

Volume 25 Number 2 July 2001

ISSN 0350-5596

Informatica

**An International Journal of Computing
and Informatics**

Special Issue:

**The Changing University,
and the Role of Information Technology**

Guest Editor:

Jan Knop

Viljan Mahnič



The Slovene Society Informatika, Ljubljana, Slovenia

Informatica

An International Journal of Computing and Informatics

Archive of abstracts may be accessed at USA: <http://>, Europe: <http://ai.ijs.si/informatica>, Asia: <http://www.comp.nus.edu.sg/liuh/Informatica/index.html>.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.

The subscription rate for 2001 (Volume 25) is

- USD 80 for institutions,
- USD 40 for individuals, and
- USD 20 for students

Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

\LaTeX Tech. Support: Borut Žnidar, Kranj, Slovenia.

Lectorship: Fergus F. Smith, AMIDAS d.o.o., Cankarjevo nabrežje 11, Ljubljana, Slovenia.

Printed by Biro M, d.o.o., Žibertova 1, 1000 Ljubljana, Slovenia.

Orders for subscription may be placed by telephone or fax using any major credit card. Please call Mr. R. Murn, Jožef Stefan Institute: Tel (+386) 1 4773 900, Fax (+386) 1 219 385, or send checks or VISA card number or use the bank account number 900-27620-5159/4 Nova Ljubljanska Banka d.d. Slovenia (LB 50101-678-51841 for domestic subscribers only).

Informatica is published in cooperation with the following societies (and contact persons):

Robotics Society of Slovenia (Jadran Lenarčič)

Slovene Society for Pattern Recognition (Franjo Pernuš)

Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)

Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)

Automatic Control Society of Slovenia (Borut Zupančič)

Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)

Informatica is surveyed by: AI and Robotic Abstracts, AI References, ACM Computing Surveys, ACM Digital Library, Applied Science & Techn. Index, COMPENDEX*PLUS, Computer ASAP, Computer Literature Index, Cur. Cont. & Comp. & Math. Sear., Current Mathematical Publications, Cybernetica Newsletter, DBLP Computer Science Bibliography, Engineering Index, INSPEC, Linguistics and Language Behaviour Abstracts, Mathematical Reviews, MathSci, Sociological Abstracts, Uncover, Zentralblatt für Mathematik

The issuing of the Informatica journal is financially supported by the Ministry for Science and Technology, Slovenska 50, 1000 Ljubljana, Slovenia.

Post tax paid at post 1102 Ljubljana. Slovenia tax Percue.

Introduction:

The changing university, and the role of information technology

"The Changing University, and the Role of Information Technology" was motto of the 7th International Conference of EUNIS - European University Information Systems Organization. The conference was held in March 28-30, 2001 at the Humboldt University in Berlin and brought together 470 participants from 39 countries. For the first time in the short history of EUNIS this conference was organized to be interdisciplinary, covering the application of Information Technology (IT) in computing centres, libraries, multimedia centres, and university management. More than 150 scientific and corporate presentations from 26 countries dealt with the dramatic changes that have taken place at our universities in the areas of research, teaching, learning, and administration as a consequence of the digital revolution. Meeting this challenge demands a continuous renewal, reassessment of old procedures, search for new solutions, abandonment of old patterns, and encouragement of professional growth. The idea is to focus on the university as a whole and to search for new frontiers for the application of IT.

By the courtesy of Professor Matjaž Gams, Executive Associate Editor of *Informatica*, we were given the possibility to publish some of the best papers presented at the Conference in a special issue of *Informatica*. Taking into account the different areas of the use of IT at universities we selected 12 papers and collected them into four groups:

- Supporting Change in Teaching and Learning,
- Information Systems and Data Warehousing,
- High Performance Computing and Multimedia, and
- IT and Network Computing Security.

The first group consists of four papers. The first, *Wizards of OZ - Change in Learning and Teaching* by Coy and Pirr, describes a distance learning environment for lectures, seminars and practice in computer science at the Humboldt University in Berlin. The environment relaxes spatial and temporal restrictions of teaching by interconnecting rooms at different campuses of the university. The second, *A Hybrid System for Delivering Web Based Distance Learning and Teaching Material* by Greenberg, describes the experience and plans of The Open University in providing courses that combine the usage of the Internet and DVD technology. The large data storage capacities and versatile functionality of DVD technology can resolve the problems of limited bandwidth to student's homes which prevents effective on-line delivery of data-intensive media such as video and high resolution still images. The third, *Learning by Experience: Networks in Learning Organizations* by Kuitinnen et al., describes a course entitled *Networks in Learning Organization* that will be delivered entirely on

the web at the Virtual University of Finland. An important aim of the course will be to teach students how to use network-based collaborative software tools for communication within a given organization in ways that contribute towards the development of that organization. The last paper in this group, *How to Learn Introductory Programming over the Web* by Haataja, Suhonen and Sutinen, describes some experience in delivering the introductory programming course at the University of Joensuu, Finland. It was found that the learning process could be improved by using interactive visual tools, adaptive learning materials, and more intensive collaboration.

Papers from the second group deal with administrative and management information systems. In the first paper, *Information System Supporting CATS*, Ryjáček et al. describe an information system that supports on-line student registration for courses and provides a support for managing information on study programs, timetables, student records, agenda of admission, alumni records etc. The system was developed at the University of West Bohemia in Plzen and is now used by six Czech universities. In the second paper, *A Data Warehouse for French Universities*, J-F. Desnos describes the French national project that aims to build a global data warehouse containing data from different administrative systems (students system, financial system, and personnel system), local applications, and data coming from outside. A brief description of the target database, software tools used, extraction, transformation and loading process as well as target queries is given. The last paper in this group, *Data Quality: A Prerequisite for Successful Data Warehouse Implementation* by Mahnič and Rožanc, explores the possibility of building a data warehouse at the University of Ljubljana, Slovenia, with an emphasis on data quality. An assessment methodology to empirically determine the data quality is described and the results of the assessment are presented.

The third group consists of three papers. The first, *'Beowulf Cluster' for High-performance Computing Tasks at the University: A Very Profitable Investment* by Galan et al., describes the advantages of the Beowulf Cluster that make it specially suitable for the university environment. An implementation of such a cluster using commodity equipment and dedicated to run high-performance computing tasks at the University of Las Palmas, Spain, is described. The second paper, *Evaluation of Codec Behavior in IP and ATM Networks* by Naegele-Jackson et al., compares a codec for teleconferencing and teleteaching applications over an IP network with a MJPEG codec for similar applications over ATM. Comparison concentrates on the Quality of Service parameters delay, jitter and subjective picture quality. The third paper, *An Environment for Processing Compound Media Streams* by Feustel et al., de-

scribes a system for processing media streams that are composed of multimedia data objects. The system is based on a hypermedia data model of reusable object components and consists of an intelligent media database, a Web-authoring tool, and a time directed presentation stream.

Papers in the last group deal with security issues. Linden et al. in FEIDHE - Integrating PKI in Finnish Higher Education describe the key issues affecting the implementation of a public key infrastructure (PKI) based identification system with smart cards in Finnish higher education, while Strachan et al. in Information Systems Delivery in a Tiered Security Environment describe how the security requirements were incorporated into the design of the network and information services at the University of Paisley, Scotland.

We hope that the selected papers give a good overview of different aspects of the use of IT at European universities. We also believe that this special issue will contribute to further promotion of activities of EUNIS with the aim of encouraging the communication and transfer of information between information system providers in higher education establishments in Europe. Once more, we would like to thank Professor Matjaž Gams for his help.

Editors of the Special Issue,

*Jan Knop
Viljan Mahnič*

Wizards of OZ –

Change in learning and teaching

Wolfgang Coy and Uwe Pirr
 Humboldt-Universität zu Berlin
 @informatik.hu-berlin.de
 Unter den Linden 6, D-10099 Berlin
 URL: waste.informatik.hu-berlin.de

Keywords: e-Learning, tele-teaching, multimedia in teaching and learning

Received: February 7, 2001

OZ is a distance learning environment for lectures, seminars, and exercises in computer science at Humboldt-University, where "OZ" means distributed in space (ortsverteilt) and independent of time (zeitversetzt). The focus is on computer-"exercises", which has certain implications on the technical setup: the back channel is very important and the teacher's presentation as well as the student's computer screens is visible on both places. Different modes of use are possible. Two rooms at different campuses of the university are connected by the universities internal high-bandwidth, but non-reserved, TCP/IP-connection. Each one is equipped with two data projectors, where one is used as white board mirroring the teacher's or a student's computer screen, while the other shows the teacher and the opposite room inhabitants. Each room has its own control computer that is also used for video or audio streaming. By the use of wireless connected laptops a comfortable degree of freedom for teachers and students is achieved. Despite this, we use low-cost hardware and standard software wherever possible. The audio-taped lectures are combined with the screen materials, mostly PowerPoint slides enhanced with QuickTime media, so that lectures as well as other stored materials are accessible via Internet. After more than two years of thorough experiments the installation has reached now a state where it is relatively stable, so that it is transferred now for regular use in our teaching activities

1 Introduction

University teaching has always been an activity reflecting two quite contrary challenges, namely the motivated acquisition of knowledge and skills versus a proof of professional abilities documented by examinations and degrees. This strange mixture of demands was met over centuries by canonical forms of teaching and learning: lectures, exercises, seminars and practical training, followed or accompanied by examinations. While societies were under a strict pressure of technological change during more than two centuries, the teaching environment was kept stable to an astonishing degree.

The main technical innovation after the invention of lectures and universities in medieval times was the use of printed books, the introduction of blackboards and quite recently copying machines and overhead projectors.

Now computers and telecommunication allow another step that may resolve spatial and temporal unity of the classroom. This is a potential that has to prove its value by experiment.

2 Technical Means of Learning & Teaching at University Levels

The use of telecommunication and computing technology meets several educational demands in post-industrial societies. First, education at academic levels often becomes a necessity in a complex world of decision-making and high-skilled technological interactions. Second education guarantees no longer life long job security. Labor flexibility, as it is demanded, means more and more the ability for life-long learning and studies. Universities still do not reflect these changing demands from society and economy.

Fragmented Studies

As training on the job is not sufficient in highly complex fields of academically trained labor more and more short sequences of training and productive work will define a job career. As a simple consequence of such an insight, school and university studies can no longer establish complete education. As a consequence these phases should be

shortened and institutions have to be opened for on-going education processes - or high level education will be delegated to other institutions.

In fact, we already have experiences with such structures though they are still not dominant. Distance learning, at the British Open University or at comparable institutions, does exist as well as the support of some of these demands in on-going education or add-on university training. Some universities also offer supplementary studies, but this all may of restricted value if only added to a traditional curriculum - it makes much more sense if it is considered as training after some job experiences. In general, we feel that Universities are not very open to such ideas. This, of course, fragments the traditional studies into smaller consequent pieces.

Part-time students

Part-time studies are another example of fragmented education. We must accept the fact that many students especially in, but no way restricted to, informatics and IT-related fields are actually working in these fields during their university studies. The curriculum must reflect the existence of these part-time students. They should be supported much more than it is actually done.

Unavoidably this means also a stronger inclusion of practical experiences into the curricula - either as an underlying assumption, namely that students are already practicing their field of studies to some degree or as an explicit part of the university studies.

Life-long learning

We actually are in the process of fragmenting education, dispersed over the student's whole lifetime. No longer can we rely on the traditional sequence of "school-university-job". This sequence will lose its character as the standard type of education. Of course self-studies are some answer to the growing demand of adult training, but they show also distinct disadvantages. So the idea of Internet based distance learning is propagated for a revised form of self-studies. It seems to be too early to judge the practicability, but there are promising field studies.

Fragmenting discipline

If education will become a life long effort, we should not expect that the sequence of education

blocks follows traditional curricula as they were constructed by different disciplines. A life long learning process will very naturally assemble different pieces of curricula from different disciplines. It is not difficult to foresee that law and economics or management studies will become standard elements in the life long process for engineers or computer scientists. Contrary many trained in social or cultural sciences will include elements from computer science into their life long curriculum.

This cannot be without influence on the disciplines as they are formed now. New disciplines may arise (like e.g. chemical engineering, media informatics, or media economics) but this will not be the final answer. After some period of adaptation, we will become accustomed to complex mixes of diverse disciplines, which will be based in a variety of disciplines.

New technologies

With the advent of new technologies there seem to be some answers to the challenge of new educational structures worth to be investigated.¹ In the „Wizards of OZ“-project we demonstrated that new technologies allow new forms of online-teaching and learning as well as related offline self studies. By the use of computing technology and telecommunications we constructed an environment where spatial and temporal restrictions of teaching were relaxed to some degree. Our first experiences show that new forms of teaching could be (and to some extent must be) founded on traditional forms.

3 The Wizards of OZ Project

OZ is a tele-learning environment for lessons, seminars, and exercises in computer science at Humboldt-University, where "OZ" means distributed in space (*ortsverteilt*) and independent of time (*zeitversetzt*).

Connecting places

We use two rooms at different campuses of the university; one is placed at the central campus in Berlin-Mitte, the other at Berlin-Adlershof. The distance between these places is one hour by public transportation. The rooms are connected by the university's internal high-bandwidth, but non-reserved, TCP/IP-over-ATM-connection. Each

¹ R. STEINMETZ, *Multimedia Technology - Fundamentals and Introduction*, Berlin: Springer-Verlag, 1993

place is equipped with two data projectors, where one is used as white board mirroring the teacher's or a student's computer screen, while the other shows the teacher and the opposite room inhabitants. Each room has its own control computer that is also used for video or audio streaming. The setup is completely symmetrical in its functions and in fact, lectures are given from Berlin-Mitte as well as from Berlin-Adlershof without special arrangements in advance. The same holds for exercises.

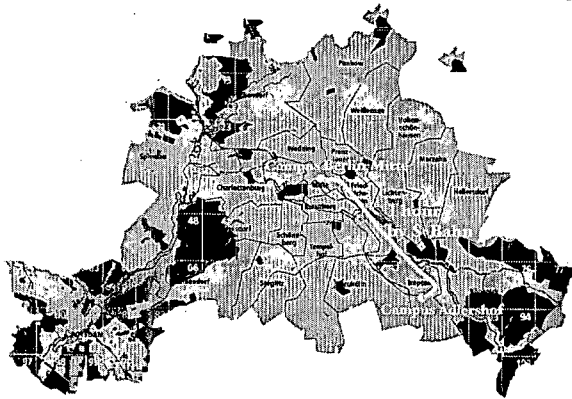


Fig.1 Connecting places; one room is at the main campus in Berlin-Mitte, the other at Berlin-Adlershof. This distance is one hour by inner-city train (S-Bahn).

By the use of wireless connected laptops a comfortable degree of freedom for teachers and students is achieved (actually only in one room for experimental purposes). In general, we use low-cost hardware and standard software wherever possible. This was a basic economic decision for the whole project design and we find it well justified after more than two years experience.

Tele Exercises

A focus is on tele-“exercises”, which has certain implications on the technical setup: the back channel is very important and the teacher's presentation as well as the student's computer screens has to be visible on both places. Different modes of use are possible.



Fig.2 Teaching in Adlershof

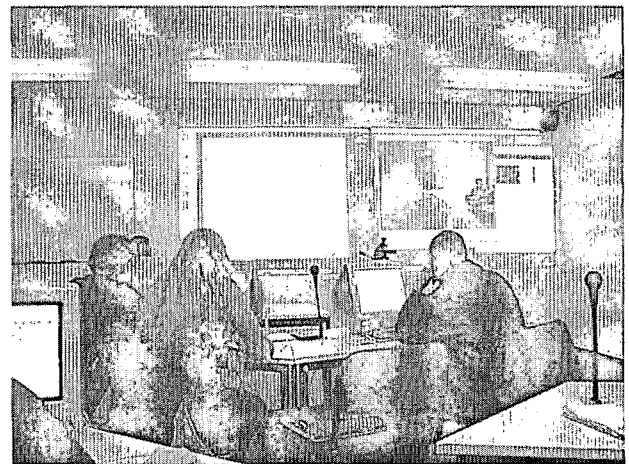


Fig.3 Remote listeners in Berlin-Mitte

Online- and Offline Multimedia-Materials

The audio-taped lectures are combined with the screen materials, mostly PowerPoint slides enhanced with QuickTime media, so that lectures as well as other stored materials are accessible via Internet and as CD-ROMs. After some experiments with video- and audio-tapes, we decided that audio-enhanced white-board will be sufficient for most purposes of training and repetition. But this decision reflected also bandwidth and storage considerations. Broadband internet connections like DSL- or cable and DVD-storage will allow us to reconsider these design decisions, as we did already in some experimental setup.²

² For a different approach cp. R. MÜLLER, T. OTTMANN, The „Authoring on the Fly“ system for automated recording and replay of (tele)presentations, *Volume 8, Issue 3, pp 158-176, Multimedia Systems*

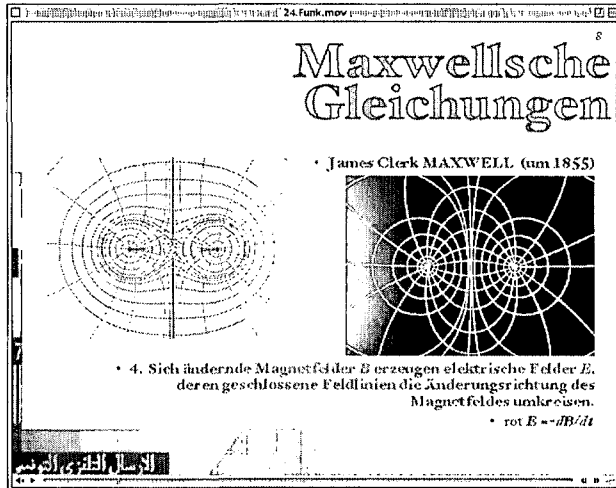


Fig.4 Online- and offline material: PowerPoint slides enhanced by audio-stream

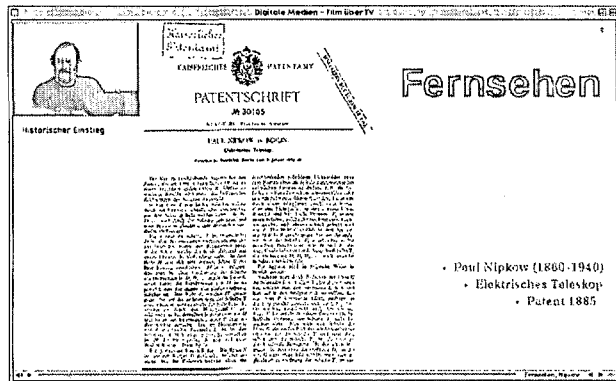


Fig.5 Online- and offline material: PowerPoint slides enhanced by video- and audio-stream

4 Some Considerations

After more than two years of practical experiences we come to some first conclusions. Though they are still mere hypothesis, they may be used as guidelines for further explorations and experiments.

- University teaching is standing in a long historical tradition. Despite all obvious problems it has demonstrated over the centuries some inherent positive qualities. Before we expect that new technologies may succeed classical forms of teaching, we should keep in mind that the printing press improved university teaching but it did not turn off lecturing. We

may note that lectures are not defined by pupils copying sentences read aloud by a teacher. Lectures exhibit a complex social and cultural context. It is much more likely that technology may enhance lectures than to extinguish it.

- Lectures and exercises must be well prepared – not only by content, but also by the used technology. This means a *substantial additional amount of preparation time* as well as knowledge and proficiency with the media used.
- There is a certain danger with *perfect media demonstrations*. Every lecture that reaches at the frontier of scientific knowledge can probably not avoid a touch of incompleteness and a flux from stable basics to yet unsettled fields still waiting for a better didactical treatment. Perfect media finish may give a wrong impression of a closed body of knowledge that may not motivate students for a deeper understanding. Therefore the emphasis on media technology must not be overstretched. It is not TV finish that must be reached. Student *motivation* is much more important than technological perfection.
- Technology should become *transparent* in the process of teaching and learning. This goal will probably never be reached in a perfect way. E.g. the pages of a book must be turned over again and again while reading: Though this introduces a break in the reading process, we are usually not aware of it. Similarly we accept some technical irritations with computers and telecommunication equipment, but they should never determine the whole work flow.
- Actual technology suffers, at least under the dictate of economic resource planning, from imperfections: insufficient bandwidth for video and audio streams, lack of storage capacities, slow computing speed, or bad codec qualities, to name a few. Even if they improve, we must cope with these imperfect situations for a long time. That means, we must try to use *cost-effective solutions*: standard equipment, standard software, preferable from public domain, or shareware origin. Open source developments may help to adapt software when necessary.

³ For a comparable but distinctly different setup cp. Ch. Zimmer, L. Meyer, V. Pipek, B. Schinzel, A. Wegerle, M. Won, V. Wulf, Erfahrungsbericht zur Telelehrveranstaltung "Informatik und Gesellschaft" im Sommersemester 1999, IIG-Bericht 1/2000, Dez. 2000, Freiburg: Universität Freiburg – Institut für Informatik und Gesellschaft

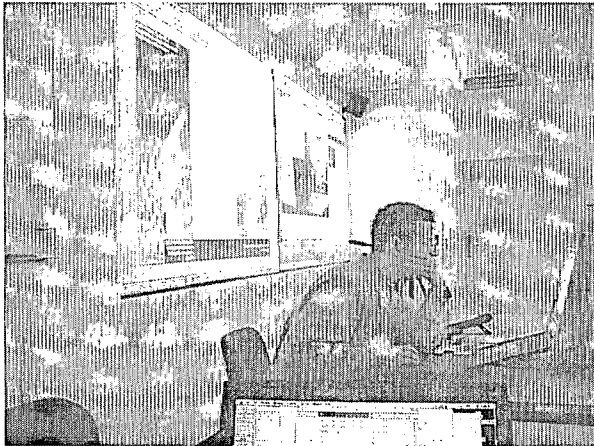


Fig.6 Exercises in Berlin-Mitte

- The process of teaching and learning is context-sensitive as well as content-sensitive. There is no general form of lectures, seminars, and exercises throughout the disciplines. That means, technological decisions for tele-teaching must be adapted to actual needs. In our case, we found synchronous bi-directional audio to be the most important factor, followed by responsive whiteboard demonstrations. Video quality for teacher close-ups were considered to be much less important. In fact, we omitted this video channel completely in the CD-ROM and web-versions of the lectures (a decision to be postponed with higher bandwidth and larger storage).

We assume that changing the structure of teaching and learning will not be without influence on the structure of the underlying disciplines. Technology demands a stricter modularization of single courses – and that in turn generates modules for better structured disciplines. Our courses were accepted by diploma students in the computer science department as well as additional courses by students in various fields from egyptology, cultural studies, mathematics, or language studies.

5 Outlook

After more than two years of thorough experiments the installation has reached a state where it is relatively stable, so that it is transferred now for regular use in our teaching activities. Students acted favorably. They are well aware of the organizational advantages of space distributed teaching and they made good use of the CD-ROMs. Last semester, by a poll result among all students of the department the lecture was given a prize for outstanding teaching. This validates our belief that the use of technology must not degrade the quality of lectures

and exercises even under otherwise less favorite conditions.

6 Acknowledgements

The research project OZ was established between the Computing Center of the Humboldt-University and the Institute of Informatics. It is financially supported by the German federal government via DFN-Verein.

We also acknowledge the cooperation of Peter Schirmbacher, Roland Kubica, Jochen Koubek, Henrik Pantle, Andreas Beck, and, as a special guest, Lena Bonsiepen.

A hybrid system for delivering web based distance learning and teaching material

Joel Greenberg
 Learning and Teaching Services, The Open University,
 Walton Hall, Milton Keynes, MK7 6AA, England
 Phone: +44 1908 653424
 j.greenberg@open.ac.uk

Keywords: web based learning, DVD-ROM, connected DVD, distance education

Received: February 4, 2001

There is a growing expectation from distance learners that their learning and teaching environment will be on-line. Limitations on bandwidth to the home has so far constrained the richness of such environments. A hybrid solution to this problem combines the immediacy of the Internet with the versatility of DVD Technology. This versatility allows the disc-based material to be integrated with web based material or used off-line as a self-contained learning environment. All content which is not computer dependant such as audio-visual material, can be viewed with a domestic DVD player from the same DVD disc.

1 Introduction

Over 150,000 students register with the Open University each year, including 5% non-UK EU students and 10% outside the EU. The University is considered by many to be the world's leading distance learning institution. All students of the University are offered a comprehensive advice, guidance and learning support service, starting from the initial point of enquiry through to completion. A full range of media has traditionally been used to support students, including a strong telephone-based advice and guidance service, student toolkits on study skills, TV programmes, group and individual face-to-face support from course tutors and residential and day schools. Learning and teaching materials have been sent to students in a number of ways including print, broadcast television and radio, videocassettes, audiocassettes, home experiment kits and CD-ROM. The Open University is delivering computer-based learning and teaching materials to over 80,000 students and has over 140,000 users of its on-line services.

The quality and effectiveness of the University's teaching is monitored through the collection, analysis and dissemination of data about the strengths and weaknesses of the materials and services provided, and the quality of the student's experience and learning outcomes.

Web-based advice and guidance, email as an advisory medium and the use of computer-media conferencing for teaching and learner support, are expanding across all services. CD-ROM has grown dramatically in the last few years as the primary distribution media for computer-based learning and teaching materials.

The Open University aims to establish the critical baseline of IT elements for all courses and programmes by 2002; build IT elements into courses to achieve compulsory IT elements for all University degrees by 2005; increase Web focused courses to at least 20 by 2002. Several key environmental factors have influenced these target levels and the way the University is responding to the challenge presented by them. These are:

- the rapid growth of the Internet and prospective student expectation that courses will be available on-line.
- limited bandwidth to the home with prevents effective on-line delivery of data intensive media such as video and high resolution still images.
- the large data storage capacities and versatile functionality of DVD technology.

Software tools are now available which take advantage of the immediacy of the Internet and the versatility of DVD technology. Hybrid developments of this kind are referred to as "Connected DVD" and web based student learning environments can be developed which use DVD-ROM media to hold up to 9 GB of learning and teaching material. The versatility of DVD technology allows the disc based material to be integrated with web based material or used off-line as a self-contained learning environment. All material which is not computer dependant (audio, video, images) can be viewed with a domestic DVD player from the same DVD disc.

The Open University is well positioned to exploit this technology in its teaching. Over 80,000 students are

already using personal computers in their course work and an increasing number have DVD-ROM drives in their home machines. This approach may offer improved media integration and a reduction in the University's learning media production costs.

2 What is DVD?

DVD originally stood for Digital Video Disc and was intended for the distribution of video. It now officially stands for Digital Versatile Disc as its uses now include standalone audio and distribution of software and data. Physically it is the same size as a CD-ROM, but the laser beams used are much finer and the rate of rotation of the disc is faster (Taylor 2001).

This data can be played by any suitable application, but a key part of the DVD standard is that there is a "standard player" which can either be in the form of a standalone (table-top) player connected to a TV or a software player on a computer. The player interprets the data on the disk in a special way to create a substantially interactive experience. This includes multiple camera angles, multiple language audio tracks, multiple sub-title tracks, still pictures and interactive screens (or video) called menus.

There are emerging standards for extensions to this player which allow display of web pages at key points and also to allow web pages to display DVD video in a convenient manner.

DVD authoring consists of encoding video, audio, text data, creating interactive screens and scripts, and organising the arrangement of all these resources on the disk so as to work appropriately with the standard DVD player. After testing this (including the writing of DVD-R discs) then a tape is created which is sent off together with label graphics etc. for the pressing of a large number of DVD discs. Additional software and files can be included on a DVD and these can include software which incorporates the playing of the encoded DVD files in ways not possible with a DVD player.

The DVD-ROM format can be used as a distribution medium, offering up to 15 times the capacity of the CD-ROM. A move from CD-ROM to DVD-ROM and VHS to DVD distribution would result in a considerable saving to the University as has a move from floppy disk to CD-ROM. The DVD-ROM format in most common use in Europe is DVD-9 which holds around 9GB of data on two layers of a single-sided disc. DVD-18 production is now underway in the USA and this format holds 18 GB on two layers on both sides of the disc.

Recent projections for DVD household penetration in Europe show a growth from 14% in 2000 to around 80% by 2005. DVD-ROM drives are now rapidly replacing CD-ROM drives in desktop PCs and an upgrade to a DVD-ROM drive is now an inexpensive option. DVD-

ROM drives are compatible with almost all CD formats including CD-ROM and CD Audio.

The Open University is currently working with two systems: the Spruce Maestro DVD authoring system and the Daikin Scenarist DVD authoring system.

3 Applying the Technology

3.1 Applications in Science

S103 Discovering Science is a wide-ranging course that introduces important scientific concepts and develops the skills needed to study science successfully. It introduces the disciplines of biology, chemistry, Earth science and physics and shows the links between them. The course is designed both as a broad foundation for students who intend to study for a science degree, and as a stand-alone course for those who want to discover more about the science of the world around them. As well as full colour books, students receive interactive CD-ROMs, videocassettes, TV programmes, and a practical kit complete with rock specimens and fossil casts. In addition, students are allocated to regional tutors, and they can participate in computer conferences and attend a residential school. All of these activities and materials form an integrated teaching package where each component is used for the purpose to which it is most suited. The course offers such a wide range of media and delivery systems, that it is ideal for piloting the Connected DVD concept with real Open University learning and teaching material. The course in its current presentation format includes:

3.1.1 Books, Study Files and Study Guides

Students receive eleven printed texts which are the main component of the teaching materials and are specifically written by academic authors at the University. Each text is accompanied by a loose-leafy *Study File* containing notes and activities to help the student track their progress through the learning materials. A *Study Guide* card is provided for each text to help students plan and co-ordinate their work.

3.1.2 Interactive Media

Students receive 25 interactive multimedia tutorials on CD-ROM to teach topics which are difficult to present in printed texts. They also receive software to assess and give practice in mathematics skills and self-assessment questions which help identify areas where students require more work to improve their understanding or skills.

3.1.3 TV and Video

There are ten 30 minute television programmes broadcast by the BBC and 5 hours of videocassette material which contains 20 video activities.

3.1.4 Practical Work

There are eight practical activities carried out at home by students.

S216 Environmental Science is planning to send students a number of interactive multimedia applications and all of the course video material on DVD-ROM in its first year of presentation (2002). The course team believes that students will benefit from a reduction in the number of delivery platforms used in the course.

4 Building the DVD-ROM

All of the learning and teaching materials produced for *S103 Discovering Science* have been put in digital format and a pilot DVD-R has been produced which contains all of the S103 software and text material. All University text material is now held in PDF format so text searching functionality is easy to implement with the Adobe Acrobat Reader. All of the video material has been digitised and various compression options are being investigated. The interactive media material has been easy to integrate and access has been provided through a common icon based interface.

The development strategy is to build a DVD-ROM which holds all of the S103 text, application software, much of the video material either broadcast or sent to students on videocassette and which has links to web based material. All of the material will be accessible from a PC with a DVD-ROM drive and the video material will be playable on a domestic DVD player. Apart from providing S103 students with an invaluable resource, the application will demonstrate the potential of this new technology to higher education. This kind of development will help overcome the concerns of many academics about the Internet and the difficulty of integrating high quality media with a web site.

5 Applications in the Humanities

A number of Arts Faculty projects have been initiated including a pilot project for *A220 Princes and Peoples: France and the British Isles 1620–1714* which looks, for example, at the structure of three contrasting church buildings of the seventeenth century. Other courses which may include DVD based material include *A207 From Enlightenment to Romanticism*, and *A218 History of Medicine*. Both of these courses do not go into their presentation phase until 2004 which gives time to evaluate the use of standalone and computer based DVD players by Arts Faculty students. A218, for example, may use CD-ROM during the first half of the course life but may then convert its material to a mixed mode format containing both DVD-Video and DVD-ROM components during the second half of its presentation phase.

6 Technical Issues

6.1 Storage Limitations

The most common DVD-ROM format in mass production today, DVD-9, consists of two data layers on one side of a disc, each layer holding around 4.5 GB. The S103 material requires over 13GB of data storage (text and software: 4.5GB, TV programmes in MPEG-1: 4GB, video programmes: 4.8GB). The amount of data storage required by video material depends on the compression format adopted. Opting for higher data rates during compression results in larger data files but better image quality and the MPEG-1 format (1 Mbits/sec) offers a reasonable data storage/image quality compromise. As there are no direct links from the text material to the broadcast programmes, these will be put on a separate DVD-5 disc which has one data layer.

6.2 Video Encoding

While MPEG encoding has yielded some impressive results, there is a problem in creating a format suitable for domestic DVD players and PC based software players. Most PCs with DVD-ROM drives are shipped with software DVD players and the various DVD authoring products have their own software players. Therefore, rather than relying on a wide range of DVD players on student's machines, we are likely to bundle our own player with the DVD-ROM based material and install it on the student's PC.

6.3 DVD Audio

There is some ambiguity as to the most compatible format for audio and the options include: Dolby (AC3) 5 channel, Dolby Stereo, Dolby Mono, MPEG Audio stereo, MPEG Audio mono, PCM and advanced PCM.

7 Conclusions

The potential of DVD technologies has been ignored by most of the higher education sector, particularly in Europe. Some impressive educational developments based around DVD technologies are currently being undertaken by the Ohana Foundation and these are currently being tested in Hawaii schools (www.ohanalearning.org). As more Open University courses are presented on-line, the bandwidth available to most students in their home will place constraints on the type of learning and teaching material which can be delivered on-line. There is a view in the University that we over produce our courses and that the constraints inherent in on-line delivery may force our course production processes to be more efficient. There is also ample evidence that non-broadcast audio-visual material and other computer-based teaching material which cannot be delivered on-line, does enhance the student's

learning process. The hybrid system described in the paper will allow course teams to offer students a richer on-line learning experience until a significant number of University students have access to broadband data networks.

8 References

- [1] Taylor J. (2001) DVD Demystified, McGraw-Hill.

Learning by experience: Networks in learning organizations

Marja Kuittinen

Dept. of Computer Science, University of Joensuu, P.O. Box 111, FIN-80101 Joensuu, Finland

Phone: +358 13 251 7935, Fax: +358 13 251 7955

E-mail: Marja.Kuittinen@cs.joensuu.fi

Erkki Sutinen

ITN - Institutionen för Teknik och Naturvetenskap, Campus Norrköping, Linköpings Universitet, 60174 Norrköping, Sweden

Phone: +46 11 363 338, Fax: +46 11 36 32 70

E-mail: erksu@itn.liu.se

Heikki Topi

Dept. of Computer Information Systems, 403 Smith Technology Center, Bentley College, 175 Forest Street, Waltham, MA 02452, U.S.A.

Phone: +1-781-891-2799, Fax: +1 781 891 2949

E-mail: htopi@bentley.edu

Marko Turpeinen

Alma Media Corporation, Eteläesplanadi 14, P.O. Box 140, FIN-00101 Helsinki, Finland

Phone: +358 9 50771, Fax: +358 9 507 8555

E-mail: Marko.Turpeinen@almamedia.fi

Keywords: learning organizations, organizational learning, communication networks, virtual university

Received: January 24, 2001

As part of the Virtual University of Finland, the Connet framework offers undergraduate courses in Cognitive Science. The studies are mainly organized as web-based courses or collaborative student projects. In the basic studies component, a student completes four one credit methodology courses and chooses a related assignment worth three credits for one of them. One of the methodology courses is entitled Networks in Learning Organization. It will be delivered entirely on the web, and it also will serve as a pilot course to help design other methodology courses. This paper will first briefly discuss the theoretical foundations of learning organizations and organizational learning. Then, we will review the role of networks as the technological foundation of a learning organization and continue by discussing the use of educational technology to support organizational learning and learn about it. Finally, we will describe the structure and the methods of the course and present topics that form the starting point for the discourse within the course.

1 Introduction

As part of the Virtual University of Finland, the Connet framework offers undergraduate courses in Cognitive Science. The studies are mainly organized as web-based courses or collaborative student projects. For example, in the basic studies component, a student completes four one credit methodology courses and chooses a related assignment worth three credits for one of them. One of the methodology courses is entitled Networks in Learning Organization. It will be delivered entirely on the web, and it also will serve as a pilot course to help design other methodology courses.

For the students, the goal of the Networks in Learning Organization course is to learn to utilize network-based

collaborative software tools for communication within a given organization in ways that contribute towards development of that organization. For example, the course introduces various CSCW (computer-supported collaborative work) and problem solving tools, which the students are asked to use to work on topics related to learning organizations.

The fact that network-based learning is learned via network-based learning tools is of particular interest. The learning by experience principle obviously fits the given context well. The students taking the course are required to establish small heterogeneous groups. These groups will form their fictitious organizations, specify their goals, communicate with organizations of their peer students, and develop their own organizations further on. Students may also

use any available open problem solving tools to intensify their learning process. To add real life flavor, the students can also belong to several organizations at the same time.

This paper will first briefly discuss the theoretical foundations of learning organizations and organizational learning. Then, we will review the role of networks as the technological foundation of a learning organization and continue by discussing the use of educational technology to support organizational learning and learn about it. Finally, we will describe the structure and the methods of the course and present topics that form the starting point for the discourse within the course.

2 Learning organizations and organizational learning

The topic of learning organizations (and the related topic of organizational learning) has been widely discussed in management literature for more than twenty years, at least since Argyris & Schön (1978) published their seminal book *Organizational Learning*. The topic has been extensively covered in books (Argyris & Schön, 1978 and 1996; Garvin, 2000; Garrat, 2000; Senge, 1990) and in comprehensive review articles (for example, Dodgson, 1993; Fiol & Lyles, 1985; Huber, 1991; Levitt & March, 1988; Robey, Boudreau & Rose, 2000). The purpose of this section is not to provide a comprehensive integrative review of the area but to highlight the most important findings that are important in forming the foundation for the course described in this paper.

As several authors (including Garvin, 1993 and Crossan, Lane, & White, 1999) point out, the active academic discussion on organizational learning and a learning organization has not been able to produce one consensus definition for this widely used term. Following Garvin (1993) and other influential authors in the field as cited below, we believe that at least the following elements of organizational learning are important in understanding the nature of genuine learning organizations (i.e., organizations that consistently exhibit effective and efficient learning behaviors):

1. In organizational learning, organizations observe their own behaviors and actions and modify them based on the feedback they receive from the environment with the natural goal of improving their performance. (Argyris, 1977; Fiol & Lyles, 1995).
2. Organizational learning is a process of sharing of "insights, knowledge, and mental models" (Stata, 1989), and it requires efficient and effective communication within the organization.
3. Organizational learning increases "the range of its potential behaviors" (Huber, 1991), i.e., it improves the organization's ability to choose a proper behavior for a particular situation.

4. Organizational learning is a continuous and never-ending process (Garratt, 2000; Garvin, 1993).
5. Organizational learning requires creativity and innovation. (Garvin, 1993).
6. Organizational learning requires that the organization is capable of encoding the results of its learning in both behaviors (Fiol & Lyles, 1985) and repositories for conceptual understanding and factual knowledge (Walsh & Ungson, 1991; Robey, Boudreau, & Rose, 2000). The latter perspective links this issue immediately to the very popular topic of knowledge management (Davenport & Prusak, 1997), and it seems obvious that any genuinely good learning organization is also competent in its knowledge management activities.

Crossan et al. (1999) provide interesting and important insights about organizational learning for our purposes because they discuss extensively the linkage between learning at the individual, group, and organizational levels. They specifically emphasize that organizational learning is a multi-level process and it takes place at all of these three levels simultaneously. According to them, the links between these levels are based on four social and psychological processes, which they call intuiting, interpreting, integrating, and institutionalizing (Crossan et al 1999, p. 525). Intuiting is an individual level process, interpreting forms a bridge between the individual and group levels, integrating links the group and organizational levels, and finally, institutionalizing takes place at the organizational level. Institutionalizing is the true organizational learning process in a sense that it is the only stage where learning will be embedded in the organizational systems, processes, and practices, but institutionalizing cannot take place without the previous stages, i.e., the processes that take place at the individual and group levels. Thus, all levels have to be taken into account also in pedagogical approaches utilizing organizational learning ideas.

3 Communication Networks as a Foundation for Learning Organizations

In their comprehensive review of the literature on the relationship between organizational learning and information technology, Robey et al. (2000) identify two main streams of research: a) use of organizational learning methods and tools to learn about information technology and b) the use of information technology to support organizational learning. From the perspective of creating a learning environment for learning about learning organizations, the latter stream is clearly more interesting. In their discussion, Robey et al identify two major (and intuitively obvious) ways information technology can be used to support organizational learning (and thus learning organizations):

1. Information technology can be used for maintaining organizational knowledge repositories and thus, supporting organizational memory. Again, the practitioner literature on knowledge management (e.g., Davenport & Prusak, 1997) is very closely linked to this topic.
2. Information and communication technology can be used both for communication and discourse between individuals and groups within the organization and for access to the knowledge repositories. In addition, communication networks support individual learning across organizational boundaries.

Organizational knowledge repositories can, naturally, be maintained with a rich variety of technologies varying from relational database management systems for structured databases used for administrative information systems to highly unstructured repositories of multimedia data maintained on a corporate intranet, which can be based either on open WWW technologies or on Lotus Notes or other similar proprietary technology. The types of technologies that are appropriate for a specific organization depend on a variety of factors (size, industry, technical expertise, knowledge intensity, etc.), but it is essential that every organization explicitly recognizes the need to formally organize and maintain their knowledge repositories so that they support the organization's learning goals. Experience from a variety of organizations suggests that particularly the efficient utilization of unstructured textual and multimedia data is very difficult, both because it is difficult to find strong enough incentives for organizational members to consistently contribute to the common repositories and because the interpretative processing of the repository contents is often difficult and insufficient.

It is important to note that the communication support that the networks provide takes place at two different levels: on one hand, they provide efficient access to factual knowledge both through individual (e.g., e-mail) and group (electronic conferencing) communication and through access to various knowledge repositories. On the other hand, they support the group level interpretative processes (Crossan et al 1999) by supporting one-to-one and group level discussion and debate. The structure information systems provide for communication can be a vitally important part of efficient support for the organization's learning processes, although it is clear that individuals and groups do not always follow the structures systems create for them - instead, they appropriate the technology in ways that best supports their personal goals (DeSanctis & Poole 1994). Therefore, evaluating the fit between the use of communication technologies and organizational learning goals is important.

4 Learning about Organizational Learning using Educational Technology

A successful implementation of virtual courses and curricula requires an explicit need and a well-defined objective. Educational technology is too complicated to be wasted for building applications or information repositories with an unspecified educational goal. Compared to many other subject areas, the motivation for building a virtual course on Networks for Learning Organization is strongly related to the skills it builds. Summing up the elements of organizational learning, it is easy to draw the goals of the course. A student should

1. experience herself as a member of a learning student organization, which is linked together by a communication network;
2. be able to evaluate the tools applied in the network; and
3. analyze the learning process of the group.

These goals are clearly hard to achieve by a regular material-based course, whether implemented in a traditional classroom setting or as a web-based information package. Hence, rather than starting from a teacher's point of view, by providing a learner with extensive material on organizational learning, we should give him an experience of participating in a organizational learning process and absorbing the course's goals from inside.

In educational technology literature, behaviorist methods have long been juxtaposed with constructivist learning environments (Boyle 1997; Jonassen et al. 2000). However, even the latter often emphasize "objective" learning: a student might learn a topic by seeking for information in the World Wide Web or taking a role in a virtual world, be it a simulation or a MUD. Even learning environments which involve a group of real learners exploring a real case, like distributed cognition (Bell & Winn, 2000), are based on the objective approach stressing the distance between a cognitive learner and the object to be studied, individually or as a collaborative group. In experience-based learning, or participatory learning, the approach is highly "subjective": one is learning the system he is in. The learning environment is not just a - potentially alienating - cognitive tool but a world to be explored and learned; like a profession in the traditional system of apprenticeship.

To learn, an apprentice needs a concrete assignment to work on. In the case of learning to use networks for organizational learning, the concrete assignment consists of the following elements:

- A learning organization. This could obviously be a group of students learning the same course at about the same time.

- A learning goal. The idea of a learning organization is to improve its performance; thus, the student group selects a topic from a given list or develops a problem of its own.
- A communication network. The natural choice is to provide the student group with CSCW software tools running over the Web.

The similarity of organizational learning to problem-based collaborative learning is evident. However, the problems of a learning organization are rarely explicitly specified, or closed, but rather blurred or fuzzy. This means that the group should not only improve the performance of an identified process or practice, but even recognize the problems by themselves. This means that they need to make use of a creative problem solving tool. It is important that the students do not only communicate with each other, sticking to their starting points, but are open to novel and innovative approaches. To intensify this process, tools like idea generators are needed; for example, the IDEGEN software. These tools should be applied to the problem solving phase as well as its specification.

However experience-oriented the organizational learning process might be, the group involved has to be able to reflect the process afterwards. Hence, the students need tools to keep track of their learning process. The purpose of the organizational learning process should not be limited to just an improved performance but even a better capability to improve performance; a more efficient way of organizational learning. Various tools for creating a collaborative learning diary could be applied; for example, Woven Stories (Harviainen et al. 1999).

5 Structure and Methods

In this section, we will discuss at a more detailed level the issues related to the implementation of the course and provide a description of the structure that the course will follow.

The course will be structured as an organizational development project in which students will be assigned into small teams that work together to solve problems closely related to the primary topic areas of the course. The problems will be presented in the forms of small open-ended case descriptions that have been designed to illustrate issues learning organizations are facing. For each instance of the course, the process will be synchronous and it will have a clear starting point and a well-defined end.

The course will be structured as follows:

1. At the time of the course registration, students provide demographic and academic background information. In addition, their pre-course attitudes towards and their initial ideas regarding solutions to organizational learning problems will be captured.
2. The background information and the data regarding the students' initial attitudes and solution models will be used to form heterogeneous teams of four or five students. The intention is to find heterogeneity not only in terms of gender, technical skills, professional experience, and academic background but also ensure that each team members' basic initial approach to solving organizational problems is not the same. We believe that some level of initial disagreement is beneficial because it provides a fruitful starting point for a fuller exploration of the solution space. The extent to which the group formation will succeed depends, of course, on the initial heterogeneity of the student population, but the interdisciplinary nature of the program should ensure sufficient diversity.
3. The students will attend 2-3 initial background modules ("lectures") that will be implemented in the virtual environment but delivered synchronously. The purpose of these modules is to ensure that the participants have a sufficient understanding of the basic concepts, fundamental goals, and the most important existing work in the area. The intention is not to provide model solutions or teach approaches to problem solving, but to provide an initial understanding of the work that the students can use as a resource while working on their own problems. To the extent it is possible, links to the most important resources will be provided in the virtual environment (see the references of this paper for examples of seminal work in the area).
4. Each team will be assigned a topic that has been identified in the organizational learning literature to be a potential problem area. The next section will briefly discuss these topics. In addition to a brief textual description, the teams will get a short case that illustrates the nature of the problem in an organizational context. For each instance of the course, the problems will be categorized into problem families that provide natural linkages between the teams.
5. The teams will work on the problems using two different types of tools: a) a distributed learning environment such as Lotus Notes, TopClass, or WebCT, and b) a problem-solving support and idea generation tool such as IDEGEN described above. The directions given to the teams will focus mostly on outcomes and not on the process because the active discovery of the process is one of the most important learning objectives of the entire course. An electronic log of the communication events will be maintained on both of the tools. In addition, the students will be asked to maintain a personal journal that includes their experiences regarding the learning process.

The outcomes of the problem-solving process will include the following elements: a) a definition and detailed identification of the aspects of the problem; b) a conceptual analysis of the factors relevant for the problem based on existing literature; c) raw results of the idea generation process; d) a description of solu-

tion alternatives; and e) a detailed description of the selected solution with a carefully developed justification for the selection. The final reports will be made available for all students in the course.

6. The analysis of the learning process is also a vitally important part of the learning process. After the teams have gone through the idea generation - solution selection process described above, they will be asked to review the electronic logs and their own personal journals in order to analyze the strengths and the weaknesses of their team as a learning organization with a strong focus on the reasons that affected the quality of the team's performance. Specifically, the teams are asked to make suggestions regarding the ways they could have improved their performance and avoided the problems they were facing during their work.
7. In addition to the two team reports (solution report and analysis of the learning organization), the course will require an individual final examination that will be administered as a mini-paper that requires the students to master not only the results of their own work but also the fundamental concepts introduced in the introductory lectures, a subset of the materials included in the course repository and, most importantly, in the work by the other teams.

6 Potential Topic Areas for the Teams

During the course, the topic areas that the project teams will be working on will be expressed as problem descriptions focusing on issues such as these:

- Only few members of the organization actively contribute to the organizational knowledge repositories.
- Large amount of data is available in knowledge repositories but it is poorly organized and not interpreted in the organizational context.
- Large amount of information is available, but the users ignore in their daily work the repositories in which the information is stored.
- The organization suffers from a very strong 'Not Invented Here' -syndrome.
- Best practices discovered within the organization are never shared with other members of the organization. In general, there is very little communication between organizations' members.
- The organization is very inflexible, reluctant to change, and demonstrates very little creativity.
- The organization performs well in familiar situations but it has a very limited ability to adapt to new circumstances.

- Emphasis on organizational learning has become a theme that surfaces occasionally as a special project, but the organization is not able to maintain a consistent focus on learning.

All these topic areas are very close to the core identity of organizational learning and learning organization (please see the discussion related to the nature of learning organizations above in Section 2).

Please note that these comprise just a small subset of possible issues, and that the limited amount of space allows only brief, cursory descriptions of each of the topics. Actual problem descriptions will be significantly longer and they will, as discussed above, be accompanied with a brief case that links the problem to the organizational context. Cases will come with instructions stating that the case illustration is not intended to provide an exhaustive description of the problem space.

7 Future Research and Conclusion

This course provides plenty of opportunities for future empirical research in this area. Although constructivist and experiential learning approaches are well-known and widely researched topics, in this case the dynamic interaction between the topic area and the pedagogical approach form a unique combination that generates a lot of questions for future investigation.

One of the interesting questions that can be explored with this course is the role of repositories and various types of repository components in this environment. How much will the materials available in the repositories be used in the formulation of the solutions to the problems and which types of materials will be used most? What is the value of materials from one course instance to the participants of later instances and are the courses able to build on the top of material developed by previous participants? With the future course instances, it will be possible to manipulate the availability of the various types of repository materials and evaluate the effect this has on the nature of the learning process.

Another interesting area for further exploration is the interteam communication: In this paper, we have only briefly referred to the interaction between the teams, but we believe that building incentive mechanisms to encourage this may have an important impact on the learning process through which we might also be able to model some of the knowledge transfer processes. What are the mechanisms that truly encourage cooperation between the teams and how will this cooperation affect the learning results? Finally, the course will provide an interesting opportunity to evaluate the relationship and the linkage between the uses of the various tools, in this case a traditional collaboration support tool and a tool for idea generation support.

The core idea of the course Networks in Learning Organizations will be the use of a learning organization to learn about organizational learning and learning organiza-

tions. In this topic area, it is vitally important and beneficial to apply the principles of constructivist and experiential learning approaches because they (particularly experiential learning) are very close to the fundamental issues of the field. In the same way passive repositories of knowledge do little good to any organization if they are not actively linked to the life of the organization, traditional repository based approaches to learning are not effective in teaching students the core concepts of learning organizations. We believe that the planned course that places the integration of intellectual activity and practical experience of the students to the center of the process will provide them with a deep and long-lasting understanding of the use of networks for building learning organizations.

References

- [1] Argyris, C. (1977) Organizational learning and management information systems. *Accounting, Organizations and Society*, 2, 2, p. 113-123.
- [2] Argyris, C. & Schön, D. A. (1978) *Organizational Learning*, Reading, MA: Addison-Wesley.
- [3] Argyris, C. & Schön, D. A. (1996) *Organizational Learning II*, Reading, MA: Addison-Wesley.
- [4] Bell, P. & Winn, W. (2000) Distributed Cognitions, by Nature and by Design. Jonassen, D. H. & Land, S. M. (Eds.) *Theoretical Foundations of Learning Environments*, Addison-Wesley.
- [5] Boyle, T. (1997) *Design for Multimedia Learning*, Prentice Hall.
- [6] Crossan, M., Lane, H. W. & White, R. E. (1999) An organizational learning framework: from intuition to institution. *Academy of Management Review*, 24, 3, p. 522-537.
- [7] Davenport, T. & Prusak, L. (1997) *Working Knowledge: How Organizations Manage What They Know*, Cambridge, MA: Harvard Business School Press.
- [8] DeSanctis, G. & Poole, M. S. (1994) Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science*, 5, p. 121-147.
- [9] Dodgson, M. (1993) Organizational learning: a review of some literatures *Organization Studies*, 14, 3, p. 375-394.
- [10] Fiol, C. M. & Lyles, M. A. (1985) Organizational learning. *Academy of Management Review*, 10, 4, p. 803-813.
- [11] Garvin, D. A. (1993) Building a Learning Organization. *Harvard Business Review*, 96, 1, p. 19-28.
- [12] Garvin, D. A. (2000) *Learning in Action: A Guide to Putting the Learning Organization to Work*, Cambridge, MA: Harvard Business School Press.
- [13] Garratt, B. (2000) *The Learning Organization*, Harper Collins Business.
- [14] Harviainen, T., Hassinen, M., Kommers, P. & Sutinen, E. (1999) Woven stories: collaboratively authoring micro-worlds via the Internet. *International Journal of Continuing Engineering Education and Life-Long Learning*, 9, 2/3/4, p. 328-341.
- [15] Huber, G. P. (1991) Organizational learning: the contributing processes and the literatures. *Organization Science*, 2, p. 88-115.
- [16] Jarvenpaa, S. L. & Leidner, D. E. (1999) Communication and trust in global virtual teams. *Organization Science*, 10, 6, p. 791-815.
- [17] Levitt, B. & March, J. G. (1988) Organizational Learning. *Annual Review of Sociology*, 14, p. 319-340.
- [18] Robey, D., Boudreau, M. & Rose, G. M. (2000) Information technology and organizational learning: a review and assessment of research. *Accounting, Management and Information technologies*, 10, p. 125-155.
- [19] Senge, P. (1990) *The fifth discipline: the art and practice of the learning organization*, New York: Doubleday.
- [20] Stata, R. (1989) Organizational learning: the key to management innovation. *Sloan Management Review*, 30, 3, p. 63-74.
- [21] Walsh, J. P. & Ungson, G. R. (1991) Organizational memory. *Academy of Management Review*, 16, 57-91.

How to learn introductory programming over the Web?

Haataja Arto

University of Joensuu, Department of Computer Science, P.O.Box 111, FIN-80101, Joensuu, Finland
Phone: +385 13 251 5272, Fax: +385 13 251 7955

arto.haataja@cs.joensuu.fi

Suhonen Jarkko

University of Joensuu, Department of Computer Science, P.O.Box 111, FIN-80101, Joensuu, Finland
Phone: +385 13 251 5272, Fax: +385 13 251 7955

jarkko.suhonen@cs.joensuu.fi

Sutinen Erkki

University of Joensuu, Department of Computer Science, P.O.Box 111, FIN-80101, Joensuu, Finland
Phone: +385 13 251 7934, Fax: +385 13 251 7955

erkki.sutinen@cs.joensuu.fi

Keywords: introductory programming, virtual university, high school, distance education

Received: February 1, 2001

As part of the Finnish Virtual University, the universities in the east of Finland offer high school students an opportunity to obtain their first 15 university credits of Computer Science over the Web. At the University of Joensuu, the courses consist of three parts: General Introduction to Computers (three credits), Introduction to Computer Science (five credits), and Programming (seven credits). Instruction is almost entirely given over the Web. The students follow a schedule given in the web site, learn the related chapters from their textbooks, and return the exercises by strict deadlines.

The students live as far as a hundred kilometers away from the university, and their local high schools have not been able to hire any qualified programming teachers. We found a solution to this problem by organizing on-line teachers at the university to answer students' questions, and assigning tutors, who had hardly any experience in programming, to encourage the students at the local schools. To intensify the learning outcomes, we are planning to use three different learning tools in the web-based course environment: Excel, Jeliot and BlueJ. Each of them can be used for understanding a given program visually. In particular, the environments serve as virtual laboratories for real problems: the students can study their own programs.

At the University of Joensuu, the course started in August 2000. Out of the 80 enrolled students, more than 65 were active after the first three months. Altogether, the course will last for 16 months. Students' activity and commitment to their studies indicate that the approach chosen to teach programming has proved to be efficient.

We have aimed at designing a solid model for building the environment.

1. Virtual Computer Science Studies

In Finland, the Ministry of Education is funding a three-year project to establish the Finnish Virtual University during the years 2001-2003. One particular goal in the project is to develop new methods for science education. The three universities in the east of Finland (University of Joensuu, University of Kuopio and Lappeenranta University of Technology) work jointly in the Virtual University project. One of the concrete objectives is to create a web-based learning environment for introductory computer science, intended for high school students. From the research perspective of educational technology and computer science education, this task is particularly challenging.

The project is called "Virtual Certificate", indicating that high school students can obtain 15 credits of Computer Science studies in one and a half years time via the Internet (Haataja et al. 2001). In the Finnish university system, each credit equals 40 hours of studying; 160 credits are required for the Master's degree. Therefore, after passing all the 15 credits of the program, a student has completed the first year of Computer Science studies. Moreover, if a student passes the program with a grade 2/3, she is free to enter the university as a Computer Science major student. Almost all teaching is done via the Internet because the students come from all around the district

of North Karelia in the east of Finland. We have minimized the need for face-to-face teaching situations so that the students do not have to waste their time and money by traveling from their home to the university.

The most important part of our project is to give students good programming skills. Hence, our emphasis is to make the learning of programming as smooth as possible. One essential part of the project is *tutor-teachers* in the high schools and *on-line tutors* in the university. Particularly in the programming courses these tutors can give a very valuable contribution to the learning process. Most of the courses, including programming courses, have a similar structure: students have material to work with, they do weekly assignments, and their learning outcomes are evaluated in an exam. However, we have tried to add some flavor to the studies by employing a few other teaching methods such as follows:

- *Group activities.* Some courses will be implemented so that the students work in small groups. We hope that this would encourage students to work collaboratively, and would bring about interaction into the learning process.
- *Learning by writing.* Students compose an essay on a certain subject. This is hopefully done in cooperation with local newspapers: the students will be reporting on what they learn.
- *Netbus consulting.* A bus with an Internet connection will visit each school for a few hours. A Question & Answer session will be organized, as well as individual consulting.
- *Days at campus.* To make the students feel themselves as members of the academic community, they are invited to the campus for four days during their studies.
- *Camp at campus.* A theoretical course on algorithm will be arranged at campus, as a one-week intensive period.

2. Description of Introductory Programming Courses

In general, the basic structure of introductory programming course is similar to the other courses. We chose Java as our teaching language because the graphical Java-applets provide more interesting and exciting ways to learn programming than standard text-based programming languages do. Basically, our method of teaching the programming is very simple. The students have got learning materials and weekly assignments to work with, and the students' own contribution to the learning process is essential. The student must have 1/3 of the assignments completed if she wants to take part in the exam. Furthermore, by doing extra exercises she can achieve bonus points for grade of the course. The purpose of the weekly assignments is to make the students understand the

basic and most important aspects of introductory programming. We think that learning by doing is the most efficient way to accomplish it (Dewey 1918). In the following chapters, we describe our methods for teaching introductory programming in more detail.

2.1 Learning Material

One of the most important principles of the teaching model is joining together the printed learning material, i.e. textbooks, and the material on the web. The *textbook* gives detailed knowledge of the domain to the students while web materials support the learning process. The *Web-material* provides an overview of the programming domain, as well as examples of programming to support the textbook material in a meaningful way. With the web-material, the student can easily work out the different concepts and their relationships in the course. We hope that in this way, the web-material will provide a specific mental map of the programming concepts. In addition, the web-material divides the domain in logical parts so the students are given the opportunity to concentrate on a certain subject at a time. Hence, one programming course consists of smaller units and in every unit there are exercises to work with. By splitting the course into smaller units, we can control the learning process in a sophisticated way so that the information flow does not get too stressful. We are using the WebCT¹ platform as our learning environment, but all web-materials are designed to function independently.

As the web-material is providing all the crucial parts of the programming domain, we use the course books to give detailed information to the students. This way we can link the web-material and textbooks together so that the potential of these two media is efficiently used to support the learning process. We can be quite sure that academic programming textbooks give accurate and deep knowledge about the subjects in hand. In the web-material, we are able to concentrate on providing examples, pictures, animations etc. to support the learning process in multiple ways.

¹ <http://www.webct.com>

This way, we can focus our attention to enriching the learning process and we do not have to write all the detailed Java programming issues ourselves. Example

of the learning material used in a programming course can be seen in figure 1.

ALKU viikko1 viikko2 viikko3 viikko4 viikko5 viikko6
viikko7 viikko8 viikko9 viikko10 viikko11 viikko12 Servletit

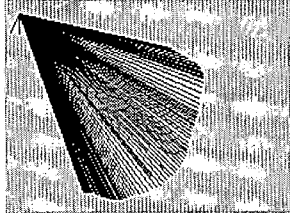
Lisäksi tehtäviä iltojen iloksi

JAVA JATKOKURSSI

VIKKO 3

Hiiren ohjelmointi
MouseListener rajat
Repaint
Painikkeet
Lahelit
Tekstikentät

Hiiri-ohjelma jossa toiminto mouseMoved ja mouseDragged



Liikuta hiirtä appletin kohdalla

Yllä olevassa appletissa on käytetty metodia mouseMoved, joka reagoi hiiren liikkeisiin. Vastaavasti on olemassa mouseDragged, joka reagoi hiiren raahaamiseen.

Mikäli halutaan toiminnot mouseMoved ja mouseDragged, niin tehdään ohjelmaan seuraavat lisäykset:

```
//Otetaan käyttöön rajapinta MouseMotionListener

//Lisätään kuuntelija addMouseMotionListener(this);
addMouseMotionListener(this);

//Lisätään metodit mouseMoved ja mouseDragged
```

Figure 1: Snapshot of the web-based learning material used in the project

2.2 Tutor-Teachers

Another crucial aspect of our teaching strategy is the tutor-teachers. These teachers work at local high schools and their main contribution to the project is to provide mental support for the students. Because the tutors are not necessarily computer science professionals, they are not capable of teaching programming to the youngsters. We rely on the ability of the tutors to contribute the learning process by providing pedagogical know-how for the use of the students. For this reason, the tutor may choose the working methods and timetables during the courses by herself. There are a few other minor obligations for the tutor, but mostly she may organize the activities in schools rather freely.

2.3 On-line Tutors

Because the high school teachers lack computer science knowledge, an on-line tutoring system had to be set up. On-line tutors work at the university and their main responsibility is to go through the students' answers to the weekly assignments. After the assessment, the on-line tutor gives feedback to the students. In the feedback, the tutor can tell all the good or bad parts of the answers. We hope that this way, the on-line tutor can give crucial hints and instructions to the students about the things to which they should pay attention in the next assignments. When all the assignments have been assessed and the feedback has been given, the tutor at the university gives out the model solutions for the assignments.

After the "correct" answers have been given out, we hope that the students would look up the solutions and compare them to their own answers. In this way (with model answers and feedback) the students can reflect the good and bad aspects of the solution to the problem in hand (Polya 1958). We have made it clear

that the model solutions are just one way to solve a given problem. All that really matters is that the answer gives an appropriate solution to the problem. In some cases, we have directly used the students' answers as model solutions. We hope that this encourages students to really concentrate on finding the solutions to the programming problems.

Another important task of the on-line tutor is to give instant help for those students who need it. Students can use e-mail or bulletin board messages to contact the tutor. In the project, our aim is to answer the questions as quickly as possible so that the upcoming problems do not disturb the learning process. After the course, the on-line tutor corrects and grades the exams. Our opinion is that the tutor gets a good idea of the students' level of understanding of the subject covered during the course.

3. What We Learned about Teaching Programming over the Net

For us the development of distance education studies has been a fairly new experience. We knew that the programming part of our studies could be a difficult one. In some cases, we have managed well and the students' opinions towards it have been mainly positive. The assignments and learning materials have inspired the students to work hard with their studies. This can be seen in the students' answers to the exercises, which in many cases have been of high quality. In fact, one of the first big surprises was that the majority of the students accomplished over 70 % of all the assignments. Under normal circumstances, university students' first aim is to get 1/3 of the exercises done so they can take the exam.

We assumed that this was to happen with our students, too. In some cases we had to ask them to slow down with the assignments because the students were complaining about the amount of assignments per week. We pointed out that there was no need for them to accomplish all of the assignments. We think the reason for the amount of returned assignments derives from the learning culture in high schools. Often all the work at high school is compulsory to the students and learning is measured, for example, by the amount of exercises completed. Therefore, students do not work with their assignments in order to learn. They do the assignments to prove that they are capable of keeping up with the studies.

Another quite disturbing discovery was the fact that the students had in some cases over 35 hours of high school studies a week. By taking our courses, the students' workload became far too heavy. For some students this led to a situation where they had to quit our studies. They simply did not have any time to concentrate on rather a complicated domain like

programming. At the same time, the tutor-teachers' resources in some schools were too low so the students did not get the necessary support during the critical periods in the courses. Still, there has been a group of students, altogether about 40, who have been capable of handling their high school studies, as well as our studies successfully.

4. How to Intensify the Learning Outcomes?

During the "Virtual Certificate" project, it has been clear that our method of teaching programming is not effective for all kinds of learners. To make the learning process as efficient as possible, we are planning to use the following methods: interactive visual tools, adaptive learning materials and intensive collaboration.

4.1 Visual Tools

Visualization has long been an important pedagogical tool in CS education. The use of the web and interactive animations provide opportunities to expand the availability of visualization-based teaching and learning tools. We have been experimenting BlueJ, Jeliot and Excel. Each of them can be used for understanding a given program visually. In particular, the environments serve as virtual laboratories for real problems: the students can study their own programs.

4.1.1 BlueJ

BlueJ¹ is a visual programming environment designed for teaching and learning object-oriented programming using Java as the implementation language (Kölling 2000). The environment is specifically designed to support object-oriented programming in the beginners' courses (Kölling & Rosenberg, 2001). BlueJ is based on the Blue system. Blue is an integrated application merging an object-oriented language and an object oriented development environment, generated at Sydney University and Monash University, Australia. The implementation language for the Blue system was C++. BlueJ is an environment that offers an almost identical platform to Blue, but with Java as the supported programming language. The system runs on all known Java 2 platforms (Kölling & Rosenberg 2001).

The environment is designed to enable the use of the "objects first" approach in Java, helping students to develop understanding of object-oriented concepts, such as objects and classes, message passing, method invocation and parameter passing. The aim is to concentrate on solving programming problems without becoming distracted by the mechanics of compiling, executing and testing Java programs. Hence, the teacher can present the object-oriented

¹ <http://bluej.monash.edu>

concepts before teaching the non-trivial facts concerning the traditional use of Java. For example, the “main” method in Java includes various complicated concepts such as *static*, *public*, *void*, *string*, *arrays* and *parameters*.

With BlueJ, the teacher can choose the appropriate timing for such concepts to be presented in the course. Furthermore, the system gives a visual presentation of classes and objects created. These are presented so that the students can actually see, for example, the relationship between different objects (Kölling & Rosenberg 2000).

In addition, the BlueJ environment provides some tools for handling Java programs: an integrated visual debugger, support for generating and executing applets and an export function that can create executable jar files.

4.1.2 Jeliot

Jeliot animates algorithms or programs written in the Java programming language by visualizing data structures as graphical objects that move smoothly. Jeliot aims at automatic animation, i.e. the user submits a plain algorithm and Jeliot generates the visual presentation (animation) of the algorithm (Lattu, Meisalo & Tarhio 2000). Jeliot is based on a theater metaphor.

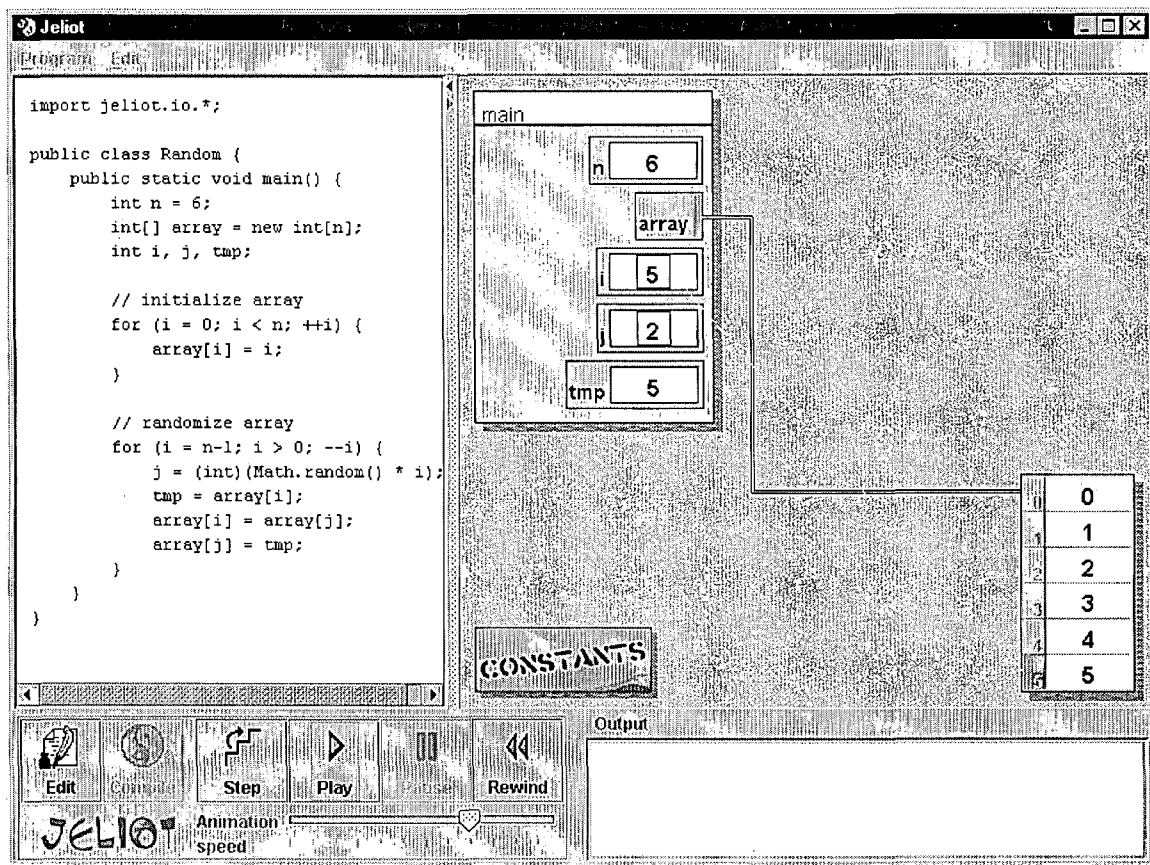


Figure 2: Jeliot 2000 animation environment

The algorithm given by the user is a script in a play where the variables are acting the roles. The user is the director of the play controlling, for example, the speed of the presented animation (Lahtinen, Sutinen & Tarhio 1998). Jeliot is used with a Java-capable Web browser and the application can be utilized in distance education context such as “Virtual Certificate.

Jeliot 2000 is the newest version of Jeliot and it is intended for teaching computer science especially to

high school students (Ben-Ari, Levy & Uronen, 2000). The emphasis of Jeliot 2000 is on program animation, which demonstrates the execution of input-output, assignment, selection and loop statements. Animation of higher lever data structures and of algorithms is not necessary on high school level. Example of the students’ interface at Jeliot 2000 can be seen in figure2.

The main features and modifications introduced in Jeliot 2000 are (Ben-Ari, Levy & Uronen, 2000):

- A single window is displayed with two panels, one for the source code and the other for the animation
- All the variables are animated uniformly
- Expression evaluation and control decisions are animated
- The only controls are the familiar VCR-like ones
- Text input-output is animated
- Jeliot 2000 is a single Java application
- Simplified user interface

Jeliot 2000 is implemented in Java using version 1.2.2 of the SDK. Jeliot 2000 is complete in that it animates all the constructs of the programs it accepts, but the current implementation is limited in the language constructs it supports. Jeliot 2000 is distributed freely at: <http://stwww.weizmann.ac.il/g-cs/benari/vis.htm>

4.1.3 Excel Animations

We have also been studying how to prepare animations of simple algorithms with Microsoft Excel spreadsheet program. Excel offers a light, adaptive and inspiring platform for creating visualizations of various needs in CS (Rautama, Sutinen & Tarhio 1997). A teacher can prepare visualizations for teaching, or the visualizations can be student assignments. Two standard features of Excel, i.e. data visualization and macro programming with VBA, provide together an easy-to-use environment for animating algorithms. The animation is seen as updating charts representing the data structures of the algorithm. However, the basic scheme can be enhanced in many ways. Besides macros, Excel has three layers: workbook, object and chart. By combining these features with a WBA program, it is possible to produce interesting animation features (Dybdahl, Sutinen & Tarhio 1998). Additionally, the drawing tools and textboxes in Excel have proved to be very useful in constructing animations.

In the Excel environment, the user starts from a rough idea and he approaches a satisfactory solution step-by-step with the help of the feedback of the system by adjusting and improving the visualization. It has been proved that Excel provides the students an easy entrance into the world of algorithms (Rautama, Sutinen & Tarhio 1997). The students start from simple programs or concepts and visualize them with Excel. After this they can compile more complicated algorithms and animate them with more advanced systems like Jeliot. Jeliot and Excel have been successfully experimented in computer science courses in the University of Helsinki¹.

4.2 Adaptive Learning Material

During the two programming courses, we noticed some difficulties in the learning processes of the students. Especially during the first programming course, we detected that our learning material was not sufficient for the needs of different learners. Students complained that the learning material was focused on wrong things. It was important for us to offer the students a smooth start. And that was the reason why we concentrated too much on the easy part of the course to get things going on. We thought that if the beginning went well we could speed up the pace. In some cases this led to a situation where some parts of programming, for example while-loops, were not understood correctly. It could be that the easy start led some students to believe that our studies were not so demanding. The exam at the end of the course showed this very clearly. There were students who managed very well, but there were also students whose performance was fairly poor. Next year we are going to emphasize more those aspects of the course that appeared to be especially difficult for the students.

One rather clear absence in our model of teaching is the fact that our web-based learning environment does not pay attention to the needs of different students. For example, the learning materials and exercises are the same to all students. In this situation, it is difficult to come up with learning materials and exercises that are suitable for the use of all kinds of students. It was very clear that certain assignments were too difficult for some students, and at the same time other students found them to be too easy and they got frustrated. This leads us to a clear conclusion; we have to provide different materials and exercises for different kind of students. One can say that the learning environment should be adaptive for the needs of different learners (Brusilovsky 1999).

4.3 Collaboration

Modern learning theories emphasize the value of discussion and collaboration in the learning process (Slavin 1990). We have found that discussing programming concepts is almost impossible in the normal bulletin board format. In face-to-face learning situations, it is easy to point out the different parts of the code. However, specifying the problem in written format only is not a trivial task. As the code gets longer, the difficulties become even more obvious. We are sure that a collaboration tool for teaching programming should provide the means for presenting or pointing out the parts of the program that are under discussion. When we think further, the next step would be an environment where students can write and compile programs together. They could discuss or point out the most important aspects of the code in hand at the same time.

¹ Excel animations, <http://www.cs.helsinki.fi/research/aaps/excel/>

5 Discussion

The purpose of this paper is to shortly describe our experiences in delivering the introductory programming studies over the net. Our project is going on firmly, and at the moment, we are planning the next phase, which will start in autumn 2001. We are aware that our method of teaching programming needs to be improved in certain issues. On the other hand, we believe that the general principle of delivering the studies is working rather well. There are some crucial issues to deal with, but our intention is to enhance the implementation of the studies as described in the previous chapter.

During the project, some rather interesting subjects for future research have arisen. One interesting discovery is that our students are eager to give frank feedback about all aspects of the project. Because the feedback has been mainly pertinent, we have been able to change certain approaches in the project. For example, the time limits of the weekly assignments have been modified according to the opinions of our students. We feel that this kind of chat-like intensive feedback can help the development and implementation of distance education projects. Another subject is how to combine face-to-face learning situations with distance education as efficiently as possible. In the future, the need for maximizing the balance between face-to-face and distance education will be of great interest for business and academic communities.

References

- Ben-Ari, M., Levy, R. & Uronen, P.A. (2000) An Extended Experiment with Jeliot 2000. In Sutinen, E. (ed.), *Proceedings of the First Program Visualization Workshop*, Porvoo, Finland, p.131-140.
- Brusilovsky, P. (1999) Adaptive and Intelligent Technologies for Web-based Education. In Peylo, C. & Rollinger, C. (ed.), *Künstliche Intelligenz, Special Issue on Intelligent Systems and Teleteaching*, Vol.4, p.19-25.
- Dewey, J. (1918) *Democracy and Education; An Introduction to the Philosophy of Education*, MacMillan, New York.
- Dybdahl A., Sutinen E. & Tarhio J. (1998) On Animation Features of Excel. *Proceedings of the ITiCSE'98 Conference*, Dublin, Ireland, p.77-80.
- Haataja, A., Kontkanen, S., Suhonen, J. & Sutinen, E. (2001) Teaching University Level Computer Science to High School Students over the Web, Accepted for publication at the ED-MEDIA 2001 Conference, Tampere, Finland.
- Kölling, M. (2000) *The BlueJ Tutorial*. Technical Report 2000-01, School of Network Computing, Monash University, November.
- Kölling, M. & Rosenberg, J. (2001) BlueJ – The Hitch Hiker's Guide to Object Orientation. To appear in *Journal of Object-Oriented Programming*, Available: <ftp://mars.pscit.monash.edu.au/pub/mik/papers/hitch-hiker.pdf>
- Lahtinen, S.-P., Sutinen, E. & Tarhio, J. (1998) Automated Animation of Algorithms with Eliot. *Journal of Visual Languages & Computing*, 9(3), p. 337-349.
- Lattu, M., Meisalo, V. & Tarhio, J. (2000) How a Visualization Tool Can Be Used – Evaluating a Tool in a Research & Development Project. *Proceedings of PPIG'00, 12th Annual Conference of the Psychology of Programming Interest Group*, Cozenza, Italy, p.19-32.
- Polya, G. (1958) *How to Solve It*, Princeton University Press.
- Rautama, E.; Sutinen, E. & Tarhio, J. (1997) Excel as an Algorithm Animation Environment. *Proceeding of ITiCSE '97, Integrating technology into Computer Science Education*, Uppsala, Sweden, p.24-26.
- Slavin, R. (1990) *Cooperative Learning: Theory, Research and Practice*, Allon and Bacon.

Information System Supporting CATS

Zdeněk Ryjáček
 University of West Bohemia, Univerzitní 8, Plzeň, Česká Republika
 Phone: +420 19 7491 146, Fax: +420 19 7429 989
ryjacek@kma.zcu.cz

Jan Rychlík
 University of West Bohemia, Univerzitní 8, Plzeň, Česká Republika
 Phone: +420 19 7421 417, Fax: +420 19 7421 419
rychlik@civ.zcu.cz

Petr Jiroušek
 University of West Bohemia, Univerzitní 8, Plzeň, Česká republika
 Phone: +420 19 7421 417, Fax: +420 19 7421 419
petr@civ.zcu.cz

Keywords: information systems, prototyping life cycle, credit transfer system, database, study agenda

Received: February 10, 2001

The U.W.B. has developed an information and database system, supporting the internal Credit Accumulation and Transfer System (CATS). The system primarily supports an on-line student registration to courses within the framework of the existing timetable, and further provides a support for designing and filing study programmes, timetables, student records, agenda of admission and alumni etc. We give a survey of basic features of the system and of the history of its development and we present basic rules of the internal CATS that have to be followed by every institution that wants to adopt the system. The system has been already installed at 5 further Universities in the Czech Republic.

1 Introduction

In 1992, after the substantial political changes in East European countries, the management of the University of West Bohemia in Pilsen (UWB) decided to reorganize the traditional system of studies at the University in terms of a credit accumulation and transfer system (CATS). The main reasons for the reorganization were to provide students more flexibility in building their individual curricula, and to meet the (expected) growing demand on student mobility and compatibility of the internal system with the EU standards (ECTS). Based on the structure of the university, which is unique in the country for its high percentage of classes shared by students from different faculties, it was decided that the internal credit accumulation and transfer system (CATS) under development shall have uniform rules across all faculties of the University. This allows much more flexibility for the faculties to develop more interdisciplinary study programmes, and for the students to build their own individual curricula. This solution also allows much better efficiency of the courses, but at the same time it brings much more demand on the administrative and organizational aspects of running such a system.

2 The goals of the project

The UWB CATS is designed to meet the following main goals:

- compatibility with ECTS,
- uniformity of rules at all faculties, providing necessary conditions for development of interdisciplinary courses and programmes,
- high flexibility, allowing to describe various study programmes offered by individual departments in a uniform language of credit transfer system,
- accessibility of all courses for students of all faculties (of course within certain prerequisite and capacity constraints),
- measurability of teaching load at individual departments as a data for internal budgeting at the University.

The main goal of the database system is to provide an organizational and administrative support to this system of study for

- students (on-line registration to courses and exams, access to timetable, curricula, syllabi and further information),

- faculties (student records, computerized check of fulfillment of the requirements by every individual student and further standard outputs),
- departments (offer and advertising of courses, information about students enrolled to courses, agenda of exams and of registration to them),
- persons responsible for timetabling (support for creating the timetable by providing free room search and collision checking),
- internal accreditation board as a part of the internal quality assurance system.
- all individual faculty members (timetable information, lists of students registered to courses, search for free rooms etc.)

3 The environment: administrative, technical as well as financial constraints

At present, the UWB has approx. 11000 full time students, enrolled at 7 faculties:

- Faculty of Applied Sciences,
- Faculty of Economics,
- Faculty of Electrical Engineering,
- Faculty of Humanities,
- Faculty of Education,
- Faculty of Law,
- Faculty of Mechanical Engineering.

The multidisciplinary of the University gives very high requirements on flexibility of the CATS such that the language of the CATS allows all departments to describe their specific requirements and needs in an uniform and algorithmizable way.

The team of the University computing centre was successful in the last years in obtaining several grant projects. Thanks to this success the University has a relatively very modern and efficient internal computer network.

It was evident already from the first outline of the designed CATS that the related administrative tasks cannot be resolved without a specialized database information system. After a short search it was clear that no eligible software product is available. That is why the database group of the University computing center commenced development of our own information system that would match the specific needs of the UWB.

4 The Project Life Cycle

System design and development is supported by Oracle CASE tool – Oracle Designer. This tool consists of four components, namely Process Modeling, Systems Modeling, Systems Design and Systems Generation, and it supports structured analysis (Yourdon 1989).

However, it was clear already from the first steps of the analysis of the problem that the classical *structured project life cycle* is not the right method in this case, since it is not possible to complete the abstract *paper model* in reasonable real time. A university with seven relatively independent faculties is a very specific environment in which a detailed description of the behavior of an information system is impossible before its implementation as a collection of programs. There are many important potential users of the system at the university, and since many of them have their own opinion on many features of the university life, it is very difficult to find an approach that is simple enough to be implementable, but at the same time general enough to yield a common algorithm, acceptable for all parties. These difficulties with implementation of information systems at Universities are described in (Vrana et. al. 1999).

The *prototyping life cycle* or also *iteration life cycle* (Yourdon 1989, Lacko 1994, Richta & Sochor 1996) is a method that is recommended in such cases, and the information system supporting CATS was developed with its use.

At first, the context diagram and the decomposition of the system must be done. Based on the results of the decomposition, the so-called *kernel* of the information system (i.e., the minimal content of the prototype, meeting the main goals of the system) is determined. The first phase of development of the system then covers creating the prototype as a functioning system that covers the kernel and possibly some further features. The second phase comprises further iterations (or, so-called *expansion*) of the prototype.

The general functions of the kernel have to be defined by the users, while details of functions are usually specified by the project manager. Oracle Designer offers three tools for system modeling:

- Entity Relationship Diagrammer,
- Function Hierarchy Diagrammer,
- Dataflow Diagrammer.

Major emphasis must be given on the analysis of the data at the prototyping stage, since changes of the data structure are much more complicated than changes in functions later on. Entity Relationship Diagrammer supports creating E-R-A models (e.g. Connolly et. al. 1995) evolved from work by Chen.

In our case, the primary task (the prototype of the system) consisted in an on-line student registration to courses within an existing timetable, with a possibility of interactive reaction of faculties and departments on the development of the registration. This goal was met by the first version of the system, which was installed in 1994. Growing demand from the academic community for further functions and features resulted in a continuous development of the system within the existing data

structure. These expansions of the prototype were based on the solid-state database model.

In 1995, the continuous growth of the amount of functions supported by the system resulted in necessity of designing a new database structure and developing a second version of the system. It is interesting that the main motivation for creating a new data model was not an immediate need for a new function, but a very long response time (which in peak hours was sometimes more than 3 minutes). Within the existing structure, new settings of Oracle system parameters helped to shorten the response time only about 30% which was insufficient (the data model of the prototype was, as the analysis of real system, correct, but it did not take into account the future functions over the database).

5 The project actual state

At present the University is running third version of the system, which has been already in routine operation for 4 years. In peak hours, the system is able to serve up to 300 users working on the data on-line in parallel with response time at most 1 second.

At present the system consists of the core module - study agenda, and two cooperating modules - agenda of admission and enrollment, procedures, and agenda of graduates and alumni.

System design and development are supported by Oracle CASE tool - Oracle Designer. System ERA model consists of 110 entities, 1132 attributes and 120 relationships.

Available functions cover the following areas:

- students (20)
- courses and syllabi (6)
- timetable (15)
- study programmes and specializations (3)
- curricula (5)
- examinations (8)
- standard print outputs (31)
- student registration to courses and to examinations (8)
- anonymous browsing of selected information (19)
- admission procedures (35)
- agenda of graduates and alumni (24)
- other system and support functions (20)

altogether: 194 functions

Under preparation is a quality evaluation module for automatic generation, evaluation and statistical analysis of student feedback questionnaires

The system development is from the beginning based on Oracle development tools and Oracle database. At UWB we are now using Oracle 8i server (version 8.1.6) and on the client side Oracle Developer (version 6). At present,

the system has about 15000 users.

System security is based on Oracle roles, grants and "fine grained access control". This means that the system administrator has a possibility to decide which functions and which data are accessible for particular users. For example, it is possible to restrict the access for users 'faculty', 'department' to their own records only. This solution is very flexible and it allows the system administrator to parametrize the system according to specific requests of the university.

Users login data including password are transmitted in encrypted form. The system itself does not encrypt data transmission, but encryption of all data transmission can be achieved by extending the system with the Oracle Advanced network security component.

System documentation is available at <http://stag.zcu.cz>

Anonymous browsing of selected information is also web-based and it is available at <http://stag.zcu.cz/prohlizeni>

6 An analysis of the results of the project and of its impact on the institution

At the beginning of the project, a part of the academic community was reluctant to the new ideas and challenges of the system. The scepticism culminated in the third year of the project (i.e. approx. in 1995), when already a lot of work had been invested (especially as concerns filling the system with data), but due to some technical problems (insufficient capacity of some connections, continuous development of the system) the impact of the system on simplification of the routine everyday work of the community was still insufficient. At present the system is an integral part of the academic life of the University and many of its features are considered by the academic community as a common and obvious part of their work.

7 Further developments

At present, a next, fourth version of the system is being developed. The need for a new version was evoked by

- the new University law of the Czech Republic, which has been valid since 1999 and which changed many circumstances,
- requirements and needs of other Universities in the Czech Republic, who are running the system in their specific conditions.

In the new version, the client-server communication will be restricted only to a few users with highest rights in the system. All functions that are accessible to a wide part of the academic community (students, teachers, departments) will be web-based. Under a cooperation with the Prague branch of Lucent Technologies, the development team is designing selected system functions to be accessible via voice-modem.

There are two database companies in the Czech Republic (Pragodata, Magion) who have developed information systems for Universities covering all fields (economics, budgeting, human resources etc.) except for study agenda. In cooperation with them, the new version is being designed to be compatible with both these systems such that there will be a complex information system for Universities available. The outcome of the cooperation with these companies will be available also on commercial basis.

8 Applicability to other institutions

Due to its flexibility and universality, the system can be used at any institution that has organized study programmes and courses in terms of a credit transfer system, or who intend to introduce an internal CATS.

The UWB is willing to provide the system to other academic institutions. At present, the system has been installed at four Universities in the Czech Republic: at Palacky University in Olomouc, at the University of South Bohemia in Ceske Budejovice, at the University of Ostrava, at the Technical University at Pardubice and at Jan Evangelista Purkyně University in Usti nad Labem

In Ostrava and in Ceske Budejovice the system is already in routine operation for second academic year, the Universities in Olomouc, in Usti nad Labem and Pardubice are filling the system with data. Four further institutions are preparing technical conditions for installation. The dissemination of the system within academic institutions at the Czech Republic is partially financially supported by the Czech Ministry of Education who contribute the purchase of Oracle licenses for the adopting institutions. There are no basic obstacles that would restrict eligibility of the system even for academic institutions in other countries. These are the two main questions that should be discussed prior to such an arrangement:

- the courses at the adopting institution have to be organized in terms of an internal ECTS-compatible CATS, or the institution has to be willing to introduce such a system,
- the adopting institution would be asked to cooperate with the UWB database development team on the translation of the system and its documentation to the respective language.

As soon as both these preconditions are resolved, the UWB is willing to provide the adopting institution a copy of the system on a cooperation basis.

9 Conclusion

The way of development of the system by the University staff was the only possibility to meet the goals envisaged. The method of prototyping life cycle was chosen because it eliminates the impossibility of specifying the abstract model of the system at the beginning of the project and because it reflects the need for parallel system development and development of needs and users requirements. However, it should be emphasized that this way of development is very expensive. Since the project commencement, the University has been employing (full-time) a team of 4-5 database specialists.

10 References

- [1] Connolly, T.M.; Begg, C.E.; & Strachan, A.D.: Database System – A Practical Approach to Design, Implementation and Management. *University of Paisley, Addison-Wesley Publ. Comp.* (1995), Paisley.
- [2] Lacko, B.: Prototyping and development of database applications (Czech). In: *DATASEM 94, Cs-Compex* (1994), Brno, Czech Republic.
- [3] Richta, K.; Sochor, J.: Software Engineering I (Czech). Textbook, *Czech Techn. Univ., 1st ed.*, (1996), Prague, Czech Republic.
- [4] Vrana, I.; Bůřil, J.; Černý, A.: Methods of implementing information systems at Universities (Czech). *Technical University of Brno* (1999), Czech Republic.
- [5] Yourdon, E.: Modern Structured Analysis. *Yourdon Press* (1989), *Prentice Hall Building, Englewood Cliffs, New Jersey*.

A data warehouse for French universities

Jean-François Desnos
 Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France

Keywords: data warehouse, executive information system

Received: January 31, 2001

This paper presents the French universities national data warehouse project, which was begun in 1999. The Source databases are the administrative systems, and the target data base, or, data warehouse, is designed to be the executive information system of the institution. Pilot universities in Amiens, Paris, Rennes, Strasbourg, and Versailles are presently testing the prototype. A first release will be done in 2002 for other interested institutions.

1 Introduction

The French national Agence de Modernisation des Universités, (AMUE), is an Information Technology Consortium with most French Universities (about one hundred and ten) as members. AMUE provides large management software applications to its members, e.g. student, financial, and personnel systems. All of them are designed with Oracle client/server technology. AMUE also acts as a consultant to its members in the fields of information technology, university management, and professional development. The Data Warehouse project is one aspect of the management improvement process conducted by AMUE.

Present applications are heterogeneous: The student system has been designed by AMUE and developed by a software company while the financial system has been bought on the market and adapted. The personnel system has been done entirely by AMUE. Because of this heterogeneity, it is presently difficult, and even impossible, to present reports that cross-reference information coming from several data bases. For example, a report mixing student, staff, and financial data is not really available automatically per request.

They are not intended for extraction of data sets within an executive information system.

A data warehouse (Inmon, 1996) achieves this last objective. A data warehouse (DWH) is built by extracting data from source data bases, verifying and transforming them, and then loading a target data base which becomes the DWH. This process of extraction, transformation and loading, done periodically, provides historical layers of data, which are snapshots of the institutional information system. This DWH is designed principally to edit reports on paper, the Intranet, or by e-mail.

A data warehouse is a set of snapshots of the institution's information system. It is designed to provide on demand indicators to all the concerned individuals of the university.

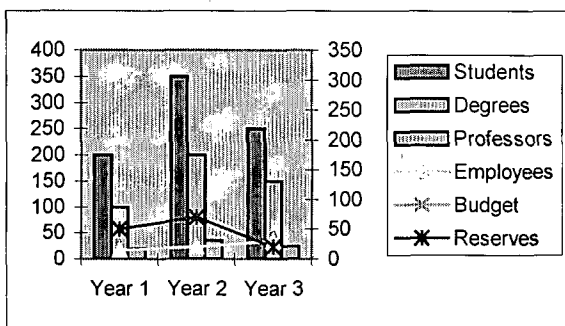


Figure 1: Some key-numbers for a given field of studies

The production data bases are continuously evolving and are organized for transactional use: for example, register a student, print a money order, or pay a new employee.

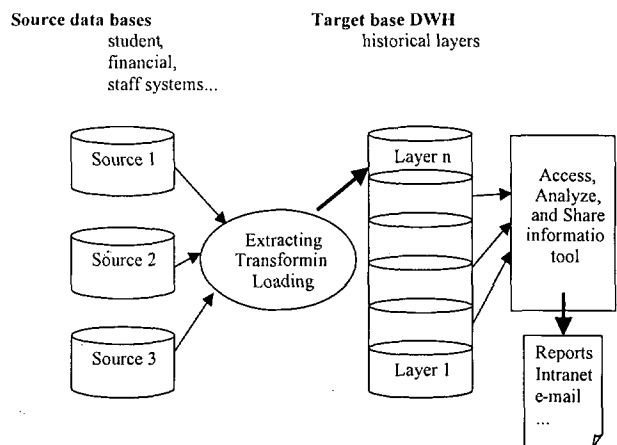


Figure 2: Principles of operations for a DWH

For French universities, the source data bases are the student, financial, and staff systems, all of them provided nationally, plus all the local applications, and also data coming from outside (for comparative reports for

example). The DWH has to take into account the complexity, variety and the dynamic character of the source data bases as well as the unsteady quality of source data.

2 The steps of the project

The project schedule was defined considering recommendations from (Kimball, 1996), and includes the following steps:

- requirements specifications,
- building the meta-dictionary,
- target data base design,
- software development,
- delivery of the first release to pilot sites,
- beta-testing,
- development of new, improved releases that will be available at no charge to all interested universities.

During the requirements specifications phase, three "scenarios" or classes of indicators were defined by a working committee, with university presidents, general secretaries, statisticians and information technology managers from nine institutions as members. The specifications were validated in March 2000.

Then a data warehouse meta-dictionary was built that describes the pieces of data extracted from source databases, how the information is treated, and then transferred to the target. This very important step of DWH design was validated by the steering committee in August 2000.

As most of the sources are Oracle 7 based databases, the Oracle 8i standard relational technology (Oracle 8 library, 2000) was chosen for target data base implementation.

During the software development phase, two kinds of procedures had to be implemented:

- procedures that extract data from source data bases, transform them, and load the data warehouse (viz. ETL – Extraction, Transformation, and Loading procedures), and
- procedures for querying data warehouse and presenting results to end users.

In order to ensure good productivity and easier maintenance of software we decided to build ETL procedures using a commercially available tool rather than full in-house software development. After a two-month comparative study, Data Stage (Ascential Software, 2000) was selected, versus Genio from Hummingbird. We also looked at Oracle Warehouse Builder, Decision Base from Computer Associates, and Sunopsis.

On the other hand, Business Objects R5 (Business Objects, 1999) were selected for target extractions, data

analysis (slice and dice), and presentation of data to end users.

The first release was delivered to pilot sites in November 2000. The five pilot universities are Jules Verne, (Amiens), P. et M. Curie, (Paris), Louis Pasteur, (Strasbourg), Rennes 1, and Versailles. Four of them: Amiens, Rennes, Strasbourg, and Versailles, use the same student, (Apogée), financial, (Nabuco), staff (Harpège), and payroll (Paye) systems. Major sources thus become the same, but of course, other locally built sources exist.

Paris does not run Apogée, and also had the ETL tool, Genio, before the project began. This pilot institution has therefore to adapt ETL procedures to their student source, and to their ETL tool.

A number of technical documents have been delivered with the software:

- Technical recommendations and specifications,
- Design and development documentation,
- Examples of outputs, (about 30),
- An installation methodology.

In 2001, a technical meeting with pilot teams is scheduled in Paris about every month. The steering committee meets every two months.

A minor corrective release was delivered in January 2001, while major new release, including evolutions of the target database and querying possibilities will be ready in April 2001.

A national seminar will be held in May hosted by Agence de Modernisation to report on this experiment to all interested French universities. In July 2001, a CD-ROM including all software and documentation of the project will be available at no charge to all interested universities.

3 The specifications

The committee, which made the specifications for the pilot issues of the data warehouse, defined three executive information boards, which have been called "scenarios".

The first scenario is about the evolution of the number of professors with regard to the research and teaching needs. The aim is to help a president's decision about recruitment. It produces a set of reports on the number of students and teachers per the field of study, and the ages of teachers, (to forecast retirement schedules). The same information is provided regarding laboratories, thesis directors, and researchers.

The second scenario is a synthetic presentation of the faculties making up the institution. Human Resources, (academics, associates, staff), and the financial means

regarding number of students, number of degrees offered, and the developments over several years are analysed and reported in this scenario.

The third scenario aims to measure the international attractiveness of the institution. The purpose is to control the institution's policy on international reputation and exchanges. All students involved in exchange programs with foreign universities, invited professors, and the exchange of scientific and collaborative programs are listed per field and academic year with the budgets involved.

4 The target database

The target or data warehouse itself is a ROLAP Oracle 8i database consisting of:

- Four principal tables: STUDENT, PERSONNEL, BUDGET, and HC, (HC is made of extractions from payroll). Each of these tables gathers information from the various operational applications.
- A number of wording and aggregate tables. Columns of aggregate tables are calculated during the loading process: for example the number of professors in a given field.
- Two "service" tables: The "structure of the university" table, and the "observation dates" table. The "structure of the university" table gives a unique code to each faculty member, and each school of the university. This is necessary because all of the operational source application has its own code and wording. The "observation dates" table permits that the Business Objects tool creates reports using data from different layers. For example, a report can be created based on the snapshot on 15/01/2001 of the student system, and the snapshot on 01/01/2001 of the financial system, etc... In these tables the value of an indicator is a number (derived from a boolean) the value of which is 0 (false) or 1 (true). Indicators will simplify the scorings with Business Objects.

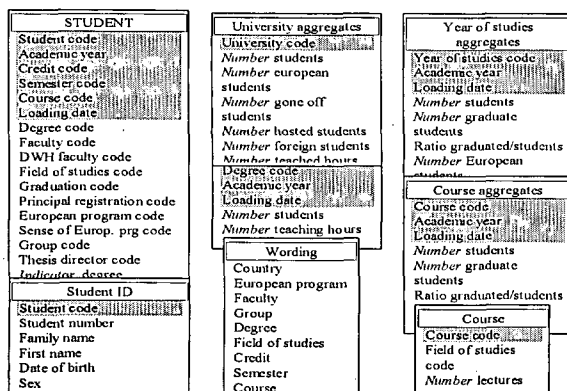


Figure 3: STUDENT and subsidiary tables. The STUDENT table and its subsidiary tables are represented in Figure 3, avoiding specific French system data.

5 Metadata

Metadata (Kimball, 1998) include definitions of all items (data fields, aggregates, fact tables) in the data warehouse, describing the way how they are extracted, calculated, and inserted in target tables and columns. Our metadata repository being a 55 page document, is listed here with only few general notions.

- Academic year: it's a key concept of the DWH. For most of extracted data, the reference year is the academic year, except for the financial data, where it is the calendar year. The existence of two reference periods is one of the difficulties of the DWH building and querying.
- Course: the educational organization is tree-structured. Only terminal elements of the tree-structure are used for calculations.
- Faculty, school: some of them do not appear in all the source application. The structure of the institution and the codes can also be different. A common "institution structure" table, with source and structure links, has been added to the DWH tables.
- Teaching units: the payment of teaching hours for an associate professor depends on what is done: lectures, tutorials, labs, and on the teacher's rank.

6 Data Stage job example

In this section an example of ETL process, viz. loading the budget, is given. Job parameters are (1) university structure code and (2) a set of calendar years. Extraction, transformation, and loading of budget data consists of the following steps (see Figure 4):

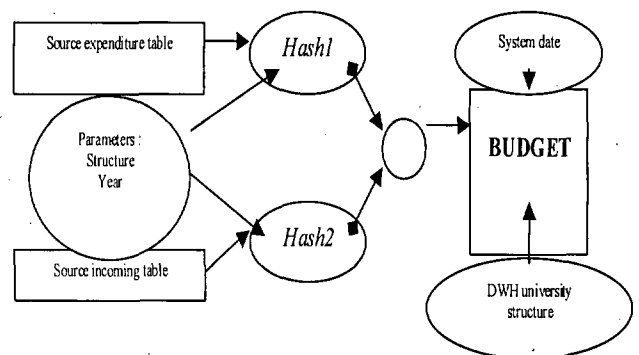


Figure 4: Loading DWH BUDGET table

- Expenditure data is extracted from the financial source database. Data is written in a hash file for performance reasons.
- Income data is extracted from the financial source database and written to the second hash file.
- The BUDGET target table is updated from hash1 and hash2, with the loading date and the university

structure code (expenditure and income do not necessarily correspond).

Why use a commercial ETL tool ?

ETL procedures could be developed using SQL and C for example, and not a specific tool, which is expensive (15,000 to 50,000 Euro). But such an ETL tool allows:

- faster development,
- more reliable corrections and evolutions,
- better adaptation to changes in sources and target bases,
- adaptability to a changing information system, and
- when operating, loading of layers ("snapshots") by a planning tool.

7 Target queries

In our implementation, a commercial software tool Business Objects (BO) is used to extract, analyze, and present data discovered in a data base (data mining). BO is designed to make these operations easier. A BO query extracts data and presents it in sheets which can be dynamic (multidimensional "slice and dice"), and published on the Intranet. With the present release of the project (March 2001) two Business Objects universes are provided.

BO is used by two categories of users:

- IS specialists, who define the universe (in other words, the architecture of data extraction), and
- End users, who can either create new queries, or simply set the parameters and run existing queries.

Navigating trough a report

Business Objects variables, also called objects, are mainly of two types:

- dimension: a dimension has a list of values,
- indicator: a number.

Let's take a simple example of a dimensional star schema which is a simplified extraction from our datawarehouse tables. It consists of a fact table with four dimensions:

- school,
- academic year,
- nationality,
- student code,

and one indicator:

- registration (0/1)

The corresponding schema is shown on Figure 5. The grain of the fact table is the student registration (for each individual) per school, year and nationality.

The values of the dimensions are:

- For *school*: Chemistry, Computer Science, Economics, etc...
- For *academic year*: 1999-2000, 2000-2001, etc...
- For *nationality*: Italy, France, etc...

- For *student*: the student registration number.

The indicator "registration" is a number, having value 0 or 1, which is to be summed.

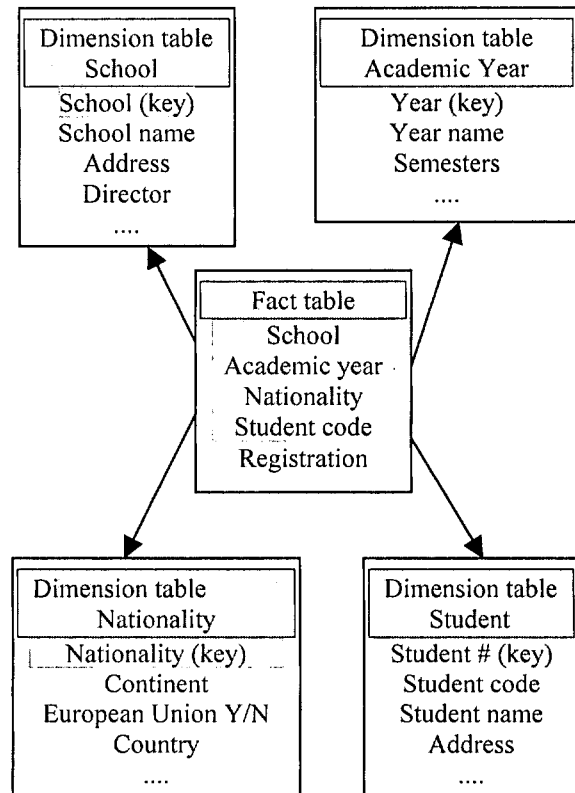


Figure 5: A simple dimensional star schema

The number of registrations can be represented on an array, here a 3-dimensional array, also called *cube* as there are three dimensions (see Figure 6).

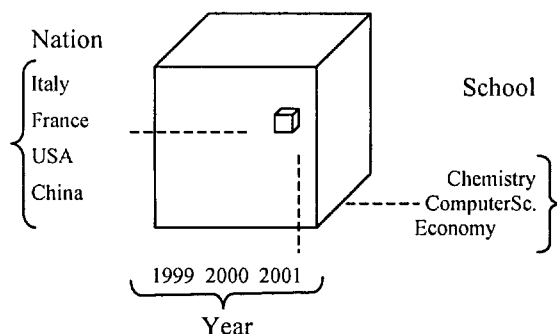


Figure 6: The Data Cube

If one of the dimensions is set as a constant, for example fixing the value of "School" to "Computer Science", we obtain a two dimensional array (spread sheet type), and in each element of the array, the number of registrations per year and nationality for the school " Computer Science". In the same way, a value of the "Academic year" or the "Nationality" could be fixed. This fixing

operation is called a dimension rotation or a "slice and dice" of the cube.

There are three different ways of performing the slice and dice with three dimensions. More generally, if an indicator is associated to n dimensions ($n > 3$) the cube becomes an hyper-cube, but is still called a cube with linguistic simplicity in mind. Several indicators are generally associated to the n dimensions. In our example, the second indicator could be the number of PhD thesis.

Drilling

Let's consider our "school" dimension. If a school can be split in "graduate" and "undergraduate" parts, and in "fields of studies" in each part, the "school" dimension can be replaced by three hierarchical dimensions (inclusion relation):

- Level 1: school
- Level 2: level of studies (graduate/undergraduate)
- Level 3: field of studies

Details can then be obtained on the two indicators (number of registrations and number of PhD thesis). The elements of the array can be detailed by level of study, or by field of study, and this is called the *drill down*, or, on the opposite, summed by level, or for the entire school, which is called *drill up*. These operations are called navigation operations, down (detailing) or up (grouping).

In the same way, "academic years" could be made of shorter periods (semesters, trimesters), and nationalities could be gathered by continents or others sets (European union for example). Navigation should then be done on the three hierarchical sets of dimensions, with a number of reports possibilities. This is called multi-dimensional navigation.

To conclude on navigation, when data are organized on proper hierarchical dimensions and indicators, there are two classical analysis methods, which are both supported by Business Objects, and used in our project:

- slice and dice, and
- drill up and down.

8 Conclusion

This project is a national attempt to give each French university a starting core model for the future development of its own Decision Support System. This was feasible because in 1992 French universities organized a software consortium which provides most members their main administrative systems. The pilot universities will test the prototype, and adapt it with the project team, until July 2001. A free delivery to volunteer universities will take place from the end of the year 2001. This project is part of a general modernization of management processes started in the universities under the auspices of the Agence de Modernisation des Universités.

9 Acknowledgements

Many thanks to all the participants to the project:

- Josette Soulas, Suzanne Maury-Silland, Christian Charrel, Sibylle Rochas (Agence management),
- Roselyne Casteloot, Jean-Pierre Finance, Paul Personne, Jacques Francois, Sylvie Robert, Michel Roignot, Marylène Oberlé, Jean-Emmanuel Rudio, Danièle Savage (members of the steering committee),
- Bernard Barbez, Emmanuelle Cravoisier, Mostafa Al Haddad, Claude Derieppe, Brigitte Perrigault, Anne Routeau, François Cadé, Elisabeth Flenet, Olivier Raunet, Valéry Vaillant, Dominique Fiquet, Nicolas Courtay, Robert Rivoire (members of the working committee),
- Marie-Hélène Glénat, Virginie Garnier and Nicolas Maume (project developers, Agence Grenoble),

10 References

- [1] Inmon W.H., Building the Data Warehouse, Wiley 1996
- [2] Kimball R., The Datwarehouse Toolkit, Wiley 1996
- [3] Kimball R., The Datawarehouse Lifecycle Toolkit, Wiley 1998
- [4] Oracle 8 library, Oracle Corp., 2000
- [5] Business Objects Designer's Manual 5.1, Business Objects Corp. 1999
- [6] Ascential Software, Data Stage Reference Manual, Ascential Software Corp., 2000

Data quality: A prerequisite for successful data warehouse implementation

Viljan Mahnič and Igor Rožanc
 University of Ljubljana
 Faculty of Computer and Information Science
 Tržaška 25, SI-1000 Ljubljana, Slovenia
 E-mail: Viljan.Mahnic@fri.uni-lj.si, Igor.Rozanc@fri.uni-lj.si

Keywords: data warehouse, information quality, total quality management, data quality assessment

Received: January 14, 2001

Building a data warehouse for a large decentralized university such as the University of Ljubljana is an attractive challenge, but also a risky and demanding task. Experience has shown that projects attempting to integrate data are especially vulnerable to data quality issues. Therefore, before embarking on a data warehouse initiative a thorough quality assessment of the source data is necessary. We describe how the assessment criteria based on the Total Quality data Management Methodology were adapted to our specific needs and used to determine the quality of student records data at two member institutions, viz. the Faculty of Computer and Information Science, and the Faculty of Electrical Engineering. The most important results of the assessment are described and proposals are given for further activities. The assessment has shown that the student records data at the Faculty of Computer and Information Science and Faculty of Electrical Engineering are good enough to be used as source for the global warehouse at the university level after some data cleansing takes place. Additionally, special attention must be devoted to the integration of such data that are replicated at many individual departments (viz. employees, subjects taught, and students). Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught be defined in the first step of the data warehouse design, and an ongoing data quality management process is established clearly defining the roles and responsibilities of all personnel involved.

1 Introduction

The University of Ljubljana is the largest university in Slovenia. It consists of 26 member institutions (20 faculties, 3 academies, and 3 colleges) and has more than 40,000 students. In the past, the member institutions had substantial autonomy regarding the usage of information technologies, which led to uncoordinated development of their information systems. Different applications were developed for the same purpose and due to this heterogeneity it is very difficult or even impossible to create reports that require cross-referencing data from different institutions or application areas [1].

In such a situation, the building of a data warehouse at the university level seems to be an appropriate solution. Therefore, a pilot project started with the aim of defining a data model for a global data base which will be fed by data from member institutions on a regular basis and which will serve as a basis for analytical processing at the university level. One of the main tasks within this project is to define the granularity of data in the data warehouse and different levels of detail that will support best the decision making processes.

However, experience from other organizations has shown that projects attempting to integrate data are especially vulnerable to data quality issues [2]. A recent study by

the Standish Group states that 83 percent of data migration projects overrun their budget (or fail) primarily as a result of misunderstandings about the source data and data definitions. Similar surveys conducted by the Gartner Group point to data quality as a leading reason for overruns and failed projects [3].

In order to avoid such a pitfall, a thorough assessment of the quality of data that are used as input to the global data warehouse is necessary. Since we decided that our data warehouse will be populated gradually, starting with student records data, the quality of these data was analyzed first. The aim of this paper is to describe in detail the assessment methodology and results obtained at two typical member institutions, viz. the Faculty of Computer and Information Science (FCIS), and the Faculty of Electrical Engineering (FEE).

The assessment criteria were defined according to English's Total Quality data Management methodology [4] and Forino's recommendations [3]. A description of these criteria is given in Section 2, while the results of the assessment are presented in section 3. Section 4 describes the proposals for further activities, and section 5 summarizes the most important conclusions.

2 Data Quality Assessment Methodology

Data quality assurance is a complex problem that requires a systematic approach. English [4] proposes a comprehensive Total Quality data Management Methodology (TQdM), which consists of 5 processes of measuring and improving information quality, and an umbrella process for bringing about cultural and environmental changes to sustain information quality improvement as a management tool and a habit:

- Process 1: Assess Data Definition & Information Architecture Quality
- Process 2: Assess Information Quality
- Process 3: Measure Nonquality Information Costs
- Process 4: Reengineer and Cleanse Data
- Process 5: Improve Information Process Quality
- Process 6: Establish the Information Quality Environment

Each process is further divided into steps that must be followed in order to achieve the desired data quality. Organizations embarking on data warehouse initiatives and that do not yet have an information quality function must conduct many of these steps, but may do so in a different sequence, based on their specific needs.

Considering our specific needs, we concentrated on Process 2 which defines two aspects of information quality: the inherent information quality and the pragmatic information quality. Inherent information quality is the correctness or accuracy of data, while pragmatic information quality is the value that accurate data has in supporting the work of the enterprise. In this paper the results of inherent information quality assessment are described.

In order to determine the quality of data a field-by-field assessment is required. However, simply having data is not enough, but the context for which the data is to exist must also be known. To put in other terms, a clear data definition or the so called meta data must be provided [3]. Generally speaking, one can find meta data in data models, data dictionaries, repositories, specifications, etc. If current meta data does not exist, then a subject matter expert is needed, and meta data is a much-desired by-product of a data quality assessment.

The extent of assessment in great deal depends on the availability of meta data. According to [3], assessments usually focus on one or more of the following types of quality criteria:

1. Data type integrity
2. Business rule integrity

3. Name and address integrity

If the assessment team knows nothing more than field names, types and sizes, then the focus is on testing the field's integrity based on its type (numeric, alphanumeric, date, etc.). If additional characteristics of the field are provided (domain, relationship with other fields, etc.), then business rule integrity is also performed. Finally, if name and address data is critical (particularly if it will be consolidated with other data), then name and testing should be performed.

On the other hand, English [4] defines the following inherent information quality characteristics and measures:

1. Definition Conformance
2. Completeness (of values)
3. Validity, or business rule conformance
4. Accuracy to surrogate source
5. Accuracy (to reality)
6. Precision
7. Nonduplication (of occurrences)
8. Equivalence of redundant or distributed data
9. Concurrency of redundant or distributed data
10. Accessibility

Considering the aforementioned quality characteristics and our specific needs we decided to measure the quality of our data using the following criteria:

1. **Completeness of values:** Users were encouraged to prioritize a subset of source fields that must have a nonnull value. For each field the percentage of records with missing values was computed.
2. **Validity, or business rule conformance:** For a chosen subset of most important fields we measured the degree of conformance of data values to their domains and business rules.
 - a) All 1:N relationships were examined for existence of foreign keys in master tables.
 - b) Since time represents an important dimension in a dimensional data warehouse a range check was performed on all date fields in order to find possible out-of-range values.
 - c) A range check was performed on all other fields from the highest priority subset (e.g. academic year, year of study, grades etc.).

d) Special cross-checks were defined for fields having relationships with other fields that further define the allowable domain set, e.g.

- when a student applies for an exam for the first time (i.e. NO_OF_ATTEMPTS=1) the date of previous examination must be blank and vice versa;
- the date of previous examination must be lesser than the date of next examination;
- in each degree record the difference between the degree date and thesis issue date must be positive and less than six months.

3. Nonduplication of occurrences and equivalence of redundant and distributed data: In our case two different tests were performed:

a) All files in the student records database were checked for eventual duplications of primary keys.

b) The student and employee files at both faculties were checked for the existence of the records that are duplicate representations of the same student or employee respectively.

3 Assessment Results

3.1 Completeness of values

The most important fields in 33 files were checked for nonnull values, and it was found, that only few values were missing. Results are summarized in Table 1, while Tables 2 and 3 show the fields and files with most missing values for each faculty, respectively. The greatest problem represent the missing student_id values in alumni records. This is a consequence of the fact that 15 years ago (when alumni records started) some students were not assigned a student identification number.

Faculty	No. of files	No. of checked attributes	No. of checked records	No. of records with missing values	% of erroneous records
FCIS	33	69	201617	43	0.021 %
FEE	33	69	405436	996	0.246 %
TOTAL	66	138	607053	1039	0.171 %

Table 1: Completeness of vaules (Summary Data)

Field name	Field description	File name	Total no. of records	No. of records with missing values	% of erroneous records
VPIS_ST	student id	DIPLOMA	961	18	1.873 %
IME	employee first name	DELAVEC	391	4	1.023 %

Table 2: Fields with most missing values (Faculty of Computer and Information Science)

Field name	Field description	File name	Total no. of records	No. of records with missing values	% of erroneous records
VPIS_ST	student id	DIPLOMA	2926	185	6.323 %
IME_D	employee first name	DELAVEC	399	4	1.003 %
DELAVEC	employee id	VAJE	152071	737	0.485 %

Table 3: Fields with most missing values (Faculty of Electrical Engineering)

3.2 Validity, or business rule conformance

Existence of foreign keys in master tables: The examination of all 1:N relationships revealed that in about a half percent of records foreign keys do not have their counterpart values in the corresponding master tables (see Table 4). However, this average is misleading, since the majority of errors appear within a

small number of relationships. The following relationships appeared to be the most problematic at both faculties (see Tables 5 and 6):

1. the relationship between entities DIPLOMA (student's degree thesis) and NACIN (a code table of possible types of study, e.g. full-time or part-time);
2. multiple relationships between entities STUD_F (student's personal data) and OBCINA (a code table

of territorial units in Slovenia) representing the territorial unit of student's residence, place of birth, secondary school finished etc.

The first of the aforementioned problems is simply a consequence of the fact that (within the alumni records database) the faculties started collecting the type of study data in 1994, while this datum is missing in older

records. The second problem is much harder and was caused by significant changes in territorial organization at the local level in Slovenia after independence, which required several consequent modifications of the corresponding code table. Although each year's enrolment data corresponded to the currently valid version of the code table, there are a lot of inconsistencies when we look at the data in time perspective.

Faculty	No. of files	No. of checked relationships	No. of checked records	No. of relationships with errors	No. of records with errors	% of relationships with errors	% of records with errors
FCIS	49	207	1361758	38	6360	18.357 %	0.467 %
FE	49	207	2702005	51	16605	24.638 %	0.615 %
TOTAL	98	414	4063763	89	22965	21.498 %	0.565 %

Table 4: Existence of foreign keys in master tables (Summary Data)

Entity (File)	No. of records	Foreign key field		No. of non existent values	% of errors
		Name	Description		
DIPLOMA	961	NACIN	type of study code	546	56.816 %
STUD_F	4062	OBCINA_SS	territorial unit code of student's secondary school	1340	32.989 %
STUD_F	4062	OBCINA_R	territorial unit code of student's place of birth	1331	32.767 %
STUD_F	4062	OBCINA_S	territorial unit code of student's permanent residence	1314	32.349 %
STUD_F	4062	OBCINA_Z	territorial unit code of student's temporary residence	603	14.845 %

Table 5: Problematic relationships (Faculty of Computer and Information Science)

Entity (File)	No. of records	Foreign key field		No. of non existent values	% of errors
		Name	Description		
DIPLOMA	2926	NACIN	type of study code	1461	49.932 %
STUD_F	8164	OBCINA_SS	territorial unit code of student's secondary school	3565	42.908 %
STUD_F	8164	OBCINA_R	territorial unit code of student's place of birth	3503	42.896 %
STUD_F	8164	OBCINA_S	territorial unit code of student's permanent residence	3502	42.896 %
STUD_F	8164	OBCINA_Z	territorial unit code of student's temporary residence	1422	17.418 %

Table 6: Problematic relationships (Faculty of Electrical Engineering)

Range checking of date fields did not reveal any serious problems. In the worst case the rate of out-of-range field values reached 0.149 % at FCIS, and 0.493% at FEE.

Range checking of other fields from the highest priority subset also yielded quite good results. The filed IME_D (viz. employee's first name) was ranked the worst at both faculties containing 1.023 % erroneous values at FCIS, and 1.003 % erroneous values at FEE.

Special cross-checks pointed out that the business rule requiring that each student completes his/her degree project in six months is sometimes violated. Namely, the difference between the degree date and thesis issue date was more than six months in 0.976 % of cases at FCIS, and in 1.805 % of cases at FEE. Error rates reported by other cross-checking tests were less than 0.5 %.

3.3 Nonduplication of occurrences and equivalence of redundant and distributed data

Nonduplication of primary keys: The student records information system at FCIS and FEE [5] (as well as at other member institutions of the University of Ljubljana) is implemented using Clipper which does not automatically force the uniqueness of primary keys. In spite of the fact that all programs have been carefully written in order to avoid duplicates, some can be introduced through the manual maintenance of data,

especially code tables. Therefore, all files in the student records database were checked for eventual duplications of primary keys. An excerpt of assessment results showing only files with most duplicates is presented in tables 7 through 9. Table 7 shows the number of duplicated primary keys in code tables that are maintained centrally by the University computing center for all member institutions, while tables 8 and 9 refer to duplicates at FCIS and FEE respectively. A relatively high percentage of errors in some code tables indicates that the maintenance of code tables at the University computing center should be improved.

File	No. of records	No. of duplicated primary keys	% of errors
VSI.DBF	1314	68	5.175 %
SS_POKLI.DBF	1566	42	2.682 %
CENTRI.DBF	51	1	1.960 %
ZAVOD.DBF	100	1	1.000 %

Table 7: Nonduplication of primary keys (Code tables maintained by the University computing center)

File	No. of records	No. of duplicated primary keys	% of errors
IZJEME.DBF	56	1	1.786 %
TEMA.DBF	978	3	0.307 %
DELAVEC.DBF	391	1	0.226 %
DVIG.DBF	922	1	0.108 %

Table 8: Nonduplication of primary keys (FCIS)

File	No. of records	No. of duplicated primary keys	% of errors
SPP.DBF	270	11	4.074 %
PRED_PR.DBF	746	8	1.072 %
PP.DBF	2557	12	0.469 %
DVIG.DBF	2825	9	0.319 %

Table 9: Nonduplication of primary keys (FEE)

Equivalence of redundant and distributed data: Given the fact that the some teachers teach at both faculties as well as that some students study at both faculties (e.g. a student can obtain his/her B. Sc. degree at FEE and enrol for graduate study at FCIS or vice versa) some data are replicated across faculties. In order to state the extent of

such a replication two measures were introduced: the number of replicated entity occurrences (viz. the same teacher or student in both databases) with the same primary key, and the number of replicated entity with different primary keys.

Entity (File)	Description	No. of replications	
		Same primary key	Different primary key
DELAVEC	Employees	388	0
STUDENT	Students	4	121

Table 10: Equivalence of redundant occurrences (FCIS and FEE)

Assessment revealed that employee files at both faculties are consistent: all replicated employees have the same primary key. This is not the case with student files. Due to the decentralized organization of the University of Ljubljana each faculty assigns its students a different

identification number regardless the fact that the student has already been enrolled at another faculty. Unfortunately, this kind of inconsistencies may be a source of major problems when integrating data into a global warehouse.

4 Proposed Further Actions

On the basis of assessment results we propose two kinds of further actions: cleansing of source data and an appropriate design of the global data warehouse at the university level.

Some erroneous source data can be cleansed automatically (e.g. missing type of study code in file DIPLOMA), while other data require manual or combined manual and automatic approach (e.g. removal of duplicated primary keys, re-establishment of relationships using territorial unit codes). Some errors (e.g. out-of-range values) can be prevented by the incorporation of appropriate controls in the program code.

Special attention must be devoted to the maintenance of code tables that are common for the whole university. A relatively high percentage of duplicate codes in code tables, maintained by the university computing center up to now indicates that the maintenance of these code tables must improve.

Additional problems could arise during the integration of those data and code tables that are at present maintained by individual departments (viz. employees, subjects taught, and students). Although our assessment did not reveal serious problems within each department, many duplications and code conflicts may occur during the integration. Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught is defined in the first step of the data warehouse design.

The data warehouse design must be based on principles of TQdM methodology. Data standards must be defined and data definition and information architecture quality assessment must take place before programming and population of the data warehouse begins. An ongoing data quality management process must be established and the roles and responsibilities of a data quality administrator, subject area champions, data oversight committee, and data owners clearly defined [6].

5 Conclusions

Our paper was intended to increase the awareness of the importance of data quality not only when building a data warehouse but also in operational environments that support transactional processing. An assessment methodology to empirically determine the data quality was described and the results of the assessment were presented.

Considering the assessment results we estimate that source data at the Faculty of Computer and Information Science and Faculty of Electrical Engineering are good enough to be used as source for the global warehouse at

the university level after some data cleansing takes place. In first place, missing student identification numbers must be provided in alumni records and the broken relationships using territorial units codes must be re-established in students' personal data.

During the design of the global data warehouse a special attention must be devoted to the integration of those data that may be replicated at many individual departments (viz. employees, subjects taught, and students). Since each department has its own policy of coding, many duplications and code conflicts may occur during the integration. Therefore, we propose that a unique coding scheme for all employees, students, and subjects taught is defined in the first step of the data warehouse design. Additionally, an ongoing data quality management process must be established and the roles and responsibilities of all personnel involved should be clearly defined.

6 References

- [1] Mahnič, V. Towards the re-integration of the University of Ljubljana information system, in J-F. Desnos & Y. Epelboin (eds.), *European Cooperation in Higher Education Information Systems EUNIS 97*, Grenoble, September 1997, pp. 250-258.
- [2] Celko, J., McDonald, J., Don't Warehouse Dirty Data, *Datamation*, October 15, 1995, pp. 42-53.
- [3] Forino, R., The Data Quality Assessment, Part 1, *DM Review Online*, August 2000, <http://www.dmreview.com>
- [4] English, L.P., *Improving Data Warehouse and Business Information Quality*, John Wiley & Sons, Inc., 1999, ISBN 0-471-25383-9.
- [5] Mahnic, V., Vilfan, B., Design of the Student Records Information System at the University of Ljubljana, in J. Knop (ed.), *Trends in Academic Information Systems in Europe, Proceedings of the EUNIS'95 Congress*, Düsseldorf, Germany, November 1995, pp. 207- 220.
- [6] Kachur, R., Data Quality Assessment for Data Warehouse Design, *DM Review Online*, April 2000, <http://www.dmreview.com>

'Beowulf cluster' for high-performance computing tasks at the university: A very profitable investment.

Manuel J. Galán, Fidel García, Luis Álvarez, Antonio Ocón and Enrique Rubio

CICEI. Centro de Innovación en Tecnologías de la Información.

University of Las Palmas de Gran Canaria.

Edificio de Arquitectura, Campus Universitario de Tafira Baja., 35017 Las Palmas de Gran Canaria, SPAIN.

E-mail: manolo@polaris.ulpgc.es

Keywords: Cluster, OSS, Linux, High performance computing, computer commodities.

Received: January 24, 2001

We describe a high performance/low price “computer cluster” named Beowulf Cluster. This kind of device was initially developed by T. Sterling and D. Becker (N.A.S.A) and it is a kind of cluster built primarily out of commodity hardware components, running an OSS (Open Source Software) operating system like Linux or FreeBSD, interconnected by a private high-speed network, dedicated to running high-performance computing tasks. We will show several advantages of such a device, among them we can mention very high performance-price ratio, easy scalability, recycling possibilities of the hardware components and guarantee of usability/upgradeability in the medium and long-term future. All these advantages make it specially suitable for the university environment and, thus, we make a description about the implementation of a Beowulf Cluster using commodity equipment dedicated to run high-performance computing tasks at our university focusing on the different areas of application of this device that range from production of high quality multimedia content to applications in numerical simulation in engineering.

1 Introduction

1.1 Concurrency and Parallelism

Regarding program execution, there is one very important distinction that needs to be made: the difference between “concurrency” and “parallelism”. We will define these two concepts as follows:

- “Concurrency”: the parts of a program that can be computed independently.
- “Parallelism”: the parallel parts of a program are those “concurrency” parts that are executed on separate processing elements at the same time.

The distinction is very important, because “concurrency” is a property of the program and efficient “parallelism” is a property of the machine.

Ideally, “parallel” execution should result in faster performance. The limiting factor in parallel performance is the communication speed (bandwidth) and latency between compute nodes.

Many of the common parallel benchmarks are highly parallel and communication and latency are not the bottleneck. This type of problem can be called “obviously parallel”. Other applications are not so simple and executing “concurrent” parts of the program in “parallel” may actually cause the program to run slower, thus offsetting any performance gains in other “concurrent” parts of the program. In simple terms, the cost of communication time must pay for the savings in computation time, otherwise the “parallel” execution of the “concurrent” part is inefficient.

Now, the task of the programmer is to determine what “concurrent” parts of the program should be executed in “parallel” and what parts should not. The answer to this will determine the efficiency of the application.

In a ‘so called’ perfect parallel computer, the ratio of communication/processing would be equal to one and anything that is “concurrent” could be implemented in “parallel”. Unfortunately, real parallel computers, including shared memory machines, do not behave “this well”.

1.2 Architectures for Parallel Computing

1.2.1 Hardware Architectures

There are three common hardware architectures for parallel computing:

- Shared memory machines, SMP, that communicate through memory, i.e. MPP (Massively Parallel Processors, like the nCube, CM5, Convex SPP, Cray T3D, Cray T3E, etc.). This kind of configuration is sustained on dedicated hardware. The main characteristics are a very high bandwidth between CPUs and memory.
- Local memory machines that communicate by messages, i.e. NOWs (Networks of Workstations) and clusters. In this category each workstation maintains its individuality, however there is a tight integration with the rest of the members of the cluster. So we can say they constitute a new entity known as “The Cluster”. In our proposal we will focus on a particular kind of cluster called “Beowulf Cluster” (see [3]).

- Local memory machine that integrate in a loosely knit collaborative network. In this category we can include several collaborative Internet efforts which are able to share the load of a difficult and hard to solve problem among a large number of computers executing an “ad hoc” client program. We can mention the SETI@home (see [4]), Entropia Project (see [5]), etc.

The formed classification is not strict, in the sense that it is possible to connect many shared memory machines to create a “hybrid” shared memory machine. These hybrid machines “look” like a single large SMP machine to the user and are often called NUMA (non-uniform memory access). It is also possible to connect SMP machines as local memory compute nodes. The user cannot (at this point) assign a specific task to a specific SMP processor. The user can, however, start two independent processes or a threaded process and expect to see a performance increase over a single CPU system. Lastly we could add several of this hybrid system into some collaborative Internet effort building a super hybrid system.

1.2.2 Software Architectures

In this part we will consider both the “basement” software (API) and the application issues.

1.2.2.1 Software API (Application Programming Interface)

There are basically two ways to “express” concurrency in a program:

- Using Messages sent between processors: A Message is simple: some data and a destination processor. Common message passing APIs are PVM (see [6]) or MPI (see [7]). Messages require copying data while Threads use data in place. The latency and speed at which messages can be copied are the limiting factor with message passing models. The advantage to using messages on an SMP machine, as opposed to Threads, is that if you decided to use clusters in the future it is easier to add machines or scale up your application.
- Using operating system Threads: They were developed because shared memory SMP designs allowed very fast shared memory communication and synchronization between concurrent parts of a program. In contrast to messages, a large amount of copying can be eliminated with threads. The most common API for threads is the POSIX API. It is difficult to extend threads beyond one SMP machine, It requires NUMA technology that is difficult to implement.

Other methods do exist, but the formers are the most widely used.

1.2.2.2 Application Issues

In order to run an application in parallel on multiple CPUs, it must be explicitly broken into concurrent parts. There are some tools and compilers that can break up programs, but parallelizing codes is not a “plug and play”

operation. Depending on the application, parallelizing code can be easy, extremely difficult, or in some cases impossible due to algorithm dependencies.

1.3 Definition of Cluster

Cluster is a collection of machines connected using a network in such a way that they behave like a single computer. Cluster is used for parallel processing, for load balancing and for fault tolerance. Clustering is a popular strategy for implementing parallel processing applications because it enables companies to leverage the investment already made in PCs and workstations. In addition, it's relatively easy to add new CPUs simply by adding a new PC or workstation to the network.

1.4 The Beowulf Cluster

Beowulf is not a special software package, new network topology or the latest Linux kernel hack. It is a kind of cluster built primarily out of commodity hardware components, running an OSS (Open Source Software) operating system like Linux or FreeBSD, interconnected by a private high-speed network, dedicated to running high-performance computing tasks.

One of the main differences between a Beowulf Cluster and a COW (Cluster of Workstations) is the fact that Beowulf behaves more like a single machine rather than many workstations. The nodes in the cluster don't sit on people's desks; they are dedicated to running cluster jobs. It is usually connected to the outside world through only a single node.

While most distributed computing systems provide general purpose, multi-user environments, the Beowulf distributed computing system is specifically designed for single user workloads typical of high end scientific workstation environments.

Beowulf systems have been constructed from a variety of parts. For the sake of performance some non-commodity components (i.e. produced by a single manufacturer) have been employed. In order to account for the different types of systems and to make discussions about machines a bit easier, It has been propose the following classification scheme:

CLASS I: This class of machines built entirely from commodity vendor parts. We shall use the “Computer Shopper” certification test to define “commodity off-the-shelf” parts. (Computer Shopper is a one-inch thick monthly magazine/catalog of PC systems and components). A CLASS I Beowulf is a machine that can be assembled from parts found in at least 3 nationally/globally circulated advertising catalogs.

The advantages of a CLASS I systems are: hardware is available form multiple sources (low prices, easy maintenance), no reliance on a single hardware vendor, driver support from O.S. commodity, usually based on standards (SCSI, Ethernet, etc.).

On the other side, the disadvantages of a CLASS I systems are: best performance may require CLASS II hardware.

CLASS II: This class is simply any machine that does not pass the Computer Shopper certification test.

The advantages of a CLASS II system are: performance can be quite good

The disadvantages of a CLASS II system are: driver quality of support may vary, reliance on single hardware vendor, may be more expensive than CLASS I systems.

One class is not necessarily better than the other. It all depends on your needs and budget. In the last times we are seeing an increment in the number CLASS II Beowulf Clusters using 64-bit ALPHA Processors due to the large performance increase that can be achieved.

1.5 Evolution of the Beowulf Cluster: State of the Art

The first Beowulf-class computers that achieved the gigaflops goal appeared at Supercomputing '96 in Pittsburgh. One of those came from collaboration between Caltech and the Jet Propulsion Laboratory and the other from Los Alamos National Laboratory. Both systems consisted of 16 200-megahertz Pentium Pro processors and were built for about \$50,000 in the fall of 1996. One year later, the same machines could be built for about \$30,000.

In a paper for the 1997 supercomputing meeting -- simply called SC97 -- Michael Warren of Los Alamos and his colleagues wrote: "We have no particular desire to build and maintain our own computer hardware. If we could buy a better system for the money, we would be using it instead." (See [8]).

Finally, we can mention that the Avalon, which is a cooperative venture of the Los Alamos National Laboratory Center for Nonlinear Studies and Theoretical Division, built as a 140 64-bit processors Alpha Beowulf Cluster machine appeared as the 265th in the list of the fastest computer systems in the word (see [9]).

There are a lot of Beowulf Clusters spread around the word dedicated to every kind of computationally intensive task. Among them we can mention:

- Stone SuperComputer Oak Ridge National Lab (ORNL) a 126 node cluster at zero dollars per node. The system has already been used to develop software for large-scale landscape analysis (see[10]).
- The SuperAbacus is an implementation in the CityU Image Processing Lab at City University of Hong Kong. to support multimedia signal processing (see[11]).
- LoBoS Supercomputer for Molecular Graphics and Simulation Laboratory, National Institutes of Health NIH, (see[12]). This cluster is dedicated to study more complex biological systems using computational methods.

Beowulf Clusters are also deployed in our country (Spain), they are used also for intensive computation. The most mature projects could be:

- HIDRA University of Barcelona's UB-UPC Dynamical Systems Group dedicated to several projects that require a huge amount of computations (i.e., numerical simulations of continuous and discrete systems, bifurcation analysis, numeric and symbolic computation of invariant manifolds, etc.) (see [13]).

- LAMA's Materials Laboratory at UPV/EHU running Monte Carlo simulations of phase transitions in condensed matter physics (see [14]).

1.6 Characteristics of the Beowulf Cluster

Commodity networking, especially Fast Ethernet, has made it possible to design distributed-memory systems with relatively high bandwidths and tolerably low latencies at a low cost.

Free operating systems, such as Linux, are available, reliable, and well supported, and are distributed with complete source code, encouraging the development of additional tools including low-level drivers, parallel file systems, and communication libraries.

With the power and low prices of today's off-the-shelf PCs and the availability of 100/1.000 Mb/s Ethernet interconnect, it makes sense to combine them to build High-Performance-Computing and Parallel Computing environment.

With free versions of Linux and public domain software packages, no commercially available parallel computing system can compete with the price of the Beowulf system.

The drawback to this system is, of course, that there will not exist any "support center" to call when a problem arises (anyway, "support centers" are many times only marketing hype and do not provide real support). We can say that the Open Source Support Center is the whole Internet, in the sense that there does exist a wealth of good information available through FTP sites, web sites and newsgroups. Besides that you can also since a maintenance agreement with any of the increasing number of companies that provide commercial support to these installations.

Another key component contributing to forward compatibility is the system software used on Beowulf. With the maturity and robustness of Linux, GNU software and the "standardization" of message passing via PVM and MPI, programmers now have a guarantee that the programs they write will run on future Beowulf Clusters, regardless of who makes the processors or the networks.

That said, the main characteristics a Beowulf Cluster can be summarized in the following points:

- Very high performance-price ratio.
- Easy scalability.
- Recycling possibilities of the hardware components.
- Guarantee of usability/upgradeability in the future.

2 Our proposal of Beowulf Cluster for the University of Las Palmas

All the good characteristics of the Beowulf Cluster justify its deployment in any organization that require high computational power. In the case of an academic institution we can say that it is not only advisable but also imperative. The Beowulf cluster is not only a wonderful tool to provide high computing power to the University but at the same time, is a very interesting object of study "per se". The evaluation of its

performance, adaptability, scalability, its behavior regarding the parallelization of procedures, etc. is a field of study that we suspect full of findings.

2.1 System Description

Our system has a hardware part and a software part. Hardware part consists in eight PCs connected by a Fast Ethernet switch at 100 Mb/s. One of this PCs is the cluster's console that controls the whole cluster and is the gateway to the outside world (master node). Nodes are configured and controlled by the master node, and do only what they are told to do.

The proposed node configuration consists of an AMD single processor based PC at 750 Mhz, with 256 megabytes of RAM and including a local IDE hard disk drive of 8GB capacity. They have also a non-expensive video card and floppy disk drive. Besides they will be provided with a Fast Ethernet network interface card.

The master node is provided with a larger hard disk drive (24 MB) and a better graphics video card, besides it has a second network interface and also CD-ROM drive, medium sized monitor and keyboard and mouse to be able to perform the controlling tasks for the cluster.

The proposed Fast Ethernet switch will have 24 autosensing ports and will be SNMP capable.

In the description of the hardware we can see that only one node needs input/output devices. The second network interface card in the master node is used to connect the Intranet to the Internet. The switch has a number of port bigger than strictly necessary to can enlarge the cluster in the future.

The logical part will be built using GNU / Linux operating system according to the distribution "Extreme Linux CD" with additional OSS software such as kernel modifications:

- PVM: Parallel Virtual Machine: PVM is a software package that permits a heterogeneous collection of Unix / Linux or NT computers hooked together by a network to be used as a single large parallel computer. It is freely-available, portable, message-passing library generally implemented on top of sockets. It is clearly established as the de-facto standard for message-passing cluster parallel computing (see [6]).
- MPI libraries: Message Passing Interface: Communication between processors on a Beowulf Cluster is achieved through the Message Passing Interface (MPI). This is a standardized set of library routines. Both the C and the Fortran programming languages are supported (see [7]).

Additionally we will be proceed to the installation of several configuration, management and monitoring tools which make the Beowulf architecture faster, easier to configure, and much more usable.

2.2 Basic Software Installation and Configuration

As previously stated the installation pathway will run along the "Extreme Linux CD". The following steps will be taken:

The first step we will be installing the Master Server, which involves the following tasks:

- Partition sizes; installing Red Hat Linux; network configuration; setting up DNS; network file configuration: "/etc/hosts", "/etc/resolv.conf" and "/etc/hosts.equiv"; local file configuration: ".cshrc"; clock synchronization.

After the installation of the master server we will precede to the installation and configuration of the client nodes:

- Installing the operating system on one client; cloning clients; configuring clients.

The third step will be installation of basic application software:

- Compilers; communication Software: PVM and MPI; conversion Software; System Monitoring Software: bWatch, httpd and CGI scripts, Netpipe, netperf, NASA parallel Benchmarks, CMS.

Finally we will attend the security concerns both in the master sever and client nodes.

3 Fields of Application

There are a lot of OSS software packages optimized to run in clusters like Beowulf. We can mention the following:

- *MP_SOLVE*, *Parallel Sparse Irregular System Solvers* solving large, irregular, sparse, indefinite systems of equations with multiple excitation vectors on distributed memory parallel computers using LU factorization (see [15]).
- *NAMD* is a parallel, object-oriented molecular dynamics code designed for high-performance simulation of large biomolecular systems (see [16]).
- *POV-RAY*, *The Persistence of Vision Raytracer* is a high-quality tool for creating stunning three-dimensional graphics (see [17]).

Nevertheless there are many others applications that can take profit when run in a Beowulf Cluster their range covers from standard numerical applications, going through high intensive physical and chemical computation, biochemical modeling and multimedia and CAD applications.

4 Conclusions

In the present paper we have made quick and succinct overview about the state of distributed computing, centering our focus on a concrete cluster configuration called "Beowulf Cluster".

The advantages of this class of cluster configuration are evident for any organization that requires high computational power "for the buck". This is, when we take into account the performance/price ratio, easy scalability and 'upgradeability' and recycling properties of the hardware components. If this is true for any

organization, we are convinced that it is imperative for an academic institution like our University. Therefore we make a proposal of deployment of such a device starting with a schematic installation to be eventually enlarged and improved.

5 References

- [1] Phil Merkey, "CESDIS", 2000, <http://cesdis.gsfc.nasa.gov/>
- [2] Jim Fischer, "ESS Project Overview", 2000, <http://sdc.gsfc.nasa.gov/ESS/overview.html>
- [3] Donald Becker and Phil Merkey, "The Beowulf Project", 2000, <http://www.beowulf.org/>
- [4] Mick Evans, "SETI@HOME Pages", 2001, <http://www.kevlar.karoo.net/seti.html>
- [5] Entropia Inc., "Entropia:High performance Internet Computing", 2000, <http://www.entropia.com>
- [6] Jack Dongarra & al., "Parallel Virtual Machine", 2000, http://www.epm.ornl.gov/pvm/pvm_home.html
- [7] Lam Team, "LAM / MPI Parallel Computing", 2001, <http://www.mpi.nd.edu/lam>
- [8] Michael Warren, "Pentium Pro Inside: I. A Treecode at 430 Gigaflops on ASCI Red, II.Price/Performance of \$50/Mflop on Loki and Hyglac", 1997, <http://loki-www.lanl.gov/papers/sc97/>
- [9] Netlib TOP 500 Supercomputer Sites 2000 <http://www.netlib.org/benchmark/top500/top500.list.html>
- [10] Forrest M. Hoffman, William W. Hargrove, and Andrew J. Schultz, "The Stone SuperComputer", 1997 <http://stonesoup.esd.ornl.gov/>
- [11] Super Abacus, "Super Abacus", 2001, <http://abacus.ee.cityu.edu.hk/>
- [12] NIH Computational Biophysics Section, "The LoBoS Supercomputers", 2000, <http://www.lobos.nih.gov/>
- [13] Joaquim Font, Àngel Jorba, Carles Simó, Jaume Timoneda, "HIDRA: a home-assembled parallel computer", 1998, <http://www.maia.ub.es/dsg/hidra/index.html>
- [14] S. Ivantchev, "LAMA BEOWULF CLUSTER", 2000, <http://lcdx00.wm.lc.edu/~svet/beowulf/>
- [15] New Mexico State University, "MP_SOLVE", 1998, http://emlab2.nmsu.edu/mp_solve/
- [16] University of Illinois, "NAMD Scalable Molecular Dynamics", 2001, <http://www.ks.uiuc.edu/Research/namd/>
- [17] POV-Ray Inc., "POV-Ray", 2000, <http://www.povray.org/>
- [18] Jacek Radajewski and Douglas Eadline, "Beowulf Installation and Administration HOWTO", 1999, http://www.beowulf-underground.org/doc_project/BIAA-HOWTO/Beowulf-Installation-and-Administration-HOWTO.html

Evaluation of codec behavior in IP and ATM networks

Susanne Naegele-Jackson, Ursula Hilgers and Dr. Peter Holleczeck
 Regional Computing Center Erlangen (RRZE), Martensstrasse 1, 91058 Erlangen, Germany
 Phone: +49 9131 852 9479, Fax: + 49 9131 30 29 41
 {Susanne.Naegele-Jackson, Ursula.Hilgers, Peter.Holleczeck}@rrze.uni-erlangen.de

Keywords: QoS, ATM, IP, Video Transmissions

Received: January 27, 2001

As multimedia applications are becoming more and more widespread, there is a need for video data to be transmitted over IP as well as ATM networks. While ATM with its Quality of Service features is able to provide adequate transmission guarantees for the strong timing and bandwidth constraints of video data, the transfer of video over IP networks is an issue wherever ATM is not available and applications are used that can tolerate longer delays and occasional impairments. The study evaluates a codec for teleconferencing and teleteaching applications over an IP network. A direct comparison is made to a MJPEG codec for similar applications over ATM with empirical measurements in a simple testbed. The main focus lies on the Quality of Service parameters delay, jitter and subjective picture quality. Network overload situations and realistic network behavior are simulated and their impact on picture quality is evaluated.

1 Preface

With an increased use of video and audio transmissions over data networks the term Quality of Service (QoS) has become a primary focus of interest. Multimedia applications with their strong time constraints cannot rely on best effort service, but are in need for a guaranteed transmission quality if video and audio sequences are to be displayed at the receiver without interruptions.

Before the video signal of a camera can be sent over an IP or ATM network, coders compress and transform the signals into packets suitable for transmission. Once the packets have reached their destination it is the responsibility of decoders to regain the video signal and make it available for display.

In the first chapter of the paper QoS parameters for video transmissions are explained. This paragraph is followed by a discussion of QoS in IP and ATM networks. In chapter three an IP codec for teleconferencing and teleteaching applications is evaluated with the main focus on delay, jitter and subjective picture quality over an IP network. A direct comparison is made to a MJPEG codec for similar applications over ATM. The influence of network impairment or overload situations on the picture quality is evaluated as well.

2 Quality of Service Parameters for Video Transmissions

Quality of Service (QoS) can be described as the collective effort of service performance and as such determines the overall degree of satisfaction of a user

with a service [12]. The ISO [9] further extends this definition of QoS and includes all characteristics which can be measured or recognized by a user. Any evaluation of the quality of a service should therefore not only consider objective measurements, but a user's subjective impressions as well. Objective and subjective quality can be described by several parameters [10, 11]. Examples for objective measurements are throughput, delay, delay variation and errors or losses. The overall transmission quality can be considered as an example for a subjective parameter.

One of the main Quality of Service requirements of video applications is to have enough bandwidth available during the transmission to make room for a continuous flow of data which allows the receiving side to display a video sequence with its original data rate. If data packets are delayed the receiver can no longer display 25 frames per second (as in PAL) to deliver a flowing and uninterrupted motion picture. Movements become jerky as frames are lost and the picture may even freeze temporarily. The delay variation must be kept within very small limits and the maximum distance between arriving packets should not be exceeded. Every packet received after the given time frame is considered late and must be discarded. In addition to timing errors transmissions may also be affected by losses due to hardware or software errors. Packet loss can be caused by congestion in network nodes with buffer overflow.

To keep bandwidth requirements at a minimum a digital video signal of 270 Mbit/s is compressed to much smaller bandwidths before transmission. Data streams typically range from 1.5 Mbit/s to 40 Mbit/s streams after

compression. This process of compression - and subsequent decompression on the receiving end - adds a significant amount of delay to the overall transmission time and is accomplished by so-called codecs (encoders/decoders). Encoding and decoding delay times are very much determined by the equipment that is used and as such have - next to the underlying quality of the network - a strong impact on the overall quality of a transmission.

3 ATM vs. IP Technology for Video Transmissions

In today's networks most of the high quality video is transmitted over ATM networks [8]. ATM technology was developed especially for multimedia traffic and therefore provides excellent mechanisms for bandwidth reservations and guaranteed Quality of Service. But with the increasing integration of all kinds of services into one infrastructure, the IETF proposed the addition of QoS functionality to the IP protocol as well.

Quality of Service in the ATM Protocol. In the ATM protocol, extensive QoS and traffic management functionalities are implemented. To achieve connections with different QoS characteristics the ATM Forum has defined several service categories [1]. A distinction is made between real-time and non-real-time service classes: Constant Bit Rate (CBR) and real-time Variable Bit Rate (rt-VBR) provide real-time characteristics. Two other classes, the non real-time Variable Bit Rate (nrt-VBR) and Unspecified Bit Rate (UBR) are non-real-time services with no QoS guarantees.

To manage the network capacities and control the flow of data traffic management functions [1] such as the Call Admission Control (CAC) and the Usage Parameter Control (UPC) are specified. During the connection setup the CAC ensures that enough resources are available in the network and rejects a request if the required capacities cannot be provided.

If a call is accepted, the necessary performance and QoS parameters are guaranteed during the lifetime of the connection. To avoid congestion in network components the UPC monitors and controls the traffic and the validity of a connection. The UPC discards cells if the sending rate is higher than the previously negotiated rate.

Quality of Service in the IP Protocol. The Differentiated Services [2] architecture is a scalable approach to add QoS to the IP protocol in order to provide different service classes to applications. IP packets are classified according to the first six bits in the Type of Service (TOS) field in the IP header, which is named Differentiated Services Code Point (DSCP). Traffic profiles are checked to ensure that they conform with the characteristics of the class and packets are dropped if the Service Level Agreement (SLA) is violated. The quality of the service class determines the queuing and

scheduling algorithms and has an impact on resource allocation in the network nodes. In the worst case scenario of a congested network, however, this approach can only provide best effort service.

Another approach of the IETF is the Integrated Services proposal [3], which is based on an admission control unit, a packet forwarding mechanism and a protocol to reserve resources in network components and end devices. One such protocol presented by the IETF is called Resource Reservation Protocol (RSVP) [3]. This protocol does not offer scalability in broadband WAN interfaces because of its complex signaling and the storage of flow states in each network node and end device. But the concept is able to offer absolute guarantees concerning the quality of a service.

4 Measurements and Tests

The study evaluates two pairs of codecs with the main focus on the parameters delay, delay variation and picture quality: A pair of CellStack Classic (KNET) [4, 5] codecs for ATM networks using MJPEG as compression algorithm, and a pair of CAMVision-2 7615 [15] Litton codecs compressing in MPEG-2 (4:2:0) MP@ML format. The Litton codec was equipped with both an ATM interface and a 10/100 Ethernet interface. For the tests only the Fast Ethernet card was used.

Delay and Jitter. To measure the delay the codecs were setup back-to-back in loopback mode (Fig. 1) without any network components involved and all parameters mentioned above were evaluated:

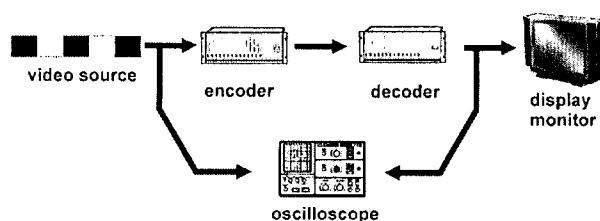


Fig. 1: Test Setup for Delay Measurements

As a video source a PAL sequence was used where 12 black and 12 white frames alternated in a nonstop loop. The video signal was connected to an oscilloscope (Tektronix 2220) [17] and at the same time fed into the encoder video input slot. After the encoding process the compressed signal arrived at the decoder and was subsequently transmitted to a second input channel on the oscilloscope. For observation and control the signal was also displayed on a monitor. With the original signal on one input channel and the delayed signal on the other channel the oscilloscope showed the time spent on the encoding and decoding process. The jitter was derived using 100 samples of measured delay values and was calculated as the difference between maximum and

minimum delay within the obtained value range (Table 1). The indicated values include both encoding and decoding times.

Codecs	Compression Format	Bandwidth	Delay	Mean Delay	Jitter
Cell-Stack Classic (ATM)	MJPEG	11.5 Mbps	90ms to 102ms	97.66ms	12ms
Litton (IP)	MPEG-2 (4:2:0) GOP=1 (I frames only)	7.2 Mbps	182ms to 222ms	202.78ms	40ms

Table 1: Delay and Jitter Measurements

The CellStack Classic codecs scored considerably better in both jitter and delay measurements compared to the Litton CAMVision CV2. According to the recommendation of the ITU [14] the delay has a disturbing impact if the video transmission is bidirectional and 150 ms are exceeded. Therefore the CellStack Classic codecs can be considered more suitable for bidirectional transfers such as in videoconferencing, for example.

Picture Quality. The quality of a video sequence can be measured objectively or subjectively [18, 7]. So far subjective tests have been used predominantly, since in the end it depends on how a human spectator evaluates the perceived quality of a service [6]. Subjective testing, however, has the disadvantage that a large number of people are required to evaluate the picture quality in a laboratory under conditions that remain the same for every test and can be reproduced at all times, if statistically relevant information is to be obtained. For this reason objective methods of evaluation have been developed which focus on high conformity with the subjective perception of a human spectator and use this degree of correlation to validate their tests [16].

The objective of this study is not to reach evaluations in a standardized laboratory that are statistically representative; instead the intention of the authors is to assess picture quality subjectively and provide some interesting results and insights into codec behavior using different compression formats over different types of networks. In the subjective evaluation the video sequences are rated according to the Mean Opinion Score (MOS) with its categories "excellent", "good", "fair", "poor" and "bad". [13]. The tests also include the simulation of WAN behavior with an impairment tool and the creation of overload situations with synthetically generated traffic.

Evaluation of Picture Quality with Impairment Tool. A camera zoomed in onto a moving metronome was used as a video source (Fig. 2). The picture qualities were

varied by introducing errors with an Impairment Tool (Interwatch 95000) to simulate WAN behavior.

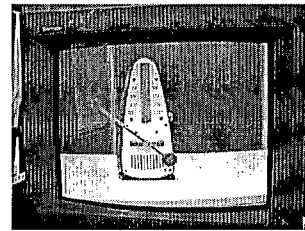


Fig. 2: Video Input Signal for Subjective Evaluation of Picture Quality

The video signal traveled from the encoder over two network components to the decoder and was then displayed on a control monitor for subjective evaluation (Fig. 3).

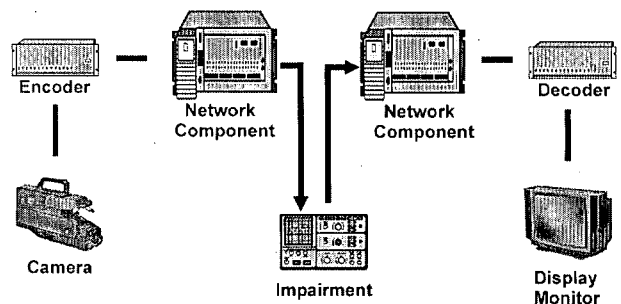


Fig. 3: Test Setup with Impairment Tool

CellStack Classic Codecs. The CellStack Classic codecs for ATM technology were connected to two FORE LE-155 ATM switches as network components. One ATM PVC (Permanent Virtual Circuit) was configured to carry the video stream. The PVC was classified as UBR (Unspecified Bit Rate) traffic, since both switches involved in the test did not carry any other traffic and the full bandwidth of 149.76 Mbit/s was available across the interfaces. In a first reference test no errors were introduced into the network connection and the picture quality was rated as excellent.

Error Rates	Metronome: 60 beats per minute
Reference Test No errors	excellent
10 ⁻⁸	excellent
10 ⁻⁵	Good (a few blocks)
10 ⁻⁴	Fair (many block errors)
10 ⁻³	Bad (picture freezes)

Table 2: Picture Quality of the CellStack Classic Under the Influence of Error Rates

As errors were introduced with the impairment tool the quality of the picture started to deteriorate (Table 2). The

video sequence turned into a frozen picture with an error rate of 10^{-3} .

Litton Codecs. The Litton codecs were connected to two Cisco 7500 routers over Fast Ethernet interfaces. The bandwidth was set to 7.2 Mbit/s. The codecs were first evaluated without the introduction of errors with the impairment tool. As soon as strong motion was added to the video input, the picture froze and the codecs locked up. For this reason the picture quality was only rated as good at best (Table 3).

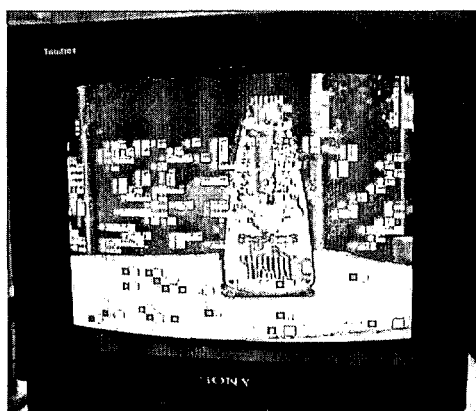


Fig. 4: Severe Blocking in Areas with Strong Movements

Error Rates	Metronome, 60 beats per minute, GOP size = 1
Reference Test No errors	Good
10^{-8}	Fair (single errored blocks)
10^{-7}	Poor (complete lines are errored, severe blocking in areas with strong movements (Fig. 4))
10^{-6}	Bad (picture starts trembling and freezes)
10^{-5}	Bad (picture freezes)

Table 3: Picture Quality of the Litton Codecs Under the Influence of Error Rates

Evaluation of Picture Quality with Background Traffic. A camera focused on a moving metronome was used again as video input. The video signal traveled from the encoder over two network components to the decoder and was then displayed on a control monitor for subjective evaluation (Fig. 3). Background traffic was generated with a HP 4200B Analyzer and caused an overload situation at the outgoing interface of the network component nearest to the encoder (Fig.5).

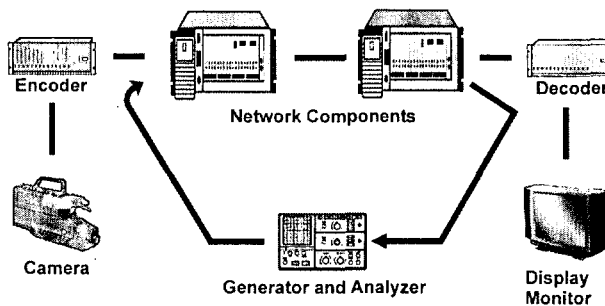


Fig. 5: Test Setup with Background Traffic

CellStack Classic Codecs. Fore LE-155 ATM switches were also used as network components for testing the CellStack Classics with background traffic. The FORE LE-155 STM-1 interfaces are capable of handling a total load of 149.76 Mbit/s. The coder was sending 6800 Protocol Data Units (PDU) per second.

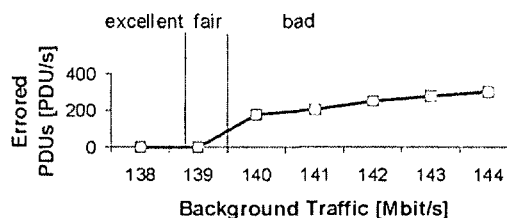


Fig. 6: Influence of Background Traffic on PDUs

Figure 6 shows the influence of background traffic on the PDUs of the video stream. Due to the increasing background traffic the demand for bandwidth of the video signal can no longer be satisfied. As a consequence the picture immediately starts deteriorating to quality ratings of "fair" and "bad".

Litton Codecs. As in the impairment test Cisco 7500 routers were used as network components. The codecs were configured to use 7.2 Mbit/s for data packets. The GOP size was set to 1 (1 frames only). The connection between the routers was implemented as an ATM connection with ATM Interface Processors (AIPs).

In accordance with the findings in the case of the CellStack Classic codecs there was also an immediate decline of the picture quality as soon as the background traffic claimed an amount of bandwidth that exceeded the minimal bandwidth requirements of the video signal.

It is obvious that for both codecs even a few losses already reduce the picture quality to the point where the spectator is no longer satisfied.

QoS Features. As was shown in the previous chapters small impairments of the transmission quality already

have a severe impact on the picture quality and thus on the satisfaction of the user.

The CellStack Classic Codecs compress the video signals in separate streams which can then be transported over ATM using PVCs. Since ATM technology provides the possibility to configure each traffic stream as Continuous Bit Rate (CBR) traffic which reserves the requested bandwidth as a maximal rate, the video signal can be shielded from any network overload situations.

The only mechanism to offer Quality of Service over IP implemented in the Litton codecs is the setting of the Type of Service (TOS) bit via the registry in the Windows NT platform [15]. Unfortunately this feature could not be tested, since the available codecs were issued with an older software version where the TOS bit capability was not enabled.

To receive guaranteed QoS in networks video codecs must be able to reserve resources in the network, for example with RSVP. At the moment the tested Litton codec is not able to compensate for the lack of Quality of Service (QoS) features of the transport protocol in IP traffic.

5 Conclusion

The tests show, that QoS guarantees are required for the transport of video over networks. If parameters such as bandwidth, delay and jitter are not guaranteed or kept within certain bounds, packets are lost and the picture quality deteriorates, even if loss rates are small. To ensure high quality bidirectional video transmissions, the two-way delay, the delay variation and the loss rates should be kept to a minimum. Since the stability of the Litton codecs was limited further tests are recommended.

High quality applications require a networking infrastructure which is able to make Quality of Service guarantees. The Internet with its best effort IP service is not yet able to meet these demands and is therefore still not suitable for the transmission of high quality video data.

6 References

- [1] ATM Forum (1996) Traffic Management Specification 4.0.
- [2] Blake S., Black D., Carlson M., Davies E., Wang Z. & Weiss W. (1998) An Architecture for Differentiated Services (RFC 2475).
- [3] Braden B., Clark S. & Shenker S. (1994) Integrated Services in the Internet Architecture: An Overview (RFC 1633).
- [4] CellStack (1997) CellStack Decoder, Version PCB: 3i, Logic: 2d, Firmware CellStack Video 1.3f over 0.7a (Master), Release: Build 588, June 18, 1997.
- [5] CellStack (1998) Cellstack Encoder, Version PCB: 3i, Logic: 2d, Firmware Cellstack Video 1.4d over 0.9e (Master), Release: Build 745, March 12, 1998.
- [6] Fenimore C. & Libert J. (1998) Perceptual Effects of Noise in Digital Video Compression. 140th SMPTE Technical Conference, Pasadena, California, 28-31 October 1998.
- [7] Fibush D. (1997) Overview of Picture Quality Measurement Methods. Contribution to *IEEE Standards Subcommittee, Committee G-2.1.6. Compression and Processing Subcommittee*, May 6, 1997.
- [8] RRZE (2000) Homepage of the Gigabit Testbed South (application projects) http://gtb.rrze.uni-erlangen.de/projekte_en.html.
- [9] International Organization for Standardization (ISO) (1985) Quality-of-Service - Basic Framework, *Technischer Bericht ISO/IEC JTC1/SC21 N9309*.
- [10] International Organization for Standardization (ISO) (1986) Quality-of-Service - Basic Framework, *Technischer Bericht ISO/IEC JTC1/SC21 N9309*.
- [11] ITU-T Recommendation (1992) E.430 - Quality of Service Framework.
- [12] ITU-T Recommendation (1994) E.800 - Terms and Definitions Related to QoS and Network Performance Including Dependability.
- [13] ITU-T Recommendation (1996) P.800 - Methods for Subjective Determination of Transmission Quality.
- [14] ITU-T Recommendation (1996) G.114 - Transmission Systems and Media. General Characteristics of International Telephone Connections and International Telephone Circuits.
- [15] Litton (2000) CamVision-2 7615 CV2, Rev. E/25 February 2000, Litton Network Access Systems.
- [16] Schertz A., Franzen N., Lu J. & Ravel M. (1997) IRT/Tektronix Investigation of Subjective and Objective Picture Quality for 2-10 Mbit/sec MPEG-2 Video: Phase 1 Results, Contribution to *IEEE Standards Subcommittee, Committee G-2.1.6. Compression and Processing Subcommittee*, October 10, 1997.

- [17] Tektronix (1986) Tektronix 2220 60 MHz Digital Storage Oscilloscope Rev. Nov. 1986.
- [18] Wolf S., Pinson M. H., Webster A. A., Cermak G. W. & Paterson E. (1997) Objective and Subjective Measures of MPEG Video Quality, *139th SMPTE Technical Conference*, New York, 21-24 November 1997.

An environment for processing compound media streams

B. Feustel, A. Kárpáti, T. Rack and T.C. Schmidt
 {feustel,karpati,rack,schmidt}@fhtw-berlin.de
 Computer Centre, Fachhochschule für Technik und Wirtschaft Berlin
 Treskowallee 8, 10318 Berlin, Germany
 www.rz.fhtw-berlin.de/MIR

Keywords: Synchronized Media, Hypermedia Modelling, Multimedia Database, Web-Authoring

Received: January 12, 2001

With today's widespread availability of networked multimedia potentials embedded in an infrastructure of qualitative superior kind the distribution of professionally styled multimedia streams has fallen in the realm of possibility. This paper presents a prototypic environment – both model and runtime system – for processing composite media streams variably composed of multimedia data objects. The system consists of an intelligent media database, a Web-authoring tool, a time directed presentation stream and is based on a hypermedia data model of reusable object components. The plug-in free runtime system is designed as a pure JAVA implementation. Further educational applications of our architecture are presented, as well.

1 Introduction

Today's standards of internet-connected computers provide easy, intuitive access to multimedia information documents within classrooms as well as students homes and thereby confront students as well as teachers with a new paradigm of knowledge transfer: Not only the offer of an unfiltered totality of the present (rapidly changing) knowledge requests for continuous (network) access, but also a formerly unknown multitude of presentation methods to the lecture hall or – in the framework of teleteaching – to students homes has come around. Nothing since the invention of blackboard and chalk, we are tempted to claim, has revolutionized teaching in a more fundamental way than networked multimedia.

Consequently web-based teleteaching and distant learning offers are by now seriously considered parts of the educational system and gain increasing importance. But by preparing educational applications the insufficiency of approaches purely grounded on html-style technologies becomes more and more evident: Design and maintenance of a website approximately reflecting the complexity of an interactive online course is on the one hand an experience of little practicability. On the other hand information streams formed of time-based media or continuously online processed data hardly incorporate into a stateless presentation layer. Consequently a growing awareness for the demand of better information models can be presently

observed within the community of educational computing [1],[3].

Learning modules request for a coherent design of interrelated portions of information being at the same time subject to structural subdivisions regarding thematic aspects such as topic, subtopic, related field etc., didactic classifications concerning complexity, order, relevance to the objective, etc., presentational attributes like positions in space and time, display contexts, ... and finally meta information regarding format, author, access rights etc. The meaningful shaping of such structural overlays belongs to the author. Therefore a desirable information system not only should exhibit capabilities of embedding its contents into flexible structuring but also needs to strongly support the process of authoring in accordance to its abilities, unique points of source editing being the most prominent feature under request.

Of equal relevance appears the support of multimedia data. Different types of media such as text, images, animations interplaying with time-based material i.e. audio, video or online data processing request for an individual specialized treatment which for the non technical oriented author is hard to fulfil. In recent times it has been widely understood that the preparation of qualitative advanced multimedia material ranges far beyond the scope of individual lecturers. Facing the demand for good multimedia supplements in teaching on the one hand and recognising the difficulties in the production of such material a 'marketplace'-type

idea of exchange and reuse appears quite natural in multimedia supported teaching.

Teaching has to account for perception being a time-dependent process. The important notion of time in teaching is one major reason for drawing a lot of attention in recent research works to World Wide Web techniques which distribute multimedia documents with temporal and spatial relations. In addition the growing demand for synchronized handling of time-based media such as video and audio serves as a general motive for introducing temporal aspects to the Web. Finally, streaming data sources invent a new level of scalability by accounting for transport timing and therefore rapidly gain quantitative importance throughout the Internet.

In this paper we present ongoing work on an environment grounded on an object oriented multimedia information model. Residing in a database management system media objects can combine to form complex documents by means of an active document structuring, allowing for temporal and event-type inter-object relations. The environmental basis is employed for the Media Objects in Time time-based runtime environment as well as in further teaching applications. All components together allow for composing complex teaching applications from media objects and streaming them to the Web.

This paper is organized as follows. In section 2 we introduce the basic ideas of our Media Object Model and exemplarily compare to related works. Section 3 presents the underlying multimedia database system and introduces to its authoring toolset. Principles, architecture and implementation properties of the MobIT runtime environment will be discussed in section 4. A brief introduction to further applications is given in section 5. Finally, section 6 is dedicated to conclusions and an outlook on the ongoing work.

2 The Media Object Model and related Works

2.1 Enabling active Document Structures

The teaching and presentation environment introduced here aims at the one hand at profoundly supporting arbitrary media data including time-based material. On the other hand we want to provide a flexible mechanism for structuring documents which not only accounts for thematic interrelations (e.g. links), but also gives rise to object compositions including temporal aspects or complex interaction events. Thus led by our object

model and guided by its web authoring toolset an author should be able to produce for instance a multimedia information stream being the composite of any heterogeneous collection of media data (text blocks, audio, images, video, ...).

Central to our object oriented information model therefore is the notion of active references as a basic composition mechanism. These interrelations not only carry the ability to refer to subordinate presentation data, but are capable of imposing event-type actions on its references. As typical notions of referential actions between documents we consider the connections in time and (presentation-) space or the predefinition of possible user interactions. Since these active references are foreseen in the data object model authors may by a simple editing of attributes build up a document structure which inherits some active processing from its information object class. The individual mark-up of active document structuring is defined with the design of an educational application. Structuring process has been kept very flexible to permit application-oriented, semantically meanings, thereby donating intuition to authors when dealing with the system.

As pointed out above an important role is dedicated to the reusability of document data. As it is of course easy to assure multiple exploitability of simple media files, reusability of complex, composite media documents is of much higher importance. These collections of interrelated documents usually play the role of knowledge modules and bury significant amounts of authoring work. They are also subject to singular change, depending on the knowledge evolution. Our model does support for reusable presentation components by providing a uniform, media independent data structure which we denote by Media Objects. Mobs serve as universal containers for embedding either subsequent Mobs or media data. By referencing one another Mobs allow for arbitrary compositions of unlimited complexity, where the atomic nodes of the resulting graph structure are formed by distinct Data Objects. Besides its uniform appearance the Media Object entities support for application reuse in restricting active links to referenced objects thereby relying on referential integrity ensured by the underlying database system.

2.2 Media Objects

Media Objects may be seen as the central constituents to comprise the data structure of our object model. As the basic design idea a Mob consists of both, the subordinate reference list and the collection of active references, the latter being

restricted to act on Mobs included in the reference list. Simultaneously arbitrary annotations may stick to these hulls, neutral with respect to applications or actual media data.

Defining an application at first requests for turning the Mob structure into a meaningful formation. Semantics can be brought onto Mobs in a twofold fashion: Dynamic content processing may ground on (mandatory) attributes assigned to the data and thereby organize content according to meta information. Media objects in this first step remain singular informational units. The powerful approach however lies in interpreting the active references native to Mobs. An application designer not only can choose from arbitrary interrelations such as trees, Petri-nets, circuits, but can dedicate operative instructions to those data links ranging from a simple automated Web-link generation over spatial and temporal construction policies up to conditional interactions within arbitrary scenes.

2.3 Related Works

Numerous activities rank around document structuring and authoring of more complex information models than HTML-formatting. From the educational area we exemplarily mention the group of Maurer [3],[4], who propose and implemented the Hypermedia Composite Model as a semantic container for learning documents. Even richer research is going on in the area of multimedia database systems. For an excellent overview we refer to [6].

As mentioned earlier several interesting research activities are presently enforcing the notion of time to the Web, the most prominent being the W3C recommendation Synchronized Multimedia Integration Language (SMIL) [10]. As a declarative language SMIL allows for synchronization of media objects in an somewhat simplistic, HTML-style fashion. Synchronization is done in object pairs, either sequential or in parallel. The appearance of any object may be bound to a duration parameter. SMIL extends the meaning of hyperlink to connecting temporal and spatial subregions.

The runtime behaviour of any SMIL interpreter thereby is more or less left open, which probably is the most important drawback of the model. Combined with the absence of a stringent handling of timelines temporal inconsistencies in more complex documents can be foreseen. Besides few reference implementations of SMIL players there is an attempt to include synchronization features into the Web browser named HTML+TIME [11]. This proposal addresses temporal extensions to HTML and incorporates basic elements of SMIL.

Both ideas however suffer from strong limitations due to the simplistic ansatz of HTML omitting any structuring for media object use. Rutledge et al [12] consequently report about severe difficulties in authoring SMIL presentations mainly due to the lack of reusability for object compositions as well as SMIL's inability to deal with complex object relations. In most recent works, the 2.0 specification of SMIL [10], the World Wide Web Consortium heads towards a realization of SMIL as a module within the framework of the XHTML language. Most of this work is presently ongoing and far too incomplete from permitting implementations.

As a completely other example more similar to our work we like to mention the Nested Context Model (NCM) of Soares et. al. [13]. With the aim of grounding a strong structure for flexible deployment of hypermedia documents the NCM provides a composite meta-structuring for different media types, called nodes, up to an arbitrary level of complexity. Those nodes may contain a reference list of denoted nodes giving rise to an arbitrary graph structure of the composed document. The model, which has been implemented in a system called HyperProp, treats hypermedia documents essentially as passive data structures. Synchronizations define through events which may occur as the result of object presentation or user interaction.

Since embedding of media objects within the NCM results in a passive mesh without further presentational meaning, an additional structure of activation, events and contexts (called perspectives), has to be superimposed. This characteristic on one hand leaves some liberty to the author (the same object structure may encounter different behaviour in different contexts), on the other hand it adds an additional level of complexity to the modelled hypermedia system and denotes the major difference to our work.

3 MIR - A Media Information Repository

3.1 The Media Object Database System

The core of our multimedia environment is formed by a media object database system. Named Media Information Repository (MIR) it combines all operations related to data storage and at the same time keeps track of information structuring ensuring referential integrity. Although MIR fully implements the media object model it remains neutral with respect to applications built on top of the database layer. The intention in designing the

media repository was to provide a robust, powerful basis, on which a multitude of educational systems may be established with rather limited effort. MIR divides into two functional groups: The Media Object Lattice and the Data Store (s. fig 1). Objects in both repositories may be addressed by symbolic names embedded in a virtual file system. Besides administrative information concerning owner, group and access rights data entities can carry arbitrary annotations by means of an open property list. Technically only distinct by data type definitions properties may contain any kind of meta information, e. g. content descriptors such as subjects and keywords, didactic annotators concerning presentation order or information depth, and technical markers being specific to the educational applications on top of the database, as well.

Media entities in our object oriented database design belong to classes, which define their properties. Any object instantiates the class its derived from and thereby inherits the property set including type and attribute definitions. Customisation of the MIR to support a new application thus limits to the set-up of appropriate object classes with the possible need for extending authoring functions (see below). Quite independent of actual exploitation the universal data structure MOB is offered for application processing.

3.2 The MIR Authoring Environment

Easy access for authors the system grants through a Web authoring tool. It is designed to guide through the different levels of complexity by means of several adapted views. As it is well known and to some extend obvious that the WYSIWYG paradigm does not hold in the case of temporal, structural or event editing [5], we attempt to relate the multiple aspects of authoring to specific, intuitive appearances of the tool, thereby relying on the semantics of structural relations mentioned above. Application design by means of an object class editor though carries no presentational meaning and remains rather formal as its use might be restricted to technical staff.

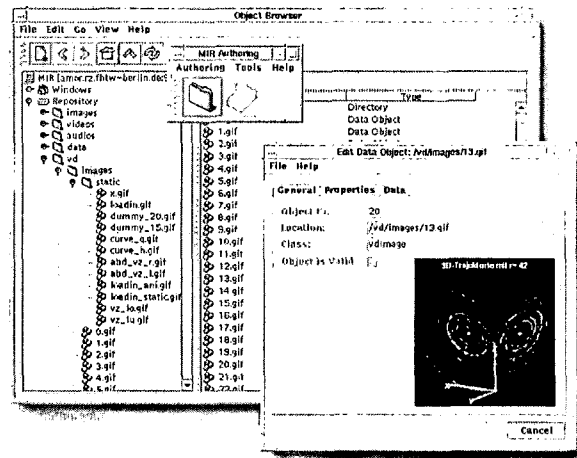


Figure 1: The MIR authoring tool

At the first stage of content authoring our tool allows for DOB upload and control. Guided by an object browser the author may organize and retrieve objects in a directory structure of a virtual file system, donate names, media types and properties to the dobs and upload actual data to the MIR data store (s. fig. 2). In general media data manipulation is not meant to be part of the application authoring process, but for the sake of simplicity a simple text editor which also supports for HTML-formatting is included in the system and permits the direct generation of written text.

Whereas the object browser in the MOB editing regime remains unchanged, dedicated support is given to the author in designing presentations. With the help of a structure view, a spatial view and an (relative) object timeline authoring of MOB-based applications receives its basic instruments. As was pointed out above, however, the specific semantic of media object structures is only fixed with application layout. The authoring requirements thus may significantly change between different fields of

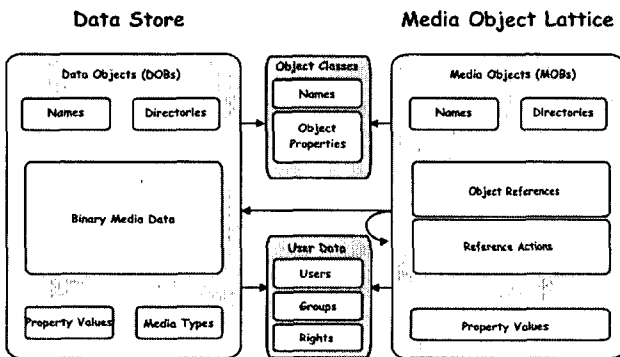


Figure 2: Media Object Database Architecture

As an advantage of the MIR data logic actual media handling separates completely from application design: The Data Objects (DOBs) reside in a Data Store together with its media descriptors, resp. mime-types. They are ready for multiple access by either MOB-based applications or directly through http-requests, where media specific treatment has to be taken care of by clients and - if necessary - by middleware components independent of the actual application.

use and specific aspects cannot be foreseen in general. Including a toolbox of methods our open system therefore provides a programming interface to allow for easy, application dependent extensions in the form of specific views.

3.3 Architecture and Implementation

The technical concept of the MIR environment is formed by an open multimedia architecture designed according to a 3-tiered principle as is shown in fig. 3. Implementation thereby followed the major goals

Functionality: The environment – meant as a uniform platform for media object processing – must provide all fundamental operations on MOBs and DOBs like load, store, search etc.

Flexible Media Handling: The system must adapt to a general range of possible media types including discrete and continuous objects.

Standard Conformance: All used or implemented components should conform to present standards established. User or application interfaces should rely as far as possible on application standards s.a. the Web protocol, mime-type handling, streaming protocol standards etc.

Performance: According to well known resource requirements of (continuous) media data specific measures concerning leanness, optimization and scalability of the system should be applied.

Encapsulation: Access to media object data should only be granted by a set of appropriate, general

operations, thereby hiding low level manipulations such as SQL-statements.

Extensibility: Characteristics of applications as well as additional media types can be expected to impose specific requirements to the environment. Besides a uniformly suitable data environment application and media processing units need to offer universal programming interfaces for adding the requested capabilities to the system.

The current implementation of the media object database runs in a relational database management engine, a Sybase adaptive server, with special tuning applied to it. This platform we chose as a robust, very fast and lean basis. For the sake of encapsulation and performance, but also to achieve an 'object oriented' data modelling all data accesses and manipulations are realised by means of stored procedures. Media specific operations such as compression/decompression, streaming or synchronisation tasks are performed by the middleware components, since middleware services are scalable, support load balancing and in our case accommodate caching.

All middleware components are written in JAVA and are primarily responsible for the session and transaction management and for a buffering cache layer which allows for latency hiding. Even though we employ a single component server solution, the Sybase Jaguar, most of the implementations fulfil the JAVA EJB specification and are therefore rather neutral with respect to the specific product. Client access is granted in a manifold way (s. fig. 3): On the standard side the natural IIOP-exchange of objects is offered to intelligent client apps complemented by standard Web protocol http for all public entities in the database, the latter being implemented by a servlet in the back of an apache/tomcat installation. As serialisation of binary large objects forms an inefficient way of transport we decided to incorporate the Sybase proprietary transaction protocol TDS, which shuffles binary data in bulk. The client programming interface for TDS-transport is hidden behind JDBC, so that proprietary code can be kept from application programming.

As an important feature of the platform introduced here may be seen its ability to deal with pluggable subservers (s. fig. 4). Subservicing not only opens up the field for application dependent media streams, but also allows for incorporation of new, complex functionality such as online data processing without fattening a thin applet client. For an overall stream oriented system it appears quite natural to include served media for streaming and such. MIR provides a flexible and simple interface for this purpose. In current applications subservers are used to incorporate the high performance optimized JAVA Wavelet video player of Cycon et al [7], a direct text sender which permits messaging to ongoing presentations and an MPEG3 server which processes audio.

The interface to include any type of subserver has been purposely designed in minimal fashion: Any subserver in

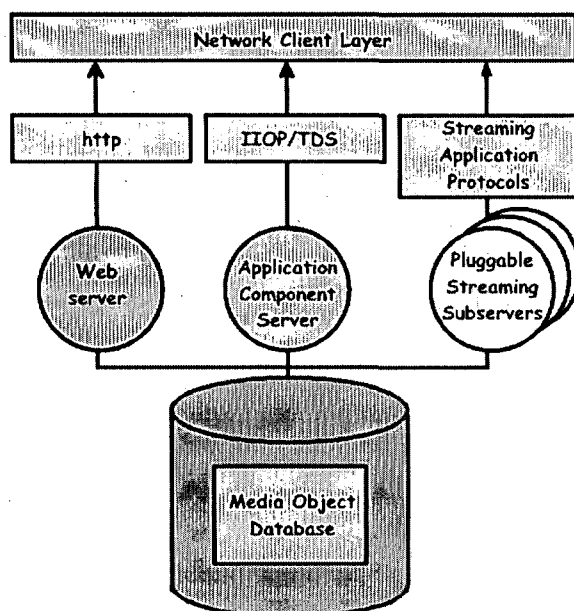


Figure 3: Networked Architecture

perspective must implement the methods `getPortCount` to allow for inquiry on requested number of ports, `setPorts` to permit port assignment, `setData` to receive data handles and the initialization. Additional information classes etc. are kept optional. The interface at the corresponding client site appeals as even simpler: `setServerInfo` and `getServerInfo` are the methods needed here. Within this open framework it should be easy to bring additional data servers to the system as for instance to include real-time visualization or live streams or ...

Implementations on the client side merely depend on application complexity than on guidelines taken from MIR environment: Clients may be applets based in Swing like our authoring tool, simple HTML-pages or Servlets running JAVA Beans in correspondence to JAVA Server Pages. Any time-based application we however have not undertaken without browser's JAVA machine.

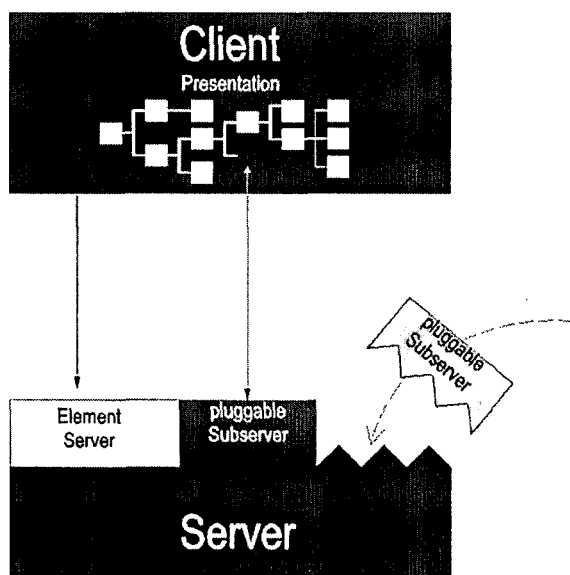


Figure 4: Subserving as a Plug-In

4 Media Objects in Time

4.1 Presenting a media stream

As one major application the teaching and presentation system Media Objects in Time (MobIT) centres about the idea of media objects synchronizable in time which may be linked to form fairly complex presentations. But at the same time any object remains self consistent and of independent use. Roughly speaking our basic concept consists of defining media object instances and lining them up in time as is shown in figure 5. MobIT intends to provide an accurate scheme for temporal and spatial placement of presentation objects, where authors neither have to take care of

interobject synchronization dependencies nor adaptation to possibly inaccurate network performance, the latter being subject to implementation of latency hiding techniques.

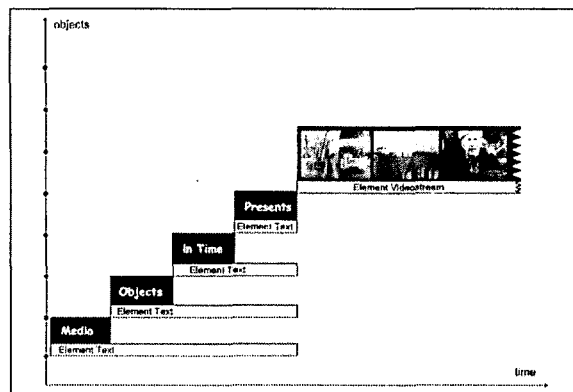


Figure 5: Media Object Instances in Time

Presenting itself on a timeline any presentation becomes a time-based data object, even if composed only from timeless media such as texts or images. Any presentation component will carry an instance of initial appearance and a moment (possibly at infinitum) for fading away from the client's screen. Within this framework of MIR any streaming media such as video or audio may be included and synchronized to the scene and the overall data stream.

Aiming at the combined utilisation in lecture rooms as well as teleteaching our model focuses on a clear, straight forward concept of reusable compound media components. Any of these will be accompanied by screenplay scripts arranging their behaviour in space and time. Thus in place of the page oriented WWW concept or the typically event driven nature of CBT products MobIT runs as a flow oriented presentation model showing for example a crash-test video combined with charts of relevant statistics and vocally explained CAD car models in subsequence.

4.2 The Compound Flow Model

In designing an educational system within our environment structuring has to be given an applicational meaning. In the context of MobIT this is done by the Compound Flow Model (CFM), which takes much care to define a simple structure of straight forward logic intuitively appealing to document authors. The CFM organises the uniform hull entity Mob in a tree structure, where any branch reference expresses a temporal and spatial

inclusion relation. (s. figure 6).

Media objects form the central construction element of the CFM data structure. As bound to the basic design idea of MIR Mobs include the subordinate object reference list and a screenplay script acting on the references, thereby describing all parameters responsible for their behaviour in time and space. Those scripts we denote as Playlists. Playlists describe the states attained by the corresponding Mobs in total.

Tightly bound to the concept of combined reference to objects and their states is the notion of generalized reusability for any component involved. Roughly speaking an object exhibits generalized reusability, iff it is self consistent, i.e. free of recursions, and parametrizable in state space. The fundamental parameters of the state space up until now are the spatial size and the duration in time. Some additional features such as background color or font type change have been implemented.

Vital to the framework of CFM is an environment for generating and controlling the flow. As media objects for a given presentation may be widely branched, each one of them equipped with a complex structural inheritance and its own synchronization demands, a flow control module needs to resolve all structural data dependencies. It thereafter has to linearize resulting bulk information, to form an ordered flow and at last addict objects to the externally provided primary timer.

Even though components of the Model are of

active, self consistent nature an additional flow generator needs to be present. Generating a flow in our context has to fulfil the task of resolving all open object dependencies, collecting the data and en passant performing co-ordinate transforms and at the core linearize data with respect to time. As a result of such linear alignment all playlists are merged to form a complete script for the screenplay the whole presentation consists of. Additionally may be observed that the flow generator as described is – if properly implemented - well suitable for transmitting presentations data collection as a sequential stream over the network. For a more detailed description of the MobIT application see [2].

5 Further Applications

5.1 Virtual Design

The design studio of tomorrow will not contain a computer anymore, but will consist of the computer network. Guided by this maxim a completely different idea of computer based educational system has been developed in collaboration with Bildo Akademie für Kunst und Medien Berlin. Interactive picture networking has been adopted as a basic co-operative internet platform for designers of digital images [8],[9]. The project has been honoured in the meanwhile with the “New Talents Award” at the direct marketing congress DIMA in Düsseldorf 1998 and the special price Multimedia Transfer at Learntec 2000.

People from art and design communicate through their visual products. As it is rather difficult to circumscribe representational and aesthetic contents in standard language terms, a specific way of expression needs to be utilized: A Language of Pictures. Like any stream of statements such a visual speech needs basic order principals, a time-line and thematic assignments at minimal.

The Virtual Design project started from the idea of supplying a networked communication platform which allows for creation of visual dialogs. Starting from a “white canvas” each participant is enabled to contribute data sets consisting of an image, a title and a textual commentary to the system. The system itself requests such contributions to be a reaction of a former entry. It thereby links entities and lines pictures in time chains, optionally branching at nodes which invoked multiple reactions. As time evolves the Virtual Design system will give rise to a tree of pictures with each branch representing a visual dialog between authors (s fig. 7).

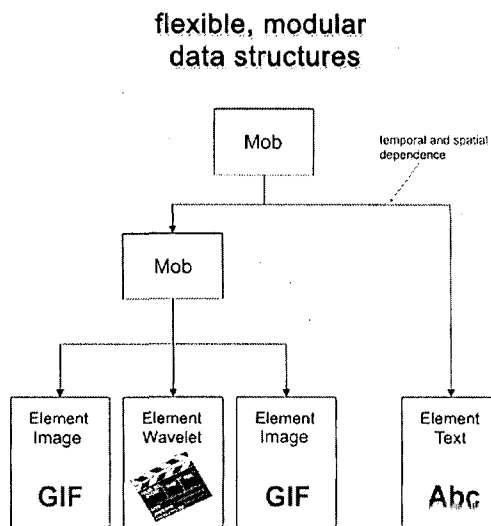


Figure 6: Media Object Hierarchy

Relying on MIR basic environment Virtual Design MOBs enclose images, thumbnails and textual complements. The media object structure in the virtual design application is defined by the virtual dialogs performed by using the system and is assigned automatically as part of the work process. Note that no separate authoring is needed since VD combines workspace and presentation.

5.2 Knowledge Market Place

As a third, much simpler application of the MIR data environment we want to introduce a small knowledge café prototype. The system ranks around pieces of information which are classified according to topics and keywords and with respect to information complexity, as well. Generating content-based meshes from Mob references the information repository not only is able to answer property related searches but will dynamically present document groups as Mob references are automatically transformed into Web links. With the use of XML/XSP-techniques this useful application could be developed in a very limited number of days by relying on the strength of the MIR technology basis.

6 Conclusions and Outlook

The multimedia information technology presented in this paper is an ongoing project in many ways: Having accomplished an efficient basic solution on structured media processing several teaching applications to be used either in the lecture hall or at students homes are to be implemented. Most exciting however we consider future developments in the area of time-based learning and presentation system.

Much work however has to be done in this ongoing project. Interactions have not been defined in CFM yet. As simple smil-type hyperlinks in our flow oriented model could only support hopping between - possibly nested - parallel timelines and as we do not intend to produce some sort of interaction programming language, our current activities concentrate on modelling an interaction paradigm. Accounting for the CFM potential to operate on self consistent media objects we are aiming at a small 'alphabet' of operations which enables authors to open up an unlimited number of navigational paths to the receptor with only a limited number of interactions defined.

Interactions will introduce an additional complexity to the treatment of network behaviour as they might contradict latency hiding techniques in some parts. This is unavoidably true for user dialog elements.

For the loading of binary large elements a careful time control will be needed. The buffered pre-reading of Mobs however may be viewed as filling an instruction cache of a processing unit. Interactions impose branches to the instruction flow

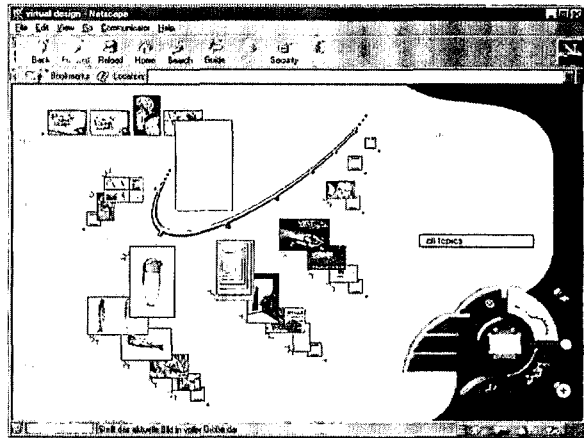


Figure 7: Virtual Design Visual Navigator

and can be buffered in parallel so that immediate system response generalizes.

Acknowledgements

We would like to thank Hans Cycon and his group for enjoyable co-operation. They developed the highly optimized Wavelet Video algorithms and the codec. Mark Palkow carefully ported the codec to JAVA.

References

- [1] Maurer, H.: *Can WWW be succesful*; IFIP - DS8; Jan. 99; 1999;
- [2] Feustel, B., Schmidt, T.C., Marpe, D., Palkow, M., Cycon, H.L.: *Compound Media Streaming in Time*; Vaclav Scala; Proc. 9-th Int. Conf. Comp. Graph., Visual and Comp. Vision WSCG'2001; 2001; Plzen; 2001;
- [3] Helic, D., Maglajlic, S., Schaerbakov, N.: *The HC-Data Model: A New Semantic Hypermedia Data Model*; Proc. WebNet'99; 1999; Charlottesville, VA, USA; AACE; ; 493 - 498;
- [4] Helic, D., Maurer, H., Scherbakov, N.: *Authoring and Maintaining of Educational Applications on the Web*; Proc. ED-MEDIA'99; ; 1999; Seatle, USA; AACE; ; 1792-1797;
- [5] Jourdan, M., Layaida, N., Cécile, R.: *Authoring Techniques for Temporal Scenarios of Multimedia Documents*; Borko Furht: Handbook of Internet and Multimedia Systems and Applications; Boca Raton, Florida; CRC Press LLC; 1999; 179 - 200;

- [6] Aberer, K., Timm, H., Neuhold, E.J.: *Multimedia Database Management Systems*; Borko Furht: Handbook of Internet and Multimedia Systems and Applications; Boca Raton, Florida; CRC Press LLC; 1999; 285-310;
- [7] Marpe, D., Cycon, H.L.: *Very Low Bit-Rate Video Coding Using Wavelet-Based Techniques*; IEEE Trans. in Circ. and Sys. for Video Techn.; 9; 1999; 1; 85-94;
- [8] Karpati, A., Löser, A., Schmidt, T.C.: *Interactive Picture Network*; IEE Electronics & Communications; 99; 1999; 109; London; 1999; 18/1-18/8;
- [9] Born, T., Hannecke, T., Heine, A., Karpati, A., Kemnitz, T., Löser, A., Schmidt, T.C.: *Virtual Design Update*; 27-99; Berlin; FHTW-Transfer; 1999;
- [10] Ayars, J. et. al. (ed.): *Synchronized Multimedia Integration Language (SMIL 2.0)*; World Wide Web Consortium Working Draft; Sept. 2000; <http://www.w3.org/TR/smil20>
- [11] Schmitz, P., Yu, J., Santangeli, P.: *Timed Interactive Multimedia Extensions for HTML (HTML+TIME)*; Note for Discussion submitted to the World Wide Web Consortium; September 1998; <http://www.w3.org/TR/NOTE-HTMLplusTIME>
- [12] Rutledge, L., Hardma, L, v. Ossenbruggen, J.: *The Use of Smil: Multimedia Research currently applied on a Global Scale*; Karmouch, A: Proc. of Multimedia Modelling 1999; Ottawa, Canada; Wolrd Scientific; 1-17;
- [13] Soares, LFG., Casanova, MA., Rodriguez, NLR.: *Nested Composite Nodes and Version Control in an Open Hypermedia System*; Intern. Journ. on Information Systems; 20; 1995; 6; Elsevier; 1995; 501-519;

FEIDHE – integrating PKI in Finnish higher education

Mikael Linden

Tampere University of Technology, P.O. box 553, FIN – 33101 Tampere, Finland
mikael.linden@tut.fi

Janne Kanner and Mika Kivilompolo

CSC – Scientific Computing Ltd., P.O. box 405, FIN – 02101 Espoo, Finland
janne.kanner@csc.fi, mika.kivilompolo@csc.fi

Keywords: Security, PKI, smart cards, authentication

Received: February 11, 2001

The FEIDHE project aims at specifying what it will take to implement a public key infrastructure (PKI) based identification system with smart cards in Finnish higher education. Main drivers of the project are data security, flexible use of electronic resources over the network and the national PKI initiative FINEID (Finnish Electronic Identification). The project is looking for a way to manage identification and authentication when accessing electronic resources and services over networks.

1 Introduction

In Finnish universities and polytechnics students and staff members have access to several information systems and services requiring user authentication. Traditionally authentication is based on username-passwords pairs. In the information systems the passwords or, alternatively, some one-way hash values deriving from them are stored in a user database. When the user is to be authenticated to the system, he is prompted to enter his username and password, which are then compared to the data stored in the database. If the username-password pair corresponds to the one in the user database, access is granted to the resources that are available for the user.

In the username-password pair authentication, the identification of the user is based on the knowledge that the user has. Using encrypted connections provides protection against attackers who listen to the communication channel in order to capture passwords. However, from the security point of view, the username-password authentication has some distinct vulnerabilities, which are mostly due to the ‘human factor’: the user may use a password that is too short or it can be easily guessed, or the same username-password pair is used in several systems at the same time. Some countermeasures can be implemented for example by rejecting passwords that are too weak or by forcing the user to change passwords frequently. Furthermore, there is no administrative tools available for preventing the user from storing the passwords on different media than one’s memory (i.e. the user may write down the passwords on a small piece of paper just next to the workstation).

An alternative approach to user authentication is to use public key cryptography (aka asymmetric cryptography)

in which there are two distinct keys, one for encryption and the other for decryption. The encryption key (the public key) is made available for everyone willing to send an encrypted message to the user. Only the user has the private key that is needed for decrypting the encrypted message.

In order to authenticate a user the computer system sends to the user a challenge, which is basically a sequence of random bits encrypted with the user’s public key. If the user has the corresponding private key, she is able to decrypt the challenge and send the decrypted bit sequence back to prove her identity.

The private key can be stored in a cryptographic token, e.g. a smart card, that will never reveal the key, but instead uses it for decryption when user authentication is needed. Now, when the information system encrypts a challenge with the user’s public key, it can be responded to only if the corresponding private key in the token is available. Furthermore, the cryptographic token is generally protected with a personal identify number (PIN) or other protection functions that prevents unauthorised users to misuse the token. Hence, the public key authentication with the cryptographic token is based on something the user knows (PIN code) and something unique the user possesses (a cryptographic token).

2 PKI and its national implementation in Finland

The generation of a private-public key pair is not enough for the comprehensive use of public key cryptography in authentication. A system called Public

Key Infrastructure (PKI) is needed for managing the keys. This chapter introduces some principles of the PKI and its national implementation in Finland.

From the system point of view, a user in the PKI is the owner of the private key. The PKI is needed to define how to distribute the public key to the authenticating information system and how to pair the public key to a particular user possessing the corresponding private key. In order to implement a large scale PKI to be used for example in the Internet, a trusted third party, called Certificate Authority (CA), is needed. The CA is exclusively responsible to ensure the identity of the private key's holder.

The CA defines its policy in a public document called Certificate Policy (CP). The Certificate Practice Statement (CPS) is a document that states specifications on the CP. Typically, the CPS defines for instance how the identity of the user is verified and who is authorised to do that when the user is applying for a certificate. Internet Engineering Task Force has specified a framework for CP and CPS for the Internet community (Chokhani & Ford 1999).

As a proof of the identity of the private key's holder the CA issues a certificate. It is a statement that by the virtue of the digital signature of the CA binds the identity of the user in a real world to the particular public key in the network. A certificate contains, typically, name of the user, the public key of the user, serial number for the certificate, and the validity period and the intended usage of the certificate. In order to make the certificate available for everyone in the network, it is stored in a public directory server. Internet Engineering Task Force has specified the certificates to be used in the Internet (Housley et. al. 1999).

In PKI authentication, the identity of the user can be verified as follows: the authenticating system fetches user's certificate from the public directory and then gives to the user a challenge that can be correctly responded to only if the user has the private key. However, commonly used PKI-enabled protocols, such as Secure Sockets Layer (Frier et. al. 1996), pass the certificate to the authenticating system in the beginning of the authentication so that the certificate directory is not needed.

The certificate has a finite validity. Also the user can lose or destroy the private key, or the key might be stolen or compromised. Due to these reasons, the CA maintains a public list of invalid certificates. This list is called in the literature as the Certificate Revocation List (CRL). Therefore, in the PKI authentication process also the validity of the certificate must ascertain from the CRL.

The market is gradually getting ready for large scale deployment of PKI. The standards are becoming stable, and commercial PKI products are entering markets.

2.1 PKI fundamentals

2.2 National PKI in Finland

In Finland the public sector has been the driving force in PKI implementation. In 1996 a common working group of three ministries issued a report (Ministry of Finance et. al. 1996) stating that the electronic identification of a citizen belongs to the infrastructure of the information society. According to the report it is necessary to develop and maintain information systems needed. A project called Finnish Electronic Identity (FINEID) was launched to specify a national PKI and to make other necessary preparations. The Population Register Center was nominated as CA. Related modifications in the legislation were made in the parliament.

From the beginning of December 1999 the citizens of Finland have been entitled to apply for an electronic ID card, FINEID card (Police of Finland 2001). The application procedure is similar to the practice for the conventional ID documents (e.g. passport); the citizen files an application at the local police office, where the officer identifies the applicant and checks the validity of the application. Electronic ID cards with the embedded chip are manufactured by Setec, a Finnish smart card vendor, and the certificate is issued by Population Register Center. The price for the electronic ID card is 27 euros (card is valid for three years), whereas the conventional ID card costs 22 euros.

The FINEID card contains two private keys and the related public key certificates. One of the keys is used only for authentication and decrypting messages and the other for digital signatures.

Until September 25th, 2000 Population Register Center had issued 6045 FINEID cards (Population Register Center 2001). The bottleneck for the rapid increase in the number of FINEID cards has been the lack of services relying on the FINEID card and PKI. Only one bank has launched a FINEID based authentication to its web services, and few public sector services, mostly utilising digital signatures, have been introduced. The private service providers seem to hesitate as long as the number of potential customers having the FINEID card is so small. Another barrier delaying the growth of the user number is the cost and availability of equipment i.e. the smart card reader and the software necessary for utilising the reader in applications.

3 FEIDHE - Electronic Identification in Finnish Higher Education

Reliable user authentication has been an issue in Finnish universities and polytechnics as well. The national FINEID project accelerated the related discussion in

universities among the people responsible for network security.

Students in Finland have national plastic student ID card, which is used for getting student prices in public transportation and other student discounts. The student cards are issued by the student unions, which have strong position in Finnish universities. For students embedding a chip to the student card is considered to be a natural step in expanding the current range of the services to the PKI based services. For staff members it would be the responsibility of the university to provide the EID cards needed. Because all the universities in Finland are owned by the state, the FINEID project formed a natural starting point for building up PKI in the universities.

A FEIDHE (Electronic identification in Finnish higher education) project was chartered in May 2000 as a common project of all the Finnish universities, the national union of the Finnish student unions and CSC – Scientific Computing Ltd., which is a non-profit company owned by the ministry of education and the maintainer of the national research and education network in Finland. Polytechnics and their student unions joined the project in October 2000. Altogether there are 50 project members and they cover approximately 250 000 potential users (including students and staff members), which is about 5% of the population in Finland.

The goal of the project is to develop a smart card based authentication system for the needs of the Finnish higher education. The project aims at enabling an introduction of an EID smart card for higher education (FEIDHE card) during year 2002.

To reach the goal the project evaluates solutions available in the markets and makes specifications and recommendations needed. Furthermore, the project estimates the costs and funding needed to build the necessary infrastructure. The project also collects and distributes information to the project members on PKI and how to build services on it.

The main interest of the project is the introduction of stronger user authentication methods in the universities and polytechnics. However, digital signatures are also a subject of interest because of the increase in service level and cost-reduction in the administration that they would imply. There has also been interest in using the FEIDHE card for various payments in the campus.

4 Technical aspects

From the technical point of view the project field can be divided into four layers (Figure 1). The security of each layer is based on the underlying layers.

Basically all the security is based on cryptography, especially on the RSA algorithm and SHA-1 hash function, that are used in FINEID PKI and cards.

Network protocols, such as Secure Sockets Layer (SSL), utilise PKI in user authentication. End user services are

Services	e.g. logging on a workstation, web services
Network protocols	e.g. SSL, Secure shell, IPSec, Kerberos
Public key infrastructure	X.509 certificates, e.g. FINEID
Cryptography	e.g. RSA, SHA-1

Figure 1. Technical layers in FEIDHE.

built on top of the network protocols, such as WWW services, which are run on SSL. This chapter examines the PKI layer, Network protocol layer, Service layer and their integration in the user management of the institute in more detail.

4.1 PKI planned in FEIDHE

The FEIDHE PKI will be based on interoperability with as many services as possible in the public and commercial sectors in Finland. To that end the PKI will lean heavily on compatibility with the existing governmental PKI and Finnish legislation on electronic identification.

The project proposes that one or more commonly trusted commercial certificate authorities should be selected for issuing the certificates for the users. CA signs the certificates with its own private key to bind the person's identity to the public key. CA is also responsible for maintaining a public directory for the certificates and the certificate revocation list.

CA does not necessarily do everything by itself. It is usual that the task of identifying the applicants and distributing the issued cards is given to a separate body that is called Registration Authority (RA). In FEIDHE the duty of RA is planned to be performed by the student administration of higher education institutes. Each applicant will be identified with a valid visual identity card such as a driving licence or a passport according to the CA's certificate policy. Then the applicant will fill an application form and give a photograph.

The project will use standard X.509 certificates (Housley et. al. 1999). No additional information is included with the possible exception of an email address. This is to avoid constantly updating the certificates to reflect changes e.g. in person's role or position. The certificates will be placed both in the smart card and in the public directory that supports Light-weight Directory Access Protocol (Wahl et. al. 1997).

Components needed in the client and server side are illustrated in Figure 2. In the client side a smart card

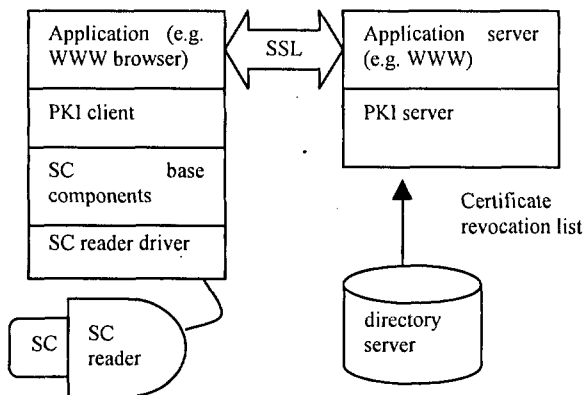


Figure 2. Client and server components in a smart card (SC) enabled PKI.

reader and some extra software is needed to enable the workstation to use smart cards. PC/SC working group, which is a consortium of the major computer operating system and smart card vendors, has published a de facto standard (PC/SC 1997) for client side software architecture. Microsoft has released the base components of the architecture for Windows operating system, and the corresponding software for Linux environment is implemented in a project called MUSCLE, Movement For The Use Of Smart Cards In A Linux Environment (MUSCLE 2001).

The PC/SC architecture as such is generic. The component specific to EID cards is called PKI client, and it is built on top of the PC/SC base components. The PKI client provides a standard interface to client applications, such as web browser and mail client, to access the services in the EID smart card.

To make user authentication PKI-enabled some extra functionality needs to be implemented in the server side as well. The signature and the validity dates of the user's certificate need to be verified. Furthermore, the server needs to check that the certificate is not in the certificate revocation list. Each institute will integrate this functionality in their centralised authentication server or set up a separate PKI server to handle it.

4.2 Network protocols utilising PKI

From the network protocol point of view the operation, in which the user is authenticated, is called client authentication. Fortunately most existing network protocols have support for PKI in client authentication implemented or at least specified.

Maybe the most important network protocol supporting client authentication is Secure Sockets Layer (SSL), securing transactions on World Wide Web. Client authentication is a feature of SSL handshake protocol and the most popular WWW browsers have a built-in support for it. During SSL handshake the server can optionally request the client to send its certificate. To ensure that the client has also control of the corresponding private key, the server uses the public key

of the certificate to encrypt a challenge to the client. The client implementation of SSL uses the underlying PKI client to utilise the attached smart card to get a proper response to the challenge.

Another protocol for secure transport is Secure Shell (Secure Shell 2001), that is commonly used in remote shells in Unix environments. Secure Shell supports client authentication based on certificates, and commercial products using the feature are entering the markets. In FEIDHE project the extensions needed in Secure Shell client and server are investigated.

Kerberos is a well-known, complex protocol used in authentication and session key distribution. Kerberos is traditionally based on symmetric cryptography. It is included in Windows 2000 operating system as the primary authentication method, but Microsoft has replaced the first Kerberos transaction with public key operation called PKINIT (Kerberos 2001). This enables the use of a smart card in Windows 2000 login.

Several commercial Virtual Private Network (VPN) products based on IP Security Protocol (IPSec 2001) have been launched, some of them supporting PKI in client authentication as well. IPSec is commonly used in establishing remote connections to corporate networks over insecure public networks such as the Internet. FEIDHE project evaluates existing VPN solutions to test their smart card support.

4.3 End user services in FEIDHE

End user service is a extensive and complex concept in FEIDHE. Logging on to the Windows network using Kerberos, establishing a remote shell using Secure Shell and setting up a remote connection using IPSec can be considered grass-root level services in the project. More advanced services can be implemented on WWW.

Because there are much more students than staff members, the large volume services are those used by students. Enrolments to periods, courses and exams are examples of administrative services for students. Educational services for students would cover the large variety of e-learning environments that have been discussed a lot in public. The FEIDHE card would replace the use of passwords in these services.

Using smart card authentication instead of passwords may increase convenience for the user, but it does not cause significant cost reduction for the institute. More savings may be obtained by introducing the digital signature in services and processes in the administration. Routine work can be reduced if documents requiring traditional print-and-sign procedure can be converted to the paperless format with digital signature. This concerns especially the administration of the institutes, e.g. travel expense reports, bookkeeping etc. In addition to the cost-savings in administration, the implementation of digital signature can benefit from the improved level of services

experienced by the end-user (e.g. faster and more flexible service, place and time independent use of services etc).

The FEIDHE project will not implement end-user services in a large scale. Only some single services are implemented to test the technology and to get some user experience. Instead the project should be considered as an initiative to build up the infrastructure on top of which services can be implemented. It is the responsibility of the institutes to develop the services when the infrastructure is fully implemented and tested.

4.4 Integrating PKI in user administration

The user database of the higher education institute contains data about the network users in the institute; their name and contacts, role (student or/and staff member and so on), username and password to the information systems on campus and usually many other attributes. The planned FEIDHE implementation separates the user authentication and the user authorisation (access control) as detached services.

When a user authenticates to the system with a FEIDHE card the validation of the certificate will be examined at first. Challenge is used to ensure that the user controls the corresponding private key. After authentication the server consults the user database to find the access rights for the presented certificate, and hence, for the user. For instance only students are authorised to enrol on an examination and only staff members to read the enrolment data. Therefore, the certificate needs to be mapped to a user and her role in the user database. To enable this one or more fields from users certificates need to be introduced for each user in the user database in advance.

We will probably end up with an FEIDHE architecture in which each user has several alternative authentication methods providing different levels of security. The weakest authentication method provided for personal services is the use of username and password. Use of certificate is considered to be significantly more secure, especially if the related private key is stored in a smart card. As the time past the PKI markets will probably grow, there will be several commercial sector's PKI systems that are launched e.g. by banks, mobile operators, trade etc. The use of PKI is however, finally defined by the level of trust that end-users and service providers have to Certificate Authority. In FEIDHE project the FINEID card and certificate are considered as the strongest authentication method, because the registration operations are made by the police.

Prospects of dynamic, secure and flexible use of national and interinstitutional services and resources among higher education institutes is the main driving force in implementing electronic identification for higher education. However, as the FEIDHE solves only the authentication part of the whole procedure, more work has to be done; interinstitutional access control is yet another

matter. Finnish higher education community is currently considering a decentralised directory approach for distributing information relevant to authorisation from one institution to another.

5 Further prospects in FEIDHE

Large scale deployment of PKI and services relying on it is not possible without piloting. To get technical experience and early user experiences on the services to be implemented, a set of pilots is to be launched during 2001 in seven universities and polytechnics.

Pilots vary with respect to the target group and service, number of users and the network protocols used. In most pilots the environment is WWW, taking advantage of the security provided by SSL. Another important pilot environment is Windows, because higher education institutes are gradually upgrading their network environments into Windows 2000 that has built in support for smart cards. The necessary modifications in the user database and administration are also implemented in the biggest pilots.

In the smallest pilot the target group consists of 10 system administrators, which are considered as an important target group, because their secure authentication to the information system is crucial. In the largest pilot 700 pilot cards are distributed to students.

6 Conclusions

Public key infrastructure makes it possible to replace authentication based on passwords with public key authentication. To implement smart card based PKI, smart card readers need to be installed in workstations. Extra software is needed both in the client and server side to make use of the PKI-enabled client authentication built in most network security protocols. Furthermore, modifications to the user administration and procedures for issuing and distributing smart cards to network users have to be implemented.

FEIDHE is a project in Finnish universities and polytechnics whose goal is to develop smart card based authentication system for the needs of the Finnish higher education. The project aims at enabling an introduction of electronic identity smart card during year 2002. In the project a PKI and network protocols relying on it are piloted to establish a secure environment on top of which services requiring high security can be implemented.

7 References

- [1] Ministry of Finance, Ministry of Transport and Communications, Ministry of the Interior. (1996) Electronic identity and Identity card. <http://www.vn.fi/vm/kehittaminen/tietoturvallisuus/vahti/sidrap10.htm> (in Finnish)

- [2] The Police of Finland. (2001) Licence services. <http://www.poliisi.fi/english/pi274en.htm>
- [3] Population Register Center. (2001) The Electronic ID card. <http://www.fineid.fi/default.asp?path=1%2CGeneral%2FNews&template=>
- [4] A. Frier, P. Karlton, P. Kocher. (1996) The SSL 3.0 Protocol. Netscape communications. <http://home.netscape.com/eng/ssl3/ssl-toc.html>
- [5] S. Chokhani, W. Ford. (1999) Internet X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework. Internet Engineering Task Force, RFC 2527. <http://www.ietf.org/rfc/rfc2527.txt>
- [6] R. Housley, W. Ford, W. Polk, D. Solo. (1999) Internet X.509 Public Key Infrastructure Certificate and CLR Profile. Internet Engineering Task Force, RFC 2459. <http://www.ietf.org/rfc/rfc2459.txt>
- [7] M. Wahl, T. Howes, S. Kille. (1997) Lightweight Directory Access Protocol (v3). Internet Engineering Task Force, RFC 2251. <http://www.ietf.org/rfc/rfc2251.txt>
- [8] PC/SC specifications version 1.0. (1997) PC/SC Working Group. <http://www.pcscworkgroup.com/>
- [9] MUSCLE. (2001) Movement For The Use Of Smart Cards In A Linux Environment. <http://www.linuxnet.com/>
- [10] Secure Shell Working Group. (2001) Internet Engineering Task Force. <http://www.ietf.org/html.charters/secsh-charter.html>
- [11] Kerberos Working Group. (2001) Internet Engineering Task Force. <http://www.ietf.org/html.charters/krb-wg-charter.html>
- [12] IP Security Protocol Working Group. (2001) Internet Engineering Task Force. <http://www.ietf.org/html.charters/ipsec-charter.html>

Information systems delivery in a tiered security environment

Anne Strachan, Tony Shaw and Donna Adams
 Network & Information Systems Management, University of Paisley
 High Street, Paisley, PA1 2BE Renfrewshire, Scotland, UK
 anne.strachan@paisley.ac.uk tony.shaw@paisley.ac.uk donna.adams@paisley.ac.uk

Keywords: application, model, network, security

Received: February 7, 2001

The University of Paisley, located in the South-West of Scotland operates across several campuses and has experienced many changes. These changes include the reorganization of academic structures, and a student population becoming more diverse.

Linking all the campuses is a Wide Area Network which has been structured into four distinct layers for security purposes. The network infrastructure has been developed using a framework model to support the delivery of information systems services and intranet developments. These developments incorporate both access and security requirements. Initially, delivery and operation of various services was centralized, but gradually this has been changing with a move to more devolved operations as well as additional services coming on stream.

This paper presents an example application that demonstrates the change in delivery of a service, partly as a response to organizational and resource issues. It also demonstrates the application of an information systems model allied to a technical model to help focus on requirements.

1 Introduction/Environment

The University of Paisley, located in the South-West of Scotland, operates across several campuses at dispersed geographical locations (Figure 1). The University was originally located only on one site at Paisley, but due to mergers etc. has now expanded to another site in Paisley itself, to Ayr, and more recently, to Dumfries as part of a joint project with Glasgow University. It has undergone rapid changes not only to its student population and the structures of the courses it offers, but also to its organisational structures, particularly on the academic side. This rapidly changing environment continues to make demands for appropriate information services.

A prime example of how changes are evolving is demonstrated by the way in which the University manages its portfolio of modules. A module is a basic taught unit that can be studied individually or can constitute an element of a named course which full-time students normally pursue. The management of this information has changed from being a central management function to one in which responsibility is devolved to the individual academic divisions, a corollary of which is the acknowledgement of different individual rights in maintaining the information content.

These changes have occurred concurrently with a rapid expansion in part-time students as the University has focused on being a major provider of education for wider

access. In fact, Paisley is the leading institution in Scotland which is committed to wider access with recent funding allocation reflecting this commitment.

2 Network and Information System Resources

2.1 Network and Information Systems Management

Network and Information Systems Management (NISM) is a technical support department of the University of Paisley and is responsible for both Information Technology and Information Systems strategy. This includes overseeing both network operations and network planning and undertaking information systems design, development and support. It also plays a key role in IT services development which includes network development and information services development whilst ensuring appropriate security. The department comprises approximately 33 staff consisting of 1 director, 9 application developers, 8 systems and network staff, 10 frontline support staff, 3 administration staff and 2 procurement staff. There is a total of 1100 staff in the University.

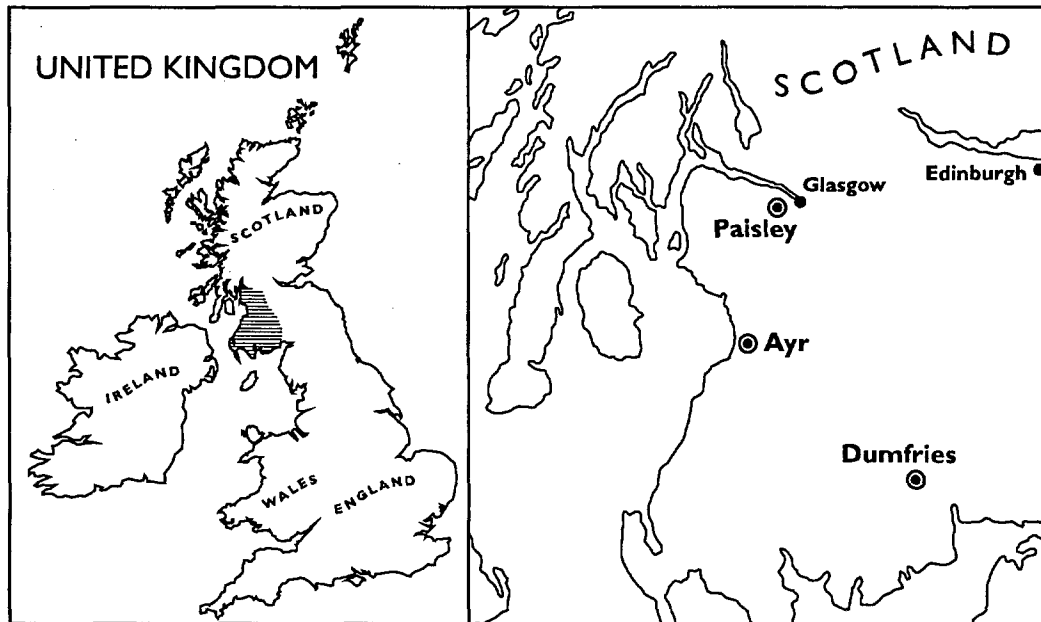


Figure 1: Location of Paisley campus

At an operational level, NISM staff are responsible for managing most of the network services including Netware application servers, a number of Unix servers used by University Administration, Unix servers that form part of the Campus-wide information service and non-departmental (i.e. not owned by specific academic departments) World Wide Web servers. This responsibility involves the routine administration of a range of Netware and Unix servers as well as the management of the University staff e-mail system. In general, NISM has no responsibility for operational aspects of student services, particularly with respect to teaching and learning, but does provide technical advice to staff working in this area.

The University's main business systems are also the responsibility of the department. These include the Admissions and Enrolment systems, all the Examination systems and other support systems that enable changes and additional information to be recorded about students. Changes to current student information is maintained separately as transactional data, and is periodically distributed as a service to satellite systems to enable them to update their local records. These satellite systems are the responsibility of the departments that developed them.

2.2 Network Overview

The University's central network and systems consists of a category 5 cabled network within buildings. Four pairs of fibres run from each building to a central computer room via building switches or hubs to the central switch. The switches provide ATM connection and the hubs 10Mbytes connection to each virtual LAN. There are approximately 3000 connected data points.

The core central network is managed by a combination of in-house management and selective outsourcing of various aspects of the service.

2.3 Systems Overview

There are approximately 39 central systems. These consist of a mix of NT, Unix and Netware server running core application and print services to staff, central network services, such as firewall, domain name service and mail systems, and student record information systems. All internal systems are managed in-house with external consultants being involved where necessary as a mechanism to enable change and provide skills transfer.

3 Paper Outline

As the network has developed with an increasing demand and need for access to appropriate information and/or use of information systems, the approach to delivering information systems has used a framework model in order to focus on requirements. In considering this approach, the paper examines:

- The development of an Information Systems model and the subsequent technical model for the network infrastructure.
- The benefits that have resulted from these framework models.
- The evolution of an example application in conjunction with the development of the framework models.

4 Information Services Model

The design of the network was influenced by several factors. Although some factors are not relevant to the current discussion, an important one was the realisation there were different categories of user with widely differing requirements, not least of which were differing security constraints. Consequently, an Information Systems model was developed initially describing the proposed services from a user perspective. This

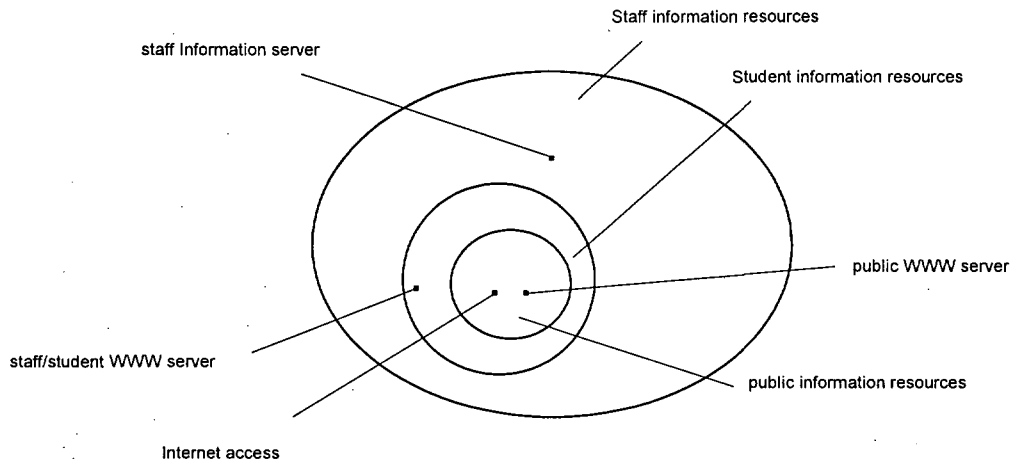


Figure 2: Information access for different user groups

influenced the subsequent development of the technical model.

The relationship between the sets of services is shown in Figure 2. Clearly staff potentially have access to a wider range of services than either students or the public. Note the initial classification of staff does not distinguish between academic and other staff; although there will be differences in terms of service requirements at a local level.

This model is intended to encompass virtually all foreseeable requirements but does not include some occasional specialist facilities. It is assumed that the establishment of any specialist facilities will be a very closely controlled process including authorisation of the connection of items to the network and appropriate security controls. This represents a significant change from previous practice where there has been very limited control over the connection of users (both internal and external to the University) to the network.

5 Tiered Network Structure

NISM were initially responsible for the original administration network, although other separate networks existed at the Paisley site. However, in 1996, NISM took over responsibility for the entire network, although at that time the campus network comprised a disjoint collection of single networks and a single Wide Area Network (WAN) linking Paisley and Ayr. At the start, much effort was put into designing an appropriate

technical model for the network infrastructure and services, whilst taking into consideration emerging requirements especially with respect to Teaching and Learning and remote access. In addition, issues relating to security and associated network management had also to be considered (DTI 1993). The result was a hierarchy of separate networks being established in 1997.

Figure 3 illustrates the principle of the technical model, the main characteristics of which are:

- It is hierarchical with the ability for users on one network to access directly devices on any other network lower down the hierarchy.
- All networks are monitored and controlled to an appropriate level.
- Server resources can be located as appropriate to the user category that they service.
- There is a public network to meet the requirement of increased availability to Internet resources without unreasonable overheads on user administration activities.

At the outset the desire was to move towards a more structured information services environment where basic good practice in the areas of security, standards and service provision could exist. This resulted in different networks being identified for administration, academic and student users, and also the public. The public network would be the home for resources such as the public web server. The resultant tiered security

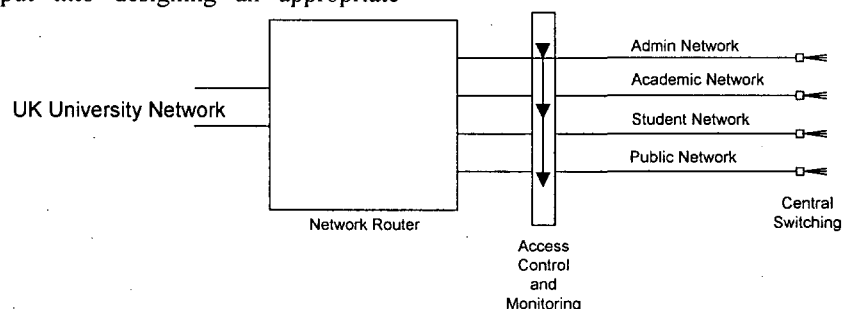


Figure 3: Network infrastructure technical model

environment has subsequently directed the deployment of information services at both LAN and WAN sites. Figure 4 illustrates how various services relate to different user groups.

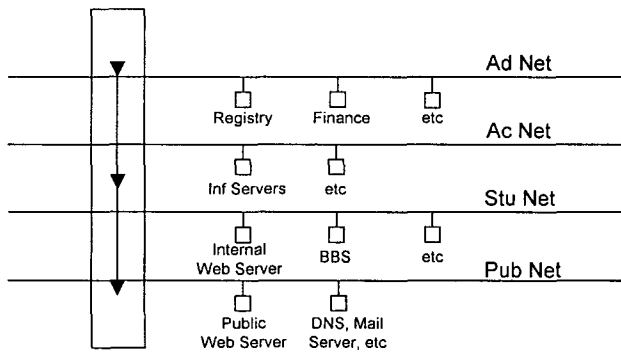


Figure 4: Service distribution in relation to the model

Development of a managed WAN, which had to support the four-tier network model including traffic prioritization, took place in late 1997, in conjunction with the establishment of a commercial firewall for protection of the network. Concurrent with these technical developments was the development of a University security policy. This ensured that access issues were carefully considered and that the different requirements relating to the four networks (and any subsets of these) could be incorporated into the policy. In planning the WAN, attention focused on what type of service was required at the various locations and which ones justified being connected to the WAN. So, for example, the needs of others including student residences would be met through remote access.

5.1 Tiered Network Benefits

There have been various benefits resulting from the network model including the following:

- The way in which it has been relatively straightforward to configure the network firewall in line with the different networks/user classifications.
- The extent to which the model has served to help structure discussions on planning and expenditure. At a time when financial resources are limited the model has made it easier to balance the technical planning and expenditure requirements associated with the different network/user classification.
- The framework has also enabled the needs of the many to take precedence over the wants of the few. That is, the model has helped ensure that emphasis has been placed on identifying the key requirements of each of the user categories rather than be dominated and distracted by, for example, individual departmental perspectives.

6 Example Application

The selected application example facilitates the maintenance of the definitive description of each module called a Module Descriptor. At the start, the original 'system' was paper-based with documents being sent from academic departments to Registry and held there as the definitive source. All updates were handled in the same way. In 1995 this was followed by a freestanding computer-based system based in Registry whose purpose was to act as a primary information collection mechanism. This system comprised a Module Administration application and a Data Entry application which allowed data from amended paper-based Module Descriptors to be entered. All new Module Descriptors and amendments to existing ones were handled centrally by Registry staff in this way. 1995 marked the beginning of various computer-based applications to tackle this issue.

Subsequently, the overall system provided for version control such that two versions of a module were allowed to exist at one time. This was handled by using two databases, one termed the Planning database, the other the Live database. The current version of a Module was held in the Live database and the new version was held in the Planning database. Information in the Planning database was transferred to the Live database at a predetermined time dictated by Registry. However, due to staff resources and volume of work due to the proliferation of modules, the task became unmanageable. This situation was exacerbated as awareness of the deficiencies of the module descriptor database content became more apparent when the application allowing Module details to be viewed became available in 1996. This awareness increased with the availability of updated browser facilities via the student intranet in September 1997. It should be noted that the process for managing the modules was handled by manual procedures under the control of Registry.

The initial applications were operated centrally under the control of one department. Despite the original requirement which had been for a central operation, no secondary mechanism having been specified, this was really the only effective way of implementing the system at the time. Security was achieved simply by locating the application within the department and relevant staff had access to the entire collection of modules. However, with the development of the network, a facility to search for and view the module information was relatively easy to provide, the main consideration being the location of the database with respect to the tiered network. The original database and application remained accessible for editing purposes by the one department. A copy of the database was located on an appropriate server accessible by both staff and students to allow read-only access. No other security provisions were needed as the view facility was read only and the database was not the original on which editing took place.

As a result of various pressures, the situation was reassessed, which also coincided with some requirements changes. In reviewing the purpose of the application, the

User Class	Database Access	Functions Required
Modules Database Administrator	Planning	Administration Data entry/edit Information retrieval/viewing
	Active	Administration Data entry/edit Information retrieval/viewing
Registry Staff	Planning	Data entry/edit Information retrieval/viewing
	Active	Information retrieval/viewing
Learning & Teaching Custodians	Planning	Data entry/edit Information retrieval/viewing
	Active	Information retrieval/viewing
Other Academic Staff	Planning	Information retrieval/viewing
	Active	Information retrieval/viewing
Students	Active	Information retrieval/viewing

Figure 5: Resulting user classification

principle established when developing the Information Systems model was applied. That is, users were classified to focus on the applications required and help prioritize the work. In this way five groups of users were identified, each requiring to access the data in some way. Three user types needed to be given explicit permission and access to carry out necessary functions. Figure 5 indicates in general terms which database each group would want access to and what they would want to do. The specific details were further elaborated during the development process.

It was acknowledged the task of maintaining the module information was too great for Registry alone and the decision was taken to allow authorized users to be able to change that information for which they are responsible. This raised issues of delivery and security on the data editing application which hitherto had not needed to be addressed. A web-based application delivered over the intranet was deemed the best way forward, with the modules database suitably located on the staff network, which is accessible by administrative users. All staff have a network login identifier. Novell Netware uses the Lightweight Directory Access Protocol (LDAP) mechanism for authenticating users and this mechanism was employed by the application to verify users. Once in the application, a secondary mechanism was needed to ensure that only those users who had the authority to edit specific modules could do so.

Currently, further changes are in progress, partly as a result of recommendations from the Quality Assurance Agency (QAA 2001) and partly as a result of streamlining procedures in response to the academic restructuring that has taken place. The net effect is that the existing database will have to be restructured and the applications rewritten. However, by developing a model of user classification it will be much easier to get the user groups to focus on their needs and to be able to prioritize the applications.

7 Conclusions from the Evolution of the Application

7.1 Factors Promoting Change

Over time, there have been various factors which have emerged to promote change. These are summarized as follows:

- A centrally managed operation was unable to cope with the amount of work that resulted.
- The academic community wanted access to reliable current information.
- There was a need for better and more reliable information for students.
- Within the University there was an apparent need for change in working practice.
- The network developments provided a means to devolve operations securely.
- There was a need for a more efficient and effective information management process.

7.2 Issues

- Conceptually, the data concerned is relatively simple. However, the organizational process is complex and generally not well understood. It is further complicated by organizational restructuring provoking a need to rethink the process.
- User classification was a difficult concept to promote, and still can be. Many academics regard it as a right to have access to any and all information the University may have.
- One overarching application versus multiple applications. Allied to user classification, this is another difficult concept for some users who find it hard to understand that multiple applications, relating

to different types of users, can run against the same underlying database. This led to a misconception that classification was a mechanism to be used to prevent users from having any access, rather than to provide the required access.

- Deployment of a service whilst maintaining security and appropriate access does require more careful thought, a necessary requirement when addressing sensitive data.
- Who the person is to drive a project forward and be responsible for the project management process, including resourcing.

7.3 Benefits

- Removal of the workload and responsibility for maintaining Module Descriptors from the Registry staff, although they still have full access.
- Devolving responsibility from a central administrative department has empowered academic staff to take ownership of their Module Descriptors with the attendant responsibility to maintain their currency.
- The information management process is more efficient and effective.
- From an application development perspective, there may be opportunities for the re-use of application components across the various user groups.
- Staff and students have more confidence in the University's module information.

8 Concluding Remarks

Security issues and attendant data protection matters are of paramount concern to IT staff, but not to most users. One of the difficulties in pursuing any project is in getting the users to appreciate there are security issues here that affect projects.

Other, less complex applications have been delivered by the intranet, and the framework model of user classification has been a useful one to employ. For example, a classification of Course Administrator was identified and a suite of applications developed to support this role. This was carried out in consultation with a small group of course administrators. Access to the application was restricted to this small group of users only. The LDAP mechanism of authentication is used in all cases where applications deal with personal data. Authorization to use a particular application is controlled by an access table. Although roles are not specifically defined as such, the notion of a role giving the right to use a particular application is implemented in this way (Henderson 2001). There may be further controls within an application, such as that within the module editor which restricts users to their own set of modules.

9 References

- [1] DTI (Department of Trade and Industry) (1993) *A Code of Practice for Information Security*

Management. British Standards Institution (DISC) PD0003. Superseded by BS 7799.

- [2] QAA (Quality Assurance Agency) (2001) *The Framework for Qualifications of Higher Education Institutions in Scotland, January 2001*. <http://www.qaa.ac.uk/crntwork/nqf/scotfw2001/contents.htm>
- [3] Henderson J. (2001) *Using Role Based Access Control to Administer Information Security Policy*. <http://iii.gla.ac.uk/scotmid/gendocs/rbac-smp.html>

Register allocation: A program-algebraic approach

R. Daniel Resler
 Dept. of Mathematical Sciences
 Virginia Commonwealth University
 Richmond, VA 23284–2014, USA
dresler@vcu.edu

James M. Boyle
 Technology Development Division
 Argonne National Laboratory
 Argonne, IL 60439, USA
boyle@anl.gov

Keywords: register allocation, code generation, program transformation, compilers, programming languages

Received: February 16, 2001

The problem of allocating a finite number of hardware registers to evaluate arbitrarily complex arithmetic expressions arises in the implementation of programming language compilers. Traditionally, register allocation has been implemented by using graph-theoretic algorithms. In contrast, we discuss an approach based on direct algebraic manipulation of the expressions for which registers are to be allocated. These manipulations employ simple identities and canonical forms from “program algebra”. The algebraic approach admits a straightforward implementation of the required identities as rewrite-rule program transformations. The use of canonical forms for the intermediate expressions makes it possible to apply the transformations automatically.

The approach we describe offers a number of advantages. The algebraic approach is easy to understand, because expressions are manipulated directly instead of being converted to graphs. Moreover, the algebraic approach can implement efficient register-allocation strategies without sacrificing this understandability, through the use of suitably chosen intermediate canonical forms. Finally, the correctness of the algebraic approach is easy to prove, because the program transformations that perform the manipulations are based on identities from program algebra.

1 Introduction

An important activity performed by compilers for programs written in typical programming languages such as C or Fortran is *register allocation*: the assignment of the high-speed registers provided by the central processor architecture to hold the operands and intermediate results of computations. For most computer architectures, access to values held in registers is faster than to values held in either random access memory (RAM) or cache. Thus, the goal of efficient register allocation is to avoid, insofar as possible, the storing and re-fetching of intermediate results and the re-fetching of operands when evaluating expressions.

In this paper we describe a *program-algebraic* approach to register allocation over expressions. In the algebraic approach, the expressions in the program are manipulated directly (without conversion to an intermediate form, typically a directed acyclic graph (DAG) (see [1], chapter 9)) according to rules of a program algebra (in this case, simple rules from the λ -calculus) to introduce the required temporaries, to eliminate common subexpressions, and then to

(approximately) minimize the number of loads and stores required.

The algebraic approach manipulates the expressions in a form closely related to their source-language form. This feature confers several advantages:

- It avoids converting the expressions to an alien intermediate form (i.e. a DAG).
- It makes the process underlying register allocation easy to understand, because each intermediate form of the expression can be understood as a source language expression.
- It scales up nicely to register allocation over basic blocks by combining multiple expressions and carefully controlling the list of available registers as you move between blocks.
- It enables register allocation to be implemented using source-to-source rewrite-rule program transformations.

- Most importantly, it opens the door to a relatively straightforward proof of correctness of the register allocation process, by using the program-algebraic identities on which the program transformations are based to prove that the transformations preserve the correctness of the program at each step.

We have implemented the program-algebraic approach to register allocation described here as a sequence of sets of program transformations that are applied by the TAMPR program transformation system (see, for example, [7, 2, 3, 4] for discussions of TAMPR and its applications). TAMPR program transformations are pure rewrite rules, expressed in terms of the syntax of the source language being transformed.

As discussed in [4], the problem of applying sets of program transformations automatically can be solved by using intermediate *canonical forms*. A distinctive feature of the TAMPR system is that it applies each set of transformations *to exhaustion*; that is, it applies each set of rules to a program until the program no longer contains any instance to which a rule in the set could apply. Thus, if the application of a set of transformations to a program terminates, then the program is in a *canonical form* defined by the set of transformations. This feature naturally leads to structuring a complex task, such as compilation or register allocation, as transformation through a *sequence* of canonical forms. In the case of register allocation, the canonical forms correspond exactly to the three steps described earlier: all temporaries created, all common subexpressions eliminated, and all temporaries allocated to hardware registers.

2 Notations and Assumptions

In this paper, we assume that the target for compilation of a program is a RISC-architecture CPU. As discussed earlier, such CPUs typically supply just two instructions that refer to memory, *load* and *store*. All other arithmetic and logical operations are performed register-to-register within the CPU; in most RISC architectures, these instructions are *three address* instructions; that is, they permit the independent specification of two operand registers and a destination register.

To simplify the exposition of the algebraic approach, we restrict the programs being compiled to contain only assignment statements whose expressions consist of any number of binary arithmetic or logical operations performed on simple constants or identifiers. Thus, for simplicity, we consider neither unary operations nor subscripted variables as operands. (One of the strengths of the algebraic and transformational approach to this problem is that transformations can be developed for this restricted problem, and then they can be easily extended incrementally to handle the omitted features.)

Because we intend to employ source-to-source program transformations to manipulate expressions at the source-

language level, we use the typical assignment statement notation of programming languages to express the hardware load, store, and arithmetic and logical operations. Thus, for the compiler we are describing, the output of the register allocation phase is a program in which (1) each complex expression has been broken into a number of assignment statements, each of which has at most a single arithmetic or logical operation on the right-hand side, and in which (2) some special notation, such as R_j , denotes those operands that are registers. Figure 1 compares generic assembly language instructions (first column) with the source language notation we use to express them (second column). Here,

load m, R_i	$R_i = m;$
store R_i, m	$m = R_i;$
$op R_j, R_k, R_i$	$R_i = R_j \ op \ R_k;$

Figure 1: Notation

registers are denoted by R_n , and all other identifiers refer to values residing in memory; *op* denotes any arithmetic or logical operation provided by the hardware.

We further assume that this generic RISC architecture CPU permits reuse of an operand register as the result register for an operation. In such an architecture, a minimum of two registers is required to compile code for an arbitrary arithmetic expression, assuming the ability to spill values to memory allowing for register reuse. (Of course, there are certain situations, for example, when $i = j = k$ in Figure 1, where only one register would be required.)

An expression and representative of those for which register allocation needs to be performed (and having operand reuse) is Expression 1; we use it for examples throughout this paper.

$$(a - (b + c)) * b - (b + c) \quad (1)$$

Under the assumption that only two registers, R0 and R1 are available, the code shown in Figure 2 would be generated from expression (1). Note that, as one would expect,

```

R0 = c ;
R1 = b ;
R0 = R1 + R0 ;
R1 = a ;
R1 = R1 - R0 ;
spill = R0 ;
R0 = b ;
R0 = R1 * R0 ;
R1 = spill ;
R1 = R0 - R1 ;
x = R1 ;

```

Figure 2: Code for expression (1)

the need to retain the value of the common subexpression $b + c$ causes its value to be spilled from register R0 to memory.

3 Algebraic Approach to Register Allocation

In this section we describe the algebraic approach to register allocation, emphasizing the mapping from an unbounded number of registers to a finite set of registers. Our goal is not to propose a new heuristic that provides more nearly optimal register allocation than what is attainable using the conventional graph-theoretic approach, but rather to discuss an approach to register allocation that provides adequate performance, is easily understandable, and whose correctness is readily amenable to proof.

One of the properties of the transformational approach is that often the same objectives can be met in different ways, by defining different intermediate canonical forms and rearranging the order in which transformations are applied to produce these forms. The approach we describe here progresses through a series of logical and self-contained transformational steps that does indeed produce high-quality code; we do not claim, however, that applying this set of transformations in the order about to be described is the only, or even the optimal, way to perform register allocation.

The algebraic approach to register allocation progresses through seven major steps:

1. **Transformation to lambda expression form**—each operand and each result of an operation in an arithmetic expression is transformed into an identity lambda expression, in which the λ -variable name is drawn from an unbounded set of temporary variable names.
2. **Common subexpression elimination**—all common subexpressions are eliminated from the expression resulting from the preceding step, thereby eliminating both operand reloads and recomputation of identical subexpressions.
3. **Register pressure reorganization**—the evaluation order of the lambda expressions in the expression is changed to allow for early calculation of subexpressions having a high “register pressure” (those requiring a large number of registers for their evaluation).
4. **Scope reduction**—bindings of temporary variables are moved closer to their first uses, if possible. (A temporary variable in a lambda expression is eventually allocated a register; therefore scope reduction reduces the number of assembly instructions over which that register is in use.)
5. **Marking free temporaries**—the last use of each temporary is marked to allow for register reuse.
6. **Allocation of registers to temporaries**—temporary variables in lambda expressions are allocated to registers; lambda expressions to implement register spilling and reloading are introduced if necessary.

7. **Transformation to assembly language**—the lambda expressions are transformed to assembly-language-like assignment statements and then to three-address assembly instructions.

The following subsections discuss the details of each of these steps.

3.1 Transformation to Lambda Expression Form

The algebraic approach to register allocation begins by making each program variable, constant, or result of an operation the value of a temporary variable drawn from an unbounded set. In the program algebra, these temporary variables are represented by lambda expressions. Thus, each program variable, constant, and arithmetic operation must be made the argument of a lambda expression.

This property is easily and correctly accomplished by representing each variable and arithmetic expression as an identity lambda expression of the form $\lambda_{a_1.a_1}(arg)$. For example, the expression ‘ $a + b$ ’ is represented by the lambda expression

$$\lambda t_3.t_3(\lambda_{a_1.a_1}(a) + \lambda_{b_2.b_2}(b)) \quad (2)$$

Thus, for this example, this step introduces the λ -variable temporaries t_3 , a_1 , and b_2 . Clearly, this step has not changed the meaning of the original expression.

Note that we use the ASCII notation

```
lambda  $\lambda$ -variable @
   $\lambda$ -body
end ( $\lambda$ -argument)
```

for lambda expressions in the remainder of this paper. Using this notation, expression (2) is

```
lambda t3 @
  t3
end (
  lambda a1 @
    a1
  end ( a ) +
  lambda b2 @
    b2
  end ( b )
)
```

Figure 3 shows expression (1) using this notation.

Each lambda expression will eventually be represented in the assembly code by either a register load, or, where required, a store instruction to spill a register to memory. To achieve this representation, the register allocation process continues through a series of transformations that produce a more-nearly-optimal ordering of arithmetic operations and memory loads.

The final code is then produced by assigning registers to each λ -variable prior to generating load or spill instructions. The transformations generate such code beginning

```

lambda t00011 @
  t00011
end (
  lambda t00007 @
    t00007
  end (
    lambda t00005 @
      t00005
    end (
      lambda a00001 @
        a00001
      end ( a ) -
      lambda t00004 @
        t00004
    end (
      lambda b00002 @
        b00002
      end ( b ) +
      lambda c00003 @
        c00003
      end ( c )
    ) *
    lambda b00006 @
      b00006
    end ( b )
  ) -
  lambda t00010 @
    t00010
  end (
    lambda b00008 @
      b00008
    end ( b ) +
    lambda c00009 @
      c00009
    end ( c )
  )
)
...
lambda t00007 @
  t00007
end (
  lambda b00002 @
    lambda c00003 @
      lambda t00004 @
        lambda t00005 @
          t00005
        end ( a00001 - t00004 )
      end ( b00002 + c00003 )
    end ( c )
  end ( b ) *
  lambda b00006 @
    b00006
  end ( b )
)
...

```

Figure 3: Identity lambda representation of expression (1)

with the outermost lambda expression and moving in towards the center of a nest of lambda expressions. Therefore, any reorganization of the lambda expressions results in a new, and perhaps more nearly optimal, ordering of register use in the assembly code.

3.2 Common Subexpression Elimination

Once each program variable and arithmetic expression has been made the argument of a lambda expression, the next step is to eliminate common subexpressions. Common subexpression elimination is easy to understand and verify in the algebraic approach (see [5] for a thorough discussion). It simply involves expanding the scope of each lambda expression until that scope encompasses all possible instances of the subexpression the lambda expression represents. Any instances of the subexpression within the scope of that lambda expression can then be replaced by instances of the λ -variable.

For example, consider the following expression (which is an intermediate form that appears when transforming the lambda expression in Figure 3 to a λ -nest in which all common subexpressions have been eliminated):

Note that program variable *b* is the value of two distinct λ -variables (*b00002* and *b00006*) in the preceding expression. To eliminate one of these common subexpressions, first the scope of lambda expression *b00002* is increased to encompass the outermost λ -*t00007* expression. The redundant λ -*b00006* expression can then be deleted, replacing it, and all occurrences of *b00006* in the expression, by *b00002*. This step produces the expression

```

...
lambda b00002 @
/* Scope of lambda b00002 expanded. */
lambda t00007 @
  t00007
end (
  lambda c00003 @
    lambda t00004 @
      lambda t00005 @
        t00005
      end ( a00001 - t00004 )
    end ( b00002 + c00003 )
  end ( c ) * b00002
  /* b00006 replaced by b00002. */
)
end ( b )
/* Scope of lambda b00002 expanded. */
...

```

The result of common subexpression elimination is a λ -nest having the property that no lambda expression has any lambda expressions in its argument; Figure 4 shows the example expression of Figure 3 after common subexpressions have been removed.

3.3 Register Pressure Reorganization

The third major step in the algebraic approach is to reorganize the λ -nest taking into consideration the *register pressures* of each subexpression. The register pressure of

```

lambda a00001 @
  lambda b00002 @
    lambda c00003 @
      lambda t00004 @
        lambda t00005 @
          lambda t00007 @
            lambda t00011 @
              t00011
            end ( t00007 - t00004 )
          end ( t00005 * b00002 )
        end ( a00001 - t00004 )
      end ( b00002 + c00003 )
    end ( c )
  end ( b )
end ( a )

```

Figure 4: Expression (1) after common subexpression elimination

a (sub)expression is defined to be the maximum number of registers required to evaluate that (sub)expression.

The λ -nest can be reorganized by calculating subexpressions having higher register pressures early. The value of this reorganization is easy to see: Suppose that the subexpression of an expression having the higher register pressure requires five registers for its calculation. If this subexpression can be calculated first, perhaps a total of only five registers will be required to calculate the entire expression (including the register required to hold the result of this subexpression); whereas if the five-register subexpression is calculated after one or more registers are in use, more than five registers will be required for the entire expression. Thus, performing this reorganization often results in a more nearly optimal use of registers over an entire expression by freeing up registers used in the evaluation of more complex subexpressions. Figure 5 illustrates the potential advantage of this reorganization for the evaluation of the expression $a + ((b * c)/d - (e * f))$; evaluating the subexpression $(b * c)/d - (e * f)$ prior to loading a register with a requires one fewer register.

Note that, while register pressure reorganization is important for many expression, Expression 1 is not complex enough for the order of its operations to be altered by the register pressure reorganization transformations.

3.4 Scope Reduction

Optimal use of registers requires delaying register loads until just before the value loaded is needed. The preceding transformations, however, especially the ones that eliminate common subexpressions, do not guarantee this; in fact, they can have quite the opposite effect. Eliminating common subexpressions requires making the scopes of lambda expressions as large as possible—loading values early—in the hope of finding commonable expressions. For example, consider the following λ -nest, representative of a form that

R0 = a ;	R0 = b ;
R1 = b ;	R1 = c ;
R2 = c ;	R1 = R0 * R1 ;
R2 = R1 * R2 ;	R0 = d ;
R1 = d ;	R0 = R1 / R0 ;
R1 = R2 / R1 ;	R1 = e ;
R2 = e ;	R2 = f ;
R3 = f ;	R2 = R1 * R2 ;
R3 = R2 * R3 ;	R2 = R0 - R2 ;
R3 = R1 - R3 ;	R0 = a ;
R3 = R0 + R3 ;	R2 = R0 + R2 ;

(a)	(b)
no reorganization requires 4 registers	after reorganization requires 3 registers

Figure 5: Register pressure reorganization of $a + ((b * c)/d - (e * f))$

occurs frequently after common subexpressions have been eliminated:

```

lambda x00001 @
  lambda y00002 @
    lambda z00003 @
      lambda t00001 @
        ...
      end ( y00002 + z00003 )
    end ( z )
  end ( y )
end ( x )

```

Note that when allocating registers moving from the outermost lambda expression (lambda x00001) into the λ -nest, variable x would be loaded into a register prematurely (it is not needed until after the expression $y00002 + z00003$ is evaluated). This would result in an unnecessary spill should there be only two registers available for calculation of the λ -nest.

This problem can be solved by reducing the scope of every λ -variable as far as possible prior to allocating registers. Generally this involves “pushing” a λ - v_1 expression into the λ -nest over all expressions that do not use the λ -variable v_1 .

Care must be taken, however, not to ‘undo’ any of the previous register pressure reorganization of the λ -nest. Preserving the evaluation order of subexpression operands while forcing lambda expressions inward suffices to preserve the register pressure reorganization. For example, suppose that the evaluation order of expression $subexpr_1 + subexpr_2$ (assuming left-to-right evaluation of subexpressions) has been changed to $subexpr_2 + subexpr_1$ as a result of determining that $subexpr_2$ has a larger register pressure. The scope reduction transformations are free to move the evaluation of operand $subexpr_2$ as close as possible to performing the addition as long as $subexpr_2$ is always completely evaluated prior to evaluating $subexpr_1$.

Applying these transformations to the commoned λ -nest of Figure 4 results in the more nearly optimal arrangement shown in Figure 6. There are two types of lambda expres-

```

lambda c00003 @
  lambda b00002 @
    lambda t00004 @
      lambda a00001 @
        lambda t00005 @
          lambda t00007 @
            lambda t00011 @
              t00011
            end ( t00007 - t00004 )
          end ( t00005 * b00002 )
        end ( a00001 - t00004 )
      end ( a )
    end ( b00002 + c00003 )
  end ( b )
end ( c )

```

Figure 6: Expression (1) after λ -variable scope reduction

sion in the canonical form in Figure 6: *simple-variable-load* lambda expressions and *binary-operation* lambda expressions. The scope-reduction transformations guarantee that this canonical form possesses an important property (an invariant) required by the register allocation transformations discussed later: at most two simple-variable-load lambda expressions separate any two binary-operation lambda expressions in a λ -nest after scope reduction.

3.5 Marking Free Temporaries

Scope reduction is used to avoid loading a value into a variable (register) too early. It is also important to avoid retaining the value of a variable (register) beyond its last use, because a variable containing a value no longer needed in subsequent operations could be reused to store a new value. To reuse registers, it is first necessary to identify the last uses of each variable in the λ -nest.

Fortunately, it is easy to identify last uses in the algebraic approach (although we have been unable to find a notation for expressing this information comparable in elegance to the scopes of lambda expressions). Because the argument of a lambda expression represents an operation that must always be performed prior to any operations in the body of the expression, the *last* use of a variable in *time* is its most deeply nested occurrence in an argument in a canonical λ -nest. This use corresponds to the *first* use of the variable *lexicographically* in the λ -nest.

For example, consider the uses of variable t00004 in Figure 6. The last use of t00004 in this expression is its most deeply nested occurrence, found in the argument of the λ -t00011 expression. This use is actually the first occurrence of t00004 in the argument of a lambda expression when performing a left-to-right textual scan of the entire λ -nest.

```

lambda c00003 @
  lambda b00002 @
    lambda t00004 @
      lambda a00001 @
        lambda t00005 @
          lambda t00007 @
            lambda t00011 @
              t00011
            end ( t00007 : $free$
                  - t00004 : $free$ )
          end ( t00005 : $free$
                * b00002 : $free$ )
        end ( a00001 : $free$ - t00004 )
      end ( a )
    end ( b00002 + c00003 : $free$ )
  end ( b )
end ( c )

```

Figure 7: Expression (1) after marking last uses

The pattern-matching capabilities of the TAMPR program transformation system greatly simplify locating and marking the last uses of variables within each expression. Figure 7 shows the example expression with the last uses of each λ -variable marked with the type qualifier *free*. This qualifier indicates that the register eventually allocated for that variable is free to be reused in subsequent operations; this mark is automatically transferred to the allocated register when it is substituted for the variable.

3.6 Allocation of Registers to Temporaries

The final step in the algebraic approach to register allocation is to map the unbounded set of λ -variables used in a λ -nest to a finite set of available hardware registers. Simply put, this mapping involves simulating the unbounded number of λ -variables by spilling a hardware register to memory when all registers are in use. The value spilled to memory must, of course, then be reloaded into a register prior to its next use. Provided the set of available hardware registers contains at least as many registers as the maximum number of operands in any hardware arithmetic operation (two for the RISC instruction sets discussed here), this simulation of an unbounded number of registers can always be performed.

Transformationally, the mapping is accomplished by allocating registers in a “wave” from the outside of the λ -nest inward. The wave leaves in its wake λ -variables allocated to registers. If the allocation wave encounters a λ -variable and no free registers are available, a register that is still in use is spilled to free it for allocation. Spilling is accomplished by making the register to be freed an argument to a lambda expression that binds its value to a new unique memory temporary (i.e., by binding its value to a λ -variable that represents the register’s contents in memory). Figure 8 shows expression (1) after register allocation with only two

registers available; in this case, spilling the value in register *r0* and reloading it later is required.

```

lambda r0 @
  lambda r1 @
    lambda r0 @
      lambda r1 @
        lambda r1 @
          lambda spill00013 @
            lambda r0 @
              lambda r0 @
                lambda r1 @
                  lambda r1 @
                    r1
                    end ( r0 - r1 )
                end ( spill00013 )
            end ( r1 * r0 )
          end ( b )
        end ( r0 )
      end ( r1 - r0 )
    end ( a )
  end ( r1 + r0 )
end ( b )
end ( c )

```

Figure 8: Expression (1) after register allocation

In the following subsection, we describe a somewhat idealized strategy for allocating a finite number of registers to an unbounded set of λ -variables. Then, in Subsection 3.6.2, we introduce some of the tactics and details that we use to implement the approach in practice. Finally, Subsection 3.6.3 discusses an example that shows how allocation proceeds.

3.6.1 Allocation concept

Register allocation begins after all common subexpressions have been eliminated, the scope of each λ -variable has been reduced, and the last uses of λ -variables have been marked. To perform register allocation, two pieces of information are required:

- the complete set of registers available for use in evaluating the λ -nest, and
- some measure of the relative costs of allocating each register in that set.

This information can be made available to the transformations by associating with the λ -nest a list representing the set of available registers and their costs-of-use. We call this list the “available register list” (ARL). The transformations can then select registers for allocation or spilling from this ARL.

One way to encode the ARL in the program text is to embed the λ -nest in a two-argument function of the form

allocate (λ -nest , register-list)

This function is semantically the identity function on its first argument. (Such notation is necessary in TAMPR because transformations must transform code segments into syntactically and semantically equivalent constructs.) The register list in the second argument is the ARL.

Broadly speaking, a register on the ARL has one of three reuse costs:

1. a zero cost, because the register is free;
2. an infinite cost, because the register contains a value that will be used in the next binary operation; or
3. a finite non-zero cost, because the register holds the value of a commoned subexpression and this value will be reused deeper in the nest.

When a register holds an operand for the next binary operation, we assign it an infinite reuse cost. If the cost of reuse of such registers were not infinite, then the register holding the value for one of the operands of the next binary operation could be selected to be spilled and allocated to hold the value of the other operand. Then, the spilled value would need to be reloaded prior to performing the binary operation, requiring the allocation of yet another register, possibly leading to an allocation loop.

The semantics of lambda expressions (and of the corresponding RISC hardware operations) allow for the reuse of either operand register in a binary operation, should that register be available, to store the result of the operation. (Reuse is possible because the values in the operand registers have been used before the result of the operation is stored.) Therefore, while the reuse cost of the registers holding the operands of a binary operation is infinite just prior to encountering the lambda expression for the operation, the cost of reusing these registers drops to a finite value before allocating a register for the result of that binary operation.

The particular finite cost value of a register on the ARL is based on a heuristic; ideally, this heuristic should give the lowest cost to the register that causes the fewest reloads if its value were to be spilled. Obviously, then, the heuristic must give free registers a cost of zero.

Provided that (1) no operation in a λ -nest requires more than two operand values, (2) the set of available registers contains at least two registers, and (3) the invariant guaranteed by the scope-reduction transformations (that at most two simple-variable-load lambda expressions separate any two binary-operation lambda expressions in a λ -nest, see Section 3.4) holds, there will always be one or more registers having a finite reuse cost on the ARL and, hence, available for allocation to load the values needed for a binary operation. Thus, regardless of the number of λ -variables used in a λ -nest, the allocation of a finite number of registers to these variables can be completed.

3.6.2 Allocation details

The register allocation concept described in the preceding section can be implemented in a number of ways. We have

chosen to maintain the ARL ordered in terms of increasing reuse cost, re-sorting the list each time a register is entered with an updated cost. We also have chosen to “represent” registers having an infinite reuse cost by deleting them temporarily from the ARL. This representation is possible because all such registers are known to be used in the next binary operation, so they can be re-entered on the list (giving them a finite cost) just prior to allocating a register for that operation. Hence, the names of the infinite cost registers are not lost when they are deleted from the ARL. Not entering them in the ARL with an infinite cost saves time by avoiding pointless re-sorting of the list.

Thus, in the transformations discussed here, all registers actually present on the ARL have finite reuse costs. Free registers have a zero cost. For the others, we use a heuristic to determine reuse cost: a combination of (1) how recently a λ -variable has been referenced and (2) the number of times a λ -variable is referenced in the body of its defining lambda expression. (The first of these measures is similar to the least-recently-used strategy often used for replacing pages in a virtual-memory management system.)

Because the transformations assume that the ARL is sorted in order of cost of reuse (cheapest to most expensive), the ARL requires careful maintenance when registers are added or their costs updated. Thus, certain maintenance tasks must be performed *every* time a register is inserted into the ordered ARL. First, any other instances of that register must be removed from the list, because the cost associated with the inserted register is the (possibly) new cost of using the register at the current point in the allocation process. The ARL must then be sorted by reuse cost in order to guarantee that the register at the head of the list is always the cheapest one to allocate.

Before specifying the behavior of the register allocation transformations in more detail, we consider what happens when a register must be spilled. As discussed earlier, when there are no zero-cost (free) registers on the ARL, a live register must be freed by spilling its value to memory (if the value is not already in memory) and setting up a later reload just prior to the next use of the register’s value.

To spill a value to memory, the unallocated portion of the current λ -nest is “wrapped” in a new lambda expression that associates the spilled register variable with a memory temporary, and then all occurrences of the spilled register in the body of the current expression are transformed to the name of the new memory temporary. Another lambda expression is then added just prior to the first use of the new memory temporary to cause re-allocation of a register for the value in memory.

As a representative example, suppose that the transformations have completed part of the allocation of a λ -nest, so that the partially transformed expression looks like:

```
...
$allocate$ ( ...
    lambda t00011 @
    ...
    end( r0 * ... )
```

```
... ,
$registers$ ( ... )
)
...

```

Assume that at this point in allocating registers for a λ -nest, the transformations need to spill register $r0$ to memory (i.e., $r0$ holds the result of an operation rather than a value already stored in memory). First the lambda expression representing the actual spill (to a memory temporary `spillTemp`) is created, taking care to transform all occurrences of $r0$ in the λ -nest to `spillTemp` and adding the spilled register $r0$ to the ARL with zero reuse cost (indicated by *free*):

```
...
lambda spillTemp @
    $allocate$ (
        ...
        lambda t00011 @
        ...
        end ( spillTemp * ... )
        ... ,
        $registers$ ( r0 : $free$ , ... )
    )
end ( r0 )
...

```

A second lambda expression is also generated to reload the spilled value (defining the λ -variable `spillLoadTemp` in this example) and pushed in as closely as possible to the first use of `spillTemp`. All occurrences of `spillTemp` in the λ -nest are then transformed to the new λ -variable. Assuming that the first use of `spillTemp` is in the λ -`t00011` expression, this step produces

```
...
lambda spillTemp @
    $allocate$ (
        ...
        lambda spillLoadTemp @
            lambda t00011 @
            ...
            end ( spillLoadTemp * ... )
        end ( spillTemp )
        ... ,
        $registers$ ( r0 : $free$ )
    )
end ( r0 )
...

```

Spilling the value of a register being freed is skipped if the value already resides in a named memory location; that is, if the value in the register is that of a program variable or constant or if it is a value that has previously been spilled. In this case, it is only necessary to perform the last two steps just shown: transform all occurrences of the name of the reused register in the body of the λ -nest to the name

of the value in memory, followed by inserting a lambda expression just prior to the first use of the substituted memory location name to cause re-allocating a register for the value in memory and renaming the memory variable to the λ -variable of this inserted expression.

With these tactics, the behavior of the register allocation transformations has a simple description. Recall that the transformations preceding register allocation place the program in a canonical form in which the register allocation transformations can encounter only three types of lambda expression, which differ only in the nature of their argument. We discuss register allocation schemes and ARL maintenance requirements for each of these types of lambda expression in turn.

```

lambda spill00013 @      lambda b00002 @
  ...
end ( r0 )                end ( b )

(a)                       (b)

lambda t00004 @
  ...
end ( r0 + r1 )

(c)

```

Figure 9: Types of lambda expression in canonical λ -nest

Case 1 (Figure 9(a)) involves a lambda expression that spills a register to a memory temporary. Because such a lambda expression specifies writing a register's contents to memory, no register need be allocated for its evaluation. However, such an expression frees a previously allocated register for reuse; freeing is accomplished by inserting the register in the ordered ARL with a reuse cost of 0.

Case 2 (Figure 9(b)) involves a lambda expression that has a variable or constant as its argument. Here a register must be allocated, and because the register currently at the head of the ARL is (by construction) the cheapest one to use at this point, it is selected for allocation. If the register selected for allocation is still alive (i.e., if it is referenced in the body of the current lambda expression, indicated by its having a non-zero reuse cost), then it first must be spilled to memory, as just discussed. Whether the selected register was previously allocated or not, all occurrences of the λ -variable being allocated must be replaced by the allocated register throughout the body of the current lambda expression.

Case 3 (Figure 9(c)) involves a lambda expression whose argument is a binary arithmetic operation. In this case also, a register must be allocated for the λ -variable. To guarantee that there are always registers available on the ARL, the already-allocated operand registers referenced in the binary arithmetic operation of this lambda expression are added to the ARL with finite reuse cost *just prior* to allocation of the register for the result of the binary operation. (The ARL

must, of course, be re-sorted to maintain least-cost order.) Again, the transformations select the register at the head of the ARL as in Case 2 and perform the appropriate spill and re-fetch, if required.

3.6.3 Allocation example

In this section, we illustrate the operation of the register allocation transformations on a simple example expression. Suppose that only two registers, r_0 and r_1 , are initially available, and that these registers are to be allocated for the λ -nest in Figure 7. Then to initiate the allocation process this λ -nest is transformed into the form:

```

$allocate$ (
  lambda c00003 @
    lambda b00002 @
      lambda t00004 @
        ...
      end ( b00002 + c00003 : $free$ )
    end ( b )
  end ( c ) ,
  $registers$ ( r0 : $free$ , r1 : $free$ )
)

```

Allocating a register for the outermost λ -variable in this example follows Case 2 and involves transforming the λ -variable $c00003$ to r_0 (the first register on the ARL) wherever $c00003$ occurs in the lambda $c00003$ body. After allocating a register for this lambda expression, the transformations move the $\$allocate\$$ function in to surround the body of that expression and remove the register just allocated from the ARL (because its reuse cost is now infinite), producing

```

lambda r0 @
  $allocate$ (
    lambda b00002 @
      lambda t00004 @
        ...
      end ( b00002 + r0 : $free$ )
    end ( b ) ,
    $registers$ ( r1 : $free$ )
  )
end ( c )

```

After the next step, in which a register is allocated for λ -variable $b00002$ according to Case 2, the expression becomes

```

lambda r0 @
  lambda r1 @
    $allocate$ (
      lambda t00004 @
        ...
      end ( r1 + r0 : $free$ ) ,
      $registers$ ( )
    )
  end ( b )
end ( c )

```

At the next step, registers `r0` and `r1` from the binary operation are added to the ARL just prior to allocating a register for `t00004` (Case 3); `r0` is allocated because it is free. The resulting expression (with the innermost ellipsis filled in to permit carrying the example further) is

```
lambda r0 @
  lambda r1 @
    lambda r0 @
      $allocate$ (
        lambda a00001 @
          lambda t00005 @
            lambda t00007 @
              lambda t00011 @
                t00011
              end ( t00007 : $free$
                  - r0 : $free$ )
            end ( t00005 : $free$
                * r1 : $free$ )
          end ( a00001 : $free$ - r0 )
        end ( a ) ,
      $registers$ ( r1 )
    )
  end ( r1 + r0 )
end ( b )
end ( c )
```

Note that no free register is available to allocate for `a00001`, because both `r0` and `r1` hold values that are used in more deeply nested computations. At this point, spilling would occur (as in the example discussed in the preceding section) if `r1` held a value not already in memory. However, its value, `b`, already resides in memory and therefore need not be spilled; `r1` is simply marked free, after introducing a lambda expression to reload its value from memory just prior to its next use and substituting the name of the λ -variable in that expression for uses of `r1`:

```
lambda r0 @
  lambda r1 @
    lambda r0 @
      $allocate$ (
        lambda a00001 @
          lambda t00005 @
            lambda load00013 @
              lambda t00007 @
                lambda t00011 @
                  t00011
                end ( t00007 : $free$
                    - r0 : $free$ )
              end ( t00005 : $free$
                  * load00013 : $free$ )
            end ( b )
          end ( a00001 : $free$ - r0 )
        end ( a ) ,
      $registers$ ( r1 : $free$ )
    )
  end ( r1 + r0 )
end ( b )
```

```
end ( c )
```

Register allocation proceeds in this manner through the remainder of the λ -nest. When a register must be allocated for `load00013`, both `r0` and `r1` contain values not held in memory; therefore, one of these values must be spilled.

Register allocation stops when the first argument of the `$allocate$` function no longer contains a lambda expression. At this point, registers have been assigned to all temporaries in the λ -nest.

3.7 Transformation to Assembly Language

Once register allocation completes, the λ -nest is ready for transformation to assembly code. Transforming lambda expressions to three-address assembly code involves assigning the argument of each lambda expression to its λ -variable (a register); for example

```
lambda r0 @
  ...
end ( r1 * r0 )
```

becomes

```
R0 := R1 * R0
```

Such transformations are trivial and do not warrant further discussion.

4 Transformations and Trusted Compilation

One of the major advantages of the algebraic approach to register allocation is the possibility of proving that the transformations that implement it *preserve the correctness* of the programs they transform. Such transformations could be used to construct a *trusted compiler*—a compiler that has been formally verified to correctly compile any correct program [6].

To formally verify a compiler constructed using conventional techniques requires proving that the program implementing the compiler is correct. Proofs of such programs tend to be massive and monolithic, even when the correctness of individual subroutines is proved independently [8]. The task is complex and labor-intensive even for simple computer programs; it is almost impossible to carry out for complex programs such as compilers that perform optimization. Even verifying only the register allocation phase of a conventional compiler would require proving the correctness of the graph-theoretic subroutine that performs register allocation, a daunting task.

In contrast, a compiler based on the algebraic approach and implemented by program transformations can be proved correct by proving that each individual transformation rule preserves the correctness of *any* program to which it applies. Because the transformations are relatively simple, so are their proofs; thereby the need to construct

massive, monolithic proofs is avoided. (Of course, one must also know that the TAMPR transformer, the program that applies the transformations, is correct. The TAMPR transformer can be used to bootstrap its own implementation from simpler versions, thereby helping to simplify its own proof of correctness.)

We have begun work on a methodology for formally proving that TAMPR transformations, such as those for register allocation, preserve the correctness of the programs they transform. Space does not permit a discussion of this methodology here (see, however, [9]).

5 Conclusions

We have discussed a program-algebraic approach to a compiler's allocation of registers for arbitrarily complex arithmetic expressions. We have demonstrated that the program-algebraic approach offers two major advantages over the traditional graph-theoretic methods:

- The algebraic approach is easy to understand, because expressions are manipulated directly instead of being converted to graphs. The program is always a program throughout the manipulation, and it is always correct. The required manipulations are implemented by a sequence of small, easily understood rewrite-rule transformations that automatically carry a program through a series of canonical forms.
- The correctness of the algebraic approach is easy to prove, because the program transformations that perform the manipulations are based on identities from program algebra. Thus, the algebraic provides an approach to constructing a trusted compiler.

References

- [1] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers, Principles, Techniques, and Tools*. Addison-Wesley, 1986.
- [2] James M. Boyle. Abstract programming and program transformation—an approach to reusing programs. In Ted. J. Biggerstaff and Alan J. Perlis, editors, *Software Reusability*, volume I, chapter 15, pages 361–413. ACM Press (Addison-Wesley), 1989.
- [3] James M. Boyle and Terence J. Harmer. A practical functional program for the CRAY X-MP. *Journal of Functional Programming*, 2(1):81–126, January 1992.
- [4] James M. Boyle, Terence J. Harmer, and Victor L. Winter. The TAMPR program transformation system: Simplifying the development of numerical software. In Erland Arge, Are Magnus Bruaset, and Hans Petter Langtangen, editors, *Modern Software Tools for Scientific Computing*, chapter 17, pages 353–372. Birkhäuser Boston, Inc., 1997.
- [5] James M. Boyle and R. Daniel Resler. A program-algebraic approach to eliminating common subexpressions. *Informatica*, 24(3):397–408, June 2000.
- [6] James M. Boyle, R. Daniel Resler, and Victor L. Winter. Do you trust your compiler? *IEEE Computer*, 32(5):65–73, May 1999.
- [7] J.M. Boyle and M.N. Muralidharan. Program reusability through program transformation. *IEEE Transactions on Software Engineering*, SE-10(5):574–588, September 1984.
- [8] F. A. De Millo, R. J. Lipton, and Perlis A. J. Social processes and proof of theorems and programs. *Communications of the ACM*, 22(5):271–280, May 1977.
- [9] Victor L. Winter and James M. Boyle. Proving refinement transformations for deriving high-assurance software. In *Proceedings, High-Assurance Systems Engineering Workshop, Niagara on the Lake, Ontario, Canada, Oct. 21-22, 1996*, pages 68–77. IEEE Computer Society Press, Los Alamitos, CA, 1997.

Neural fields: An approach to infinite-dimensional systems for information processing

Alexander D. Linkevich
 Polotsk State University
 P.O. Box 105, 211440 Novopolotsk-3, Belarus
 linkevich@psu.unibel.by; adlinkevich@tut.by

Keywords: information encoding, information processing, neural field, neural network, mental phenomena, meaning of information, cognitive system, learning rule, neural-quantum similarity, supersymmetry, evolution equation

Received: July 20, 1999

We consider main properties of a neural field whose state is determined by a vector of some topological space (such as Banach or Hilbert ones). Dynamics of the field (time evolution of its state) is associated with information processing and mental phenomena. We formulate the learning problem for neural fields and offer several solutions that are generalizations of learning algorithms proposed for neural networks before. The state vector of a neural field is represented in a form like in quantum theories, which gives a clue to a kind of supersymmetry, viz symmetry between bosonic and fermionic modes of excitation of the neural field. An evolution equation for a field function is suggested. A minimal architecture of a cognitive system is proposed which comprises perceptual, lexical and semantic subsystems. It is suggested to treat the meaning of a piece of information as a code associated with a settled field function produced by a semantic neural network.

1 Introduction

It is a burning issue how nonlinear dynamical effects could be used for information encoding (IE) and information processing (IP). The advance in this direction may appear fruitful to deepen and broaden the understanding of general principles of IE and IP, to gain new insights into dynamical mechanisms of brain activity and mental phenomena, to develop new methods, algorithms and hardware implementation for artificial systems.

The prevailing type of systems whose dynamics is currently exploited for IP tasks is neural networks (NNs) (for an introduction see, e.g., [Amit, 1989; Frolov, Murav'ev, 1987, 1988; Hertz et al, 1991; Peretto, 1989; Vedenov, 1988]). In the broad sense, a NN may be defined as any complex dynamical system composed of interacting units called neurons (or, more strictly, neuron-like elements). Nonlinearity may occur both in dynamics of a single (isolated) neuron and in interneuronal (synaptic) connections. Undoubtedly, it would be desirable to deal with realistic models in which formal neurons resemble real biological nerve cells and their interactions are also close to real synaptic couplings as much as possible. This way leads us to the notion of NNs in the narrow sense as models of the brain which emerges as a nonlinear dynamical system exhibiting complex, highly ordered spatiotemporal behavior at various time and length scales. However, the tremendous number of neurons and essential heterogeneity and nonhomogeneity of the brain hampers description of processes underlying mental phenomena. This suggests to develop an approach in which information is set into a neural field

(NF) which may, in general, be defined as a system whose state is determined by a vector of some linear topological (infinite-dimensional) space (such as Banach or Hilbert ones). Dynamics of the NF (time evolution of its state) is associated with information processing and mental phenomena.

Various ways for NFs to occur may actually be sketched as follows.

(1) NFs may be conceived just as intended for IP. Then this is the only advantage that can justify particular models, software and hardware relied on this approach.

(2) NFs can be generated by NNs. So, probability distributions and geometrical characteristics of phase spaces may easily be treated as NFs.

(3) NFs may be introduced as a continuous approximation of NNs (similar to the hydrodynamical description in statistical physics).

(4) It seems also worthwhile to study electromagnetic fields radiated in the course of the brain activity.

(5) NFs under consideration can be connected with quantum fields. So, it is often supposed that only quantum physics can, in principle, enable explanation of mind-brain interactions (for an introduction see, e.g., [Eccles, 1993; Peres, 1996, 1997a, 1997b] and references therein). However, the ordinary quantum mechanics seems to be too poor for this purpose, and only nonlinear quantum fields that incorporate self-interaction and possess complex intrinsic dynamics could provide, being associated with mind, a physical foundation for links between mental phenomena and brain activity. But a huge number of gauge theories and other nonlinear quantum models meet this general re-

quirement and may turn out to be promising. It is reasonable therefore to put forward, as an oncoming approach, a generic mathematical description of NFs that can be relevant for information processing and mental phenomena.

(6) NFs, being interpreted in a quantum fashion, may be incorporated in the wave function of the Universe, which is in line with some interpretations of quantum measurements.

(7) Also, one can speculate about NFs in the spirit of esoteric traditions and parapsychology interpreting NFs as prana, qi, sansa, subtle energy, vital energy, biofield, Ψ -field, information and energetic flows (streams), etc.

Gestalt psychology and the field theory due to W. Köhler, K. Lewin and others may be viewed as a precursor of contemporary studies of NF's variously treated (see, e.g., [Wilson, Cowan, 1973; Amari, 1977, 1983; Ingber, 1991; Liang, 1995; Bressloff, 1996; Jirsa, Haken, 1996; Kistler et al, 1998]).

It is worth noting that NFs can appear essential for IE and IP. So, the primary property which enables information to be useful for an individual is its meaning. Hence, thinking, as a transformation of information, turns out to be operations with meanings of objects instead of manipulations with the real objects.

It is significant that the meaning of any thought is, generally speaking, inexpressible completely in a finite number of words (or signs of another kind), and consequently meanings are inevitably infinite-dimensional entities. In particular, if a person would like either to convey or to understand a meaning more precisely, one can sequentially clarify it in dialogs, in conversations or repeated attempts to read and comprehend texts.

It is attempts to handle meaning of information that are the direct reason to deal with NF here. To treat NFs, we exploit quantum theory to take advantage as a clue. (In this methodological sense we follow the spirit of the known books [Arbib, 1972; Hofstadter, 1979; Pribram, 1971].)

We try to develop a fairly general framework such that ordinary (deterministic)

NNs, their stochastic counterparts and proper quantum systems appear to be particular cases of the subject. In a sense, our general setting should embrace both the NNs machinery and the world of systems treated in quantum mechanics and quantum field theory, and it bridges the gap between these two realms. (The second quantization is not incorporated yet into the approach, but this seems to be attainable with the aid of the functional integration technique.)

A part of the results given in the paper was presented before, in a preliminary form, at meetings [Linkevich 1996, 1998b, 1998c, 1999a, 1999b, 1999c].

The paper is organized as follows. In the next section we consider main properties of a NF and introduce basic quantities to describe IE and IP associated with time evolution of the field. In Sect.3 we formulate the learning problem for NFs and give solutions of the problem that are direct generalizations of learning algorithms proposed before for NNs in [Kapelko, Linkevich, 1996; Kartynnick, Linkevich,

1994; Linkevich, 1992, 1993a, 1993b]. Sect.4 is devoted to the representation of the state of the NF in a form similar to that accepted in quantum theories, which lead us to a kind of supersymmetry. An evolution equation for a field function is suggested in Sect.4. A structure of a cognitive system is considered in Sect.5 where the meanings of information are treated as codes determined by a settled neural field function.

This paper is conceived as the first in a series of works devoted to NFs. In subsequent papers, we are going to consider in more detail such issues as measures generated by NNs and other dynamical systems, connections of NFs with NNs as well as with quantum fields, implications of NFs for dynamical foundations of mental phenomena and for mind-brain interactions.

2 Main Properties of Neural Fields

Let us consider a neural field (NF) whose *state* at time moment t is described by a vector $|\phi(t)\rangle$ of some linear topological space Φ . Specifically we will deal with Banach and Hilbert spaces. (For an introduction into notions of functional analysis used in this section see, e.g., [Edwards, 1965; Kantorovich, Akilov, 1982; Reed, Simon, 1972; Yosida, 1968].)

Further we define a *field function* (FF) $\phi(x, t)$ which can variously be viewed depending on what the state space Φ is. So, let Φ be a kind of functional space, i.e. $\Phi = \{\phi : (x, t) \rightarrow y, x \in X, t \in \mathbf{R}, y \in Y\}$. Then it is natural to define the quantity $\phi(x, t)$ as the value of a function from $\Phi : \phi(x, t) \in Y$. If Φ is regarded only as abstract space, then one can involve its dual space Φ' (i.e. the space of linear continuous functionals defined on Φ) and interpret the FF $\phi(x, t)$ as either the value of a functional $x \in \Phi'$ at the point $|\phi(t)\rangle \in \Phi$, or, vice versa, as the value of the functional $|\phi(t)\rangle \in \Phi$ at a point $x \in \Phi'$ (the latter variant becomes possible if we appropriately redefine Φ and Φ' so that Φ' be proper space and Φ its dual space).

If Φ is separable Hilbert space, then there exists a finite or countable infinite complete set of orthonormal basis vectors. Taking such a set of vectors $|x\rangle$ as a basis, we may assign the scalar product $\langle x|\phi(t)\rangle$ to be the FF: $\phi(x, t) = \langle x|\phi(t)\rangle$.

If Φ is infinite-dimensional (that is the case of our main concern), then one can use not only a countable set of orthonormal basis vectors, but also an uncountable (continuous) set of orthogonal vectors normalized on the Dirac's delta function. It is often convenient to choose a basis composed of eigenvectors $|x\rangle$ of a linear operator, and then the basis is countable or uncountable depending on whether the operator spectrum is discrete or continuous. It is significant that the eigenvectors $|x\rangle$ belong, generally speaking, not to the space Φ , but to another space [Gelfand, Shilov, 1967].

We assume that information is set into the FF $\phi(x, t)$. Then time evolution of the NF generates IP, and the result

of computation (the output of IP) may be extracted from the NF. Specifically, we suppose that any piece of information (PI) is derived from some *settled* neural field function (SNFF) $\psi(x)$ which emerges under suitable conditions after a time interval. Thereby we announce an operator $\Gamma : \phi \rightarrow \psi$.

A natural way is to define $\psi(x) = \lim_{t \rightarrow T} \phi(x, t)$ where T is set to be equal to either infinity or a certain finite value which is preassigned in advance or adjusted during a session of IP. Another variant is to average the FF $\phi(x, t)$ with some weight $K(t)$ so as $\psi(x) = \int K(t) \phi(x, t) dt$, which has a certain neurobiological motivation as well.

It is obvious that the inverse transformation Γ^{-1} is, generally speaking, not unique, which may play a role in the NF learning because a significant freedom remains in determining the function $\psi(x)$.

The crucial issue is how to read out information from the SNFF $\psi(x)$. To put forward a feasible approach to this point, we suppose that there exists an operator $\Lambda : \psi \rightarrow \mu$ to transform $\psi(x)$ into a Borel measure $\mu(x)$ supported on X . (For an introduction into the theory of measure see, e.g., [Halmos, 1974; Reed, Simon, 1972; Royden, 1968].)

Introduce further the integral

$$Q(f, A, \mu) = \int_A f(x) d\mu(x) \quad (2.1)$$

for any Borel measurable function $f(x)$ defined on X and for any Borel measurable set A from the σ -algebra $\sigma(X)$ on X . The functional Q may be assigned to convey information embedded in the NF. Indeed, given an appropriate function f and set A , the value of Q depends on the NF state and can be viewed as a carrier of some PI. Variation of f and/or A yields then different PIs sunk into the given NF. Having chosen in advance a family of functions f and/or a family of sets A , one can evaluate, for each of them (and for a given NF state), the quantity Q and treat the family of its values as codes of the corresponding PIs stored in the NF. (The quantity Q can, in particular, be viewed also as the average $\mu_A(f)$ of the function $f(x)$ with the measure $\mu(x)$ over the space A .)

To put the above suggestion in slightly more detail, let us consider a family of measurable functions $f^* = \{f^\alpha(x), x \in X, \alpha \in I_f\}$ and/or a family of measurable sets $A^* = \{A^\beta \in \sigma(X), \beta \in I_A\}$. Here the indices α and β may, generally speaking, take both discrete and continuous values from some sets I_f and I_A respectively (and these I_f and I_A may be admitted to be unions of finite, countable and uncountable sets). As elements of the family f^* , one can use both functions given by different equations and algorithms, and functions which are determined in just the same manner but differ only in numeric values of some parameters. A similar consideration can be offered for the family of sets A^* as well. Actually, dependence of Q on A may easily be incorporated into dependence of Q on f since $Q(f, A, \mu) = Q(f\chi_A, X, \mu)$ where the indica-

tor $\chi_A(x) = 1$ for any $x \in A$ and $\chi_A(x) = 0$ if $x \notin A$. It may appear rather advisable to distinguish, however, these two factors in some cases (e.g., if fractal sets are taken to be elements of A^*).

The value of the functional $Q(f^\alpha, A^\beta, \Lambda \psi_k)$ obtained for a given function f^α and a given set A^β provided that the NF occurs in a state resulting in a given SNFF ψ_k can be interpreted as a code $U_k^{\alpha\beta}$ of the corresponding PI which is stored in the NF.

There is an essential distinction between the scheme stated above and prevailing approaches to neural and quantum computations. So, in the attractor neural networks paradigm, what is really used to encode a PI is only the number (label) of an attractor. As such an attractor has been retrieved, nothing else may be obtained in respect to the corresponding PI. Similarly, in quantum IP systems, only labels of quantum states are commonly used to represent PIs. In contrast, if, for given f and A , the values of $Q(f, A, \Lambda \psi)$ do not provide an appropriate representation of PIs, one can enrich the codes by making use of additional functions f and/or sets A .

A family of functions f^* together with a family of sets A^* act like a measuring device which determines, with a certain accuracy, data describing the NF, viz the values of the functional Q . Using different elements of f^* and/or different elements of A^* enables one to read out different PIs from just the same state of the NF. Moreover, the resolution of this "device" may, in principle, be improved as much as required, and a code $U_k^{\alpha\beta}$ associated with a PI may be obtained as precise and detailed as necessary with the aid of additional "measurements" exploiting, as a matter of fact, additional functions f^α and/or additional sets A^β chosen to be more and more "subtle".

Let us mention in passing that with a more general and abstract point of view the function $\psi(x)$ looks like an "embryo" of a measure, and this can easily be cast in the definition of an appropriate mathematical object as follows.

Let ψ be a function defined on space X , and there exists an operator Λ such that $\mu = \Lambda \psi$ is a Borel measure supported on X . Then ψ is called a *generator* of a Borel measure associated with the operator Λ .

The operator Λ may variously be chosen depending on both specific IP tasks to be solved and implementation facilities. So, one can simply identify $\psi(x)$ with the probability distribution density, which leads to our previous approach [Linkevich, 1998]. This scheme emerges naturally if one deals, e.g., with a NN governed by stochastic differential equations.

For a pure (coherent) quantum system, it is natural to regard $\psi(x)$ as the wave function of the system in the x -representation, and define accordingly $\mu(dx) = |\psi(x)|^2 dx$. Another possibility is to treat the measure $\mu(x)$ as the Wigner distribution function.

A special case is when the measure $\mu(x)$ is absolutely continuous with respect to another measure $\nu(x)$ taken to be fixed. In other words, we suppose that for any measure $\mu(x)$ that can occur during IP and for a given measure

$\nu(x)$, there exists such a function $\rho(x)$ called the Radon - Nikodym derivative that $d\mu(x) = \rho(x) d\nu(x)$. Then one can assume existence of an operator $\Lambda : \psi \rightarrow \rho$ and cast Q into the form

$$Q = \int_A \rho(x) dF(x) \tag{2.2}$$

where the measure $dF(x) = f(x) d\nu(x)$ appears to be absolutely continuous with respect to the "reference" measure $\nu(x)$ as well.

In the particular case when $d\nu(x) = dx$, we have the form

$$Q = \int_A \rho(x) f(x) dx \tag{2.3}$$

similar to a scalar (inner) product of the functions $\rho(x)$ and $f(x)$.

It is quite reasonable to choose the functions f to be elements of an orthonormal basis of the space Φ . Then eq.(2.3) provides the Fourier coefficients of the expansion of the function $\rho(x)$ over this basis.

Yet another justification for the quantities (2.2) to be introduced appears if the space Φ comprises all bounded continuous complex functions $\psi(x)$ vanishing at infinity, i.e. such that $\psi(x) \rightarrow 0$ as $\|x\| \rightarrow \infty$. Then any continuous linear functional defined on Φ may uniquely be represented in the form (2) with $F(x)$ taken to be a bounded complex-valued measure (see, e.g., [Jost, 1965]).

Time evolution of the NF may mostly be put on display as follows (a particular case is considered in more detail in Sect.5). Let $U(t, t_0)$ be the evolution operator for the NF state, i.e.

$$|\phi(t)\rangle = U(t, t_0) |\phi(t_0)\rangle \tag{2.4}$$

Then the time evolution of the field function

$$\phi(x, t) = \int dx_0 G(x, t|x_0, t_0) \phi(x_0, t_0)$$

is governed by the Green function (propagator) $G(x, t|x_0, t_0) = \langle x|U(t, t_0)|x_0\rangle$.

A large class of NFs meet evolution equations

$$\frac{d}{dt} |\phi(t)\rangle = H_t(\phi) \tag{2.5}$$

with, generally speaking, nonlinear continuous operators $H_t(\phi)$. It is significant that, under certain conditions, the above evolution operator $U(t, t_0)$ appearing in (2.4) is expressed through $H_t(\phi)$ with the aid of a nonlinear generalization of the Feynman path integrals [Maslov, 1976].

If a basis of the space Φ comprises eigenvectors of a time-dependent operator, then they evolve in time: $|x\rangle = |x(t)\rangle$. This may appear appropriate if X is the state space of a dynamical system. For the case when Φ is some Hilbert space one can hence put $\phi(x, t) = \langle x(t)|\phi(t)\rangle$, or even $\phi(x, t) = \langle x(t)|\phi\rangle$ with the NF state taken to be constant. Thus the three interpretations of $\phi(x, t)$ may, in principle, be offered, viz $\langle x|\phi(t)\rangle$, $\langle x(t)|\phi\rangle$, \langle

$x(t)|\phi(t)\rangle$. However, the two latter cases can easily be reduced to the first with the aid of the relevant evolution operator $V(t)$ such that $|x(t)\rangle = V(t)|x(0)\rangle$. Therefore we can, without loss of generality, restrict ourselves to the case when $\phi(x, t) = \langle x|\phi(t)\rangle$.

If Φ is actually finite-dimensional (i.e. the real dynamics of the NF makes the state vector $|\phi(t)\rangle$ confined in some finite-dimensional subspace $\Phi' \subset \Phi$), then the NF is reduced to a finite number of interacting elements governed by relevant dynamical equations. Such a system can be viewed as a kind of the NN appropriately defined. Our main concern is, however, the case when the NF is an infinite-dimensional system.

It is worth also noting that we have to admit measures that are not positive definite in order to incorporate proper quantum systems into our setting.

3 Some Learning Rules for Neural Fields

3.1 Mathematical Formulation of the Learning Problem for Neural Fields

If a NF can occur in states specified by settled neural field functions (SNFFs) ψ_k , $k \in I_\psi$ and one uses to read out information a family of measurable functions $f^* = \{f^\alpha(x), x \in X, \alpha \in I_f\}$ and a family of measurable sets $A^* = \{A^\beta \in \sigma(X), \beta \in I_A\}$, then the NF contains PIs whose codes are $Q(f^\alpha, A^\beta, \Lambda\psi_k)$, $\alpha \in I_f, \beta \in I_A, k \in I_\psi$. Accordingly, the learning problem is formulated as the inverse problem: given a set of codes $U_k^{\alpha\beta}$, $\alpha \in I_f, \beta \in I_A, k \in I_\psi$, find SNFFs ψ_k , measurable functions f^α and measurable sets A^β such that

$$Q(f^\alpha, A^\beta, \mu_k) = U_k^{\alpha\beta} \tag{3.1}$$

with $\mu_k = \Lambda\psi_k$ for all α, β and k .

It is easy to observe that this task falls apart into the following two problems: (i) given a set of codes $U_k^{\alpha\beta}$, find measures μ_k , functions f^α and sets A^β such that the above condition (3.1) holds; (ii) given a set of known (obtained) measures μ_k , construct an operator $\Lambda^{-1} : \mu_k \rightarrow \psi_k$.

It is obvious that the second problem has no general solutions for lack of a general constructive definition of the operator Λ . In the simplest case, μ and ψ are simply the same. For quantum systems, usually one has $d\mu(x) = |\psi(x)|^2 dx$ so that ψ is easily reconstructed up to an arbitrary phase factor.

In the present work we are concentrated on the first problem restricting ourselves by the case when $\mu(x)$ is absolutely continuous with respect to a "natural" Lebesgue measure defined on the same space X : $d\mu(x) = \rho(x)dx$. Here $\rho(x)$ is the Radon-Nikodym derivative (density). Thus instead of $Q(f, A, \mu)$ we will deal in what follows with func-

tionals of a more special type:

$$q(f, \rho) = \int_X f(x) \rho(x) dx \tag{3.2}$$

(Here dependence on the set A is incorporated into dependence on the function f as discussed above.) Accordingly, the learning problem is reduced to finding measure densities ρ_k and functions f^α such that

$$q(f^\alpha, \rho_k) = u_k^\alpha, \quad \alpha \in I_f, \quad k \in I_\psi \tag{3.3}$$

for a given set of codes u_k^α .

3.2 Reduction of the Learning Problem to an Algebraic Problem

Further we will consider the customary case when the domains of indices I_f and I_ψ are finite sets of discrete values which always (after an appropriate renumeration if necessary) may be written so as $I_f = \{1, \dots, M\}$, $I_\psi = \{1, \dots, K\}$. Let us expand the measure density $\rho_k(x)$ over a basis $h_l(x)$, $l = 1, 2, \dots$:

$$\rho_k(x) = \sum_{l=1}^L \rho_{kl} h_l(x), \quad \forall k = 1, \dots, K \tag{3.4}$$

Here the number L of terms of the series should be, generally speaking, infinite, but we will traditionally restrict ourselves by finite values assuming thereby that the chosen value L ensures a required accuracy of approximation of the measure. If in the course of real information processing it appears that this is not the case, then the value of L should appropriately be increased and therefore the method under consideration must enable us to carry out additional calculations in a convenient enough manner. This pertains also to situations when it is demanded to increase the number M of functions in order to get more precise values of codes of PIs.

Substitution of (3.4) and (3.2) into (3.3) yields the following system of equations:

$$\sum_{l=1}^L \rho_{kl} v_l^\alpha = u_k^\alpha, \quad k = 1, \dots, K, \quad \alpha = 1, \dots, M, \tag{3.5a}$$

or in the matrix form

$$\rho v^\alpha = u^\alpha, \quad \alpha = 1, \dots, M \tag{3.5b}$$

Here

$$v_l^\alpha = \int f^\alpha(x) h_l(x) dx \tag{3.6}$$

and we have introduced the $K \times L$ -matrix $\rho = \|\rho_{kl}\|$ and vectors $v^\alpha = (v_1^\alpha, \dots, v_L^\alpha)$ and $u^\alpha = (u_1^\alpha, \dots, u_K^\alpha)$.

Thus the learning problem is reduced to finding the $K \times L$ -matrix ρ which satisfies eq.(3.5) for given sets of vectors $v^1, \dots, v^M \in \mathbf{R}^L$, $u^1, \dots, u^M \in \mathbf{R}^K$.

3.3 Finding a Family of Solutions of the Learning Problem

The learning problem for neural fields represented in the form (3.5) turns out to be a straightforward generalization of the corresponding task for neural networks for which one has $K = L$. This only difference does not cause any significant obstacles and allows us to extend results obtained previously for networks [Kapelko, Linkevich, 1996; Kartynnick, Linkevich, 1992, 1993a, 1993b] in the case of fields. Therefore we will mention only key points of the procedure of constructing the matrix ρ .

Since (3.5) is a linear system, its general solution can be cast in the form

$$\rho = R + BH$$

where R is a particular solution of (3.5), H is a solution of the corresponding homogeneous system, and B is an arbitrary real $K \times K$ -matrix.

The matrix R can be found as follows. Let R^μ be a solution of the system (3.5) for the first μ pairs of vectors v^1, \dots, v^μ and u^1, \dots, u^μ respectively, i.e.

$$R^\mu v^\kappa = u^\kappa, \quad \kappa = 1, \dots, \mu \tag{3.7}$$

Then

$$R^{\mu+1} = R^\mu + (r^\mu, v^{\mu+1})^{-1} (u^{\mu+1} - R^\mu v^{\mu+1}) \otimes r^\mu, \tag{3.8a}$$

or

$$R_{kl}^{\mu+1} = R_{kl}^\mu + \left(\sum_{i=1}^L r_i^\mu v_i^{\mu+1} \right)^{-1} \left(u_k^{\mu+1} - \sum_{i=1}^L R_{ki}^\mu v_i^{\mu+1} \right) r_l^\mu \tag{3.8b}$$

Here and below $(a, b) = \sum_{l=1}^L a_l b_l$ is the scalar product of vectors a and b ; the notation $a \otimes b$ is used for their direct (tensor) product, i.e. $(a \otimes b)_{ij} = a_i b_j$. The vector $r^\mu = (r_1^\mu, \dots, r_L^\mu)$ should be orthogonal to vectors v^1, \dots, v^μ , i.e.

$$(r^\mu, v^\kappa) = 0, \quad \kappa = 1, \dots, \mu, \tag{3.9}$$

and, in addition, the condition

$$(r^\mu, v^{\mu+1}) \neq 0 \tag{3.10}$$

must be satisfied.

Accordingly, the matrix H is determined by the relation

$$H^{\mu+1} = H^\mu - (r^\mu, v^{\mu+1})^{-1} (H^\mu v^{\mu+1}) \otimes r^\mu, \tag{3.11a}$$

or

$$H_{kl}^{\mu+1} = H_{kl}^\mu - \left(\sum_{i=1}^L r_i^\mu v_i^{\mu+1} \right)^{-1} \left(\sum_{i=1}^L H_{ki}^\mu v_i^{\mu+1} \right) r_l^\mu, \tag{3.11b}$$

which again may easily be verified by direct inspection.

The above receipt of constructing the matrices R and H holds in such a form only if the vectors v^1, \dots, v^M are linearly independent, which is necessary for existence of the vector r^μ obeying the condition (3.10) at any value $\mu = 1, \dots, M$. If this is not the case for some μ then the algorithm needs a modification. Specifically, let us assume first that vectors $v^{\mu+1}$ and $u^{\mu+1}$ are the same linear combinations of vectors v^1, \dots, v^μ and vectors u^1, \dots, u^μ respectively, i.e.

$$v^{\mu+1} = \sum_{\kappa=1}^{\mu} c_{\kappa} v^{\kappa}, \quad u^{\mu+1} = \sum_{\kappa=1}^{\mu} c_{\kappa} u^{\kappa}, \quad (3.12)$$

where c_1, \dots, c_{μ} are some real constant coefficients. From (3.8), (3.11) and (3.12) we obtain the relations

$$R^{\mu} v^{\mu+1} = u^{\mu+1}, \quad H^{\mu} v^{\mu+1} = 0 \quad (3.13)$$

Thus in the case under consideration one has $R^{\mu+1} = R^{\mu}$, $H^{\mu+1} = H^{\mu}$.

Suppose now that the first equation (3.12) holds at some values of the coefficients c_1, \dots, c_{μ} , whereas the second is not satisfied. Then, as is known (see, e.g., [Albert, 1972]), there does not exist an exact solution of the linear system (3.5). Nevertheless in this case we can put again $R^{\mu+1} = R^{\mu}$, $H^{\mu+1} = H^{\mu}$, which yields an approximate solution of eqs.(3.5).

Thus it is required to construct a vector r^μ obeying the conditions (3.9), (3.10) only when vector $v^{\mu+1}$ is linearly independent of vectors v^1, \dots, v^μ . As a result, we arrive at the following algorithm: calculate the vector $\Delta^\mu = H^\mu v^{\mu+1}$; if $\Delta^\mu = 0$ then put $R^{\mu+1} = R^\mu$, $H^{\mu+1} = H^\mu$ and take the next pair of vectors v^κ, u^κ from the sets $\{v^1, \dots, v^M\}$, $\{u^1, \dots, u^M\}$, otherwise find a vector r^μ that meets the conditions (3.9), (3.10).

In this manner the learning problem is reduced to constructing a family of vectors r^μ , $\mu = 1, \dots, M$, obeying (3.9), (3.10) for all $\mu = 1, \dots, M$ provided that vectors v^1, \dots, v^M are linearly independent. To this end, just as in the case of NNs, one can use a number of algorithms as partly described below.

a. Sequential Learning

The simple expression

$$r_i^\mu = \sum_{k=1}^K u_k^{\mu+1} H_{kl}^\mu, \quad l = 1, \dots, L, \quad \mu = 1, \dots, M, \quad (3.14a)$$

or in the matrix form

$$r^\mu = (H^\mu)^T u^{\mu+1}, \quad \mu = 1, \dots, M, \quad (3.14b)$$

has the following remarkable property. Equation (3.14) together with (3.8), (3.11) provides with not only a solution of the learning problem, but also a solution of the sequential learning problem which may be posed as follows. Let R^μ and H^μ be matrices that ensure storage (memorizing)

of μ pieces of information, i.e. they yield a solution of the learning problem (3.5) for μ pairs of vectors v^κ, u^κ :

$$R^\mu v^\kappa = u^\kappa, \quad H^\mu v^\kappa = 0, \quad \kappa = 1, \dots, \mu$$

It is required to find matrixes $R^{\mu+1}, H^{\mu+1}$ that are a solution of the learning problem for $\mu + 1$ pieces of information so that

$$R^{\mu+1} v^\kappa = u^\kappa, \quad H^{\mu+1} v^\kappa = 0, \quad \kappa = 1, \dots, \mu + 1,$$

and, moreover, the matrices $R^{\mu+1}, H^{\mu+1}$ should be expressed only through the matrices R^μ, H^μ and vectors $v^{\mu+1}, u^{\mu+1}$, but not through the vectors $v^1, \dots, v^\mu, u^1, \dots, u^\mu$.

In the case of NNs $K = L$ and therefore one can also put $r^\mu = (H^\mu)^T v^{\mu+1}$ as in [Linkevich, 1992, 1993b].

b. Using Outer Products of Vectors

Algebra of tensors and outer products of vectors enables us to construct a family of vectors r^μ for a given set of vectors v^1, \dots, v^μ as follows [Linkevich, 1993a; Kapelko, Linkevich, 1996]. (For an introduction into the mathematical methods used here see, e.g., [Efimov, Rozendorn, 1970].)

Let us introduce a skew-symmetric tensor \tilde{V} whose components ($m = L - \mu$)

$$\tilde{V}_{i_1 \dots i_m} = \sum_{k_1, \dots, k_\mu=1}^L \varepsilon_{i_1 \dots i_m k_1 \dots k_\mu} v_{k_1}^1 \dots v_{k_\mu}^\mu$$

are expressed through the coordinates of the vectors v^1, \dots, v^μ with the aid of the fully antisymmetric tensor $\varepsilon_{j_1 \dots j_L}$. For the case of the linearly independent vectors v^1, \dots, v^μ the tensor \tilde{V} is nonzero and consequently there exists at least one nonzero component, say, $\tilde{V}_{n_1 \dots n_m}$ with $n_1 < \dots < n_m$.

We construct a vector b taking arbitrary values as coordinates b_{n_1}, \dots, b_{n_m} and determining the other components by the relation

$$b_j = (\tilde{V}_{n_1 \dots n_m})^{-1} \sum_{i=1}^m b_{n_i} \tilde{V}_{n_1 \dots n_{i-1} j n_{i+1} \dots n_m}$$

where $j \in \{1, \dots, L\} \setminus \{n_1, \dots, n_m\}$. Every such a vector is orthogonal to the vectors v^1, \dots, v^μ and therefore can be used as the vector r^μ .

It is easy to see that the vector $r^\mu = b$ is determined up to m arbitrary parameters given explicitly. These degrees of freedom may be exploited to control information processing like it appears in the case of NNs [Kapelko, Linkevich, 1996].

c. Using the Gram-Schmidt Orthogonalization

This known method (see, e.g., [Gantmacher, 1959; Strang, 1976]) enables us to construct the vectors r^μ as follows:

$$r^\mu = v^{\mu+1} - \sum_{\alpha=0}^{\mu-1} \frac{(r^\alpha, v^{\mu+1})}{(r^\alpha, r^\alpha)} r^\alpha, \quad \mu = 0, 1, \dots, M$$

Here the value $\mu = 0$ is added because the auxiliary vector $r^0 = v^1$ is necessary for calculations. For NNs this learning rule was considered in [Linkevich, 1993a; Kartynnick, Linkevich, 1994].

4 Neural Fields vs Quantum Fields: Supersymmetric Neurodynamics?

Here we address ourselves again to separable Hilbert space Φ associated with a system whose state at time moment t is described by a vector $|\phi(t)\rangle \in \Phi$. We offer a consideration along lines of quantum theory (for an introduction see, e.g., [Bogolyubov, Shirkov, 1959; Emch, 1972; Itzykson, Zuber, 1980; Ryder, 1985; Umezawa et al, 1982; Ziman, 1969]) though we do not imply any proper quantum objects to be involved further. On the contrary, all this proves to be nothing but only a mathematical framework that can be used for a wide range of situations (so, in [Linkevich, 1999b] a similar treatment is suggested in slightly more detail for signal processing).

As Φ is separable, there exists a finite or countable infinite complete set of basis vectors $|\alpha\rangle, |\alpha_1, \alpha_2\rangle, \dots, |\alpha_1, \dots, \alpha_m\rangle, \dots$ where any index α_i takes discrete values. We can assume, just as in quantum theory, that there exist a *unique cyclic unit* vector $|0\rangle \in \Phi$ called the *vacuum* state vector and operators a_α such that any basis vector arises as

$$|\alpha_1, \dots, \alpha_m\rangle = a_{\alpha_1}^+ \dots a_{\alpha_m}^+ |0\rangle, \tag{4.1}$$

whereas $a_\alpha |0\rangle = 0$. Here and below the sign $+$ denotes the Hermitian conjugation. It is convenient to deal with such operators a_α, a_α^+ that are permutable so that

$$a_\alpha a_\beta^+ = c_{\alpha\beta} a_\beta^+ a_\alpha + d_{\alpha\beta} \tag{4.2}$$

where $c_{\alpha\beta}$ and $d_{\alpha\beta}$ are some c -numbers. Then the orthonormality condition $\langle \alpha | \beta \rangle = \delta_{\alpha\beta}$ yields $d_{\alpha\beta} = \delta_{\alpha\beta}$, while $c_{\alpha\beta}$ remains arbitrary.

In quantum theory, only two values of the quantity $c_{\alpha\beta}$ are commonly considered, viz $c_{\alpha\beta} = \pm 1$, so that (4.2) takes the form: $[a_\alpha, a_\beta^+]_{\mp} = \delta_{\alpha\beta}$. The first (upper) case corresponds to the Bose-Einstein statistics, while the second pertains to the Fermi-Dirac one.

In contrast, in our scheme any value of $c_{\alpha\beta}$ may be taken, and the theory should, strictly speaking, be invariant with respect to such a choice. In this way, we immediately meet a kind of *supersymmetry*, or symmetry between bosonic and fermionic modes of excitation of the NF.

Thus we conclude here that any state vector $|\phi\rangle \in \Phi$ can be represented in the form

$$|\phi\rangle = \sum_{m=0}^{\infty} \sum_{\alpha_1, \dots, \alpha_m} \phi_{\alpha_1 \dots \alpha_m} a_{\alpha_1}^+ \dots a_{\alpha_m}^+ |0\rangle \tag{4.3}$$

where the term with $m = 0$ is simply $\phi_0 |0\rangle$. In quantum theory, the quantity $\phi_{\alpha_1 \dots \alpha_m}$ is referred to as the wave func-

tion of the system in the α -representation. The above expansion is ultimately adopted in most contemporary models of the quantum physics including such modern trends as supersymmetric quantum theories, string models, etc.

Further we can introduce the operators

$$q_\alpha = -iq_\alpha(a_\alpha^+ - a_\alpha), \quad p_\alpha = p_\alpha(a_\alpha^+ + a_\alpha)$$

with c -number constants q_α, p_α such that these operators obey the *canonical commutation relations*: $[q_\alpha, p_\beta]_{\mp} = i\hbar \delta_{\alpha\beta}$.

One can also show that the so-called harmonic oscillator appears the simplest system in the sense that its time evolution is governed by linear differential equations for the operators a_α, a_α^+ . No wonder that the harmonic oscillator is the prevailing model system in various areas. This is nothing but the first approximation of dynamical equations.

An interpretation and possible implications of bosonic and fermionic modes of excitation of the NF will be given in a subsequent paper where a further advance in examination of connections between NFs and quantum fields is supposed to be presented as well.

5 An Evolution Equation for the Neural Field

The aim of this section is to specify appropriately the FF $\phi(x, t)$ and suggest a particular dynamical equation to describe its time evolution. Namely, we suppose that a real d -dimensional variable $x = (x_1, \dots, x_d)$ determines a point inside a neural system (either brain or an artificial computing system), whereas a real M -dimensional function $\phi(x, t) = (\phi^1(x, t), \dots, \phi^M(x, t))$ represents the state of the neuron located at point x at time t . (More precisely, it is usually implied a short-time average and average over neurons placed around the position x .) The FF is chiefly chosen to be the postsynaptic somatic membrane potential of the neuron (so that $M = 1$). (For more detail see [Wilson, Cowan, 1973; Amari, 1977, 1983; Ingber, 1991; Liang, 1995; Bressloff, 1996; Jirsa, Haken, 1996; Kistler et al, 1998].)

This continuous approximation can be justified by sufficiently dense disposition of neurons such that the distance between two nearest neurons is significantly smaller than the characteristic length scale of a function $W^{\alpha\beta}(x, y)$ that describes the synaptic interaction of neurons located at points x and y . This enables us to replace the real discrete arrangement of neurons by a continuous line ($d = 1$), surface ($d = 2$) or medium ($d = 3$) formed by the nerve tissue.

Relying on data acquired in neuroscience we suggest describing dynamics of the NF by equations

$$\partial_t \phi^\alpha(x, t) = F^\alpha(\phi(x, t)), \quad \alpha = 1, \dots, M, \tag{5.1}$$

where the operator

$$F^\alpha(\phi(x, t)) = A^{\alpha\beta}(x) \phi^\beta(x, t) + \partial_{x_i} \partial_{x_j}$$

$$(B_{ij}^{\alpha\beta}(x) \phi^\beta(x, t)) + J^\alpha(\phi(x, t)) + I^\alpha(x, t) \quad (5.2)$$

contains two quasilinear terms, viz the first and the second, which correspond to weak excitations of the NF and generalized diffusion effects respectively, the nonlinear contributions

$$J^\alpha(\phi(x, t)) = K^\alpha(\phi(x, t)) + \int dy W^{\alpha\beta}(x, y) f^\beta(\phi(y, t)) \quad (5.3)$$

to incorporate self-interaction of the NF and interaction between its different parts, and an external input signal $I^\alpha(x, t)$ entering the neuron located at point x at time t . Summation over repeated indices is assumed hereafter, and we use the notations $\partial_t = \frac{\partial}{\partial t}$, $\partial_{x_i} = \frac{\partial}{\partial x_i}$.

The matrices $A^{\alpha\beta}(x)$ and $B^{\alpha\beta}(x)$ are taken, generally speaking, to be dependent on the variable x in order to take heterogeneity and spatial nonhomogeneity of the system into account. If, instead, the NF may be regarded as spatially uniform, then these matrices are constant.

The matrix $A^{\alpha\beta}(x)$ is responsible for the dynamical behavior of the NF under its weak rippling excitations (when the values of $\phi(x, t)$ are small, the self-interaction term $K^\alpha(\phi)$ can be omitted, the transfer function $f^\beta(\phi)$ is close to zero and entails cross-interaction vanishing, and, in addition, the spatial distribution of disturbances is nearly flat so that diffusion turns out to be negligible).

The second term in (5.2) makes the system to be of the reaction-diffusion type to model effects of the spatiotemporal organization of the brain in a nonsynaptic diffusion neurotransmission field in accord with evidence of diffusion through extracellular fluid and across membranes (see [Liang, 1995] and references therein).

The self-interaction term $K^\alpha(\phi)$ is the chief novelty of our model and it may appear requisite to originate and maintain ordered structures in the NF. So, it is widely believed that the nerve tissue manifests itself as an excitable medium. In models of NNs, this peculiarity is usually provided by an appropriate architecture inspired by the structure of the visual cortex that includes a short-range excitation and a long-range inhibition of interacting neurons. Such systems exhibit a surprisingly rich repertoire of patterns including traveling waves, rotating spirals, concentric expanding rings, etc. (see, e.g., [Ermentrout, 1998; Kistler et al, 1998]). Meanwhile, there exists neurobiological evidence that a single isolated neuron is by itself an excitable element, which result, in particular, in existence of sustained complex oscillations (see, e.g., [Kapelko, Linkevich, 1996] and references therein).

To incorporate this feature of the NF, we follow the FitzHugh neuronal model [FitzHugh, 1961] and its modifications [Kapelko, Linkevich, 1996; Linkevich, 1997, 1998] choosing the form of $K^1(\phi)$ and putting all the other $K^\alpha(\phi) = 0$, $\alpha = 2, \dots, M$. To wit, we introduce the notation

$$\tilde{K}(\phi) = A^{11}\phi + K^1(\phi) \quad (5.4)$$

and suggest the following variants:

$$\tilde{K}(\phi) = a\phi - b\phi^3, \quad b > 0, \quad (5.5)$$

$$\tilde{K}(\phi) = \begin{cases} -h\phi + (g + h) & \phi \geq 1 \\ g\phi & -1 < \phi < 1 \\ -h\phi - (g + h) & \phi \leq -1 \end{cases} \quad (5.6)$$

$$\tilde{K}(\phi) = -a\phi + b \tanh(g\phi), \quad a > 0, \quad (5.7)$$

with constants a, b, g, h . It is significant that these functions ensure asymptotic description of experimental data on the neuron current as a function of the membrane potential and the remarkable N -like shape of $\tilde{K}(\phi)$ which contains a negative resistance region essential to maintain oscillations in bursting neurons [Wilson, Wachtel, 1974].

Lastly, the function $W^{\alpha\beta}(x, y)$ yields the strength of the influence of the output firing rate $\omega^\beta = f^\beta(\phi(y, t))$ of the neuron located at point y on the state of the neuron at point x . Here the transfer input-output function $f^\beta(\phi)$ of the neuron is taken to be of a sigmoid shape, i.e. strictly monotonically increasing and bounded. A simple prevalent choice is $\omega = c \tanh(g\phi)$ with positive constants c, g (see, e.g., [Linkevich, 1997] and references therein for other forms).

Particular cases of our model are diverse and can be obtained along various lines of simplification. So, if we put $M = 1$, $A^{11}(x) = A = const$, $B^{11}(x) = 0$, $K^1(\phi) = 0$, then from the above equations one gets

$$\partial_t \phi(x, t) = A \phi(x, t) + \int dy W(x - y) f(\phi(y, t)) + I(x, t) \quad (5.8)$$

where $W(x - y) = W^{11}(x - y) = W^{11}(x, y)$ in view of the homogeneity assumption adopted here. It is just the equation that is mainly used to model the nerve tissue (see, e.g., [Amari, 1977, 1983; Heiden, 1980; Murray, 1990]).

A stochastic counterpart to the model (5.1) is established so as

$$\partial_t \phi^\alpha(x, t) = F^\alpha(\phi(x, t)) + \xi^\alpha(x, t) \quad (5.9)$$

where $\xi^\alpha(x, t)$ is Gaussian white noise with zero means $\langle \xi^\alpha(x, t) \rangle = 0$ and the autocorrelation functions

$$\langle \xi^\alpha(x, t) \xi^\beta(y, s) \rangle = 2\Gamma^{\alpha\beta}(x, y) \delta(t - s)$$

6 A Minimal Architecture of a Cognitive System, Dynamical Background of Representation of Meanings, and Mental Phenomena

Here we sketch the general structure of a cognitive system relying on data accumulated in modern science (for an introduction into cognitive psychology see, e.g., [Anderson, 1990; Barsalou, 1992; Johnson-Laird, 1983; Pylyshyn, 1986; Shepard, 1990; Solso, 1988]) and outline ways to

implement basic elements of such a system in the framework of the developed approach. A more detailed description and results of computer simulations will be presented in a subsequent paper. It is worth also noting that a NF appears rather a kind of informational machine [Zeleznikar, 1995, 1997] that cannot be reproduced completely by digital computers.

Objects of the environment produce stimula that influence on receptors and, after processing by sense organs and the brain, generate mental images called *percepts*. Such images may be stored in the memory and then they become known as memorized images or memorized patterns.

Images can also be created. To *imagine* something means to produce an image of an object that is not perceived here and now, i.e. the object is outside the area of direct perception, or the object does not exist either yet or at all. Such images are referred to as *imagined (created) images*.

Language provides facilities to develop, maintain and exploit an elaborate system of *meanings* and associate them with objects of the environment. Therefore a cognitive system should be capable to deal with words and sentences during IP.

Thus a cognitive neural system (CNS) should comprise at least three interconnected subsystems referred to as *perceptual* neural system (PNS), *lexical* neural system (LNS) and *semantic* neural system (SNS).

In our previous approach [Linkevich, 1997] three interconnected NNs are treated as perceptual, lexical and semantic NNs (PNN, LNN and SNN) and their activity spaces A_p, A_l, A_s are used as *perceptual* space (PS), *lexical* space (LS) and *semantic* space (SS) so that (i) any *image* is encoded by an attractor ω_p^κ in the *perceptual* space; (ii) any known *word* is given by an attractor ω_l^λ in the *lexical* space A_l ; (iii) any *meaning* understandable to the system is represented by an attractor ω_s^μ in the *semantic* space A_s .

However, such a uniform description is deficient because meanings are rather infinite-dimensional and a kind of continuous medium seems more appropriate. Therefore we introduce a *neural field* and treat it as a *semantic field* to implement meanings. Of course, such a field is only a sketchy substitute for real complicated structures and interactions between various parts of the brain and other organs.

Thus our refined approach to representation of information looks as follows.

(i) Any *image* is still viewed as an attractor ω_p^κ in the activity space A_p of an analog NN called the PNN.

(ii) It may be more appropriate to exploit a discrete (e.g., binary) NN to encode words so that any *word* is given by an attractor ω_l^λ in the activity space of a LNN.

(iii) It is suggested to treat the meaning of a piece of information as a code $\tilde{\omega}_s^\mu = Q(f^\alpha, A^\beta, \Lambda \psi_k)$, $\mu = \{\alpha, \beta, k\}$, associated with a settled field function ψ_k generated by a NN referred to as the SNN.

Here and below index q stands for p, l or s (i.e. perceptual, lexical or semantic), the sign \sim denotes

steady states of neural systems used for IE, the state vector of the NN $X_q(t) = (x_{q1}(t), \dots, x_{qN_q}(t))$ is composed of the state vectors of neurons $x_{qi}(t) = (x_{qi}^1(t), \dots, x_{qi}^{M_q}(t))$, $i = 1, \dots, N_q$, and the activity vector is $\omega_q(t) = (\omega_{q1}(t), \dots, \omega_{qN_q}(t))$.

The learning of the PNS, LNS, SNS may be carried out as follows. A supervisor provides input signals I_p^α and I_l^α , $\alpha = 1, 2, \dots$, to the PNS and LNS. Attractors $\tilde{\omega}_p^\alpha$ and $\tilde{\omega}_l^\alpha$ are formed due to an appropriate adjusting of the synaptic couplings, relevant settled states emerge due to interconnections between the three NSs. Thus, connections between $\tilde{\omega}_p^\alpha, \tilde{\omega}_l^\alpha$ and $\tilde{\omega}_s^\alpha$ are produced.

Performance of the system in response to an external signal can be of the two kinds: (i) As an input I_p is close to some sample input I_p^α , the attractor $\tilde{\omega}_p^\alpha$ is retrieved. Besides, the patterns $\tilde{\omega}_l^\alpha$ and $\tilde{\omega}_s^\alpha$ are restored due to interconnections between the NSs. (ii) As a signal I_l appears close to some I_l^α , the corresponding $\tilde{\omega}_l^\alpha$ and, hence, $\tilde{\omega}_p^\alpha$ and $\tilde{\omega}_s^\alpha$ emerge.

Performance in the no-signal condition when activity of the NS evolves toward a settled state during IP, whereas no input signal exists: (iii) Appearance of a word (attractor) $\tilde{\omega}_l^\alpha$ brings $\tilde{\omega}_p^\alpha$ and $\tilde{\omega}_s^\alpha$. (iv) If a state φ occurs close to some meaning, the NSs approach the corresponding $\tilde{\omega}_s^\alpha, \tilde{\omega}_p^\alpha$ and $\tilde{\omega}_l^\alpha$.

A system composed of the PNS, LNS and SNS could provide a mathematical proving ground for studies of mental phenomena along the lines of the dual-coding hypothesis, radical imagery hypothesis, conceptual-propositional hypothesis and other assumptions of cognitive psychology concerning representation of information in the mind.

Thus the scheme of the CNS can be sketched as follows:

- perceptual neural system (PNS)

kind: neural network
 to represent: images
 state: $X_p(t) \in S_p \subset \mathbf{R}^{M_p N_p}$
 carrier: activity $\omega_p(t) \in A_p \subset \mathbf{R}_+^{L_p}$
 codes: attractors $\tilde{\omega}_p^\kappa \in A_p$,
 $\kappa = 1, \dots, K_p$

- lexical neural system (LNS)

kind: neural network
 to represent: words, sentences
 state: $X_l(t) \in S_l \subset \mathbf{R}^{M_l N_l}$
 carrier: activity $\omega_l(t) \in A_l \subset \mathbf{R}_+^{L_l}$
 codes: attractors $\tilde{\omega}_l^\lambda \in A_l$,
 $\lambda = 1, \dots, K_l$

- semantic neural system (SNS)

kind: neural field
 to represent: meanings
 state: $|\varphi(t)\rangle \in \Phi$
 carrier: settled field function $\psi(x)$
 codes: $\tilde{\omega}_s^\mu = Q(f^\alpha, A^\beta, \Lambda \psi_k)$,
 $\mu = \{\alpha, \beta, k\} = 1, \dots, K_s$

7 Conclusion

Thus we put forward a rather general approach in which information is set into a kind of infinite-dimensional topological vector space referred to as the NF, and IP is associated with time evolution of the field due to its intrinsic dynamics.

The epistemology which is in tune with our framework can be phrased as follows [Gombrich, 1960; Popper, 1963; Popper, Eccles, 1977].

Knowledge is always a modification of previous knowledge, and it is gained due to learning for all the life. The senses challenge us to make our hypotheses and match them, and a hypothesis precedes an observation.

In broad outline, we accept that knowledge is embedded into a NF, and learning is the NF updating. Input signals disturb spontaneous dynamics and make the NF undergo the forced feedback time evolution treated as IP. The output of IP is extracted from the NF after a time interval.

NFs seem to be particular dynamical systems that belong to the class of informational machines that are characterized by non-computability so that they cannot be imitated completely by digital computers but should operate by themselves as informing entities [Zelevnikar, 1995, 1997] and possess certain primary meanings embedded into them [Linkevich, 1997]. The infinite dimension of NFs appear essential to handle information taking its meaning into account.

As a NF is not a set of discrete elements, but rather a kind of continuously distributed systems, it is quite natural to try condensed mediums for implementation of the suggested approach. In this respect, mediums obeying reaction-diffusion equations could be especially promising due to a rich variety of remarkable nonlinear phenomena such as self-organization, dissipative and fractal structures, phase transitions, spatio-temporal chaos, etc. [Prigogine, 1990].

In subsequent paper we are going to proceed investigation of these and other issues.

Acknowledgement

The author is thankful to A.B. Antonevich, C.C.A.M. Gielen, A.M. Krot, G.G. Krylov, V.I. Kuvshinov, P.V.E. McClintock, M. Perus, N.M. Shumeiko, J.A. Starzyk and A.T. Vlassov for useful discussions and interest in the work.

References

- [1] Albert, A. 1972. Regression and Moore-Penrose Pseudo-inverse. Academic. New York.
- [2] Amari, S.-I. 1977. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol. Cybern.* **27**: 77-87.
- [3] Amari, S.-I. 1983. Field theory of self-organizing neural nets. *IEEE Trans. Syst. Man. Cybern.* **SMC-13**: 741-748.
- [4] Amit, D.J. 1989. *Modelling Brain Functions*. Cambridge Univ. Press. Cambridge.
- [5] Anderson, J.R. 1990. *Cognitive Psychology*. Freeman. San Francisco.
- [6] Arbib, M.A. 1972. *The Metaphorical Brain*. Wiley. New York.
- [7] Barsalou, L.W. 1992. *Cognitive Psychology*. Erlbaum. Hillsdale. NJ.
- [8] Bogolyubov, N.N & D.V. Shirkov. 1959. *Introduction to the Theory of Quantized Fields*. Interscience. New York.
- [9] Bressloff, P.C. 1996. New mechanism for neural pattern formation. *Phys. Rev. Lett.* **76**: 4644-4647.
- [10] Eccles, J.C. 1993. *How the Self Controls Its Brain*. Routledge. London.
- [11] Edwards, R.E. 1965. *Functional Analysis*. Holt. New York.
- [12] Efimov, N.V. & E.R. Rozendorn. *Linear Algebra and Multidimensional Geometry*. Nauka. Moscow. *In Russian*.
- [13] Emch, G.G. 1972. *Algebraic Methods in Statistical Mechanics and Quantum Field Theory*. Wiley. New York.
- [14] Ermentrout, B. 1998. Neural networks as spatio-temporal pattern-forming systems. *Rep. Progr. Phys.* **61**: 353-430.
- [15] FitzHugh, R. 1961. Impulses and Physiological States in Theoretical Models of Nerve Membrane. *Biophys. J.* **1**: 445-466.
- [16] Frolov, A.A. & I.P. Murav'ev. 1987. *Neuronal Models of Associative Memory*. Nauka. Moscow. *In Russian*.
- [17] Frolov, A.A. & I.P. Murav'ev. 1988. *Information Characteristics of Neural Networks*. Nauka. Moscow. *In Russian*.
- [18] Gantmacher, F.R. 1959. *The Theory of Matrices*. Chelsea. New York.
- [19] Gel'fand, I.M. & G.E. Shilov. 1967. *Generalized Functions*. vol.3. Academic Press. New York.
- [20] Gombrich, E. 1960. *Art and Illusion*. Panteon Books. New York.
- [21] Halmos, P.R. 1974. *Measure Theory*. Springer. New York.

- [22] Heiden, an der U. 1980. *Analysis of Neural Networks*. Springer. New York.
- [23] Hertz, J.A., A. Krogh & R. Palmer. 1991. *Introduction to the Theory of Neural Computation*. Addison-Wesley. Redwood City.
- [24] Hofstadter, D.R. 1979. *Gödel, Escher, Bach: an Eternal Golden Braid*. Basic Books. New York.
- [25] Ingber, L. 1991. Statistical mechanics of neocortical interactions: A scaling paradigm applied to electroencephalography. *Phys. Rev. A* **44**: 4017-4060.
- [26] Itzykson, C. & J.B. Zuber. 1980. *Quantum Field Theory*. McGraw-Hill. New York.
- [27] Jirsa, K. & H. Haken. 1996. Field theory of electromagnetic brain activity. *Phys. Rev. Lett.* **77**: 960-963.
- [28] Johnson-Laird, P.N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press. Cambridge, MA.
- [29] Jost, R. 1965. *The General Theory of Quantized Fields*. American Math. Soc. Providence, RI.
- [30] Kantorovich, L.V. & G.P. Akilov. 1982. *Functional analysis*. Pergamon. Oxford.
- [31] Kapelko, V.V. & A.D. Linkevich. 1996. Chaos and associative generation of information by networks of neuronal oscillators. *Phys. Rev. E*, **54**: 2802-2806.
- [32] Kartynnick, A.V. & A.D. Linkevich. 1994. Retrieval of memorized patterns by non-symmetric neural networks: Local Lyapunov functions and local contracting maps methods. *Optical Memory and Neural Networks*. **3**: 329-342.
- [33] Kistler, B.M., R. Seitz & J.L. van Hemmen. 1998. Modeling collective excitations in cortical tissue. *Physica D*. **114**: 273-295.
- [34] Liang, P. 1995. Neurocomputation by reaction diffusion. *Phys. Rev. Lett.* **75**: 1863-1866.
- [35] Linkevich, A.D. 1992. A sequential pseudo-inverse learning rule for networks of formal neurons. *J. Phys. A*. **25**: 4139-4146.
- [36] Linkevich, A.D. 1993a. Learning neural networks with the aid of projection techniques. Proc. Second Seminar "Nonlinear Phenomena in Complex Systems". Polatsk. February 15 - 17. 1993. edited by V.I. Kuvshinov and D.W. Serow. PIYaF. Saint-Petersburg. 373 - 382.
- [37] Linkevich, A.D. 1993b. A sequential learning rule for analogous neural networks. *Optical Memory and Neural Networks*. **2**: 111 - 116.
- [38] Linkevich, A.D. 1997. Anticipation, perception, language, mind and nonlinear dynamics of neural networks. *Informatica*. **21**: 435-463.
- [39] Linkevich, A.D. 1998a. Some FitzHugh-like neuronal models. *Nonlinear Phenomena in Complex Systems*. **1**: 105-108.
- [40] Linkevich, A.D. 1998b. Representation of information in dynamical systems: measures, entropy and fractals. Proc. Seventh Seminar "Nonlinear Phenomena in Complex Systems". April 14-17, 1998. Minsk. In press.
- [41] Linkevich, A.D. 1998c. Neural fields: systems allied to neural networks and quantum realms. Proc. Fourth Joint Conf. Information Sciences. Research Triangle Park. North Carolina. USA. October 23-28. 1998. **2**: 182-185.
- [42] Linkevich, A.D. 1999a. Neural fields and dynamical foundations of mental phenomena. Proc. to VIII Intern. Seminar 'Nonlinear Phenomena in Complex System'. May 17-20, 1999. Minsk. Belarus. In press.
- [43] Linkevich, A.D. 1999b. Entropy-like functionals and analogs of quantum operators for signal processing. Proc. to VIII Intern. Seminar 'Nonlinear Phenomena in Complex System'. May 17-20, 1999. Minsk. Belarus. In press.
- [44] Linkevich, A.D. 1999c. Some learning algorithms for neural fields. In A.M. Krot, Ed. *Intelligent Systems*. Vol.2. Inst. Engineering Cybernetics of Natl. Acad. Sci. of Belarus. Minsk. In press. In Russian.
- [45] Maslov, V.P. 1976. *Complex Markov Chains and Continual Feynman Integral for Nonlinear Equations*. Nauka. Moscow. In Russian.
- [46] Murray, J.D. 1990. *Mathematical Biology*. Springer. Berlin.
- [47] Peretto, P. 1989. *The Modelling of Neural Networks*. Editions de Physique. Les Ulis.
- [48] Perus, M. 1996. Neuro-quantum parallelism in mind-brain and computers. *Informatica* **20**: 173-183.
- [49] Perus, M. 1997a. Neuro-quantum coherence and consciousness. *Noetic J.* **1**: 108-113.
- [50] Perus, M. 1997b. System-theoretical backgrounds of consciousness. *Informatica* **21**: 491-506.
- [51] Popper, K.R. 1963. *Conjectures and Refutations*. Routledge & Kegan Paul. London.
- [52] Popper, K.R. & J.C. Eccles. 1977. *The Self and Its Brain*. Springer. New York.
- [53] Pribram, K.H. 1971. *Languages of the Brain*. Prentice-Hall. Englewood Cliffs, NJ.

- [54] Pylyshyn, Z.W. 1986. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press. Cambridge. MA.
- [55] Reed, M. & B. Simon. 1972. *Methods of Modern Mathematical Physics. vol.1. Functional Analysis*. Academic Press. New York.
- [56] Royden, H. 1968. *Real Analysis*. Macmillan. New York.
- [57] Ryder, L.H. 1985. *Quantum Field Theory*. Cambridge Univ. Press. Cambridge.
- [58] Shepard, R.N. 1990. *Mind Sights*. Freeman. San-Francisco.
- [59] Solso, R.L. 1988. *Cognitive Psychology*. Allyn and Bacon. Boston.
- [60] Strang, G. 1976. *Linear Algebra and its Applications*. Academic. New York.
- [61] Umezawa, M., H., Matsumoto & M. Tachiki. 1982. *Thermo-Field Dynamics and Condensed States*. North-Holland. Amsterdam.
- [62] Vedenov, A.A. 1988. *Modelling Elements of Thinking*. Nauka. Moscow. *In Russian*.
- [63] Wilson, H.R. & J.D. Cowan. 1973. A Mathematical Theory of the Functional Dynamics of Cortical and Thalamic Nervous Tissue. *Kybernetick*. **13**: 55-80.
- [64] [64] Wilson, W.A. & H. Wachtel. 1974. Negative Resistance Characteristic Essential for the Maintenance of Slow Oscillations in Bursting Neurons. *Science*. **186**: 932-934.
- [65] Yosida, K. 1968. *Functional Analysis*. Springer. Berlin.
- [66] Zeleznikar, A.P. 1995. A concept of informational machine. *Cybernetica*. **38**: 7-36.
- [67] Zeleznikar, A.P. 1997. Informational Theory of Consciousness. *Informatica*. **21**: 345-369.
- [68] Ziman, J.M. 1969. *Elements of Advanced Quantum Theory*. Cambridge Univ. Press. Cambridge.

An overview of mobile agents in distributed applications: Possibilities for future enterprise systems

Seng Wai Loke
CRC for Enterprise Distributed Systems Technology
Monash University, Caulfield Campus,
PO Box 197, Caulfield East, Victoria 3145, Australia.
E-mail: swloke@dstc.monash.edu.au

Keywords: mobile agents, enterprise, distributed applications

Received: November 17, 2000

Mobile agents can be regarded as software components which can move from host to host to perform computations. Research over the past half-decade has found the mobile agent paradigm to be useful for many applications. This paper aims to show the broad applicability of mobile agents for enterprise distributed applications. It first provides an overview of mobile agent usage in three types of enterprise applications: intra-organizational, inter-organizational and customer-to-business, with focus on data management, distributed parallel processing, computer-supported collaborative work, virtual enterprises, and customer-facing systems. Then, we discuss issues and future work.

1 Introduction

Information technology has had broad impact on enterprises and commerce. The trend has been bringing the utility of computers from back-office applications to the desks of employees, and more recently to the global market. According to [2], there are three types of enterprise applications:

- **intra-organizational** which concerns applications within a business such as corporate knowledge management, and collaborative work support;
- **inter-organizational** which concerns business-to-business applications such as supply-chain management involving virtual enterprises, and knowledge sharing with shared networks called Extranets; and
- **customer-to-business** which concerns the customer's experience with the business such as transactions (e.g., purchases, orders, etc), and services.

Numerous software systems have been built in support of the above applications. With resources and operations within an enterprise being typically distributed in nature (e.g., multiple staff; multiple computers, multiple departments and sub-departments) and with homes and enterprises harnessing the Internet, the World Wide Web, Intranets, and Extranets, *distributed computing* has and will play an important role in these applications. The client-server model with message-passing of data and remote procedure calls (RPCs) has been a dominant paradigm of enterprise computing for many years. More recently, the client-middleware-server model has become very popular with the introduction of middleware component frameworks such as CORBA [58], and DCOM [69].

Meanwhile, agent technology has been exploding in popularity.¹ Numerous definitions have been given for *intelligent software agents*. In essence, as the word implies, an agent is software which performs a task on behalf of a person. In [91], an intelligent agent is defined to be software which can be described with attributes normally associated with human intelligence such as autonomy, proactiveness, reactivity, and communicative (with other agents and users) ability. Further attributes and characteristics of agent software are discussed in [57]. Stronger notions of agency employ mentalistic notions such as knowledge, beliefs and intentions.

The use of agent technology has resulted in flexible and useful software in the enterprise, for example, in business process enactment (e.g., [41]), financial portfolio management and organizational decision-making (e.g., [76]), intelligent manufacturing (e.g., [72]), product design systems (e.g., [17]), and Intranet and Internet-based knowledge retrieval and synthesis (e.g., [19, 50]). A roadmap of agent research and applications is presented in [42].

An aspect of agent systems growing in popularity in recent years is mobility [68, 67, 61]. The first language for programming *mobile agents* called TeleScript [88] was introduced in 1994. Mobile agents are software components which can move (i.e., not only is data transferred, but also code and computation state), on their own volition or invited, from one *place* to another. A *place* is a server which can receive agents, and is where an agent executes. A machine may host several places. Figure 1 shows two interconnected places with an agent at one place and two at the other. When a place is interfaced to services such as

¹For example, see the online survey on intelligent software agents at <http://www.sics.se/ps/abc/survey.html>.

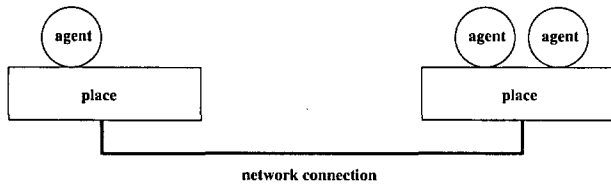


Figure 1: Two places with three agents.

databases, an agent running at that place can interact with the database. For this paper, we define a *mobile agent infrastructure* as a network of places equipped with basic facilities for mobile agents such as agent execution, agent transfer, and inter-agent communication.

There has been a proliferation of mobile agent systems in recent years such as D'Agents [26], Jinni [78], and many others in the Java language (e.g., as described in [90]). Standardization by the Object Management Group (OMG) to enable different mobile agent systems to inter-operate is being carried out [56]. The first system conforming to the existing standard has been released [4]. FIPA (Foundations of Intelligent Physical Agents)² is an organization formed to produce standards to allow software agents to inter-operate. Several FIPA-compliant agent systems have emerged. For example, Grasshopper³ is one which is both FIPA and OMG MASIF compliant. Recent work has begun to add mobility to CORBA objects [11].

Current research has involved finding killer applications for mobile agent technology. Current thoughts [44, 84] are that applications implemented with mobile agents could just as well be implemented using other techniques albeit perhaps with poorer performance. Instead of one niche application, the mobile agent paradigm has been found useful for numerous distributed applications. A challenge posed in [44] is:

“... researchers must present a set of applications and argue that the *entire set* can be implemented with much less effort (and with that effort spread across many different programming groups).”

In the above three types of enterprise applications, mobile agents have been used. For example, the mobile agent paradigm has been used or proposed for

1. **intra-organizational data management** such as distributed data retrieval and data warehousing,
2. **intra-organizational distributed parallel processing** where hosts within an Intranet are collectively utilized for distributed processing,
3. **intra- and inter-organizational computer supported collaborative work (CSCW)** such as business process enactment, team awareness, and networked device control,

4. **virtual enterprises** such as dynamic supply-chain creation and product-tracking, and
5. **customer-facing systems** which delivers business services to customers such as buying and selling, auctioning and advertising.

In this paper, we nominate enterprise applications as a set of applications in the spirit of the above quotation. We show that the mobile agent paradigm is being applied in the above five areas. Our focus is on applications. The reader is referred to [68, 67, 61, 63, 43] for an overview of mobile agent research.

2 Data Management

Data is an important asset for an organization. Technology is required to enable efficient data sharing and distance learning with up-to-date data and to permit distributed and mobile workers access to information on-demand. Mobile agents have been explored for managing data within an enterprise. We look at querying and mining distributed data resources, data warehousing, and transactions involving databases.

2.1 Querying and Mining Distributed Resources

Both querying and mining databases involve computation on data. A sophisticated query might include processing of results returned from SQL database queries. Mining of databases involves analysis of large databases to extract rules capturing observed regularities in data.

In the system described in [59], Java agents move from one database to another interacting with databases via the Java Database Connectivity (JDBC) interface. In [62], a framework for accessing databases using mobile agents and JDBC has been developed called the DBMS-Aglet Framework. A DBMS mobile agent (called a *DBMS-aglet*) carries an itinerary of database servers to visit, security certificates and queries. The DBMS-aglet is sent from a Java applet [28] to a place residing on the database server's host. On arrival, the DBMS-aglet initiates loading of appropriate JDBC drivers for communicating queries to the database server, and is then parked at the host. Additional queries can be sent from the applet to the parked DBMS-aglet via messenger aglets. Instead of a mobile agent (or DBMS-aglet) acting as mediator between the client applet and the database server, Dale *et al* [14], describes a mobile agent architecture where mobile agents interact with static resource agents which interface to databases.

Papyrus [65] is a distributed data mining system where mobile agents (built with AgentTcl) are used to transport queries to data mining services and clusters of data, and to return results. The agents arrange for the parallel execution of queries on the clusters and is capable of operations such as selection and averaging of query results. In [93], a

²<http://www.fipa.org>

³<http://www.grasshopper.de>

mobile agent contains a document classification algorithm which is sent to a remote host. Once relevant documents obtained from the classifier are sent back to the local site, the agent dies.

The use of mobile agents permits different kinds of computations to be sent to the server, and so, the server's design need not have anticipated all possible kinds of processing. In addition, using agents in these scenarios reduces network communication costs. Instead of moving huge amounts of data from the database's server to the client where computations occur, computation in the form of the agent is sent to the database's server host. Processing with the data then happens locally. Experiments comparing the use of mobile agents with client-server techniques for data mining showed improvements in performance with mobile agents [38]. Other experiments recorded in [81] showed that mobile agents can be used to reduce communication costs in distributed information filtering. Similar results in [40] showed that mobile agents are better than client-server calls for information retrieval when mobile agent code is not too large and if the wireless link over which the agent travels is error-prone. Also, in a computation involving frequent interactions with the database, moving the computation to the server minimises latencies due to network communications as demonstrated in [18].

Mobile agents are also useful for accessing databases from mobile devices. Mobile agents could be sent from a mobile device to a database server and picked up later, removing the need for maintaining the (wireless) connection between the mobile device and database.

However, it is not always the case that moving the agent to the database is more efficient, particularly when the size of the agent is large compared to the size of messages. A combination of factors are involved in optimizing agent performance such as network conditions, and the size of information transferred. An agent interacting with a network of databases might perform optimally by using a mixture of migrations, RPCs, and message-passing. For example, in experiments described in [10], a combination of message-passing and agent migration gave optimal performance compared to using message-passing or migration alone.

In [1], an extension of Java is implemented to allow programs to monitor network latency. Using the latency estimations, an agent implemented in the language can estimate the time to transfer data, and only if this time is too large, the agent moves to the data source. Hence, agents can be programmed with heuristics for optimizing network resources.

With multiple information sources, where the probability of each site containing the desired information and estimations of network latency between sites are known, a planning problem emerges of finding the sequence of sites the agent must visit in order to find the desired information in minimum time. Investigations have been carried out to find optimal planning algorithms in such scenarios [5].

2.2 Data Warehousing and Integration

In data warehousing, an information repository called the *data warehouse* stores information extracted from multiple distributed databases. The data warehouse (on one host) can be queried without going back to the original databases (on other hosts). An issue is keeping the data warehouse up-to-date whenever the databases change. Maamar [52] proposes the use of mobile agents to automate this process. The agents extract data from different sources and transfer them to the data warehouse. Extraction of data might involve data mining and so the functionality in the agents presented in the previous subsection could be utilized here.

Software agents not necessarily mobile have also been a key metaphor for software which extracts information from Web-based sources (e.g., the systems mentioned in [19, 50]). Agents can (albeit with proper engineering) retrieve (if necessary, after moving to resources) information from distributed heterogeneous information sources including structured databases and unstructured Web documents, and synthesize a report for the user, thereby providing an integrated front-end to heterogeneous information sources.

Beyond information retrieval is the management of knowledge. The proposed Distributed Knowledge Networks (DKNs) [32] includes the use of mobile agents to move among distributed information resources to access, analyze and synthesize knowledge to support distributed problem-solving and decision-making under real-time constraints. Due to these constraints and the effort required to represent knowledge from different sources in a consistent format, DKNs are more feasible on Intranet-scale than on Internet-scale.

2.3 Transactions on Distributed Resources

Besides simply sending an agent to perform a task at a remote database, a mobile agent could be used to implement an entire transaction involving several distributed resources.

In [16], an architecture is proposed where places are extended with transaction support. An agent moves from one place to another and implements a transaction by performing the tasks of the transaction at the places it visits. An advantage of using mobile agents in such transactions is support for distributed environments and mobile devices. The latter is because mobile agents can perform transaction tasks asynchronously without maintaining a continuous connection between the mobile device and task places.

Instead of using one agent to implement a transaction, multiple agents are employed in [30]. The advantage over the single agent approach is the ability to react to dynamic changes in the environment. For example, in the single agent approach, while the agent is moving from place to place, changes might occur in previously visited places which would cause the transaction to later abort. In the multiple agent approach, an originator agent, while going

from place to place, would leave monitoring agents at previously visited places. Changes are detected and forwarded by the monitoring agent to the originator agent which could then take immediate action. In response to changes, the monitoring agent might also move to another place (or resource) and inform the originator agent. When the originator agent reaches the last place in its itinerary, a two-phase commit protocol involving all the agents is used to finalize the transaction.

These mobile agent based transactions are usable not only within the enterprise but also for electronic transactions between customers and businesses. For example, using the multiple agent approach, a transaction can be implemented to buy all components of a computer or none, where each component is bought from a different place and the stock of components can vary during the agents movements.

3 Distributed Parallel Processing

Mobile agents are also being investigated for distributed parallel processing (e.g., [23, 89]). This is useful for utilizing idle CPU cycles, particularly in organizations with a large number of networked computers. For example, at night when CPU load in the company's computers are low, mobile agents can move to these hosts to perform computations. Tasks encapsulating both code and data can be added to a tasks pool for execution by "worker" mobile agents.

An organization might also rent out idle CPU time but not without proper protection of its resources and mechanisms for charging for resources used (e.g., using software described in [66]). Other organizations use these resources by sending jobs encapsulated in mobile agents. Such resource rental would have to be governed by appropriate service level agreements.

4 CSCW

Two areas to which mobile agents have been applied in CSCW are workflow and team awareness. We also discuss, as a CSCW application, multi-user control of networked devices via mobile agents.

4.1 Workflow

A *workflow* (often called a business process) consists of a set of activities (or tasks) which need to be executed in some controlled order to realize a business purpose. Workflow management systems aim to automate and streamline business process enactment. For example, opening a bank account involves information and control flow among entities (e.g., customer, bank representative, and supervisor) within an organization, where each involved entity must perform a specific task. Static (or non-mobile) agents, each performing a specific role or representing an entity, and communicating by message-passing, have been extensively

used to realize business processes (e.g., the systems described in [41, 79, 94]).

A different approach to workflow utilizes mobile agents (e.g., the DartFlow [9] and Autopilot [22] software). The key difference from the static agents approach is the transfer of business logic from entities into the mobile agent. For instance, in the example about opening a bank account in [9], a mobile agent encapsulates customer details and knowledge about how the details need to be processed for an account to be opened. The agent moves from place to place performing some task at each place towards the goal of opening an account. These places might be interfaced to databases or the agent might present a graphical user interface to interact with a person.

The places representing entities visited by the mobile agents become passive compared to the static agents representing entities in the static agents approach. Also, the path taken by an agent might be determined at runtime from interactions at places or by querying a service broker. Hence, each request for opening an account would have its own thread of control providing flexibility and specialization. The agent's movements correspond naturally to the flow of work (or sequence of tasks).

For a robust and efficient workflow system, runtime changes to the agent's itinerary, and congestion due to a large number of agents simultaneously working, must be managed. These issues have been dealt with in part in Autopilot.

In [92], a tool for distributed project management is developed using mobile and static agents. Project activities are represented by static agents and mobile *service agents* can be sent along critical paths of static agents to retrieve activity information and to calculate overall project duration. Agents can clone themselves to calculate the overall duration of parallel activities. A resource agent represents a resource (e.g., material, machines) and can also move when the resource is transferred to another location. In the tool, mobility is a useful attribute for agents representing movable resources. When required, cloning and code mobility enables parallel and distributed processing to happen at specified locations.

Lightweight CSCW systems involving simple coordination among multiple entities can easily be built with mobile agent systems. For example, one can build meeting scheduler agents which move among potential participants and interact with them [86]. These agents can tailor their user interfaces according to information obtained from previous participants.

A recent idea integrates mobile agents with distributed event notification services aiming to combine their advantages. For example, in the PROSYT process support system [13], entities subscribe to events they are interested in by sending a message to a notification server, and receive notifications about these events when other entities publish event notifications. The event system is used to coordinate business processes and cooperation among entities. PROSYT utilizes *reactive objects* which can migrate be-

tween hosts in response to event notifications.

A notable feature of PROSYT's event system (as well as others such as Elvin [71]) is that event publishers do not need to know the identity or location of event notification subscribers, i.e. we have decoupling of event publishers from subscribers and *undirected messaging*. The notification server acts as an intermediary between event publishers and subscribers maintaining the locations of subscribers and forwarding notifications to the right places. The decoupling is useful when the locations of subscribers are continually changing (e.g., when the subscribers are mobile agents). Although more work is still required to investigate the advantages of such an integrated approach, with the likely presence of event systems in future office environments (as indicated by the success of Elvin in CSCW [21] and the increasing interest in event-based systems [37]) and the use of mobile agents in business processes, such integrations could become more prevalent.

Besides intra-organizational workflow, mobile agents have also been explored for inter-organizational workflow. For example, based on the Common Open Service Market (COSM) architecture [55], mobile agents have been proposed for coordinating workflow across organizational boundaries. The mobile agent approach suits the lack of a central coordination mechanism, and permits applications (encapsulated in mobile agents) to move into organizations which do not have those applications. The latter is particularly useful when the inter-organizational cooperation is transient. What is required is only that the cooperating organizations provide a basic infrastructure for the execution of the mobile agents, rather than permanent changes to the organizations' systems for temporary inter-operation. Such systems have to be carefully engineered so that agents can perform tasks as needed. For instance, the agent may need to access different kinds of databases in different organizations. Such inter-organizational workflows are essential activities of virtual enterprises discussed in the next section.

We conclude this subsection by listing some potential benefits of mobile agents for workflow:

- *workflow extensibility and adaptability*: These attributes refer to a workflow system which permits *component-based extensibility*, where new places representing new business entities or interfaced to new databases (or software) can be added to the existing network of places, new workflows can be introduced by creating new agents at run-time to extend workflows on-the-fly, and agents can uptake or replace its components at run-time to cope with task changes (e.g., to interact with different databases).
- *workflow integration*: Two or more workflows which are simultaneously running (e.g. each workflow happening in different organizations) can be integrated via communication between the agents enacting the workflows.

- *workflow lifecycle monitoring*: A workflow being enacted by an agent could be monitored by tracking the agent's activities.
- *workflow for mobile device users*: Mobile agents can be launched from mobile devices into a fixed network (or to other mobile devices) to perform tasks, without requiring the mobile device to be continuously connected, and can reduce network traffic over narrowband wireless networks by moving computation to databases instead of moving huge amounts of data to where computations are carried out.
- *remote installation of workflow components*: In inter-organizational workflows where there is heterogeneous computational infrastructure across different organizations, lack of central management and high costs of integrating systems across organizations, mobile agents support ad-hoc workflows by being moved as software components to places as needed. Further benefits of mobile agents for inter-organizational workflow are detailed in [55].

Despite this conceptual advantage, a more rigorous comparison is needed between workflows implemented using mobile agents and the corresponding static agents implementation.

In general, we can envision more complex inter-organizational workflows where the agent's task at an organization is only part of the organization's internal workflow. Other tasks within an internal workflow might be carried out by static agents, mobile agents, or by non-agent workflow systems, i.e., we can imagine each organization as having their own workflow system but integrated with other organizations' workflow systems via the tasks performed by mobile agents moving among places. Tools will, however, be needed to model and build such large scale workflows. For example, using Petri-Nets to model mobile agents for coordinating inter-organizational workflows is proposed in [49].

4.2 Team Awareness

Mobile agents can be used to support team awareness in CSCW, acting as observers and reporters of events within a distributed environment. For example, with the MOLE Office tool [7, 8] users (or team members) can send remote users (also running the MOLE Office tool) mobile observer agents which stay at the remote ends to listen for and filter events. The mobile agent sends back relevant events to the originating user. This approach has a number of advantages including reducing network traffic since not all events are transmitted and filtering takes place at the source end, and flexibility, since agents can be sent to observe newly connected users or retracted from disconnecting users. Also, users can launch observer agents from their mobile devices into a fixed network and collect the agents when they next reconnect.

4.3 Networked Device Interfaces

Mobile agents can be used as intermediaries between users and devices (such as a slide projector or a printer) on a network. Mobile agents can be downloaded to a host on demand much like Java applets. A person who wants to control a device but does not have the software to do so can download an agent which knows the protocol for communicating with the device. The agent provides a graphical user interface for the person on one hand and communicates with the device on the other. Two advantages are:

- **software on-demand:** the user can download agents to control devices whenever required, and so, does not need pre-installed software, and
- **passing of control and coordination:** passing of control from one person to another is easily carried out by passing the agent; the agent maintains state which is also transferred. Similarly, when the agent moves from one place to another, control is shifted from one party to another, and so the agent's movement can be used for control coordination among parties.

Similar mechanisms have been used in [31] where a user can download to his/her mobile device the user interface for interacting with devices in a seminar room, and in the Jini framework [75], where a device wishing to use another device downloads a *service object* from a registry.

5 Virtual Enterprises

Virtual enterprises enable a business organization to outsource or utilize services of other organizations without the need for expanding. This collaboration of organizations might only be temporary and once-off for the production of a new product or service. The set of collaborating organizations (or parts thereof) is called a *virtual enterprise* [80].

Key issues for enabling the virtual enterprise outlined in [80] and [83] include inter-operability of systems across organizational boundaries, and the ease and efficiency of setting up a virtual enterprise. In [80], a global infrastructure spanning organizational boundaries is proposed as a key technology for virtual enterprises. With recent work on applying mobile agents to virtual enterprise applications [55, 64, 77, 6, 60, 15], the mobile agent infrastructure might become one such infrastructure.

Three research prototypes [6, 60, 15] have been built which use mobile agents to enact a supply-chain in manufacturing. In these systems, the buyer places an order for an item by interacting with a mobile agent directly or indirectly via another agent. The mobile agent then moves to the suppliers' hosts with the order. If the order cannot be satisfied internally, another level of the supply-chain would be created when the agents move to other suppliers' hosts. Similar to the workflow agents in Section 4.1, the agents encapsulate both data and business logic. The agents might initiate actual physical processes to create a product. These

systems have all been prototyped using IBM's Aglet Java mobile agent toolkit [46, 36].

Conceptual parallels between physical production processes (e.g., assembly of a PC, garment, or food) and information production processes (e.g., workflow) are being examined with the aim of using mobile agents to coordinate and track the process of product creation [24, 77].

In [12], an architecture is proposed for establishing temporary inter-organizational workflows using mobile agents (called *adlets*). An adlet might roam an area to gather information and connect related workflows. Adlet based workflows are similar to the inter-organizational workflows described in Section 4.1 but with greater emphasis on the evolving, temporary, and ad hoc nature of virtual enterprises. For example, a workflow can be expanded on a need-to basis by dynamically spawning new adlets. Plans for a prototype implementation using IBM's Aglet toolkit are reported.

6 Customer-Facing Systems

Agents have been extensively investigated for various stages in consumer buying behaviour [29, 51]. These stages include brokering to determine what to buy and who to buy from, negotiation for terms of a transaction, purchasing and delivery of products.

More recently, mobile agents are being explored for electronic marketplaces and electronic auctions:

- **electronic marketplace for agents:** an electronic marketplace functions as a brokering environment (also called a *digital bazaar* [53]) where customers send agents to find services or sellers of desired goods. An example is *e-Marketplace* [35] a middleware for electronic marketplace supporting three kinds of agents: customer/buyer agents, shop/seller agents, and advertising agents which monitor activities at an e-Marketplace server to seek out customer agents. An electronic marketplace can be formed by a federation of e-Marketplace servers. The precursor to e-Marketplace is the TabiCan Web site (<http://www.tabican.ne.jp/index.html>) which supports mobile agents for determining travel arrangements.
- **electronic auctions:** an electronic auction house permits customers to bid for advertised goods. An electronic auction house could be part of an electronic marketplace. Although several electronic auction houses exist on the Web, the *Nomad* system which works with the *eAuctionHouse* servers appears to be the first use of mobile agents in electronic auctions [33]. The mobile agents are used for bidding, information collection, price distribution learning, and setting up auctions.

The advantages of mobility in these consumer environments are similar to that for databases such as support for mobile devices, and reducing network traffic, especially in

the case where the cost of transferring the agent is less than the amount of messages which need to be transferred. Mobile agents are an attractive means for monitoring an auction or the auction house and informing the consumer of new bids or new goods instead of the consumer repeatedly polling the auction house.

To support electronic commerce applications, mobile agents need mechanisms for negotiation, accountability, and reliability.

If the agents are to perform transactions on their own, they must have negotiation capabilities. In [82], an architecture for plugging negotiation protocols into mobile agents is proposed. The aim is that the agents dynamically take up appropriate negotiation protocols as needed. The use of such mobile agents for negotiating contracts among businesses or customers is proposed in [27], where an agent carrying a contract moves from one collaborating party to another, interacting with the parties, and recording information obtained from the parties. Four features of this approach are (1) the agent can check consistency of the contract's contents as it is being modified by the different parties, (2) no concurrency control on the contract is necessary since the agent visits one party at a time, (3) on arrival the agent shows a user interface customized to the contract's current contents, and (4) the agents would migrate over international boundaries at the appropriate working hours.

Accountability and reliability mechanisms are important for electronic commerce transactions especially since mobile agents can get lost due to network or host errors, or be attacked by remote hosts. Several such mechanisms have been developed. For example, in [87], a trust service is provided between customer and business hosts. When an agent is transferred from the customer host to a business host and back, messages are sent to the trust service to create a history log. Also, in [74], an aspect of reliability dealt with is the exactly once property, i.e. that the agent moves to a place and performs a task exactly once. A protocol is presented to guarantee this property. This property is particularly important for electronic commerce transactions, for example, when the agent is making a flight or hotel reservation.

7 Conclusions and Speculations on Future Enterprise Applications

We have described a set of enterprise applications for mobile agents: data management, distributed parallel processing, CSCW, virtual enterprises, and customer-facing systems. This set of applications spans all three types of enterprise applications: intra-organizational, inter-organizational and customer-to-business. While still under research, further developments of the ideas mentioned could lead to more robust industrial strength systems.

We have also noted how the benefits mobile agents outlined in [47] such as reducing network traffic, encapsulating protocols, distributed software installation, and asyn-

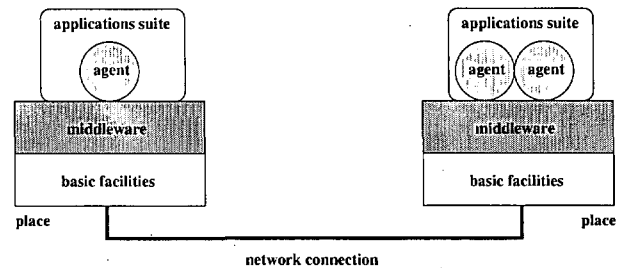


Figure 2: Three layered conceptual architecture illustrated with two places.

chronous and autonomous computation are being exploited in these applications. Although not all components of enterprise software will be mobile, mobile agents will likely play an integral role in many distributed applications, particularly to enable flexible software systems for dynamic and/or virtual enterprises.

In the final three subsections, we discuss possible developments and ongoing issues in using mobile agents in the enterprise.

7.1 Enterprise-Wide Agent Platform and Application Suites

To use mobile agents for the set of applications we have surveyed, we can envision a mobile agent based applications suite for the enterprise, in the spirit of Microsoft Office and Lotus eSuite.⁴ Such a mobile agent based office tool will enable its user to send agents, receive agents, and interact with agents (e.g., provide input to agents). The tool functions as an interface to a myriad of “intelligent mobile assistants”. The mobile agent the user interacts with might be an interface to the corporate’s distributed databases, an agent enacting a business process, an agent organizing a meeting, an agent for communicating with a device, an advertising agent, an observer agent for team awareness, or an agent managing a supply chain.

The applications suite could be supported by a software infrastructure whose conceptual architecture spreading over two places is illustrated by Figure 2. We envision this architecture for large enterprises to spread over tens or even hundreds of places. The architecture consists of three layers:

1. **a common basic mobile agent infrastructure for the entire enterprise:** different distributed applications would use the same mobile agent infrastructure which provides the basic facilities for all agents. Although the agents have to be carefully engineered for different applications, the agents would use the common infrastructure. This would enable the same agent

⁴Lotus eSuite consists of eSuite applets each either an application such as spreadsheet or word processor, or a building block for applications. See <http://www.esuite.lotus.com/>.

to utilize different resources and different agents to utilize the same resources.

For instance, if a database server and software interfacing with manufacturing equipment are located within the same infrastructure, an agent enacting a business process could visit the database *and* communicate with manufacturing equipment. Other agents used in a virtual enterprise or managing data warehouses could run on the same infrastructure. It would not be cost effective if three different mobile agent infrastructures have to be set up: one for database access, one for workflow, and another for a virtual enterprise, especially when agents access common resources.

Existing mobile agent toolkits [90, 63] already provide the basic infrastructure for mobile agents.

2. **mobile agent middleware:** The mobile agent infrastructure might need to be augmented with additional software components to support specific applications. For example, to support transactions involving several places, each of these places would need to be augmented with a component for implementing transactions [16]. To support an electronic marketplace over a collection of places, each designated place would be augmented with components for hosting marketplace agents. Places might be interfaced to corporate databases or other existing corporate software. Places which function as gateways between organizations would need to be appropriately augmented (e.g., with directories of places within each organization and security software). A place might be augmented with components for different applications, and it is not necessary for all places to have the same augmentations (unless required by the applications). We call such components augmenting places *mobile agent middleware*.

Such middleware (e.g., transaction support) would be useful for more than one application (e.g., database transactions and workflow), and so could reduce implementation effort for applications through software reuse. Mobile agents could be used to install such middleware at the appropriate places.

Further work will be needed to identify and implement appropriate mobile agent middleware.

3. **a mobile agent based enterprise applications suite:** the mobile agent based office tool sits at the top of the conceptual architecture.

Apart from an interface to agents, such a tool could also be an environment for developing applications: agents could be created by different members of the organization and injected into the mobile agent infrastructure. Tools will be needed for constructing or configuring application agents.

Application sub-suites could also be explored. For example, an enterprise database application sub-suite

would consist of a collection of mobile agents with database-related functionality, similar to the library of database aglets proposed in [62].

There is also opportunity for reusing agents. For example, an agent which knows about databases could be used by a workflow agent.

7.2 Mobile Agents for Nomadic Users and Mobile Platforms

Finally, we note that as the workforce is not only becoming increasingly distributed but also mobile (e.g., as argued in [54]), we can expect mobile agents, with their advantages for asynchronous and autonomous computation, to play an important role in enterprises. Intensive research has already begun on mobile agents for mobile platforms (e.g., [25, 39, 70, 45, 48]) and to support nomadic users that move from one (not necessarily mobile) device to another. The Magnitude⁵ project aims to investigate mobile agents as intermediaries between nomadic users and agents on the fixed infrastructure; negotiation is investigated to control mobile agent access to resources and for cooperation with other agents. The LEAP (Lightweight Extensible Agent Platform) project⁶ aims to develop an agent platform which is lightweight and executable on small devices such as PDAs and phones.

7.3 On-going Technical Issues

We have seen in this paper a number of research proposals and prototypes where mobile agents are used, and there are many industry-released mobile agent toolkits.⁷ However, there is hardly any commercial-off-the-shelf (or industrial strength) mobile agent based enterprise applications.

Indeed, mobile agent technology is still maturing. Research is underway to tackle existing technical hurdles for widespread adoption of mobile agent technology [44]. We briefly mention four hurdles: security, execution performance, standardization, and robustness.

Security and resource control are crucial when agents executing on a local machine come from foreign hosts and when agents cross organizational boundaries. Security issues must be dealt with to encourage the uptake of mobile agent systems and is more difficult for the open Internet than for an Intranet within a firewall. Problems of protecting hosts from agents, agents from other agents, and agents from hosts are being tackled (e.g., [85]), and economic models for agents to purchase resources from sites are being explored [3].

Execution performance is important (e.g., for parallel processing) but often sacrificed in order to achieve portability via code interpretation. Just-in-time compilation is one way towards better performance.

⁵See <http://www.ecs.soton.ac.uk/~lavm/magnitude/case.html>

⁶See <http://leap.crm-paris.com>

⁷See <http://mole.informatik.uni-stuttgart.de/mal/preview/preview.html>

Standardization is important to enable the mobile agent systems of different organizations to inter-operate. Apart from OMG's standardization efforts, standards efforts on inter-agent communication languages (e.g., KQML) could be exploited for inter-operability between agents, and between agents and mobile agent middleware [20].

Adding robustness involves fault-tolerance schemes (e.g., coping with hosts crashing) and is key for trust and acceptance of the technology in companies.

Solutions to these hurdles could be mobile agent middleware or be incorporated into the underlying agent infrastructure.

Work aiming to tightly integrate mobile agent applications into the workplace has already begun with the integration of mobile agent places with Web browsers (e.g., the Fiji applet [34] and the PESOS browser [73]). The browser then becomes a front-end to both mobile agent applications and the Web.

Mobile agents of the future will also exhibit more intelligent behaviour building on AI research. We speculate that, then, mobility will not be seen as a luxury but as a readily available option to be taken on the basis of the agent's rational decision-making.

Acknowledgements

The work reported in this paper has been funded in part by the Co-operative Research Centre Program through the Department of Industry, Science & Tourism of the Commonwealth Government of Australia.

References

- [1] A. Acharya, M. Ranganathan, and J. Saltz. Sumatra: A Language for Resource-Aware Mobile Programs. *Lecture Notes in Computer Science*, 1222, 1997.
- [2] N.R. Adam, O. Dogramaci, A. Gangopadhyay, and Y. Yesha. *Electronic Commerce: Technical, Legal and Business Issues*. P T R Prentice-Hall, Englewood Cliffs, NJ 07632, USA, 1999.
- [3] J. Bredin, D. Kotz, and D. Rus. Economic Markets as a Means of Open Mobile Agent Systems. In *Proceedings of the Workshop on Mobile Agents in the Context of Competition and Cooperation at Autonomous Agents '99*, Seattle, U.S.A., May 1999. Available from <<http://mobility.lboro.ac.uk/MAC3/>>.
- [4] M. Breugst, I. Busse, S. Covaci, and T. Magedanz. Grasshopper – A Mobile Agent Platform for IN Based Service Environments. In *Proceedings of IEEE IN Workshop 1998*; pages 279–290, Bordeaux, France, May 1998. Available at <<http://www.ikv.de/download/grasshopper/IEEE-INWS-98.pdf>>.
- [5] B. Brewington, R. Gray, K. Moizumi, D. Kotz, G. Cybenko, and D. Rus. Mobile Agents in Distributed Information Retrieval. In M. Klusch, editor, *Intelligent Information Agents*, chapter 15. Springer-Verlag, 1999.
- [6] D. Brugali, G. Menga, and S. Galarraga. Inter-Company Supply Chain Integration via Mobile Agents. In *The Globalization of Manufacturing in the Digital Communications Era of the 21st Century: Innovation, Agility and the Virtual Enterprise*. Kluwer Academic Pub., 1999. Available at <<http://www.cim.polito.it/Articles/Art-Agents/98PROLAMAT.ps>>.
- [7] C. Burger. Team Awareness with Mobile Agents in Mobile Environments. In *Proceedings of the 7th International Conference on Computer Communications and Networks (IC3N'98)*, pages 45–49, October 1998.
- [8] C. Burger. Festival - MOLE-Office. 1999. Web page at <<http://www.informatik.uni-stuttgart.de/ipvr/vs/projekte/Festival/MOLE-Office.engl.html>>.
- [9] T. Cai, P.A. Gloor, and S. Nog. DartFlow: A Workflow Management System on the Web Using Transportable Agents. Technical Report PCS-TR96-283, Department of Computer Science, Dartmouth College, May 1996. Available at <<ftp://ftp.cs.dartmouth.edu/TR/TR96-283.ps.Z>>.
- [10] T. Chia and S. Kannapan. Strategically Mobile Agents. In K. Rothermel and R. Popescu-Zeletin, editors, *Mobile Agents. Proceedings of the First International Workshop on Mobile Agents 97 (MA'97)*, Lecture Notes in Computer Science No. 1219, Berlin, Germany, April 1997. Springer.
- [11] S. Choy and M. Breugst. A CORBA Environment Supporting Mobile Objects. In *Proceedings of the 6th International Conference on Intelligence in Services and Networks*, Barcelona, Spain, April 1999. Available at <<http://www.italtel.it/drsc/marine/corbatmat-isn99.pdf>>.
- [12] P.K. Chrysanthis, T.F. Znati, S. Banerjee, and S.K. Chang. Establishing Virtual Enterprises by Means of Mobile Agents. In *Proceedings of the Research Issues in Data Engineering Workshop*, Sydney, Australia, March 1999. Available at <ftp://ftp.cs.pitt.edu/panos/MDBS/ride_99.ps.gz>.
- [13] G. Cugola and C. Ghezzi. The Design and Implementation of PROSYT: an Experience in Developing an Event-based, Mobile Application. In *Proceedings of the 13th IEEE International Conference on Automated Software Engineering (ASE'98)*, October 1998.
- [14] J. Dale and D.C. DeRoure. A Mobile Agent Architecture for Distributed Information Management.

- In *Proceedings of the International Workshop on the Virtual Multicomputer*, March 1997. Available at <<http://www.mmrg.ecs.soton.ac.uk/publications/archive/dale1997b/>>.
- [15] P. Dasgupta, N. Narasimhan, L.E. Moser, and P.M. Melliar-Smith. MAgNET: Mobile Agents for Networked Electronic Trading. *IEEE Transactions on Knowledge and Data Engineering, Special Issue on Web Applications (to appear)*, 1999. Available at <<http://alpha.ece.ucsb.edu/~pdg/research/papers/MAgNEThtml/index.html>>.
- [16] F.M. de Assis Silva and S. Krause. A Distributed Transaction Model Based on Mobile Agents. In K. Rothermel and R. Popescu-Zeletin, editors, *Mobile Agents. Proceedings of the First International Workshop on Mobile Agents 97 (MA'97)*, LNCS 1219, Berlin, Germany, April 1997. Springer-Verlag.
- [17] A. Deshmukh, A. Krothapalli, T. Middelkoop, and C. Smith. Multi-Agent Design Architecture for Integrated Design Systems. *AIAA Journal of Aircraft*, 1999. Revised version to appear. Available at <http://farm.ecs.umass.edu/~mtim/papers/1999/aiaa_aircraft>.
- [18] H. Detmold and M.J. Oudshoorn. Using Mobile Objects as Ambassadors to Minimize Latency in World-wide Distributed Systems. Technical Report 97-05, Department of Computer Science, The University of Adelaide, 1997. Available at <<http://www.cs.adelaide.edu.au/users/michael/papers/tr97-05.ps>>.
- [19] O. Etzioni. Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web. In *Proceedings of the 13th National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence Conference*, pages 1322–1326, Menlo Park, U.S.A., August 1996. AAAI Press / MIT Press. Available at <<ftp://ftp.cs.washington.edu/pub/etzioni/softbots/a96.ps.gz>>.
- [20] T. Finin, Y. Labrou, and Y. Peng. Mobile Agents Can Benefit from Standards Efforts on Interagent Communication. *IEEE Communications*, 36(7):50–56, July 1998.
- [21] G. Fitzpatrick, T. Mansfield, S. Kaplan, D. Arnold, T. Phelps, and B. Segall. Instrumenting and Augmenting the Workaday World with a Generic Notification Service called Elvin. In *Submitted to ECSCW'99*, 1999. Available at <<http://www.dstc.edu.au/Elvin/doc/papers/ecscw99/ecscw99.pdf>>.
- [22] S.S. Foster, D. Moore, and B.A. Nebesh and M.J. Flester. Control and Management in a Mobile Agent Workflow Architecture. In *Proceedings of the 3rd International Conference on Autonomous Agents (Agents'99)*, Seattle, U.S.A., May 1999.
- [23] R. Ghanea-Hercock, J.C. Collis; and D.T. Ndumu. Heterogeneous Mobile Agents for Distributed Processing. In *Proceedings of the Workshop on Agent-based High Performance Computing at Agents'99*, May 1999. Available at <<http://www.cs.cf.ac.uk/User/O.F.Rana/agents99/papers/okpapers/order/collis.ps>>.
- [24] J.B.M. Goossenaerts, A.T.M. Aerts, and D.K. Hammer and. Merging a Knowledge Systematization Framework with a Mobile Agent Architecture. In *Proceedings of the 3rd Conference on the Design of Information Infrastructure Systems for Manufacturing*, Texas, U.S.A., May 1998. Available at <http://www.wis.win.tue.nl/~wsinatma/Agents/ks_ag.ps>.
- [25] Robert S. Gray, David Kotz, Saurab Nog, Daniela Rus, and George Cybenko. Mobile Agents for Mobile Computing. Technical Report TR96-285, Department of Computer Science, Dartmouth College, May 1996. Available at <<ftp://ftp.cs.dartmouth.edu/TR/TR96-285.ps.Z>>.
- [26] R.S. Gray. Agent Tcl: A Flexible and Secure Mobile Agent System. In M. Diekhans and M. Roseman, editors, *Fourth Annual Tcl/Tk Workshop (TCL 96)*, pages 9–23, Monterey, CA, July 1996. Available at <<http://www.cs.dartmouth.edu/~agent/papers/tcl96.ps.Z>>.
- [27] F. Griffel, T. Tu, M. Mönke, M. Merz, W. Lamersdorf, and M. Mira da Silva. Electronic Contract Negotiation as an Application Niche for Mobile Agents. In *1st International Workshop on Enterprise Distributed Object Computing*, 1997.
- [28] D. Gulbrandsen, K. Rawlings, and J. December. *Creating Web Applets with Java*. Sams Publishing, 1996.
- [29] R. Guttman, A. Moukas, and P. Maes. Agent-Mediated Electronic Commerce: A Survey. *Knowledge Engineering Review*, 13(2), June 1998. Available at <<http://ecommerce.media.mit.edu/papers/ker98.pdf>>.
- [30] A. Buchmann H. Vogler. Using Multiple Mobile Agents for Distributed Transactions. In *Proceedings of the 3rd IFICIS Conference on Cooperative Information Systems (CoopIS'98)*, New York, USA, August 1998. Available at <<http://www.ito.tu-darmstadt.de/publs/papers/coopis98.ps.gz>>.
- [31] T.D. Hodes, R.H. Katz, E. Servan-Schreiber, and L. Rowe. Composable Ad-hoc Mobile Services for Universal Interaction. In *Proceedings of the 3rd ACM/IEEE International Conference on Mobile*

- Computing and Networking*, 1997. Available at <<http://daedalus.cs.berkeley.edu/publications/services-mobicom97.ps.gz>>.
- [32] V. Honavar, L. Miller, and J. Wong. Distributed Knowledge Networks. In *Proceedings of the IEEE Information Technology Conference*, Syracuse, N.Y., U.S.A., 1998. Available at <<http://www.cs.iastate.edu/~honavar/Papers/it98-dkn.ps>>.
- [33] Q. Huai and T. Sandholm. Mobile Agents in an Electronic Auction House. In *Proceedings of the Workshop on Mobile Agents in the Context of Competition and Cooperation at Autonomous Agents '99*, Seattle, U.S.A., May 1999. Available from <<http://mobility.lboro.ac.uk/MAC3/>>.
- [34] IBM. Fiji - Running Aglets in Web Pages. (updated) 1999. Web page at <<http://www.trl.ibm.co.jp/aglets/infrastructure/fiji/fiji.html>>.
- [35] IBM. IBM Aglets Software Development Kit - e-Marketplace. (updated) 1999. Web page at <<http://www.trl.ibm.co.jp/aglets/emplace/emplace.html>>.
- [36] IBM. IBM Aglets Software Development Kit - Home Page. (updated) 1999. Web page at <<http://www.trl.ibm.co.jp/aglets/index.html>>.
- [37] Irvine Research Unit in Software, editor. *Workshop on Internet Scale Event Notification (WISEN'98)*, July 1998. Available at <<http://www.ics.uci.edu/IRUS/twist/wisen98/>>.
- [38] L. Ismail and D. Hagimont. A Performance Evaluation of the Mobile Agent Paradigm. In *Proceedings of the International Conference on Object-oriented Programming, Systems and Applications (OOPSLA'99)*, 1999. Available at <<http://sirac.imag.fr/Interne/doc/sirac/publications/INTERNE/Soumis/soumis-oopsla-agents.ps.gz>>.
- [39] K. Jacobsen and D. Johansen. Mobile Software on Mobile Hardware - Experiences with TACOMA on PDAs. Technical Report 97-32, Department of Computer Science, University of Tromso, December 1997. Available at <<http://www.cs.uit.no/Lokalt/Rapporter/Reports/9732.htm>>.
- [40] R. Jain, F. Anjum, and A. Umar. A Comparison of Mobile Agent and Client-Server Paradigms for Information Retrieval Tasks in Virtual Enterprises. In *Proceedings of the 2000 Academia/Industry Working Conference on Research Challenges (AIWoRC 2000) on 'Next Generation Enterprises: Virtual Organizations and Mobile/Pervasive Technologies'*, Buffalo, U.S.A., April 2000. IEEE Press.
- [41] N.R. Jennings, P. Faratin, T.J. Norman, P. O'Brien, and B. Odgers. Autonomous Agents for Business Process Management. *Journal of Applied Artificial Intelligence*, 1999. Available at <<ftp://ftp.elec.qmw.ac.uk/pub/isag/distributed-ai/publications/aaaj991.ps.gz>>.
- [42] N.R. Jennings, K. Sycara, and M. Wooldridge. A Roadmap of Agent Research and Development. *Autonomous Agents and Multi-Agent Systems*, 1:7-38, 1998.
- [43] N. Karnik and A. Tripathi. Design Issues in Mobile Agent Programming Systems. *IEEE Concurrency*, pages 52-61, July-September 1998. Available at <<http://www.cs.umn.edu/Ajanta/papers/ieeconc.ps>>.
- [44] D. Kotz and R. Gray. Mobile Code: The Future of the Internet. In *Proceedings of the Workshop on Mobile Agents in the Context of Competition and Cooperation at Autonomous Agents '99*, Seattle, U.S.A., May 1999. Available from <<http://mobility.lboro.ac.uk/MAC3/>>.
- [45] E. Kovacs, K. Rohrlé, and M. Reich. Integrating Mobile Agents into the Mobile Middleware. In *Proceedings of the 2nd International Workshop on Mobile Agents (MA'98), Lecture Notes in Computer Science 1477*, pages 124-135, September 1998.
- [46] D.B. Lange and M. Oshima. *Programming and Deploying Java Mobile Agents with Aglets*. Addison-Wesley, 1998.
- [47] D.B. Lange and M. Oshima. Seven Good Reasons for Mobile Agents. *CACM*, 42(3):88-89, March 1999.
- [48] S. Lazar and D. Sidhu. Laptop Docking Support for Mobile Agent Based Applications. Technical report, Maryland Center for Telecommunications Research, Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, 1998. Available at <http://discovery.mctr.umbc.edu/Papers/Laptop_Support.pdf>.
- [49] S.W. Loke and S. Ling. Mobile Agent Itineraries and Workflow Nets for Analysis and Enactment of Distributed Business Processes. In *Proceedings of the International Symposium on Multi-Agents and Mobile Agents in Virtual Organizations and E-Commerce*, pages 459-466, Wollongong, Australia, December 2000. ICSC Academic Press.
- [50] S.W. Loke, L. Sterling, and L. Sonenberg. A Knowledge-based Approach to Domain-specialized Information Agents. *Journal of Internet Research: Electronic Networking Applications and Policy*, 9(2):140-152, 1999.

- [51] M. Ma. Agents in E-commerce. *Communications of the ACM*, 42(3):78–80, March 1999.
- [52] Z. Maamar. A Data Warehousing Environment Based on Software and Mobile Agents. In *Proceedings of the 9th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management*, Newark, Delaware, U.S.A., October 1998.
- [53] P. Maes, R.H. Guttman, and A.G. Moukas. Agents That Buy and Sell. *Communications of the ACM*, 42(3):81–91, March 1999.
- [54] T. Makimoto and D. Manners. *Digital Nomad*. John Wiley & Sons, 1997.
- [55] M. Merz, B. Liberman, and W. Lamersdorf. Using Mobile Agents to Support Interorganizational Workflow Management. *Applied Artificial Intelligence*, 11(6):551–572, 1997.
- [56] D. Milojicic, M. Breugst, I. Busse, J. Campbell, S. Covaci, B. Friedman, K. Kosaka, D. Lange, K. Ono, M. Oshima, C. Tham, S. Virdhagriswaran, and J. White. MASIF: The OMG Mobile Agent System Interoperability Facility. In *Proceedings of the 2nd International Workshop on Mobile Agents (MA'98), Lecture Notes in Computer Science 1477*, pages 50–67, September 1998.
- [57] H.S. Nwana. Software Agents: An Overview. *Knowledge Engineering Review*, 11(3):205–244, September 1996. Available at <<http://www.cs.umbc.edu/agents/introduction/ao/>>.
- [58] R. Orfali, D. Harkey, and J. Edwards. *Instant CORBA*. Addison Wesley, Bonn, 1998.
- [59] T. Papaioannou and J. Edwards. Mobile Agent Technology Enabling the Virtual Enterprise: A Pattern for Database Query. In *Proceedings of the Agent Based Manufacturing Workshop at Autonomous Agents '98*, May 1998. Available at <<http://luckyspc.lboro.ac.uk/Docs/Papers/Agents98.html>>.
- [60] T. Papaioannou and J. Edwards. Manufacturing System Performance and Agility: Can Mobile Agents Help? *Special Issue of Integrated Computer-Aided Engineering*, 1999. Available at <<http://luckyspc.lboro.ac.uk/Docs/Papers/Icae98.pdf>>.
- [61] T. Papaioannou and N. Minar, editors. *Workshop on Mobile Agents in the Context of Competition and Co-operation at Agents '99*, May 1999. Available at <<http://mobility.lboro.ac.uk/MAC3/>>.
- [62] S. Papastavrou, G. Samaras, and E. Pitoura. Mobile Agents for WWW Distributed Database Access. In *Proceedings of the 15th International Conference on Data Engineering*, Sydney, Australia, March 1999. Available at <<http://ada.cs.ucy.ac.cy/~cssamara/DBMS-Agents/Paper/papastavrous.ps>>.
- [63] V.A. Pham and A. Karmouch. Mobile Software Agents: An Overview. *IEEE Communications*, 36(7):26–37, July 1998.
- [64] R.J. Rabelo and L.M. Spinosa. Critical Research Issues in Agent-based Manufacturing Supply Webs. In *Proceedings of the Workshop on Supply-chain Management at Agrosoft'97*, 1997. Available at <<http://www.gsigma-grucon.ufsc.br/english/files/agro3.zip>>.
- [65] A.T. Ramu. Incorporating Transportable Software Agents into a Wide Area High Performance Distributed Data Mining System. Master's thesis, Electrical Engineering and Computer Science, University of Illinois, 1998.
- [66] D. Reed, I. Pratt, P. Menage, S. Early, and N. Stratford. Xenoservers: Accountable Execution of Untrusted Programs. In *Hot Topics in Operating Systems*, 1999. Available at <<http://www.cl.cam.ac.uk/Research/SRG/netos/xeno/hotos1/index.html>>.
- [67] K. Rothermel and F. Hohl, editors. *MA '98: Proceedings of the 2nd International Workshop on Mobile Agents*, number 1477 in LNCS. Springer-Verlag, September 1998.
- [68] K. Rothermel and R. Popescu-Zeletin, editors. *MA '97: Proceedings of the 1st International Workshop on Mobile Agents*, number 1219 in LNCS. Springer-Verlag, April 1997.
- [69] W. Rubin, M. Brain, and R. Rubin. *Understanding DCOM*. P T R Prentice-Hall, 1998.
- [70] A. Sahai and C. Morin. Mobile Agents for Enabling Mobile User Aware Applications. In Kattia P. Sycara and Michael Wooldridge, editors, *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, pages 205–211, New York, May 9–13, 1998. ACM Press. Available at <<http://www.acm.org/pubs/articles/proceedings/ai/280765/p205-sahai/p205-sahai.pdf>>.
- [71] B. Segall and D. Arnold. Elvin has Left the Building: a Publish/subscribe Notification Service with Quenching. In *Proceedings of AUUG97*, Brisbane, Australia, September 1997. Available at <<http://www.dstc.edu.au/Elvin/doc/papers/auug97/AUUG97.html>>.
- [72] W. Shen and D.H. Norrie. Agent-Based Systems for Intelligent Manufacturing: A State-of-the-Art Survey. *Knowledge and Information Systems*, 1(2):129–156, 1999. Available at

- <<http://imsg.enme.ucalgary.ca/publication/abm.htm>>.
- [73] P. Stanski. Private Communication. June 1999.
- [74] M. Strasser and K. Rothermel. Reliability Concepts for Mobile Agents. *International Journal of Cooperative Information Systems*, 7(4):355–382, 1998. Available at <<http://www.informatik.uni-stuttgart.de/ipvr/vs/Publications/Publications.html#1998-strasser-02>>.
- [75] Sun Microsystems Inc. Jini(tm) Connection Technology. 1999. Web page at <<http://www.sun.com/jini/>>.
- [76] K. Sycara, K. Decker, A. Pannu, M. Williamson, and D. Zeng. Distributed Intelligent Agents. *IEEE Expert*, December 1996. Available at <<http://www.cs.cmu.edu/~softagents/papers/iecc-agents96.ps.gz>>.
- [77] N. Szirbik, J.B.M. Goossenaerts, D. Hammer, and A.T.M. Aerts. Mobile Agent Support for Tracking Products in Virtual Enterprises. In *Proceedings of Workshop on Agents for Electronic Commerce and Managing the Internet-Enabled Supply Chain*, Seattle, Washington, U.S.A., May 1999. Available at <<http://www.research.ibm.com/CoopDS/Agents99/szirbik.ps>>.
- [78] P. Tarau. Jinni: Intelligent Mobile Agent Programming at the Intersection of Java and Prolog. In *Proceedings of the Practical Application of Intelligent Agents and Multi-Agent Technology*, London, U.K., April 1999. Available from <<http://www.cs.unt.edu/~tarau/research/99/jpaper.html>>.
- [79] K. Taveter. Business Rules' Approach to the Modelling, Design, and Implementation of Agent-Oriented Information Systems (Poster). In *Proceedings of the Workshop on Agent-oriented Information Systems at Agents '99*, Seattle, U.S.A., 1999. Full paper available at <<http://www.vtt.fi/te/samba/staff/tav/aois99-kt.ps>>.
- [80] S. Terzis, P. Nixon, V. Wade, S. Dobson, and J. Fuller. The Future of Enterprise Groupware Applications. In *Proceedings of the 1st International Conference on Enterprise Information Systems*, March 1999. Available at <http://www.cs.tcd.ie/Virtues/Papers/Groupware_iceis99.ps>.
- [81] W. Theilmann and K. Rothermel. Efficient Dissemination of Mobile Agents. In W. Sun, S. Chanson, D. Tygar, and P. Dasgupta, editors, *Proceedings of the 19th International Conference on Distributed Systems Workshop*, pages 9–14, June 1999.
- [82] M. T. Tu, F. Griffel, M. Merz, and W. Lamersdorf. A Plug-in Architecture Providing Dynamic Negotiation Capabilities for Mobile Agents. In K. Rothermel and F. Hohl, editors, *Proc. 2nd Int. Workshop on Mobile Agents*, volume 1477 of *Lecture Notes in Computer Science*, pages 222–236, Stuttgart, Germany, 1998. Springer-Verlag, Berlin.
- [83] D. Veeramani, P. Joshi, and V. Sharma. Critical Research Issues in Agent-based Manufacturing Supply Webs. In *Proceedings of Workshop on Agents for Electronic Commerce and Managing the Internet-Enabled Supply Chain*, Seattle, Washington, U.S.A., May 1999. Available at <<http://www.research.ibm.com/CoopDS/Agents99/veeramani.ps>>.
- [84] B. Venners. Solve Real Problems with Aglets, a Type of Mobile Agent. *JavaWorld*, May 1997. Available at <<http://www.javaworld.com/javaworld/jw-05-1997/jw-05-hood.html>>.
- [85] G. Vigna, editor. *Mobile Agents and Security*, number 1419 in *Lecture Notes in Computer Science*. Springer-Verlag, 1998.
- [86] L. Vinke, M. Heitz, and N. Swoboda. A Distributed Appointment Scheduler. December 1997. Web page at <http://www.cs.indiana.edu/hyplan/nswoboda/b629/aglets_meeting/tiny/readme.html>.
- [87] H. Vogler, T. Kunkelmann, and M.-L. Moschath. Distributed Transaction Processing as a Reliability Concept for Mobile Agents. In *Proceedings of the 6th IEEE Workshop on Future Trends of Distributed Computing Systems (FTDCS'97)*, Tunis, Tunisia, October 1997. Available at <<http://www.ito.tu-darmstadt.de/publs/papers/ftdcs97.ps.z>>.
- [88] J.E. White. Telescript Technology: The Foundation for the Electronic Marketplace. White paper, General Magic, Inc., 2465 Latham Street, Mountain View, CA 94040, 1994.
- [89] C. Wicke, L.F. Bic, M.B. Dillencourt, and M. Fukuda. Automatic State Capture of Self-Migrating Computations in MESSENGERS. In *Proceedings of the 2nd International Workshop on Mobile Agents (MA'98)*, LNCS 1477, September 1998. Available at <<http://www.ics.uci.edu/~bic/messengers/MA98.ps>>.
- [90] D. Wong, N. Paciorek, and D. Moore. Java-based Mobile Agents. *Communications of the ACM*, 42(3):92–102, March 1999.
- [91] M. Wooldridge and N. Jennings. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, 10(2), 1995. Available at <<http://www.elec.qmw.ac.uk/dai/pubs/KER95/>>.
- [92] Y. Yan, T. Kuphal, and J. Bode. Applications of Multiagent Systems in Project Management. In *Proceedings of the 2nd International Conference*

on *Autonomous Agents*, May 1998. Available at <<ftp://fireflow.com/Agent98.zip>>.

- [93] J. Yang, P. Pai, V. Honavar, and L. Miller. Mobile Intelligent Agents for Document Classification and Retrieval: A Machine Learning Approach. In *Proceedings of the European Symposium on Cybernetics and Systems Research*, 1998. Available at <<http://www.cs.cmu.edu/afs/cs.cmu.edu/user/honavar/www/Papers/emcsr98.ps>>.
- [94] L. Yu and B.F. Schmid. A Conceptual Framework for Agent Oriented and Role Based Workflow Modeling. In *Proceedings of the Workshop on Agent-oriented Information Systems at CAiSE'99*, Heidelberg, Germany, June 1999. Full paper available at <<http://www.mcm.unisg.ch/people/lyu/yuAOIS99.pdf>>.

A pattern recognition approach to the prediction of price increases in the New York Stock Exchange Composite Index

William Leigh
 Department of Management Information Systems, University of Central Florida
 Orlando, Florida 32816-1400 USA
 Phone: 407 823 3173, Fax: 407 823 4166
 leigh@pegasus.cc.ucf.edu

Mario Paz
 Dept of Civil Engineering, University of Louisville
 Louisville, Kentucky USA

Noemi Paz
 Science Applications International Corp.
 Orlando, Florida USA

Russ Purvis
 Department of Management, Clemson University
 Clemson, South Carolina USA

Keywords: Stock market forecasting, pattern recognition, heuristics, financial decision support, efficient markets hypothesis, technical analysis

Received: February 10, 2001

We match a template depicting a "bull flag" 60-day price behavior to closing prices of the New York Stock Exchange Composite Index to detect buying opportunities at a 20 trading day horizon. The results of the experiment indicate that the technique is capable of achieving results which are superior to those attained by random choice. Statistical results are significant. The paper constitutes evidence that the stock markets are not efficient.

1 Introduction

Many academic studies promulgate the thesis that the prediction of stock market prices using historical price data is futile. This point of view, codified as the efficient markets hypothesis (Fama 1970, Haugen 1997), states that market prices follow a random walk and cannot be predicted based on their past behavior. According to the efficient markets hypothesis there are three degrees of market efficiency. The *strong form* states that *all information that is knowable* is immediately factored into the market's price for a security. If this is true, then all of those stock analysts are definitely wasting their time, even if they have access to private information. In the *semi-strong form* of the efficient markets hypothesis, *all public information* is considered to have been reflected in price immediately as it became known, but possessors of private information can use that information for profit. The *weak form* holds only that

any *information gained from examining the security's past trading history* is immediately reflected in price. Of course, the past trading history is public information, which implies that exceptions and counter-examples to the weak form also apply to the strong and semi-strong forms.

A great many of the professional stock market price predictors practice technical analysis, which is based on the assumption that patterns in past price and volume history can be used to predict future price behavior. By using technical analysis, these individuals choose to ignore the weak form of the efficient markets hypothesis.

Technical analysis is concerned solely with the dynamics of the market price and volume behavior as a basis for price prediction. The dynamics are described in terms of metrics and patterns (called "stock price chart patterns"). Charles Dow developed the original theory of technical analysis in 1884, and

modern explications are published periodically (Edwards & Magee 1977). Papers and books on technical analysis appear frequently in the practitioner literature (Martinelli & Hyman 1998, Plummer 1990), but in the past there has been little in the academic literature and then usually in an indirect, pejorative, or defensive form (Treynor & Ferguson 1985, Irwin & Uhrig 1984). However, studies which offer evidence of market inefficiencies and some support for technical approaches to market forecasting are beginning to appear (Gencay 1998, Hong et. al. 2000, Hong & Stein 1999, Chan et. al. 1996, Caginalp et. al. 2000). These studies concentrate on the “momentum” strategy (that is, past winners continue to perform well and past losers tend to continue to lose) and use numerical metrics rather than the recognition of stock price chart patterns.

The results of the experiment reported in this paper constitute strong confirmation of one bit of technical analysis lore. Specifically, we take one stock chart price pattern heuristic from technical analysis that is thought to be a signal for a price increase; implement a recognizer for this stock chart price pattern; and use the output of the recognizer to predict whether there will be a price increase or not. Our results are statistically significant and fail to confirm the implied null hypothesis that predictions using the method are no better than random choice (which is implied by the efficient markets hypothesis.)

2 Pattern Recognition

The process that we use is an example of template matching (Duda & Hart 1973), a pattern recognition technique used to match an image to a template.

Our work concentrates on one technical analysis pattern, the bull flag, which is considered to signal an imminent increase in price. The definition of “flag” (Downes and Goodman 1988 p.212): “a technical chart pattern resembling a flag shaped like a parallelogram with masts on either side, showing a consolidation within a trend. It results from price fluctuations within a narrow range, both preceded and followed by sharp rises or declines.” A bull flag pattern is a horizontal or downward sloping flag of “consolidation” followed by a sharp rise in the positive direction, the “breakout.” The template that we use for the bull flag pattern, shown in Figure 1, is represented with a ten-by-ten (10X10) grid, the cells of which contain weights ranging from -2.5 to $+1.0$. The pattern of positive and negative weighting defines areas in the template for the descending consolidation and for the upward-tilting breakout

portions of this bull flag heuristic pattern. Note that the weights in each column of the template sum to 1.0, reflecting that no column is weighted higher than another. We tried several bull flag templates with weights similar to this one, but which did not perform as well as this one.

We fit this 10X10 bull flag template grid to the closing price for each trading day in the time series of price information. We map the 10X10 grid to 60 trading days at a time. We tried several windows other than 60 days, but none worked as well as 60 days with this template.

The leftmost time series data point in the 60 day window represents the trading day which precedes the current day by 59 trading days, and the rightmost time series data point in the window corresponds to the trading day which is being analyzed. Values for the earliest 10% of the trading days (6 trading days) are mapped to the first, leftmost, of the ten columns of the grid, values for the next-to-earliest 10% of the trading days are mapped to the second-from-the-left column of the grid, and so on until the most recent 10% of the trading days are mapped to the rightmost column.

Within each 60 days of data, we ‘windsorize’ (Sargent 1993 p.150) to remove the worst noise by replacing every observation which is beyond two standard deviations from the mean with the respective two standard deviation boundary value.

The fitting process is adaptive in the vertical dimension: the highest value in the window is made to correspond with the top of the grid, and the lowest value in the window is made to correspond with the bottom of the grid, and the intervening vertical cells are made to correspond with linearly with the values in between. For a fitting to the time series data, we enter into each cell of a column the percentage of price values which fall into the respective cell in the column; for example, if all of the values in a column were between the lowest value in the 60 trading day window and a value which is higher than the lowest value by 10% of the difference between the window lowest and highest values in the window, then that column’s lowest cell would be coded with 100%, and the rest of the cells in the column would be coded with 0%.

To compute the degree of match between the bull flag template and the grid of values derived from the time series data, the percentage of values which falls in each cell of a column is multiplied by the weight in the corresponding cell of the bull flag template. This

cross-correlation computation is done for the 10 cells in the column and summed, resulting in a fit value for the column. Thus, 10 column fit values for price are computed for each trading day. Summing all 10 values for a trading day results in a total fit for the trading day.

For example: There will be 6 trading days represented in each column of the fitting of a single 60 trading day window. If all 6 of these trading days have price values that are in the lowest decile of the 60 price values for the day, then 100% (6 values out of a total of 6 in the column) will be the value in the lowest cell of the ten cells in the column. If this column is the leftmost of the columns in the window, then this 100% will be multiplied by the value in the corresponding cell in the bull flag template (which is the one in the lowest left-hand corner), which has the value of -1.0 (See Figure 1), to result in a cell fit value of $-1.0 \times 100\% = -1.0$. This is done for the 10 cells in the column and summed, resulting in a fit value for the column of -1.0, since in this example there will be 0.0% of the values in the other 9 cells of the column.

3 Implementation and Results

This study focuses on the prediction of increases in the New York Stock Exchange Composite Index with a 20-day horizon for the period from 08/06/80 to 09/15/99. The price data used are obtained from the Standard & Poor's DRI database. We programmed the template matching procedure in Microsoft Corporation's Excel spreadsheet.

Figures 2 through 7 report the results of applying this bull flag price chart pattern recognizer.

Figure 2 plots the 60 trading day bull flag template total fit value for each trading day in the sample against the 20-day price change experienced following the trading day. Each point represents a single fitting of the 60-day bull flag template. The template fit value is the result of the cross-correlation calculation between the template and the price data in the window. Note the apparent positive correlation of fit and price increase which shows on the right side of the chart (toward the better fits).

Figure 3 shows the cumulative distribution of trading days by their fit value. Notice that approximately 1000 of the almost 5000 trading days in the period of the study have a 60-day bull flag template fit value of 2 or better. This implies that if a template fit of 2 or better were used as a trading rule, then purchases

would be indicated on about 1000 trading days. The distribution appears to be approximately normal with a mean slightly below zero and a negative skew.

Each point on the plot in Figure 4 represents the mean value of price increase for all trading days having the indicated fit or better. Figure 4 shows that the cumulative mean 20-day price change increases dramatically for fits of 1 to 6. After 6 there is a marked falling off. We suspect that this falling off is a symptom of some defect in our template, and we are currently investigating why this is so

The cumulative standard deviation for all fit values for a given fit value and greater is plotted on Figure 5. The standard distribution decreases, showing less risk, for fit values greater than about 4.

Figure 6 shows one-tailed t-test probabilities for the difference between a) the overall mean 20-day price change and b) the mean 20-day price change for days with a certain fit value or better. Note that fit values of 2 and greater have statistically significant t-test probability values. The dip in t-test probability values in the negative fit values is due to the markedly low profit percentages experienced, making the cumulative t-test probability significant but for a lower than average overall profit. This effect can be examined on Table 1.

Table 1 presents the results summarized for 100 trading day groups ordered by fit value. The first row contains summaries for the 100 trading days with best template fit values, and so forth. The 20 day profit for each 100 day group is compared with the overall 20 day profit average for the complete 4816 day test period, and the cumulative 20 day profit to that point is compared with the overall profit.

Figure 7 is an attempt to present the results in terms of the time line. Pairs of points for each trading day represent (1) the 20 day profit percent (small round symbols ranging between 0 and about .05) for all buys which would have been indicated by a fit value of 6 or better in that day or in the 249 preceding days and (2) the one-tailed t-test probability (triangle symbol ranging from 0.0 to 0.5) of that average compared with random buys in that 250 day period. Pairs are plotted only when trading days with fits of 6 or better occurred in that day or in the previous 249 trading days. We choose the fit value of 6 and the rolling window width of 250 days because these values result in a display that clearly shows the effectiveness of the method. We observe good profitability and good t-test significance in active buying periods around Feb-82, just before Nov-84,

around Mar-86, Dec-88, Oct-95, and just before Jul-98. Performance is not so good in active buying periods around May-90 and Jan-93.

Figure 7 embodies a cross-validation of this work through time. The face validity of this cross-validation is a defense against charges that the reported results are the result of a search of parameter values until good results are achieved. All forecasting and statistical studies may be accused of this "data snooping." It is the unreported procedures and results that must be examined, and therein is a paradox. Always, the reader must judge the likelihood that the reported results are repeatable and robust, regardless of the statistical significance reported. Highly significant results, which are demonstrated to hold across many cuts and slices of the data, are excellent evidence for the defense. It is not difficult to defend the opinion that the rolling window cross-validation reported in Figure 7 has considerable face validity and constitutes strong evidence that the technique is valid.

4 Conclusion

These results indicate that there is support for the technical analysis approach to stock price prediction using the bull flag price chart pattern heuristic. Much further work, however, is needed to examine fitting windows other than 60 trading days, price horizons other than 20 trading days, and other pattern heuristic templates in order to assemble enough response data to begin to speculate on the mechanism behind this phenomenon. Explanations may lie in the effects of information diffusion and/or bounded rationality of traders.

5 References

- [1] G. Caginalp, D. Porter, and V. Smith (2000) Momentum and Overreaction in Experimental Asset Markets. *International Journal of Industrial Organization*, 18, p. 187-204.
- [2] L. Chan, N. Jegadeesh, and J. Lakonishok (1996) Momentum Strategies. *Journal of Finance*, LI, 5, p. 1681-1713.
- [3] J. Downes and J. Goodman (1988) *Dictionary of Finance and Investment Terms*. Barron's Educational Series, Inc., New York.
- [4] R. Duda and P. Hart (1973) *Pattern Classification and Scene Analysis*. John Wiley, New York.
- [5] R. Edwards and J. Magee (1973) *Technical Analysis of Stock Trends*, Amacom, New York.
- [6] E. Fama (1970) Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25, p. 383-417.
- [7] R. Gencay (1998) Optimization of Technical Trading Strategies and the Profitability in Security Markets. *Economics Letters*, 59, p. 249-254.
- [8] R. Haugen (1977), *Modern Investment Theory*. Prentice Hall, Upper Saddle River, New Jersey.
- [9] H. Hong, T. Lim, and J. Stein (2000) Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies. *Journal of Finance*, LV, 1, p. 205-295.
- [10] H. Hong and J. Stein (1999) A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *Journal of Finance*, LIV, 6, p. 2143-2184.
- [11] S. Irwin and J. Uhrig (1984) Do Technical Analysts Have Holes in Their Shoes? *Review of Research in Futures Markets*, 3, p. 264-277.
- [12] R. Martinelli and B. Hyman (1998) Cup-With-Handle and the Computerized Approach. *Technical Analysis of Stocks and Commodities*, 16, 10, p. 63-66.
- [13] T. Plummer (1990) *Forecasting Financial Markets*. John Wiley & Sons, New York.
- [14] T. Sargent (1993) *Bounded Rationality in Macroeconomics*. Oxford University Press, Oxford.
- [15] J. Treynor and R. Ferguson (1985) In Defense of Technical Analysis. *Journal of Finance*, XL, p. 757-775.

Figure 1: Bull Flag template used in this experiment. The first seven columns depict a consolidation and the last 3 columns represent a breakout.

0.5	0.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	-1.0	0.0
1	0.5	0.0	-0.5	-1.0	-1.0	-1.0	-1.0	-0.5	0.0
1	1	0.5	0.0	-0.5	-0.5	-0.5	-0.5	0.0	0.5
0.5	1	1	0.5	0.0	-0.5	-0.5	-0.5	0.0	1
0.0	0.5	1	1	0.5	0.0	0.0	0.0	0.5	1
0.0	0.0	0.5	1	1	0.5	0.0	0.0	1	1
-0.5	0.0	0.0	0.5	1	1	0.5	0.5	1	1
-0.5	-1.0	0.0	0.0	0.5	1	1	1	1	0.0
-1.0	-1.0	-1.0	-0.5	0.0	0.5	1	1	0.0	-2.0
-1.0	-1.0	-1.0	-1.0	-0.5	0.0	0.5	0.5	-2.0	-2.5

Figure 2: Bull flag template fit value compared to 20-day price change. Each point represents the fit value and price change for a single trading day.

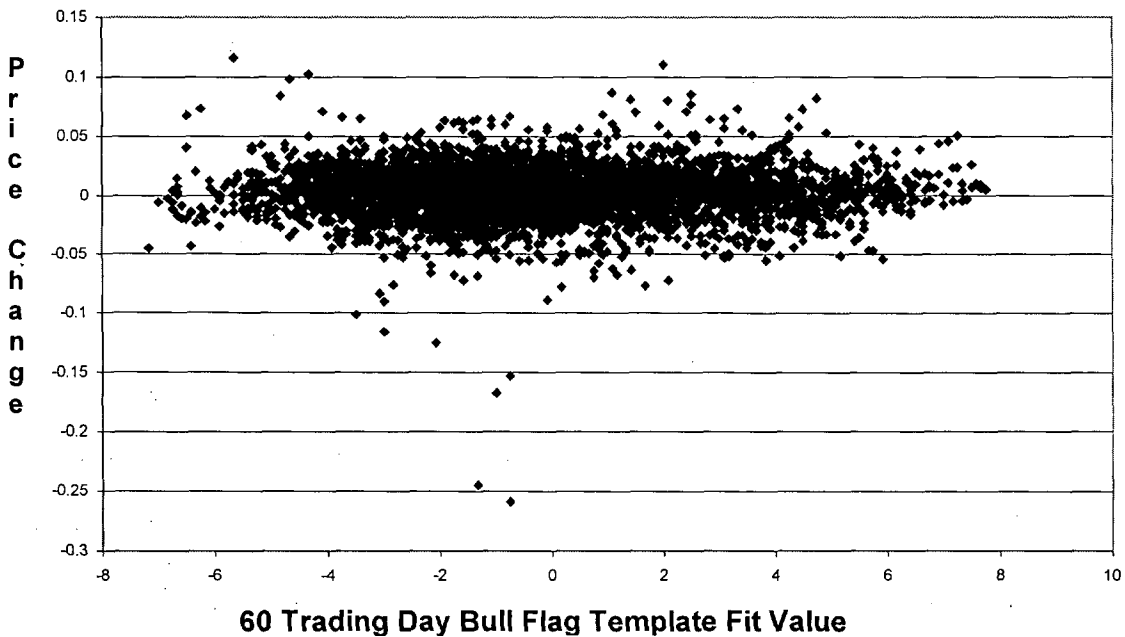


Figure 3: Cumulative distribution of better fitting trading days.

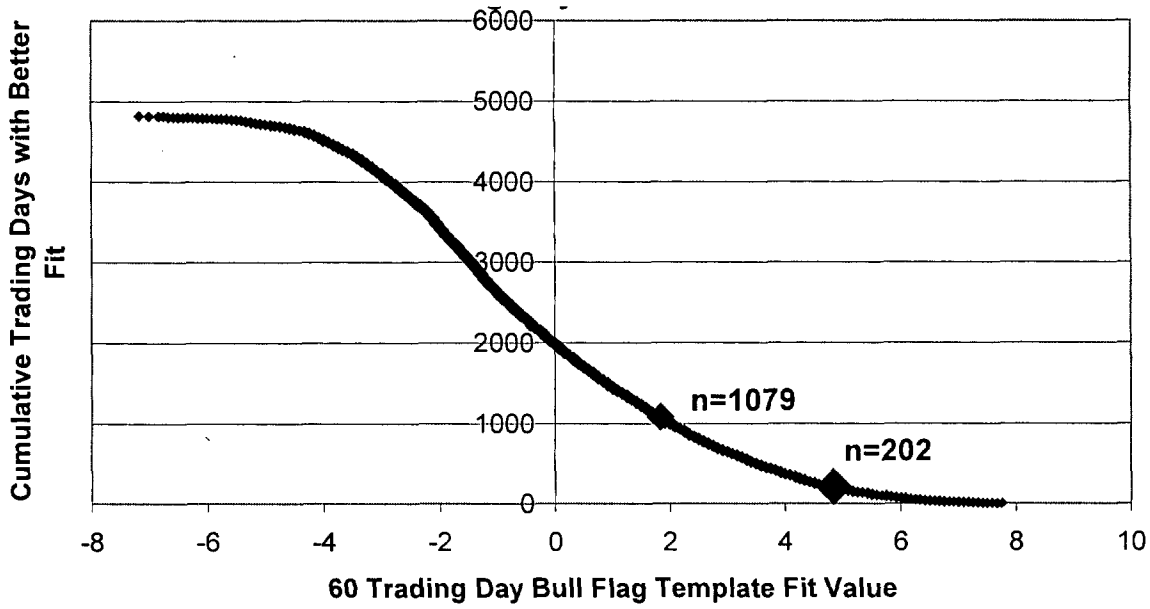


Figure 4: Bull flag template fit value compared to mean 20-day price change for all better fitting trading days.

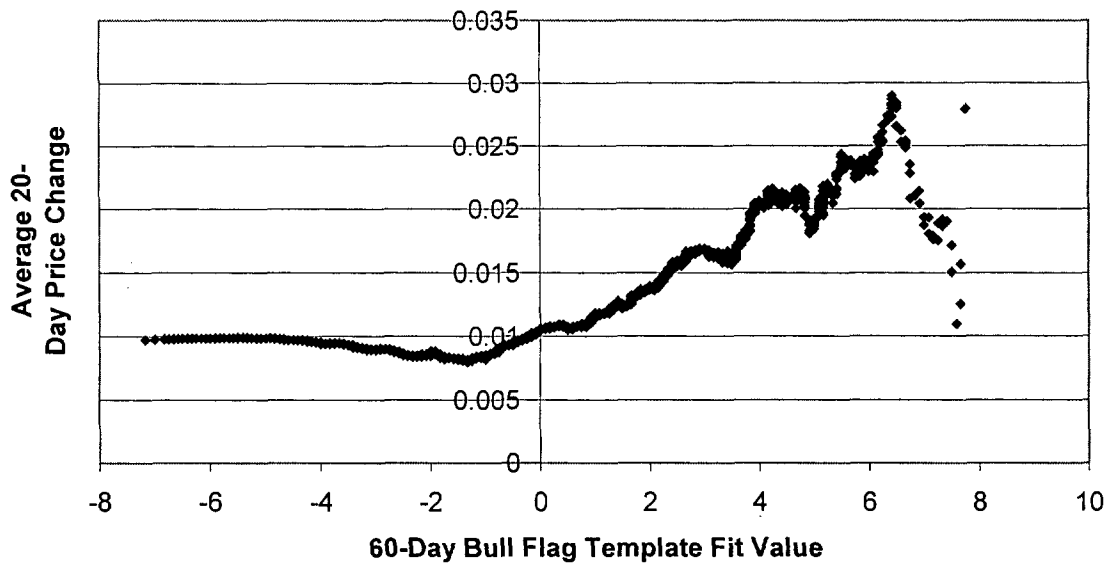


Figure 5: 60-day bull flag template fit compared to cumulative standard deviation of 20-day price change for better fitting trading days.

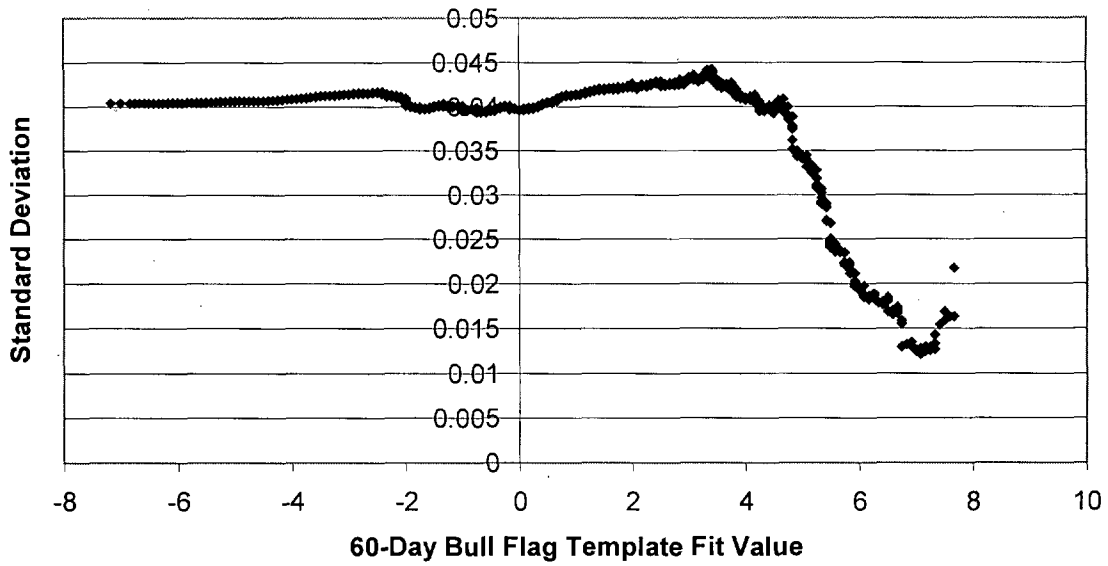


Figure 6: 60-day bull flag template fit compared to cumulative t-test probability of difference in means between a) 20-day price change for better fits and b) price change overall.

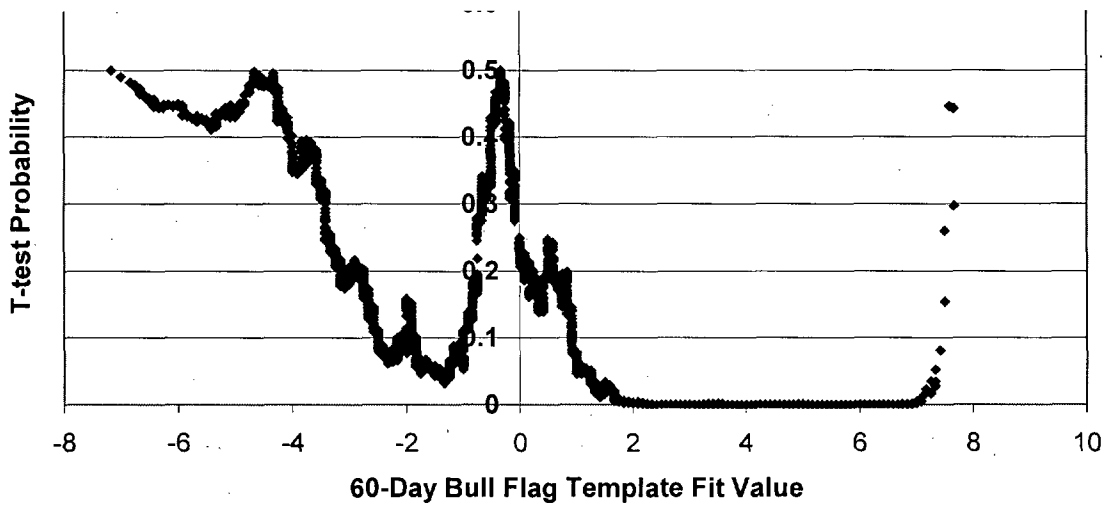
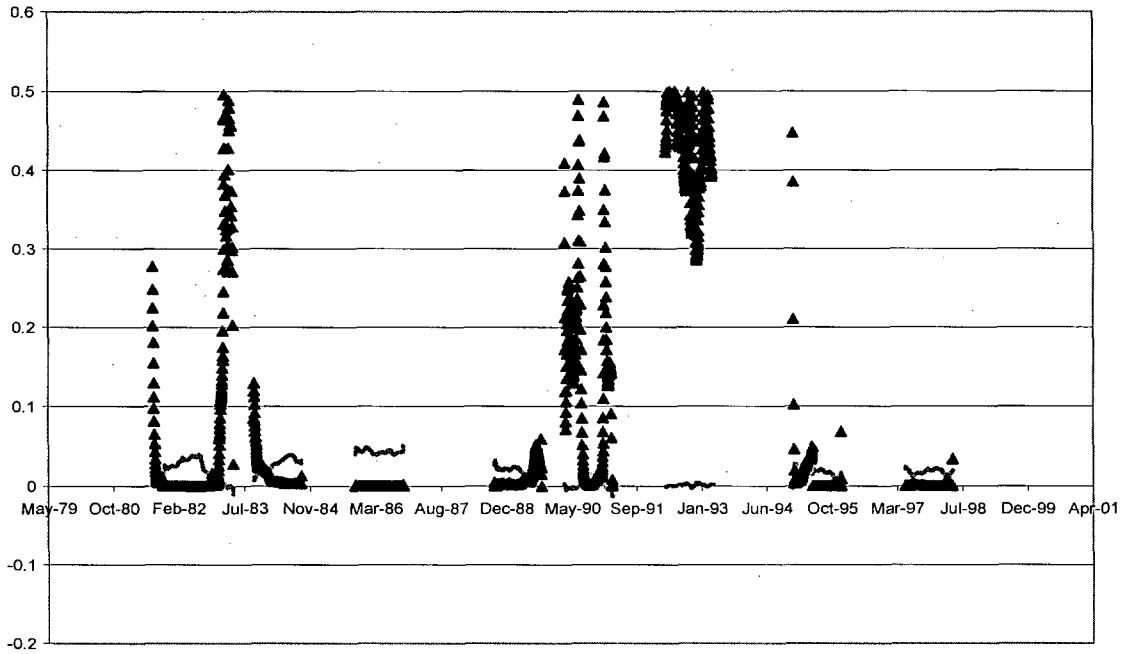


Table 1: Results summarized for 100 trading day groups ordered by fit value. The first row contains summaries for the 100 trading days with best template fit values, and so forth. The “100 Trading Days” 20 day profit is compared with the overall 20 day profit average for the complete 4816 day test period.

Rank in Fit		100 Trading Days		Cumulative	
begin	end	20 day profit	ttest prob.	20 day profit	ttest prob.
1	100	2.36%	0.0003	2.36%	0.0003
101	200	1.75%	0.0294	2.05%	0.0001
201	300	2.25%	0.0009	2.12%	0.0000
301	400	1.34%	0.1834	1.92%	0.0000
401	500	0.39%	0.0772	1.62%	0.0004
501	600	1.61%	0.0606	1.62%	0.0001
601	700	1.88%	0.0136	1.65%	0.0000
701	800	1.12%	0.3583	1.59%	0.0000
801	900	0.44%	0.0959	1.46%	0.0005
901	1000	0.59%	0.1766	1.37%	0.0024
1001	1100	0.67%	0.2279	1.31%	0.0068
1101	1200	0.50%	0.1243	1.24%	0.0204
1201	1300	1.05%	0.4273	1.23%	0.0230
1301	1400	0.57%	0.1638	1.18%	0.0464
1401	1500	0.19%	0.0282	1.11%	0.1192
1501	1600	0.58%	0.1663	1.08%	0.1778
1601	1700	0.81%	0.3488	1.06%	0.2089
1701	1800	1.37%	0.1636	1.08%	0.1629
1801	1900	0.74%	0.2834	1.06%	0.2010
1901	2000	0.76%	0.3007	1.05%	0.2379
2001	2100	0.01%	0.0090	1.00%	0.4003
2101	2200	0.24%	0.0360	0.96%	0.4690
2201	2300	0.24%	0.0359	0.93%	0.3487
2301	2400	0.58%	0.1696	0.92%	0.2944
2401	2500	-0.20%	0.0022	0.87%	0.1594
2501	2600	0.33%	0.0580	0.85%	0.1102
2601	2700	0.21%	0.0315	0.83%	0.0693
2701	2800	0.52%	0.1330	0.82%	0.0530
2801	2900	0.26%	0.0408	0.80%	0.0334
2901	3000	1.57%	0.0707	0.82%	0.0568
3001	3100	0.89%	0.4158	0.83%	0.0572
3101	3200	1.02%	0.4506	0.83%	0.0630
3201	3300	1.27%	0.2339	0.85%	0.0814
3301	3400	2.00%	0.0061	0.88%	0.1520
3401	3500	-0.19%	0.0026	0.85%	0.0860
3501	3600	0.54%	0.1485	0.84%	0.0712
3601	3700	0.99%	0.4810	0.84%	0.0760
3701	3800	1.23%	0.2624	0.85%	0.0934
3801	3900	1.85%	0.0155	0.88%	0.1488
3901	4000	1.62%	0.0568	0.90%	0.2012
4001	4100	0.91%	0.4390	0.90%	0.2001
4101	4200	1.11%	0.3719	0.90%	0.2146
4201	4300	1.94%	0.0087	0.93%	0.3037
4301	4400	1.74%	0.0286	0.95%	0.3822
4401	4500	0.67%	0.2315	0.94%	0.3542
4501	4600	1.91%	0.0105	0.96%	0.4505
4601	4700	1.96%	0.0078	0.98%	0.4485
4701	4800	0.90%	0.4299	0.98%	0.4564
4801	4816	-1.58%	0.0047	0.97%	0.5000

Figure 7: Results expressed in terms of a 250 day rolling average. Pairs of points for each trading day represent (1) the 20 day profit percent (smaller symbols ranging between 0 and about .05) for all buys which would have been indicated by a fit value of 6 or better in that day or in the 249 preceding days and (2) the one-tailed t-test probability (larger triangle symbol ranging from 0 to .5) of that average compared with random buys in that 250 day period. Pairs are plotted only when trading days with fits of 6 or better occurred in that day or in the previous 249 trading days.



A digital multi-layer-perceptron hardware architecture based on three dimensional massively parallel optoelectronic circuits

Klaus D. Maier^{1,3}, Clemens Beckstein², Reinhard Blickhan¹, Dietmar Fey², Werner Erhard²

¹ Friedrich-Schiller-University, Institute of Sports Science, D-07740 Jena, Germany

² Friedrich-Schiller-University, Institute of Computer Science, D-07740 Jena, Germany

³ Infineon Technologies AG, CMD MCU ME ACE, P.O.Box 80 09 49, D-81609 Munich, Germany

Phone: +49-89-234-81597, Fax: +49-89-234-717822; Email: klaus.maier@infineon.com

Keywords: digital neural networks, optoelectronics, hardware architecture

Received: July 12, 2000

A digital neural network architecture is presented which is based on three-dimensional massively parallel opto-electronic circuits. A suitable optical interconnect system and the structure of the required electronic circuits is specified. For this system general formulas for the performance of such a neural network architecture are determined. A parameter study using current technological limitations and timing values from electronic implementation is carried out. Based on this analysis it is shown that this novel type of neuro-architecture that is using 3D massively parallel opto-electronic circuits shows performance rates of up to one magnitude higher than systems using digital neurochips based on fully electronic implementation.

1 Introduction

Massively parallel processing units like multi-layer perceptron neural networks (MLPs) are powerful tools used for a wide range of purposes from pattern classification over modelling to implementations for control. For some of these tasks timing matters only to a limited extend. For others real time performance is not only desired but of vital importance. In such cases simulation of the required MLP on mainframe computers or micro-controllers might not provide sufficient performance. Employing dedicated hardware is necessary. Hardware implementations of MLPs that use purely electronic implementations suffer from shortcomings: Digital hardware has to cope with problems related to high Fan-Outs of each neuron and the fact that a very large number of interconnects between neurons result in large routing areas. Analogue hardware is also faced with the problem of high Fan-Outs as well as that of possibly low precision of data processing.

In this paper we propose a hardware architecture that uses dedicated digital circuits for high precision data processing combined with optical interconnects for transmission of data between neurons. The art of chip-level optical interconnects is still in the development phase, nevertheless advances have been made. Experimental data from current work can be used to evaluate the performance to be expected from hardware using VLSI-circuits with optical interconnects. We intend to apply this hardware to control of movement in walking or hopping autonomous systems [1].

2 Multi-Layer-Perceptrons

MLPs consist of a number of neurons (or perceptrons) that have inputs and generate an output using a nonlinearity [2, 3]. Neurons in a MLP can be categorised in input neurons, output neurons and neurons that are neither of the two – so called hidden neurons. An MLP network is grouped in layers of neurons, i.e. input layer, output layer and hidden layers of neurons that can be seen as groups of parallel processing units. Each neuron of a layer is connected with all neurons of the following layer. These connections are directed (from the input to the output layer) and have weights assigned to. The operation of a MLP can be divided into two phases:

1. The training phase: Here the MLP is trained for its specific purpose using learning algorithms (e.g. Backpropagation training [2]).
2. The retrieve phase: The previously trained MLPs are used to generate outputs.

2.1 Structure of a neuron

A neuron n has a number of inputs and one output no , the so called activation state of the neuron. The activation states of the R neurons $p(1), \dots, p(R)$ from the previous layer that are connected to n are multiplied with their respective weights $w(1), \dots, w(R)$ and then summed up by the neuron in order to generate the neural input ni .

$$ni = \sum_{i=1}^R p(i) \cdot w(i) \quad (1)$$

To the neural input ni a bias value b is added. The output of the neuron no is determined using the transfer function T . This transfer function is usually sigmoid (s-shaped, [3]). Typical transfer functions are tangens hyperbolicus (2a) and the logistic function (2b):

$$T(x) = \tanh(x) \quad (2a)$$

$$T(x) = \frac{1}{1 + e^{-x}} \quad (2b)$$

The output no of the neuron is defined as

$$no = T(ni + b) \quad (3)$$

The values $w(1), \dots, w(R)$ of the connection weights and the bias b are determined during the training phase and used in the retrieve phase. For a supervised learning approach for control purposes, building the respective learning algorithm into hardware is not necessary. The training data is generated by numerically or analytically solving the control task [1]. For this study we will be focusing on networks that are trained by software simulation.

2.2 Layout of the MLP

The neurons are grouped in layers. The MLP consists of an input layer, an output layer and one or more hidden layers that are neither input layer nor output layer. In the input layer the inputs for the first hidden layer are generated. These inputs are connected to all neurons in the first hidden layer. The input layer itself is not used for processing, it is only generating the inputs for the first hidden layer. Each output in the first hidden layer is again connected with all neurons of the second layer and so on. The final output of the MLP is generated using the output layer. The neurons in the output layer differ from the previously introduced neurons by the transfer function used. The transfer function T used by output neurons usually is

$$T(x) = x \quad (4)$$

3 Mapping the MLP onto an optoelectronic 3D architecture

Our hardware implementation of an MLP neural network is build up with VLSI circuits using a three-dimensional optical interconnect system (see Figure 1). An optical realisation is favourable because of its high connection density. For processing data in each neuron an electronic design offers high integration density and good handling to implement logic functions.

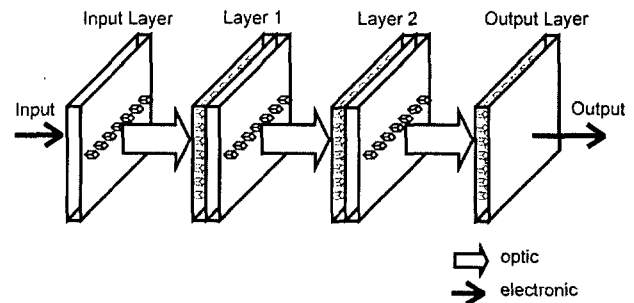


Figure 1: MLP as three dimensional system using optical interconnects with two hidden layers.

The transfer of the activation state of one neuron into the following layer is done optically, the processing of the activation state in each neuron electronically. In each neural cell the input signals (activation states of all neurons in the previous layer) are multiplied with their respective weights, then summed up and have the bias added to. Eventually the transfer function is applied to this sum yielding the activation state of the neuron. The processing is taking place alternating between electronic processing and optical transmission. In the input layer the scaled input signals are converted from electronic into optic form and then transmitted to the first hidden layer. In the output layer the optical signals are transformed into electronic signals and are then made the output of the neural network for further processing. The layered structure suggests to build one dedicated optoelectronic chip for each layer.

3.1 Specification of the optical interconnection system

Figure 2 shows a structure that can be used to connect two layers. The following notation is used:

- $\#N_{SL}$: number of physical neurons in the sending layer.
- $\#N_{RL}$: number of physical neurons in the receiving layer.
- b : number of bits used for representing the activation state and weights of each neuron.
- bl : transmitted wordlength – number of bits that are transmitted in one step. bl is determined as $bl = b/ac$ with the division factor $ac \in (1, 2, 4, 8, \dots)$ and $b \leq bl$.

The neurons in the sending layer are arranged horizontally (Figure 2). The light emitted by one transmitting unit is split into multiple beams by means of a diffraction grid that is adjusted on the transmitting unit. The receiving units are vertically placed at the respective bending maxima. Using this interconnection system $\#N_{RL}$ receiving units can be connected with one transmitting unit.

The sending layer consists of $\#N_{SL} \cdot bl$ transmitting units. The Fan-Out for each unit is $\#N_{RL}$. The receiving layer consists of $\#N_{RL} \cdot \#N_{SL} \cdot bl$ receiving units. The transmission for each package with bl bits is repeated ac times in order to transmit the b bit wide activation function.

The maximum Fan-Out FOM for each transmitting unit is limited by technological constraints (see section 4.1).

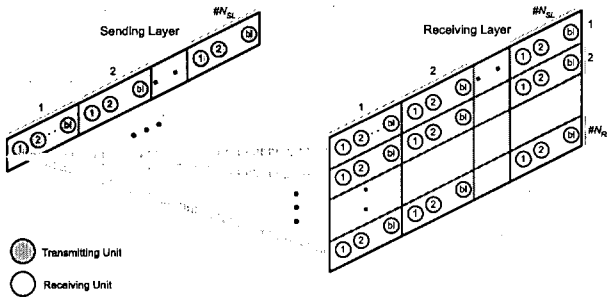


Figure 2: Structure of the optical interconnection system

3.2 Structural Design and Timing

The physical neurons can be divided into three functional blocks: a block that receives the optical signals and then electronically stores these signals (optic to electronic block, OEB), a block that is used for the electronic processing of the data (data processing block, DPB) and a block that can be used to transmit electronic data to optic data and to send it to the following layer (electronic to optic block, EOB).

3.2.1 Optic to electronic block.

The structure of the OEB is shown in Figure 3. The optical signals representing the activation states of all neurons from the previous layer are received by $bl \cdot \#N$ receiver cells and fed into a local memory. $\#N$ is the physical number of neurons. To further increase the number of neurons the processing elements and interconnects of the physical neurons can be used multiple times with employing multiplexing techniques. For this purpose the multiplication factor nc is introduced.

Incoming signals are stored in their respective place in the memory. The memory has an overall size of $nc \cdot \#N$ times b bits. It is used to store the activation states for further processing. Receiving the signals and storing them in the local memory requires time

$$T_{OEB} = ac \cdot nc \cdot (T_E + T_O). \tag{5}$$

T_E is the duration of one electronic processing step (including the conversion of optical to electrical). T_O is the duration of one optical transfer step (photon time of flight, T_O is much smaller than T_E , therefore T_O is set to be equal to zero). Hence

$$T_{OEB} \approx ac \cdot nc \cdot T_E. \tag{5a}$$

3.2.2 Data Processing Block.

Figure 4 shows the DPB. The b -bit wide activation states that are stored in the memory introduced in the OEB are transferred one after another to a multiplier (the overall time duration for this is $nc \cdot \#N \cdot T_E$). In the multiplier the activation states are multiplied with their respective weight value (requires $1 \cdot T_E$ time units). Eventually the product is accumulated by an adder (this also takes $1 \cdot T_E$ time). The memory contains an additional memory cell that is used to add the bias to the resulting sum. This

input activation state memory cell has the value 1 and the respective bias is stored in a memory cell in the weight memory. Furthermore the memory contains weight values for all nc logical neurons represented by this one physical neuron.

Processing takes place in parallel, i.e. the overall result accumulation is completed $3 \cdot T_E$ time units after the last activation state is looked up from the memory (or $2 \cdot T_E$ time units after the bias value) and is multiplied with the last weight value. Then the result of the accumulation is fed into the transfer function unit.

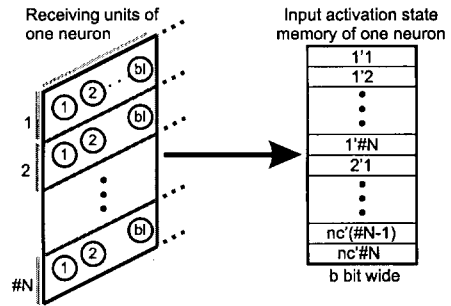


Figure 3: Optic to electronic block. The parameter in the activation state memory before the apostrophe indicates the number of repetitions of transfers due to multiplication and the parameter after the apostrophe indicates the number of the physical neurons.

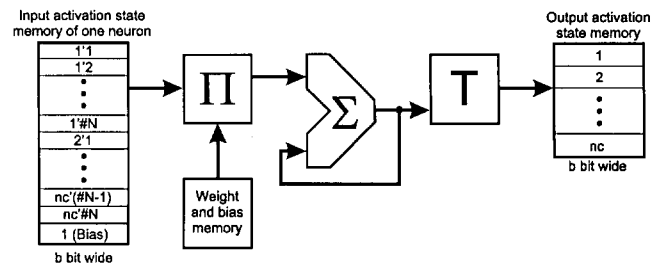


Figure 4: The data processing block

Here either a look-up table or function computation methods are used to determine the value of the transfer function.

This is done for the first neuron's activation state and then repeated for the other nc neurons using the remaining weights stored. At the end the resulting activation states of the nc logical neurons are fed into the electric to optic block. This gives an overall processing time in the DPB block of

$$T_{DPB} = nc \cdot (nc \cdot \#N + 3) \cdot T_E. \tag{6}$$

Using parallelising methods for the multiplication/accumulation parts can further enhance processing speed and performance.

3.2.3 Electronic to optic block.

The structure of the EOB can be seen in Figure 5. The memory states have a wordlength of b bit and are altogether nc states. For the transmission of one bl -bit

packet of the output activation state memory is sent. Hence, the time necessary for the repeated transfer of all packets is:

$$T_{EOB} = ac \cdot nc \cdot T_E \tag{7}$$

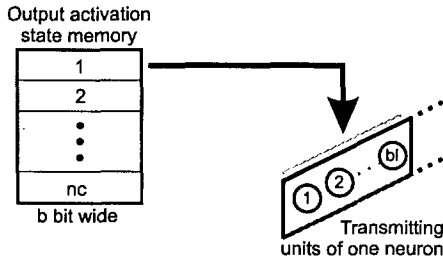


Figure 5: The electric to optic block

3.2.4 Overall processing time in one hidden or output layer.

The overall processing time T_H for one layer therefore is the sum of the times for optic to electronic block (5), electronic to optic block (7) and data processing block (6):

$$T_H = T_{OEB} + T_{DPB} + T_{EOB} = (2 \cdot ac + nc \cdot \#N + 3) \cdot nc \cdot T_E \tag{8}$$

3.3 Time for one complete propagation

The processing time for evaluating the complete network, i.e. the overall processing time T_{NN} in the whole MLP is

$$T_{NN} = \sum_{i=0}^{\#HL+1} T_i = (ac \cdot nc + (\sum_{i=1}^{\#HL+1} (2 \cdot ac + nc \cdot \#N_i + 3)) \cdot nc) \cdot T_E = ((2 \cdot ac + 3) \cdot (\#HL + 1) + nc \cdot \sum_{i=1}^{\#HL+1} \#N_i + ac) \cdot nc \cdot T_E \tag{9}$$

In this formula T_0 and $T_{\#HL+1}$ denote the propagation time of the input and output layer and the remaining ones are for the hidden layers of the network. The index $\#HL+1$ specifies the output layer, index $i \in (1, 2, \dots, \#HL)$ the hidden layer number i and index 0 states the input layer. Figure 6 shows the overall structure of the MLP.

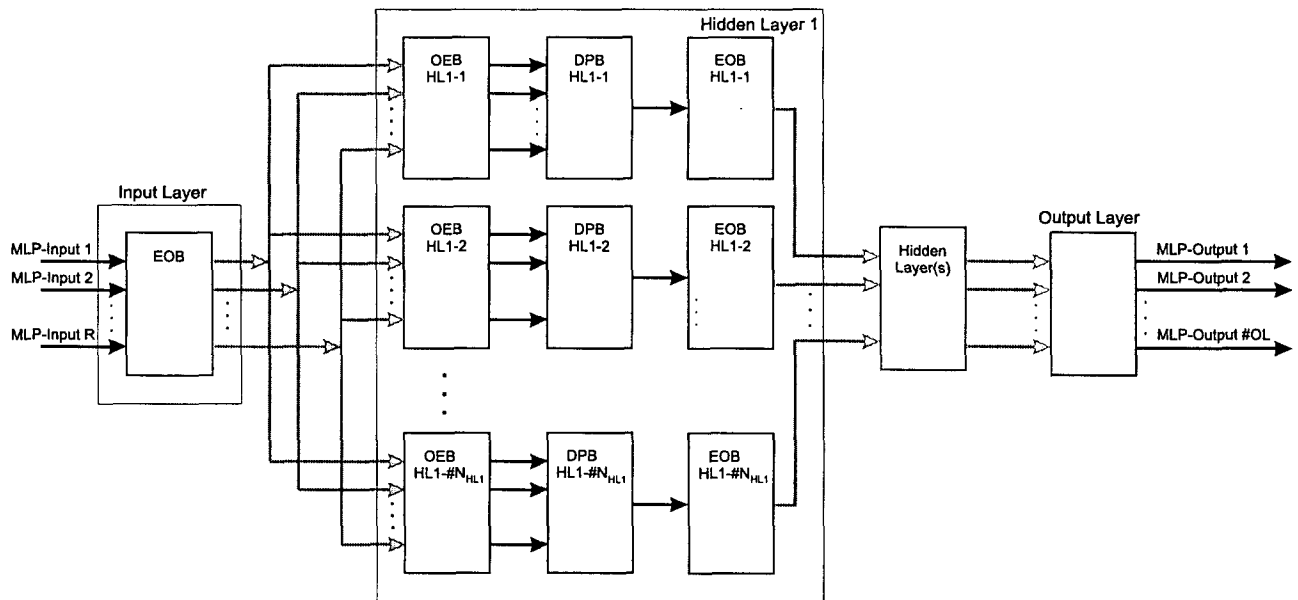


Figure 1: Overall structure of the MLP with detailed view of the first hidden layer.

4 Performance Estimation

The performance of a neural network hardware architecture can be characterised by its size and its speed. The size is usually expressed by the number of connections (#connections) in the network and the speed is described by the 'Connections per Second' (CPS) measure. To determine the number of connections in the network first the number of connections (#connections) in each layer *l* is calculated

$$\#connections_l = \#neurons_l \cdot \#synapses_l \quad (10)$$

with $l \in (1, 2, \dots, \#HL)$. The number of neurons (#neurons) of a layer *l* is

$$\#neurons_l = nc \cdot \#N_l \quad (11)$$

and the number of connections (#synapses) that originate from one neuron in the layer *l* is

$$\#synapses_l = nc \cdot \#N_{l+1} \quad (12)$$

with $l \in (1, 2, \dots, \#HL)$. #synapses_{*l*} states how many neurons in the following layer are connected to one neuron. With fully connected MLP-networks these are all neurons of the following layer.

From (11) and (12) follow the number of connections of one layer

$$\#connections_l = nc^2 \cdot \#N_l \cdot \#N_{l+1} \quad (13)$$

To eventually determine the number #connections of connections in the whole network the sum of all number of connections in the single layers is build

$$\#connections = \sum_{l=0}^{\#HL} \#connections_l \quad (14)$$

$$= nc^2 \cdot \sum_{l=0}^{\#HL} (\#N_l \cdot \#N_{l+1}).$$

The speed of the network can now be determined by the number of connections and the processing time of the whole network

$$CPS = \frac{\#connections}{T_{NN}} \quad (15)$$

With (9) and (14) the speed becomes

$$CPS = \quad (16)$$

$$nc^2 \cdot \sum_{l=0}^{\#HL} (\#N_l \cdot \#N_{l+1})$$

$$\frac{((2 \cdot ac + 3) \cdot (\#HL + 1) + nc \cdot \sum_{i=1}^{\#HL+1} \#N_i + ac) \cdot nc \cdot T_E}{}$$

Here we estimate the best case performance. When weight values are set to zero the performance is reduced accordingly. Still a general performance evaluation can be done.

4.1 Technological limits to implementation

The maximum number of neurons and layers are limited by the implementation technology: the former is the maximum number *RM* of sender and receiver cells per area and the latter the maximum Fan-Out *FOM* for each transmitting unit. The limits are

$$\#IO \leq RM \quad (17)$$

and

$$\#N_{RL} \leq FOM \quad (18)$$

The number #IO of sender and receiver cells per area is determined as

$$\#IO = bl \cdot \#N_{SL} \cdot (1 + \#N_{RL}) \quad (19)$$

Values from literature for minimisable structures indicate minimum pitch sizes (size transmitter or receiver cells) of between 62.5µm and 125µm [4].

For switching a gate in 0.8 µm CMOS technology a current *I_{ph}* of 20 µA is sufficient. This value was determined by our own SPICE simulation experiments. Using (20) allows us to determine the maximum number *FOM* of diodes for receiving such a signal [5, 6].

$$P_{opt} = \frac{I_{ph}}{R} \quad (20)$$

R denotes the responsivity of the receiving photodiodes, a typical value of *R*=0.35A/W [7] results in an optical transmission power *P_{opt}* of 57 µW. VCSEL (Vertical-Cavity Surface-Emitting Laser) diodes are capable of generating a *P_{opt}* of up to 5mW [8]. Without amplification of the signal 87 diodes can receive such a signal at a time.

$$FOM \leq 87 \quad (21)$$

Using the power output of VCSELs and respective amplification circuits the number of receiving diodes can be multiplied, e.g. with using fingered photodiode structures in average light pulses can be received with a 3dB rate of 300 Mbit/s by requiring only 5.6 µW light input power (λ=860 nm) for one diode [9].

We are only looking at the non amplified case in this study.

For implementing a higher number of neuronal cells multiplying the data transmission between the layers is possible: Several neurons are combined and the data for each of these neurons will be transferred one after another as described in chapter 3.1. By increasing the number of neurons using multiplexing the transmission time is increasing as well as the area needed for implementing the activation state storage, weight storage, and additional control circuitry. Still, by using multiplexing techniques, a large enough number of neurons per layer can be implemented when necessary.

4.2 Number of hidden layers

Using a MLP with two hidden layers is sufficient to approximate every nonlinear and continuous function. Therefore the number of hidden layers #HL can be set to two:

$$\#HL = 2 \quad (22)$$

For reasons of modularity it is advisable to build each layer in a hardware component and to implement the neural network by combining several single layers. Therefore an identical maximum number of neurons will be implemented for each layer (23). Neurons that are not required can be 'switched off' by setting their corresponding multiplication weights to zero.

$$\#N = \#N_i \quad (23)$$

with $i \in (1, 2, \dots, \#HL + 1)$.

Hence, by using (22) and (23) the propagation time T_{NN} of the network and the number $\#connections$ of connections can be simplified:

$$T_{NN} = (7 \cdot AC + 9 + 3 \cdot NC \cdot \#N) \cdot NC \cdot T_E \quad (24)$$

$$\#connections = 3 \cdot nc^2 \cdot \#N^2 \quad (25)$$

By using equations (24) and (25) the speed becomes

$$CPS = \frac{3 \cdot nc \cdot \#N^2}{(7 \cdot ac + 3 \cdot nc \cdot \#N + 9) \cdot T_E} \quad (26)$$

4.3 Division factor ac of the bit length

The maximum number of receiving cells is limited by technology (see chapter 4.1). Looking at the nonamplified case we see that the maximum Fan-Out FOM is 87. Therefore a maximum of 87 receiver cells can be attached to one sending cell.

Suitable values for ac are 1, 2, 4 and 8 (using an activation function represented with the wordlength $b=8$). A wordlength of 8 bit allows for sufficient precision for a MLP used in retrieve mode [10]. To achieve a natural number result for the division with ac the maximum value of Fan-Outs is assumed to be 80. With a division factor of $ac=8$ a maximum number of 80 receiving neurons can be connected to one sending neuron. With (19) this results in 6,400 sending and receiving cells. By using a fixed number of sending cells in the area of the FOM and with lower division factors the number of receivers can be further reduced, which facilitates realisation. At the same time the number of neurons decreases.

With these values the maximum number of physical neurons and the number of sender and receiver cells are as shown in Table 1.

Table 1: Maximum number of physical neurons per layer for different division factors

Division factor ac	1	2	4	8
Maximum number of neurons per layer $\#N$	10	20	40	80
Number $\#IO$ of sender and receiver cells per layer	880	1,680	3,280	6,480

4.4 Overall propagation time

The overall propagation time of the neural network was determined in (24). Furthermore we now have to determine T_E . Its minimum value depends on the technological implementation. The following components are used for implementing:

- an input activation state memory in the optical to electronic block,
- an output activation state memory in the electrical to optic block,
- the weight / bias storage,
- the multiplier,
- the accumulation adder,

- a look-up table for the transfer function in the processing unit and
- the overall control logic for the layer.

Estimating the timing for these components using [11] and [12] indicates that the multiplier is the component which is likely to have the longest propagation time. An 8x8-bit Pezaris-multiplier [11, 12] was implemented using VHDL code. This is a very fast and size efficient multiplier architecture. Using the Synergy package of CADENCE for synthesis with the AMS Hit Kit standard cell 0.8 μm CMOS process, the maximum propagation time for the longest path in the multiplier is 5.24 ns.

Also all other components have been implemented using VHDL and synthesised for an 8 bit wide signal representation. This was done to study the implementation feasibility of the universal architecture presented in this paper: A prototype hardware for one layer of ten neurons was designed using a novel implementation approach based on standard cells [13].

The respective longest path propagation of the above components is well below the 5.24 ns of the Pezaris multiplier. For further estimation purposes T_E is now set to 6 ns.

In this case the propagation times T_{NN} for different values of ac and nc according to (24) become as in Table 2.

4.5 Performance

With using the estimate $T_E = 6ns$ CPS as well as $\#connections$ can be determined for different values of ac and nc . Figure 7 shows $\#connections$ and Figure 9 shows CPS for various values of ac and nc .

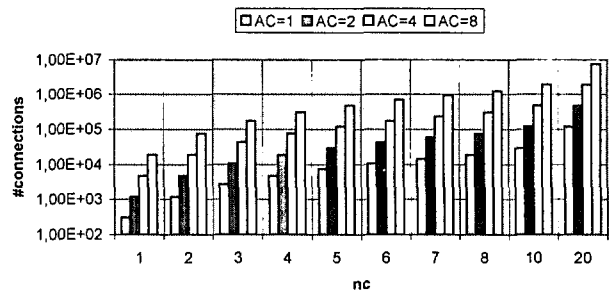


Figure 7: $\#connections$ for different values of ac and nc with $T_E = 6$ ns.

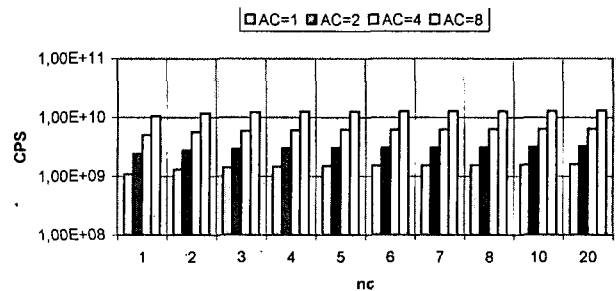


Figure 8: CPS for different values of ac and nc with $T_E = 6$ ns.

Table 2: Propagation times in seconds for different values of ac and nc in seconds.

$ac, \#N$	1, 10	2, 20	4, 40	8, 80
$nc=1$	2,76E-07	4,98E-07	9,42E-07	1,83E-06
$nc=2$	9,12E-07	1,72E-06	3,32E-06	6,54E-06
$nc=3$	1,91E-06	3,65E-06	7,15E-06	1,41E-05
$nc=5$	4,98E-06	9,69E-06	1,91E-05	3,80E-05
$nc=6$	7,06E-06	1,38E-05	2,73E-05	5,42E-05
$nc=8$	1,23E-05	2,41E-05	4,79E-05	9,53E-05
$nc=9$	1,90E-05	3,74E-05	7,42E-05	1,48E-04
$nc=20$	7,39E-05	1,47E-04	2,92E-04	5,84E-04

4.6 Comparison with existing hardware

Figure 9 illustrates the performances of several neural processing architectures. Performances of serial implementations, implementations based on conventional, analog and digital neurochips are drawn as shaded surfaces (the values are derived from publications [14]).

For comparison four biological implementations are also included in the figure. The crosses indicate the performances of the presented digital optoelectronic neurohardware for different values of ac and nc . The figure shows that the performance is on the level of the most powerful type of electronic implementations (analog neurochips). The optoelectronic hardware is up to one magnitude higher in speed (CPS) than the presented digital electronic implementations. Hence the optoelectronic implementation shows a performance that is in the range of analog neurohardware with a precision that is equal to digital neurochips. For up to 10^4 connections the processing speed is superior even to analog neurochips.

These values could be further improved by parallelising methods mentioned in section 3.2, by means of amplifier circuits for the optic receiver cells and by using higher integrated electronic standard cell packages.

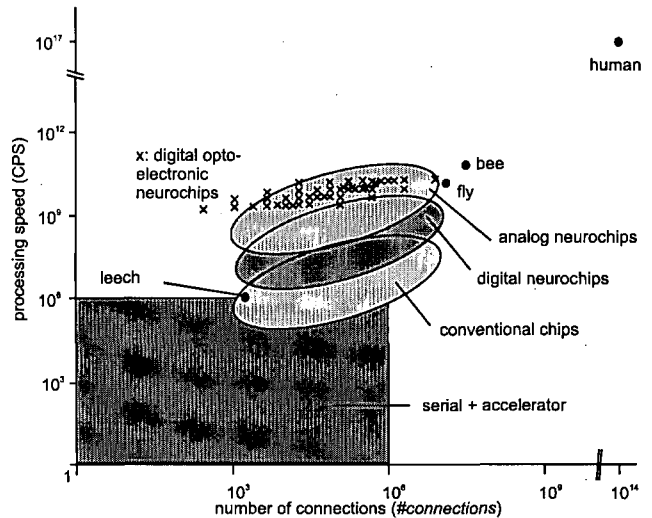


Figure 2: Performances of neurocomputers (after [14], altered)

5 Conclusion

A novel digital neural network hardware based on MLPs has been introduced. For this purpose a suitable optical interconnect system and an electronic VLSI architecture for one single layer was presented. Performance values for variable parameters have been calculated to study the potential of such an optoelectronic hardware and actual performance rates using current technology have been determined. The results indicate that this new hardware has a performance potential that is up to one magnitude higher than the performance of conventional digital electronic neurochips.

References

- [1] K.D. Maier, V. Glauche, C. Beckstein, R. Blickhan. Controlling one-legged dynamic movement with MLPs, International ICSC / IFAC Symposium on Neural Computation (NC'98), Vienna, September 23rd-25th, 1998, pp. 784-790.
- [2] D. Rumelhart, J. McClelland and the PDP Research Group. Parallel Distributed Processing, Volume 1, The MIT Press, 1986.
- [3] S. Haykin. Neural Networks, Macmillan Publishing Company, 1994.
- [4] A.V. Krishnamoorthy et. al., "3-D integration of MQW modulators over active sub-micron CMOS circuits: 375 Mbit/s transimpedance receiver-transmitter circuit", IEEE Photonics Technology Letters, 7, 11, 1995, pp. 1288-1290.
- [5] R. Paul. Optoelektronische Halbleiterbauelemente. B.G. Teubner, Stuttgart, 1992.
- [6] C.R. Pollock. Fundamentals of optoelectronics. Chicago, Irwin, 1995.
- [7] H. Berger, J. Sturm, et. al., "Contactless Function Test of Integrated Circuits on The

- Wafer - General Survey", Proceedings Microsystems 95, 1995
- [8] Data sheet Mitel Semiconductor for VCSEL Array 4D469, 1998
 - [9] M. Kuijk, D. Coppée, R. Vounckx. Specially modulated light detector in CMOS with sense-amplified receivers operating at 180 Mbits/s for optical datalink applications and parallel optical interconnects between chips. *IEEE Journal of Selected Topics in Quantum Electronics*, 4, 6, 1998, pp. 1040-1045.
 - [10] J.L. Holt, J.-N. Hwang. Finite Precision Error Analysis of Neural Network Hardware Implementations, *IEEE Transactions on Computers*, Vol. 42, No. 3, 1993.
 - [11] P. Pirsch. *Architekturen der digitalen Signalverarbeitung*, Teubner, 1996.
 - [12] K. Hwang. *Computer Arithmetic. Principles, Architecture and Design*, John Wiley and Sons, 1979.
 - [13] K.D.Maier, C. Beckstein, R. Blickhan, W. Erhard. Standard cell-based implementation of a digital optoelectronic neural network hardware, *Applied Optics*, Vol. 40, No. 8, in print.
 - [14] M. Glesner, M. Pöschmüller. *Neurocomputers. An overview of neural networks in VLSI*. Chapman & Hall, Neural Computing Series, London, 1994.

Information technology: The lie groups defining the filter banks of the compact disc

Ernst Binz
Lehrstuhl für Mathematik I, Universität Mannheim,
D-68159 Mannheim, Germany

Walter Schempp
Lehrstuhl für Mathematik I, Universität Siegen,
D-57068 Siegen, Germany

Keywords: IT, CD

Received: November 16, 2000

The versatility of the compact disc (CD) has quickly become apparent to manufacturers and users alike. Exceeding the expectations of even its most ardent supporters, the CD holographic disc storage system has become one of the most successful consumer electronics products ever introduced. The phenomenal success of the audio CD on the eager worldwide marketplace has encouraged rapid development of CD technology and spawned entirely new high tech applications for the dimpled disc. The Mini Disc (MD), for instance, occupies about one-fourth the area of the standard CD-Digital Audio (CD-DA) format yet provides an identical playing time through efficient data reduction. The essence of digital audio lies in its numerical basis. It is the aim of the present paper to elaborate the mathematical principles underlying the audio CD as far as they are concerned to the format's electronic and holographic principles.

1 Introduction

An important aspect of all information transmission is the storage and detection of encoded information. Information technology (IT) deals with the implementation of these modalities. Specifically storing audio information places great demands on a digital medium. A 60-minute musical program recorded in stereo channel modality with pulse-code modulation (PCM) at a standard time sampling frequency of 44.1 kHz and with 16-bit amplitude quantization level every 23 μ s, for instance, generates over 5 billion bits of information in all. Because error correction, synchronization and modulation are obligatory requirements for successful audio information storage and detection, the total capacity required with random access capability is over 15 billion bits.

The original Compact Disc-Digital Audio (CD-DA) format was developed to meet these demands at low costs. The CD-DA system has become one of the most successful high precision microelectronic devices ever introduced. Specifically, more than one billion CD-DAs are sold every year, and LPs have all but vanished. Because it contains the same audio information, bit for bit, that was recorded in the studio, the CD-DA has been growing rapidly in popularity since it was launched in the early 1980s. The commercially available CD players represent prime examples of the benefits of digital microelectronic chips and integrated optoelectronic systems. They are perhaps the most sophisticated and microelectronically subtle high tech pieces of

audio equipment to ever reach the consumer.

Today, the compact disc family encompasses alternative format specifications such as the Compact Disc-Read Only Memory (CD-ROM) for professional databases and mass storage for computer-related applications, Digital Versatile Disc-Read Only Memory (DVD-ROM) for advanced high density storage of high fidelity audio-video frames, and various other types of CD formats among which the interactive compact disc (CD-I) format is a special specification of the CD-ROM format ([4], [21]). Actually CD-ROMs which store 600 megabytes on one side of the 5 inch disc are the logical extension of the CD-DA format toward the broader application of information storage on a digital medium. The CD-ROM standard, unlike the CD-DA standard, does not link CD-ROM to any specific application of IT. The format is thus transparent and offers a cost-effective way of distributing large amounts of information, especially information not requiring frequent updating. Of course, various more advanced concepts such as the computationally highly demanding Mini Disc (MD) Adaptive Transform Acoustic Coding (Atrac 3) format, supported by the technological *and* signal theoretic progress made since the launching of the original CD-DA format, are under the way to open a new market for the information storage and retrieval industry.

Elementary signal theoretic techniques such as the time sampling process and amplitude quantization modes of IT form the basis of digital audio. Without them, the CD-DA system would not be a viable reality. Both sampling

and quantization are parameters which determine the limitations of an audio digitization transducer. Therefore all digital audio system architectures use these parameters to record and reproduce signals.

There are various different ways to encode serial strings of digital data ([16]). Modulation is the process of encoding source information prior to transmission and detection via information channels and storage. Among these techniques, PCM is one of the most efficient high performance encoding methods. The PCM hardware design is routinely used for telemetry of images from space vehicles and forms the most popular digital audio system architecture, owing to its error-free properties. PCM is a modulation process in which the instantaneous amplitude of an analog signal is converted to a binary number by a A/D converter and then transmitted as a serial string of bits. The encoded signal is fully compatible with digital circuitry which is usually designed to operate with a binary code. Because of its efficient use of bandwidth and its compatibility with off-the-shelf circuitry, PCM has proven to be an expedient means of representing audio data for recording and retrieval.

The PCM format like various other coding schemes of IT requires wider bandwidth than the corresponding analog signal. However, PCM data is easily multiplexed, that is, several data channels may be merged to form *one* channel of data. Therefore the majority of digital recordings are mastered on PCM digital audio recorders. At the output of a CD player which provides access to any part of the audio program within a second or less, the data returns to its PCM format at the digital-to-analog (D/A) converter. A D/A converter does the opposite IT job of an analog-to-digital (A/D) converter. It takes a train of binary-coded words as its input and produces a continuous-time output proportional to the value of the digital input by means of the impulse response of the Heaviside zero-order hold. The input sample-and-hold (S/H) circuit, sometimes called an aperture circuit, which is the next integrated circuit following the D/A converter and installed on the same D/A micro-electronic chip, performs a hold function to buffer instability in the analog signal and correct for high-frequency roll-off. When the D/A output voltage is stable and any glitches have passed, the S/H output forms a pulse amplitude staircase signal.

The S/H circuit is essentially made of a capacitor and switch. It tracks the signal until the sample command causes the switch to open, isolating the capacitor from the signal. The capacitor holds this analog voltage during conversion. The timing of the sample command must be carefully controlled to prevent jitter, the phenomenon of imprecise sample times. Then the pulses in the output are the width of a sampling period. Reconstruction requires pulses of infinitely short duration. This is impossible to achieve because it would require infinitely large current amplitude flow. Because of the finite duration of the output samples, a filtering effect occurs in which the amplitude response declines to zero at the sampling frequency. This is beneficial because image spectra are attenuated.

To summarize the hardware design which realizes an audio PCM digitization transducer, its recording section consists of input amplifiers, a dither generator, input anti-aliasing low-pass filters, S/H circuitry, A/D converters, a multiplexer, digital processing and modulation circuitry, and a storage medium such as optical disc. On the digitization transducer's output side are demodulation and processing circuits, a demultiplexer, D/A converters, S/H aperture circuitry, output anti-imaging low-pass filters, and output amplifiers. From the mathematical point of view, the transducer's input-output reflection symmetry is of particular importance.

Usually, the mathematical interest of digital audio storage media such as the CD-DA format is restricted to the error correction which is performed by the cornerstones of error correction: interleaving and parity. Interleaving is employed to guard against the occurrence of burst errors. The parity bit added to every data word represents the redundancy contained in the correction codes. The parity bit is chosen so that the total numbers of ones and zeros in the data word plus parity bit is even or odd. Due to extra data created from the original data to help detect errors, the chance to correct errors is easier with digital data than with acoustic analog signals. The particular algorithm used in the CD-DA information transmission channels is the Cross Interleave Reed-Solomon Code (CIRC). The CIRC circuit uses two correction codes for additional correcting capability and three interleaving stages to encode data before its placed on the disc. Similarly, CIRC performs error correction while decoding the serial string of data during playback.

Upon playback, following demodulation, data is transmitted to a CIRC decoder for de-interleaving, error detection, and correction. The CIRC decoding process utilizes parity from two Reed-Solomon decoders and scatters consecutive errors by de-interleaving. In this way, errors become more likely random errors which are more easily corrected.

In contrast to the treatments of coding algorithms, the present paper deals with the *temporal* data readout of the CD-DA. It shows that the timing defined by the real Heisenberg nilpotent Lie group G is behind the readout procedure and that the metaplectic group leaving the one-dimensional center C of G pointwise fixed dictates the optical focusing as well as the discrete time data sampling process of IT ([23], [29]). Although there are, independently of the summation formulae approach, very short proofs of the time sampling theorem available, the approach based on a *central* projection has to be given preference, at least from the methodological *and* the epistemological point of view, because G is a relatively elementary non-compact non-abelian Lie group which allows to define in a *conceptual* way the filter banks of IT.

Next to audio information, visual information plays a major role in human communication and orientation. It is estimated that about 80 % of all information received is of visual nature. It is not surprising, therefore, that with the advent of electronic data processing the desire arose

for the acquisition, processing and analysis of pictures and sequences of images by digital computers. Since the first trials some 30 years ago, image processing has developed into a broad scientific discipline which intensively interacts with several other topics such as phase coherent optics, quantum statistics of radiation, information theory and signal processing, pattern recognition, and artificial intelligence ([5], [6], [7], [8]). As an outlook to advanced phase coherent summation imagery, the paper refers to a two-dimensional imaging implementation of this system which leads to the non-invasive diagnostic modality of clinical magnetic resonance tomography ([24], [25], [26]). In fact, an understanding of the CD-DA system forms an excellent preparation for the understanding of the much more sophisticated modality of clinical magnetic resonance imaging (MRI) and synthetic aperture radar (SAR) imaging ([8], [20]).

The theory of Lie groups and Lie algebras is a fundamental part of mathematics because it allows to investigate basic symmetry principles. According to Wolfgang Ernst Pauli (1900-1958), symmetry forms the fundamental organising principle of physics and the natural sciences. In signal processing, symmetries are used to implement *fast* processing algorithms by sophisticated special-purpose processors. Among these, spectrum analyzers implementing the fast Fourier transform (FFT) are particularly popular ([16]). Albert Einstein's intuitive treatment of relativity was followed shortly by a more sophisticated treatment by Hermann Minkowski (1864-1909) in which Lorentz transformations were shown to constitute a Lie group of rotational collineations. Similarly, shortly after Werner Karl Heisenberg (1901-1976) introduced his famous Commutation Relations in quantum physics, which underlie his Uncertainty Principle, Hermann Weyl used the Lie commutator bracket $[\cdot , \cdot]$ to show that they could be interpreted as the structure relations

$$\left[\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \right] = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for the canonical basis of the real Lie algebra $Lie(G)$ of the real Heisenberg nilpotent Lie group G . This paper presents an introduction of harmonic analysis on the "almost abelian" Heisenberg group G to information theory with an outlook to the enchanting area of theta identities, such as the Jacobi and the Landsberg-Schaar identities, and the field of phase coherent summation imagery which is conceptually based on the notion of filter bank well known from multirate signal analysis or subband coding.

2 The data readout procedure of it

Many methods of audio storage and detection have evolved since Thomas Alva Edison (1847-1931) made the first audio recording in 1877 on a cylinder covered with tin foil. Early acoustical recordings were made on wax cylinder and

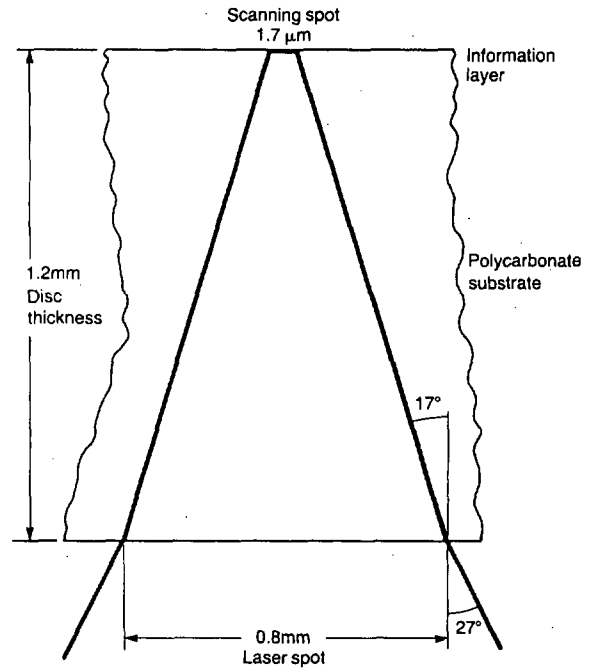


Figure 2: For data readout, the semiconductor laser beam passes the disc substrate. The refractive index of the substrate contributes to the optical focusing of the laser beam. The transparent plastic substrate forms most of the CD's 1.2 mm thickness.

shellac disc. Subsequently, numerous magnetic tape formats were developed. However, all of these audio systems recorded and reproduced analog signals by using a mechanical pickup.

In IT the CD is certainly one of the most advanced storage media available. The CD-DA format stores its information digitally and uses a laser optoelectronic pickup. The length of its data represents the binary bits which represent the original audio signal. A laser beam of carrier frequency ν is focused to read the data stream. The data is physically contained in the disc's pits which are impressed along its top surface and are covered with a 50 to 100 nm metal layer. The data storage in pits on a flat surface is not directly visible to the naked eye. A scanning electron microscope is needed to get a sufficiently good look on the track of pits (Figure 1) arranged in a continuous spiral running from the inner circumference to the outer one. Another 10 micrometers to 30 micrometers plastic layer protects the metalized pit surface. The laser beam is focused on the metalized data surface embedded inside the disc and passes through the transparent plastic substrate and back again.

The fact that the laser beam passes the disc substrate provides one of the significant assets of the CD system. The plastic substrate has refractive index of 1.55 whereas air has normalized refractive index 1.0. The speed of light slows from $c = 3 \cdot 10^8$ m/s to $c' = 1.9 \cdot 10^8$ m/s and changes the carrier frequency ν to the fraction νy . It is the *local*

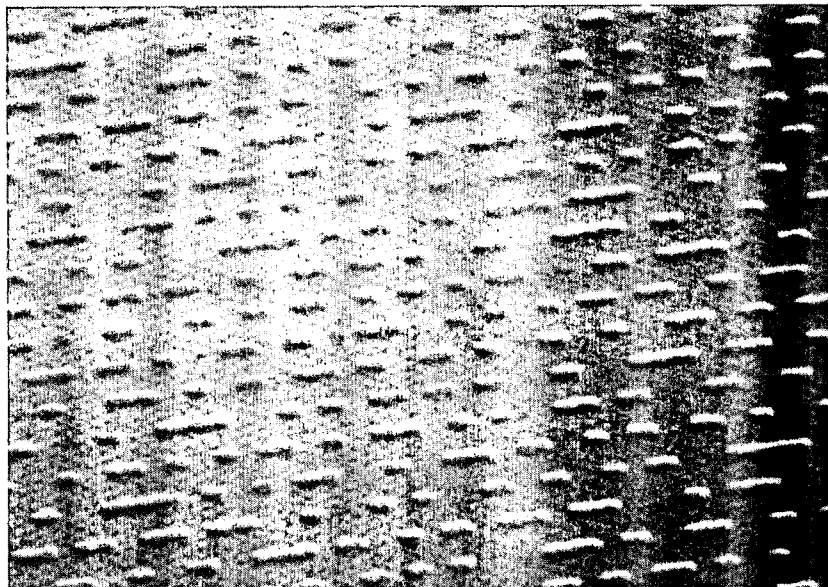


Figure 1: Visualization of the tracks of pits on the metalized CD surface by use of a scanning electron microscope. The horizontal line of the scan indicates a scale of $5\ \mu\text{m}$. The track pitch, the distance between adjacent laps of the pit spiral, is $1.6\ \mu\text{m}$ so that there are about 600 tracks to a mm. Each pit has a width of about $0.5\ \mu\text{m}$. In comparison, the cross-section of a human hair has a width of $75\ \mu\text{m}$. The minimum pit length is $0.833\ \mu\text{m}$ to $0.972\ \mu\text{m}$, the maximum pit length is $3.05\ \mu\text{m}$ to $3.56\ \mu\text{m}$ so that a track of pits might contain about 3 billion pits precisely arranged on a spiral. Unspiraled, the track would stretch about 3.5 miles. Because each outer track revolution contains more pits than each inner track revolution, the CD must be slowed down as it plays in order to maintain a constant rate of data. Based on the timing of the master clock, the CD player automatically regulates the disc rotational speed to maintain a constant bit rate of 4.3218 MHz.

frequency νy which characterizes the frequency modulation of the carrier frequency ν induced by the substrate. Because of the refractive index, the thickness of the CD, and the numerical aperture of the objective lens, the size of the laser beam on the disc surface is approximately $2\ \mu\text{m}$. Hence, the laser beam is focused to a point slightly larger than a pit width but does not overlap the tracks of pits (Figure 1).

The reflective flat surface, called land (Figure 2), causes almost ninety percent of the laser light to be reflected into the optoelectronic pickup. When considered from the laser's perspective, the pits are viewed as tracks of bumps. The height of each bump is between $0.11\ \mu\text{m}$ and $0.13\ \mu\text{m}$. This height is slightly smaller than the semiconductor laser's wavelength $\lambda = 780\ \text{nm}$ in air. Inside the polycarbonate substrate with a refractive index of 1.55, the laser's wavelength is about $\lambda' = 500\ \text{nm}$. The height of the bumps is therefore approximately $\frac{1}{4}$ of the laser's wavelength λ' inside the disc substrate.

Notice that light striking land travels a distance $\frac{1}{2}\lambda$ further than light striking a bump. This creates a *phase* difference

$$x = \frac{1}{2}\lambda$$

between the part of the beam diffracted from the bump and the part reflected from the surrounding land (Figure 3). The phase difference x causes the two parts of the beam

destructively interfere with and cancel each other. Actually each pit edge, whether leading or trailing, is a one and all areas in between, whether inside or outside a pit, define zeros. This is a more efficient storage technique than coding the binary bits directly with pits. Combinations of the varying lengths of the pits encode the binary data stream to be read by the semiconductor laser beam. Because the readout procedure is non-invasive, the CDs are completely immune to damage from repeated playing.

A CD might contain 3 billion pits precisely arranged on a spiral track. The optoelectronic pickup has to focus on, track, and read that data spiral. To achieve sharp focus within a $\pm 2\ \mu\text{m}$ tolerance on the data surface, and proper frequency modulation for the definition of the disc data y , a monochromatic illumination of the CD data surface is required. Otherwise the phase interference between the direct and reflected laser light is lost along with the audio data, as well as the tracking information, and, ironically, the focusing information itself. The objective lens must therefore be able to refocus as the disc surface deviates vertically. A servo-driven auto-focus system manages this control problem.

Auto-focus control is an absolute prerequisite in a laser optoelectronic pickup system. Disc warpage and other irregularities would place the data out of the pickup's depth of focus, making it impossible to create the necessary phase interference pattern with the pit height and land. Specifi-

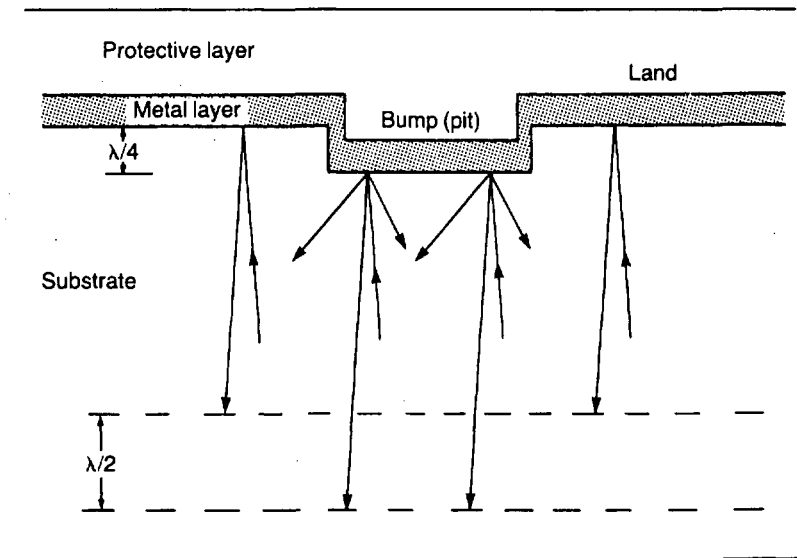


Figure 3: Data is physically contained in pit tracks which are impressed along the CD top surface and are covered with a thin 50 to 100 nm metal layer. Another thin 10 to 30 μm plastic layer protects the metalized pit surface. A semiconductor laser beam is used to read the data. In the laser illumination, a pit height causes a wavelength path difference of $x = \frac{1}{2} \lambda$ relative to surrounding land. The laser optoelectronics, servo system for automatic focusing and tracking, control micro-processors, and integrated output D/A circuitry are all high, high tech.

cally, phase coherence is vital to implement phase cancellation in the near-infrared light beam produced by disc pits so that disc data x can be read. Both specifications can be accomplished by a laser diode placed at the focus of the collimator lens with a long focal length. A monitor diode stabilizes the semiconductor laser's output.

Any optoelectronic pickup system must, of course, control both tracking and focusing simultaneously. When the auto-focus is not operative, the system pulls the objective lens back to prevent damage to the lens or CD.

Because the parameters x and y have timelike character, a spindle motor is used to rotate the CD with constant linear velocity, a phase coherency condition in which a uniform relative velocity is maintained between the disc and the pickup. To achieve this, the rotation speed of a CD has to vary depending on the position of the pickup underneath the surface. Because each outer track revolution contains more pits than each inner track revolution, the CD must be slowed down as it plays in order to maintain a constant rate of data.

When the laser beam is reflected at the revolving disc surface during playback, the response is detected by a photodiode sensor. The optoelectronic pickup's servo loops use electric signals to control motors to mechanically adjust the pickup's position horizontally and vertically, relative to the disc surface. In another servo loop, information from the data itself is used to determine the disc's precise rotational speed, and maintain the proper data stream rate. It is the voltage stemming from the sensor which is ultimately transformed into the analog audio signal output from the

CD player. The encoded data from the pickup must first be decoded.

In the CD-DA player, the numerical aperture of the objective lens, wavelength of the semiconductor laser, thickness and refractive index of the disc, and size and height of the pits all work together to allow data to be read from the disc. The various subsystems in a CD player are closely inter-related with a tightly interlocked timing relationship: The audio data rate is 176.4 kbytes per second. Because there are 24 audio bytes in a frame, the frame rate is 7350 Hz and the master clock 4.3218 MHz. A single master clock is employed for all the signal processing circuitry, including the *oversampling* filter bank and D/A converters. The master clock establishes that the CD forms a *temporal* device. As a result, the data stream of the CD-DA player is *synchronous*, preventing any internal beating.

While CD-ROM uses a data format similar to that of the CD-DA format, the players are not compatible. A CD-ROM player contains laser optics, modulation, and error correction, but D/A conversion and audio output sections are replaced with a computer interface to output the ROM data to a host computer. Data is transmitted to the host computer in blocks of 2 kbytes. Because they are not tied to one specific operating system or data processor, CD-ROM devices can be interfaced with all existing computer systems. CD-ROM is limited only by the capabilities of the operating system and microprocessor of the host computer.

A DVI (Digital Video Interactive) all-digital optical disc is a CD-ROM format containing DVI specific data of reproducing full-motion, full-screen video, computer gen-

erated video graphics, and digital audio via a CD-ROM drive ([28]). Although data on DVI discs is formatted to CD-ROM specifications and can be played on a CD-ROM drive, special DVI decoding technology is required. The DVI format is incompatible with the CD-I format and diverse other CD-ROM implementations. Interestingly, CD-I was the first large volume application of the Moving Picture Experts Group (MPEG) international standard for coded representation of moving pictures, associated audio, and their combinations when used for storage and retrieval on digital storage media. The MPEG audio and video coding algorithms allow use of video sequences coded with a variety of CD-I picture formats, as well as a variety of CD-I audio formats.

3 Harmonic analysis on the real heisenberg lie group

Taking into account the aforementioned parameters detected by the laser optoelectronic pickup, the real Heisenberg group G collects the phase difference x , the local frequency νy , and the real variable z dual to the carrier frequency ν into an upper triangular matrix with real entries

$$g = \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix}.$$

The transversal trace of the element $g \in G$ is given by

$$g_0 = \begin{pmatrix} 1 & x & 0 \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \quad (x \in \mathbf{R}, y \in \mathbf{R}).$$

The group law of G is matrix multiplication

$$g_1 \cdot g_2 = \begin{pmatrix} 1 & x_1 & z_1 \\ 0 & 1 & y_1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & x_2 & z_2 \\ 0 & 1 & y_2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x_1 + x_2 & z_1 + z_2 + x_1 y_2 \\ 0 & 1 & y_1 + y_2 \\ 0 & 0 & 1 \end{pmatrix},$$

so that G forms a non-commutative real Lie group of dimension 3. The inverse of the element $g \in G$ is given by

$$g^{-1} = \begin{pmatrix} 1 & -x & -z + xy \\ 0 & 1 & -y \\ 0 & 0 & 1 \end{pmatrix}.$$

The center $C \hookrightarrow G$ is formed by the subgroup of matrices

$$\begin{pmatrix} 1 & 0 & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (z \in \mathbf{R}),$$

and is therefore isomorphic to the real line \mathbf{R} . The matrix exponential, which maps the Lie algebra $\text{Lie}(G)$ onto G ,

projects the one-dimensional center of $\text{Lie}(G)$ of all matrices

$$\begin{pmatrix} 0 & 0 & \zeta \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (\zeta \in \mathbf{R})$$

onto the real line C . If the elements $g \in G$ are written as triples of real numbers (x, y, z) , then the elements of C are identified with the real numbers $z \in \mathbf{R}$. The non-trivial unitary characters of C are then given by the functions

$$z \rightsquigarrow e^{2\pi i \nu z} \quad (\nu \neq 0).$$

For the next step of reasoning it is important to note that the physical meaning of the elements $g \in G$ and its applications to IT becomes apparent only by unitarily representing the Lie group G .

Let $\mathcal{S}_C(\mathbf{R})$ denote the Schwartz space of infinitely differentiable rapidly decreasing complex-valued functions ψ on the real line \mathbf{R} . For each real number $\nu \neq 0$, the Lie group G acts time generating on the waveform $\psi \in \mathcal{S}_C(\mathbf{R})$ according to the temporal rule

$$\rho_\nu \left(\begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \right) \psi(t) = e^{2\pi i \nu (z + yt)} \psi(t + x) \quad (t \in \mathbf{R}).$$

Then there is a unique unitary linear extension of ρ_ν from $\mathcal{S}_C(\mathbf{R})$ to the standard complex Hilbert space $L^2_C(\mathbf{R})$. The unitary linear representation ρ_ν of G in $L^2_C(\mathbf{R})$ is irreducible. It is called the linear Schrödinger representation of G ([23]) associated to the carrier frequency $\nu \neq 0$. Because the irreducible unitary linear representation ρ_1 is square integrable mod C , it admits a reproducing kernel K allied to the global Frobenius reciprocity theorem ([18], [19]). In terms of multirate signal analysis or subband coding, the transversal linear mappings

$$\rho_1 \left(\begin{pmatrix} 1 & x & 0 \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \right) : L^2_C(\mathbf{R}) \longrightarrow L^2_C(\mathbf{R}) \quad (x \in \mathbf{R}, y \in \mathbf{R})$$

defined by the transversal traces $g_0 \in G$ are called filter bank operators associated to the linear Schrödinger representation ρ_1 of G .

Often the linear Schrödinger representation of G is confused with the Schrödinger equation of quantum mechanics. Whereas the Schrödinger equation is related to the probabilistic detection of signals and therefore not invariant under the action of the Lorentz group, the irreducible linear Schrödinger representation of G describes the time modi of information transmission by phase coherent signal processing. An important consequence of the irreducibility is that the unitary linear representations ρ_ν and $\rho_{\nu'}$ of G are inequivalent for $\nu \neq \nu'$ ([23]). As a consequence, the associated diffraction patterns do not interfere so that audio signals can be conveyed from one device to another with a minimum amount of confusion. Specifically in the DVD-ROM format this consequence of the fundamental Stone-von Neumann theorem ([23]) is used to increase the information content by stacking several transparent slices of pit

tracks. The transmission of digital data streams, however, is a great deal more complicated on account of the potential disagreements of the sampling frequency, the synchronization method used, and the block length. In this case, a time-sharing multiplexing transmission channel transmits or receives frames, each containing left and right channel data alternatively. The transmission rate corresponds exactly to the source sampling frequency. When the time sampling frequency is 44.1 kHz, the CD-DA format allows to transmit 44.100 frames per second. One frame consists of two subframes, labelled left and right stereo channel, each containing 32 bits of audio information.

In the case of a unique slice as in the CD-DA format it is convenient to normalize the carrier frequency scale such that $\nu = 1$ holds. Then the Levi-Civita mapping

$$J : \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & -y & z \\ 0 & 1 & x \\ 0 & 0 & 1 \end{pmatrix} \quad (z \in \mathbf{R})$$

forms an automorphism of period 4 of G which leaves the center $C \hookrightarrow G$ pointwise fixed. Of course, the elements

$$\begin{pmatrix} 1 & 0 & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \in C$$

are the only fixed points of the mapping J . It is not difficult to establish that the Fourier cotransform $\bar{\mathcal{F}} : \mathcal{S}_C(\mathbf{R}) \rightarrow \mathcal{S}_C(\mathbf{R})$ which is defined by the assignment

$$\bar{\mathcal{F}}\psi(s) = \int_{\mathbf{R}} e^{2\pi i s t} \psi(t) dt \quad (s \in \mathbf{R})$$

satisfies the intertwining identity

$$\bar{\mathcal{F}}^{-1} \circ \rho_1(g) \circ \bar{\mathcal{F}} = \rho_1(J(g))$$

for all matrices $g \in G$. It follows that the linear Schrödinger representation ρ_1 is isomorphic to the unitary linear representation $\rho_1 \circ J$ of G . The unitary isomorphism is given by the Fourier cotransform $\bar{\mathcal{F}}$. It is immediate that its inverse is associated with the automorphism

$$J^{-1} : \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \rightsquigarrow \begin{pmatrix} 1 & y & z \\ 0 & 1 & -x \\ 0 & 0 & 1 \end{pmatrix} \quad (z \in \mathbf{R})$$

of G , and the Fourier transform $\mathcal{F} : \mathcal{S}_C(\mathbf{R}) \rightarrow \mathcal{S}_C(\mathbf{R})$ where

$$\mathcal{F}\psi(t) = \int_{\mathbf{R}} e^{-2\pi i t s} \psi(s) ds \quad (t \in \mathbf{R}).$$

It is immediate that both $\bar{\mathcal{F}}$ and \mathcal{F} admit unique extensions to $L^2_C(\mathbf{R})$. The mappings J and J^{-1} are elements of the metaplectic group and can be associated with the symplectic matrices

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

respectively. It makes sense to denote them also by J and $J^{-1} = -J$, respectively.

The symplectic matrices J and J^{-1} represent turns of 90 degrees of opposite orientation in the affine Euclidean plane $\mathbf{R} \oplus \mathbf{R}$ (Figure 4). They suggest a complexification of the plane $\mathbf{R} \oplus \mathbf{R}$ in order to identify the matrices J and J^{-1} with i and \bar{i} in \mathbf{C} , respectively. Because the sampling process of IT has the symplectic structure of the Levi-Civita matrices, the CD-DA system is actually two-dimensional. This aspect does not only contribute to the high precision, it also makes the understanding of the CD-DA system an excellent preparation of the clinical MRI modality which has enjoyed, along with the art of electronics, an explosive development in the last decades ([24], [16]).

4 The time sampling process of it

Sampling means dividing a signal into evenly spaced discrete points in time and smoothing by linear convolution. If the two-dimensional lattice $\mathbf{Z} \oplus \mathbf{Z}$ is considered as a subgroup of $\mathbf{R} \oplus \mathbf{R}$, periodization of $\rho_1 \bmod \mathbf{Z} \oplus \mathbf{Z}$ transforms

$$G \ni \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \rightsquigarrow \rho_1 \left(\begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \right) \psi(t) \in \mathbf{C}$$

$$(\psi \in \mathcal{S}_C(\mathbf{R}))$$

where $t \in \mathbf{R}$, into the equivalent mapping

$$G \ni \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \rightsquigarrow \sum_{n \in \mathbf{Z}} e^{2\pi i(z+ny)} \psi(x+n) \in \mathbf{C}.$$

It is actually sufficient to assume that the function ψ is continuous on an interval of periodicity and to understand convergence in the distributional sense. The central projection $z = 0$ provides the periodized filter bank operator, and subsequently an application of the mapping J under which the two-dimensional digitization lattice $\mathbf{Z} \oplus \mathbf{Z}$ is invariant provides the Poisson summation formula in its symmetrized form due to G.H. Hardy ([11])

$$e^{\pi i x y} \sum_{n \in \mathbf{Z}} e^{2\pi i n y} \bar{\mathcal{F}}\psi(n+x) = e^{-\pi i x y} \sum_{n \in \mathbf{Z}} e^{2\pi i n x} \psi(n-y)$$

$$((x, y) \in \mathbf{R} \oplus \mathbf{R}).$$

The projection $y = 0$ of the two-dimensional digitization lattice $\mathbf{Z} \oplus \mathbf{Z}$ yields the Poisson summation identity ([29])

$$\sum_{n \in \mathbf{Z}} \bar{\mathcal{F}}\psi(n+x) =$$

$$\sum_{n \in \mathbf{Z}} \psi(n) e^{2\pi i n x}$$

Suppose that $\bar{\mathcal{F}}\psi \in L^1_{\mathbf{C}}(\mathbf{R})$ holds. Since

$$\int_{\mathbf{R}} |\bar{\mathcal{F}}\psi(s)| ds = \sum_{n \in \mathbf{Z}} \int_n^{n+1} |\bar{\mathcal{F}}\psi(x)| dx = \sum_{n \in \mathbf{Z}} \int_0^1 |\bar{\mathcal{F}}\psi(n+x)| dx$$

is finite, the series on the left hand side of the Poisson summation identity converges for almost all $x \in \mathbf{R}$ to a periodic integrable function. Invoking a mathematical principle first explicitly enunciated and systematically exploited by Erich Hecke: "A periodic function should always be expanded in a Fourier series", the k th Fourier coefficient is given by resonance

$$\sum_{n \in \mathbf{Z}} \int_0^1 e^{-2\pi i k x} \bar{\mathcal{F}}\psi(n+x) dx = \int_{\mathbf{R}} e^{-2\pi i k x} \bar{\mathcal{F}}\psi(x) dx = \psi(k) \quad (k \in \mathbf{Z}),$$

the term-by-term integration of the Laurent joined series being justified by the dominated convergence theorem. Hence the Laurent joined series on the right hand side of the Poisson summation identity is the Fourier series of the function on the left hand side. Multiplication by the character

$$s \rightsquigarrow e^{-2\pi i x s} \quad (x \in \mathbf{R})$$

of the additive group \mathbf{R} , and integration over the symmetric unit interval $[-\frac{1}{2}, +\frac{1}{2}]$ gives

$$\sum_{n \in \mathbf{Z}} \int_{-\frac{1}{2}}^{+\frac{1}{2}} e^{-2\pi i x s} \bar{\mathcal{F}}\psi(n+s) ds = \sum_{n \in \mathbf{Z}} \psi(n) \int_{-\frac{1}{2}}^{+\frac{1}{2}} e^{2\pi i(n-x)s} ds = \sum_{n \in \mathbf{Z}} \psi(n) \frac{\sin \pi(x-n)}{\pi(x-n)}$$

and so

$$\sum_{n \in \mathbf{Z}} \psi(n) \frac{\sin \pi(x-n)}{\pi(x-n)} = \int_{-\frac{1}{2}}^{+\frac{1}{2}} e^{-2\pi i x s} \bar{\mathcal{F}}\psi(s) ds.$$

But

$$\psi(x) = \sum_{k \in \mathbf{Z}} \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} e^{-2\pi i x s} \bar{\mathcal{F}}\psi(s) ds = \int_{-\frac{1}{2}}^{+\frac{1}{2}} e^{-2\pi i x s} \bar{\mathcal{F}}\psi(s) ds,$$

and so

$$\psi(x) - \sum_{n \in \mathbf{Z}} \psi(n) \frac{\sin \pi(x-n)}{\pi(x-n)} = \sum_{k \in \mathbf{Z}} (1 - e^{2\pi i k x}) \int_{k-\frac{1}{2}}^{k+\frac{1}{2}} \bar{\mathcal{F}}\psi(s) ds \quad (x \in \mathbf{R}).$$

If ψ is a bandlimited function so that the symmetric spectral condition

$$\bar{\mathcal{F}}\psi(s) = 0 \quad (|s| \geq \frac{1}{2})$$

indicates the cut out procedure performed by the stop-band, the cardinal series representation follows ([22])

$$\psi(x) = \sum_{n \in \mathbf{Z}} \psi(n) \frac{\sin \pi(x-n)}{\pi(x-n)} = \sum_{n \in \mathbf{Z}} \psi(n) \text{sinc}(x-n) \quad (x \in \mathbf{R}).$$

To reproduce ψ from its bi-infinite sequence of samples $(\psi(n))_{n \in \mathbf{Z}}$, the amplitude response function

$$\text{sinc } x = \begin{cases} \frac{\sin \pi x}{\pi x} & \text{for } x \neq 0 \\ 1 & \text{for } x = 0 \end{cases}$$

denotes the *sinus cardinalis* filter of spline theory. It is important to note that the canonical product expansion of the window function

$$\text{sinc} : z \rightsquigarrow \prod_{n \geq 1} (1 - \frac{z^2}{n^2})$$

extends to an even entire holomorphic function of exponential type. Therefore the discrete time data sampling process of IT is closely allied to Carlson's theorem of complex variables which states that the trivial function is the only entire holomorphic function of exponential type $< \pi$ that vanishes at the set of integers (Section 5 below).

The cardinal series representation permits to spot the *multiresolution* flavor of IT. Each sample value is multiplied by the appropriate sinc coefficient corresponding to its contribution to the overall impulse response of the filter. The products are summed to produce the output filtered sample. It thus digitally simulates the impulse response of an analog filter. The time sampling theorem dictates that the frequency content of the audio signal be less than or equal to the half-sampling frequency. The input signal may contain frequencies greater than the half-sampling frequency. A low-pass filter removes high frequencies to produce a spectrum of frequencies below the half-sampling frequency. Using the procedure repeatedly, the final approximation space obtains.

From this procedure another technique of multirate signal analysis called *oversampling* design becomes immediate. The oversampling filter bank is utilized in today's CD

players in which additional sample values are computed by interpolating between original sample values on board a dual-channel linear-phase finite impulse response (FIR) digital filter chip. In view of the fact that additional samples have been generated, the sampling rate of the output signal is greater than the input signal. The spectrum of the signal is changed, with the images appearing at multiples of the oversampled sampling rate. Because the distance between the baseband and sidebands is larger, a gentle analog filter bank design can be used to remove the images without causing phase shift or other artifacts ([21]).

5 A reproducing kernel Hilbert space

Filtering is a fact of life for digital audio systems. An input anti-aliasing filter must precede the sampler to uphold the symmetric spectral condition for bandlimited and thus lossless sampling. Similarly, the output anti-imaging filter must filter out all frequencies above the half-sampling frequency.

The time sampling identity of the filter specified by the sinc function allows an extension to the Paley–Wiener space ([30]). The entire holomorphic functions of exponential type at most π that are square integrable on the real axis forms a complex vector space $\mathcal{PW}(\mathbf{C})$. Under its natural scalar product, the complex Hilbert space $\mathcal{PW}(\mathbf{C})$ is isometrically isomorphic to $L^2_{\mathbf{C}} \left(\left[-\frac{1}{2}, +\frac{1}{2}\right] \right)$. Let the reflection

$$w \rightsquigarrow \bar{w}$$

denote the involutory automorphism of \mathbf{C} with fixed point set \mathbf{R} given by complex conjugation. The uniquely determined reproducing kernel ([13], [19], [27]) of the Paley–Wiener space $\mathcal{PW}(\mathbf{C})$ is defined by the holomorphic-antiholomorphic function of positive type

$$K : (z, w) \rightsquigarrow \text{sinc}(z - \bar{w})$$

on the space $\mathbf{C} \times \mathbf{C}$. The function K reflects the global Frobenius reciprocity by incorporating production and reproduction simultaneously. As a consequence, the convolution representation

$$\psi(z) = \int_{\mathbf{R}} \psi(s) \text{sinc}(z - s) ds \quad (z \in \mathbf{C})$$

holds for every function $\psi \in \mathcal{PW}(\mathbf{C})$. From the convolution representation specified by the sinc filter it is immediate that the Paley–Wiener space $\mathcal{PW}(\mathbf{C})$ of reproducing kernel K is closed under the operation of differentiation.

6 Basic theta identities

A classical example of an application of the Poisson summation identity referred to above is the Jacobi identity for

the theta null function

$$\sum_{n \in \mathbf{Z}} e^{-\pi n^2 \tau} = \frac{1}{\sqrt{\tau}} \sum_{n \in \mathbf{Z}} e^{-\frac{\pi n^2}{\tau}} \quad (\tau > 0)$$

which has also the computational merits of convergence acceleration for small values of the parameter $\tau > 0$ ([2]). As a matter of fact, one of the main benefits of the Poisson summation formula is the systematic supplying of analytic and arithmetic approximations.

Due to its quantum mechanical background, harmonic analysis on the real Heisenberg nilpotent Lie group G is, of course, *not* restricted to the derivation of the time sampling process of IT and the reproducing kernel K of $\mathcal{PW}(\mathbf{C})$ as a nice by-product. An analysis based on the *dual* stochastic aspects of quantum physics ([7], [8]) and the Maslov index, however, leads via the longitudinal evolution operator associated to the Schrödinger equation or the method of path integration of the longitudinally driving Lévy stochastic process to the deep Landsberg–Schaar identity for quadratic Gaussian sums ([1], [9])

$$\frac{1}{\sqrt{p}} \sum_{0 \leq n \leq p-1} e^{2\pi i \frac{n^2 q}{p}} = \frac{1}{\sqrt{2q}} e^{\frac{\pi i}{4}} \sum_{0 \leq n \leq 2q-1} e^{-\pi i \frac{n^2 p}{2q}}$$

$$(p > 0, q > 0),$$

valid for positive integers p and q . The standard proof of the Landsberg–Schaar identity is by putting

$$\tau = 2i \frac{q}{p} + \varepsilon \quad (\varepsilon > 0)$$

and then letting $\varepsilon \rightarrow 0+$ in the Jacobi identity. This method invokes an example of another mathematical principle first explicitly enunciated and systematically exploited by Hecke: "Exact knowledge of the behaviour of a holomorphic function in the neighbourhood of its singularities forms a source of arithmetic theorems". In view of the Lévy–Hinčin spectral trace formula ([3], [17]), the Landsberg–Schaar identity may be considered, however, as *dual* to the Jacobi identity for the theta null function. The geometry of the Lévy–Hinčin formula strikingly disproves that the one-dimensional unitary representations of G are "substantially uninteresting" ([14]). Actually these unitary characters of G are of substantial interest for the *detection* procedure because they represent the collapsed states of phase coherent quantum field theory ([8], [10]). In elementary number theory the Landsberg–Schaar identity plays a central role underpinning key results relating to the law of quadratic reciprocity in terms of Legendre symbols

$$\left(\frac{p}{q}\right) \left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2} \frac{q-1}{2}} \quad (p > 0, q > 0)$$

for *odd* integers p and q , and characters. Indeed, the concept of Heisenberg group G may serve "à une démonstration de la loi de réciprocité quadratique, apparentée à celle

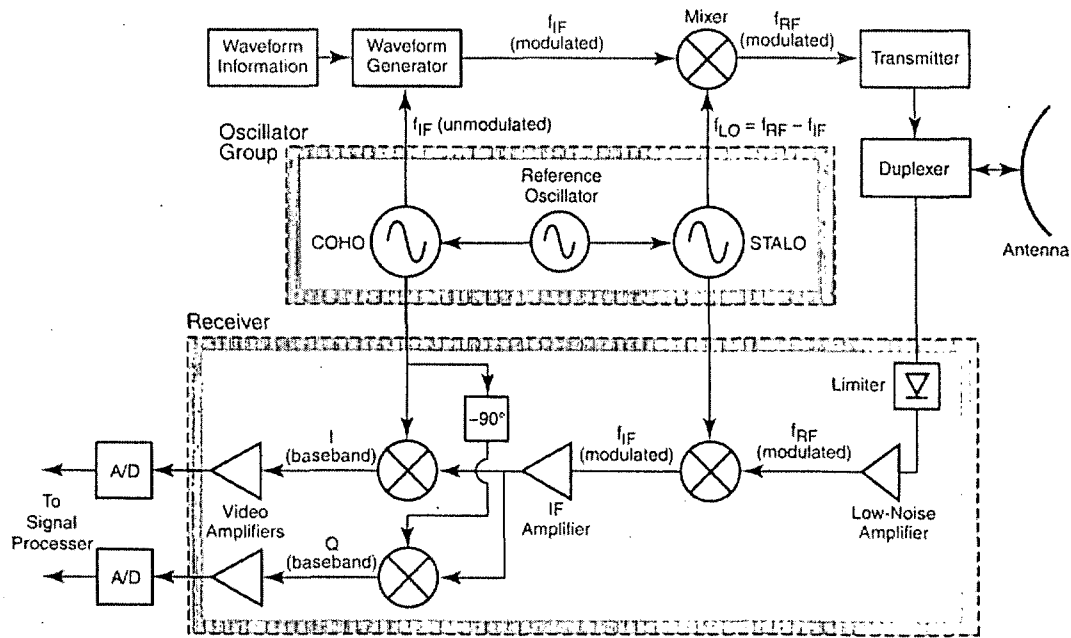


Figure 4: Simplified block diagram of a coherent radar system. The metaplectic group provides coherent signals via mixers which are indicated by the symbol \otimes . In the receiver, the phase delay of -90 degrees implements the element of the metaplectic group associated to the matrix $-J$ (f_{IF} = intermediate frequency, f_{RF} = radio frequency, COHO = coherent oscillator, LO = local oscillator, STALO = stable local oscillator). The two output channels of the receiver feed A/D converters which are the most critical and costly components in a reliable audio digitization transducer.

qui figure au dernier chapitre du livre classique de Hecke sur les corps de nombres algébriques” ([29]). Shor’s algorithm for quantum computing suggests a close interrelation between elementary number theory and quantum information, too.

7 Outlook: phase coherent summation imagery

A central goal of signal processing is to describe real life signals by the concept of filter bank. Filter banks represent coherent arrangements of lowpass, bandpass, and highpass filters used in IT for the spectral decomposition and synthesis of signals. They play an important role in modern signal processing applications because they easily allow the extraction of spectral components of a signal while providing very efficient implementations.

A symplectic extension of the summation formulas leads to filter banks which are at the basis of clinical magnetic resonance tomography (MRI) and synthetic aperture radar (SAR) imaging. Both of these imaging modalities are based on the hologram idea in the radiofrequency and the microwave range, respectively ([24]). In contrast to the cardinal series filter implemented by the CD-DA, the construction of a filter bank performs the recovery of the image. The output of the A/D converter installed in the CD-DA player is parallel data in which all 16 bits of the data

word appear at once on 16 lines. Yet electronical storage devices permit storage only of serial string data in which the bits appear one after another. Data is therefore converted from parallel to serial format. The symplectic extension to optical holography, however, allows the parallel processing modes of the SAR and MRI modalities by optically implemented filter banks. The recovery of the image is then performed by a symplectic Fourier transform of the modulation transfer function ([20], [24]). The intrinsic symmetry of the symplectic Fourier transform allows to accelerate the image reconstruction without severe degradation of the picture.

Clinical MRI which is based on intrinsic differences between normal and abnormal tissues, provides a multitude of image contrasts. Due to this advantage of spin dynamics, MRI is the *non-invasive* imaging modality of choice in the majority of all cases of clinical diagnosis. Differences in longitudinal and transversal relaxation, spin density, macromolecular composition, diffusive motion, and bulk flow can be underscored by a variety of specifically designed pulse trains of suitable duration, orientation, and frequency. The subject of pulse train design is excitingly helpful and the clinical imaging results are cute.

8 Conclusion

Godfrey Harold Hardy (1877–1947), “the purest of the pure” mathematicians, thought that the existence of mathe-

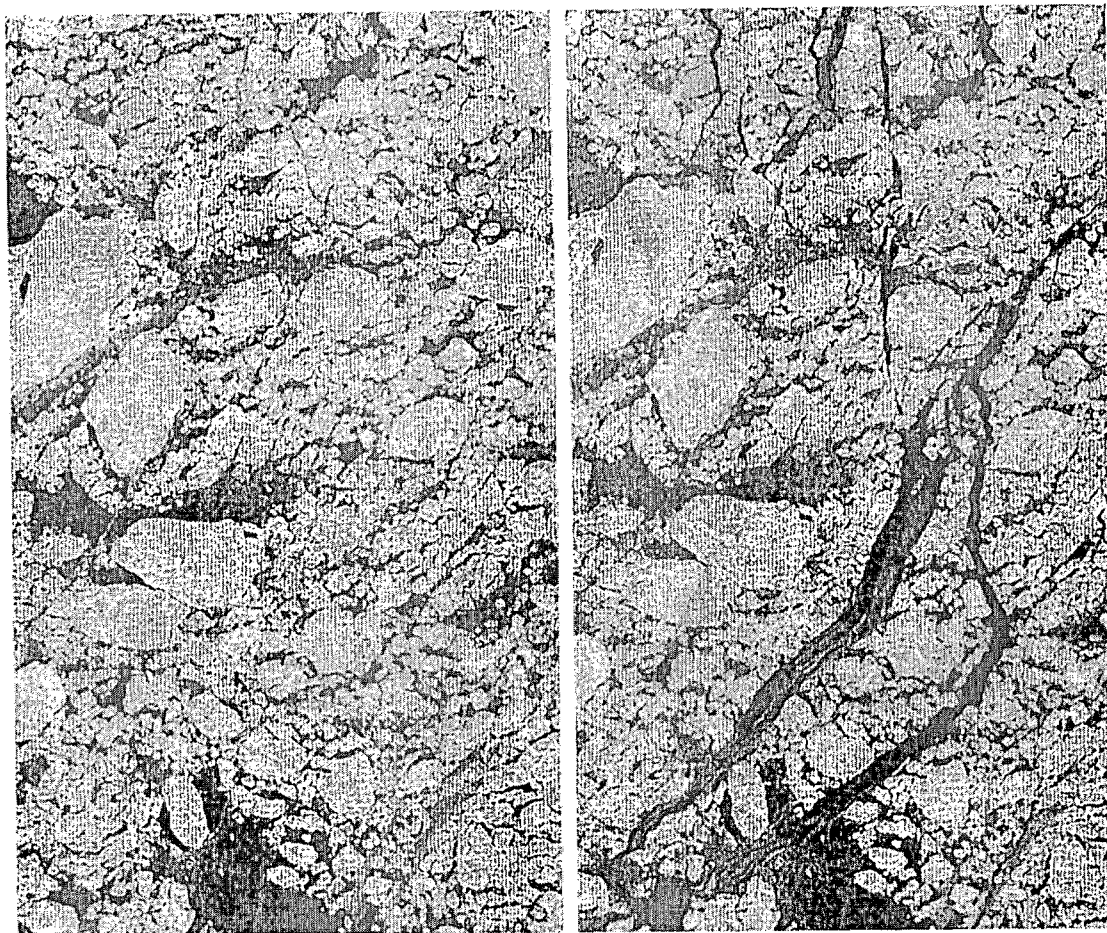


Figure 5: Radar imaging: The dynamics of pack ice in the Beaufort sea, located in the north of Canada and Alaska, has been observed and updated every three days by Radarsat. The channels, so called polynjas, have been created in the arctic ocean within nine days. Radarsat telemetry established that the lengths of some of the polynjas is up to 2000 kilometers.

matics could only be justified as *art* if it could be justified at all ([15], [12]). Although he did not specify his understanding of the metaphysics of art, he insisted on the fact that his own mathematical achievements have been based on pure thoughts and did not admit any real life application. However, the grand master's elegant symmetric version of the Poisson summation formula which has been put, independently, in its general group theoretical context by André Weil (1906–1998), projects from the two-dimensional digitization lattice $\mathbf{Z} \oplus \mathbf{Z}$ onto the time sampling theorem which, ironically, forms the base of the most successful consumer electronics products, the CD–DA, ever introduced. The CD–DA player forms the most sophisticated piece of audio electronics to reach the home. Because all CDs and players offer considerable advantage over other audio media, the CD has proved to be a technological wunderkind in the highly sophisticated and competitive field of music and data storage. Due to their versatility which has quickly become apparent to manufacturers and users alike, more than a billion CDs are sold every year. In the extremely storage-hungry market of IT, which forced the MD Atrac 3 format to increase the maximum

playing time of a conventional CD–DA from 74 minutes to 320 minutes, the annual worldwide demand for CD–DAs, CD–ROMs, DVIs, CD–Is, and DVD–ROMs is still rapidly climbing.

A central extension of the symplectic structure hidden by the *projected* summation formulae allows a powerful application to phase coherent summation imaging modalities of IT such as SAR and clinical MRI. Different from Hardy's and Weil's view of the Poisson summation formula, the *temporal* approach justifies the projection approach to the sampling processes and quantization modes of signal theory and opens a new perspective to the innovative field of IT. In this context, basic theta identities and the law of quadratic reciprocity appear as nice by-products of harmonic analysis on the Heisenberg nilpotent Lie group G which reveals itself as the *universal* mathematico-temporal structure of multirate signal analysis.

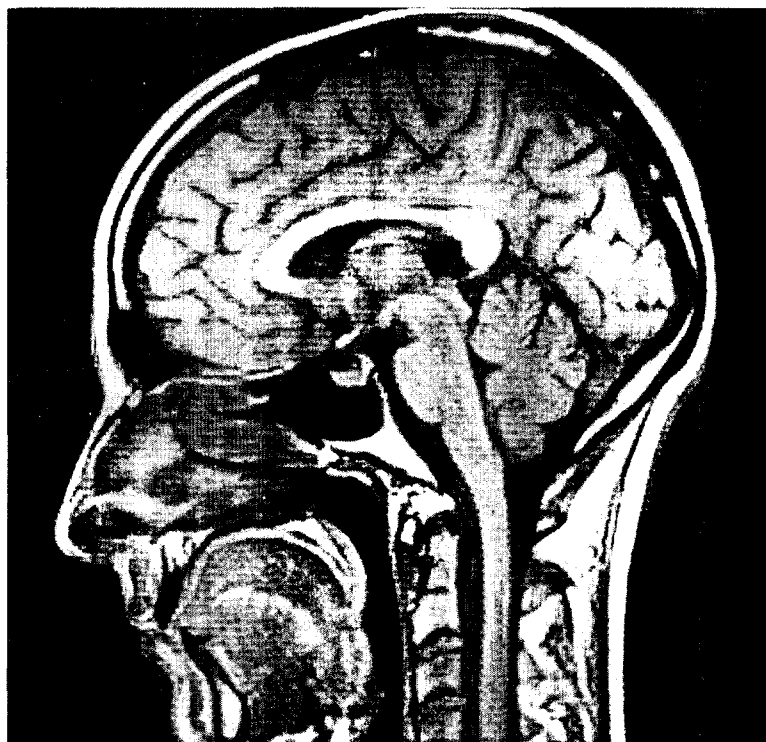


Figure 6: High resolution clinical magnetic resonance tomography: Sagittal cross-section of the neurocranium along the falx cerebri within the longitudinal interhemispheric fissure demonstrating midline sagittal neuroanatomy of the outwardly rounded gyri and inwardly invaginating fissures and sulci of the human brain. The various portions of the corpus callosum shown include the rostrum, genu, body and splenium, pineal gland, quadrigeminal plate, infundibulum, third and fourth ventricle, pituitary gland, cerebellar vermis, pons, aqueduct of Sylvius, prepontine space, and craniocervical junction. High resolution MRI scans approximate the same level of detail as cut specimens to non-invasively depict neuroanatomy even in the deepest recesses of the brain.

References

- [1] J.V. Armitage, A. Rogers, Gauss sums and quantum mechanics. *J. Phys. A: Math. Gen.* 33, 5993–602 (2000)
- [2] R. Bellman, *A Brief Introduction to Theta Functions*. Holt, Rinehart and Winston, New York 1961
- [3] J. Bertoin, *Lévy Processes*. Cambridge University Press, Cambridge, New York, Melbourne 1998
- [4] C. Biaesch-Wiebke, *CD-Player und R-DAT Recorder*. Second Edition, Vogel Buchverlag, Würzburg 1989
- [5] E. Binz, W. Schempp, Quantum teleportation and spin echo: A unitary symplectic spinor approach. In: *Aspects of Complex Analysis, Differential Geometry, Mathematical Physics and Applications*, S. Dimiev, K. Sekigawa, editors, pp. 314–365, World Scientific, Singapore, New Jersey, London 1999
- [6] E. Binz, W. Schempp, Vector fields in three-space, natural internal degrees of freedom, signal transmission and quantization. *Result. Math.* 37, 226–245 (2000)
- [7] E. Binz, W. Schempp, Quantum hologram and relativistic hodogram: Magnetic resonance tomography and gravitational wavelet detection. *Proc. CASYS 2000, Fourth International Conference on Compting Anticipatory Systems*, Liège, Belgium (in print)
- [8] E. Binz, W. Schempp, *Space-Time Geometry and Quantum Information: Transmission, Encoding and Detection*. Manuscript (to appear)
- [9] E. Binz, W. Schempp, The Landsberg-Schaar identity and the real Heisenberg nilpotent Lie group (to appear)
- [10] E. Binz, W. Schempp, The Lévy intensity measure and the Third Keplerian Law of planetary motion (to appear)
- [11] R.P. Boas, Jr., Summation formulas and band-limited signals. *Tôhoku Math. J.* 24, 121–125 (1972)
- [12] A. Borel, *Mathematik: Kunst und Wissenschaft. Themen-Reihe der Carl Friedrich von Siemens*

- Stiftung XXXIII, München 1982. In: *Collected Papers*, Vol. III, 685–701, Springer–Verlag, Berlin, Heidelberg, New York 1983
- [13] W.F. Donoghue, Jr., *Monotone Matrix Functions and Analytic Continuation*. Springer–Verlag, Berlin, Heidelberg, New York 1974
- [14] K.I. Gross, On the evolution of noncommutative harmonic analysis. *Amer. Math. Monthly* 85, 525–548 (1978)
- [15] G.H. Hardy, *A Mathematician's Apology*. Foreword by C.P. Snow, Cambridge University Press, Cambridge, New York, New Rochelle 1988
- [16] P. Horowitz, W. Hill, *The Art of Electronics*. Second edition, Cambridge University Press, Cambridge, New York, Melbourne 1995
- [17] I. Karatzas, S.E. Shreve, *Brownian Motion and Stochastic Calculus*. Second edition, Springer–Verlag, New York, Berlin, Heidelberg 1999
- [18] R.A. Kunze Positive definite operator–valued kernels and unitary representations. In: *Functional Analysis, Proc. of a Conference at the University of California, Irvine*, B.R. Gelbaum, editor, pp. 235–247, Thompson Book Company, Washington, D.C. 1967
- [19] R.A. Kunze, On the Frobenius reciprocity theorem for square–integrable representations. *Pacific J. Math.* 53, 465–471 (1974)
- [20] E.N. Leith, Synthetic aperture radar. In: *Optical Data Processing*, D. Casasent, editor, pp. 89–117, Springer–Verlag, Berlin, Heidelberg, New York 1978
- [21] K.C. Pohlmann, *The Compact Disc Handbook*. Second edition, Oxford University Press, Oxford, New York, Toronto 1992
- [22] H. Pollard, O. Shisha, Variations on the binomial series. *Amer. Math. Monthly* 79, 495–499 (1972)
- [23] W. Schempp, *Harmonic Analysis on the Heisenberg Nilpotent Lie Group, with Applications to Signal Theory*. Pitman Research Notes in Mathematics Series, Vol. 147, Longman Scientific and Technical, London 1986
- [24] W.J. Schempp, *Magnetic Resonance Imaging: Mathematical Foundations and Applications*. Wiley–Liss, New York, Chichester, Weinheim 1998
- [25] W. Schempp, Sub–Riemannian geometry and clinical magnetic resonance tomography. *Math. Meth. Appl. Sci.* 22, 867–922 (1999)
- [26] W. Schempp, Wavelet modelling of clinical magnetic resonance tomography: An ensemble quantum computing approach. In: *Inverse Problems, Tomography, and Image Processing*, A.G. Ramm, editor, pp. 129–176, Plenum Press, New York, London 1999
- [27] L. Schwartz, Sous–espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Analyse Math.* 13, 115–256 (1964)
- [28] G.E. Thomas, Future trends in optical recordings. *Philips Technical Review* 44, 51–57 (1988)
- [29] A. Weil, Sur certains groupes d'opérateurs unitaires. *Acta Math.* 111, 143–211 (1964); In: *Collected Papers*, Vol. III (1964–1978), pp. 1–69, Springer–Verlag, New York, Heidelberg, Berlin 1980
- [30] R.M. Young, *An Introduction to Nonharmonic Fourier Series*. Pure and Applied Mathematics Series, Academic Press, New York, London, Toronto 1980

INFORMATION SOCIETY 2001
INFOS, Cankarjev dom, Ljubljana, Slovenia
22.-26. October 2001

Members of programme committee

Cene Bavec, chair
 Tomaž Kalin, co-chair
 Jozsef Györkös, co-chair
 Marko Bohanec
 Jaroslav Berce
 Ivan Bratko
 Dušan Caf
 Saša Divjak
 Tomaž Erjavec
 Matjaž Gams
 Marko Grobelnik
 Nikola Guid
 Marjan Heričko
 Borka Jerman Blažič Džonova
 Gorazd Kandus
 Marjan Krisper
 Andrej Kuščer
 Jadran Lenarčič
 Dunja Mladenič
 Franc Novak
 Marjan Pivka
 Vladislav Rajkovič
 Ivan Rozman
 Niko Schlamberger
 Franc Solina
 Stanko Strmčnik
 Tomaž Šef
 Jurij Tasič
 Denis Trček
 Andrej Ule
 Tanja Urbančič
 David B. Vodušek
 Baldomir Zajc
 Blaž Zupan

Members of international programme committee

Vladimir Bajic
 Heiner Benking
 Se Woo Cheon
 Howie Firth
 Vladimir Fomichev
 Alfred Inselberg
 Jay Liebowitz
 Huan Liu
 Henz Martin
 Marcin Paprzycki
 Karl Pribram
 Claude Sammut
 Jiri Wiedermann
 Xindong Wu
 Yiming Ye
 Ning Zhong

Organizational committee

Matjaž Gams, chair
 Damjan Demšar
 Benjamin Jošar
 Aleksander Pivk
 Mili Remetic
 Maja Škrjanc

You are kindly invited to cooperate on multi-conference Information Society – IS 2001, which will be held under INFOS from 22nd to 26th of October 2001 in Cankarjev dom in Ljubljana. The multi-conference will include important achievements on the fields mentioned below. Emphasis will be given on the exchange of ideas and particular suggestions, which will be included in the final paper of individual conferences.

IS 2001 exists of nine carefully chosen conferences:

Collaboration and information society
 Data mining and warehouses
 Development and reengineering of information systems
 Education in information society
 Intelligent systems
 Management and information society
 Medical and cognitive science
 Speech technologies
 New information technologies in fine arts

Further information is available at <http://is.ijs.si/> or <http://ai.ijs.si/is/is2001/index01.html>.

Institutions, enterprises and donators are invited to present interesting new developments on their fields of work as 'normal' contributions. They can make a review of new developments and existing situation in their institutions and talk about problems of development in Slovenia, attitude of governmental institutions, and about the way Slovenia should be developing in the direction of information society. They can grant certain interesting activities, related to their work (please turn to the organizer, for example matjaz.gams@ijs.si).

The emphasis is on development, new ideas and trends in information society. If you have something interesting to tell or show to Slovenia, Information Society is the right place to be.

Invited are primarily all those, who have some knowledge about information society. Presentations of enterprises are welcome, especially from the functional point of view. To summarize, we will meet to tell what can we do in Slovenia, to exchange our experiences and to help Slovenia make a step forward in the direction of information society.

You are kindly invited to make a presentation and actively take part in the open exchange of ideas with your knowledge and achievements. The submission deadline is fall 2001.

Pictures from the IS 2000 conference can be found at <http://ai.ijs.si/IS/is2000/index00.html>.

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan-Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 700 staff, has 500 researchers, about 250 of whom are postgraduates, over 200 of whom have doctorates (Ph.D.), and around 150 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of Slovenia (or S^onia). The capital today is considered a crossroad between East, West and Mediter-

ranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

In the last year on the site of the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

At the present time, part of the Institute is being reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project is being developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park will take the form of a shareholding company and will host an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of Economic Relations and Development, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 219 385
Tlx.:31 296 JOSTIN SI
WWW: <http://www.ijs.si>
E-mail: matjaz.gams@ijs.si
Contact person for the Park: Iztok Lesjak, M.Sc.
Public relations: Natalija Polenec

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please submit three copies of the manuscript with good copies of the figures and photographs to one of the editors from the Editorial Board or to the Contact Person. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible directly on the manuscript, from typing errors to global philosophical disagreements. The chosen editor will send the author copies with remarks. If the paper is accepted, the editor will also send copies to the Contact Person. The Executive Board will inform the author that the paper has been accepted, in which case it will be published within one year of receipt of e-mails with the text in Informatica L^AT_EX format and figures in .eps format. The original figures can also be sent on separate sheets. Style and examples of papers can be obtained by e-mail from the Contact Person or from FTP or WWW (see the last page of Informatica).

Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the Contact Person.

QUESTIONNAIRE

Send Informatica free of charge

Yes, we subscribe

Please, complete the order form and send it to Dr. Rudi Murn, Informatica, Institut Jožef Stefan, Jamova 39, 1111 Ljubljana, Slovenia.

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than five years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

ORDER FORM – INFORMATICA

Name:

Title and Profession (optional):

Home Address and Telephone (optional):

Office Address and Telephone (optional):

E-mail Address (optional):

Signature and Date:

Informatica WWW:

<http://ai.ijs.si/informatica/>
<http://orca.st.usm.edu/informatica/>

Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Mohamad Alam, Dia Ali, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Vladimir Bajič, Grzegorz Bartoszewicz, Catriel Beeri, Daniel Beech, Fevzi Belli, Francesco Bergadano, Istvan Berkeley, Azer Bestavros, Andraž Bežek, Balaji Bharadwaj, Ralph Bisland, Jacek Blazewicz, Laszlo Boeszoermyeni, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Jerzy Brzezinski, Marian Bubak, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Netiva Caftori, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepielewski, Vic Ciesielski, David Cliff, Maria Cobb, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Deborah Dent, Andrej Dobnikar, Sait Dogru, Georg Dorfner, Ludoslaw Drelichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Jozo Dujmović, Pavol Ďuriš, Johann Eder, Hesham El-Rewini, Warren Fergusson, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Hugo de Garis, Eugeniusz Gatnar, James Geller, Michael Georgiopolus, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Inman Harvey, Elke Hochmueller, Jack Hodges, Rod Howell, Tomáš Hruška, Don Huch, Alexey Ippa, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Sabhash Kak, Li-Shan Kang, Orlando Karam, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Kevin Korb, Gilad Koren, Henryk Krawczyk, Ben Kroese, Zbyszko Krolkowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Phil Laplante, Bud Lawson, Ulrike Leopold-Wildburger, Joseph Y-T. Leung, Barry Levine, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Matija Lokar, Jason Lowder, Kim Teng Lua, Andrzej Małachowski, Bernardo Magnini, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Daniel Memmi, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Gautam Mitra, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Hari Narayanan, Rance Necaie, Elzbieta Niedzielska, Marian Niedq'zwiadziński, Jaroslav Nieplocha, Jerzy Nogieć, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, William C. Perkins, Warren Persons, Mitja Peruš, Stephen Pike, Niki Pissinou, Aleksander Pivk, Ullin Place, Gustav Pomberger, James Pomykalski, Dimithu Prasanna, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Luc de Raedt, Ewaryst Rafajłowicz, Sita Ramakrishnan, Wolf Rauch, Peter Rechenberg, Felix Redmill, David Robertson, Marko Robnik, Ingrid Russel, A.S.M. Sajeev, Bo Sanden, Vivek Sarin, Iztok Savnik, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer, Zhongzhi Shi, William Spears, Hartmut Stadler, Olivero Stock, Janusz Stokłosa, Przemysław Stpicyński, Andrej Stritar, Maciej Stroinski, Tomasz Szmuc, Zdzisław Szyjewski, Jure Šilc, Metod Škarja, Jiří Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Wiesław Traczyk, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Zygmunt Vetulani, Olivier de Vel, John Weckert, Gerhard Widmer, Stefan Wrobel, Stanisław Wrycza, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Zonling Zhou, Robert Zorc, Anton P. Żeleznikar

EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

Executive Editor – Editor in Chief

Anton P. Železnikar
Volaričeva 8, Ljubljana, Slovenia
s51em@lea.hamradio.si
<http://lea.hamradio.si/~s51em/>

Executive Associate Editor (Contact Person)

Matjaž Gams, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Phone: +386 1 4773 900, Fax: +386 1 219 385
matjaz.gams@ijs.si
<http://www2.ijs.si/~mezi/matjaz.html>

Executive Associate Editor (Technical Editor)

Rudi Murn, Jožef Stefan Institute

Publishing Council:

Tomaž Banovec, Ciril Baškovič,
Andrej Jerman-Blažič, Jožko Čuk,
Vladislav Rajkovič

Board of Advisors:

Ivan Bratko, Marko Jagodič,
Tomaž Pisanski, Stanko Strmčnik

Editorial Board

Suad Alagić (Bosnia and Herzegovina)
Vladimir Bajić (Republic of South Africa)
Vladimir Batagelj (Slovenia)
Francesco Bergadano (Italy)
Leon Birnbaum (Romania)
Marco Botta (Italy)
Pavel Brazdil (Portugal)
Andrej Brodnik (Slovenia)
Ivan Bruha (Canada)
Se Woo Cheon (Korea)
Hubert L. Dreyfus (USA)
Jozo Dujmović (USA)
Johann Eder (Austria)
Vladimir Fomichov (Russia)
Georg Gottlob (Austria)
Janez Grad (Slovenia)
Francis Heylighen (Belgium)
Hiroaki Kitano (Japan)
Igor Kononenko (Slovenia)
Miroslav Kubat (USA)
Ante Lauc (Croatia)
Jadran Lenarčič (Slovenia)
Huan Liu (Singapore)
Ramon L. de Mantaras (Spain)
Magoroh Maruyama (Japan)
Nikos Mastorakis (Greece)
Angelo Montanari (Italy)
Igor Mozetič (Austria)
Stephen Muggleton (UK)
Pavol Návrat (Slovakia)
Jerzy R. Nawrocki (Poland)
Roumen Nikolov (Bulgaria)
Marcin Paprzycki (USA)
Oliver Popov (Macedonia)
Karl H. Pribram (USA)
Luc De Raedt (Belgium)
Dejan Raković (Yugoslavia)
Jean Ramaekers (Belgium)
Wilhelm Rossak (USA)
Ivan Rozman (Slovenia)
Claude Sammut (Australia)
Sugata Sanyal (India)
Walter Schempp (Germany)
Johannes Schwinn (Germany)
Zhongzhi Shi (China)
Branko Souček (Italy)
Oliviero Stock (Italy)
Petra Stoerig (Germany)
Jiří Šlechta (UK)
Gheorghe Tecuci (USA)
Robert Trappl (Austria)
Terry Winograd (USA)
Stefan Wrobel (Germany)
Xindong Wu (Australia)

Informatica

An International Journal of Computing and Informatics

Introduction		147
Wizards of OZ — change in learning and teaching	W. Coy, U. Pirr	149
A hybrid system for delivering web based distance learning and teaching material	J. Greenberg	155
Learning by experience: networks in learning organizations	M. Kuittinen, E. Sutinen et al.	159
How to learn introductory programming over the Web?	H. Arto, S. Jarkko, S. Erkki	165
Information system supporting CATS	Z. Ryjáček et al.	173
A data warehouse for French universities	J.-F. Desnos	177
Data quality: a prerequisite for successful data warehouse implementation	V. Mahnič, I. Rožanc	183
'Beowulf cluster' for high-performance computing tasks at the university: a very profitable investment	M.J. Galán, F. García et al.	189
Evaluation of codec behavior in IP and ATM networks	S. Naegelé-Jackson, U. Hilgers et al.	195
An environment for processing compound media streams	B. Feustel, A. Kárpáti, T. Rack, T.C. Schmidt	201
FEIDHE - integrating PKI in Finnish higher education	M. Linden, J. Kanner, M. Kivilompolo	211
Information systems delivery in a tiered security environment	A. Strachan, T. Shaw, D. Adams	217
<hr/>		
Register allocation: A program-algebraic approach	R.D. Resler et al.	223
Neural fields: An approach to infinite-dimensional systems for information processing	A.D. Linkevich	235
An overview of mobile agents in distributed applications: Possibilities for future enterprise systems	S.W. Loke	247
A pattern recognition approach to the prediction of price increases in the New York Stock Exchange Composite Index	W. Leigh, M. Paz, N. Paz, R. Purvis	261
A digital multi-layer-perceptron hardware architecture based on three dimensional massively parallel optoelectronic circuits	K.D. Maier, C. Beckstein, R. Blickhan et al.	271
Information technology: The lie groups defining the filter banks of the compact disc	E. Binz	279
Reports and Announcements	W. Schempp	293