Disco: Running Commodity Operating Systems on Scalable Multiprocessors

Edouard Bugnion, Scott Devine, and Mendel Rosenblum

Computer Systems Laboratory
Stanford University,
Stanford, CA 94305
{bugnion, devine, mendel}@cs.stanford.edu
http://www-flash.stanford.edu/Disco

In this paper we examine the problem of extending modern operating systems to run efficiently on large-scale shared memory multiprocessors without a large implementation effort. Our approach brings back an idea popular in the 1970s, virtual machine monitors. We use virtual machines to run multiple commodity operating systems on a scalable multiprocessor. This solution addresses many of the challenges facing the system software for these machines. We demonstrate our approach with a prototype called Disco that can run multiple copies of Silicon Graphics' IRIX operating system on a multiprocessor. Our experience shows that the overheads of the monitor are small and that the approach provides scalability as well as the ability to deal with the non-uniform memory access time of these systems. To reduce the memory overheads associated with running multiple operating systems, we have developed techniques where the virtual machines transparently share major data structures such as the program code and the file system buffer cache. We use the distributed system support of modern operating systems to export a partial single system image to the users. The overall solution achieves most of the benefits of operating systems customized for scalable multiprocessors yet it can be achieved with a significantly smaller implementation effort.

1 Introduction

Scalable computers have moved from the research lab to the marketplace. Multiple vendors are now shipping scalable systems with configurations in the tens or even hundreds of processors. Unfortunately, the system software for these machines has often trailed hardware in reaching the functionality and reliability expected by modern computer users.

Operating systems developers shoulder much of the blame for the inability to deliver on the promises of these machines. Extensive modifications to the operating system are required to efficiently support scalable machines. The size and complexity of modern operating systems have made these modifications a resource-intensive undertaking.

SOSP 16.

In this paper, we present an alternative approach for constructing the system software for these large computers. Rather than making extensive changes to existing operating systems, we insert an additional layer of software between the hardware and operating system. This layer acts like a virtual machine monitor in that multiple copies of "commodity" operating systems can be run on a single scalable computer. The monitor also allows these commodity operating systems to efficiently cooperate and share resources with each other. The resulting system contains most of the features of custom scalable operating systems developed specifically for these machines at only a fraction of their complexity and implementation cost. The use of commodity operating systems leads to systems that are both reliable and compatible with the existing computing base.

To demonstrate the approach, we have constructed a prototype system targeting the Stanford FLASH shared memory multiprocessor [17], an experimental cache coherent non-uniform memory architecture (ccNUMA) machine. The prototype, called Disco, combines commodity operating systems not originally designed for such large-scale multiprocessors to form a high performance system software base.

Disco contains many features that reduce or eliminate the problems associated with traditional virtual machine monitors. Specifically, it minimizes the overhead of virtual machines and enhances the resource sharing between virtual machines running on the same system. Disco allows the operating systems running on different virtual machines to be coupled using standard distributed systems protocols such as NFS and TCP/IP. It also allows for efficient sharing of memory and disk resources between virtual machines. The sharing support allows Disco to maintain a global buffer cache transparently shared by all the virtual machines, even when the virtual machines communicate through standard distributed protocols.

Our experiments with realistic workloads on a detailed simulator of the FLASH machine show that Disco achieves its goals. With a few simple modifications to an existing commercial operating system, the basic overhead of virtualization is at most 16% for all our uniprocessor workloads. We show that a system with eight virtual machines can run some workloads 40% faster than on a commercial symmetric multiprocessor operating system by increasing the scalability of the system software, without substantially increasing the system's memory footprint. Finally, we show that page placement and dynamic page migration and replication allow Disco to hide the NUMA-ness of the memory system, improving the execution time by up to 37%.

In Section 2, we provide a more detailed presentation of the problem being addressed. Section 3 describes an overview of the approach and the challenges of using virtual machines to construct the system software for large-scale shared-memory multiprocessors. Section 4 presents the design and implementation of Disco and Section 5 shows experimental results. We end the paper with a discussion of related work in Section 6 and conclude in Section 7.

2 Problem Description

This paper addresses the problems seen by computer vendors attempting to provide system software for their innovative hardware. For the purposes of this paper, the innovative hardware is scalable shared memory multiprocessors, but the issues are similar for any hardware innovation that requires significant changes in the system software. For shared memory multiprocessors, research groups have demonstrated prototype operating systems such as Hive [5] and Hurricane [25] that address the challenges of scalability and fault containment. Silicon Graphics has announced the Cellular IRIX kernel to support its shared memory machine, the Origin2000 [18]. These designs require significant OS changes, including partitioning the system into scalable units, building a single system image across the units, as well as other features such as fault containment [5] and ccNUMA management [26].

With the size of the system software for modern computers in the millions of lines of code, the changes for ccNUMA machines represent a significant development cost. These changes have an impact on many of the standard modules that make up a modern system, such as virtual memory management and the scheduler. As a result, the system software for these machines is generally delivered significantly later than the hardware. Even when the changes are functionally complete, they are likely to introduce instabilities for a certain period of time.

Late, incompatible, and possibly even buggy system software can significantly impact the success of such machines, regardless of the innovations in the hardware. As the computer industry matures, users expect to carry forward their large base of existing application programs. Furthermore, with the increasing role that computers play in today's society, users are demanding highly reliable and available computing systems. The cost of achieving reliability in computers may even dwarf the benefits of the innovation in hardware for many application areas.

Computer hardware vendors that use "commodity" operating systems such as Microsoft's Windows NT [9] face an even greater problem in obtaining operating system support for their ccNUMA multiprocessors. These vendors need to persuade an independent company to make changes to the operating system to support the new hardware. Not only must these vendors deliver on the promises of the innovative hardware, they must also convince powerful software companies that running on their hardware is worth the effort of the port [20].

Given this situation, it is no small wonder that computer architects frequently complain about the constraints and inflexibility of system software. From their perspective, these software constraints are an impediment to innovation. To reduce the gap between hardware innovations and the adaptation of system software, system developers must find new ways to develop their software more quickly and with fewer risks of incompatibilities and instabilities.

3 A Return to Virtual Machine Monitors

To address the problem of providing system software for scalable multiprocessors, we have developed a new twist on the relatively old idea of virtual machine monitors [13]. Rather than attempting to modify existing operating systems to run on scalable shared-memory multiprocessors, we insert an additional layer of software between the hardware and the operating system. This layer of software, called a virtual machine monitor, virtualizes all the resources of the machine, exporting a more conventional hardware interface to the operating system. The monitor manages all the resources so that multiple virtual machines can coexist on the same multiprocessor. Figure 1 shows how the virtual machine monitor allows multiple copies of potentially different operating systems to coexist.

Virtual machine monitors, in combination with commodity and specialized operating systems, form a flexible system software solution for these machines. A large ccNUMA multiprocessor can be configured with multiple virtual machines each running a commodity operating system such as Microsoft's Windows NT or some variant of UNIX. Each virtual machine is configured with the processor and memory resources that the operating system can effectively handle. The virtual machines communicate using standard distributed protocols to export the image of a cluster of machines.

Although the system looks like a cluster of loosely-coupled machines, the virtual machine monitor uses global policies to manage all the resources of the machine, allowing workloads to exploit the fine-grain resource sharing potential of the hardware. For example, the monitor can move memory between virtual machines to keep applications from paging to disk when free memory is available in the machine. Similarly, the monitor dynamically schedules virtual processors on the physical processors to balance the load across the machine.

The use of commodity software leverage the significant engineering effort invested in these operating systems and allows ccNUMA machines to support their large application base. Since the monitor is a relatively simple piece of code, this can be done with a small implementation effort as well

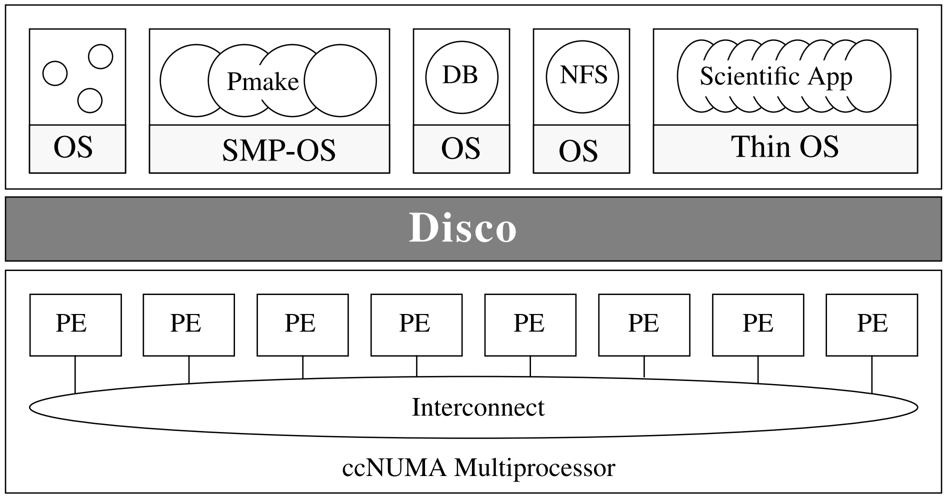

FIGURE 1. Architecture of Disco: Disco is a virtual machine monitor, a software layer between the hardware and multiple virtual machines that run independent operating systems. This allows multiple copies of a commodity operating system to coexist with specialized "thin" operating systems on the same hardware. The multiprocessor consists of a set of processing elements (PE) connected by a high-performance interconnect. Each processing element contains a number of processors and a portion of the memory of the machine.

as with a low risk of introducing software bugs and incompatibilities.

The approach offers two different possible solutions to handle applications whose resource needs exceed the scalability of commodity operating systems. First, a relatively simple change to the commodity operating system can allow applications to explicitly share memory regions across virtual machine boundaries. The monitor contains a simple interface to setup these shared regions. The operating system is extended with a special virtual memory segment driver to allow processes running on multiple virtual machines to share memory. For example, a parallel database server could put its buffer cache in such a shared memory region and have query engines running on multiple virtual machines.

Second, the flexibility of the approach supports specialized operating systems for resource-intensive applications that do not need the full functionality of the commodity operating systems. These simpler, specialized operating systems better support the needs of the applications and can easily scale to the size of the machine. For example, a virtual machine running a highly-scalable lightweight operating system such as Puma [24] allows large scientific applications to scale to the size of the machine. Since the specialized operating system runs in a virtual machine, it can run alongside commodity operating systems running standard application programs. Similarly, other important applications such as database and web servers could be run in highly-customized operating systems such as database accelerators.

Besides the flexibility to support a wide variety of workloads efficiently, this approach has a number of additional advantages over other system software designs targeted for ccNUMA machines. Running multiple copies of an operating system, each in its own virtual machine, handles the challenges presented by ccNUMA machines such as scalability and fault-containment. The virtual machine becomes the unit of scalability, analogous to the cell structure of Hurricane, Hive, and Cellular IRIX. With this approach, only the monitor itself and the distributed systems protocols need to scale to the size of the machine. The simplicity of the monitor makes this task easier than building a scalable operating system.

The virtual machine also becomes the unit of fault containment where failures in the system software can be contained in the virtual machine without spreading over the entire machine. To provide hardware fault-containment, the monitor itself must be structured into cells. Again, the simplicity of the monitor makes this easier than to protect a full-blown operating system against hardware faults.

NUMA memory management issues can also be handled by the monitor, effectively hiding the entire problem from the operating systems. With the careful placement of the pages of a virtual machine's memory and the use of dynamic page migration and page replication, the monitor can export a more conventional view of memory as a uniform memory access (UMA) machine. This allows the non-NUMA-aware memory management policies of commodity operating systems to work well, even on a NUMA machine.

Besides handling ccNUMA multiprocessors, the approach also inherits all the advantages of traditional virtual machine monitors. Many of these benefits are still appropriate today and some have grown in importance. By exporting multiple virtual machines, a single ccNUMA multiprocessor can have multiple different operating systems simultaneously running on it. Older versions of the system software can be kept around to provide a stable platform for keeping legacy applications running. Newer versions can be staged in carefully with critical applications residing on the older operating systems until the newer versions have proven themselves. This approach provides an excellent way of introducing new and innovative system software while still providing a stable computing base for applications that favor stability over innovation.

3.1 Challenges Facing Virtual Machines

Unfortunately, the advantages of using virtual machine monitors come with certain disadvantages as well. Among the well-documented problems with virtual machines are the overheads due to the virtualization of the hardware resources, resource management problems, and sharing and communication problems.

Overheads. The overheads present in traditional virtual machine monitors come from many sources, including the additional exception processing, instruction execution and memory needed for virtualizing the hardware. Operations such as the execution of privileged instructions cannot be safely exported directly to the operating system and must be emulated in software by the monitor. Similarly, the access to I/O devices is virtualized, so requests must be intercepted and remapped by the monitor.

In addition to execution time overheads, running multiple independent virtual machines has a cost in additional memory. The code and data of each operating system and application is replicated in the memory of each virtual machine. Furthermore, large memory structures such as the file system buffer cache are also replicated resulting in a significant increase in memory usage. A similar waste occurs with the replication of file systems for the different virtual machines.

Resource Management. Virtual machine monitors frequently experience resource management problems due to the lack of information available to the monitor to make good policy decisions. For example, the instruction execution stream of an operating system's idle loop or the code for lock busy-waiting is indistinguishable at the monitor's level from some important calculation. The result is that the monitor may schedule resources for useless computation while useful computation may be waiting. Similarly, the monitor does not know when a page is no longer being actively used by a virtual machine, so it cannot reallocate it to another virtual machine. In general, the monitor must make resource management decisions without the high-level knowledge that an operating system would have.

Communication and Sharing. Finally, running multiple independent operating systems made sharing and communication difficult. For example under CMS on VM/370, if a virtual disk containing a user's files was in use by one virtual machine it could not be accessed by another virtual machine. The same user could not start two virtual machines, and different users could not easily share files. The virtual machines looked like a set of independent stand-alone systems that simply happened to be sharing the same hardware.

Although these disadvantages still exist, we have found their impact can be greatly reduced by combining recent advances in operating system technology with some new tricks implemented in the monitor. For example, the prevalence of support in modern operating systems for interoperating in a distributed environment greatly reduces the communication and sharing problems described above. In the following section we present techniques that allow the overheads to be small compared to the benefits that can be achieved through this approach.

4 Disco: A Virtual Machine Monitor

Disco is a virtual machine monitor designed for the FLASH multiprocessor [17], a scalable cache-coherent multiprocessor. The FLASH multiprocessor consists of a collection of nodes each containing a processor, main memory, and I/O devices. The nodes are connected together with a high-performance scalable interconnect. The machines use a directory to maintain cache coherency, providing to the software the view of a shared-memory multiprocessor with non-uniform memory access times. Although written for the FLASH machine, the hardware model assumed by Disco is also available on a number of commercial machines including the Convex Exemplar [4], Silicon Graphics Origin2000 [18], Sequent NUMAQ [19], and DataGeneral NUMALiine.

This section describes the design and implementation of Disco. We first describe the key abstractions exported by Disco. We then describe the implementation of these abstractions. Finally, we discuss the operating system requirements to run on top of Disco.

4.1 Disco's Interface

Disco runs multiple independent virtual machines simultaneously on the same hardware by virtualizing all the resources of the machine. Each virtual machine can run a standard operating system that manages its virtualized resources independently of the rest of the system.

Processors. To match the FLASH machine, the virtual CPUs of Disco provide the abstraction of a MIPS R10000 processor. Disco correctly emulates all instructions, the memory management unit, and the trap architecture of the processor allowing unmodified applications and existing operating systems to run on the virtual machine. Though required for the FLASH machine, the choice of the processor was unfortunate for Disco since the R10000 does not support the complete virtualization of the kernel virtual address space. Section 4.3.1 details the OS changes needed to allow kernel-mode code to run on Disco.

Besides the emulation of the MIPS processor, Disco extends the architecture to support efficient access to some processor functions. For example, frequent kernel operations such as enabling and disabling CPU interrupts and accessing privileged registers can be performed using load and store instructions on special addresses. This interface allows operating systems tuned for Disco to reduce the overheads caused by trap emulation.

Physical Memory. Disco provides an abstraction of main memory residing in a contiguous physical address space starting at address zero. This organization was selected to match the assumptions made by the operating systems we run on top of Disco.

Since most commodity operating systems are not designed to effectively manage the non-uniform memory of the FLASH machine, Disco uses dynamic page migration and replication to export a nearly uniform memory access time memory architecture to the software. This allows a non-NUMA aware operating system to run well on FLASH without the changes needed for NUMA memory management.

I/O Devices. Each virtual machine is created with a specified set of I/O devices, such as disks, network interfaces, periodic interrupt timers, clock, and a console. As with processors and physical memory, most operating systems assume exclusive access to their I/O devices, requiring Disco to virtualize each I/O device. Disco must intercept all communication to and from I/O devices to translate or emulate the operation.

Because of their importance to the overall performance and efficiency of the virtual machine, Disco exports special abstractions for the SCSI disk and network devices. Disco virtualizes disks by providing a set of virtual disks that any virtual machine can mount. Virtual disks can be configured to support different sharing and persistency models. A virtual disk can either have modifications (i.e. disk write requests) stay private to the virtual machine or they can be visible to other virtual machines. In addition, these modifications can be made persistent so that they survive the shutdown of the virtual machine or non-persistent so that they disappear with each reboot.

To support efficient communication between virtual machines, as well as other real machines, the monitor virtualizes access to the networking devices of the underlying system. Each virtual machine is assigned a distinct link-level address on an internal virtual subnet handled by Disco. Besides the standard network interfaces such as Ethernet and FDDI, Disco supports a special network interface that can handle large transfer sizes without fragmentation. For communication with the world outside the machine, Disco acts as a gateway that uses the network interfaces of the machine to send and receive packets.

4.2 Implementation of Disco

Like most operating systems that run on shared-memory multiprocessors, Disco is implemented as a multi-threaded shared memory program. Disco differs from existing systems in that careful attention has been given to NUMA memory placement, cache-aware data structures, and interprocessor communication patterns. For example, Disco does not contain linked lists or other data structures with poor cache behavior. The small size of Disco, about 13,000 lines of code, allows for a higher degree of tuning than is possible with million line operating systems.

To improve NUMA locality, the small code segment of Disco, currently 72KB, is replicated into all the memories of FLASH machine so that all instruction cache misses can be satisfied from the local node. Machine-wide data structures are partitioned so that the parts that are accessed only or mostly by a single processor are in a memory local to that processor.

For the data structures accessed by multiple processors, very few locks are used and wait-free synchronization [14] using the MIPS LL/SC instruction pair is heavily employed. Disco communicates through shared-memory in most cases. It uses inter-processor interrupts for specific actions that change the state of a remote virtual processor, for example TLB shootdowns and posting of an interrupt to a given virtual CPU. Overall, Disco is structured more like a highly tuned and scalable SPLASH application [27] than like a general-purpose operating system.

4.2.1 Virtual CPUs

Like previous virtual machine monitors, Disco emulates the execution of the virtual CPU by using direct execution on the real CPU. To schedule a virtual CPU, Disco sets the real machines' registers to those of the virtual CPU and jumps to the current PC of the virtual CPU. By using direct execution, most operations run at the same speed as they would on the raw hardware. The challenge of using direct execution is the detection and fast emulation of those operations that cannot be safely exported to the virtual machine. These operations are primarily the execution of privileged instructions performed by the operating system such as TLB modification, and the direct access to physical memory and I/O devices.

For each virtual CPU, Disco keeps a data structure that acts much like a process table entry in a traditional operating system. This structure contains the saved registers and other state of a virtual CPU when it is not scheduled on a real CPU. To perform the emulation of privileged instructions, Disco additionally maintains the privileged registers and TLB contents of the virtual CPU in this structure.

On the MIPS processor, Disco runs in kernel mode with full access to the machine's hardware. When control is given to a virtual machine to run, Disco puts the processor in supervisor mode when running the virtual machine's operating system, and in user mode otherwise. Supervisor mode allows the operating system to use a protected portion of the address space (the supervisor segment) but does not give access to privileged instructions or physical memory. Applications and kernel code can however still be directly executed since Disco emulates the operations that cannot be issued in supervisor mode. When a trap such as page fault, system call, or bus error occurs, the processor traps to the monitor that emulates the effect of the trap on the currently scheduled virtual processor. This is done by updating some of the privileged registers of the virtual processor and jumping to the virtual machine's trap vector.

Disco contains a simple scheduler that allows the virtual processors to be time-shared across the physical processors of the machine. The scheduler cooperates with the memory management to support affinity scheduling that increases data locality.

4.2.2 Virtual Physical Memory

To virtualize physical memory, Disco adds a level of address translation and maintains *physical-to-machine* address mappings. Virtual machines use *physical addresses* that have memory starting at address zero and continuing for the size of virtual machine's memory. Disco maps these physical addresses to the 40 bit *machine addresses* used by the memory system of the FLASH machine.

Disco performs this physical-to-machine translation using the software-reloaded translation-lookaside buffer (TLB) of the MIPS processor. When an operating system attempts to insert a virtual-to-physical mapping into the TLB, Disco emulates this operation by translating the physical address into the corresponding machine address and inserting this corrected TLB entry into the TLB. Once the TLB entry has been established, memory references through this mapping are translated with no additional overhead by the processor.

To quickly compute the corrected TLB entry, Disco keeps a per virtual machine *pmap* data structure that contains one entry for each physical page of a virtual machine. Each pmap entry contains a pre-computed TLB entry that references the physical page location in real memory. Disco merges that entry with the protection bits of the original entry before inserting it into the TLB. The pmap entry also contains backmaps pointing to the virtual addresses that are used to invalidate mappings from the TLB when a page is taken away from the virtual machine by the monitor.

On MIPS processors, all user mode memory references must be translated by the TLB but kernel mode references used by operating systems may directly access physical memory and I/O devices through the unmapped segment of the kernel virtual address space. Many operating systems place both the operating system code and data in this segment. Unfortunately, the MIPS architecture bypasses the TLB for this direct access segment making it impossible for Disco to efficiently remap these addresses using the TLB. Having each operating system instruction trap into the monitor would lead to unacceptable performance. We were therefore required to re-link the operating system code and data to a mapped region of the address space. This problem seems unique to MIPS as other architectures such as Alpha can remap these regions using the TLB.

The MIPS processors tag each TLB entry with an address space identifier (ASID) to avoid having to flush the TLB on MMU context switches. To avoid the complexity of virtualizing the ASIDs, Disco flushes the machine's TLB when scheduling a different virtual CPU on a physical processor. This approach speeds up the translation of the TLB entry since the ASID field provided by the virtual machine can be used directly.

A workload executing on top of Disco will suffer an increased number of TLB misses since the TLB is additionally used for all operating system references and since the TLB must be flushed on virtual CPU switches. In addition, each TLB miss is now more expensive because of the emulation of the trap architecture, the emulation of privileged instructions in the operating systems's TLB-miss handler, and the remapping of physical addresses described above. To lessen the performance impact, Disco caches recent virtual-to-machine translations in a second-level software TLB. On each TLB miss, Disco's TLB miss handler first consults the second-level TLB. If it finds a matching virtual address it can simply place the cached mapping in the TLB, otherwise it forwards the TLB miss exception to the operating system running on the virtual machine. The effect of this optimization is that virtual machines appear to have much larger TLBs than the MIPS processors.

4.2.3 NUMA Memory Management

Besides providing fast translation of the virtual machine's physical addresses to real machine pages, the memory management part of Disco must also deal with the allocation of real memory to virtual machines. This is a particularly important task on ccNUMA machines since the commodity operating system is depending on Disco to deal with the non-uniform memory access times. Disco must try to allocate memory and schedule virtual CPUs so that cache misses generated by a virtual CPU will be satisfied from local memory rather than having to suffer the additional latency of a remote cache miss. To accomplish this, Disco implements a dynamic page migration and page replication system [2,7] that moves or replicates pages to maintain locality between a virtual CPU's cache misses and the memory pages to which the cache misses occur.

Disco targets machines that maintain cache-coherence in hardware. On these machines, NUMA management, implemented either in the monitor or in the operating system, is not required for correct execution, but rather an optimization that enhances data locality. Disco uses a robust policy that moves only pages that will likely result in an eventual performance benefit [26]. Pages that are heavily accessed by only one node are migrated to that node. Pages that are primarily read-shared are replicated to the nodes most heavily accessing them. Pages that are write-shared are not moved because they fundamentally cannot benefit from either migration or replication. Disco's policy also limits the number of times a page can move to avoid excessive overheads.

Disco's page migration and replication policy is driven

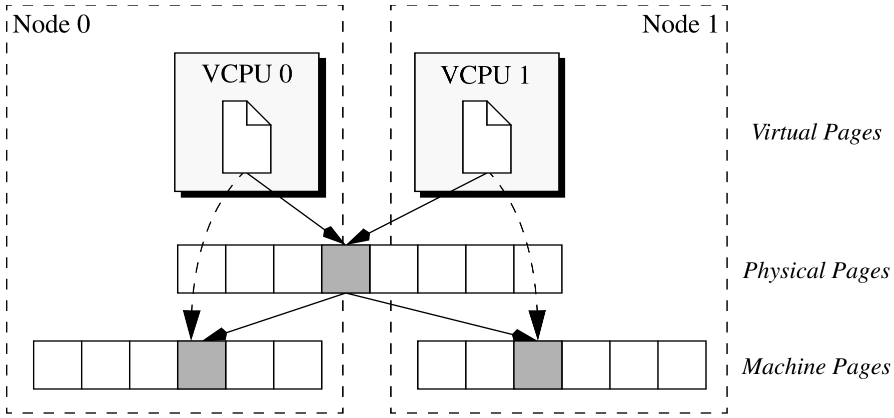

FIGURE 2. Transparent Page Replication. Disco uses the physical to machine mapping to replicate user and kernel pages. Virtual pages from VCPUs 0 and 1 of the same virtual machine both map the same physical page of their virtual machine. However, Disco transparently maps each virtual page to a machine page replica that is located on the local node.

by the cache miss counting facility provided by the FLASH hardware. FLASH counts cache misses to each page from every physical processor. Once FLASH detects a hot page, the monitor chooses between migrating and replicating the hot page based on the cache miss counters. To migrate a page, the monitor transparently changes the physical-to-machine mapping. It first invalidates any TLB entries mapping the old machine page and then copies the data to a local machine page. To replicate a page, the monitor must first downgrade all TLB entries mapping the machine page to ensure read-only accesses. It then copies the page to the local node and updates the relevant TLB entries mapping the old machine page. The resulting configuration after replication is shown in Figure 2.

Disco maintains a *memmap* data structure that contains an entry for each real machine memory page. To perform the necessary TLB shootdowns during a page migration or replication, the memmap entry contains a list of the virtual machines using the page and the virtual addresses used to access them. A memmap entry also contains pointers to any replicated copies of the page.

4.2.4 Virtual I/O Devices

To virtualize access to I/O devices, Disco intercepts all device accesses from the virtual machine and eventually forwards them to the physical devices. Disco could interpose on the programmed input/output (PIOs) from the operating system device drivers by trapping into the monitor and emulating the functionality of the hardware device assumed by the version of the operating system we used. However we found it was much cleaner to simply add special device drivers into the operating system. Each Disco device defines a *monitor call* used by the device driver to pass all command arguments in a single trap.

Devices such as disks and network interfaces include a DMA map as part of their arguments. Disco must intercept such DMA requests to translate the physical addresses specified by the operating systems into machine addresses. Disco's device drivers then interact directly with the physical device.

For devices accessed by a single virtual machine, Disco only needs to guarantee the exclusivity of this access and translate the physical memory addresses of the DMA, but does not need to virtualize the I/O resource itself.

The interposition on all DMA requests offers an opportunity for Disco to share disk and memory resources among virtual machines. Disco's copy-on-write disks allow virtual machines to share both main memory and disk storage resources. Disco's virtual network devices allow virtual machines to communicate efficiently. The combination of these two mechanisms, detailed in Section 4.2.5 and Section 4.2.6, allows Disco to support a system-wide cache of disk blocks in memory that can be transparently shared between all the virtual machines.

4.2.5 Copy-on-write Disks

Disco intercepts every disk request that DMAs data into memory. When a virtual machine requests to read a disk block that is already in main memory, Disco can process the request without going to disk. Furthermore, if the disk request is a multiple of the machine's page size, Disco can process the DMA request by simply mapping the page into the virtual machine's physical memory. In order to preserve the semantics of a DMA operation, Disco maps the page read-only into the destination address page of the DMA. Attempts to modify a shared page will result in a copy-on-write fault handled internally by the monitor.

Using this mechanism, multiple virtual machines accessing a shared disk end up sharing machine memory. The copy-on-write semantics means that the virtual machine is unaware of the sharing with the exception that disk requests can finish nearly instantly. Consider an environment running multiple virtual machines for scalability purposes. All the virtual machines can share the same root disk containing the

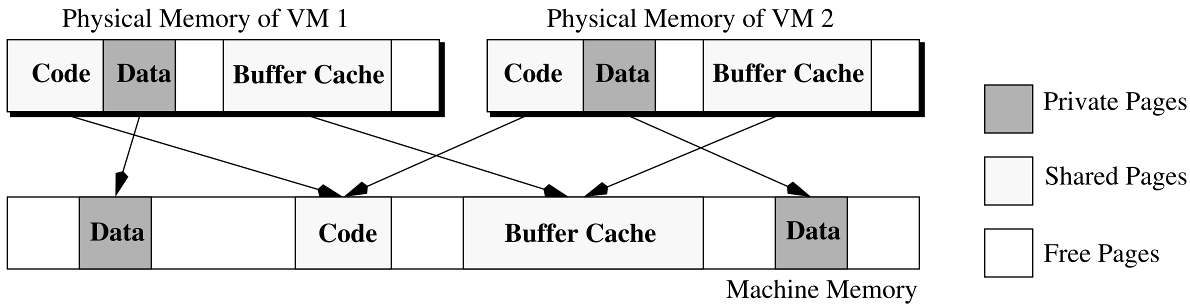

FIGURE 3. Transparent Sharing of Pages. Read only pages brought in from disk such as the kernel text and the buffer cache can be transparently shared between virtual machines. This creates a global buffer cache shared across virtual machines and helps reduce the memory footprint of the system.

kernel and application programs. The code and other read-only data stored on the disk will be DMA-ed into memory by the first virtual machine that accesses it. Subsequent requests will simply map the page specified to the DMA engine without transferring any data. The result is shown in Figure 3 where all virtual machines share these read-only pages. Effectively we get the memory sharing patterns expected of a single shared memory multiprocessor operating system even though the system runs multiple independent operating systems.

To preserve the isolation of the virtual machines, disk writes must be kept private to the virtual machine that issues them. Disco logs the modified sectors so that the copy-on-write disk is never actually modified. For persistent disks, these modified sectors would be logged in a separate disk partition managed by Disco. To simplify our implementation, we only applied the concept of copy-on-write disks to non-persistent disks and kept the modified sectors in main memory whenever possible.

The implementation of this memory and disk sharing feature of Disco uses two data structures. For each disk device, Disco maintains a B-Tree indexed by the range of disk sectors being requested. This B-Tree is used to find the machine memory address of the sectors in the global disk cache. A second B-Tree is kept for each disk and virtual machine to find any modifications to the block made by that virtual machine. We used B-Trees to efficiently support queries on ranges of sectors [6].

The copy-on-write mechanism is used for file systems such as the root disk whose modifications as not intended to be persistent or shared across virtual machines. For persistent disks such as the one containing user files, Disco enforces that only a single virtual machine can mount the disk at any given time. As a result, Disco does not need to virtualize the layout of the disk. Persistent disks can be accessed by other virtual machines through a distributed file system protocol such as NFS.

4.2.6 Virtual Network Interface

The copy-on-write mechanism for disks allows the sharing of memory resources across virtual machines, but does not allow virtual machines to communicate with each other. To communicate, virtual machines use standard distributed protocols. For example, virtual machines share files through NFS. As a result, shared data will end up in both the client's and server's buffer cache. Without special attention, the data will be duplicated in machine memory. We designed a virtual subnet managed by Disco that allows virtual machines to communicate with each other, while avoiding replicated data whenever possible.

The virtual subnet and networking interfaces of Disco also use copy-on-write mappings to reduce copying and to allow for memory sharing. The virtual device uses ethernet-like addresses and does not limit the maximum transfer unit (MTU) of packets. A message transfer sent between virtual machines causes the DMA unit to map the page read-only into both the sending and receiving virtual machine's physical address spaces. The virtual network interface accepts messages that consist of scattered buffer fragments. Our implementation of the virtual network in Disco and in the operating system's device driver always respects the data alignment of the outgoing message so that properly aligned message fragments that span a complete page are always remapped rather than copied.

Using this mechanism, a page of data read from disk into the file cache of a file server running in one virtual machine can be shared with client programs that request the file using standard distributed file system protocol such as NFS. As shown in Figure 4, Disco supports a global disk cache even when a distributed file system is used to connect the virtual machines. In practice, the combination of copy-on-write disks and the access to persistent data through the specialized network device provides a global buffer cache that is transparently shared by independent virtual machines.

As a result, all read-only pages can be shared between virtual machines. Although this reduces the memory footprint, this may adversely affect data locality as most sharers will access the page remotely. However, Disco's page replication policy selectively replicates the few "hot" pages that suffer the most cache misses. Pages are therefore shared whenever possible and replicated only when necessary to improve performance.

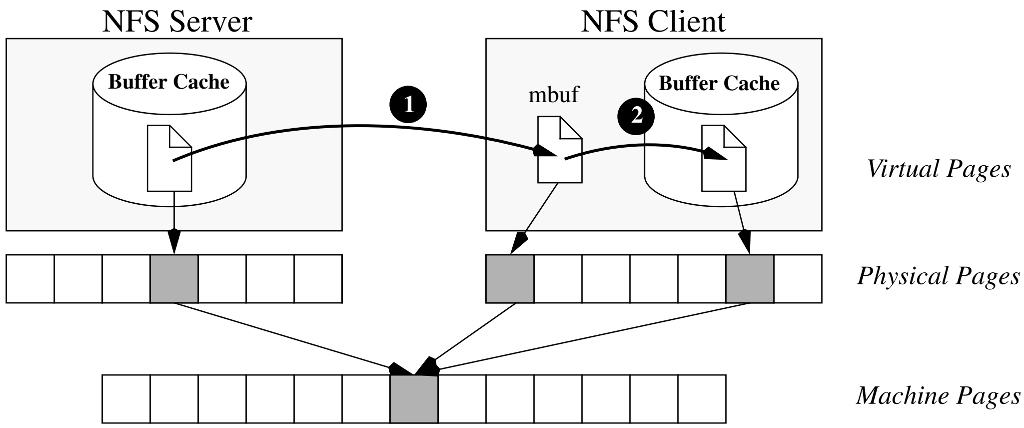

FIGURE 4. Transparent Sharing of Pages Over NFS. This figure illustrates the case when the NFS reply, to a read request, includes a data page. (1) The monitor's networking device remaps the data page from the source's machine address space to the destination's. (2) The monitor remaps the data page from the driver's mbuf to the clients buffer cache. This remap is initiated by the operating system through a monitor call.

4.3 Running Commodity Operating Systems

The "commodity" operating system we run on Disco is IRIX 5.3, a UNIX SVR4 based operating system from Silicon Graphics. Disco is however independent of any specific operating system and we plan to support others such as Windows NT and Linux.

In their support for portability, modern operating systems present a hardware abstraction level (HAL) that allows the operating system to be effectively "ported" to run on new platforms. Typically the HAL of modern operating systems changes with each new version of a machine while the rest of the system can remain unchanged. Our experience has been that relatively small changes to the HAL can reduce the overhead of virtualization and improve resource usage.

Most of the changes made in IRIX were part of the HAL[1]. All of the changes were simple enough that they are unlikely to introduce a bug in the software and did not require a detailed understanding of the internals of IRIX. Although we performed these changes at the source level as a matter of convenience, many of them were simple enough to be performed using binary translation or augmentation techniques.

4.3.1 Necessary Changes for MIPS Architecture

Virtual processors running in supervisor mode cannot efficiently access the KSEG0 segment of the MIPS virtual address space, that always bypasses the TLB. Unfortunately, many MIPS operating systems including IRIX 5.3 place the kernel code and data in the KSEG0 segment. As a result, we needed to relocate the unmapped segment of the virtual machines into a portion of the mapped supervisor segment of the MIPS processor. This allowed Disco to emulate the direct memory access efficiently using the TLB. The need for relocating the kernel appears to be unique to MIPS and is not present in other modern architecture such as Alpha, x86, SPARC, and PowerPC.

Making these changes to IRIX required changing two header files that describe the virtual address space layout, changing the linking options, as well as 15 assembly statements in *locore.s*. Unfortunately, this meant that we needed to re-compile and re-link the IRIX kernel to run on Disco.

4.3.2 Device Drivers

Disco's monitor call interface reduces the complexity and overhead of accessing I/O devices. We implemented UART, SCSI disks, and ethernet drivers that match this interface. Since the monitor call interface provides the view of an idealized device, the implementation of these drivers was straightforward. Since kernels are normally designed to run with different device drivers, this kind of change can be made without the source and with only a small risk of introducing a bug.

The complexity of the interaction with the specific devices is left to the virtual machine monitor. Fortunately, we designed the virtual machine monitor's internal device driver interface to simplify the integration of existing drivers written for commodity operating systems. Disco uses IRIX's original device drivers.

4.3.3 Changes to the HAL

Having to take a trap on every privileged register access can cause significant overheads when running kernel code such as synchronization routines and trap handlers that frequently access privileged registers. To reduce this overhead, we patched the HAL of IRIX to convert these frequently used privileged instructions to use non-trapping load and store in-

1. Unlike other operating systems, IRIX is not structured with a well-defined HAL. In this paper, the HAL includes all the platform and processor-specific functions of the operating system.

structions to a special page of the address space that contains these registers. This optimization is only applied to instructions that read and write privileged registers without causing other side-effects. Although for this experiment we performed the patches by hand to only a few critical locations, the patches could easily be automatically applied when the privileged instruction first generates a trap. As part of the emulation process, Disco could overwrite certain instructions with the special load and store so it would not suffer the overhead of the trap again.

To help the monitor make better resource management decisions, we have added code to the HAL to pass hints to the monitor giving it higher-level knowledge of resource utilization. We inserted a small number of monitor calls in the physical memory management module of the operating systems. The first monitor call requests a zeroed page. Since the monitor must clear pages to ensure the isolation of virtual machines anyway, the operating system is freed from this task. A second monitor call informs Disco that a page has been put on the operating system's freelist without a chance of reclamation, so that Disco can immediately reclaim the memory.

To improve the utilization of processor resources, Disco assigns special semantics to the reduced power consumption mode of the MIPS processor. This mode is used by the operating system whenever the system is idle. Disco will deschedule the virtual CPU until the mode is cleared or an interrupt is posted. A monitor call inserted in the HAL's idle loop would have had the same effect.

4.3.4 Other Changes to IRIX

For some optimizations Disco relies on the cooperation of the operating system. For example, the virtual network device can only take advantage of the remapping techniques if the packets contain properly aligned, complete pages that are not written. We found that the operating system's networking subsystem naturally meets most of the requirements. For example, it preserves the alignment of data pages, taking advantage of the scatter/gather options of networking devices. Unfortunately, IRIX's *mbuf* management is such that the data pages of recently freed mbufs are linked together using the first word of the page. This guarantees that every packet transferred by the monitor's networking device using remaps will automatically trigger at least one copy-on-write fault on the receiving end. A simple change to the mbuf freelist data structure fixed this problem.

The kernel implementation of NFS always copies data from the incoming mbufs to the receiving file buffer cache, even when the packet contained un-fragmented, properly aligned pages. This would have effectively prevented the sharing of the file buffer cache across virtual machines. To have clients and servers transparently share the page, we specialized the call to *bcopy* to a new *remap* function offered by the HAL. This remap function has the semantics of a bcopy routine but uses a monitor call to remap the page whenever possible. Figure 4 shows how a data page transferred during an NFS read or write call is first remapped from the source virtual machine to the destination memory buffer (mbuf) page by the monitor's networking device, and then remapped into its final location by a call to the HAL's remap function.

4.4 SPLASHOS: A Specialized Operating System

The ability to run a thin or specialized operating system allows Disco to support large-scale parallel applications that span the entire machine. These applications may not be well served by a full function operating system. In fact, specialized operating systems such as Puma [24] are commonly used to run scientific applications on parallel systems.

To illustrate this point, we developed a specialized library operating system [11], "SPLASHOS", that runs directly on top of Disco. SPLASHOS contains the services needed to run SPLASH-2 applications [27]: thread creation and synchronization routines, "libc" routines, and an NFS client stack for file I/O. The application is linked with the library operating system and runs in the same address space as the operating system. As a result, SPLASHOS does not need to support a virtual memory subsystem, deferring all page faulting responsibilities directly to Disco.

Although one might find SPLASHOS to be an overly simplistic and limited operating system if it were to run directly on hardware, the ability to run it in a virtual machine alongside commodity operating systems offers a powerful and attractive combination.

5 Experimental Results

We have implemented Disco as described in the previous section and performed a collection of experiments to evaluate it. We describe our simulation-based experimental setup in Section 5.1. The first set of experiments presented in Sections 5.2 and 5.3 demonstrate that Disco overcomes the traditional problems associated with virtual machines, such as high overheads and poor resource sharing. We then demonstrate in Sections 5.4 and 5.5 the benefits of using virtual machines, including improved scalability and data locality.

5.1 Experimental Setup and Workloads

Disco targets the FLASH machine, which is unfortunately not yet available. As a result, we use the SimOS [22] machine simulator to develop and evaluate Disco. SimOS is a machine simulator that models the hardware of MIPS-based multiprocessors in enough detail to run essentially unmodified system software such as the IRIX operating system and the Disco monitor. For this study, we configured SimOS to resemble a large-scale multiprocessor with performance characteristics similar to FLASH. Although SimOS contains simulation models of FLASH's MIPS R10000 processors,

Workload	Environment	Description	Characteristics	Execution Time
Pmake	Software Development	Parallel compilation (-J2) of the GNU chess application	Multiprogrammed, short-lived, system and I/O intensive processes	3.9 sec
Engineering	Hardware Development	Verilog simulation (Chronologics VCS) + machine simulation	Multiprogrammed, long running processes	3.5 sec
Splash	Scientific Computing	Raytrace from SPLASH-2	Parallel applications	12.9 sec
Database	Commercial Database	Sybase Relational Database Server decision support workload	Single memory intensive process	2.0 sec

Table 1. Workloads. Each workload is scaled differently for the uniprocessor and multiprocessor experiments. The reported execution time is for the uniprocessor workloads running on IRIX without Disco. The execution time does not include the time to boot the operating, ramp-up the applications and enter a steady execution state. This setup time is at least two orders of magnitude longer and performed using SimOS's fast emulation mode.

these simulation models are too slow for the workloads that we chose to study. As a result, we model statically scheduled, non-superscalar processors running at twice the clock rate. These simpler pipelines can be modelled one order of magnitude faster than the R10000. The processors have the on-chip caches of the MIPS R10000 (32KB split instruction/data) and a 1MB board-level cache. In the absence of memory system contention, the minimum latency of a cache miss is 300 nanoseconds to local memory and 900 nanoseconds to remote memory.

Although SimOS allows us to run realistic workloads and examine their behavior in detail with its non-intrusive annotation mechanism, the simulation slowdowns prevent us from examining long running workloads in detail. Using realistic but short workloads, we were able to study issues like the CPU and memory overheads of virtualization, the benefits on scalability, and NUMA memory management. However, studies that would require long running workloads, such as those fully evaluating Disco's resource sharing policies, are not possible in this environment and will hence have to wait until we have a real machine.

Table 1 lists the workloads of this study together with their base simulated execution time. The workloads were chosen to be representative of four typical uses of scalable compute servers. Although the simulated execution times are small, the SimOS environment allowed us to study the workload's behavior in great detail and determine that the small execution regions exhibit similar behavior to longer-running worklaods. We also used the fast mode of SimOS to ensure that the workloads did not include any cold start effects.

5.2 Execution Overheads

To evaluate the overheads of running on Disco, we ran each workload on a uniprocessor, once using IRIX directly on the simulated hardware, and once using Disco running IRIX in a single virtual machine on the same hardware. Figure 5 shows this comparison. Overall, the overhead of virtualization ranges from a modest 3% for Raytrace to a high of 16% in the pmake and database workloads. For the compute-bound engineering and Raytrace workloads, the overheads are mainly due to the Disco trap emulation of TLB reload misses. The engineering and database workloads have an exceptionally high TLB miss rate and hence suffer large overheads. Nevertheless, the overheads of virtualization for these applications are less than 16%.

The heavy use of OS services for file system and process creation in the pmake workload makes it a particularly stressful workload for Disco. Table 2 shows the effect of the monitor overhead on the top OS services. From this table we see the overheads can significantly lengthen system services and trap handling. Short running services such as the IRIX quick page fault handler, where the trap overhead itself is a significant portion of the service, show slowdowns over a factor of 3. Even longer running services such as execve and open system call show slowdowns of 1.6.

These slowdowns can be explained by the common path

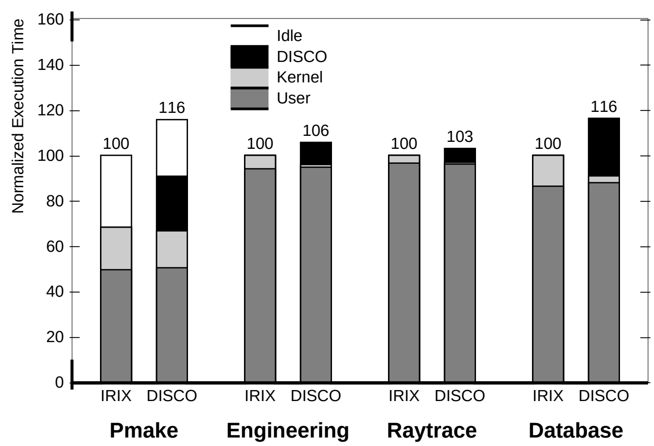

FIGURE 5. Overhead of Virtualization. The figure compares, for four uniprocessor workloads, the execution time when running IRIX directly on the simulated hardware with IRIX running in a Disco virtual machine. The execution time is separated between the time spent in user programs, the IRIX kernel, Disco, and the idle loop.

Operating System Service	% of System Time (IRIX)	Avg Time per Invocation (IRIX)	Slowdown on Disco	Relative Execution Time on Disco				
				Kernel Execution	TLB Write Emulation	Other Privileged Instructions	Monitor Calls & Page Faults	Kenrel TLB Faults
DEMAND_ZERO	30%	21 µs	1.42	0.43	0.21	0.16	0.47	0.16
QUICK_FAULT	10%	5 µs	3.17	1.27	0.80	0.56	0.00	0.53
open	9%	42 µs	1.63	1.16	0.08	0.06	0.02	0.30
UTLB_MISS	7%	0.035 µs	1.35	0.07	1.22	0.05	0.00	0.02
write	6%	12 µs	2.14	1.01	0.24	0.21	0.31	0.17
read	6%	23 µs	1.53	1.10	0.13	0.09	0.01	0.20
execve	6%	437 µs	1.60	0.97	0.03	0.05	0.17	0.40

Table 2. Service Breakdown for the Pmake workload. This table breaks down the overheads of the virtualization for the seven top kernel services of the pmake workload. DEMAND_ZERO is demand zero page fault, QUICK_FAULT, is slow TLB refill, UTLB_MISS is a fast TLB refill. Other than the UTLB_MISS service, the IRIX and IRIX on Disco configurations request the same number of services of each category. For each service, the execution time is expressed as a fraction of the IRIX time and separates the time spend in the kernel, emulating TLB writes and privileged instructions, performing monitor call and emulating the unmapped segments. The slowdown column is the sum of the relative execution times and measures the average slowdown for each service.

to enter and leave the kernel for all page faults, system calls and interrupts. This path includes many privileged instructions that must be individually emulated by Disco. A restructuring of the HAL of IRIX could remove most of this overhead. For example, IRIX uses the same TLB wired entry for different purposes in user mode and in the kernel. The path on each kernel entry and exit contains many privileged instructions that deal exclusively with this feature and are individually emulated.

We also notice the relatively high overhead of servicing kernel TLB-faults that occur since Disco runs IRIX in mapped addresses rather than the unmapped addresses used when running directly on the machine. This version of Disco only mapped 4KB page pairs into the TLB. The use of larger pages, supported by the MIPS TLB, could significantly reduce this overhead. Even with these large slowdowns, the operating system intensive pmake workload with its high trap and system call rate has an overhead of only 16%.

Figure 5 also shows a reduction in overall kernel time of some workloads. Some of the work of the operating system is being handled directly by the monitor. The reduction in pmake is primarily due to the monitor initializing pages on behalf of the kernel and hence suffering the memory stall and instruction execution overhead of this operation. The reduction of kernel time in Raytrace, Engineering and Database workloads is due to the monitor's second-level TLB handling most TLB misses.

5.3 Memory Overheads

To evaluate the effectiveness of Disco's transparent memory sharing and quantify the memory overheads of running multiple virtual machines, we use a single workload running under six different system configurations. The workload consists of eight different instances of the basic pmake workload. Each pmake instance reads and writes files from a different disk. In all configurations we use an eight processor machine with 256 megabytes of memory and ten disks.

The configurations differ in the number of virtual machines used and the access to the workload file systems. The first configuration (IRIX) runs IRIX on the bare hardware with all disks local. The next four configurations split the workload across one (1VM), two (2VMs), four (4VMs), and eight virtual machines (8VMs). Each VM has the virtual resources that correspond to an equal fraction of the physical resources. As a result, the total virtual processor and memory resources are equivalent to the total physical resources of the machine, i.e. eight processors and 256 MB of memory. For example, the 4VMs configuration consists of dual-processor virtual machines, each with 64 MB of memory. The root disk and workload binaries are mounted from copy-on-write disks and shared among all the virtual machines. The workload file systems are mounted from different private exclusive disks.

The last configuration runs eight virtual machines but accesses workload files over NFS rather than from private disks. One of the eight virtual machines also serves as the NFS server for all file systems and is configured with 96 megabytes of memory. The seven other virtual machines have only 32MB of memory. This results in more memory configured to virtual machines than is available on the real machine. This workload shows the ability to share the file cache using standard distributed system protocols such as NFS.

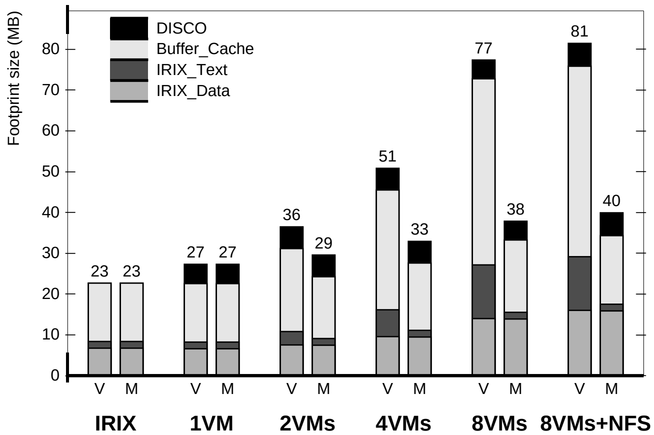

FIGURE 6. Data Sharing in Disco. This figure compares the memory footprints of the different configurations of Section 5.3 which run the pmake workload. For each configuration, "V" breaks down the virtual footprint of the system and "M" and actual machine memory footprint. The virtual footprint is equivalent to the amount of memory required in the absence of memory sharing optimizations.

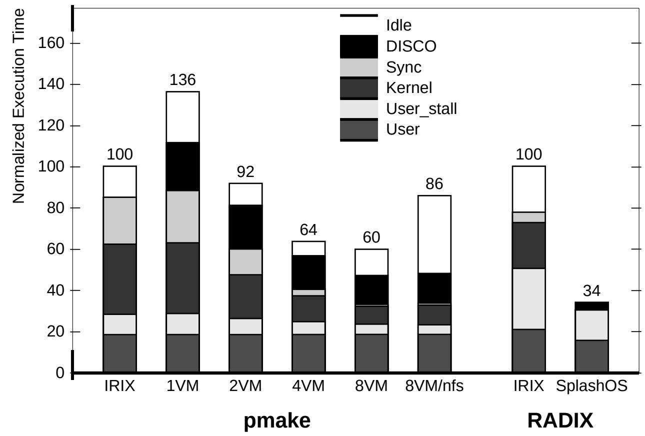

FIGURE 7. Workload Scalability Under Disco. The performance of the pmake and radix workloads on an eight-processor ccNUMA machine is normalized to the execution time running IRIX on the bare hardware. Radix runs on IRIX directly on top of the hardware and on a specialized OS (SPLASHOS) on top of Disco in a single virtual machine. For each workload the execution is broken down into user time, kernel time, time synchronization time, monitor time, and the idle loop. All configurations use the same physical resources, eight processors and 256MB of memory, but use a different number of virtual machines.

Figure 6 compares the memory footprint of each configuration at the end of the workload. The virtual physical footprint (V) is the amount of memory that would be needed if Disco did not support any sharing across virtual machines. The machine footprint (M) is the amount of memory actually needed with the sharing optimizations. Pages are divided between the IRIX data structures, the IRIX text, the file system buffer cache and the Disco monitor itself.

Overall, we see that the effective sharing of the kernel text and buffer cache limits the memory overheads of running multiple virtual machines. The read-shared data is kept in a single location in memory.

The kernel private data is however not shareable across virtual machines. The footprint of the kernel private data increases with the number of virtual machines, but remains overall small. For the eight virtual machine configuration, the eight copies of IRIX's data structures take less than 20 megabytes of memory.

In the NFS configuration, the virtual buffer cache is larger than the comparable local configuration as the server holds a copy of all workload files. However, that data is transparently shared with the clients and the machine buffer cache is of comparable size to the other configurations. Even using a standard distributed file system such as NFS, Disco can maintain a global buffer cache and avoid the memory overheads associated with multiple caching of data.

5.4 Scalability

To demonstrate the scalability benefits of using virtual machine monitors we ran the pmake workload under the six configurations described in the previous section. IRIX5.3 is not a NUMA-aware kernel and tends to allocate its kernel data structures from a single node of FLASH causing large hot-spots. To compensate for this, we changed the physical memory layout of FLASH so that machine pages are allocated to nodes in a round-robin fashion. This round-robin allocation eliminates hot spots and results in significantly better performance for the IRIX runs. Since Disco is NUMA-aware, we were able to use the actual layout of machine memory, which allocates consecutive pages to each node. To further simplify the comparison, we disabled dynamic page migration and replication for the Disco runs.

Figure 7 shows the execution time of each workload. Even at just eight processors, IRIX suffers from high synchronization and memory system overheads for system-intensive workloads such as this. For example, about one quarter of the overall time is spent in the kernel synchronization routines and the 67% of the remaining kernel time is spent stalled in the memory system on communication misses. The version of IRIX that we used has a known primary scalability bottleneck, *memlock*, the spinlock that protects the memory management data structures of IRIX [23]. Other operating systems such as NT also have comparable scalability problems, even with small numbers of processors [21].

Using a single virtual machine leads to higher overheads than in the comparable uniprocessor Pmake workload. The increase is primarily due to additional idle time. The execution of the operating system in general and of the critical regions in particular is slower on top of Disco, which increases the contention for semaphores and spinlocks in the operating system. For this workload, the increased idle time is due to additional contention on certain semaphores that protect the virtual memory subsystem of IRIX, forcing more processes to be descheduled. This interaction causes a non-linear effect in the overheads of virtualization.

However, partitioning the problem into different virtual machines significantly improves the scalability of the system. With only two virtual machines, the scalability benefits already outweigh the overheads of the virtualization. When using eight virtual machines, the execution time is reduced to 60% of its base execution time, primarily because of a significant reduction in the kernel stall time and kernel synchronization.

We see significant performance improvement even when accessing files using NFS. In the NFS configuration we see an increase in the idle time that is primarily due to the serialization of NFS requests on the single server that manages all eight disks. Even with the overheads of the NFS protocol and the increase in idle time, this configuration executes faster than the base IRIX time.

The other workload of Figure 7 compares the performance of the radix sorting algorithm, one of the SPLASH-2 applications [27]. Radix has an unfortunate interaction with the lazy evaluation policies of the IRIX virtual memory system. IRIX defers setting up the page table entries of each parallel thread until the memory is touched by the thread. When the sorting phase starts, all threads suffer many page faults on the same region causing serialization on the various spinlocks and semaphores used to protect virtual memory data structures. The contention makes the execution of these traps significant in comparison to the work Radix does for each page touched. The result is Radix spends one half of its time in the operating system.

Although it would not have been difficult to modify Radix to setup its threads differently to avoid this problem, other examples are not as easy to fix. Rather than modifying Radix, we ran it on top of SPLASHOS rather than IRIX. Because it does not manage virtual memory, SPLASHOS does not suffer from the same performance problems as IRIX. Figure 7 shows the drastic performance improvements of running the application in a specialized operating system (on top of Disco) over using a full-blown operating system (without Disco). Both configurations suffer from the same number of page faults, whose processing accounts for most of the system time. This system time is one order of magnitude larger for IRIX than it is for SPLASHOS on top of Disco. The NUMA-aware allocation policy of Disco also reduces hot spots and improves user stall time.

5.5 Dynamic Page Migration and Replication

To show the benefits of Disco's page migration and replication implementation, we concentrate on workloads that exhibit poor memory system behavior, specifically the Engineering and Raytrace workloads. The Engineering workload consists of six Verilog simulations and six memory system simulations on eight processors of the same virtual machine. The Raytrace workload spans 16 processors. Because Raytrace's largest available data set fully fits in a 1MB cache, we ran the Raytrace experiments with a 256KB cache to show the impact of data locality.

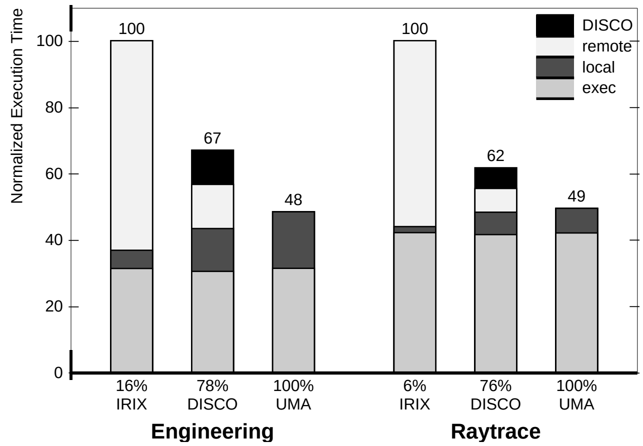

FIGURE 8. Performance Benefits of Page Migration and Replication. For each workload, the figure compares the execution time of IRIX on NUMA, IRIX on Disco on NUMA with page migration and replication, and IRIX on an bus-based UMA, The execution time is divided between instruction execution time, local memory stall time, remote memory stall time, and Disco overhead. The percentage of cache misses satisfied locally is shown below each bar.

Figure 8 shows the overall reduction in execution time of the workload. Each workload is run under IRIX, IRIX on Disco with migration and replication, and IRIX on a UMA memory system. The UMA memory system has a latency of 300ns equivalent to the local latency of the NUMA machine. As a result, the performance on the UMA machine determines a lower bound for the execution time on the NUMA machine. The comparison between Disco and the NUMA IRIX run shows the benefits of page migration and replication while the comparison with the UMA IRIX run shows how close Disco got to completely hiding the NUMA memory system from the workload.

Disco achieves significant performance improvements by enhancing the memory locality of these workloads. The Engineering workload sees a 33% performance improvement while Raytrace gets a 38% improvement. Both user and kernel modes see a substantial reduction in remote stall time. Disco increases data locality by satisfying a large fraction of the cache misses from local memory with only a small increase in Disco's overhead.

Although Disco cannot totally hide all the NUMA memory latencies from the kernel, it does greatly improve the situation. Comparing Disco's performance with that of the optimistic UMA where all cache misses are satisfied in 300 nanoseconds, Disco comes within 40% for the Engineering

Action	Engineering		Raytrace	
	num / sec	avg time	num / sec	avg time
Migration	2461	67 µs	909	102 µs
Replication	2208	57 µs	2671	73 µs

Table 3. Action taken on hot pages. This table shows the number of migrations and replications per second and their average latency for the two workloads.

workload and 26% for Raytrace.

Our implementation of page migration and replication in Disco is significantly faster than a comparable kernel implementation [26]. This improvement is due to Disco's streamlined data structures and optimized TLB shootdown mechanisms. As a result, Disco can be more aggressive in its policy decisions and provide better data locality. Table 3 lists the frequency and latency of page migrations and replications for both workloads.

6 Related Work

We start by comparing Disco's approach to building system software for large-scale shared-memory multiprocessors with other research and commercial projects that target the same class of machines. We then compare Disco to virtual machine monitors and to other system software structuring techniques. Finally, we compare our implementation of dynamic page migration and replication with previous work.

6.1 System Software for Scalable Shared Memory Machines

Two opposite approaches are currently being taken to deal with the system software challenges of scalable shared-memory multiprocessors. The first one is to throw a large OS development effort at the problem and effectively address these challenges in the operating system. Examples of this approach are the Hive [5] and Hurricane [25] research prototypes and the Cellular-IRIX system recently announced by SGI. These multi-kernel operating systems handle the scalability of the machine by partitioning resources into "cells" that communicate to manage the hardware resources efficiently and export a single system image, effectively hiding the distributed system from the user. In Hive, the cells are also used to contain faults within cell boundaries. In addition, these systems incorporate resource allocators and schedulers for processors and memory that can handle the scalability and the NUMA aspects of the machine. This approach is innovative, but requires a large development effort.

The virtual machines of Disco are similar to the cells of Hive and Cellular-IRIX in that they support scalability and form system software fault containment boundaries. Like these systems, Disco can balance the allocation of resources such as processors and memory between these units of scalability. Also like these systems, Disco handles the NUMA memory management by doing careful page migration and replication. The benefit of Disco over the OS intensive approach is in the reduction in OS development effort. It provides a large fraction of the benefits of these systems at a fraction of the cost. Unlike the OS-intensive approach that is tied to a particular operating system, Disco is independent of any particular OS, and can even support different OSes concurrently.

The second approach is to statically partition the machine and run multiple, independent operating systems that use distributed system protocols to export a partial single system image to the users. An example of this approach is the Sun Enterprise10000 machine that handles software scalability and hardware reliability by allowing users to hard partition the machine into independent failure units each running a copy of the Solaris operating system. Users still benefit from the tight coupling of the machine, but cannot dynamically adapt the partitioning to the load of the different units. This approach favors low implementation cost and compatibility over innovation.

Like the hard partitioning approach, Disco only requires minimal OS changes. Although short of providing a full single system image, both systems build a partial single system image using standard distributed systems protocols. For example, a single file system image is built using NFS. A single system administration interface is built using NIS. System administration is simplified in Disco by the use of shared copy-on-write disks that are shared by many virtual machines.

Yet, unlike the hard partitioning approach, Disco can share all the resources between the virtual machines and supports highly dynamic reconfiguration of the machine. The support of a shared buffer cache and the ability to schedule all the resources of the machine between the virtual machines allows Disco to excel on workloads that would not perform well with a relatively static partitioning. Furthermore, Disco provides the ability for a single application to span all resources of the machine using a specialized scalable OS.

Digital's announced Galaxies operating system, a multi-kernel version of VMS, also partitions the machine relatively statically like the Sun machine, with the additional support for segment drivers that allow applications to share memory across partitions. Galaxies reserves a portion of the physical memory of the machine for this purpose. A comparable implementation of application-level shared memory between virtual machines would be simple and would not require having to reserve memory in advance.

Disco is a compromise between the OS-intensive and the OS-light approaches. Given an infinite OS development budget, the OS is the right place to deal with issues such as resource management. The high-level knowledge and greater control available in the operating system can allow it to export a single system image and develop better resource management mechanisms and policies. Fortunately, Disco is capable of gradually getting out of the way as the OS improves. Operating systems with improved scalability can just request larger virtual machines that manage more of the machine's resources. Disco provides an adequate and low-cost solution that enables a smooth transition and maintains compatibility with commodity operating systems.

6.2 Virtual Machine Monitors

Disco is a virtual machine monitor that implements in software a virtual machine identical to the underlying hardware. The approach itself is far from being novel. Golberg's 1974 survey paper [13] lists over 70 research papers on the topic and IBM's VM/370 [15] system was introduced in the same period. Disco shares the same approach and features as these systems, and includes many of the same performance optimizations as VM/370 [8]. Both allow the simultaneous execution of independent operating systems by virtualizing all the hardware resources. Both can attach I/O devices to single virtual machines in an exclusive mode. VM/370 mapped virtual disks to distinct volumes (partitions), whereas Disco has the notion of shared copy-on-write disks. Both systems support a combination of persistent disks and temporary disks. Interestingly, Creasy argues in his 1981 paper that the technology developed to interconnect virtual machines will later allow the interconnection of real machines [8]. The opposite occurred and Disco benefits today from the advances in distributed systems protocols.

The basic approach used in Disco as well as many of its performance optimizations were present in VM/370 and other virtual machines. Disco differs in its support of scalable shared-memory multiprocessors, handling of modern operating systems, and the transparent sharing capabilities of copy-on-write disks and the global buffer cache.

The idea of virtual machines remains popular to provide backward compatibility for legacy applications or architectures. Microsoft's Windows 95 operating system [16] uses virtual machines to run older Windows 3.1 and DOS applications. Disco differs in that it runs all the system software in a virtual machine and not just the legacy applications. DAISY [10] uses dynamic compilation techniques to run a single virtual machine with a different instruction set architecture than the host processor. Disco exports the same instruction set as the underlying hardware and can therefore use direct execution rather than dynamic compilation.

Virtual machine monitors have been recently used to provide fault-tolerance to sensitive applications [3]. Bressoud and Schneider's system virtualizes only certain resources of the machine, specifically the interrupt architecture. By running the OS in supervisor mode, it disables direct access to I/O resources and physical memory, without having to virtualize them. While this is sufficient to provide fault-tolerance, it does not allow concurrent virtual machines to coexist.

6.3 Other System Software Structuring Techniques

As an operating system structuring technique, Disco could be described as a microkernel with an unimaginative interface. Rather than developing the clean and elegant interface used by microkernels, Disco simply mirrors the interface of the raw hardware. By supporting different commodity and specialized operating systems, Disco also shares with microkernels the idea of supporting multiple operating system personalities [1].

It is interesting to compare Disco with Exokernel [11], a software architecture that allows application-level resource management. The Exokernel safely multiplexes resources between user-level library operating systems. Both Disco and Exokernel support specialized operating systems such as ExOS for the Aegis exokernel and SplashOS for Disco. These specialized operating systems enable superior performance since they are freed from the general overheads of commodity operating systems. Disco differs from Exokernel in that it virtualizes resources rather than multiplexes them, and can therefore run commodity operating systems without significant modifications.

The Fluke system [12] uses the virtual machine approach to build modular and extensible operating systems. Recursive virtual machines are implemented by their nested process model, and efficiency is preserved by allowing inner virtual machines to directly access the underlying microkernel of the machine. Ford et al. show that specialized system functions such as checkpointing and migration require complete state encapsulation. Like Fluke, Disco totally encapsulates the state of virtual machines, and can therefore trivially implement these functions.

6.4 ccNUMA Memory Management

Disco provides a complete ccNUMA memory management facility that includes page placement as well as a dynamic page migration and page replication policy. Dynamic page migration and replication was first implemented in operating systems for machines that were not cache-coherent, such as the IBM Ace [2] or the BBN Butterfly [7]. In these systems, migration and replication is triggered by page faults and the penalty of having poor data locality is greater due to the absence of caches.

The implementation in Disco is most closely related to our kernel implementation in [26]. Both projects target the FLASH multiprocessor. Since the machine supports cache-coherency, page movement is only a performance optimization. Our policies are derived from this earlier work. Unlike the in-kernel implementation that added NUMA awareness to an existing operating system, our implementation of Disco was designed with these features in mind from the beginning, resulting in lower overheads.

7 Conclusions

This paper tackles the problem of developing system software for scalable shared memory multiprocessors without a massive development effort. Our solution involves adding a level of indirection between commodity operating systems and the raw hardware. This level of indirection uses another old idea, virtual machine monitors, to hide the novel aspects of the machine such as its size and NUMA-ness.

In a prototype implementation called Disco, we show that many of the problems of traditional virtual machines are no longer significant. Our experiments show that the overheads imposed by the virtualization are modest both in terms of processing time and memory footprint. Disco uses a combination of innovative emulation of the DMA engine and standard distributed file system protocols to support a global buffer cache that is transparently shared across all virtual machines. We show how the approach provides a simple solution to the scalability, reliability and NUMA management problems otherwise faced by the system software of large-scale machines.

Although Disco was designed to exploit shared-memory multiprocessors, the techniques it uses also apply to more loosely-coupled environments such as networks of workstations (NOW). Operations that are difficult to retrofit into clusters of existing operating systems such as checkpointing and process migration can be easily supported with a Disco-like monitor. As with shared-memory multiprocessors, this can be done with a low implementation cost and using commodity operating systems.

This return to virtual machine monitors is driven by a current trend in computer systems. While operating systems and application programs continue to grow in size and complexity, the machine-level interface has remained fairly simple. Software written to operate at this level remains simple, yet provides the necessary compatibility to leverage the large existing body of operating systems and application programs. We are interested in further exploring the use of virtual machine monitors as a way of dealing with the increasing complexity of modern computer systems.

Acknowledgments

The authors would like to thank John Chapin, John Gerth, Mike Nelson, Rick Rashid, Steve Ofsthun, Volker Strumpen, and our shepherd Rich Draves for their feedback. Our colleagues Kinshuk Govil, Dan Teodosiu, and Ben Verghese participated in many lively discussions on Disco and carefully read drafts of the paper.

This study is part of the Stanford FLASH project, funded by ARPA grant DABT63-94-C-0054. Ed Bugnion is supported in part by an NSF Graduate Research Fellowship. Mendel Rosenblum is partially supported by an NSF Young Investigator Award.

References

[1] Michael J. Accetta, Robert V. Baron, William J. Bolosky, David B. Golub, Richard F. Rashid, Avadis Tevananian, and Michael Young. Mach: A New Kernel Foundation for UNIX development. In *Proceedings of the Summer 86 USENIX Conference*. pp. 99-112. Jun. 86.

[2] William J. Bolosky, Robert P. Fitzgerald, and Michael L. Scott. Simple But Effective Techniques for NUMA Memory Management. In *Proceedings of the 12th Symposium on Operating Systems Principles (SOSP)* pp. 18-31. Dec. 1989.

[3] Thomas C. Bressoud and Fred B. Schneider. Hypervisor-based Fault-tolerance. In *Proceedings of the 15th Symposium on Operating Systems Principles (SOSP)*. pp. 1-11. Dec. 1995.

[4] Tony Brewer and Greg Astfalk. The evolution of the HP/Convex Exemplar. In *Proceedings of COMPCON Spring '97*. pp. 81-96. 1997

[5] John Chapin, Mendel Rosenblum, Scott Devine, Tirthankar Lahiri, Dan Teodosiu, and Anoop Gupta. Hive: Fault containment for shared-memory Multiprocessors. In *Proceedings of the 15th Symposium on Operating Systems Principles (SOSP),* pp. 12-25. Dec. 1995.

[6] Thomas H. Cormen, Charles E. Leiserson and Ronald L. Rivest. Introduction to Algorithms. McGraw-Hill. 1990.

[7] Alan L. Cox and Robert J. Fowler. The Implementation of a Coherent Memory Abstraction on a NUMA Multiprocessor: Experiences with Platinum. In *Proceedings of the 12th Symposium on Operating Systems Principles (SOSP),* pp. 32-44. Dec. 1989.

[8] R. J. Creasy. The Origin of the VM/370 Time-Sharing System. *IBM J. Res. Develop* 25(5) pp. 483-490, 1981.

[9] Helen Custer. Inside Windows NT. Microsoft Press. 1993.

[10] Kermal Ebcioglu and Erik R. Altman. DAISY: Dynamic Compilation for 100% Architectural Compatibility. In *Proceedings of the 24th International Symposium on Computer Architecture (ISCA)*. pp. 26-37. Jun. 1997.

[11] Dawson R. Engler, M. Frans Kaashoek, and J. O'Toole Jr. Exokernel: An Operating System Architecture for Application-level Resource Management. In *Proceedings of the 15th Symposium on Operating Systems Principles (SOSP)* pp. 251-266. Dec. 1995.

[12] Bryan Ford, Mike Hibler, Jay Lepreau, Patrick Tullmann, Godmar Back, Stephen Clawson. Microkernels meet Recursive Virtual Machines. In *Proceedings of the 2nd Symposium on Operating System Design and Implementation (OSDI)*. pp. 137-151. Oct. 1996.

[13] Robert P. Goldberg. Survey of Virtual Machine Research. *IEEE Computer Magazine* 7(6), pp. 34-45, Jun. 1974.

[14] Maurice Herlihy. Wait-free synchronization. In *ACM Transactions on Programming Languages and Systems (TOPLAS)* 13(1) pp. 124-149. Jan. 1991.

[15] IBM Corporation. IBM Virtual Machine /370 Planning Guide. 1972.

[16] Adrian King. Inside Windows 95, Microsoft Press, 1995.

[17] Jeffrey Kuskin, David Ofelt, Mark Heinrich, John Heinlein, Richard Simoni, Kourosh Gharachorloo, John Chapin, David Nakahira, Joel Baxter, Mark Horowitz, Anoop Gupta, Mendel Rosenblum, and John Hennessy. The Stanford FLASH Multiprocessor. In *Proceedings of the 21st International Symposium on Computer Architecture (ISCA)*, pp. 302-313, Apr. 1994.

[18] Jim Laudon and Daniel Lenoski. The SGI Origin: A ccNUMA Highly Scalable Server. In *Proceedings of the 24th International Symposium on Computer Architecture (ISCA)*. pp. 241-251. Jun. 1997.

[19] Tom Lovett and Russel Clapp. STiNG: A CC-NUMA Computer System for the Commercial Marketplace. In *Proceedings of the 23rd International Symposium on Computer Architecture (ISCA)*. pp. 308-317. Jun. 1996.

[20] Mike Perez, Compaq Corporation. Interview "Scalable hardware evolves, but what about the network OS?" PCWeek. Dec. 1995.

[21] Sharon E. Perl and Richard L. Sites. Studies of Windows NT using Dynamic Execution Traces. In *Proceedings of the 2nd Symposium on Operating System Design and Implementation (OSDI)*, pp. 169-183. Oct. 1996.

[22] Mendel Rosenblum, Edouard Bugnion, Scott Devine and Steve Herrod. Using the SimOS Machine Simulator to study Complex Computer Systems, *ACM Transactions on Modelling and Computer Simulation (TOMACS)*, 7(1), pp. 78-103. Jan. 1997.

[23] Mendel Rosenblum, Edouard Bugnion, Steven A. Herrod, Emmett Witchel and Anoop Gupta. The Impact of Architectural Tends on Operating System Performance. In *Proceedings of the 15th Symposium on Operating Systems Principles (SOSP)*, pp. 285-298. Dec. 1995.

[24] Lance Shuler, Chu Jong, Rolf Riesen, David van Dresser, A. B. Maccabe, L.A. Fisk and T.M. Stallcup. The Puma Operating System for Massively Parallel Computers. In *Proceedings of the Intel Supercomputer User Group Conference,* 1995.

[25] Ron Unrau, Orran Krieger, Benjamin Gamsa and Michael Stumm. Hierarchical Clustering: A Structure for Scalable Multiprocessor Operating System Design. *Journal of Supercomputing*, 9(1), pp. 105-134. 1995.

[26] Ben Verghese, Scott Devine, Anoop Gupta, and Mendel Rosenblum. Operating System Support for Improving Data Locality on CC-NUMA Compute Servers. In *Proceedings of the 7th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pp. 279-289. Oct. 1996.

[27] Steven Cameron Woo, Moriyoshi Ohara, Evan Torrie, Jaswinder Pal Singh, and Anoop Gupta. The SPLASH-2 programs: Characterization and Methodological Considerations. In *Proceeedings of the 22nd Annual International Symposium on Computer Architecture (ISCA)*, pp. 24-36. May 1995.