Final Report for SNF Project no. 9502830

Networks and Paradigms for the Next Generation of Distributed Systems

Eric Jul[‡] Povl Koch[‡] Michael Svendsen[‡] Robin Sharp[†] Arne Skou[¶]
Niels Christian Juul[§]
Birger Andersen[§ £]

[‡] Department of Computer Science, University of Copenhagen, DIKU
[†] Department of Information Technology, Technical University of Denmark, DTU
[¶] Basic Research in Computer Science, Univerity of Aalborg, AAU
[§] Department of Informatics, Copenhagen Business School, CBS
[£] Earlier affiliated with DIKU and AAU

October 16, 1998

Overall Project Summary

The SNF funded framework grant "Networks and Paradigms for the Next Generation of Distributed Systems" has been a joint project (so-called *rammebevilling*) between Computer Science departments at the University of Copenhagen, the Technical University of Denmark, Aalborg University, and Copenhagen Business School. The grant has provided support for a number of activities concerning the iterative development of new network paradigms for optimal utilization of modern network hardware. The major financial support was for a Ph.D. student, some programmer support, and money for arranging workshops and travel to conferences.

The grant period started January 1st 1996 and lasts until December 31st 1998. During this period two SNF workshops and three workshops at international conferences have been organized. Participation and presentations at many international conferences have also been funded by the grant.

A major part of the grant money was support for a Ph.D. student. Recruiting such a student was not easy but one was found by November 1st 1996 when Michael Svendsen began his three year Ph.D. study.

A previous SNF equipment grant provided ATM networking hardware at AAU, DTU, and at DIKU. This equipment has been used extensively as an experimental testbed for various distributed networking projects including wide area networking in that during the project period, it has been possible to get access to an ATM connection between AAU and DTU via a high-speed link (34 Mbit/s). Much of the research has focused on lower level protocols such as network striping (using multiple ATM connections as if they were one), distributed shared memory implemented using ATM, efficient support for memory-mapped networks, distributed parallel programming using the Linda paradigm, using ATM for video communication, and developing reliable protocols for ATM.

Part of the ATM research at DIKU has been performed in cooperation with Olicom A/S and CIT.

In the following, the activities at each of the four departments are described separately. The activities described are those that have been initiated since the start of the grant and up to now (September 1998). Hardware that has been used in the research under this grant came partly from the earlier SNF grant no. 9400353 and partly from other grants.

1 Common Activities

We have arranged two SNF workshops in the autumns of 1996 and 1997. Both workshops were attended by researchers from the four departments under this grant and by invited people from the Danish industry. At the workshops both groups presented current work used as a basis for further discussions. The workshops were a success where both researchers and the representatives of the industry gained knowledge of current work and trends.

DIKU and CBS have cooperated in organizing workshops on Mobility and Replication at the European Conference on Object-Oriented Programming (ECOOP) [2, 5] and at Object-Oriented Programming Systems, Languages and Applications (OOPSLA) [4] in the years 1996-98. The workshops at ECOOP were developed in cooperation with researchers in Portugal.

Parts of the project supported by the SNF grant was research of WAN (Wide Area Network) technology. DTU and AAU have been connected via a high-speed (34 Mbit/s) ATM connection. The connection has been used to experiments with distributed systems and video conferencing. The link has been provided by the Danish research net on an experimental basis. Research and measurements of the WAN connection between DTU and AAU are described later in this report.

2 Activities at DIKU

The research at DIKU has focussed on the following areas:

- High-speed networking over ATM with measurements of the characteristics for the employed hardware.
- Network striping over ATM which has been a joint project CIT.
- TCP over ATM including the design and implementation of a TCP protocol directly on top of ATM.
- TCP connection handoff giving better support for cluster servers and allowing an aggregate bandwidth otherwise not obtainable.
- OS support for memory mapped network based on the new SCI network technology.
- Wireless networks with focus on disconnected operation and changes in the quality of service (QoS) provide by the network subsystem (in cooperation with CBS).

In the grant period DIKU have had several visitors: **Willy Zwaenepoel** from Rice University as opponent on Povl Koch's dissertation [15], **Norm Hutchinson** from University of British Columbia as opponent on two Masters thesises [25, 14], **Christian Ionitoiu** a Ph.D. student from the Technical University of Timisoara in Romania wrote a paper with Birger Andersen [1] published in the Workshop Reader of ECOOP'96 [3], **Sylvain Loubtin** from University of Dublin and **Phillippe Bonnét** from INRIA in Grenoble both as potential candidates for a postdoc position at DIKU.

DIKU has been represented at many international conferences. **Eric Jul** has been on the ECOOP PC for the 1997, '98, and '99 ECOOP conferences and was PC Chair in 1998. He has also been on the OOPSLA '98 PC and on the TOOLS '98 and '99 PC. Under this grant he has also continued to develop a highly successful series of tutorials on distributed computing using objects [?] presented at ECOOP '96 and at OOPSLA '96. At OOPSLA '97, and '98 it was (will be) given twice due to its popularity. Eric Jul is a member of the SOSP'99 organizing committee. Eric attended the (EUROPEAN SIGOPS WORKSHOP) [?]. **Jørgen Sværke Hansen** presented a paper [7] at the USENIX Windows NT Workshop in 1997 and attended the SIGCOMM conference in 1996 with **Povl Koch**. Povl presented a paper [17] at EUROPAR'98. **Michael Svendsen** attended the USENIX Annual Technical Conference in 1998 and will attend ASPLOS VIII in October 1998 with an accepted paper [19]. The research group and three Masters students (7 persons in total) attended the SOSP conference in 1997.

2.1 High-speed Networking over ATM

Two students, Jacob J. Schrøder and Thomas Stormark, performed bandwidth and latency meassurements [24] on the acquired ATM network. Their tests showed that full theoretical bandwidth was obtained for both TCP and UDP using half-duplex, but problems occurred for both protocols using full-duplex lines (communication goes both ways). The problems came because of CPU saturation caused by especially handing interrupts from incoming packets and copying buffers. Measurements of latency in the same project showed that ATM, in spite of the higher bandwidth, is not faster than the Ethernet technology.

Michael Svendsen and Peter S. Knudsen utilized the higher bandwidth of the ATM network to experiment with redesigning an existing DSM system. The basic assumption in modern DSM systems (distributed shared memory systems) is that write-invalidate protocols are generally better performing since network bandwidth is a limited resource. However, with ATM this is no longer an obvious assumption. In their Master thesises [25, 14] they document the design, implementation and tests with DSM system using a hybrid protocol that is a mix between write-update and write-invalidate. The results showed that performance could be improved in cases where latency caused by invalidations is a problem and where network bandwidth is no longer a bottleneck in the system.

2.2 Network Striping over ATM

The Software-based Parallel ATM Networks (SPAN) project deals with the issues of using multiple network adapters in parallel (also called network striping) to provide (1) higher aggregate network bandwidths, (2) lower message latencies, and (3) more reliable communication than a single network adapter in each node.

We have successfully implemented a striping prototype [22, 23] in Windows NT, which shows that it is possible to almost double the network bandwidth at the driver level using two ATM adapters. Due to the performance of the Windows NT TCP/IP protocol stack this bandwidth increase cannot be fully utilized by user programs. Our implementation of ATM striping introduces two new concepts; the **Multiple Virtual Connection** (MVC) and the **Trigger Round Robin** (TRR) striping algorithm [7].

[1] Birger Andersen was assistant professor at DIKU until August 1996.

The MVC is a new software abstraction which provides a uniform interface to striping. The abstraction can be considered an extension of the PVC/SVC to accommodate striping while upper software layers retain the simplicity of a single connection abstraction. Striping is *not* directly specified for MVCs, rather a particular Quality of Service is specified. It is then the responsibility of the MVC software to provide this level of service, e.g., by using multiple, striped connections, if necessary. TRR is designed to efficiently fulfil the service guarantees of ATM. It is based on an existing striping algorithm, but has been improved in two ways: (1) we have added extra mechanisms to ensure in-order delivery of packets, and (2) the algorithm now adapts to dynamically varying bandwidth of different stripes.

2.2.1 Multi-Path Routing

Michael Svendsen, the Ph.D. student funded by the grant, is working on multipath routing. Multipath routing is basically the same as striping, but where the SPAN project worked with the problem of dedicated lines between two hosts the multipath project worked with the problem of striping over non-dedicated lines with highly dynamic and very different properties. Multipath routing is already possible in the existing Internet protocols but is close to not used for bandwidth aggregation. His Ph.D. work includes investigating the problems arising when striping over lines with changing bandwidth and latency.

The original idea was to use the ATM WAN connections between DIKU, DTU and AAU. However, ATM has never been established internally within UNIVERSITY OF COPENHAGEN. The work has therefore been redesigned to make the necessary tests locally.

Michael will continue this work after returning from a 7 month visit to Rice University[2] Previously Michael has had a one half year paternity leave; thus his stipend will continue through April 2000.

2.3 Efficient Operating System Support for Memory-Mapped Networks

Traditional network interfaces for clusters of workstations are designed for generality, not performance. This seriously limits the performance of many distributed applications that have increasing communication demands. The next generation network interfaces, e.g., IEEE's Scalable Coherent Interface (SCI), are based on memory-mapped interfaces and promise much higher bandwidths and very low latencies. But the good performance of the new networks can only be efficiently exploited with appropriate simple, low-overhead programming interfaces.

We have constructed a prototype, called SciOS [16, 17] that provides global memory management on an SCI cluster. The main techniques are based on the hardware's possibility of directly accessing remotely shared memory through normal load/store instructions, but also protocols to dynamically migrate and replicate shared memory pages are used to enhance locality of memory references.

The research activity was initially performed at INRIA Rhone-Alpes (France) by Povl Koch — supported in part by the EU-TMR program and the Danish SNF. During the project, close contact has been maintained with DIKU's distributed systems group. The project will continue at DIKU which has established a small SCI cluster testbed for future research in the area.

2.4 TCP over ATM

ATM is a connection oriented network technology. In many research papers it has been suggested

[2]The Danish Ph.D. program requires Ph.D. candidates to visit to a different research environment for one half year.

that TCP (that is a connection-oriented transport protocol) could be implemented directly on top of ATM. This would, according to the suggestions, make the TCP implementation faster and simpler due to less load of the host CPU and utilization of the ATM properties. However, no implementations of TCP over ATM exists. DIKU has two Masters Students who are implementing TCP over ATM.

2.5 TCP Connection Handoff

While visiting Rice University (during the Spring of 1998) Michael Svendsen participated in the ScalaServer project developing a distributed Internet cluster server. The goal of the project was to build a distributed server from commodity hardware. Since most Internet traffic is asymmetric the resources used for requesting data is much less than the resources used replying to those requests. The project developed a model where a single machine accepts all incoming requests and based on the request forward the connection to one of several servers as depicted in Figure 1 in Appendix A. The project has several interesting components in the fields of distributed computing: (1) Load distribution and balancing, (2) cache and bandwidth aggregation, (3) fault tolerance, and (4) TCP optimization and updates of functionality.

In the working implementation of the cluster server as it is presented in [19] an update of the TCP protocol was necessary for the system to run transparent to the client. Work with this modification is continuing. The change to the protocol has enabled support for TCP connection distribution of the load from the Internet. The main results from this work are the successful load distribution and the aggregation of file caches, outbound network bandwidth and secondary storage bandwidth. It is possible to get a higher outbound network bandwidth from the server than can be provided by any single network interface/machine.

We have found that TCP is not prepared for the extreme loads that is seen at several Internet servers. We changed TCP in two ways: (1) we enabled distributed handling of connections as if handled by only a single server and (2) rewrote several algorithms used in TCP to enable the server to handle $> 100,000$ connections simultaneously and up to 5000 connections/second at a sustained rate.

Our analysis of persistent HTTP connections showed that the load of the server drops because fewer connection establishments are needed. Based on traces from different websites we found that correlation between files in a persistent connection is less than expected. This resulted in an overall performance degradation since caching in the cluster server is then harder to control.

The results of this work will be presented in Michael's Ph.D. dissertation expected in April 2000.

2.6 Wireless Networks

The TrAnsparent COmmunication (TACO) [10, 8] layer is a system for mobile computing, that hides the complexity of using and switching between different means of communication, supports semi-connected operation, and makes scheduling of data possible on low bandwidth communication channels. In our system we consider mobile hosts, that can either be connected directly to a LAN, or be connected by a dial-up serial connection to a Mobility Support Gateway (MSG) on the LAN. When the mobile host is connected directly to the LAN, regular TCP/IP communication is used. When the mobile host disconnects from the LAN, a TCP connection is split into two separate connections — one on the serial link from the mobile host to the MSG and one on the wired network from the MSG to the peer host. In this case both the MSG and the mobile host can schedule the data to utilize the possible low capacity of the serial link most efficiently. This work was performed in cooperation with Birger Andersen who was originally at DIKU and is now at CBS.

2.7 EU Grant Proposals

We have also put considerable effort into trying to improve our international cooperation by applying for EU funding.

2.7.1 COPERNICUS Application

In cooperation with universities in Poland, Rumania, Belgium, and U.K., we applied for funding from EU's COPERNICUS program for work on distributed object systems. Our application received favorable reviews and was scheduled to receive funds, but the EU commission had insufficient funds at the end.

2.7.2 ESPRIT Application

In cooperation with companies in Germany, Norway, and France and the University of Lund, DIKU participated in a two phase ESPRIT application concerning distributed shared memory using high speed network interfaces. The application received favorable reviews and was admitted to a second application round where it unfortunately did not receive funding.

3 Activities at DTU

At the Department of Information Technology at DTU, work has concentrated on four topics:

1. Measurements of performance in a local area ATM network and over a wide area ATM network.
2. Distributed implementation of the parallel programming paradigm Linda in a local area ATM network.
3. Implementation of a Work Pool system based on a local area ATM network.
4. Investigation of the parallel programming system Calypso in a local area ATM network.

3.1 Performance measurements

The ATM network installed at the Department of Information Technology consists of a cluster of 4 Digital Alpha workstations, running the Digital Unix operating system, attached to a Digital Gigaswitch ATM switch. Due to a series of delays in the delivery of the equipment from the supplier, the network was in fact not finally installed until November 1995. This delayed our programme of measurements, which finally started in Spring 1996.

3.1.1 Measurements on an Local Area ATM Network

Measurements of throughput and latency for a variety of traffic patterns were carried out in the cluster of workstations at DTU. This work was performed partly in collaboration with IRESTE, Nantes, from where we had a visiting Master's thesis student.

The results, mostly presented in [21], can be summarised as follows: The TCP/IP overhead is considerable, both due to the large header sizes and the stream-oriented protocol, that the buffer copying mechanism causes a large overhead which is especially marked at small window sizes, and that the use of Nagle's algorithm has a deleterious effect on the throughput in several ranges of window size. These results are in general accordance with other investigations into the behaviour of TCP/IP over ATM, such as [18]. We also noted that the implementation in Digital Unix (version 3.2C) also seems to be unable to use the full bandwidth of the ATM network in full-duplex operation.

3.1.2 Measurements on a Wide Area ATM Network

In the Spring of 1997, after extensive negotiations with UNI-C, who had set up an ATM-based high-performance network as a prototype for a Danish Research Network, an experimental direct ATM connection from DTU to AAU was established. The available bandwidth for this connection was by agreement limited to 1 Mbit/s in the first instance. Measurements of latency and jitter were made for a loopback connection starting from a workstation at DTU, passing through the ATM network to the switch at Aalborg and back again to DTU, and were compared with similar measurements over the ATM LAN at DTU with loopback in the switch at DTU. Sets of measurements were made:

- Using the AAL5 protocol alone, and using IP over AAL5.
- For workstations under "normal load" (running a full complement of daemons), and for workstations with essentially all daemons shut down.

An example is shown in Figure 2 in Appendix A. The general conclusions of these measurements are:

- The AAL5 protocol (and also IP) gives only a small contribution to jitter.
- The WAN network contributes to latency to an extent which corresponds more or less directly to the signal propagation over the distance involved, but it contributes very little to jitter.
- The daemons of the OS — especially the periodic ones — give a marked contribution to jitter.
- OS scheduling needs to be genuine Real-Time scheduling to achieve accurate synchronisation.

3.2 Linda

A distributed implementation of the parallel programming paradigm Linda has been developed, initially during a Master's thesis project [20]. In the Linda paradigm, parallel processes have access to a large, conceptual common store, known as *tuple space*. This basically offers a non-blocking write operation, $\mathbf{out}(t)$, which stores a named tuple of values, t, and a blocking read-and-remove operation, $\mathbf{in}(s)$, which retrieves a value for a tuple matching the template s.

This paradigm can be implemented in a large number of ways in a distributed system. The implementation at DTU uses partial distribution of tuples, so that the tuple space is partitioned among the available systems. The operation **out** stores tuples locally. An **in** operation searches initially in the local part of the tuple space; if this fails, the template s (known as an *anti-tuple*) is broadcast to the other systems via the network. If a match is found, the tuple which matches s is transferred to the originating system. If no match can be found, the anti-tuple waits on the remote systems for a time, say t, for a tuple to be inserted there; if this does not occur, the anti-tuple is discarded and the broadcast procedure is repeated after a suitable delay, say Δ.

This method has the disadvantages that it needs broadcasting (which may be expensive without special hardware support) and it uses a relatively complicated **in** operation, compared to, say a fully replicated distributed implementation as used by Ahuja, Carriero and Gelernter [1]. However, it has the advantages of offering a high degree of distribution, and of giving fast **out** operations (and in fact also fast **in/read** operations when the tuple is found locally). Unlike full replication it does not require a multi-system delete operation to remove all the copies of a retrieved tuple, and the **in** operation, although expensive in number of messages, only broadcasts relatively small anti-tuples rather than the much larger tuples.

To support this implementation, it was considered necessary, in the light of the measurements discussed above, to develop a special light-weight protocol, designed specifically to exhibit low latency for small message sizes, in order to avoid using TCP/IP over ATM. The protocol, known as XSTP, is a simple stop-and-wait protocol implemented over the AAL5 adaptation layer to ATM, offering resilience to PDU loss and corruption. For technical reasons, XSTP is a kernel subsystem within Digital Unix. It has a latency for the exchange of small messages (length about 16 bytes) of about $155\mu s$, markedly lower than the typical TCP/IP latencies observed in the system (10–$50ms$).

Basing the Linda implementation on this protocol, a series of measurements of the efficiency of the Linda system have been made. A simple "Linda latency" test, where a client on one system issues a small tuple which is read by a server on another system, after which the procedure is reversed, gives a time of $1420\mu s$ for the round trip. This can be improved by further optimisations in the implementation.

3.3 Work Pools and Calypso

In addition to the use of Linda, a series of more complex experiments, which use *agenda parallelism*, have been carried out. These have been based on two techniques: The Work Pool paradigm, and the use of the system Calypso. Systems for both paradigms were implemented on the Alpha workstations under Version 4.0 of the Digital Unix operating system.

The Work Pool system was implemented as part of an M.Sc. thesis project. The basic idea was to implement a general API which could be used to run a variety of different algorithms in a distributed Work Pool environment. An schematic view of the system design is shown in Figure 3 in Appendix A. The implementation uses a multi-threaded design following the Posix Pthreads standard and a TCP Sockets API to the communication system, in order to achieve as portable a system as possible. It is the intention in due course to port the system to the cluster at AAU, so that experiments can be carried out on both the AAU and DTU clusters, connected by an ATM WAN. In the initial implementation, a simple balancing algorithm, which does not take the current level of activity in the individual workers into account, is used to distribute jobs between the workers in the pool.

A series of experiments was performed in the DTU cluster to investigate the performance of algorithms in the Work Pool. The Travelling Salesman problem was chosen as a typical asynchronous parallel algorithm, and Matrix Multiplication as a typical synchronous algorithm. Measurements were made using first the ATM network and then a 10 Mbit/s Ethernet as the basic communication system in the cluster.

These measurements demonstrate that an ATM network gives a marked reduction in the total execution time for the algorithm in relation to Ethernet, thus indicating that communication is an important factor. However, even with an ATM network, the communication overhead is considerable, and it is necessary to use fairly coarse-grained parallelism to achieve the optimum results — each subsidiary task should involve at least $175ms$ of computation in order to show a marked speedup relative to the centralised case. This corresponds in general terms to the experiences accumulated in the performance measurements. However, further work is needed in order to investigate the exact conditions under which it is possible to reduce the granularity as far as possible.

3.4 Visits etc.

During January 1998, Robin Sharp took part in the workshop organised by the Danish Ministry of Research to present ideas and discuss plans for the setting up of the Danish Research Network.

¿From February 1998 to July 1998, Robin Sharp was an academic visitor at the Computing Laboratory, Oxford University, UK. During this period, he also visited Cambridge University Computer Laboratory (Dr. Ian Leslie) and Engineering Laboratory (Prof. Andy Hopper), and the University of Kent at Canterbury (Prof. Peter Linington) to discuss projects related to networks and distributed system.

3.5 Ongoing Activities

In the Autumn of 1998, a series of activities is taking place whose aim is to investigate some aspects of the transmission of multi-modal interactive traffic via ATM networks. These activities are related to work which is going on at DTU in the area of interactive distributed multimedia. A collaboration with Alcatel Danmark has resulted in a donation of equipment based on SDH technology, which will be used in conjunction with the existing ATM switch, funded by SNF, at IT-DTU in order to investigate the behaviour of composite high-performance local area networks in this context.

Of particular interest is the development of multi-cast protocols suitable for distribution of compressed audio and video among multiple participants in an interactive multimedia session. To maintain an appropriate ordering of the messages in a multi-party exchange, causal protocols based on raw ATM, ATM with AAL5 and IP.V6 are to be investigated.

4 Activities at Aalborg University (AAU)

The activities at AAU have been concentrated on design and experiments with high-speed protocol techniques and techniques for video transfer.

4.1 Performance analysis of Video Communication on ATM-networks

In [11] a performance analysis is presented for protocols and compression techniques when applying ATM-networks for video communication. The main conclusions are that

- ATM networks provide a substantial improvement with respect to throughput and jitter compared to traditional local networks like Ethernet. However, there is still a substantial delay caused by standard operating systems software.

- MPEG-2 provides better compression than MPEG-1. However, for live-video sequences, the CPU demand for encoding is still too high for general purpose CPU's - as opposed to the CPU demand for decoding.

4.2 Layered Coding/Decoding of Video Transmission

In [13, 12], a new technique is presented for encoding of video traffic — based on the idea of splitting the video representation into layers of varying importance. The technique has been implemented by taking advantage of the dedicated Visual Instruction Set of SUN's UltraSparc platform. This gives a performance of a factor 3-4 compared to optimized C-code.

4.3 Reliable Protocols for ATM Networks

When applying high speed networks for video communication, it is normal to accept a certain (minimal) error rate. On the other hand, when transferring high-volume data files, the transmission must be completely reliable. Current techniques for reliable transfer (e.g. sliding window) are not well suited for high-speed networks like ATM. In [6], we confirm this thesis through experiments with an existing technique (TCP), and by developing a new reliable protocol. It turns out that the new protocol outperforms TCP — due to the use of a modern retransmission technique and to the reduction of buffer copying.

5 Activities at CBS

The distributed computing group at CBS consists of Associate Professor Lars Frank, Associate Professor Niels Christian Juul, and visiting Associate Professor Birger Andersen[3]. During the project the local affiliation has changed from the Department of Informatics and Management Accounting to the recently established Department of Informatics. The group is now part of the IT systems development group.

Beside participation in the over all project, the groups activities falls mainly with the areas of:

- AMIGOS - Advanced Mobile Integration in General Operating Systems
- ACDC - Availability versus Consistency in Distributed Computing

5.1 Advanced Mobil Integration in General Operating Systems (AMIGOS)

One main activity of the AMIGOS project has been knowledge exchange between researchers working in the field of mobility and replication. This was realized by taking part in organizing three workshops: *III Workshop on Mobility and Replication* [5], *Workshop on Object Replication and Mobile Computing* [4], and *II Workshop on Mobility and Replication* [2]. Hereby we got a lot of inspiration and met researchers working on issues related to our issues. Furthermore, the AMIGOS project became more widely known to the community. In 1998 the workshop was organized with two Portuguese colleagues, and in 1996 the two workshops were organized with Portuguese colleagues and Eric Jul at DIKU.

Another activity has been on the design and implementation of an Internet protocol extension for automatic adaption in mobile environments [9, 10]. This protocol extension makes it possible to physically disconnect from Internet still keeping TCP/IP connections active (but with communication suspended). Later on, a mobile computer can physically connect to Internet using a modem or mobile phone and continue using the kept TCP/IP connections. This design serves for making it more easy using a mobile computer running applications not developed for the use in mobile environments. This work was done with two students at DIKU working on their M.Sc. thesis and Eric Jul at DIKU as described earlier.

The last AMIGOS activity at CBS is the design of a replicated object model with lazy consistency [3]. The goal was to develop an object implementation paradigm working well in mobile environments with low degree of connectivity as well as in large distributed environments with high communication latencies. Our goal was achieved by using lazy consistency known from several DSM implementations but this time in a distributed object memory model. The model still needs to be

[3]Birger Andersen joined the group in July 1997.

implemented for final evaluation. This work was done with Cristian Ionitoiu, at M.Sc. student from Romania.

5.2 Availability versus Consistency in Distributed Computing (ACDC)

In this project we are working on defining and implementing methods for choosing best combinations of availability and consistency in distributed computing systems. Maximum availability and consistency of data requires high bandwidth, low error rate, and low latency on the network. However, by loosing the consistency requirements, high availability of data can be implemented using a network with much lower Quality of Service (QoS) parameters. This had led to investigations into relaxation of ACID properties in distributed databases and other distributed systems.

Further investigation into network and QoS, especially on the Internet, has been done to facilitate CSCW, multi-media, and heavy bandwidth usage in general on networks like the Internet. These application may each contain several streams of data with higher or lower real-time constraints. The idea of filtering out excessive multi-media data using priorities within the stream has been investigated and in this light the state of the art in Internet protocols has been surveyed.

The study of reliable DSM systems by Niels Christian Juul in cooperation with University of California, Riverside has moved from Mirage to general issues in reliable DSM systems.

5.3 Small Office, Home Office (SOHO)

Within the framework of the grant, an infrastructure has been build based on ISDN telephone lines and ISDN-routers. The connectivity of the project members at DIKU and INF has, in general, been enlarged to a level where the technical difference between working with the common resources at night (from home) and in the offices are less than 1.4 ms delay.

References

[1] S. Ajuha, N. Carriero, and D. Gelernter. Linda and friends. *Computer*, 19(8):26–34, August 1986.

[2] Birger Andersen, Carlos Baquero, Eric Jul, and Rui Oliveira. *Proceedings of ECOOP'96 II Workshop on Mobility and Replication.* Available at http://www.diku.dk/distlab/workshops/wmr96/, July 1996. Workshop held at Linz, Austria.

[3] Birger Andersen and Cristian Ionitoiu. Replicated objects with lazy consistency. In Max Műhlhäuser, editor, *Special Issues in Object-Oriented Programming.* dpunkt Verlag, 1997.

[4] Birger Andersen, Eric Jul, Francisco Moura, and Rui Oliveira. *Proceedings of OOPSLA'96 Workshop on Object Replication and Mobile Computing.* Available at http://www.diku.dk/distlab/workshops/ormc96/, October 1996. Workshop held at San Jose, Califonia.

[5] Birger Andersen, Niels Christian Juul, Carlos Baquero, and Francisco Moura. *Proceedings of ECOOP'98 III Workshop on Mobility and Replication.* Available at http://www.econ.cbs.dk/people/birger/WMR98/, July 1998. Workshop held at Brussels, Belgium.

[6] Mikkel Christiansen, Jesper L. Hagen, Brian Nielsen, Arne Skou, and Kristian Q. Skov. The Design and Implementation of ZCRP Zero Copying Reliable Protocol. Technical Report R-98-5011, Aalborg University, Department of Computer Science, 1998.

[7] Jørgen Sværke Hansen and Eric Jul. A scheduling scheme for network saturated NT multiprocessors. In USENIX, editor, *The USENIX Windows NT Workshop 1997, August 11–13, 1997. Seattle, Washington*, pages 39–46, Berkeley, CA, USA, August 1997. USENIX.

[8] Jørgen Sværke Hansen and Torben Reich. Semi-Connected TCP/IP in a Mobile Computing Environm ent. Master's thesis, Department of Computer Science, University of Copenhagen, June 1996. DIKU Project no. 95-6-11.

[9] Jørgen Sværke Hansen, Torben Reich, and Birger Andersen. Semi-Connected TCP/IP in a Mobile Computing Environment. In Heidi Schumann, Jose L. Encarnacao, Thomas Kirste, and Jorge Bonigk, editors, *Information Visualization and Mobile Computing IMC'96*, 1996. Available at ftp://ftp.diku.dk/diku/distlab/amigos/.

[10] Jørgen Sværke Hansen, Torben Reich, Birger Andersen, and Eric Jul. Dynamic Adaption of Network Connections in Mobile Environments. *IEEE Internet Computing*, 2(1), January/February 1998.

[11] Thomas Husfeldt, Brian Nielsen, Finn N. Pedersen, and Dao V. The. Experiments with Video Communication on ATM–network. Technical Report R-97-5002, Aalborg University, Department of Computer Science, 1997.

[12] Morten V. Jensen and Brian Nielsen. Design and Implementation of an Efficient, Layered Video Codec for Heterogeneous Networks. Technical Report R-98-5008, Aalborg University, Department of Computer Science, 1998.

[13] Morten V. Jensen and Brian Nielsen. Real-Time Layered Video Compression using SIMD Computation. Submitted for publication, 1998.

[14] Peter Skov Knudsen. ORDER — an efficient all-software DSM system with a dynamic, adaptable unit of coherence. Master's thesis, Department of Computer Science, University of Copenhagen, August 1996.

[15] Povl T. Koch. *Message-Driven Consistency on a Network of Workstations.* PhD thesis, Department of Computer Science, University of Copenhagen (DIKU), Universitetsparken 1, DK-2100 Copenhagen, DENMARK, April 1996.

[16] Povl T. Koch, Emmanuel Cecchet, and Xavier Rousset de Pina. Global management of coherent shared memory on an SCI clusters. In *Proceedings of SCI Europe 98*, Bordeaux (France), September 1998. To appear.

[17] Povl T. Koch and Xavier Rousset de Pina. SciOS: Flexible operating system support for SCI clusters. In *Proceedings of the 4th International Euro-Par Conference*, number 1470 in Lecture Notes in Computer Science, pages 601–609, Southampton (UK), September 1998.

[18] K. Moldeklev and P. Gunningberg. How a large ATM MTU causes deadlocks in TCP data transfers. *IEEE/ACM Transactions on Networking*, August 1996.

[19] V. Pai, M. Aron, G. Banga, M. Svendsen, P. Druschel, W. Zwaenepoel, and E. Nahum. Locality-aware request distribution in cluster-based network servers. To appear in Proceedings of ASPLOS VIII, October 1998. V. Pai, M. Aron, G. Banga, P. Druschel, and W. Zwaenepoel are from Rice University. M. Svendsen is from University of Copenhagen. E. Nahum is from IBM T.J. Watson.

[20] T. Pontoppidan. Linda in an ATM network. Master's thesis, Department of Information Technology, DTU, January 1996.

[21] D. Raimbault. Evaluation of a Digital ATM network: TCP/IP over ATM. Master's thesis, Department of Information Technology, DTU and IRESTE, Nantes, June 1996.

[22] Jacob Jul Schrøder. Design and performance of striping over atm. Master's thesis, Department of Computer Science, University of Copenhagen, February 16th 1998.

[23] Thomas Stormark. Protocols for OS-based ATM network striping: Design, implementation and evaluation. Master's thesis, Department of Computer Science, University of Copenhagen, April 1998.

[24] Thomas Stormark and Jacob Jul Schrøder. Ydelsesmålinger på ATM-netværk med DEC GIGA switch og ether-netværk med IBM etherstreamer switch. Technical report, Department of Computer Science, University of Copenhagen, March 1996.

[25] Michael Svendsen. ORDER — an efficient all-software DSM system with a dynamic, adaptable unit of coherence, *performance evaluation*. Master's thesis, Department of Computer Science, University of Copenhagen, August 1996.

A Figures

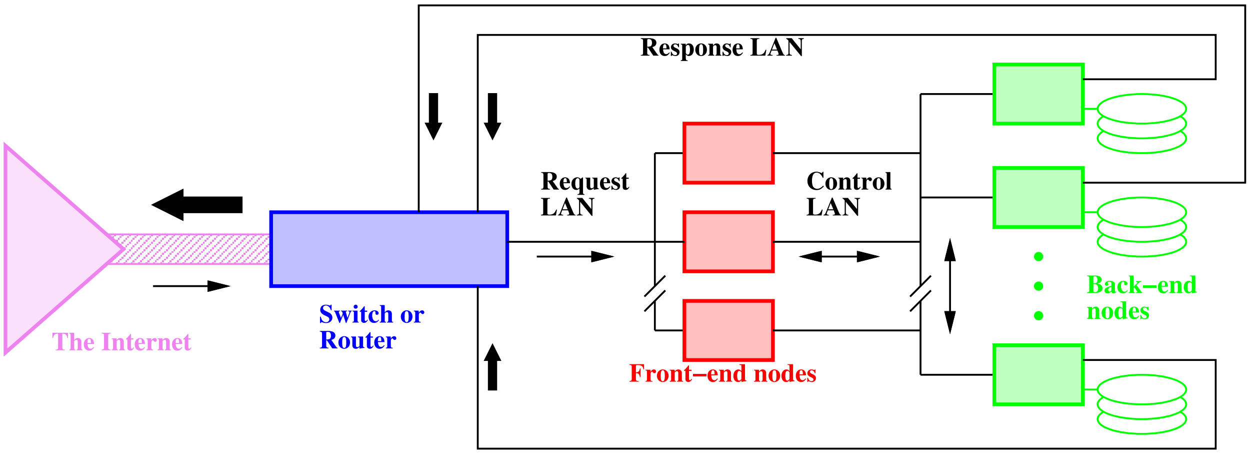

Figure 1: The model of the implementation of the ScalaServer.

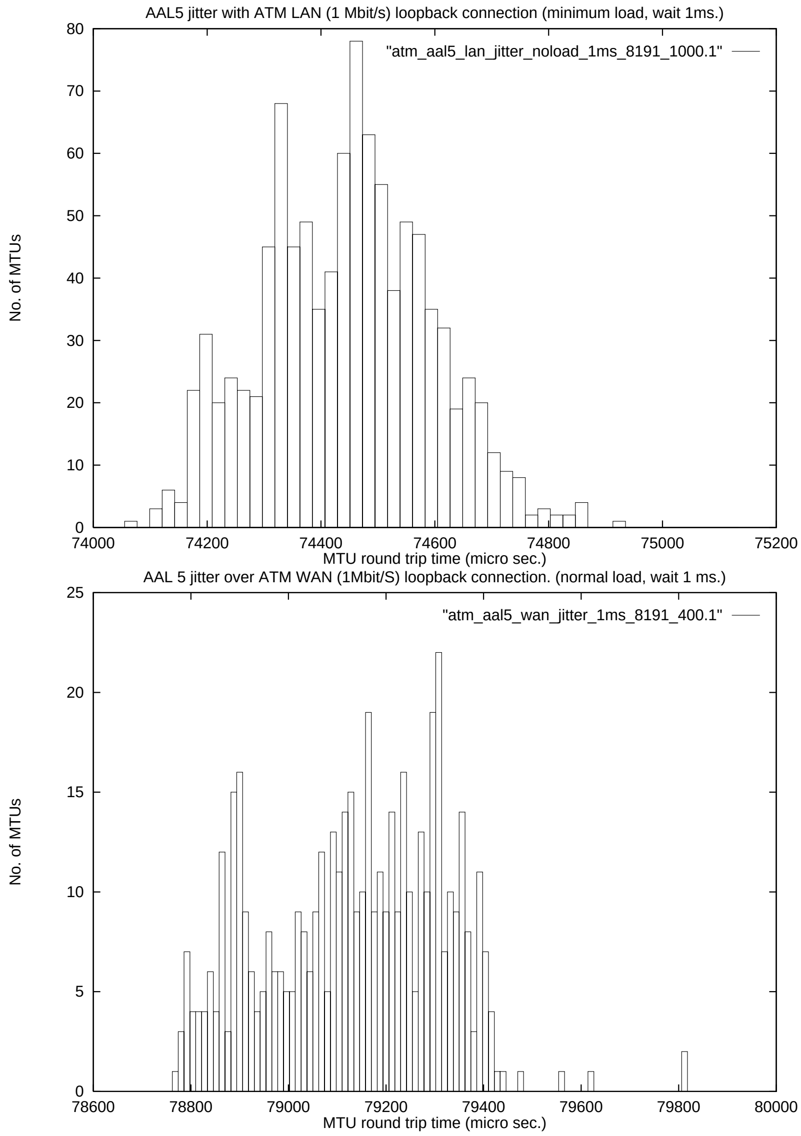

Figure 2: Round-trip jitter over ATM LAN (left) and WAN (right)

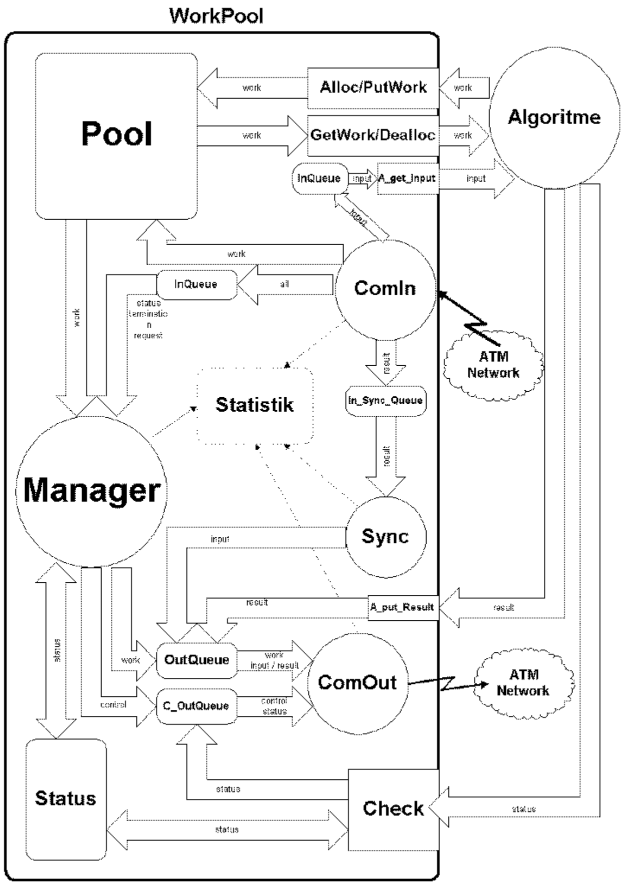

Figure 3: Structure of the Work Pool system

The five circles represent the threads running in each system.