Implementing Operating Systems without Kernels [1]

Dave Probert and John Bruno

Computer Science Department
University of California
Santa Barbara, California 93106

contact: probert@cs.ucsb.edu

UCSB Computer Science — TRCS95–24 — December 1995

[1]This research is supported in part by a UC MICRO grant and the Xerox Corporation.

Implementing Operating Systems without Kernels

Introduction

Most contemporary operating systems are structured around a central monolithic kernel (or micro-kernel). An alternative is to construct operating systems without a kernel. In this approach, operating system abstractions (processes, threads, virtual memory, communication) are implemented as application-level code, with the attendant flexibility and extensibility that many user applications require.

We are building such a kernel-less operating system, using optimized cross-domain calls which can reduce cross-domain invocations cost below the current cost of kernel invocations by exploiting properties that kernel-less operating systems have with respect to robustness, extensibility, security, and performance. The first three of these properties are generally accepted as benefits of the kernel-less approach. One of the primary contributions of this work is to achieve these benefits while also achieving gains in performance.

In previous work[1, 2] we have established a model for building efficient cross-domain mechanisms, which we call SPACE. SPACE achieves its efficiency through a generalized exception model that includes dynamic synthesis of custom cross-domain calls, capability-based accessing of system services, and co-location of hardware protection domains.

The goal of this project is to evaluate SPACE when applied to building kernel-less system services to support a real application development environment. In particular we propose to investigate our approach when applied to parallel computing environments. This is an area where there are critical performance and functional requirements that are unmet by existing systems.

1 Operating Systems Approaches

Operating system design is generally inseparable from hardware architecture. The success of systems like UNIX in the 1970s has brought about a convergence of processor architectures. Most recent machines support multiple paged address spaces, user/supervisor processor modes, and memory mapped I/O. Systems based on segmentation, hardware capabilities or multiple rings of protection are no longer feasible. The hardware provides domains based on page-level protection within flat address spaces. These domains are used to build process address spaces. They are also used to structure operating systems, as we explore in the following.

1.1 Kernel-based Operating Systems

Most operating systems are constructed around a central monolithic piece of software which controls the machine at the lowest levels. This software is called the *kernel* or *micro-kernel*, depending on the level of the abstractions it implements.

There are two primary advantages of employing a kernel. First, details of the hardware architecture can be encapsulated and mapped into the appropriate operating system abstractions. Secondly, the set of heavily-used primitive abstractions deemed critical can be insolated in the kernel and be thoroughly tested and tuned for performance.

Evolving application requirements are turning the advantages of kernel-based design into liabilities. Applications require direct access to low-level hardware for networking and graphics. Even seemingly primitive kernel mechanisms, such as threads and address space management, may be too general and thus too expensive to satisfy applications driven by fine-grain events (e.g. parallel machine communication and *futures* mechanisms). Novel uses of memory management hardware have been proposed for garbage collection and pointer swizzling, but are only feasible if overheads are extremely low.

The limitations imposed by kernel-based structures are exactly the flip-side of their advantages. Details of the hardware architecture often need to be encapsulated in an application-specific way. Secondly, the primitive operating system abstractions provided by the kernel are not suited for every application. Applications with performance or functional requirements not served by existing abstractions must be able to replace them with new ones, not just build on top of the old ones.

1.2 Kernel-less Operating Systems

Using the protection domains in processor architectures that support kernel-based systems, it is possible to build an operating system without relying on a central kernel. What is required is some simple extensions to the hardware trap mechanisms. Before discussing these mechanisms, we first examine some of the goals of the kernel-less approach:

- **Robustness** The more complex software is, the more likely it is to have bugs. This rule also applies to kernels. The problem is that bugs in the kernel are fatal in most systems. Bugs in services outside the kernel are much less likely to bring down the *entire* system, and can usually be dealt with by restarting the daemon or subsystem.

  If kernel-based services are implemented in application code, robustness will similarly improve. When a particular service crashes, it can be more easily restarted. Multiple implementations of a particular type of service can coexist, further insulating the entire system from the failure of one component.

- **Extensibility** Support for multiple implementations of a particular service is critical for extensibility. Systems which only provide extensibilty through refinement can only deliver functionality, not performance.

  Allowing multiple implementation instances to coexist allows applications to replace a service with one better suited, or even specifically tailored to their requirements. A single application may take advantage of multiple implementations (e.g. different virtual memory implementations for different regions of addresses).

  Coexistence is also desirable when developing services, because it is no longer necessary to risk breaking a working service while testing out the next version.

- **Security** Extensibility requires a tailorable interface to system services. Kernel-based systems typically export a single *system call interface* to all applications. Verification of access rights is handled internally to the services. Because only a fixed set of kernel services are available, there has not been the incentive to provide customizable interfaces.

  System call interfaces are implemented by mapping exceptions (i.e. trap instructions) into protected calls on system services. Such an interface can be customized on a per-application basis. Access to a particular cross-domain call represents a capability, analogous to the capabilities to access memory pages managed by the hardware memory management unit (MMU).

  These types of capabilities provide a run-time alternative to schemes based on uid or other authentication checks. There are also performance advantages to checking access when the capability is granted rather than every time the service is accessed.

- **Performance** Performance is *the* critical goal. Application-based extensibility and capability-based security contribute to this goal. But performance also benefits from the role extensibility plays in furthering operating systems research. New ideas can be tested side-by-side with old ideas. The kernel-less approach diminishes the need to build a new system from scratch or heavily cannibalize an existing system when experimenting with new systems ideas.

  To achieve performance requires that cross-domain invocations of services can be performed cheaply. Performance is a primary goal of many other operating system research projects which we will examine in the next section. Many of these systems have attempted to minimize the number for domain crossings rather than the cost of domain crossings.

  In sections 4 and 5 we will present our approach to domain crossing which allows cross-domain invocations to be performed even more efficiently than kernel invocations are today.

2 Related Work

Interest in operating systems research has accelerated in recent years, with large numbers of projects underway[3]. Increasing application requirements for efficient system services is in conflict with the overhead associated with traditional operating system implementation strategies.

The primary approach to resolving this conflict is to allow applications to build custom operating system services. A common strategy is to allow applications to dynamically load systems code into a central kernel. The Cache Kernel[4] defines low-level system objects: address spaces, threads, and kernels. These objects can be loaded into the kernel which will execute the performance critical parts. The SPIN Operating System[5] allows application code written in Modula 3 to be linked into the kernel, using language and loader mechanisms to build software protection domains. A similar approach is taken by Bridge[6] using Software Fault Isolation[7] to allow dynamic linking of application code written in standard languages, like C.

The Flexible Mach[8] effort and Mach development at OSF[9] are co-locating operating system services into applications, at user-level and in the kernel, depending on protection requirements. A more limited strategy is adopted by Protected Shared Libraries[10], which uses protected memory regions to build co-located domains for implementing operating system services using segmentation features of the RS/6000 architecture.

An alternative to implementing new system services in the kernel is to streamline the services that are already there through dynamic code synthesis. Synthesis[11] and Synthetix[12] takes this approach, as does Scout[13].

All these approaches still maintain a central kernel (or microkernel). One of the primary considerations in their design is minimizing the crossing of hardware protection boundaries. Cross-domain calls have been a critical issue in the performance of previous systems. However evolving hardware architectures have reduced the cost of cross-domain calls to the same order as system calls to monolithic kernels. Dynamic kernel approaches gain their performance by reducing the number of cross-domain calls at the expense of somewhat lower performance within the kernel, due to the increased abstraction overhead required to support software protection. As cross-domain calls become less expensive, the increased abstraction expense reduces the performance that can be gained. When these techniques are applied to existing applications written to a specific operating system API, the number of cross-domain calls becomes difficult to reduce and the kernel abstraction costs reduce performance. Centralized kernels represent centralized abstractions, which become imposed on all system services.

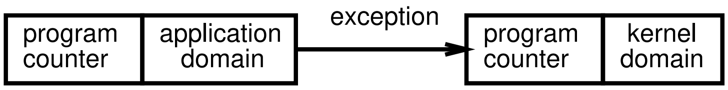
The Exokernel approach is being developed to specifically address the problem of burdening applications with a rigid set of abstractions maintained by the kernel. Originally the Exokernel[14] took a very similar approach to the Cache Kernel, but more recent work indicates that its developers are moving toward the kernel-less model that we have been advocating[1], though with less emphasis on mechanisms realizable in hardware.

This is the fundamental difference between our approach and the Exokernel and Cache Kernel approaches. The Cache Kernel provides abstractions that seem beneficial to operating systems, namely the caching of operating system objects. In our approach the SPACE abstractions are generalizations of the available hardware trap mechanisms. With continued evolution of processor architectures, our generalized framework is realizable in the hardware itself. The abstractions in the Exokernel and Cache Kernel will likely continue to require software or microcode implementation.

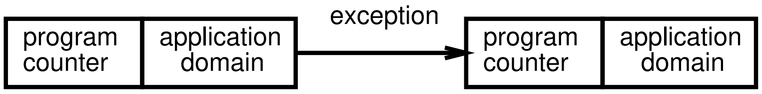
3 Building Kernel-less O/S Abstractions

In this section we describe how some of the common operating system abstractions, typically found in kernels, can be implemented as applications using generalized hardware trap handling.

Invocation of kernel services requires that a program generate a exception, e.g. executing a *trap instruction*. The hardware trap architecture changes the processor mode to *supervisor*, saves minimal state and performs an indirect branch through the trap vector table.



To implement a cross-domain mechanism requires a slight generalization of this hardware mechanism:



Instead of handling the exception in the kernel domain, a handler is invoked in another application domain. For architectures that support tagged TLBs, this difference is whether the supervisor mode bit is set in the processor state register or a new context id is inserted into the MMU context register.

We refer to the mapping from an exception in one domain to a handler in another a *portal*. A *cross-domain call* is then defined as *passing through* a portal.

In later sections we will examine the internal details of portals and cross-domain calls, particularly addressing performance and implementation issues. There we will address issues of how and when state is saved, how hardware exceptions are demultiplexed, and how the trap vector table is generalized. For now we assume that the internal details of cross-domain calls are represented by defining a set of portal types:

- **simple-trap** The previous *program counter* (PC), *MMU context id* (CTXID) and hardware *trap vector address* (TRAPVECTOR) are saved in a LIFO queue, and the hardware is updated with a new PC, CTXID, and TRAPVECTOR. Other register state must be saved in the new domain. Parameters are passed based on the nature of the exception.
- **resume** The PC, CTXID, and TRAPVECTOR at the head of the LIFO are restored. Depending on the exception the last instruction is re-executed or skipped.
- **suspend** Like a simple-trap, except that the processor registers are saved in the current LIFO which is itself then saved in a table and an empty LIFO is established. A token is passed to the handler which can be used to retrieve the saved LIFO.
- **unsuspend** The current LIFO is discarded and replaced by the saved LIFO identified by a token passed as an argument. The previous state is restored.

We have implemented these portal types on a 40MHz SuperScalar SPARC, and measured the performance. To pass through a portal of type *simple* and then back through again (*simple-resume*) to the original domain takes $1.6\mu S$, over half of which is due to hardware trap overhead. To do pass through a series of portals and then returning takes $2.5\mu S$ per portal. The suspend/unsuspend pair of portal types require $4.9\mu S$ to pass through and immediately return. In comparison Solaris 2.3 running on the same hardware requires $9.0\mu S$ to perform a null-system call (*getpid*). These initial measurements suggest that the performance of a kernel-less operating system is able to exceed the performance of a kernel-based system on the same hardware. In a later section we will examine the reasons for higher costs in kernel-based systems, and also explore the applicability of these observations to other processor architectures.

3.1 A Simple Kernel-Less Thread Mechanism

To build a thread mechanism from the simple cross-domain invocation uses the *suspend* portal type, which allows a domain to be entered from one chain of cross-domain calls, yet resume a different one. Each of the LIFO queues used to implement the generalized trap mechanism represent a history of cross-domain calls. Portals of the *suspend* portal type switch to a new LIFO, while saving the previous one where it can later be retrieved by the portals of type *unsuspend*. A token is associated with each saved LIFO, both for identification and authentication of the domain that later resumes the suspended LIFO.

Using the suspend portal type we can build thread operations which execute in an application domain, for instance consider the handlers for thread create and suspend:

```
thread_create(portal,token)
    ENQUEUE(token)                        ! suspends calling thread
    INVOKE(portal)                        ! invokes 'portal' as a new thread
  // reached when the invoked domain executes the instruction: trap T_RESUME
    INVOKE T_UNSUSPEND(DEQUEUE()) ! invokes unsuspend version of RESUME
  // reached only on error

thread_suspend(token)
    ENQUEUE(token)                        ! suspends calling thread
    INVOKE T_UNSUSPEND(DEQUEUE()) ! invokes unsuspend version of RESUME
```

Each create/exit of a thread takes 7.3μS on our SPARC platform. The ability to invoke threads across protection domains comes essentially for free. There are more complicated versions of the thread mechanisms that can be built (e.g. with support for signal mechanisms and priority scheduling). Applications using more complex versions will pay a commensurate higher price, but applications that do not require additional functionality can continue to use the simpler fast-thread mechanism detailed here.

We measured trivial thread create/exit operations under Solaris.

```
null_thread() {}

...
thr_create(stack, sizeof(stack), null_thread, 0, 0, 0)
...
```

Each create/exit of a thread takes 512μS in Solaris, 50 times slower than in the kernel-less implementation, and the Solaris threads execute within only a single protection domain.

The extra cost in Solaris is due to the need to support some external abstractions related to kernel scheduling and locking. The Solaris threads mechanism also incorporates many more features, which involve some cost even when they are not used. A simpler thread package could be implemented in Solaris, though to approach the performance of our simple kernel-less threads would require significant modifications to the Solaris kernel.

3.2 Kernel-less Virtual Memory Abstractions

Most processor architectures support virtual memory using in-memory data structures to represent the mappings between *virtual addresses* and *physical addresses*. For performance reasons these mappings are usually cached in a *memory management unit* (MMU), using *translation look-aside buffers* (TLBs). To avoid flushing the MMU when switching between address spaces, (e.g. processes) TLB entries are often tagged with an MMU context identifier (CTXID).

There is great variation between processors in the implementation details of these structures, particularly in the memory representations of page maps, the management of TLBs, and protection. Because these structures are primarily designed to implement protection domains for processes and to support virtual memory, the details are typically encapsulated in several layers of kernel code which adapt the hardware abstractions to the desired operating system abstractions[15].

Performance of virtual memory can be highly dependent on page replacement policies. Some virtual memory abstractions give applications hooks to control or influence these policies[16]. These extensions can be used to build new types of virtual memory at the application level, but only allow coarse grain access to the hardware paging mechanisms.

However virtual memory is not the only interesting use of the processor memory management structures. Appel & Li[17] have examined a number of other uses for virtual memory primitives, including concurrent garbage collection and persistent stores. To effectively use the MMU in these applications requires efficient access to these primitives without incurring the overhead of the operating system abstractions.

We measured an operation, suggested by the Appel & Li paper, in which an application causes a page protection fault and invokes an application-level handler which fixes the protection on the faulting page, and then protects a second page and returns. Under Solaris this is implemented with *sigaction* and *mprotect*. The time from the fault to the successful execution of the memory reference took $430\mu S$ under Solaris, and only $20\mu S$ in our kernel-less implementation.

The application-level handler which implements the Appel & Li example can instead implement general purpose pagefault handling as part of a kernel-less implementation of virtual memory. Different address spaces (or even regions of address spaces) need not necessarily invoke the same pagefault handler, allowing multiple virtual memory implementations to coexist.

3.3 Kernel-less Operating System Services

Kernel-less operating system services have been successfully built using the microkernel/server approach, as in Mach 3.0. The goals of robustness and extensibility have been achieved for the services that were removed from the kernel. However microkernels still contain significant abstractions for threads, virtual memory, and inter-process communication. The result is that performance of Mach servers is much worse than for systems based on full-kernels. After significant efforts to improve performance, Mach researchers have retreated from the microkernel/server approach and are now dynamically loading server code directly into application address spaces (*co-location*).

There are two problems that co-location solves. The first is the cost of the Mach abstractions for communication. Even after extensive tuning, remote procedure calls (RPC) are still four times slower[9] than in systems like Spring[18]. Dynamic loading of services allows RPC to be avoided, as in Bridge/SFI[7] and systems like SPIN[5].

The second problem is that services implemented in distinct address space require parameters to be copied between address spaces. Services in the application address space can directly access parameters (given sufficient protection mechanisms).

Co-location is one of the advantages that kernel-based systems seem to have. Hardware memory management supports two protection domains within every address space: *user* and *supervisor*. Code executing in the supervisor domain can typically access all data in the user domain.

We have proposed an alternative implementation of co-location[1], which uses overlapping hardware address spaces to build an arbitrary number of domains within each address space. Services in co-located domains are accessed via cross-domain calls through portals. Services can directly reference any page within the common address space for which they have been granted permission.

On architectures with tagged TLBs and hierarchical page table structures, the cost of this implementation is the additional page tables that must be added to describe the distinct domains, as we show in[2].

In our kernel-less implementation, invocation of a null operating system service (co-located or not) takes only the cost of the cross-domain call ($1.6\mu S$ on our SPARC platform). A *getpid()* call to a kernel-based system, Solaris, takes $9.0\mu S$ on the same platform, yet this is on hardware designed to support the kernel-based approach.

The overhead in Solaris is due, indirectly, to its complexity. There is an very large amount of code in the kernel, and to make its management tractable, various abstractions are maintained: *proc* table entries, the *user* structure, run-time environment (register windows, the kernel stack), and LWP (light-weight processes). Requesting a kernel service normally requires passing through an *abstraction-adaptation* layer (i.e. *locore.s*), even if few of the abstractions are important for a particular service.

Because there is so much sharing of abstractions, kernel-based systems achieve robustness by insuring that the abstraction assumptions are maintained. If an inconsistency is detected at a time it is unexpected, the result is that the kernel has difficulty recovering. Generally this requires that the entire system reboot. Robustness in kernel-less operating systems can instead be achieved by executing groups of services in protected domains — each with appropriate restart mechanisms to recover from internal errors.

There have been several attacks on the overhead problem in the context of kernel-based systems. One has been to write special case code in locore.s for services whose performance is too critical to pass through the abstraction-adaptation layer. Another has been to dynamically synthesize kernel-code at run-time[11].

A kernel-less system does not eliminate the abstraction-adaptation layer. To do so would make it infeasible to construct service-rich operating systems. Where a kernel-less operating system gains in performance is by providing extensibility that allows customized versions of operating system services to be built that can take advantage of the specific circumstances of a particular application. Kernel-less systems are likely to still have large monolithic implementations of operating system services. The difference is that these services execute as applications, not a central kernel. There can be multiple instances, as in Mach, and there can also be alternative implementations of particular services.

4 SPACE

A significant perception, that has driven much recent operating systems research, is that crossing domain boundaries is too expensive, and thus other ways of solving the functionality and performance problems must be found. One direction that researchers have looked is embedding application code within the kernel domain, using language and linking techniques to build protection domains in software[5, 6].

These techniques are applicable to kernel-less systems just as much as to kernel-based systems. Instead of dynamically installing application code in the kernel, they can be used to install application code into servers (essentially the inverse of the Mach co-location approach).

Two factors are key for deciding when software protection domains are valuable. The first is how greatly the number of cross-domain calls can be reduced by the scheme. If taken to the extreme, the software protection approach would run all applications in a single hardware protection domain. Systems like the B6700 directly supported this approach in the context of a segmented architecture. However without segmentation, single-address space systems have problems with program linking, address-space allocation, and limitations on the manipulation of the address space (e.g. unix fork is impossible)[19]. Short of building single-address space systems, hardware domain crossings will continue to occur. Moving application functionality into the kernel can reduce the frequency of cross-domain calls, but will not eliminate them — particularly for existing applications which use a standard API.

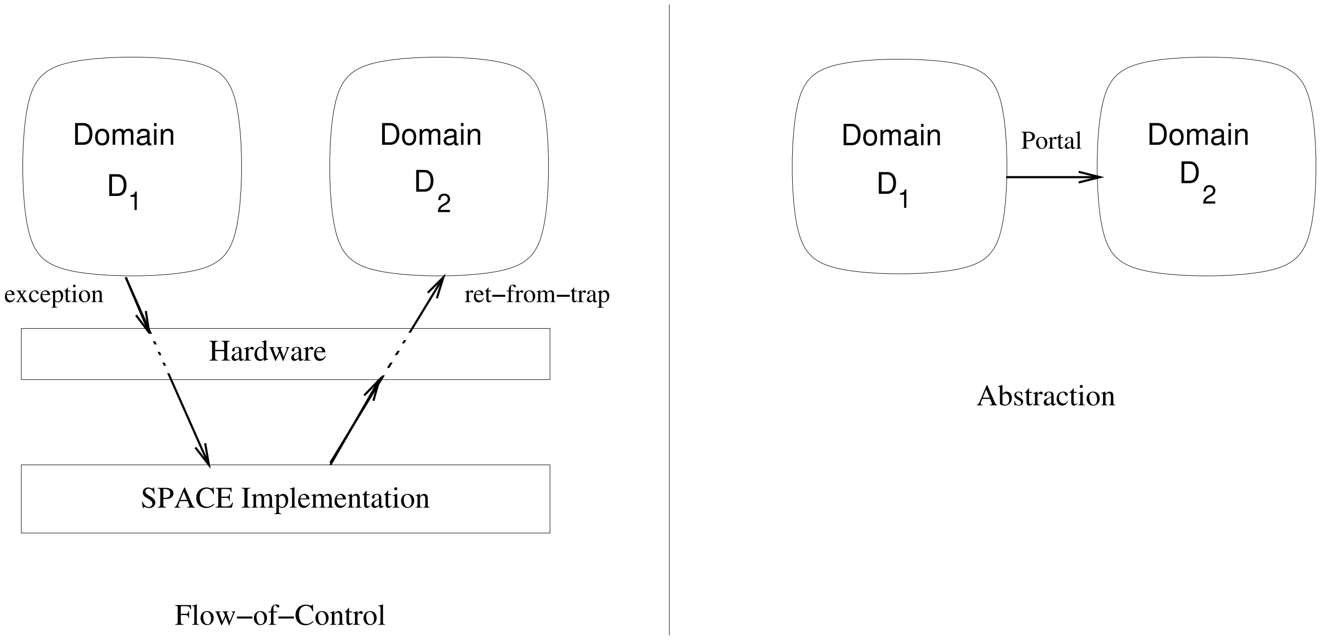
The second factor is the cost of hardware domain crossings. If cross-domain calls cost on the order of a procedure call, software protection domains would be unnecessary. The reduction of these costs is not only important for crossing between domains, but also for handling exceptions. Exceptions incur very similar costs to cross-domain calls.

Where software protection domains attempt to minimize the frequency of crossing hardware domains, the kernel-less approach assumes that the costs of crossing hardware domains can be dramatically reduced. In this section we present the SPACE approach which we used to make the measurements in section C.3. In the next section we examine the possibility of changes to hardware trap architectures which will allow direct implementation of cross-domain calls without the specialized code required in SPACE

SPACE is implemented as synthesized sequences of instructions that begin execution in the hardware trap vector and implement cross-domain calls. We have implemented SPACE on the SPARC processor, but the technique can be used on most processors that exist today. Ultimately we expect that the generalized trap architecture presented by SPACE will be integrated into the hardware trap architectures of a future generation of processors, resulting in direct support for generalized cross-domain calls. We examine the needed changes after presenting SPACE in its current form.

4.1 The SPACE mechanisms

The figure below depicts the Flow-of-Control in a cross-domain control transfer initiated by an exception which takes the control flow from domain D_1 into the SPACE Implementation. The SPACE Implementation determines the domain and the address of the handler based on the domain D_1 and the exception. The cross-domain control transfer is complete when the SPACE Implementation executes a hardware *return-from-trap* instruction to being executing in the domain D_2.



The abstraction of this control transfer is called a *portal*. Different portals can be defined for each domain by parameterizing a set of *portal types*. Portal types determine the details of a specified cross-domain control transfer, including how state is saved/restored and what parameters are passed.

C.4.2 Details of Cross-Domain Calls

To move from one domain to another requires that some state be preserved (analogous to saving the PC on a procedure call stack). Minimally SPACE needs to preserve the PC, the MMU context-id, and trap table pointer. SPACE also records the continuation address within the SPACE Implementation. This is the address of the code which will restore execution in the original domain when the invoked domain returns.

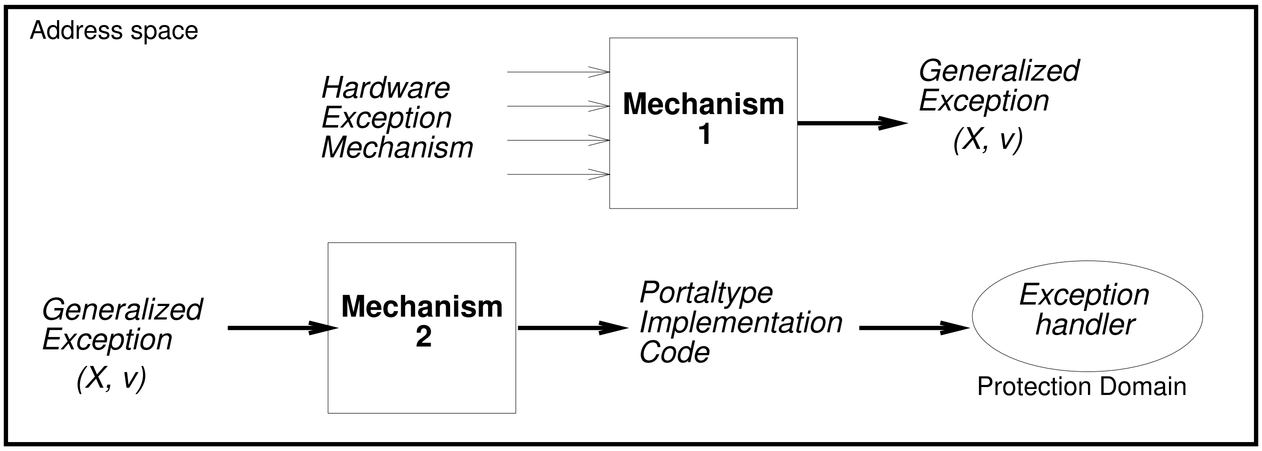
If trap registers are available, this minimal state can be saved without any memory references. But if the new domain performs yet-another cross-domain call, the available trap registers will not be available and a data structure (PCB) must be allocated to hold the processor state.

A sequence of cross-domain calls and returns will result in a LIFO chain of PCBs topped by state stored in the trap registers.

Cross-domain calls are implemented as portals which instantiate a particular portal type. The portal types are implemented as parameterized code templates and assembly code sequences. These templates are used to synthesize code when a portal is initialized. This synthesis allows parameters (such as MMU context id and location of the domain's trap table) to be bound when the domain is created. Use of static code and parameter lookup would add memory references, slowing execution.

5 The SPACE Model

SPACE generalizes the processor architecture by abstracting the hardware notions of *address spaces*, *privilege modes*, and *exceptions*. These mechanisms are used to construct software using *spaces*, *domains*, and *portals*. We describe the nature of the underlying mechanisms used by SPACE, and explain how these mechanisms provide fundamental extensibility. In a subsequent section we suggest how these mechanisms can be directly implemented in hardware.



Generalized Exception and Portal Type Mechanisms

5.1 Generalized Exceptions

Exceptions are extended in SPACE through the mechanisms shown in figure 2.

The first mechanism translates a hardware exception and related information (e.g. processor exception registers) into a generalized exception and parameter pair *(X,v)*. The parameter *v* normally identifies the particular instance of the *generalized exception X* (e.g. trap number). However for exceptions generated by memory references, *v* is the virtual address that caused the hardware exception.

This generalized exception mechanism decouples the hardware view of exceptions from the requirements of the application. Hardware exceptions which directly satisfy application requirements are mapped into generalized exceptions without overhead. Examples of exceptions which may be directly mapped include trap-instructions and arithmetic exceptions.

In other cases a hardware exception may not be suitable for the application. Perhaps the application was originally designed for another architecture. Perhaps the hardware exception needs to be demultiplexed into several cases that are handled differently. A generalized exception can be synthesized from a hardware exception and auxiliary processor state (e.g. special registers and access bits in page table entries).

5.2 Portals

The second mechanism maps generalized exceptions into portals. A *portal* associates an *(X,v)* pair with a *generalized-exception handler*. It specifies a new value for the program counter, as well as a new address space and privilege mode in which to execute. For memory exceptions, where *v* is a virtual address, different handlers can be associated with different parts of an address space. This supports the coexistence of distinct virtual memory implementations within an application. Portals are generalizations of the trap vector mechanism used in most processors.

5.3 Portal Types

Portals also specify the *portal type* implementation to use for traversing the portal to the exception handler and later returning. The portal type determines the conventions for saving state, synthesizing parameters, and resuming the interrupted execution upon return.

Portal type implementations do not necessarily involve cross-domain calls. Some portal types may execute to completion without ever entering a new protection domain, *short-circuiting* the normal portal mechanism. Such portal types are useful for extension of the processor instruction set (e.g. compare&swap), as well as implementation of short operations critical to application performance. Short-circuit portal types are also used to implement specialized operations, such as *return-from-exception*.

5.4 Address-based Capabilities

The use of a single kernel in conventional operating systems simplifies issues of protection. The kernel trusts itself. In a kernel-less operating system, with users writing system services, a more general protection mechanism is needed, such as capabilities.

Implementation of low-level capabilities can be built into portal types that need it. *The ability to traverse a portal is a capability*. This mechanism allows portals to be used to access protected services from one privilege mode without granting access to all privilege modes executing within the address space. Capabilities can be created dynamically, for example when a file is opened. In many cases capabilities eliminate the need for an operating system service to independently validate the right to access the service, saving overhead.

6 Hardware trap mechanisms for SPACE

SPACE does incur additional costs over Solaris, but these are primarily due to the cost of a round-trip crossing of the hardware domain boundary. A single trap/return-from-trap on the SuperSPARC takes 18 clocks (450nS@40MHz). Because traps into application domains are not directly supported, SPACE requires two crossings of the hardware trap boundary, while Solaris requires one.

As other researchers have observed, the cost of low-level hardware O/S support has not scaled with the improvements in overall processor mips rate. In the past these hardware costs were obscured by poor O/S performance. With SPACE these costs are exposed as being significant, which will lead to improvements in trap mechanisms by computer architects, even as switching address spaces has already been improved through the use of tagged TLBs and physical or tagged-virtual caches.

On the SuperSPARC processor we used for our experiments, the minimal cost of a round-trip through the trap mechanism (i.e. the trapvector contains a *rett/jmp* pair) is 18 clocks. The cost is much higher than the costs of the jmp/returns involved (6 clocks). The additional cost is due to the need to avoid traps caused by instructions before the exception being registered after supervisor mode is entered. The 12 clocks is normally not significant, as the software overhead in kernel-based systems is far greater. However for the fastest cross-domain mechanisms in SPACE, the hardware trap cost is uncovered and represents 37% of the cost of making a round-trip between domains.

The work with generalized trap mechanisms in SPACE suggests that hardware trap costs are more critical. There are a number of improvements that can be made, leading to the implementation of the critical SPACE operations in hardware.

The SPARC hardware trap mechanism performs an indexed branch into the trap vector, determined by the *trap base register* (TBR). The trap mechanism also switches to the trap register set, saves the program counters, and changes the processor mode to supervisor (with traps disabled):

$$(PC_{old}, REGS_{user}, MODE_{user}, TRAPS_{enabled})$$

$$\xrightarrow{TBR[EXCEPTION]} \; > \; (PC_{new}, REGS_{trap}, MODE_{super}, TRAPS_{disabled})$$

The generalized mechanisms in SPACE implement the following transition (from domain D_0 to D_1):

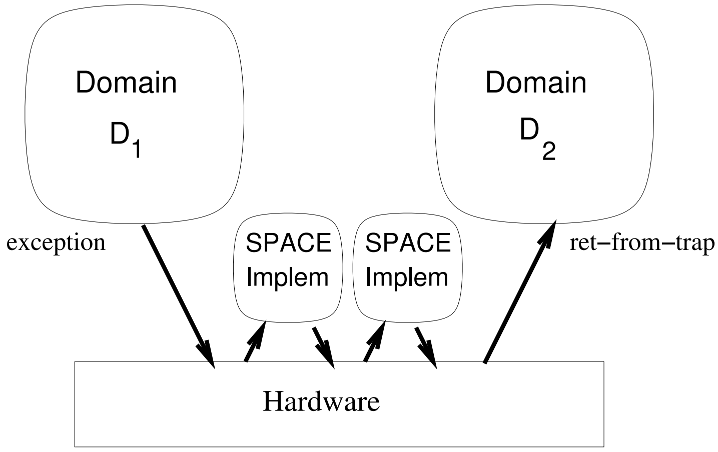
$$(PC_{old}, REGS_{D_0}, MMUCTXID_{D_0}, TRAPS_{enabled}, TBR_{D_0})$$
$$\xrightarrow{TBR_{D_0}[EXCEPTION]} \; > \; (PC_{new}, REGS_{*}, MMUCTXID_{D_1}, TRAPS_{enabled}, TBR_{D_1})$$

The differences from the hardware implementation are the following:

- The registers are saved/modified in an exception-dependent way. In the simplest case they are not modified at all, and state saving is left to the application handler.
- Instead of changing between hardware modes, the MMU context id is changed changing between hardware protection domains.
- The set of mappings are themselves changed by substituting a new value in the trap base register.

To implement the SPACE mechanisms in hardware requires that the MMU CTXID and new TBR value be available in the trap table as well as the new program counter. However the SPACE mechanisms are actually much more general than can be implemented by hardware. Hardware exceptions may need to be demultiplexed (synthesized exceptions). The transition between domains may depend on more than just the exception type (e.g. the faulting address). There may be specific requirements for saving register state generating parameters.

If the cost of the hardware mechanism is sufficiently reduced, it becomes feasible to eliminate the abstraction adaptation layer in the current version of SPACE in favor of one that invokes application level handlers to provide for more complex domain transitions:



7 A Kernel-less O/S for Parallel O/S Research

Our project is centered on building a kernel-less operating system as a platform for conducting research in the area of parallel platforms. Initially we plan to implement this operating system on the 64-node Meiko CS-2, which is the machine where we have already implemented the SPACE mechanisms. We propose to build a suitable O/S environment for running application code, including experiments involving operating system extensions.

7.1 Parallel O/S Requirements

Operating system services for a single-node computers can be categorized according the resources controlled and the level of the abstractions implemented:

	Low-level	**High-level**
Processor	Thread scheduling & synchronization	Process management, Interprocess communication, Authentication
MMU	Address space management	Virtual memory, Buffer management
Disk I/O	Disk Scheduling & DMA	Filesystems
Network I/O	Network interfacing	Network protocols
Special I/O	Timers, Graphics, Keyboard/mouse, Serial lines	Alarm clocks, Block transfers, Terminal line disciplines

Low-level services tend to focus on a single resource, while high-level services can involve several different types of abstractions (e.g. process management involves virtual memory).

Parallel computers are built out of single-node computers connected with a network. Of particular interest are machines where the network is fast (latency in microseconds, bandwidth in tens of megabytes/second).

Two common types of operating systems are used with parallel computers. One approach is to run an operating system that implements only low-level abstractions on most of the nodes, reserving a single node (or a few nodes) to provide higher-level functions[20, 21, 22]. One advantage is that memory and processor cycles are conserved for application use on the compute nodes.

The alternative approach is to run a full single-node operating system on each node, and add additional functionality to allow cooperation and communication between nodes[23].

These two approaches exist because to a large degree we do not yet know how to build an operating system which effectively provides services across a parallel machine with appropriate functionality and performances. The problem is not just that there are multiple processors, but that there are potentially many, many processors, and their communication performance is an order of magnitude slower than their memory access speeds.

To achieve high-performance from a parallel machine requires that the nodes be kept busy. This requires that communication and I/O effectively overlap node computations. It also requires that overhead be kept to a minimum.

Many parallel algorithms require that data computed on one subset of the nodes be available on another subset of nodes before they can proceed. Anything that interferes with consistent progress on all nodes can delay the progress in the computation across the entire machine. Thus a parallel machine operating system must provide much finer control over the scheduling of physical resources (processor, memory, communication, and I/O). Thread scheduling on individual nodes must be coordinated with other nodes. Page faults must be limited in their effect on progress in the computation. Communication between nodes and with disks must have low overhead and as well achieve overlap with computation without creating congestion in the network.

Higher-level abstractions require even more cooperation between nodes. Processes may migrate to balance loads. Node memory may be used to cache and stage I/O for other nodes. Filesystems may be built out of multiple disks attached to multiple nodes, yet provide both performance and consistency.

These issues and many more are being actively studied. At UC Santa Barbara researchers are working in many areas of parallel computer system research, including communication and hybrid shared memory systems, and parallel languages and filesystems, as well as parallel operating systems.

We propose to build a development environment in which these researchers can effectively conduct experiments with operating system services for high-performance computing. The kernel-less approach appears to be the appropriate approach for several reasons. First, a parallel operating system cannot be built out of a monolithic kernel, simply because there is no memory shared across all the nodes. Kernels on each node have to be tied together through cross-node communication, negating many of the advantages of using a monolithic kernel in the first place.

8 Summary

In previous research we have developed the SPACE mechanisms to the point where we are able to perform significant experiments that demonstrate order-of-magnitude performance improvements of a kernel-less implementation of standard mechanisms over a conventional kernel-based operating system.

The advantages of kernel-less operating systems are suggested by microkernel/server operating systems, but these systems have been unable to overcome the penalties of the abstractions contained in the microkernel. In a kernel-less operating system, even the most fundamental mechanisms can be replaced on a per-application basis.

The critical requirement is that cross-domain calls can be made very inexpensive. Otherwise approaches, such as software protection domains, are required which minimize the frequency of cross-domain calls.

The SPACE mechanisms exceed the performance of kernel-based mechanisms because modularity and the use of capabilities for security, eliminate the need to support the abstraction adaptation layer found in monolithic kernels.

Bibliography

[1] Dave Probert, John Bruno, and Murat Karaorman. Space: A new approach to operating system abstraction. In *Proceedings of the International Workshop on Object Orientation in Operating Systems*, pages 133–137, October 1991.

[2] Dave Probert and John Bruno. Building fundamentally extensible application-specific operating systems in space. Technical Report TRCS95-06, Computer Science Dept., UC Santa Barbara, March 1995.

[3] Patrick Bridges. Current operating systems projects and os-related research. http://www.cs.arizona.edu/people/bridges/oses.html.

[4] David R. Cheriton and Kenneth J. Duda. A caching model of operating system kernel functionality. In *Proceedings of the 1st USENIX Symposium on Operating Systems Design and Implementation*, pages 179–193, November 1994.

[5] Brian Bershad, Stefan Savage, and et. al. Extensibility, safety and performance in the spin operating system. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, December 1995.

[6] Steve Lucco. Adaptable binary programs. Technical Report CMU-CS-94-137, CMU, 1994.

[7] R. Wahbe, S. Lucco, T. Anderson, and S. Graham. Efficient software-based fault isolation. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles*, pages 203–216, July 1993.

[8] In-Kernel Servers on Mach 3.0: Implementation and Performance. Jay lepreau and mike hibler and bryan ford and jeff law. In *Proc. of the 3rd Usenix Mach Symposium*, Santa Fe, NM, April 1993.

[9] Michael Condict, David Mitchell, and Franklin Reynolds. Optimizing performance of mach-based systems by server co-location: A detailed design, August 1993.

[10] Arindam Banerji and David L. Cohn. Protected shared libraries. To appear in IWOOOS, December 1994.

[11] C. Pu, H. Massalin, and J. Ioannidis. The synthesis kernel. *Computing Systems*, 1(1):11–32, Winter 1988.

[12] Calton Pu and et. al. Optimistic incremental specialization: Streamlining a commercial operating system. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, December 1995.

[13] A. B. Montz and et. al. Scout: A communications-oriented operating system. Technical Report 94-20, Dept. of Computer Science, Univ. of Arizona, June 1994.

[14] Dawson R. Engler, M. Frans Kaashoek, and James W. O'Toole Jr. Exokernel: an operating system architecture for application-level resource management. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, December 1995.

[15] Berny Goodheart and James Cox. *The Magic Garden Explained*. Prentice Hall, 1994.

[16] Richard Rashid, Avadis Tevanian, Jr., Michael Young, David Golub, Robert Baron, David Black, William J. Bolosky, and Jonathan Chew. Machine-independent virtual memory management for paged uniprocessor and multiprocessor architectures. *IEEE Transactions on Computer*, 37(8):896–908, August 1988.

[17] Andrew W. Appel and Kai Li. Virtual memory primitives for user programs. In *ASPLOS IV*, pages 96–107, April 1991.

[18] J. Mitchell and et. al. An overview of the spring system. In *Proceedings of Compcon Spring 1994*, February 1994.

[19] Jeffrey S. Chase, Henry M. Levy, Edward D. Lazowska, and Miche Baker-Harvey. Lightweight shared objects in a 64-bit operating system. In *OOPSLA*, October 1992.

[20] Thinking Machine Corporation. Cm-5 technical summary, 1991.

[21] Arthur B. Maccabe, Kevin S. McCurley, Rolf Riesen, and Stephen R. Wheat. Sunmos for the intel paragon: A brief user's guide. In *Proc of the Intel Supercomputer Users' Group*, pages 245—251, June 1994.

[22] Stephen R. Wheat and et. al. Puma: An operating system for massively parallel systems. In *Proc of the 27th Hawaii Intl Conf on System Sciences*, pages 56—65, 1994.

[23] Meiko Scientific Corporation. Meiko cs-2 documentation set, 1995.