Interposition as an Operating System Extension Mechanism

Douglas P. Ghormley Steven H. Rodrigues David Petrou Thomas E. Anderson

{ghorm,steverod,dpetrou,tea}@cs.berkeley.edu
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720

March 19, 1997

Abstract

Modern production operating systems are large and complex systems developed over many years by large teams of programmers, containing many hundreds of thousands of lines of code. Consequently, it is extremely difficult to add significant new functionality to these systems [16, 1, 42]. In response to this problem, a number of recent research projects have addressed the issue of extensible operating systems; these include SPIN [7, 5], VINO [45, 41], Exokernel [20], Lipto [17], and Fluke [21]. This paper addresses the problem of providing extensibility for existing production operating systems such as Solaris, through the technique of interposition on existing kernel interfaces. Interposition is useful for extensions because it is transparent, it permits the incremental addition of functionality to an interface, and it enables the easy composition of multiple extensions.

We have designed and implemented a prototype extension mechanism, SLIC, which utilizes interposition to efficiently insert trusted extension code into a production operating system kernel. We have used SLIC to implement a number of useful operating system extension, such as a patch to fix a security hole described in a CERT advisory, an encryption file system, and a restricted execution environment for arbitrary untrusted binaries. Performance measurements of the SLIC prototype show that interposition on existing kernel interfaces can be accomplished efficiently.

1 Introduction

Modern production operating systems are large and complex systems developed over many years by large teams of programmers, containing many hundreds of thousands of lines of code. To make matters worse, in order to run on SMP's, much of the code must be multi-threaded, compounding its complexity and requiring extensive revalidation even for the smallest of changes. It is common for major releases of production operating systems to be riddled with flaws introduced in developing the features in the release, usually requiring multiple "bug fix" releases that in turn introduce their own flaws.

Consequently, in practice, it is extremely difficult to add significant new functionality to modern production operating systems [16, 1, 42]. This does not diminish the need to continue to modify these systems. For example, security flaws are routinely discovered and reported by organizations such as Carnegie-Mellon's Computer Emergency Response Team (CERT) and the Department of Energy's Computer Incident Advisory Capability (CIAC). Despite the need for immediate repair to prevent wide exploitation of the flaw, the required patches often take weeks to become available [13]. In addition, there is a large catalog of value-add functionality that has not been widely deployed, in part because of the difficulty of modifying existing systems: load sharing [56], process migration [48, 16], fast communication primitives [6, 49], upcalls [14], distributed shared memory [32], and user-level pagers [54].

The goal of this paper is to simplify the process of evolving and extending existing production operating systems. A consequence of accomplishing this would be to enable independent software vendors (ISV's) to develop and deploy innovative operating system features. By contrast with operating systems where relatively few successful ISV's exist (Transarc being a notable exception), robust ISV markets exist in other areas, such as databases [18], Web software [24, 33], and desktop publishing [38, 9] — in each case, because extensibility has been designed into the system.

Prior approaches to extending operating systems can be roughly divided into three categories: (i) re-engineer the operating system from the ground up, in the process making it easier to extend, (ii) incrementally re-engineer selected portions of the kernel, and (iii) add extensions to existing systems without significant modification to either the operating system or its applications.

Over the years, a number of systems have attempted to reduce the cost of adding new kernel functionality by restructuring the operating system with extensibility as a design goal. Systems built using this approach include Hydra [53], Mach [1], SPIN [7, 5], VINO [45, 41], Exokernel [20], Lipto [17], and Fluke [21]. Many of these systems have successfully demonstrated greatly reduced costs of adding new functionality. However, the cost of starting over from scratch can be prohibitive; for example, Microsoft has spent over $300M developing Windows NT [55]. Thus, it is likely that the need to extend existing production operating systems will persist for the foreseeable future.

A small number of projects have taken the alternative approach of re-engineering existing kernel interfaces to reduce the complexity of adding new functionality at those interfaces. The `vnode` interface [30, 39] is a prime example of this approach. While this interface is extremely useful for adding complete new file systems, it does not support incrementally adding functionality. Implementation details of the interface also vary widely across operating system vendors [51]. Applying this technique to make existing operating systems more extensible would require substantially modifying and exposing all interfaces where additional functionality is desired, effectively re-engineering the majority of the operating system.

We take the alternative approach of adding functionality with no modifications of application code and only minor modifications to the underlying operating system. We differ with earlier efforts in that our solution — kernel-level insertion of trusted extension code — is simple to implement, efficient, requires no hardware support, simplifies composability among extensions, and protects extensions from malicious or buggy applications. We believe that no other system provides this powerful combination of features for extending existing production operating systems.

Specifically, Interposition Agents [27] leverages the Mach system call redirection facility to transparently insert extensions at the system call interface. However, since the extension code runs in the application's address space and protection domain, this solution cannot enforce security guarantees or share resources among distrustful processes. Software Fault Isolation (SFI) [50] can be used to rewrite application binaries to protect extensions running at user level from the application; unfortunately, it is difficult to apply SFI to arbitrary application programs without prohibitive implementation complexity. Protected Shared Libraries [4] has the same capability as SFI with less software effort, but requires specialized hardware support that is not available on most architectures.

To investigate these issues, we have developed SLIC, a prototype system for efficiently inserting trusted extension code into existing operating system kernels without application or kernel source code. Conceptually, SLIC is a wrapper around the operating system kernel, potentially capturing events at many different interfaces — system calls, exceptions, page faults, and device interrupts (although only system calls and certain signal-generating exceptions are supported in the prototype). This wrapper transparently invokes extension code, enabling an extension to export new functionality to applications while the underlying kernel remains oblivious to the extension. SLIC dynamically loads extensions into the kernel, where they are installed either by modifying jump tables or by patching the kernel's event handlers, as appropriate. The prototype currently runs on Solaris 2.5; a Linux 2.0 port is in progress.

We have used the SLIC prototype to implement a number of extensions that would have been significantly more difficult to accomplish by other means. One extension patches a security flaw publicized by CERT [11]. A second enables process execution with restricted capabilities, while a third is an encrypted file system.

The rest of this paper is organized as follows. Section 2 provides background on interposition. In section 3, we describe the design, implementation, and performance of SLIC, our prototype interposition system. Three sample extensions and their performance are presented in section 4. In section 5 we discuss our experience with interposition as an extension tool, and the lessons we have learned on building system interfaces to support interposition effectively. Section 6 discusses related work while sections 7 and 8 close with future work and conclusions.

2 Interposition Background

Interposition is the process of capturing events crossing an interface boundary and forwarding those events to an *interface extension*. The extension performs some processing on the event, and then passes the event on to its original destination. Figure 1 demonstrates interposition on an interface by first one, and then a second, extension. The original interface is maintained by the inserted extension code. Interposition is thus *transparent*; user applications and the kernel are oblivious to extension code.

Interposition enables *incremental extensions*. Extensions need only capture the events that they are interested in, instead of all events crossing the interface boundary, and they can leverage the functionality of the existing interface. This means that extension writers only have to implement the desired extension functionality, instead of the functionality of the entire interface.

Because interposition maintains the original interface above and below the extension, it can be applied recursively; this is *composition*. The right-hand side of Figure 1 shows a second extension being added to the extension stack. In the diagram, extensions A and B are oblivious of each other's presence, just as the application and kernel are oblivious to the presence of any extensions.

The features of transparency, incrementality, and composability make interposition uniquely suited to the extension of existing interfaces. Specifically, the transparency of interposition implies that interfaces not originally designed for ex-

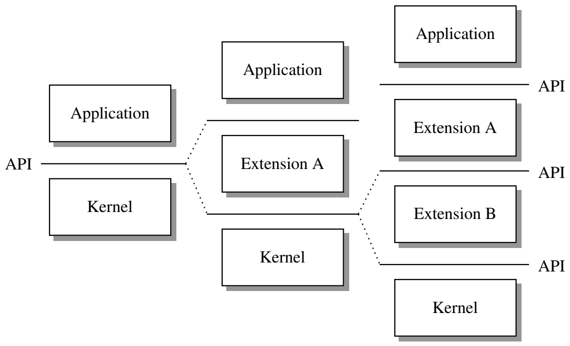

Figure 1: Interposing extensions on an interface. The dotted lines illustrate an interface which was interposed on by an extension. Events are intercepted as they cross the original interface and are routed through extension code. In this diagram, the original API on the left is maintained at each level on the right, making the interposed extension transparent to the applications, the kernel, and even the other extensions.

tensibility can be extended. Incrementality ensures that extensions need only provide the functionality desired, without re-implementing substantial portions of the kernel. Composability means that any number of extensions provided by independent vendors can be applied to a single interface.

Using kernel-level interposition, extensions have a broad range of capabilities. Extensions can provide security guarantees (for example, patching security flaws or providing access control lists), virtualize resources (providing a cluster-wide process identifier), modify data (transparently compressing or encrypting files), re-route events (sending events across the network for distributed systems extensions), or inspect events and data (tracing, logging).

However, interposition does have a number of limitations. First, interposition requires a well-defined interface on which to capture events. System with poorly decomposed functionality may have few such interfaces. Second, new functionality can only be implemented in terms of existing functionality. Extensions cannot, for example, add new abstractions to the system, but only compositions and variants of existing abstractions. This can limit extension functionality; for example, a cache-coherent file system can only be constructed through interposition if underlying layers expose a cache-management mechanism in the file system interface [28]. This limitation on new functionality being expressible only as a composition or variant of existing functionality is more strict: interposition cannot add new events to interfaces. New events can only be added by overloading existing events (as routinely occurs with the Unix `ioctl()` interface).

Despite these limitations, interposition's power and flexibility has led to its wide-spread use throughout modern computing systems. Forms of interposition can be found in the 'pipe' construct used in Unix shells, in the extensibility mechanisms of programming language systems [29], in distributed file systems such as NFS [40], in distributed shared memory systems such as TreadMarks [2], in World Wide Web proxy caches [8], and in MS-DOS terminate-and-stay-resident utilities and Macintosh toolbox extensions.

3 SLIC Design and Implementation

To investigate the suitability of interposition for adding new functionality to existing operating systems, we have designed and implemented SLIC, an interposition mechanism for production Unix operating systems. The primary purpose of SLIC is to leverage the transparency, incrementality, and composability of interposition to extend existing operating systems with minimal kernel modifications. Within this framework, SLIC was designed to provide extensions with the following features:

Security: Extensions should be able to make decisions that cannot be subverted. This enables extensions which affect physical resources, such as file systems, and extensions which enforce security mechanisms.

Efficiency: The interposition mechanism should impose minimal overhead on the system and per-extension overhead should be a few times the cost of a procedure call, when performance is a concern. The user-level extension code used in systems such as Interposition Agents [27] is costly to access; our mechanism should enable in-kernel or mixed in-kernel/user-level extensions for performance [47].

Ease of Development: Extension writers should be able to use state-of-the-art programming tools such as symbolic debuggers and performance analysis tools.

SLIC assumes that extensions are trusted. For untrusted extensions, kernel code and data can be protected from malicious or faulty extensions through technologies such as Software Fault Isolation [50, 41], or by writing extensions in a safe language such as Modula-3 or Java [44, 26].

3.1 SLIC Architecture

SLIC is comprised of multiple *dispatchers* and *extensions* as well as various *support routines*. Dispatchers are responsible for intercepting system interface events and for routing those events to interested extensions. Extensions receive events from the dispatcher and provide additional functionality to the operating system. Support routines provide a simple, consistent interface to useful functionality such as memory allocation and synchronization primitives. These routines enable extensions to be portable across implementations of SLIC for various operating systems. Each dispatcher

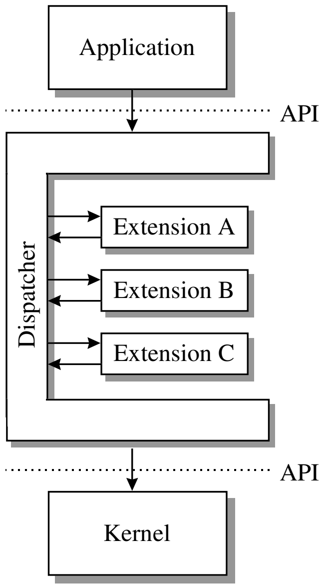

Figure 2: The basic design of SLIC. The dotted lines represent the interposed interface. Events crossing this interface are captured by a *dispatcher* and forwarded to one or more *extensions*.

may provide additional support routines as appropriate for the interface. For example, the system call dispatcher may provide routines to determine the children of a given process.

3.1.1 Dispatchers

Each dispatcher is responsible for capturing events on a single system interface (such as system calls or signals). SLIC dispatchers use two different techniques to intercept interface events. For those interfaces which are implemented using jump tables, such as the system call, `vnode`, and virtual memory interfaces, SLIC records the original function address from the jump table and stores the address of its own interception routine. For procedural interfaces which are called directly from various locations in the kernel, such as the Solaris signal delivery and page replacement policy interfaces, dispatchers intercept events on these interfaces using binary patching. The first few instructions of the relevant procedure is saved and replaced with instructions to jump to the dispatcher whenever the procedure is called. After the dispatcher runs, if the event has not been aborted by an extension, the saved instructions from the interposed procedure are executed and control is returned to that interposed procedure right after the binary patch. Using these techniques, SLIC dispatchers can intercept interface invocations for the cost of a procedure call.

Once an event has been captured by a dispatcher, that event is sent to interested extensions for processing. Figure 2 depicts the relationship between dispatchers and extensions. Extensions express interest in events using a small number of predicates defined by the dispatcher. For example, the system call dispatcher filters system call events based on process identifier and system call type; an extension that only wishes to trace the `open()` system call from process 4191 can specify this easily. The signal dispatcher provides similar functionality, enabling filtering on process id and signal type. These predicates are roughly analogous to the *guards* found in SPIN [35].

Table 1 presents a simplified portion of the system call dispatcher interface. Upon receiving an event, an extension has a number of options available: the extension can pass along the event unmodified, the extension can modify the event parameters in `struct Slic_SyscallInfo` and then pass it along, or the extension can complete the event with an arbitrary return value or error condition (e.g., when a system call would exploit a known security hole) by setting fields in `struct Slic_ReturnInfo`. When an event is completed, extensions further down the call chain (and the kernel) never see the event for processing. Additionally, the extension may initiate other events using `Slic_IssueSyscall()` with arbitrary parameters, capturing the return values. For example, a system call tracing extension must periodically dump collected tracing data to stable storage. A limitation of our current system call dispatcher is that additional events generated by extensions appear to have been generated by the user application, and have the same privileges (and limitations) of the user process.

If an extension passes an event on for further processing, the dispatcher routes that event to other interested extensions. For this purpose, the dispatcher maintains a chain of extensions through which events flow. Events can be modified or completed at any point along this chain. The dispatcher treats the underlying kernel as the last extension on the chain, passing it the event only if the event was not completed by an extension earlier on the chain. When an extension marks an event as completed, the return value then flows up the chain in reverse order, supporting inspection or modification of the value.

On initialization, extensions register with one or more dispatchers using `Slic_RegisterExtension()`. The two predicates for the system call interface, a per-process tracing flag and a bit-mask of intercepted system calls, are manipulated using the `Slic_TraceProc()`/`Slic_UntraceProc()` and `Slic_RegisterHandler()` functions. An extension is only invoked for a given system call event if the process which generated the system call is marked for tracing by this extension and the extension has a handler registered for this system call. Similar functionality is provided by the signal dispatcher.

```
struct Slic_SyscallInfo {
       int syscallNum;
       int args[];
};

struct Slic_ReturnInfo {
       bool forceReturn;
       int returnValue;
       int errno;
};

Slic_RegisterExtension(int dispatcherId);
Slic_RegisterHandler(int syscallNum,
             void (handler)(Slic_SyscallInfo *syscallInfo,
             Slic_ReturnInfo *returnInfo));
Slic_TraceProc(int pid, bool traceAllChildren);
Slic_UntraceProc(int pid, bool untraceAllChildren);
Slic_IssueSyscall(Slic_SyscallInfo *syscallInfo,
                  Slic_ReturnInfo *returnInfo);
```

Table 1: Simplified interface for the system call dispatcher.

3.1.2 Extensions

The SLIC architecture enables extensions to be structured in two ways, supporting a tradeoff between performance and ease of extension development. Extensions can be loaded as *user-level servers* or as *in-kernel extensions*, as shown in Figure 3, or as a combination of these types.

Extension code executing as a *user-level server* enables extension development to proceed as with normal user programs, with access to user-level libraries (such as communication libraries) and state-of-the-art development tools (such as symbolic debuggers and performance analysis tools such as Purify [36] and Quantify [37]). A user-level extension must not register its own events or events from its development tools with any dispatchers. The extension is protected against modification by user programs by virtue of running in a separate address space. The drawback of this approach is that invoking the extension from the dispatcher requires costly context switches and kernel-user boundary crossings. This organization is similar to that employed by micro-kernels such as Mach [1].

An *in-kernel extension* is loaded directly into the kernel. When events are frequent, this organization has considerably better performance than the user-level approach, since the extensions are directly invoked by a procedure call from the dispatcher. Extensions are protected against modification by user applications because they are in the protected kernel region of the address space. There are a number of limitations to this approach: the kernel is not protected from malicious or faulty extension code and there is no support for user-level development tools. Methods of protecting the kernel are well-known [35, 26, 43, 41]. By supplying extensions with the same interface, whether at the user level or in the kernel, extensions can be safely developed at the user level and then inserted into the kernel. This enables development in a safe environment without sacrificing the potential for good performance.

SLIC extensions can also use both models simultaneously. Performance-critical sections of an extension can be located in the kernel, while functionality that is rarely used or which requires access to user-level libraries can be located in a user-level server.

3.2 SLIC Implementation

The current implementation of SLIC provides dispatchers on the system call and signal interfaces of Solaris 2.5 running on UltraSPARC workstations; work is in progress for a Solaris page replacement policy dispatcher and a port to the Intel 80*x*86 version of Linux 2.0, an operating system whose internals share no heritage with Solaris. At this time, the signal dispatcher is not yet fully functional and only supports simple extensions used for tracing. All SLIC components — dispatchers, extensions, and support routines — are dynamically loaded by the system administrator into the kernel as loadable device drivers.

Solaris system calls are routed through the `sysent` table, which contains function pointers to the appropriate system call routines. The system call dispatcher intercepts system call events by replacing entries in this table with pointers to its own dispatch function. Solaris signal delivery proceeds through the `sigaddq()` and `sigaddqa()` functions; the signal dispatcher intercepts signals by modifying the in-memory images of these two functions. To enable the in-memory modification of these functions, one line of Solaris source code was changed to make the kernel text image writable by root processes.

The current implementation catches events within the kernel, rather than at the machine level (interrupts); for example, system calls are caught at the `sysent` table rather than upon execution of the `trap` instruction. While the two approaches are conceptually similar, catching events within the kernel considerably simplified the implementation of our prototype.

3.2.1 System Call Dispatcher

System calls transfer data in the form of pass-by-value arguments and pass-by-address memory buffers. Modifying the pass-by-value arguments in an extension is straightforward. Modifying in-memory data is more difficult, as the data is located in a user-level page. When user programs are multi-threaded, in-place modification of data buffers renders the extension's operation vulnerable to inspection or subversion by the user program. Thus, only the kernel can be allowed to access the modified memory buffer. However, the Solaris kernel has protection checks on memory buffers that require that these buffers be located in the application's ad-

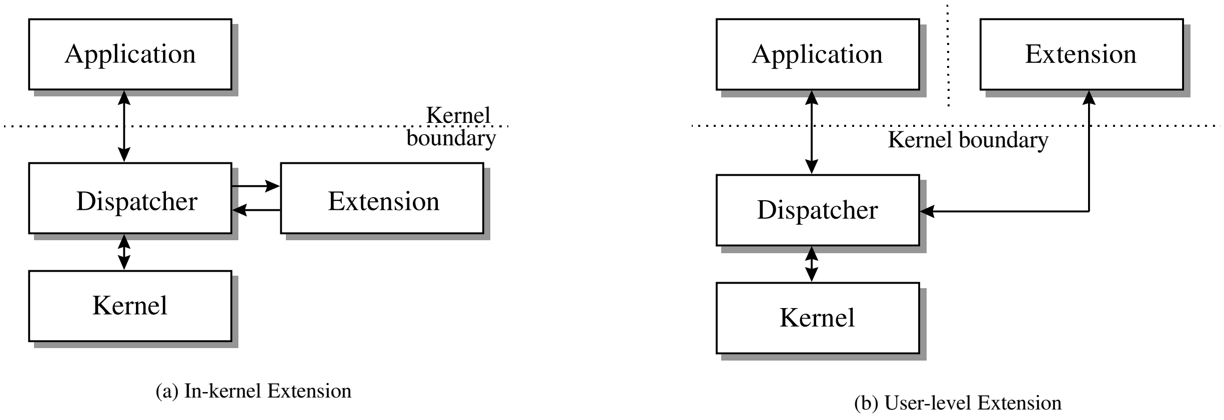

Figure 3: Organization alternatives in SLIC. In 3(a), extension code is located within the kernel's address space, while in 3(b), extension code is run as a separate user level process. Dotted lines represent protection domain boundaries. Passing the event to a user level extension (as in 3(b)) involves multiple context switches and protection boundary crossings.

dress space. (In general, applications cannot pass kernel address as memory arguments to system calls.) These checks are performed in the kernel's `copyin()` and `copyout()` routines.

Our solution to this problem is to mark regions of the kernel address space as *valid* for system call memory buffers and to modify the `copyin()` and `copyout()` routines to recognize these regions. This is a potential security hole, as user applications could pass in the addresses of these buffers into the kernel and the modified `copyin()`/`copyout()` routines would not catch this violation. To ensure security, the in-kernel system call buffers are marked valid on a per-thread and per-event basis. This ensures that in-kernel buffers can only pass security checks when an extension has explicitly requested such a buffer. Even when such a buffer exists, other user threads cannot access it.

3.3 Micro-benchmarks

We have measured the overhead imposed by SLIC on system calls and signals. All micro-benchmark numbers reported are from a 167MHz UltraSPARC running Solaris 2.5.1.

There are three micro-benchmarks; we run each operation one million times and average the result. The first micro-benchmark performs a `getpid()` call, which in Solaris is essentially a null system call. This micro-benchmark measures the raw overhead of the system call dispatcher, but does not invoke the modified copy routines. In order to quantify the overhead resulting from our modification to the copy routines, the second micro-benchmark performs a `sigprocmask()` system call. This system call involves a memory copy of 16 bytes from a kernel data structure into user space. The `kill()` micro-benchmark measures the overhead of the signal dispatcher. It involves a single process sending itself a `SIGUSR1` signal. We use a single process to avoid context switches, thus maximizing the effect of our overhead. This micro-benchmark measures the overhead of the signal mechanism and the cost of interposing on the signal mechanism.

We tested our micro-benchmarks with various configurations of SLIC: unloaded (i.e., a bare system), loaded and active with no extensions, and loaded and active with multiple extensions. The extensions used for these measurements are null extensions which merely catch events and re-issue them. For system call events, after an extension re-issues a system call, the extension also receives the return value from this system call. The results are presented in Table 2. The implementation for user-level extensions of the signal interface is not yet complete.

The base overhead of SLIC is very low, adding between 200 and 600 nanoseconds to the overhead of a system call. Loading an extension adds between 300 nanoseconds and 3 microseconds to the overhead; the reason for the discrepancy is that `getpid()` does not pass in any memory arguments, while `sigprocmask()` invokes the modified copy routines. Note that the cost of the copy routines is paid regardless of whether any extensions are active or not. The primary source of per-extension overhead is the allocation of register windows in the UltraSPARC. The current implementation forces a register window to be allocated for every extension that wishes to examine the return value from a system call. Measurements indicate that extensions which do not capture return values, and thus do not require a register window, add an average of only 220 nanoseconds to the dispatcher overhead.

	getpid()	sigprocmask()	kill()
Unmodified system	2.82	3.90	80.83
SLIC, dispatcher loaded, no extensions	3.08	4.53	89.09
SLIC, dispatcher loaded null extension (disabled)	3.22	4.44	88.99
SLIC, dispatcher loaded null extension (enabled)	8.27	9.69	99.28
SLIC, dispatcher loaded $2 \times$ null extension	10.79	12.28	103.27
SLIC, dispatcher loaded $3 \times$ null extension	13.37	14.84	108.01
SLIC, dispatcher loaded null extension at user level	57.45	62.53	

Table 2: Micro-benchmark performance of SLIC. All numbers are in microseconds and are averaged over one million iterations. The `getpid()` and `sigprocmask()` tests were run with the system call dispatcher loaded while the `kill()` test was run with the signal dispatcher. The `getpid()` test is effectively a measure of a null system call; the `sigprocmask()` test measures the overhead of SLIC's interposition on the kernel copy routines. The `kill()` benchmark measures the time required for a process to signal itself. The implementation for user-level extensions of the signal interface is not yet complete.

4 Extending the Operating System

To demonstrate the functionality and performance of SLIC, we have implemented prototypes of a variety of extensions: a solution to a recent CERT advisory, an encryption file system, and a restricted execution environment. Without SLIC, these extensions would have required kernel source code modifications to achieve similar functionality and performance. Further, without SLIC, changes to add these features to the existing kernel would likely be *ad hoc* rather than a general solution that can be leveraged for future extensions.

4.1 CERT Advisory extension

The Computer Emergency Response Team (CERT) of Carnegie Mellon University's Software Engineering Institute regularly provides the Internet community with information regarding system security problems. Whenever possible, these advisories include information on how to resolve the problem reported. However, due to the lack of extensibility in existing systems, frequently this advice is to completely disable the insecure feature [11, 10, 12]. Though operating system vendors do provide patches for software found to be insecure, these patches can take weeks to become available [13]. Since CERT does not have access to source code for many systems, they are unable to directly provide patches for these problems. However, using SLIC, many of these advisories could be accompanied by small extensions which would resolve the problem.

To demonstrate this, we have implemented an extension to patch a recent security hole discovered in the Solaris `admintool` [11] which allowed unprivileged users to delete arbitrary files in the system in certain circumstances. Our 25 line extension monitors file operations, denying the operations which cause the problem. Operation of the `admintool` is maintained.

4.2 Encryption File System

In a distributed file system, maintaining file security is often a concern. For example, in a networked environment with a central file server, traditional Unix file protections can be easily circumvented by monitoring the network traffic. To protect sensitive files, users may use encryption tools such as PGP [22]. However, stand-alone encryption tools can be time consuming and are not easily integrated with existing application binaries. A more effective method of ensuring file security is to support file encryption directly in the file system, transparently encrypting file writes, and decrypting file reads, when communicating with the server.

We have implemented a simple encryption file system. This extension implements a trivial *exclusive-or* encryption algorithm similar to that implemented in VINO [41]. The extension watches for `open()` and `creat()` system calls of files with a particular suffix and then records the process id and the file descriptor returned to the application. On subsequent `read()` or `write()` system calls to these file descriptors, the extension applies a byte-wise `xor` on the data. This operation adds an extra copy step to the data transfer.

Implementing an encryption file system using the standard `vnode` interface found in most production systems would have required implementing a full file system. In contrast, our extension implements only the additional functionality of encryption and can be applied to any mounted file system.

4.3 Restricted Execution Environment

Under UNIX, processes run by a user has access to all of the resources granted to that user. There are many cases, however, in which the user does not trust the program being run. For example, programs downloaded from untrusted sources like the Internet may actually be Trojan horses designed to steal or destroy information [52, 15]. In addition, there are cases in which the user trusts the program, but not the data being processed, as in the case of helper applications used by web browsers to display various data formats. Input data could potentially exploit bugs in helper applications to insert Trojan horses into the system, a process similar to that used by the Internet Worm `fingerd` attack [19].

A common method for constructing a restricted execution environment is to use the tracing facility of the standard `/proc` file system to selectively deny or change those system calls which would violate security [23]. This approach suffers from a number of shortcomings. First, the Solaris 2.5 `/proc` file system only allows system calls to be denied with `EINTR` as the error code, a code normally used to indicate that the attempted operation was interrupted and should be retried. Returning this code causes many applications to loop endlessly rather than receiving a permission violation error. Second, the `/proc` file system cannot trace programs which are marked `setuid`; in Solaris, this includes programs such as `crontab`, `ps`, and `ping`. Finally, intercepting system calls using `/proc` is expensive. We measured an added 150 microseconds to the base system call overhead. This is especially problematic for system-call intensive applications. Using SLIC we have implemented a restricted execution environment extension that does not have these limitations.

This extension provides the user with a configurable security environment. For example, applications can be given a subset of read/write/execute access to any number of directory subtrees. The right to `fork()` can be disabled. Any attempts by the application to perform a proscribed operation results in an `EPERM` error. The extension monitors only the subset of system calls necessary to maintain the security guarantees, providing low overhead. When traced applications invoke restricted system calls, the extension checks the arguments to the call and determines if the call should be allowed or denied. For example, applications running under the X Window System can send X events to other running X applications, which could potentially corrupt the user's trusted applications (such as `netscape`). A solution is to run the untrusted application in a secure, restricted X server such as `Xnest`[1]. Our restricted environment extension can check `write()` calls to ensure that they are going to the `Xnest` server instead of the standard X server. Because it runs under SLIC, this extension cannot be circumvented and thus maintains its security guarantees. The restricted environment used for benchmarking denies 45 system calls outright (for example `chown()`), and performs security checks, such as checking the path or file access permissions, for 21 additional system calls (for example `rmdir()`.

[1] `Xnest` is part of the standard X11R6 distribution.

4.4 Performance

To evaluate the impact of these extensions on system performance, we ran the extensions under three benchmarks: the Modified Andrew Benchmark [25, 34], a TEX compilation of a 262-page (760KB) document, and a `gcc` compilation of `emacs-19.34` without support for X Windows. The Modified Andrew Benchmark consists of multiple phases which (i) create directory subtrees, (ii) copy files, (iii) search file attributes via `find` commands, (iv) search files for a text string via `grep`, and (v) compile files. While the benchmark fits entirely in the file cache of modern systems and is therefore useless for measuring file system performance, this will expose the overhead imposed by SLIC. The TEX benchmark was chosen to be representative of a document processing workload. Lastly, we chose the `gcc` benchmark because it performs significant I/O. Table 3 reports some relevant statistics for each benchmark.

Table 4 presents the results of running the benchmarks on each extension as well as on all extensions simultaneously. To decrease variability due to disk latency, all data files for the benchmarks were placed in a memory-mounted `/tmp` file system. The TEX benchmark shows very little change in performance due to SLIC, as expected from the relative infrequency of system calls listed in Table 3. We believe that variance in the measurements accounts for the fact that the TEX benchmark appears to run faster under the Restricted Execution Environment than with no extensions loaded. Though SLIC imposes a certain amount of overhead on applications, the last line in Table 4 illustrates that the much of the overhead experienced by the benchmarks is due to the SLIC infrastructure, a cost which is only paid once; the per-extension overhead is small for common workloads.

5 Evaluating Interposition

This section describes our experiences in implementing SLIC and presents a number of general principles for developing interfaces that are conducive to interposition. Although SLIC was designed for and will work with existing operating systems, there are a number of improvements that can be made to make these systems more interposition-friendly. We have drawn these lessons from our implementations of the system call and signal dispatchers for Solaris 2.5.1, as well as preliminary analyses of interposing on the virtual memory mechanism and page replacement policy for Solaris and the scheduler interface for FreeBSD 2.1.6, and our experiences in an initial port of SLIC to Linux 2.0.

	Procs	Total System Calls	System calls Caught CERT		Encrypt		Rexec	
AFS	471	40500	5760	14%	2732	7%	9246	23%
TEX	3	2457	126	5%	850	35%	188	8%
gcc	379	140656	43720	31%	11031	8%	47888	34%

Table 3: Benchmark characterization. For each benchmark, this table presents the total number of processes created during a run, the total number of system calls issued by those processes, and the portion of those system calls which are caught by each extension.

	AFS Time (s)	Overhead	TEX Time (s)	Overhead	gcc Time (s)	Overhead
Baseline	15.81		6.36		158.30	
No extensions	16.65	5%	6.69	5%	164.14	4%
CERT	17.51	11%	6.65	5%	163.83	3%
Encrypt	17.45	10%	6.42	1%	166.87	5%
Rexec	17.14	8%	6.48	2%	166.35	5%
CERT + Encrypt + Rexec	17.92	13%	6.78	7%	168.14	6%

Table 4: Benchmark performance on sample extensions. All measurements were run on a 167MHz UltraSPARC running Solaris 2.5. Each column for each benchmark contains both the total elapsed run time in seconds and the percent slowdown. "Baseline" represents a machine without any SLIC dispatchers or extensions loaded. "No extensions" represents the system call dispatcher loaded, but no extensions. The rows labels "CERT", "Encrypt" and "Rexec" (Restricted Execution Environment) present the benchmark elapsed times with a single extension loaded. The last line presents benchmark performance with all three extensions interposing simultaneously.

The problems that we have encountered can be divided into four categories:

1. The asymmetric trust mechanisms of the system call interface;
2. The lack of explicit information in the interfaces we interposed on;
3. An incomplete decomposition of functionality in the system; and
4. Other implementation issues.

5.1 Asymmetric Trust Mechanisms

The system call interface is an untrusted interface; the kernel views all data provided by user applications as potentially malicious and performs security checks to ensure that user processes cannot access secure kernel state. System call extensions must take the same distrustful view of application-supplied data. For example, when an application issues an `open()` system call, it provides the kernel with the address of a buffer containing the file name. Before actually reading data from this buffer, the kernel validates that the buffer is located in the process' address space. However, extensions such as the CERT patch need to perform their own security checks. Security would be violated if an application were able to modify the file name after the CERT extension validated the file name but before the kernel actually performed the `open()`. To prevent this, the CERT extension must first copy the file name into a buffer that the user application cannot access; once validated by the CERT extension, the address of this secure buffer is then passed to the kernel to perform the `open()` operation.

However, because the CERT extension is interposed on the system call interface, the kernel views the extension as part of the untrusted user application. Consequently, the kernel's protection mechanisms will fail and reject this secure buffer address supplied by the CERT extension, aborting the system call with an error.

Resolving this problem requires either modifying or circumventing the kernel's security checks. SLIC interposes on the kernel's `copyin()` and `copyout()` routines and checks for secure buffers allocated via SLIC's support routines, bypassing the kernel security checks when necessary. Naïvely disabling the kernel's security checks can create a potential security hole, as malicious user applications could pass the kernel a pointer to one of these buffers. To prevent this, secure buffers are allocated on a per-thread and a per-event basis. If an extension copies user data into a secure buffer, the `Slic_CopyIn()` routine validates the user pointer before marking the secure buffer as valid for the kernel `copyin()` routine. If no extension modifies the user data, no buffers are marked as secure and the kernel `copyin()` routine will flag any pointer to a kernel address. The overhead imposed by these additional checks is negligible.

5.2 Lack of Explicit Information

Many of the interfaces in today's operating systems are not fully *explicit*: some information is not passed directly as an argument to the event but rather is stored in global data structures. For example, when a system call is performed, the process id of the calling process is not part of the system call invocation. System call error conditions are also not explicit; they are instead stored directly in the kernel's process structure and loaded into the user's registers upon return to user-mode. The Solaris virtual memory system relies heavily on global data structures to determine information such as page ownership.

Our solution to this scattering of information has been to develop utility functions which provide extensions with access to important information such as process identifiers. A number of these utility functions are reasonably complex, and have to walk through internal operating system data structures. A well-formed interface would make this information readily accessible when an event is raised.

While the system call interface is generally quite conducive to interpose on (despite its lack of explicit information), the `ioctl()` call is a notable exception. Originally designed as a way to manage a particular device, `ioctl()` has evolved into the generic method of extending Unix functionality. An `ioctl()` call can pass in arbitrary memory buffers, which may contain pointers to other memory buffers; the structure of each buffer is defined by the particular device driver. This means that an extension cannot know, in advance, how to handle the arguments to an `ioctl()` call. This is problematic for extensions such as system call tracers or security extensions which must understand the arguments of an `ioctl()`. We are currently experimenting with more elaborate interposition on the `copyin()` and `copyout()` routines to capture all reads from user space, in order to identify the location of every buffer read for the `ioctl()` call.

5.3 Separation of Policy and Mechanism

Extensible operating systems papers often argue that the separation of policy and mechanism is essential for maintaining extensibility. Our experiences agree with this contention. While experimenting with extending the scheduler interface in FreeBSD 2.1.6, we found that the FreeBSD scheduler routine (`cpu_switch()`) implements both the policy of selecting the next process to run as well as the actual context switch mechanism. To enable interposition on the scheduler policy, we relocated the code which selects the next process to run into a separate routine, creating a procedural interface which can be interposed on.

This problem also occurs in the kernel `copyin()` and `copyout()` routines, which combine the kernel's security policy with a mechanism for bringing in data from user space. In Solaris, we can interpose on these two routines and implement our own security mechanism for SLIC secure buffers. However, the Linux versions of these routines (`getuser()` and `putuser()`) are normally inlined, and interposing on inlined routines is significantly more difficult than on standalone functions (which need only be modified once). Because of this increased difficulty, our Linux port modifies the Linux source to disable inlining of the copy routines.

5.4 Other Issues

Our method of using binary patching to intercept procedural invocations was simple to implement. However, intercepting procedural interface events using binary patching requires a writable kernel text. Unfortunately, Solaris is loaded such that the kernel text is read-only. By modifying a single line of Solaris source, we were able to make the text writable. In earlier versions of Solaris, a system configuration file enabled kernel writability, but this functionality has recently been removed.

The dispatchers and extensions often need a way of recording state that persists across event invocations. For instance, the system call dispatcher needs to keep track of which processes are marked for tracing and a virtual memory extension may need to store per page information. The traditional place to store this information is in the process table or the page structures, but in many operating systems, administrative tools rely on the size and organization of these structures to remain constant. Consequently, we currently implement shadow structures of the kernel's process and thread structures to store information for the system call and signal dispatchers and the extensions.

Extensions operate with the privileges of the calling thread when they are invoked. Consequently, an extension cannot write to files owned by root or signal root-owned processes, except when invoked from a root process. We are currently investigating this problem and solutions to it.

5.5 Lessons for Kernel Developers

Given these experiences, there are a number of lessons which we learned about making operating system interfaces conducive to interposition. First, when an interface defines a boundary between trusted and untrusted code, it should be possible to keep the two distinct. For example, isolating the security checks into a collection of well-defined functions (which Solaris does but Linux does not) allows these checks to be interposed on and augmented.

Second, it is important to distinguish between mechanism and policy, and to enforce this distinction even in the compiled binary. This enables extensions to replace system policies while still leveraging the mechanisms of the system. Although this solution may impact performance slightly, the costs involved with most mechanisms and policy decisions in today's production operating systems considerably outweigh the relatively low costs of a procedure call.

Third, a full set of functions should be provided which perform the side effects required by the system. For instance, Solaris provides a function to set the error value for a system call, but does not provide a function to clear that error, which some extensions may need to do. Currently dispatchers provide such functionality in utility functions, but doing so requires an understanding of many details of the system. Providing these functions in the underlying kernel will make dispatcher implementations simpler and more robust.

Finally, a simple method should be provided to enable dispatchers to write to the kernel code image. This is necessary to enable interposition on procedural interfaces using binary patching.

6 Related Work

There has been a considerable amount of recent work [17, 45, 5, 20, 21] that has focused on building highly-extensible operating systems. Because these systems have built new operating systems kernels and new kernel structures, they have not focused on the problem of extensibility for existing systems. Of these systems, SPIN [5] and VINO [45, 41] are the closest in concept to our work. Both offer extensibility through interposition on a number of kernel interfaces. These interfaces have been explicitly designed for extensibility, rather than enabling extensibility on existing interfaces, which SLIC supports. SPIN and VINO also aggressively focus on ensuring kernel protection from extensions, SPIN by using a type-safe language [44, 26], and VINO through software fault isolation [50] and in-kernel transactions [41]. We assume trusted extensions.

Interposition Agents [27] demonstrated that it is useful to construct interposition extensions in terms of the underlying abstractions of the interposed interface, rather than in terms of the physical events crossing that interface. The interposition technology used in [27] bounced system calls to extensions linked into an application's address space. This mechanism has two disadvantages relative to SLIC. First, extensions are not protected from applications and thus cannot implement security extensions or share data between distrustful applications. Second, the multiple protection boundary crossings limit the performance of the system. Our interposition technology enables high-performance interposition that is both enforced upon, and protected from applications, enabling a larger class of extensions. The actual toolkit presented in [27] could easily be constructed on our interposition platform and could simplify the process of extension development.

COLA [31] enables interposition at the system call interface, but without any modification of the operating system kernel. It operates through interposition at the library level and consequently suffers from the same security drawbacks as the interposition technology used in [27].

Protected Shared Libraries [4] enables extensions to be securely loaded into an application's address space, so that user programs cannot access or modify extension code or data. Protected Shared Libraries is primarily used for adding new interfaces to a system, although they could be combined with the interposition mechanisms used in SLIC to enable modification of existing interfaces. Their protection techniques rely on hardware features of the IBM RS/6000 architecture, while the principles in SLIC are generally applicable across a variety of operating system platforms.

7 Future Work

Work on SLIC is proceeding in a variety of directions. We are focusing on extending the functionality of the base system, such as adding support for interposition at the page replacement policy and page replacement mechanism interfaces of Solaris 2.5. In addition, we are working toward making the base system portable to a number of other platforms, and on developing a large number of useful extensions.

An area of active research is in managing conflicts between extensions. While interposition enables composable extensions, these extensions may behave in ways that negatively impact system stability. Prior experiences with with conflicts among MS-DOS terminate-and-stay-resident (TSR) utilities and Macintosh toolkit extensions indicate that a method of managing conflicts among extensions is sorely needed.

A Linux 2.0 port of the base SLIC system is nearly complete. Because the base system encapsulates a significant amount of functionality, extensions written for the Solaris version of SLIC should port with few or no modifications to the Linux version.

We are also exploring the design space of useful extensions to be built with SLIC. Among these are transparent remote execution, a single system image in a network of workstations, extending the semantics of NFS to include full cache-coherency for simultaneous access [46] and supporting scheduler activations [3] through interposition on signals or directly on the scheduler.

8 Conclusion

This paper has examined the utility of interposition as a mechanism for adding extensibility to production operating systems. We have shown that interposition is suitable to a number of useful extensions, and we have presented a prototype system, SLIC, which enables operating system extensions through interposition in Solaris with minimal kernel source modifications. SLIC demonstrates that extending an existing operating system can be done efficiently and securely, enabling a larger class of extensions than previous work in this area.

We have also examined the problems found in transparently extending operating system functionality, such as the asymmetric trust found across the system call interface. Drawing from experiences with these problems, we presented a number of lessons that can be used by operating systems designers to provide interfaces which are conducive to interposition. Foremost is the imperative to maintain clear procedural barriers between operating system policy and mechanism. Additionally, in order to reduce the effort necessary in implementing an interposition mechanism, extension interfaces should be explicit and expose all information related with an event.

We believe that the techniques and technology we have described in this paper can provide substantial benefits to users of existing operating systems, enabling a viable third-party industry in operating system extensions. The resulting competition will stimulate innovation and increase the rate of technology transfer from operating systems research into production systems.

Availability

SLIC is implemented on Solaris 2.5.1 and a port to Linux 2.0 is in progress. Current status and source code is available at `http://now.cs.berkeley.edu/Slic/`

References

[1] M. Accetta, R. Baron, W. Bolosky, D. Golub, R. Rashid, A. Tevanian, and M. Young. Mach: A New Kernel Foundation For UNIX Development. In *Proceedings of the 1986 USENIX Summer Conference*, pages 93–112, June 1986.

[2] C. Amza, Alan L. Cox, Sandhya Dwarkadas, Peter Keleher, H. Lu, R. Rajamony, Weimin Yu, and Willy Zwaenepoel. TreadMarks: Shared memory computing on networks of workstations. *IEEE Computer*, 29(2):18–28, February 1996.

[3] Thomas Anderson, Brian Bershad, Edward Lazowska, and Henry Levy. Scheduler Activations: Effective Kernel Support for the User-Level Management of Parallelism. In *ACM Transactions on Computer Systems*, pages 53–79, February 1992.

[4] Arindam Banerji, John M. Tracey, and David L. Cohn. Protected Shared Libraries — a new approach to modularity and sharing. In *Proceedings of the 1997* USENIX *Technical Conference*, pages 59–76, Anaheim, CA, January 1997.

[5] B. N. Bershad, S. Savage, E. G. Sirer P. Pardyak, M. Fiuczynski, D. Becker, C. Chambers, and S. Eggers. Extensibility, Safety and Performance in the SPIN Operating System. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, December 1995.

[6] Brian Bershad, Thomas Anderson, Edward Lazowska, and Henry Levy. Lightweight Remote Procedure Calls. In *ACM Transactions on Computer Systems*, pages 37–54, February 1990.

[7] Brian N. Bershad, Craig Chambers, Susan Eggers, Chris Maeda, Dylan McNamee, Przemysław Pardyak, Stefan Savage, and Emin G. Sirer. SPIN—An Extensible Microkernel for Application-Specific Operating System Services. Technical report, University of Washingtion, 1994.

[8] C. Mic Bowman, Peter B. Danzig, Darren R. Hardy, Udi Manber, and Michael F. Schwartz. The Harvest information discovery and access system. In *Proceedings of the Second International World Wide Web Conference*, pages 763–771, October 1994.

[9] Kraig Brockschmidt. *Inside OLE 2*. Microsoft Press, Redmond, WA, 1994.

[10] Vulnerability in expreserve. CERT Advisory CA-96.19, CERT, August 1996.

[11] Vulnerability in Solaris admintool. CERT Advisory CA-96.16, CERT, August 1996.

[12] Vulnerability in WorkMan. CERT Advisory CA-96.23, CERT, October 1996.

[13] Vulnerability in talkd. CERT Advisory CA-97.04, CERT, January 1997.

[14] David D. Clark. The Structuring of Systems Using Upcalls. In *Proceedings of the 10th ACM Symposium on Operating Systems Principles*, pages 171–180, December 1–4 1985.

[15] Chaos Computer Club. CCC: Microsoft security alert. http://berlin.ccc.de/radioactivex.html, March 1997.

[16] Fred Douglis and John Ousterhout. Transparent Process Migration: Design Alternatives and the Sprite Implementation. *Software - Practice and Experience*, 21(8):757–85, August 1991.

[17] Peter Druschel, Larry L. Peterson, and Norman Hutchinson. Beyond Micro-Kernel Design: Decoupling Modularity and Protection in Lipto. In *Proceedings of the 12th International Conference on Distributed Computing Systems*, pages 512–520, June 1992.

[18] Timothy Dyck. Informix slices its way to front. *PC Week*, 14(9):1, March 1997.

[19] Mark W. Eichin and Jon A. Rochlis. With Microscope and Tweezers: An Analysis of the Internet Virus of November 1988. In *Proceedings of the 1989 IEEE Symposium on Research in Security and Privacy*, 1989.

[20] D. R. Engler, M. F. Kaashoek, and Jr J. O'Toole. Exokernel: An Operating System Architecture for Application-Level Resource Management. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles*, December 1995.

[21] Bryan Ford, Mike Hibler, Jay Lepreau, Patrick Tullman, Godmar Back, and Steven Clawson. Microkernels Meet Recursive Virtual Machines. In *Proceedings of the 2nd USENIX Symposium on Operating Systems Design and Implementation*, October 1996.

[22] Simon Garfinkel. *PGP: Pretty Good Privacy*. O'Reilly and Associates, Sebastopol, CA, first edition, December 1994.

[23] Ian Goldberg, David Wagner, Randy Thomas, and Eric Brewer. A Secure Environment for Untrusted Helper Applications. In *Proceedings of the Sixth USENIX Security Symposium*, July 1996.

[24] James Gosling and Henry McGilton. The Java(tm) Language Environment: A White Paper. http://java.dimensionx.com-/whitePaper/java-whitepaper-1.html, 1995.

[25] J. Howard, M. Kazar, S. Menees, D. Nichols, M. Satyanarayanan, R. Sidebotham, and M. West. Scale and performance in a distributed file system. *ACM Transactions on Computer Systems*, 6(1):51–81, February 1988.

[26] Wilson Hsieh, Marc Fiuczynski, Charles Garrett, Stefan Savage, David Becker, and Brian N. Bershad. Language support for extensible operating systems. In *Proceedings of the Workshop on Compiler Support for System Software*, February 1996.

[27] Michael B. Jones. Interposition Agents: Transparently Interposing User Code at the System Interface. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles*, pages 80–93, December 1993.

[28] Yousef Khalidi and Michael Nelson. Extensible File Systems in Spring. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles*, pages 1–14, December 1993.

[29] Gregor Kiczales, Jim des Rivières, and Daniel G. Bobrow. *The Art of the Metaobject Protocol*. The MIT Press, Cambridge, MA, 1991.

[30] Steven R. Kleiman. Vnodes: An architecture for multiple file system types in Sun UNIX. In *Proceedings of the 1986 Summer* USENIX *Technical Conference*, pages 238–247, Atlanta, GA, 1986. Sun Microsystems.

[31] Eduardo Krell and Balachander Krishnamurthy. COLA: Customized overlaying. In *Proceedings of the Winter 1992* USENIX *Technical Conference*, San Francisco, CA, January 1992. AT&T Bell Laboratories, USENIX Association.

[32] Kai Li and Paul Hudak. Memory Coherence in Shared Virtual Memory Systems. *ACM Transactions on Computer Systems*, 7(4):321–359, November 1989.

[33] Netscape Communications Corporation. *Netscape Navigator*, 1994. http://www.netscape.com.

[34] John Ousterhout. Why aren't operating systems getting faster as fast as hardware? In *Proceedings of the 1990 Summer* USENIX *Conference*, pages 247–256, Anaheim, CA, June 1990.

[35] Przemysław Pardyak and Brian N. Bershad. Dynamic binding in an extensible system. In *Proceedings of the Second Symposium on Operating Systems Design and Implementation (OSDI '96)*, pages 201–212, Seattle, WA, October 1996. University of Washington, USENIX Association.

[36] Pure Software, Inc. *Purify User's Guide*, 1996.

[37] Pure Software, Inc. *Quantify User's Guide*, 1996.

[38] Xtensions.com:the ultimate source for quark xtensions. `http://www.xtension.com/`, 1997.

[39] David S. H. Rosenthal. Evolving the vnode interface. In *Proceedings of the 1990 Summer* USENIX *Technical Conference*, pages 107–117, Anaheim, CA, June 1990. Sun Microsystems.

[40] R. Sandberg, D. Goldberg, Steven Kleiman, D. Walsh, and B. Lyon. Design and implementation of the Sun network filesystem. In *Proceedings of the Summer 1985* USENIX *Technical Conference*, pages 119–130, Dallas, TX, June 1985. Sun Microsystems.

[41] Margo I. Seltzer, Yasuhiro Endo, Christoper Small, and Keith A. Smith. Dealing With Disaster: Surviving Misbehaved Kernel Extensions. In *Proceedings of the 2nd USENIX Symposium on Operating Systems Design and Implementation*, October 1996.

[42] Margo I. Seltzer, Yasuhiro Endo, Christopher Small, and Keith A. Smith. Issues in extensible operating systems. Publication pending from IEEE Press, 1997. Available from `http://www.eecs.harvard.edu/~vino/vino/papers/acm-96.ps`.

[43] Emin Gün Sirer, Marc Fiuczynski, Przemysław Pardyak, and Brian N. Bershad. Safe dynamic linking in an extensible operating system. In *Proceedings of the Workshop on Compiler Support for System Software*. University of Washington, February 1996.

[44] Emin Gün Sirer, Stefan Savage, Przemysław Pardyak, Greg DeFouw, Mary Ann Alapat, and Brian N. Bershad. Writing an operating system using Modula-3. In *Proceedings of the Workshop on Compiler Support for System Software*. University of Washington, February 1996.

[45] Christopher Small and Margo Seltzer. VINO: An Integrated Platform for Operating System and Database Research. Technical Report TR-30-94, Harvard, October 1994.

[46] V. Srinivasan and Jeffrey C. Mogul. Spritely NFS: Experiments with cache-consistency protocols. In *Proceedings of the 12th ACM Symposium on Operating Systems Principles*, pages 45–57, Litchfield Park, AZ, December 1989. ACM.

[47] David C. Steere, James J. Kistler, and M. Satyanarayanan. Efficient user-level file cache management on the Sun vnode interface. In *Proceedings of the Summer 1990* USENIX *Technical Conference*, pages 325–332, Anaheim, CA, June 1990. Carnegie Mellon University, USENIX Association.

[48] Marvin Theimer, K. Landtz, and David Cheriton. Preemptable Remote Execution Facilities for the V System. In *Proceedings of the 10th ACM Symposium on Operating Systems Principles*, pages 2–12, December 1985.

[49] Thorsten von Eicken, David E. Culler, Steh C. Goldstein, and Klaus Erik Schauser. Active Messages: a Mechanism for Integrated Communication and Computation. In *Proc. of the 19th Int'l Symposium on Computer Architecture*, May 1992.

[50] Robert Wahbe, Steven Lucco, Thomas Anderson, and Susan Graham. Efficient Software-Based Fault Isolation. In *Proceedings of the 14th ACM Symposium on Operating Systems Principles*, pages 203–216, December 1993.

[51] Neil Webber. Operating System Support for Portable Filesystem Extensions. In *Proceedings of the 1993 USENIX Winter Conference*, pages 219–228, January 1993.

[52] Nick Wingfield. ActiveX used as hacking tool. `http://www.news.com/ News/Item/0%2C4%2C7761%2C00.html`, February 1997.

[53] W. Wulf, E. Cohen, W. Corwin, A. Jones, R. Levin, C. Pierson, and F. Pollack. HYDRA: The Kernel of a Multiprocessor Operating System. *Communications of the ACM*, 17(6):337–344, June 1974.

[54] M. Young, A. Tevanian, R. Rashid, D. Golub, J. Eppinger, J. Chew, W. Bolosky, D. Black, and R. Baron. The Duality of Memory and Communication in the Implementation of a Multiprocessor Operating System. In *Proceedings of the 11th ACM Symposium on Operating Systems Principles*, pages 63–76, November 1987.

[55] G. Pascal Zachary. *Showstopper! The Breakneck Race to Create Windows NT and the Next Generation at Microsoft*. Macmillan, Inc., 1994.

[56] Sognian Zhou, Jingwen Wang, Xiaohn Zheng, and Pierre Delisle. Utopia: A Load Sharing Facility for Large, Heterogeneous Distributed Computing Systems. Technical Report CSRI-257, University of Toronto, 1992.