A Comparison of Two Paradigms for Distributed Shared Memory

Levelt, W.G.; Kaashoek, M.F.; Bal, H.E.; Tanenbaum, A.S.

published in
Software, Practice and Experience
1992

DOI (link to publisher)
10.1002/spe.4380221105

document version
Publisher's PDF, also known as Version of record

A COMPARISON OF TWO PARADIGMS FOR DISTRIBUTED SHARED MEMORY

Willem G. Levelt
M. Frans Kaashoek
Henri E. Bal
Andrew S. Tanenbaum

Department of Mathematics and Computer Science
Vrije Universiteit
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

ABSTRACT

This paper compares two paradigms for Distributed Shared Memory on loosely coupled computing systems: the shared data-object model as used in Orca, a programming language specially designed for loosely coupled computing systems and the Shared Virtual Memory model. For both paradigms two systems are described, one using only point-to-point messages, the other using broadcasting as well.

The two paradigms and their implementations are described briefly. Their performances on four applications are compared: the travelling-salesman problem, alpha-beta search, matrix multiplication and the all-pairs shortest paths problem. The relevant measurements were obtained on a system consisting of 10 MC68020 processors connected by an Ethernet. For comparison purposes, the applications have also been run on a system with physical shared memory. In addition, the paper gives measurements for the first two applications above when Remote Procedure Call is used as the communication mechanism.

The measurements show that both paradigms can be used efficiently for programming large-grain parallel applications, with significant speed-ups. The structured shared data-object model achieves the highest speed-ups and is easiest to program and to debug.

KEYWORDS:
Amoeba Distributed shared memory Distributed programming Orca
Shared data-objects Shared Virtual Memory

1. INTRODUCTION

As computers become cheaper, there is an increasing interest in using multiple CPUs to speed up individual applications. There are basically two design approaches to achieve this goal of high performance at low cost: multiprocessors and multicomputers. Multiprocessors contain physical shared memory; processors in a multiprocessor can communicate by reading and writing words in this memory. Multicomputers, on the other hand, do not contain physical shared memory; processors communicate by exchanging messages. In this paper, we compare two hybrid approaches that allow processors in a multicomputer to communicate through *Distributed Shared Memory*. We shall discuss the implementation of these two hybrid approaches, describe some applications we have written on both, and present measurements of the performance.

The key property of a multiprocessor is that each processor has a consistent view of the physical shared memory. When any processor writes a value to memory, any other processor that subsequently reads the word will retrieve the value just written. To achieve these semantics special and complicated hardware (such as snoopy caches or fast switching networks) is needed to build a multiprocessor. The software for multiprocessors, however, is straightforward. All processes run within a single shared address space, so they can easily share data structures and variables. When one process updates a variable, other processes will immediately see this change. The methods to allow processes to synchronize their activities are well understood. Multiprocessors are hard to build but are relatively easy to program.

In a multicomputer each processor has its own private memory which it alone can write and read. These processors can be connected using standard networking technology. The software for multicomputers, on the other hand, is much more complicated. Processors in multicomputer must communicate using message-passing. Although much effort is put into hiding the message-passing from the programmer it is difficult to make it entirely transparent. A very popular approach to hide the communication is remote procedur call. The idea is to make the communication look like an ordinary procedure call. The programmer, however, still has to be aware that the semantics of remote procedure calls is different from a local procedure call. For example, passing pointers as parameters in an RPC is difficult, and passing arrays is costly. This makes programming on a multicomputer difficult.

To combine the advantages of multiprocessors (easy to program) and multicomputers (easy to build), communication paradigms that simulate shared data on a multicomputer have become popular. These mechanisms are implemented with message-passing, but they provide the illusion of shared data. They provide a *Distributed Shared Memory* (DSM). Processes on different processors run in the same address space. All processes can access the variables in the shared address space directly. They communicate through these variables.

In this paper we compare two approaches to Distributed Shared Memory: The Shared Virtual Memory model (SVM)[1, 2] and the shared data-object model.[3] Both models provide logical shared

memory, but use different abstraction techniques. Shared Virtual Memory provides the illusion of true physical shared memory. The shared data-object model encapsulates shared data in user-defined objects.

Shared Virtual Memory

The Shared Virtual Memory model simulates true physical shared memory on a loosely-coupled system. A number of processes share a single address space. This address space is divided into pages, which are distributed among the processes. Processes either have **no**, **read** or **write** access to a page. **Read**-pages can be replicated on multiple processors to reduce access times. The system provides a coherent address space: a read operation always returns the value of the most recent write to the same address. Mutual exclusion synchronization can be implemented by locking pages.

The SVM paradigm can be viewed as a low-level *unstructured* DSM approach. The address space is divided into fixed sized pages with no relation to the structure of the stored data. The SVM system is completely transparent to the processes that use it. There is no distinction between shared and non-shared data. Even the operating system itself can use it for communication between processors.

A disadvantage of this low-level approach is the lack of support for the application programmer. The SVM can only be accessed with primitive operations, such as load, store and lock. When two independent and heavily used variables reside on the same page, this will lead to contention. To avoid unacceptable performance penalties, such variables have to be placed on different pages. This has to be done by the programmer, because a compiler has not enough knowledge to decide this. The compiler does not know how variables are mapped onto pages and it can not decide at compile time if two objects that are possibly accessed through pointers are independent. For popular languages, like C, this problem becomes even harder for memory that is allocated during run-time. The compiler has no way of knowing how this memory is going to be used and whether two blocks of dynamically allocated memory will be used independently of each other. Therefore the application programmer must be constantly aware of how the data-structures are accessed and where they are placed in the SVM address space, or suffer an unacceptable performance penalty. Thus in practice the Shared Virtual Memory is not really transparent to the user, unless the user does not care about performance. Furthermore the SVM is just one global flat address space; no access protection or type-security is enforced by the system. This makes distributed programming difficult.

Shared data-object model

The shared data-object model is a high-level, *structured*, approach to Distributed Shared Memory. In contrast to the SVM model, which is implemented by the kernel using hardware support, the shared data-object model is implemented outside the kernel completely in software. A Run Time System (RTS), using information generated by the compiler, keeps the DSM coherent.

In a shared data-object language, shared data are encapsulated in objects. A shared data-object is an instance of a user-defined abstract data type and can only be accessed through operations defined in the object's specification. These operations are executed indivisibly and the RTS ensures that all processes that share the object see the result. Partitioning of the DSM address space is not defined by the system, as in the SVM approach, but implicitly by the application programmer. The unit of programmer-defined sharing is the shared object, not the page. As an object is an instance of an abstract data type, variables that are independent of each other will typically reside in different objects. False sharing as in SVM will be less of a problem in the shared data-object model.

One way to share objects in a shared data-object language is by dynamic creation of processes. The parent process can pass any of its objects as a shared object to its children, which communicate through these shared data-objects. The RTS may replicate these objects on more than one processor, to reduce the access time. Other sharing mechanisms are also possible.

Also in contrast to the SVM approach, the Distributed Shared Memory is not treated as a flat address space that can be accessed in an arbitrary way. The semantics of the language restrict the scope of shared variables, similar to scope rules in a sequential language. The processes can only access those shared data-objects they are supposed to access. Furthermore, the shared data can only be accessed with high-level operations, which the programmer defines in the abstract data type. Because the execution of these operations is indivisible, mutual exclusive access is provided implicitly. Figure 1 summarizes the differences between these two paradigms.

	Shared Virtual Memory	**Shared data-object model**
Implementation level	In kernel, using hardware support	Completely in software
Unit of sharing	System defined page	User defined object
Unit of synchronization	Machine instruction	Procedure
Data placement	Explicit	Implicit
Address space	Flat	Structured

Fig 1: Differences between the SVM and shared data-object paradigm

Related work

Several systems that provide shared data on distributed computing systems have been designed. Most of them do not provide a structured address space, but just a flat one. Examples of these are page-based systems like IVY,[4] Shiva,[5] Mirage,[6] Mether,[7,8] and Munin.[9] Each one has its own coherency technique, but they are all based on the ideas Li put forward in his thesis1. For example, Mirage uses a technique similar to Li's fixed distributed management algorithm, but as an extra feature it

maintains timers to prohibit pages to be paged out too soon after paging in. Other unstructured Distributed Shared Memory systems include the Agora shared memory,[10] where the shared data consists of immutable data elements, which can be accessed through mutable maps. Data are changed by adding new data elements and changing the map. In Agora, however, read operations can return stale data.

An example of a structured DSM system is the Tuple Space used in Linda.[11, 12] The Tuple Space contains tuples, which are similar to records. A tuple can only be changed by taking it out of the Tuple Space, modify it locally, and reinsert it. Another example of a structured DSM system is Sloop, which supports a shared object space.[13] A discussion of these and other Distributed Shared Memory paradigms is provided by Bal and Tanenbaum.[14]

Outline for this paper

The purpose of this paper is to compare two paradigms for Distributed Shared Memory: a structured and an unstructured one. Does it pay off to implement an application oriented DSM system, like the shared data-object model, or is the general approach of the Shared Virtual Memory equally good? To make a direct comparison possible, we have implemented these two DSM paradigms on the same distributed computing system, each in two different ways. Using these systems, we have measured the performance of four applications: the travelling-salesman problem, alpha-beta search, matrix multiplication and the all-pairs shortest paths problem. To place these measurements in perspective, we also give the measurements obtained by running these applications on a physical shared memory system and we supply measurements for TSP and alpha-beta using pure message-passing as communication mechanism. As a side-effect of this research we can present real-time performance figures of two SVM implementations. To the best of our knowledge, nobody (including Li) has ever published such figures. We will, furthermore, discuss some programmability aspects of these two paradigms.

The remainder of this paper is structured as follows: the next section describes the hardware and the operating system (Amoeba) we have used for our experiments. Then we discuss the two SVM systems we have implemented. The following section describes the shared data-object model implementations. Then we describe the applications we have implemented, after which we discuss the measurements and other aspects of the two paradigms. In the last section we present our conclusions.

2. THE AMOEBA DISTRIBUTED OPERATING SYSTEM

Amoeba is an operating system specially designed for loosely-coupled computing systems.[15] The Amoeba architecture consists of four principal components.

First are the workstations, one per user, on which users can carry out editing and other tasks that require fast interactive response. We are currently using SUN-3s as workstations. Second are the pool processors, a group of CPUs that can be allocated dynamically as needed, used, and then returned to the pool.

Third are the specialized servers, such as directory servers, file servers, data base servers, bank servers, boot servers, and various other servers with specialized functions.

Fourth are the gateways, which are used to link Amoeba systems at different sites and different countries into a single, uniform system. The main function of the gateways is to isolate users from the peculiarities of the protocols that must be used over the wide-area networks.

Amoeba is an object-based system. The system can be viewed as a collection of objects, on each of which there is a set of operations that can be performed. The list of allowed operations is defined by the person who designs the object and who writes the code to implement it. Both hardware and software objects exist.

Associated with each object is a *capability*, a kind of ticket or key that allows the holder of the capability to perform some (not necessarily all) operations on that object. Capabilities are protected cryptographically to prevent users from tampering with them.

The Amoeba kernel runs on each processor. It provides communication service and little else. Communication between processes is done with Remote Procedure Calls (RPC). An RPC consists of the following steps: A client process sends a request to a server process, and blocks. The server accepts the request, processes it, and sends a reply message back. At the arrival of this reply an acknowledgement packet is sent, the client is unblocked again and the RPC has succeeded. A message consists of a header of 32 bytes and a buffer of up to 30,000 bytes.

In addition to RPC, the Amoeba kernel provides reliable, order preserving, broadcast communication.[16,17] Even in the presence of communication failures, the kernel guarantees that if two processes broadcast a message simultaneously, one of them goes first and its message will be processed at all processors before the other one. In normal operation, this protocol needs two messages per broadcast.

A process consists of a collection of threads that run on the same processor. These threads share a single address space, but they all use a dedicated portion of this address space for their private stack. In this version of Amoeba (3.0) threads are not preempted; conceptually they run until they block.

3. UNSTRUCTURED DSM: THE SHARED VIRTUAL MEMORY PARADIGM

The Shared Virtual Memory model simulates true physical shared memory on loosely-coupled systems. This simulation is accomplished by dividing the SVM address space into fixed-size pages, which are distributed among the processes forming an application. The SVM system coordinates the movement and validity of the pages. Pages with read-only access can be replicated at multiple processors to reduce access times. An invalidation scheme is used to keep the address space coherent. For example, when a process tries to write on a page, all other copies of that page are first invalidated, then the write is permitted.

To synchronize multiple requests for the same page, each page is owned by a specific process

(which may change in time). This owner process has a valid copy of the page, and all requests for that page are sent to it. To be able to keep the address space coherent, the SVM system must, for each page, keep track of the owner process and of all the processes containing a valid copy.

Li has proposed several methods for organizing and maintaining this ownership and copy-list data.[1] He first made a distinction between centralized and distributed management of ownership data:

- With Centralized Management, one process, the Central Manager, keeps track of all pages. All requests for a page are first sent to the central manager, which in turn sends it to the current owner of the page. In the regular *Central Manager* method, the Central Manager keeps track of the copy-list data too. In the *Improved Central Manager* method, the copy-list information is decentralized; it is stored at the process that owns the page.

- With distributed management, the ownership data is distributed among the processes. This can be done in a fixed or dynamic way.

  Fixed Distributed Management means that each process manages the ownership data of a predetermined subset of the pages, which remains fixed throughout the whole lifetime of an application. Copy-list data can either be stored at the manager or at the owner.

  In the *Dynamic Distributed Manager* approach, ownership (and copy-list) data is transferred together with the page itself. Only the owner process knows for sure who the owner is.

An important design choice is the page size. The page size is a tradeoff between the cost of sending a message and memory contention. Because of set-up costs, the time to send a message is not proportional to the size of the message. This favors large page sizes. On the other hand, the bigger the page, the greater the chance that two or more key variables used by different processors will accidentally reside there, leading to contention.

Which page size is optimal is an open question, it will probably vary from system to system and from application to application. IVY uses 1K, Shiva uses 4K, Mirage uses 512 bytes and in Mether the application can choose between two page sizes: 32 bytes or 8K. The Amoeba operating system internally uses a page size of 4K for the partitioning of memory space. We therefore decided to use 4K as the page size too.

We have implemented two SVM systems: the first using Centralized Management, the second using Dynamic Distributed Management. The first system uses RPC for interprocess communication. The second uses broadcasting as well as RPC. We now describe each system in turn.

Central Manager System

In our implementation, every application runs in its own private SVM address space, managed by its own Central Manager. The system consists of two parts: one Central Manager process and a collection of SVM handler threads, one associated with each process in the application. The Central

Manager is an ordinary user program that handles the bulk of the protocol. All requests for pages are sent to the Central Manager, which keeps track of page-ownership and copy-list data. Each SVM handler manages the page-access rights for its associated process in the application. The SVM handler threads run in kernel space, because they have to access kernel data structures (e.g., the page table). This is shown in Figure 2.

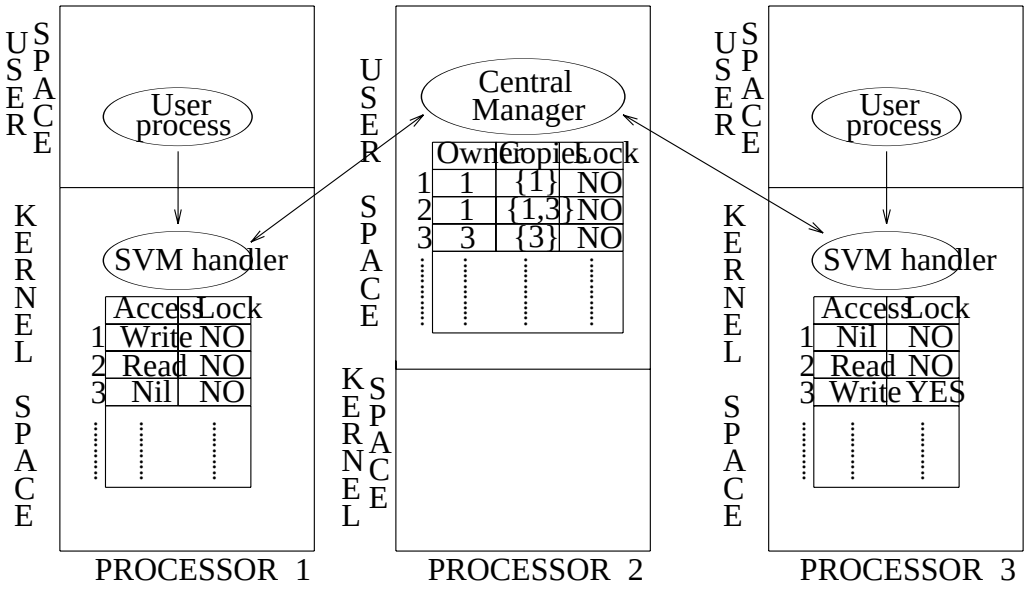

Fig 2: processor outline in the Central Manager system

The Central Manager maintains a table with copies of pages to which no process has write access. Requests for such a page can be handled without contacting the owner. The Central Manager can reply immediately with the copy it stored earlier.

On receiving a write request, the Central Manager has to invalidate all other copies of that page. If this were handled by a single thread, the copies would be invalidated one by one, which is slow. Therefore, the Central Manager has a thread for each SVM handler in the application. These threads are used to send the invalidation messages in (pseudo) parallel. In this way, the next invalidation RPC can be started before the previous one has been completed.

When the Central Manager receives a write-page request from a process that still has a valid read-copy, the Central Manager does not send a copy of the page. Thus a read fault, followed immediately by a write fault on the same processor, causes only one page transfer.

A typical communication pattern caused by a write fault in one process, followed by a read fault in a different process (on the same page), is shown in Figure 3. At first (top of the figure), processes 2 and 3 have a read-copy of the page. Process 1 tries to write to the page, which causes a write fault. The SVM handler of process 1 sends a write request to the Central Manager. The Central Manager has a valid copy available, so it only has to invalidate the other copies. When both copies are

invalidated, it can send the page to process 1.

Shortly after the read-copy at process 3 is invalidated, a read fault occurs, and a read-page request is sent to the Central Manager. Because the Central Manager is still handling the request from process 1, this second request is blocked. When the acknowledgement from process 1 arrives, the Central Manager can handle the blocked request. It requests a read-copy from process 1, stores it locally, and sends it to process 3.

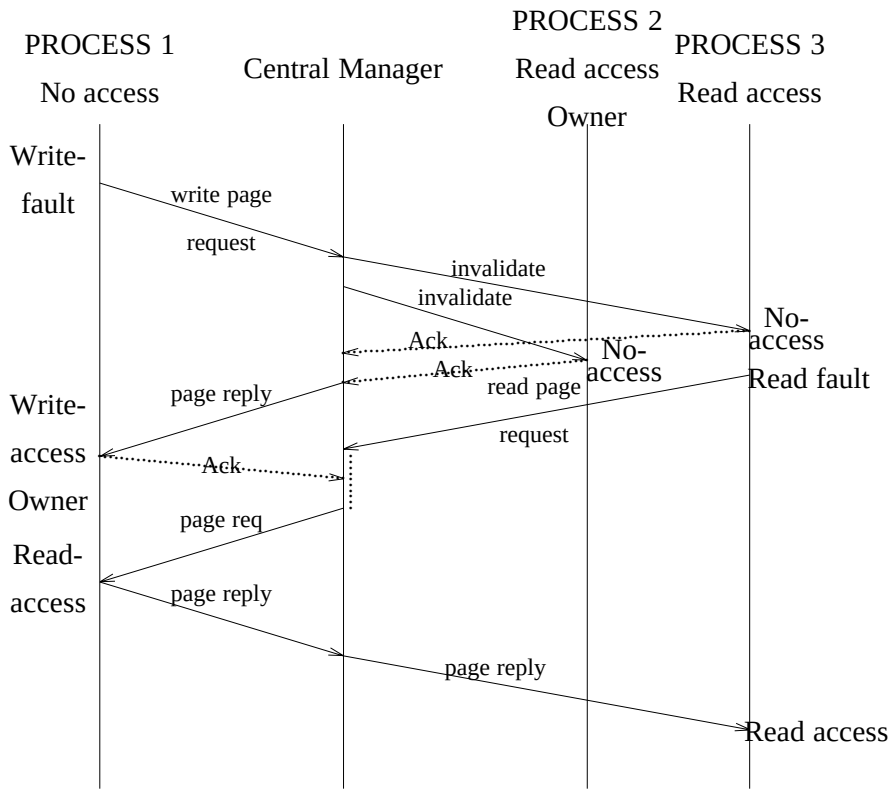

Fig 3: A write fault in process 1, followed by a read fault in process 3 in the Central Manager system

For mutual exclusion synchronization a page-locking mechanism is provided. A process can lock a page for reading or for writing. A read-locked page will not be invalidated until it is unlocked again. Write locking is more powerful: It provides a process with the sole copy of a page. Upgrading of a read-lock to a write-lock is not possible.

When a page is not locked, a page fault takes two RPCs to complete: the first from the requester to the Central Manger, the second from the Central Manager to the owner. When the Central Manager has a valid copy, just one RPC is needed. In addition, on a write fault, an extra RPC is necessary for each copy of the page.

There are two important differences between this Central Manager system and the system proposed by Li. First, in Li's Central Manager system, one of the processes in the application also works as the Central Manager. In our system, the Central Manager is a distinct process, which does not

work on the application.

Our approach has the advantage that the Central Manager can run entirely in user-space. The kernel is therefore not enlarged by this code. This fits the Amoeba principle of keeping the kernel small. However, implementing the Central Manager as a distinct process in user-space has a disadvantage too, which is caused by the non-preemptive scheduling of processes in Amoeba 3.0.

In Amoeba, user-processes are rescheduled twice a second, or when they execute a system call. When the Central Manager has to service a request, it is unacceptable that it has to wait for rescheduling. Therefore, the Central Manager must be the only process running on a processor, so no other user-processes can delay it.

The second difference is that Amoeba uses Remote Procedure Call as communication mechanism, whereas Li uses straight message-passing. The communication pattern in Li's (improved) Central Manager system, in case of a page fault, is as follows:

1. A message from the faulting process to the Central Manager.
2. A message from the Central Manager to the owner process.
3. A message from the owner to the faulting process.

This is not appropriate for RPC, where a process always waits for a reply. The faulting process contacts the Central Manager with an RPC, and it will block until the Central Manager sends a reply. The process cannot accept a reply from another process. The communication pattern therefore is as follows:

1. A request from the faulting process to the Central Manager.
2. A request from the Central Manager to the owner process.
3. A reply from the owner to the Central Manager.
4. A reply from the Central Manager to the faulting process.

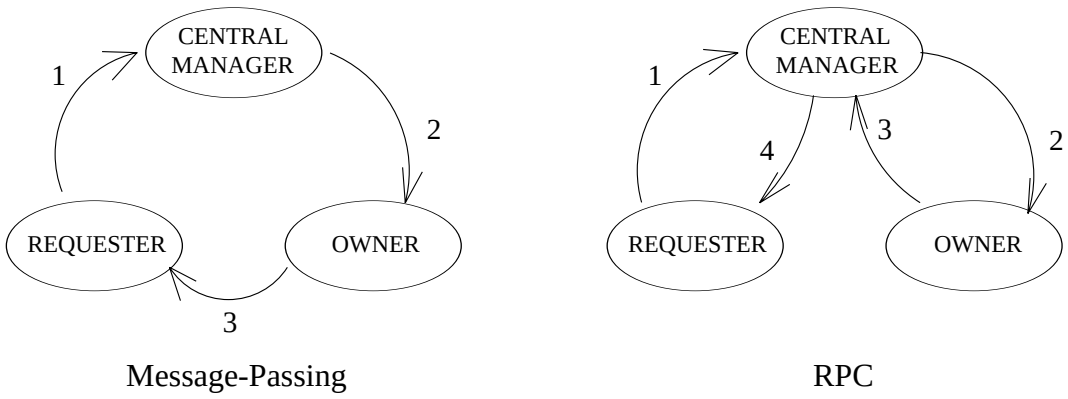

Fig 4: Communication patterns in IVY and our Amoeba implementation

These communication patterns are depicted in Figure 4. Here, the RPC uses four messages while Li's

scheme uses three. An Amoeba RPC uses three packets, a request, a reply and an acknowledgement. If the messages in Li's scheme use two packets (one for the message and one for the acknowledgement), both approaches use the same number of physical packets. When a copy of a page has to be transferred (needing a large message), Li's scheme only transmits the page once, while the RPC approach has to transmit it twice. Thus, our implementation of the Central Manager performs worse (due to the RPC) than Li's implementation in this respect.

Dynamic Distributed Management System

In our Distributed Management method each process has its own SVM handler, like the processes in the Central Manager method. But in contrast to the Central Manager method, there is no Central Manager. In the event of a page fault, the SVM handler associated with the faulting process broadcasts a request for the page. All other SVM handlers receive this broadcast. One of them is the owner and it will send the page to the requester using an RPC. From this point on, the requesting process is the new owner, and it will respond to the next request for this page. Thus in contrast to the Central Manager method, where ownership only changes on a write fault, in the Distributed Management method ownership also changes on a read fault.

A write-request broadcast causes all SVM handlers to invalidate their copy of the page, so no copy-lists are needed. This scheme only works because each process gets the broadcasts in the same order, and no broadcasts are lost. All SVM handlers have the same view of the system and there can be no confusion over which process is the owner and which processes have a valid copy.

In this system, pages can be locked too. The SVM handlers try to handle a broadcast affecting a locked page as soon as possible. For instance: on receiving a write-request for a read-locked page, the SVM handler will send the page right away. The invalidation is done as soon as the page is unlocked again.

A page fault (read and write) takes one broadcast and one RPC to complete. When a read-copy is upgraded to a write-copy, no RPC is used. A typical communication pattern in case of a write fault is shown in Figure 5.

4. STRUCTURED DSM: THE SHARED DATA-OBJECT MODEL

The shared data-object model was proposed by Bal and Tanenbaum[14] to ease the programming of parallel applications. They and Kaashoek designed and implemented a new parallel programming language *Orca*[18, 3, 19, 20] which uses the shared data-object model as communication paradigm.

In the shared data-object model, shared data are encapsulated in *data-objects**, which are variables of user-defined abstract data types. An abstract data type has two parts:

---

* We shall sometimes use the term "object" as a shorthand notation for data-objects. Note, however, that unlike the situation in most parallel object-based systems, objects in this model are purely passive.

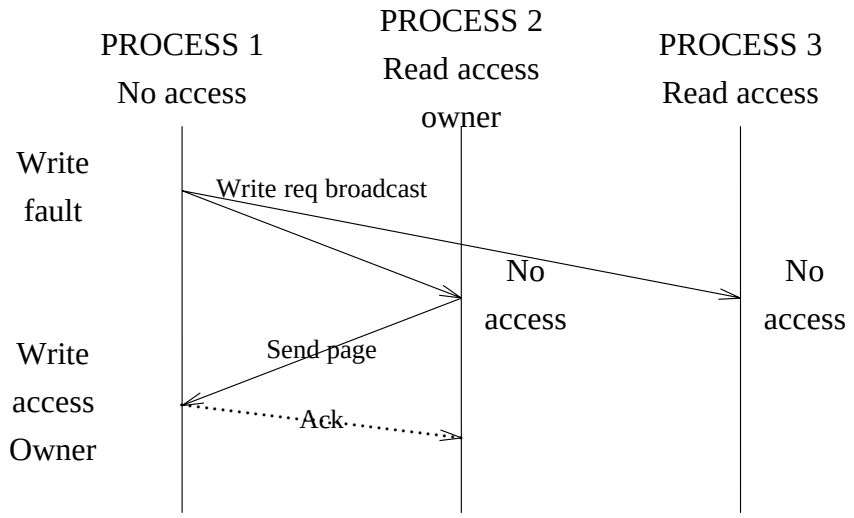

Fig 5: A write fault in the Distributed Management method

- A specification of the operations that can be applied to objects of this type.
- The implementation, consisting of declarations for the local variables of the object and code implementing the operations.

Instances (objects) of an abstract data type can be created dynamically. Each object contains the variables defined in the implementation part. These objects can be shared among multiple processes, typically running on different machines. Each process can apply operations to the object, according to their listing in the specification part of the abstract type. In this way, the object becomes a communication channel between the processes that share it.

The shared data-object model uses two important principles related to operations on objects:

1. All operations on a given object are executed *atomically* (i.e., *indivisibly*). To be precise, the model guarantees *serializability*[21] of operation invocations: if two operations are applied simultaneously to the same data-object, then the result is as if one of them is executed before the other; the order of invocation, however, is nondeterministic.

2. All operations apply to *single* objects, so an operation invocation can modify at most one object. Making *sequences* of operations on different objects indivisible is the responsibility of the programmer.

Linguistic support for the shared data-object model

The new programming language Orca gives linguistic support for the shared data-object model. Orca is a simple, procedural, type-secure language. It supports abstract data types, processes, a variety of data structures, modules, and generics. It does not support global variables and pointers, but provides a new type−constructor *graph* that allows the programmer to build any datastructure that can be constructed with pointers. Parallelism in Orca is based on explicit creation of sequential processes. Processes are similar to procedures, except that procedure invocations are serial but newly created processes run in parallel with their creator.

The parent can pass any of its *data-objects* as a shared parameter to its children. The parent and child can communicate through this shared object, by executing the operations defined by the object's type. The children can pass the objects to *their* children, and so on. This mechanism can be used for sharing objects among any number of processes. If any of these processes performs an operation on the object, they all observe the same effect, as if the object were in shared memory, protected by a lock variable.

Processes in a parallel program sometimes have to synchronize their actions. This is expressed in Orca by allowing operations to *block* until a specified predicate evaluates to *true*. A process that invokes a blocking operation is suspended for as long as the operation blocks. The data-structuring mechanism in Orca is type-secure.

Implementations of the run time system

A run time system for Orca is, among other tasks, responsible for managing processes, shared data-objects, and (un)marshalling of data structures. The RTS works closely together with the compiler to perform its task. The compiler generates *descriptors* to allow the RTS to marshal complex objects, like graphs, and to perform extensive run-time type checking. These descriptors describe the layout of an object and the sizes of the components of the object. For example, for a graph, such a descriptor contains a table with pointers to the actual nodes in the graph. The nodes in a graph are represented as offsets in this table, rather than as pointers, to make marshalling of a graph easy. Using the descriptors the RTS can marshal any object and check all array and graph references. If a node is deleted from a graph and subsequently referenced, a run-time error will be given.

Another important task that the compiler performs and that is used by the RTS, is the classification of operations. The compiler tells the RTS which operations change the internal data of an object (a write-operation) and which operations do not change the internal data (a read-operation). Using this information the RTS can implement efficient replication schemes. Bal et. al. have experimented with two implementations of the RTS on a distributed computing system using different replication schemes.[22] We shall describe these two implementations briefly below.

Structure of the multicast RTS

The first implementation replicates all objects on all processors and uses a distributed update protocol based on reliable multicast messages for keeping all copies up to date. The RTS contains the same protocols as used above (for distributed management) for implementing reliable multicast on top of an unreliable network.

The multicast RTS is best thought of as a new kind of operating system kernel designed specifically for parallel applications. Shared data-objects are created and updated by the kernel. User processes can read local copies of objects directly, without using the kernel. If a user process applies a write operation to a shared object, the user process traps into the kernel; the local kernel multicasts the operation and its parameters to the kernels of all processors; each kernel then applies the operation to its local copy of the object. As the multicast primitive is reliable and indivisible, write operations are executed by all kernels in the same order.

Structure of the remote procedure call RTS

The second RTS is implemented on top of the Amoeba distributed operating system. It replicates and migrates objects selectively based on statistical information gathered during run time. It updates the replicas by using a 2-phase primary-copy protocol, using point-to-point communication only. For each object, one processor contains the primary copy of the object and zero or more other processors contain secondary copies. Read operations are applied to the local copy, if available, and write operations are forwarded to the processor with the primary copy.

Updating replicas through point-to-point messages is more expensive than through multicast messages. Furthermore, the communication overhead with point-to-point messages grows linearly with the number of replicas. The RPC RTS therefore replicates objects *selectively*. A given processor only keeps a copy of an object if it reads the object frequently. Run-time statistics are maintained for deciding where to store the primary and secondary copies of each object. Application programmers are not aware of this; it is completely automatic.

There is one incarnation of the RTS on each processor. Each incarnation of the RTS starts a number of *listener tasks* that wait for requests from remote machines. A request can either be:

1. An update of a replicated object.
2. An operation to be performed on an object whose primary copy resides on this machine, on behalf of a remote machine.
3. A request to create a new Orca process.

When a message needs to be sent to another machine, the task wishing to send the message deposits it in a per-machine queue. For each queue—and thus for each remote machine—there is a *talker task* that handles the messages in the queue. A talker repeatedly waits for a message, sends it to the remote machine (using Amoeba RPC), and optionally returns a reply to the task that requested the

message to be sent.

With this approach, the replicas of an object can be updated in parallel by depositing update messages in multiple queues. This programming model is similar to the *promises* model.[23] If each Orca process performed the RPC calls itself, parallel updating would not be possible, since RPC calls are blocking. As another advantage of our approach, multiple objects residing on the same machine can be updated in parallel.

5. EXPERIMENTS AND RESULTS

The most important properties of a Distributed Shared Memory system are its performance and its ease of programming. In order to judge DSM-systems these properties must be compared. The way of comparing programmability aspects is easy: just program a few applications. However, the interpretation is subjective. In contrast, performance comparison can be done objectively (using the system clock). But the question of what to compare is not straightforward.

Because the Shared Virtual Memory system and our shared data-object system use different methods for replication and updating, it is not possible to compare low level operations. We therefore decided to compare them by running four parallel applications: the travelling-salesman problem, alpha-beta search, matrix multiplication and the all-pairs shortest paths problem. To make a direct comparison possible, for all applications both systems used the same input, the same division of work and the same algorithm.

We do not include the time to start the processes. Time is measured from the moment all processes have been started, to the moment that they all have finished their computations. The SVM programs are written in C, using two extra systems calls to lock and unlock the page(s) containing a shared variable.

- lock(variable, sizeof(variable), lock_type)
- unlock(variable, sizeof(variable), lock_type).

In the shared data-object model different shared variables are handled independently. In the Shared Virtual Memory model, variables residing on the same page are always handled together. When the variables are independent, this can lead to unnecessary contention. To get good performance, it is necessary to place independent variables on different pages and variables that are mostly referenced together on the same page.

To emphasise the importance of good placement, we ran each test in the SVM system twice, once with all variables optimally distributed, and once with the shared variables packed together on as few pages as possible. We shall give performance measurements of both cases.

As described earlier, the Central Manager has to run on a separate processor. We counted this processor when calculating the speed-ups. The Central Manager measurements therefore start at 2 processors.

We ran all measurements of the SVM model at least four times. The variances of the measurements were all below 1.0 and mostly around 0.25. The measurements of the shared data-object model were run three times. For comparison purposes we also ran all applications on a multiprocessor with physical shared memory. In addition, we supply measurements for TSP and alpha-beta, programmed using Remote Procedure Call as communication mechanism. How these were programmed is described by Bal et al.[24]

Appendix A lists all these timings. To keep the tables compact, we supply the mean values only. In addition to these tables there are a few graphs. They depict the speed-ups achieved by each of the six systems. For the SVM systems, we used the measurements achieved with good placement.

Hardware environment

All measurements were obtained on a distributed system consisting of up to 10 MC68020 CPUs, interconnected through Lance chips to a 10Mbit/s Ethernet. Each processor has 2 MB of private memory. For the physical shared memory system we used the same type of processors but now connected by a VME-bus to 8 MB of shared memory. Unfortunately, we only had 9 working processors, so no measurements for 10 processors can be supplied. This system does not have snoopy caches, so every access to shared memory goes through the VME-bus.

In our Amoeba version, a 4-byte and a 4K RPC take respectively 1.4, and 6.7 msec to complete.[15] A 20 byte broadcast to 10 processors, using hardware multicast, takes 1.5 msec to complete.[16]

Basic Operations in the Shared Virtual Memory System

In addition to taking measurements of applications, we also measured the timing of some basic operations in our two SVM systems. They are shown in Figure 6. The time to get a page is the time as seen by the user process on a page fault. Hence, it includes the time spent in the kernel and the time to enter and leave the kernel. These measurements were taken on an otherwise almost idle system.

Travelling Salesman Problem

In the travelling-salesman problem (TSP), a set of cities and the distances between these cities are supplied. Given an initial city, the problem is to find the shortest route that visits every other city exactly once. This can be solved with a branch-and-bound algorithm in which a tree of possible solutions is built. Each node of the tree depicts a city, and a path in the tree describes a route through these cities that are depicted by the nodes on the path. A path from the root-node to a leaf-node describes a route that visits every city exactly once.

The bounding rule uses the length of the shortest route found so far. If the length of a path to a node exceeds this minimum length, all paths through this node will be longer than the shortest route, so they do not have to be searched.

Central Manager

Operation	Time (msec)
Get page, stored at manager	7.1
Get page, not stored at manager	14.0
Additional time to invalidate copies	
2 copies	4.4
4 copies	7.4
6 copies	10.4
8 copies	13.9

Distributed Management

Operation	Time (msec)
Get read or write page	10.2
Invalidate all copies of a page	3.0

Centralized & Distributed Management

Operation	Time (msec)
Lock & unlock an in-memory page	0.275
Page-fault schedule time to SVM-handler and back	0.267

Fig 6: Basic operations in the Shared Virtual Memory systems

The sequential branch-and-bound algorithm applies Depth First Search on the tree, using the nearest-city-first heuristic. This can easily be parallelised. The search tree is static and subtrees can be searched independently by multiple processes. The only dynamic quantity is the length of the shortest route found so far. This variable (the *minimum*) is not updated very often, perhaps 10 times during the whole search. It will be read very frequently, but that causes no overhead, because reading is a local operation.

We used as input an 12×12 distance matrix, and a starting city. The work is divided by a distribution process, which constructs the top two levels of the search tree, i.e. generating $11 \times 10 = 110$ partial paths. These partial paths are searched further (with sequential DFS) by worker processes (and the distribution process) until all paths have been inspected.

We used three randomly generated distance matrices. The measurements we supply are the mean values for these three graphs. Figure 7 depicts the speed-ups. Both systems achieve close to linear speed-ups. This is explained by the fact that, as soon as one of the workers discovers a new minimum, it updates the shared variable and all workers can use this minimum to prune their subtree earlier. Because in parallel search a low minimum is sometimes found earlier than in sequential search, it can happen that fewer nodes are searched. For an input matrix where a good minimum is found quickly, the speed-ups in some cases superlinear. The RPC implementation performs worse, because each worker process can only accept a new value for the minimum when it has finished its

previous job.

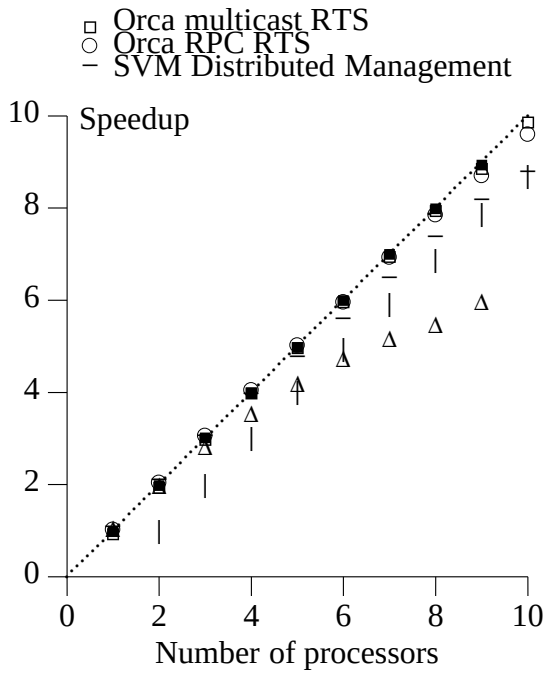

Fig 7: Travelling-salesman problem

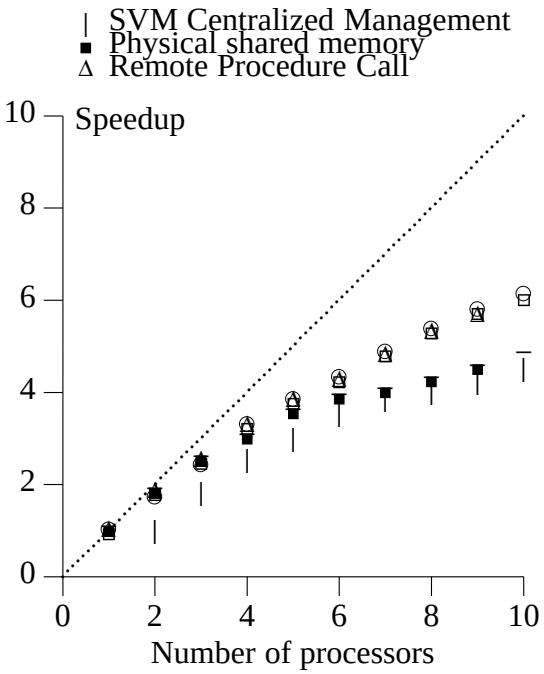

Fig 8: Alpha-beta

Alpha-beta search

Alpha-beta search is typically used for two-player, zero-sum, board games like chess and checkers. It decides on the next move by evaluating the current board-position.[25] To evaluate a board-position, a tree is built with the current board-position as the root-node and for every possible move, a child with the board-position after that move. The child with the worst possible board-position (which is a position for the opponent) indicates the best possible move. Before a value can be assigned to the root-node, all its children have to be evaluated. This is done by first evaluating their children, etc. This process must stop somewhere, so at a certain level (*ply*) a *static evaluation* is done on a position. This assigns a value (weight) to a position.

The whole procedure can be programmed compactly by using positive values at odd levels, and negative values at even levels. The value of a node is expressed as follows: assign to a node the value of the child with the lowest value, negated.

To reduce the search space, the search tree is pruned using an upper and lower bound, *alpha* and *beta*.

In the parallel implementation that we have used, the work is divided in a way similar to the TSP approach: Each process searches a part of the search tree. In contrast to TSP, where reaching a leaf-node and finding a shortest route is enough to improve the solution, in alpha-beta search the value of a node can only be changed by combining the values of its children. To be able to combine the values

of children that are evaluated by distinct processes, these values are stored in shared memory: the top part of the tree down to, and including, the nodes that are evaluated by worker-processes, is built explicitly in shared memory.

The 'leaf' nodes of this explicit tree are evaluated by worker-processes using sequential alpha-beta search without building the sub-trees explicitly. When one of the workers has evaluated its board-position (sub-tree), it will propagate this new value in the explicitly constructed tree, so that the next worker to start evaluating a new position will have better alpha and beta bounds.

Both implementations use a fanout of 38 and a search-depth of 6. Only the top two levels of the search tree are built explicitly in shared memory, so that there are 38 subtrees to search.

The measurements were obtained using three different static evaluation functions. We supply the mean values of the measurements for these three functions. The speed-ups for alpha-beta, which are depicted in Figure 8, are clearly less than those for the travelling-salesman problem. This is because, in contrast to the TSP algorithm, a parallel implementation of alpha-beta always searches more nodes than the sequential implementation. This is caused by the fact that at the start of the search, there are no good alpha and beta values available to prune the search tree. In sequential alpha-beta, only the first subtree suffers from this weak alpha and beta value. The next subtree can be pruned with a much stronger alpha and beta value (produced by the first subtree), so it can be pruned earlier. In the parallel implementation, each process evaluating a subtree starts with the weak alpha and beta value. On top of that, updating the top of the tree is only useful for jobs that still have to be started. Processes that are already working on a subtree cannot use these new values of their top-node to reduce the search space. Thus the more processes are used, the more nodes are searched relative to sequential alpha-beta.

Matrix multiplication

We parallelised matrix multiplication by dividing the rows of the matrix among the available processes. When the two input matrices are known to each process, each process can compute the results for its rows independently. In the Orca program, each process generates the same (pseudo random) input matrices locally: they are not shared. The output is produced by printing the result rows in turn. Therefore the result is not shared either. In the SVM-system program, we used the same approach. Thus both systems only use shared memory for synchronization and work division, not for sharing of input and output.

The algorithm has a constant factor: the generation of the matrices. The speed-ups the number of processes increases, lag more and more behind. The measurements were obtained by multiplying two 250×250 integer matrices. The speed-ups are shown in figure 9.

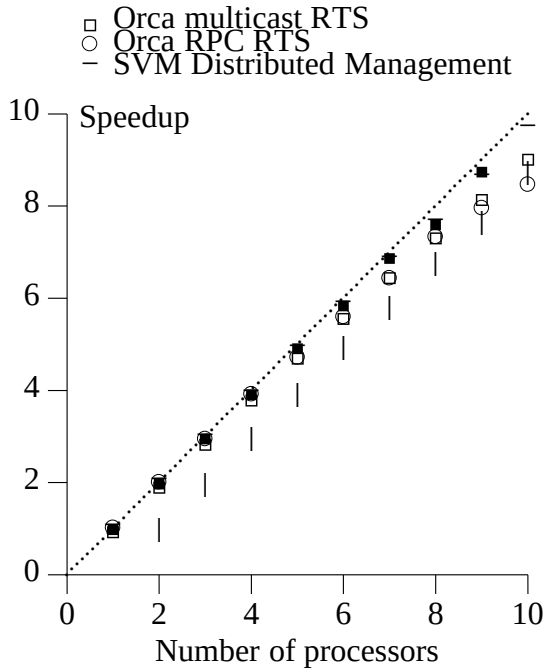

Fig 9: Matrix multiplication

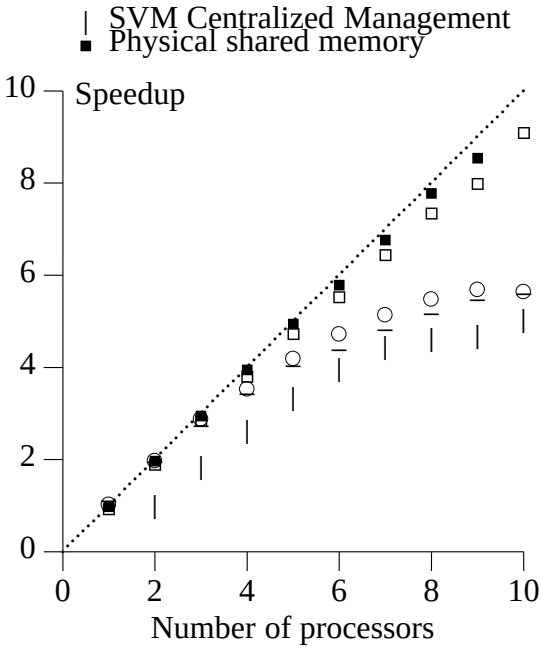

Fig 10: All pairs shortests paths

All pairs shortest paths problem

The all-pairs shortest-paths (ASP) algorithm computes the shortest distance from every node to every other node in a graph. A sequential algorithm is due to Floyd.[26] For an $N \times N$ matrix representing the graph, it uses N iterations to compute the distance matrix. In iteration k, for every node combination $1 \le i,j \le N$:

$$\delta^k(i,j) = \text{Min } \{ \delta^{k-1}(i,j), \delta^{k-1}(i,k) + \delta^{k-1}(k,j) \}$$

For all k, $\delta^k(i,j)$ is the length of the shortest path between node i and node j that passes through node 1,...,k only. Matrix δ^N contains the desired result.

This algorithm can also be parallelised by dividing the rows to compute among the processes. But for a process to compute a row in iteration k, it needs $\delta^{k-1}(k,j)$ $1 \le j \le N$, that is row k of iteration k–1. These rows are shared and their availability synchronizes the processes. Speedups for a 200×200 matrix are shown in figure 10.

The computation of iteration k can only proceed when row k of iteration k–1 is available. During the computations, all processes are waiting for the same data. If these data, when they become available, are distributed over the processes one by one, each row becomes a sequential bottleneck. So, the only system (besides physical shared memory) that performs well is the multicast shared data-object system. In this system, the new value is broadcast to all processes in one message.

6. DISCUSSION

Both the structured and the unstructured systems achieve good speed-ups for most applications. Overall, the speed-ups for Orca programs are better. This can be attributed to three facts.

- In Orca programs, the granularity of the shared data is inherently tailored to the application. A message only has to be as large as the data it contains. The Run Time System can even choose between sending the new value or sending the operation, depending on the costs. In the SVM system, shared data are always transferred in chunks of 1 page. Pages of 4K, as we use, are mostly too big: just a small portion of the page is of interest, the remainder wastes time and bandwidth. Although this does not occur in our applications, 4K could be too small and multiple requests have to be sent to get the whole data-structure. What does occasionally happen is that a data-structure (e.g. a row of a matrix) spans a page boundary. To access it, two pages have to be referenced, instead of one.

  This effect of wasted time and bandwidth is reflected in the results for the travelling-salesman problem (TSP) and alpha-beta search. Getting a new job requires a small message in the shared data-object systems, while in the SVM systems a message of 4K must be sent. In the TSP problem, updating the minimum leads to the same situation.

- The structured paradigm uses updating whereas the unstructured paradigm uses invalidation to keep the address space coherent. When a process changes some shared data, and these data will be referenced by most other processes before further changes, the update approach performs better. This is the case in all applications except alpha-beta.

- Because of the overhead of the Orca Run Time System, which still is a prototype, sequential Orca programs (except alpha-beta) are 3 to 4 times slower than the sequential C programs. Therefore, the communication channel is relatively 3 to 4 times faster for the Orca programs than for the C programs; it will not become a bottleneck as quickly as in the unstructured systems. When a better Orca compiler is available, we shall be able to make more realistic measurements.

Multicast RTS versus RPC RTS

For the most communication intensive application, ASP, the broadcast protocol performs very well. As explained earlier, the distribution of the newly generated data is no sequential bottleneck. The positive effect of broadcasting a new value versus updating through a 2-phase primary copy protocol can also be seen in the TSP measurements. Because the shared variable *minimum* is read very frequently by all processes, they all have a local copy. Hence the RPC RTS has no advantage of partial replication. A multicast update is more efficient in this situation.

In the alpha-beta algorithm the shared top-part of the tree is only referenced when a job is

completed. Before the next reference, this shared variable is updated a few times by other processes. The RPC RTS therefore does not replicate it at every processor, but just maintains a primary copy at one processor. In contrast to the full replication multicast RTS, where updating the tree involves all processors, in the RPC RTS only two processors are involved. This effect can be seen in the results of alpha-beta which is the only application where the RPC RTS is equally fast as the multicast RTS. Which approach performs better depends on the access patterns a specific application exhibits.

Distributed versus Centralized management

For all applications the speed-ups in the Central Manager system keep in step with the speed-ups in the Distributed Management system. Clearly, for 10 processors, the Central Manager is not a bottleneck. The measurements for TSP even show better speed-ups for the Central Manager system than for the Distributed Management system. This can be explained as follows: When the shared variable minimum is updated, all copies are invalidated. In the Central Manager system, every process contacts the Central Manager to get a read-copy again. For the first of these requests, the Central Manager has to contact the new owner with an RPC. But all other requests can be handled without an RPC to the owner, because the Central Manager has a valid copy available. Therefore these requests need one RPC only.

In the Distributed Management system, each process broadcasts a request and waits for the page to arrive, so a broadcast and an RPC are needed. Furthermore, the owner process, which could otherwise work on the application, is interrupted to send the page. In this way not only the requester, but also the current owner of the page, lose computation time. Lastly, each broadcast is received by all processors, which means that each processor is interrupted at every broadcast, even though a given processor can ignore most broadcasts. The main properties that affect performance are summarized in Figure 11.

	Centralized Management	**Distributed Management**	**RPC RTS**	**Multicast RTS**
Unit of sharing	Page	Page	Data-object	Data-object
Replication	Partial	Partial	Partial	Full
Invalidation	YES, one by one	Yes broadcast	NO	NO
Updating	On demand, one by one	On demand, one by one	Always, one by one	Always, multicast

Fig 11: The four most important properties that affect performance

Good versus bad placement in the SVM systems

The difference in performance between SVM programs where the shared variables are cleverly distributed over the pages, and programs where they are not, can only be seen in the TSP program. The other applications do not suffer from poor placement for the following reasons. Matrix multiplication does not do much communication at all. The ASP program communicates through a shared matrix

occupying 40 pages, where only the first 5 rows of the matrix are positioned on a page containing other shared data. In alpha-beta, once a process gets a job, it will not reference shared memory further until it has finished that job completely. References to shared memory are therefore infrequent, and pages being paged out do not delay the computation.

Only with the TSP problem do the speed-ups lag when the data are poorly distributed. This is explained by the fact that every time a process takes a new job, a counter in shared memory must be updated and the shared-variable minimum, which lies on the same page, is inaccessible for all other processes. These processes cannot continue their computation until the page is paged in again. The effect on the speed-ups can be seen in Figure 12.

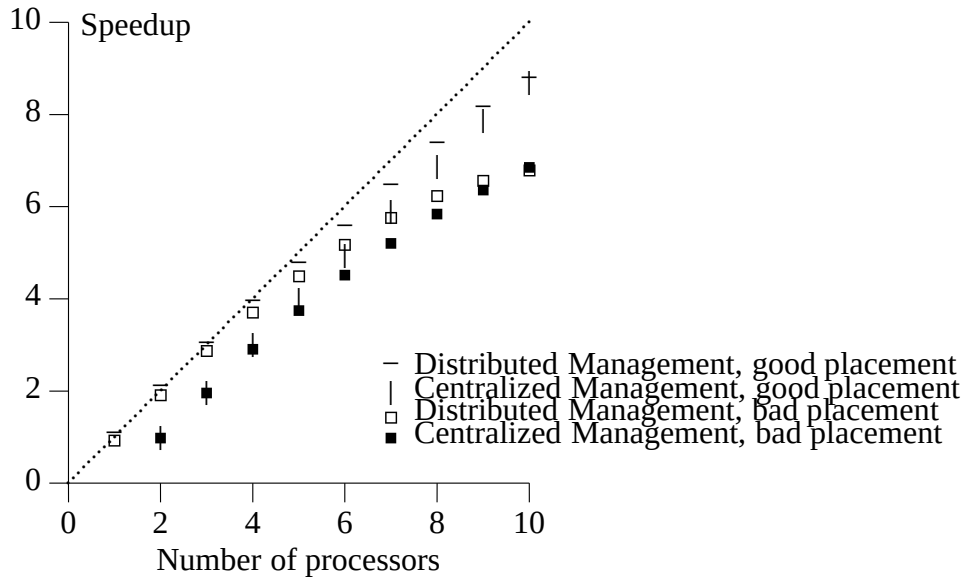

Fig 12: Travelling Salesman Problem with good and bad placement

Programmability aspects

In general, the programming of the structured DSM model is much easier, for the following reasons:

- Orca provides type security. This means that incorrect use of variables is detected, either by the compiler or by the RTS. For instance: assigning an integer value to a boolean variable, or supplying an out of range index to an array. Because parallel programs mostly are non-deterministic they are difficult to debug. The extra support type security provides is very useful. The type security offered by Orca is very helpful with debugging.

- The implicit mutual exclusion synchronization, supplied with operations on shared data-objects, makes the model easy to understand and less burdensome to program than the SVM system, where most accesses to shared memory must be protected by locks explicitly. Another

advantage of implicit mutual exclusion synchronization is that programs are more compact and easier to read.

- In the SVM system, the programmer can only use the shared memory efficiently when he or she understands (at least part) of the working of the system. The placement of variables is especially important. Another consideration is what lock to use: a read or a write-lock, or possibly no lock at all. Because the shared data-object model provides a structured approach to DSM, the programmer is shielded from the low-level functionality of the system. The system itself takes care of efficient replication of objects.

Scalability

To achieve the goal of high performance at low cost one needs to know if the applications will scale to a larger number of processors than 10. There are a number of factors that influence the scalability of both DSM models: the access patterns of the applications to DSM, the granularity of the operations, and the size of the messages that are sent to keep DSM consistent. Although we can not make any hard statements about these factors, we shall discuss the influence on scalability of each of them. Our assumption is that if network traffic can be avoided, applications will scale well.

The access patterns that an application exhibits determine how well the underlying DSM model will scale. Applications that perform a few write operations but perform many read operations will scale to a large number of processors, because they will not generate any network traffic due to the replication schemes that DSM systems use. This can already be seen at 10 CPUs. For TSP, the physical shared memory system performs much worse than any of the DSM implementations, because it does not have any hardware scheme to replicate shared memory.

The second important factor is the granularity of the operations. If operations of the DSM are low-level operations, such as "move a word from a register to main memory", these operations can lead to a large amount of network traffic. Consider the case that a shared record is updated continuously by two processors. In an unstructured DSM implementation this will generate many messages, if the programmer does not lock the record explicitly. It might happen that for each write on a word of the record, a message has to be sent. In a structured DSM implementation the operation on the record is packed in one user-defined procedure call. In this case, only one message has to be sent.

The third important factor is the size of the messages that are sent. In an unstructured DSM, each time a write operation on a page that is not located on the local processor has to be performed, the DSM system has to get the complete page across the network, even if the write required only four bytes of the page to be changed. Furthermore, if such a 4 byte object crosses page boundaries, two complete pages have to be sent over. In a structured DSM, the RTS only sends over the operation code or moves the object. In both cases, typically only a limited number of bytes has to be transferred.

Both unstructured and structured DSM rely on replication schemes and dynamic placing of

shared memory to scale to a large number of processors. However, the unstructured DSM will in general need more messages of a larger size than a structured DSM to keep shared memory consistent. In general, one might expect that structured DSM would scale better than unstructured DSM. Ultimately, however, scaling will be determined by the application.

7. CONCLUSIONS

In this paper we have compared two Distributed Shared Memory paradigms, a structured and an unstructured one. This comparison was mainly done on the basis of measured execution times for four parallel applications. The unstructured Shared Virtual Memory paradigm showed better absolute execution times, whereas the structured shared data-object paradigm showed better speed-ups. A more efficient implementation for the Orca RTS than the current prototype could make this difference in absolute times smaller. It remains to be seen if the speed-ups for such a faster RTS will stay at their current high level. However, it is our expectation that they will be better than the speed-ups for the SVM system, because of the structured application-dependent division of shared memory.

In terms of programmability and readability, programming applications for the shared data-object model is clearly superior to programming applications for the Shared Virtual Memory model; the clean semantics of the shared data make it easy to write and debug programs. In addition, the programs execute efficiently, even for users who do not understand the underlying system.

An advantage of the SVM system is that every programming language supporting global data could be extended with a few system calls to make use of distributed shared memory. Implementing the shared data-object model within an existing language is less easy.

In conclusion, both paradigms are useful for distributed programming and can be implemented efficiently. Although the SVM paradigm currently has the faster implementation, the shared data-object paradigm is easier to understand, to program and to debug. On top of that, it is more promising in terms of performance, as soon as a better Orca compiler is available.

REFERENCES

1. K. Li, ‘‘Shared Virtual Memory on Loosely Coupled Multiprocessors,’’ Research Report 492 (Ph.D. dissertation), Yale University, New Haven, CT (Sept. 1986).
2. K. Li and P. Hudak, ‘‘Memory Coherence in Shared Virtual Memory Systems,’’ *ACM Trans. Comp. Syst.* **7**(4), pp. 321-359 (Nov. 1989).
3. H.E. Bal, M.F. Kaashoek, and A.S. Tanenbaum, ‘‘Orca: A Language for Parallel Programming of Distributed Systems,’’ *IEEE Transactions on Software Engineering* **18**(3), pp. 190-205 (March 1992).

4. K. Li, ‘‘IVY: a Shared Virtual Memory System for Parallel Computing,’’ *Proc. 1988 International Conference on Parallel Processing*, St. Charles, IL **2**, pp. 94-101 (Aug. 1988).

5. K. Li and R. Schaefer, ‘‘A Hypercube Shared Virtual Memory System,’’ *Proc. 1989 International Conference on Parallel Processing*, St. Charles, IL, pp. 125-132 (Aug. 1989).

6. B.D. Fleisch and G.J. Popek, ‘‘Mirage: a Coherent Distributed Shared Memory Design,’’ *Proc. Twelfth Symposium on Operating Systems Principles*, Litchfield Park, AZ, pp. 211-223 (Dec. 1989).

7. R.G. Minnich and D.J. Farber, ‘‘The Mether System: Distributed Shared Memory for SunOS 4.0,’’ *USENIX Summer’89*, pp. 51-60 (1989).

8. R.G. Minnich and D.J. Farber, ‘‘Reducing Host Load, Network Contention, and Latency in a Distributed Shared Memory System,’’ *Proc. Tenth International Conference on Distributed Computing Systems*, Paris, pp. 468-475 (May 1990).

9. J.K. Bennett, J.B. Carter, and W. Zwaenepoel, ‘‘Munin: Distributed Shared Memory Based on Type-Specific Memory Coherence,’’ *Proc. Second Symposium on Principles and Practice of Parallel Programming*, Seattle, WA, pp. 168-176 (March 1990).

10. R. Bisiani and A. Forin, ‘‘Architectural Support for Multilanguage Parallel Programming on Heterogeneous Systems,’’ *Proc. Second International Conference on Architectural Support for Programming Languages and Operating Systems*, Palo Alto, CA, pp. 21-30 (Oct. 1987).

11. S. Ahuja, N.J. Carriero, and D.H. Gelernter, ‘‘Linda and Friends,’’ *IEEE Computer* **19**(8), pp. 26-34 (Aug. 1986).

12. D.H. Gelernter, ‘‘Generative Communication in Linda,’’ *ACM Trans. Prog. Lang. Syst.* **7**(1), pp. 80-112 (Jan. 1985).

13. S.E. Lucco, ‘‘Parallel Programming in a Virtual Object Space,’’ *SIGPLAN Notices (Proc. Object-Oriented Programming Systems, Languages and Apllications 1987)*, Orlando, FL **22**(12), pp. 26-34 (Dec. 87).

14. H.E. Bal and A.S. Tanenbaum, ‘‘Distributed Programming with Shared Data,’’ *Computer Languages* **16**(2), pp. 129-146 (1991).

15. A.S. Tanenbaum, R. van Renesse, H. van Staveren, G. Sharp, S.J. Mullender, A. Jansen, and G. van Rossum, ‘‘Experiences with the Amoeba Distributed Operating System,’’ *Commun. ACM* **33**(12), pp. 46-63 (Dec. 1990).

16. M.F. Kaashoek, A.S. Tanenbaum, S. Flynn Hummel, and H.E. Bal, ‘‘An Efficient Reliable Broadcast Protocol,’’ *Operating Systems Review* **23**(4), pp. 5-20 (Oct. 1989).

17. M.F. Kaashoek and A.S. Tanenbaum, ‘‘Group Communication in the Amoeba Distributed Operating System,’’ *11th Int’l Conf. on Distributed Computing Systems*, Arlington, Texas, pp. 222-230 (20-24 May 1991).

18. H.E. Bal, *Programming Distributed Systems,* Silicon Press, Summit, NJ (1990).

19. H.E. Bal, M.F. Kaashoek, and A.S. Tanenbaum, ‘‘Experience with Distributed Programming in Orca,’’ *Proc. IEEE CS 1990 Int. Conf. on Computer Languages*, New Orleans, LA, pp. 79-89 (March 1990).

20. A.S. Tanenbaum, M.F. Kaashoek, and H.E. Bal, ‘‘Parallel Programming using Shared Objects and Broadcasting,’’ *IEEE Computer* (Aug. 1992).

21. K.P. Eswaran, J.N. Gray, R.A. Lorie, and I.L. Traiger, ‘‘The Notion of Consistency and Predicate Locks in a Database System,’’ *Commun. ACM* **19**(11), pp. 624-633 (Nov. 1976).

22. H.E. Bal, M.F. Kaashoek, A.S. Tanenbaum, and J. Jansen, ‘‘Replication Techniques for Speeding up Parallel Applications on Distributed S ystems,’’ *Concurrency Practice & Experience* **4**(5) (Aug. 1992).

23. B. Liskov and L. Shrira, ‘‘Promises: Linguistic Support for Efficient Asynchronous Procedure Calls in Distributed Systems,’’ *Proc. SIGPLAN 88 Conf. on Progr. Lang. Design and Impl.*, Atlanta, GA, pp. 260-267 (June 1988).

24. H.E. Bal, R. van Renesse, and A.S. Tanenbaum, ‘‘Implementing Distributed Algorithms Using Remote Procedure Calls,’’ *Proc. AFIPS Nat. Computer Conf.*, Chicago, IL **56**, pp. 499-506, AFIPS Press (June 1987).

25. D.E. Knuth and R.W. Moore, ‘‘An Analysis of Alpha-Beta Pruning,’’ *Artificial Intelligence* **6**, pp. 293-326 (1975).

26. R.W. Floyd, ‘‘Algorithm 97: Shortest Path,’’ *Commun. ACM* **5**, p. 345 (1962).

APPENDIX A

This appendix lists the execution times for our four applications. The tables contain the mean values over a number of runs. All times are in seconds. The measurements for the SVM Central Manager system start at two processors because there is always one processor occupied by the Central Manager. We have no measurements with 10 processors for the physical shared memory system because we only had 9 working processors.

Travelling-salesman problem (execution times in seconds)										
Nr. processors	1	2	3	4	5	6	7	8	9	10
SVM, Centralized, bad placement		166.0	84.0	56.7	44.1	36.6	31.7	28.3	26.0	24.1
SVM, Distributed, bad placement	166.3	83.4	56.4	44.0	36.4	31.7	28.5	26.4	25.0	24.2
SVM, Centralized, good placement		159.6	79.8	52.7	39.8	32.2	27.0	23.2	20.2	18.3
SVM, Distributed, good placement	159.9	78.7	53.8	41.1	34.0	29.0	25.0	21.9	19.8	18.3
Orca Multicast	587.7	291.7	192.6	144.7	116.6	97.3	83.7	73.2	65.8	59.1
Orca RPC	660.8	327.7	217.1	164.2	132.3	111.3	95.6	84.4	76.1	69.0
Shared Memory	169.6	84.6	55.8	42.2	33.9	28.2	24.2	21.2	18.9	
Amoeba RPC	145.6	75.6	52.2	41.5	35.1	31.0	28.4	26.8	24.5	

Alpha-beta search (execution times in seconds)										
Nr. processors	1	2	3	4	5	6	7	8	9	10
SVM, Centralized, bad placement		1159.4	634.9	457.1	385.4	326.4	299.0	289.2	273.7	256.9
SVM, Distributed, bad placement	1160.5	635.4	458.0	386.8	327.6	300.0	290.3	274.6	258.0	243.3
SVM, Centralized, good placement		1159.2	634.7	457.0	385.7	326.2	299.0	289.1	273.8	256.7
SVM, Distributed, good placement	1160.5	635.1	458.1	386.8	327.5	300.2	290.4	274.6	258.2	243.5
Orca Multicast	1801.5	972.1	713.4	547.5	470.8	418.5	371.0	336.0	311.6	296.4
Orca RPC	1816.1	1063.7	752.6	551.8	474.5	421.8	373.9	338.7	314.1	296.6
Shared Memory	970.6	530.7	382.2	322.4	272.9	249.9	241.7	228.6	214.8	
Amoeba RPC	1813.4	978.5	718.3	552.2	475.0	424.5	377.1	342.3	319.2	

Matrix multiplication (execution times in seconds)										
Nr. processors	1	2	3	4	5	6	7	8	9	10
SVM, Centralized, bad placement		170.8	85.5	57.5	43.6	34.5	29.2	25.2	22.3	20.0
SVM, Distributed, bad placement	171.0	86.0	57.7	43.5	34.7	29.1	25.2	22.1	19.6	17.8
SVM, Centralized, good placement		170.8	85.9	57.6	43.5	34.5	29.3	25.2	22.3	19.5
SVM, Distributed, good placement	171.0	85.8	58.0	43.7	35.0	29.3	25.1	22.4	19.9	17.7
Orca Multicast	810.3	410.6	279.1	209.8	169.9	143.8	124.2	109.8	98.7	89.1
Orca RPC	780.1	392.9	265.8	200.0	166.0	139.8	121.6	106.5	98.2	92.3
Shared Memory	180.5	90.5	60.8	45.9	36.6	30.8	26.2	23.7	20.6	

All pairs shortest paths problem (execution times in seconds)										
Nr. processors	1	2	3	4	5	6	7	8	9	10
SVM, Centralized, bad placement		68.8	37.1	26.0	20.8	17.2	16.1	14.6	14.7	13.0
SVM, Distributed, bad placement	68.6	37.0	26.0	20.6	17.6	15.6	15.0	13.5	12.6	12.7
SVM, Centralized, good placement		68.6	37.1	26.1	20.4	17.3	15.4	14.8	14.6	13.6
SVM, Distributed, good placement	68.7	37.0	26.0	20.7	17.5	16.0	14.6	13.6	12.8	12.5
Orca Multicast	432.1	218.9	148.0	111.4	90.0	77.1	66.3	58.2	53.6	47.1
Orca RPC	400.0	204.7	140.0	114.1	95.9	85.1	78.2	73.2	70.5	71.2
Shared Memory	68.6	34.4	23.1	17.3	13.8	11.8	10.1	8.8	8.0	