal dimension together with the phase of the speech signal to ensure consistent intonation contours, natural –sounding speech synthesis was achievable with word level speech. In order to extend the flexibility of this framework, we focused on the filtering and compression of the phase to maintain and produce natural sounding speech.

Introduction

In an ideal world, a speech synthesizer should be able to synthesize any arbitrary word sequence with complete intelligibility and naturalness. The trade-off schematic in Figure 1 illustrates how current synthesizers have tended to strive for flexibility of vocabulary and sentences at the expense of naturalness (i.e., arbitrary words can be synthesized, but do not sound very natural). This applies to articulatory, rule-based and concatenative methods of speech synthesis [3,4,5,6].

An alternative strategy is one, which seeks to maintain naturalness by operating in a constrained domain. There are potentially many applications where this mode of operation is perfectly suitable. In conversational systems for example, the domain of operation is often quite limited, and is known ahead of time [7].

Past work by others have examined how unit selection algorithms can be formulated, and what constraints must be maintained [3,5,6].

In this work, we develop a framework for natural-sounding speech synthesis using fractal dimension. The developmental philosophy that we have adhered to throughout the work, places naturalness as a paramount goal. In our preliminary work involving word fractal dimension, the vocabulary size is relatively small, but naturalness is very high. Evaluation of the Fractal Dimension

Fractal dimension as parameter is important because it can be defined in terms of real-world data, and can be measured approximately by means of experiment [8-10]. Fractal dimension is a real number that in general, falls between the limits of 1 and 5 and can be calculated in a number of ways.

In order to test the validity and accuracy of the different methods for computing the fractal dimension, a fractal signal of size 256 with a known fractal dimension D was created by filtering white Gaussian noise with the filter $1/k^q$ where $q = (5-2D)/2$. The fractal dimensions of the fractal signal were then evaluated using the Walking Divider Method (WDM), the Box Counting Method (BCM), the continuous box counting Method (CBCM) and the Power Spectrum Method (PSM). The results are given in Table 1.

From Table 1, it is clear that the Power Spectrum Method provides the most consistently accurate results throughout the range 1 to 2. The box counting method provides good results for fractal dimensions with a value below 1.5. After this, the fractal dimension is below the original value used to synthesize the data; for a value of 2.0, the box counting method returns a value of 1.6. However, this does not mean that the Box Counting Method will give results as low as this for all fractal signals used as the test procedure considered here uses a signal generator based on filtering white Gaussian noise. The Walking-Divider Method provides a good approximation of the fractal dimension for values below 1.5, returning results that are slightly higher than those produced by the Box Counting Method and the continuous Box Counting Method. For an original value of 2.0 the Walking-Divider Method returns a value of 1.604. Of the four methods tested, the PSM is the fastest as it is based on a non-iterative approach based on the least square estimate which relies on the use of an FFT.

The accuracy, efficiency and the versatility of the PSM lead naturally to its use in many areas of signal processing. The test discussed above provides confidence in the realization that the PSM is the most appropriate technique for applications to speech processing.

In this paper the fractal dimension computation is obtained from the PSM only.

Original value of D	WDM	BCM	CBCM	PSM
1.0	1.220	1.097	1.187	1.006
1.1	1.268	1.147	1.268	1.138
1.2	1.287	1.165	1.255	1.219
1.3	1.351	1.265	1.309	1.273
1.4	1.404	1.307	1.344	1.382
1.5	1.451	1.380	1.417	1.495
1.6	1.501	1.418	1.451	1.599
1.7	1.542	1.482	1.507	1.705
1.8	1.592	1.537	1.612	1.832
1.9	1.617	1.554	1.621	1.942
2.0	1.604	1.561	1.650	1.997

Table 1. Evaluation and Comparison of fractal dimension

Power Spectrum Method

For the case of fractal speech signals and curves the fractal dimension lie between 1 and 2. The Power Spectrum Method (PSM) [11] has been used as an application of the Fourier power spectrum technique to calculate the fractal dimension of speech phonemes. The speech signal is Fourier Transformed by means of an FFT and the power spectrum is computed, $P_i = \mathrm{Re}(k_i)^2 + \mathrm{Im}(k_i)^2$. Assume that P_i is the measured power spectrum then $\hat{P}_i$ is the expected form of the fractal power spectrum, $\hat{P}_i = \mathrm{c}\ |k_i|^{-\beta}$, where c is a positive constant and β the positive spectral exponent [12].

Applying the Least Square approach to calculate the spectral exponent β and c yields to the following equation:

$$\beta = \frac{N \sum_{i=1}^{N} (\ln P_i)(\ln|k_i| - (\sum_{i=1}^{N} \ln P_i)(\sum_{i=1}^{N} \ln|k_i|)}{(\sum_{i=1}^{N} \ln|k_i|)^2 - N \sum_{i=1}^{N} (\ln|k_i|)^2} \qquad (1)$$

&

$$C = \frac{\sum_{i=1}^{N} \ln P_i - \beta \sum_{i=1}^{N} \ln|k_i|}{N} \qquad (2)$$

Where C= ln c. Using the relationship:

$$D = \frac{5-\beta}{2} \qquad (3)$$

Provides a simple formula for computing the fractal dimension from the power spectrum of a signal.

The implementation of the PSM consists of applying the FFT to the speech signal in order to obtain a spectral representation of the phoneme. A pre-filter step is then used to adjust the estimated values of the fractal dimension to fit within the range 1 and 2. The power spectrum of the pre-filtered signal is computed then the least square approach is applied to calculate the power exponent β (Eq. 1). Hence the fractal dimension D (Eq. 3) is obtained.

It is important to mention that without the pre-filtering step, the values of the fractal dimension were not satisfying the range of the fractal model. However the use of the Pre-filter $(\frac{1}{w})$ has the effect of confirming the speech data to fit the range of the fractal dimension for speech signal which lies between the range 1 and 2.

Synthesising Speech with Fractals

In the previous section we have discussed how the power spectrum of a signal's Fourier transform can be used to extract the fractal dimension. This followed from assuming the power spectrum, $\hat{P}_i$, was related to the dimension in the following form, $\hat{P}_i$= c $|k_i|^{-\beta}$, where β=5-2D.

To create a synthetic fractal is then the process of filtering white noise of the required size with a low pass filter, q whose Fourier transform is $Q(k)=|k|^{-\beta/2}$, where β=5-2D. Using this principle we will start by creating a fractal signal. The process consists of four steps explained bellow:

Step 1: Compute a random Gaussian distributed array $G_i, i=0,1,...,N-1$ using a conventional Gaussian random number generator, with zero mean and unit variance. Compute a random number sequence of uniform distributed numbers $U_i\ ,i=0,1,...,N-1$ in the range zero to one.

Step 2: Calculate the real and imaginary parts; $N_i=G_i\cos 2\pi U_i$ and $M_i=G_i\sin 2\pi U_i$. This defines G_i as the amplitude and U_i as the phase.

Step 3: Filter N_i, M_i with $W_i=\dfrac{1}{K_i{}^{\beta/2}}$ to create N' and M'.

Step 4: Inverse DFT the result using a FFT to obtain

$$n_i=\mathrm{Re}(\hat{F}^{-1}\{N'+iM'\}).$$

The exponent is $\beta/2$ to ensure that the power spectrum, P_k, satisfies:

$$P_k=(N'_k)^2+(M'_k)^2\propto k^{-\beta}$$

By using the same random noise for *U* and *G*, we can see in Figure 1 how changes to *D* affect the signal. The analytic signal is important because it is from this signal that the amplitude, phase and frequency modulations of the original real valued signal can be determined.

If $f(x)$ is a real valued signal with spectrum $F(k)$, then $f(x)$ can be computed from $F(k)$ via the inverse Fourier transform

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{+\infty}F(k)\exp(ikx)dk$$

This involves integrating over k from $-\infty$ to $+\infty$. The analytic signal is obtained by integrating only over the positive half of the spectrum, which contains the physically significant frequencies (i.e. integrating over k from 0 to $+\infty$).

If s is used to denote the analytic signal of f, then by definition:

$$s(x)=\frac{1}{\pi}\int_{-\infty}^{+\infty}F(k)\exp(ikx)dk \qquad (4)$$

From this result it is possible to obtain an expression for s in terms of f which is done by transforming s into Fourier space and analysing the spectral properties of the analytic signal.

The Analytic Signal and the Hilbert Transform

From the last result it is clear that the analytic signal associated with a real valued function *f* can be obtained by computing its Hilbert transform to provide the quadrature component. This process is called quadrature detection.

The analytic signal is a complex function and therefore contains both amplitude and phase information. The important feature about the analytic signal is that its spectrum (by definition) is zero for all values of *k* less than zero. This type of spectrum is known as a single sideband spectrum because the negative half of the spectrum is zero. An analytic signal is therefore a single sideband signal. This result provides another way of computing the Hilbert transform of a function $f(x)$:

- Compute the Fourier transform of $f(x)$
- Set of the component of the complex spectrum $F(k)$ in the negative half space to zero.
- Compute the inverse Fourier transform, which will have real and imaginary parts $f(x)$ and $q(x)$ respectively.

Attributes of the analytic signal

As with any other complex function, the behavior of the analytic signal can be analysed using an argand diagram and may be written in the form:

$$s(x) = A(x)\exp[i\theta(x)] \qquad (5)$$

where

$$A = \sqrt{f^2 + q^2} \qquad (6)$$

$$\theta = \tan^{-1}\left(\frac{q}{f}\right) \qquad (7)$$

The parameter A describes the average dynamical behaviour of the amplitude modulations of f. For this reason, it is sometimes referred to as the amplitude envelope. The parameter θ measures the phase of the signal at an instant in time and is therefore known as the instantaneous phase.

$$s = A\exp[i(\theta + 2\pi n)]; \quad n = 0, \pm 1, \pm 2, \ldots (8)$$

If we confine the value of the phase to a fixed period (i.e. we compute the phase using only one particular value of n), then it is referred to as the wrapped phase.

The next section will illustrate the new algorithm used to reproduce a natural sounding speech synthesis system, which make use of the fractal dimension of the word and its unwrapped phase.

Algorithm Used

The main objective of this algorithm is to create natural sounding speech and to ensure consistent intonation contours.
The work carried out was based on two hypotheses:

First, that the phase of the speech signal carries important information, hence intelligibility of the synthesis speech and second, that the fractal dimension characteristics, if used in the energy of the signal, will reproduce a natural sounding speech.

The algorithm steps are summarised in the block diagram given in Figure 2.

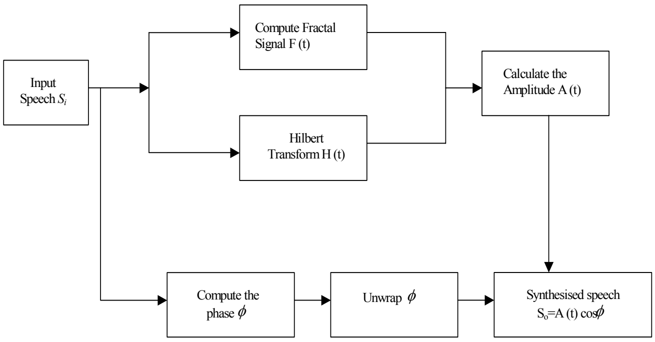

Input Speech S_i
Compute Fractal Signal F (t)
Hilbert Transform H (t)
Calculate the Amplitude A (t)
Compute the phase ϕ
Unwrap ϕ
Synthesised speech S_0=A (t) cosϕ

Fig. 2. Block diagram of speech synthesis using Fractal

Experiments &Discussion

Three experiments have been conducted in the simulation process involving four different words namely “test”, “best”, “Open” and “zone”.
In the first experiment, only the unwrapped phase of the word has been used along with the fractal signal, in the second one the phase is then 65% compressed from the original and finally in the third experiment the phase is low pass filtered and the remaining signal is replaced by a white random noise. The three experiments gave good quality and intelligibility of the synthesised words and they all sounded very natural, however, among the three the best natural sounding speech was enhanced when the phase of the speech signal was low pass filtered and white noise added. It is important to mention here, that the use of the fractal signal in the energy of the reconstructed speech signal has for effect to control the naturalness sounding of speech synthesis.
Figure 3 shows the unwrapped phase of the word “best”, the amplitude, as well as the original word and its reconstructed one. We can clearly notice the similarities in the waveforms of the input speech word with the reconstructed one.

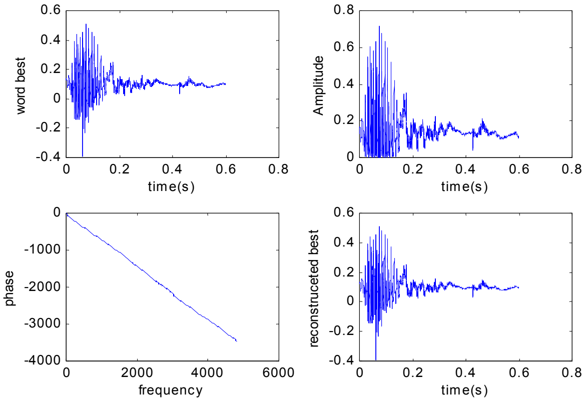


Fig. 3. Synthesis results of word “Best”

To test the validity of our results 10 people have been listening to the synthetic speech and their evaluation is elaborated in Figure 4

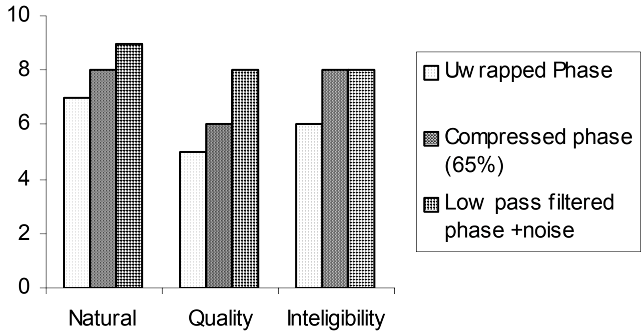


Fig. 4. Evaluation of the synthesis

Conclusion

A new algorithm based on fractals has been used for the synthesis of speech words.

The synthesis process involved three cases of experiments, which increased the quality and the intelligibility of the synthesised speech.

The naturalness level we placed as paramount in our work was highly achieved as a result of the fractal characteristic used in the synthesis process.

Despite the small size of vocabulary we used, the naturalness is very high an as the pursuit of naturalness dominates, human listening provided the best feedback.

References

[1] Kleijn K., Paliwal K., “Speech coding and synthesis”, Elsevier Science B.V., The Netherlands, 1998.

[2] Santen J., Sproat R., Olive J., Hirschberg J., “Progress in speech synthesis”, Springer-Verlag, New York Inc., 1997.

[3] Campbell N., “CHATR: A high-definition speech re-sequencing system,” *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, Dec. 1996.

[4] Huang X., Acero A., Adcock J., Hon H., Goldsmith J., Liu J., and Plumpe M., “Whistler: A trainable text-to-speech system,” in *Proc. ICSLP*, Philadelphia, PA, pp. 2387–2390, Oct. 1996.

[5] Hunt A. J. and Black A.W., “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. ICASSP*, Atlanta, GA, pp. 373–376, May 1996.

[6] Sagisaka Y., “Speech synthesis by rule using an optimal selection of nonuniform synthesis units,” in *Proc. ICASSP*, New York, NY, pp. 679–682, Apr. 1988.

[7] Jon R.W. Yi.and James R. Glass, “ Natural-sounding speech synthesis using variable-length units”, ICLSP98.

[8] McDowell P.S and Datta S. “ A fractal approach to the characterisation of speech ”, Acoustic Letters, Vol.17, No.1, 1993.

[9] Maragos P. “ Fractal Aspects of Speech Signals: Dimension and Interpolation ”, Proceedings *IEEE* international Conference on Acoustic Speech and Signal Processing (ICASSP), Vol.1, pp417-424, 1991.

[10] Fekkai S. and Al-Akaidi M. “ A Novel approach to measure fractal dimension of speech phonemes ”, Euromedia, Belgium, 1999.

[11] Fekkai S., Al-Akaidi M., “ Word recognition based on fractal techniques ”, Proceeding of international conferences on image, science, system and technology, Las Vegas, USA, pp589-595, 1999.

NETWORK APPLICATIONS

NEW APPROACH FOR ONLINE CASE-BASED REASONING WEB-SERVICES IN KNOWLEDGE MANAGEMENT APPLICATIONS

H.-JOACHIM NERN
Aspasia-Knowledge Systems
D-40210 Duesseldorf, Germany
nern@knixmas.de

TANJA ATANASOVA
Institute of Information Technologies
BAS, Sofia, Bulgaria
t.atanasova@iit.bas.bg

FRIEDBERT PAUTZKE
University of Applied Science
D-44707 Bochum, Germany
friedbert.pautzke@fh-bochum.de

SUSAN JOHNSON
Clicktivities AG
D-40212 Duesseldorf, Germany
susan.johnson@clicktivities.net

KEYWORDS

Case Based Reasoning, Decision Making, Web Services,
Decision Supported Web Service Composing,
Information Retrieval, Knowledge Management

ABSTRACT

The findings presented in this paper are the results of a European project in which the authors participated [1]. Consequently, and on the basis of this former RTD actions, [2, 3, 4] the authors present in this paper an enhancement on a specific type of decision support methodology - a new approach in Case Based Reasoning (CBR) for general application in decision supported information retrieval. The realisation trait of this system is the use of a closed loop approach considering and taking into account statistically evaluated user and experience feedback information implicitly. This closed loop structure assures stable, yet dynamically growing internal case bases and optimal adaptation to the user's systems needs. Furthermore, a description will be given for introducing the CBR methodology as decision- making and inferencing unit in web-service composing techniques.

INTRODUCTION

Due to the fact that, nowadays, new knowledge creation and generation is rapidly accelerating, the development of efficient tools and systems for knowledge management and reasoning is essential. A main application field in the wide area of knowledge management is designated to decision support systems - systems that provide comprehensive support in gathering, organizing, refining and distributing knowledge.

Within a thematic overview the main and significant terms and definitions of CBR are given, whereas in the following part of this paper the different application fields and realisation areas are touched upon [5]. Exemplary, the approach of a CBR-Wrapper [6] composed for product configuration and used in a personal computer sales system will be discussed. In [8] the new term "utility" [9] is introduced in the realm CBR methodology. The authors refer to this introduction in a section and briefly describe the proposed methodology. Alternatively, the authors present a further approach to CBR that is modelled and designed as a closed loop process. As a further, meanwhile more important, application field – web-service technology is discussed. In this context the authors propose and introduce the CBR approach as the fusion of inferencing capabilities and a web service Knowledge Base in service composing processes.

Case Based Reasoning (CBR)

Generally, problem solving related to CBR is based on the main assumption, that "similar problems have similar solutions." Accordingly, the general structure of a CBR-System is illustrated in (Fig. 1), where the relationship between user oriented task resp. problem description and the recommended solution (knowledge object, artefact) is traditionally based on similarity measures given in a so-called "case base". In this case base the problem descriptions and solutions are stored as coupled object pairs.

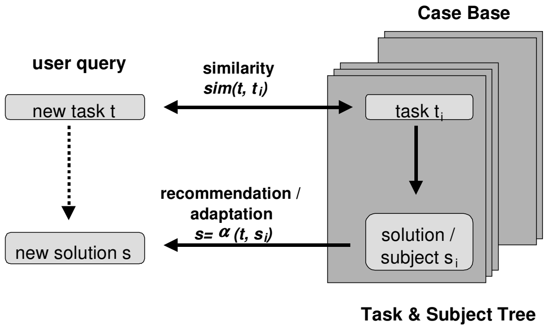


Fig. 1: Traditional CBR model

Hence, cases resp. solutions / subjects are retrieved from the case base using a similarity relationship. In this sense, the decision making process is mainly dependent on the quality of determining a more likely precise similarity measure – some kind of function that assesses this similarity and represents it by a numeric value (in simplest case a similarity coefficient). Generally the case with the highest similarity is selected for recommendation. CBR can be formalised as a four-step process:

Retrieve: Given a target problem, retrieve cases from the case base that are relevant to solving it.

Reuse: Map the solution from the previous case to the target problem, and try to adapt the solution as needed to fit the new one.

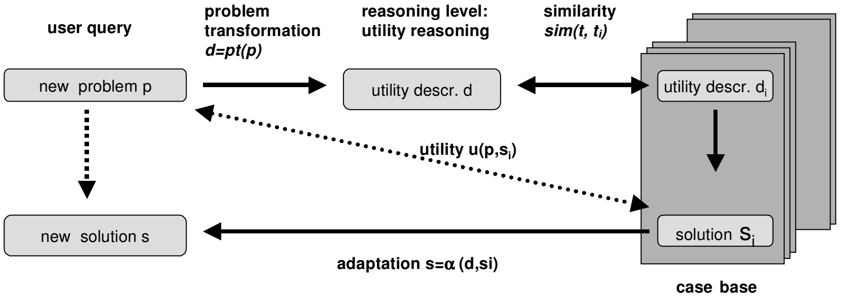


Fig. 2: CBR Model with utility level [8]

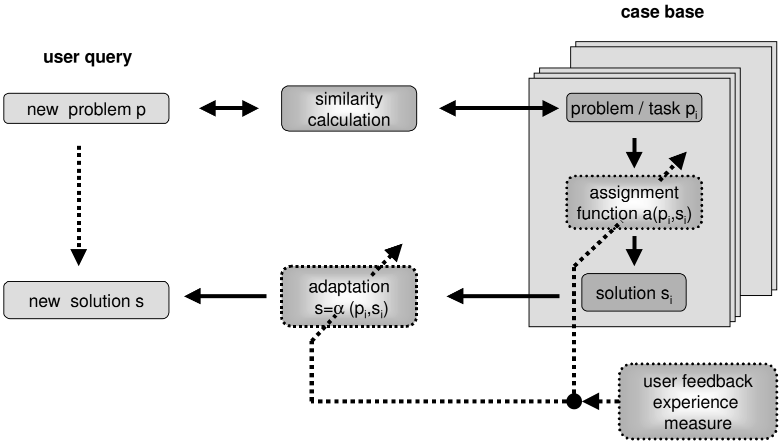


Fig. 3: Approach and enhancement of the CBR model [1, 2, 3, 10, 11]

Revise: After mapping the previous solution to the target one, the new solution should be evaluated – in case of bad functioning revise it.

Retain: After positive adaptation of the solution to the target problem, store the resulting experience as a new case in the case base.

In contrary to rule based decision making procedures the CBR-approach has its advantages in all fields, where the knowledge about the decision making related facts are not clearly and precise describable and/or are elusive.

Application Fields of CBR

The CBR methodology is often implemented in areas, where some kind of “reuse” is appropriate, like electronic design reuse, reuse of SW packages and, component based SW development. Therefore, applications of the CBR approach can be found in different fields and areas such as electronic commerce, planning issues (project management), diagnosis, CBR-supported planning Design - generally spoken, in all areas where reasoning and decision making might be based on existing “similar” solutions and cases. At present one of the most important commercial applications is in e-commerce, for example the handling of distributed information and sales support for customised e-commerce products.

In [6] a framework for product configuration is introduced. It consists of a configurator and a wrapper for the configurator. The application is a specific personal computer sales system within the e-commerce field, whereby the goal of product configuration is to assemble PC products and components by arranging a predefined set of product parts, so that the end product satisfies the user’s requirements by “observing the connectivity of the parts and limitations on the resources necessary for assembling the products” [7]. The CBR wrapper generates a task, namely a problem description, according to the user’s query. The problem description itself consists of a parts description and a requirements description, whereas the wrapper generates only a requirements description using similar cases for a given query. The solver then performs the actual task of product configuration based on the problem description given.

Utility Oriented Matching in CBR

The authors in [8] propose in their work to “overcome” the traditional view of CBR, namely “similar problems have similar solutions”. They want to extend the CBR view to situations in which “we do not have cases that contain pairs of problem descriptions and solutions in the usual sense”. However, in [8] the new term "utility" [9] is introduced in CBR methodology, based on the cognition, that "a similarity measure always tries to approximate some form of utility" [8]. Therefore the authors in [8] proposed a utility-oriented case matching approach by applying a utility reasoning level (Fig. 2). They overcome the assumption that cases consist of a problem description and a solution description – in their approach “only a solution description is required”. From this solution description, they compute the so-called “utility description”, which means that some kind of “utility reasoning” has to be performed. Hereby a “utility function“ is introduced, reflecting “elementary references”. The authors do not provide a solution that considers user feedback – or implicitly internal feedback – it seems that this approach is a pure static approach, because the “utility-function” is determined and loaded by the administrator (expert, maintainer) of the CBR system. This utility oriented - approach to reasoning is solely a static approach and due to this fact, it may be identified as an open-loop system, which is in the sense of cybernetics not dynamic and not capable of handling and managing user influences and experiences and accordingly, does not enable an unsupervised stable growth of the case base.

New Approach of Closed-Loop CBR

As an alternative, the authors of this paper present an approach to CBR modelling designed as a closed-loop process. One of the characteristics of this new approach is to consider and evaluate user feedback and "satisfying" data and applying this data and user information in a closed-loop cycle for dynamically optimising the relationship and assignment functions within the problem/solution pairs.

The procedure, using statistically evaluated feedback data (fitness function), assures implicitly the consistency of the case base (Fig.3) and furthermore optimises the adaptation of known "old" solutions to current problems resp. the generation of new solutions. This presented CBR methodology is discussed and illustrated as an enhancement of a decision support strategy and decision making process already proposed in several papers [1, 2, 3, 10, 11] at the Euromedia conferences 2000-2002.

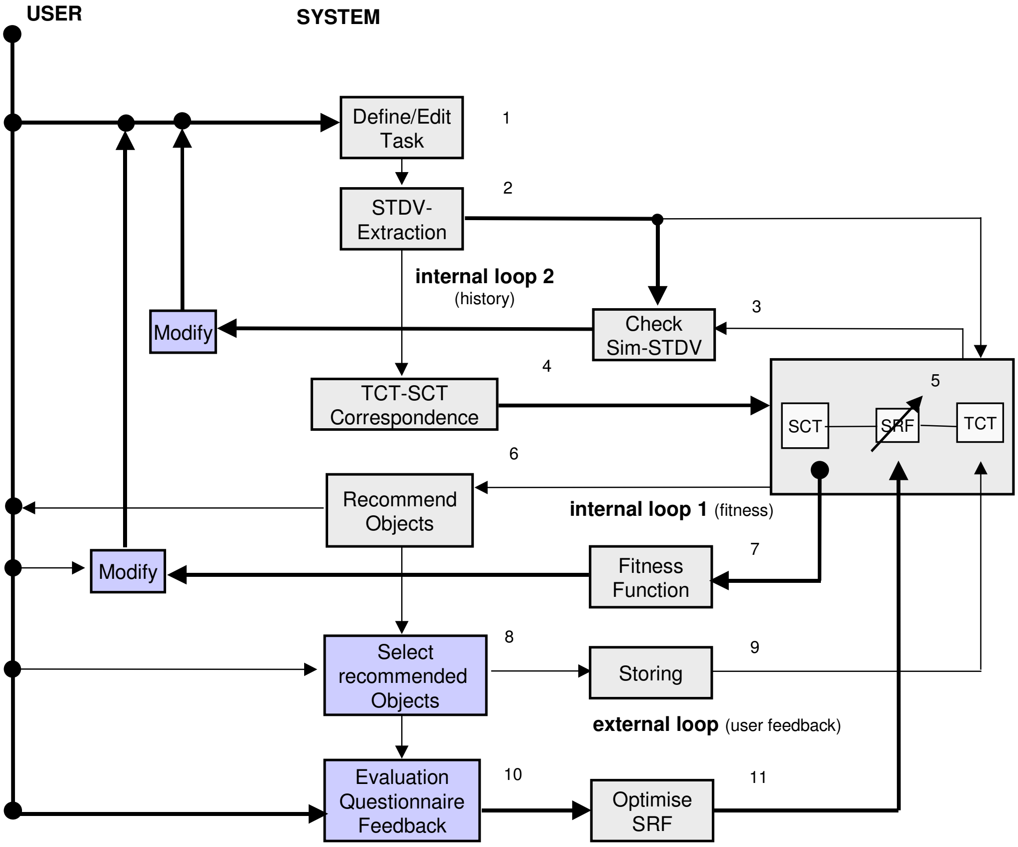
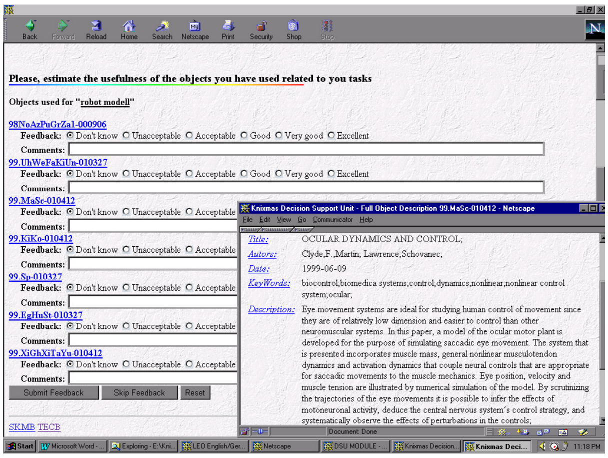


Fig. 4: Modelling of the CBR approach as a Decision Support Unit

The main approach of this CBR modelling (in the following called DSU – Decision Support Unit) consists of a dynamic relationship between subject trees (SCT – the case base, consisting of solutions) and task trees (TCT – the problem base – consisting of problems resp. tasks), whereby the TCT is structured in the same way as the SCT [see 1, 10, 11 for details]. As illustrated in (Fig.3) the correspondence between SCT and TCT is realised as a structured relationship (assignment) function SRF, which implies similarity terms and coefficients, feedback terms, calculated out of the feedback data given by the user and further coefficients, calculated out of the statistical analysis of the system history data such as the accumulated satisfying factor, task counter etc (ASF, TC, SF) [2, 3]. The dynamic growth and expansion of the TCT is referenced to the SCT by the varying relationship (assignment) function SRF - the TCT is organically mapped to the SCT. Each object in the TCT is formalized as a Structured Task Description Vector - STDV. A permanently running module, the relationship module, calculates the relationship function out of the data given within questionnaires, filled out by the user, similarity estimations using the node representatives and further statistical history data calculations

A further main, characteristically significance, feature of the DSU is the use of task related multi-loop feedback procedures. The multi-loop feedback structure ensures stable generation of dynamic knowledge pools embedded in coupled subject trees - and this in relation to given task-profiles of the users. The correspondences between subject (object) trees and user specified task (- profile) trees are adapted (and varied) according to multi-loop feedback evaluations. This guarantees a dynamically growing and implicitly supervised corresponding knowledge trees (the case bases – for problems/tasks and solutions).

Internal feedback loops based on fitness functions evaluate and determine which parts of the user specified task/problem description are fitting well (and in which degree) to the objects recommended by the system. That means that the user is assisted and supported in optimising his task (problem description). This fitness function (loop in Fig. 4, interface in Fig. 5) also gives implicitly advise to the user to avoid generation of inconsistencies and indicates probable contradictions. A further internal loop considers statistical history data to facilitate modifications of the task description by the user.



Please, estimate the usefulness of the objects you have used related to you tasks

Objects used for "robot modell"

98NoAzPuGrZa1-000906
Feedback: ● Don't know ○ Unacceptable ○ Acceptable ○ Good ○ Very good ○ Excellent
Comments:

99.UhWeFaKiUn-010327
Feedback: ● Don't know ○ Unacceptable ○ Acceptable ○ Good ○ Very good ○ Excellent
Comments:

99.MaSc-010412
Feedback: ● Don't know ○ Unacceptable ○ Acceptable
Comments:

99.KiKo-010412
Feedback: ● Don't know ○ Unacceptable ○ Acceptable
Comments:

99.Sp-010327
Feedback: ● Don't know ○ Unacceptable ○ Acceptable
Comments:

99.EgHuSt-010327
Feedback: ● Don't know ○ Unacceptable ○ Acceptable
Comments:

99.XiGhXiTaYu-010412
Feedback: ● Don't know ○ Unacceptable ○ Acceptable
Comments:

Knixmas Decision Support Unit - Full Object Description 99.MaSc-010412 - Netscape

Title: OCULAR DYNAMICS AND CONTROL;
Autors: Clyde,F.,Martin; Lawrence,Schovanec;
Date: 1999-06-09
KeyWords: biocontrol;biomedica systems;control;dynamics;nonlinear;nonlinear control system;ocular;
Description: Eye movement systems are ideal for studying human control of movement since they are of relatively low dimension and easier to control than other neuromuscular systems. In this paper, a model of the ocular motor plant is developed for the purpose of simulating saccadic eye movement. The system that is presented incorporates muscle mass, general nonlinear musculotendon dynamics and activation dynamics that couple neural controls that are appropriate for saccadic movements to the muscle mechanics. Eye position, velocity and muscle tension are illustrated by numerical simulation of the model. By scrutinizing the trajectories of the eye movements it is possible to infer the effects of motoneuronal activity, deduce the central nervous system´s control strategy, and systematically observe the effects of perturbations in the controls;

Document: Done

Fig.5: External feedback: the user evaluates the usefulness of recommended objects with respect to the predefined task resp. problem description

Within the external feedback loop the user evaluates recommended objects with respect to and in relation to the user specified task. The user estimates the usefulness of the objects by filling out a questionnaire and gives feedback evaluations. This evaluation and feedback data are employed implicitly for optimising and dynamically adapting the relationship and assignment function (SRF) between the TCT (Task/Problem Classification Tree) and the SCT (Subject/Solution Classification Tree). This yields optimised correspondences between the task-profiles and the knowledge objects stored and handled in the SCT. In this sense the DSU provides facilities for a knowledge (information) retrieval process which is instantiated and supervised by a decision making process - supporting and assisting the user in determining and classifying his information requirements related to his given task/problem description. The main output of the DSU is to recommend for the given task a set (list) of approaches, methods, documents, software libraries and other relevant knowledge objects that satisfy the given criteria determined by the user – and furthermore give recommendations for modifications of the task description itself

The retrieval procedure is undertaken stepwise in an iterative way, so that the user is guided in optimising his task description and accordingly optimising his information retrieval. The system extracts from the user's - predefined tasks knowledge patterns (-profiles) and accordingly recommends relevant information objects. These objects are automatically rated with respect to their task/problem related relevancy.

According to the illustration given in Fig. 4 the procedure is as follows:

1 - Define/Edit Task: The DSU provides the user with a multi-step request cycle in which the information request is undertaken stepwise in an iterative approach. The user determines in a 10-step cycle his task and provides the system with data concerning: task name, subject, keywords, narrative task description, aim description, related authors, methods, time constraints, exclude subjects, exclude authors, etc.

2 - STDV Extraction: The system extracts from this data a knowledge pattern and generates a user specific task profile, called the Structured Task Description Vector (STDV, see previous paragraph). This profile is stored in the automated generated Task (problem) Classification Tree (TCT) – the problem case base.

3 - Check Sim-STDV: The user specified task is checked, in an internal loop, with respect to already stored and processed history data. If a similar task exists (within a predefined range or threshold) the user accesses the relevant objects and modifies accordingly his task description (internal loop).

4/5 - TCT-SCT-Correspondence and assignment function: The STDV given by the user is checked regarding its similarity and fuzzy matched with the task description tree (TCT). The given structured relationship and assignment function (SRF) establishes a weighted reference to the subject tree.

6 - Recommend Objects: According to the weight of the relationship function a set of knowledge objects is recommended to the user. The objects are rated with a relevancy factor.

7 - Fitness Function: Within a further internal loop an implicit fitness function (relevancy checking) evaluates and determine which parts (components) of the user specified task/problem description (STDV) are fitting well (and to which degree) to the recommended objects. Accordingly the user modifies his task description.

8 - Select Objects: The user selects or discards recommended objects with respect to his given task description.

9 - Storing: The recommended objects are stored, referenced and assigned to the given STDV.

10 - Evaluation-Feedback: The user determines and evaluates the usefulness of recommended objects / solutions with respect to the user's given task description.

11 - Optimise SRF: As an external feedback loop the feedback data are exploited to optimise and adapt the relationship function SRF (as correspondence function) between task and subject trees. The main scientific approach here is that a dependency is established between the object ratings and the predefined task.

Web-Services

The main application target of Web-Services is to establish and enhance the interoperability between different information providing and information demanding entities. An essential attribute of Web-Services should be platform and language independence, which guarantees the useful integration in heterogeneous environments. In this context several languages define standards for service discovery, description and messaging protocols [12, 13, 14]

UDDI – Universal Description, Discovery and, Integration
WSDL – Web Service Description Language
SOAP – Simple Object Access Protocol

These given web-service standards are useful and applicable for generating single and static web-services. In the case of dynamically composing existing web-services, the use of these standards is not sufficient. Accordingly, further enhancing languages have been proposed and developed, which extend the current web technologies with the focus of the semantic web. Due to the lack of handling of dynamic processes in web, especially in semantic web technology, the OWL – the web ontology language has been specified in 2001 [15, 16] by the W3C et al. Therefore, in combination with DAML+OIL (DARPA Agent Mark-up Language plus Ontology Inference Language) & DAML-S (DAML service language) the standards and languages are presented for handling and managing web-services, including under aspects of dynamical composing.

In [17] a concept and some tools are described for the semi-automatic composition of web-services. DAML-S decomposes the semantic description of a web-service into three components: the service profile (input / output types), process model (how the service works) and grounding (details of how an agent can access the service). The service composition prototype of [17] has two basic components: a composer and an inference engine, whereby the composer is the user interface that handles the communication between the human operator and the engine. The inference engine, which is realised in [17] as an OWL reasoner built in Prolog, stores the information about known services in its Knowledge Base.

At this stage, the present authors propose a further new approach and concept: the application of hybrid case based reasoning as a fusion of the Knowledge Base and the inferencing process. The web-service Knowledge Base is designed as a typical case (task and/or problem) base, reflecting the abilities and features of stored web-services. In conjunction with a simple structured rule base for supervising the matching process between existing and new designed web-services the more complicated structure of separate inference engine and Knowledge Base is eliminated. A further advantage is the direct access and the comparison of existing cases (web-services) as the assumption for building new (similar) web-services. In this sense the composing process is decision supported by using the similarity phenomena, explained in the previous sections.

OUTLOOK

A European consortium has been formed to continue the ESPRIT/INCO project KNIXMAS in the current 6th Framework within the IST strategic objective: Semantic Based Systems in Networked Businesses and E-Governments. This project will be based on the scientific results of the KNIXMAS project and furthermore will enhance the former project achievements [18] with modern and contemporary techniques and methods, especially the modification of the CBR approach and the introduction and implementation of web-services using DAML + OIL, OWL, OOA and DAML-S [18]. Within this new project the focus will be on establishing a framework for intelligent web-services created by web service composers accessing a decision supported organisational memory.

References

[1] Nern H,-J. et al;
"Final Report - Knowledge Shared XPS-Based Research Network Using Multi-Agent Systems", European project: ESPRIT HPCN : 977113 KNIXMAS; http://www.cordis.lu/ esprit/src/977113.htm, 2001

[2] Grancharova A., Nern H.-J., et al;
"Decision Support Unit as Part of a Global Knowledge-Shared Research Network", Euromedia 2000, May 8-10, 2000, Antwerp, Belgium

[3] Nern H.-J., Grancharova A., et al;
"Decision Strategies for Rating Objects in Knowledge-Shared Research Networks", Proc. of 4th World Multi-Conference on Circuits, Systems, Communications & Computers (CSCC 2000), Athen, Greece, July 10-15, 2000,

[4] Atanasova T., Nern H.-J., et al;
"Distributed Heterogeneous Knowledge Data Base for Control System Design: Multi-Agent Development and Support", IEEE-ISIC 2000, Int. Symposium on Intelligent Control, Rio, Patras, Greece, July, 2000

[5] Bergmann R., Breen S., Goeker M.;
"Developing Industrial Case Based Reasoning Applications", the INCREA Methodology, Springer 1999

[6] Inakoshi H., Okamoto S., Ohta Y.
"CBR Wrapper: New Framework for Product Configuration", Proc. 1. Conference on Professional Knowledge Management, Baden-Baden, March 2001, Shaker, Aachen, 2001

[7] Sabin, D. Weigel, R.
"Product configuration frameworks – a survey", IEEE Intelligents Systems13(4):39-88, 1992

[8] Bergmann R., Richter M.M., Schmitt S.;
"Utility-Oriented Matching: A New Research Direction for Case-Based Reasoning", Proc. 1. Conference on Professional Knowledge Management, Baden-Baden, March 2001, Shaker, Aachen, 2001

[9] Althoff K.D., Richter M.;
"Similarity and Utility in Non-Numerical Domains", in Mathematische Methoden der Wirtschaftswissenschaften, Physika-Verlag, 1999

[10] Nern H.-J., Rusc T.;
„Filtering and Classification of Meta-Data-Information Objects for Knowledge Sharing within an AI-Based Research Network" , Euromedia 2001, 18-20th April 2001, Valencia, Spain

[11] Atanasova T., Nern H.-J., Pautzke, F.;
"Multi-Agent Approach for Task Related Decision Supported Information Retrieval", Euromedia 2002, April 2002, Modena, Italy

[12] Christensen E., et al
"Web Services Description Language (WSDL)", http:// www.w3.org/TR/2001/NOTE-WSDL-20010315, 1.1, 2000

[13] "The UDDI Technical White Paper",
http://www.uddi.org/ , 2000

[14] "W3C.SOAP 1.2",
http://www.w3.org/TR/2002/CR-soap12-part0-20021219/

[15] T.Berners-Lee, J.Hendler, O.Lassila ,
"The Semantic Web", Scientific American, 284(5), p34-43, May 2001

[16] DAML Services Coalition
" DAML: Web Service Description for the Semantic Web", First International Semantic Web Conference (ISWC), June 2002

[17] Evren Sirin, James Hendler, Bijan Parsia,
"Semi-automatic Composition of Web Services using Semantic Descriptions", "Web Services: Modeling, Architecture and Infrastructure" workshop in conjunction with ICEIS2003, 2003

[18] Atanasova T.; Nern H.-J., Johnson S. et al
"Introducing CBR-Approaches to Web Service Design and Composing", * 7th WSEAS Int.Conf. on COMPUTERS, Corfu 2003, Greece

DESIGNING AND DEVELOPING AN EFFECTIVE RECOMMENDATION SYSTEM FOR ECOMMERCE STORES

Penelope Th. Markellou
Maria I. Rigou
Athanasios K. Tsakalidis
Department of Computer Engineering
and Informatics, Campus,
University of Patras,
26500 Patras, Greece
E-mail: markel@ceid.upatras.gr
rigou@ceid.upatras.gr
tsak@ceid.upatras.gr

Konstantinos Th. Markellos
Spiros P. Sirmakessis
Research Academic Computer
Technology Institute, Internet and
Multimedia Technologies Research Unit,
61 Riga Feraiou str.,
26110 Patras, Greece
E-mail: kmarkel@cti.gr
syrma@cti.gr

KEYWORDS
eCommerce, Personalization, Recommendation.

ABSTRACT

Based on the continuing fundamental changes of eCommerce the last years, customers need tailored e-stores to support them in completing their purchases online. A variety of services have been developed to fulfil this requirement. The most widespread such services apply various techniques that aim to personalize the online shopping experience. Recommendation systems feature as a big category of personalized services. The scope of this paper is to identify the necessary requirements of a recommendation system and describe the architecture and supported functionalities of the implemented e-store that deploys recommendation features in order to foster customer satisfaction and loyalty, and increase sales.

INTRODUCTION

How can someone create a usable eCommerce store? How to recommend products that the customer wants? When does the customer buy? How to ensure that the customer will keep doing business with the store? What is the most important content to customize/personalize? What types of data are needed to stimulate transactions and predict lifetime value? What store features make the customer revisit, repurchase or recommend the e-store or specific products?

These are only few of the questions that we try to answer in this paper. A lot of discussion has been made on the significance and importance of eCommerce. eCommerce is more than just handling purchase transactions and funds transfer over the Internet. It includes the tasks that support buying and selling goods and services, as well as the interactions among them. More specifically, eCommerce embodies the entire business process from promotion, advertisement, marketing, through sales, ordering, distribution, customer support, turnover and market share, to interactions within the company (processes and organization), as well as with external partners (customers and suppliers) (Shaw et al. 2000). However, the majority of eCommerce sites (e-stores) fail to fulfil the customer's needs, desires and requests. The secret of success in eCommerce today is building and sustaining an effective relationship with the customers (EC Enterprise DG 1999).

The goal is too simple. The store should give to the potential customers a compelling reason to buy products or services. Rationally, there are differences between the electronic store and the physical store. An electronic store is not necessarily a substitute for the physical store. In most cases, it works as a complement for the physical store and aims to gain the confidence of the consumers. Obviously, there is no easy answer to the way someone can maintain or even increase customers' e-loyalty (Smith 2000). However, the first key is to understand what the customer wants (Seybold 1988). The second key is to build a close personal relationship. This is not too different from the relationship that a store has in real world, since web site "relationships" work in the same way. Based on this assumption, the e-store should provide the appropriate tools ensuring the creation of lasting relationships between the web site and the customers. This means that the store should use customers' data to develop and maintain an intelligent interaction. So, when the customer gives information, the store should respond with more personalized services, products, and information (Nielsen 2000), (Pearrow 2000) in return.

Personalization and recommendation systems are only two of the tools that may improve the customer's shopping experience by offering advanced services and functionalities (Chen and Chen 2001), (Ohsugi et al. 2002), (Sarwar te al. 2000), (Schafer et al. 1999). The aim is to have customers revealing themselves. The e-store uses this information, builds and maintains customer's profiles about their particular needs, interests, preferences, etc. The web site is then personalized, so that every time they come back it adapts itself to better suit their profile. This behaviour gives customers a friendly feeling every time they come back, increases their loyalty and trust about the e-store and of course improves their satisfaction (Serco Usability Services 2000).

In this paper we address the issue of offering recommendation services and the way we implemented their underlying mechanism within the framework of a research project. More specifically, we examine how we can actively associate a product or products to the customers of an e-store. We explain how the recommendation system, while a customer is examining a product, can suggest a set of other products. We refer to all available methods (e.g. online catalogue, banners, offers, cross-selling and up-selling tactics, etc.) an e-store can deploy in order to offer successful recommendations. Another issue under consideration is the promotion effectiveness of having items featuring on the home page of an e-store.

This work is organized as follows; the next section discusses the requirements of a recommendations system under both the customer and the e-store owner (seller) perspective; the third section describes the supported functionalities of the implemented recommendations system; recommendations system architecture is briefly outlined in the forth section, while the last section concludes with some general thoughts and suggestions on future developments in the area.

REQUIREMENTS

Customers expect to be able to make their online purchases easily and effectively. e-Store owners expect to sell as many products and/or services as possible easily and cost-effectively. The requirements of a recommendation system should be examined from both the perspective of the customer and the store owner.

From the customer's point of view, the recommendation system should ensure that:

- Efficient recommendations about new products, product discounts, special offers, coupons, sweepstakes, etc. are produced and delivered.
- All second-time visitors receive personalized content to meet their needs and that this content is embedded into recommendations about relevant product.
- Browsing through the products of the online catalogue is easy.
- Useful assistance is available during all steps of the shopping process.
- Customers' time is respected and irritation is minimized.
- Customers' individuality is respected.
- Customers are allowed to specify and modify own profile settings.

As regards the store owner point of view, a recommendation system should:

- Promote the products.
- Create customer profiles. This means that the system should capture customer behavioural information from login procedure, completed purchases, response to questionnaires, etc. The data should be updated at predefined time intervals.
- Generate consumers demographics.
- Manage navigation data.
- Analyse sales data.
- Gather statistical data.
- "Remember" and analyse everything the store needs to "know" about the customer.
- Encourage customers to "leave something of themselves behind" by different and "invisible" methods.
- Increase of second-time purchases.

All information about customers is gathered in a specific database called *customer profile database*. A customer profile is where customers 'store' their preferences and where the system notes important things it has learned about them. For example, name, phone number, address, what credit card the customer uses, what kind of products the customer prefers, etc. This database is updated and queried constantly. Customers transaction information is stored into another database called *transaction database*. The system combines the data and generates appropriate and 'relevant' recommendations. Figure 1 depicts the types of information customers supply directly or indirectly to the system databases.

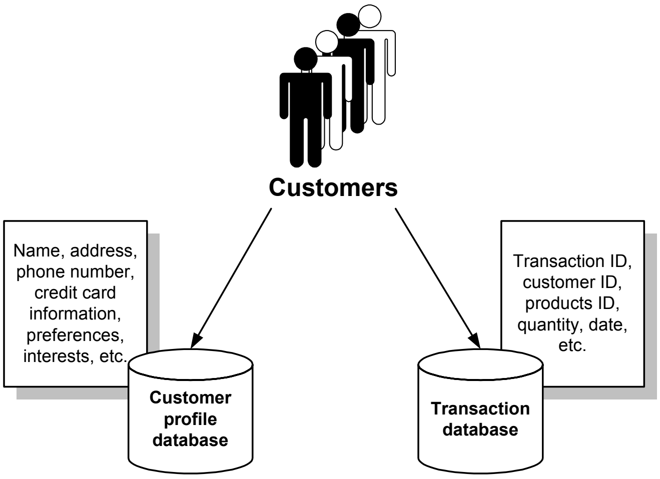

Customers
Name, address, phone number, credit card information, preferences, interests, etc.
Customer profile database
Transaction ID, customer ID, products ID, quantity, date, etc.
Transaction database

Figure 1: Information on Customers

SUPPORTED FUNCTIONALITIES

The ability to suggest to customers to make additional purchases is a powerful eCommerce sales tool. The user should be able to find products of interest in the e-store in various different ways. Figure 2 shows a state diagram describing the behaviour of a customer engaged in selecting products to purchase. Indeed, the state diagram is a pattern since it describes the entire range of different sequences (multiple sequences for each customer). Following, we provide the explanation of every state.

The diagram has one initial and one final state:

- *Start*: a solid circle indicates the initial state where the customer enters the procedure of identifying and selecting a desirable product.
- *Select:* the final state is indicated by a bull's eye, labeled 'select'. The customer reaches this state upon having selected the desirable products.

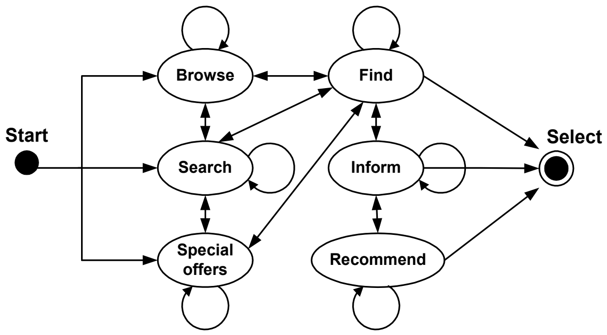


Figure 2: State Diagram of Product Selection

On figure 2 we can identify several ways for the customer of the e-store to browse through the set of available products. The most obvious method is to browse inside the store and when finding the product simply click on it in order to select it. In most cases, the store is structured in departments that are available on the storefront (main page). When the customer selects a department the list of available products is displayed. Another option is to use the search option (it can be in the form of simple or advanced search) to locate a particular product in the online catalogue. Moreover, products can be selected from the special offers section. All the ways to spot products mentioned are depicted in the corresponding states:

- *Browse:* navigation inside the store.
- *Search:* identification of specific products using appropriate keywords.
- *Special offers:* locating products that have a discount.

When the customer finds the product then state 'find' has been reached. There detailed information about the product can be found (state 'inform') or recommendations can be delivered (state 'recommend'). The customer may also jump to a specific item through promotional links, such as those that are supplied by the cross-selling or up-selling options.

- *Find:* the customer has located the desirable product.
- *Inform:* presents in detail all the features of the product including name, price, other characteristics (like number, weight, height, material), photograph, etc.
- *Recommend:* access a specific item through promotional links, such as those that are supplied by the cross-selling or up-selling options.
    - *Cross-selling:* suggests products related to the one(s) the user is currently viewing. In many cases, these are complementary products. For example, proposing a music CD with a book or batteries with toys.
    - *Up-selling:* suggests products perhaps more expensive or advanced to the one(s) the user has chosen. The customer will be informed about products available in the next price level, which he/she may not know about. This tactic depends on the type of products.
- *Other customers' opinions:* suggests additional products that the customer may like, based on what other (probably like-minded) customers believe (since they already purchased them).
- *History data:* analysing the history of purchased data (stored on transaction database), the store is able to offer customers an extremely targeted range of choices that are most likely to fit their profile.

Before deploying either technique it is absolutely vital to ensure that it will be of benefit to the customer and will not interfere with the process of completing the current purchase. Indeed, it is clear that what online shoppers want most is convenience as they are looking for an easy way to acquire goods. Therefore, the recommendation system should be fast, efficient, reliable and discrete.

ARCHITECTURE

As mentioned before, the system saves data about customers' shopping habits, in order to make effective suggestions. This kind of technology offering sales recommendations when the visitor is examining a product, as well as offering the visitor a way of saying "What do you recommend to me?".

The proposed system recommendation technology is based on examining the real-world buying habits of its customers. The problem for implementing this kind of technology is that it needs a lot of orders for creating an enviable base of data on which to run the queries. Another problem is that in many cases some recommendations are often based on assumptions made a priori. Also, statistical information that has not been gathered directly from the customer may lead to inaccurate and misleading recommendations. Finally, people do not usually like the fact that the system collects information about them. So, they avoid supplying the system with personal information.

Figure 3 depicts the architecture of the proposed recommendation system.

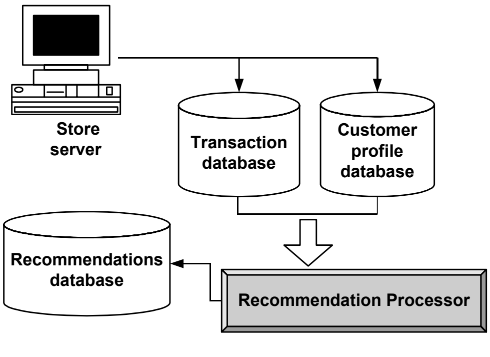


Figure 3: Recommendation System Architecture

Appropriate data is supplied to the recommendation processor by both the customer profile and the transaction databases. The recommendation processor aggregates all the orders ever placed for each individual product of the online catalogue. The results are ordered and recorded to the recommendation database.

The first stage in the recommendation process is to take a snapshot of the orders currently recorded in the system and pass them to the recommendation processor. The processor also has a snapshot of the current version of the online products catalogue. The recommendation process is then triggered in order to apply the cross-selling tactic:

1. First, it scans through the transaction database for orders containing the selected product.
2. Once all orders are processed the system has sorted the products according to their popularity. Therefore, it has estimated the most common other product that is bought by customers together with the selected one.
3. At the end of this process, the system has dynamically produced a set of recommended items based on the real-world ordering habits of its customers.

To maximize data gathering opportunities, the system collects data from every customer touch point, *online* and *offline*.

Online customer touch points include:

- Registration: the store asks some basic information about the customer (e.g. name, address, phone number, fax, interests, preferences, etc.), including the e-mail address and the password. Being a registered user makes future purchases faster, easier and friendlier.
- Transactions: purchase data or request for information.
- Sign-ups: newsletters, email notifications, samples, coupons, partner offers, etc.
- Customer profiles or user preferences.
- Customer surveys: research-related and entertaining content surveys.
- Customer service.
- Web log files: pages viewed, categories searched, links clicked, etc.
- Incoming and outgoing URLs (URL linking to the store, and links leading outside the store).
- Advertising banners.
- Sweepstakes and other promotions requiring customer data.

Offline customer touch points comprise:

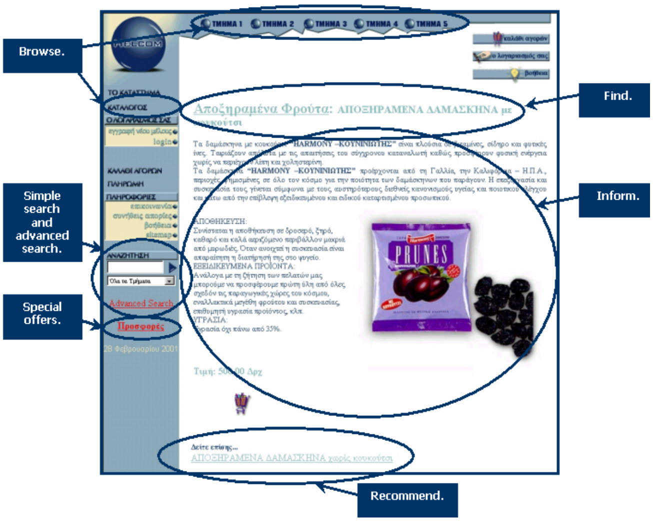
- Customer service by phone, stored in the customer profile database.
- In-store transactions (meaning physical store purchases).
- Various surveys.
- Paper submissions (e.g. sweepstake or promotion entries).

The most important data collection source is the initial registration. In most cases this registration process is more important than the first transaction, in that the act of registering indicates that a customer wants to start a 'conversation' or a relationship and that he/she is giving the store permission to begin this process. When adequate data is collected subsequent interactions with the store may well exceed the visitor's expectations. Ensuring that the store allows customers to update and modify own profile data not only will keep the customer information up to date but shall also engender more trust because customers know what information is maintained about them by the e-store.

Another equally effective way to gather data about the customer is when the system does not explicitly ask for any information at all; many successful web sites use cookies and unique identifiers to make collecting customer-specific data invisible to the customer.

It is worth mentioning that the front page of the e-store is a good place to put items that the store wants to promote as it is a place where user attention is by default drawn. This page should be constantly updated to keep people coming back. A good e-store design should provide an option (in the administrative page) that allows the store to set whether a product is to be a featured product appearing on the homepage or not.

Finally, figure 4 presents the implemented recommendation system as a component of a business-to-consumer (b2c) eCommerce application where the focus is placed on the different ways to select products (as presented in the state diagram of figure 2).



Αποξηραμένα Φρούτα: ΑΠΟΞΗΡΑΜΕΝΑ ΔΑΜΑΣΚΗΝΑ με κουκούτσι

Τα δαμάσκηνα με κουκούτσι "HARMONY –ΚΟΥΝΙΝΙΩΤΗΣ" είναι πλούσια σε βιταμίνες, σίδηρο και φυτικές ίνες. Ταιριάζουν απόλυτα με τις απαιτήσεις του σύγχρονου καταναλωτή καθώς προσφέρουν φυσική ενέργεια χωρίς να περιέχουν λίπη και χοληστερίνη.
Τα δαμάσκηνα "HARMONY –ΚΟΥΝΙΝΙΩΤΗΣ" προέρχονται από τη Γαλλία, την Καλιφόρνια – Η.Π.Α., περιοχές φημισμένες σε όλο τον κόσμο για την ποιότητα των δαμάσκηνων που παράγουν. Η επεξεργασία και συσκευασία τους γίνεται σύμφωνα με τους αυστηρότερους διεθνείς κανονισμούς υγιεινής και ποιοτικού ελέγχου και κάτω από την επίβλεψη εξειδικευμένου και ειδικού καταρτισμένου προσωπικού.

ΑΠΟΘΗΚΕΥΣΗ:
Συνίσταται η αποθήκευση σε δροσερό, ξηρό, καθαρό και καλά αεριζόμενο περιβάλλον μακριά από μυρωδιές. Όταν ανοιχτεί η συσκευασία είναι απαραίτητη η διατήρηση της στο ψυγείο.
ΕΞΕΙΔΙΚΕΥΜΕΝΑ ΠΡΟΪΟΝΤΑ:
Ανάλογα με τη ζήτηση των πελατών μας μπορούμε να προσφέρουμε πρώτη ύλη από όλες σχεδόν τις παραγωγικές χώρες του κόσμου, εναλλακτικά μεγέθη φρούτου και συσκευασίας, επιθυμητή υγρασία προϊόντος, κλπ.
ΥΓΡΑΣΙΑ:
Υγρασία όχι πάνω από 35%.

Τιμή: 500,00 Δρχ

Δείτε επίσης...
ΑΠΟΞΗΡΑΜΕΝΑ ΔΑΜΑΣΚΗΝΑ χωρίς κουκούτσι

Figure 4: Example of Application for the Selection of a Product

CONCLUSIONS

We started off this paper by discussing the importance of a recommendation component in an eCommerce application promoting additional relevant items while the customer is shopping around the e-store. Besides, effective recommendations are a valuable service to the customer and a profitable service to the seller. We then saw how we can build such a system, its architecture and the way it stores and then analyses transaction data and customers' habits and preferences in order to generate successful and accurate product recommendations.

For achieving better results out of the recommendation process, the company's marketing department should be involved in the design and delivery of the service. Equally necessary is the continuous evaluation of the produced recommendations, bringing us to the natural question "when is a recommendation system successful?". A quick but rather too strict answer could be that "it is successful when the customer buys every product that the recommender suggests". But this is not realistic though since customers do not necessarily buy the suggested product even though it was interesting.

There remains lot of work for defining the metrics for measuring the performance of our recommendation system. The evaluation basic directions include assessment of a) system's ability to "build" customers' models that correctly reflect their preferences and b) the "value" of the produced recommendations. This will be part of our future work.

REFERENCES

Chen, H. and A. Chen. 2001. "A Music Recommendation System Based on Music Data Grouping and User Interests", In *Proceedings of the 10th International Conference on Information and Knowledge Management, CICM'01* (Atlanta, Georgia, USA, Nov. 5-10), 231-238.

European Commission Enterprise DG. 1999. *Best Business Web Sites October 1999*.

Nielsen, J. 2000. *Designing Web Usability*, New Riders Publishing, Indianapolis, Indiana USA, 100-160.

Ohsugi, N.; A. Monden; and K. Matsumoto. 2002. "A Recommendation System for Software Function Discovery", In *Proceedings of the 9th Asia-Pacific Software Engineering Conference, APSEC20002* (Gold Coast, Queensland, Australia, Dec. 4-6), 248-257.

Pearrow, M. 2000. *Web Site Usability Handbook*, Charles River Media, Rockland, Massachusetts, USA, 295-315.

Sarwar, B.M.; G. Karypis; J.A. Konstan; and J. Riedl. 2000. "Analysis of Recommendation Algorithms for E-Commerce", In *Proceedings of the 2000 ACM Conference on Electronic Commerce, EC'00* (Minneapolis, MN, USA, Oct. 17-20), 158-167.

Schafer, J.B.; J.A. Konstan; and J. Riedl. 1999. "Recommender Systems in E-Commerce", In *Proceedings of the ACM 1999 Conference on Electronic Commerce, EC'99* (Denver, Colorado, USA, Nov. 3-5), 158-166.

Serco Usability Services. 2000. "How to Design a Customer-Friendly On-Line Store: Usability Guidelines", Serco Ltd, available at http://www.usability.serco.com/research/susshopguide.pdf.

Seybold, P.B. 1988. *customers.com: How to Create a Profitable Business Strategy for the Internet and Beyond*. Crown Business, New York, USA, 52-62.

Shaw, M.; R. Blanning; T. Strader; and A. Whinston (Eds.). 2000. *Handbook on Electronic Commerce*. Springer-Verlag, Berlin, Heidelberg, Germany, 3-24.

Smith, E.R. 2000. *e-loyalty: How to Keep Customers Coming Back to Your Website*. HaperBusiness, New York, USA, 3-27.

AUTHORS' BIOGRAPHIES

KONSTANTINOS MARKELLOS has an MSc in Computer Science, in the area of Integrated Software and Hardware Systems. He is currently working in the Internet and Multimedia Technologies Research Unit of the Research Academic Computer Technology Institute and as a researcher in the Computer Engineering and Informatics Department of the University of Patras. His research areas include e-business and e-learning systems, personalization and web mining techniques. He has published several research papers and is co-author of the book "eBusiness" (available in Greek). eMail: kmarkel@cti.gr.

PENELOPE MARKELLOU has a MSc in Computer Science, in the area of designing and evaluating e-business systems. She is working in the Computer Engineering and Informatics Department of the University of Patras and as a researcher in the Internet and Multimedia Technologies Research Unit of the Research Academic Computer Technology Institute. Her research areas include personalization techniques in e-business and e-learning systems, text and web mining. She has published several research papers and is co-author of the books "Multimedia and Networks", "Usability Models for eCommerce Applications" and "eBusiness" (available in Greek). eMail: markel@ceid.upatras.gr.

MARIA RIGOU has a MSc in Computer Science, in the area of evaluating interactive systems. She is working at the Computer Engineering and Informatics Department of the University of Patras and as a researcher of the Internet and Multimedia Technologies Research Unit of the Research Academic Computer Technology Institute. She has worked on personalization techniques and use of web usage mining in adapting content and structure. She has several publications in the area of web mining, e-learning and the use of user communities in the learning process. eMail: rigou@ceid.upatras.gr.

SPIROS SIRMAKESSIS is an Adjunct Assistant Professor in the Computer Engineering and Informatics Department of the University of Patras, Greece. He is also the Manager of the Internet and Multimedia Technologies Research Unit of the Research Academic Computer Technology Institute (http://www.cti.gr). He is the coordinator of NEMIS, a network of excellence on text mining, funded by the European Community, and responsible for the web mining research areas of the network. He is author of three books and several research papers published in international journals and conferences. eMail: syrma@cti.gr.

ATHANASIOS TSAKALIDIS is a Computer-Scientist, Professor of the University of Patras. Born 27.6.1950 in Katerini, Greece. Studies: Diploma of Mathematics, University of Thessaloniki (1973), Diploma of Informatics (1980) and Ph.D in Informatics (1983), University of Saarland, Germany. Career: 1983-1989, researcher in the University of Saarland. 1989-1993, Associate Professor and since 1993 Professor in the Department of Computer Engineering and Informatics, University of Patras. 1993-1997 and 2001-today Chairman of the same Department. 1993-today, member of the Board of Directors of the Computer Technology Institute (CTI) and 1997-today, Coordinator of Research & Development. He is one of the contributors of the "Handbook of Theoretical Computer Science" (Elsevier and MIT-Press 1990). He has published many scientific articles, having an especial contribution to the solution of elementary problems in the area of data structures. Scientific interests: data structures, computational geometry, information retrieval, computer graphics, databases, bio-informatics. eMail: tsak@cti.gr, URL: http://www.tsakalidis.gr.

Database Design for Multiuser Collaboration in VRCASE Tool

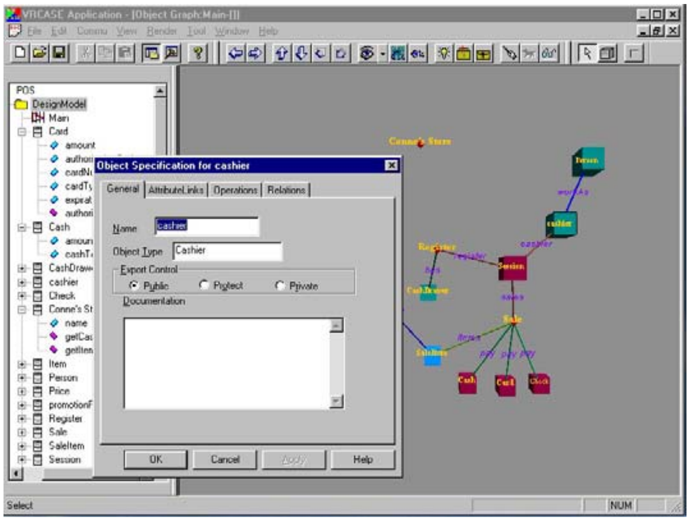
Qingping Lin, Jim Mee Ng, Chor Ping Low, Juan Bu, Xiaohua Liu
Information Communication Institute of Singapore
School of Electrical and Electronic Engineering
Naynang Technological University
Singapore 639798
Email: iqplin@ntu.edu.sg

Abstract

VRCASE is a CASE tool incorporating form-Object-to-Class based approach with collaborative virtual environment. It provides a 3D multi-user software modeling environment with automatic object-class abstraction, class diagram generation, and C++ skeleton generation facilities for assisting Object-Oriented software development. In this paper, we present the database design for achieving efficient collaborative software development in VRCASE. An overview of the system architecture of the multi-user collaborative support in VRCASE will be introduced before discussing the concurrency control consideration in design of the database system, followed by detailed implementation of the database. Our experiment results show that our database design can support efficient collaborative software design with good system responsiveness.

1. Introduction

Current Computer Aided Software Engineering (CASE) tools rely on software engineers' experience to abstract class relationships from the software objects identified at requirement capturing stage. The formidable technical notations of analysis and design result also make it hard for the end-user to comprehend and evaluate. To overcome the identified problem, we have developed VRCASE – a CASE tool incorporating Object-to-Class based approach with collaborative virtual environment. It provides a 3D multi-user software-modeling environment with automatic object-class abstraction facility for assisting Object-Oriented software development. It automatically generates UML (Unified Modeling Language) class diagram and corresponding C++ skeleton codes based on software objects and their relationships modeled in the 3D virtual environment. The system is designed based on client/Server architecture. There are two subsystems in this architecture: client and server. Client and server communicate via TCP/IP over the Internet. The server is the core part of the whole system that keeps track of the system state information and acts as a router transmitting the events of user interactions. It maintains all of the data of the VRCASE system, manages the communication among clients and guarantees the consistency of the whole system. All communication between users goes via server. The client side undertakes most of the users' software modeling and design work. It provides a 3D editor to enable users to visualize their software development in a VR environment. Project managers can add or delete projects, recover database, monitor system activities and project information etc. Actually, the server graphic user interface acts as the system administrator tool in VRCASE. Once the software modeling is completed in VRCASE, users can use its Object-Class Abstraction and Code Generation facilities to generate UML class diagram and corresponding C++ skeleton codes. Sample screen capture of client GUI and class diagram generated by VRCASE are illustrated in Figure 1 and Figure 2. Detailed algorithms and implementations of VRCASE can be found in reference [1]. This paper only focuses on database design aspect of VRCASE. The following sections will discuss the concurrency control consideration in Section 2, Database design and implementation in Section 3, system responsiveness experiment results in Section 4, and summary in Section 5.



Name: cashier
Object Type: Cashier
Export Control: ● Public ○ Protect ○ Private

Figure 1 VRCASE Graphic User Interface

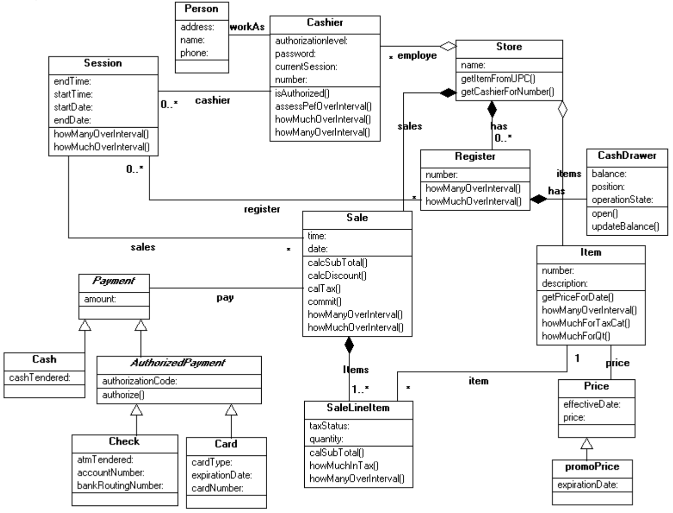


Figure 2 Class diagram generated from VRCASE.

2. Concurrency Control Consideration

To support concurrent engineering in a multi-user collaborative software development environment, concurrency control is the core issue. Traditional concurrency control techniques for database systems have been successful in many multi-user setting, but these techniques are inadequate in open, extensible and Computer Supported Co-operative Work (CSCW) systems supporting collaborating users. Transactions in collaborative applications have several characteristics that distinguish them from transactions in database systems requiring concurrency control. Collaborative applications are typically interactive, for example, they may be of long duration; they are often unstructured and more sensitive to response time performance than total system throughput. Thus, concurrency control's role of preventing inconsistent concurrent actions must compete with the need of users to share their partial results with each other. Hence, collaborative applications have their own peculiar concurrency control requirements such as support for long transactions, user control over transaction execution, relaxed consistency criteria, and integration with access control.

Though database systems support multiple users, however they are not collaborative since they provide little notification. Because their transactions are usually fixed programs, they could not respond to interface if it did happen. If one user performs some action, other users are not normally notified of the action and may only learn of it by explicitly querying the system.

Traditional database concurrency control is generally restrictive for collaborative applications. The semantics of the shared data object of collaborative application is usually much more complex than the read/write semantics of the database data models. As a result, the concurrency control it supports is more conservative than necessary. A database using read/write semantics will not allow concurrent insertions but will block one to let the other proceed. Traditional database systems do not allow concurrent transactions to mutually depend on each other, whereas in collaborative systems users whose workspaces are coupled may be expected to influence each other in mid-transaction. Before adding a new item, a user would observe the actions of the other user to ensure that duplicate items are not inserted in the database. In collaborative systems users may wish to temporarily allow conflicting actions and delay their resolution until some later time. Conventional database systems, however, do not allow the database to remain in an inconsistent state for indefinite periods. The action that a conventional database system will take when conflict is identified is to throw away all work that led to the conflict and return the database to a prior consistent state. For a user about to commit a large number of changes to a document, this would not only be very un-pleasant, but is probably unnecessary; the user may only need to discard changes to one paragraph [2].

The early efforts on CASE addressed mainly form and representation issues of software development methodology. Program support tools such as translator, compilers, assemblers, linkers and loaders were developed. Later the range of support tools began to expand with the development of program editors, debuggers, code analyzers etc. Large-scale software development demanded enhanced support of the entire software development process from CASE tool developers. Assistance was required for the requirements definition, design and implementation phases of the software development lifecycle, testing, documentation and version control [3]. Thus, the focus of CASE moved from implementation phase to cover the entire software developing process.

There exist a wide range of CASE tools to support different methods and different stage of the software development lifecycle [4, 5]. CASE tools make the software architecture and design clearer and easier to be understood and modified because users are enabled to abstract away from the entanglement of the source code. The graphic notation makes the communication between the software development teams easier and the skeleton code generation motivates development teams to construct a good design before coding. Apart from helping software engineers to use some particular method of application development, CASE tools also play an important role in software engineering methodology research. As each new method is devised, the provision of CASE tools that specifically support that method is a valuable way of promoting the new method within the software engineering community. Some CASE tools even support roundtrip engineering with the ability of keeping model and code consistent and the ability of deriving class diagrams or database models from source code [6]. Extensive possibilities for maintaining models and producing documentation are also provided.

CASE tools [7, 8] provide automated assistance for software development. One facet of modern CASE tools that is significantly underdeveloped is support for group-based development work. Although most modern software development takes place as a group process, tools designed to assist in that development are more oriented to a single user working in isolation. [9] Traditional CASE tools fail to improve development or do not fully support the cooperative nature of software development in their framework. The full potential of CASE technology can only be utilized if the technology supports both distributed development tasks and the coordination of these tasks [10,11]. This can be obtained by augmenting the CASE tools with CSCW facilities. The combination of CASE tools with CSCW principles serves to create a tool much more reflective of the reality of the multi-user collaborative software development process. Consequently, the acceptance and usefulness of CASE tools should be significantly improved. However, literature review on this area shows that no mechanism can cater for all kinds of concurrency control in collaborative group work. Most of existing multi-user collaborative CASE tools adopt the object-level locking mechanism[12, 13,14], file-level locking mechanism [15,16] and the floor control mechanism[17] for concurrency control. Some tools constrain the edit right only to one person (for example, the object owner) strictly to reduce the opportunities of concurrent conflict [14]. All of these mechanisms cannot provide a fine-grained sharing of system data and tend to cause long blocking.

In VRCASE, we propose a new concurrency control strategy that is synthesized from mechanisms used in traditional database management and approaches used for concurrency control in CSCW systems. Fine-grained locking and notification, and Visual Indicator mechanism are proposed to work together as the concurrency control mechanism in VRCASE system. VRCASE provides contention resolution at the level of attributes in nodes or links. In effect, VRCASE reduce concurrent conflicts and notifications from the traditional object level to attribute level. The system allows any number of users simultaneously to read and display the data field of a given node in a separate window on the screen. However, permissions to make modifications to the data field of a particular node are restricted to one user at a time. Detailed description of the fine-grained locking and notification, and Visual Indicator mechanism can be found in reference [18].

3. VRCASE Database Design and Implementation

3.1 VRCASE Server Software Architecture

VRCASE server software has been implemented to meet the design goal of multi-user VRCASE that supports collaborative design of OO software. Based on the server's functional requirements, the architecture of VRCASE server application is established as shown in Figure 3 :

- Central Database works as a data source to store all of the system data and project information.
- ClientThread processes all system messages corresponding to various events in the Message Processor. Message Translator dissembles received messages and assembles out-going messages. Lock Manager manages locks in the concurrency control. Message Reader, Message Buffer, Message Sender and Client Socket take active roles in system communication and work as parts of Communication Manager.
- SchedulerThread controls all of the ClientThreads through ClientThread manager and backs up system data through DataBackUp Manager. Listening Socket and Socket Acceptor act as parts of the Communication Manager in system communication.
- The Communication Manager is the interface to underlying network. The main task of the Communication Manger is to maintain two-way communication between the server and the clients.

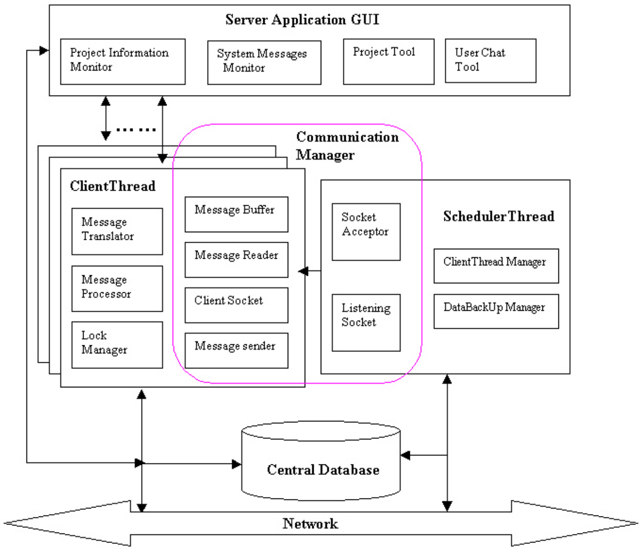


Figure 3 Structure of VRCASE Server Software

- Sever Application GUI is used to construct the user interface for the administrator. It consists of a Project Information Monitor that retrieves and displays all project information, a System Messages Monitor that monitors and displays all events sent to server, a project tool that helps the administrator to manage projects and a user chat tool that helps project team members to communicate with each other in an explicit way.

3.2 Database Design and Implementation

In the development of VRCASE server application, information concerning with the VRCASE record is stored in a Microsoft SQL server 6.5 based relational database server that employs the Microsoft Open Database Connectivity (ODBC) protocols. Figure 4 shows the structure of VRCASE server using SQL server through ODBC. The VRCASE Server Application requests a data source to be opened by specifying a data source name. The data source name provides the next layer, the driver manager, with enough information to load the appropriate ODBC driver and initialize it for a connection. The driver manager passes on any subsequent calls to the ODBC API, from the application to the driver. The driver actually implements the call, doing the work required to retrieve, accept data and other operations.

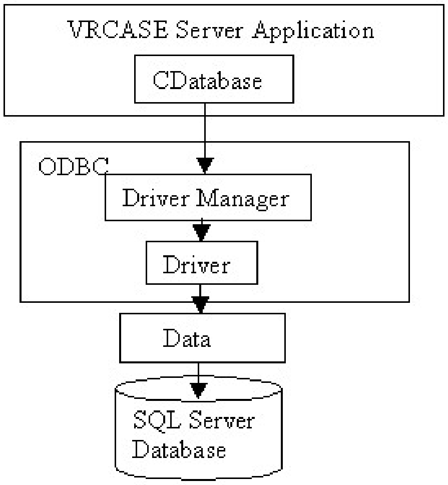


Figure 4 Structure of the VRCASE using SQL server through ODBC

3.2.1 Central Database

The Central Database is crucial in making the entire system work logically and consistently. It provides the organization structures of the system data as well as a full service for recording local and remote users' work areas. The Central Database organizes both textual and graphical information according to their structural relationship. There are two parts in the Central Database: one is the Information Database that is used to record project and user information and the other one is the System Non-Graph Database that is used to record system data.

3.2.1.1 Information Database

Information Database includes Project Information table and User Information table. Project Information table records all of the project information such as ProjectID, ProjectName and predefined Password etc. It is initialized and managed by project manager. Project manager initializes the project name, predefined password and username list through the administrator interface of server application. Server grants a user the project access right and loads all initial data to the user only if he can provide valid user name and password. All users who will work in the same project collaboratively are recorded in a name list in the project information table. The structure of Project Information table is shown in Table 1.

Table 1 Data Structure of Project Information table

Field Name	**Description**
ID	The project ID
Name	The project name
Password	The predefined project password
DBName	The name of database which contains this project data
ShareList	A name list that records all names of predefined users who are working in the same project.

User Information table records the user information such as user name, password and project name. User can change his password periodically. The newest password is recorded in this table. Table 2 shows the data structure of User Interface table.

Table 2 Data Structure of User Information table

Field Name	**Description**
UserName	The name of the user
Password	User defined password
ProjectName	The name of the project in which the user joins

3.2.1.2 System Database

System Database is used as the whole system data container. Every project has its own database that includes multiple tables, which are used to record different data structure of VRCASE system data. Figure 5 shows the overview of the structure of the system database. System Database includes two kinds of databases: Graph database and Non-Graph Database.

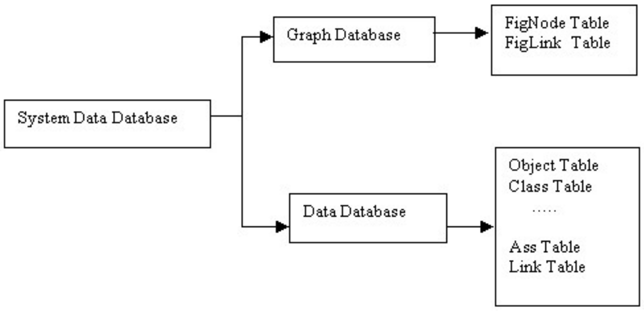


Figure 5 System Database Structure

(1) Graph Database

The Graph Database is used to record the 3D graphic model data. FigNode and FigLink are defined to represent 3D-node model and the relations between nodes respectively. Table 3 shows the structure of the FigNode. The position of 3D graphic object in the 3D display is described using x, y, z coordinates. The relationships need to be stored in the Graphic Database are described as following:

(a) The relationship between the graphic nodes and the project that contains them. *ParentID* describe the ID of the project.
(b) The relationship between graphic model and software model. *RelatedElementID* specifies which object the FigNode represents.

(c) The relationship among FigNodes. Table 4 shows the data structure of FigLink. The *SrcLinkEndID* and *DesLinkID* specify the source and destination 3D FigNodes of the FigLink, thus the relationships between the 3D nodes are described.

(d) The enclosing relationship between 3D FigNodes. The enclosing relationship is described by the *EnclosingID* field that specifies the parent which encloses the 3D node (Table 3).

Table 3 Data structure of FigNode table

Field Name	Description
FigNodeID	The Identification of FigNode
ParentID	The ID of project that contains the FigNode.
RelatedElementID	The ID of object that the FigNode represents
Posx	The x coordinate of the FigNode position
Posy	The y coordinate of the FigNode position
Posz	The z coordinate of the FigNode position
Lock	Specify the lock state
Property	Specify the detailed data of the FigNode
Owner	The name of the users who create the FigNode
ShareList	A name list that records all of the names of users who have the edit right to the FigNode.
EnclosingID	The ID of the parent that encloses the FigNode

Table 4 Data Structure of FigLink table

Field Name	Description
FigLinkID	The identification of FigLink.
ParentID	The ID of project that contains the FigLink
RelatedElementID	The ID of object that the FigLink represents
Lock	Specify the lock state
Property	Specify the detailed data of the FigLink
Owner	The name of the user who creates the FigLink
ShareList	A name list that records all of the names of users who have the edit right to the FigLink.
EnclosingID	The ID of the parent that encloses the FigLink
SrcFigNodeID	The ID of the source FigNode from which the Figlink begins.
DesFigNodeID	The ID of the Destination FigNode to which the Figlink ends.

(2) Non-Graph Database

Non-Graph Database is used to record non-graphic type of system data. VRCASE classifies the system data in several different data types and different data types have different table structures. The relationships stored in the Non-Graph Database include:

1) The relationship between the system element and the parent element or project. *ParentID* describes the parent ID.
2) The relationship between the class and its instance. For example, Table 5 shows the structure of the Object table. *ClassID* describe the class from which the object is instantiated.
3) The relationship between Objects. Objects relate to each other via Links. Figure 6 shows the link relationship between two objects. As shown in Link Table 6, *SrcLinkEndID* and *DesLinkEndID* describe the source and destination of the Link respectively. In the LinkEnd table (Table 7), the *OwnerObjectID* and *ConnectLinkID* present the connected Object and Link respectively. Thus the relationship between Link and LinkEnd is also presented. Figure 7 shows the retrieval procedures.
4)

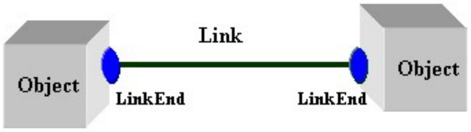


Figure 6 Link Between Objects

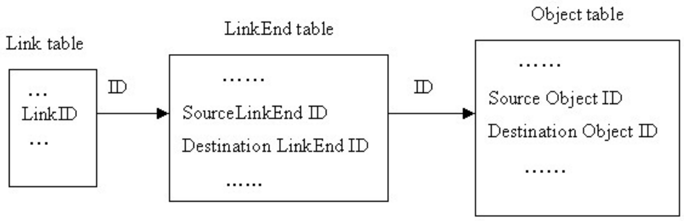


Figure 7 Relationship Retrieval procedures

Table 5 Data structure of Object table

Field Name	Description
ObjectID	The Identification of Object.
ParentID	The ID of project that contains the Object
ClassID	The ID of class from which the object is instantiated
Lock	Specify the lock state
Owner	The name of the users who create the Object
ShareList	A name list that records all of the names of users who have the edit right to the object.
Property	Specify the detailed data of the object

Table 6 Data structure of the Link table

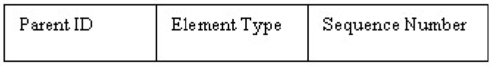
Field Name	Description
LinkID	The Identification of Link.
Property	Specify the detailed data of the Link.
Lock	Specify the lock state
Owner	The name of the users who create the Link
ShareList	A name list that records all of the names of users who have the edit right to the Link.
ParentID	The ID of project that contains the Link
AssociationID	The ID of class from which the Link is instantiated.
SrcLinkEndID	The ID of the source LinkEnd from which the Link begins.
DesLinkEndID	The ID of the Destination LinkEnd to which the Link ends.

Table 7 Data Structure of the LinkEnd table

LinkEnd ID	The Identification of LinkEnd.
ParentID	The ID of project that contains the LinkEnd
Property	Specify the detailed data of the LinkEnd
AssociationEndID	The ID of class from which the LinkEnd is instantiated.
OwnerObjectID	The ID of the object which contains the LinkEnd
ConnectLinkID	The ID of the Link with which the LinkEnd connects.

4.3.3 Data Retrieval

VRCASE system uses ID as the global index. The server grants an ID to system data item and manages the ID numbers to work as a global index to retrieve data in the central database. As shown in Figure 8 , each element ID includes three parts: ParentID that indicates the parent of the new item, ElementType that indicates the data type of the new item and SequenceNumber. Whenever the user creates a new system element, the client software will first send the new data to the server, and the server will then grant a corresponding ID to this new element through ID Generator. ID Generator is responsible for generating the ID for a new system item.



Parent ID	Element Type	Sequence Number

Figure 8 ID format

Figure 9 shows a rough ID Index Tree that is used as the path to retrieve VRCASE data using the ID number. Server retrieves the data directly through the ID index tree. Client applications also use the ID to identify system elements.

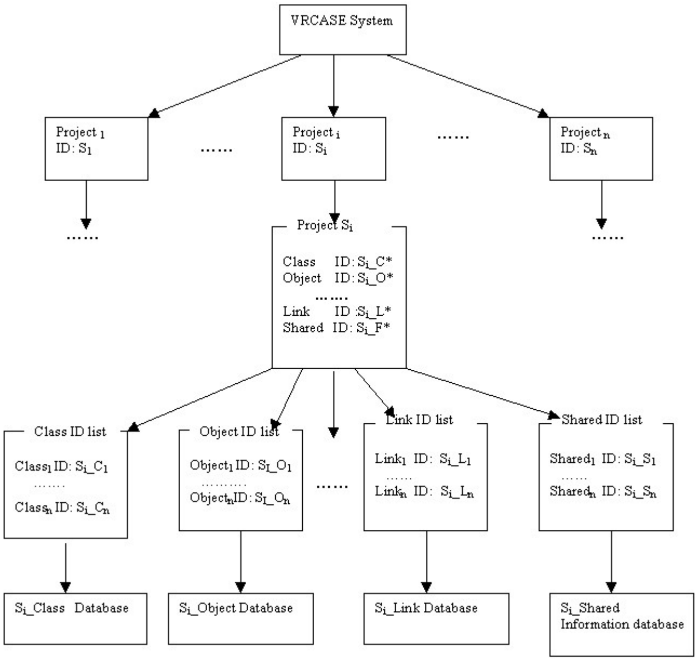


Figure 9 ID Index Tree

4 System Responsiveness

To evaluate the system responsiveness, we tested the VRCASE in LAN (Local Area Network) environment and measured the time lag between user action upon the VRCASE client interface (e.g. change on a object property, or add an object) and the corresponding system information update at the client end. As shown in Figure 10, the time latency increases approximately linearly with the size of the group. *L1* shows the time lag when clients work in the same project with system data shared by all clients and *L2* is the result when clients work in two projects. The time latency varies within 135-155msec and latency differences varies within 20msec. The results show that the system responsiveness is sufficient for collaborative work in VRCASE. Although our experiment was conducted in a LAN environment, it should also work well over the Internet since the size of the data packet for maintaining the system consistency is very small (about 61 bytes per message in our approach).

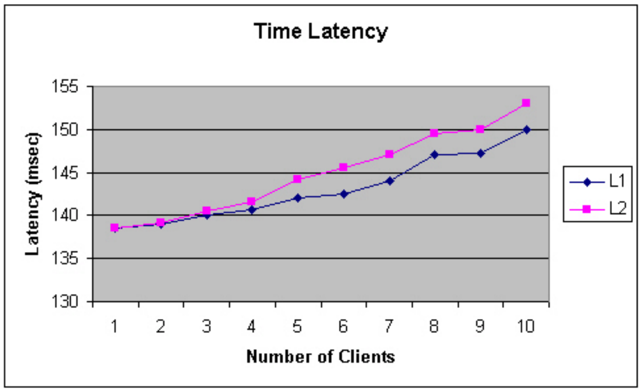


Figure 10 System Responsiveness Test Results

5. Summary

In this paper we have presented the database design in VRCASE to support multiuser collaborative software design in a 3D shared virtual environment. The needs of concurrency control support in collaborative VRCASE have been considered in the database design. It has been designed to enable fine-grained locking and notification mechanisms incorporating with Visual Indicator mechanism to be used to reduce concurrent conflicts and notifications from the traditional object level to attribute level. Our experiment results show that our database design can support efficient collaborative software design with good system responsiveness.

Acknowledgement

This research work was carried out under the research grant AcRF 20/98 provided by Ministry of Education, Singapore. We would also like to thank those postgraduate research students from Information Communication Institute of Singapore, School of Electrical & Electronics Engineering, Nanyang Technological University, who had participated and helped in evaluating VRCASE.

Reference

1. Lin Q, Low CP, Ng J M, 2002, “A Virtual Environment Based System for Assisting Object-Oriented Software Development”, Final Report for AcRF project 20/98, Nanyang Technological University of Singapore.
2. Chengzheng Sun, Rok Sosic, “Consistency maintenance in web-based real-time group editors”, Proceedings 19th International Conference on Distributed Computing Systems, IEEE, Pages 15 –22,1999.
3. Page.D, Mehandjiska.D, Griffin.D and Usherwood. L, “Methodology Independent OO CASE Tool: Supporting Methodology Engineering”, Proceedings of International Conference on Software Engineering: Education and Practice, IEEE, Pages 373 –380,1998.
4. Fugetta.A, “A classification of case technology”, IEEE Computer, Vol 26, No 12, Pages 25-38, Dec1993.
5. Norman.R.J, M.Chen, "Working Together to Integrate CASE", IEEE Software, Vol. 9, No. 3, Pages 13-16, Mar 1992.
6. XiaoHua Liu, “Multi-user Virtual Environment Based Object-Oriented CASE Tool”, thesis for Master of Engineering, Nanyang Technological University of Singapore, 2000.
7. Software Engineering Institute in Carnegie Mellon University, “What is a CASE Environment?”, http//www.sei.cmu.edu/legacy/case/case_whatis.html.
8. Diane Lending, Norman L.Chervany, “The Use of CASE Tools”, Proceedings of the ACM 1998 Conference on Computer Personnel Research, Pages 49 – 58,1998.
9. Tony Pittarese, “Supporting Collaborative Software Development Using CASE: Team-Based Issues in CASE Success” http://www.pittarese.com/Auburn/cse625/Supporting%20Collaborative%20Software%20Development%20Using%20CASE.html.
10. Sorensen.C, “Why CASE tools do not support coordination”, IEE Colloquium on CSCW and the Software Process (Digest No.1995/036), IEEE, Pages 4/1 -4/3,1995.
11. Purvis.M, Jones.P, “A group collaboration tool for software engineering projects”, Proceedings of International Conference on Software Engineering: Education and Practice, IEEE, Pages 362 –369,1996.
12. Sterling Software, ObjectTeam(Now the COOL:Jex) product document, Cayanne, 2000.
13. Advanced Software Technology, GDPro on line document, http:// www.advancedsw.com, 2001.
14. Tae-Hoon Kim, Woo-Chang Shin, Geun-Duk Park, Tae-Heun Lee, Tae-Gyun Kim, Yeong Gil Shin and Chi-Su Wu, “DOOD Distributed Object-Oriented Software Development Environment”, Proceedings of APSEC '97 and ICSC '97, IEEE, Pages 427 –434,1997.
15. Rational Rose online document, Rational software, http://www.rational.com/,2001.
16. Gintell.J.W, Houde.M.B and McKenney.R.F, “Lessons Learned By Building and Using Scrutiny, a Collaborative Software Inspection System”, Proceedings of 7th International Workshop on Computer-Aided Software Engineering, IEEE, Pages 350 –357,1995.
17. Brothers.V, Sembugamoorthy and M.Muller, “ICICLE: Groupware for Code Inspection”, Proceedings of the ACM 1990 Conference on Computer-Supported Cooperative Work, Pages 169-181,1990.
18. Juan Bu, “Multi-user collaborative Work in Virtual Reality based CASE Tool”, thesis for Master of Engineering, Nanyang Technological University of Singapore, 2001.

Optimizing the Flow of Organizational Information using a Groupware Application

Roland Haas, Arun Kumar Krishnamoorthy, Bharath Chandrasekaran
DaimlerChrysler Research and Technology India
137, Infantry Road
Bangalore 560001
India
E-mail: (roland.h.haas, arunkumar.krishnamoorthy, bharath.chandrasekaran)@daimlerchrysler.com

KEYWORDS

Information and knowledge management, groupware, Lotus Notes, organizational processes

ABSTRACT

This paper presents a case study which shows how the information flow and the accessibility of data within an organization can be streamlined and optimized. The optimization is based on a careful analysis of people's information needs. All relevant data is kept in the central data storage of a Lotus Notes based application named eDP. While these databases provide easy and secure access to the information, Lotus Notes' workflow capabilities offer a straightforward mechanism to automate complex organizational workflows. The paper gives an introduction to the basic functionality and system architecture of eDP. One crucial process – leave application for employees – is studied in detail.

INTRODUCTION

DaimlerChrysler Research and Technology India is a 100% subsidiary of the DaimlerChrysler group delivering research, IT and engineering services to DC business units worldwide. The center is part of DaimlerChrysler's central research division and works on challenging research projects in the areas of image and signal processing, telematics, and communication sytems. This work is mostly done in global, distributed teams where team members in India work together with their German and American colleagues. In addition to the research activities the center also has a diverse development and service portfolio ranging from software development to CAx (CAD, CAE, CAM) support. Most of these projects are also excuted in a mixed on-site/off-site (off-shore) mode where some of the project team members are at the customer site while the rest of the team is based in India.

Modern collaborative tools and video conferencing equipment help to streamline information flow and optimize communication between the different locations.

MOTIVATION

Seamless availability of information and short response times in addressing customer needs are crucial to any service organization. To illustrate this, let us look at a typical business scenario: A new project is signed off, requiring a team on-site. The team members have to be selected on the basis of their skills, experience and availability. Once it is decided who will be deputed to the customer site, the travel arrangements have to be made. A crucial step here is the application for the appropriate VISA. While the process is clear and well defined, the success depends on the availability of all relevant pieces of information and documents.

If, for example, a document for the VISA application form is missing, or the passport validity is not tracked, significant delays and even a loss of the project could be the result.

eDP

eDP is an abbreviation for 'electronic Data and Processes'. Its aim is to provide seamless access to employee-related and organizational information, and to support key processes and workflows.

In the first phase, the development team concentrated on handling employee-realted data like date of recruitment, skill profile, and leave data. All employees and managers can access this data through the browser, for a quick overview or to edit and update it.

The ultimate goal is to provide in a phased manner all necessary organizational data to those who may need it and to keep it up to date. Data security, of course, is crucial and needs to be guaranteed by proper authentication and authorization mechanisms.

The eDP manages data like:

- employee data (particulars of employees)
- organizational information (like org. charts, etc.)
- policy documents (travel, leave, allowances, etc.)
- customer related data
- project information, and
- library information

Based on this data, typical organizational processes can be supported by workflow (this constitutes the 'e' part in the name of the application). Some of the corresponding organizational workflows are:

- leave application and approval
- foreign travel and VISAs
- training
- issue tracking
- library
- recruitment

SYSTEM ARCHITECTURE

eDP runs on Lotus Notes Domino. Lotus Notes is an ideal platform for setting up such an Information Portal rapidly. Lotus Notes offers robust features to deliver secure communication, collaboration and business applications. IF Lotus Notes is used, the GUI can be either proprietary (this would need a Notes client) or Browser-based.

The GUI of eDP is web-based and is a part of the company's Intranet. All users can access their information using the browser and change it if necessary. Each module in eDP is developed as a separate Notes database. This enables parallel development and enhancement of modules. The following illustration (Fig. 1) is a high-level block-diagram of the system.

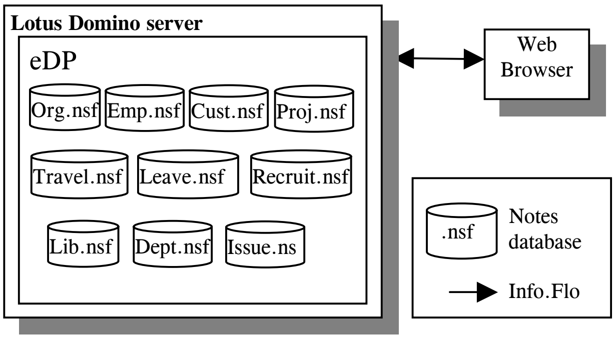


Fig. 1 High-level block diagram of DP

The various modules in eDP can be classified under two broad heads – data repositories that manage organization-wide data and organizational processes that implement various organizational workflows. Each eDP data repository in Notes maintains a specific category of data and generates various reports of interest to the user. Functionalities offered by these components in the repository are driven by Access Control Lists (ACL) specified for that repository. ACLs for each module are specified by the respective stakeholders of that module. For example, the data repository called Emp.nsf maintains employee-related data and its ACL is specified by the Human Resources (HR) department. The following illustration (Fig. 2) is a high-level block diagram of a typical Notes database in eDP that serves as a data repository.

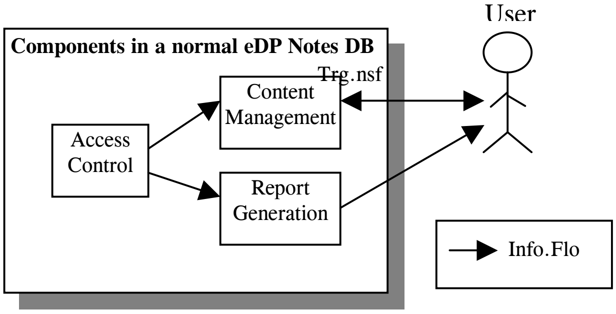


Fig. 2 High-level block diagram of an eDP data repository

Each module in eDP that automates organization processes has a workflow logic that routes a task-related document in the organization, and determines how the state of the document changes. Workflow logic also incorporates the necessary access control and the rules for information flow. Mail notification is sent as per the specified rules, typically when the state of a document changes, and to the people who m this change affects. Reports generated from these modules mainly give statistical information on activities that went through the workflow. The following illustration (Fig. 3) is a high-level block diagram of various components in an eDP database that automate organization processes.

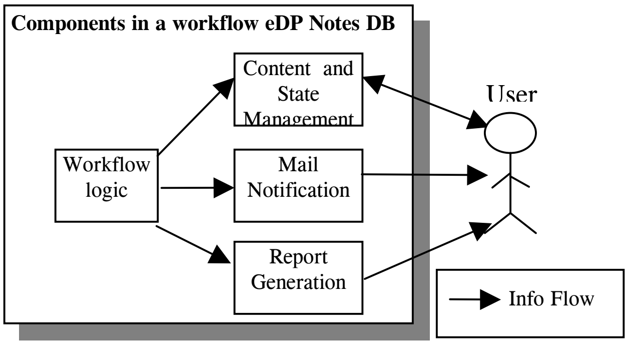
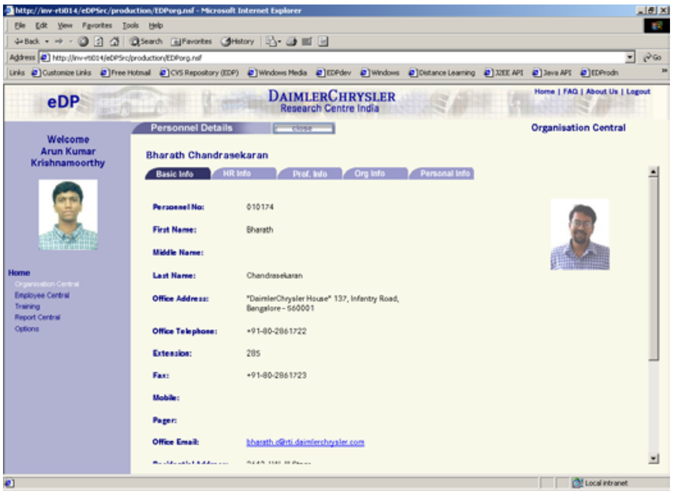

Components in a workflow eDP Notes DB
Content and State Management
User
Workflow logic
Mail Notification
Report Generation
Info Flow

Fig. 3 High-level block diagram of eDP workflow database

Lotus Notes offers several inbuilt features to realize the components discussed above. The access control mechanism provided by Lotus Notes comes in handy for incorporating security and confidentiality in each of the modules. In Lotus Notes, access can be controlled at the database level right down to the level of a field in a form. This adds the desired granular security to the application. Design elements provided by Lotus Notes help in rapid realization of other components in the module.

Fig. 4 shows a snapshot of details maintained in the Employee database of the eDP.


http://inv-rti014/eDPSrc/production/EDPorg.nsf - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://inv-rti014/eDPSrc/production/EDPorg.nsf
Go
Links Customize Links Free Hotmail CVS Repository (EDP) Windows Media EDPdev Windows Distance Learning J2EE API Java API EDProdn
eDP
DAIMLERCHRYSLER
Research Centre India
Home | FAQ | About Us | Logout
Personnel Details
close
Organisation Central
Welcome
Arun Kumar
Krishnamoorthy
Home
Organisation Central
Employee Central
Training
Report Central
Options
Basic Info
HR Info
Prof. Info
Org Info
Personal Info
Local intranet

Bharath Chandrasekaran

Personnel No: 010174

First Name: Bharath

Middle Name:

Last Name: Chandrasekaran

Office Address: "DaimlerChrysler House" 137, Infantry Road, Bangalore - 560001

Office Telephone: +91-80-2861722

Extension: 285

Fax: +91-80-2861723

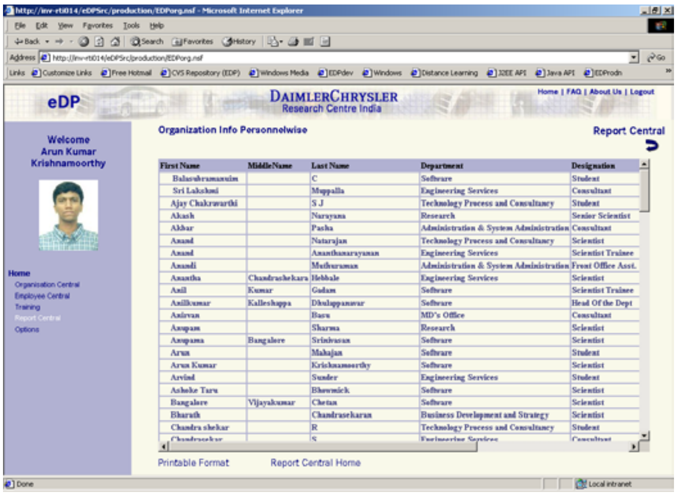
Mobile:

Pager:

Office Email: bharath.c@rti.daimlerchrysler.com

Fig. 4 Snapshot of Employee database in eDP

Fig. 5 shows a snapshot of a report generated from the Employee database.


http://inv-rti014/eDPSrc/production/EDPorg.nsf - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://inv-rti014/eDPSrc/production/EDPorg.nsf
Go
Links Customize Links Free Hotmail CVS Repository (EDP) Windows Media EDPdev Windows Distance Learning J2EE API Java API EDProdn
eDP
DAIMLERCHRYSLER
Research Centre India
Home | FAQ | About Us | Logout
Welcome
Arun Kumar
Krishnamoorthy
Home
Organisation Central
Employee Central
Training
Report Central
Options
Printable Format
Report Central Home
Done
Local intranet

Organization Info Personnelwise

Report Central

First Name	MiddleName	Last Name	Department	Designation
Balasubramanuim		C	Software	Student
Sri Lakshmi		Muppalla	Engineering Services	Consultant
Ajay Chakrawarthi		S J	Technology Process and Consultancy	Student
Akash		Narayana	Research	Senior Scientist
Akbar		Pasha	Administration & System Administration	Consultant
Anand		Natarajan	Technology Process and Consultancy	Scientist
Anand		Ananthanarayanan	Engineering Services	Scientist Trainee
Anandi		Muthuraman	Administration & System Administration	Front Office Asst.
Anantha	Chandrashekara	Hebbale	Engineering Services	Scientist
Anil	Kumar	Gadam	Software	Scientist Trainee
Anilkumar	Kalleshappa	Dhulappanavar	Software	Head Of the Dept
Anirvan		Basu	MD's Office	Consultant
Anupam		Sharma	Research	Scientist
Anupama	Bangalore	Srinivasan	Software	Scientist
Arun		Mahajan	Software	Student
Arun Kumar		Krishnamoorthy	Software	Scientist
Arvind		Sunder	Engineering Services	Student
Ashoke Taru		Bhowmick	Software	Scientist
Bangalore	Vijayakumar	Chetan	Software	Scientist
Bharath		Chandrasekharan	Business Development and Strategy	Scientist
Chandra shekar		R	Technology Process and Consultancy	Student
Chandrasekhar		S	Engineering Services	Consultant

Fig 5 Snapshot of a report generated from Employee database in eDP

Leave Application Process – A Case study

The Leave Application Process is discussed here to demonstrate how a typical organization process, namely, applying for and processing leave, optimized in eDP using Lotus Notes. To start with a basic foundation for this process, a simple leave application process with a single approval cycle is built into eDP.

1. An employee logs into eDP and submits an application for leave. A mail notification is sent to the approver (Head of Department concerned)
2. The approver logs into eDP and approves/rejects the leave application. A mail notification is sent to the applicant regarding the status of his/her leave application (approved/rejected).
3. HR and Finance departments get various reports from the leave process system

Fig. 6 depicts the workflow involved in the leave application process stated above.

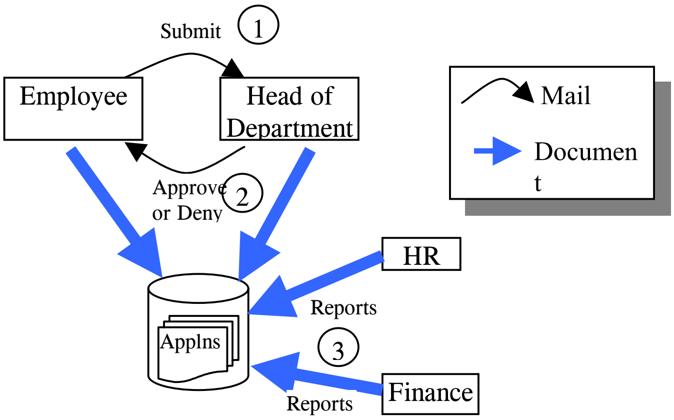


Fig. 6 Workflow in a simple leave application process

In a typical organization, roles are clearly specified for each person in the organization. The reporting structure is well-defined. In such a case, the workflow for processing a leave application should be maped to the reporting structure. The following description shows how the roles and reporting structure that are characteristics of a typical project are used to route the leave application workflow in eDP.

The following roles are usually present in a typical project.

1. Team Member
2. Project Leader
3. Project Manager
4. Department Head

A Team Member can be part of more than one project. Hence, he/she might have to report to more than one Project Leader. Similarly, Project Leaders who lead more than one project might have to report to more than one Project Manager. Consider how the workflow in a leave application process propagates in this multi-level reporting structure.

1. A Team Member logs in to eDP and submits leave application. A mail notification is sent to all Project Leaders concerned.
2. Each of these Project Leaders logs in to eDP and approves/rejects the leave application.
3. If any of the Project Leaders does not approve a leave application within a certain period of time, it gets escalated to the next level.
4. If any of the Project Leaders rejects a leave application, it will be rejected and a mail notification sent to the applicant.
5. If all the approvers approve the leave application, then the status of the leave application is set to 'approved'. A mail notification is sent to the applicant.
6. HR and Finance departments get various reports from the leave process system.

Fig. 7 depicts the workflow involved in the multi-level leave application process stated above.

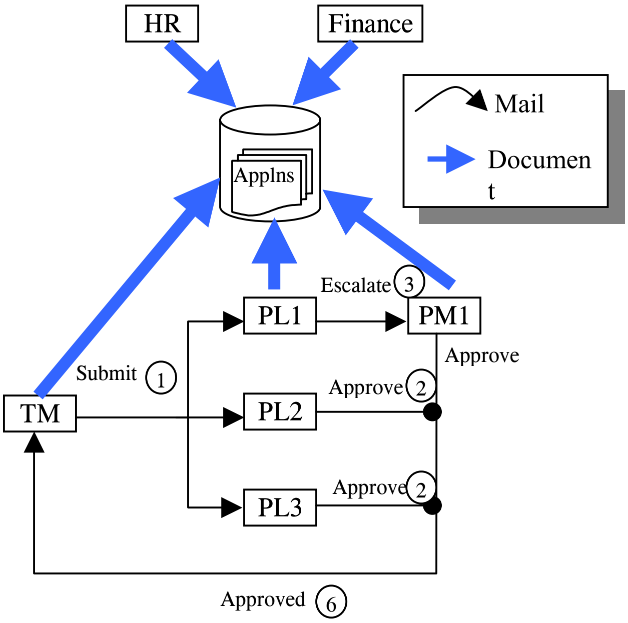


Fig. 7 Workflow of a multi-level leave application process

The leave process can also be configured to specify the maximum number of days of leave that a person at a certain level in the reporting structure is authorized to approve/reject.

The roles of employees considered in the leave application process are picked from the Project database. Department, designation and other employee details are taken from the

Employee database. This demonstrates how information in an organization can be optimized in eDP using Lotus Notes.

SUMMARY AND OUTLOOK

Quick response times and seamless access to all relevant information and knowledge is a must in any competitive business area.

The paper presents a groupware system that is designed to support all sorts of organizational processes.

It provides a portal to crucial information about all employees and helps to keep the information up to date.

Future versions will support futher processes and will also include complex workflows.

Support for workflows is a strength of Lotus Notes. Once a workflow process has been designed and is well understood, the effort needed to implement it is minimal.

eDP will be an important tool in the company's quality initiative. Currently DCRTI is preparing for CMM-3 certification for its software development. At this level, the coordination between organizational units like HR and Finance plays an important role. The information flow will benefit significantly from eDP's deployment.

Currently the eDP development team is also studying the integration of this application with the existing Knowledge Management (KM) system in the company, with the ultimate goal of building up a consisten KM repository across the organization.

With more and more critical information stored in eDP, the security featuresof the application will become increasingly important. This will be studied in detail to assess the risk and possible security breaches. Connectivity (e.g. ERP systems) may also play an important role in future versions.

REFERENCES

1. Boy, G., and Durstewitz, M. (1998). "Organizational traceability and hypertext functionality", In ICIS Workshop on Hypertext Functionalities, Helsinki
2. Davenport T., Prusak L. (1998). Working Knowledge - how organizations manage what they know. Harvard Press, Boston
3. Hawryszkiewycz, Igor. (1997). Designing the Networked Enterprise, Artech House Inc., Boston
4. R. Haas (2000): Corporate Intranets – technologies and applications, Proceedings ICE2000, Toulouse.
5. R. Haas (2000): Engineering Knowledge Management - Current status and future challenges, Proceedings ICE2000, Toulouse.
6. Drucker (2000) People and Performance.
7. Volker Bach, Petra Vogler, Hubertus Oesterreich (Eds.), "Business Knowledge Management", Springer 1999.
8. Roland Haas, Wilfried Aulbur, Sunil Thakar: Enabling Communities of Practice at EADS Airbus in Sharing Expertise – Beyond Knowledge Management. M. Ackermann, V. Pipek, and V. Wulf (Eds). MIT Press 2003
9. R. Orfali, D. Harkey, J. Edwards: Client/Server Survival Guide. 3rd ed., Wiley, New York, 1999.

DISTRIBUTED ENVIRONMENTS

A WEB SERVICE BASED APPROACH TO MONITOR AND CONTROL A DISTRIBUTED COMPONENT EXECUTION ENVIRONMENT

Anja Schanzenberger and D. R. Lawrence
Middlesex University, School of Computing Science
Bounds Green Road, London, N11 2NQ, United Kingdom
E-mail: anja.schanzenberger@gfk.de, dave7@mdx.ac.uk

1 ABSTRACT

Many modern enterprises have already discovered and deployed component based distributed systems. These kinds of systems and certainly the Internet or Intranet as a world-wide network enable companies to expand their organisations over national boundaries and to internationalise them. It is an attractive approach to use component based distributed systems.

This technology allows companies to grow without the burden of costly changes in approaches, because these systems are flexible and can be easily extended by the relatively simple insertion of additional components. However, distributed systems have a disadvantage which is of great interest to our research. There is an inherent reduction or loss of control, if the number and the distribution of components becomes excessive.

A further complication is that the expanded system's personnel knows only small parts of the whole system (for example only those parts for which they are responsible). Few people know the interplay of the complex structure of the system and this is why the reaction to failures or system breakdowns becomes more and more difficult with a growing distributed system. This paper focuses on potential possibilities to automatically control and monitor distributed component based systems.

A classic example is the Planning-, Control- and Monitoring System of GfK Marketing Services. This system is largely at a conceptual stage and forms a case study for our research. This paper introduces the concept for this system and discusses issues related to the design of such a system. The use of web services forms part of the conceptual solution. Research issues, and key points arising from our initial research are briefly discussed.

2 INTRODUCTION

2.1 Distributed Systems and Components

The concept for developing distributed systems is not new. The research began in the 80's and 90's (Sharp 1987)(Mullender 1993)(Champine 1980)(Gomer 1990). Today, commercial suppliers of this technology (e.g. (Sun 2003), (Microsoft 1998), etc.) profit from that previous research and offer sophisticated tools and platforms to support and develop such systems. There has been a large increase in the number of companies deploying distributed systems (Linnhoff-Popien 2000) and there are many justifications for this increase. Some of the more essential ones can be found in the following list.

- The trend for internationalisation of companies (globalisation)
- The effective possibilities of distribution of resources and the associated utilisation of these resources
- With the usage of such systems, reliability (failsafe performance) can be achieved through 'standby' components.
- Likely increase in productivity due to parallel uses of multiple instances of components.

The considered distributed environments include so-called components. At the moment there are already two leading technologies for developing components. The first are JAVA and CORBA (Java 2002). The second are Microsoft and the COM+ technology (Pattison 2000). Both of these have advantages and disadvantages but their common aim is to offer a platform for developing components, which can be used beyond the borders of networks. Whereas JAVA is usually independent of an operating system, COM+ is based on Microsoft operating systems.

2.2 Introduction of an Example for Monitoring and Controlling

2.2.1 The Case Study

To investigate monitoring and controlling in component-based distributed environments, GfK Marketing Services and their planned PCMS (Planning-, Control- and Monitoring System) is used as a classic case study. However, the examinations are of common interest and relevant to component based distributed systems in general. The following describes GfK in short, to give an impression of a typical environment, some background of the system, and the plans for development.

GfK Group is a leading market researcher. GfK Marketing Services, one of four main divisions of the GfK Group, produces reports from retailers. They create retailer statistics and offer these reports to their customers world-wide. Local branches are available in more than 40 countries (GfK 2002). GfK Marketing Services has already created a very large component-based data production system as you can see in the rough sketch figure 1.

In each branch there are many participating components which collaborate together, as a given workflow. Data is gathered and formatted into a GfK internal uniformed format at each of these local branches. After this process the data is transmitted to the central branch at Nuremberg. There, the data of all countries are collected and the extrapolation to reports follows.

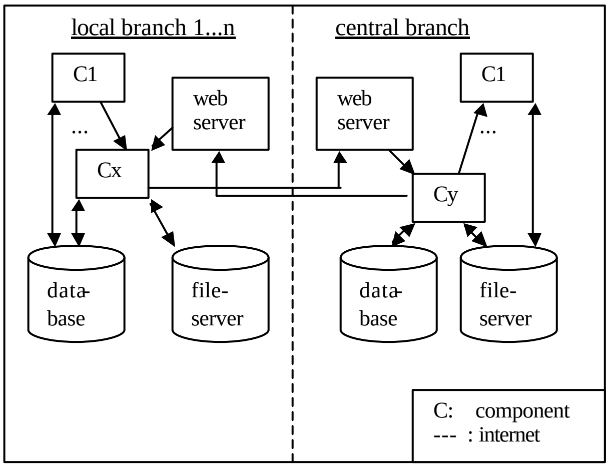


Figure 1: A distributed component execution environment

Data consisting of sold articles from one retailer within a certain delivery period (monthly, bimonthly, etc.) corresponds to the term of a job. The workflow definition prescribes the way a job flows from a local to the central branch. At the central branch all data is gathered. Also, activities of manipulation as aggregation or splitting occur during these production steps. After the data is processed and gathered, the emphasis switches to reporting where evaluation periods and product groups are examined. From the viewpoint of a database expert, this implies permanent changing of the identification keys for the jobs and extremely complicates the task of control (e.g. delivery periods are aggregated to evaluation periods, data of discrete retailers are aggregated to product groups, etc.).

The reason for the redevelopment of the whole production system was the desire to eliminate an old mainframe, because of cost, efficiency considerations and non-parallelism.
The new production system is able to handle jobs parallel and is thus more flexible, less expensive (more data in less time), extendible (components, partitioning on several CPUs, etc.) and faster.

For the new component based production system, Microsoft's Componentware (Eddon 1998; Microsoft 1998) is used. The fully developed production system is called the ***'execution environment'***. It is responsible for all process steps from data acquisition (local) to reporting (central) to cerate accurate marketing reports. The workflow definition of the production system is not expected to support excessive changes and is therefore designated as (more or less) fixed. This is why an adequate control method in this scenario should not have to cope with permanent changing workflow definitions.

The need as well as the reasons for an automated supervision was realised early. Firstly, they involved only the German local and the central branch. Step by step the former path to the mainframe was replaced by an increasing amount of loosely coupled components. Expeditiously, questions are arising about the existence and the quality of particular data or the behaviour in failure situations. Problems with responsibilities (for example: who has the competence to correct the failure at a particular component?) and the control about the whole business process were identified during early experiments with this new environment. Other considerations were that the staff operating the current production system has important knowledge about when and how retailer data has to be processed periodically. Due to the large amount of personnel and departments involved, the management decided the integration of this knowledge to an as large as possible extend is advisable.
The fulfilment of contracts with their customers and the guarantee of delivering reports on time are the most important targets of GfK's business. However, the conclusion for GfK was, that without an automated control over all participating resources (humans, machines, software, etc.) this target could not be reached effectively. This is mainly why the company intends to establish a management tool (the so-called PCMS, which has not yet been developed).

2.2.2 Benefits for PCMS Users

The expected benefits of the PCMS can be seen in the following list.

1. General aspects:
   - *Optimisation of the whole business process*

   One part of a PCMS are statistics concerning the average utilization. This helps to identify a component which is perhaps a bottleneck in the workflow. An adequate reaction would be to replace this single component with two or more instances of it. The detection is only possible if a tool is available for controlling such scenarios. However, it is not just the system technical viewpoint that must be optimised. Priority flags and rules for giving preference to important jobs can be profitable to reduce production costs. Peek workloads can be reduced for all resources (human and others) as can be seen in figure 2.

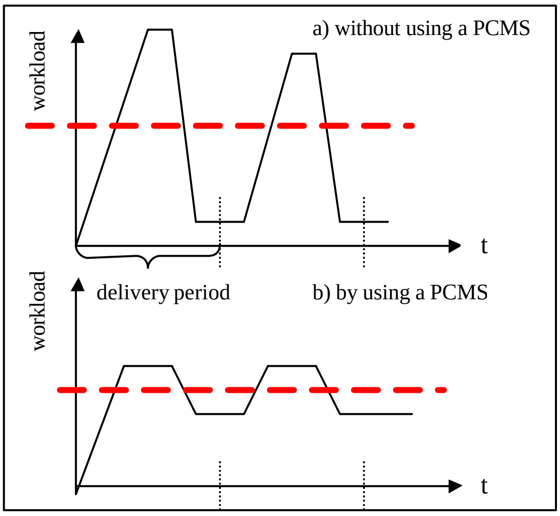


Figure 2: Fluctuation margin reduction of workload

Another advantage of using priorities is the possibility to reduce duration time for important jobs. This increases the topicality of the reports and higher throughput can be deduced.

- *Parallelism of the whole production process*

To involve more instances of components shortens the processing times and more efficient utilization of the production system is guaranteed.

2. Planning aspect:

- *Distribution of priorities*

By using priorities as mentioned in the general aspects in point 1, it is possible to deliver the job priorities to all participating departments. Work lists can be generated. This helps the personnel to plan their actions more precisely and to meet the final delivery date of the reports. Without a PCMS this has to be done manually and it is not possible to obtain an overview of all jobs, departments and involved personnel in every situation.

3. Monitoring aspect

- *To overview the whole system as well as it's parts*

It is important to respect the independence of the involved departments, just as it is to guarantee the common target to produce the reports in time.

- *To get information about system states*

For example, it is important to understand why job X from the branch in Great Britain has not been delivered to the central branch, although the data is needed to meet a report's deadline.

- *To get information about processed data to a certain state*

Adequate quality-control summaries (e.g. about sales volumes, etc.) are necessary. These summaries help to evaluate the quality of the included data in a job.

- *To involve organisational aspects*

Every participant is able to view his or her own responsibilities as well as those of others. This is useful in delineation of responsibility zones and for coping with failures over more than one area. One result is, downtimes can be reduced because responsibilities are defined and are instantly addressable.

- *Operational aspects*

It is possible to notify, for example, any bottlenecks and problems with load balancing. After a proactive and automatic notification, actions can be taken and performed by the user.

- *Possibility to log production steps*

Two types of log information are required. The first one is to log system technical events (e.g. a component is not working well).

The second type is more important. It logs information about the content of the data (e.g. retailer X for the delivery period Y has passed component Z successfully). With this second log type a historical documentation of the jobs can be achieved. However, it is more important to be able to measure the current progress of the jobs (e.g. 80% of a job is processed) with this log information.

The plan can be compared finally with the current state and be evaluated through, for example comparison with the preceding month or years.

4. Controlling aspect:

- - *To control components in the whole workflow*

For example, to involve, delete, start, or stop components.

- *To set breakpoints at execution time*

If the production can be paused at a well-defined state, repairs (of components or data) will be much easier.

- *To set job priorities*

This is a more important function. Jobs run with a default priority, but this priority can be changed manually at anytime. After a change, a notification must be send to the place of current execution (e.g. the job is then assigned a new preference in a message queue).

- *Error handling*

The aim is to reset a computer, a component or data to a defined state before a failure happened.

2.2.3 Content of a PCMS

After realising the need for a PCMS the question arises as to the intrinsic contents of such a system? Our research showed that a supervisory instrument must be able to represent a current state as well as a planned state of the production system in a suitable way. This is why an accurate analysis is necessary for the definition of the system states and the calculation rules used to reach this target. Important is to define methods for handling differences between current and planned states in the system and how they can be estimated. Equally significant is the reaction to failures which must be enabled (e.g. failure messages of components or system break downs). In addition, other necessary content must be defined such as an organisational overview of individual responsibilities or work-lists for all participating (e.g. human) resources.

The challenging task to design a PCMS

The previous research shows, there are two focal points needed for a PCMS:

(a) First, a central located data inquiry system mainly responsible for querying the logging and the state of information at runtime. Its information can be transferred to every local location per Web technology.

(b) Secondly, a hierarchical oriented management system is needed (e.g. responsible for tasks like starting, stopping, asking for response of components and computer workstations etc.). Some parts of the system are local and some are central located.

For both portions of PCMS it is not possible to integrate them in a usual information system which could be easy developed. There is much more involved as it is not only a monitoring system and thus justifies our research. For example, relevant to research are the following reoccurring problems:

- A PCMS is not only a simple enquiry system, but a fully functional controlling system. It must be possible to intervene actively in the workflow to react immediately to events such as errors and failures in the production.
- The complexity of the PCMS increases, due to the fact that it is a control instrument for a distributed data production system. World-wide all participants will want to access the PCMS. This means staff of a local branch are interested in a successful computation of data delivered to the central branch and vice versa personnel from the central branch are interested in the state of execution for one particular job at a local branch.
- Not only data packages must be easily controlled with a PCMS, but also the content of the jobs must be monitored and controlled (information contained within the data packages). This is an important argument against existing, commercial and traditional Workflow Management Systems, due to the lack of support for these special requirements.
- Additionally, the usage of changing identification keys makes the problem more difficult. The identification keys change several times within the whole workflow (e.g. delivery periods are transformed into reporting periods; data of discrete retailers are aggregated to product groups, etc.)
- Not only changing identification keys are a problem. Also permanent data transformation has to be considered (aggregation and splitting data) and is a requirement for the PCMS.
- The flow through the production cycle is almost the same, nevertheless the workflow must be configurable. For example, the insertion of a new component into the workflow should not become an obstacle for a flowing process.
- Priorities for jobs are necessary. The order of processing must be derivable. Important jobs have to be served first.
- One speciality of GfK's business is that for the completion of a report e.g. processing of 80% of the required data deliveries can be enough. The gap can be filled through statistical extrapolation. The consideration of this fact is another requirement to the research work of our PCMS.

2.3 The Technological Background

The most eligible approach for data exchange in distributed systems as described in the case study involves Web technologies.

Alternatives are, for examp le, point-to-point RPCs (Remote Procedure Calls). This is a well-known technology for activating functions over a network. But many point-to-point connections between sources and their targets quickly become unmanageable and inflexible (cp. Linthicum 2001) and they are a security risk and thus mostly blocked by firewalls.

The reasons for using Web technologies are that they enable the company to exchange data over networks, using comfortable and well-proven, standardised technologies like HTML, ASP, FTP or e-mail.

Web technologies also allow large (some with permanent line) and small local branches (only equipped with modem) to connect to centralized data without any differences in access strategies. Only the transfer time (bits per second) for data is different if slow modems are used. Other reasons for Web technologies are more global considerations like economy, scalability (reduction of the functional complexity), ability to communicate (ability to collaborate), future-proof for plans and investments and user-friendliness with uniformed GUIs. The global trend towards an e-business society must not be ignored by the industry or by GfK (Dwyer 2000).

It appears that the best approach to provide these services for a PCMS are web services. Sun Microsystems describe a Web Service as a representation of a unit of business, application or system functionality that can be accessed over the Web (Manes 2001). Another description is stated in IBM's Tutorial (Vasudevan 2001): "Web services are a new breed of Web application. They are self-contained, self-describing, modular applications that can be published, located, and invoked across the Web. Web services perform functions, which can be anything from simple requests to complicated business processes...Once a Web Service is deployed, other applications (and other web services) can discover and invoke the deployed service.". (Please find more information about web services at (web-services.org 2002)).

This technology is very young. Based on common standards like the accepted XML ('Extensible Markup Language') (W3C 2002), SOAP ('Simple Object Access Protocol') (W3C 2002) and WSDL ('Web Services Description Language') (WSDL 2002), it is no longer necessary to care about conventions of the message (e.g. header, body, etc.), of the transport (e.g. open port) and it is possible to structure the included functions and data well. What makes web services so appealing is that it works via HTTP, the traffic goes right through port 80 (HTTP standard port). But this can create problems from a security viewpoint. Web services and especially SOAP are being controversially discussed at this time and a lot of action is expected concerning these issues in the near future (Farley 2000).

However, there are several reasons to use web services for GfK's project. To use a future-proof technology for transporting the control-information of the PCMS through the world-wide system is important. Web services are not revolutionary, but they offer a giant step forward in evolution. It is only a question of time till they become the standard as W3C demonstrates (W3C 2002). Web-GUIs are not always needed. One advantage of the web services is the power to work with and without user interfaces. Web services are characterised by the ability to transport data correctly and timely over the Web. In this example, it is proposed to use web services mainly for the very special task of transporting control information within the PCMS to control the component based and distributed data production system.

3 THE APPROACH TO MONITOR AND CONTROL A DISTRIBUTED COMPONENT EXECUTION ENVIRONMENT

The logical architecture of PCMS plays a vital role for the choice concerning the usage of web services. It consists, in essence, of two principles. These principles will be introduced in this chapter. The first one is the principle of the service based processing. Supervision of orders and additional supervision of the control flow can be achieved through the call of services between the participating components. The second principle is the one of a level based approach. A separation of different tasks in the system can be achieved through the usage of different levels by the modelling of PCMS. For example, the data level is uncoupled from the system-components level. The consequence is the transfer of very big data packages (data deliveries, jobs) of GfK can be uncoupled and transferred between the spatially separated organisation parts cleverly e.g. per FTP. The advantage is that the system components don't have to communicate between the different organisation parts.
(There are other features of the system which are not described in this paper to reduce the complexity for the reader. For example a method for tracking single items through the whole workflow is described at (Schanzenberger et al. 2003))

3.1 Principle of a Service Based Processing

The original idea is based on an orthogonal method on the basis of single services. The control of the production is reached by exchanging messages between the components. Due to asynchronous processing of the messages between the participating components a decoupling is achievable. Only persistent queuing systems should be used on the system level. These are the reasons why messages cannot be lost or (transactional) duplicated in the future. In essence, two kinds of messages should be applied:

- <u>*Order messages*</u>

Initially, an order is given from the end of the process chain, the reporting-side. It is propagated backwards to the beginning of the process, the data entrance. Additionally, basic values are adjusted appropriate to the production plan to meet the changing primary keys during the process steps. Beside the content-based information an order message contains time-based and production-based information, which must be followed to satisfy a deadline of one data-order:

- Information in percentage: According to a relative quantity a given amount of data (to a special deadline) must be processed.
- Content-based information: Deliveries of special shops must be processed to satisfy an order.

The order message includes the *planned* and the *actual state* of an order. The planned state implies that special data areas are produced to a planned deadline. The actual state of an order message represents to what extend an order is processed (based on estimation). On the one hand, this gives the user of the PCMS the possibility to query the current state of one order. On the other hand it gives resources additional information to prioritise orders. The actual state is calculated periodically with the help of a statistic component per branch and is based on the log information and the data stock.

- <u>*Processing messages*</u>

A processing message is subject of the control flow which is exchanged between the components. It describes the local order. A change in priority can be achieved if the relationship between the amount of the data from the processing message and the data stock of the order message is calculated.

3.2 Principle of a Level-Based Approach

The architectural approach of the PCMS is based fundamentally on the division into six levels. And, through this, a clear separation of the participating components is achieved. For example, it does not permit components to transfer data implicit with messages (as payload of a processing message) and others to transfer data explicit through the mention of filenames or tablenames. In the following six levels (see figure 3) are introduced briefly.

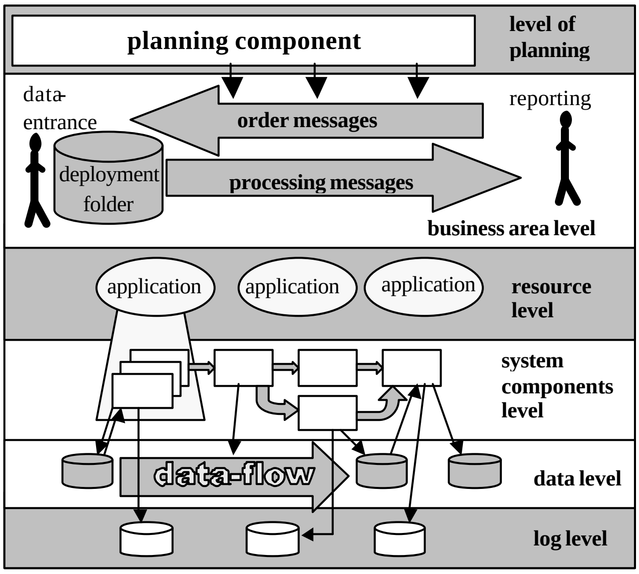


Figure 3: Principle of a level-based approach

<u>(1) Level of Planning (Global Over All Business-Areas)</u>
The creation and processing of order messages serve as an instrument for planning and include the demand for data quality (e.g. correctness and completeness of the facts within the data) and deadline. The orders are propagated from reporting to the data entrance (as it is done on a calendar today). The addressed data flows through the production chain by leading them from the first production step to the next. The planning of the deadlines occurs (a)

with a manual insertion and (b) through the comparison with past scenarios.

(2) Business Area Level

A business area represents one step in execution (with all necessary sub-steps) with one common deadline (e.g. the representation of one participating department in the process). The allocation is oriented on the established production process.

The business area is essentially an information system. In this information system a deployment folder exists in which all resources (see Nr. 3 below "Resource Level") are registered which belong to this business area. Incoming order messages are forwarded to all participating resources. This is the precondition for taking one order into account locally (web-postings as work lists, e-mail allocation, etc.) or incoming processing messages can get the right priority concerning to their execution. The content of a business object represents a business area and includes the following content-based signatures, beside lifecycle methods:

- *Acceptance of order messages*
  Orders given from the planning component are introduced to the business object. This business object is leading these orders to the participating resources.
- *Information about the control flow*
  Resources are able to find out their successor component with this function. The information about their successor is deposited in the control flow. The deployment folder includes the information about the resources and their possibilities for ramification after their underlying components have finished execution.

Therefore, the deployment folder is of importance in this process. From the technological viewpoint it has to be designed in a way that external resources (in particular staff) can be included in the process as well.

(3) Resource Level

Different resource types represent different production factors, which are locally administered through the corresponding resource. Examples for resource types are:

- *Web-page updating or e-mail service*
  Incoming orders are published in the Intranet as a kind of >*work list*< or announced to one or more staff members.
- *Application*
  An application-resource is a logical representation of a system component (and of an existing software-module).

(4) Level of System Components

Software components involve all software modules which are responsible for one single step within the production chain. More than one instance per component type can run on different computers. From the higher viewpoint of the business area this is transparent for and controlled by the associated resource.

In the deployment folder of the corresponding business area exists a list of successor components (complete list of the order within the workflow for one business unit). Every system component kicks off the start process of its successor by asking what its successor components are. After that, it sends a processing message to the successor to inform it about the pending job.

(5) Data Level

On the data level all systems or files for the storage of data can be found. Every system component (more detailed: system component type) is logically assigned to a particular data storage system. If instances of the same component type are able to access the same data storage at once, the aspect of multithreading has to be considered on the data level and for example a transactional protection has to be guaranteed.

(6) Log Level

In addition to the components of the production system, log-components and the statistic component are arranged. Resources log with a logging service in different log structures:

- *Service-Log*
  The service-log contains information about the runtime of components and system based failures. It is separated spatially in one local (in every country available) and one central log (central branch Nuremberg).
- *Application-Log*
  The application-log includes specific information about the applications. For example amount of data sets, time span for the processing of particular product groups. The application-log is the basis for the statistic tool and can be used from human resources as well. With it the execution of particular steps in the process can be announced (with a Web-GUI). This log is also separated spatially in a local and a central location.

The level of system-components and the data level are together the data production system. All other levels are original parts of the PCMS.

In principle the statistic tool is an ordinary resource of a business object. On the one hand, it receives a copy of all incoming orders. On the other hand, the statistic tool uses the application-log and the process data. With information from both, the statistic tool is able to proof a current state of execution. Existing orders can be updated manually (perhaps with suggestions) and then (analogous to the planning component) are introduced to the business object. The business object forwards these orders to all its participating resources for the distribution of priorities. As a result, the statistic module is basis for a qualitative control.

4 OPERATIONAL AREAS OF WEB SERVICES IN THIS ENVIRONMENT

In the introduced system architecture there are several possibilities to use Web Service technology to support a smooth production and a clever supervision of these business processes. Possible operational areas are discussed in this section.

- *Control flow in the data production system*
  The control flow has to be organised per business unit, which represents for example one single department.

Additionally, transitions have to be created between the business units to be able to continue the control flow from the beginning to the end of the production process.

The mentioned deployment folder shall be used for finding suitable services (components) within a business unit as well as between them. Its implementation can be a standard like UDDI (Repository for 'Universal Description, Discovery and Integration', a technical reference information (OASIS 2002)). Alternatively, an own database solution can be chosen to fulfil the need to find services. In addition, a process order of the single component types is needed, to simplify the task of finding the next following process step. The combination of the deployment folder and the process order enables the service-based principle. Both include location independence of the successor services, arbitrary parameter possibilities and openness for changes in the process order.

- *Data flow in the data production system*
  The investigations have shown, that the immense amount of data, which flows through this production process, is not well manageable through web services. Better techniques for data transportation are well known standards, as for example FTP (File Transport Protocol). This is comprehensible, because of needed compression algorithms and well developed methods for data handling in databases.
- *Control flow and data flow of the PCMS for supervision of the data production system*
  Both, control flow and data flow are useful operational areas of web services. Both are vertical hierarchical organised. The data flow consists of the log information. The log information has to be transmitted over the corresponding business areas to a central location. All participants can access these gathered information over ordinary web pages, to enable an overview over the whole workflow for local and central located users.
  The control flow of the PCMS reports all remarkable events in the production process to its hierarchical overlying level. In the end, the control flow is responsible for forwarding the gathered data to the central location for preparation and a well-structured view on screen, globally.
- *Extended protocol functionality*
  To record the log information is also a good usage area for web services. Once implemented, every component can use these services independent from its location.

5 LESSONS LEARNED AND RESEARCH ISSUES

The increasing complexity of such very large distributed systems causes a raise in the need for an adequate surveillance and control. Additionally, an increasing amount of involved technologies (distributed components, different databases, FTP, etc.) intensifies again the complexity of a good supervision. Therefore, approaches for a powerful control should be as easy as possible and nevertheless as complex as necessary. Further investigations in this area will follow in our research project to confirm these statements through gained facts.

There exists no standard solution for the control of all kinds of distributed systems which could be used in every case similarly. In systems with changing primary keys it is more advisable to develop an own solution like our approach. However in a goods-oriented distributed production system a standard solution like workflow management systems could be possibly used. The categorisation of when to use which control instrument is one aim of our research.

In addition, it is not the main focus of a producing industry to control their distributed systems. For example, it is not sensible to build control systems which again have to be controlled as well and hence initiate an endless recursion-chain with control systems. In the contrary to a distributed system itself, a supervision system is not a very productive system. Its only purpose to exist is to gain transparency and surveillance over the business processes. The power of our described approach is in addition to the control functionality to be able to reach also an optimisation of the production processes. Our method achieves it through a clever distribution of deadlines and priorities.

Beside the architectonical side of the approach there are several more investigations which have to be done in the near future. For example, it is interesting which kind of statistics are needed and can be achieved through these kinds of surveillances. Beside the supervision over the sequential control an historical backtracking of the content specific quality of the information contained in the data is expected so far. Also, classification numbers for the management are conceivable. The difference between the current state and a planned state within the distributed system can be calculated through the use of our approach. For example, out of it measures could be deduced for a raise in efficiency. Likewise deceivable could be the question if a raise in production could be achieved, because problems in the production process are automatically recognised, documented and can be remedied earlier. Another vision is to generate financial statements about the production. For example, processing times of the single processing steps are gathered. The exertion of resources could be handled cleverer with these facts. For the management could this mean a calculation basis for the disposition of resources substantiated through financial classification numbers. The exploration about the power and dimensions of statistics, which can be achieved knowledgeably with our approach, are following in the near future.

6 CONCLUSION

Many component based distributed systems lack clearly structured and defined supervision, due to the volume of involved components and the problem of manual management, as the system develops. This paper describes a web service based approach for the supervision of such systems.

GfK's supervisory system (PCMS - Planning, Controlling and Monitoring System), has been introduced here, and it is an excellent example for motivating and substantiating further research in this field. Early experiments with GfK's data production system have shown a lack of transparency and surveillance and thus the decision for a widely automated supervisory system for controlling and monitoring the data production system is supported. (Still in a conceptual stage).

Multifaceted benefits for PCMS users are presented in this paper. For example, the support of parallelism in the controlled system and the possibilities to distribute priorities to important jobs, demonstrate these advantages. Also, the possible composition of a PCMS has been explored. It consists in essence, of a centrally located data inquiry system and a hierarchical oriented management system with local and central parts. However, arguments are introduced as to why it is challenging to design a supervisory system (e.g. control over data packages as well as data content or changing the identification keys in addition to constant data transformation within the data production system).

After assessing the PCMS, a technological background regarding web technology followed, and in particular that the classification of web services is relevant to our approach.

A system architecture has been proposed for a suitable PCMS, which uses service and level based principles for smooth surveillance of world-wide distributed data production systems. One essential element and one of the main advantages of our approach is the integrated method for production optimisation. Thus, out of a more or less sequential processing a parallel production will be created, which can be optimally handled and meet management requirements. As a result, this is achievable through an intelligent distribution of deadlines and priorities.

Our plans for the future are to continue our experiments on our prototype. In the first step we will implement the monitoring aspect of the system. This will lead to initial results as to how effectively the data production system works without the additional automated production optimisation, and priorities. Those results will then be utilized as the basis for eventual comparison with automated optimisation after the planning functions and priority checking have been developed and finalized. Additional work must still be performed to track and evaluate the multiple levels of variation between the planned and the current states over longer time periods, in order to obtain even more consistent results. Following these steps, control possibilities for potential interventions will be created and implemented. We will continue to examine the management classification numbers (e.g. cycle times, productivity, etc.) which can be attained and determine the available options for visualising PCMS's output in a graphical way.

We expect that the development of the prototype at GfK will lead to further research issues and will help to extend our knowledge in this important area.

ACKNOWLEDGEMENTS

We are grateful to Colin Tully for excellent comments on an initial draft, and to the entire team of GfK Marketing Services for many inspiring discussions.

REFERENCES

Champine, George A., 1980, Distributed Computer Systems, chapter 2,3,8,9,10,13, Impact on Management, Design and Analyses, North-Holland Publishing Company.

Dwyer, Tom, 2000,e-Business Integration Drives EAI, An Interview with Aberdeen's Tom Dwyer, [Online], Available: http://www.eaijournal.com/PDF/Analyst_Perspective-July_1.pdf, eAI Journal, [2002,Aug, 28].

Farley, Jim, 2000, .NET from the Enterprise Perspective: SOAP [Online], Available: http://java.oreilly.com/news/soap_0900.html.

GfK Marketing Services, 2002, [Online], Available: http://www.gfkms.com, [2002, Aug., 23].

Gomer, Thomas, et al, 1990, Heterogeneous Distributed Database Systems for Production Use, ACM Computing Surveys, vol.22, no.3.

Guy Eddon and Henry Eddon, 1998, Inside Distributed COM, Microsoft Press, Redmond, Washington 98052-6399

Java, 2002, 'Java and Corba', [Online], Available: http://www.java.sun.com, [2002, Apr.13].

Linnhoff-Popien, Claudia and Heinz-Gerd Hegering (Eds.), 2000, Trends in Distributed Systems: Towards a Universal Service Market, Third International IFIP/GI Working Conference, USM 2000, Preface, Sept. 2000, Lecture Notes of Computing Science 1890, Springer Verlag, Heidelberg.

Linthicum, David S., 2001, 'B2B Application Integration, e-Business-Enable Your Enterprise', Addison-Wesley.

Manes, Anne Thomas, Sun Microsystems, Inc., 2001, Enabling Open, Interoperable, and Smart Web Services, [Online], Available: http://www.w3.org/2001/03/WSWS-popa/paper29, [2002, Aug., 18].

Microsoft COM-Technology, 1998, 'Microsoft Component Service, A Technical Overview', [Online], Available: http://msdn.microsoft.com/library, [2002, Jun. 09].

Mullender, Sape, 1993, Distributed Systems, chapter 1,2, Addison-Wesley, ACM Press.

OASIS, UDDI-Universal Description, Discovery and Integration of Web Services, [Online], Available: http://www.uddi.org, [2002,Dez.,31].

Pattison, Ted, 2000, Programming Distributed Applications with COM+ and Visual Basic 6.0, 2nd edn, Microsoft Press, Redmond, Washingtion 98052-6399

Schanzenberger, Anja, Colin Tully and Dave R. Lawrence, 2003, Überwachung von Aggregationszuständen in verteilten komponentenbasierten Datenproduktionssystemen, 10. GI-Fachtagung Datenbanksysteme für Business, Technologie und Web, BTW 2003, Lecture Notes in Informatics, Koelln Druck Verlag

Sharp, John A., 1987, An Introduction to Distributed and Parallel Processing, chapter 5,6,7, Blackwell Scientific Publications.

SOAP, 2001, SOAP 1.1 W3C, [Online], Available: http://www.w3.org/TR/#Notes, [2001, Dez. 20].

Sun Microsystems, 2003, Homepage, [Online], Available: http://www.sun.com, [2003, Jan. 05]

Vasudevan, Venu, 2001, "Web Services Primer", [Online], Available: http://www.xml.com/pub/a/04/04/2001/webservices/index.html[2002, Aug., 18].

W3C, 2002, Web Services Activity, [Online], Available: http://www.w3.org/2002/ws, [2002, Aug., 18].

Web services, 2002, [Online], Available: http://www.webservices.org, [2002, Mai, 31].

WSDL, 2002, WSDL 1.2 W3C, [Online], Available: http://www.w3.org/TR/2002/WD-wsdl12-20020709/, [2002, Aug. 28].

XML, 2001, W3C XML, [Online], Available: http://www.w3.org/TR, [2001, Dez. 20].

MODELLING OVERHEAD IN JAVASPACES

Frederic Hancke
Gunther Stuer
David Dewolfs
Jan Broeckhove
Frans Arickx
Tom Dhaene
Department of Mathematics and Computer Science
University of Antwerp
2020, Antwerp
Belgium
frederic.hancke@ua.ac.be

KEYWORDS
Distributed Computing, JavaSpaces, Performance Modelling.

ABSTRACT

In this paper a theoretical model is developed to compare different distributed platforms on their performance in computational problems. In order to compare different platforms in a heterogeneous environment on the same basis, the use of tasklists in XML is introduced. In this paper some initial results of the performance and overhead of JavaSpaces are presented. The resulting data set of the tests is statistically analyzed using the presented theoretical model, leading to a first conclusion on the use of JavaSpaces for computational problems. In a first stage, the data set is investigated on the presence of outliers, and in a second stage some basic statistics are applied.

INTRODUCTION

In the research group CoMP (*Computational Modelling and Programming*) research is done in the field of computational science, aimed at understanding physics and engineering problems through modern modelling techniques, using new software development paradigms and advanced mathematical techniques.

Many problems in the area of quantum physics often involve very intense and large computations. Therefore distributed platforms, such as JavaSpaces [FHA99], MPICH [MPI] (an MPI [GES99] implementation), etc. could be very useful. The goal of our project is to find out what platform is most suited for a certain kind and number of calculations. Thus, to classify computationally large problems.

First of all, we are interested in the possibilities of some existing platforms. In this paper, we consider a Linda Tuple Space [Yal] implementation, JavaSpaces. This includes the platforms architecture, functionality and performance against a theoretical model.

Next, the platform needs to be tested with a wide variety of fictitious, though representative, problems, each divided in subproblems so that the calculations can be distributed. As we don't want the testcases to be problem specific, we consider tasks that simulate execution time only.

Finally, the results of a first test of JavaSpaces against this theoretical model should provide a comparison of different distributed platforms, which is the main goal.

DISTRIBUTED SYSTEMS

In distributed systems, three kinds of models can be distinguished: the *push* model, the *pull* model and the *push-pull* model. In the discussion of these models in the rest of this section X will refer to the entity having some tasks to be performed and Y to the entity performing the tasks.

The push model is used when each X distributes tasks to specific chosen Ys. The pull model is just the opposite: each Y scans the tasks waiting to be performed at their respective X and performs the tasks it can. When an X pushes its tasks into a medium where they are pulled to be performed by a Y, we speak of a push-pull model. A visualization of the three models is shown in figure 1.

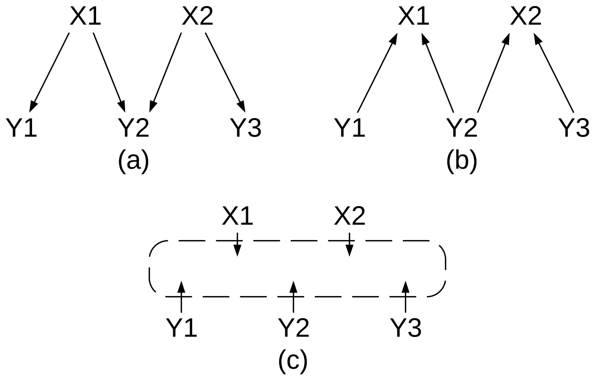


Figure 1: $X1$ has workers $Y1$ and $Y2$. $X2$ has workers $Y2$ and $Y3$. (a) push model; (b) pull model; (c) push-pull model

The *farmer-worker* model is hierarchically situated one level higher than the models discussed above. The rule is simple: there is one *master* (the farmer) and one or more *slaves* (the workers) doing the jobs the farmer orders them to do. Orders are given in the form of a *task description*, the result of the task in a *result* *description*. The farmer-worker model can thus be implemented as a full push model, such as MPI, or as a push-pull model, such as the Linda Spaces implementations. The pull model could also be used, but in this case the farmer would not be the master anymore.

THEORETICAL MODEL BUNDCHEN

The main reason for constructing a theoretical model, where the aim is to obtain a formula for calculating the total time needed to solve a given problem, is to compare the different distributed platforms with each other quantitatively. JavaSpaces could be compared to, say MPICH, but the comparison would only be *between* the two. If a third platform is included in the test, this should be compared to both JavaSpaces and MPICH to decide what is better in which situation. The central theoretical model proposed to allow for such comparison, will be referred to as *Bundchen*.

Building this theoretical model can be done in a number of steps. At each step a number of parameters is added.

In the first step, the only model parameters are:

- the number of workers N_{work},
- the number of tasks N_{task}, and
- the calculation time (in ms) of one task on an ideal system T_{task}.

Thus, supposing every task requires the same calculation time, a first step to a general formula for the total time needed could be written as

$$T_1 = \lceil \frac{N_{task}}{N_{work}} \rceil T_{task} \tag{1}$$

In the second step we also take into account:

- the communication overhead (in $bytes$) of one task description C_{task},
- equally, the communication overhead (in $bytes$) of one result description C_{res},
- the speed (in $bytes/s$) of the network infrastructure S_{net}, and
- the load of the network infrastructure L_{net} ($0 \leq L_{net} \leq 1$).

Supposing the first two parameters remain constant for all task and result descriptions, the total time now becomes

$$T_2 = T_1 + \frac{N_{task}}{S_{net}(1 - L_{net})}(C_{task} + C_{res}) \tag{2}$$

Of course, this model is still not satisfying. The aim is to have a formula or algorithm that calculates the minimum time spent on the job with the best possible distribution of tasks. Therefore we need to consider more parameters, such as

- the speed (measured against the speed of an ideal machine ($S_{ideal} = 1$)) of the farmers processor $S_{proc,0}$ ($0 \leq S_{proc,0} \leq 1$),
- the load of the farmers processor $L_{proc,0}$ ($0 \leq L_{proc,0} \leq 1$),
- the speed (also measured against the speed of an ideal machine) of each workers processor $S_{proc,w}$ ($0 \leq S_{proc,w} \leq 1$ and $1 \leq w \leq N_{work}$), and
- the load of each workers processor $L_{proc,w}$ ($0 \leq L_{proc,w} \leq 1$ and $1 \leq w \leq N_{work}$).

The meaning of these parameters is also shown in figure 2.

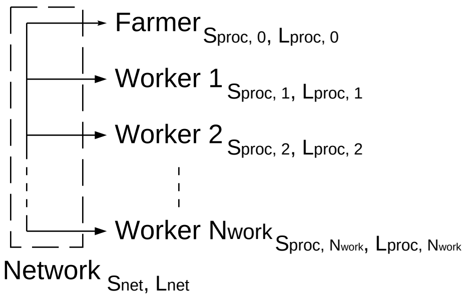

Farmer Sproc, 0, Lproc, 0
Worker 1 Sproc, 1, Lproc, 1
Worker 2 Sproc, 2, Lproc, 2
Worker Nwork Sproc, Nwork, Lproc, Nwork
Network Snet, Lnet

Figure 2: S_{net}, L_{net}, $S_{proc,w}$ and $L_{proc,w}$ ($0 \leq w \leq N_{work}$)

Finally, the above assumptions have to be cleared out:

- not every task generally requires the same amount of time to be computed, thus T_{task} becomes task dependent $T_{task,t}$ ($1 \leq t \leq N_{task}$),
- the task description can not be assumed constant for all tasks, thus C_{task} becomes $C_{task,t}$ ($1 \leq t \leq N_{task}$), and
- equally, the result description can not be assumed constant for all tasks, thus C_{res} becomes $C_{res,t}$ ($1 \leq t \leq N_{task}$).

To include the latter parameters into our proposed model, Bundchen, and clearing out the above assumptions, something stronger than a formula is needed. This means that an algorithm should do the job of distributing the tasks perfectly among the available workers, hereby using the knowledge of the speeds and loads of the available workers and the computational complexity of the tasks. Thus, an optimization problem replaces the simple formula.

Although different algorithms exist for scheduling loads [BGMR96], the easiest, but surely the slowest, way to solve the problem, is just to check all possible solutions. The one that then returns the smallest value is the one we need. The number of possibilities (with $N_{work} > 1$) is given by

$$N_{poss} = \sum_{i_1=0}^{N_{task}} \sum_{i_2=0}^{N_{task}-i_1} \sum_{i_3=0}^{N_{task}-(i_1+i_2)} \ldots$$

$$\sum_{i_j=0}^{N_{task}-\sum_{k_j=1}^{j-1} i_{k_j}} \ldots \sum_{i_{N_{work}}=0}^{N_{task}-\sum_{k_{N_{work}}=1}^{N_{work}-1} i_{k_{N_{work}}}}$$

$$\left(\begin{array}{c} N_{task} \\ i_1 \end{array} \right) \left(\begin{array}{c} N_{task} - i_1 \\ i_2 \end{array} \right) \left(\begin{array}{c} N_{task} - (i_1 + i_2) \\ i_3 \end{array} \right) \ldots$$

$$\left(\begin{array}{c} N_{task} - \sum_{l_j=1}^{j-1} i_{l_j} \\ i_j \end{array} \right) \ldots$$

$$\left(\begin{array}{c} N_{task} - \sum_{l_{N_{work}}=1}^{N_{work}-1} i_{l_{N_{work}}} \\ i_{N_{work}} \end{array} \right)$$

with $3 < j < N_{work}$, and where

$$\forall n, j \in \mathbb{N}_0, n \geq j : \left(\begin{array}{c} n \\ j \end{array} \right) = \frac{n!}{j!(n-j)!}$$

is the binomial of Newton.

Clearly, this approach is rather simplistic. The algorithm does not distribute the tasks as any existing platform would do. The goal of the algorithm is to distribute tasks efficiently using its explicit properties. Distribution platforms that offer resource management for distributing the tasks will of course perform better against this model.

XML TASKLIST GENERATOR

Instead of testing a platform with real computationally complex problems, worker processes simulate task execution time only and are implemented by a *sleep*. This way it is easier to compare different platforms with Bundchen and the pool of workers becomes almost perfectly homogeneous. So, each task description takes one value that represents the duration of the worker to sleep. A tasklist consists of a number of such task descriptions where each tasks duration is randomly chosen using a probability distribution.

To implement this idea, a program was written to generate these tasklists in XML [XML, McL01] format. This was done for two reasons:

1. as a storage medium for tasklists, and
2. to be able to easily use the same tasklists over and over again in different settings of different testcases.

The XML tasklist generator currently has two input parameters: the number of task durations to generate and the probability function to use. The currect version supports two types of probability distributions: the constant and the normal, or *Gauss*, distribution. Depending on this choice the program needs one extra parameter m for the former — the duration m for all tasks — or two extra parameters m (the mean of the normal) and sd (the standard deviation of the normal) for the latter distribution.

In future versions of the generator more distributions and functionality will be added. The tasklists will also be stored in an XML database as well as logfiles of the results of tests.

TESTCASE FOR JAVASPACES

The aim of this testcase is to find out whether JavaSpaces is a good potential candidate to solve high performance calculations, such as the quantum physics problems mentioned in the introduction. Other platforms such as MPICH, TSpaces [TSp, Wyc98] and GigaSpaces [Gig] will be tested the same way in the near future. Besides the platforms functionality, we are also interested in its performance measured against Bundchen.

JavaSpaces

JavaSpaces is one of many implementations of the so called Linda Spaces distributions concept. The underlying idea is that objects can be thrown into a virtual space and taken out, or simply read, by any object connected with the space. Many distributed platforms have been built using this idea. Other implementations, besides JavaSpaces, are TSpaces and GigaSpaces.

JavaSpaces was built on top of Jini [Jin, Edw01] as a service of the Jini technology. Its functionality was kept very simple, but nevertheless is very powerful. In fact there are only three basic actions on the JavaSpace itself:

- **write**: to write an object *into* the space,
- **read**: to read an object from the space, but leaving the object *in* the space, and
- **take**: to take an object *out* of the space.

There is a fourth operation possible, but this one does not really perform on the space: **notify**, which notifies an object of objects being added to the space.

Setup

Table 1 shows the machines that have been used to perform the tests, as well as their characteristics.

PC name	Processor	OS	Java	Jini
smurf	Intel PII 400	SuSE Linux 7.3	1.4.1	1.2.1
drone1of1	Intel PIV 1.7	SuSE Linux 8.0	1.4.1	1.2.1
drone2of1	Intel PIV 1.7	SuSE Linux 8.0	1.4.1	1.2.1
drone3of1	Intel PIV 1.7	SuSE Linux 8.0	1.4.1	1.2.1
drone4of1	Intel PIV 1.7	SuSE Linux 8.0	1.4.1	1.2.1

Table 1: Testcase setup

The HTTP server, RMI Activation Daemon, Lookup service, the JavaSpace service and the farmer were run on *smurf*. Each of the four workers was run on a separate *drone*.

Measures

The tests have covered different values (in the near future, tests will be performed with more different values for w and t) for four different parameters. These are:

- w: the number of workers ($w \in \{1, 2, 3, 4\}$),

- t: the number of tasks ($t \in \{10, 50, 100\}$),
- m: the mean (in ms) of the Gauss distribution for tasks ($m \in \{1, 10, 100, 500, 1000, 2000, 5000, 10000\}$), and
- v: the standard deviation (as a percentage of m; $0 \leq v \leq 100$) of the Gauss distribution for tasks ($v \in \{1, 2, 5, 10, 20, 30\}$).

Thus, for every combination of t, m and v, an XML file has been generated, except for the combinations of $m = 1$ for all v, and $m = 10$ with $v \in \{1, 2, 5\}$. For these combinations mv becomes real. Instead, XML files for $m = 1$ and $m = 10$ were generated with task duration m for all tasks. So a constant distribution was used instead of the normal distribution.

This means 123 XML files were generated. Execution of the farmer-worker process with a given XML file resulted in one value representing the duration wct (wallclock time) of the whole process. Every test ran 10 times for each XML file and for each w, yielding 4920 data points.

Statistical Analysis

As the data set for evaluating JavaSpaces is, as mentioned in the previous subsection, not yet complete, we will give a first brief analysis of it in this subsection.

First, the mean duration of all tasks in one tasklist (XML file) was calculated. This was done to prevent using the theoretical mean that was used to generate the tasklist, because this would corrupt further calculations. Formula (1) of Bundchen was used to get a first indication of the performance of JavaSpaces. In the rest of this section we work with the overhead of distributing one task in JavaSpaces, which is given by

$$\frac{wct - T_1}{t}$$

As robustness of data sets [HMT00] is not self-evident, we first took out a number of potential outliers. *Boxplots* and *Stem-and-Leaf Plots* are two possible methods to identify potential outliers. Each of these does not necessarily produce the same results. The decision which outliers will finally be discarded, is up to the user. Each potential outlier must be investigated carefully and may only be discarded if there is a good reason. In our case, occasionally high network or processor load might be good reasons.

The method we used is that of boxplots. Using these, one can distinguish potential *mild* outliers from potential *extreme* outliers. A data point is marked as a mild outlier if [Wei02]

$$d < Q1 - 1.5IQR$$

or

$$d > Q3 + 1.5IQR$$

with d the value of the data point, $Q1$ and $Q3$ respectively the first and the third quartile and $IQR = Q3 - Q1$ the interquartile range. A data point is marked as an extreme outlier if (remark that extreme outliers are also mild outliers)

$$d < Q1 - 3IQR$$

or

$$d > Q3 + 3IQR$$

This technique was applied on our data set. Table 2 shows a comparison of the resulting statistics. The table shows a significant difference between the mean, standard deviation, minimum and maximum using the complete data set (ALL) and using the complete data set discarding extreme outliers (ALL - EO). The median, $Q1$ and $Q3$ shrink only slightly, which means that the complete data set was corrupted by only a few (171) data points. The difference between discarding extreme outliers and discarding mild outliers confirms this. As it is better to discard as few data points as possible, we will use the complete data set discarding only the extreme outliers in the rest of this subsection.

	ALL	ALL - EO	ALL - MO
N	4920	4749	4589
N (%)	100.00	96.52	93.27
Mean	50.5536	41.1575	39.8202
Median	40.0600	39.2000	38.7000
Std. Deviation	184.1168	30.8262	26.5052
Minimum	-559.60	-75.00	-34.30
Maximum	11859.77	168.03	113.20
$Q1$	20.3550	20.1300	20.2450
$Q3$	57.4950	56.1300	55.0000

Table 2: Statistics on the overhead (in ms) using JavaSpaces for distributing one task, using respectively the complete data set (ALL), the complete data set discarding extreme outliers (ALL - EO) and finally the complete data set discarding mild outliers (ALL - MO)

The graph for the overhead using JavaSpaces for distributing one task is shown in figure 3. The mean is marked with a solid line at $41.1575ms$. The graph is divided in four columns grouped by the number of workers. So, from left to right we have first the data points based on 1 worker, then 2 workers, etc.

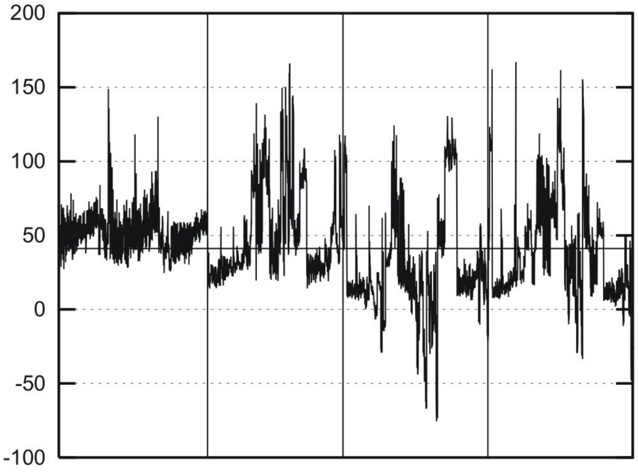


Figure 3: Graph representing the overhead using JavaSpaces for distributing one task. The X axis represents the data points, the Y axis the overhead (in ms)

There are still a few problems with this visualization. As mentioned earlier, the graph does not represent data measured with more than 4 workers. Another, more important, problem is the fact that some data points in the graph have negative values, which influence the mean badly. The reason for these negative values lies in the proposed theoretical model. Indeed, the first, simplest, formula for Bundchen was used, which supposes all tasks to have the same duration (the mean of the tasklist, not the theoretical mean). In the calculations with 3 or more workers, the situation in figure 4 may occur.

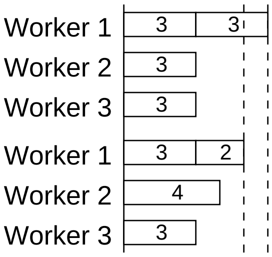


Figure 4: The problem with using formula (1) of Bundchen for 3 or more workers. The upper half represents the way Bundchen distributes the tasks, the lower half the way it could be distributed by JavaSpaces

Suppose 4 tasks must be distributed with each a length of respectively 3, 4, 3 and 2. The mean of this tasklist is clearly 3. So, Bundchen will theoretically distribute 4 tasks of length 3. Thus, situations in which tasks are distributed in such a way that they outperform Bundchen, might occur. However, using the complete Bundchen algorithm, which has not been fully implemented as yet, should solve this problem.

CONCLUSION

In this paper a theoretical model, called Bundchen, was introduced. This should provide for a better way to compare different distributed platforms on their performances for computational problems, ranging from simple ones to very complex ones. In the near future, the implementation of Bundchens algorithm will be completed.

The XML tasklist generator, which is used to generate tasklists in XML format using a given distribution, will also be updated with much more functionality. A second XML file format will be developed in the near future to store the complete *result* set generated during tests. This will allow the use of an XML database for both tasklists and result sets.

When more detailed data sets will become available, more thoroughgoing analysis tools will be used, such as analysis of variance (ANOVA), to obtain better predictions of the overhead.

REFERENCES

[BGMR96] Veeravalli Bharadwaj, Debasish Ghose, Venkataraman Mani, and Thomas G. Robertazzi. *Scheduling Divisible Loads in Parallel and Distributed Systems*. IEEE Computer Society Press, 1996.

[Edw01] W. Keith Edwards. *Core Jini Second Edition*. Prentice Hall, 2001.

[FHA99] E. Freeman, S. Hupfer, and K. Arnold. *JavaSpaces Principles, Patterns, and Practice*. Addison Wesley, 1999.

[GES99] William Gropp, Lusk Ewing, and Anthony Skjellum. *Using MPI*. The MIT Press, second edition, 1999.

[Gig] Gigaspaces. URL: http://www.j-spaces.com.

[HMT00] David C. Hoaglin, Frederick Mosteller, and John W. Tukey. *Understanding Robust and Exploratory Data Analysis*. Wiley, 2000.

[Hup00] Susanne Hupfer. The nuts and bolts of compiling and running javaspaces programs. Technical report, Sun Microsystems, Inc., 2000.

[Jin] Jini. URL: http://www.jini.org.

[McL01] Brett McLaughlin. *Java & XML*. O'Reilly & Associates, Inc., 2001.

[MPI] Mpich. URL: http://www-unix.mcs.anl.gov/ mpi/ mpich/.

[NKNW96] John Neter, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. WCB/McGraw-Hill, fourth edition, 1996.

[NZ01] Michael S. Noble and Stoyanka Zlateva. Scientific computation with javaspaces. Technical report, Harvard-Smithsonian Center For Astrophysics and Boston University, 2001.

[NZ02] Michael S. Noble and Stoyanka Zlateva. Distributed scientific computation with javaspaces. Technical report, Boston University, 2002.

[SM99] Inc. Sun Microsystems. Javaspaces: Innovative java technology that simplifies distributed application development. Technical report, Sun Microsystems, Inc., 1999.

[TSp] Tspaces. URL: http://www.almaden.ibm.com/ cs/ TSpaces/.

[Wei02] Neil A. Weiss. *Introductory Statistics*. Addison Wesley, sixth edition, 2002.

[Wyc98] P. Wyckoff. Tspaces. Technical report, IBM Almaden Research Center, 1998.

[XML] Xml. URL: http://www.xml.com.

[Yal] Yale linda group. URL: http://www.cs.yale.edu/ Linda/ linda.html.

DYNAMIC REPLICATION OF CONTENT IN THE HAMMERHEAD MULTIMEDIA SERVER

Jonathan Dukes
Jeremy Jones
Department of Computer Science
Trinity College Dublin, Ireland
Email: Jonathan.Dukes@cs.tcd.ie

KEYWORDS

Multimedia servers, server clusters, video-on-demand, group communication, replication.

ABSTRACT

In a clustered multimedia server, by periodically evaluating client demand for each file and performing selective replication of those files with the highest demand, the files can be distributed among server nodes to achieve load balancing. This technique is referred to as *dynamic replication*. Several dynamic replication policies have been proposed in the past, but to our knowledge, our work is the first to describe in detail the implementation of dynamic replication in a server cluster environment. In this paper, we describe the architecture of the HammerHead multimedia server cluster. HammerHead has been developed as a cluster-aware layer that can exist on top of existing commodity multimedia servers – our prototype takes the form of a plug-in for the multimedia server in Microsoft Windows Server 2003$^{\text{TM}}$. Replicated state information is maintained using the Ensemble group communication toolkit. We briefly describe our own dynamic replication policy, Dynamic RePacking, and its implementation in the HammerHead server. Finally, we present early performance results from a prototype version of the HammerHead server.

INTRODUCTION

The server cluster model is widely used in the implementation of web and database servers. High-availability and scalability are achieved by combining commodity hardware and software, server clustering techniques, storage technologies such as RAID and storage area networks (SANs). However, the implementation of large-scale on-demand multimedia servers in a cluster environment presents specific problems.

Advances in storage technology have made it possible for a single commodity server to supply soft real-time multimedia streams to clients across a network. Such servers, however, exhibit poor scalability and availability (Lee, 1998). One solution is to “clone” the servers, mirroring available data on each node. This approach increases the bandwidth capacity and availability of the service and is common in web server clusters, where the volume of data stored on the server is small. Cloned servers can be grouped to form a network load-balancing cluster and client requests are distributed among cluster nodes according to their capabilities and current workload. However, the volume of data that is typically stored on a multimedia server usually prohibits this form of complete server replication. In addition, since the majority of multimedia files are rarely requested by clients, replication of low-demand files is wasteful. (We use this model as a performance baseline for our server.)

The use of storage area network (SAN) technology in multimedia servers has also been investigated in the past (Guha, 1999). One approach is to provide a cluster of front-end streaming nodes with access to shared SAN storage devices, such as disks or RAID storage systems. Scalability and availability problems are, however, merely moved from front-end nodes to SAN storage devices, since the aggregate server bandwidth is likely to exceed the bandwidth of the SAN storage devices. In this case, any solution that can be applied to servers with directly attached storage devices may equally be applied to SAN architectures.

Server striping has been used in the past to share workload between multimedia servers. The concept is similar to RAID-0 (Patterson et al., 1988) – multimedia files are divided into equal size blocks, which are distributed among server nodes in a predefined order (Lee, 1998). Implicit load-balancing across server nodes is achieved, while only storing a single copy of each file. The degree of node interdependence caused by server striping is high, however, because each server node is used in parallel to supply each individual multimedia stream. Node interdependence has several disadvantages. First, the process of stream reconstruction is expensive. Secondly, as nodes are added to a server, existing content must be redistributed. Many architectures also require that each sever node has an identical hardware configuration (Bolosky et al., 1997). Finally, the failure of any single node will lead to the loss of all streams, unless redundant data is stored (Wong and Lee, 1997).

We argue that partial, selective replication of content is a more appropriate technique for distributing content among nodes in a clustered multimedia server. Client demand for individual multimedia files is periodically evaluated and this information is used to allocate a subset of the files to each node. The subsets are chosen to approximate an even distribution of server workload across all server nodes. Some files may be replicated to satisfy client demand or facilitate load balancing. The popularity of individual files will change over time, so the assignment of files to nodes must be reevaluated periodically.

Since each node can independently supply streams of the files it stores to clients, the degree of node interdependence is minimal. Thus, servers can be constructed from nodes with varying

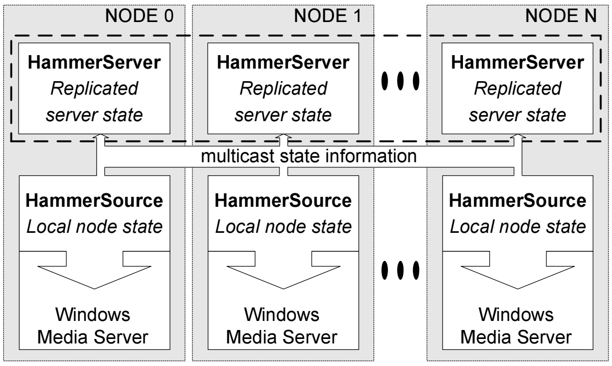


Figure 1: HammerHead Architecture

storage and bandwidth capacities and nodes an be added or removed without redistributing all existing content. Although partial replication may require more storage capacity than server striping, the additional cost is predictable and is minimised by replication policies such as the Dynamic RePacking policy implemented by our HammerHead multimedia server.

Although past research has produced other dynamic replication policies, we believe our work is the first to examine in detail the implementation of a dynamic replication policy in a cluster environment. This paper describes the architecture of the HammerHead clustered multimedia server. Rather than develop our own multimedia streaming software, Hammerhead has been designed as a cluster-aware layer that integrates with existing multimedia server implementations. Our prototype server uses a plug-in to integrate with the multimedia server in Microsoft Windows Server 2003^{TM}, although it could easily be adapted to integrate with other commodity multimedia servers. We have used the Ensemble group communication toolkit to provide reliable inter-node communication, facilitate maintenance of replicated server state and detect node failure.

In the next section, we describe the architecture of the HammerHead clustered multimedia server. We then briefly describe the implementation of our content replication policy in the context of the HammerHead server and we present performance results. We will also briefly discuss other related work.

HAMMERHEAD ARCHITECTURE

A HammerHead server cluster consists of three main components (illustrated in Figure 1).

Multimedia Server

Rather than develop our own multimedia streaming server, we have designed HammerHead as a cluster-aware layer on top of an existing commodity multimedia server. For our prototype server, we have chosen to use the multimedia server in Microsoft Windows Server 2003^{TM}.

HammerSource

This component was developed as a plug-in for the multimedia server described above, but could easily be adapted for use with other multimedia servers. It captures the state of the multimedia server (source) on a single node and publishes the state in the HammerServer layer, described below. The captured state includes the server's capabilities, properties of any stored multimedia content and the state of any client streams being supplied by the server. Each HammerSource functions independently of HammerSources on remote nodes.

HammerServer

Each HammerServer instance maintains a replica of the combined state information published by the HammerSources in the group. Client requests are initially directed to any available HammerServer, which uses the replicated server state to redirect the client request to the multimedia server on a suitable node. The choice of node will depend on the load balancing policies implemented by the server. The HammerServer is also responsible for implementing the Dynamic RePacking content placement policy described later in the paper.

It is intended that the HammerServer component will eventually present an image of a single virtual multimedia server to both streaming clients and to other geographically distributed multimedia servers. This would allow, for example, a HammerHead cluster to participate in a content distribution network (CDN).

Although Figure 1 shows a HammerServer instance on each node, this is not a requirement. Only one HammerServer instance must exist to perform client redirection and manage the HammerSources present in the cluster. Any additional HammerServers will increase the availability of the client redirection service and provide additional replicas of the global state.

Cluster Communication

We have used the Ensemble group communication toolkit (Hayden and van Renesse, 1997) to provide reliable unicast and multicast communication between HammerSources and HammerServers. Ensemble provides application programmers with mechanisms for performing reliable message-based communication between a group of processes and detecting the failure of processes in the group. Specifically, we have used the Maestro open toolkit (Vaysburd, 1998), a set of tools that runs on top of Ensemble to to allow application programmers to work with object-oriented group communication abstractions.

An implementation model facilitated by Maestro that corresponds to the HammerHead architecture described above is the *client/server model with state transfer*. In this model, group members are designated as either *server members* or *client members*. Only server members participate in the state transfer protocol. Both clients and servers can reliably send messages to any single group member, to a subset of members, to the set of all servers in the group or to the entire group. HammerServers are *server* group members and HammerSources are *client* group members.

When a HammerSource joins a group containing one or more HammerServers, it multicasts its current state information to the HammerServers in the group. Each HammerServer will then merge the HammerSource's state into the global server state. When an event (for example, the start of a stream or the addition of a new file) causes a change in the state of a HammerSource, a message describing the event is multicast to the set of HammerServers, which will update the global state accordingly.

Similarly, if after executing the Dynamic RePacking content placement policy, a HammerServer determines that a file must be copied or removed, it uses Ensemble to send a message to the corresponding HammerSource, which is then responsible for making the required changes.

When a HammerServer joins a group already containing one or more HammerServers, the Maestro state transfer protocol is initiated. During the execution of the state transfer protocol, Maestro will prompt one of the existing HammerServers to send its state to the new server. When this state information has been received and stored, the new server informs Maestro that the state transfer is complete. Any messages sent during the state transfer, which are not related to the state transfer itself, will be delayed and sent when the transfer is complete. Thus, when a new HammerServer joins the group, every HammerServer, including the new one, will contain the same server state and will receive exactly the same set of messages, resulting in a consistent replicated state.

It is worth noting that to maintain a reasonable level of performance and avoid locking the server state for long periods, HammerHead uses an asynchronous communication model. For example, when a HammerServer sends a request to a HammerSource to obtain a copy of a file, it does not wait for a response. Instead, the server will only see the result of the request when the source informs it of the presence of the new file.

In the following section, we describe how the HammerHead server cluster implements the Dynamic RePacking content placement policy.

DYNAMIC REPACKING

The HammerServer component performs three related functions: evaluation of client demand for each file on the server, assignment of a subset of those files to each node and redirection of client requests to suitable nodes for streaming.

Demand Evaluation

Usually information about the demand for individual files will not be available from an external source, and needs to be evaluated by the HammerServer. When evaluating the demand for a file, several variables need to be considered:

Server load Changes in overall server utilisation may occur on a short-term basis (e.g. between morning and evening). Usually a dynamic replication policy should ignore these fluctuations and only respond to longer-term changes in the relative demand for files, for example, over several days. For this reason, we define the demand, D_i, for a file i as the *proportion* of server bandwidth required to supply the average number of concurrent streams of that file over a given period, which is independent of the actual load on the server.

File popularity The relative popularity of individual multimedia files will change over time, often decreasing as files become older, resulting in changes in the relative demand, D_i, of files. The server should adapt quickly to these changes.

Stream duration The average duration of streams of different files may vary and our definition of demand takes this into account by evaluating the average number of concurrent streams of each file.

Bit-rate The multimedia files stored on a server will usually have different bit-rates, depending for example on the media type (video, audio, etc.) and the level of compression used. Again, our definition of demand takes bit-rate into account.

To evaluate the demand for a file, the HammerServer needs to calculate the average number of concurrent streams of the file, over a period of time of length τ. This value, L_i, for a file i, can be calculated by applying Little's formula (Little, 1961) ($L_i = \lambda_i . W_i$). The arrival rate, λ_i, can be calculated by dividing the number of requests for the file, received over a period, by the length of the period, τ. Similarly, the average stream duration, W_i, can be calculated by dividing the cumulative duration of all streams of the file, Q_i, by the number of requests. Substituting for λ_i and W_i in Little's formula, L_i may be evaluated as follows:

$$L_i = \frac{Q_i}{\tau}$$

To calculate D_i for each file, L_i is scaled by the bandwidth, B_i, required to supply a single stream of the file. The following formula can then be used to express the demand for a file as the proportion of server bandwidth required by the expected number of concurrent streams of the file:

$$D_i = \frac{Q_i . B_i}{\sum_{k=0}^{K-1} (Q_k . B_k)}$$

where K is the number of files stored on the server. Thus, the only measurement required by a HammerServer to evaluate the demand for each file is the cumulative duration of all streams of each file, Q_i, over the period τ. The value of D_i for each file can be used at the end of each period as the input to the file assignment algorithm. To obtain an average value over a longer period, a "sliding window" approach is used, where the demand for the last T measurement periods is saved. The average demand for a file i, D'_i, is estimated by taking a weighted average of all of the measurements of demand in the sliding window.

Each time a stream of a file begins on a node, the HammerSource on that node will multicast information about the new stream to the set of HammerServers. Similarly, when the stream ends, the HammerSource will multicast an end-of-stream message to each HammerServer. When a stream ends, each HammerServer will calculate the duration of the stream and add it to the cumulative stream duration for the corresponding file. At the end of each period, τ, each HammerServer uses the cumulative stream duration of each file to estimate the demand using the technique described above. This information is then used to assign files to nodes, replicating files if necessary, as described in the next section.

File Assignment

In this section, we give a brief overview of the Dynamic RePacking file assignment algorithm. A more comprehensive description can be found in our prior work on Dynamic RePacking (Dukes and Jones, 2002). Our policy is based on a number of modifications to the MMPacking policy described in (Serpanos et al., 1998). We have modified MMPacking to handle nodes with varying bandwidths and storage capacities. We have also modified the original algorithm to reduce the cost of adapting the server to changes in client demand. For example, even a small change in demand can cause MMPacking to move every file from one node to another. In contrast, our algorithm attempts to repack files where they were previously located.

Consider a server which is to store K multimedia files, $M_0 \ldots M_{K-1}$ on N nodes, $S_0 \ldots S_{N-1}$. Each file M_i has demand D_i, each node S_j has bandwidth B_j and it is assumed that $K \gg N$.

We begin by evaluating the *target cumulative demand*, G_j for each node j, which represents the bandwidth of the node as a proportion of the total server bandwidth:

$$G_j = \frac{B_j}{\sum_{n=0}^{N-1} B_n}$$

We then define for each node the target shortfall, H_j, which is the difference between the cumulative demand for the files packed on the node and the target cumulative demand. As the packing algorithm executes and files are assigned to nodes, this value will decrease. Formally, if the cumulative demand for the files assigned to node j is C_j, then the target shortfall is $H_j = G_i - C_j$.

The assignment of files to nodes takes place in rounds. The nodes are sorted by *decreasing* target shortfall, H_j^l, at the beginning of each round $l + 1$ of assignments and the files are initially sorted by *increasing* demand. During each round of file assignments, we begin with the first node. The file selected for assignment to that node is the first file in the list *that was previously stored on the same node*. If no suitable file exists, then the first file on the list is assigned to the node. After assigning a file to a node during round l, the target shortfall, H_j^l, of the node is reduced by the demand associated with the assigned file. If assigning the file to the node completely satisfies its demand, the file is removed from the list of files remaining to be assigned. However, if the demand for the file exceeds the target shortfall for the node, the demand for the file is decreased by the node's target shortfall and the file remains on the list. In this case, the node is removed from the list of nodes, preventing further assignment of files to that node. The assignment of files to nodes is illustrated in Figure 2. A round of assignments ends if the target shortfall of the current node is still greater than the target shortfall of the next node in the list, or if the end of the list has been reached. This is required to force replication for only the most demanding files.

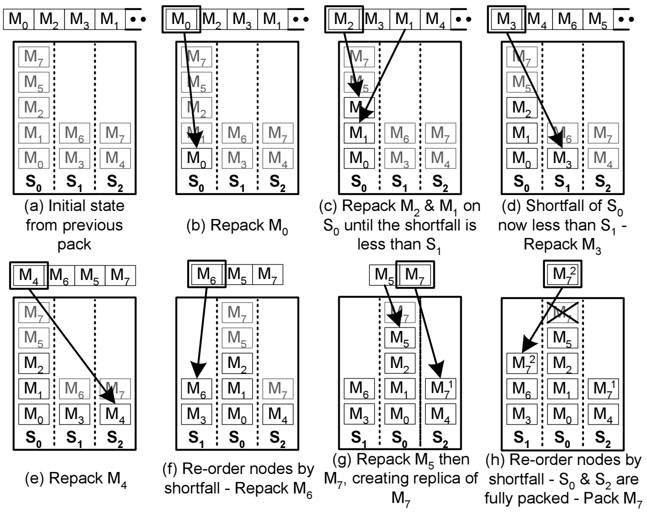

(a) Initial state from previous pack
(b) Repack M_0
(c) Repack M_2 & M_1 on S_0 until the shortfall is less than S_1
(d) Shortfall of S_0 now less than S_1 - Repack M_3
(e) Repack M_4
(f) Re-order nodes by shortfall - Repack M_6
(g) Repack M_5 then M_7, creating replica of M_7
(h) Re-order nodes by shortfall - S_0 & S_2 are fully packed - Pack M_7

Figure 2: Basic Dynamic RePacking Algorithm

Every period of length τ, the Dynamic RePacking algorithm described above needs to be executed by one of the HammerServers to determine if any change is required to the current assignment of files to nodes. Since each HammerServer will see the same ordered list of members of the Maestro server group, we simply select the first member in the group to evaluate the file assignment. The algorithm is executed using a snapshot of the replicated server state, resulting in a list of files to be either removed from nodes or added to them. A list of required changes is sent to each HammerSource.

Client Request Redirection

To perform load-balancing, clients must be directed to the least-loaded node in the cluster. Our prototype server uses the RTSP protocol (Schulzrinne et al., 1998), making the redirection of requests trivial. When an RTSP request is received, the most suitable node to supply the stream is evaluated and the HammerServer sends an RTSP redirect response back to the client, redirecting it to the selected node. The initial RTSP requests are distributed among HammerServer instances using a commodity network load-balancing solution.

PERFORMANCE

In this section, we provide early results from our prototype HammerHead server implementation. Detailed results of a simulation of the Dynamic RePacking policy can be found in our previous work (Dukes and Jones, 2002).

To perform the tests, we constructed a four node cluster. A workload generator was developed to generate client requests and receive the resulting streams. Request inter-arrival times were generated from an exponential distribution. Individual files were requested with a frequency determined from a Zipf distribution with the parameter $\theta = 0$ as described by Chou et al (Chou et al., 2000).

Table 1: Server Test Parameters

	Test 1	Test 2	Test 3	Test 4
Number of nodes (N)	4			
Number of files (K)	100			
File size (MB)	3.1			
File bit-rate (Kbps)	266			
Stream duration (seconds)	94			
Node storage (MB)	unrestricted			
Node bandwidth (Kbps)	25600			
Evaluation period (seconds)	200			
Number of sliding window periods	8			
Simulation time (hours)	2		10	
Popularity rotation period (hours)	–		1	
Requests per hour	6000			

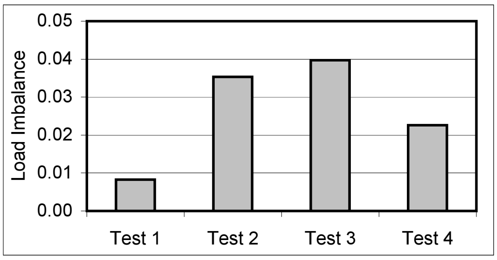


Figure 3: Server Load Imbalance

We performed four tests. Each test used a set of 100 uniform files, with identical duration, bit-rate and size. In the first test, we placed a copy of every file on every node and did not perform Dynamic RePacking. This allowed us to determine a performance baseline for future tests. In the second test, the set of files was assigned to the server nodes in round-robin order and Dynamic RePacking was enabled. In the third test, we periodically rotated the popularity of the files, so the least popular file would suddenly become the most popular, and the most popular files would gradually become less popular. This test was designed to examine the server's response to changing client behaviour. For our final test, we repeated test three, but altered the Dynamic RePacking policy to create a minimum of two copies of the ten most popular files, giving greater flexibility to perform load balancing at the expense of increased storage utilisation. The parameters for each of the four tests are summarised in table 1.

In each of the tests, we recorded the bandwidth utilisation of each node every 15 seconds. For each set of samples, we used the standard deviation of the bandwidth utilisation as an expression of the degree of load-balancing at the time when the samples were taken. The average of this value over the duration of each test is shown in Figure 3. Figure 4 shows the average storage requirement (in units of files) for each test.

Test one achieves the best load-balancing, but at the expense of a high storage capacity overhead, since each node stores the entire set of files. Since most client requests are for a small number of popular files, replicating the remaining unpopular files is wasteful. Test one is our performance baseline.

In test two, with Dynamic RePacking enabled, the load imbalance is only slightly higher than that for the baseline configuration. On average, there was a standard deviation of 3.5% in node utilisation, compared with 0.8% for the baseline. The advantage of selective replication is clear from Figure 4 – the storage capacity required for test two was approximately 25% of that for test one, since only a small number of files are replicated.

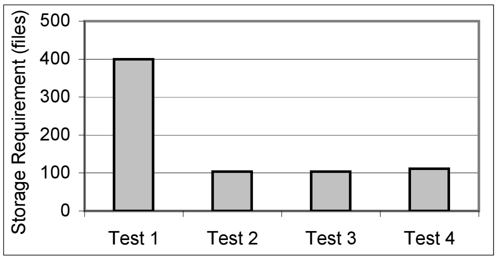


Figure 4: Server Storage Requirement

The load imbalance for test three is slightly higher than that for test two (4.0%), illustrating the cost of adapting the server to changes in file popularity over time. The storage capacity requirement is similar to that for test two.

Finally, in test four, where we created at least two copies of the most popular 10% of the files, the load-balancing is significantly improved, with an average 2.2% standard deviation in node utilisation. This improvement is at the expense of only a marginal increase in the average storage requirement.

In summary, Dynamic RePacking, as implemented in the HammerHead server, achieves a level of load balancing that is only slightly less than the ideal baseline performance, while significantly reducing the storage requirement. The storage cost for the baseline configuration is proportional to $N \times K$, where N is the number of cluster nodes and K is the number of files. The minimum storage cost, which may be achieved using a cluster based on striping without fault-tolerance, is proportional to K. Both the tests described here and a simulation study of our Dynamic RePacking policy (Dukes and Jones, 2002) have shown that the storage cost for our policy is approximately proportional to $K + 2N$, which is significantly better than the baseline configuration.

RELATED WORK

Several dynamic replication policies have been proposed in the past, however, we feel that none of these existing policies were suitable for the HammerHead multimedia server cluster. The DASD dancing policy (Wolf et al., 1995) attempts to assign file to nodes such that the ability of the server to move active streams from one node to another is maximised, without performing accurate static load-balancing based on expected demand. The Bandwidth to Space Ratio (BSR) policy (Dan and Sitaram, 1995b) attempts to balance the ratio of used bandwidth to used storage across server nodes. Our policy assumes bandwidth alone is the primary constraint and storage capacity is only considered as a secondary constraint. Another existing policy (Venkatasubramanian and Ramanathan, 1997) creates the most replicas of the least demanding files. In contrast, our policy creates the most replicas of the most demanding files. Intuitively, this increases the ability of the server to perform load balancing as client requests arrive. Another policy, called Dynamic Segment Replication (DSR) (Dan and Sitaram, 1995a) creates partial replicas of files and may be used to complement our Dynamic RePacking policy. Another threshold-based replication policy (Chou et al., 2000) uses threshold values to determine if replication is required for a particular file. Our simulation results from a previous study (Dukes and Jones, 2002), however, suggest that this leads to higher storage utilisation than our policy. Finally, the MMPacking policy (Serpanos et al., 1998) has been used as the basis for our Dynamic RePacking policy.

Group communication systems have been used before to manage multimedia server clusters (Anker et al., 1999). However, the focus of this work was on fault-tolerance, rather than replication and load-balancing. This work would complement our own.

CONCLUSIONS

Software for providing multimedia streaming services using commodity hardware is widely available. However, the use of such software in a cluster environment presents us with specific problems. If we assume that replicating all multimedia content on every server node will be prohibitively expensive, then we need to distribute the content among the nodes, without compromising the scalability and availability of the server.

In this paper, we have described the HammerHead multimedia server cluster and, in particular, its use of the Dynamic RePacking policy to perform selective replication of content and provide load-balancing. The HammerHead architecture has been designed to integrate with existing multimedia server software, providing a cluster-aware layer which is responsible for estimating the demand for each file, distributing content among server nodes and redirecting client requests to suitable server nodes. We have used the Ensemble group communication toolkit to provide reliable communication between server components and cluster membership detection.

Test results indicate that the HammerHead server compares favourably with our chosen baseline – a server cluster that replicates all content on every node. We significantly reduced the storage cost for the server, with only a small increase in load-imbalance. The performance results presented in this paper represent preliminary results and further, extensive performance analysis is required.

We are also undertaking further development of the server. In particular, we are investigating server clusters with more complex storage hierarchies, how these hierarchies might be reflected in the server state information and the resulting impact on the use of Dynamic RePacking.

ACKNOWLEDGEMENTS

The research described in this paper is part funded by Microsoft Research, Cambridge, UK.

REFERENCES

Anker T., Dolev D. and Keidar I., 1999, “Fault Tolerant Video on Demand Services”. In *Proceedings of the 19th International Conference on Distributed Computing Systems*, Austin, Texas, USA.

Bolosky W.J., Fitzgerald R.P. and Douceur J.R., 1997, “Distributed Schedule Management in the Tiger Video Fileserver”. In *Proceedings of the Sixteenth ACM Symposium on Operating System Principles*, Saint-Malo, France, 212–223.

Chou C., Golubchik L. and Lui J., 2000, “Striping Doesn’t Scale: How to Achieve Scalability for Continuous Media Servers with Replication”. In *Proceedings of 20th International Conference on Distributed Computing Systems*, Taipei, Taiwan, 64–71.

Dan A. and Sitaram D., 1995a, “Dynamic Policy of Segment Replication for Load-Balancing in Video-on-Demand Servers”. *ACM Multimedia Systems*, 3, no. 3, 93–103.

Dan A. and Sitaram D., 1995b, “An Online Video Placement Policy based on Bandwith to Space Ratio (BSR)”. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose, California, USA, 376–385.

Dukes J. and Jones J., 2002, “Dynamic RePacking: A Content Replication Policy for Clustered Multimedia Servers”. In *Proceedings of the Microsoft Research Summer Workshop*, Cambridge, England.

Guha A., 1999, “The Evolution to Network Storage Architectures for Multimedia Applications”. In *Proceedings of the IEEE Conference on Multimedia Computing Systems*, Florence, Italy, 68–73.

Hayden M. and van Renesse R., 1997, “Optimising Layered Communication Protocols”. In *Proceedings of the 6th IEEE International Symposium on High Performance Distributed Computing*.

Lee J.Y.B., 1998, “Parallel Video Servers: A Tutorial”. *IEEE Multimedia*, 5, no. 2, 20–28.

Little J.D.C., 1961, “A Proof of the Queuing Formula: $L = \lambda W$”. *Operations Research*, 9, no. 3, 383–387.

Patterson D.A., Gibson G. and Katz R.H., 1988, “A Case for Redundant Arrays of Inexpensive Disks (RAID)”. In *Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data*, Chicago, Illinois, USA, 109–116.

Schulzrinne H., Rao A. and Lanphier R., 1998, “Real Time Streaming Protocol (RTSP)”. IETF RFC 2326 (proposed standard), available at http://www.ietf.org/rfc/rfc2326.txt.

Serpanos D.N., Georgiadis L. and Bouloutas T., 1998, “MMPacking: A Load and Storage Balancing Algorithm for Distributed Multimedia Servers”. *IEEE Transactions on Circuits and Systems for Video Technology*, 8, no. 1, 13–17.

Vaysburd A., 1998, *Building reliable Interoperable Distributed Objects with the maestro Tools*. Ph.D. thesis, Cornell University.

Venkatasubramanian N. and Ramanathan S., 1997, “Load Management in Distributed Video Servers”. In *Proceedings of the International Conference on Distributed Computing Systems*, Baltimore, Maryland, USA.

Wolf J.L., Yu P.S. and Shachnai H., 1995, “DASD Dancing: A Disk Load Balancing Optimization Scheme for Video-on-Demand Computer Systems”. In *Proceedings of ACM SIGMETRICS ’95*, Ottawa, Ontario, Canada, 157–166.

Wong P.C. and Lee Y.B., 1997, “Redundant Arrays of Inexpensive Servers (RAIS) for On-Demand Multimedia Services”. In *Proceedings ICC 97*, Monréal, Québec, Canada, 787–792.

A DESIGN FOR NETWORKED MULTIPLAYER GAMES: AN ARCHITECTURAL PROPOSAL

Stefano Ferretti, Stefano Cacciaguerra
Department of Computer Science, University of Bologna
Mura A. Zamboni 7, 40127 Bologna, Italy
E-Mail: {sferrett, scacciag}@cs.unibo.it

ABSTRACT

This paper presents a general distributed architecture that supports networked multiplayer games. By moving server functionality to the boundaries of the network it is possible to provide robustness, congestion control, scalability and cheating prevention. Our architecture replicates the game state on the system, filters messages and uses application access points so as to optimize the communication with the clients. Our approach allows the distribution of the characters' information over the network and at the same time the replication of the virtual world's state only where necessary.

KEYWORDS

Multiplayer Games, Distributed Architectures, Scalability, Responsiveness, Message Filtering.

1. INTRODUCTION

Developing distributed multiplayer games is becoming one of the upcoming challenges in research communities (EU-Gamenet). A number of architectural proposals, algorithms, group communication systems relating to Internet gaming have been presented in last years. If we desire to understand networked multiplayer games and networked multiplayer game design, we must determine the fundamental characteristics of all games and briefly describe the families of games, from the user point of view. A game is a closed formal system that represents a subjective and deliberately simplified representation of reality. It has explicit rules and is a collection of parts which interact with each other, often in complex ways. A multiplayer game is a game with several users that interact in some form. With the advent of the Internet, it is possible to play a multiplayer game connecting to other users through the net. Today a lot of multiplayer games are commercially available. These games present a bewildering array of properties. Given this large set of games, it is possible to highlight the main features of gaming design by establishing a taxonomy of multiplayer networked games. This is only a way of organizing a large number of related objects. Many taxonomies are admissible. Our taxonomy is composed only by four families of multiplayer games, to avoid useless repetitions and not significant families. It is possible that a game belongs to more families: in this case we call it a "hybrid one". The main multiplayer game families are (see Table 1 and 2): Role Playing Game, Shoot 'em up, Sport Game and Strategic/Simulative Game:

- In a *Role Playing Game* users play the role of a character. During his virtual life the character, making experiences, evolves and gains more powers and skills. The character spends his life in a virtual environment.
- In a *Shoot 'Em Up* users should shoot to theirs enemies with several kinds of weapons.
- *Sport Games* are a reproduction of real existing sport games or a fantastic review of these ones.
- In a *Strategic/Simulation Game* users manage resources to design and implement empires, towns or communities.

A game is characterized by three features (see Tables): characters, space and time. As we will see, these features enable a set of considerations that help in the design of a distributed architecture supporting different types of multiplayer games. The word *character* means everything the user or eventually an artificial intelligence can pilot. For example, warriors in a battlefield, cars in a racing or buildings in a simulation game are characters. Characters move, make action and interact in a *space*. Examples of spaces are battlefields, circuits or towns. Finally, characters move, make action and interact under particular *time* constrains. These constrains include real time requirements, time dilatation or contraction. The reminder of this paper is organized as follows. In Section 2 we describe general design issues that influenced our work. Section 3 presents the proposed architecture designed to support several types of multiplayer games. Section 4 reports some related works. Finally, Section 5 provides some conclusions.

2. DESIGN ISSUES

2.1 Architectural Issues

In order to support game applications with distributed clients, several architectural solutions are possible. We can distinguish among: i) Client-Server architectures, ii) Peer-to-Peer architectures and iii) Distributed architectures.
Concerning clients, the application receives inputs from the player, notifying them to the server which is connected, and receives events from the server. In essence, a multiplayer game architecture is composed of I/O control entities and service entities which maintain the state of the game. If server functionalities are installed and executed directly on the client machine, as a daemon, a Peer-to-Peer architecture is obtained. Vice versa, in the case of a single server, we will obtain a Client-Server architecture. Typically, a server provides support for:

Table 1. Four families of networked multiplayer games

GAME	RPG	SEU	SPORT	S/S
CHARACTER	Fantastic Characters (warriors, wizards, …).	Soldiers, Mechs or Monsters, …	Sportive Characters (soccer players, F1 cars, …)	Different Typologies (pawn, workers, soldiers, buildings, …).
SPACE	Fantastic World: it is possible to move and to interact with the environment modifying it.	Narrow spaces (room) and realistic scenarios.	Sport grounds, sport tracks	Different Typologies (chessboard, geographic place, …).
TIME	Dilatation (ex: when a character makes a fine action) or Contraction (ex: when a character rests) of time.	It has real time constrains.	Usually it is a contraction of the time match (ex: a soccer match of 90 minutes becomes one of 3).	The duration and the management of time are strictly correlated to the rules of the game (ex: if we play chess or a real time strategic game).
EXAMPLES	Ultima Online, Dark Age of Camelot…	Doom, Quake, Half-Life…	FIFA 2002, Gp4…	The Sims, Age of Empires, Civilization…

- State maintenance: the server maintains the game state that is composed of the virtual world representation and of the state of all the involved characters.
- Consistency maintenance: due to latency, timing relationships may be distorted in the simulated world causing anomalies and inconsistencies. Functionalities are needed so that players perceive the same events in the virtual world.
- Group Management: not all the players are interested to receive the same messages. This observation enables the grouping of players and message filtering. Grouping and message filtering typically relate to the semantic of the application (e.g. players in the same room receive same messages), but from our point of view it could depend also on the client configuration (e.g. clients with low bandwidth connection or with low computation capacity may receive less details than clients with high speed connection and high computational capacity).
- Message Multicast Delivery: each generated event has to be sent to other players.
- Accounting: before starting the game, the client configures and negotiates parameters with the server to better enjoy the service.
- Legality and cheating control: each event generated by a player on the virtual world has to be legal and authorized.

The Client-Server architecture is the classic solution in commercial products, e.g. Quake, Ultima Online, and Everquest. A single copy of the game state is maintained at the server-side. Each event generated by a client is sent to the server that processes it and forwards the new game state to all clients. The server controls the event's legality. In this scenario, clients have only to take inputs from the players and render the output state. Consistency is easy to maintain as well as cheating avoidance, because only the server has the game state. Thus, illegal manipulations of the game state are extremely difficult (Mauve et al. 2002; Bauer et al 2002). However, a centralized server is the bottleneck of the system. Besides, users with different network delays are treated unfairly. Therefore, pure Server based solutions require powerful (or cluster of) servers with a high bandwidth connection. In Peer-to-Peer architectures, each client maintains a copy of the game state and there is no central server. The replicated game state has to be consistent; therefore, each event produced by a player is directly sent to all others without the intervention of a central entity. An example of a game that uses a Peer-to-Peer solution is MiMaze (Gautier and Diot 1999). The main advantages of a fully-distributed architecture are low latency and robustness. In fact, messages are not relayed by any centralized server and there is no single point of failure. However, since interactions and accounting information are not verified by any server, cheating is possible. Moreover, pure Peer-to-Peer approaches are not scalable limiting the number of players that interact together in a group. A multicast connection among clients is needed so as to reduce bandwidth requirements: the total amount of traffic grows with the square of the number of clients. Therefore, P2P architectures are suitable for small-scale games. Distributed architectures use different servers geographically distributed over the network that maintain the game state (Cronin et al. 2002; Griwodz 2002; Openskies). This solution moves intelligence and functionalities to the border of the network. A distributed architecture is a hybrid structure that tries to join the advantages of the two previous approaches. Each client connects directly to the closest server and communicates in the same way as in a centralized Client-Server architecture. Servers are connected in a Peer-to-Peer fashion. The main advantages of these architectures are that there is no central bottleneck, they alleviate the problem of congestion and augment the robustness of the system. As in Peer-to-Peer, distributed architectures require special synchronization mechanisms so as to provide a consistent game state to all players. Having distributed servers replicated in different geographic areas, an appropriate feature is to insert access point functionalities in order to manage communications on the client side (Aarhus et al. 2002; Mauve et al. 2002). Another important expedient to improve the scalability is to send messages only to the interested clients (Bharambe et al. 2002; Fiedler et al. 2002; Funkhouser 1996). Finally, we are able to conclude that, in order to obtain networked multiplayer games, the use of a distributed architecture seems the best solution. It enables large-scale games, with administrative control.

2.2 Model Issues

Each game has a set of rules that model and define the game itself. A game state is composed of a set of characters in a virtual world. Each character has a set of variables that represent its state, like for example, position and score. The state of the game is the union of all the states related to the players in the game. Depending on the game, the virtual world could evolve. An example is a Role Playing Game where players interact with the world modifying it.

Table 2. Features of networked multiplayer games (from the user view point)

		FEATURES	RPG	SUE	SPORT	S / S
Character	A	Number of characters piloted by a user	1	1	1	More than 1
	B	Number of users in the same match	Not Limited	Limited	Limited	Not Limited
	C	Number of characters in the same room	Not Limited	Limited	Variable	Variable
Space	D	Spatial Granularity	Variable	Real	Real	Variable
	E	Dimension of the scenario in comparison to the character's movement	Great	Little	Little	Variable
	F	Possible permanent modification of the scenario	Yes	No	No	Yes
Time	G	Time Granularity	Variable	Real	Real	Variable
	H	Need of action/reaction	No	Yes	Yes	No
	I	Duration of the match	Not Limited	Limited	Limited	Not Limited
	J	The character can enter after the beginning of the match	Yes	Yes	No	No
	K	Saving match	Yes	No	No	Yes
	L	The character death stops the match	Yes	Yes	No	Yes

Each event generated by a character on the virtual world has to be legal with respect to the character's capabilities, the world state and timing constrains (e.g. the character could not be able to modify the virtual environment at a certain time). In essence, the state of a player, the state of the world and time management have to be controlled, stored and managed by the server, in order to interpret the rules of the model. Besides, several multiplayer games (e.g. Role Playing Games) need the capability of saving the game state so as to freeze it and restart the match in a second moment. We propose to control in a separate way each entity involved in the game (i.e. characters and the virtual world representation), time management and storing capabilities. Therefore, we obtain a set of *managers* which control the already mentioned issues:

- *Space Manager (SM)*: it maintains the game map and manages the updates. These updates could be temporary (after a certain time interval the state is restored) or permanent (i.e. a permanent modification of the map). It also controls the coherency and the legality of the events generated in the world (e.g. an avatar cannot move through a wall)
- *Characters Manager (CM)*: it manages characters' evolution in the game. In a Role Playing Game for example, each character evolves during its virtual life; this manager controls the state of the character. It also maintains the list of players, and related characters, that are present in the current game session and controls the legality of player's actions.
- *Storage Manager (STM)*: it provides the capability of saving the game state in storage, in order to allow the restart of the game session in a second moment. A manager is needed so as to provide different saving typologies: if a player wants to leave the game and re-join later, the CM has to save the character's state. Vice versa, if the whole session has to be freezed, then the game state has to be stored. Moreover, other settings are possible, e.g. storing of configurations, scores, skills, etc. etc.
- As concerns *time management*, each manager has to face it. Distributed players have to perceive the same simulation time at the same wall-clock time or with a bounded time difference. A synchronization algorithm is thus necessary. Moreover, in distributed architectures these algorithms are mandatory so as to maintain consistency. A number of techniques have been proposed in literature, ranging from conservative approaches (Bryant et al. 1977; Chandy and Misra 1978) to optimistic approaches. Optimistic algorithms seem to be suitable for interactive applications like on-line games. Time Warp (Jefferson 1985) and its evolutions (Gafni 1988; West 1988, Madisetti et al. 1993; Steinman et al. 1993; Steinman 1995; Cronin et al. 2002) are examples of optimistic techniques. In the specific game context, MiMaze implements an optimistic version of the conservative bucket synchronization algorithm.
- In the definition of an architecture that supports distributed on-line gaming, it is interesting to observe that not all the players are interested to receive the same messages. This enables the grouping of players that need the same information. Such approach makes an efficient use of bandwidth enabling message filtering. In fact, the broadcast of every event will consume all the available bandwidth. To this aim, the best solution is the introduction of a *Group Communication Manager (GM)*, that takes the decision about who needs data and who does not.

3. THE PROPOSED ARCHITECTURE

In this Section, we present the proposed architecture designed to support different families of multiplayer games. Due to the problems related to client-server and peer-to-peer approaches, our solution is a hybrid structure that tries to join the advantages of the two previous approaches. We then propose a distributed architecture, in which the game state is replicated in different geographical areas.. We introduce an Application Access Point (AAP) that takes care of managing the optimized communication from the client-side. The AAPs are distributed in different geographical areas. Besides, we designed a support mechanism, called Group Communication Manager that allows message filtering so as not to waste computational and network resources. In the following we describe how our architecture works (see Figure 1).

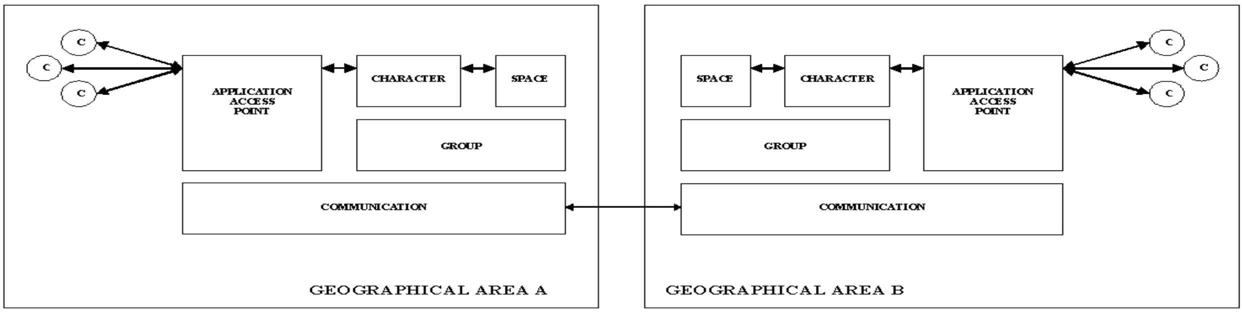


Figure 1. Architectural scheme

Players send messages to the AAP. The AAP puts these messages in a queue, marking them with a time-stamp. Then, the AAP forwards the messages to the CM which is connected. Each CM maintains only informations about characters piloted by players in its geographical area; therefore the CM is distributed but not replicated. In particular the CM is the guarantor of the character's state. The CM carries on messages, verifies inconsistencies and asks for the legality of character's actions to the SM. This last communication updates the game state. The SM is distributed and replicated, and it takes care of the local view of the characters in its geographical area. For example, if in a geographical area "A" there are no players connected to a SM that pilot characters in a virtual room "X", the SM does not maintain information about that room. Vice versa, if a user is in "A" and his character is in the virtual room "Y", the SM in "A" takes care of maintaining consistent the game state of room "Y" with other SMs (see Figure 2). The SM uses GM functionalities to communicate the change of the state to all involved players and others involved SMs. The GM takes care of filter messages sending them to the appropriate receivers. The information's transmission, added with a synchronization mechanism performed by the SM enables the consistency maintenance of the overall game state. In order to send information, the GM uses a Communication Delivery Service that is in charge of transmitting the messages to the appropriate AAPs and to GMs of other areas (look at the double arrow that links the two geographical areas in Figure 2). Eventually, a storage manager could take care of saving the game state when necessary. The importance of this manager depends on the game. For example RPG and S/S games may require several saving operations, while Sport and SEU typically does not need the storing of information. The main advantages of this architecture are:

- There is no centralized server that maintains the game state. In this way the system becomes *robust* and *scalable*, because different game's copies are held in different geographical areas.
- The *computational load* is divided among different geographically distributed systems, avoiding bottlenecks and any sort of *congestion*.
- Placing different AAPs at the border of the network we have a higher level of accessibility to the system. In fact, every client connects to its closest AAP. This approach provides *fairness.* Besides, each GM filters messages sending information only to interested receivers. This approach reduces *latency delay* and *congestion* in the system.
- Each AAP optimizes the communication between clients and the system. From the client-side, the AAP maintains the communication active also in critical situations, sending only messages with high priority. From the system-side, it manages messages of different clients at the same time and with the same priority.
- The system avoids *cheating* having the CMs that control the characters' state.

4. RELATED WORK

Online gaming is a complex research topic that poses several issues and involves different disciplines. In this work we pose our attention on architectures. It is interesting to observe that at the moment, commercial solutions differ from research proposals: commercial architectural solutions use Client-Server architectures while scientific works typically propose distributed architectures. Our approach allows the distribution of the characters' information over the network and at the same time the replication of the virtual world state only where necessary. The distinction between characters and virtual world information is not present in other works. In fact, in Client-Server approaches a single state of the game is maintained. Distributed architectures (Cronin et al. 2002; Mauve et al. 2002; Griwodz 2002) typically replicate the game Server, adopting a particular consistency mechanism and message filtering. Finally, Peer-to-Peer architectures replicate the state of the game directly at each peer. In the following, we report some recent related works, in order to compare our proposal with others. (Gautier and Diot 1999) presents MiMaze, a Peer-to-Peer distributed multiplayer game on the Internet based on a RTP communication system. In order to provide consistency, an optimistic version of the conservative bucket synchronization algorithm has been implemented. (Cai et al. 2002) suggests a spatial partition of the virtual world, so as to assign each partition to a specific server. The server is composed of two main parts: a front-end server that interacts with clients, and a back-end server which processes interactions, accounting information and update messages. (Fiedler et al. 2002; Bharambe et al. 2002) propose approaches for a communication architecture based on the publish-subscribe model. In (Aarhus et al. 2002), the authors propose a distributed architecture to filter events.

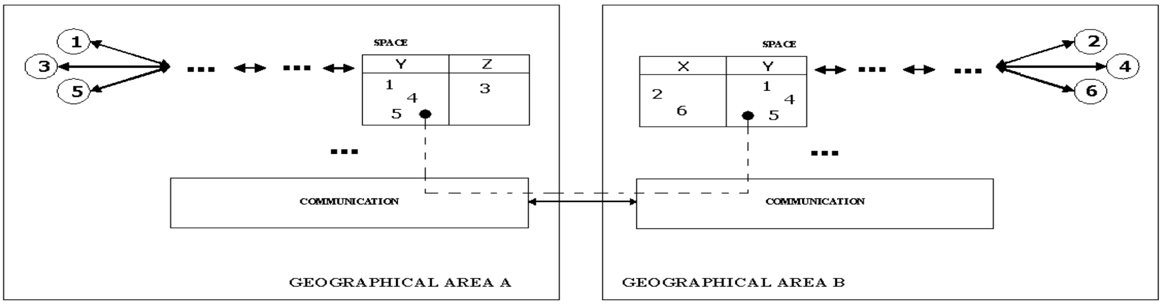


Figure 2. An example of game state replication and communication for maintaining consistency

The server side of the architecture is organized in two tiers: a server tier and the concentrator tier. (Mauve et al. 2002) presents a proxy system for distributed games that provides congestion control, robustness, fairness and that minimizes the impact of network delay. (Griwodz 2002) uses distributed proxy architecture and assigns to events an urgency value to indicate the requirement for low latency, and a relevance value to indicate the requirement for high reliability. In (Bauer et al. 2002) is proposed a filtering mechanism that acts at network level, with the use of booster boxes by monitoring and measuring network traffic. (Cronin et al. 2002) presents a new synchronization mechanism called *trailing state synchronization* (TSS) in order to provide consistency in a mirrored distributed server architecture.

5. CONCLUSIONS

In this paper we presented a general multiplayer game architecture for a work-in-progress research project on networked computer games. We claim that neither centralized Client-Server architectures nor Peer-to-Peer applications are able to solve all of the problems encountered in the networked multiplayer games. By moving server functionality to the boundaries of the network we believe that it is possible to solve important problems of networked multiplayer games, including robustness, congestion control, latency minimization, providing fairness, scalability and cheating prevention. In our distributed architectural proposal we can replicate the game state in different geographical area. Besides, we insert application access points to optimize the communication between clients and the system. Again, we improve the performance sending messages only to the interested clients. Finally, we introduce a CM that manages every single character's state, preventing cheat. We are actually implementing this architecture and we are currently planning an experimental campaign.

BIBLIOGRAPHY

Aarhus L., Holmqvist K., Kirkengen M., "Generalized Two-Tier Relevance Filtering of Computer Game Update Events", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

Bauer D., Rooney S., Scotton P., "Network Infrastructure for Massively Distributed Games", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

Bharambe A. R., Rao S., Seshan S., "Mercury: A Scalable Publish-Subscribe System for Internet Games", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

Bryant R. E., "Simulation of Packet Communication Architecture Computer Systems", Computer Science Laboratory, Massachusetts Institute of Technologies, Cambridge, Massachusetts, 1977.

Cai W., Xavier P., Turner S. J., Lee B. S., "A Scalable Architecture for Supporting Internet Games on the Internet", in *Proc. of 16th Workshop on Parallel and Distributed Simulation (IEEE)* Washington, D.C., May 12 - 15, 2002.

Cronin E., Filstrup B., Kurc A., Jamin S., "An Efficient Synchronization Mechanism for Mirrored Game Architectures", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

EU-GamenetWeb Site: http://www.scit.wlv.ac.uk/~cm1822/eugames.htm

Fiedler S., Wallner M., Weber M., "A Communication Architecture fro Massive Multiplayer Games", *in Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

Funkhouser T. A., "Network Topologies for Scalable Multi-User Virtual Environment", 1996.

Gautier L., Diot C., "A Distributed Architecture for Multiplayer Interactive Applications on the Internet" in *IEEE Networks Magazine 13 (4)* July/August 1999.

Gafni A., "Rollback mechanism for optimistic distributed simulation systems", in *Proc. of the SCS Multiconference on Distributed Simulation*, 19: 61-67, 1988.

Griwodz C., "State replication for multiplayer games", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

Mauve M., Fisher S., Widmer J., "A Generic Proxy System for Netwoked Computer Games", in *Proc. of NetGames2002*, Braunschweig, Germany, April 16-17, 2002.

"Openskies Network Architecture", 2002, http://www.openskies.net/files/Openskies_Network_Architecture.pdf

Steinman J. S., Bagrodia R., Jefferson D., "Breathing time warp", in *Proc. of the 1993 Workshop on Parallel and Distributed Simulation*, pages 109-118, May 1993.

Steinman J. S., "Scalable Parallel And Distributed Military Simulations Using The Speedes Framework", in *Proc. of Object-Oriented Simulation Conference*, Pg.3-23, 1995.

West D., "Optimizing Time Warp: Lazy rollback and lazy re-evaluation", Computer Science Department, University of Calgary, Alberta, Canada, 1988.

E-LEARNING AND TRAINING

The value chain of the learning organization. The L(E)RP system integrating all its subfunctions: an architecture.

Jeanne Schreurs, Rachel Moreau, Ivan Picart,
Limburgs Universitair Centrum
Universitaire Campus
B-3590 Diepenbeek, Belgium
E-mail: jeanne.schreurs@luc.ac.be
rachel.moreau@luc.ac.be
ivan.picart@luc.ac.be

KEYWORDS

e-learning, e-blended learning, content management, learning object, value chain, warehouse

ABSTRACT

The end product of the learning organization is the learning process or the creation and organization of an e-blended learning course. The learning organization process will be structured as a value chain, consisting of 4 primary subprocesses and 1 supporting subprocess. The individual activities and sub IS have been identified. The architecture of an integrated L(E)RP system for the learning organization process has been set forward.

INTRODUCTION

Till to now in e-learning applications research and development focus is mostly on some parts of the learning process. The last years research and development has been concentrated on authoring systems, on content management systems, on asynchronic discussion forum, on online testing, ... Special focus is also on the personalisation and the reusablity of e-learning materials and courses.
Our intention now is to look at the entire value chain of the learning organisation. A system integrating all its subfunctions, a L-(E)RP system, will optimise the learning process or will create the highest value for the users, being from one side the learners and from the other side the teachers who are organising courses and learning materials for the learners.

THE VALUE CHAIN OF THE LEARNING ORGANISATION

The end product of a learning organisation

The end product of a learning institute is the learning process or the creation and organization of an e-blended learning course.
The realisation of an e-blended learning course implies:

- The organization of the learning process: an e-blended learning process as a mix of several and several kind of activities
- Making available the learning content. We distinguish different kinds of content documents:
    - Multimedia documents (all kind of files)
    - Webcourse modules, free available on the web or available from suppliers on a commercial base.
    - Textbook + its website
- Organization of the learning process environment
    - Teacher/tutor organizes, supports and evaluates the learners in the learning process
    - Individual students are taken part in the organized activities. Often students collaborate in team activities.
    - Some activities are organized as classroom activities, sometimes as virtual classes.

The learning organization process structured in the value chain

The organization of the learning process has to been seen on three levels:

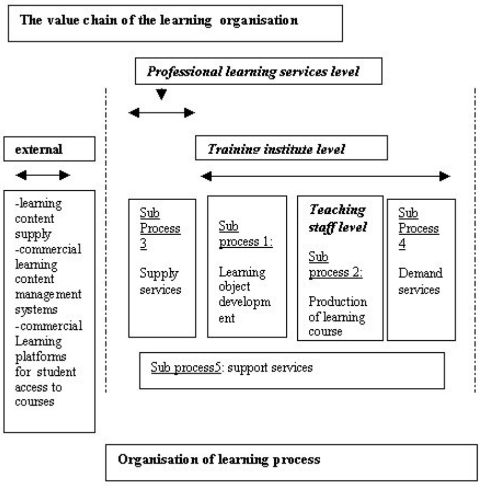
- on teaching staff level
- on training institute level
- on professional learning services level (a special service unit in the organization *or* an ASP commercial organization)

The learning organization process can be split into five sub processes:

- sub process 1: the learning object creation
- sub process 2: the production of the learning course or course development
- sub process 3: of supply services
- sub process 4: of demand services
- sub process 5: of support services

Two external organization units don't belong to the process but will be linked with it.

Figure 1: the value chain


The value chain of the learning organisation
Professional learning services level
external
Training institute level
-learning content supply -commercial learning content management systems -commercial Learning platforms for student access to courses
Sub Process 3
Supply services
Sub process 1:
Learning object developm ent
Teaching staff level
Sub process 2:
Production of learning course
Sub Process 4
Demand services
Sub process5: support services
Organisation of learning process

VALUE CHAIN ACTIVITIES AND LEARNING FUNCTIONS INFORMATION SYSTEMS (L-IS)

The value chain has been structured in 5 sub processes. The following activities and supporting L-IS can be discerned: Sub processes from 1 to 4 are the primary activities and sub process 5 consists of the support activities.

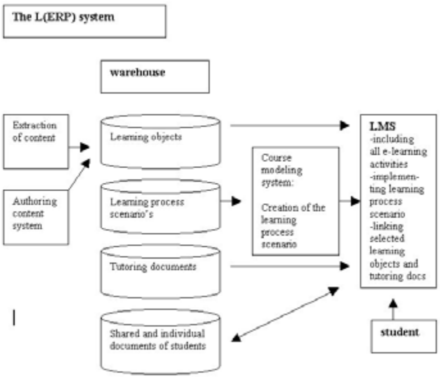
- Sub process 1: learning object development
  - selection and extraction of content from the institutional content warehouse
  - creation/selection of the learning object, taken into account the characteristics of the learner.
  - selection and extraction of learning models/scenario's
- Sub process 2 : production of the course (modules)
  - design of scenario of e-blended learning process
  - selection and implementation of personalised learning objects
- Sub process 3 : supply services
  - Supply of relevant content : acquisition of knowledge in several knowledge domains
  - Web based authoring of learning content elements
  - creation of an institutional warehouse and management of content
- Sub process 4 : demand services : Implementation of a course in a learning platform/Learning Management System(LMS).
  - E-learning courses,
  - E-blended learning courses
  - virtual classrooms

  including e-learning activities :
  - online textbook, testing and exercises/ assignments
  - asynchronic discussion forum
  - sharing online info/knowledge between students
  - question and answer facility
  - realtime tutoring session
  - interactive live classroom session for distance students
- Sub process 5 :support services
  - ICT user support
  - Administration : user admin, SLA(service level agreement), …
  - Quality assurance

THE L(E)RP SYSTEM INTEGRATING ALL SUBFUNCTIONS OF THE LEARNING ORGANISATION

The centralisation of the learning objects in a warehouse forms the basis for the integration.
The easiest way to guarantee the seamless co-operation of all the individual IS as supporting the sub processes, is building them using compatible components and to join them into one common learning organisation system.
More realistic is to link the different systems via the development of middleware modules. This last option is preferable because it is more flexible in that the organisation can use its own preferred information systems.
Our L(E)RP system consists of a set of modules integrated upon the central warehouse of learning objects and accessible via a common interface for the user.
The Scorm specifications are followed for standardising the re-usability and interoperability of learning content. Scorm divides the learning system into two functional components : the reusable learning objects from one side and an advanced LMS from the other side. The latter one is keeping learner info and is responsable for implementing the personalised learning process scenario of activities and linked learning objects.

Figure 2 : The L(E)RP system


The L(ERP) system
warehouse
Extraction of content
Authoring content system
Learning objects
Learning process scenario's
Tutoring documents
Shared and individual documents of students
Course modeling system:
Creation of the learning process scenario
LMS
-including all e-learning activities
-implemen-ting learning process scenario
-linking selected learning objects and tutoring docs
student

CONCLUSION

The professionalisation of the learning organisation does require the implementation of a L(E)RP system integrating all IS of the subfunctions of the learning organisation. The kernel of this system is the warehouse of the learning objects. The overall learning process will be optimised in this professional learning organisation.

REFERENCES

A Lotus Development Corporation White Paper: Lotus and IBM knowledge management strategy. September 2000.

Schreurs, J. 19-20 October 2000. *Development of a Virtual Learning Environment and Implementation of Courses in ICT and in ICT Management*. Conference Proceedings, Bolton International Conference.

Schreurs, J. a.o. 9-10 May 2000. *A multimedia warehouse supporting on line learning via internet.* Euromedia 2000: Building a global business. Ed. F.Broeckx & L.Pauwels.

Cuppers L. and J. Schreurs. *The implementation of a virtual learning environment at the Limburgs Universitair Centrum*. Euromedia 18-20 april 2001.

Schreurs, J : *The learning portal of a collaboration learning system.* Euromedia 18-20 april 2001.

Picart, I .and Moreau,R. and Schreurs,J. *Learning content management in a university.* Euromedia , Modena 15-17 april 2002.

http://www.fcw.com/fcw/articles/2001/0122/tec-xml-01-22-01.asp

Applied Multimedia System for Analysis and Mechanical Design of Robot

M. H. Korayem
S. Baghaei
Robotic Research Laboratory,
University of science and technology
Iran
E-mail: hkorayem@iust.ac.ir

KEYWORDS
Robot design, Multimedia system.

ABSTRACT

Usage of scientific and tutorial systems are widely used in recent decade. Besides, these packages are currently used in robotic field for conceptual understanding and robot analysis. For this purpose, initial activities have been done, but not deeply for applied research purposes. With respect to importance of robot design and its influence on robot construction, as well as affection of multimedia systems for informing and learning, these two subjects are used to create an applied multimedia package for improving the robot design and tutorial aspects. To present an engineering package based on multimedia system for applied goals, instruction and robot design is the subject of this paper. Finally, a computer program is developed and used to display the extensive analysis formation in a compact form. This helps a robot designer develop an intuitive feel for his problem through reinforced display.

INTRODUCTION

ENGINEERING Multimedia systems are powerful tools for effective grasp of scientific concepts. These softwares can be widely used in robotic field. Therefore the robot design subjects can be easily and efficiently studied with considerable multimedia facilities.

As we know robotic systems have several mechanical components which include: stiffness, elasticity, inertia, specific strength and etc. If mechanical design of robot is not appropriate, some critical problems such as endurance, accuracy, repeatability, stability, work space, safety and so on will occur. Robot somehow has to be designed in order to act efficiently without any difficulties for avoiding these problems. Since nonlinearity of dynamical parameters is basic characteristic of robot. Designer should predict the effectiveness of changing the geometrical factors for robot dynamic performance. So designer should know the dynamical behaviour of robot before construction. Hence we avoid wasting considerable amount of capital for robot action test. On the other hand presence of advanced computer technology can help to analyse and design the mechanical systems. Therefore classification of robot design subjects after that presentation and analysis with the multimedia software can be valuable.

Application of multimedia systems in science and technology has a short time history. The activities for robot design and multimedia systems creation in recent years are summarily as below :

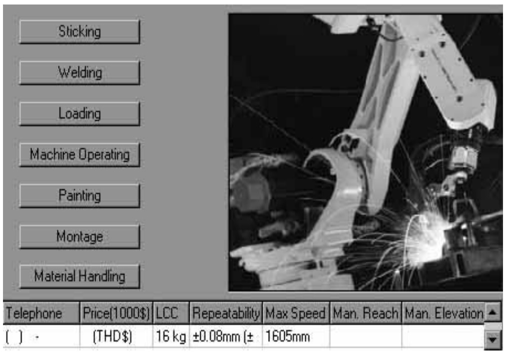
The design of spherical 4R linkages for specified orientations was investigated by Douglas, Ruth and J. Michael. The Clifford algebra of double Quaternions and the optimization of TS robot design was presented by Ahlers, Michael and Mccarthy. Besides, Mccartthy published the paper of Mechanisms synthesis theory and the design of robots. Moreover, the scientific multimedia system which is named VIRTUAL HOSPITAL was designed by Brandser, Natisha, Busick and Sandra. This package is perfect reference for instruction and information on Anatomy.

Robotic multimedia software for conceptual understanding was created by Korayem, Baghaei and Azmoudeh. Swain published an article on improving instruction and importance of far distance education for effective teaching.

ROBOTS BANK

It is possible to access the information and technical specifications of industrial robots with this software. The environment of this program is a guide and appropriate tool for acquiring the robots properties and intelligent selection of them. This part of multimedia system contains welding, pasting, machining, palletizing and surgical robots operations. There are technical specifications in this package, for example maximum speed, accuracy, repeatability, payload, number of axes etc. (Fig. 1). This Section has many options and facilities as below:

Robot in the world, Robot in Iran, Search, Tutorial, Delete the expired information of robot, Update expired information of robot, Report, Exit.



Telephone	Price(1000$)	LCC	Repeatability	Max Speed	Man. Reach	Man. Elevation
( ) -	(THD$)	16 kg	±0.08mm (±	1605mm		

Figure 1: Robots bank

MODELLING AND SIMULATION

Design of robot before construction is an important issue, because this procedure is logical for financial aspects and optimize the time. This purpose can be done by specific engineering softwares that can combine the technology of motion and finite element analysis in unique area. The Cartesian robot which is used for assembly line of internal combustion engines was analysed by using Working Model 4D. Finally the results and outputs were organized and classified with slides, animations and explanations in multimedia system. A robot designer can easily predict and foresee the impact of changing the operational conditions or any of the kinematic parameters on the performance characteristics of a robotics manipulator. Fig. 2 shows the kinematical analysis of Bridge robot during picking and placing of work piece. With this analysis the position, velocity and acceleration of the end effector in x, y and z directions can be determined.

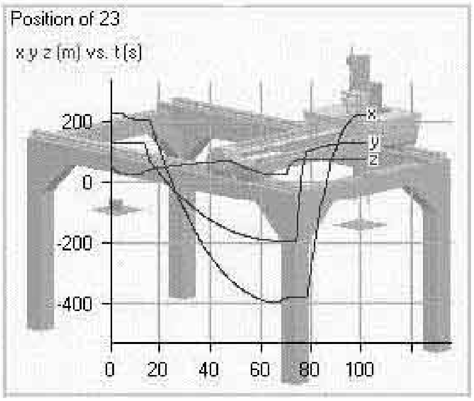


Figure 2: Kinematical analysis of Bridge robot

DYNAMICAL BEHAVIOUR UNDER DIFFERENT TRAJECTORIES

The most important property of robot is nonlinear dynamical parameters such as inertial forces, centrifugal, coriolis effects, gravity and coupling. This multimedia system has an option to predict the robot behaviour. In this part, during design process, recognition of robot dynamic behaviour under certain work condition is important. This part of system is an appropriate tool for acquiring the correct dynamical model of SCARA robot links.

Study of dynamical behaviour

Three types of velocity are used in order to examine different characteristics of the dynamic trajectories in this step. They contain:
1-Polynomial trajectories, 2-Numerical Control trajectories, and 3- compound trajectories.
The driving torques can be obtained based on lagrangian method as follows:

$$\tau_i = \sum_{j=1}^{N} D_{ij}\ddot{q}_j + \sum_{j=1}^{N}\sum_{k=1}^{N} D_{ijk}\dot{q}_j\dot{q}_k + D_i \quad (i = 1,2,3,\ldots,N) \qquad (1)$$

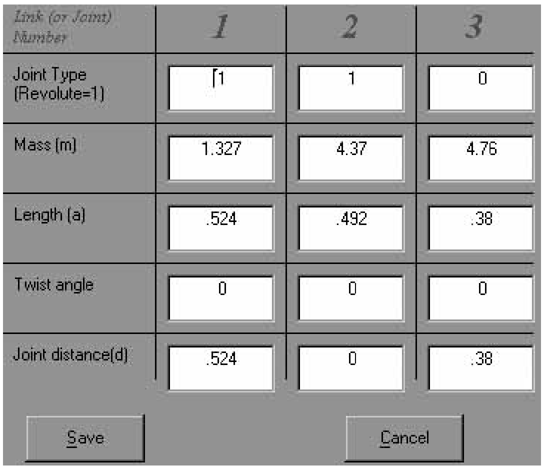
where

$$D_{ij} = \sum_{p=\max(i,j)}^{n} T_r\left(\frac{\partial T_p}{\partial q_j} j_p \left(\frac{\partial T_p}{\partial q_i}\right)^T\right) \qquad (2)$$

$$D_{ijk} = \sum_{p=\max(i,j,k)}^{n} T_r\left(\frac{\partial^2 T_p}{\partial q_j \partial q_k} j_p \left(\frac{\partial T_p}{\partial q_i}\right)^T\right) \qquad (3)$$

$$D_i = \sum_{p=i}^{n} - m_p g^T \left(\frac{\partial T_p}{\partial q_i}\right) r_p \qquad (4)$$

T_p, m_p, and r_p are transform matrix, mass of link p and the distance of the mass center of link p with respect to link p's coordinate.
After inputting initial parameters such as inertial matrix, pay load, geometrical parameters, analysis begins and system plots the velocities, accelerations and driving torques for each link as shown in Figs.3 and 4.



Link (or Joint) Number	1	2	3
Joint Type (Revolute=1)	1	1	0
Mass (m)	1.327	4.37	4.76
Length (a)	.524	.492	.38
Twist angle	0	0	0
Joint distance(d)	.524	0	.38

Figure 3: The input kinematic parameters for analysing of dynamic behaviour.

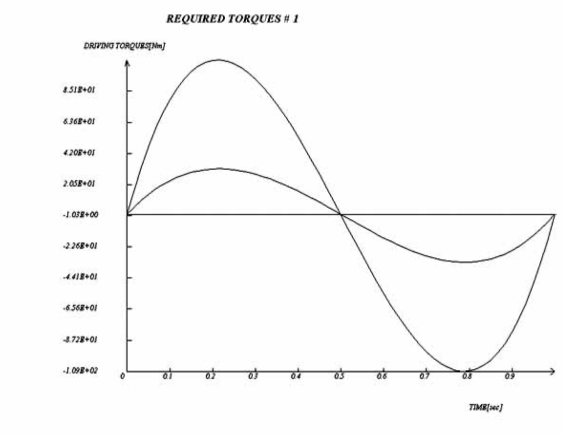


Figure 4: Required driving torque for SCARA robot in polynomial trajectory

The driving torques for link No.1 is presented and calculated under polynomial trajectory in Fig. 4. The curve is continuous, because its trajectory is polynomial function of time. This trajectory is appropriate for doing the industrial job with high accuracy, but there is no optimum time for operation. On the other hand NC2 trajectory has optimum time, But it causes the vibration on the robot end effector. This multimedia system is able to estimate the driving torques for each link under polynomial, NC2 and compound trajectories.

PERFORMANCE

An executive program under multimedia system is presented based on lagrangian dynamic for obtaining the dynamic properties of PUMA560. Moreover, this Section is able to change the geometrical parameters and express their effectiveness on dynamic performance.

Performance indicator

This part presents a criterion to measure the dynamic performance of a robotic arm is largely dependent on its inertia terms. The performance indicator is based on the logarithmic function of the sensitivity of the eigenvalues of the inertia matrix to change robot geometrical parameters.

To examine the effect of geometrical parameters on the dynamic performance, it is essential to develop a performance index. In equation 2 inertia matrix has 9 components, as follows:

$$D(q) = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{33} \\ D_{31} & D_{32} & D_{33} \end{bmatrix} \qquad (5)$$

The inertia matrix D(q) is similar to diagonal matrix. The eigenvalues of the inertia matrix of the three links can be obtained in the following form:

$$A = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \qquad (6)$$

By using Euclidean norm, the average response of the three eigenvalues can be written in the following form:

$$\mu = \sqrt{\sum_{j=1}^{N} \lambda_j^2} \qquad (7)$$

Since the analysis is concentrating on the minimum sensitivity irrespective of the derivative's sign, the absolute values of the derivatives are applied to determine the performance measure. These induced positive values for the eigenvalues' norm permits usage of their logarithm as a performance measure. We define the sensitivity of the inertia matrix's eigenvalues of geometrical parameters ξ_i in the following form:

$$\Gamma_{\xi_i} = \log_{10} \left\| \frac{\partial \mu}{\partial \xi_i} \right\| \qquad i=1,\ldots,N \qquad (8)$$

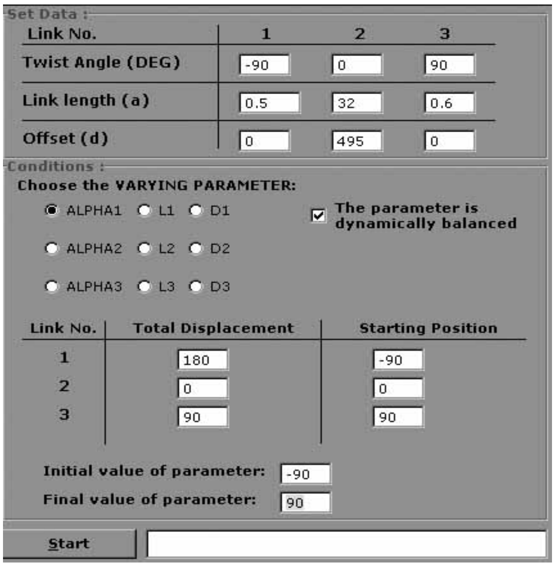
where

$$\xi_i = \{\alpha_i, a_i, d_i\} \qquad (9)$$

α_i, a_i and d_i are twist angle, length and offset of link i.

Running the performance division

After clicking on performance icon, a menu will appear as shown in Fig. 5. First user should input the geometrical parameters such as twist angle, length and offset for each link in top of the menu. Then, the type of variable for determining the dynamical sensitivity of this variable should be selected. The option of dynamically balance of manipulator is on the right side of the menu. In next step the position of each link in the space and its range should be defined. Now we can begin the process of analysis by clicking on start icon afterward 3D graphs which are very helpful for optimum design will be plotted. It is seen that the minimum area of these curves is appropriate location for design with high efficiency.



Set Data :

Link No.	1	2	3
Twist Angle (DEG)	-90	0	90
Link length (a)	0.5	32	0.6
Offset (d)	0	495	0

Conditions :

Choose the VARYING PARAMETER:

● ALPHA1 ○ L1 ○ D1 ☑ The parameter is dynamically balanced

○ ALPHA2 ○ L2 ○ D2

○ ALPHA3 ○ L3 ○ D3

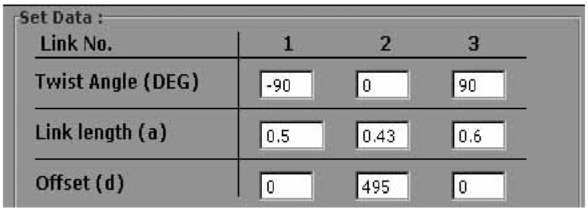
Link No.	Total Displacement	Starting Position
1	180	-90
2	0	0
3	90	90

Initial value of parameter: -90

Final value of parameter: 90

Figure 5: Main Menu for PERFORMANCE

Fig. 7 shows the dynamical sensitivity for robot performance with respect to twist angle of third link. This curve was obtained under specific space position as shown in Fig.6. Fig.7 expresses that $\alpha_3 = 0$ is the best location for optimum design, because this point is minimum sensitivity of dynamic performance, hence the robot control is more convenient during motion. The multimedia system can determine the dynamic performance for each link. The results of this part are 3D curves and numerical data which are helpful for mechanical design.



Set Data :

Link No.	1	2	3
Twist Angle (DEG)	-90	0	90
Link length (a)	0.5	0.43	0.6
Offset (d)	0	495	0

Figures 6: Space position of each link

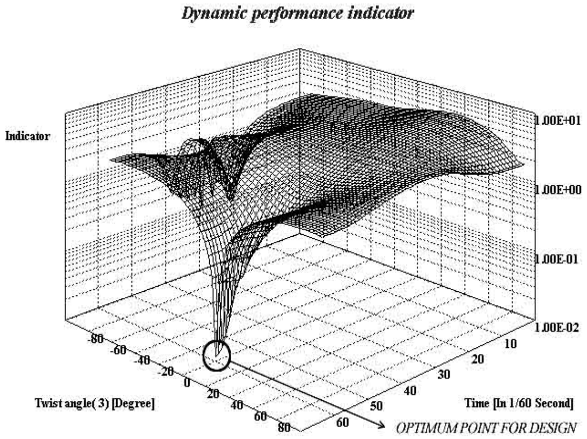
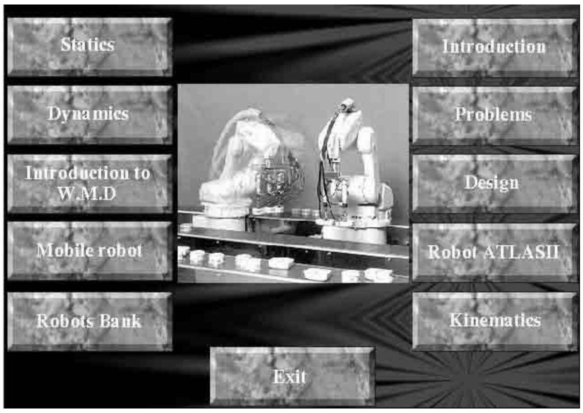


Figure 7: Dynamic performance indicator versus α_3 and time

MULTIMEDIA SYSTEM FOR UNDERSTANDING OF ROBOTIC CONCEPTS

This package, not only has applied purposes for design but also can instruct the basic and advanced concepts of robotic field to understand efficiently robot science using multimedia technology. All of the subjects which are presented in this region apply the multimedia facilities such as text, voice, slide, instruction, movie, animation. Therefore the scientific subjects are easily taught. Furthermore, both amateur and professional users can work with this package to analyse and predict the robot performance. Now each part of tutorial part will be summarily explained. (Fig. 8)



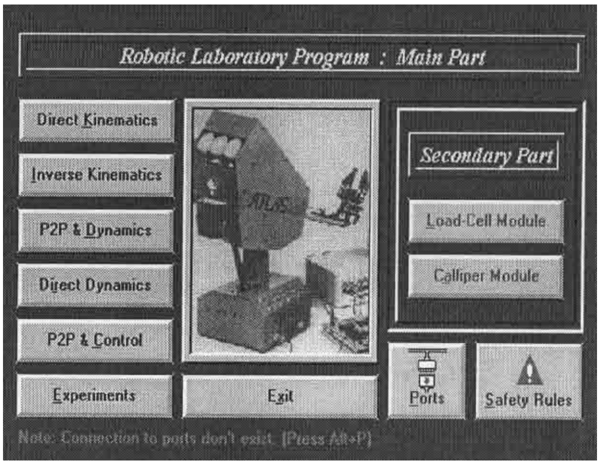
Figures 8: Main Menu of the robotic multimedia system

Introduction to automation

After loading this program, the main menu appears as shown in Fig.8. We can access the introduction to robots and automation by clicking on it. It deals with history of robots technology improvement and describes about automation and affection of robots in this process, robots classifications, components and their specifications.

AtlasII Robot

Educational robots have important role to study of robotic concepts. We can provide the conditions of design and construction with investigation and evaluation of robot. In this part, scheduled robot motions are simulated as graphical object. For betterment of action, an interface for connecting to PC was designed in order to run the simulated motions by experimental robot. After kinematic and dynamic modeling of robot, a group of experiments has been designed in order to instruct skillfully. (Fig. 9)



Note: Connection to ports don't exist. [Press Alt+P]

Figure 9: Main Menu of Robot ATLASII

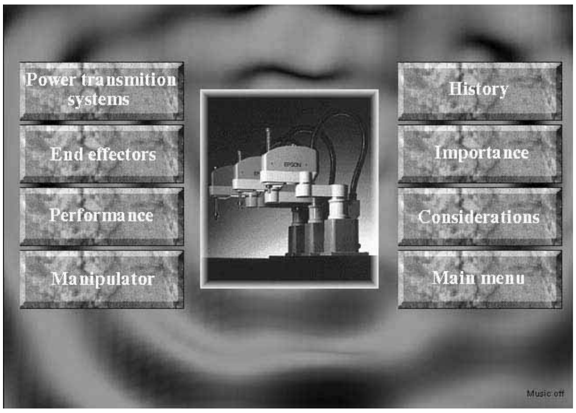
Kinematics, Statics and Dynamics

In kinematics, user can study on geometric of robots motions, direct kinematics, inverse kinematics and differential motion, furthermore user can investigate the forces, moments and Jacobean matrix of robots in static section.

Finally in dynamic part, Lagrange's method, inertial tensors of robots, equations of motions, inverse dynamics, Newton - Euler method and closed form dynamics equations are discussed.

Design

Now, it is possible to enter the section with known kinematics and dynamics information in order to analyse the design equations and robots structures. By clicking on this icon, submenu appears that we explain in next step.



Figures 10: Design section of robotic multimedia system

History

It discusses the history of mechanical design of robots in the world and its development. Besides, this part introduces important industrial robots which are used in universities, experimental and industrial centers.

Importance and goals

The reason and necessity of design science is illustrated in this part. Furthermore, influence of this knowledge in production of robots is explained.

Design considerations

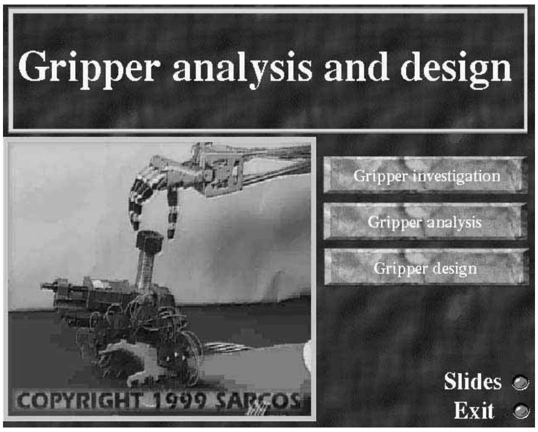
It deals the important factors of robots design. This Section comments about some essential parameters in design process, such as weight, work space, degrees of freedom, accuracy and etc.

Transmission systems

It discusses design of components in power transmission which includes gears, chains, cables, harmonic drivers, belts, ball screws.

End effectors

By clicking on this icon, trainee can investigate different kinds of grippers, mechanisms of end effectors, modeling and fundamentals of grippers design as shown in Fig. 11.



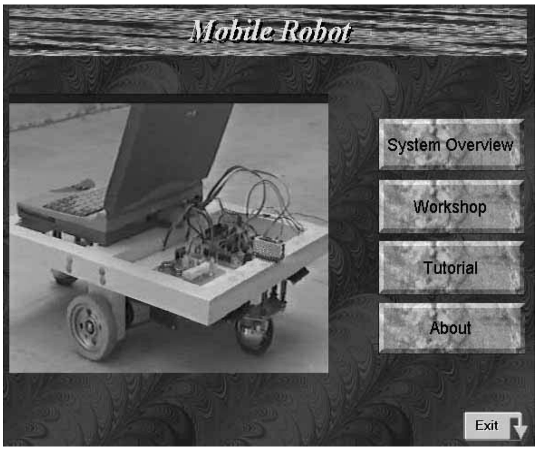
Figures 11: Part of end effectors

Manipulators

In this section, structures of manipulators, mechanical parts and components, selection of appropriate materials for robot structure, stiffness of mechanical components, impression of cross sectional area, nonlinear loading and calculation of stiffness, effectiveness of preloading, types of bearings such as revolute and reciprocating bearings, fatigue analysis are investigated.

Mobile robot

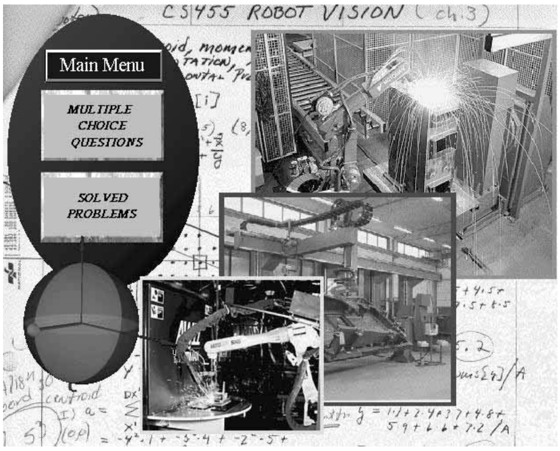
By clicking on Mobile robot in main menu, user is able to study about mobile robot analysis. This Section encompasses kinematics and dynamics, simulation, path finding, description for electrical and mechanical parts, movies and pictures of this robot and principles of robot design. Besides an interface has been provided for running the simulated motion by mobile robot. (Fig. 12)



Figures 12: Mobile robot

Test

If trainees peruse all of the subjects and understand all of concepts of this package, they can evaluate their knowledge in test division. This part has two Sections. The first part is multiple choice questions and the second is solved problems. As a matter of fact solved problems are used in order to study some practical and applied cases in robotic field for better concepts understanding. User can choose the type of questions in the test menu. In other words there are two types of questions, first type is modeling and fundamental of robots and second type is design and structural analysis of robots. In multiple choice questions, computer gives the limited time to users for answering the questions so user should not only answer correctly but think and response fast. Therefore it is useful for improving the learning and robotic knowledge. (Fig.13)



Figures 13: Test section

CONCLUSION

In this paper, multimedia system for applied and tutorial purposes was discussed. This package can improve the communication between instructor and student. On the other hand this system has practical tools like robots bank for intelligent selection of industrial robot and obtaining technical parameters. Dynamical analysis under different trajectories for determining the driving torques was presented in order to select the driver motor for each link. Dynamic performance with respect to geometrical parameters and finding the optimum point for design with 3D curves and numerical data was discussed. It is needed for designer to predict the robot properties before construction.

Finally this system is almost perfect group of robotic concepts and applications for multi purposes. On the other hand capabilities of multimedia systems and their impression are brilliant. So combination of science of robot design and multimedia technology creates the powerful tool for analysis and design of robot. In fact multimedia systems can make interesting creativities. Therefore this magical system can be used in other sciences, industrial centers and tutorial locations and universities.

REFERENCES

Asada and H.Stidies. 1979 .“In Prehension and Handling By Robot With Elastic”. Kyoto University.

Asada, H and J.E.Slotine. 1986. ”Robot analysis and control”. John Wiley & Sons Publication.

Bgartner, E. 2000. “NASA JPL Uses Working Model2D and 3D Asteroid Mission with Japanese Space Institute”. paper of NASA JPL.

Brandser, E and N Busick and A Sandra. 1995. ” Software of Virtual Hospital”. University of Iowa College of Medicine, Department of Anatomy and Cell Biology.

Chen, F.Y. 1982. ”Gripping Mechanisms for Industrial Robots –An overview”. Mechanism and Machine Teory, 17 No.5, pp299-311.

Douglas and A. Ruth and J.M. McCarthy, 1998. ”The Design of Spherical 4R Linkages for four specified Orientations”. Computational Method in Mechanisms, Ed, J, Angeles, Spring-Verlag.

Habibnejad, M and S Baghaei and A. Azmoudeh. 2000 “Robotics Multi Media Software For Conceptual Understanding”. Australia Proc 7th annual conf on mechtronics and machine vision in practice pp11-16.

Habibnejad, M. 2000 .“Analysis of robot dynamic behaviour under different trajectories”. International Journal Of Engineering Science, vol11,pp107-117.

Hall and Hollowenko and Laughlin. 1982. “Machine design”. Schaum-Hll, Hollowenko, Laughlin, McGraw-Hill.

Korayem, M.H. 1999. “THE INFLUENCE OF CHNAGING GEOMETRICAL PARAMETERS ON ROBOT OPTIMAL DYNAMIC PERFORMANCE”. International Journal of Engineering Science, Vol 10,no 5,pp107-121.

McCarthy, J.M. 2000. ”Mechanisms Synthesis Theory and the Design of Robots”. Proc. Int. Conf. Robotics and Automation, San Francisco, CA, April 24-28.

Refaat, M H and S A Meguid. 1998 .“Accurate modeling of compliant grippers using a new method”. RoboticaVolume 16,part2.

Sawn G.Ahlers and J.Michael McCarthy, 2000. ”The Clifford Algebra of double Quaternions and the Optimization of TS Robot Design”. Applications of Clifford Algebras in computer science and engineering,(E.Bayro and G.Sobczyk, editors), klewer.

Schigley, J.E . 1986.”Mechanical Engineering Design”. McGraw-Hill.

Swain. C. 2002.” Improving Traditional Teaching Using Findings from Far Distance Education”. EFFECTIVE TEACHING (January 9)

“Document of Working Model 4D”. Published and printed by Knowledge Revolution, 2000.

Web presentations of 3D civil engineering projects

Gerardo Silva Chandía
Associated Professor, U. of Santiago, Chile
gsilva@lauca.usach.cl

Abstract

Senior civil engineering (c.e.) students, developed 3D models of civil engineering projects as part of a last grade course since 1998 to 2002. Final presentations would be located in their own web pages so interested students or professionals, could easily appreciate them from any place in the world.

After five years of student's evaluated experience, a methodology can be proposed to do web presentations of 3D civil engineering projects.

This methodology doesn't require expensive hardware or software or computational knowledge other than CAD user knowledge and practice.

1. Introduction

Traditionally, civil engineering projects presentations have been accomplished by the blue prints exposure over a table or wall hanged.

This way of presentations works fine with a technical audience, but it is absolutely undesirable with a non-technical audience.

As civil engineering projects must be approved by non-engineering decision groups, such as ambient and ecological authorities, a non-technical way of presentations is needed.

TV presentations based on short high quality videos are used to gain the social acceptation from the community when a big c.e. project begins its building stage.

Web presentations have the advantages of being widely accepted by the non-technical communities and the ability of being accessed any time and anywhere.

Web pages can include two different file types as c.e. projects presentations:

- Videos (high, medium, low quality)
- VRML worlds

2. Methodology

The work must be break down in three steps:

- 3D modeling of the c.e. project
- Animations and VRML files generation
- Web publishing

3. Modeling of the c.e. project

3.1. 3D modeling of the site

The 3D modeling of the c.e. project optionally includes the site modeling.

This process is done by using DTM (Digital Terrain Modeling) software giving as result a CAD file generally a dxf or a dwg file.

The 3D model of the site reflects the original site before the earthworks included in the c.e. project are done and the final site aspect after earthworks are done.

See Figures 1 and 2.

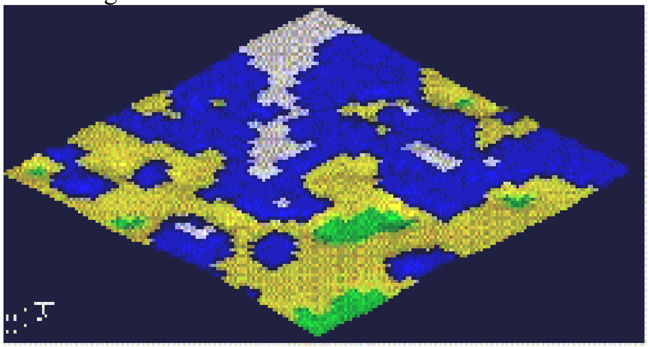


Figure 1: Original site, no earthworks.

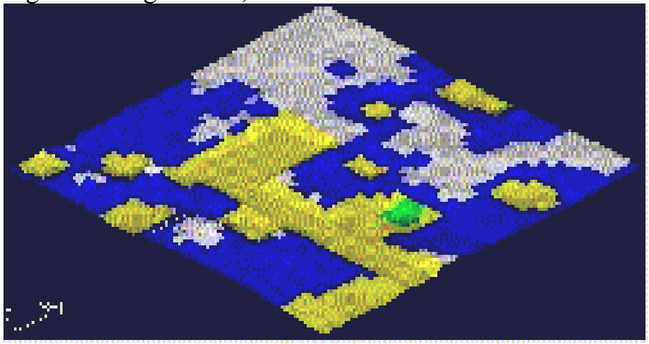


Figure 2: Projected site, earthworks reflected.

The site model is texture enriched only if animations are considered, because texture use generates time delay in the VRML worlds access via Internet.

3.2. 3D modeling of the building project

This step can be done in three alternative ways:

- "Simple model"
- "Complex model"
- "Complete model"

In the simple models, the third dimension is accomplished by the “extrusion” of 2D entities or objects, such as lines, arcs, circles, etc.

Commonly in these models, extruded lines represent walls. There aren’t moveable elements.

In the complex models, the elements are generated by the inclusion of 3D objects such as boxes, spheres, pyramids, etc.

Commonly in these models, walls are represented by polylines with a width equal to the wall width. 3D blocks are used and they correspond to moveable elements.

In the complete model, there is additional information added to the graphical 3D model.

See Figure 3.

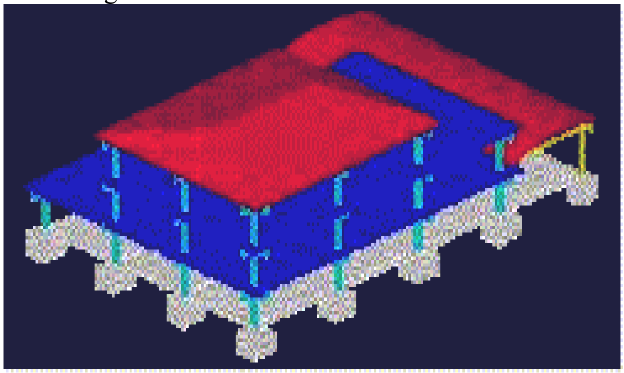


Figure 3: Industrial building 3D complete model.

This additional information, commonly named “attributes”, is exported from the 3D model to worksheets or databases, to generate bills of materials.

This procedure is not object oriented as proposed by F. Marir et al (1998) but is very simple to use by the students.

Any 3D model is texture enriched only if animations are considered, as explained before.

3.3. 3D modeling of the environment

Commonly environment is represented by any number of trees of different heights and geometries.

In an animation-oriented model, trees geometries are simple extruded lines with different applied textures.

In a VRML oriented model, trees geometries are more complex and have not textures.

See Figure 4.

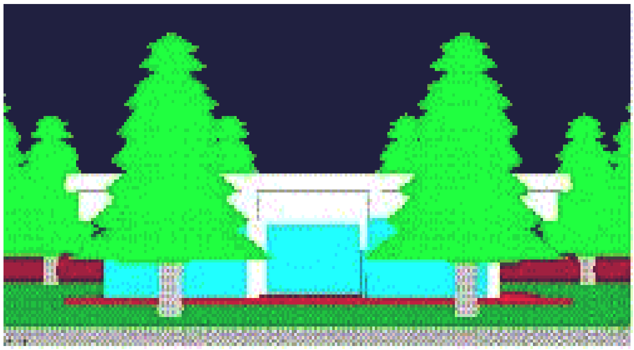


Figure 4: Vectorial 3D trees.

4. Animations and VRML files generation

4.1. Animations generation

By using software such as 3D Studio by example, students can generate very impressive animations, but to avoid the complexities of the use of this software to them, WalkThrough, a very simple and easy to use software was used in the experience.

WalkThrough has the advantage of being a ”near VR” software, so students explore the c.e. project in a freely way by using only the mouse to fly or walk inside the 3D model. This software doesn’t have the ability to recognize collision detection nor Level of Details (LOD) capacity.

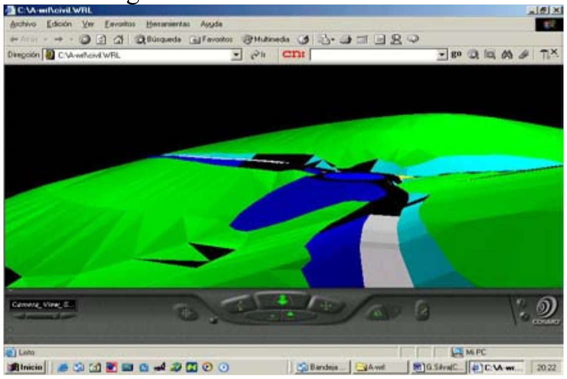
Textures are applied and the better animations defined by the students are saved in avi format files.

4.2. VRML files generation

As the experience was realized with c.e. students, no attempt was done to introduce them in the VRML language.

Instead, 3D Studio software was used simply as a “translator” from CAD dwg files to VRML wrl files.

This translation procedure sometimes doesn’t work fine. See Figure 5:



C:\A-wrl\civil.WRL
Archivo Edición Ver Favoritos Herramientas Ayuda
Búsqueda Favoritos Multimedia
Dirección C:\A-wrl\civil.WRL
Camera_View_0
Listo
Mi PC
Inicio
2022

Figure 5: Site VRML file with errors.

Textures were not used to obtain better response times in the web.

This procedure generates big sized VRML files normally under 4 to 5 MB.

5. Web publishing

In Chile, wide band Internet connections, only last year 2002 are growing up significantly, so files sizes are determinant in the user attitude.

By using video editing software, big avi files were converted to mpg and streaming formats.

Mpg files can be downloaded by the user and can be seen directly from the hard disk.

Real media rm format was used as streaming video format with 320 x 240 size preferred. This option requires the availability of a streaming video server.

No attempt was done to optimize the VRML files size, so download process was suggested to the web page visitors previous to access them.

VRML browser plug in Cosmo Player was suggested to the students. This plug in allows an easy examination of the 3D model by using the named views defined in the development stage in the CAD software.

The plug in has the ability to change the speed of the walk through so it is easy to move inside big models or models with a big extension.

Once the VRML file is downloaded by the user, the power of the current PC (Pentium 4, 1 GB RAM, etc.) is sufficient to handle very big size files.

Text contents with summarized information about estimated cost and duration of the building project, was included in the same web page.

Many students included graphical information such as Gantt charts of their projects.

Some examples of the students' work are presented in: http://universidaddesantiago.cl/doocc/ceweb.htm

6. Students evaluation of the experience

At the end of each academic period, a survey was conducted with the students.

This survey had only one open question:

"Give us your opinion about the use of VR in civil engineering education"

In precedent years this author has published student's answers (G. Silva 2000, 2001, 2002).

As in T. Sulbaran (2001) and in T. Sulbaran and N. Baker (2000), results always were positive considering the general students attitude.

The aspect considered in this paper was not relevant to the students in the survey results, probably because was the last step in the design process and their priority were to grade the course.

In the 2002 experience, almost all the problems appeared in precedent years were solved.

Availability of more servers' space and better Internet access speed were positive factors compared with previous years.

7. Author's evaluation of the experience

The aspect presented in this paper is the last step in a large process of c.e. design developed by the students.

It begins with the study of the site and ends with the Internet project's presentation.

This "new" way of communication is assumed by the c.e. students naturally and represents a new ability of the future civil engineers professionals.

This kind of presentation normally is delivered in a CD showing one or more animations, but rarely or never includes a VRML file, so user can't examine the 3D model of the building or the site in a freely way.

With the experience of the c.e. students gained in this work, they could construct 3D models of relevant c.e. projects without subcontract the web design with a designers firm.

Different authors, they think the 3D models of c.e. projects will aid to minimize most of the misunderstanding that occur in the construction stage.

Finally is remarkable that c.e. students can't do 3D models if they don't know exactly the steps and order that must be considered in the real project.

So their 3D models reflect their project's knowledge better than the traditional 2D blue prints and the rare physical models used in the large projects.

8. References

T. Sulbaran (2001). Impact of Distributed Virtual Reality on Engineering Knowledge Retention and Student Engagement. Georgia Institute of Technology, School of Civil and Environmental Engineering, Thesis Proposal, Summer 2001.

T. Sulbaran and N. Baker (2000). Enhancing Engineering Education Through Distributed Virtual Reality. FIE 2000, 30th ASEE/IEEE Frontiers in Education Conference, October 18 - 21, 2000 Kansas City, MO.

G. Silva (2002). VR in Civil Engineering Education. Which is the Students Expectancy? 6th International Conference on Information Visualisation, IV'02, London, England, July 10-12, 2002, Poster presentation.

G. Silva (2001). VR and WEB Page Support in Civil Engineering Education: A Recent Experience, EUROMEDIA 2001, APTEC-1, UPV, Valencia, Spain, April 18-20, 2001.

N. Murray, T. Fernando, G. Aouad (2000). A Virtual Environment for Building Construction, 17th ISARC, pp1137-1142, 18-20th September, 2000.

F. Marir, G. Aouad and G.S. Cooper (1998). OSCONCAD: A Model - Based CAD System Integrated With Computer Applications, ITcon Vol. 3 (1998).

MOBILE AGENT BASED SOLUTIONS FOR KNOWLEDGE ASSESSMENT IN eLEARNING ENVIRONMENTS

Mihaela Dinsoreanu
Cristian Godja
Claudiu Anghel
Computer Science Department
Technical University of Cluj-Napoca
Baritiu 26-29, RO-3400
Cluj-Napoca, Romania
E-mail:mihaela.dinsoreanu@cs.utcluj.ro

Ioan Salomie
Tom Coffey
Department of Electronic and Computer Engineering
University of Limerick, Ireland
E-mail: {Ioan.Salomie, Tom.Coffey}@ul.ie

KEYWORDS:
Virtual Learning Environments, mobile agents, Agent-Oriented Software Engineering, Student Assessment Service.

ABSTRACT

E-learning is nowadays one of the most interesting of the ”e-“domains available through the Internet. The main problem to create a Web-based, virtual environment is to model the traditional domain and to implement the model using the most suitable technologies. We analyzed the distance learning domain and investigated the possibility to implement some e-learning services using mobile agent technologies.

This paper presents a model of the Student Assessment Service (SAS) and an agent-based framework developed to be used for implementing specific applications.

A specific Student Assessment application that relies on the framework was developed.

INTRODUCTION

Almost every domain we know has nowadays its “e-“ Internet-based counterpart. We talk about e-commerce, e-banking, e-learning etc. Each “e-“domain emulates the traditional one in a new, virtual, Web-based environment. The major problems of creating the virtual environment involve traditional domain modeling and implementing the model using the most suitable technologies.

Our research is concerned with creating Web-based services for Virtual Learning Environments (VLE). This involves a complete analysis of the learning domain. The outcome of the analysis is the identification of the main concepts and relationships and building a conceptual model of the domain. On the other hand, the most appropriate technologies for implementing the model have to be analyzed and decided upon.

In this paper we focus on one aspect related to VLE, the Student Assessment. One of the most important educational components is the assessment of the student’s acquired knowledge. There are several issues related to assessment that should be considered: communication issues, security issues, evaluation types, student answer analysis and grading.

This paper is structured as follows: an analysis of the Student Assessment domain is presented in Section 2, considering the most important concepts, constraints etc and building the conceptual model of the domain.

Section 3 presents a possible solution based on the Mobile Agents Technology. The concepts in the Application Domain are mapped in the Solution Domain, providing therefore a computational model. The computational model was further developed as a framework that can be used for implementing specific applications.

Section 4 presents a specific Student Assessment application that was built using the framework mentioned above.

We end with a discussion of some conclusions and possible developments in Section 5.

VLE

VLEs have to provide all the necessary resources for overcoming time and space limitations existent in traditional f2f environments. Students and Instructors involved in a VLE can be located world-wide, they don’t have to synchronize their communication, and their number is not limited.

Therefore, the services provided by VLEs should be designed considering issues like accessibility, scalability, security, communication etc. In our paper we will focus on one of the services of a VLE: the Assessment Service (AS).

Student Assessment

AS provides the means of evaluating the students’ acquired knowledge. It also provides the means for a student to get valuable feedback regarding his progress. AS is a highly dynamic component of the VLE, involving both synchronous and asynchronous communication between students and instructor. In order to build a model of AS, we analyzed the assessment process, different possible scenarios, and different assessment types. Based on the analysis we identified the main concepts involved and the relationships between them.

Main Concepts Identification

Analyzing the main concepts involved in student evaluation we identified the following:

- Learning entity (the Student)
- Teaching authority (the Instructor)

- Assessment type (Compulsory Examination, Self-Assessment)
- Test
- Question Type
- Question
- Correct Answer
- Assessment procedure (as an Evaluation Engine)

The relationships between the concepts are depicted in a simplified manner in Figure 1.

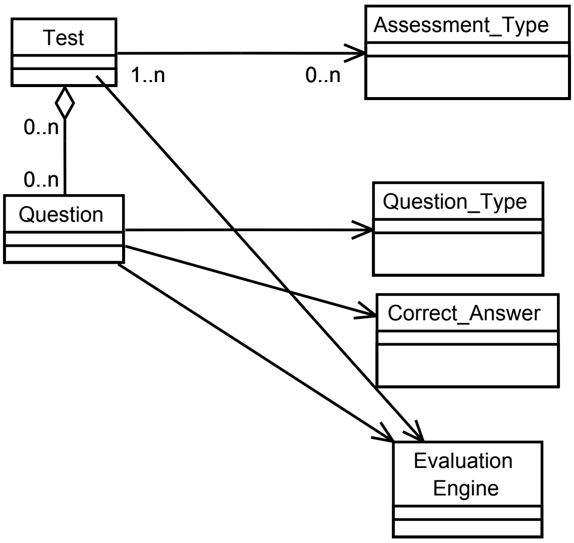


Figure 1: Assessment Concepts and Relationships

The Teaching Authority provides the tests. A test may belong to different assessments (Compulsory examinations, self-assessments, etc) and contains a set of questions. A question is associated to a question type, to one or more correct answers and also to an assessment procedure (implemented as an evaluation engine).
The evaluation engine provides the knowledge for evaluating the Student's answer against the correct answer(s) associated to a question. Student answers can be in a limited range, long natural language essays can not be analyzed.
The Student is able to access available tests, to run a test and provide his answers. The student should also receive feedback regarding his performance (the grade, the correct answers etc).

Main Functional Tasks

The next step in the analysis of the system is the representation of the functional requirements of the system. Considering the main external actors that interact with the system, the Student and the Instructor, we modeled the functional requirements in an UML-based manner as use-cases associated to the external actors.
The main functionalities of the system provided for the Student are: Visualization of Tests, Start a new Self-Assessment Test, View Test Results.
The functionalities provided for the Instructor are mainly: View Tests, Add new Test, Modify existing Test, Delete existing Test, View taken Tests, Schedule Tests
Other administration related functionalities are also considered

Non-functional Requirements

Besides the main concepts and functionalities described above, when modeling the AS service, some additional constraints have to be considered: independence of the VLE, independence of the implementation technology, scalability and accessibility. Another constraint is related to question/answer types. As mentioned above we did not consider Student answers as essays, therefore AS is more suitable for technical disciplines where the correct answers are in a limited range.

MOBILE AGENTS – AN EFFICIENT SOLUTION

Since we are dealing with a highly distributed system and considering the constraints mentioned above, we investigated the possibility to provide a solution based on agent technology. Our goal was to design a multi-agent system that fulfills the functional requirements described respecting also the discussed constraints.
In our approach the multi-agent system is considered an organization of agents. The organization knowledge and capabilities are larger than the sum of knowledge and capabilities of the individual agents (Wooldridge et al. 2000), (Zambonelli et al. 2000), (Zambonelli et al. 2001).
In modeling organizations the following factors should be generally modeled at some level of detail (Weiss 1999):

- Agents comprising the organization
- The organization's design (structure)
- Tasks that should be carried out
- The environment the organization exists in
- Stressors acting on the organization and

We started our design by mapping the functional requirements represented as use-case diagrams to a set of tasks the system has to perform.
In order to build a complete task model we considered a top-down approach decomposing more general tasks to specific subtasks. Next, we identified the necessary agent roles to perform the tasks. Agent roles define the position of the agent in the organization.
The organizational design actually consists of a set of models, each addressing one facet of the organization:

- Environment Model
  The Environment model represents the available resources and also access protocols to resources.
- Interaction Model
  The Interaction model represents the communication structure between agents.
- Role Model
  The Role model is actually the authority structure in the organization. It links also tasks to roles.

Task Decomposition

Analyzing the use case diagrams that model the functional requirements of the system, we considered the following main tasks:

Communication Tasks

Communication is a key issue from both internal and external viewpoints. The organization obviously does not exist in isolation so it has to communicate to the exterior world. On the other hand we talk about an organization, so agents are supposed to communicate in order to achieve their goals. Therefore, we considered the two main communication types:

- Communication to external actors (Student, Instructor, VLE)
- Communication inside the system (modeled by Interaction Protocols)

To provide efficient communication to human external actors a **Personal Assistant Agent** was considered. The Personal Assistant (PA) is a stationary agent living on the client machine and providing the communication interface between the external actor (Student, Instructor) and the system.

Coordination Tasks

Besides communication, coordination of the organization is also a key issue. Coordination tasks involve: handling self-assessment requests, handling compulsory examinations, generating evaluation engines, performing evaluation etc.

The system was designed as a centralized coordinated system, the core of the coordination module being a **Server Agent.** The Server Agent (SA) is a stationary agent that lives on the AS machine and is responsible with handling self-assessment requests, examinations set by Instructors, generating corresponding evaluation engines etc.

For the evaluation itself we considered an **Evaluation Agent.** The Evaluation Agent (EA) is a mobile agent that migrates on the client (Student) machine and is able to perform the evaluation. EA is loaded with an Evaluation Engine containing the complete Test (questions, answer options, correct answer) and the assessment procedure.

The creation of the EA is also SA's responsibility.

We also considered other dependencies between tasks (Weiss 1999): pooled (results of one or more tasks jointly needed to perform another task), sequential (two or more subtasks should be performed in a specific sequence), reciprocal (two tasks depend jointly on each other) .

Organizational Model

As previously stated, modeling an organization involves several concepts comprised in different sub-models of the organization. Our approach models a closed organization (no alien agents are allowed), containing benevolent, cooperative agents. We considered the following sub-models as components of our organizational model.

Environment Model

The Environment Model represents the resources available to the agents and the associated access protocols to them. We consider as resources both data and knowledge storage structures and other components (objects, servers etc) that provide specific services to agents. The design of the agents is independent of any specific resources. The Environment Model is represented by several UML-based package and class diagrams.

Role Model

This model contains the agent roles in terms of their tasks, interactions and accessible resources. As mentioned above we identified three agent roles like depicted in Figure 2.

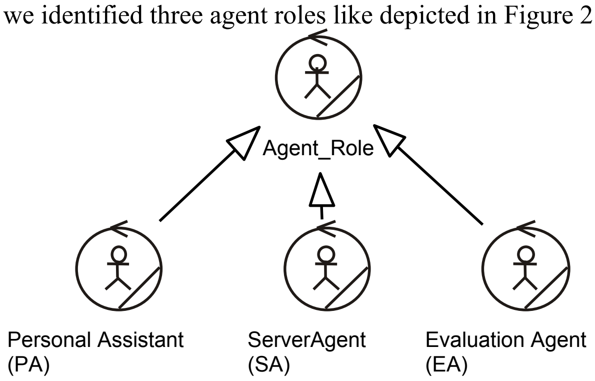


Figure 2: Agent Roles

Each role is associated to the set of tasks it's responsible for. We modeled the tasks as UML-type use-cases. In Figure 3 EA and its associated tasks is represented. EA is therefore responsible for traveling to the Student's site, for cooperating with the existing PA in order to perform the evaluation, for displaying the questions via a friendly graphical interface to the Student, for allowing the Student to enter his answers, for evaluating the answer and choosing accordingly the next question (adaptive behavior) and finally providing a result of the evaluation.

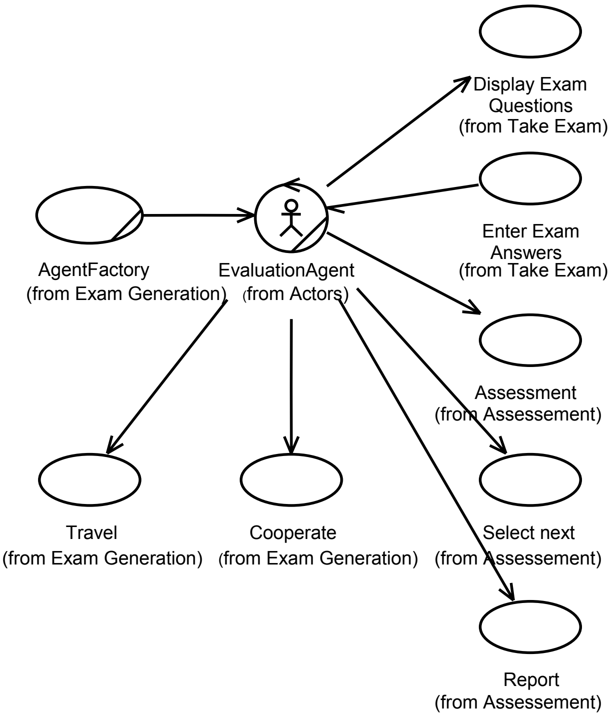

Display Exam Questions (from Take Exam)
Enter Exam Answers (from Take Exam)
AgentFactory (from Exam Generation)
EvaluationAgent (from Actors)
Assessment (from Assessement)
Travel (from Exam Generation)
Cooperate (from Exam Generation)
Select next (from Assessement)
Report (from Assessement)

Figure 3: EA and associated tasks

The agent is therefore responsible for performing several concurrent tasks. Each task defines a behavior of the agent. Agent behaviors are represented in our approach as State Chart diagrams (DeLoach et al. 2001). These diagrams contain possible states of the agent and the transitions between states. A state may have a set of associated activities defined as functions (DeLoach 2000), (Sparkman et al. 2001):

result = activity_name(param1, param2, …, paramn)

A transition occurs if the following conditions are true:

- the current state of the task is the initial state of the transition
- the trigger event occurred
- the guard has the logical value *true*
- all the activities of the initial state were performed

The general syntax of a transition is:

Trigger [guard]/ transmission(s)

A transition may generate transmissions. A transmission is either an external message sent to another agent, or an internal event sent to another task of the same agent.

In our approach, each concurrent task is modeled as a state-chart diagram associated to the agent. In Figure 4 an example of a task model for EA is shown.

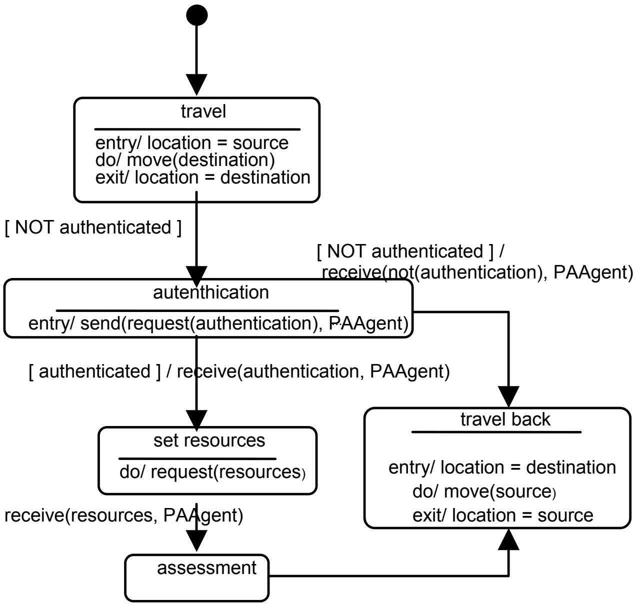

travel
entry/ location = source
do/ move(destination)
exit/ location = destination
[NOT authenticated]
[NOT authenticated] /
receive(not(authentication), PAAgent)
autenthication
entry/ send(request(authentication), PAAgent)
[authenticated] / receive(authentication, PAAgent)
set resources
do/ request(resources)
travel back
entry/ location = destination
do/ move(source)
exit/ location = source
receive(resources, PAAgent)
assessment

Figure 4: Evaluation Task of EA

Interaction Model

Interactions between agents are represented by communication protocols. These protocols are part of the social rules of the organization. The communication protocols details are represented as sequence diagrams (Bauer et al. 2001), (Bergenti and Poggi 2000), (Van Dyke Parunak and Odell 2002). We defined communication protocols between external actors and the system, and also between the agents inside the organization.

CASE STUDY

Considering the models developed in the analysis phase, we designed a general architecture of the application. The general architecture was developed as a framework in order to be used by specific applications. The main considered issues were: distribution, reliability, scalability, platform independence, data storage independence, error proof. The architecture has to be therefore well structured and layered.

We considered a multi-layered structure containing well delimited, independent modules. Modules on lower levels provide services to modules on the upper levels.

The main modules of the system are:

- GUI – User Interface Module contaning three submodules:
  - Instructor Interface Module
  - Student Interface Module
  - Admin Interface Module
- BL – Business Logic Module being together with MA (Mobile Agent Module) the core of the system
- MA – Mobile Agent Module
- DAO – Data Access Module – provides primitive data access operations (store, retrieve, update). Isolates the system from the data storage suport assuring independence.
- Utility Server – provides services to other modules. Allows for different configuration settings.

The general architecture is presented in Figure 5.

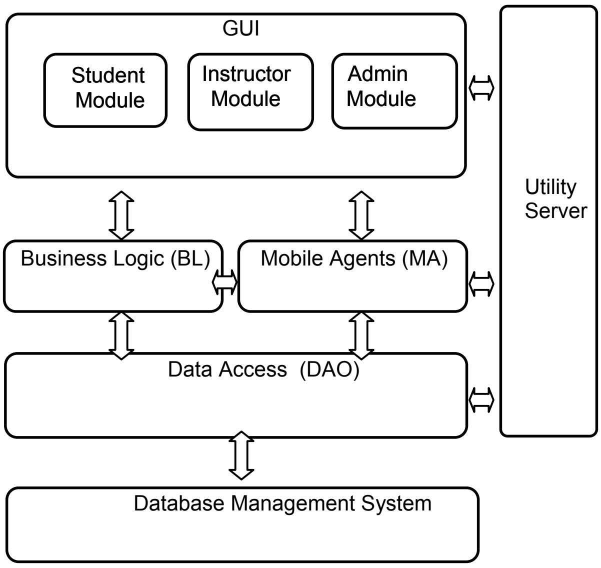


Figure 5: Framework Architecture

The implemented framework uses Java technologies (JSP, Java Beans, etc) and JADE mobile agent platform (JADE 2002). Based on the general framework presented above, a specific Student Assessment Service (SAS) was implemented and integrated in an already existing VLE. In order to define the functionalities of SAS we identified two main functional approaches:

- A Pull (Self-Assessment) Scenario initiated by the Student, who learned a certain section of a specific matter and wants to evaluate his/her knowledge. In this case the Test type is configured by the Student and no record of the assessment is registered in VLE.
- A Push (Exam) Scenario initiated by the Teacher, who enforces a certain Test type for evaluating the students' knowledge level. In this case the configuration is done by the Teacher and the result of the evaluation is recorded in VLE.

We will describe the Self-Assessment module, the Exam module being treated the same way. The Self-assessment module can be further detailed considering sub-modules that can be mapped on some of the social tasks of our system like Assistance, Exam Generation, Taking Exam etc.

For accomplishing the Assistance task we considered the need of a Personal Assistant Agent (PAA) providing an interface for the student to interact with the system. The PAA would be a stationary agent residing on the student's machine. He interacts on the other side with the SAS, communicating the student's requests and providing access to the local resources for the assessment.

We considered on the SAS side the need of a Server Agent (SA) role. The SA should be responsible with managing PAA's requests, initiating the creation of a specific Evaluation Engine corresponding to the configuration received and also initiating the creation of an Evaluation Agent (EA) responsible with the actual evaluation. SAs are also static agents residing on the SAS's site.

The evaluation has two main phases:

- an offline phase where specific domain knowledge is acquired creating the domain knowledge base. In this phase also the expert answers of the test are analyzed and structured.
- an online phase where the students' answers are analyzed and matched against the expert answers structures.

The off line phase takes place before any assessment is performed.

The need of other components of the SAS is obvious: an Evaluation Engine Factory that should create specific Evaluation Engines for specific assessment configurations. The Evaluation Engine is attached to an EA and provides its ability to analyze the student's answer and to match it against the expert answer, therefore being able to evaluate it.

Another important component is an Agent Factory that actually creates EA's. The EA is a mobile agent, loaded with assessment knowledge (the Evaluation Engine), with a set of questions and expert answers. The EA travels to the student's site and co-operates with the PAA in order to get the assessment done. The EA has an adaptive behavior depending on the student's answers.

We will not focus in this paper on the structure of the Evaluation Engine. The Evaluation Engine will be able to manage different test types like: multiple choice tests, short answers using natural language etc. using a natural language processor based on latent semantic analysis approach.

CONCLUSION

We are interested in our work to analyze and model specific areas of Virtual Learning Environments and to investigate the most suitable technologies to implement the developed models, particularly mobile agent-based technologies since we are dealing with a distributed and complex environment.

We believe that the methodology we used can be considered a foundation for modeling multi-agent systems. It takes advantage of a goal-driven approach, considers agent-specific issues like roles, tasks and interactions in the analysis phase and can be supported by a well-known modeling language as UML, therefore several off-the–shelf CASE tools being appropriate to be used. We developed a general framework that provided the foundation for a specific Student Assessment Service.

We considered future developments concerning more efficient knowledge representation models for integrating in Evaluation Engines that are suited to be ported by mobile agents.

REFERENCES

Weiss G. 1999. "Multi-Agent Systems, A Modern Approach to DAI", *MIT Press.*

Bauer B.; J.P.Mueller and J. Odell. 2001. "Agent UML: A Formalism for Specifying Multiagent Interaction", in P. Ciancarini and M. Wooldridge, editors, *Agent-Oriented Software Engineering*. Springer-Verlag Lecture Notes, Berlin, pp.91-103.

Bergenti F., A. Poggi. "Exploiting UML in the Design of Multi-Agent Systems". 2000. In *A. Omicidi, R. Tolksdorf, F. Zambonelli, eds., Engineering Societies in the Agents World - Lecture Notes on Artificial Intelligence*, volume 1972, pp.106-113, Berlin, Germany, Springer Verlag.

DeLoach S.A.; M.F.Wood and C.H.Sparkman 2001, "Multiagent Systems Engineering", *IJSEKE*, Vol.11, No. 3 231-258.

DeLoach S.A. 2000. "Specifying agent Behavior as Concurrent Tasks: Defining the Behavior of Social Agents", *AFIT/EN-TR-00-03*. Technical Report.

Sparkman C.H.; S.A. DeLoach and A.L. Self. 2001 "Automated derivation of Complex Agent Architectures from Analysis Specifications", *AOSE – 2001*, Montreal, Canada.

Van Dyke Parunak H. and J. Odell. 2002. "Representing Social Structures in UML", *Agent-Oriented Software Engineering Workshop II*, Michael Wooldridge, Paolo Ciancarini, and Gerhard Weiss, eds., Springer, Berlin, pp. 1-16.

Wooldridge M., N.R. Jennings and D. Kinny. 2000. „The Gaia Methodology for Agent-Oriented Analysis and Design", *Autonomous Agents and Multi-Agent Systems*, 3(3): 285-312, September.

Zambonelli F., N.R. Jennings, A. Omicini, M. Wooldridge. 2000. „Agent-Oriented Software Engineering for Internet Applications". In *Coordination of Internet Agents: Models, Technologies and Applications*. Springer-Verlag.

Zambonelli F, N.R. Jennings, M. Wooldridge. 2001. „Organisational abstractions for the Analysis and Design of Multi-Agent Systems", in P. Ciancarini and M. Wooldridge, editors, *Agent-Oriented Software Engineering*. Springer-Verlag Lecture Notes in AI Volume 1957, January 2001.

JADE 2002. http://sharon.cselt.it/projects/jade/

COLLABORATIVE VIRTUAL ENVIRONMENT WITH PERSONALIZED SERVICES

Valentin Cristea
Stefan Trausan-Matu
Octavian Udrea
University “Politehnica” of Bucharest
Splaiul Independentei 313
Bucharest, Romania
E-mail: valentin@cs.pub.ro; trausan@cs.pub.ro; uoctav@mymail.ro

KEYWORDS
virtual environments, collaborative systems, e-Learning, multi-agents systems, semantic Web, knowledge-based systems, personalized learning

ABSTRACT

The paper presents a collaborative virtual environment for training, whose facilities can be adapted to the needs and skills of the employees of any organization. The system has a multi-agent architecture where human and artificial agents collaborate to achieve the training and learning tasks. There are two types of agents in the environment: high-level agents, dedicated to interact with the user and to perform high-level tasks on their behalf, and low-level agents that represent the agent-based technology used to develop the collaborative environment. Among the high-level agents used in the environment, we can mention: the personal agent of the user, the information retrieval agent and the collaborative agent. These agents are used in the environment for doing collaborative training activities. The environment uses known e-Learning standards such as IMS, is based on open technologies and standards, and is developed using component based design. One of the main features is personalized training using enterprise / institutional knowledge repositories developed with knowledge management tools based on Web services and XML technologies. The system is able to integrate and use public knowledge repositories, ontologies and annotated documents. The environment is under development at the National Center for Information Technology, University Politehnica of Bucharest.

INTRODUCTION

The paper refers to personalized training offered to company’s employees as a component of their work. The proposed training environment is a framework for the integration of ontology-based technologies on the Semantic Web, with Web services technology for distributed processing, and with tools for distance and distributed training. The framework includes modules for content creation and reuse from the Internet (according to standards for eLearning, such as IMS - http://www.imsproject.org, ARIADNE - http://www.ariadne-eu.org, SCORM, AICC), intelligent search of learning materials on the web, knowledge extraction, and summarization. Intelligent tutoring technology, assuring the best possible personalization is also provided.

The modules of the project are developed as an integrated collection of web services that allows a flexible access of any trainee to the most relevant and recent knowledge and learning resources. The accomplishment of these purposes is achieved through the integration of knowledge management techniques with XML-based access to heterogeneous and distributed databases, and user friendly, intelligent interfaces available on various devices, including mobile systems.

The tools may be used in various training scenarios, ranging from simple support of courses and lectures, to virtual classes and even complex intelligent tutoring processes. The tools may be classified in: collaborative and knowledge-based tools, the latter including intelligent tutoring, a knowledge server and generation of didactic material (content creation and management).

The next sections of the paper present the collaborative services, continue with the knowledge-based processing tools and also include some conclusions.

COLLABORATIVE SERVICES

Collaborative tools are divided in two levels. The lower level includes audio-video conferencing, whiteboard, shared display, shared electronic notebook, synchronized browser, chat, and discussion forum. The higher level includes the collaborative agents that are responsible for planning and coordinating collaborative activities. Once a team is formed for doing a collaborative assignment, the collaborative agent assists the team in developing and carrying out plans. It may use a plan library from which it autonomously retrieves similar plans and associated actions, proposing them to the team, and records the new instance of the current plan as the plan is modified or developed by the team, for further use. The collaborative agent may also assist in task decomposition.

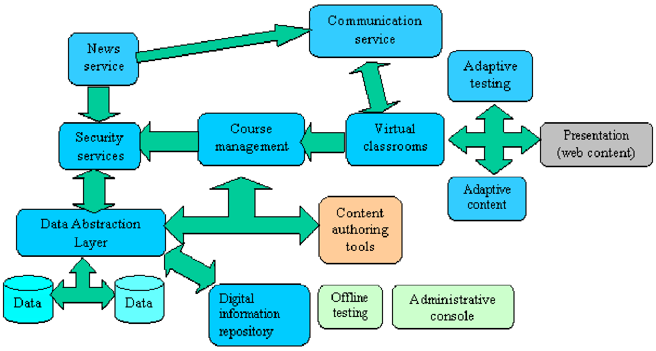


Figure 1: Platform architecture

The application as a whole is a container of tightly coupled modules as shown in Figure 1. It offers a complete environment containing:

- Communication tools for tutors and students,
- Virtual classrooms for course presentation, assessment engines and other classroom related activities such as homework, projects, discussions, etc.
- Web-based and stand-alone authoring tools
- Connection to a digital information repository
- News service
- Complex user authentication and authorization service

The application has been designed in conformity with the IMS standards. IMS together with ADL SCORM and AICC are well-known and widely implemented standards in the eLearning world. We have implemented the IMS standard also in Question and Test Interoperability, which allows imports and exports of questions, tests and other related resources from other IMS-compliant applications. Two other standards being implemented are:

- The IMS metadata specification, which relates to the packaging and exchange of course content between applications
- The IMS LIP (Learner Information package) specification, which refers to representation and exchange of user profile and "knowledge level"; this permits users migrating between LIP-compliant applications to retain their profile and knowledge level information

The communication module consists of JNDI-based discussion lists and "synchronous" tools such as chat rooms, whiteboards, etc. The service is complemented by a web-based news service and by email newsletters and discussions transcripts

The environment also contains a connection to a digital information repository, which has already been developed in our laboratories. Documents from this repository will be available to the students and tutors based on their authorization level.

Probably one of the most important parts of the application, the virtual classroom environment is designed to integrate all services normally offered to students, and more. It integrates course presentation, assessments, homework, projects, grading and communication services for both tutors and students. Obviously different actors have different roles in this scenario. While files are being submitted to the tutors as homework or projects, the tutor grades and replies to these submissions. The testing engine is fully automated both regarding the presentation and grading. This environment also allows tutors and teachers to shape the on-line courses from the existing materials with regard to the content presented, the course timeline, and the student responsibilities.

The different modules described above have been developed by different teams using a global application design and project management. Therefore, third party modules can be integrated with ease as long as they are conformant with the overall "philosophy"

A data abstraction layer module is used, which offers the single access point to databases. The rationale of introducing it in the architecture was:

- permit parallel development of the environment processing modules
- support legacy databases
- permit modifications in the database structure without affecting already implemented processing modules.

Each processing module defines the structure of the data it intends to access. The description is specified in XML, and represents a definition of the database used by this module. Additional elements can be defined, such as:

- data types (Java classes) that map the data fields in the database. This way, a request for data from DAL returns an object that is defined by the processing module.
- types of entities that are common to several databases.

During the development of the module, several advantages became evident and very useful:

- transparent access to multiple databases
- modifications of databases without recompiling or re-deploy of the processing module; just a modification of the data descriptors is sufficient.
- Mapping of data on data types defined by the user
- Offers Connection pooling for JDBC drivers that not support internally this mechanism
- Offers an automatic mechanism for logging errors
- assures a basic support for the logical correctness of the applications.

KNOWLEDGE-BASED PERSONALIZATION ON THE WEB

For providing personalized services, *users' models* (Kobsa and Whasler, 1989) are needed, that contain knowledge about the user profile and context of interaction, together with knowledge regarding possible paths of interaction and knowledge allowing selecting among them. Representation and processing of these types of knowledge may vary from simple preference and log files, where no artificial intelligence is involved, to complex knowledge processing environments.

In personalized, adaptive learning, knowledge based tools for the development of *student models* are needed (Sleeman and Brown, 1982). These tools track student's activity and interactions with the learning material, analyze his answers and texts he writes, identify needs or interests and evaluate his psychological profile and learning style. Socio-emotional intelligence issues should also be considered for tailoring the learning process.

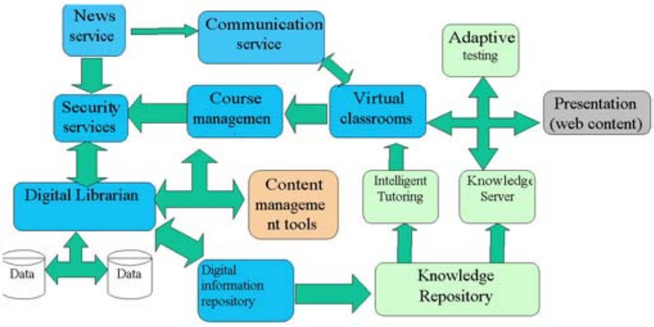


Figure 2: Knowledge support architecture

One important component of the student model is what knowledge he has, what knowledge he does not have and what knowledge he has wrongly (Dimitrova et al., 2000). This facts are derived from answers at questions, from the analysis of essays (texts) written by the student, from student's interactions. The facts may be further used for dynamic Web pages and test generation and lesson planning. This knowledge attributed to the student is put in correspondence with the general knowledge about the considered domain (the *ontology* of the domain) and inferences are made towards planning future interactions (e.g. what knowledge items must be learned further by the student).

A group of tools permit the automatic generation of multiple answers tests and of systematic collections of Web pages (Trausan-Matu and Negreanu, 1996, Trausan-Matu et al., 2002). These may be also used in connection with the intelligent tutoring system or for knowledge navigation on the knowledge server

Our environment includes a collection of tools and repositories that integrates collaborative techniques on the Web with knowledge-based methods, agents, and multiple purpose XML-based annotation (metadata, exchange of reusable components, personalization, and knowledge representation) that empowers personalization.

Knowledge management in our environment uses general ontologies combined with ontologies specific to the domain of interest of the trainees. It is able to integrate newly extracted knowledge by text and data mining methods. Graphical knowledge visualization and editing is also provided. Specific ontologies can be developed considering also the existing databases. Intelligent, personalized retrieval of latest information is provided.

Content creation is facilitated by semantics-oriented graphical editors that enable the authoring of learning components annotated according to metadata standards IEEE-LTSC (http://ltsc.ieee.org), ARIADNE, and Dublin Core (http://purl.org/DC). A direct effect will be also the possibility of reusing learning components from the Web. Learning components are stored also locally in information repositories. Specialized editors and other interfaces are provided for the management of the intelligent tutoring services (e.g. ontology editor). Standard ontology and knowledge representation are used (e.g. OIL, see http://www.ontoknowledge.org/oil/).

Our personalized learning system was designed for conformance with widely known eLearning standards such as IMS or the above mentioned metadata standards. This approach offers several advantages including, but not limited to:

- Facilitation of content exchange with platforms conforming to the same specifications
- Opening the architecture to allow external module integration

Our platform is currently applying these principles, therefore allowing exchange of the following types of information with other similar applications:

- Exchange of user profile and background information, including estimated preparation level and full training history is accomplished through the implementation of the IMS Learner Profile Information (LIP) specification
- Exchange of several types of learning content (e.g. lecture notes, practical exercises, course support materials) is accomplished through the use of IMS Metadata and Content Packaging specifications
- Exchange of test information, including questions, tests, grading and evaluation information, as well as full result history is accomplished through the use of IMS Question & Test Interoperability (QTI) specification.

One of the main ideas of our approach is the use of the Semantic Web for personalized training in a given domain. That means the extraction and adaptation of knowledge from the Web, for a given learner, in a given context. Knowledge bases (ontologies), documents (texts), and learning objects in standard XML-based languages are the sources for content creation.

The Web is a huge, permanently changing hypermedia on Internet, browsable with very simple, direct manipulation interfaces. Its explosive growth in only several years is the best proof of its usefulness. Two of the causes of this phenomenon are, probably, the ease of "publishing", of communicating something through text and/or images on the Web. From the other direction of the communication process, it is very easy for everybody to explore the network of Web pages. As a consequence, we notice the extremely dynamic character of information nowadays, the availability of the Web today having definitely changed the information scenario. The time between the appearance of new information in some domain and the use of this information by people has extremely shortened comparatively with some years ago. Therefore, information may become obsolete very quickly or be replaced by some other information. A good tutoring system should consider this scenario, and consequently be able to update its information continuously.

The process of extracting and using the most relevant information from the Web involves three phases: information acquisition performed by searching the Web, knowledge identification by semantic editing and the usage of this knowledge (Trausan-Matu et al., 2002).

The domain ontology (a declarative knowledge base containing concepts and relations among them) plays important roles in each of the three mentioned phases. The keywords used by a Web spidering agent for the search of relevant documents are obtained from the domain ontology. The process controlled by this agent searches on the Web by means of activating a number of search engines. During this phase, data mining techniques may be applied in order to better select automatically the fit between the requested information and the retrieved one.

The knowledge-based services are also used for the intelligent retrieval of relevant documents on the Web and for text mining. For these purposes statistical natural language techniques and the lexical ontology WordNet (http://www.cogsci.princeton.edu/~wn/) are used. The information retrieval agent is able to travel in Internet and access multiple information sources for seeking relevant data. In contrast to traditional Web search engines, the information agent is capable of a semantic interpretation of the retrieved information, of filtering this information according to user's preferences and criteria, and of a heuristic classification of data based on the user's profile.

Our environment is in the process of including also a multi-agent system that will provide another dimension of the collection of personalized services. For example, the search of relevant documents on the Web is performed by agents (Trausan-Matu et al., 2002). An automatic tutor agent will be used to guide the student through the on-line course material and to give different hints to questions that are to be answered by the student.

The platform was designed for conformance with widely known eLearning standards such as IMS. This approach offers several advantages including, but not limited to:

- Facilitation of content exchange with platforms conforming to the same specifications
- Open architecture that allows external module integration

Our platform is currently applying these principles, therefore allowing exchange of the following types of information with other similar applications:

- Exchange of user profile and background information, including estimated preparation level and full training history is accomplished through the implementation of the IMS Learner Profile Information (LIP) specification
- Exchange of several types of learning content (e.g. lecture notes, practical exercises, course support materials) is accomplished through the use of IMS Metadata and Content Packaging specifications
- Exchange of test information, including questions, tests, grading and evaluation information, as well as full result history is accomplished through the use of IMS Question & Test Interoperability (QTI) specification.

These knowledge-based services are also used for the intelligent retrieval of relevant documents on the web and for text mining. For these purposes statistical natural language techniques and the lexical ontology WordNet are used. The information retrieval agent is able to travel in Internet and access multiple information sources for seeking relevant data. In contrast to traditional Web search engines, the information agent is capable of a semantic interpretation of the retrieved information, of filtering this information according to user's preferences and criteria, and of a heuristic classification of data based on the user's profile.

An automatic tutor agent is used to guide the student through the on-line course material and to give different hints to questions that are to be answered by the student.

The knowledge server will provide web-based access to the above ontologies and knowledge bases. This is a mean of examining the knowledge used by the intelligent tutoring system for the learning regimes.

A group of tools will permit the automatic generation of multiple answers tests and of systematic collections of web pages. These may be also used in connection with the intelligent tutoring system or for knowledge navigation on the knowledge server.

CONCLUSIONS

The training environment:

- provides a flexible and easy to use environment for both students and tutors
- uses adaptive content based on user preferences and preparation level, both for course and test preparation and analysis
- adapts easily to a specific domain by incorporating an adequate specific ontology
- provides interoperability with other applications conforming to a similar set of standards contain both presentation and content authoring services Flexible, standardized, adapted to enterprise needs and to trainees profiles (including emotional intelligence)

The environment is under development at the National Center for Information Technology (University Politehnica of Bucharest). The design aimed at obtaining platform independent components that permits a rapid deployment on different premises. The service has been tested and deployed on two platforms using the following configurations:

- an IBM Netfinity machine running Linux, Websphere Application Server and IBM DB2 UDB
- a Sun Enterprise 10000 machine running Solaris, iPlanet Application Server, Oracle 8i

The application is still under development and thanks to its modularity it can be extended to provide more advanced features. The virtual classrooms environment, the assessment engine, the user security service, the student-tutor communications, and the content authoring tools are currently undergoing their test stages.

REFERENCES

Baker S. 1997. *CORBA Distributed Objects Using Orbix*. Addison Wesley.

Cerri, S., Gouarderes, G., Paraguacu. 2002. *Intelligent Tutoring Systems.* Springer.

Conati, C., Zhou. 2002. "Modeling students' emotions from cognitive appraisal in electronic games". In *Intelligent Tutoring Systems,* Cerri, S., Gouarderes, G. Paraguacu (Eds.). Springer, 944-954.

Cristea V., A.M.Florea, A.M.Stanescu. 2001. Collaborative Enterprise University Pilot Architectural Model. In *Proceedings of the ICE 7th International Conference on Concurrent Enterprising*, Bremen, Germany, 27-29 June 2001

Cutcosky, M.R., et al. Madefast. 1996. "Collaborative Engineering on the Internet". In *Commmunications of the ACM* 39, No. 9 (Sept.).

Detmer, W.M., Shortliffe, E.H. 1997. "Using the Internet to Improve Knowledge Diffusion in Medicine". In *Communications of the ACM* 40, No. 8 (Aug.).

Dimitrova, V., Self, J., Brna. 2000. "Maintaining a Joinly Constrcted Student Model". In S.A.Cerri (ed.), *Artificial Intelligence, Methodology, Systems, Applications 2000*, Springer-Verlag, ISBN 3-540-41044-9, pp.221-231.

Evans E., D.Rogers. 1997. "Using Java applets and CORBA for Multi-User Distributed Applications". In *IEEE Internet Computing* 1, No.3 (May/June).

Fatoohi R., D. McNab, D. Twenten. 1997. "Middleware for Building Distributed Applications Infrastructure". In *NAS Technical Report*, Dec 1997.

Filman R, S.Pant.1998 "Searching the Internet". In *IEEE Internet Computing* 2, No.4 (Jul/Aug)

Kobsa, A., Wahlster, W. (eds.). 1989. *User Models in Dialog Systems*, Springer Verlag, 1989.

McFall. 1998. "An Object Infrastructure for Internet Middleware". In *IBM on Component Broker, IEEE Internet Computing* 2, No.2 (March/April)

Sleeman, D., Brown, J.S. 1982. *Intelligent Tutoring Systems*, Academic Press, 1982.

Trausan-Matu, St, D. Maraschi, S. Cerri. 2002. Ontology-Centered Personalized Presentation of Knowledge Extracted From the Web. In S.Cerri, G.Gouarderes (eds.), *Intelligent Tutoring Systems 2002*, Lecture Notes in Computer Science 2363, Springer. 259-269.

Trausan-Matu, St.,. Negreanu, L. 1996. *Sistem inteligent de asistare a instruirii,* Research report RACAI, RR-14, Romanian Academy, June 1996.

AUTHORS BIOGRAPHIES

VALENTIN CRISTEA is a graduate of the Control and Computers Faculty, "Politehnica" University of Bucharest. In 1980 he received the Ph.D. degree with a thesis on Resource Management in a Conversational Computer System. Since 1993 he is professor of the Computer Science and Engineering Department, "Politehnica" University of Bucharest. His main fields of expertise are Computer Network Software, Parallel and Distributed Processing, and Communication Protocols. He is the director of the National Center for Information Technology – CoLaborator, which is based on environments for collaborative education and research. Valentin Cristea is member of the IEEE, and of the ACM Society. Email: valentin@cs.pub.ro.

STEFAN TRAUSAN-MATU was born in Bucharest, Romania, where he obtained the engineer (1983) and PhD (1994) degrees, at Bucharest Politehnica University (PUB). He was the chief of Artificial Intelligence Laboratory of the Institute of Informatics until 1994, now is professor at the Computer Science and Engineering Department of PUB and principal researcher I at the Romanian Academy Institute for Artificial Intelligence. His research interests are knowledge-based systems, semantic web, text mining, e-Learning, human-computer interaction. Web page: http://www.racai.ro/~trausan. Email: trausan@cs.pub.ro.

OCTAVIAN UDREA is a graduate of the Control and Computers Faculty, "Politehnica" University of Bucharest. Since 2001 he has been the technical project leader and application architect for the National Center for Information Technology's eLearning platform. He is also the technical project leader for several other projects as part of NCTI and IBM E-Business Academy laboratories of the Control and Computers Faculty, as well as an instructor for courses held by the IBM E-Business Academy laboratory. His main research interests are Distributed Computing, Real-time Operating Systems and Java Enterprise technologies. E-mail: uoctav@mymail.ro.

SECURITY ISSUES WITH E-LEARNING

Matthew Warren and Michelle McDougall,
School of Information Technology,
Deakin University,
Victoria, 3217,
Australia
E-mail: mwarren@deakin.edu.au

KEYWORDS
E-learning, Security and Information Technology.

ABSTRACT

Within Australian there has been a growth in the use of E-learning technologies to help overcome the issues of distance in such a large country. The paper looks at the security issues and problems that have arisen in regards to an Australian University and their use of E-learning technologies.

INTRODUCTION

Deakin University, Australia is a relatively young institute known for its innovative approaches to teaching and learning. Communications technology is used extensively by students and staff, both for administration and within courses, enhancing the learning experience and preparing future graduates for life and employment in an e-world.

Deakin was the first Australian university with an Internet address and it has provided free IT and email accounts for students since 1987. Deakin has become the primary provider of off campus (i.e. distance) courses to undergraduate and postgraduate students within Australia (Raitman and Zhou, 2002) and in 2001, 40% of enrolments were for courses being taken in off-campus mode (Deakin University, 2001).

Of the 1059 computers in general purpose laboratories, 70% are less than one year old and 95% are less than two years old. There are networked computers in the student residences, and there are currently 3439 computer conferences for teaching and student life (Deakin University, 2003a).

Deakin first used computer mediated communication (CMC) in 1981 and has since developed, tested and mainstreamed further approaches supporting the use of CMC and other online technologies for teaching. Teleconferencing and video conferencing are used in many subjects and students have Web access to unit learning materials in many subjects for all Faculties and Schools (Deakin University, 2003b).

Over half of student enrolments are for units in which information and communications technologies are used for core educational processes. Two thirds of all Deakin students make use of virtual classrooms in any given year (5).

All students use sophisticated Web and CD-ROM multimedia materials to learn key aspects of first-year units in the Faculties of Health and Behavioural Sciences, Science and Technology, Business and Law and Education (5).

All Faculties and Deakin Australia (Deakin's commercial arm) make significant use of computer based objective testing (5).

Deakin staff and students mainly use the Learning Management systems WebCT, TopClass and FirstClass as the basis for its online teaching and learning environment.

E-LEARNING

E-learning involves more than the availability of texts or lecture notes online or merely the use of the Internet within a course. Different electronic media may be used including computers, the Internet, intranet, CD-ROMS, DVDs, audio and video tapes and virtual environments.

E-learning provides a convenient and flexible learning experience that can complement or replace traditional face-to-face teaching for on campus students.

For off campus students, e-learning is a more engaging and interactive method of learning than conventional approaches such as the post and telephone. Students can work in an environment that seems much less isolated from their lecturers and fellow students. Accessibility is also improved to students with geographical or time issues.

Students can work at their own pace, have up to date course information that can revisited at any stage and have synchronous or asynchronous contact with the unit's staff or other students.

KEY FEATURES

Some of the important groups of key features that have been identified by previous research (Smissen, 2002) for online teaching and learning include:

- Easy to use;
- Platform and browser compatibility;
- Synchronous communication;
- Asynchronous communication;
- Collaborative work;
- Online assessment;
- Result management;
- Assignment submission.

Synchronous communication can include text chat, live audio or video, shared whiteboard, file sharing, application sharing, viewing another's desktop, remote control of another's desktop or private chat groups (Smissen, 2002).

Asynchronous communication can include bulletin board, threaded discussion, email, seeing who's read messages, adding attachments, redirecting messages and file sharing (Smissen, 2002).

LEARNING MODELS

The traditional instructor-centric teaching and learning environment (figure 1) and the learner-centric model (figure 2) emulating e-learning are provided by Tronsden (Tronsden, 1998) and explained further by Raitman and Zhou (Raitman and Zhou, 2002).

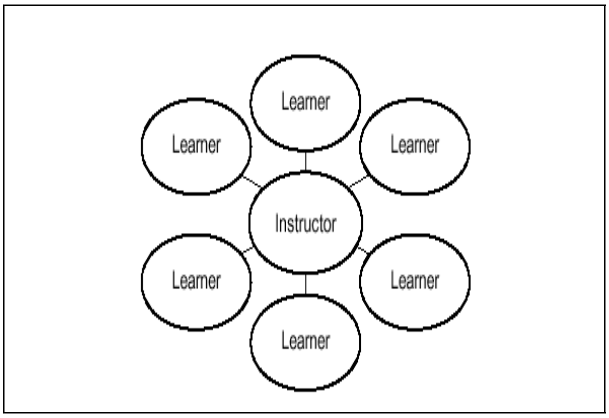
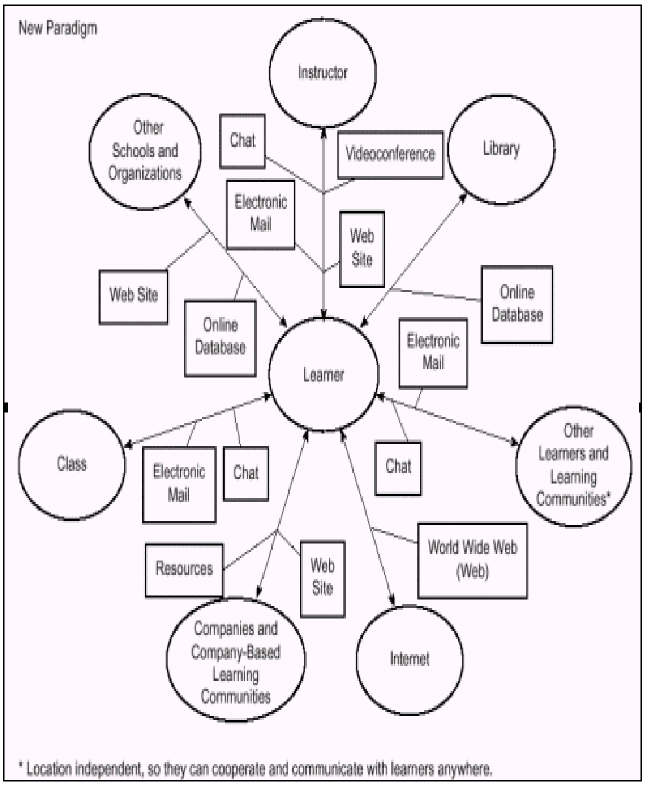


Figure 1. Traditional Paradigm

Many Australian Universities are now using technology to move from the traditional teaching paradigm to the new E-Learning Paradigm. Figure 2 shows how the impact of different technologies allow the E-Learning paradigm to develop and how the different technologies can impact upon the learner.



* Location independent, so they can cooperate and communicate with learners anywhere.

Figure 2. New E-Learning Paradigm

E-LEARNING TECHNOLGIES

Deakin University uses a number of E—learning technologies and the following are a description of those technologies:

WebCT

The School of Information Technology in the Faculty of Science and Technology makes extensive use of WebCT.

WebCT is a web-based tool that incorporates the school's intellectual and technical resources to serve as a store of information and facilitate a more flexible learning experience for students.

Course notes, class news, results and other relevant materials are kept online for individual units and updated when required. Additional features include: facilities for asynchronous or synchronous discussions, a whiteboard, student tracking, internal course email and quizzes.

The WebCT folder for a particular unit is accessible only by the teaching staff and students associated with that unit (Trondsen, 1998).

FirstClass

FirstClass is an internet based communication system which provides real-time chat, group conferences with threaded messaging, email, a community directory, file sharing and many other useful features (Deakin University, 2003c).

Staff and students involved in a unit have access to meet online and communicate with each other in small, private, study groups or in more public discussions.

TopClass

TopClass is a more recently utilised online teaching and learning management system. It is used to deliver web-based course material and assessment and to facilitate communication and collaboration among staff and students, in the form of mail, announcements and discussion lists (Deakin University, 2003d).

SECURITY CONSIDERATIONS

Security problems have arisen through the use of E-learning technologies in regards to security. A number of different situations have developed which show the weaknesses associated with the security mechanisms underpinning the E-learning systems.

The problems related to:

- Passwords – a single default method is used to generate students passwords. The problem is that the same simple method is used to generate thousands of students logins. If the method of password creation was made public it would compromise thousands of passwords at the start of the semester. The other problem is that students are then expected to change their default password. The problem is that there is not a mechanism to force the students to change their password and some students use the default password for the entire semester.

- Guessing URLS – students can bypass the security authentication mechanisms by directly guessing a URL address. For instance if a lecturer calls an assignment solution – solution.htm and stores it in the SCC209 (unit) directory, the student can directly go to SCC209/solution.htm page avoiding the security authentication mechanism. The only solution to this problem is to use abstract names for the naming of pages eg, instead of solution.htm use ssdddccc.htm.

- Problems with Software – the E-learning systems do not log students out of the system when they finish. This means that another student can use the computer after the authorised student has finished their session and gain access to their E-learning account. This allows some one to masquerade as an authenticate user and then for example post offensive messages to other students or staff.

- Non Repudiation - the E-learning systems allow students to submit their assignments, The problem is that the systems do not allow non repudiation to be carried out. Staff cannot prove that assignments submitted were submitted by the correct student. The system also do not show that a student tried to submit an assignment and failed e.g. a session crashing during the upload process.

- Complexity of E-learning systems – due to the complex nature of the E-learning systems novice users may mis-configure the system they are using. This means that staff may not set-up the security features correctly which allow students to gain access to information that they should not have e.g. assignment solutions.

Many of the security problems and weaknesses described are often not discussed by E-learning researchers. The authors experiences have shown that E-learning systems are not secure and can easily be open to abuse.

The Security implication of E-learning needs to be fully considered. What the authors will be conducting is the development of security guidelines that focus upon E-learning security. These guidelines will be focused towards the protection of systems from the user viewpoint and administrator viewpoints. These guidelines will be reported upon in future research papers.

CONCLUSIONS

The major advantage of E-learning out weighs any disadvantages relating to security problems. Previous research Irving (Irving, 1995) & Simon (Simon, 2000) have identified the advantages that students believe that E-learning has. These include:

- On-line courses were more convenient for them than traditional courses;

- The virtual classroom gave them the chance to work on the materials at times of their own choice;

- The communication by computer and modem was very comfortable for students and teachers;

- Working at their own pace allowed them freedom in the process of learning;

- Positive learning experiences were felt by the students because they did not have to go to the classroom;

- They had better access to their teacher in the virtual classroom;

- Gained greater knowledge was gained about computers, finding information, and working independently.

E-learning is a powerful tool that all Australian Universities are now using. E-learning is here to stay but, the security issues and problems need still to be addressed.

REFERENCES

Deakin University (2001) *Deakin University – Pocket Statistics*. 2001, Deakin University Planning Unit: Australia. Online:
http://www.deakin.edu.au/planning_unit/statistics/2001/stats/#student_enrolments
Accessed: Jan 3rd, 2003.

Smissen, I. (2002). "Requirements for Online Teaching and Learning at Deakin University: A Case Study", Deakin University, Australia. Online:
http://www.deakin.edu.au/~ismissen/ausweb02/paper.html
Accessed: Jan 3rd, 2003.

Deakin University (2003a). Deakin University website: 'Online learning Tools' Online:
http://www.deakin.edu.au/visitors/welcome/tools2.html
Accessed: Jan 3rd, 2003.

Deakin University (2003b). Deakin University website: 'Online learning' Online Access'
http://www.deakin.edu.au/visitors/welcome/access.html
Accessed: Jan 3rd, 2003.

Deakin University (2003c) Deakin University website: 'Online teaching and learning - Firstclass' Online:
http://www.deakin.edu.au/students/teach_learn/firstclass.html
Accessed: Jan 3rd, 2003.

Deakin University (2003d) Deakin University website: 'Online teaching and learning- Topclass' Online:
http://www.deakin.edu.au/students/teach_learn/topclass.html
Accessed: Jan 3rd, 2003.

Irving, Richard (1995), *A Study of Computer-Modem Students: A Call for Action.* American Educational Research Association.

Raitman R. and Zhou W. (2002) *E-learning: A Literature Review*, Deakin University, School Of Information Technology, Technical Reports TR C 02/21, Australia.

Simon, Steven J. (2000), The relationship of learning style and training method to end-user computer satisfaction and computer use: A structural equation model. Information Technology, Learning, and Performance Journal, Morehead, Spring 2000, Vol. 18:1, Pg. 41-59.

Trondsen, E. 1998. "The New World of Technology-Based Learning" SRI Consulting Business Intelligence.

NETWORK MONITORING AND MODELLING

An approach to a methodology and implementation of a Network Monitoring Analysis Tool

Elias Aravantinos

Dr. Petros Ganos

Aristidis Ilias
Research Academic Computer
Technology Institute

Riga Feraiou 61

GR- 26221 Patras, Greece
E-mail: eliasara@cti.gr
E-mail: ganos@cti.gr
E-mail: ilsadis@cti.gr

Dr. Christos Bouras
Computer Engineering and
Informatics Department
University of Patras
GR-26500, Patras, Greece

Research Academic Computer
Technology Institute
E-mail: bouras@cti.gr

KEYWORDS
Network monitoring, IP Based Networks, SNMP, Real Time System, Network resources, Databases.

ABSTRACT

In this paper we describe the design and implementation of a network monitoring analysis tool. The network resources were configured to support SNMP under various operating systems, thresholds definitions and events. The data were collected by a network monitoring system and handled according to the values of the variables or the event type. A real- time report to a database was established via adaptive scripts. A Network Node Manager (NNM) will inform an external, remote database about the network status by sending real time data containing alerts and events. The whole tool is called Network Management Analysis Tool (NMAT).

INTRODUCTION

A fundamental division among network monitoring systems is related to whether the monitoring is done off-line (while only test messages are flowing), on-line (while user traffic is on the network) or both. Techniques for monitoring networks, when they are out of service, involve the generation of test sequences, monitoring and analysis of the results, to determine measurable levels of performance and protocol conformance. Such tests may be implemented in a test bench environment, manufacturing floor, network operations centre or during an outage in an operational network.

The process of on-line monitoring and analysis may be performed on a continuous basis, scheduled at various times of day or invoked only when circumstances (such as load changes or quality concerns) demand a closer scrutiny. Continuous monitoring (intrusive testing) is the most costly alternative in terms of resources and time, but is also the most beneficial from a user's viewpoint.

Monitoring may be divided into two categories according to the layers involved:

- Monitoring and analysis of the physical and network layers referring to the OSI model
- Monitoring and analysis of the hardware platforms and operating systems

A very large number of access points (network nodes) may be required and they would usually be dedicated to the monitoring system. The amount of data that needs to be collected and transported to a local/remote database may also be very high. The benefit, however, is that proactive management and dynamic prediction may be a reality.

Network performance measurement is an extremely broad area and we can only briefly mention some of the more relevant work. (Carter and Crovella 1996a, b) present two tools "bprobe" and "cprobe" to measure the bottleneck link speed and competing traffic respectively, on a path using ICMP ECHO packets. Since these tools do not use TCP, they are not able to capture any TCP related effects that an application might experience. Van Jacobson proposes the tool "pathchar" (Jacobson 1997) that estimates bandwidth on all hops of a path and hence can take a very long time. It also requires root access, making it less desirable for grid environments. The tool "Treno" (Mathis amd Mahdavi 1996) emulates an idealized TCP, which makes the measurements independent of host-specific TCP implementations but not representative of what applications would experience. "Treno" also needs root access. Topology-d (Obraczka 1998) uses "ping" and "netperf" to make measurements between all pairs within a group and then computes a minimum-cost logical topology. The Network Weather Service (NWS) (Wolski 1998) uses TCP to send small, fixed-size probes measured in kilobytes, with a frequency that is tuneable, but typically ranges from tens of seconds to several minutes. Performance measurement systems, such as the National Internet Measurement Infrastructure (NIMI) project (Paxson 1998) are designed for arbitrary Internet hosts. The Cooperative Association complements this work for Internet Data Analysis. The goal is to develop metrics and tools for analyzing traffic across the Internet. Additionally, the data from such tools intend to be used by "higher-level" systems. Lowecamp (Lowecamp 1998) also provides an API whereby applications can pose flow and topology-based queries. AppLeS (Application

Level Scheduler) (Su 1998) uses information from the NWS to schedule distributed resource-intensive applications.

Finally other components will use "Gloperf" information in similar resource discovery and allocation functions. We note (Lee 1999) that "Gloperf" is designed to enhance portability. It makes end-to-end TCP measurements. Storing "Gloperf" data in a directory service provides data discovery and access. However, there are missing some measurements to allow applications to probe network resources and getting fresh data.

The purpose of this paper is to approach a methodology and implementation of a Network Monitoring Analysis Tool (NMAT) during the operation of an intranet network containing hundreds of terminals established in enterprise buildings, airplanes, ships etc. The aim is to collect automatically and in real-time values of monitoring data in a storing schema, a database for instance and finally make decisions at a human level.

This paper is organized as follows: First the tool design is presented in main principles. The next section describes the proposed phases of functionalities, containing techniques, specifications and results of network monitoring. Finally future work and concluding remarks are provided.

ARCHITECTURE

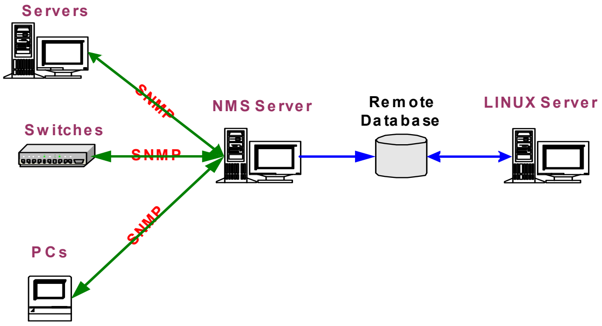
The Network Management System (NMS) used, is HP Open View Network Node Manager (NNM). It is a commercial, well-known and reliable platform that implements the Simple Network Management Protocol (SNMP). The NNM collects by polling in tuneable time intervals the SNMP data from the managed network devices of the intranet and automatically sends the required values of data to a database. The tables are filled by these values, which could be exported in various metadata formats in order to realize a data statistical analysis. Then the analysis will be assessed and will also be the input to form the predictions of network behaviour and performance. The last step concerns decisions and reconfiguration processes.

The whole process is described as following:

- The NMS is fully configured.
- The general database is created (installation, table creation).
- The network resources are configured to support SNMP (agent) under Linux and Windows operating system.
- Thresholds are defined and events are configured at the NMS.
- Data are collected by the NMS and handled according to the values of the variables or the event type.
- Real-time report is imported to the database via a script language.
- The events are classified by an appropriate value for later data manipulation and analysis.
- Prediction and decision analysis about the network components performance are made.

The tested network is an Intranet, supported by a Fast Ethernet network with bus topology. The Fast Ethernet network cable is Unshielded Twisted Pair category 5 (UTP cat 5). The tested network consists of a certain number of terminals, servers, Local Area Network (LAN) router and LAN switches. Each LAN segment is located in a different physical place in the Intranet. The LAN switches are Fast Ethernet network devices that provide each sender/receiver pair with full 100 Mbps capacity. Each port on the switch gives full bandwidth to a single server or client station.

Furthermore, there is a switch-router connected to the Internet to serve the Intranet. The terminals and the servers are connected to different LAN segments. It was supposed that the tested terminals are located in different physical LANs instead of VLANs. All the terminals are connected to different switches instead of VLANs of the same switch.



Figures 1: System Architecture

Figure 1 describes the tested system architecture. The network monitoring process was based on the SNMP, managing all devices, like servers, switches and PCs. The Remote Database was hosted on a LINUX Server and NMS established a one way communication link, each time that appeared the need of sending data to the database.

The NMS server supported Windows 2000 Server operating system and the terminals Windows 2000 and LINUX.

An SNMP agent resides in each terminal machine in order to be remotely managed by NMS. The managed devices run software (SNMP agent) that enables them to send alerts (traps) when they recognize problems (for example, SNMP link failure). Upon receiving these alerts, management entities are programmed to react by executing one, several or a group of actions (referring to scripts) including event logging, changes in device configuration and system shutdown.

The management entities poll the end stations to check the values of certain variables in Management Information Bases (MIB). A MIB is a SNMP structure that describes the particular device being monitored. Polling can be automatic or user-initiated and agents in the managed devices respond to all polls. Agents are software modules that first compile information about the managed devices in which they reside, then store this information in a management database and finally provide it (proactively or reactively) to management entities within NNM via SNMP.

IMPLEMENTATION ISSUES

In our tested network an external database is used instead of an internal NNM database. The reason is to create applicable tables with different relations compared to the compact, complicated and inflexible NNM's database model. As a result we propose a scheme, which could appear database platform independency. Moreover the use of an external database provides interoperability, because it is possible to comply with several software tools, like JAVA-based platforms or other experimental tools developed for similar purposes.

The NNM collects by polling in tuneable time intervals the SNMP data from the managed network devices of the intranet and automatically sends the required values of data to a database.

Any corporate network has two types of end users: the organization's typical employee and the network operator who is responsible for maintaining the end-to-end service. The organization's typical end users and the success of the business applications that live on the network are the ultimate beneficiaries of all the various network monitoring activities that are performed within the network. However, the end users are least aware of the infrastructure behind the service. They simply need a quality of service that meets their business needs and achieves certain invisibility. It is the goal of achieving this "invisibility" of consistently good service that makes the added cost of monitoring and analysis tools justifiable.

The MIBs are used to monitor and access network and system information variables. They are either vendor MIBs or experimental MIBs. The vendor MIBs concern mainly the network switches while the experimental ones are used to monitor special parameters like the hostmib.mib which monitors system information (for example CPU load).

The SNMP polling interval was generally configured at 60 seconds. The NMS polls each critical variable every 10 seconds receiving traps and events. This polling interval concerns mostly the servers and the switches of the Intranet. The rest of the nodes are controlled from NNM every 1 minute.

The Network Management Analysis Tool (NMAT) reports all events in the external database with the help of two script-files (Brown 1999) (Events.ovpl and Events_value.ovpl) developed in Perl scripting language. These files are useful for the automatic actions, in order to insert the network variables values polled from NNM into the Database. Both scripts are running with the help of Open View Perl installed with NNM and configured with the help of the Event Configuration module of NNM.

The main problem we had to solve was the handle of consequential events, which are events that occurred exactly the same time. After making several experiments, we noticed that there was a unique variable generated by every event. This variable helped to distinguish the consequent events as discrete instants and store them into a remote database.

The events are separated in two categories:

- Events with value
- Events without a measured value

Events with value

The configuration of the events that do contain value should include the following automatic action command:
OVHIDESHELL cmd.exe /c "Events_value.ovpl $s $ar $N $3 $8"

Through this command the Perl file is called and rest of the symbols present the specific event variables based on NNM's manual.

```
#!/opt/OV/bin/Perl/bin/perl
my $param_1 = $ARGV[0];
my $param_2 = $ARGV[1];
my $param_3 = $ARGV[2];
my $param_5 = $ARGV[4];
my $param_4 = 100;
my $param_7 = 100;
my
$SUMMARY_TEXT="tmp_event_".$param_5.".sql"
;
open(SUMMARY,">$SUMMARY_TEXT");
$command1 = "insert into
nms(severity,source,event,value) values ('$param_1',
'$param_2', '$param_3', '$param_4');";
printf SUMMARY ("$command1");
close(SUMMARY);
$command2 = "psql.exe -h 150.140.21.70 -f
$SUMMARY_TEXT system nms";
$command3 = "del $SUMMARY_TEXT";
system($command2);
system($command3);
exit 0;
```

Figures 2: Listing of Events_value.ovpl

Events without a measured value

The configuration of the events that do not contain value should include the following automatic action command:
OVHIDESHELL cmd.exe /c "events.ovpl $s $ar $N 100 $3"

```
#!/opt/OV/bin/Perl/bin/perl
my $param_1 = $ARGV[0];
my $param_2 = $ARGV[1];
my $param_3 = $ARGV[2];
my $param_4 = $ARGV[3];
my $param_5 = $ARGV[4];
my
$SUMMARY_TEXT="tmp_event_".$param_4.".sql"
;
open(SUMMARY,">$SUMMARY_TEXT");
$command1 = "insert into
nms(severity,source,event,value) values ('$param_1',
'$param_2', '$param_3', '$param_5');";
printf SUMMARY ("$command1");
close(SUMMARY);
$command2 = "psql.exe -h 150.140.21.70 -f
$SUMMARY_TEXT system nms";
$command3 = "del $SUMMARY_TEXT";
system($command2);
system($command3);
exit 0;
```

Figures 3: Listing of Events.ovpl

The scripts use some NNM variables that are generated with the event, open a connection with the Postgres database using the "psql" client and insert data to the NMS table. By the time of an event, a trap occurs, an automatic action occurs, the appropriate script is executed reporting the status of the nodes (up or down) and certain data are imported to the NMS table of the database. The data could be event description, severity of the event (normal, major) and special values such as CPU load and temperature. When the event does not contain value like node status events, Events.ovpl is running. When the event does contain value like CPU load Events_value.ovpl is running. Additionally, PSQL client helps to connect to the database by using the IP of the LINUX server instead of the DNS to avoid a possible failure of DNS server.

Then, the events should be activated (this is an easy procedure). Once the configuration is created, everything is included in the trapd.conf file. The events and the specifications are:

Table 1: Events Configuration

Specifications	**Name of the event**
Polled by the SNMP manager every 60 sec	OV_NODE_UP OV_NODE_DOWN
Checked on all machines every 60 sec OVER the limit in case load > 70% REARM the limit in case load < 60%	OV_CPU_OVER OV_CPU_REARM
OVER if T° > 41°C REARM if T° <= 40°C	OV_TEMP_OVER OV_TEMP_REARM
Link is DOWN if connection of cable is removed	SNMP_LINK_UP SNMP_LINK_DOWN

The thresholds concern the control of CPU load (OVER in case load > 70%, REARM in case load < 60%) of all monitored nodes and also the control of CPU temperature (OVER if T° > 41°C, REARM if T° <= 40°C) at terminals side. When the threshold value is exceeded, NNM reports to the database through an automatic action the new value. Everything is included in the SnmpCol.conf file.

In order to keep NMS CPU level in low level, a task is added in Windows task manager that restarts the "SNMPcollect" procedure every 5 minutes, running cpuNMS.bat file. This task is necessary because the CPU is overloaded due to the short SNMP polling intervals. The suggested regular NNM polling interval from HP is 5 min.

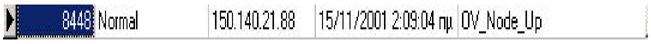
The external database could be Postgres, MySQL etc. The database management system selected was a Postgres, hosted on a LINUX server. The configuration of DBMS concerned only trusted "users", specifically only the IPs of NMS server represented a trusted machine, which could communicate with the appropriate port of the LINUX server and with the database. The above option was selected to ensure the security of database and its data.

NNM is able to send and store data to the central database. The interface uses the Postgres client to connect to the database and some SQL commands to insert data into the NMS table. All the modules of this interface are stored in a temporary folder of hard disk.

Table 2: NMS Table Architecture

ID	Event Identifier
Event	Short description
Source	The IP Address of the Node
Severity	Event Importance
Value	Variable value of measured events

A different script is executed according to the type of the event that occurs, in order to store CPU load, node status, temperature etc. into the database table. The script is embedded in the Events configuration module and exports several variable values of NNM. All actions are automated and occur in real time. The following table shows a sample record in the NMS table:



8448	Normal	150.140.21.88	15/11/2001 2:09:04 μμ	OV_Node_Up

Figures 4: Sample record

The first field is an auto-counter, next the severity importance, next the source indicating the IP address, next a field that contains the local date-time and last a field that contains a short description of the event. Postgres generates automatically the field date-time. In the field "Value" is stored data from measured events like temperature and CPU load.

Other variables that NNM monitors and measures are:

- Interface Utilization of TCP/IP port
- Interface errors
- Bandwidth
- Hardware fails of the monitored devices
- CPU load
- Temperature value of the terminal
- Node status
- Thresholds to critical values
- SNMP traffic, etc.

Table 3 presents some samples of database records during a certain time period of about three months. It is obvious that the desired network and system variables were collected according to the scenario. The Event with ID 11298 that was stored in the table referred to a very important, major situation affecting the CPU of the machine with IP 150.140.21.63. The CPU load of the machine at the specific date time was over the specific limit, about 80%. Then the Event with ID 11302 refers to a situation where the CPU operates normally according to pre-configured limits. The network administrator checks the table and finds out which are the 'weak' points of the network. Then he makes decisions about network and system improvements, in order to avoid some similar situations or combination that already occurred. The main and critical issue is to take decisions about the whole information system containing network expansion, bandwidth issues, devices with hardware and software specifications etc. concluding sometimes to a brand new system 'refresh', after taking into account the different parameters.

Table 3: Sample of experimental results

ID	EVENT	SOURCE	SEVERITY	VALUE	DATE AND TIME
8356	OV_Node_Up	150.140.21.39	Normal		27/08/2002 01:34:32
11298	OV_Cpu_Over	150.140.21.63	Major	80	17/10/2002 01:41:15
11299	OV_Node_Down	150.140.21.30	Normal		17/10/2002 01:41:43
11302	OV_Cpu_Rearm	150.140.21.63	Normal	20	17/10/2002 01:51:01
13825	SNMP_Link_Down	150.140.21.38	Major		20/11/2002 10:35:08

FUTURE WORK

The future work concerns the extension of the monitored SNMP variables in order to cover all the possible network events. The final expectation is to focus on the real network status, show its weak points and improve the network performance.

The prediction and decision analysis of the data collection could become another issue of our future work. We intend to select some standard models related to data analysis and do further work close to this problem. This work will enhance the decision analysis and prediction parts of our tool.

Another future goal is to apply this tool to several networks and assess the variables based on their criticality.

CONCLUSIONS

For network operators, network monitoring and analysis provides the means to become proactive (i.e. to detect faults prior to receiving a user's complaint). It also allows them to manage service level contracts, to be assured of day-to-day operations and to validate system changes.

The result of our work is a methodology and an implementation of a network monitoring analysis tool, which could improve the network performance. This could be achieved via the prediction and decision analysis of the data collection.

REFERENCES

Brown, M. C., 1999, "The Complete Reference Perl", Osborne/McGraw-Hill.

Carter, R. and M. Crovella. 1996. "Dynamic server selection using bandwidth probing in wide-area networks", *Technical report*, *Boston U.*, TR-96-007.

Carter, R. and M. Crovella. 1996. "Measuring bottleneck link speed in packet-switched networks", *Technical report*, *Boston U.*, TR-96-006.

Jacobson, V. 1997. "A tool to infer characteristics of Internet paths", *Technical report*, Lawrence Berkeley Lab.

Lee, C. A., J. Stepanek, R. Wolski, C. Kesselman and I. Foster, November 1999, "A Network Performance Tool for Grid Environments", *Technical Paper presented at Super Computing '99*, Portland, Oregon, USA.

Lowecamp, B. et al., August 1998, "A resource query interface for network-aware applications", *7th IEEE Symposium on High Performance Distributed Computing*, pages 189–196.

Mathis, M. and J. Mahdavi, 1996, "Diagnosing Internet congestion with a transport layer performance tool", *Proc. INET '96.*

Obraczka, K. and G. Gheorghiu. August 1998. "The performance of a service for network-aware applications", *2nd Sigmetrics conference on parallel and distributed tools*.

Paxson, V., J. Mahdavi, A. Adams, and M. Mathis. August 1998. "An architecture for large-scale internet measurement", *IEEE Communications*, 36(8): 48–54.

Su, A., F. Berman, R. Wolski and M. M. Strout, November 1998, "Using apples to schedule a distributed visualization on the computational grid".

Wolski, R., N. Spring, and H. Hayes. 1998. "The network weather service: A distributed resource performance forecasting service for meta-computing", *Future Generation Computing Systems*.

Self-Similar Traffic: burst size effects on packets delay behaviour

C.D'Apice*, R. Manzo*, N. Likhanov**
*Dipartimento di Ingegneria dell'Informazione
e Matematica Applicata Università di Salerno -Italy -
email: {dapice,manzo}@bridge.diima.unisa.it
**Institute for Problems of Information Trasmission,
University of Moscow, 19 Bol'shoi Karentny, GPS-4 Moscow.
e-mail likhl@online.ru

Abstract

In this paper we study asymptotic delay distribution at network node driven by Self-Similar traffic with Poisson arrivals and sessions (bursts) lengths distributed by Pareto. It was shown, recently, that for this kind of traffic delay distribution function can decrease very slowly. For this reason, to guaranty the Quality of Service (QoS) in communication networks, burst size is usually bounded by some value using, for example, leaky-bucket mechanism. The main contribution of this paper is the analysis of how this burst size bound will change delay distribution behavior. It is also found the critical value of the burst size at which delays start increase considerably.

1. Introduction

Packets delay distribution and packets loss probability are today the most important parameters of network performance. Typically packet delay or loss arise at network node. As a simple model, network node can be represented as a $G/D/1$ queueing system. The main problem is to find appropriate distribution for input process G for this queueing model. Basically, there are two approaches to this problem: the so-called many sources asymptotic and the large buffer asymptotic. The many sources asymptotic has been studied by many authors [1], [4], [15], [8] to name just a few. The many sources asymptotic is important when a large number of sources access a buffer which is drained by a server of very high capacity. The large buffer asymptotic is relevant when there are some sources which utilize a significant amount of server capacity. This has been studied by a number of authors under various hypotheses on the source characteristics, see [5] for the case of light-tailed sources. Empirical studies by Willinger *et al* [7] and [14] have shown the presence of sources with heavy-tailed characteristics. It means, that sources transmit for long periods of time when they come on and the activity periods have heavy-tailed characteristics in time.

The large buffer asymptotics for the case of heavy-tailed source activity periods has been studied under various assumptions on the input streams. These range from queues with fractional Brownian motion inputs [12], general Gaussian processes with negative drifts [3]; ON-OFF inputs with long-tailed ON periods [6] and $M/G/\infty$ type of inputs with long-tailed G distributions [13], [11], [10]. An excellent account can be found in the survey [1].

In this paper we will focus on the model proposed and studied in [9]. In this model input process for network node is considered as a sum of sessions (bursts) arrived by Poisson but with length distributed according to power law. There are many type of sources which can differ by exponent of length distribution, intensity of arrival, the rate in active period. Considering the sum of different types of sources makes this model sufficiently general and give us possibility to analyze how parameters of the sources can change performance characteristics of the server. In the paper [9] was found how the rate of active source period and the exponent of the length distribution can influence system performance. In this work we will find how the value of maximun length of active period (burst length) can change performance characteristics of the server.

The organization of the paper is as follows: in Section 2 we give notations, formulation of the model, and asymptotic definition. Section 3 presents a main result and its proof in brief form. In Section 4 we consider simple homogeneous case and we give interpretation of the developed results.

2. Model Formulation

We consider a discrete-time queue with M types of independent input processes $Y_{t,i}$. Each input processes $Y_{t,i}$ is the aggregation of sessions (bursts) arrived by Poisson with intensity λ_i. Sessions arrive independently of each other and a session of type i transmits at the rate $r_i > 0$ for a duration of time τ_i. For large values of x, the distribution of τ_i is given by

$$\Pr\{\tau_i > x\} \sim \begin{cases} \alpha_i x^{-1-\beta_i} \text{ if } x < B \\ \\ 0 \text{ if } x \geq B \end{cases}$$

where α_i, $\beta_i > 0$ are some constants and maximum burst size B will be defined later. If $\beta_i \in (0,1]$ and $B = \infty$ the input process $Y_{t,i}$ will be asymptotically Self-Similar.

It is assumed that the buffer is drained at a rate of C units per time. Let $\theta_{t,i}$ denote the number of sessions of type i which arrive into the system. We have

$$\Pr\{\theta_{t,i} = n\} = \frac{{\lambda_i}^n}{n!} e^{-\lambda_i}; \quad \lambda_i > 0 \tag{2.1}$$

The input process of the queue will be equal to

$$Y_t = \sum_{i=1}^{M} Y_{t,i}$$

and each component $Y_{t,i}$ can be expressed as

$$Y_{t,i} = \sum_{n=t}^{-\infty} \sum_{j=1}^{\theta_{n,i}} r_i I(\tau_{n,i,j} \geq t - n) \tag{2.2}$$

where $I(A)$ denotes the indicator function of the event A.

We denote the average input load of the system by

$$\rho = \mathbf{E}[Y_t] = \sum_{i=1}^{M} \lambda_i r_i \mathbf{E}[\tau_{t,i,1}]. \tag{2.3}$$

We will assume that $\rho < C$.

Denote by W_t the stationary buffer occupancy. W_t is given by the following formula

$$W_t = \max(W_{t-1} + Y_t - C, 0). \tag{2.4}$$

In this article we are interested in studying the behavior of the tail probability $\Pr\{W_t > z\}$ for large z i.e. as $z \to \infty$, with maximum burst size $B\tilde{}bz$, where $b > 0$ is some constant. More precisely we are interesting to find the probability for the system to reach overload period. We define overload period a period when packets delay is greater than z/C. Taking into account that for large z this probability is sufficiently small, we will define it as

$$F_{ov} = \Pr\{W_t > z, \quad W_i \leq z, \quad t - \delta z < i < t\}$$

where $\delta > 0$ is some sufficiently small constant.

To state the main result we need define the following random variables.

J denotes a set $(j_1, j_2, \cdots, j_M)$ of M integers;

$\kappa_J = \sum_{i=1}^{M} \beta_i j_i$ corresponds to decay exponent corresponding to the set J;

$R_J = \sum_{i=1}^{M} r_i j_i$ is the rate corresponding to the set J of sessions;

and

$$J_0 = \arg\min_J \{\kappa_J \ : R_J - (C - \rho) > \frac{1}{b}\}.$$

3. Overload Probability

Now we are ready to formulate main result. For large z overload probability of the system can be find as

$$F_{ov}\tilde{}z^{-\kappa_{J_0}+1} \prod_{i=1}^{M} P_i^{j_i^{(0)}} / j_i^{(0)} \tag{3.1}$$

where

$$P_i = \frac{\lambda_i \alpha_i}{\beta_i} \left((R_{J_0} - (C - \rho))^{-\beta_i} - b^{-\beta_i} \right).$$

In order to prove this result, first, for $b > \varepsilon > 0$ we define the processes Y_t^l and Y_t^h as follows. The process Y_t^l corresponds to the number of active sessions which have session lengths at most εz. This is given by:

$$Y_t^l = \sum_{n=t}^{t-\varepsilon z}\sum_{i=1}^{M}\sum_{j=1}^{\theta_{n,i}} r_i I(\varepsilon z \geq \tau_{n,i,j} \geq t-n). \tag{3.2}$$

The process Y_t^hcorresponds to the number of active sessions which have session lengths greater than εz. This is given by:

$$Y_t^h = \sum_{n=t}^{-\infty}\sum_{i=1}^{M}\sum_{j=1}^{\theta_{n,i}} r_i I(\tau_{n,i,j} \geq t-n, \tau_{n,i,j} > \varepsilon z) \tag{3.3}$$

We can see that the processes Y_t^l and Y_t^h are mutually independent and $Y_t = Y_t^l + Y_t^h$. Since we put $\varepsilon < b$ process Y_t^l will have the same properties as in unbounded case when $B = \infty$.

We start our analysis separately for the processes Y_t^l and Y_t^h and then combine them to get final result. For the process Y_t^l we can use directly two lemmas from the paper [9]. These Lemmas can be formulated in the following form.

Let

$$X_k^l = \sum_{n=t-k}^{t} Y_n^l. \tag{3.4}$$

Lemma 1. For any given $\delta_2 > 0$, $c_1 > 0$, $0 < \varepsilon < \min\left\{\frac{\delta_2}{\rho}, \frac{\beta_{\min}}{c_1}, b\right\}$ and sufficiently large z

$$\Pr\{\sup_{k\geq 0}\{X_k^1 - \rho k\} > \delta_2 z\} \leq z^{1-c_1\tilde{\delta}_2}$$

where $\tilde{\delta}_2 = \delta_2 - (\rho + \delta_1)\varepsilon$, $\rho = \mathbf{E}[Y_t]$.

Lemma 2. For any given $\delta_1 > 0$, $\delta_2 > 0$, and sufficiently small $b > \varepsilon > 0$, as $z \to \infty$

$$\Pr\{\inf_{k\geq 0}\{X_k^1 - (\rho - \delta_1)k\} < -\delta_2 z\} \leq \ e^{-O(z)}$$

First we need to find asymptotic for the probability $\Pr\{W_t > z\}$.

To prove upper bound, we consider process Y_t^l as an input to a queue with service rate ρ and process Y_t^h as an input to another queue with service rate $C - \rho$. Then both queues are stable and let W_t^l denote the stationary workload for the first queue and W_t^h denote the stationary workload for the second queue. If we define $X(-t,k) = \sum_{j=-t}^{k} Y_j$, and $X^l(-t,k) = \sum_{j=-t}^{k} Y_j^l$, we get

$$W_t = \sup_{k\geq 0}\{X(t-k,t) - Ck\},$$

$$W_t^l = \sup_{k\geq 0}\{X^l(t-k,t) - \rho k\},$$

$$W_t^h = \sup_{k\geq 0}\{X^h(t-k,t) - (C-\rho)t\}.$$

It is easy to see that

$$\Pr\{W_t > z\} \leq \Pr\{W_0^l > \delta z\} + \Pr\{W_t^h > (1-\delta)z\} \tag{3.5}$$

Now from Lemma 2 we get

$$\Pr\{W_t > z\} \leq \Pr\{W_t^h > (1-\delta)z\}(1 + o(z))$$

In this way, the upper bound is established by the system W_t^h.

Now, to get lower bound, with similar notation as above we have

$$\sup_{k\geq 0}\{X(t-k,t) - Ck\} \geq$$
$$\sup_{k\geq 0}\{X^h(t-k,t) - (C-\rho)k\} + \inf_{k\geq 0}\{X^l(t-k,t) - \rho k\}$$

and again, as $z \to \infty$ and using Lemma 1, we are get

$$\Pr\{W_t^h > z\}(1 + o(z)) \leq \Pr\{W_t > z\}$$

establishing that probability $\Pr\{W_t > z\}$ is determined by the system W_t^h.

To analyze W_t^h we will also use lower and upper bounds arguments. To compute lower bound we just need to find any configuration of sessions which arise system overflow: $W_t^h > z$. To construct upper bound we need to find what kind of busy periods will dominate. Let us start with lower bound. Define by $A_{J,t}$ the event that at time t there are j_i active sessions of type i. Using Poisson sessions arrivals, we have

$$\Pr\{A_{J,t}\} \sim const.\ z^{-\kappa_J},$$

or more exactly

$$\Pr\{A_{J,t}\} \sim z^{-\kappa_J}\prod_{i=1}^{M}\frac{(P_{J,i})^{l_i}}{j_i!},$$

where

$$P_{J,i} = \frac{\lambda_i \alpha_i}{\beta_i} \left((R_J - (C-\rho))^{-\beta_i} - b^{-\beta_i} \right).$$

Since we compute $\Pr\{A_{J,t}\}$, probability of the busy period with given configuration of J_0 active sessions which are simultaneously active during time interval at least equal to

$$l_0 = (R_{J_0} - (C-\rho))\, z$$

can be computed.

To get upper bound, first we find what it will be the typical busy period. Following the proof of the Lemma 3.2 from [9] we can find that typical busy period at which overflow occur will be isolated and will not contain any session except the sessions from configuration J_0 and

$$\Pr\{W_t^h > z,\ t \in ITBP\} = O(z^{-\kappa_{J_0}}),$$

$$\Pr\{W_t^h > z,\ t \notin ITBP\} = o(z^{-\kappa_{J_0}}).$$

Combing this fact with lower bound and taking into account above arguments concerning dominating of W_t^h system in overflow probability, we get equation (3.1).

4. Homogeneous Case

In the simple case, when we have only one class of sources, easy calculations can be made to understand how finite length of the sessions can change overflow probability. In homogeneous case we have $r_i = r$, $\beta_i = \beta$, $J_0 = \{j_0\}$, $R_{J_0} = rj_0$, $\kappa_{J_0} = \beta j_0$. Thus

$$j_0 = \lfloor \frac{(C-\rho) + \frac{1}{b}}{r} \rfloor + 1$$

and

$$F_{ov} = const.\ z^{-\beta j_0}.$$

As we can see, in sense of overflow probability, critical value of burst size bz is about $\frac{z}{C-\rho}$: critical delay value divided by $C-\rho$. It means, that increasing of the burst size more than the value of bz will not change significantly critical delay probability or overflow probability. In this case only the average rate in session r will take important role. Meanwhile, bounding the burst size to the value less than bz we can significantly decrease overflow probability or probability to exceed critical delay. Roughly speaking, the absolut value of the delay distribution exponent will increase in two times if we decrease the maximum burst size in two times.

5. Conclusion

In this paper we have investigated how the limit of the traffic burst size can change the overload probability on the node in the data network. We got our result for the so called Poisson/Pareto traffic model, where sessions (bursts) arrived by Poisson and session length distributed by power law, which implies that the probability of the long length sessions has significant value. Overload on the node can be interpreted as the queue length on the node buffer exceeds some level. In this situation packets loss will arise or the delay of the packets starts to be larger than the maximum value which guaranteed by the system to specify quality of service requirements. In the frame of our model we found how the different values for the limit of the burst size (limit on burst size means that distribution function of the burst length is cut at some point to zero with appropriate normalization) will change the situation with overload probability on the network node. Let us discuss briefly the derived results. In the paper we consider large buffer asymptotic. In this situation, as it was shown analytically, typical overload scenario will arise due to the certain number of the bursts with large length which are active in the same period of time. It means, that typical overload will arise not due to the fluctuation of the number of sources or bursts in the system but due to the length of the bursts. As a result, overload probability decrease according to the burst length distribution (by power law). While in the case of the overload due to the fluctuations of the number of sources, overload probability will decrease exponentially. Now, if put limit to the maximum burst length, we will have the following situation. If this limit to burst length is bigger than the maximum accepted delay value, overload probability asymptotically will be not changed since typical scenario will still the same. Otherwise, when the value of the limit is less than the maximum delay value, to reach overload in the system we need extra sources arrivals with larger bursts length. This will essentially change the overload probability. In the case of homogeneous sources it will decrease exponentially over the limit on burst size. In this way, maximum burst size starts to play important role for overload probability evaluation. Another parameter of the traffic, in the frame of considered model, which also plays critical role, is peak rate of the source. Peak rate of the sources means the rate inside the burst which is generated by this source. As it was shown previously and also in this paper, overload probability will decrease exponentially with the decreasing of the peak rate. In some sense, for the overload probability it is not so important peak rate or length of the burst. It is important the product of the burst length by the peak rate. Of course, this is true only in the case when we keep burst length less than maximum accepted delay value. In practical sense it means, that we can increase peak rate of the sources keeping the burst size (expressed in bits) constant and this will not increase overload probability on the node if the lengths of the bursts (in units of time) is less than our critical value. Let us now discuss how is practical the model and results considered in this paper. First of all, we should keep in mind that presented results are asymptotical. We consider asympotic while buffer size and burst length go to infinity (so called large buffer asympotic). It means, that for not limited values derived results can be considered only as some approximations which can be good or not depend on how is paticular system is close to limited one. For example, overload scenario in the real system will arise not only due to the long lengths sessions like in asympotics case, but also due to fluctuations of active sources number. Next, we should keep in mind, that it is also possible to construct different models for network traffic. For example, consider asymptotic when number of sources in the system goes to infinity. In this case situation with overload probability will be quite different. Overload probability will decrease exponentially over the buffer size and typical overload scenario will be due to the sources number fluctuation. Practically, overload will arise when average load will exceed service rate of the system. Validity of one or another model should be considered from statistical properties of particular network traffic.

References

[1] A. Botvich, A. and N. G. Duffield.; *Large deviations, economies of scale and the shape of the loss curve in large multiplexers*, Queueing Systems, 20, 1995, pp. 293-320

[1] O. J. Boxma and V. Dumas; *Fluid queues with long-tailed activity period distributions*, Computer Communications, 1998, 21, pp. 1509–1529.

[3] J. Choe and N.B. Shroff; *On the supremum distribution of integrated stationary Gaussian processes with negative linear drift*, Advances in Applied Probability, March 1999, 31, pp. 135–157.

[4] C. Courcoubetis and R. Weber; *Buffer overflow asymptotics for a switch handling many traffic sources*, J. Appl. Prob., Vol. 33, No. 3, 1996, pp. 886-903

[5] N. G. Duffield and N. O'Connell; *Large deviations and overflow probabilities for the general single-server queue with applications*, Math. Proc. Camb. Phil. Soc., 118(1), 1995, pp. 363-374.

[6] P. R. Jelenkovic and A. A. Lazar; *Asymptotic results for multiplexing subexponential on-off sources*, Advances in Applied Probability, 1999, 31, pp. 394–421.

[7] W. E. Leland, M.S. Taqqu, W. Willinger and D. V. Wilson; *On the self-similar nature of Ethernet traffic (extended version)*, IEEE/ACM Trans. on Networking, Vol. 2, No. 1, 1994, pp. 1-15.

[8] N. Likhanov and R. Mazumdar; *Cell loss asymptotics in buffers fed with a large number of independent stationary sources*, Journal of Applied Probability, March 1999, 36, pp. 86–96.

[9] N. Likhanov and R. Mazumdar; Loss asymptotics in large buffers fed by heterogeneous long-tailed sources, Advances in Applied Probability, 32, 2000, pp. 1168-1189.

[10] N. Likhanov; *Bounds on the buffer occupancy probability with self-similar input traffic*, in Self-similar network traffic and performance evaluation, K.Park and W.Willinger eds., Wiley, 2000, pp. 193-214.

[11] Z. Liu, P. Nain, D. Towsley and Z-L. Zhang; *Asymptotic behavior of a multiplexer fed by long-range dependent process*, Journal of Applied Probability, March 1999, 36, pp. 105–118.

[12] I. Norros; *A storage model with self-similar input*, Queueing Systems, **16**, 1994, pp. 387-396.

[13] M. Parulekar and A. M. Makowski; *Tail probabilities for $M/G/\infty$ input processes*, Queueing Systems, 27, pp. 271–296.

[14] V. Paxon and S. Floyd; *Wide area traffic: the failure of Poisson modeling*, IEEE/ACM Trans. on Networking, **3**, 1993, pp. 226-244.

[15] A. Simonian and J. Guibert; *Large deviations approximation for fluid queues fed by a large number of ON/OFF sources*, IEEE J. Sel. Areas Commun., Vol. 13, N0. 6, 1995, pp. 1017- 1027

GRAPHICAL RTP SESSION MONITOR USING JAVA MEDIA FRAMEWORK (JMF)

Fernando Boronat Seguí, Salvador Llopis Torres, J. Carlos Guerri Cebollada, Manuel Esteve Domingo
Área de Ingeniería Telemática, Departamento de Comunicaciones
Universidad Politécnica de Valencia - Escuela Politécnica Superior de Gandia
Ctra. Nazaret-Oliva S/N, 46730 Grao de Gandía (VALENCIA)
Telf: +34 962 849 341, Fax: 962 849 313
E-mail: {fboronat,jcguerri,mesteve}@dcom.upv.es, salloto@epsg.upv.es

KEYWORDS
Multimedia, RTP session monitor, JMF.

ABSTRACT

A number of multicast conferencing applications have already been developed for the MBone using the Real-Time Transport Protocol (RTP). In this paper, we describe MonitorRTP, an application for displaying control information associated with an RTP session. It displays in near real-time a table-based overview of the control feedback generated by all the participants in the RTP session and gives chance of analyzing it graphically live or later.

1. INTRODUCTION

The Real-Time Transport Protocol (RTP) (Schulzrinne et al. 1996) is an application level protocol that is intended for delivery of delay-sensitive content, such as audio and video, through different networks. In this paper, we present a java-based application to Monitor RTP sessions that displays some interesting control information associated with them. It displays in near real-time a graphical or table-based overview of the control feedback generated by active and passive users in the RTP session and gives the possibilities to analyze it graphically live or later.

To develop this application, we have used the Java^{TM} Media Framework API (JMF, 2000). This optional package, which can capture, playback, stream and transcode multiple media formats, extends the multimedia capabilities on the J2SE^{TM} platform, and gives multimedia developers a powerful toolkit to develop scalable, cross-platform technology.
The rest of the paper is organised as follows. Section 2 presents a survey of the monitor applications we have studied. Section 3 presents the importance of RTP/RTCP in multimedia communications. Next, JMF API utilities for RTP are presented in section 4. The application is presented in section 5 and, finally, in section 6, we show several graphical presentations of statistic data collected by our tool.

2. RELATED WORK

We have studied several RTP Session monitors to develop our tool:

- *RtpMonitor* (Afonso, 1999) is a Java application that collects statistics about RTP Sessions. It is implemented using Sun's JMF version 2.0 Beta, and it has the following capabilities: global session statistics, individual stream statistics and feedbacks from all participants. Statistic data are recorded in text files continuously. It also permits the user to play the received streams (audio and video) and, if desired, he can participate in the session by sending RTCP packets.
- The *SDR Monitor* (Sarac et al., 1999) tracks, manages, and presents information about the availability of world-wide SDR sessions. SDR is a session directory tool designed to allow the advertisement and joining of multicast conferences on the Mbone.
- *Rtpmon* (Bacher et, al., 1996) can be used to monitor the control information exchanged between applications that implement RTP. Feedback from receivers, including the loss rate and jitter, are displayed in a table that can be sorted in various ways to help isolate and diagnose multicast distribution problems. It has no provision for displaying data from multiple sessions.

- *Mtrace* (Fenner et al., 2000), *MultiMON* (Robinson et al.), and *Mhealth* (Makofske, et al., 1999) are multicast monitors.

We can also find monitors included in RTP tools such as vic, rat, etc, including all the statistic data obtained from RTCP packets. We have studied all these monitors to add their main characteristics to our tool.

3. RTP/RTCP PROTOCOL (Schulzrinne et al. 1996),

RTP, proposed by IETF in RFC 1889, has been accepted as a standard for real time multimedia data transmission. It supports the transmission of time-dependent media, such as audio and video, over wide-area networks (WANs), by adding synchronization and quality-of-service (QoS) feedback capabilities to the existing transport protocol. RTP has been widely used in the Multicast Backbone (MBone), a virtual network that has become a shared worldwide medium for Internet multicast communications. In both modes (multicast or unicast), data transmission is monitored by a complementary control protocol called Real Time Control Protocol (RTCP), which allows monitoring the data delivery in a scalable manner to large multicast networks, and provides minimal control and identification functionality. RTP is very useful because it is independent of the underlying protocol and can work on any type of network like TCP/IP, ATM, Frame-Relay etc.

The Transport Control Protocol (TCP) cannot support the real time services like interactive video, conferencing etc., The reason behind this is the fact that TCP is rather a slow protocol, requiring three way hand-shaking. Hence UDP is used over IP as a better option than TCP over IP. But UDP inherently is an unreliable protocol, which does not support retransmissions, upon packet loss. Still, UDP has some features like multiplexing and checksum services, which favors the real time services. To overcome the drawbacks of UDP, RTP is proposed at the application layer level. Among the services offered by RTP we can find payload type identification, sequence numbering, timestamps and delivery monitoring information. RTP can sequence those packets, which arrive out of order at the receiver. Sequence numbers can also be used to identify lost packets. Timestamping is used for calculating the proper playout point of each media stream.

Despite all we mentioned above, RTP does not, by itself, provide any mechanism to ensure timely delivery or provide other quality-of-service guarantees. It relies on lower-layer services to do so. RTP does not also guarantee delivery or prevent out-of-order delivery, nor does it assume that the underlying network is reliable and delivers packets in sequence. RTP is primarily designed to satisfy the needs of multiparticipant multimedia conferences. It is also applicable for services like storage of continuous data, interactive distributed simulation, active badge, and control and measurement applications.

On the other hand, RTCP monitors the quality of service and also conveys information about the participants in an on-going session. The received data is continuously monitored by RTCP, which informs the RTP layer that can adjust its coding and transmission parameters for the proper delivery of data, accordingly with the network conditions. For example if the RTCP layer detects severe packet loss, it may inform the RTP layer to slow down the rate of transmission.

If one user is transmitting multiple media during a session, such as audio and video, separate RTP sessions are opened for each one of them. Hence there is no multiplexing of media at RTP level. It is up to the lower layers to multiplex the packets from various media and send them in a single channel. RTCP maintains one identifier called CNAME, which is the same for all the media initiated by the same user. Hence, CNAME is the only identifier that can identify the media originated from a user.

3.1.- RTCP

RTP receivers provide reception quality feedback using RTCP report packets. There are two kind of report packets: sender or receiver type. Participants only send SR (Sender Reports) packets if they are participating actively in the session, otherwise they will send RR (Receiver Reports) packets. In addition to theses two packet types, RTCP defines other types of control packet: SDES (Source Description), BYE and APP (Application Defined). Next we are going to explain the most useful RTCP packets for our tool.

Our tool extracts statistic data from the reports SR and RR using the JMF API, reading the following fields from the SR and RR packets.

- *Sender's packet count* (in SR)*:* total number of RTP data packets transmitted by the source.
- *Sender's octet count* (in SR): total number of payload octets sent from the beginning of the session.
- *Fraction lost* (in RR and SR): fraction of RTP packets lost since the previous SR or RR packet was sent.
- *Cumulative number of packets lost* (in RR and SR): total number of RTP data packets lost.
- *Extended highest sequence number received* (32 bits, in RR and SR): Low 16 bits contain the highest sequence number received in an RTP data packet and the most significant 16 bits extend that sequence number with the corresponding number of sequence number cycles.
- *Inter-arrival jitter* (in RR and SR): estimation of the statistical variance of the RTP data packet arrival time measured in timestamp units.

On the other hand, the RTCP SDES packet contains useful information to identify each user in the session: fields CNAME (Canonical name), NAME (User Name), EMAIL, PHONE, LOC, TOOL, NOTE and PRIV. All they are explained in RFC 1889.

The RTCP BYE packet is sent by a user when leaves the RTP session.

4. JAVA^{TM} MEDIA FRAMEWORK (JMF)

Our tool has been developed using JMF 2.1.1a. This is an API for incorporating time-based media into Java applications and applets. It supports several features: capturing media data, enable the development of media streaming and conferencing applications in Java, enable advanced developers and technology providers to implement custom solutions based on the existing API and easily integrate new features with the existing framework, provide access to raw media data, enable the development of custom, downloadable demultiplexers, codecs, effects processors, multiplexers, and renderers (JMF *plug-ins*), and maintain compatibility with JMF 1.0.

This API include several APIs among which we have mainly used the JMF RTP API. The high-level JMF RTP architecture is shown in figure 1.

4.1 JMF RTP APIs

The RTP APIs in JMF 2.0 supports the reception and transmission of RTP streams and addresses the needs of application developers who want to use RTP to implement media streaming and conferencing applications.

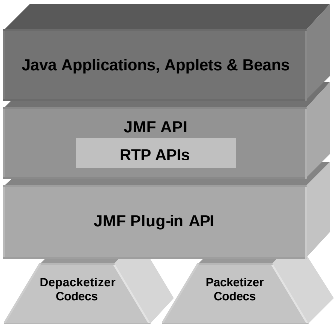


Figure 1. High-level JMF RTP architecture

In JMF v.2.1.1a, RTP packages have been reorganized and some classes, interfaces, and methods have been renamed to make the API easier to use. RTP event classes that were in *javax.media.rtp.session* are now in *javax.media.rtp.event*. RTCP-related classes that were in *javax.media.rtp.session* are now in *javax.media.rtp.rtcp*. The rest of the classes in *javax.media.rtp.session* are now in *javax.media.rtp* and the *javax.media.rtp.session* package has been removed.

The above APIs are designed to enable the development of media streaming and conferencing applications in Java, support media data reception and transmission using RTP and RTCP, support custom packetizer and depacketizer plug-ins through the JMF 2.1.1a plug-in architecture and be easy to program.

4.1.1. RTPManager

In JMF 2.1.1a the old *SessionManager* has been replaced by the *RTPManager* to coordinate an RTP session. This manager maintains track of the session participants and the streams being transmitted. It keeps the state of the session as viewed from the local user and, in fact, it is a local representation of a distributed entity that is the RTP session. It also handles the RTCP packets, and supports RTCP for both senders and receivers. *RTPManager* interface defines methods that enable an application to initialise and start participating in a session, remove individual streams created by the application and close the entire session.

4.1.2. Session Statistics

The *Manager* keeps statistics on all of the RTP and RTCP packets sent and received in a session. It tracks statistics for the entire session on a per-stream basis. It also provides access to global reception and transmission statistics through the following interfaces: *GlobalReceptionStats*, which maintains global reception statistics for the session, and *GlobalTransmissionStats*, which maintains cumulative transmission statistics for all local senders. Developers can also obtain statistics for a particular user or outgoing stream from *ReceptionStats*, which maintains source reception statistics for an individual participant, and *TransmissionStats*, which maintains transmission statistics for an individual send stream.

4.1.3. Session Participants

The manager keeps track of all of the participants in a session. It creates a Participant whenever an SDES RTCP packet arrives that contains a source description with a canonical name (CNAME) that has not been seen before in the session (or has timed-out since its last use). Each participant is represented by an instance of a class implementing the Participant interface. There are two types of participants: passive (sending control packets only) or active (also sending one or more RTP data streams). A participant can receive more than one stream, each of which is identified by the synchronization source identifier (SSRC) used by the source of the stream.

4.1.4. Session Streams

For each stream of RTP data packets in the session, the Manager keeps an *RTPStream* object. There are two types of RTP streams: *ReceiveStream* which represents a stream that is being received from a remote participant, and *SendStream* which represents a stream of data coming from the *Processor* or input *DataSource* that is being sent over the network. A *ReceiveStream* is constructed automatically whenever the session manager detects a new source of RTP data. To create a new *SendStream*, developers can call the RTPManager *createSendStream* method.

5. MonitorRTP APPLICATION

MonitorRTP is mainly based on RtpMonitor tool [2]. It includes all the facilities of this tool but it adds some new features, obtained from other tools studied in section 2, for example, later graphical analysis. In contrast to RtpMonitor, we have used JMF 2.1.1a *RTPManager* Class while it used the interface *SessionManager* (JMF 2.0 beta).

For the graphical user interface (GUI) of our application, we have used java´s Swing API. This is an API for the Java platform for user interface design. RtpMonitor uses AWT that is the original user interface for the first Java release, based on the native windows look-and-feel. Java emulates the look-and-feel of the native platform (or platforms) without making use of the native systems call itself. All widgets in Swing are available on all platforms in which Java runs whether the native system windows toolkit has this feature or not. Swing’s lightweight architecture allows dynamic look-and-feels to be *hot-swapped* simultaneously as a Java application is running.

Despite of the above, the main contribution with our tool is that we have added new classes for drawing graphics in near real-time of the statistic data recorded in a session.

5.1. RTPManager

As mentioned before, *RTPManager* is the main java class used by our tool. Using its methods and other related methods and classes, MonitorRTP obtains statistic data and other information about of a RTP session. Figure 2 displays this class and some of its methods. Using the *RecordTask* class it records statistic data in text files that will be processed later to draw graphics using java classes for graphical drawing.

RTPManager()

- · initialize(SessionAddress address)
- · initialize(SessionAddress[] localAddresses, SourceDescription[] sourceDescription, double rtcpBandwidthFraction, double rtcpSenderBandwidthFraction, EncryptionInfo encryptionInfo)
- · getActiveParticipants()
- · getAllParticipants()
- · getGlobalReceptionStats()
- · getGlobalTransmissionStats()
- · getPassiveParticipants()
- · getLocalParticipant()
- · getRemoteParticipants()

Figure 2. RTPManager class and methods

Figures 3 and 4 show the main window of MonitorRTP and the configuration window, respectively. As it can be seen, it contains several fields that show statistic data of the session in near real-time and other information such as the number of participants, the streams being transmitted, the active and passive participants, etc.

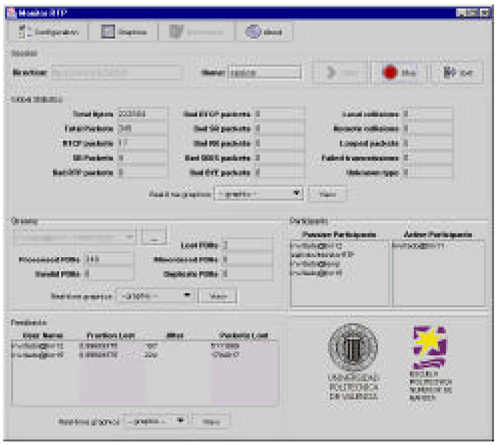
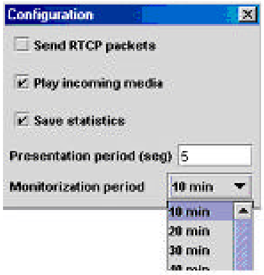


Figure 3. Main Window


Configuration

☐ Send RTCP packets

☑ Play incoming media

☑ Save statistics

Presentation period (seg) 5

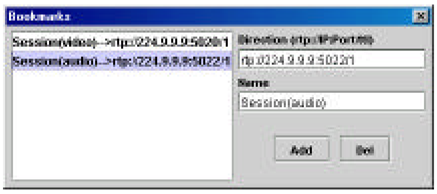
Monitorization period 10 min

Figure 4. Configuration window

In MonitorRTP the user can decide if he wants the monitor application to be something more than only a monitor. It can become another passive receiver in the session (application will have to send RTCP RR packets) to the sender. It can also display the streams received in the session that it is monitoring.

If we decide to monitor the session, we can decide if we only want to watch the statistic data in the GUI or to store it for future analytical or graphical analysis, from 10 minutes to 24 hours.

Figure 5 shows the address list implemented in our tool to store the most common sessions to monitor. Using this window we can maintain a list of session addresses by adding or removing entries and giving them a logical name easy to remember. So we don't have to remember nor write all the entire address (*rtp:// + IP Address + : + port + TTL*) each time we want to start a monitoring session and we only have to select an item of the list in this window.


Bookmarks

Session(video)-->rtp://224.9.9.9:5020/1
Session(audio)-->rtp://224.9.9.9:5022/1

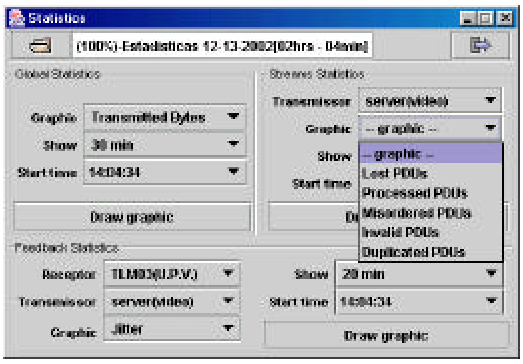
Direction (rtp://IP:Port/ttl) rtp://224.9.9.9:5022/1

Name Session(audio)

Figure 5. Address list

We considered the statistic file structure of RtpMonitor no appropriate for our expected graphical analysis. Our tool store all the statistic data in one file per session, despite of its duration. Moreover, we store data with reference to the canonical names (CNAME) of the session participants, not to the SSRC (as rtpMonitor does) that is difficult to remember after the session when we do the graphical analysis. With all these changes it is easier for our graphical analysis block to obtain the statistic data to show.

Figure 6 shows the window for selecting statistic data to be graphically presented. It is possible to obtain graphical representation for all the parameters showed in the main window for all the senders and receivers, in intervals from ten minutes to 24 hours. In the next section we present several of these graphical presentations.


Statistics

(100%)-Estadisticas 12-13-2002[02hrs - 04min]

Global Statistics

Graphic Transmitted Bytes
Show 30 min
Start time 14:04:34

Streams Statistics

Transmissor server(video)
Graphic -- graphic --
Show
Start time

Feedback Statistics

Receptor TLM03(U.P.V.)
Transmissor server(video)
Graphic Jitter
Show 20 min
Start time 14:04:34

Figure 6. Graphical Statistics window

6. MONITORING RESULTS

6.1. Measurement scenario

In order to evaluate the correct performance of the MonitorRTP, we have implemented a measurement scenario to transmit real time video using Mbone tools, such as *vic* (McCanne and Jacobson 1995) and rat, and have collected statistics that we present in the following figures. In the scenario (figure 7), all the workstations are connected by an Ethernet LAN . The sender transmits a video clip (30 minutes) from a videotape to three receivers. Specifications of audio and video flows are shown in table 1.

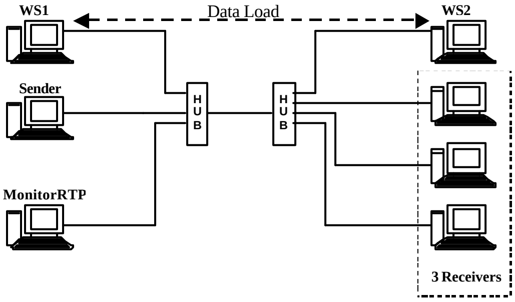


Figure 7. Experimental system.

Table 1. Specifications of audio and video.

Configuration	Video	Audio
Average bit rate	900 bps (24 fps) aprox.	8 kbps
Encoder	H261	GSM

Workstations WS1 and WS2 are utilised to introduce a traffic flow of interference in the experiment. WS1 sends fixed size data messages of 1472 bytes each to WS2 under the UDP protocol at exponentially distributed intervals. We adjust the amount of the interference traffic by changing the average of the interval. We have repeated the experiment 10 times and in each repetition of the experiment, they have introduced a constant load in the network from 1 Mb/s to 10 Mb/s. Global statistics collected by our application for the audio and video sessions are shown in table 2.

Figure 8 displays the Lost Packets during a session obtained from one of the video receivers when the data load introduced by WS1 and WS2 was 9 Mb/s. Figure 9 shows the inter-arrival jitter reported from one of the audio receivers during the same experiment.

Table 2. Global Statistics

Item	Video	Audio
Monitoring Period (min)	30	30
Participants	5	5
Receptors	4	4
Transmissors	1	1
Total Transmitted Bytes	308822704	3687116
Total Transmitted RTP Packets	327239	47504
Total Transmitted RTCP Packets	1439	2412

Figure 10 shows the results of the ten repetitions, for the mean Lost Packets versus data load introduced by WS1 and WS2, from 1 Mb/s to 10 Mb/s. As we expected, for the video stream, RTP loss rate is 0 while data load doesn't exceed 9 Mb/s because there is bandwidth enough for the successful delivery of video frames. When it exceeds 9 Mb/s the number of lost packets increases drastically and we noticed jerkiness in the reproduction and bad reception. For the audio stream, the number of lost packets obtained in all cases is very low because of the bit rate needed of 8 Kb/s.

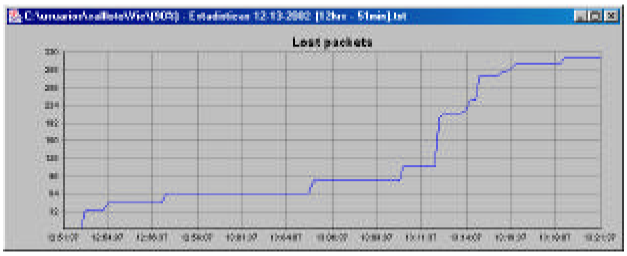


Figure 8. Session Lost Packets

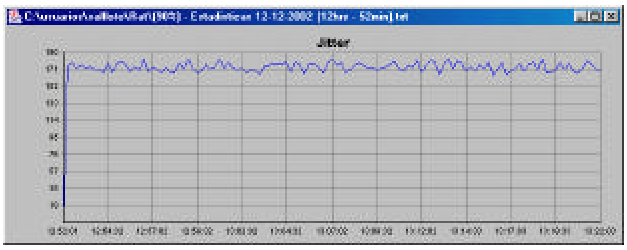


Figure 9. Inter-Arrival Jitter

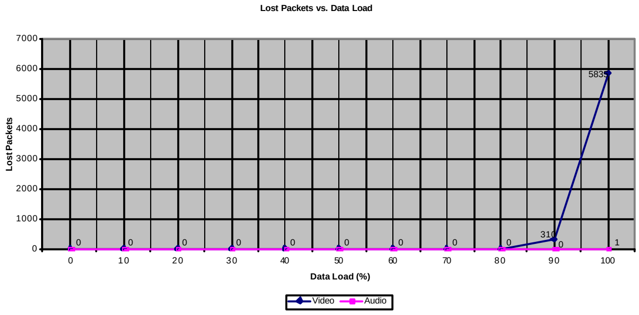


Figure 10. Lost Packets versus Load

BIBLIOGRAPHY

Afonso F.C., "Virtual Reality Transfer Protocol (VRTP): Implementing a monitor application for the Real-Time Transport Protocol (RTP) using the JMF". Naval Postgraduate School of Sao Paulo, 1999.

Bacher, D., Swan, A. and Rowe, L. A. "rtpMon: A Third-Party RTCP Monitor". Computer Science Division – EECS University of California, Berkeley, 1996.

Fenner, W. and Casner, S, "A traceroute facility for IP Multicast", draft-ietf-idmr-traceroute-ipm-07.txt, 2000.

Java^{TM} Media Framework API. 2000. [http://java.sun.com/products/java-media/jmf]

Makofske, D. and Almeroth, K., "MHealth: A Real-Time Multicast Tree Visualization and Monitoring Tool", *NOSSDAV '99*, Basking Ridge New Jersey, June 1999.

McCanne, S. and Jacobson, V. "vic: A Flexible Framework for Packet Video". *ACM Multimedia*, November 1995, San Francisco, CA, pp. 511-522

Robinson, J.L. and Stewart, J.A., "Multimon - an IP multicast monitor", Communications Research Centre, Ottawa. [http://www.merci.crc.ca/mbone/MultiMON/]

Sarac, K. and Almeroth, K., "Sdr global session monitoring effort", 1999

Schulzrinne, H.; Casner, S.; Frederick, R. and Jacobson, V. "RTP: A Transport Protocol for Real-Time Applications". RFC 1889. January 1996.

COMMUNICATION TECHNOLOGIES

Beamforming and Band Allocation for Multiband CDMA Systems

D. Alnsour, M. Al-Akaidi
School of Engineering and Technology
De Montfort University, Leicester
LE1 9BH,UK
e-mail: mma@dmu.ac.uk

Abstract

Multiband (or multi-carrier) CDMA is a promising approach to increasing the capacity of CDMA systems, while maintaining compatibility with existing systems. A joint band allocation, beamforming and load sharing algorithm is proposed in a Multiband CDMA system. The Distributed constrained power control (DCPC) algorithm, which is a quality based power control that maximizes the power of weak users in the system, is investigated in this paper. The proposed algorithm will solve the problem of degradation in performance at high loads.

Keywords— **Multiband CDMA, power control, beamforming, band allocation, DCPC**

I. Introduction

Due to the demand for increased capacity of code division multiple access (CDMA) systems Multiband CDMA is introduced. The system is a combination of two well known techniques for modulation and Multiple Access (MA), Frequency Division Multiple Access and CDMA. In FDMA the available spectrum is divided into distinct bands, and each connection is allocated to a band. CDMA provides a frequency reuse of 1, and the guard bands are much smaller than pure FDMA.

Power control is also a crucial component in cellular systems. It plays an important role in determining the interference and capacity of the uplink of a code division multiple access (CDMA) system. It is evident that equitable sharing of resources among users can be achieved only through the introduction of power control.

In a multi-cell system, the fact that users near the edge of the cell must transmit at higher power levels leads to considerable inter-cell interference, and it is in general difficult to formulate the optimal rule for the power level. Thus, distributed algorithms that adjust power levels based on quality of service are essential.

Power control aims to give all users the same signal to interference ratio (SIR). Since all users within a cell experience the same interference in the single band case, this is equivalent to having equal received signal powers. However, for users in different cells or different bands, the two are not interchangeable. Separating near and far users is only beneficial if the different groups are allowed to be received at different power levels. This occurs automatically if power control equalizes SIR directly, since the interference will be lower in the bands containing weaker users [1]. In Multiband CDMA Systems the near-far effect is dealt with by assigning strong users to a specific band and weaker ones to another.
In this paper, the performance of the system using different power control based algorithms is studied. The combination of Resource Allocation Algorithms (RAA) and power control was found to significantly increase the capacity of the system.

In Section III, two power control algorithms are described in details. Section IV discusses power control based load sharing, band allocation for Multiband CDMA and joint base station assignment and beamforming algorithms. These algorithms proved to have the best performance in relative to simple power control techniques as shown in Section V. In section VI a new algorithm is proposed and the anticipated results are described.

II. Multiband CDMA

The available wideband spectrum is divided into smaller number of spectra with smaller bandwidths. Each of these smaller subchannels becomes a narrowband CDMA system having a processing gain lower than the original CDMA system. This hybrid scheme has an advantage that different users can be allocated different subspectrum bandwiths depending upon their requirements.

Multiband CDMA provides the opportunity to address the near-far effect problem more directly by separating strong ("near") and weak ("far") users into separate bands. Thus weak users are not competing with strong users, and a lower received power is acceptable. This translates to a lower transmit power from the mobile. Another important advantage to Multiband CDMA is its backward compatibility for the IS-95A standard [1].

III. Power Control Algorithms

The distributed constrained power control algorithm (DCPC)[2] has become one of the most widely accepted power control algorithms. It frequently appears incorporated in resource management [3], [4], [5] and [6]. DCPC has a property that the power reaches the maximum level when a user is experiencing degradation of channel quality. Unfortunately, using maximum transmitter power may not necessarily lead to sufficient improvement of channel quality and will thereto generate severe interference, hitting other users.

A. Distributed Constrained Power Control

DCPC was developed with the goal of equalizing (CIR) among all connections, thereby providing maximum capacity for a given link quality requirement. Here the CIR (γ_i) of the i^{th} user as measured at the base station it connects to (before despreading in the case of CDMA system) is expressed as [2]

$$\gamma_i = \frac{C_i}{I_i} = \frac{PiG_{ii}}{\sum_{j=1,j\neq i}^{M} P_j G_{ij} + N}, 1 \leq i \leq M \qquad (1)$$

where,

$$G_{ij} = \frac{A_{ij}}{d_{ij}^{\beta}}$$

and C_i/I_i is the carrier to interference ratio, P_i is the transmission power of the i^{th} user, G_{ij} is the path gain from user i to base station j, A_{ij} is the attenuation factor, d_{ij} is the distance between user i to base station j, β is a correction factor, N is the receiver thermal noise, and M is the number of users operating on the same band.

The situation of the interference expressed in Eq. 1 is shown in Figure 1. The algorithm assumes that the allocated bands are sufficiently spaced such that there is no interference between bands. 'CIR balancing' is achieved by adjusting mobile transmitter powers through use of feedback loop between the base station and mobile. As in [2], at the n^{th} iteration, the transmission power of the i^{th} mobile is

$$P_i^n = \min\left\{P_{max}, \gamma_t \frac{P_i^{n-1}}{\gamma_i^{n-1}}\right\} \qquad (2)$$

where P_{max} is the maximum transmission power in the system, and γ_t is the target CIR. Results in [2] show that the algorithm converges after five iterations for a system with ten users operating over two channels in ten different cells.

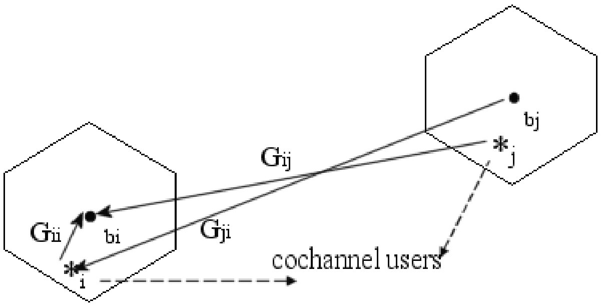


Fig. 1. Configuration and links of a cellular network

The performance of DCPC is shown to be nearly optimal; it supports the maximum number of users with CIR greater than or equal to the target. However, the algorithm does not provide an alternative for users required to transmit at maximum power. Band allocation as proposed by [3] provides such an alternative. Upon reaching P_{max}, connections are switched to another band, the transmission power is lowered to that at call initialization, and the power continues to be adjusted by DCPC [3].

IV. Antenna Array and Beamforming

An antenna array consists of a set of antennas, designed to receive signals radiating from some specific directions and attenuate signals radiating from some specific directions and attenuate signals radiating from other directions of no interest. The output of array elements are weighted and added by a beamformer as shown in Figure 2 to produce a directed main beam and adjustable nulls. In order to reject the interferences, the beamformer has to place its null in the directions of sources of interference, and steer to the direction of the target signal by maintaining constant gain at this direction. A sample antenna array pattern, which is depicted in Figure 3 shows this effect [6].

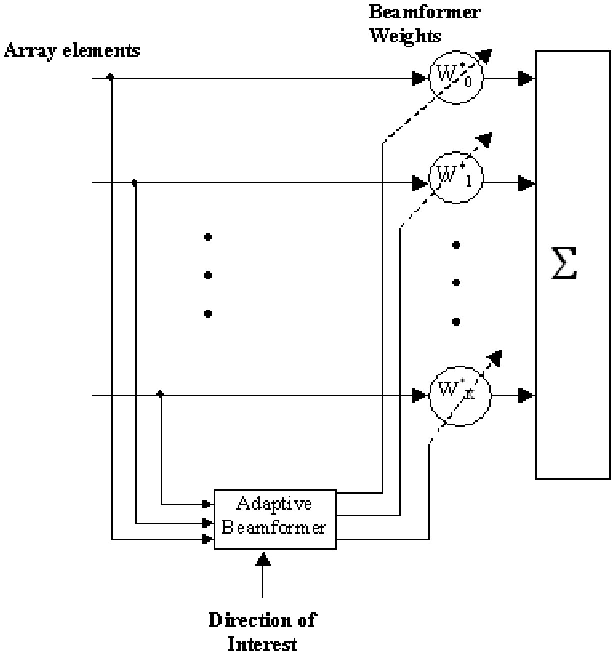


Fig. 2. Antenna Array and Beamformer

Assume that an antenna array each consisting of K elements. Consider a cochannel set consisting of M transmitter and receiver pairs and denote the signal from the j^{th} source by $s_j(t)$ and its power by P_j. The received signal at the i^{th} receiver can be written as follows [6]

$$x_i(t) = \sum_{j=1}^{M} \sqrt{P_j} s_j(t) \mathrm{a}_{ji} + \mathrm{n}_i(t) \qquad (3)$$

where vector a_{ji} is called the spatial signature or response of the i_{th} array to the j^{th} source, and $\mathbf{n}_i(\mathrm{t})$ is the thermal noise vector at the input of this array.

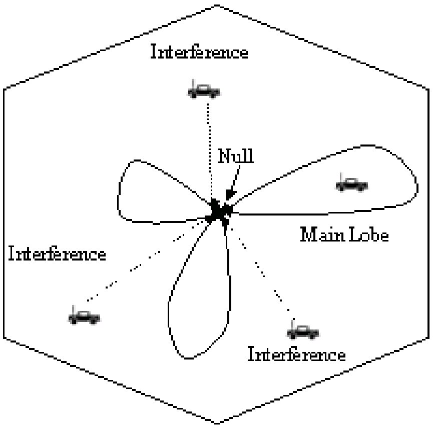


Fig. 3. Sample Antenna Area Pattern

Beamforming works in such a way that CIR is maximized for a specific link, which is equivalent to minimizing the interference at the receiver of that link. In order to minimize the interference, the variance or average power at the output of the beamformer is minimized subject to maintaining unity gain at the direction of the desired signal. The output of the beamformer at the i^{th} receiver can be written as [6]

$$\mathbf{e}_i = \mathbf{w}_i^H \mathbf{x}_i \tag{4}$$

where $\mathbf{w}_i$ and $\mathbf{x}_i$ are the beamforming weight vector and the received signal vector at the i_{th} receiver, respectively and the superscript H denotes conjugate transpose. The average output power is given by

$$\begin{aligned} \varepsilon_i &= \mathrm{E}\{\mathbf{w}_i^H \mathbf{x}_i \mathbf{x}_i^H \mathbf{w}_i\} \\ &= \sum_{j \neq i} P_j G_{ji} \mathbf{w}_i^H \mathbf{a}_{ji} \mathbf{a}_{ji}^H \mathbf{w}_i + N_i \mathbf{w}_i^H \mathbf{w}_i \end{aligned} \tag{5}$$

where N_i is the noise power at the input of each array element. In order to minimize the interference we minimize ε_i subject to maintaining unity gain at the direction of the desired signal it can be shown that the optimal weight vector is given by [7]

$$\hat{\mathbf{w}}_i = \frac{\Phi_{in}^{-1} \mathbf{a}_{ii}}{\mathbf{a}_{ii}^H \Phi_{in}^{-1} \mathbf{a}_{ii}} \tag{6}$$

where,

$$\Phi_{in} = \sum_{j \neq i} P_j G_{ji} \mathbf{a}_{ji}^H \mathbf{a}_{ji} + N_i \mathbf{I}, \tag{7}$$

where $\mathbf{I}$ is an identity matrix. The maximum CIR at the i_{th} receiver is given by

$$\gamma_i = \frac{C_i}{I_i} = \frac{PiG_{ii}}{\sum_{j \neq i} P_j G_{ij} G_{ai} + N \hat{\mathbf{w}}_i^H \hat{\mathbf{w}}} \tag{8}$$

where $G_{ai}(\mathbf{w}_i, \mathbf{a}_{ji}) = \mathbf{w}_i^H \mathbf{a}_{ji} \mathbf{a}_{ji}^H \mathbf{w}_i$ is the i_{th} antenna gain towards the j_{th} source.

V. Resource Allocation Algorithms

A. Band Allocation

An important consideration when designing Wide band CDMA systems are their compatibility with other systems. Bandwidth resource is very limited resource the same band may have to be shared. This Multiband system implement frequency reuse by alternating band assignments between base stations, and cochannel interference is decreased.

The algorithm by Shrader [3] proposes the use of DCPC to balance the received carrier to interference ratio (CIR) for all users. Each base station orders the bands from the preferred band down to the most protected band. An initial band usage layout is used as shown in Figure 4 for the case of 2 bands. All users connecting to the same BS initially utilize the same band according to the shading in Figure 4.

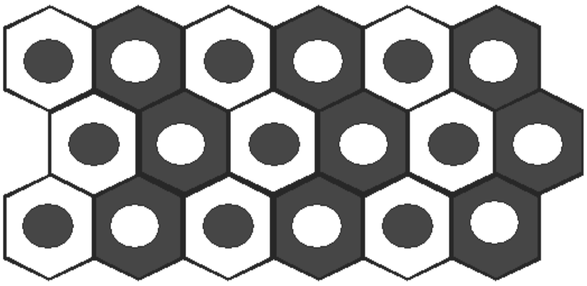


Fig. 4. Initial band allocation layout for two bands

The results obtained in Figure 5 by Shrader[3] indicate that this algorithm performed better than both a wideband system and least load algorithm [8] at low loads. The degradation of the performance at high loads is due to the fact that more users than necessary are being switched to the protected band or dropped from the system. A more conservative switching of users between bands can provide a solution for this problem. In the algorithm proposed here users are switched according to the least load algorithm which is defined as follows:

Calls are assigned to cells with the minimum load (link counts) [8]. If multiple locations achieve the minimum load the assignment is random among them each having equal probability.

B. Joint Power control, Beamforming and Base Station Assignment Algorithm

In this algorithm an adaptive antenna array is used. It consists of a set of antennas designed to attenuate signals from other directions and to attenuate signals from other directions of no interest. The outputs of array elements are weighted and added by a beamformer to produce a directed main beam [9]. A detailed analysis is given in [6].

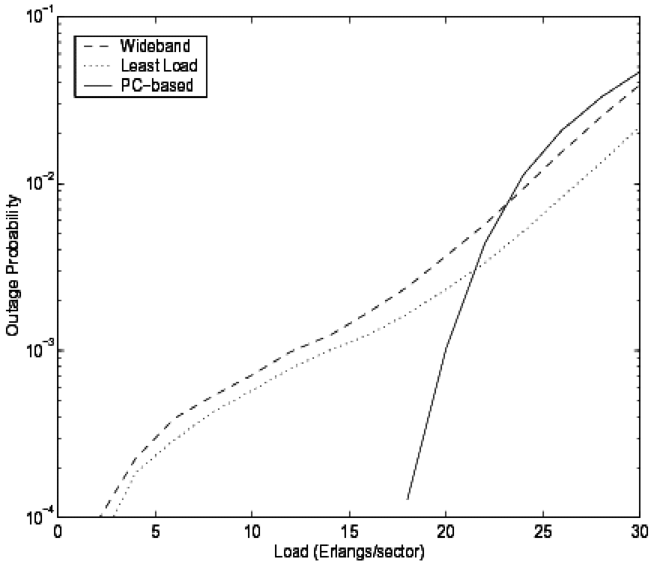


Fig. 5. Outage probability versus load for wideband CDMA, multiband CDMA with least load band allocation, and multiband CDMA with the power-control-based band allocation.

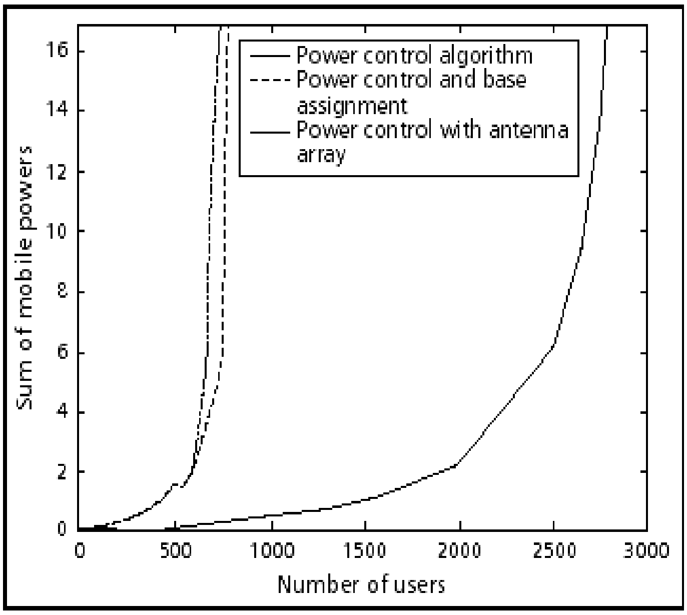


Fig. 6. Total mobile power vs. the number of users

The mobile will access base station (BS) equipped with antenna arrays. Starting from any initial power vector, the algorithm converges to optimal power allocation. The mobile power is updated based on the current beam forming and base station assignment.

The performance of the joint power control base station and beamforming is presented in Figure 6 [6]. The total mobile power versus the number of users is shown. In this case, the capacity increased to 2800 users which is almost 5 times better than the fixed station power control algorithm [9].

VI. Proposed Algorithms

Our proposed algorithm modifies the power control based band allocation algorithm in such a way that it includes both beamforming and a load balancing algorithm. The steps of the algorithm are described below:

1. **Initialization:** Users are assigned to a base station and begin transmitting on the initial band for that base station at transmission power, P_o
2. **Power Control**: DCPC algorithm is run through 5 iterations for all users. Beamforming is performed for the base station at each iteration and calculating the mobile transmitted power for the next iteration. After choosing the BS with the least required power, the mobile i updates the power based on the current beamforming and base station assignment as follows,

$$P_i^{n+1} = \min_{wij, j \in B_i} \{ \gamma_i \sum_{k \neq i} \frac{G_{kj} G_{ai}(\mathrm{w}_{ij}^{n+1}, a_{kj})}{G_{ij}} P_n^k \quad (9)$$

$$+ \frac{\gamma_i N_j \mathrm{w}_{ij}^{{n+1}^H} \mathrm{w}_{ij}^{n+1}}{G_{ij}} \}, i = 1, 2, ..., M \quad (10)$$

subject to $\mathrm{w}_{ij}^{{n+1}^H} a_{kj} = 1$.
3. **Band Switching**: users are switched between bands according to the load sharing algorithm where strongest users are allocated to the preferred band.

An upper bound is implemented so that the number of users does not exceed the allowed capacity for the base station. The strongest mobile users are allocated according to the BS with the least load.

The strongest user in the protected band is moved to the preferred band in order to increase other users CIR and then allocated to the BS with the least load.
4. **Call Dropping**: The number of users above the allowed capacity are denied access to the band and under poor conditions are dropped from the system. According to the specified DCPC users transmitting at P_{max} that have cycled through all bands are dropped from the system.
5. **Algorithm Completion**: The process starts again at step 2 until no users are transmitting at power P_{max} and DCPC has converged.

This algorithm make use of the power control based RAA stated in [6] by solving the problems encountered in [3].

The simulation will be performed using RUNE toolbox [10]. The algorithm will be evaluated using snapshot analysis of the uplink. Simulations for the case of two bands will be compared to the algorithm presented in [3]. Figure 7 shows the system used in the simulation.

VII. Conclusions

This paper presented different RAA based on power control. The proposed algorithm suggested a compromise between them in order to minimize the problems faced in the power control based band allocation algorithm at

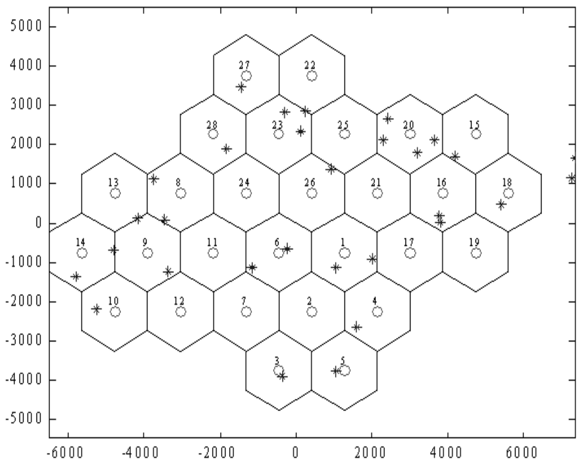


Fig. 7. A 28 cell layout with a cell radius of 1 km

high loads. It has been shown in [9] that the integration of power control and beamforming offers a high performance and can covert some highly loaded infeasible systems into feasible ones. Users are allocated using the least load algorithm, thus the switch between bands is more conservative.

There will always be a tradeoff between complexity and capacity improvements when working on power control. Owing to time constraint obtaining results for the proposed algorithm were not possible. Results will be available in time for the conference. However it is anticipated that performance will improve at higher loads due to the implementation of load sharing algorithm. Using antenna array and beamforming improves the capacity in terms of users that can be supported.

References

[1] Lachlan L. H. ANDREW, "Measurement-based band allocation in multiband cdma," *in Proc. Infocom '99, New York*, pp. 1364–1371, 1999.

[2] Grandhi S.A, Zander J., and Yates R., "Constrained power control," *Wireless Personnal Communications*, vol. 2, pp. 257–270, 1995.

[3] B.E. Shrader, R.S. Karlsson, L.L.H. Andrew, and J. Zander, "Power-control-based band allocation in multiband cdma," *IEEE Globecom '2001*, November 2001.

[4] J. Qiu and J. Mark, "A dynamic load sharing algorithm through power control in cellular cdma," *Proc. The 9th IEEE International Symposium on Personal Indoor and Mobile Radio Communications*, pp. 1280–1284, September 1998.

[5] Rashid-Farrokhi Liu, "Downlink power control and base station assignment," *IEEE Communications Letters*, vol. 1, pp. 102–104, July 1997.

[6] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu, "Joint power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Comm.*, vol. 46, no. 10, pp. 247–256, 1998.

[7] R. A. Monzingo and T. W. Miller, *Introduction to Adaptive Arrays*, John Wiley and Sons Ltd., New York, 1980.

[8] Tony Dean, Phil Fleming, and Alexander L. Stolyar, "Estimates of multicarrier (cdma) system capacity," in *Winter Simulation Conference*, 1998, pp. 1615–1622.

[9] Dejan M. Novakovic and Miroslave L. Dukic., "Evolution of the power control techniques for ds-cdma toward 3g wireless communication systems," *IEEE Communications Surveys*, vol. fourth Quarter, 2000.

[10] J. Zander and S.-L. Kim, *Radio Resource Management for Wireless Networks*, Artech House Publishers, 2001.

Performance Evaluation of MQAM in Fading Channels with Diversity

O. Alani, M. Al-Akaidi, D Alnsour

School of Engineering and Technology
De Montfort University, Leicester
LE1 9BH, UK
Email: mma@dmu.ac.uk

Abstract

The radio spectrum is a finite resource and it is important that all users exploit it efficiently. Consequently modulation scheme used for mobile environment should utilise the RF channel bandwidth and the transmitted power as efficiently as possible.
The introduction of adaptive MQAM (M-ary Quadrature Amplitude Modulation) has enhanced the performance of the communication system as shown in [4]. The impact of using diversity techniques in such systems is investigated. The gain, obtained from utilizing this combination of adaptive MQAM and diversity techniques in the system is shown.

1 Introduction

In real world wireless communication systems, the link channels are not simple AWGN (additive white Gaussian noise), in fact they are characterised by different kinds of fading due to environmental factors and interference from other users that adversely affect signals propagate in these channels. Therefore we cannot rely on the formula set by Shannon that relates the capacity of channel (hence spectral efficiency) to SNR (signal to noise ratio) since this relation is meant for AWGN channels. LEE [1] derived a relation which gives the average channel capacity of flat fading channel and it's obtained by averaging the capacity of AWGN channel, $C = W \log_2(1+\gamma)$, over the distribution of the received SIR (signal to interference ratio), γ.

To enhance the spectral efficiency, adaptive transmission schemes are used. The main idea about these schemes is real time balancing of the link budget through adaptive variation of the transmitted power level, symbol rate, constellation size, coding rate/scheme, or any combination of these parameters [2].

Adaptive transmission schemes provides much higher average spectral efficiency by taking advantage of the "time varying" nature of the propagation channel without sacrificing BER. Other fading compensation techniques (such as increased link budget margin or interleaving with channel coding) are designed for the worse case channel conditions which result in a poor utilisation of the full channel capacity for majority of the time (under negligible or shallow fading conditions) [2].

Shannon's channel capacity theorem states that, for an arbitrary small probability of error, the maximum possible bandwidth efficiency is limited by the noise in the channel, and is given by: $C/W = \log_2(1+\gamma)$. This theorem applied to an AWGN time-invariant channel with constant γ, in this case the noise of the channel rather than interference is present; where C and W are the capacity and bandwidth of the channel respectively. In cellular system where interference is a major issue, the performance of a modulation scheme in terms of spectral efficiency is interference limited rather than noise limited. For fading channels, LEE [1] has derived the capacity of the channel to consider the variation of γ due to different propagation impairment in a Rayleigh channel. The authors in [3] have derived an expression for the capacity of fading channels for different adaptation schemes when information about the channel conditions was provided for both transmitter and receiver or one of them. Goldsmith [2] used the theory that developed in [3] to obtain a closed-form expression for the capacity of Rayleigh fading channels under different adaptive transmission and diversity combining techniques.

However, if the channel fade level is known to the transmitter, then Shannon capacity is achieved by adapting the transmit parameters (power, data rate, and coding schemes) relative to this fade level.

In [4], the authors' proposed adaptive variable rate-variable power uncoded MQAM in which they derived the spectral efficiency for their proposed system. They have shown that a power loss of K depending on BER and independent of the fading distribution has been introduced when using MQAM relative to the optimal transmission scheme. In this paper, the same system is considered with adaptive rate and fixed power transmission combined with diversity at the receiver to observe the improvement that a diversity-combining scheme would introduce to the spectral efficiency.

2 Channel model

We assume a slow varying channel at a rate much slower than the data rate and a Rayleigh fading channel with exponentially distributed PDF of carrier to noise ratio, CNR

$$p(\gamma) = \frac{e^{-\gamma/\bar{\gamma}}}{\bar{\gamma}} \tag{1}$$

Where $\bar{\gamma}$ is the average received CNR.

In this work, the adaptive rate fixed-power MQAM system is combined with a well-known diversity combining techniques. In particular we use the maximal ratio combining (MRC) and selection combining (SC) of the received signal. The former requires the M signals to be weighted proportionately to their CNR and then summed coherently. Perfect knowledge of the branch amplitudes and phase is assumed. The disadvantage of MRC is that it requires knowledge of the branch parameters and independent processing of each branch. The PDF of the received CNR at the output of a perfect M-branch MRC is derived in [5] to be:

$$P^{mrc}(\gamma) = \gamma^{M-1} e^{-\gamma/\bar{\gamma}} /(M-1)\bar{\gamma}^{-M} \tag{2}$$

In the SC technique only one of the M receivers having the highest baseband CNR is connected to the output. Unlike the MRC it does not require coherent reception.

The PDF of the received CNR at the output of M-branch is again derived in [5] and it is given by:

$$P^{SC}(\gamma) = \frac{M}{\bar{\gamma}}(1 - e^{\gamma/\gamma})^{M-1} e^{-\gamma/\bar{\gamma}} \tag{3}$$

3 Adaptive MQAM

The spectral efficiency of a communication link can be increased by using a multilevel modulation scheme, such as MQAM, which tends to send a multiple bits per symbol.

The radio channels in a wireless mobile communication system are affected by different types of fading (multipath, shadowing, etc), therefore, they will have negative effect on signals carried on these channels.

To compensate for these channel impairments imposed by fading, adaptive modulation scheme is used. In adaptive MQAM, information about the channel conditions at the receiver is fed back to the transmitter so that it will adjust its' transmitted modulation level (constellation size) accordingly. This channel information is usually acquired by using a pilot signal or inserting a training sequence into the stream of MQAM data symbol to extract the channel induced attenuation and phase shift [6].

The BER of a coherent MQAM over an AWGN channel assuming a perfect clock and carrier recovery can be well approximated by [4]

$$BER(M,\gamma) \approx 0.2e^{-\frac{3\gamma}{2(M-1)}} \tag{4}$$

This expression for the BER approximation is invertible [7] as it provides a closed form for the spectral efficiency of MQAM as a function of CNR and BER.

For given CNR and assuming ideal Nyquist pulses the spectral efficiency of a continuous rate MQAM can be approximated by inverting Eq.4 giving;

$$\frac{R}{W} = \log_2(M) = \log_2(1 + \frac{3\gamma}{2K}) \tag{5}$$

Where R is channel bit rate, and K= -ln (5BER).

In practice the CNR is not fixed but fluctuates due to channel impairments, therefore the spectral efficiency is calculated by integrating the RH side of Eq. 5 over the distribution of CNR i.e. PDF of the received CNR as shown below:

$$\frac{R}{W} = \int \log_2\left(1+\frac{3\cdot\gamma}{k}\right)\cdot P(\gamma)d\gamma$$

Hence for the MRC diversity case, we integrate over MRC distribution function as given in Eq. 2:

$$\frac{R}{W} = \int \log_2(1+\frac{3\gamma}{2K})\frac{\gamma^{M-1}e^{\frac{\gamma}{\overline{\gamma}}}}{(M-1)!\overline{\gamma}^M}d\gamma \qquad (6)$$

To yield the following

$$\frac{C^{mrc}}{W} \approx [\log_2(e)]\times \mathrm{P}_M(\frac{-1}{\overline{\gamma_o}})\left(-E+\ln\overline{\gamma_o}+\frac{1}{\overline{\gamma_o}}\right)$$

$$+\sum_{k=1}^{M-1}\frac{\mathrm{P}_k(\frac{-1}{\overline{\gamma_o}})-\mathrm{P}_{M-k}(\frac{-1}{\overline{\gamma_o}})}{k} \qquad (7)$$

Where $\overline{\gamma_o} = \frac{3}{2K}\overline{\gamma}$, and $\overline{\gamma}$ is the average received CNR and k is an integer.

The achievable spectral efficiency is compared with that achieved by an M-array independent AWGN channels, optimal combining (MRC), and to that of optimal rate and constant power without diversity i.e. a normal Rayleigh channel, which are given by

$$C_{AWGN} = W\log_2(1+M.\gamma) \qquad (8)$$

$$C_{opt\text{-}rate} = \int_0^\infty W\log(1+\gamma)\,\rho(\gamma)\,d\gamma, \qquad (9)$$

where $\rho(\gamma)$ is given in Eq.1. In the case of SC diversity we integrate Eq.5 over Eq.3 to yield the following approximation [2].

$$\frac{C^{Sc}}{W}\approx M\log_2(e)\sum_{k=0}^{M-1}\frac{(-1)^k}{1+k}\binom{M-1}{k}e^{(1+k)/\overline{\gamma}}$$

$$\left[E+\ln(\frac{1+k}{\overline{\gamma}})-(\frac{1+k}{\overline{\gamma}})\right] \qquad (10)$$

Where E is Euler constant (E = 0.57721566). However our results will be restricted to those of MRC diversity scheme.

4 Results & Discussions

In Figure 1, the spectral efficiency for different diversity levels of MQAM system in the case of MRC is shown. The spectral efficiency of Rayleigh channel with optimal rate and constant power, and the Shannon capacity are shown as well.

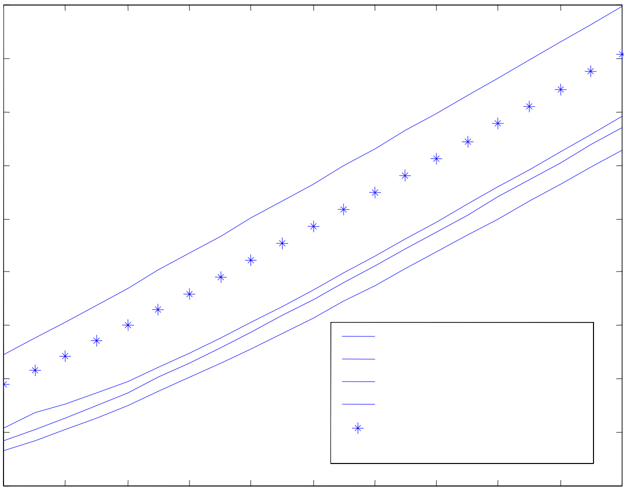


Figure 1: Capacity of MQAM with MRC diversity system for average combined received SIR value.

We notice that when the value of CNR was taken as the average combined received CNR an increase of about 1/4 bits/sec/Hz as the diversity goes from M=1 to M=2, and another 1/4 bit/sec/Hz as it increases to M=4.

The results are compared to those of optimal rate and constant power Rayleigh fading channel and to the Shannon AWGN channel capacity $C/W = \log_2(1+\gamma)$.

It could be realised that when the value of CNR was taken as the average combined received CNR, the obtained results from MRC diversity were identical to those obtained in Nakagami fading channels with the level of diversity corresponding to the value of the Nakagami fading parameter (m). This is an indication that less fading is being exposed to the channel (increasing m) when the diversity increased. It could be noticed that as the diversity increased the spectral efficiency approaches that of the Shannon capacity $C/W = \log_2(1+\gamma)$.

When the value of CNR was taken as the per branch power, then we can see that as the diversity level increases the spectral efficiency approaches that of the Shannon capacity of an M-array of independent AWGN channels $C_{AWGN} = W \log_2(1+M.\gamma)$ as shown in Figure 2. More significant improvement can be seen when the calculations are based on the per branch value of CNR. An increase of 1.3 bits/sec/Hz achieved when the diversity increased from M=1 to M=2, and 1.2 bits/sec/ Hz achieved as the diversity increased from M=2 to M=4.

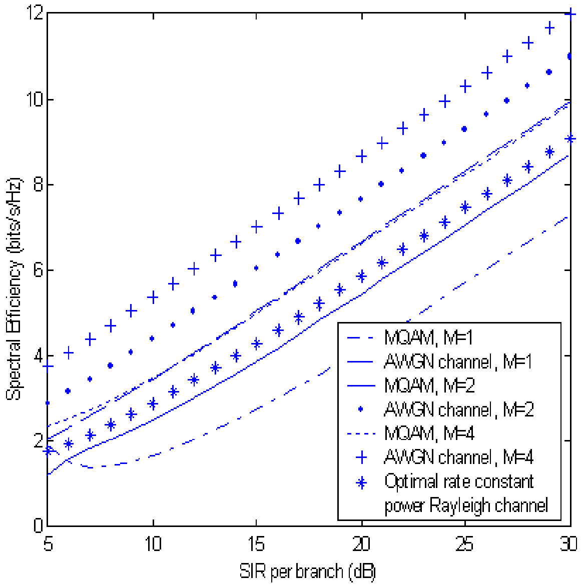


Figure 2: Capacity of MQAM system with MRC diversity for per branch SIR value

It could be noticed that when M=4, the spectral efficiency of the MQAM system has exceeded that of the optimal rate constant power Rayleigh channel and it coincides with the M=2 of an M array AWGN channel (Eq.8).
The above explanation could be applied for the results obtained when SC diversity technique was used with the results showing less performance in this case.

5 Conclusion

As the work in [4] shows that there is a constant gap between the channel capacity and the maximum efficiency of adaptive MQAM which is a simple function of BER. We have combined diversity schemes in order to see the effects on the spectral efficiency of adaptive MQAM communication system. The results show slight improvement in the spectral efficiency as the diversity level increases. These results indicate that the complexity introduced to the system by the diversity technique is unjustified as they have very little significance. However, if the per branch CNR were taken in consideration the results have improved but still cannot justify the complexity of using diversity technique.

References

1- William. C. Lee, "Estimate of channel capacity in Rayliegh fading environment" IEEE transaction. Veh. Technology, vol. VT-39, pp187-190, August 1990.

2- Mohammed Alouini, A Goldsmith, "Capacity of Rayliegh fading channels under different adaptive transmission and diversity combining techniques" IEEE Transaction on Vehicular Technology, vol. 48, NO. 4, pp. 1165-1181, July 1999.

3- A. Goldsmith, P. Varaiya, "Capacity of fading channels with channel side information," IEEE trans. Inform. Theory, Vol. IT-43, pp. 187-190, November 1997.

4- A. Goldsmith, and S. Chua, "Variable rate variable power M-QAM for fading channels", IEEE trans. On communication, Vol. COM_45, pp.1218-1230, October 1997.

5- William Jakes, *Microwave mobile communications*, John-Wiley & sons, 1974.

6- S. Sampei and Sunga, "Rayleigh fading compensation for QAM in Land mobile Radio communications." IEEE trans.Veh. Technology, Vol. VT-43, pp. 294-302, May 1993.

7- Mohammed Alouini, A Goldsmith, "Adaptive Modulation over Nakagami Fading channels", Kluwer Journal on Wireless Communications, Vol. 13, No. 1-2, pp.119-143, May 2000.

AN IMPROVED COMMUNICATION CHANNEL IN DYNAMIC RECONFIGURABLE DEVICE FOR MULTIMEDIA APPLICATIONS

Yuuki Soga, Takafumi Yuasa, Tomonori Izumi, Takao Onoye, and Yukihiro Nakamura

Department of Communications and Computer Engineering,
Graduate School of Informatics, Kyoto University
Yoshida-hon-machi, Sakyo, Kyoto, 606-8501 Japan
E-mail: {soga, tyuasa}@easter.kuee.kyoto-u.ac.jp,
{izumi, onoye, nakamura}@kuee.kyoto-u.ac.jp

KEYWORDS

Programmable Logic Device(PLD), Dynamic Reconfigurable Device, Communication Channel, Granularity

ABSTRACT

Plastic Cell Architecture (PCA) is proposed as an extension of Programmable Logic Device (PLD). PCA supports dynamic self-reconfiguration, where a circuit can be reconfigured by a control from another circuit inside the same device. On PCA, a target function is realized by asynchronous cooperation of small circuits connected by communication channels based on a hand-shake mechanism. In order to realize PCA, several types of hardware architectures of PCA are proposed and some of them are implemented as VLSI chips. This paper presents an improvement of our prototype PCA device named PCA-Chip2 from the point of view of the efficiency of the reconfiguration. Based on trial implementations for application examples by using our design tool, the pipeline system of the communication elements and the granularity of logic elements is also optimized. The proposed optimization realizes approximately four times faster data transmission and 8.6% better logic implementation.

INTRODUCTION

Nowadays *Field Programmable Gate Arrays (FPGAs)* are frequently used in various electronic products, because they have an ability to modify the functions implemented on them and contribute to achieve short design time and low cost. Recently, programmable devices which support *dynamic reconfiguration* are proposed, where the circuits can be reconfigured at run-time, and get into the limelight as key devices for *reconfigurable computing* systems.

Aiming for higher reconfigurability, *Plastic Cell Architecture (PCA)* is proposed[Nagami et al. 1998]. The key features are (1) asynchronous cooperation of small circuits called *objects* connected each other by pipelined communication paths and (2) dynamic self-reconfiguration where an object can be reconfigured by control signals generated by another object in the same device. By utilizing the function, PCA is expected to be a key device to realize a new computational mechanism where a system reconfigures itself adoptively depending on the dynamic change of the situation.

In order to realize PCA, hardware architectures of PCA have been proposed and developed as VLSI chips[Nakada et al. 2001, Tsutsui et al. 2001]. Together with these device, several studies are promoted in the fields of design automation tools and applications [Murakami et al. 2000, Okamoto et al. 2001]. We have designed our PCA device named *PCA-Chip2* [Tsutsui et al. 2001] and are developing a design automation environment for it [Okamoto et al. 2001]. Our trial implementations of some target functions on PCA-Chip2 with the tools reveal that the size of configuration data to implement a function on PCA-Chip2 designed with our tools is larger than that of commercially available FPGAs [Altera 2000, Xilinx 2001] by one order of magnitude. The hardware architecture should be improved for efficient configuration process by reducing the size of configuration data and the time for configuration process. The objective of the present paper is an architectural improvement of PCA-Chip2.

In this paper, the hardware architecture of PCA-Chip2 is examined from the viewpoints of communication systems and logic resources. To reduce the time for configuration process, the communication resource of PCA-Chip2 is enhanced considering the trade-off between data transmission rate and the hardware size of a communication resource. First, a discussion about the relation between the bit width of communication paths and the efficiency for layout process is presented. Second, an improved communication mechanism is proposed to accelerate data transmission speed. As for logic resources, the size of basic logic blocks is examined by implementing some benchmark circuits on the device and analyzing efficiency of logic implementation for each size of logic element. The experimental result shows that the proposed architecture contributes more efficient reconfiguration.

The rest of the paper is organized as follows. In Section 2, the outline of PCA and the architecture of PCA-Chip2 are presented. Improvements of the communication system and the logic element are described in Section 3 and 4, respectively. Finally, Section 5 concludes the paper.

PCA AND PCA-CHIP2

PCA realizes the characteristic of dynamic self-reconfiguration equipped with two types of modules, one to to realize logics and the other to control dynamic reconfiguration of logics and communications. This section describes the outline of PCA and our prototype device named *PCA-Chip2*.

Basic Architecture of PCA

PCA consists of basic cells referred to as *plastic cells* arranged in a two-dimensional array and connected by local connections each other as illustrated in Fig.1 (left-top). A plastic cell consists of two modules, a *plastic part* and a *built-in part*. A plastic part works as a programmable logic device and also works as a Random Access Memory (RAM) by utilizing the memory to store configuration data. A built-in part works as a node for communication paths and as a controller for configuration of logics and communications.

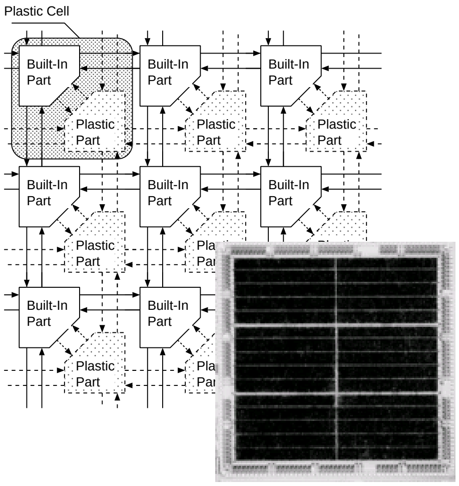


Fig. 1: a block diagram of PCA and a micrograph of our prototype device PCA-Chip2

On PCA, a target function is realized by cooperation of small circuits referred to as *objects* each of which is configured within adjacent plastic parts. Objects are connected each other via communication paths configured in built-in parts. Objects and Communication paths are configured by a sequence of commands and configuration data for the target resources conducted by built-in parts.

Hardware Architecture of PCA-Chip2

PCA-Chip2 is our prototype device (Fig.1 right-bottom) [1]

The plastic part of PCA-Chip2 is an extension of array-style PLDs which originate from Programmable Logic Arrays (PLAs). Fig.2 illustrates the structure of the plastic part. The plastic part of PCA-Chip2 adopts Look-Up Tables (LUTs) instead of programmable AND devices in PLA. A term for a target logic expression is generated by a series of cascaded LUTs and terms are gathered to generate the target logic by SUM array. Each plastic part has 24 series of 8 cascaded LUTs. Plastic part also has 16 flip-flops to realize sequential circuits and an address decoder to access the memories of LUTs.

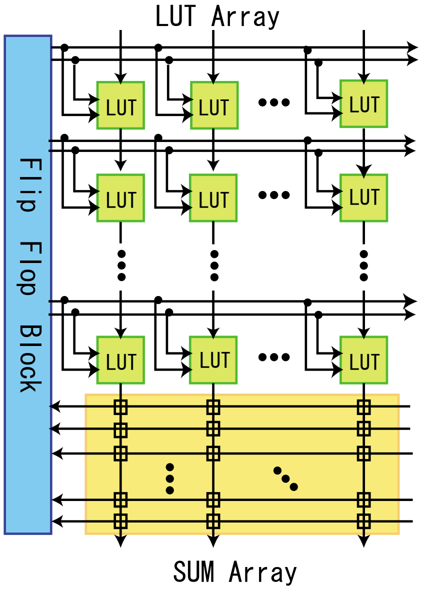


Fig. 2: the plastic part of PCA-Chip2

A built-in part works as a node for communication paths and as a controller for configuration of logics and communications. A built-in part of PCA-Chip2 has one communication channel connected to each direction of four built-in parts in the adjacent plastic cells and the plastic part in the same plastic cell. Each channel can transmit 10 bit data at a time, referred to as a *particle*. Each communication channel is equipped with a hand-shake mechanism and a series of connected channels works as a pipelined communication path. Communication channels in PCA-Chip2 adopts *2τ-pipeline* system where one particle is transmitted every two cycles.

[1] The VLSI chip in this study is designed with Cadence, Mentor, and Synopsys CAD tools through the chip fabrication program of VLSI Design and Education Center(VDEC), the University of Tokyo. The VLSI chips in this study have been fabricated in the chip fabrication program of VDEC with the collaborations by Rohm Corporation and Toppan Printing Corporation, by Hitachi Hokkai Semiconductor Ltd. and Dai Nippon Printing Corporation, and by On-Semiconductor, Nippon Motorola LTD., HOYA Corporation, and KYOCERA Corporation.

Table 1: Command set of PCA-Chip2

commands	description
ROUTING	assign communication path to the adjacent cell
CLEAR	clear the path
END CFG	the end of configuration data
END Data	the beginning of data for calculation
NEXT Read /Write	read/write configuration data form/to an adjacent plastic cell
CFG Write	write configuration data
CFG Read	read configuration data

Objects and communication paths are configured by a *stream* of commands and configuration data for the target resources. Table 1 summarizes the commands of PCA-Chip2. A stream for PCA-Chip2 is exemplified in Fig.3. Given a ROUTING command, a communication channel sets a route from the channel which sends the command to the channel in the direction designated by the command. Once the route is set, the built-in part forwards the following particles to the direction until it receives a CLEAR command. Thus, a sequence of ROUTING commands configures a communication path in a series of connected channels. Given a CFG Write (configuration-write) command, a channel configures the associated plastic part with the following configuration data until the end of the configuration data. After the configuration completes input/output data for the target calculation flow the communication paths. Given a clear command, a channel forwards the command to the direction and clears the route in it.

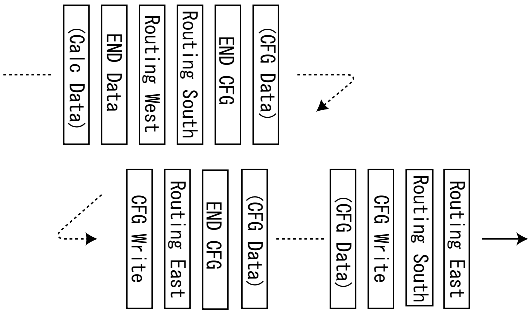


Fig. 3: An example of stream for PCA-Chip2

Together with the device, a computer aided design tools have been developed. With the tools the stream for configuration and calculation is generated from a functional description in C, through processes of partitioning, logic synthesis, layout, and so on. Our trial implementations of some target functions on PCA-Chip2 with the tools reveal that shortage of communication channels causes detours of communication paths and waste of unused plastic-parts. Furthermore, there are unused LUTs even within used plastic parts because the capacity of the plastic parts is not fit for the size of an object in many cases. These problems affects the performance of configuration process. To improve configuration processes on PCA-Chip2, reduction in the size of configuration data for a target function and acceleration of data transmission are effective. As for the built-in part, we propose an enhancement of communication channels to solve the shortage of communication resources and an improvement of the pipeline system to accelerate the transmission. As for the plastic part, we propose an optimization of the granularity (size) of the plastic part for more efficient logic implementation.

IMPROVEMENT OF BUILT-IN PART

Communication Channels

A plastic cell has an ability to be configured as an object in the plastic part and a part of communication paths simultaneously. However, there are plastic cells used only for communication because of shortage of communication resources. Our trial implementations of FFT and viterbi decoder with our design tools have revealed that the total length of routes for communications is 1.76 times longer than that of an ideal case where each path is assumed to be the shortest path between terminals. The fact implies an increase in the amount of used resources and the size configuration data. Therefore, an increase of communication channels for each direction is examined here to dissolve the congestion of communication requests.

In addition to the number of communication channels, a connection mechanism between channels must be decided taking a trade-off between its connectivity and hardware size into account, because higher connectivity leads to larger hardware size in general. Types of connection mechanism can be categorized as follows.

- fixed: Connections among channels are fixed. A particle in a channel can be transmitted only to the channel with the same ID as the channel (Fig.4 above).
- command-select: A channel to be connected to is specified by a routing command.
- auto-select: A vacant channel to be connected to is automatically selected at runtime (Fig.4 below).

In this paper, the command-select mechanism is omitted because it needs considerable changes in the command set and the design tools.

A built-in part including an increased number of channels with each connection system is described in Verilog HDL and synthesized to estimate its hardware size. Table2 (left) shows the sizes of the built-in part for the cases where the number of channel is 1, 2, 3 or 4 and the connection mechanism is fixed or auto-select. The larger number of channels are equipped

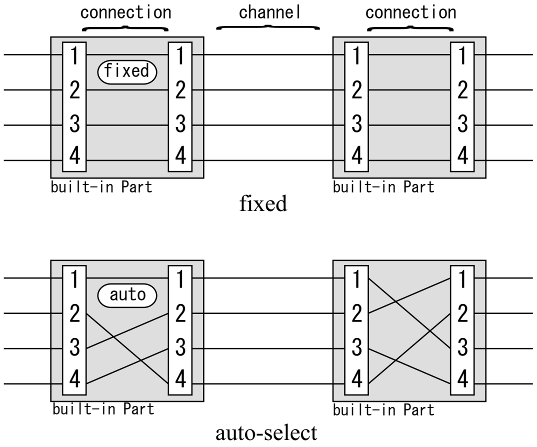


Fig. 4: connection mechanisms between channels

with, the larger hardware size is needed and the greater the difference between the size of the two mechanisms becomes.

To evaluate the effect of multiplication of channels, the total length of routes for communications is examined by trial implementations of some functions such as bitonic sorter, FFT, and viterbi decoder by using our design tools. The result is summarized in Table.2 (right), which lists the averaged total length of communication paths normalized by the ideal length. Table2 demonstrates that increasing the number of channels to two results in almost the ideal length for both connection mechanism, and the difference in the lengths is less than 1%. Thus, we adopt the 2-channel fixed connection mechanism since it achieves almost the ideal path length while the increase of the hardware size is restrained up to 42%.

Table 2: hardware sizes and total lengths categorized by the number of channels and connection system

#channel	hardware size		total length	
	fixed	auto	fixed	auto
1	4235		1.758	
2	6000	6568	1.011	1.006
3	8265	11793	1.001	1.001
4	10000	17269	1.000	1.000

Hand-shake System

Each channel transfers a particle based on a hand-shake protocol with request and busy signal and works as a node of a pipelined communication path. PCA-Chip2 adopts a 2τ-pipeline communication system where a particle in a channel can be transmitted to an adjacent channel only when it has no particle. Therefore, one particle can be transmitted to the next channel every two cycles at the maximum rate as exemplified in Fig.5(left).

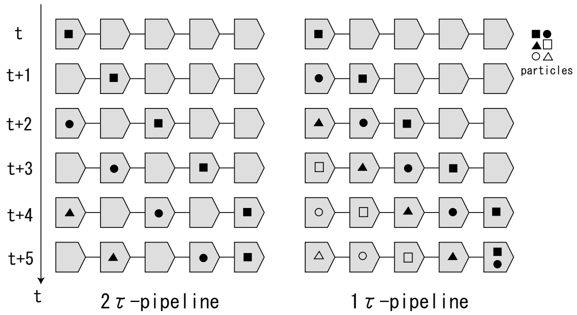


Fig. 5: 2τ- and 1τ-pipeline systems

On the other hand, a 1τ-pipeline system can transmit a particle to the next channel at every cycle. Block diagrams of the two mechanisms are illustrated in Fig.6. To adopt a 1τ-pipeline system, additional hardwares are needed, such as data registers to store a particle in the case that the pipeline is stalled.

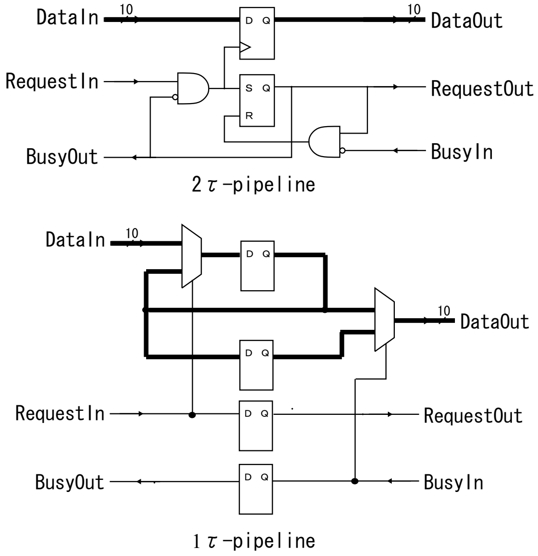


Fig. 6: block diagrams of 2τ- and 1τ- pipeline systems

Built-in parts equipped with each pipeline system are described in Verilog HDL and synthesized to estimate the hardware size of each system. The result is shown in Table 3. A channel adopting the 1τ-pipeline realizes 2 times faster data transmission than the 2τ-pipeline at the cost of only 20% larger hardware size.

Table 3: size of each pipeline system

	2τ-pipeline	1τ-pipeline
size of built-in part	6000	7123

Proposed Architecture

According to the discussions above, we propose a built-in part adopting 2 channel fixed connection system and 1τ-pipeline system. The hardware size of built-in part, which is 4235 gates originally, becomes 6000 by adopting 2-channel fixed connection and 7123 by adopting 1-τ pipeline system. Consequently, this improvement accelerates data transmission 4 times faster at the maximum rate by accepting 70% larger hardware size.

IMPROVEMENT OF PLASTIC PART

In this section, the size of a plastic part, in other word, the granularity is examined. In the configuration process of PCA-Chip2, configuration data for an object is loaded into all the LUTs included in plastic parts used for the object, even if they aren't used for logic implementation. Such unused LUTs are referred to as invalid LUTs and cause redundant configuration time. Our experimental results of trial implementation of some benchmark circuits on PCA-Chip2 demonstrates that about half of the LUTs are invalid and implies that a reduction in invalid LUTs is effective to reduce the size of configuration data.

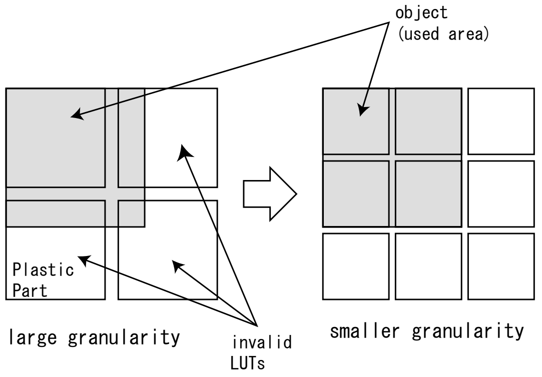


Fig. 7: granularity of plastic part and invalid LUTs

Although a plastic part which consists of fewer LUTs should reduce invalid LUTs(Fig.7), it increases the overhead of peripheral circuits such as flip-flops and an address decoder. Taking the tradeoff into account, total areas to implement functions are examined by trial implementations on a plastic part of given size. Total area is calculated by multiplying the hardware size of a plastic part by the number of used ones. MCNC benchmark circuits[Yang 1991] are used as target functions, and our design tools are used for the implementation. The size of a plastic cell is given as follows.

- The number of cascaded LUTs in a term : 8, 16, and 32.
- The number of terms: multiples of 3 between 3 to 30.

The number of terms is restricted to multiples of 3 because of our addressing system of the memory for LUTs.

The experimental result is summarized in Fig.8. The horizontal axis represents the number of terms and the vertical one does the averaged total area for target functions for each number of cascaded LUTs in a term. The area is normalized by the area in the case of the original PCA-Chip2. The result demonstrates that a plastic part with 9 terms each of which consist of 8 cascaded LUTs gives the minimum area. In the case, the rate of valid LUTs is 55.0%, which is 8.6% better than that of the original PCA-Chip2.

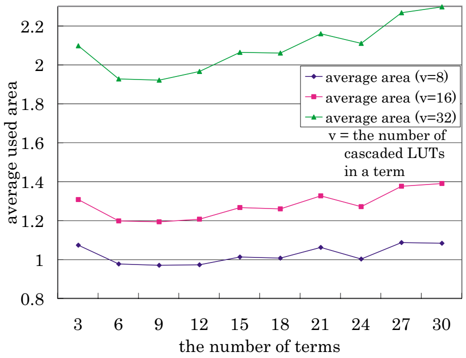


Fig. 8: Granularity of Plastic Part v.s. area usage

CONCLUSION

In order to accelerate the dynamic reconfiguration process on PCA-Chip2, the hardware architecture of the built-in part and the plastic part are improved. As for the built-in part, the number of communication channels is multiplied into 2 and the hand-shake mechanism for pipelined communication paths is improved. The finer granulated plastic part is adopted for more efficient logic implementation.

Table 4 summarizes the specifications of the original PCA-Chip2 and the proposed architecture integrated with all the improvements above.

Table 4: specifications of the original PCA-Chip2 and proposed architecture

	PCA-Chip2	proposed architecture
built-in part		
bit width of a channel	10	10
# commands	8	8
# channel	1	2
connection system	—	fixed
pipeline system	2τ	1τ
plastic part		
#LUTs in a cell	8×24	8×9
configuration data size(bit)	2331	891

The proposed one achieves smaller configuration data and faster data transmission compared to the original one. Currently, we are in the process of fitting our tools for the new architecture and plan to evaluate the total effect of the improvements by trial implementations of functions.

In addition to the further improvement of the architecture of the device, there are future works to develop and to improve design tools, operating systems, design methodologies, and so on, in order to realize a reconfigurable computing system with PCA.

REFERENCES

Altera Corporation, “AN116 configuring APEX20K, FLEX10K and FLEX 6000 devices,” May 2000.

D. Murakami, T. Izumi, T. Onoye, and Y. Nakamura, “A hardware algorithm of dynamic area allocation to circuits for plastic cell architecture,” in Proc. of *the Euromedia Conference*, May 2000, pp.85–89.

H. Nakada, H. Ito, R. Konishi, A. Nagoya, K. Oguri, T. Shiozawa, and N. Imlig, “Self-reorganizing systems on VLSI circuits,” in Proc. of *the International Symposium on Circuits and Systems (ISCAS2001)*, May 2001, pp.IV310–313.

H. Tsutsui, A. Tomita, S. Sugimoto, K. Sakai, T. Izumi, T. Onoye and Y. Nakamura, “LUT-Array-Based PLD and Synthesis Approach Based on Sum of Generalized Complex Terms Expression,” *IEICE Trans. on Fundamentals*, Vol.E84-A, No.11, pp.2681–2689, 2001.

K. Nagami, K. Oguri, T. Shiozawa, H. Ito, and R. Konishi, “Plastic cell architecture: A scalable device architecture for general-purpose reconfigurable computing,” *IEICE Trans. on Electronics*, Vol.E81-C, No.9, pp.1431–1437, 1998.

S. Yang, “Logic Synthesis and Optimization Benchmarks User Guide Version 3.0,” Jan. 1991.

T. Okamoto, K. Sakai, A. Tomita, S. Sugimoto, T. Izumi, T. Onoye, and Y. Nakamura, “C-based design automation environment for Plastic Cell Architecture,” in Proc. of *the Workshop on Synthesis And System Integration of MIxed technologies (SASIMI2001)*, pp.45–49, Oct. 2001.

Xilinx, “Virtex 2.5v field programmable gate arrays product specification ds003-1(v2.5),” April 2001.

ACKNOWLEDGEMENT

This work is supported in part by the 21st Century COE Program (Grand No. 14213201).

AUTHOR BIOGRAPHY

YUUKI SOGA received his B.E. degree in Electrical and Electronic Engineering from Kyoto University in 2001. Presently, he is a master course student at the Department of Communications and Computer Engineering, Kyoto University.

TAKAFUMI YUASA received his B.E. degree in Electrical and Electronic Engineering from Kyoto University in 2002. Presently, he is a master course student at the Department of Communications and Computer Engineering, Kyoto University.

TOMONORI IZUMI received his B.E. degree in Computer Science and M.E. and Ph.D. degrees in Electrical and Electronic Engineering all from Tokyo Institute of Technology, in 1992, 1994, and 1998, respectively. Since 1998, he has been with Kyoto University, where he is a research associate in the Department of Communications and Computer Engineering. He has also been with Synthesis Corporation since 1998, where he is a senior research scientist.

TAKAO ONOYE received B.E. and M.E. degrees in electronic engineering, and Ph.D. degree in information systems engineering all from Osaka University, Japan, in 1991, 1993, and 1997, respectively. He joined the Department of Information Systems Engineering, Osaka University in 1993 as a research associate, where he was promoted to a lecturer in 1998. He has been with the Department of Communications and Computer Engineering, Kyoto University as an associate professor since 1998. Presently, he is an associate professor in the Department of Information Systems Engineering, Osaka University. He has also served a principal research scientist of Synthesis Corporation since 1998 and Arnis Sound Technologies, Co., Ltd. since 1999.

YUKIHIRO NAKAMURA received his B.S., M.S. and Ph.D. degrees in Applied Mathematics and Physics from Kyoto University, in 1967, 1969 and 1995, respectively. From 1969 to 1996, he was with the Electrical Communications Laboratories, NTT. Concurrently, he was a guest professor at the Graduate School of Information Systems, University of Electro-Communications. Since 1996, he has been a professor in the Department of Communications and Computer Engineering at Kyoto University. He has also served a coordinator of Synthesis Corporation since 1998 and Arnis Sound Technologies, Co., Ltd. since 1999. He received the best paper award of IPSJ, the Okochi memorial technology prize, the Minister’s prize of the science and technology agency and the achievement award of IEICE in 1990, 1992, 1994 and 2000, respectively.

SECURITY

A COMPREHENSIVE APPROACH TO ELIMINATING SPAM

Pawel Gburzynski
Department of Computing Science
University of Alberta
Edmonton, AB, CANADA T6G 2E8
pawelg@sheerness.cs.ualberta.ca

Jacek Maitan
appHome, Inc.
1301 Parkinson Ave.
Palo Alto, CA 94301
jacekm@sheerness.cs.ualberta.ca

KEYWORDS
Electronic Mail, Spam, Privacy

ABSTRACT

We introduce an effective and reliable method of eliminating spam, i.e., unsolicited electronic mail, and illustrate it through a publicly available prototype implementation. Its basic idea consists in creating multiple restrictive aliases personalized for senders. This can be viewed as a reversal of the traditional paradigm whereby an E-Mail recipient is identified via a conceptually single global address accessible to all potential senders.

INTRODUCTION

One problem with the traditional electronic mail system is the ease with which the senders of unsolicited junk mail, the so-called spammers, can exploit this medium at no cost to them, but at a considerable annoyance to its serious users. Owing to the world-wide scope of the Internet, no legislative measures can be effective against this kind of abuse; thus, the only viable ways of fighting spam must involve modifications to the mail handling systems, i.e., E-Mail servers (also called Mail Transport Agents or MTA'a) and (possibly) E-Mail clients (also called Mail User Agents or MUA's). Due to the large infrastructure of already deployed MTA's and MUA'a, any solution of this kind must be compatible with that infrastructure, i.e., it must assume the standard mail delivery protocol (Klensin 2001; Postel 1982).

Most existing anti-spam tools are based on the concept of mail filtering. There seems to be no generally acknowledged, viable, reliable, and effective scheme aimed at eliminating address harvesting, i.e., the practice of collecting E-Mail addresses from various publicly exposed places with the intention of using them for spamming.

In this paper we outline a comprehensive solution to the problem of address harvesting and suggest ways of deploying it without revolutionizing, or even retiring, the present infrastructure. In the following discussion, we shall present our concept by describing the specific features available in a prototype system, called the Remailer, accessible at `http://sheerness.cs.ualberta.ca/remailer/`. Its role is to illustrate the raw power of our proposed scheme which, in a more "professional" implementation will be covered by a more friendly and more "flashy" user interface.

THE CONCEPT

From the viewpoint of its subscriber, the main function of the Remailer is easy generation of limited-time limited-accessibility alternative E-Mail addresses, which can be viewed as synonyms (aliases) for the (fixed) permanent address of the subscriber. Thus, the Remailer is usable in combination with any standard E-Mail service and can be accessed from any standard MUA.

The Remailer provides a Web interface whereby its subscribers can ask it for an unlimited number of unique aliases to form the "username" part of their E-Mail addresses. When requesting a new alias, the subscriber may specify a collection of attributes indicating for how long the alias should be valid, how many messages can be received on the alias before it is invalidated, and who is allowed to use this alias for sending messages. In a situation when the alias is exposed (e.g., by being inserted into a merchant's form on the Web), it can be made short-lived and thus useless for harvesting. Whenever a message addressed to the alias is received by the Remailer, the system checks whether the message meets the alias's reception criteria. If this is the case, the message is delivered, i.e., forwarded to the permanent address of the subscriber. Otherwise, it is bounced to the sender.

Th Remailer offers a means for its subscriber to safely publish an E-Mail address to be used by any sender trying to reach the subscriber for the first time. A message arriving on such an alias is never forwarded to the subscriber. Instead, the Remailer sets up a new short-lived alias restricted to the sender, and bounces the message. The returned message is preceded by a note explaining that the sender has to re-submit it along with a response to a subscriber-defined challenge. Having received a correct response, the Remailer delivers the message to the subscriber, who can then *validate* the new alias.

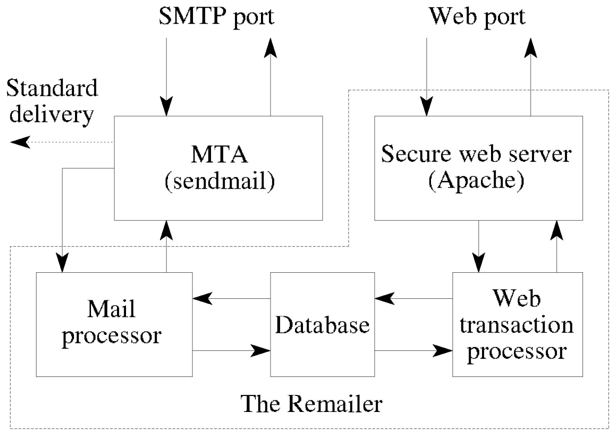


Figure 1: Remailer configuration

STRUCTURE OF THE REMAILER

The Remailer has been programmed in Tcl (Ousterhout 2001) and installed under Linux as an extension to sendmail (Costales and Allman, 2002)—see Fig. 1. All E-Mail addresses (aliases) created by the Remailer necessarily fall into the domain handled by the MTA. For example, the MTA domain of our prototype installation is `sheerness.cs.ualberta.ca`, and all addresses that can be assigned within that system are of the form *username*`@sheerness.cs.ualberta.ca`. Thus, the Remailer allocates solely the *username* component, with the domain part being determined by the MTA.

The Remailer's database stores records describing subscribers and their aliases, as well as some other information needed to implement transparent replies (described further in the paper). To sign up, a new subscriber has to specify his/her permanent E-Mail address, which will be used as the (globally unique) user Id, and a password. In response to this step, the Remailer sends to the provided permanent address a random secret code that must be entered by the subscriber once, at the first sign-on, to validate the new account.

ALIASES

A regular alias (functionally equivalent to an E-Mail address within the MTA's domain) is created manually by the subscriber using a dynamic Web form. Its most important attribute is the name, i.e., the *username* component of the E-Mail address. The subscriber has three options: to let the Remailer generate the alias name automatically, to specify a prefix and let the Remailer complete the name, or to provide the complete exact alias name. Alias names generated by the Remailer are random and obfuscated (such that they cannot be automatically targeted by spammers), although not impossible to memorize. In all cases, the Remailer makes sure that the name of the new alias is unique within the domain.

The remaining attributes of an alias include the expiration date, the message count, the pattern set, and the magic phrase. The *expiration date* indicates when the alias should expire. It can be specified as an explicit calendar date or as the number of days from now. The *message count* puts a limit on the number of messages that can be received on the alias. Unless this attribute is `infinite`, the specified number will be decremented by one after every message reception, and when it gets down to zero, the alias will become inactive and unusable.

PATTERN SETS

An alias provides room for three sets of patterns to be matched against the sender's address (the "from" patterns), the message subject (the "subject" patterns), and the message body (the "body" patterns). Within each set, there are *positive* patterns, i.e., ones that should be matched, and *negative* patterns, i.e., ones that should not be matched by the corresponding message component. Each pattern is individually ranked with an integer number between 1 and 9 and can be of one of the following four types: **substring** (representing directly a piece of string that must occur literally as a substring in the matched text), **exact** (representing an exact sequence of characters that must occur as a word in the matched text), **wildcard** (every asterisk within the pattern matches an arbitrary sequence of any characters), and **regexp** (a regular expression in its Tcl flavor).

The pattern matching test of an incoming message works as follows. Let m_{pos} denote the maximum rank of a matched positive pattern or 0, if the message does not match any of the positive patterns (including the case when no positive patterns are defined at all). Similarly, let m_{neg} stand for the maximum rank of a matched negative pattern or 0, if the message does not match any negative pattern. If $m_{pos} \geq m_{neg}$, we say that the message has passed the matching test with the rank of m_{pos}. Otherwise, the message has failed the test with the rank of m_{neg}. A message must pass the matching test to be delivered, unless its subject line contains the *magic phrase* (see below).

THE CHALLENGE

In some circumstances, when a message arriving at an alias is returned to the sender, the subscriber may want to challenge the sender to re-submit the original message inserting a "magic phrase" (of subscriber's choice) into the subject line. An incoming message that contains the correct magic phrase will be forwarded to the subscriber, and, as a side effect, the sender's address will be added to the positive "from" patterns associated with the alias (so that subsequent messages from the same sender will be delivered without problems).

One more attribute related to the magic phrase is the preamble to precede a bounced message and explain the challenge to the sender. One idea, to be implemented in the next version of the Remailer, is to send the magic phrase as an image. The image will be randomly disturbed in a way rendering the encoded text unrecognizable by a program but clearly legible to a human being.

QUICKCODES

A quickode can be viewed as a named template from which new aliases can be created automatically, without an explicit action of the subscriber. An alias created this way is called a *quick* alias.

Except for a slightly different interpretation of the expiration date, quickcode attributes have the same meaning as the corresponding alias attributes. When an alias is created from a quickode, those attributes are simply copied to the alias. The name of an alias created from a quickcode is always generated automatically by the Remailer.

The expiration time of a quickode can be `infinite`, or it can be an integer number of days. In contrast to the corresponding alias attribute, the subscriber cannot indicate here a specific date on which the quickcode is supposed to expire. This is because quickcodes (being alias templates rather than aliases themselves) never expire. When an alias is created from a quickode whose expiration time is finite, the expiration date of the alias is set to current time plus the indicated number of days.

MASTER ALIASES

The following attributes of a master alias: message count, expiration date, pattern set, have the same meaning as for a regular alias, and refer to E-Mail messages addressed to the master alias itself. Such a message is not meant to be delivered to the subscriber but represents a request for a temporary quick alias. To this end, the procedure of ranking and accepting such a message determines whether this request should be granted.

One required extra attribute of a master alias is the quickode to be used as the template for creating quick aliases in response to incoming requests. Two more attributes are the *initial message count* and the *initial expiration time* to be assumed by the quick alias until it is *validated* by the subscriber.

Whenever a message arrives at a non-expired master alias, and is not rejected by the standard set of criteria, the following actions are performed by the Remailer: 1) A new quick alias is created based on the quickode associated with the master alias. 2) The parameters of the new alias are copied from the quickcode, except for the message count and expiration time, which are set from the corresponding *initial* attributes of the master alias. 3) The new alias is internally marked as "non-validated" and associated with the master alias responsible for its creation. The sender's E-Mail address is added (as an "exact" pattern) to the list of positive "from" patterns of the alias with the rank of 7. 4) The message is bounced to the sender with the challenge of the master alias. Intentionally, the preamble of this challenge explains that the message has to be re-sent to the newly created alias with the proper magic phrase in the subject line.

When a message is received on a non-validated quick alias (whose message count is still nonzero and which has not expired), the Remailer first carries out the standard pattern matching test. If the message passes that test, its subject line is examined for the magic phrase. If the correct magic phrase is present, the message is forwarded to the subscriber. The subscriber can validate the alias by responding to that message.

HANDLING REPLIES

In addition to forwarding legitimate messages to the subscriber, the Remailer also handles replies to such messages, in a way that gives the communicating parties a consistent perception of their identities without revealing the subscriber's permanent address. Notably, the Remailer carries out these tasks assuming no cooperation from the client (MUA) used by the subscriber.

Whenever an E-Mail message arrives at a non-master alias and is deemed acceptable, the Remailer reorganizes a bit its headers before forwarding it to the subscriber's permanent address. When it reaches the subscriber, the message appears as if it were sent by the Remailer itself, from an address that consists of the name of the receiving alias combined with another string, the so-called *message tag*. The address of its true sender, as well as the possible addresses of its other recipients, are removed from the standard headers (`From`, `Reply-To`, `Cc`), and presented in two nonstandard headers: `X-Originally-From` and `X-Originally-Cc`. Also, to make it more conspicuous, the original sender address is included as a comment along with the substituted sender address.

For example, suppose that a message with the following headers is received by the Remailer:

```
...
From: "Kris Kelvin" <kris_kelvin@excite.com>
Subject: I am a sample message
To: "Pawel" <pgaglzom@sheerness.cs.ualberta.ca>
Cc: "John" <jsmith@somewhere.com>,
     "Susan" <susan@cs.ualberta.ca>
...
```

where `pgaglzom` is an active alias. When the message is delivered to the alias's owner (`pawel@cs.ualberta.ca`), its headers will be changed to something like this:

```
...
From: "kris_kelvin@excite.com"
      <pgaglzom_bywtyhexd@sheerness.cs.ualberta.ca>
Subject: I am a sample message
To: <pawel@cs.ualberta.ca>
```

```
Cc: "jsmith@somewhere.com, susan@cs.ualberta.ca"
    <pgaglzom__bywtyhexd@sheerness.cs.ualberta.ca>
...
X-Originally-From: <kris_kelvin@excite.com>
X-Originally-Cc: <jsmith@somewhere.com>,
                 <susan@cs.ualberta.ca>
...
```

Note that when the recipient replies to that message, the response will be addressed to pgaglzom_bywtyhexd@sheerness.cs.ualberta.ca and, if the reply is to "all," carbon-copied to pgaglzom__bywtyhexd@sheerness.cs.ualberta.ca, with both copies being intercepted by the Remailer. Based on the alias identifier (pgaglzom) and the message tag (bywtyhexd), the Remailer will address the response to the proper recipients and also remove the permanent address of the sender replacing it with the alias.

As a side effect of the feature discussed above, the subscriber can easily tell which particular alias was responsible for delivering every message reaching him/her through the Remailer.

TRANSIENT REQUESTS

With quickcodes and quick aliases, it is possible to combine the operations of sending a message and creating an alias, while making the message appear to the recipient as if it were sent from that alias. To illustrate this feature, suppose that we want to send a message to sales@robmeblind.com and pass it through a short-lived and restrictive alias. For this purpose, we need a quickcode with some restrictive attributes, e.g., message count = 1, expiration time = 2 days. Assume that this quickcode is labeled qq2347. We send the message addressing it to remailer@sheerness.cs.ualberta.ca and putting the following string at the beginning of its subject line: +qq2347 sales@robmeblind.com+. The Remailer will create a new (quick) alias based on the indicated quickcode and redirect the message to sales@robmeblind.com using the new alias as its sender address. The special string will be removed from the subject line, and whatever remains will be passed on as the subject of the forwarded message.

The command passed to the Remailer in the subject line can be generally quite elaborate. The more arcane possibilities cover redefinitions of practically all attributes of the new alias, including patterns confining the sender's domain and/or matching fragments of the subject line. While they may not appear extremely friendly from the viewpoint of a user accessing the Remailer via a standard MUA, they hint at the desirable features of a Remailer-aware MUA that would make the parametrization of transient quick aliases much easier.

SALVAGING THE EXISTING INFRASTRUCTURE

The primary concern of an institution contemplating a transition to the new E-Mail paradigm represented by the Remailer will be the fate of the existing infrastructure of E-Mail addresses that have been heavily harvested and put onto numerous lists sold to unscrupulous spammers. At first sight, there seems to be no alternative to scrapping them and starting the game from scratch. Fortunately, the open-ended character of the Remailer offers a reasonable solution to this problem.

If the Remailer is installed within the E-Mail domain of the existing infrastructure, the old addresses can be re-declared as master aliases. This will let them retain their official and traditional publishable status, while freeing them completely of unsolicited E-Mail, no matter how heavily they have been abused in the past. Owing to the fact that the Remailer's MTA need not give up its traditional duties, this solution can be adopted gradually, as the users become convinced that they really want to switch to the new type of service. Those of them who for one reason or another will be reluctant to subscribe to the Remailer will be able to continue using their old addresses exactly as before.

Even if the Remailer is installed in a different E-Mail domain, the old (possibly harvested) addresses can still be used as permanent addresses for the Remailer. To de-spam them, the users can deploy aggressive filters, e.g., blocking all incoming messages except for the ones arriving from the Remailer. Such messages are easy to identify through the so-called *filter cookie*. If declared by the subscriber, the cookie string will be presented as a non-standard header (labeled X-Filter-Cookie) in all messages sent by the Remailer to the subscriber. In a slightly more refined setup, the filters may additionally let through all local E-Mail, i.e., originating and entirely passed within the domain of the local organization, which property is usually easily verifiable through a superficial examination of message headers (e.g., see http://sheerness.cs.ualberta.ca/RabidFire/). This is how one of us (PG) handles all E-Mail, personal and business alike, enjoying life without spam.

RELATED WORK

A publicly available commercial service somewhat reminiscent of our approach is available as Spamex (http://www.spamex.com/). However, an alias acquired from Spamex cannot be restricted in time or confined to a narrow population of senders. Besides, Spamex offers no tool similar to our master aliases that could be used as safe publishable points of contacts for its subscribers. Although Spamex service is useful in many circumstances (for incidental short-lived contacts with untrusted parties), it cannot be reliably used for more serious and long-lived communication, as the only remedy to abuse detected by the subscriber is alias revocation.

Spamex implements transparent replies through aliases, using a method somewhat similar to the one used in our Remailer, but it does not handle group replies, i.e., those addressed to the 'Cc' recipients. A group reply through Spamex is going to reveal to those recipients the permanent address of the subscriber, while hiding the alias. This is because Spamex only archives information about the senders of the re-mailed messages, its official purpose being the possibility to track down abuse.

A solution aimed at associating attributes (policies) with E-Mail aliases has been recently proposed in (Ioannidis 2003). With that approach, the attributes are encoded in the alias itself, which is then encrypted to make the policy tamper-resistant. The primary advantage of this scheme is the state-less and memory-less processing at the MTA, which requires no database to keep track of the aliases and their attributes. However, there is no way to use a system based on state-less aliasing without a special (aware) E-Mail client (MUA) that either generates the aliases by itself or knows how to reply to a message arriving on an alias. A state-less system is also unable to provide the type of service offered by master aliases in our Remailer. Moreover, the state-less approach limits the population and types of alias attributes and makes it impossible to modify them once the alias has been created and handed out. In particular, dynamic attributes, (like the message count in our system), cannot be accommodated by this method.

FINAL REMARKS

The amount of unwanted junk E-Mail in the Internet has been increasing steadily since the network went global, and now seems to be approaching a critical point after which a drastic solution will have to be adopted. Our proposed paradigm offers one possible remedy to this problem.

One of our most serious concerns was the feasibility of smooth deployment of the new service and its seamless coexistence with the legacy infrastructure. Our experience with the prototype implementation of the Remailer strongly suggests that despite its obvious shortcomings (typical of a prototype), the system is viable, reliable, effective, and safe.

We do not share the worries expressed in (Ioannidis 2003). that the database of a state-based aliasing system will have a tendency to "grow without limits." Even if the database of the Remailer has some natural tendency to grow as new users subscribe to the service, that tendency is considerably less pronounced than in a typical database of a public E-Mail service, which tends to be heavily polluted with throw-away mailboxes. Although the Remailer can easily generate large numbers of aliases, practically all those aliases are short-lived and disappear after a while without a trace. On the other hand, an alias explicitly set up as long-lived is likely to be treated as a serious E-Mail address not to be discarded, abandoned, or forgotten. This is especially true if the Remailer service is deployed in a serious environment, e.g., at a scale of one institution, rather than being offered as a cheap and "spamvertised" source of disposable aliases for everybody.

Our rough and pessimistic estimate indicates that the stable amount of database space per subscriber, under a heavy usage pattern, is much less than 0.5MB. With contemporary pricing of disk storage, this amount is hardly prohibitive. Its bulk (more than 90%) is used to accommodate the archived header extracts needed to implement transparent replies from unaware MUA's. As the system will evolve toward more collaborative MUA software, the need for this space will be reduced and eventually eliminated.

The general idea of a challenge-response protocol for establishing the first contact with an E-Mail recipient is sometimes criticized as cumbersome, unfriendly, or impolite. Interestingly, one of us (PG) has had a chance to observe how the attitudes of people toward filter challenges evolve with time, or rather with the amount of spam that those people are forced to dig through every day. A few years ago, a message bounced with a challenge would occasionally meet with an objection from a mildly upset sender. These days, instead of objections, we are receiving words of appreciation and requests for our spam prevention tools. To put it in the right perspective, there is nothing wrong about a politely worded challenge after which the correspondence becomes absolutely noiseless, spam-less, and smooth.

REFERENCES

[1] Costales, E. and E. Allman. *Sendmail.* O'Reilly and Associates, 2002.

[2] Ioannidis, J. Fighting spam by encapsulating policy in email addresses. In *Proceedings of NDSS'03*, San Diego, CA, Feb. 2003. To appear.

[3] Klensin, J. Simple Mail Transfer Protocol. Request for comments 2821, Internet Engineering Task Force, 2001.

[4] Ousterhout, J. *Tcl and the Tk toolkit.* Addison-Wesley, 2001.

[5] Postel, J. Simple Mail Transfer Protocol. Request for comments 821, Internet Engineering Task Force, 1982.

Evaluating the reliability of commercially available biometric devices

V. Dimopoulos, J. Fletcher, S.M.Furnell

Network Research Group, Department of Communication and Electronic Engineering
University of Plymouth, Plymouth, United Kingdom
E-mail: info@network-research-group.org
Web: www.network-research-group.org

KEYWORDS
Biometrics, authentication, security.

ABSTRACT

The need for secure and accurate authentication at the entry point of every network is becoming increasingly evident. This paper considers commercial-grade biometric technologies and investigates whether their performance is sufficient to warrant their use as replacements for current user authentication methods. For this purpose, a variety of commercial-grade biometric devices were tested and their characteristics (mainly their accuracy) were assessed. From this assessment, optical fingerprint technology proved to be generally the most reliable while other techniques (such as thermal fingerprint scanning and voice verification) demonstrated good performance characteristics; although they were subject to a number of false rejections, they still did not tolerate any impostor access the network. Keystroke analysis, face recognition and dynamic signature verification all displayed average performance characteristics, with occurrences of false rejections as well as false acceptances.

INTRODUCTION

The rapid evolution of e-commerce and of the Internet has meant that networks previously inaccessible by most people can now be accessed on-line. This has introduced a variety of new security vulnerabilities, and increased problems like hacking, identity theft, malicious impersonation and data theft [Furnell 2001]; subsequently it introduced a requirement for better network security.

First in the procedure of securing a network is strengthening the point of entry by providing secure and reliable authentication. There are three common ways to verify the identity of an individual who is attempting to access valuable company assets, sensitive data and private/personal information [Smith 2002]. The first is by verifying something that ideally only the legitimate user **knows**, for example a password or a PIN. However this sort of information is often vulnerable to becoming compromised, as it can be guessed, shared or written down. [Monrose et al 1999]. A second method of authenticating someone's identity is by making use of something that ideally only this user **has** possession of. This can be for example an electronic identity card or some other form of physical token. This method however still has the disadvantage that the token can be shared or even stolen, and non-legitimate users can gain unauthorised access to the protected resources. Finally, the third method of authenticating someone's identity is provided by biometrics. This method uses something that the user **is** to verify an identity. As their definition indicates, *biometrics is the automated use of behavioural and physiological characteristics to determine user identity* [Dye et al 2001]. A person's behavioural characteristics are the way he signs, talks or even types a sentence, whereas the shape and features of a users' face, fingerprint or eye, hand can be categorised as physiological characteristics. The concept of identifying someone by utilising these characteristics is not new; human beings have the ability to identify each other by simply looking at a face or hearing a voice, while there are reports of signatures and fingerprints being used as a means of identity verification from as early as the 14th century [Novell 2001].

The basic principles of operation are similar for all biometrics. There are essentially two stages: the first is the registration of a user with a device and the creation of an individual biometric template and, from then on, the second stage is the verification/identification of a users' claimed identity, by comparing an acquired sample against the template already held.

Using a biometric based authentication method introduces a number of advantages; in verification systems, none of the measured characteristics can be stolen, forgotten or shared, resulting in an approach that is theoretically more secure, and also more convenient, since the user does not need to remember long passwords.

However, despite these advantages biometrics have yet to become widespread in commercial everyday applications. In fact according to a CSI/FBI survey biometrics are only being used in 10% of computer systems [Power 2002]. The aim of this paper is to investigate the reasons why, in spite of all their inherent advantages, biometrics have been largely confined within the environment of labs and test facilities. This is established by evaluating the performance criteria of the different biometric techniques in order to assess whether they can reliably find applications in authentication.

BIOMETRIC ASSESSMENT CRITERIA

There are many different biometrics techniques available today. Some are commercially available and others are still at a research level. The commercially available biometric products that were obtained and evaluated for the purposes of this investigation are based on the following techniques:

- Fingerprint Verification: This is a method of authenticating a person's identity by obtaining an image of their fingerprint.
- Facial Recognition: As the name suggests this technique authenticates someone by obtaining a picture of the persons' face and comparing it against a reference profile it has stored in its database to determine if there is a match or not.
- Keystroke Analysis, is a biometric technique that analyses the way a user types words on a keyboard or keypad in order to identify characteristic rythms.
- Dynamic Signature Verification: this biometric method utilises distinctive characteristics in the way a user signs to authenticate a claimed identity.
- Speaker Verification: this technique is based on the analysis of certain unique characteristics of a person's voice that can be used to establish an identity.

There are several major criteria upon which biometric techniques and products can be judged. Primarily these are a technique's technical, operational and economic characteristics, as well as the level of support provided by the manufacturer [Polemi 1997].

The main technical characteristic of a biometric is its accuracy (i.e. the rates of false acceptances and false rejections that a device is capable of accomplishing) so that it can achieve the required security levels. The operational criteria are for it to be convenient to use (ease of use, speed of enrolment and authentication), acceptable to the public by not being invasive and independent of any influences to its performance by the environmental conditions. The economic aspect of a biometric is its purchase, licensing, installation and staff training costs. For this investigation however the devices under test were chosen to be from the lower-cost end of the market (under £150 per unit). What this selection aims to achieve is to make this assessment more realistic to the majority of companies that are relatively satisfied with the security they have achieved (with minimum costs) with passwords and do not have a huge budget to spend for increasing security.

INVESTIGATION OF BIOMETRIC TECHNOLOGIES

The experimental procedure followed in this investigation involved registering a set of users with the devices, who firstly attempted to access their legitimate accounts (so that the False Rejection Rate can be calculated) and subsequently attempted to 'fool' the devices and gain access to each others' accounts (so that the False Acceptance Rates for the techniques can be evaluated).

Fingerprint scanning (physiological)

Fingerprints are recognised to have distinctive characteristics, which has led to their use in various criminology applications throughout the world to identify individuals. There are several techniques for scanning a fingerprint to extract those features that can be used for successfully identifying an individual; namely, optical, silicon, ultra-sound, and thermal technology [IBG 2002]. For the purposes of this investigation tools based on the optical and the thermal were tested.

Among the advantages of this technique, is the fact that it is proved to be capable of reliable identification of individuals. Also, many of the fingerprint sensors are of small size and minimum power requirements, which allows them to be integrated into other hardware such as keyboards and mice. The biggest disadvantage for this technique is the requirement for the purchase of a specialised reader; this elevates the costs of deployment for finger scan.

Technical: Optical finger scan

Optical technology is the most mature and commonly used finger scan technology. It is based on the use of a scanner (essentially a camera) that records images of fingerprints held against a coated glass or a plastic platen, which are then transformed into a template by the underlying software. [Ndlangisa 2001]. For the accuracy assessment of this method seven users enrolled four fingers each and followed the experimental procedure described earlier. The results are in table 1.

Security level	False Rejection Rate	False Acceptance Rate
1 (Low)	4/105 = 3.8%	0/90 = 0%
2 (Medium)	4/105 = 3.8%	0/90 = 0%
3 (High)	5/105 = 4.7%	0/90 = 0%

Table 1: Accuracy of optical finger scan

Thermal finger scan

The second method evaluated is based on capturing a full size image of a fingerprint by sweeping the finger over a thermal CMOS sensor. This technique uses infrared to measure the minimal temperature differences between the ridges and valleys of the users' heated finger. [Smallback 2002]. Seven users registered four fingers with the device and the accuracy results for the experiment are displayed in table 2.

FRR	FAR
45/180 = 25%	0/1260 = 0%

Table 2: Thermal finger scan accuracy

Operational

As far as the operational aspect of fingerprint scanning is concerned, the optical method's ease of use, good speed of enrolment (approximately 15 seconds) and authentication (approximately 3 seconds) favoured it considerably. The thermal finger scanning method took a significant amount of time to enrol and authenticate a user because of unsuccessful scanning attempts (i.e. failure to acquire a sample). This made it slightly inconvenient to use. When tested under changing environmental conditions, i.e. various temperatures, the thermal sensor did not suffer any effect from temperature variations and the optical sensors' performance was not influenced when in a poorly lit environment. It was however degraded when under extreme light.

Facial recognition (physiological)

Another method of biometric authentication is by measuring and comparing unique features that exist in peoples' faces such as the distances between the eyes, the nose and the mouth. Several methods have been developed to scan a face, the most common being eigenfaces, feature analysis, neural networks and automatic face processing [facial-scan.com 2002]. For the purposes of this assessment, a product that uses a neural network technique was selected and tested.

Advantages of this technique are its ability to integrate with existing imaging equipment (e.g. webcams) and its ability to obtain the required images transparently i.e. without disrupting the user. As a disadvantage, however, its performance can be degraded by poor background lighting, as well as by the users' positioning against the camera.

Technical

Neural network systems use algorithms to determine the similarity between an 'on the spot' image of a user's face with the one that was stored during registration. Neural network systems are capable of learning and adjusting themselves over time according to which features they judge to be more effective for matching [Nanavati and Thieme 2002]. The product tested determines the degree of similarity between the acquired and stored images on a scale from 1 to 10. The security administrator can then set the threshold, essentially the minimum degree of similarity that would still allow a user access to the system, according to the requirements of an application. Table 3 illustrates the accuracy of this similarity rating. Essentially, if the strictness threshold is set at 5 (which is the manufacturers default setting), then the FRR for this method is 46% while the calculated FAR for this technique at the same security setting reaches 3%.

User	Attempt 1	Attempt 2	Attempt 3	Attempt 4	Attempt 5	Attempt 6	Attempt 7	Attempt 8	Attempt 9	Attempt 10
A	3.82	5.67	5.21	4.17	7.86	6.21	7.12	5.65	4.69	4.89
B	5.96	7.43	5.32	2.35	5.61	5.23	3.34	6.16	2.72	4.63
C	7.52	6.31	8.14	4.23	3.24	3.51	4.84	2.56	6.51	5.42
D	5.83	5.39	7.25	4.59	3.32	5.38	7.31	3.46	5.92	5.83
F	3,76	4.69	6.28	4.62	4.83	5.93	2.89	8.12	4.35	4.17

Table 3: Degree of match between acquired and stored sample as evaluated by the facial scan device

Operational

When evaluating the operational aspects of this technique, it demonstrated a good ease of use characteristic, an average speed of enrolment (approximately a minute) and good speed of authentication (around 15 seconds) although there was some negative feedback by those subjected to the test on the functionality of its interface with the user. When tested with the users wearing glasses or growing facial hair the technology appeared to be unaffected and displayed fairly similar results. Varying background lighting however did have an effect in increasing the devices' False Rejection Rate. This device was also fooled by a photograph and granted access to the impostor almost 50% of the attempts.

Keystroke analysis (behavioural)

This technique is based on the concept that every user types characters on a keyboard or keypad in a distinctive way. Consequently it verifies a claimed identity by analysing the users' typing patterns. There are several data acquisition techniques, and different typing metrics, upon which keystroke analysis can be based. Specifically it can be static at login, periodic dynamic, continuous dynamic, keyword specific and application specific [Dowland et al 2002]. The device under testing in this investigation is based on the static at login approach (which means that the technique is applied when users enter their username and password, and the system looks not only at what they typed, but how they typed it). The advantages offered by this technique are firstly the combination of the knowledge of secret information (password) with a biometric to increase the security levels, and secondly that there is no need for the purchase of any additional costly hardware. A disadvantage is that it does not improve user convenience since the user still has to remember this secret information, and also that authentication occurs as a one-off judgement (unlike the case of the continuous dynamic technique for example).

Technical

The keystroke analysis product under test offers the option for the administrator to set the security level (from 1 to 10) that is most appropriate for a specification. In theory, setting the security level higher should reduce the false acceptance rates but simultaneously increase the false rejection rates. The recommended minimum password length to be used for optimal results is eight characters. To assess the false rejection characteristic for this technique, seven users were enrolled and they all attempted to access the system ten times on each security setting and for various password lengths as illustrated in the table below:

Password length **Security setting**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0%	0%	0%	10%	0%	0%	10%
2	0%	0%	10%	10%	0%	10%	10%
3	0%	0%	20%	20%	0%	0%	0%
4	0%	10%	20%	20%	10%	0%	10%
5	10%	30%	30%	30%	30%	40%	30%
6	10%	30%	20%	30%	40%	20%	40%
7	30%	40%	30%	30%	40%	50%	40%
8	40%	50%	50%	50%	40%	70%	40%
9	50%	70%	60%	80%	50%	70%	50%
10	70%	90%	70%	70%	70%	90%	80%

Table 4: Keystroke analysis FRR

The calculated results for the False Acceptance Rates of the keystroke analysis technique are displayed in table 5. From these last two tables it can be gathered that when operating at the default security level 5, the average False Rejection Rate for this device is 28.5% while the False Acceptance Rate is 14.7%.

Password length **Security setting**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	93%	80%	80%	86%	83%	72%	60%
2	91%	53%	68%	51%	78%	53%	51%
3	83%	23%	51%	35%	58%	33%	18%
4	71%	11%	25%	20%	25%	15%	8%
5	66%	5%	6%	5%	15%	3%	3%
6	48%	0%	2%	0%	6%	0%	2%
7	23%	0%	0%	0%	3%	0%	0%
8	11%	0%	0%	0%	0%	0%	0%
9	5%	0%	0%	0%	0%	0%	0%
10	5%	0%	0%	0%	0%	0%	0%

Table 5: Keystroke analysis FAR

Operational

From the operational point of view, this technique is characterised by low speed of enrolment, which can be from 3 to 5 minutes, as it requires the user to re-type the username and password several times so that the device can 'learn' the users' typing patterns. The speed of authentication is exclusively dependant on the users' typing speed and length of username and password. This technique's performance does not suffer during environmental changes due to its software-only nature but factors such as fatigue and injury could affect the user's performance.

Dynamic signature recognition (behavioural)

This type of technology is based on authentication through the analysis of distinctive characteristics in someone's' writing, and particularly the way a user produces a signature [The biometrics institute 2002]. It is one of the most widely acceptable biometrics since the majority of users are already accustomed to using their signature to authorise transactions and to verify their identity.

The examination of a user's signature is achieved either by the statistical analysis of characteristics such as the duration, pressure and acceleration of the signing or by a sequential method where the signature is uniformly divided and each piece is examined individually.

The products based on this technique can be either pen based (the mechanism that captures the information is a specialised pen) or tablet based (a writing tablet with a special surface that collects the data). This technology can find applications for accessing personal computers, PDA and for authorising transactions over the Internet [CIC 2002]. While the advantages of this technique can be considered to be its low price and availability for direct purchase and download over the Internet, disadvantages are that users can find it inconvenient to sign accurately on a pad and that false rejection rates increase with inconsistent signatures.

Technical

For this investigation a tablet was used as the acquisition mechanism for a verification program that analyses both the sequential stroke patterns and the timing elements of a signature. The accuracy results for the false acceptance characteristics came after enrolling eight users and three additional "simple" signatures and attempting to forge the legitimate signatures twenty times and are illustrated in table 6. The average rate of false rejections according to these results is 2.5%, and the average percentage of false acceptances is 2.5% as well.

Signature samples	User A Posing as	FAR	User H Posing as	FAR	User D Posing as	FAR	User G Posing as	FAR
A	-	-	A	20%	A	0%	A	20%
B	B	5%	B	0%	B	0%	B	5%
C	C	0%	C	0%	C	0%	C	5%
D	D	0%	D	0%	-	-	D	0%
E	E	5%	E	0%	E	0%	E	10%
F	F	0%	F	5%	F	0%	F	0%
G	G	0%	G	0%	G	0%		-
H	H	0%	-	-	H	0%	H	5%
	Simple 1	45%	Simple 1	100%	Simple 1	15%	Simple 1	80%
	Simple 2	85%	Simple 2	90%	Simple 2	35%	Simple 2	60%
	Simple 3	90%	Simple 3	90%	Simple 3	45%	Simple 3	50%

Table 6: False acceptance assessment for signature verification

To observe the false rejection characteristics of this technique six registered users attempted verification twenty times. The results are exhibited in table 7 below.

User	Successful Attempts (out of 20)	FRR
A	20	0%
B	20	0%
C	20	0%
D	19	5%
E	20	0%
F	19	5%

Table 7: FRR for signature scan.

Operational

The ease of use of dynamic signature technique is good. Moreover it has a good speed of enrolment (typically 30 seconds) and authentication (around 10 seconds). Unfortunately, the performance of this biometric is effected to a great extend by all the elements that prevent a user from signing properly, for example hand injuries and tense emotional states as well as the positioning of the signing pad.

Speaker Verification (Behavioural)

Another biometric authentication technique based on a behavioural characteristic is speaker verification. Since humans can distinguish each other by their voices, this suggests voice has distinctive characteristics. This is the concept behind voice recognition authentication techniques. Voice verification biometric products are essentially software that use a standard microphone (or telephone for certain applications) as the medium to obtain samples through. Among the distinctive features utilised by the various speaker verification products to authenticate users are the fundamental frequency and pitch of a voice, the short time spectrum of speech, and the formant frequencies [Nanavati and Thieme 2002].

Among the advantages of using voice as a means of authenticating individuals is the nature of the hardware necessary for acquiring the voice samples. The potential to use existing microphones and telephone devices makes it suitable for a large variety of applications including, for example, its integration in call centres. This technique also has the potential of operating without the users being aware. A disadvantage is that the sample acquisition and the identity verification processes can be disrupted by environmental effects such as the background noise. Two different speaker verification products were obtained and tested for the purposes of this investigation both using proprietary algorithms to generate and compare templates (voiceprints).

Technical: Speaker verification product A

The first speaker verification device that was tested, offers the administrative option to set the desired levels of FAR over FRR. A series of tests was performed on every security level available by enrolling six users, who then attempted authentication ten times on each setting. Table 8 bellow contains the calculated false acceptance rates for this product.

Selected FRR Level	FRR Achieved
0.28	1.6%
0.4	3.3%
0.55	3.3%
0.73	0%
0.93	5%
1.19	0%
1.51	0%
1.84	6.6%
2.24	5%
2.76	3.3%
3.43	3.3%

Table 8: comparison between set and measured FRR

When investigating its rate of false acceptances, this product proved to be rather accurate since there were no occurrences of such an event apart from when it was run on the lowest security levels, where the device would grant access to literally everyone.

Speaker verification product B

With six enrolled users attempting access to their legitimate accounts ten times each, the calculated False Rejection Rate for this product is illustrated in table 9.

User	Number of attempts	FRR
A	10	20%
B	10	0%
C	10	20%
D	10	20%
E	10	10%
F	10	0%
Total	60	11.6%

Table 9: FRR for speaker product B

When assessing the devices False Acceptance characteristic, in 450 attempts to fool the product by having 6 users attempting to gain access to each other's accounts there were zero occurrences of a false acceptance.

Operational

Evaluating the operational attributes for the speaker verification biometric technique, it was found to have good ease of use. The speed of the enrolment is generally low (about 4 minutes) since the process of the user having to repeat a phrase several times so as to train the device is time-consuming. However the authentication process only lasts a few seconds. Regarding the environmental effects upon the operation of the device, it was found to be unaffected by small variations of the background noise levels but, as mentioned earlier, performance is degraded critically under extreme noise conditions.

DISCUSSION

Even though behavioural biometrics did not prove to be as accurate as one of the established physiological techniques (fingerprint), the nature of the features extracted by the behavioural biometrics makes the authentication process more acceptable and less invasive. However the process of enrolment for the behavioural methods is generally more time consuming and complicated, since the devices need training, they can sometimes be perceived as not being user friendly [Silverman 2001]. Behavioural biometric techniques are also cheaper to deploy in terms of the price of the hardware required. Nevertheless, since what most companies would look for in an ideal authentication method would probably be accuracy for a reduced cost, the most accurate techniques are the physiological techniques that also require expensive hardware. During the assessment of the techniques, finger scan displayed high levels of accuracy, especially when using the optical method. Even the thermal scan that had many occurrences of false rejections still did not allow any impostors in. Thus, even if it increased the inconvenience for the user, it still did not put any valuable assets at risk. The keystroke analysis method displayed minimum FAR only above security level 6 with over 6 letter passwords while the FRR at the same security level remained as low as 20% which could be tolerable in the sense that legitimate users might get asked to re-login once in every five attempts, but many people are used to this anyway as a result of mistyping their passwords. Moreover, this technique has the advantage that it combines secret information with the biometric to increase security in the case that a user's password is compromised. Thus it can tolerate a certain number of false acceptances since it is very likely that only the legitimate user will have the knowledge of the password. As the results from this technique demonstrate, when the device's 'strictness' threshold is set to 5, the legitimate user is being rejected approximately 46% of the time. This is a very high average. It is likely that more sophisticated products would provide very different results but as was explained earlier, this investigation is based on the low price, commercially available products. Signature verification was tested to have lower FRR than most of the techniques, but its false acceptance characteristic would probably make it an unacceptable solution for many applications. When the technique was tested with very simple signatures, it was established that people with very basic signatures would find that impostors could easily forge them. Finally, for the voice verification biometric technique, there were no occurrences of any false acceptances, and a small but noticeable percentage of false rejections. Nevertheless, the drawbacks of this technique that make it less favourable from some of the others is the long period it takes for a user to train the program and the fact that both the products tested were significantly more expensive than any other of the biometrics obtained for the purposes of this investigation. Table 10 summarises the results of the assessment.

Product	Lowest **FRR**	Average **FRR**	Highest **FRR**	Lowest **FAR**	Average **FAR**	Highest **FAR**
Optical fingerprint	3.8%	4.1%	4.7%	0%	0%	0%
Thermal Fingerprint	25%	25%	25%	0%	0%	0%
Keystroke analysis	2.8%	28.5%	40%	3%	14.7%	66%
Dynamic Signature	0%	2.5%	5%	0%	2.5%	20%
Voice verification	1.6%	5%	3.3%	0%	0%	0%
Voice verification	11%	11%	11%	0%	0%	0%
Face recognition	30%	46%	70%	0%	3%	10%

Table 10: Summary of the assessment results at the default security level for each device

Technique	FAR	FRR
Face Recognition	0.25%	25%
Fingerprint scan (chip)	0.025%	10%
Fingerprint scan (chip) (Same reader as previous but with different software)	0.003%	6.5%
Fingerprint scan 2 (optical)	0.15%	22%
Speaker Verification	0.012%	12%
Signature Verification	0.4%	0.7%
Keystroke Analysis (Bleha et al)	2.8%	8.1%
Keystroke Analysis (Joyce and Gupta)	0.25%	16.67%

Table 11: Accuracy results reported in other research papers

The results of this investigation suggest that biometrics are not as accurate as they have been reported in previous research papers that evaluated the same techniques. [Mansfield et al 2001] performed an assessment of some of the major biometric techniques, namely face, fingerprint and voice scan. According to the results from their experiments, which were held under 'normal office conditions' with participants from both genders and all age groups, the measured performance of these techniques with the decision threshold set at the default (optimal) level indicated fingerprint to be more reliable both in terms of false acceptances as well as false rejections. More analytically the results are shown in table 11. According to the results of a similar report published by the Biometrics Consortium [biometrics.org 1995], when tested, signature verification technologies achieve accuracy rates of 0.7% false rejection rate and 0.4% false acceptance rate, while in an actual implementation the measured rates of false rejections were 0.1%. Finally, several papers. [e.g. Bleha et al 1990, Joyce and Gupta 1990] report keystroke analysis as being able to achieve the error rates that are displayed in table 11 as well. There are many reasons behind this dissimilarity of the results, the main being that this investigation assessed the accuracy of the commercially available products that are available for the enterprises that are not willing to invest a large amount of money to upgrade their existing authentication system. Secondly, this assessment was not performed under the ideal lab testing and operating conditions but under conditions that were varied in order to test the relative stability of the devices operation. Moreover the users that assisted with the evaluation of the products did not have any experience with such technologies and concepts, but were chosen to be ordinary PC and network users, as would be the case with any real-life implementation of such a technique.

CONCLUSION

This study used a very small group of test subjects (users), which should have helped to control error rates, nonetheless significant error rates were still observed for some of the methods. An emerging area of biometrics that could produce products with significantly improved accuracy and reliability is multiple biometrics. The combination, for example, of fingerprint and face scan can boost security levels radically, while a combination of face and voice scan would improve accuracy while maintaining low invasiveness levels, since they can both operate without the knowledge of the user.

The results from this evaluation of the commercially available biometric products clearly do not represent the entire range of products available in the industry. They should however be considered by any security administrator looking to implement biometric authentication since they are results from the evaluation of the commercially available products. This assessment established that the majority of the low cost commercially available biometrics are not suitable for those applications that require high accuracy levels such as government or military use. They can however provide increased convenience and additional security in other cases.

REFERENCES

Bleha S., Slivinsky C., Hussien B., (1990) 'Computer-access security systems using keystroke dynamics', Transactions on pattern analysis and machine intelligence Vol 12, No 12, 1990.

Communication Intelligence Corporation (2002), ‘Enterprise solutions, implemented applications’ www.cic.com

Dowland P.S. Furnell S.M, Papadaki, M. (2002) ‘Keystroke analysis as a method of advanced user authentication and response’ Proceedings of IFIP/SEC 2002 17th International Conference on Information Security, Cairo, Egypt, 7-9 May.

Dye B. Gerttula J. Kerner J. O'Hara B. (2001), ‘An introduction to biometrics’ 20 November 2001 http://www.stanford.edu/~bjohara/ introduction.htm

Furnell S. (2001) ‘Cybercrime: Vandalizing the Information Society’, Addison-Wesley Publishing Company.

International Biometric Group (2002), ‘Optical – Silicon – Ultrasound’ http://www.ibgweb.com/ reports/public/reports/ finger-scan_optsilult.html

Joyce R., Gupta G., (1990) ‘Identity authentication based on keystroke latencies’, Communications of the ACM, Volume 33, February 1990.

Liu S., Silverman M., (2001) ‘A Practical Guide to Biometric Security Technology’ IEEE Computer society, January 2001 http://www.computer.org/itpro/homepage/jan_feb01/security3.htm

Monrose F., Reiter M., Wetzel S (1999), ‘Password hardening Based on Keystroke Dynamics’ Proceedings of the 6th ACM computer and communication Security conference.

Nanavati S., Thieme M.,Nanavati R. (2002) ‘Biometrics, Identity Verification in a Networked World’ John Wiley & sons Inc.

Ndlangisa N (2001). ’Biometric Authentication using fingerprints and evaluating fingerprint readers’, November 2001, www.cs.ru.ac.za/ research/g9610159/Documents/Writeup-Final.pdf

Novell (2001), ‘Overview of biometrics’ 1 July 2001, http://developer.novell.com/research/ appnotes/2001/july/01/a0107013.htm

Power R. (2002), ‘Computer Security Issues and Trends’, CSI/FBI Computer Crime and Security Survey, Volume 8 No 1, Spring 2002, https://wow.mfi.com/csi/order/publications.html

Polemi D. (1997) “Biometric Techniques: review and evaluation of biometric techniques for identification and authentication, including an appraisal of the areas where they are most applicable”, April 1997, http://www.cordis.lu/infosec/src/stud5fr.htm

Smallback, R. C. Jr., (2002) ‘Security Access using biometric fingerprint Technology’ 28 May 2002, http://www.biometritech.com/features/

Smith R. E. (2001), Authentication from Passwords to Public Keys, Addison-Wesley Publishing Company.

The biometric consortium (1995), ‘Electronic Benefits Transfer: Use of Biometrics to Deter Fraud in the Nationwide EBT Program’, 29 September 1995 http://www.biometrics.org/ REPORTS/OSI-95-20.html

The Biometrics Institute (2002), ‘Signature Recognition’, http://www.biomet.org/ signature.html

The Facial Scan Homepage (2002), ‘Primary Facial-Scan’ www.facial-scan.com.

MOBILE
AND
WIRELESS
TECHNOLOGY

MULTI-CLIENT COOPERATION AND WIRELESS PDA INTERACTION IN IMMERSIVE VIRTUAL ENVIRONMENT

Elisabetta Farella, Davide Brunelli, Luca Benini, Bruno Riccò
DEIS - Dept. of Electronics, Computer Science and Systems - University of Bologna.
Viale Risorgimento, 2 –40136 - Bologna - ITALY
E-mail: { efarella| dbrunelli|lbenini| bricco}@deis.unibo.it

Maria Elena Bonfigli
Vis.I.T. Lab - CINECA Supercomputing Centre
Via Magnanelli 6/3 – 40033 Casalecchio di Reno (BO) – ITALY
E-mail: e.bonfigli@cineca.it

KEYWORDS
Collaborative Virtual Environment, Palmtop Computers, Wireless Networks, Multi-User Interaction, Multimedia application, Real-Time Interactive System

ABSTRACT

Virtual Environments (VEs) present many advantages for Collaborative Work. Users are immersed in simulation reproducing real world situation, thanks to stereoscopic glasses and large projection walls, maintaining natural human communications, sharing visualization of the same environments and objects, exchanging discoveries, ideas and suggestions. A main challenge to enhance cooperation is providing real-time interaction in the VE and among people, through multi-user support, wireless devices and multimedia features. This paper describes the effort to achieve these goals introducing Palmtop Computers (PDAs) in Collaborative Virtual Environment as a platform of interaction with the VE, exploiting wireless networks and their increasing storage and processing capabilities and usability.

INTRODUCTION

Virtual Reality is evolving towards multi-user shared virtual worlds, an emerging trend on the Internet. This kind of three-dimensional graphical environment, usually inhabited by avatars, i.e. digital users, needs real-time multi-user interaction support. This trend is also coming from hugely successful 3D multi-player networking games, propelling the effort of development and enhancement of enabling technologies.

Technology evolution opens new possibilities for developing collaborative Immersive Virtual Environments (IVEs), which have huge potential not only for entertainment but also for a large number of professional applications (e.g. medicine, learning, training, etc.). In our work we focus on virtual heritage, as an evolution of our previous project (Benini et al. 2002), where the cooperation of archaeologists, virtual model developers, art experts is fundamental for a correct reconstruction of an ancient artifact or a building. We designed our system basing it on the idea that the mix of Immersive Virtual Reality (IVR) and portable terminals enhances cooperation in a multi-user workplace.

One of the most interesting features of VR is that it provides users with a realistic immersion in a 3D environment, experiencing presence in a real world situation. This goal is reached through the third dimension, which is more natural, if compared with 3D graphics on 2D displays, and through the stimulation of human senses not only with 3D visualisation but also sound, and if needed, touch (Ragusa et al. 2001) (Loftin 2001, Smith 2001) (Su et al. 2001).

Usage of PDAs in Computer-Supported Cooperative Work (CSCW) has particular relevance in single display groupware (SDG), where people are physically close to each other and share the same output display. PDAs are often preferred to laptop and other devices for their limited dimension, and, at the same time, their increasingly advanced features. Moreover they turn on instantly, batteries last for days and they connect very easily to larger computing devices. Furthermore PDAs are more powerful compared to devices like mobile phones or pagers. In fact collection, organization, interaction and storage of different kind of data is possible and simple.

As already stated, our work introduces PDAs in VEs to support collaborative work and shared visualization. This choice combines the advantages of immersivity and sense of presence with those deriving from use of PDA as a support to work cooperation. Displaying on PDAs a Graphical User Interface (GUIs) for interacting with VR is less invasive and avoids the risk of occluding the user view of the VE (Wloka and Greenfield 1995), instead of displaying a User Interface directly on the projection wall or using 3D widgets or virtual devices. To reach this goal we developed a Java client-server application for a multi-client VE. We applied our system to Virtual Heritage, Virtual Reality applied to Cultural Heritage. The field of application influences the choice of features, even if most of them are of general interest. In its highest level the application is a GUI, displayed on each user terminal that interacts with the VE. The GUI offers the possibility to access many tools and media (taking notes, surfing the

Web) thereby increasing the sense of immersion inside the VE.

The paper is organized as follows. In section 1 we first survey previous work, then we describe basic design choices and the enabling hardware. Afterwards we delineate system features and implementation details. Finally we present results of performance tests, concluding with future research directions.

1. RELATED WORK

The mix of portable terminals, physical or virtual, and VR technologies has been already explored in previous research studies (Fitzmaurice et al. 1993) (Wloka and Greenfield, 1995) (Lindeman et al, 1999) (Rorke and Bangay, 1999) (Watsen et al. 1999), (Hill and Cruz-Neira 2000). A recent work refers to the use of a wireless PDA to manage annotation on a 2D map representing the 3D Virtual Harlem projected in a CAVE (Park et al, 2001). At Siggraph 2002 two interesting works were presented. First VRAC researchers presented Tweek, a middleware tool offering an extensible Java graphical user interface (GUI) for communication with VR applications. Tweek runs on desktop computers, on palmtop computers in projection-based virtual reality systems, or in a three dimensional virtual space (Hartling et al, 2002). The second work is about the ARS BOX and the Palmist project, a PC-based VR System combined with a CAVE-like environment, where a handheld PC works as its interaction interface (Hörtner et al, 2002). This solution offers a low-cost system, compared to traditional ones, and, at the same time, it enhances interaction and contents presentation through separate management of three projection screens from a unique device, a handheld PC. A limitation is that the interface is not platform independent and only one person at a time can experience interactivity. The latter is the major limitation also for Tweek.

Use of PDAs in single display groupware (SDG) has been explored in different kind of work. The Pebbles project (Myers 2001) is a set of applications for connecting PDAs (Palm Pilot or Pocket PCs) to a main computer, such as a PC, allowing people to provide mouse and keyboard input through their PDAs without leaving their seats or to share a drawing program, that allows everyone to draw simultaneously. A similar example is M-Pad system (Rekimoto 1998), where multiple-users collaborate through PDAs and a shared digital whiteboard as a support to informal meetings. The handheld computer serves as a tool palette and data entry for the large whiteboard.

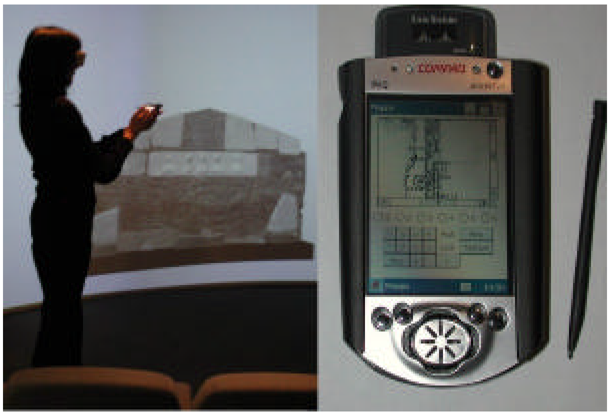


Fig.1 Interactions and PDA User interface for "Ancient Appian Way 3D Web Virtual GIS" project

As stated in the introduction, our work adds to previous researches the use of PDA in Collaborative Virtual Environment. We explored the mix of immersivity, sense of presence in Virtual Reality and computer support to collaborative work and shared vision, exploiting real-time multi-user interaction.

2. SYSTEM REQUIREMENTS

Our first step was the definition of the main design choices to guide the development of the interaction system. We chose as a priority the presence of more than a user at a time, real-time interaction, platform independence, and wireless connectivity. Moreover, a guideline was usability. This section describes our design choices in detail:

(i) **Multi-client communication**. A priority in designing our application is enabling multi presence on a networked IVE. This feature is of great interest in edutainment, entertainment and collaborative work. Therefore current work aims to support concurrent multi-user interaction for networked environment, providing exchange and collection of multimedia data (images, texts, voice, sounds, video/audio streaming). This feature will imply the development of peer-to-peer interaction among clients and scheduling protocols to organize the priority in controlling the VR exploiting the mix of using Java and RMI.

(ii) **Real-Time Interaction**. To provide the illusion of smooth continuous movement, visual and other stimuli have to be updated at regular intervals. In the visual domain this rate must be in excess of 17Hz. Moreover, to experience interactivity the user must be able to influence the application in a natural way. The time delay from user input until the application output response, otherwise known as latency, is critical. Two are the possible bottlenecks: the real-time rendering engine and the communication among the different system parts and actors. The satisfaction of real-time constraints for the 3D graphics rendering can be obtained through high-performance graphical libraries and VR toolkits and powerful graphics workstation with dedicated hardware components. On the other side, guaranteeing real-time performance in communication implies a trade-off among the number of clients, the amount of data exchanged per second, and the communication reliability.

(iii) **Platform independence**. It is important to support different kinds of platform on which the interface could be downloaded when entering the Virtual Theatre, and to give the chance to visitors to use their own handheld device (which could be a Pocket PC (PPC) or a laptop, and in future a cellular phone, of any brand with different hardware capabilities). The choice of the controls and interaction features for the interface is strongly influenced by the processor, the memory, and the multimedia capabilities of the hardware.

(iv) **Wireless connectivity**. Eliminating wires on the interaction devices helps to augment sense of presence and immersion in the Virtual World. We chose Pocket PC as preferred platform; because it is fully supported with expansion for different communication interfaces i.e. IEEE 802.11b High Rate standard GSM, Bluetooth or GPRS.

(v) **Usability**. Another important achievement is that the application implemented implies a short learning curve and low costs. That's why we chose a widespread and familiar platform such the Pocket PC. We argued about the importance of usability in Virtual Heritage field in a previous work (Benini et al. 2002).

3. SYSTEM FEATURES

It is possible to group the features implemented in our system under some basic tasks: interaction techniques; integration of tools; collaborators awareness; communication among clients.

Interaction techniques. Navigation in the virtual world, selection and manipulation of virtual objects, change in system status, are the main interaction techniques in the IVE. Obviously real-time constraints must be respected and interaction techniques should not affect the navigation smoothness or the effect of immediate reaction of the Virtual World to users requests. As an example, in our application navigation is provided both as active (guiding the motion directly through arrows) and passive transport (selecting the target point on a 2D map). Moreover, selection and manipulation (rotation, zoom, outlining details) of object are possible through menus or buttons, and both are synchronized with related text and images displayed on the PDA.

Integration of tools. The use of platforms such as laptop and palmtop computers helps the sense of immersion. The idea is to enable the users to access many different applications and media, without leaving the Virtual Environment (e.g. a text editor to take notes, a mail application or a browser). In fact many media are needed for a more effective collaboration, since each way to present and organize data has its own special benefits. In our application we exploit the Pocket PC multi-tasking feature. Opening external applications can be performed 'separately' from our interface or directly from a menu in the GUI.

Awareness of other clients. While clients are not actively communicating with each other, their awareness (i.e. their knowledge of collaborators actions, status and proprieties) can be obtained in different ways. For example, a map and some pointers can visually show where other users are in the shared IVE and, possibly, what they are doing. Alternatively, the use of effects such as reverberation, sounds and sound's levels (Yamazaki and Herder, 2000) can notify the users of different kinds of actions. In our implementation the spatial awareness of other clients is performed through a 2D map on the server and concurrently on the clients through colored dots representing each client. The dot's color gives information of what each client is doing at that moment.

Communication among clients. Another important feature in multi-clients environment is the exchange of data among clients. Requested data could be voice, sounds, music, video, graphical files, charts, images or text, i.e. basically anything useful to help the collaboration within the work environment.

In this direction we implemented a whiteboard. By using it the users can write notes, impressions, comments by PPC pen having them displayed both on his terminal and on the one of the client intentionally contacted, by peer-to-peer interaction.

Another example of communication among clients in our implementation is the request of VR control. This communication is performed through a client-server organisation. As a result, in our application, users can concurrently interact with a Map and with multimedia data and other tools (Pocket Word or Pocket Explorer as an example), but just synchronously with VR models or environment (turn-based) if only one is present at a time.

4. IMPLEMENTATION

The real-time interactive multi-user system is the combination of the Virtual Theatre at CINECA (fig. 2) and the portable terminals, i.e. the Pocket PCs connected through a wireless communication interface, the IEEE 802.11b High Rate standard wireless LAN interface working at 11 Mb/s.

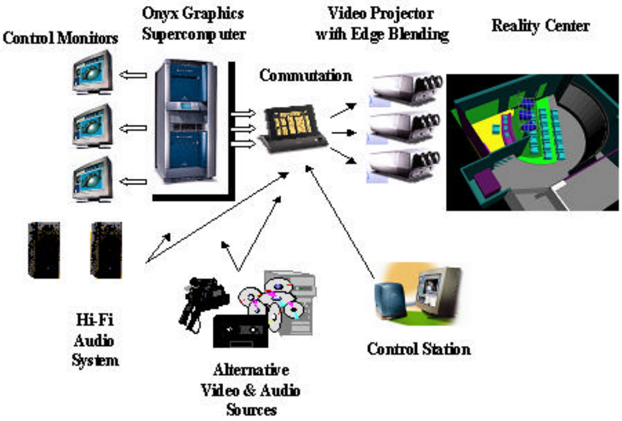

Control Monitors
Onyx Graphics Supercomputer
Video Projector with Edge Blending
Reality Center
Commutation
Hi-Fi Audio System
Alternative Video & Audio Sources
Control Station

Figure 2: Components of the Virtual Theatre at CINECA:

The current software implementation consists of a client component and a server component. Both, in their highest level, are Java applications exploiting distributed computing libraries.

At server-side some of the values, returned from clients through Remote Method Invocation, are given as parameters to different specific graphical functions (e.g. OpenGL Performer and Vega) that manage at server side the real-time visualization. The choice of Java technology was made as an evolution of previous work [Benini et al. 2002]. A first implementation of the interaction system provided the client side in Java and the server side in C, to make easy graphical function calls. Communication between them was realized through sockets. The limitation of sockets direct use was the low-level of the software architecture, which makes hard to manage a multi-user environment.

Our new solution was to exploit the rich set of libraries provided by Java. We chose Remote Method Invocation (RMI) technology suitable for distributed object computing. The greatest advantage of this choice is that details of communication between remote objects are handled by RMI; to the programmer, remote communication looks like a standard Java method invocation. RMI is supported by the JVM that runs on PPC. This choice implies the use of Java language also at the server side, implementing the call of the graphical C libraries through Java Native Interface technology.

To develop the application we used JavaTM 2 Software Development Kit (SDK), v.1.3.1. At both side a Java Virtual Machine allowed the running of the application: at client side the JVM is Jeode Runtime from Insignia Solution, optimized for WinCE platform. This JVM supports AWT classes, RMI and JNI technologies. At server side JavaTM Runtime Plug in v1.3.1 for SGI IRIX®

Our application uses both the two RMI mechanisms to obtain references to remote objects. The server registers its remote objects with RMI's simple naming facility, the rmiregistry, but also, in more than a case, the application pass and return remote object references as part of its normal operation.

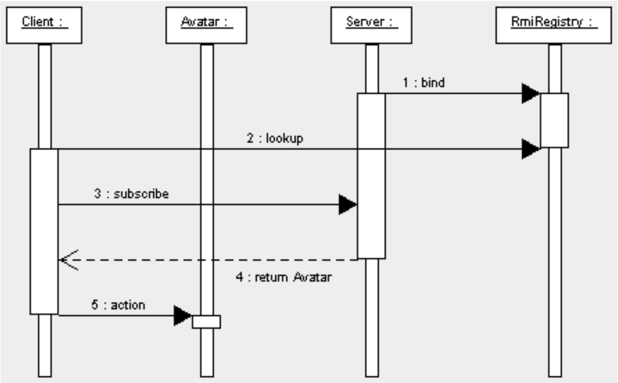


Figure 3: Start of application and subscription of the first client

In Figure 3 it is shown the start of the application at the server side and what happens when the first client enters in the IVE to interact with it. (1) The server calls the registry to associate (or bind) a name with a remote object (2) The client looks up the remote object by its name in the server's registry (3) and then invokes a method on it to tell to the server he is in, that is to subscribe itself; (4) the server returns to client a remote Avatar object reference, created by need. (5) From now on, the client can control the theatre and ask for actions on the VR (e.g. navigation, objects manipulation).

From (2) onwards those are the steps each client does entering the Virtual Reality application. The server assigns a unique identifier to each client when the client subscribes itself.

At user level what is shown by the application is a GUI, which in different cases can display a touch-sensitive 2D map of the 3D world, arrows-buttons, speed slide bar, menus. The UI on the client, realized with AWT classes, is designed to emphasize usability and short learning curve. Touching with the pen on the map, clicking on a button, moving the slide bar or selecting an item on the menu are all actions generating events handled by method locally or remotely.

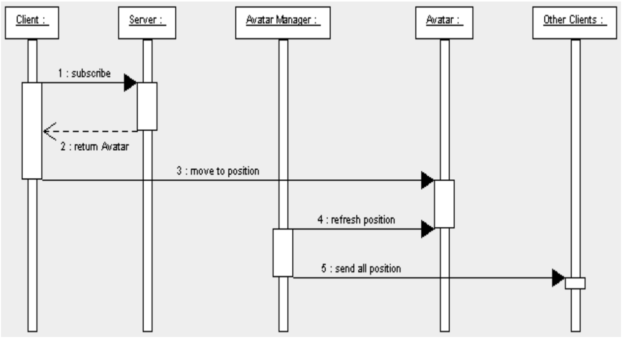


Figure 4 An Avatar moves

Fig 4 describes how a client request of action is processed and distributed to all the other clients. We refer as an example to the movement of a client on the 2D map, representing the 3D world, which is displayed on each client and on the server. Any time a client moves itself on the map, the motion is shown also on the map of all the other users.

Step 1 and 2 are the same of the previous figure, start of the client and subscription. Touching with the pen in a target point on the 2D map an event is generated and processed so that the coordinates of the new position are returned (3) to the Avatar Manager through RMI. The Avatar Manager periodically refreshes the position (4) of all the Avatars and after every refresh the new positions are notified to every client (5).

In this way even the client responsible for the motion request uses a remote method, not a local one, to move his avatar on the map. The consequence is that the avatar moves at the same speed on the maps displayed on each client.

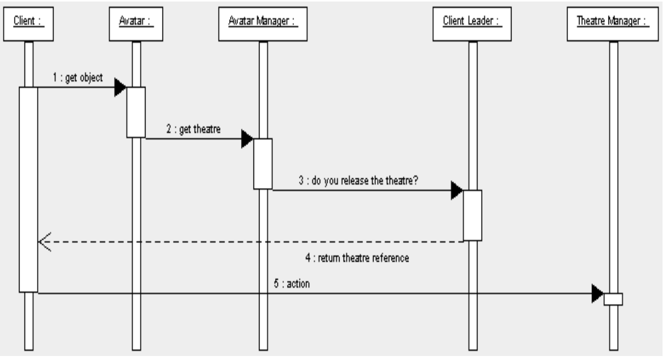


Figure 5 Control Request by a client

In Figure 5 a client asks for the control of the theatre, while it is leased to another client.

First, the client notifies his intention to take control of the theatre (1). The remote Avatar object communicates the request to acquire the control to the Avatar Manager (2), which monitors the status of all the Avatars and knows which client owns the control at a given time. The Avatar Manager asks to the current owner to release the control (3). Depending on the answer, the control is released and reassigned or simply maintained and the owner, old or new, can interact with the Virtual Reality.

The Java application at server-side is responsible not only for communication with clients and Theatre control management, but also for real-time visualization of the IVE in the Virtual Theatre and execution of user interaction requests. This duty is accomplished through Vega function calls and JNI technology. JNI enables to operate with applications and libraries written in other languages such as C, but also C++ and assembly.

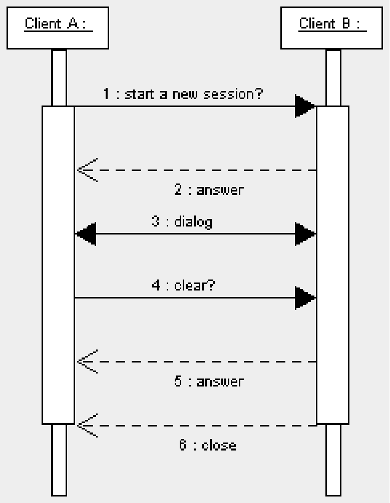


Figure 6 Peer to peer interaction

One of the features of our application is the ability to communicate between clients. As an example, figure 6 shows what happens when two clients exchange information through a whiteboard. The first client decides to communicate with another. He/She proposes a collaboration touching with the pen the 2D map where is located the avatar representing the client he/she wants to contact (1). This action implies a message on the terminal of the selected client, who answers (2). If the answer is positive a communication begins on the shared whiteboard (3). (Fig.7)

During communication one client can inform the other of his intention to clean the whiteboard (4). If the other agrees the whiteboards on both client terminals are cleared (5). Finally each client at any time can decide to close the communication (6). The other can keep on using the whiteboard, but she/he is no more in cooperation mode.

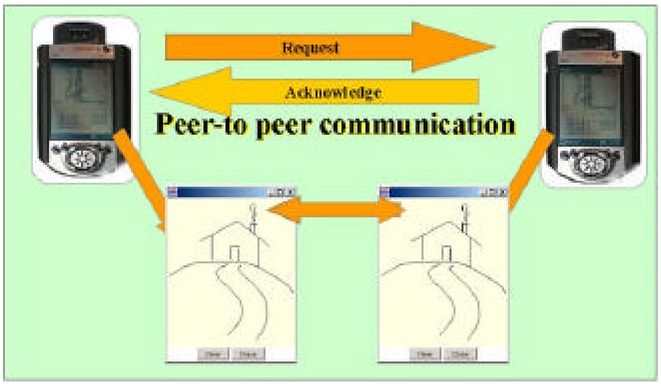


Figure 7 The Whiteboard displayed in two different clients workspace.

5. PERFORMANCE ASSESSEMENT

One of the main requirements for an effective VR system is real-time performance. Real-time rendering increases sense of immersion and presence inside the VE. We therefore tested our interaction system in order to verify

how implementation choices affect real-time rendering and interaction.
Since it is well known that RMI has non-negligible overhead with standard TCP/IP connection (Campadello et al., 2000), we tested the average time needed to RMI to call a method on remote host through the wireless network.

The first test was done in order to study how the server manages an increasing number of events per second depending on number of clients. We tested the event generated by the motion of a client on the 2D map in the GUI (on the PDA). In this case, the server manages the event in order to display the motion on all the clients' 2D maps. Tests were performed in two different ways:

1- Sequential version: any update event generates a thread on the server, which is in charge to contact sequentially every client to refresh its status.

2- Concurrent version: any update event changes every client status at the server side. On each client a dedicated thread exists, which concurrently calls the remote method for status refresh. (2) is the best testing solution because it describes better the behaviour of our application. Moreover, Concurrent update gives better performance results.

The diagram in figure 8 shows the results for the concurrent test solution (number of events per second depending on number of clients – continuous line - or the average time per event depending on number of clients – dashed line). This test gives information about application scalability. As we can see average times are acceptable up to 10 clients moving at the same time.

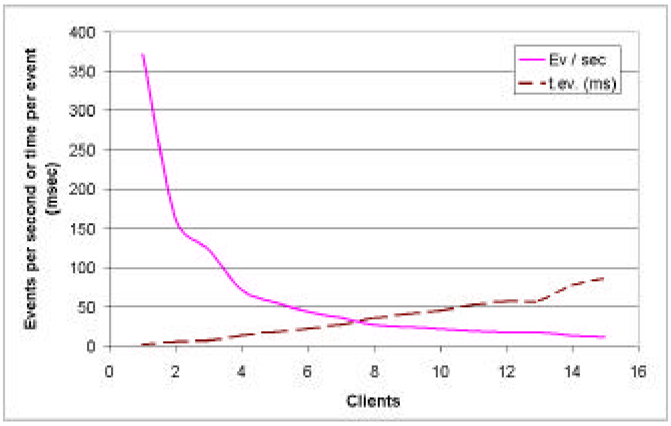


Figure. 8: Concurrent solution
X Axis: Number of clients Y Axis: Event per second or Time per event (msec)

The second test measures wireless network performance. The test is performed calling a huge number of "dummy" functions such as an echo function, a void function without any parameters or return values, a function which just increases number of parameters needed at any call, but without return values (we called this function “one way”).

The test is performed downloading the client on different platform to compare the results: (i) a notebook running Windows2000, (ii) the iPAQ where we implemented our application, in both cases (i and ii) through wireless LAN, (iii) on a desktop PC running Windows NT connected to the server workstation through wired LAN. Tests show (fig.9) that the application has better performance on the wired LAN thanks to a wider bandwidth. Wireless performance on the notebook is better than on the PPC because of computing capabilities.

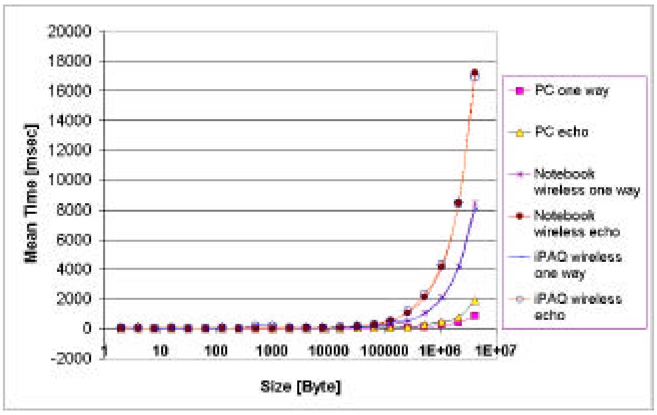


Figure 9: X Axis: Size of array sent Y Axis: Mean Time (msec). on different platform and with different methods

We verify in all cases degradation when parameters dimensions exceed 32Kbytes sent because data marshalling over this limit cannot be comprised in a unique TCP/IP packet.
Last test measures mean time needed to send an array of Java point type elements from a client to others (indicative dimensions [(1byte Boolean + 4byte + 4byte) x array length] + 4byte) and to wait for writing on the remote whiteboard. Measurements were taken changing number of clients and array dimensions. Each client tries to write concurrently to the others for 50 times. This test shows what happens when more than a remote method is invoked on the same PDA. Diagram shows average time dependent from number of clients (the parameter is the array length) or average time dependent from number of points to plot (the parameter is the number of clients). Diagrams show test performed with 1, 3, 5 and 10 clients from 30 to 1000 points sent. The results show that mean time becomes excessive when clients are more than 10 or for data packages corresponding to more then 1000 points sent concurrently by each client. (Figures 10 and 11)

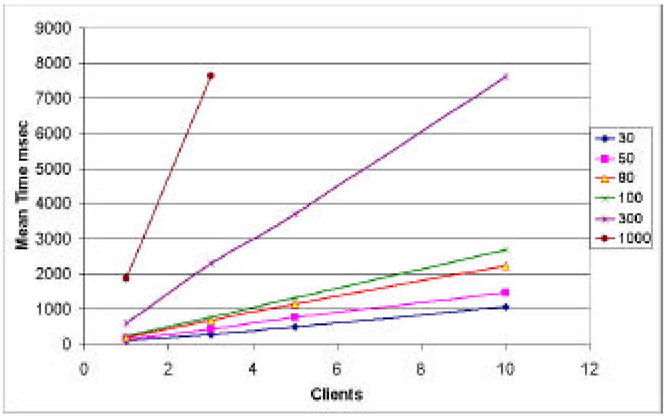


Figure 10 X axis: Number of Clients Y axis: Mean Time

As a conclusive summarizing, our quantitative approach shows that our system has good performance in a range of a limited number of users, in the neighbourhood of ten people interacting at the same time in the IVE.

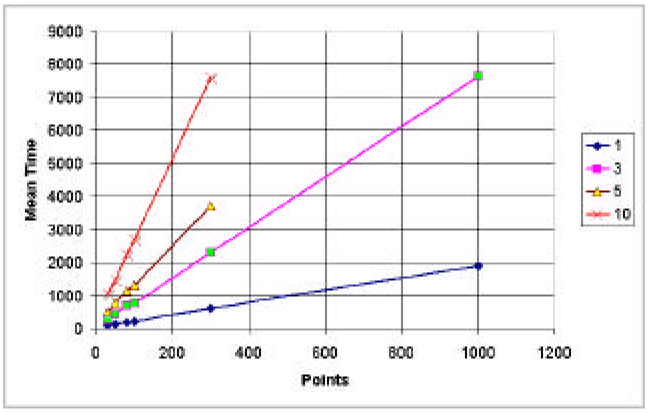


Figure.11 X axis: Points
Y axis: Mean Time

7. CONCLUSION AND FUTURE WORK

This paper presents the integration of wireless PDAs as interactions device in a semi-immersive multi-user shared virtual environment to support collaborative work and shared visualization.

Future works will explore different direction. First, we will compare our performance evaluation approach with alternative techniques. We will be involved in developing and enhancing multi-user features, to augment peers awareness, peer to peer exchange of data, to exchange audio and video data. This work will be combined with performance optimisation.

Moreover we want to explore different platform to run our client-server application, both in input and output. Finally we will perform usability test to improve the application in accessibility by the inexperienced user.

AKNOWLEDGMENTS

Authors wish to thank Prof. Marco Gaiani (PoliMi), scientific director of the “Ancient Appian Way 3D Web Virtual GIS” project on which we based our VR applications.

REFERENCES

Benford S. Grrenhalgh C., Rodden T., Pycock J.; "Collaborative Virtual environments", Communications of the ACM, July 2001/Vol.44.No7 pp.79-85

Benini L., Bonfigli M.E., Calori L., Farella E., Riccò B.; " Palmtop Computers for managing Interaction with Immersive Virtual Heritage", in Proceedings of EUROMEDIA2002, pp. 183-189.

Brooks F., "What's real about virtual reality?", IEEE Computer Graphics and Applications, November-December 1999

Campadello S., Helin H., Koskimies O., Raatikainen K., "Performance Enhancing Proxies for Java2 RMI over Slow Wireless Links", in Proceedings of the Second International Conference on the Practical Application on Java, 2000, pp.76-79

Campadello S., Koskimies O., Raatikainen K., Helin H. "Wireless Java RMI" Enterprise Distributed Object Computing Conference, 2000. EDOC 2000. Proceedings. Fourth International, 2000 pp. 114 -123

Damer B., "Inhabited Virtual Worlds", Interaction, September+October 1996, pp.27-34

Damer B., Bruckman A. (Moderators), Panel, "Peopled Online Virtual Worlds: A New Home for Cooperating Communities, a New Frontier for Interaction Design" Proceedings of ACM Conference on CSCW, 1996, pp.441-442.

Darken R.P., Peterson B. (2001) "Spatial Orientation, Wayfinding and Representation", Handbook of Virtual Environment Technology. Stanney, K.ed

Elvins T., Nadeau D., Kirsh D., “Wordlets – 3D thumbnails for Wayfinding in Virtual Environments”, Proceedings of UIST’97, 1997, pp.21-30.

Fitzmaurice G.W., Zhai, S., Chignell, M. H., “Virtual Reality for Palmtop Computers,” ACM Transactions on Information Systems, Vol. 11, No. 3, July 1993, pp 197-218, 1993.

Greenberg, S., Boyle, M., and Laberg, J. "PDAs and shared public displays: Making personal information public, and public information personal". Pers. Techs. 3, 1 (Mar. 1999), 54–64.

Hartling P., Bierbaum A., Cruz-Neira C. “Virtual Reality Interfaces Using Tweek”, in SIGGRAPH 02 Conference Abstracts and Applications, ACM SIGGRAPH2002, San Antonio, Texas, July 2002

Hill L., Cruz-Neira C., "Palmtop interaction methods for immersive projection technology systems," Fourth International Immersive Projection Technology Workshop (IPT 2000), 2000.

Hörtner H., Lindinger C.,Praxmarer R.,Riedler A. “ARS BOX with Palmist - Advanced VR-System based on commodity hardware ”, in SIGGRAPH 02 Conference Abstracts and Applications, ACM SIGGRAPH2002, San Antonio, Texas, July 2002

Kwon T. Choy Y., "A new navigation method in 3D VE (2D Map-based navigation)," International Conference on Virtual Systems and Multimedia, 2000.

Lindeman R.W., Silbert J.L., Hahn J.K. "Hand-Held Windows: Towards Effective 2D Interaction in Immersive Virtual Environments”, Proceedings of IEEE Virtual Reality '99, pp. 205-212, 1999.

Loftin R.B., "Design Engineering in Virtual Environments", Communications of the ACM, December 2001/Vol.44, No.12, pp.49-50.

Myers B.A. "Using Hand-Held Devices and PCs Together," Communications of the ACM. Volume 44, Issue 11. November, 2001. pp. 34 - 41.

Park K., Leigh, J., Johnson A., Carter B., Brody J., Sosnoski J., "Distance Learning Classroom Using Virtual Harlem" in Proceedings of the Seventh International Conference on Virtual Systems and Multimedia 2001, (VSMM 2001) pp.489-498.

Ragusa J.M., Bochenek G.M., "Collaborative Virtual Design Environments", Communications of the ACM, December 2001/Vol.44, No.12, pp.40-43

Rekimoto J., "A multiple-device approach for supporting whiteboard-based interactions", Proceedings of CHI'98, 1998.

Rorke M., Bangay S., "The Virtual Remote Control - An Extensible, Virtual Reality, User Interface Device", Proceedings of South African M. & PhD. Conference, June 1999, pp 39-43

Smith R.C., "Shared Vision", Communications of the ACM, December 2001/Vol.44, No.12, pp.45-48

Su S., Loftin R.B., "A shared Virtual Environment for exploring and designing molecules", Communications of the ACM, December 2001/Vol.44, No.12, pp.57-58.

Watsen K., Darken D. P., Capps W.V., “A Handheld Computer as an Interaction Device to a Virtual Environment,” Third International Immersive Projection Technology Workshop (IPT 1999), Stuttgart, Germany, 1999.

Wloka M.M., Greenfield E., “The Virtual Tricorder: A Uniform Interface to Virtual Reality”, UIST’95 Proceedings, 1995.

Yamazaki Y., Herder J., Exploring Spatial Audio Conferencing Functionality in Multiuser Virtual Environments Proceedings CVE00.

Dynamic Adaptation of Streaming MPEG-4 Video for Mobile Applications

Robbie De Sutter,
Sam Lerouge,
Jeroen Bekaert,
and Rik Van de Walle.
Multimedia Lab
Department of Electronics and Information Systems
Ghent University
Sint-Pietersnieuwstraat 41
B-9000, Ghent,
Belgium
E-mail: {Robbie.DeSutter,Sam.Lerouge,Jeroen.Bekeart,Rik.VandeWalle}@rug.ac.be

KEYWORDS
Mobile Multimedia, Software Framework, Universal Multimedia Access, Time-Dependent Metadata, MPEG, MPEG-4, MPEG-21

ABSTRACT

With the arrival of high-speed mobile communication networks and the increase of available mobile devices, the interest in new and advanced mobile multimedia applications is rapidly increasing. However, due to the large variety of mobile terminals (such as mobile phones, laptops, …) and, because of this, a huge collection of different terminal capabilities and terminal characteristics, it is hard to create generic mobile multimedia applications that can be used on mobile devices of different types. In this paper, we propose a mobile multimedia application that adapts its content to the capabilities of the mobile terminal and bears the typical fluctuating terminal resources into account, such as battery capacity and CPU load. To make this possible, a software framework is set up to enable negotiation between the mobile terminal and the content server. This is made possible by letting the framework be compliant to the Universal Multimedia Access framework, using Time-Dependent Metadata, and use fundamental concepts of MPEG-21. On top of that, the newly created framework is flexible, scalable, generic, and extendible. It makes it possible that multimedia applications can interact with the content provider in order to deliver an optimal multimedia presentation for any arbitrary mobile terminal at any given time, even during the actual playback of the presentation.

INTRODUCTION

Mobile communication systems have recently known an explosive growth (Becher et al. 2001). This continuing expansion will incorporate new services in the near future, such as mobile multimedia applications. These applications are designed to deliver multimedia content – a combination of audio, still and/or moving images, speech and other kinds of data – to a mobile device. However, to ensure a breakthrough of such applications, there is the need for further and extensive research and support to develop or improve design tools.
The huge variety of mobile devices (e.g., mobile phones, personal digital assistants, laptops, …) result in a collection of very different device capabilities, device characteristics, and bandwidth availability. Normally, a mobile device is – at best – optimized for handling one or a few specific kinds of multimedia data. For example, a mobile phone its main purpose is to make telephone calls (audio data) and, to a lesser degree, send and receive text messages. Mobile phones are consequently optimized in handling audio and text data, but they are not the ideal devices for processing other multimedia data, such as video. Moreover a mobile phone is, like most other mobile devices, very limited in battery capacity, processing capabilities, display possibilities, and bandwidth. Some devices (such as laptops) have better displays; others are optimized to have long battery autonomy.

The resources available on mobile devices during a multimedia presentation are more likely to change than on a desktop personal computer. Because the given fact that handling multimedia presentations is usually not the primary or sole goal of the device, the resources for executing them do not have the highest priority and are limited. For example the audio of a telephone call on a mobile phone has usually a higher priority than the audio belonging to a multimedia application. The same is true for processing capability and memory availability of the device.

METADATA

What is metadata

Metadata is generally defined as “data about the data”. Metadata can be used to describe any resource in order to make it possible to catalogue and structure this resource. It can be used to create electronic libraries; it can easily be integrated into applications as an aid or configuration; it helps to retrieve documents and so on. The key characteristic of metadata is that it is extra information that an application can use in order to facilitate the solution to the problem the application addresses.

When applied to multimedia data, metadata can be seen as extra information associated with the multimedia data describing the content of this data. The MPEG-7 standard (also known as “the Multimedia Content Description Interface”) of the Motion Pictures Expert Group (MPEG) group is one of the latest and most extensive efforts to create a standard for describing multimedia data content.

The content information described by MPEG-7 is rather static: once the audio and/or video content is finalized, the information is not likely to change drastically. The information can be refined and adjusted to reflect minor changes, but this is typically done over a relatively long period of time. When a user is viewing the content, the metadata information usually does not change during that relatively short period of time.

Time-Dependent Metadata

Time-Dependent Metadata is metadata associated with a multimedia application at a certain point in time during the execution of the multimedia application. It expresses the current preferences of the user, the current status of the end-user device, including but not limited to the available memory, CPU load, battery charge, and the network bandwidth. Research is being conducted in order to formally describe all facets of all possible involved parts (De Sutter et al. 2002).

The language which is used to write the Time-Dependent Metadata is the Extensible Markup Language (XML) (Achmed et al. 2001). Choosing XML implies that the Simple Object Access Protocol (SOAP) can be used to exchange messages between involved parts. SOAP is formally described as “a simple and lightweight mechanism for exchanging structured and typed information between peers in a decentralized, distributed environment using XML” (Box et al. 2000). SOAP is used to transmit XML data over a network connection. The protocol does not impose any restrictions about the content of the XML data and because it is fundamentally a one-way transmission from a sender to a receiver, it is ideal to use this protocol to send Time-Dependent Metadata.

Time-Dependent Metadata is used throughout the following application in order to allow the content to be dynamically adapted to the current status of the involved parts.

APPLICATION

The developed application is a Video-on-Demand (VoD) application which is capable to dynamically adapt the content of a requested video to changing usage environment. In order to achieve this goal, the application is UMA-compliant (Perkis, 2001), uses Time-Dependent Metadata, and is MPEG-21 compliant. The offered videos are encoded in MPEG-4 (Puri et al. 2000) (Walsh 2002) ISMA Profile 1 (also known as Advanced Simple Profile Level 3, Audio Profile Level 2) (ISMA). As such the application can use the MPEG-4 scalability features to adapt the content of the video to the end user environment.

There are two parts involved in the application namely the client part and the content server part. The latter is split up in 2 other parts, that is the broker part and the streaming server part.

Figure 1 depictures the architectural overview of the application and the relation between the different parts.

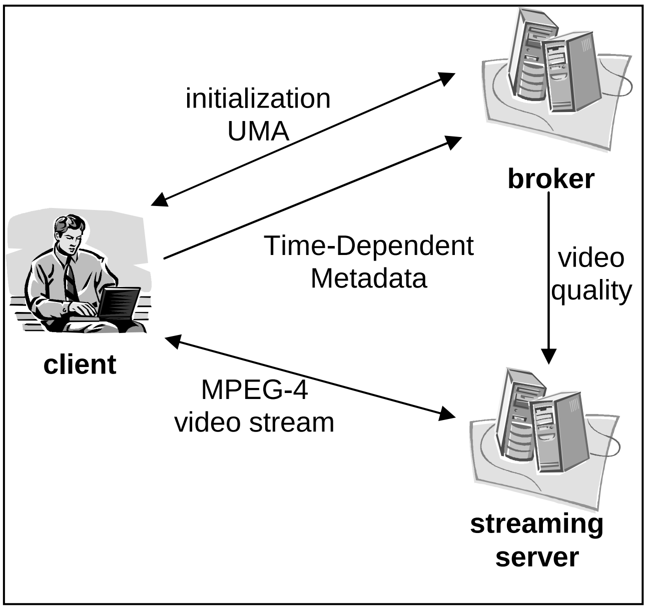


Figure 1: Architecture Overview

The client

The client encloses all important objects at the end user side. These objects are the end user terminal (including information about the hardware configuration, the software configuration, the network accessibility, …), the end user preferences, the end user location, the end user local time, and so on.

The client starts the application by communicating with the broker using SOAP. All communication between the client and the broker is initiated by the client and handles about information of the usage environment and the available videos.

The client also communicates directly to the streaming server, but only to control basic operations on the video playback stream (such as start, stop and pause), and not to neogiate about the quality of the stream.

The broker

The broker is the in-between controller and the heart of the application. Its duty is to handle all incoming requests from a client and respond to them in an appropriate way. There are two distinct types of requests:

- the requests about the application itself and the available videos such as starting a new session, request a list of available videos, request information about a specific video.

- the Time-Dependent Metadata requests that informs the broker about changing usage environment of a specific client.

It is also the broker its responsibility to communicate with the streaming server to notify on how to send a particular requested video to a specific client.

The streaming server

The streaming server sole task is to send the requested MPEG-4 video in the correct quality format to the client. It uses the information received from the broker to adapt the video stream to the changing usage environment of the client and this by using the scalability features of MPEG-4.

THE APPLICATION IN ACTION

The application has the ability to serve many concurrent sessions. A session can be seen as a single video request from an end user to the system. Each session is in either one of the two following states: the initial state or the playback state. A new session is created when the client part initiates communication with the broker. All sessions are started in the initial state.

Initial state

The initial state handles all necessary negotiation and bookkeeping before the requested video can actually be transmitted to the client part

First, the client starts the communication with the broker, hence creating a new session. The message to start a new session consists of the empty XML-element <GetVideoList> and is send to the broker in a SOAP message.

The broker will – upon arrival of this message – identify it as the start of a new session. It will reply with an overview of available videos. This overview is a MPEG-21 Digital Item made up of a Container grouping Items which represent the available videos.

Figure 2 shows the actual MPEG-21 Digital Item of the created application, containing two available videos (“Trailer Civilization III” and “Trailer Lord Of The Rings 2 – The Two Towers”).

```
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-NS"
      xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS">
  <Container id="Movies">
    <Item id="civ3">
      <Descriptor id="civ3_Title">
        <Statement mimeType="plain/text">
          Trailer Civilization III
        </Statement>
      </Descriptor>
      <Descriptor id="civ3_DII">
        <Statement mimeType="text/xml">
          <dii:Identifier>civ3</dii:Identifier>
        </Statement>
      </Descriptor>
    </Item>
    <Item id="lotr2">
      <Descriptor id="lotr2_Title">
        <Statement mimeType="plain/text">
          Trailer Lord Of The Rings 2 - The Two Towers
        </Statement>
      </Descriptor>
      <Descriptor id="lotr2_DII">
        <Statement mimeType="text/xml">
          <dii:Identifier>lotr2</dii:Identifier>
        </Statement>
      </Descriptor>
    </Item>
  </Container>
</DIDL>
```

Figure 2: MPEG-21 Digital Item of Available Videos.

The Digital Item is send to the client by using SOAP. Currently the Container – or better the individual Items within the Container – does not contain extended information, only basic information is given. It is absolutely possible that the Container is extended and contains other information, such as purchase price, content rating, required software and so on. Current information stored in the application consists of the descriptive name and it’s Identification. For the latter, we have used the Digital Item Identification (DII) framework of MPEG-21.

The end user terminal of the client part uses the information in the Container to display a list of available videos to the end user (Figure 3). If the Container also contains information about required software, the end user terminal can pick out those videos that can be played on the device. Other additional information can be modeled and responded to in similar ways. Eventually the end user is presented a list of available videos from which he has to choose one.

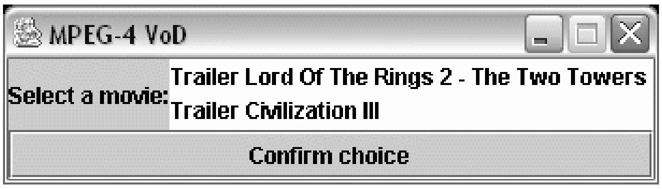


Figure 3: List of Available Video's, Presented to the End User

Selecting a video will trigger the client to send a new message to the broker, notifying it of the chosen video. The message encloses the selected video identification such that the broker can identify the desired video.

Once again this message can be extended to support particular needs of the broker. It can for example contain information about the payment or preferences on selectable parameters contained in the orginal MPEG-21 Container.

Now the broker knows the desired video, it can enter the last stage of the initial state. In this stage the broker gives the client extensive information about the selected video. It sends a new MPEG-21 Digital Item via SOAP containing three important parts (see Figure 4):

- *terminalID*: the broker will generate an unique identification string for the client. This ID is locked down in the session and must be used in all further communication initiated by the client part. It is used internally by the broker and the streaming server to know the status of the session.

- *videoURL*: this element denotes the location of the streaming server and the selected video.

- *videoParameters*: these lists the parameters and their corresponding values which can be set to adapt the video stream to the terminals condition. In the example code, two parameters can be tuned (“Power” and “CPU Load”) having their own independent set of values (respectively “Outlet”, “Battery” and “75% - 100%”, “0% - 75%”). The parameters can be anything that a content provider finds useful and necessary to guarantee the user the best possible video. As such they are dependent on the video and differ from video to video. Furthermore, it can contain a mixture of end user selectable preferences (for example full screen or windowed video) and automatic selectable terminal depending parameters (for example the remaining battery capacity). It are these parameters that will determine the final video quality.

```
<?xml version="1.0" encoding="UTF-8"?>
<soap-env:Envelope xmlns:soap-
env="http://schemas.xmlsoap.org/soap/envelope/">
<soap-env:Header/>
<soap-env:Body>
<DIDL xmlns="urn:mpeg:mpeg21:2002:01-DIDL-NS"
        xmlns:dii="urn:mpeg:mpeg21:2002:01-DII-NS">
    <Container id="lotr2">
        <Item id="terminal">
            <Statement mimeType="text/xml">
                <dii:Identification>
                    47296272c74dd69dc15346
                </dii:Identification>
            </Statement>
        </Item>
        <Item id="movieURL">
            <Statement mimeType="plain/text">
rtsp://multimedialab.elis.rug.ac.be/thetwotowers-
tlr_480.mp4
            </Statement>
        </Item>
        <Container id="movieParameters">
            <Item name="Power">
                <Item allowedValue="Outlet"/>
                <Item allowedValue="Battery"/>
            </Item>
            <Item name="CPU Load">
                <Item allowedValue="75% - 100%"/>
                <Item allowedValue="0% - 75%"/>
            </Item>
        </Container>
    </Container>
</DIDL>
</soap-env:Body>
</soap-env:Envelope>
```

Figure 4: SOAP Message with the MPEG-21 Digital Item Containing Detailed Video Information.

During this final stage, the broker also informs the streaming server about the client, the selected video and the default quality parameters for the video.

By sending this Digital Item, the session leaves the initial state and enters the playback state.

Playback State

Using the data in the last received digital item, the client has all required information to start the video playback. It can do so by connecting to the streaming server with the given URL accompanied with its terminalID. The streaming server on the other hand was informed by the broker to expect such a request and can stream the video upon request from the client.

Figure 5 displays the user interface while a movie is playing.

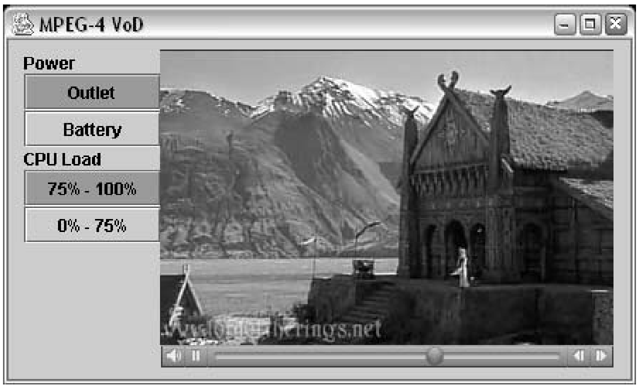


Figure 5: Playback User Interface

It is also possible that the client changes a value of one of the videoParameters because the default values don’t match the situation at end user side. For example it is possible that the “CPU Load” is “0% - 75%” instead of the value “75% - 100%”. The client can set the new value by sending a small SOAP message to the broker (see Figure 6). The message contains, beside the terminal ID, the name of the parameter and the new value for that parameter.

```
<?xml version="1.0" encoding="UTF-8"?>
<soap-env:Envelope xmlns:soap-
env="http://schemas.xmlsoap.org/soap/envelope/">
<soap-env:Header/>
<soap-env:Body>
    <setParameter
    terminalID="47296272c74dd69dc15346">
        <name>CPU Load</name>
        <value>0% - 75%</value>
    </setParameter>
</soap-env:Body>
</soap-env:Envelope>
```

Figure 6: SOAP Message Switching Parameter “CPU Load” to “0% - 75%”

Upon arrival of the message at the broker, the broker will translate it into a video quality setting. For example, our application translates a value “75% - 100%” for the parameter “CPU Load” into “only send I-frames”. The idea behind this scheme is to lower the CPU load by decoding less frames.

See Table 1 for an overview of the translation of the parameters for the video “Trailor Lord Of The Rings 2 – The Two Towers”.

Table 1: Translation "Movie Parameters" of the Video "Trailer Lord Of The Rings 2 - The Two Towers" to Video Quality Settings

Movie Parameter	**Value**	**Translation**
CPU Load	75% - 100%	Bandwidth: I-frame only
CPU Load	0% - 100%	Bandwidth: full framerate
Power	Outlet	Volume enabled: true
Power	Battery	Volume enabled: false

The translation is sent by the broker to the streaming server along with the client’s terminal identification. The streaming server can interpret this information quickly and adapts the video stream to the new situation at end user side.

Setting the videoParameters can be done at all time and is independent of the video stream. This means that the user part can set them before playback, during playback, whilst paused and so on. As such specific implementations of this application can guarantee at all times that the video playback is optimal fit for the client.

THE APPLICATION AS EXTENSIBLE FRAMEWORK

The described application can be seen as an implementation of an extensible generic framework.

It is extensible in a treble way. First of all, the framework does not impose any restrictions on the content. MPEG-4 is useful because of its inherent scalability features. Using these features makes it simple and straightforward to adapt content to different requirement, like reducing frames if required. While other video coding schemes perhaps do not allow these kinds of scalability, it remains possible to use them by implementing the streaming server otherwise. This means that existing video libraries still can be used without the huge cost of converting it to a new format.

The three part structure of the framework does not tell anything about the physical hardware itself. It is trivial that the client part consists of a possibly huge number of individual end users and their terminals. Also the streaming server part can group many geographically scattered genuine servers, each containing it own independent set of videos. Finally also the broker can actually group multiple machines to allow for example load balancing, and enable region dependent content. On the other hand all parts can even coexists on the same machine.

Lastly the framework allows content providers to specify how they want the quality of the content to be delivered. They can easily enforce minimal constraints and tune the scalability features. By incorporating a Digital Rights Management System, content providers can even impose restrictions how the content is consumed, how long the content is available for the end user, and so on.

The framework is also generic. The framework does not imply restriction whatsoever how the client, broker and streaming server part are implemented. It is necessary that the different parts can send XML to each other, but by using SOAP these messages can be send over a HTTP channel. However it must be noted that the use of SOAP is not obliged, other transport methods remain possible. It is even feasible that the communication between the client part and the broker part is different from the communication between the broker part and the streaming server part. As a final remark, the application requires that the client part must be able to handle streaming content, but how this is implemented is not imposed by the framework.

Because the framework can support all kinds of content, it is also generic in the aspect that it is independent of the provided content.

CONCLUSION AND FUTHER WORK

Given the fact of the huge diversity in functionality, capability, and available resources of mobile devices, it is obvious that there is a need for a concept to overcome the problems forthcoming out of this enormous distinct collection. While broadband wireless solutions are becoming more and more available and reasonable priced, the limitation of the device resources – especially battery limitation – are more important than ever before. Furthermore, mobile devices are more than desktop PC solutions subject to variable available resources, directly influencing the user experience.

We created a scalable, generic, and extendible framework to allow mobile multimedia application developer to reckon with the nature of mobile device. The framework is UMA compliant, uses Time-Dependent Metadata, and uses fundamental concepts from MPEG-21.

Our futher research will test the framework and the reactions of it in different network situations such wireless network, broadband networks, and smallband networks. Therefor we will use a network simulation such as NistNET to do the experiments. We also want to explorer further opportunities of scalable video, specificly MPEG-4 video, such as reducing resolution of the video.

The application is available online at http://multimedialab.elis.rug.ac.be. The client application – a Java Swing GUI – can be downloaded and can be used to connect to the Multimedia Lab content server. There are currently two videos, whereby the quality can be influenced by the client application and events can be simulated.

ACKNOWLEDGEMENTS

The research activities that have been described in this paper were funded by Ghent University, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the Belgian Federal Office for Scientific, Technical and Cultural Affairs (OSTC).

REFERENCES

Ahmed, K. and Ayers, D. and Birbeck, M. and Cousins, J. and Dodds, D. and Lubell, J. and Nic, M. and Rivers-Moore, D. and Watt, A. and Worden, R. and Wrightson, A., *Professional XML Meta Data*, WROX, Birmingham, 2001

Becher, R. and Dillinger, M. and Haardt, M. and Mohr, W. “Broad-Band Wireless Access and Future Communication Networks”, *Proceeding of the IEEE*, vol. 89, pp. 58–75, IEEE, 2001

Box, D. and Ehnebuske, D. and Kakivaya, G. and Layman, A. and Medelsohn, N. and Nielsen, H.F. and Thatte, S. and D. Winer, “SOAP v1.1 Specification”, *Technical Report*, W3C, 2000.

De Sutter, R. and Lerouge, S. and Bekaert, J. and Rogge, B. and Van De Ville, D. and Van de Walle, R. , “Dynamic adaptation of multimedia data for mobile applications,” in *Internet Multimedia Management Systems III*, John R. Smith, Thomas J. Watson, Sethuraman Panchanathan, and Tong Zhang, Eds., vol. 4862 of Proc. of SPIE., pp 240-248, 2002

Englander, R., *Java and SOAP*, O’Reilly, USA, 2002

Internet Streaming Media Alliance, ISMA, http://www.isma.tv

Maremaa, T. and Stewart, W. and Steward, B., *QuickTime for Java – A developer refrence*, Morgan Kaufmann Publishers, USA, 1999

MPEG, Moving Pictures Expert Group, http://mpeg.telecomitalialab.com

Perkis, A. and Abdeljaoued, Y. and Christopoulos, C. and Ebrahimi, T. and Charo, J., “Universal Multimedia Access from Wired and Wireless Systems”, *Birkhauser Boston Transactions on Circuits, Systems and Signal Processing*, Special Issue on Multimedia Communications, vol. 20, pp. 387 – 402, 2001

Puri, A. and Chen, T., *Multimedia Systems, Standards, and Networks*, Marcel Dekker, New York, 2000

Walsh, E. and Bourges-Severnier, M., *MPEG-4 Jump Start*, Prentice Hall PTR, New York, 2002

ROBBIE DE SUTTER received his Computer Science degree in 1999 at Ghent University, Belgium, and is currently working in the field of multimedia applications and their applicability for mobile device. He conducts this research as a research assistant at the Department of Electronics and Information Systems, Multimedia Lab, Ghent University, Belgium.

WLAN HIERACHICAL COOPERATIVE POSITIONING

Román García
Miguel Sánchez
Carlos Turró
Universidad Politécnica de Valencia
Spain
E-mail: {roman, misan}@disca.upv.es, turro@cc.upv.es

ABSTRACT

Wireless LANs on a campus-wide network are built with Access Points that act as a bridge for wireless users to access campus wired backbone. Beyond data services, Wireless LANs may be used to provide a geo-location service within the wireless network coverage area.

Previous work is based on nodes measuring signal level to three or more access points and then using some form of triangulation with or without a previous site survey.

Access Points are not deployed in a regular pattern along a campus. There are some places where a higher number of users is expected. Those areas get a higher density of access points while other areas have a much lower density.

This paper presents a new method that uses the help of neighboring nodes to create a geo-location scheme that relaxes the need of a node being covered by three or more access points too be able to know its current location. The prossed scheme is built as a client/server application mobile nodes are expected to run.

INTRODUCTION

Wireless LANs (WLAN) have many attractive applications. Campus LANs have included this kind of network to provide connectivity in libraries, and meeting rooms and almost everywhere. Roaming inside some campus area is now possible and the tendency goes toward more and more small, powerful and cheap portable computers (laptop, PDA, wearable...) that take advantage from this freedom of moving that WLAN offers.

WLANs are based on current standards IEEE 802.11 (2Mbps), 802.11b (11Mbps) and 802.11a (54Mbps). Campus WLANs usually work on infrastructure-mode supported by access point (AP) devices. A client (laptop, PDA, or any other wireless device) can communicate with other clients or with any other device in the wired campus LAN through an AP. So, it is necessary that a client is under, at least, one AP radio link coverage. If a client detects more than one AP, it will associate with the strongest signal AP. With 100mW of maximum transmitter power, the coverage of WLAN devices (with omnidirectional antenna), is in the range of tens to hundreds of meters, depending on the propagation characteristics of the environment. Indoors propagation is normally worse than outdoors due to obstacles attenuation. Furthermore, wireless bandwidth is mostly required indoors. So, we will find a higher density of APs inside campus building than outside with the highest density near press, conference or meeting rooms in order to provide higher bandwidth.

In order to ease roaming along the campus, many campus infrastructures provide outdoors connectivity. So, this paper assume that there is an outer campus area under only one AP coverage meanwhile areas near libraries, conferences or meeting rooms are under multiple APs coverage. The closer you are to these areas the higher the number of APs you are covered by.

WLAN may provide a new valuable service: mobile user computer geo-location. Global Positioning System (GPS) uses satellites to pinpoint mobile user location. But GPS can be used only outdoors. And you need special hardware to use GPS on a mobile computer. This paper presents a protocol that allows to estimate wireless user location anywhere within campus WLAN coverage.

This paper presents a hierarchical cooperative positioning approach (named HCooP) providing different levels of accuracy. Higher accuracy can be obtained based on the number of APs and other cooperating nodes nearby.

There are many uses of this service: user tracking, security, location based services, etc. A campus visitor that arrives in may start its laptop and be informed about its current location, campus map and how to get where he wants to go. Location information may be not very accurate when the user is in areas of low density AP coverage (i.e.: campus entrance, parking lot) and it gets better when he is approaching to a an area of higher AP density (i.e.: meeting room).

The rest of this paper is structured as follows: Section 2 explains the positioning protocol. Section 3 shows some simulation results.

HIERARCHICAL COOPERATIVE POSITIONING (HCOOP)

HCooP is a simple protocol to locate a mobile user in a WLAN. The protocol is based on each user periodically transmitting a message (i.e.: 5 seconds) to a central server. Messages sent by the mobile user contain the measured signal level from all the nodes in the sender's neighborhood and their respective MAC addresses (being those APs or other mobile users).

Based on the received messages, HCooP server software performs the following tasks:

1. Filters out those signal levels from mobile users that are not cooperating.

2. Collected signal level information from APs is processed first: APs have a fixed location, therefore they provide the most valuable information.

3. Collected signal level information from cooperating mobile users is used to improve estimated location in the previous step.

4. HCooP server software has a propagation model of the campus that is used to rule out possible ambiguity in the estimated locations.

5. Mobile users requesting a location estimate will get an answer from the HCooP server.

Some information is required to use HCooP: All APs precise location is know by HCooP server. Campus map is also stored on the server. Map information is used to estimate a given path attenuation. A campus APs signal survey was conducted and this information is also available at the HCooP server. Instead of using a fixed grid for signal survey we used a dBm scale so the number of measurement points could be reduced without loosing significant accuracy.

Positioning zones

Previous papers (Bahl et al. 2000) (Pahlavan et al. 1998) (Ladd et al. 2001) have presented solutions where location was inferred using APs signal measurement only and where a uniform estimate accuracy is obtained. However, those methods require at least a mobile user to be covered by three APs in order to be positioned.

HCooP allows a mobile user to be positioned if, at least, it is in range of one AP. We use other nearby mobile users location as a complementary information to position such a node.

Our scheme defines different areas depending on the number of APs that are in range from them. Level 3 areas are those where three or more APs can be reached. In these areas other mobile users information is not required as good results can be obtained as found in the literature above. Level 2 areas are those where only two APs can be reached. And Level 1 areas are under only one AP coverage. Both Level 1 and 2 areas will need mobile users cooperation to be able to provide a position estimate.

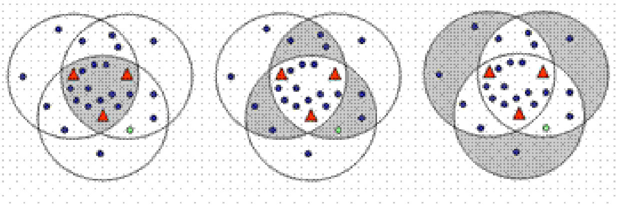


Figure 1. (a) Level 3 area (b) Level 2 area (c) level 1 area.

Level-3 Positioning

HCooP knows a mobile user is in a level 3 area if the location periodic message sent by this node contains three or more AP MAC addresses. An accurate positioning of this mobile user is possible based only on the signal levels reported from APs and the campus coverage map information.

Location estimation procedure is as follows:

Mobile user i sent a message containing three or more APs measured signal level $W = (w_1, w_2, w_3, ...)$. From the site survey we have a set of tuples of measured signal level collected in the form $L_k = (p_1, p_2, p_3, \ ...)$, these values where measured and tied to known locations X_k.

The estimated location is the X_k that makes $min(\| W - L_k \|)$.

We name mobile users positioned in this way level 3 users.

Level-2 Positioning

Mobile users are in a level 2 area when they can only hear two different APs. HCooP server cannot provide an accurate location estimate when only two APs are heard. Extra information has to be provided by cooperative mobile users.

HCooP server receives a message from mobile user $W = (w_1, w_2, w'_1, w'_2, ...)$ where only two of these measured signal levels come from APs. We split these values on two parts. Please note that it is the server software (and not the mobile user) who filters and sorts these values and knows which ones are APs and which ones are cooperative mobile users. So Q is the list of of APs signal levels and C is the list of cooperative mobile user signal level.

Location estimation procedure uses APs measure signal levels, $Q=(w_1,w_2)$, to select a set of possible locations by calculating those $\{X_k\}$ that makes $(\|Q-L_k\|)<\rho$. However, we are getting a list of different points $\{X_k\}$. ρ is the signal increment between two adjacent points of the site survey.

The final estimate is obtained by choosing the X_k that is more likely based on the measured signal level to the other mobile users $C=(w'_1,w'_2,...)$ which location $(X'_1,X'_2,...)$ is known.

Be the matrix A_{ij} the estimated power at location X_i from a mobile user located at X'_j. This matrix is calculated by using a signal propagation model and campus map information. Then given that cooperative mobile users location is known, the estimated signal levels (from the different positioned mobile users) at the different possible locations on the list $\{X_k\}$ can be obtained. For example, the estimated power levels at X_1 will be $(A_{11},A_{12},...)$

The more feasible location of the mobile user is X_h where the difference between measured and estimated signal levels is minimum, $min_h(\|C-A_h\|)$

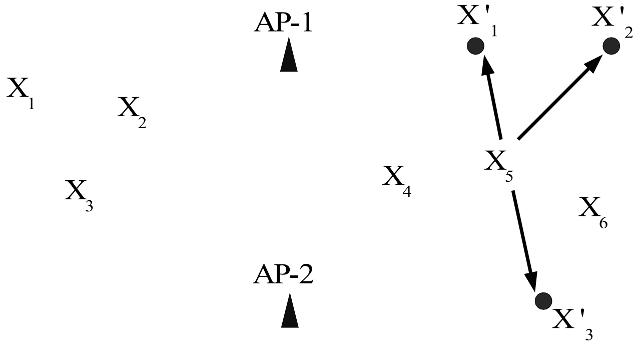


Figure 2. Level 2 positioning estimation.

Using this algorithm, level 2 mobile users are positioned. The algorithm has a significant advantage over other algorithms previously described in the literature as it takes into account propagation environment conditions. Not only Euclidean distance is considered but obstacles are taken into account too. This fact is why HCooP provides higher accuracy indoors than distance-only based algorithms.

Level-1 Positioning

Mobile user sends signal info of all nodes it hears from, as usual. HCooP server realizes there is only one AP in user message. Rest of info is from other nearby mobile users. As in level 2 procedure, non cooperating mobile users are removed from the list sent by the requesting mobile user. HCooP server obtains a list of possible locations, $\{X_k\}$, consistent with the AP measured signal level. $\{X_k\}$ is much longer than that of level 2. The other detected cooperating mobile users are used, again, to choose which element of the the $\{X_k\}$ set is the best location estimate.

Level-0 Positioning

It is technically possible to locate a mobile user without any AP (level 0) using only other previously located mobile users. Similar to the problem of positioning on an ad-hoc network (Savarese et al. 2001) . But this paper focuses on a server-based approach where mobile users limit to periodically send a message that is used to update location estimates. Mobile users perform no calculations. So to communicate with HCooP server a given mobile user has to be, at least, on a level 1 area.

RESULTS

We have conducted some test in a reduced part of Polytechnic University of Valencia main campus. HCooP test and site survey have been conducted in a controlled environment. Only some areas of our campus have WLAN coverage, but future deployment is planned to extend it soon.

Figure 3 shows the test area. It's a departmental area of approx 100 meters long by 25 meters wide. There are three APs shown as triangles 1,2,3 in the figure. These APs create a level 3 area (shadowed) and level 2 areas inside the building. There are level 1 areas outside the building that are not represented on the figure.

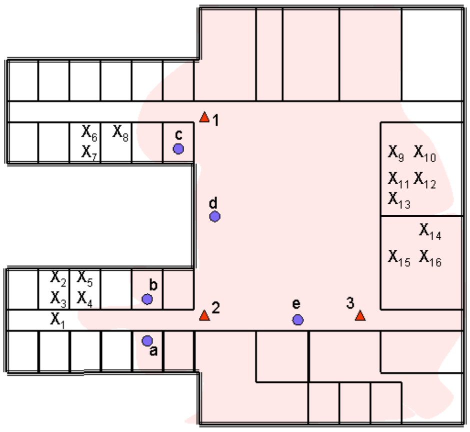


Figure 3. Test area.

Figure 3 shows a level 2 positioning sample case. A laptop is located on a level 2 area and it requests HCooP server to estimate its location. Real node location, known in the test is x_1. The laptop sends a message to the server with the signal measured from all the nodes in its neighborhood (either APs or other mobile users). HCooP server analyzes the request and it detects two of the measured signals come from APs, so it applies the level 2 algorithm. Server software knows that nodes a,b,c,d, and e have been positioned at level 3.

A list of possible locations for the mobile user is obtained at the server. A total of 16 different possible locations, from x_1 to x_{16}, satisfy the measured signal levels to the APs and a 5% error margin. Next, server software discards all the points in the result set that are in a level 3 area, x_{10} to x_{16}. This is because if the laptop were in on of these points it would have heard from the third AP.

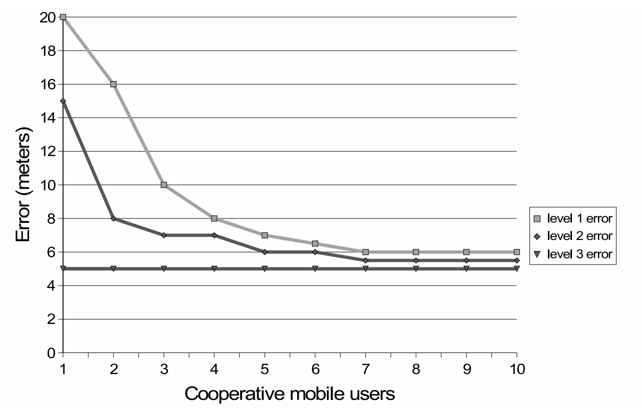


Figure 4. HCooP average location error.

For the rest of possible points, HCooP software calculates the distance to each of the mobile users present in requester's message (i.e.: requester's neighbor nodes).

The obtained results are shown in Figure 4. It is shown that for level 2 and 1 positioning results accuracy is improved as the number of mobile users is increased.

4. CONCLUSIONS

A new positioning protocol has been presented. HCooP does provide mobile users location estimate even when they are covered by only one AP. This is a significant improvement over existing methods that require a higher AP density. Only a few cooperative mobiles are needed.

Attractive features of HCooP are its simplicity and easy deploying over any wireless network and. It works in a centralized way, so it is easy to add additional features and more accurate services to the protocol.

HCooP defines three positioning levels. Higher accuracy is obtained at level 3. Each positioning level accuracy is suited to the kind of environment where such a level happens: Higher accuracy is required, more often, in those places where a higher AP density takes place.

Future work will deal with a campus-wide implementation of HCooP on a larger scale real Wireless LAN at the Polytechnic University of Valencia.

REFERENCES

P. Bahl, V. N. Padmanabhan, and A. Balachandran. A software system for locating mobile users: Design, evaluation, and lessons. Technical report, Microsoft Research, MSR-TR-2000-12, April 2000.

K. Pahlavan, P. Krsihnamurty, and J. Beneat. Wideband radio channel modeling for indoor geolocation application. IEEE Communications Magazine, 36(4):6065, Apr 1998.

H. Hashemi. Impulse Response Modeling of Indoor Radio Propagation Channels. IEEE Journal on Selected Areas in Communications (JSAC), 11:967978, September 1993.

Lance Doherty, Algorithms for position and data recovery in wireless sensor networks, M.S. thesis, University of California at Berkeley, 2000.

C. Savarese, J.M. Rabaey. Locationing in Distributed Ad Hoc Wireless Sensor Networks. 2001

A. M. Ladd, K.E. Bekris, G. Marceau, A. Rudys. Using Wireless Ethernet for Localization 2001

ACKNODLEDGEMENTS

This work was partially supported by the Generalitat Valenciana under grant CTIDIB/2002/29

MOBILITY CONTROL DAEMON: AN ARCHITECTURE SUPPORTING SEAMLESS HANDOVER BETWEEN HETEROUGENOUS NETWORKS

Jian Wu, Paul Havinga, Gerard Smit
Department of Computer Science
University of Twente
7500 AE, Enschede
The Netherlands
{jian, havinga, smit}@cs.utwente.nl

Niels Van Der Zwan
IBM Nederlands BV
1423 ND, Uithoorn
The Netherlands
niels_vanderzwan@nl.ibm.com

ABSTRACT

Seamless handover between heterogeneous networks has received considerable attention in recent years. In this paper we added a seamless handover mechanism between heterogeneous network should be dealt with at the network layer, which provide mobility over all types of media and does not need any changes in the higher protocol layers. In this paper, we propose a Seamless Handover Architecture, which is based on the implementation of Mobile IP. In order to have minimum impact on the whole system, we choose to only modify the mobile terminal part of the system, which has multiple wireless interfaces and can access multiple heterogeneous networks at the same time.

1. INTRODUCTION

With the emerging new technologies for mobile communication, computers will become a ubiquitous part of our life. We want to use them anytime and anywhere to provide us with information, communication facilities and entertainment. To satisfy this "anytime and anywhere" scenario it is expected that the mobile part of future network architectures (Becher et al. 2001) can be viewed as consisting of a set of overlapping tiers, each with its own specific characteristics. Satellite, Macro, Micro and Pico-cellular segments (Adacpi 1998; Stemm 1996) will each cover widely varying geographic areas and support different data rates for mobile terminals within them. Some tiers will be privately operated, others publicly operated. The handover between wireless cells of different type is characterized as heterogeneous handover. As the future wireless access networks become All-IP, the fundamental mobility problem becomes apparent: IP protocols were designed for stationary end systems. When the mobile host moves from one to another wireless access network, unfortunately, ongoing TCP connections break since the IP address is part of the TCP connection identifier and used at TCP connection setup. The problem of connection breakdown may be solved in different layers of the network architecture and may involve changes on all levels of the system.

This paper explores and analysis the potential technologies supporting transparent and automatic handover between dissimilar mobile networks without the interruption of ongoing connections and which can be a unified solution for different applications. This work is performed as a part of the European “Seamless Service” project on the construction and use of future network information services for mobile and non-mobile users (Seamless Service Project 2003).

2. PROBLEM DESCRIPTION

In our analysis, we stick to the existing networking technology and assume wireless network operators to have a full range of network coverage. It is not always possible and easy for us to modify some components of the network. For instance, although we could change and experiment with the lower and middle level networks, the higher layer is a public wide area network which is owned and administrated by a third party. So we normally cannot directly control the network infrastructure, which limits the changes we can make to support seamless handover.

The main characteristic of future mobile networks is the integration of the different types of wireless access network within a hierarchical cell structure. The access network types are classified into layers. Another important vision of further mobile network is the ALL-IP architecture (Ramjee et al. 2000). From the end user’s point of view, the wireless network will become a natural extension of the Internet sub-network but with an access and delivery mode that is optimized for the wireless environment. This concept will allow the deployment of a unified IP-based backbone, federating different access technologies.

In the TCP/IP world, two processes communicate via TCP sockets. If two processes are communicating over TCP, they have a logical network connection that is uniquely identifiable by the two sockets involved, that is by the combination < local IP address, local port, remote IP address, remote port>, which is called the socket pair. If any of the 4-quadruple is changed, the logical network connection will breakdown and data will be lost during the interruption.

In order to be able to connect with multiple radio access networks, we would expect the mobile terminal to have multiple radio network interfaces. When it moves around at the wireless overlaid network, handover inevitably occur from one of the radio interfaces to another. This means the mobile terminal moves from one IP subnet to another and obtains a new IP address in the new point of attachment. For all the current TCP connections, the unique identifier of the logical connection will at the same time be destroyed. Thus all ongoing communications between a mobile terminal and all its correspondent nodes would have to be terminated and new connections have to be initiated by the mobile node at its new IP address. Subsequently, every TCP network connection breaks down and all the data during connection reestablishment will be lost. Because the "always-on" concept is essential for the survival of future All-IP mobile network, the solution to enable seamless handover between various access networks would be imminent.

The network layer is concerned with controlling the operation of the subnet. A key design issue is determining how packets are routed from source to destination. It attempts to deliver packets from a node on one network segment to another node that may be on another network segment. The Internet Protocol is at the network layer. IP selects routes (determines paths) through a loosely confederated association of independent network links. IP offers routing from one network to another, in addition to some minor services such as fragmentation and reassembly, and check summing. Moving from one place to another can be modeled as changing the network node's point of attachment to the Internet, which is identified by the geographically related IP address. Supporting mobility at this layer is therefore naturally modeled as changing the routing of diagrams destined for the mobile node so that they arrive at the new point of attachment. This turns out to be a very convenient choice, and it was the option chosen by the Mobile IP working group. The major goal of Mobile IP protocol design was to handle mobility at the network layer and to leave transport and other higher layers unaffected, so that the existing routing infrastructure, non-mobile hosts, and current applications would not be required to change. Mobile IP (Mobile IP working group 2003) is a standard developed by IETF for the purpose of providing macro mobility across a set of different radio access technologies. Mobile IP defines a Home Agent as the anchor point with which the mobile client always has a relationship, and a Foreign Agent, which acts as the local tunnel-endpoint at the access network where the mobile client is visiting. Depending on which network the mobile client is currently visiting its point of attachment may change. At each point of attachment, Mobile IP either requires the availability of a standalone Foreign Agent or the usage of a Co-located care-of address in the mobile client itself.

Obviously Mobile IP is the favorable choice. Additionally, there are several reasons to choose Mobile IP as the solution. First, the mobile node can communicate with other nodes after changing its link-layer point-of-attachment to the Internet, which means a smooth handover between GPRS, WLAN or Bluetooth network. Second, the mobile node can communicate using only its home (permanent) IP address, regardless of its current link-layer point of attachment. This property will greatly facilitate the usage of corporate intranet. Third, the mobile node can communicate with other computers that do not implement the Mobile IP mobility functions, which means it is transparent to the correspondent node. Finally, several mechanisms can be deployed to enforce the security of the mobile node.

3. SYSTEM DESIGN

Although currently there are many research groups working on Mobile IP implementation on different operation systems (Dynamics-HUT Mobile IP 2003; MosquitoNet 2003), none of them has addressed the issue of seamless handover between heterogeneous networks. In this chapter, we propose a system architecture to support mobile node, which has multiple network interfaces, to seamlessly roam between different wireless access networks without any interruption of connections and to be able to select the best access network based on application and user requirements.

3.1. Overview

Mobile IP is a solution for mobility on the global Internet which is scalable, robust, and secure. It provides a mechanism for routing IP packets to mobile nodes which may be connected to any link while using their permanent IP address. Still we need a mechanism to mange the multiples interfaces on the mobile terminal, in order to be able to select the best available network connection and switch between heterogeneous networks seamlessly based on the Mobile IP implementations.

3.2. Architecture

The Seamless Handover Architecture consists of four functional entities, which are the *applications*, the *User*, the *Network Interfaces* and the *Mobility Control Daemon,* as showed in *Figure 1.*

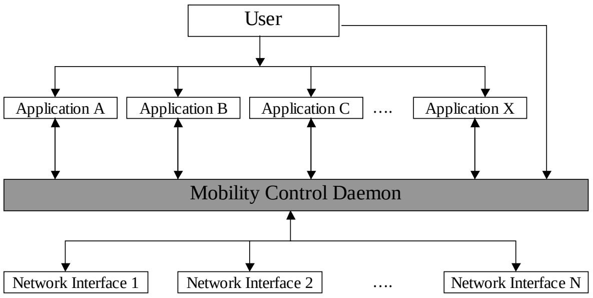


Figure 1 Architectural Overview

The *Mobility Control Daemon* is the central component of the Seamless Handover Architecture, which interacts with the other components in the mobile node. The Mobility Control Daemon has interfaces to the User, Applications and Network Interfaces. It collects information about different handover criterion, monitors the available network connections, makes choice of the best access network and executes the transition from one network to another. The Mobility Control Daemon also delegates the function of *Mobile Node* in the Mobile IP architecture. It has the necessary functionality of Agent Discovery, Registration and Tunneling.

The *Network Interfaces* are the multiple wireless network interfaces on the same mobile terminal. They could be used to connect to different wireless network. Each network interface is to have a complete IP protocol stack configured in addition to its specific wireless network protocol stack. This general TCP/IP protocol stack is the basis for IP network connectivity. At the same time, we assume that each network interface provides necessary Application Program Interfaces (API) to other applications. Through those APIs, the operational parameters of the individual network interface can be read and written, which provide the capability of outside monitoring and control. According to different application characteristics, the *Applications* can control the *Mobility Control Daemon* though the API it provides. Thus the *Application* might have influence on the choice of best wireless access network. For example, some real-time multimedia applications need to have large bandwidth to support the live streams. Then they can set the preference for higher bandwidth and lower latency access network. As a *User* of the system, he might also have his preference of the access network. His inclinations focus much more on the costs of the connectivity and the performance of the system. As such, he too has the ability to adjust the *Mobility Control Daemon* at his will.

The components of the *Mobility Control Daemon* are showed in *Figure 2*. The *Network Interface (NI)* is the class of objects representing the real Network Interface in the mobile terminal. *NI* corresponds to the specific wireless network interface and has the connections to the API of underlining real Network Interfaces. These *NIs* read the specific link properties from the correspondent wireless network interfaces, such as the Signal to Noise ratio, the Bit Error Rate and the bandwidth. Through those API, it can also switch on or off the real network interfaces when needed.

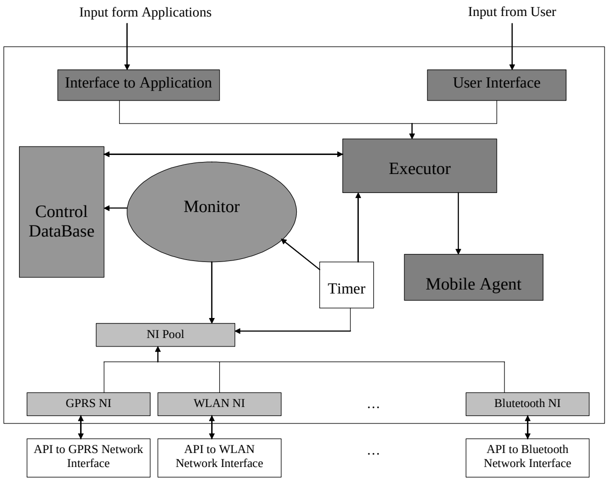


Figure 2 Components of Mobility Control Daemon

While the *NI* is powered on, it has two distinctive states, *Active* or *Sleep*. When a *NI* is in state *Sleep*, it means that the *NI* has not found any wireless access network of its type or has not yet established Network Layer connection with the Base Station of its wireless access network. Only after the *NI* has obtained an IP address from the wireless network and has been able to carry out the traffic from the upper protocol layers, can it switch to state *Active*.

The *NI Pool* is the collection of all *NI*s that are already powered on. It maintains a list of *NI*s (an example is showed in *Table 1*), through which the other system components to query for the real time parameters of each and every available network interface. Triggered by the *Timer* (showed in *Figure 2*), every two or three seconds it will probe the Operating System for newly plugged in or powered on network interfaces and add their information to the list of *NI*s. On the other hand, it deletes any *NI* that has been powered off or plugged off from the system during that time period.The information, which is stored in the NI Pool, is considered to be rather static when compared with other database in the *Mobility Control Daemon*. It acts as an index for querying the individual *NI*.

No.	Type	State	IP
1	GPRS	Active	130.89.23.22
2	WLAN	Sleep	0.0.0.0
3	Bluetooth	Active	9.168.45.3
...	...	...	...

Table 1 Example of NI List in the NI Pool

Another central function of *Mobility Control Daemon* is to monitor the available wireless network resources and to collect the data of network connections, which is the foundation for the selection of the best access network. There are two function units, which collectively accomplish this goal, as shown in *Figure 3*. The *monitor* has the responsibility of gathering the parameters of each active access network on a periodic basis. The polling interval is controlled by the *Timer* as shown in *Figure 2*, which triggers the monitoring event for every one tenth of a second. After receiving the trigger from the *Timer*, the *Monitor* first reads in the list of *NI*s from the *NI Pool* and selects all the *NI*s which are in the state *Active*. Then the *Monitor* probes each of these *NI*s for its link parameters, which includes delay, bandwidth, S/N and BER etc.

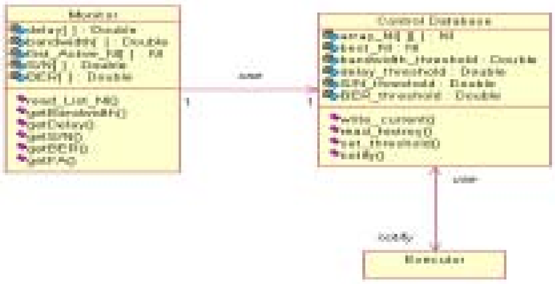


Figure 3 the Monitor and the Control Database

Another important task during this probe is Agent discovery, which is part of the functionality of the *Mobile Node*. As the Foreign Agent on each foreign link continuously broadcasts Agent Advertisements on its subnet work, the *NI*s in an *Active* state are able to intercept those messages. And if no Agent Advertisement message is received during the polling interval, an Agent Solicitation will be sent to actively discover the possible Foreign Agent on this link. As long as any FA advertisement is received, the *monitor* can learn the information about this foreign link such as FA IP address.

Each time the *monitor* finishes the data collection, it writes the current time information to the *Control Database*, which is the database for the whole *Mobility Control Daemon*. It stores not only the most recent polling results from the *monitor* but also the history of several round of polling. The information is stored in tables. Each table records all the information about a single Network Interface, as shown in *Table 2*. The aim of maintaining the history data is to cope with the problem of so called "ping-pong" effect when the mobile node moves back and forth between the borders of two overlaid wireless networks.

Time offset (ms)	FA IP	network parameters
0	130.89.34.29	dealy, bandwidth, S/N. BER
-100	130.89.37.35	dealy, bandwidth, S/N. BER
-200	130.89.37.35	dealy, bandwidth, S/N. BER
-300	130.89.37.35	dealy, bandwidth, S/N. BER
-400	None	dealy, bandwidth, S/N. BER
-500	Inactive	Inactive
…	…	…

Table 2 Record Table from One Network Interface in the Control Database

The computation of best access network in the *Executor* as shown in *Figure 4* can be triggered by the Control Database on several different circumstances. After each time new data is input from the *Monitor*, the Control Database checks whether these circumstances occur. If the answer is positive, a notification will be sent to the Executor to trigger the selection of the best access network. Those special circumstances are:

a) A special record in the Control Database is the *Current Best Access Network*, which is identified by the Network Interface it is connected. Whenever the newly inputs show that this Network Interface (called the *Current Access Network Interface*) is powered off or become inactive, a notification will be sent to the *Executor*.

b) A set of network thresholds, which set the link quality limits for the *Current Best Access Network* is compared with the real time values. In case one of the parameter thresholds is broken, it will notify the *Executor*. For example, we assume the threshold for the lower limit of bandwidth of the *Current Best Access Network* is 70k/s. Every time the new data about the bandwidth of the *Current Access Network Interface* is compared with this threshold. If it is less than 70kb/s, a new round selection will be triggered.

c) Whenever the new data shows that the known Foreign Agent IP address is different form the old one, the Control Database will notify the Executor. This happens when the Mobile Node needs to handover between the cells of the same access network, but across different subnet domains, which we always call *Homogeneous Handover*.

d) In case some of the non-current access *NI*s goes back to state *Active* and a Foreign Agent Advertisement is received on this link, a notification will be sent out. This circumstance stands for a new access network option is added to the choices and a new election round will include this new opportunity.

Several of these situations could happen in the same polling interval but only one notification will be sent to the Executor. The new selection result could choose a new access network and procedures of seamless handover have to be taken subsequently. Of course, when none of the above situations occurs in the polling, we assume the mobile node moves around in the same access network and no handover is needed.

The *Executor* is responsible for election of the *Current Best Access Network* based on the information collected from different sources. Several data sources could influence the computation of the best access network in the *Executor* as shown in *Figure 4*.

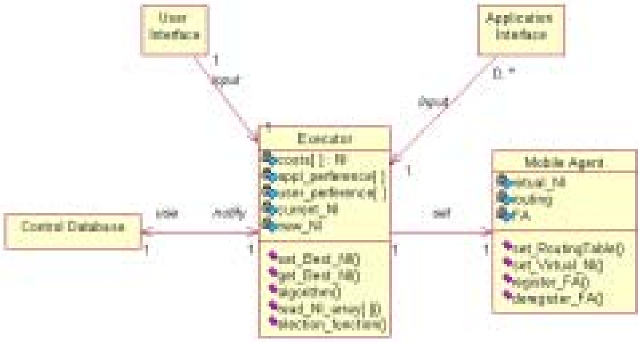


Figure 4 the Executor and the Mobile Agent

First of all, the information stored in the *Control Database* provides raw data about the operational parameters of all the *Active* Network Interfaces. It includes records, which span several continuous polling intervals of the *Monitor*. The *Executor* will read the data in when a fresh election is expected. Additionally, the *User* and *Applications* could control the election process by adjusting the preferences of the best access network. For example, the network service *User* who pays for the connection traffic, would properly very much prefer to select the lowest cost among the options. Other users may be more concerned with the link speed and they accept higher costs. The user can input his preference to the *Executor* through the User Interface and it influences the election results. On the other hand, we also propose an *Application Interface*, which opens options for specific applications to exert their influences, although it is unspecified in this paper. The election process could be very simple such as from the active access networks, choose the one that has bandwidth from 70k/s to 150k/s and the cost lower than three cents per kilobits. However, more advanced election approach has been proposed here to meet the user requirements.

Based on the data collected from the sources, the *Executor* applies the *Selection Function* on each of the current active network interfaces. This function has the inputs of various network parameters recorded in the Control Database, such as the S/N and BER, and the preferences set by the *User* and *Applications*. Through the computation of a specific algorithm, the *Executor* outputs an evaluation score for every active network interface. In the simplest case, the one that has the highest score, wins this election and is chosen to be the *Current Best Access Network Interface*. However a standard approach to affecting a tradeoff between call quality and expected the number of handoffs has been through an ad hoc algorithm based on hysteresis (Prakash et al. 2000). The implicit measure of best link quality used is simply the average value of the *Selection Function* on a certain time interval, which could easily be derived from the Control Database. The hysteresis algorithm is designed so that handover is made when the value from the new access network exceeds that from the current access network by a hysteresis level. Several more enhanced handover algorithms (Hatami et al. 1999; Akar 2001), such as Fuzzy Logic based algorithm could also be used here to improve the handover efficiency.

When the newly elected *Current Best Access Network* is different from the old one, the *Executed* invokes another functional component, the *Mobile Agent* to perform the necessary handover between different Foreign Agents. Firstly, the *Mobile Agent* sends registration request to the new Foreign Agent. During the registration, Home Agent notices this as a simultaneous binding request. After the confirmation of registration and authentication from the new Foreign Agent, the mobile node begins to receive packets through the new Current Best Access Network Interface, which are forwarded by the Home Agent. At this point, the mobile node receives tunneling packets from both the old and new best *NI*s. Then the Mobile Agent sends a de-registration message through the old best *NI*, so that the Home Agent will stop the simultaneous binding and only tunnel packet to the new best NI. Finally the Mobile Agent sets the routing table and adjusts the *Virtual Interface*, so that all the packets sent out by the mobile agent begin to flow out from the new best *NI*s instead of the old one.

4. CONCLUSIONS

In this paper, we propose a Seamless Handover Architecture, which modifies the mobile terminal and provides seamless handover between heterogeneous access networks. This architecture is based on the Mobile IP implementation, which changes the routing of datagram destined for the mobile node so that they arrive at the new point of attachment.

We propose several entities in this architecture, which collectively accomplish the overall system functionality. It allows the handover to have minimal data loss. From the upper protocol and application point of view, this solution is completely transparent. Any application based on the TCP/IP protocol stack is able to maintain its network connection during this handover, which means our architecture fulfill the generality requirement. To have minimum impact on the whole system, we only modify the mobile terminal part of the system. No modification is needed to change the third party infrastructure, which makes its implementation simple.

Nevertheless, from our analysis we still have to minimize the amount of additional network traffic used to implement handovers, which consumes bandwidth in the form of beacon packets and handover messages. On the other hand, security and QoS management remain critical issues to solve. Further work is expected to consider these implementation issues.

REFERENCES

Adacpi, F. 1998. “Wideband DS-CDMA for Next-Generation Mobile CommunicationSystems”, *IEEE Communications*, vol. 6, pp. 56-69, September 1998.

Akar, M. and U. Mitra. 2001. “Variations on Optimal and Suboptimal Handoff Control for Wireless Communication Systems”, *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, VOL. 19, NO. 6, June 2001,

Becher, R.; M. Dillinger; M. Haardt; and W.mohr. 2001. “Broad-Band Wireless Access and Future Communication Networks”, *Proceedings of the IEEE*, pp. 58-75, January 2001.

Deering, S. 1991. “ICMP Router Discovery Messages”, *RFC 1256*, September 1991.

Dynamics - HUT Mobile IP. 2003. http://www.cs.hut.fi/Research/Dynamics/.

Hatami, A.; P. Krishnamurthy; K. Pahlavan; M. Ylianttila; J. Mäkelä; R. Pichna; and J. Talvitie. 1999. “Fuzzy Logic, Neural Network and Other Algorithms for Handoff in Wireless Networks”, *Proceeding Of Third International ICSC Symposium on Fuzzy Logic and Applications*, Rochester, June 1999.

Mobile IP working group's Internet drafts and RFC documents. 2003. http://www.ietf.org/html.charters/mobileip-charter.html

MosquitoNet. 2003. The Mobile Computing Group at Stanford University, http://mosquitonet.stanford.edu/.

Prakash, R. and V.V. Veeravalli. 2000. “Accurate Performance Analysis of Hard Handoff Algorithms”, *Department of Electrical Engineering, Cornell University*, Ithaca, NY-14853,2000.

Ramjee, R.; T. La Porta; L. Salgarelli; S. Thuel; K. Varadhan. 2000. “IP-based Access Network Infrastructure for next Generation Wireless Data Networks”, *IEEE PERSONAL Communications*, TBD, July 2000.

Seamless Service Project. 2003. http://seamless.itek.norut.no/

Stemm, M. 1996. “Vertical Handoffs in Wireless Overlay Networks”, *Technical Report of University of California*, Berkeley, Number CSD-96-903, p. 29, May 1996

LATE
PAPERS

Effect of Different Parameters on Spectral Efficiency in Cellular Systems

D. Alnsour, M. Al-Akaidi
School of Engineering and Technology
De Montfort University, Leicester
LE1 9BH, UK.
e-mail: mma@dmu.ac.uk

Abstract

The performance of time division multiple access (TDMA) is affected by various factors. Accurate coverage prediction, modulation and coverage control techniques can significantly improve the capacity. The capacity of a cellular system is directly related to spectrum efficiency and is directly proportional to the cluster size, so that any decrease in the size indicates that co-channel cells are located much closer together. This allows more frequency channels to be reused taking into account minimum co-channel interference (CCI).

The above factors have an impact on the capacity, it is therefore crucial to include them in the evaluation to minimize the assumptions made. This paper shows development of a Monte Carlo simulation model, which includes these factors, to accurately estimate the spectral efficiency for average interference conditions.

I. Introduction

The radio spectrum is a finite resource and it is important that it is exploited efficiently by all users. Accordingly modulation scheme used for mobile environment should utilise the RF channel bandwidth and the transmitted power as efficiently as possible. This is due to the fact that the mobile radio channel is power and bandwidth limited, which implies the need to investigate different digital modulation schemes under different propagation channels. A typical application of such results would be the choice of modulation technique for a digital mobile radio system in a specific environment.

This paper discusses the impact of the choice of a modulation scheme on the reuse distance. Since the value of the reuse distance has a decisive effect on the performance of the spectral efficiency, a relation is set between the chosen modulation level and the spectral efficiency. Assuming the downlink case, the power at the mobile receiver from the desired base station (BS) and the co-channel BS's was calculated according to the propagation model used.

The CCI associated with a certain bit error rate (BER) for a particular modulation scheme will define the reuse distance D. The value of D is the maximum value that complies to the constraint set by CCI ratio. As detailed in [2], selected propagation model for both desired and interfering signals are applied. Since frequencies are reused at distance D, the area covered by the service of one set of these reused frequencies is roughly $\pi(D/2)^2$ [3]. The relation of the reuse distance D with radius R is given by the reuse distance factor, $Ru = D/R$ as shown in Figure 1.

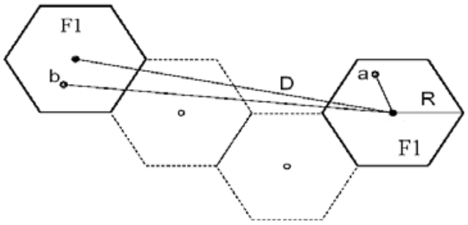


Fig. 1. The relation of reuse distance D and radius R where a and b are mobile users.

Manhattan microcell, Cost231-Hata and Lee models were used to assess the capacity potential for high-level modulation schemes. In order to provide a comparative capacity evaluation, 120^o and 60^o vs. omni directional configurations where applied in different propagation models. These coverage control schemes are used to increase capacity by reducing the co-channel interference, thus the reuse distance D [4].

The spectrum efficiency evaluation technique described studies the effective area of each cell in a co-channel interference limited environment. By specifying the bit error rate (BER) performance, the corresponding carrier-to-interference ratio (CIR) can be obtained for a specific modulation scheme, thus the reuse distance is evaluated as in [9].

The following sections describe the capacity evaluation method and the simulation results obtained using different propagation models.

II. Spectral and power efficiency of modulation system

The primary objective of spectrally efficient modulation technique is to maximise bandwidth efficiency. Performance of modulation schemes also measured in terms of power efficiency [5].

Large variety of modulation techniques have been studied for use in mobile radio communication. In this paper, the M-ary PSK and M-ary QAM was used for evaluating the spectral efficiency.

The system capacity is directly related to the bandwidth efficiency of the modulation scheme. Although that Hass and Jean have showed in their paper [6] that this is only true to a certain extent, work is still needed to be done in the region that the spectrum system efficiency increases with modulation efficiency. As the spectral efficiency increases the constellation lattice becomes more dense, hence, detection at the receiver becomes more difficult and BER may raise significantly. For this reason, greater power transmission is needed in order to maintain a specific quality of service. The higher transmitted power would raise the interference level in the system, which suggest larger reuse distance. Therefore by letting the CIR obtained from the modulation scheme for a certain BER to determine the size of the cell, will lead to optimum results.

In our work, the measurement of spectral efficiency is adapted in a cellular mobile radio system as [bits/s/Hz/km^2], therefore, it can be estimated by the following equation:

$$E_s = \frac{B_T/B_u}{\pi B_T(D/2)^2} \qquad (1)$$

Where D is the re-use distance, B_T is the total bandwidth of the system, B_u is the single user bandwidth. Hence, we have E_s [bits/s/MHz/Km^2] defined as a function of re-use distance.

III. Capacity Parameters

In this section, propagation models and modulation schemes are discussed separately:

A. Propagation Models

The cell radii is assessed from the received signal in a specific environment. The received signal using a developed $1/d^2$ Rayleigh propagation model is expressed as:

$$P_r = \sqrt{P_{rI}^2 + P_{rQ}^2}$$

where,

$$P_{rI} = P_t G_t G_r \left[\frac{\lambda}{4\pi(d+d'}\right]^2 \sum_{k=1}^{N} C_k \cos(2\pi f_c d \cos\theta_k + \varphi_k)$$

$$P_{rQ} = P_t G_t G_r \left[\frac{\lambda}{4\pi(d+d'}\right]^2 \sum_{k=1}^{N} C_k \sin(2\pi f_c d \cos\theta_k + \varphi_k)$$

P_t is the transmitter power, G_t is transmitter antenna gain, G_r is the receiver antenna gain, C_k is the reflection coefficient, f_c and λ are the carrier frequency and its wavelength, φ is the random phase introduced during the reflection process, d is the distance from the transmitter (direct path) the receiver and d' is the difference between the direct path and the reflected path [2].

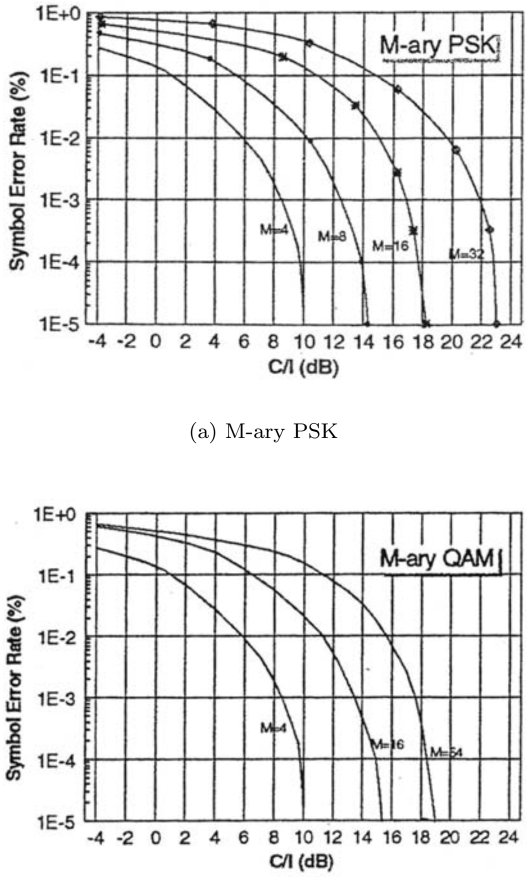
Using Rayleigh model, the cell radius is determined by the predefined system threshold, below which the mobile receiver may not be able to detect the desired signal from noise. The evaluated cell radius R_1 is 152 meters when the threshold signal level is set at -65 dBm and R_2 is 164 meters when it is set at -70 dBm. The path loss exponent 20 dB/dec can also be easily derived from Rayleigh model.

With reference to the required signal to interference ratio, the six co-channel cells are moved towards the center cell. The minimum reuse distance, D, corresponds to the point where the co-channel interference ratio is satisfied [7].

Using the evaluated cell radius, R, the reuse distance, D is obtained by estimating for each required co-channel interference level which is determined by the bit error rate (BER) requirement. The optimal reuse distance is based on the worst case configuration.

B. Modulation Scheme

In cellular systems, modulation has a significant impact on the system capacity. The capacity can be obtained by estimating for each required CIR level which is determined by the bit error rate (BER) requirement. In this paper all the simulation was performed for a BER performance of 0.001, the required co-channel interference ratio was found using Figure 2. It is clear from Figure 2 that M-ary QAM is a power efficient while M-ary PSK modulation techniques are spectral efficient. The above techniques are attractive for use in mobile communication systems as the spectrum becomes limited.



(a) M-ary PSK

(b) M-ary QAM

Fig. 2. BER performance under the co-channel interference.

The total bandwidth is determined from the following equation,

$$B_w = \frac{R_b(1+\alpha)}{\beta}$$

where R_b is the bit rater, α is the excessive bandwidth often taken as 0.35 and β is the bandwidth efficiency measured in bits/sec/Hz.

IV. Simulation

The simulated model takes into account the effect of users random location in their respective cells, the impact of propagation parameters, cell size, and cell sectorization is also considered. The spectral efficiency was computed for the worst case interference configuration, which corresponds to the case where all cochannel interferers are on the near boundary of their respective cells, at a distance $r_i = D - R$ from the desired mobile's base station. The analysis was also carried on for the best where the cochannel interferers are at a distance $r_i = D + R$ from the desired mobile's base station. In the practical case the user and the interferer were randomly allocated.

A. Lee model

Lee's empirical propagation model is considered accurate and simple to use for macrocellular systems [7]. Since this model is a semi statistical model, the results are also considered statistical. The model is used to predict the field strength of the received signal P_r which can be expressed as

$$P_r = P_o - \beta \log(\frac{d}{d_o} - \eta \log(\frac{f_c}{900}) + \alpha_o$$

where, f_c is the carrier frequency, P_o and η are the parameters found from empirical measurements at a 1.6 km point of interception [7], -64 dBm and -43.1 dBm respectively. In this paper, these were taken for an urban area (Newark, USA), where β is the path loss exponent, d is distance in km from the transmitter, d_o is 1.6 km, and α_o is the correction factor for a different set of conditions.

Using Lee's model, the radius (R=2.5 km) was determined for a maximum path loss of 65 dBm, and carrier frequency of 900 MHz. From the approach mentioned in the previous section, the reuse distance as a function of CIR was evaluated.

It should be noted from the results shown in Figures 3 , 4 and 5 that capacity has decreased considerably as the level of modulation increased and that M-ary QAM modulation scheme performed better than the M-ary PSK. As stated earlier, the improved performance of M-ary QAM is because this modulation technique is power efficient. As anticipated the results also proved an increase when introducing directional antenna. In the case of omnidirectional antenna pattern, all six interferers where considered, only two are considered when using 120^o directional antenna. The best results where obtained when introducing the 60^o directional antenna since only one interferer is taken into account.

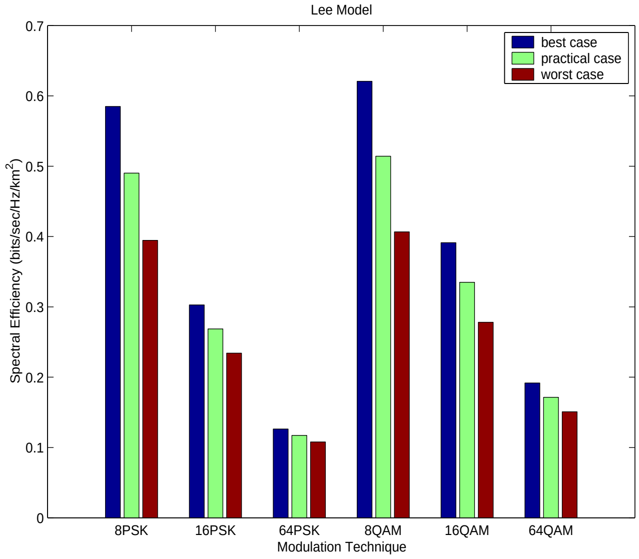


Fig. 3. Modulation effects on system capacity in macrocell environment for the case of omni directional antenna

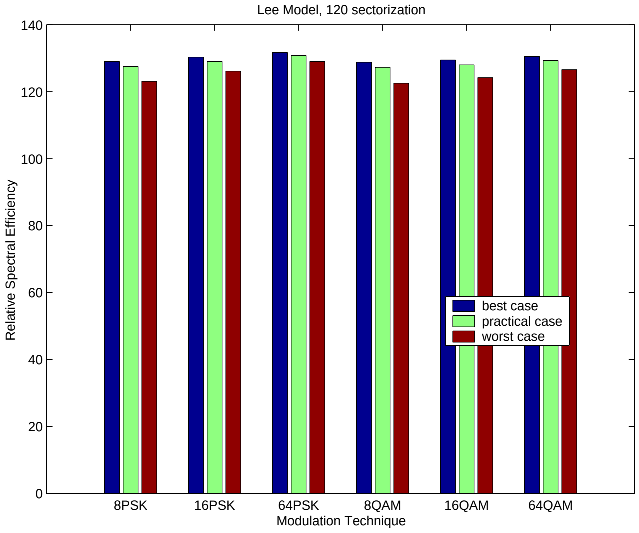


Fig. 4. Modulation effects on system capacity in macrocell environment for the case of 120^o directional antenna

B. COST231-Hata model

This model can distinguish between three different environmental types. The model is expressed in terms of the carrier frequency f_c (in MHz), the base station antenna height h_b (between 30 and 200 meters), the mobile station antenna height h_m (between 1 and 10 meters), and the distance d (between 1 and 20 km) between transmitter and receiver. The urban path loss LU is given as $A + Blog_{10}d$ for urban areas with some correction factors for suburban areas $(LSU) = LU - 15.11$ and for open areas $(LO) = LU - 30.23$. The terms A and B are expressed as follows [8]:

$$\begin{aligned} A &= 46.3 + 33.9log_{10}f_c - 13.82log_{10}h_b - a(hm), \\ B &= 44.9 - 6.55log_{10}h_b, \end{aligned}$$

where $a(hm)$ depends on the city type:

small and medium cities:

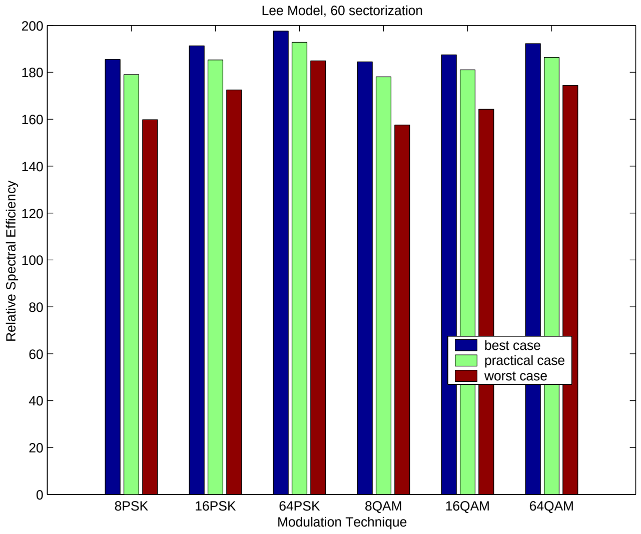


Fig. 5. Modulation effects on system capacity in macrocell environment for the case of 60^o directional antenna

$$a(hm) = (1.1log_{10}f - 0.7)hm - 1.56log_{10}f + 0.8$$

large cities:

$$a(hm) = 3.2(log_{10}(11.75h_m))^2 - 7.97$$

In our simulation, environment in large cities and urban areas is considered. The radius (R=0.78 km) was determined for a maximum path loss of 115 dBm. It is evident from the results shown in Figures 6, 7 and 8 that the accuracy of the model used and the accurate prediction of the signal strength improved the spectral efficiency. Using this model, which provides an accurate presentation of a microcell, the evaluated spectral efficiency of the 60^o directional antenna pattern 16-QAM is 3 times the efficiency estimated using omni directional model.

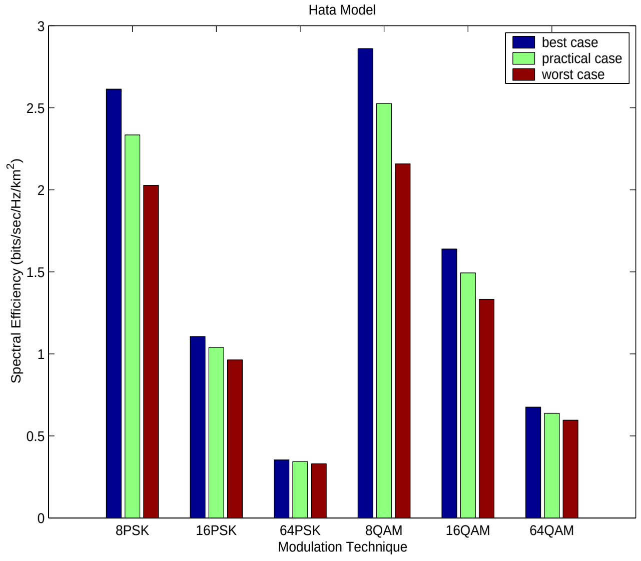


Fig. 6. Modulation effects on system capacity in microcell environment

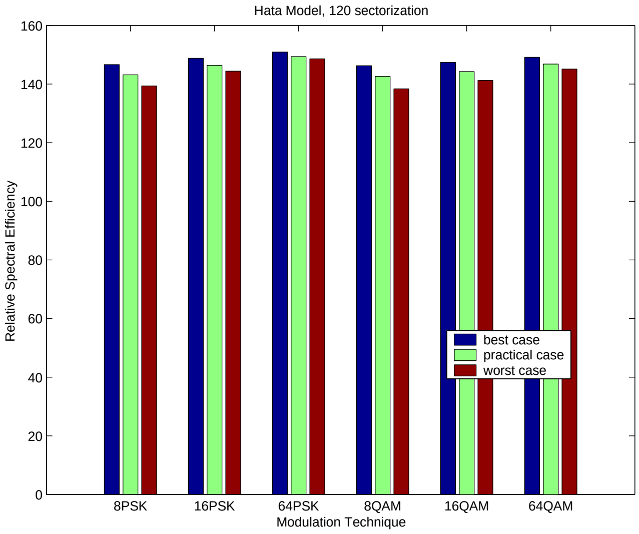


Fig. 7. Modulation effects on system capacity in microcell environment for the case of 120^o directional antenna

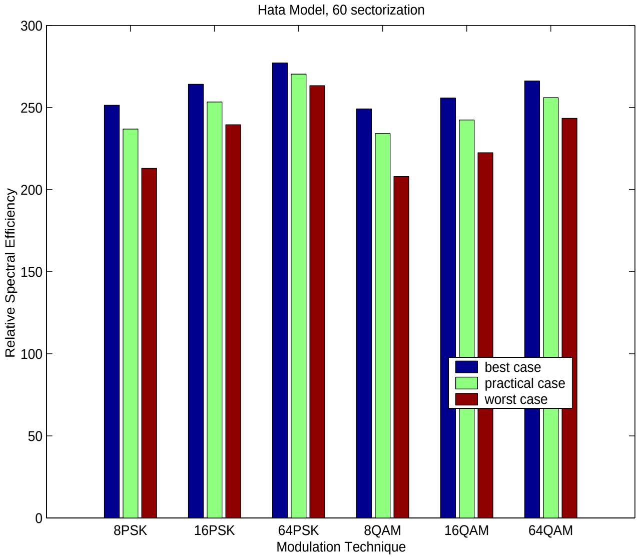


Fig. 8. Modulation effects on system capacity in microcell environment for the case of 60^o directional antenna

C. *Manhattan model*

The line of sight path loss (L_{LOS}) is defined using this model for the microcell scenario, for a distance $d \leq 300$, [3]

$$L_{LOS} = 82 + 40\log(\frac{d}{300})$$

Using the above mentioned model, the radius (R=213 meters) was determined for a maximum path loss of 115 dBm, and carrier frequency of 900 MHz. From the approach mentioned in the previous section, the reuse distance as a function of co-channel interference was evaluated. It should also be noted from the results shown in Figures 9, 10 and 11 that capacity has decreased considerably as the level of modulation increased and that M-ary QAM modulation scheme performed better than the M-ary PSK.

V. Conclusion

Higher modulation and different antenna pattern were investigated using the developed capacity evaluation technique. The obtained results demonstrated the possibility

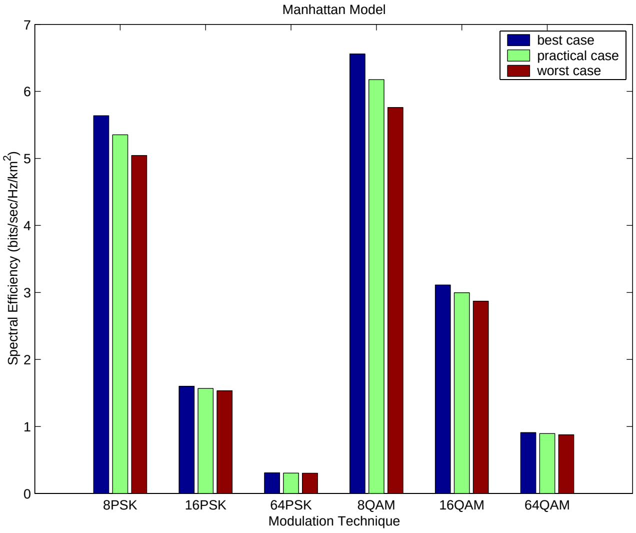


Fig. 9. Modulation effects on system capacity in microcell environment for the case omni directional antenna

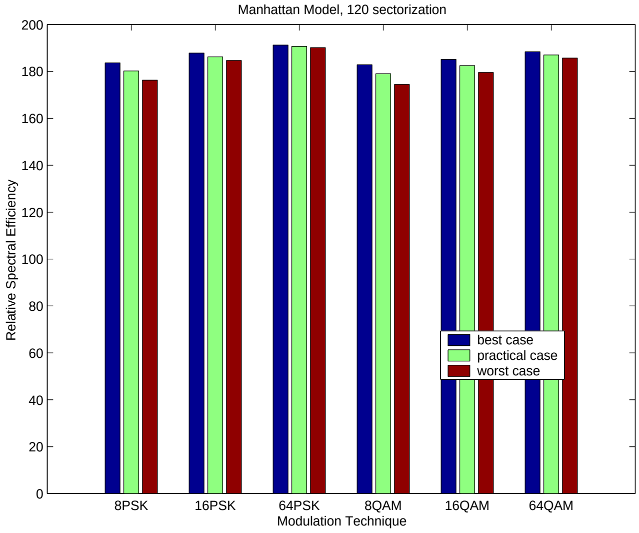


Fig. 10. Modulation effects on system capacity in microcell environment for the case of 120^o directional antenna

to use this technique to further investigate the impact of other coverage control and prediction techniques. In order, to improve system capacity, the factors, which are limiting the system performance, must be identified. It is also evident, that the best performance was obtained when a microcell is considered. Thus, the radius of the cell is inversely proportional to the spectral efficiency.

In a microcell, the performance of M-PSK and M-QAM degrades as the level increases. Despite better bandwidth efficiency at higher level modulation, more power is needed to maintain energy to noise (E_b/N_o) ratio to achieve a fixed BER at the receiver. The increase in power will require larger cell sizes to mitigate the increased interference level, thus reducing the spectral efficiency of the system. When introducing a 120^o directional antenna an increase of approximately 200% compared to omni directional was seen. The introduction of 60% caused an increase of approximately 500%.

In a macrocell, performance increases with higher modulation level since the cell size is large enough for the signal to be attenuated sufficiently before it could contribute significantly to the interference level. Therefore, the increase of bandwidth efficiency due to higher modulation level, improves the spectral efficiency, hence adaptive modulation is suggested. In other words, a higher level modulation scheme may be used to take advantage of higher transmitted power around the base site. A lower level modulation can be accommodated when the receiving power level is low. According to distance from the base station, different modulation schemes are adapted to site coverage, thus improving the system performance.

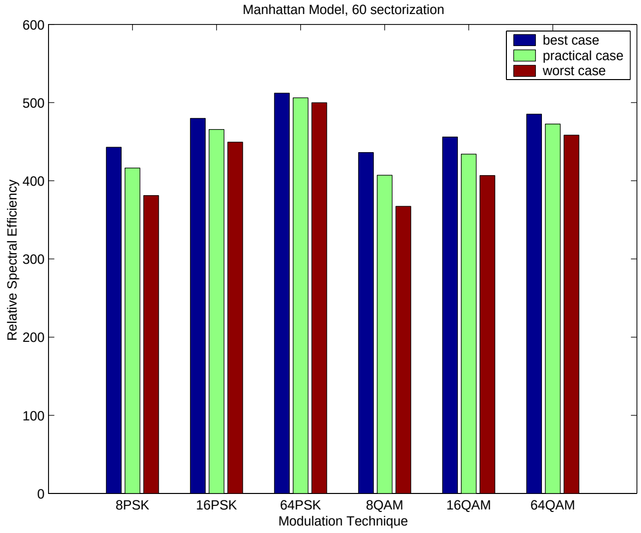


Fig. 11. Modulation effects on system capacity in microcell environment for the case of 60^o directional antenna

Finally, the improvement on the factors mentioned in this paper and the investigation of different parameters, results in an overall capacity improvement.

References

[1] M. Al-Akaidi, D. Alnsour, P. Urwin, H. Hammuda, “Capacity Evaluation in Cellular Systems”,IEE 3G2001,pp.175-179, London, March 2001.

[2] David Parsons, *The Mobile Radio Propagation Channel*, Pentech Press, 1992.

[3] Holma, Harri, Toskala, Anti, *WCDMA for UMTS*, Wiley, 2000.

[4] William C. Y. Lee, “Spectrum Efficiency in Cellular”, *IEEE Transactions on Vehicular Technology*, Vol. 38, No.2, May 1989,pp. 1-25.

[5] J. G. Proakis, *Digital Communications*, McGraw Hill, 1989.

[6] Ralf Hass and Jean-Claude Belfoire, “Specrum efficiency limits in mobile cellular systems”, IEEE transactions on vehicular technology,Vol 45, No. 1, pp. 33–40. Feb. 1996.

[7] W. C. Y. Lee, *Mobile Cellular Telecommunication System*, McGraw Hill, 1990.

[8] Nathan Blaunstein, *Radio Propagation In Cellular Networks*, Artech House, 1999.

[9] Lee, William C. Y., *Mobile Communication Design Fundementals,* Wiley and Sons, 1993.

Development and Implementation of an Internet-based Student Registration System using PHP/4 and MySQL Database

Jens Lichtenberg, Jorge Marx Gómez, Patrick Stiefel

Department of Computer Science
Technical University of Clausthal
Julius-Albert-Straße 4
D-38678 Clausthal-Zellerfeld
Germany
Phone: +49-5323-72 7115
www.in.tu-clausthal.de/~eve

ABSTRACT

Student registration procedures are very complicated and labor-intensive. For this purpose eVE [„electronic enrol procedure“ (German expression: **e**lektronisches **V**erfahren zur **E**inschreibung → eVE)] was developed for the enrollment of students for specific courses at the University of Clausthal. The aim of the present paper is to give an overview of the technical resources needed to create a system like eVE. The MySQL [„MySQL is the worlds most popular open source database, recognized for its speed and reliability.” [URL1] Unfortunately it has not the functionality of other database management systems. For a complete comparison see [Kline and Kline 2001]] database schemas will be presented and the functionality of the PHP [„PHP is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML.” [URL2]] interface shown. To get an impression there are several forms captured and presented in this paper [The shown captures are in German because that is the official language at Clausthal].

1 INTRODUCTION

The background of developing eVE was to reduce the costs of administrative overhead. Before the implementation of eVE, students had to apply in put up lists in order to register themselves. These lists had to be analyzed by secretaries, assistants or professors. This results in a problem of time and expenses. It has been solved with the implementation of eVE. The registration can be analyzed with a minimum of personal which frees them for more important purposes and saves a lot of money that way.

During the development the idea of creating a system as easy to control and handle as possible was followed. A professor or instructor enters his course informations via eVE into a database, the course data can be visited by the student to register for courses or examinations. eVE gives the instructors the possibility to view and print lists of their enlisted students or write an email to all of them, further the administrators have the right to delete students or cancel their registration from specific courses. They can also determine which students have registered for more than one course, although they only can visit one at the same time, for example 'seminars' where students might be taking away spaces for other ones.

2 THE DATABASE

2.1 Database Connection

The Institute of Computer Science (IfI) Clausthal is running an Apache HTTP-Server with MySQL and PHP support. Based on this technical standard eVE was developed on PHP scripts resulting in the structure shown below.

The Apache HTTP Server Project is an effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT. The goal of this project is to provide a secure, efficient and extensible server that provides HTTP services in sync with the current HTTP standards. Apache has been the most popular web server on the Internet since April of 1996. The Apache HTTP Server is a project of the Apache Software-Foundation [URL3].

With PHP one can easily connect to a database, this happens with the script class mysql.class.php. Embedding this file into other ones, the implementation of the connection parameters is only needed once. After having declared these parameters, the function mysql_connect connects to the Database Server on which a, in mysql.classl.php stated database, is selected by using mysql_select_db. There are some more practical functions for the use of MySQL in PHP implemented in the eVE scripts like mysql_query() or mysql_fetch_array() [Krause 2001].

2.2 Database Structure

The eVE database consists of 3 tables.

At first there is the user part of the database called 'userdates', where the information about each registered student is stored. The following relational expression represents the table 'userdates' in relational schema [Brass 2002]:

```
userdates(ID, REGNUMBER, LAST,
FIRST, STUDY, SEMESTER, EMAIL,
NICKNAME, PASSWORT)
```

A student has a first name ('FIRST'), a last name ('LAST') and a unique (at the university and in the database) registration-number, given to him when he was accepted at the university, called 'REGNUMBER'. With this data every student can be uniquely identified and named. While both name fields cover with the datatype varchar(50) most of the names, 'REGNUMBER' is limited to 6 digits with int(6), which is enough because of a 6 digit registration code at the university. Furthermore the amount of semesters and the study are stored for future use. See Outlook for more information on this topic.

If a student enlists for a course, eVE sends him an email with his registration data, to legally inform him what of his data has been stored. This information has to be sent because of the German Federal Data Privacy Law [Nagel 2001]. Instead of storing provider and email nick separately they are stored as a single string, so that an easier access is possible. The email address is in contrast to the 'REGNUMBER' no unique field because it is possible that students use the same email address. Additional to the information needed for the registration process, eVE gives the user the possibility to choose his personal unique nickname with that the user can be identified. Next to this nickname a user must choose a password. The password and nickname are stored with the other personal data in the ‘userdates’ table. The nickname is 20 characters long, but to preserve security aspects the password field has to be 100 characters long because of MD5-Encryption. MD5 was developed by Professor Ronald L. Rivest of MIT. The MD5 algorithm takes as input a message of arbitrary length and produces as output a 128-bit "fingerprint" or "message digest" of the input [URL7].

The registration data must be entered, because otherwise other students might enlist a student for a course, who does not know he was enlisted. Last but not least “userdates” has an id field, which per se is obsolete because a student can be uniquely identified through his ‘REGNUMBER’. This field was inserted nevertheless, because it provides the possibility to sort students by their registration date, which can be interesting for statistic purposes.

The next major table is called ‘courseoffer’ and holds the information about all courses. The table is represented by this scheme:

courseoffer(ID, COURSEID, PROF, COURSE, SPACES, DEADLINE^0, ATTACHMENT^0, TYPE, GROUPE)

DEADLINE^0 states that the field DEADLINE can store the NULL value. An underlined word in the brackets is the key of the given table. For further information on relational databases please refer to [Ullman et al. 2002].

To identify a course it needs a unique number. This number is called ‘COURSEID’, that is at least a 3 to 6 digit smallinteger number. The first digits represent the number of the course the last 2 define the type of the course:

8**00**	Course itself
8**1x**	Exercises
8**2x**	Tests/Exams
8**3x**	Study trip

Table 1: Example COURSEID’s for course 8

The field ‘PROF’ states the person who teaches the course by name. Furthermore an offer needs a name so that the students can detect what the course is about. This description field is called ‘COURSE’ itself, it is 255 digits long and of the type ‘varchar’, so that longer course names are possible. A course has a limited number of open seats. This number is stored in the field ‘SPACES’. It is a 6 digit smallint number, what is enough to cover all the university rooms. The field ‘DEADLINE’ states the date until which the students can enlist for the course. Being a date the field is defined as such with the MySQL specific date type. Next there is the field ‘TYPE’, which defines whether a course is a lecture, an exercise, a seminar (a course that is directly connected to the exercises and where all topics are discussed within a short compact period of time), a study trip, an examination/ test or something else in between the other 5. Also being stated before ‘TYPE’ in the scheme the field ‘ATTACHMENT’ can not be understood without the knowledge of ‘TYPE’. ‘ATTACHMENT’ is specifically for exercises or examinations belonging to a specific course mostly a lecture. The number stored in this field is the ‘COURSEID’ of the lecture the course is attached to. For better information the field ‘GROUPE’ states the difficulty level of a course, whether it is for beginners, advanced learners or for both.

The table ‘registrations’ is represented by this schema :

```
registration(REGNUMBER,
COURSEID)
```

Using the SQL Standard, both fields would be Foreign Keys referencing ‘userdates’ and ‘courseoffer’ but unfortunately MySQL does not support Foreign Keys up to Version 3.22 [URL 6], so that possible conflicts must be prohibited with the PHP-scripts. For further information on the possibilities of Foreign Keys please refer to [Ullman et al. 2002]. The table stores the enlistments, which means a student-course combination. Example: If the student with his stored ‘REGNUMBER’ 123456 enlists in course ‘Business Informatics 1’ with the ‘COURSEID’ 100, eVE stores the dataset (123456, 100) into ‘registrations’.

3 IMPLEMENTING THE DIFFERENT FUNCTIONS

eVE is divided into two sections. One section includes the student registration area, it can be found at http://www.in.tu-clausthal.de/index.html (see. Fig. 1). The part is the administration zone with the possibilities to evaluate the registrations. Its url is http://www.in.tu-clausthal.de/index2.html (see. Fig. 2)

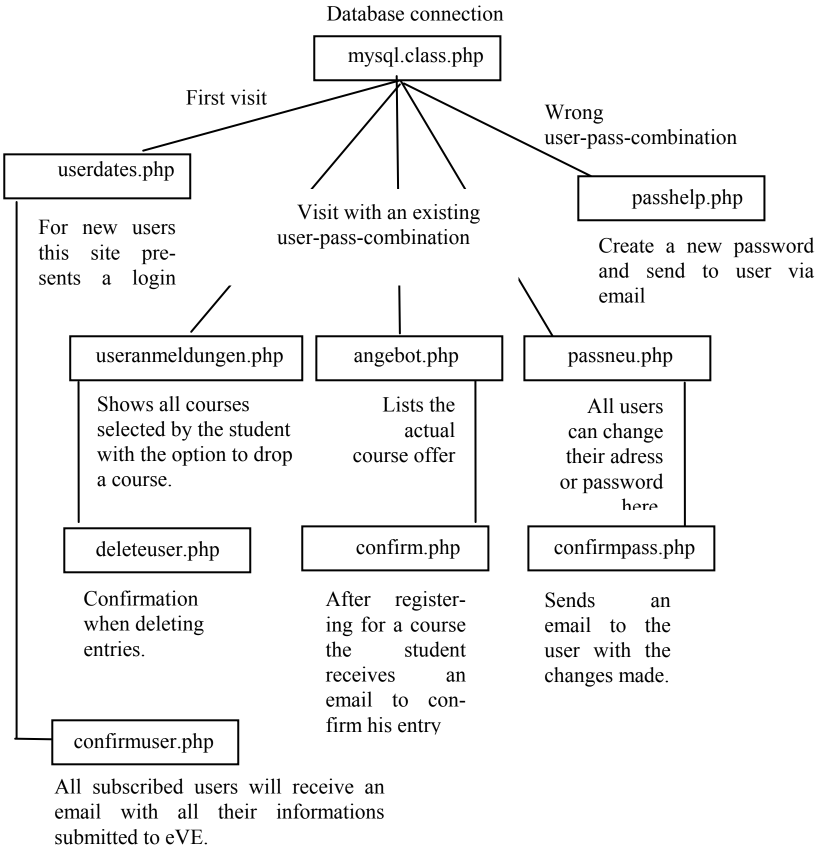
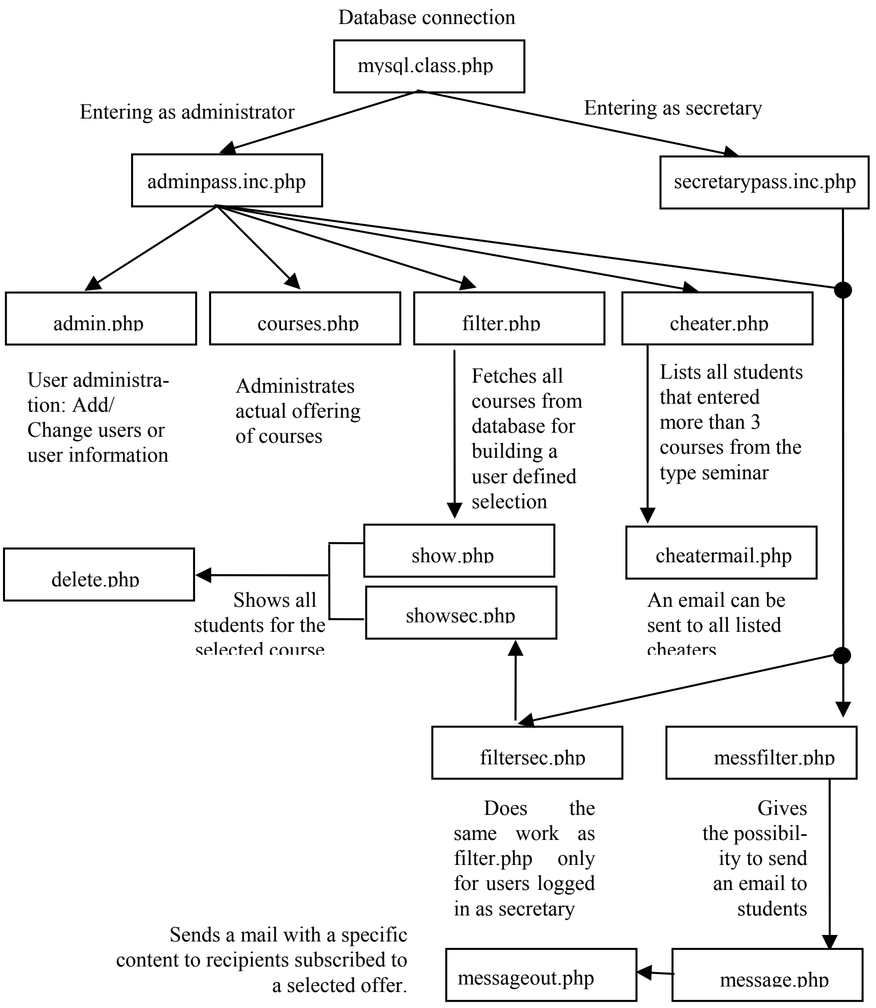

Database connection
mysql.class.php
First visit
Wrong user-pass-combination
userdates.php
passhelp.php
Visit with an existing user-pass-combination
For new users this site presents a login
Create a new password and send to user via email
useranmeldungen.php
angebot.php
passneu.php
Shows all courses selected by the student with the option to drop a course.
Lists the actual course offer
All users can change their adress or password here
deleteuser.php
confirm.php
confirmpass.php
Confirmation when deleting entries.
After registering for a course the student receives an email to confirm his entry
Sends an email to the user with the changes made.
confirmuser.php
All subscribed users will receive an email with all their informations submitted to eVE.

Fig. 1. Structure of the student registration page



Database connection
mysql.class.php
Entering as administrator
Entering as secretary
adminpass.inc.php
secretarypass.inc.php
admin.php
courses.php
filter.php
cheater.php
User administration: Add/ Change users or user information
Administrates actual offering of courses
Fetches all courses from database for building a user defined selection
Lists all students that entered more than 3 courses from the type seminar
delete.php
show.php
showsec.php
cheatermail.php
Shows all students for the selected course
An email can be sent to all listed cheaters
filtersec.php
messfilter.php
Does the same work as filter.php only for users logged in as secretary
Gives the possibility to send an email to students
Sends a mail with a specific content to recipients subscribed to a selected offer.
messageout.php
message.php

Fig.2 Structure of the administration site

3.1 Login Zone

All registered users have a login zone which they can access with an existing user-pass-combination. There they have three possibilities for their selection:

- *Change of user information:* (passneu.php), see 3.2
- *Listing of the actual course offer*: (angebot.php), see 3.3
- *List of all selected courses joined by the user*: (useranmeldungen.php), see 3.4

The Login in eVE is realized via header authorization.

```
header('WWW-Authenticate: Ba-
sic realm="Login Zone"');

header('HTTP/1.0 401 Unauthor-
ized');
```

The user receives a login box, where he can enter his username and his password, he chose during the registration. Both entry-fields can now be used to filter the table 'userdates' for a valid user-pass-combination. Although the password is transmitted as plain text, this method of authentification is appropriate for this purpose, because this basic authentification is fast in comparison to more secure methods [Welling 2001] and it is very easy to connect to our SQL database. The username is stored in the variable $PHP_AUTH_USER, the entered password in $PHP_AUTH_PW [URL5]. Its clear that the selected password at first time login is not stored in plain text into the database. The password submitted via the login box is encoded with md5 and compare with the stored password, belonging to the stated nickname, from the database. The encrypted password is stored in a variable, with which the database query is done.

```
$PWMD5 = md5($PHP_AUTH_PW);
```

In the variable 'log' the query results are stored:

```
$log=$database->query("SELECT
ID FROM <table> WHERE NICKNAME
LIKE '$PHP_AUTH_USER' AND
PASSWORD LIKE '$PWMD5'");
```

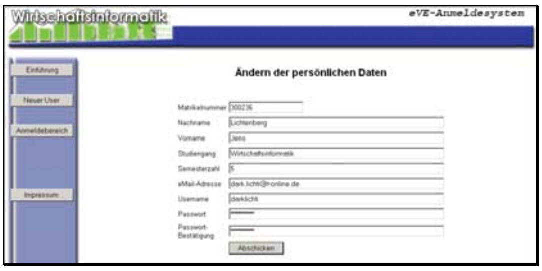
Num_Results now returns the number of rows within the results:

```
$log_num = $database-
>num_results($log);
```

If the query above is empty and no existing user-pass-combination has been found in the database, eVE replays the login box 2 more times to give the user a chance to correct possible spelling mistakes and shows at last a new form, where the user can let the server create a new password that is sent to the stored email address belonging to the used nickname (passhelp.php). The security aspect is still alive. Even if an unauthorized eVE visitor tries to enter the system with any known username and without success, he can change the users password but will never get to know the new changed password because that is emailed to the 'real' user. To transmit all needed variables from one form to an other, all Submit-Processes in eVE are realized with the form method post [URL4].

3.2 Change Of User Information

All given information at the registration to eVE is presented here (see Fig. 3) to the user with the option to change this data, even his user-password combination. After a successful change eVE will setup the new entries in the database and sent a confirmation via email. Only the most essential data for the system purposes is stored, so that the privacy of the user is not violated. This storage of only the essential data is once again in reference to the "German Federal Data Privacy Law" [Nagel 2001].



Ändern der persönlichen Daten

Matrikelnummer 300236
Nachname Lichtenberg
Vorname Jens
Studiengang Wirtschaftsinformatik
Semesterzahl 5
eMail-Adresse dark.licht@t-online.de
Username darklicht
Passwort
Passwort-Bestätigung

Fig.3 Editing the user information

3.3 List Of The Actual Offer

A few weeks before the semester starts, the course offer has to be set up. Now all registered students have the possibility to subscribe to one or more courses until the deadline (e.g. 30 days after inserting or the start of the semester) or the user limit has been reached. These to limitation have been implemented to give the instructors room for maneuver. Knowing the number of attending students before the lecture starts gives the ability to get a sufficient large enough room or enough copies of lecture materials. The offer can simply be a lecture, an exercise, a trip with a restricted number of users or also tests and examinations. Once clicked on the 'Register' button (see Fig. 4), the student will receive an email with the course he has joined and what data has been stored for that purpose. An example email would look like:

You have registered for: 100,
Business Informatics 1

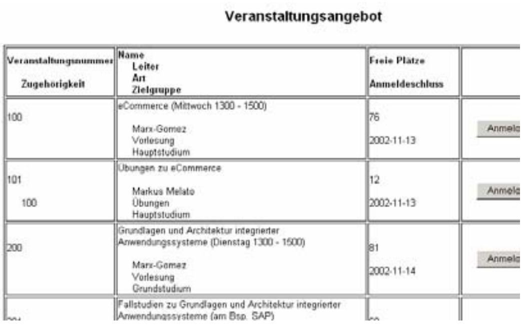
Following data has been
stored:
Register number:
123456 Courseid: 100

If one this data is falsified
please reply to this message
with the correct data.

Best regards

Your eVE-Team

Clausthal-Zellerfeld, January
22nd



Veranstaltungsangebot

Veranstaltungsnummer Zugehörigkeit	Name Leiter Art Zielgruppe	Freie Plätze Anmeldeschluss	
100	eCommerce (Mittwoch 1300 - 1500) Marx-Gomez Vorlesung Hauptstudium	76 2002-11-13	Anmeld
101 100	Übungen zu eCommerce Markus Melato Übungen Hauptstudium	12 2002-11-13	Anmeld
200	Grundlagen und Architektur integrierter Anwendungssysteme (Dienstag 1300 - 1500) Marx-Gomez Vorlesung Grundstudium	81 2002-11-14	Anmeld
	Fallstudien zu Grundlagen und Architektur integrierter Anwendungssysteme (am Bsp. SAP)		

Fig.4 Selecting a course

If the actual offer does not have enough places for an examination date, the course leader is able to put new dates, when possible, online with the help of the system administrator.

3.4 Selected Courses

Every student can take a look at his selected courses or examinations. If the end time has not been reached yet, you can deselect your courses, but not your inscriptions into examinations.

This overview of the selected courses will give the students a personal schedule once eVE is implemented throughout the university.

3.5 Admin-Zone

eVE has two types of administrators. First there are system-administrators, they have control over all functions, and then those users, that only need to get an overview about course participants like instructors and/or secretaries. This separation is necessary because users who only need to know who or how many student in a specific course are do not need the right to set new courses up, they are primarily responsible for time and room schedule planning and the information about this. Entering as system-administrator you can:

- *Add/Change user information*: see 3.6
- *Administrate actual offering of courses:* see 3.7
- *Create filter based results of inscribed students, usage grad of courses, etc.:* see 3.8
- *Show students that want to visit more than 3 courses:* see 3.9
- *Send emails to all course users*: see 3.10

The secretary mode includes only the view of filter-based results and eVEMail due to the responsibilities stated above.

3.6 Add/Change User Information

First the administrator gets a list with all registered users in the system, he then has the option to change the information of any selected user or even delete him if he is no longer a member of the Technical University of Clausthal. The possibility of an administrative user password change is given, should some students encounter problems with the automatic password generator. Like in the student registration form (see 3.2) the email address is checked with a regular expression, whether it is a correct one or not. Being incorrect the user is reminded to state a correct one. All the other fields of the form are checked whether they are valid too.

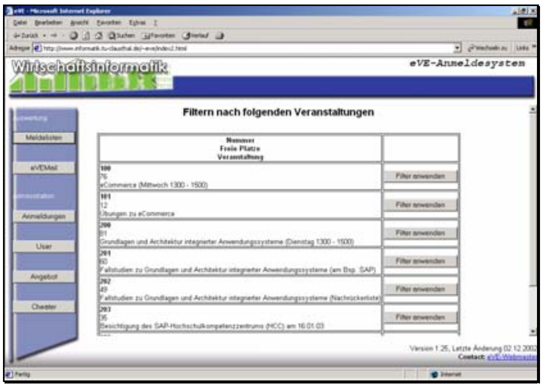
3.7 Administrate Actual Offering Of Courses

Every new semester most of the courses will change, so it is necessary to update eVE's Database with a predefined form. Like in the User Information form the administrator has the right to either edit a course or delete it completely. This form also features the possibility of changing the deadline or student limitation, named above. The inserted course are presented according to their relation to one another, so that courses attached to a major course are together in their logical order.

3.8 Create Filter Based Results

As an information for instructors or for printing out lists, eVE has integrated filters that show all subscribed users in the different courses or examinations with all needed information (see. Fig. 5). So you can plan course rooms and study material. Filter Functions are MySQL SELECT instructions with a differing WHERE clauses depending on the chosen filter number. [Welling 2001]:

```
SELECT * FROM registrations
WHERE PRUEFID='$filternr'
ORDER BY
PRUEFID,MATRIKELNUMMER
```


eVE - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Suchen Favoriten Verlauf
Adresse http://www.informatik.tu-clausthal.de/~eve/index2.html
Wechseln zu Links
Wirtschaftsinformatik
eVE-Anmeldesystem
Auswertung
Meldelisten
eVEMail
Administration
Anmeldungen
User
Angebot
Cheater
Fertig
Internet

Filtern nach folgenden Veranstaltungen

Nummer Freie Plätze Veranstaltung	
100 76 eCommerce (Mittwoch 1300 - 1500)	Filter anwenden
101 12 Übungen zu eCommerce	Filter anwenden
200 81 Grundlagen und Architektur integrierter Anwendungssysteme (Dienstag 1300 - 1500)	Filter anwenden
201 60 Fallstudien zu Grundlagen und Architektur integrierter Anwendungssysteme (am Bsp. SAP)	Filter anwenden
202 49 Fallstudien zu Grundlagen und Architektur integrierter Anwendungssysteme (Nachrückerliste)	Filter anwenden
203 35 Besichtigung des SAP-Hochschulkompetenzzentrums (HCC) am 16.01.03	Filter anwenden

Version 1.25, Letzte Änderung 02.12.2002
Contact: eVE-Webmaster

Fig.5 Creating filter-based results

3.9 Cheater Buster

Mostly there is a limit of users for a course so that after the subscription not all persons will get a place and the list hast to be cut down to the allowed number of students. Many students inscribe in lists only for having more possibilities to choose, so the Cheater Buster shows these persons. The eVE-Admins can send emails to those persons and warn them to deselect the courses the subscribed to twice. If these mails are ignored the users can be deselected by the administrators so that the normal functionality of eVE is secured.

3.10 eVEMail

To contact all students in a selected course, the instructor can sent emails to them. This gives the professors the possibility to inform the students if a lecture has to be cancelled or the homework is due a week later than planned. This for of messaging is even more secure then the old one used at the university, where lists were passed through in the lectures and everyone one could read everyone’s email addresses, which is a violation of every privacy law there is. eVEMail gets the addresses for every student out of the database and sends a mail to every student alone, so that there is no possibility that anyone can get access to this address except the receiver of the message.

The eVE interface is very easy to understand. It has a big text box for the message content and a small one for the senders name. The rest of the message like the user address is generated automatically.

4 CONCLUSION AND OUTLOOK

After only one semester in use at the Institute for Informatics department for Business Informatics eVE is bound to be expanded to the rest of the Institute. Due to this expansion eVE must become more flexible. For that purpose eVE^2 is being developed. It will include following features:

- expand eVE on any other information sections. eVE is ready to manage the whole lessons and examinations offer of the institute of information technology.
- implement more information functions for students and administrators in addition to users demand.
- implement a semester and study specific course offer for the students.
- connect the professor/ instructor selection and any other possible selection field to the institutes main page for easier maintenance and a better data integrity.

- implement a Java-Script menu for an easier structure of the administration part.

- check the possibilities of an authentification via personal keys. Only as an option, because not everyone has such a person key.

- administration through user rights. That would make the 2 administrators obsolete and allow a sorough definition of rights for each user.

- department and institute specific designs through XSLT and XML.

Being in use for just one semester, eVE is not only accepted by the professors and assistants due to the smaller administrative overhead, but also by the students, who can now comfortably enlist to courses from home.

REFERENCES

[URL1] http://www.mysql.com
01-22-2003

[URL2] http://www.php.net
01-22-2003

[URL3] www.apache.org
01-22-2003

[URL4] German guidance, references, tutorials and platform for PHP
http://www.phpbox.de/php_tutorials/formularversenden1.php
01-22-2003

[URL5] written By the PHP Documentation Group, Edited by Stig Sæther Bakken et al.
http://www.php-center.de/en-html-manual/
01-22-2003

[URL6] Guido Stepken; My-SQL Datenbankhandbuch
http://www.rent-a-database.de/mysql/
01-22-2003

[URL7] The MD5 Message-Digest Algorithm
http://theory.lcs.mit.edu/~rivest.rfc1321.txt
01-29-2003

[Brass 2002] Brass, S. : database lecture notes, Clausthal 2002 [http://www.informatik.uni-giessen.de/staff/brass/db_w02]

[Ullman et al. 2002] Garcia-Molina, H.; Ullman J. D.; Widom J.: Database Systems – The Complete Book, Prentice Hall, New Jersey 2002

[Kline and Kline 2001] Kline, K.; Kline D.: SQL IN A NUTSHELL, O'Reilly, Köln 2001

[Krause 2001] Krause, J.; PHP 4 – Webserverprogrammierung unter Windows und Linux, Hanser, Berlin 2001

[Nagel 2001] K.Nagel: Informationsbroschüre zum Bundesdatenschutzgesetz („Information about the German Federal Data Privacy Law", in German), Oldenbourg 2001

[Welling 2001] Welling, L.; Thomson L.; PHP and MySQL Webdevelopment, Sams Publishing, 2001

AUTHOR LISTING

AUTHOR LISTING

AUTHOR LISTING