Cluster Computing 6, 227–236, 2003

Starfish: Fault-Tolerant Dynamic MPI Programs on Clusters of Workstations *

ADNAN AGBARIA and ROY FRIEDMAN **
Department of Computer Science, The Technion, Haifa 32000, Israel

Abstract. This paper reports on the architecture and design of *Starfish*, an environment for executing dynamic (and static) MPI-2 programs on a cluster of workstations. Starfish is unique in being efficient, fault-tolerant, highly available, and dynamic as a system internally, and in supporting fault-tolerance and dynamicity for its application programs as well. Starfish achieves these goals by combining group communication technology with checkpoint/restart, and uses a novel architecture that is both flexible and portable and keeps group communication outside the critical data path, for maximum performance.

Keywords: checkpoint/restart, fault-tolerance, distributed system, high performance, MPI

1. Introduction

Employing clusters of workstations as a cost effective alternative to parallel computers has been the goal of much research in the past few years [5,11], especially in light of the remarkable advances in both computing power of PCs and networks speed. While the idea of building such clusters is very appealing, the realization of this goal is quite complicated, due to the numerous non-trivial issues that have do be dealt with. These include maintaining the promised performance at the application level, manageability of the cluster, and fault-tolerance. Other issues like job/process scheduling must also be considered, although these are well studied problems, and the usage of clusters vs. parallel computers does not significantly alter known solutions from the parallel world.

One major obstacle in the way of building such clusters, which has been successfully dealt with recently, lies in the fact that legacy operating systems present high overhead in accessing the network. Thus, naive usage of fast networks fails to deliver on the promised network speed. U-Net [7], NOW [5], Fast-Messages [30], VI Architecture [40] and other projects have been able to overcome this problem by developing user-level network interfaces, that bypass the operating system kernel, offering applications near optimal bandwidth and latency.

Thus, largely speaking, obtaining low network latency and high bandwidth can be considered a solved problem (at least from the research point of view). However, the problems of achieving cluster manageability, high-availability and checkpoint/restart (hereafter, C/R) without hurting the performance of applications under normal conditions are still far from resolved. In particular, the distributed nature of these clusters and the relatively high probability of partial failures that is inherent in distributed environments make these problems hard.

In this work we describe Starfish, a system that tries to tackle these issues using a novel architecture that combines group communication technology and C/R. Starfish as a system is manageable, highly available, and adjusts to dynamic changes in the cluster. For its applications, Starfish provides hooks to handle dynamic cluster changes, as well as C/R facilities. Starfish was initially implemented on Linux. Recently, we added support for other operating systems, and in particular for heterogeneous clusters and heterogeneous checkpointing at the virtual machine level, based on our work reported in [2]. For data communication, Starfish supports both Myrinet [27] (for performance) and IP (for convenience). Porting to other fast networks like ServerNet^{TM} [37] is planned, and only requires writing a thin layer of code, as described later in the paper. The bulk of the code is written in OCaml [39]; this part is fully portable to most variants of Unix, as well as Windows NT. The only parts we had to rewrite is the immediate interface to Myrinet. For native checkpointing on operating systems other than Linux, one needs to rewrite the specific code that dumps a process to disk and then restores it from disk. However, virtual machine level checkpointing already works across platforms.

Each Starfish node runs a *Starfish daemon* (or simply a *daemon*), as illustrated in figure 1. All Starfish daemons form a *process group* [8], using the Ensemble group communication toolkit [20,38]. As described later in this paper, these daemons are used to interact with clients, spawn MPI programs to which we refer to as *application processes*, track and recover from failures, and to maintain the configuration of the system. In particular, daemons utilize a lightweight group mechanism ala [35], in a manner similar to the group daemon proposed in [9], for keeping track of and reporting partial application and system failures.

Each application process is composed of 5 major components, as illustrated in figure 1, and as reported in section 2. These include, a *group handler*, which is responsible for communicating with the daemon, an *application part*, which in-

* A shorter preliminary version of this work appeared in HPDC '99.

** Corresponding author.
E-mail: roy@cs.technion.ac.il

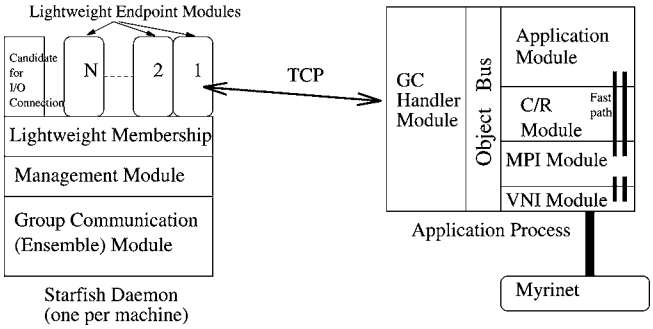

Figure 1. Starfish architecture.

cludes the user supplied MPI code, a *checkpoint/restart module*, an *MPI module*, and a *virtual network interface* (or VNI for short). These modules communicate internally using an object bus based listener model. Also, the application part has a separate *fast data path* to and from the MPI module, which guarantees low latency and minimal impact on performance.

This architecture is also very flexible and portable, in the sense that it allows us to implement several different distributed C/R protocols, both coordinated and uncoordinated checkpointing [14,32], and to run them side by side. In particular, we can run the same application with two different C/R protocols, and compare them. Moreover, Starfish supports both homogeneous and heterogeneous C/R in each process. Thus, with heterogeneous checkpointing Starfish becomes a heterogeneous system where its nodes can consist of different architectures and operating systems. Also, we can easily port Starfish to various networks, since all that is required is to provide a relatively thin interface layer inside our VNI.

Another interesting feature of Starfish relates to its API. The additional functionality of Starfish is supported through additional downcalls and upcalls beyond the standard MPI. For each of the upcalls, there is some default handling procedure, and hence applications that do not wish to handle these upcalls can simply ignore them. Similarly, applications are not required to issue any downcalls. This allows Starfish to run regular MPI programs, without any modifications. Naturally, such programs will only enjoy part of Starfish capability, e.g., system initiated checkpointing, but not all the potential benefits of the system, e.g., user initiated checkpointing and dynamic reconfiguration. Conversely, programs that use the additional API calls can be automatically, or semi-automatically, transformed back into standard MPI programs by eliminating all Starfish specific downcalls. Since these calls only deal with checkpointing and reconfiguration, such programs will then run correctly on any standard MPI implementation.

The rest of this paper is organized as follows: The general architecture underlying Starfish is presented in section 2. Section 3 elaborates on the fault-tolerance and high-availability aspects of Starfish. Section 4 describes the support in Starfish for heterogeneous checkpointing. An initial performance evaluation of the current release is presented in section 5. We compare our work to related work in section 6, and conclude with a discussion in section 7.

2. Starfish architecture

As discussed in the introduction, and as illustrated in figure 1, each Starfish node runs a daemon, where all Starfish daemons are members of the same process group, called the *Starfish group*. This group is managed by the Ensemble group communication system [20,38]. The collection of these daemons form Starfish' *parallel environment*, and they are responsible for spawning application processes, keeping track of applications health, managing the configuration and settings of the cluster, communicating with clients, and for providing the hooks necessary to provide fault tolerance for applications. In particular, all configuration and control messages, including those related to applications (but not data messages) are sent by the daemons using Ensemble.[1] The internal structure of daemons is described in detail in subsection 2.1.

Given the parallel nature of MPI programs, each application is expected to be divided into several concurrent processes, each potentially running on a separate node. As described shortly, each application process consists of a Starfish run-time environment and user supplied code. Starfish run-time library is responsible for interacting with the daemons, for checkpoint and restart, and for implementing MPI, whereas the user code is any given MPI C and OCaml program. Subsection 2.2 elaborates on the internal structure of the application process.

Finally, the user process and the daemon communicate with each other through a local TCP connection. The description of this protocol appears in subsection 2.3.

2.1. Daemon internals

Starfish daemons need to maintain some application specific data for each application processes running on the same machine, as well as some shared state that defines the current cluster configuration and settings. We maintain this data by employing the design illustrated in figure 1. Daemons are composed of four main modules: *Ensemble* (the group communication system we use), the *management module*, the *lightweight membership module*, and several instances of the *lightweight endpoint module* [19]. Daemons may exchange *control messages* among themselves; these messages may be generated by either module, and are sent through Ensemble, but are not passed to application processes (see table 1).

In our design, a single management module is used to maintain the configuration and shared settings of the cluster. Each application that runs on Starfish is associated with a lightweight process group, whose members are the daemons on the nodes that run the corresponding application processes. For each application process we instantiate a lightweight endpoint module. This module is the one that is responsible for the connection with the application process, and for passing messages and events to and from that process. The lightweight membership module is responsible for deducing the lightweight group membership from the entire Starfish process group, and for translating membership and message events between the main Starfish group and each of the lightweight groups (see also figure 2).

Table 1
Summary of message types in Starfish.

Message type	Sent between
Control	Starfish daemons
Coordination	Application processes through daemons
Data	Application processes through MPI and VNI modules using fast path
Lightweight membership	Lightweight endpoint module and application processes
Configuration	Local daemon and application processes
Checkpoint/restart	Checkpoint/restart modules through daemons

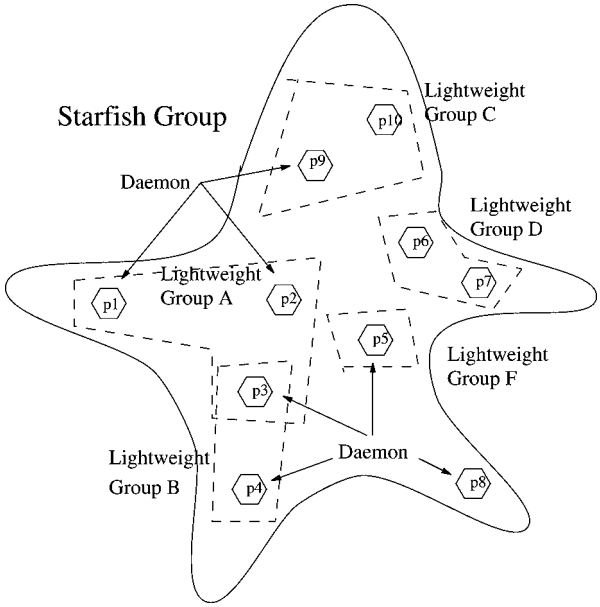

Figure 2. Lightweight groups within a Starfish group. Here, all daemons are members of the same Starfish group. Additionally, daemons p1, p2, and p3 share the same lightweight group, indicating that there is an application that spans all three machines. Similarly, p3 and p4 share a lightweight group, etc. Note that p8 does not appear in any lightweight group, indicating that no application process is currently running on the corresponding machine.

Note that typically the lightweight groups are only subsets of the main Starfish process group. This is because on a large cluster, each application spans less than the entire cluster. Thus, not every change in the Starfish group needs to propagate to every lightweight group. Similarly, if an application process on one node terminates, but the machine itself continues to run, then this should result only in a membership change of the corresponding lightweight group, and need not be reported to all members of the main Starfish group, or in other lightweight groups. Also, messages that are sent inside a lightweight group should only be delivered to members of the same lightweight group.

Of course, it would have been possible to allocate a separate full blown process group for each application. But as indicated in [19], the lightweight group approach is more efficient. Also, the structure we described allows us to maintain both consistent cluster wide information, which is independent from a specific application, and manage multiple applications on top of this cluster, mimicking a parallel computer.

2.2. Application process internals

Each application process consists of the following components, as illustrated in figure 1: a group handler, an application module, a checkpoint/restart module, an MPI module, and a virtual network interface, hereafter called VNI. All modules communicate by posting events on an object bus that invokes the corresponding event handlers at each of the listening module. Using an object bus allows us to completely decouple the modules, and also to potentially post the same events to more than one module. Finally, in order to orchestrate these modules, we have implemented our own scheduler.

Application processes are involved in five types of messages: *data messages*, *coordination messages*, *C/R messages*, *lightweight membership messages*, and *configuration messages*[2] (see table 1). There is a significant difference between data messages and other types of messages: Data messages result from the user supplied MPI code, and have strict performance requirements, while other messages are generated by Starfish itself and do not require the same responsiveness as data messages. Thus, we employ a fast data path between the MPI implementation and the application module, that does not go through the object bus. This ensures the required low latency for data messages, while still being able to provide manageability and strong guarantees for the other types of messages.

To send other types of messages, the generating module post an event for the group handler module and the group handler translates this event to a message on the TCP connection with the daemon. In the case of coordination messages and C/R messages the daemon broadcast them in the relevant lightweight group using Ensemble. In the other direction, messages received by the group handler module are translated to events on the object bus, to be invoked at the corresponding modules in the application process. Using the daemon, and therefore Ensemble, for disseminating coordination messages greatly simplifies our code, and enjoys the strong delivery guarantees of Ensemble, which also simplify our protocols.

Lightweight membership messages and configuration messages are part of the protocol executed between the application processes and the daemons, and are discussed subsection 2.3. Coordination messages are sent among application processes (potentially) located on different nodes for general coordination tasks, while C/R messages are used by the various C/R protocols. The set of C/R messages seems to be rich enough to express all C/R protocols we have encountered. Since lightweight membership messages and C/R messages are exchanged by application processes, and daemons only serve as a reliable middle communication layer, these messages are opaque to daemons.

2.2.1. Optimizing receives

As in the MPI standard, Starfish supports both *blocking* and *non-blocking* send and receive operations. In blocking mode, a send or receive operation does not return until the message has been fully sent or received, while in non-blocking, operations return immediately. The eager way of implementing sends is to immediately send a message to its destination [18]. This, however, requires that the destination be prepared to read a message from the network, even if the application process there has not yet performed a matching receive operation.

In Starfish we overcome this problem by introducing a low priority thread, called the *polling thread*. This thread continuously polls the network, so whenever a message arrives, the polling thread receives the message and puts it in a queue of received messages, for further handling by the application at a later time.

A nice feature of the polling thread is that it eliminates much of the runtime overhead of issuing a receive operation at the application level. A receive operation, particularly in blocking mode, pauses the process execution in order to receive an in-transit message. Moreover, when using the regular TCP/IP stack, receiving a message from the network involves a system call and user-level/kernel interaction, which is costly. When using the polling thread, the time required for kernel interaction is interleaved with other operations, yielding fast receive operations.

2.3. Interaction between daemon and application process

The interaction between an application process and its corresponding daemon is performed via the TCP connection between the GC handler module of the process and its corresponding endpoint module in the daemon. Figure 1 illustrates such interaction between these modules.

As mentioned in subsection 2.2, two types of messages are designated for the interaction between daemons and application processes. These are configuration messages and lightweight membership messages. Lightweight membership messages inform application processes about new views, and are used by an application process to terminate its membership in a lightweight group. Configuration messages, on the other hand, are used to inform the application process of various configuration parameters, and used to synchronize between the application process and the daemon upon initialization and termination of the application process.

3. Manageability, dynamicity, fault tolerance, and high availability

In this section we elaborate on how we achieve cluster manageability and high-availability, how we support dynamic changes in the environment at both the clusters and the application, and provide fault-tolerance for the application. We split the discussion into general cluster aspects, which are reported in subsection 3.1, and to application aspects, which are reported in subsection 3.2.

3.1. Starfish manageability, dynamicity, and high availability

3.1.1. Manageability

Starfish can be managed from any computer connected to the LAN on which the cluster runs, either directly, or through the Internet. Managing the cluster is done by opening a TCP connection to one of the daemons, on which an ASCII based protocol is used. Through this connection, the cluster administrator can add or remove nodes from the cluster, disable and (re)enable nodes, and control the parameters of the cluster. The management protocol starts with a login session, in which the client side has to authenticate itself as an administrator to the cluster, and identify the connection as a *management connection*.

The management module of Starfish handles management connections, and takes care of forwarding configuration commands to all other daemons in the system. The use of ensemble's reliable and totally ordered delivered mechanism is instrumental here, in maintaining coherent state between all cluster daemons [8].

A similar protocol that also employs a TCP connection is used between clients and any of the cluster nodes in order to submit applications for either interactive or batch execution. This protocol also begins with a login session, but is identified as a *user session*, and is thus limited to submitting, suspending, resuming, and deleting applications. (A user can only suspend, resume, and delete its own applications.)

Note that user commands regarding an application have to be propagated to all daemons that manage the corresponding application processes, and in some cases should be forwarded from these daemons to the application processes themselves. This is done by reliably multicasting the corresponding messages in the appropriate lightweight group; the lightweight membership module at a daemon that receives such a message directs it to the corresponding lightweight endpoint module. The lightweight endpoint module then takes appropriate actions, and if necessary, passes the message to the application process as well.

GUI: There is a Java GUI program for interaction between Starfish and users. The program enables the user to inspect, modify, and generally manage the cluster, making all Starfish management features easily accessible. The GUI connects to the any of the daemons remotely, using a TCP connection, and running, at low level, the textual protocol described above to configure and manage the cluster and its applications. Of course, when using the GUI, the low level protocol is transparent to the user. The GUI was written in Java, which is platform-independent language, to ensure portability and interoperability to other operating systems and architectures.

3.1.2. Dynamicity

Starfish supports dynamic changes in the clusters. These changes can be the result of a node being added or deleted from the cluster, or might happen due to failures and recoveries of nodes. Each such change causes the group communication module to generate a new *view* event, which reports

this change to all members of the cluster. Also, changes that affect only lightweight groups are reported in the lightweight group only, by the lightweight membership module. One of the main benefits of using an underlying group communication system is that our code does not need to explicitly keep track of such changes, since we can rely on Ensemble to report these changes for us.

3.1.3. High-availability

Starfish is a highly-available system, in the sense that a failure of a few nodes does not cause the entire system to crash or hang. Instead, the system continues to run applications and to be available to clients. In particular, if none of the application processes of a given application was located on a failed node, then this application continues to run transparently. Similarly, as discussed in subsection 3.2, even if the application had one of its processes on a failed node, Starfish provides enough hooks to allow the application to overcome this failure, either by restarting the failed process on a different node, or by restructuring the computation.

Note that although the system as a whole does not stop due to a failure of a single node, client connections to this node might be lost. However, if the client reconnects to the system, he/she can continue the disrupted session from the point where the connection was cut off. At the moment clients have to explicitly choose the server they wish to connect to. However, in the future we plan to make this more transparent, using a *one-IP* type of solution [12].

3.2. Applications dynamicity and fault tolerance

3.2.1. Dynamicity

Many applications can benefit from having more nodes added to them on-the-fly, while some applications might be able to accommodate loss of a few nodes on-the-fly. This is typical in applications that have trivial parallelism, since in this case usually each node works independently on a given subset of the computation space. Thus, changing the number or nodes dynamically simply requires restructuring the computation subspace on which each node computes so that the entire compute space is covered with no duplicates.

Another feature of Starfish that allows applications to cope with dynamic changes is the C/R capability. Specifically, C/R allows Starfish to migrate application processes from one node to another, e.g., if a better node becomes available, or a new node is added to the cluster.

3.2.2. Fault tolerance

Starfish offers two forms of fault-tolerance for applications: The main fault-tolerant mechanism employed by Starfish is C/R. The C/R module of Starfish is capable of performing both coordinated and uncoordinated checkpointing of a distributed application. Locally, checkpoint can be done either at the native process level, which is OS dependent, or for pure OCaml byte code programs, at the virtual machine level, which allows restarting on a different OS (see section 4). When a node failure occurs, Starfish can automatically restart the application from the last checkpoint, or *recovery line* [14,32]. Also, in some versions of uncoordinated checkpointing, it is enough to only start the failed process from its last stored checkpoint. As we mentioned earlier in this paper, Starfish can run multiple C/R protocols side by side, which enables comparing various C/R protocols on the same platform.

The other form of fault-tolerance offered by Starfish is more application dependent, and is suitable mostly to applications that can be trivially parallelized. For such applications, whenever a node that runs one of the application processes crashes, a view event is delivered to all surviving application processes. This is done by having the application process registering a listener handler with the object bus for membership events. (Note that applications that cannot utilize view changes simply do not register listeners for membership events, and their programming model remains the conventional MPI model.) Once the surviving members learn about the failure of a node, they can repartition the data sets on which each process computes, and continue to run without interruption.

When an application is submitted to Starfish, the client can also determine the fault-tolerant policy that should be applied to this application, i.e., should automatic restart or view notifications be used, and some rules regarding how to choose the node on which a process will be started after a partial failure. For compatibility, there is also an option to kill an application whenever one of the nodes dies in the middle of its execution, which mimics non fault tolerant systems.

4. Heterogeneous checkpointing in Starfish

OCaml is a platform independent functional language. The use of this language simplified the design of Starfish and its implementation as a system supporting heterogeneous clusters. Providing C/R for heterogeneous systems is more complicated than in homogeneous ones. In the latter, checkpoint can simply be done by dumping the process core. There are optimized implementations that attempt to greatly reduce the amount of saved data [32]. But even they can rely on the fact that the architecture and operating system of the computer in which the failed application is restarted are the same as the ones in the computer in which the state was saved. In particular, such implementations can assume that the data representation, the machine registers, the stack, heap and data segments, as well as the machine's native instruction sets are all the same. Yet, in heterogeneous environments the above assumptions do not hold, and more sophisticated techniques are needed.

We provide heterogeneous checkpointing in Starfish by implementing the C/R at the OCaml virtual machine. In [2] we described the heterogeneous checkpointing implementation in OCaml, and the issues involved in doing so. In order not to hurt the performance of heterogeneous checkpointing, data is saved in the machines native representation, with a concise indication of what that representation is. During

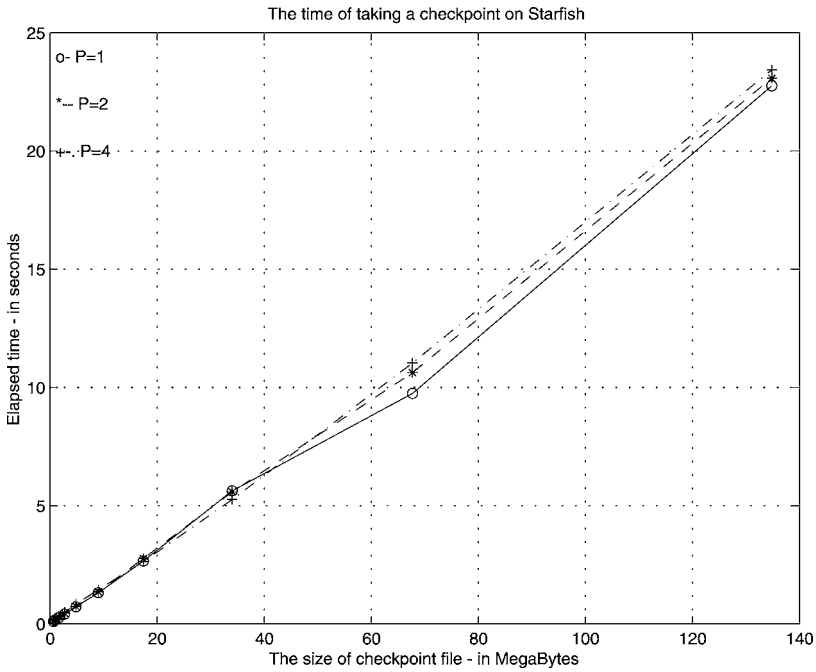

Figure 3. Native (homogeneous) checkpointing in Starfish. The smallest data point is 632 KB, which takes 0.104061 seconds for one node, 0.131898 seconds for two nodes, and 0.149219 for four nodes.

Table 2
A list of different machine types that have been tested with heterogeneous C/R.

Architecture type	OS	Representation	Word length
Intel P-II 350 MHz, i686	RedHat 6.1 Linux	little-endian	32-bit
Sun Ultra Enterprise 3000	SunOS 5.7	big-endian	32-bit
RS/6000	AIX 3.2	big-endian	32-bit
Intel P-I, 160 MHz	FreeBSD 3.2	little-endian	32-bit
Intel P-II, 350 MHz	Win NT	little-endian	32-bit
Dual Alpha DS20 500 MHz	RedHat 6.2 Linux	little-endian	64-bit

restart, the checkpointed data is converted to the machine in which the application is restarted. Table 2 lists all the machines that we have used for performing heterogeneous C/R in OCaml virtual machine. The work in [2] also provides a detailed performance analysis of our heterogeneous C/R mechanism.

5. Performance

In this section we report on some initial performance measurements conducted on the prototype of Starfish. The performance measurements were obtained using 300 MHz Pentium II computers, connected by both Ethernet and Myrinet. In the case of Ethernet, we used the regular IP stack, while with Myrinet we used the BIP user-level interface [6].

Checkpoint time: Figure 3 shows the homogeneous checkpointing time on Starfish using the stop-and-sync protocol [14]. On the other hand, figure 4 shows the heterogeneous checkpointing using the same protocol. Note that as expected, the checkpoint time grows linearly with the size of the checkpointed data. The checkpointing time is on the order of seconds. Hence, if a checkpoint is taken once every hour, it would only slow down the entire execution time by less than 1%. Also, the hardware used for these measurements is not the most advanced, and employs regular IDE bus and controller. Newer and faster hardware is likely to result is faster saving times.

The smallest data point in figure 3 is for a checkpoint file of size 632 KB, while in figure 4 is for checkpoint file of size 260 KB. These sizes correspond to checkpointing an empty program in the native process level and in the virtual machine level, respectively. In other words, this indicates the checkpoint overhead imposed by our system. We attribute this low number to our architecture, in which the run-time system on each node is divided between the application process and the daemon. The daemon, which accounts for most of the code, is shared between all processes on the same node, and is written in a way that we never have to save or recover its state. The only part of the run-time system that needs to be saved is the one that is included in the application process, and that part is relatively small. Moreover, in the heterogeneous case, the checkpointed data does not contain the virtual machine data, which is saved in the homogeneous case. Similarly, note that the largest checkpoint file reported for the same application in figure 3 is 135 MBytes, while in figure 4 it is only 96 MBytes. This is again a result of not having to save the virtual machine, and employing a more sophisticated checkpointing approach, as reported in [2].

Round-trip delay: In order to measure the application level round-trip latency, we have implemented a simple ping-style application. That is, one node sends a short message to another node, who immediately replies. We then measure

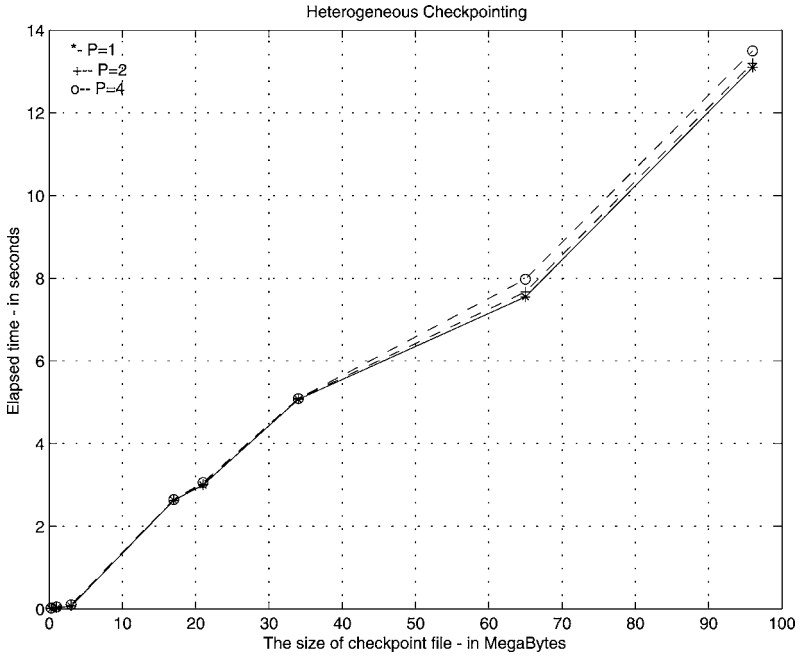

Figure 4. Virtual machine level (heterogeneous) checkpointing in Starfish. The smallest data point is 260 KB, which takes 0.0077 seconds for one node, 0.0205 seconds for two nodes, and 0.052 seconds for four nodes.

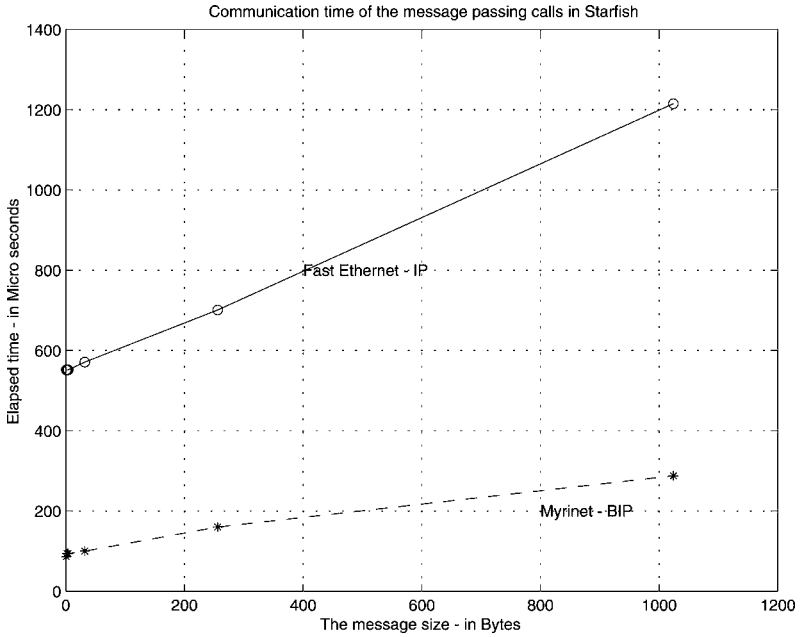

Figure 5. Round-trip delay vs. data size. The smallest data point is 1 Byte, at 86 microseconds for BIP/Myrinet and 552 microseconds for TCP/IP.

the elapsed time between sending the message and receiving the reply at the application level. This is done repeatedly a hundred times to get the average round-trip latency. We have measured the round-trip delay with both TCP/IP and BIP/Myrinet. The results of these measurements are reported in figure 5. It can be seen that the round-trip delay grows linearly with the size of the data.

The round-trip time for an empty message is 86 microseconds using BIP/Myrinet and 552 microseconds using TCP/IP, or in other words, roughly 46 microseconds one way with BIP/Myrinet and 226 microseconds one-way using TCP/IP. This time is the net overhead imposed by our system in handling a message, plus the network latency. It includes getting the message from the application and putting it on the network, and retrieving the message from the network, and then all the way back to the application. Also, these measurements were obtained on a non optimized prototype, running as bytecode. From our experience, the actual time for the optimized native code are expected to be much smaller.

System overheads: Figure 6 reports on the time a message spends in each layer of our code. Here again, we refer to the

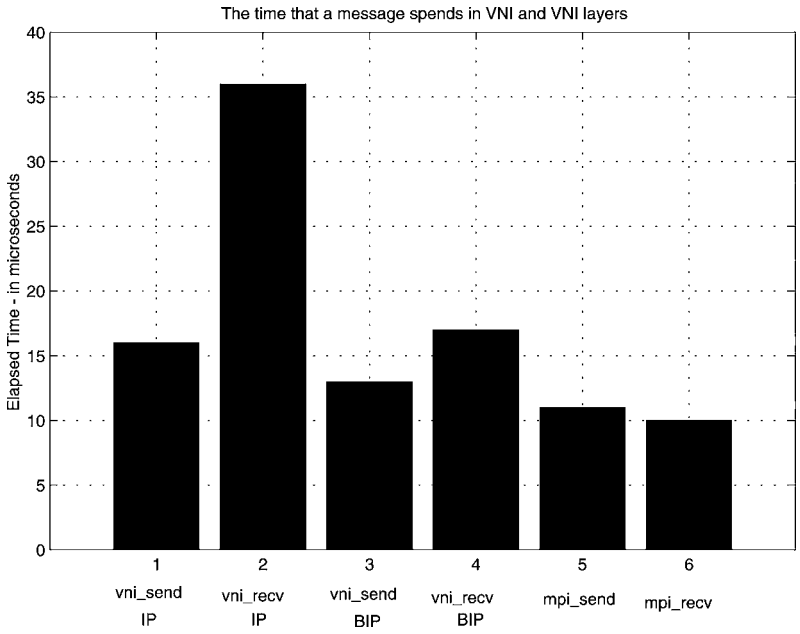

Figure 6. Layers overhead for sending and receiving messages.

non-optimized prototype, running as byte-code. Also, note that the time spent in each layer is independent of the message size, since messages are never copied in our code.

6. Related work

As discussed earlier in this paper, Starfish supports several approaches to C/R and employs group communication for providing fault-tolerance and high availability. In addition, Starfish allows dynamic changes in the number of running processes. Most of these features and other properties of Starfish have been studied. Thus, there are several distributed systems, in academia as well as in industry, that have some of these features.

There are several systems that offer C/R capabilities, e.g., Condor [23], Manetho [13], and LoadLeveler [25], and quite a few protocols and techniques for C/R have been proposed. Generally, C/R protocols can be categorized as either coordinated, in which case all processes coordinate their checkpointing to form a global consistent state [10,15,32], or as uncoordinated, in which case every process can perform checkpointing independently [1,29,32,34,41]. One of the important aspects of Starfish architecture is that it enables us to implement and study both coordinated and uncoordinated checkpointing within a single framework.

Given the large number of high-performance distributed computing systems built, it is impossible to mention all of them. Here we discuss the ones we feel are more related to our work, and examine their architecture and functionality compared to Starfish.

Condor is a distributed system that runs on a cluster of workstations [23]. Condor provides an environment for executing serial and parallel applications on clusters. Moreover, it supports C/R in order to provide fault tolerance and process migration [24]. Condor employs the sync-and-wait protocol [31], which is coordinated, and imposes several restrictions on C/R in programs. As we mentioned before, our architecture allows us to implement, side-by-side, both coordinated and uncoordinated protocols. Also, we believe that using Starfish architecture we can remove most of the restrictions imposed in Condor.

Manetho is a distributed system that runs on a cluster of workstations [13]. This system uses a novel combination of rollback-recovery and process replications to provide fault tolerance and high availability; Manetho uses coordinated checkpointing protocol as described in [10], and uses process replication to provide high availability to servers in the system [8]. Aside from supporting both coordinated and uncoordinated protocols, Starfish high availability mechanisms are somewhat different than Manetho, as we use a group communication system to manage our cluster.

Libckpt is a transparent checkpointing library on uniprocessors running UNIX [33]. It provides a mechanism for enabling fault-tolerance for long-running programming. Libckpt implements most optimizations that have been proposed to improve the performance of checkpointing, including, e.g., incremental checkpointing, forked checkpointing, copy-on-write checkpointing [32]. However, libckpt is merely a library, whereas Starfish is a complete system. Also, libckpt does not address high availability and dynamicity as we do.

LoadLeveler is a distributed system that runs on a cluster of workstations [25]. It provides an environment for executing serial and parallel applications with dynamic scheduling. In addition, it supports C/R only with serial jobs in order to balance workload and provide process migration, and is thus incomparable to Starfish.

Legion is an object-based meta-system [17]. It has been built on a collection of connected hosts to provide a virtual

computer that can access all types of data and physical resources. Legion is designed to be a worldwide virtual computer, while Starfish is designed to be a reliable and highly available distributed system for executing message-passing applications on clusters.

HPVM is a distributed system that runs on a cluster of PCs with Windows NT [11]. This system achieves high performance communication by using modern processors (300 MHz Pentium II) and FM protocol for communication on Myrinet [4,27]. In addition, the system includes efficient implementation of standard scientific computing APIs such as MPI [26]. However, HPVM does not support fault tolerance or high availability.

Millipede is a Distributed Shared Memory (DSM) system that runs on a cluster of workstations. It supports various consistency models of DSM [16], as well as thread migration inside the cluster for load-sharing and to improve the locality of memory references [22]. Millipede however, does not support fault tolerance, parallel I/O, security, and more [21]. On the other hand, Starfish system supports process migration to provide load-balancing, and in other cases to provide fault tolerance [32].

7. Discussion

Clusters of workstations offer a potential for cost effective high-performance computing. However, building usable clusters is an inherently difficult task. Successful implementations of such clusters must retain high-performance, while addressing issues like manageability, fault-tolerance, high-availability, and coping with dynamic changes in the environment.

In this paper we report on the design and architecture of Starfish, a system that tries to tackle these issues. We believe that Starfish architecture is novel in the specific way it addresses all of the above concerns. Starfish serves as a good testbed for new C/R protocols and is also a fairly portable system in its design. We have an initial release of Starfish, available from [36] and a small user community. Also, some of our ideas and techniques have been used to improve existing systems, as is the case with NetSolve [3,28], which is a client-server system for solving scientific problems remotely.

Looking into the future, developing newer and faster C/R protocols, in particular ones that utilize fast networks, is a natural research direction. Also, trying to eliminate many of the restrictions that typical checkpoint/restart systems impose would also be desirable.

Acknowledgements

Supported by Israeli Ministry of Science and Technology, Grant Number 1628-3-01.

Notes

1. Ensemble ensures reliable ordered message delivery, as well as consistent automatic failure and recovery detection. See section 3.1 for more details.
2. As discussed earlier, there are also *control messages*, but in our terminology these messages are exchanged solely by daemons, and are therefore not discussed in this section.

References

[1] A. Agbaria, H. Attiya, R. Friedman and R. Vitenberg, Quantifying rollback propagation in distributed checkpointing, in: *Proc. IEEE 20th Symposium on Reliable Distributed Systems*, October 2001, to appear.

[2] A. Agbaria and R. Friedman, Virtual machine based heterogeneous checkpointing, Technical report CS-2000-11, Technion, Israel Institute of Technology, 2000.

[3] A. Agbaria and J.S. Plank, Design, implementation, and performance of checkpointing in NetSolve, in: *Proc. IEEE of the 1st Conference on Dependable Systems and Networks*, June 2000, pp. 49–54.

[4] Y. Amir, L.E. Moser, P.M. Melliar-Smith, D.A. Agarwal and P. Ciarfella, Fast message ordering and membership using a logical token-passing ring, in: *Proc. of the 13th International Conference on Distributed Computing Systems*, May 1993, pp. 551–560.

[5] T.E. Anderson, D.E. Culler and D.A. Patterson, A case for NOW (network of workstations), IEEE Micro (February 1995).

[6] Basic Interface for Parallelism, http://lhpca.univ-lyon1.fr/bip.html.

[7] A. Basu, V. Buch, W. Vogels and T. von Eiken, U-net: A user-level network interface for parallel and distributed computing, in: *Proc. of the 15th ACM Symposium on Operating Systems Principles*, December 1996, pp. 40–53.

[8] K. Birman, The process group approach to reliable distributed computing, Communications of the ACM 36(12) (1993) 37–53.

[9] K. Birman, R. Friedman and M. Hayden, The Maestro Group manager: A structuring tool for applications with multiple quality of service requirements, Technical report TR96-1619, Department of Computer Science, Cornell University, March 1996.

[10] K.M. Chandy and L. Lamport, Distributed snapshots: Determining global states of distributed systems, ACM Transactions on Computer Systems 3(1) (February 1985) 63–75.

[11] A. Chien, M. Lauria, R. Pennington, M. Showerman, G. Ianello, M. Buchanan, K. Hane, L. Giannini, G. Koenig, S. Krishnamurthy, Q. Liu, S. Pakin and G. Sampemane, The design and evaluation of an HPVM-based Windows-NT supercomputer, Unpublished manuscript (1999).

[12] O.P. Damani, P.Y. Chung, Y. Huang, C. Kintala and Y.M. Wang, One-IP: Techniques for hosting a service on a cluster of machines, in: *Proc. of the 6th World Wide Web Conference*, April 1997.

[13] E.N. Elnozahy, Manetho: Fault tolerance in distributed systems using rollback-recovery and process replication, Ph.D. thesis, Houston University, October 1993.

[14] E.N. Elnozahy, L. Alvisi, Y.M. Wang and D.B. Johnson, A survey of rollback-recovery protocols in message-passing systems, Technical report CMU-CS-99-148, Department of Computer Science, Carnegie Mellon University, June 1999.

[15] E.N. Elnozahy, D.B. Johnson and Y.M. Wang, A survey of rollback-recovery protocols in message-passing systems, Technical report CMU-CS-96-181, Department of Computer Science, Carnegie Mellon University, October 1996.

[16] R. Friedman, M. Goldin, A. Itzkovitz and A. Schuster, Millipede: Easy parallel programming in available distributed environments, Software: Practice and Experience 27(8) (1997) 929–965.

[17] A.S. Grimshaw and W.A. Wulf, The legion vision of a Worldwide virtual computer, Communications of the ACM 40(1) (1997).

[18] W. Gropp and E. Lusk, Mpich working note: Creating a new mpich device using the channel interface, Technical report ANL/MCS-TM-000, Argonne National Laboratory.

[19] K. Guo and L. Rodrigues, Dynamic light-weight groups, in: *Proc. of the 17th International Conference on Distributed Computing and Systems*, May 1997, pp. 33–42.

[20] M. Hayden, The ensemble system, Technical report TR98-1662, Department of Computer Science, Cornell University, January 1998.

[21] A. Itzkovitz, A. Schuster and L. Shalev, The Millipede Virtual Parallel Machine for NT/PC Clusters, http://www.cs.technion.ac.il/Labs/Millipede/millipede.html.
[22] A. Itzkovitz, A. Schuster and L. Wolfovich, Thread migration and its applications in distributed shared memory systems, The Journal of Systems and Software (1998), to appear; also available as Technion CS Technical report LPCR #9603.
[23] M. Litzkow, M. Livny and M. Mutka, Condor: A hunter of idle workstations, in: *Proc. of the 8th International Conference on Distributed Computing Systems (ICDCS'88)* (1988).
[24] M. Litzkow, T. Tannenbaum, J. Basney and M. Livny, Matchmaking: Distributed resource management for high throughput computing, Technical report 1346, University of Wisconsin-Madison Computer Sciences, April 1997.
[25] LoadLeveler home page, http://www.austin.ibm.com/software.
[26] Message Passing Interface Forum, MPI-2: Extensions to the Message-Passing Interface, http://www.mcs.anl.gov/mpi (July 1997).
[27] Myricom Home Page, http://www.myri.com.
[28] NetSolve Home Page, http://www.cs.utk.edu/netsolve.
[29] R.H.B. Netzer and J. Xu, Adaptive independent checkpointing for reducing rollback propagation, Technical report CS-93-25, Department of Computer Science, Brown University, September 1993.
[30] S. Pakin, V. Karamcheti and A.A. Chien, Fast messages (FM): Efficient, portable communication for workstations clusters and massively parallel processors, IEEE Concurrency 5(2) (1997) 60–73.
[31] J.S. Plank, Efficient checkpointing on MIMD architectures, Ph.D. thesis, Princeton Unversity, January 1993.
[32] J.S. Plank, An overview of checkpointing in uniprocessor and distributed systems, focusing on implementation and performance, Technical report UT-CS-97-372, Department of Computer Science, Tennessee University, July 1997.
[33] J.S. Plank, M. Bech, G. Kingsley and K. Li, Libckpt: Transparent checkpointing under UNIX, in: *Usenix Winter 1995. Technical Conference*, New Orleans, January 1995, pp. 220–232.
[34] B. Randell, System structure for software fault tolerance, IEEE Trans. on Software Engineering SE-1(1) (June 1975) 220–232.
[35] L. Rodrigques, K. Guo, A. Sargento, R. van Renesse, B. Glade, P. Verissimo and K. Birman, Reducing interprocessor dependence in recoverable distributed shared memory, in: *Proc. of the 13th International Symposium on Reliable Distributed Systems* (1994) pp. 34–41.
[36] Starfish Home Page, http://dsl.cs.technion.ac.il/Starfish.
[37] Tandem Home Page, http://www.tandem.com.
[38] The Ensemble Home Page, http://www.cs.cornell.edu/Info/Projects/Ensemble.
[39] The OCaml Home Page, http://pauillac.inria.fr/ocaml.
[40] Virtual Interface (VI) Architecture Home Page, http://www.viarch.org.
[41] Y.M. Wang and W.K. Fuchs, Scheduling message processing for reducing rollback propagation, in: *Proc. IEEE Fault-Tolerance Computing Symposium*, July 1992, pp. 204–211.

Adnan Agbaria received his B.A. and M.A. degrees in 1994 and 1997, respectively, both in mathematics and computer science, from Haifa University. He is working toward the Ph.D. degree in computer science at the Technion – Israel Institute of Technology. His research interests include reliable distributed systems, checkpoint/restart, group communication, distributed and parallel algorithms, and computer networking. He is also a student member of the ACM.

Roy Friedman is a senior lecturer with the Department of Computer Science at the Technion – Israel Institute of Technology. He received his B.Sc. and D.Sc. in computer science from the Technion in 1990 and 1994, respectively, and was a researcher for three years at Cornell University before joining the Technion in 1997. His research areas include reliable distributed systems, high-availability, group communication, middlewares, and mobile computing. He is also one of the founders of PolyServe, Inc. Friedman is a member of the IEEE Computer Society.