Processes in KaffeOS:
Isolation, Resource Management, and Sharing in Java

Godmar Back   Wilson C. Hsieh   Jay Lepreau

Department of Computer Science, University of Utah
http://www.cs.utah.edu/flux/

Technical Report UUCS–00–010

April, 2000

Abstract

Single-language runtime systems, in the form of Java virtual machines, are widely deployed platforms for executing untrusted mobile code. These runtimes provide some of the features that operating systems provide: inter-application memory protection and basic system services. They do not, however, provide the ability to isolate applications from each other, or limit their resource consumption. This paper describes KaffeOS, a system that provides these features for a Java runtime. The KaffeOS architecture takes many lessons from operating system design, such as the use of a user/kernel boundary.

The KaffeOS architecture supports the OS abstraction of a *process* in a Java virtual machine. Each process executes as if it were run in its own virtual machine, including separate garbage collection of its own heap. The difficulty in designing KaffeOS lay in balancing the goals of isolation and resource management against the goal of allowing direct sharing. Overall, KaffeOS is up to 11% slower than the freely available JVM on which it is based, which is an acceptable penalty for the safety that it provides. KaffeOS is substantially slower than commercial JVMs, but exhibits much better performance scaling in the presence of uncooperative code.

1 Introduction

The need to support the safe execution of untrusted programs in runtime systems for type-safe languages has become clear. Language runtimes are being used in environments for executing untrusted code: for example, applets, servlets, active packets [38], database queries [13], and kernel extensions [5]. Current systems (such as Java) provide *memory protection* through the enforcement of type safety and *secure system services* through a number of mechanisms, including namespace and access control. Unfortunately, malicious or buggy applications can deny service to other applications. For example, a Java applet could generate excessive amounts of garbage and cause a Web browser to spend all of its time collecting dead objects.

To support the execution of untrusted code, type-safe language runtimes need to provide mechanisms to isolate and manage the resources of applications, analogous to those provided by operating systems. Although other resource management abstractions exist [3], the classic OS *process* abstraction seems appropriate. A process is the basic unit of resource ownership and control; it provides isolation between applications. On a traditional operating system, untrusted code can be forked as its own process; CPU and memory limits can be placed on the process, and the process can be killed if it is uncooperative. Current type-safe language runtimes do not support such functionality.

A number of approaches to isolating applications in Java have been developed by others over the last few years. An *applet context* [7] is an example

This research was largely supported by the Defense Advanced Research Projects Agency, monitored by the Air Force Research Laboratory, Rome Research Site, USAF, under agreements F30602–96–2–0269 and F30602–99–1–0503.

of an application-specific approach. It provides a separate namespace and a separate set of execution permissions for untrusted applets. Applet contexts do not support resource management, and cannot defend against denial-of-service attacks. In addition, they are not general: applet contexts are specific to applets, and cannot be used easily in other environments.

Several general-purpose models for managing resources in Java exist, such as the J-Kernel [21] or Echidna [19]. However, these solutions superimpose an operating system kernel on Java without changing the underlying virtual machine. As a result, it is impossible in those systems to account for resources spent on behalf of a given application: for example, CPU time spent while garbage collecting a process' heap.

An alternative approach to achieve isolation between different applications is to give each one its own virtual machine, and run each virtual machine in a different process in an underlying OS [23, 27]. For instance, most operating systems can limit a process's heap size or CPU consumption. Such mechanisms could be used to directly limit an entire VM's resource consumption, but they depend on underlying operating system support.

Designing JVMs to support multiple processes is a superior approach. First, it reduces per-application overhead. For example, applications on KaffeOS can share classes in the same way that an OS allows applications to share libraries. Second, communication between processes can be more efficient in one VM, since objects can be shared directly. (One of the reasons for using type-safe language technology in systems such as SPIN [5] was to reduce the cost of IPC; we want to keep that goal.) Third, embedding a JVM in a larger process, such as a web server or web browser, is difficult (or impossible) if the JVM relies on an operating system to isolate different activities. Fourth, embedded or portable devices may not provide OS or hardware support for managing processes. Finally, a single JVM uses less energy than multiple JVM's on portable devices [17].

Our work consists of supporting processes in a modern type-safe language, Java. Our solution, KaffeOS, adds a process model to Java that allows a JVM to run multiple untrusted programs safely, and still support the direct sharing of resources between programs. The difficulty in designing KaffeOS lay in balancing conflicting goals: process isolation and resource management versus direct sharing of objects between processes.

A KaffeOS process is a general-purpose mechanism that can easily be used in multiple application domains. For instance, KaffeOS could be used in a browser to support multiple applets, within a server to support multiple servlets, or even to provide a standalone "Java OS" on bare hardware. We have structured our abstractions and APIs so that they are as broadly applicable as possible, much as the OS process abstraction is. Because the KaffeOS architecture is designed to support processes, we have taken lessons from the design of traditional operating systems, such as the use of a user/kernel boundary.

Our design makes KaffeOS's isolation and resource control mechanisms comprehensive. We focus on the management of CPU time and memory, although we plan to address other resources such as network bandwidth. The runtime system is able to account for and control all of the CPU and memory resources consumed on behalf of any process. We have dealt with these issues by structuring the KaffeOS virtual machine so that it separates the resources used by different processes.

To summarize, this paper makes the following contributions:

- We describe how lessons from building traditional operating systems can and should be used to structure runtime systems for type-safe languages.

- We describe how software mechanisms in the compiler and runtime can be used to implement isolation and resource management in a Java virtual machine.

- We describe the design and implementation of KaffeOS. KaffeOS implements our process model in Java, which isolates applications from each other, provides resource management mechanisms for them, and also lets them share resources directly.

- We show that the performance penalty for using KaffeOS is reasonable, compared to the freely available JVM on which it is based. Even though KaffeOS is substantially slower than commercial JVMs, it exhibits much better performance scaling in the presence of uncooperative code.

Sections 2 and 3 describe the design and implementation of KaffeOS, respectively. Section 4 provides some performance measurements of KaffeOS, and compares its performance with that of some commercial Java virtual machines. Section 5 describes related work in more detail, and Section 6 summarizes our conclusions and results.

2 Design Principles

The following principles drove our design of KaffeOS, in decreasing order of importance:

- **Process separation**. We provide the "classical" property of a process: each process is given the illusion of having the whole virtual machine to itself.

- **Safe termination of processes**. Processes may terminate abruptly due to either an internal error or an external event. In both cases, we ensure that the integrity of other processes and the system itself is not violated.

- **Direct sharing between processes**. Processes can directly share objects in order to communicate with each other.

- **Precise memory and CPU accounting**. The memory and CPU time spent on almost all activities can be attributed to the application on whose behalf it was expended.

- **Full reclamation of memory**. When a process is terminated, its memory must be fully reclaimed. In a language-based system, memory cannot be revoked by unmapping pages: it must be garbage-collected. We restrict a process' heap writes to avoid uncollectable memory in the presence of direct object sharing.

- **Hierarchical memory management**. Memory allocation can be managed in a hierarchy, which provides a simple model for controlling processes.

The interaction between these design principles is complex. For expository purposes, we discuss these principles in a slightly different order in the remainder of this section.

Process separation. A process cannot accidentally or intentionally access another process' data, because each process has its own heap. Each process is given its own name space for its objects, as well. Type safety provides memory protection, so that a process cannot access other processes' objects.

To ensure process separation, an untrusted process is not allowed to hold onto system-level resources indefinitely. For instance, global kernel locks are not directly accessible to user processes. Violations of this restriction are instances of bad system design. Similarly, faults in one process must not impact progress in other processes.

Safe termination of processes. KaffeOS is structured such that critical parts of the system cannot be damaged when a process is terminated. For example, a process is not allowed to terminate when it is holding a lock on a system resource.

We divide KaffeOS into user and kernel parts [1], an important distinction used in operating system design. A user/kernel distinction is necessary to maintain system integrity in the presence of process termination. Others have suggested that depending on language-level exception handling is insufficient. We disagree, because exceptions interact poorly with code in critical sections.

Figure 1 illustrates the high-level structure of KaffeOS. User code executes in "user mode," as do some of the trusted runtime libraries and some of the garbage collection code. The remaining parts

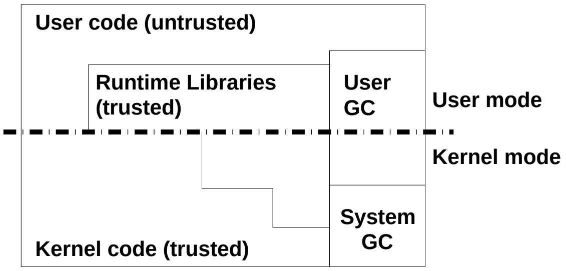

Figure 1: Structure of KaffeOS. System code is divided into kernel and user modes; user code all runs in user mode. In user mode, code can be terminated arbitrarily; in kernel mode, code cannot be terminated arbitrarily.

of the system (the rest of the runtime libraries and the garbage collector, as well as the virtual machine itself) must run in kernel mode to ensure their integrity. Note that "user mode" and "kernel mode" do not indicate a change in hardware privileges. Instead, they indicate different environments with respect to termination and resource consumption:

- Processes running in user mode can be terminated at will. Processes running in kernel mode cannot be terminated at an arbitrary time, because they must leave the kernel in a clean state.

- Resources consumed in user mode are always charged to a user process, and not to the system as a whole. Only in kernel mode can a process consume resources that are charged to the entire system, although typically such use is charged to the appropriate user process.

Such a structure echoes that of exokernels [16], where system-level code executes as a user-mode library. Note that a language-based system allows the kernel to trust user-mode code to a great extent, because type safety prevents user code from damaging any user-mode system code.

The KaffeOS kernel is structured so that it can handle termination requests and internal errors cleanly. Termination requests are deferred, so that a process cannot be terminated while manipulating kernel data structures. Kernel code does not abruptly terminate due to internal exceptions, for the same reason. Violations of these two restrictions are considered kernel bugs.

Full reclamation of memory. Since Java is typesafe, it does not provide a primitive to reclaim memory. Instead, unreachable memory is freed by a garbage collector. We use the garbage collector to recover all the memory of a process when it terminates. Therefore, we must prevent situations where the collector cannot free a terminated process' objects because another process still holds references to them.

We use techniques from distributed garbage collection schemes [29] to restrict cross-process references, although we use them to very different effect. Distributed GC seeks to overcome the physical separation of machines and create the impression of a global shared heap. We use distributed GC mechanisms to manage multiple heaps in a single address space, so that they can be collected independently.

We use *write barriers* [40] to restrict writes. A write barrier is a check that happens on every pointer write to the heap. As we show in Section 4, the cost of using write barriers, although non-negligible, is reasonable. Illegal cross-references are those that would prevent a process' memory from being reclaimed: for example, references from one user heap to another. Since those references cannot exist, it is possible to reclaim a process' heap as soon as the process is terminated. Writes that would create illegal cross-references are forbidden, and raise exceptions. We call such exceptions "segmentation violations." Although it may seem surprising that a type-safe language runtime could throw such a fault, it actually closely follows the analogy to traditional operating systems.

Unlike distributed garbage collection, in KaffeOS inter-heap cycles are not an issue. The only form of inter-heap cycles that can occur are due to data structures that are split between a user heap and the kernel heap, since there can be no cycles that span multiple user heaps. Writes of user-heap references to kernel objects can only be done by trusted code. The kernel is coded so that it only writes a user-heap reference to a kernel object whose lifetime equals that of the user process: for example, the object that represents the process itself.

KaffeOS is intended to run on a wide range of systems. We do not assume that the platforms on which it runs will necessarily have a hardware memory management unit under the control of KaffeOS. Neither do we assume that the host has an operating system that supports virtual memory. For example, a Palm Pilot violates both of these assumptions. Without these assumptions, memory can not simply be revoked by unmapping it.

Precise memory and CPU accounting. We account for memory and CPU on a per-process basis, so as to limit their consumption by buggy or possibly malicious code. In order to prevent denial-of-service attacks, it is necessary to minimize the amount of time and memory spent servicing kernel requests.

Memory accounting is complete. It applies not only to objects at the Java level, but to all allocations done in the VM on behalf of a given process. In contrast, bytecode-rewriting approaches that do not modify the virtual machine, such as Jres [11, 12], can only account for object allocations.

We try to minimize the number of objects that are allocated on the kernel heap through careful coding of the kernel interfaces. For instance, consider a system call that creates a new process with a new heap: the process object itself, which is large, is allocated on the new heap. The handle that is returned to the creating process to control the new process is allocated on the creating process' heap. The kernel heap only maintains a small entry in a process table.

We increase the accuracy of CPU accounting by minimizing the time spent in non-preemptible sections of code. In addition, distributed GC also separates the garbage collection of the user heaps and the kernel heap. We again use write barriers to detect cross-references from a user to the kernel heap, and vice versa. For each such reference, we create an *entry item* in the heap to which it points [29]. In addition, we create a special *exit item* in the original heap to remember the entry item created in the destination heap. Unlike distributed object systems such as Emerald [24], entry and exit items are not used for naming non-local objects; we only use them to decouple the garbage collection of different heaps.

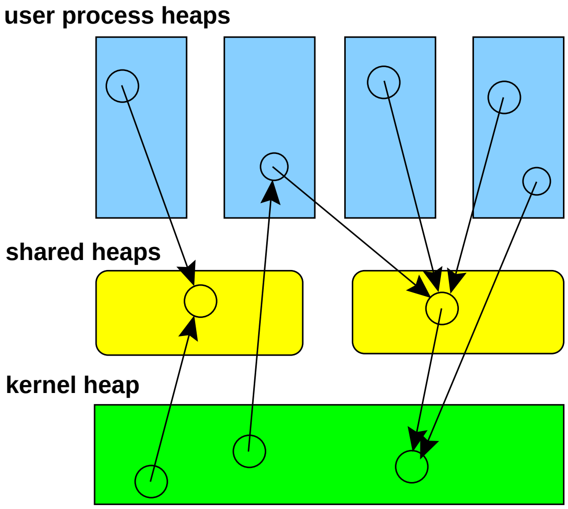

Figure 2: Heap structure in KaffeOS. The *kernel heap* can contain pointers into the user heaps, but the *shared heaps* and other user heaps cannot. User heaps can contain pointers into the kernel heap and shared heaps.

Entry items are reference counted: they keep track of the number of exit items that point to them. The reference count of an entry item is decremented when an exit item is garbage collected. If an entry item's reference count reaches zero, the entry item is removed, and the referenced object can be garbage collected if it is not reachable through some other path.

A process' memory is reclaimed upon termination by merging its heap with the kernel heap. All exit items are destroyed at this point and the corresponding entry items are updated. The kernel heap's collector can then collect all of the memory, including memory on the kernel heap that was kept alive by the process. User-kernel cycles of garbage objects can be collected at this time.

Direct sharing between processes. One of the reasons for using a language-based system is to allow for direct communication between applications. For example, the SPIN operating system allowed kernel extensions to communicate directly through pointers to memory. The design of KaffeOS retains this design principle. Figure 2 shows the different heaps in KaffeOS, and the kinds of inter-heap pointers that are legal.

In KaffeOS, a process can dynamically create a shared heap to communicate with other processes. Objects allocated in a shared heap are restricted: their non-primitive fields (i.e., pointer fields) cannot be reassigned after initialization. Only the primitive fields in a shared object are mutable. This restriction is enforced by write barriers.

A shared heap has the following lifecycle. First, one process picks one or more shared types out of a central shared namespace, creates the heap and loads the shared class or classes into it. While the heap is being created, the creator is charged for the whole heap. After the heap is populated with classes and objects, it is frozen and its size remains fixed for its lifetime. If other processes look up the shared heap, they are charged that amount. In this way, all sharers are charged for the heap. Processes exchange data by writing into and reading from the shared objects and by synchronizing on them in the usual way.

If a process drops all references to a shared heap, all exit items to that shared heap become unreachable. After the process garbage collects the last exit item to a shared heap, that shared heap's memory is credited to the sharer's budget. When the last sharer drops all references to a shared heap, the shared heap becomes orphaned. The kernel garbage collector checks for orphaned shared heaps at the beginning of each GC cycle and merges them into the kernel heap.

This model guarantees three properties:

- All sharers are charged in full for a shared heap while they are holding onto the shared heap. As a result, sharers do not have to be charged asynchronously if another sharer exits. (If n sharers were each to pay only $1/n$ of the cost of a shared heap, when one sharer exited the others would have to be asynchronously charged $(1/n - 1) - (1/n)$ of the cost.)

- One process cannot use a shared object to keep objects in another process alive. Making the non-primitive fields of shared objects immutable is the simplest means of enforcing this restriction.

- Sharers are charged accurately for all metadata, such as internal class data structures. The metadata is also allocated on the shared heap.

Although we ensure that process heaps can be scanned independently during GC, thread stacks still need to be scanned during GC for inter-heap references. Incremental schemes could be used to eliminate repeated scans of a stack [10], and a thread does not need to be scanned more than once while it is suspended. Some "GC crosstalk" between processes is still possible, because a process could create many threads in an effort to get the system to scan them all. We decided that the benefit of allowing direct sharing between processes is worth leaving open such a possibility.

Hierarchical memory management. We provide a simple hierarchical model for managing memory. Each heap is associated with a *memlimit*, which consists of an upper limit and a current use. Memlimits form a hierarchy: each one has a parent, except for a root memlimit. All memory allocated to the heap is debited from that memlimit, and memory collected from that heap is credited to the memlimit. This process of crediting/debiting is applied recursively to the node's parents.

A memlimit can be *hard* or *soft*. This attribute influences how credits and debits percolate up the hierarchy of memlimits. A hard memlimit's maximum limit is immediately debited from its parent, which amounts to setting memory aside. Credits and debits are therefore not propagated past a hard limit. A soft memlimit's maximum limit, on the other hand, is just a limit—credits and debits of a soft memlimit's current usage are reflected in the parent.

Hard and soft limits allow different memory management strategies. Hard limits allow for memory reservations, but incur inefficient memory use if the limits are not used. Memory consumption matters, because we do not assume there is an underlying operating system; as a result, KaffeOS may manage physical memory. Soft limits allow the setting of a summary limit for multiple activities without incurring the inefficiences of hard limits. They can be used to guard malicious or buggy applications where temporarily high memory usage can be tolerated.

Another application of soft limits is during the creation of shared heaps. Those heaps are initially associated with a soft memlimit that is a child of the current process heap's memlimit. In this way, they are separately accounted but still subject to their creator's memlimit, which ensures that they cannot grow to exceed their creator's ability to pay.

3 Implementation Issues

The KaffeOS VM is built on top of the freely available Kaffe virtual machine, version 1.0b4 [39], which is roughly equivalent to JDK 1.1. In this section, we describe the Java-specific issues that had to be dealt with in implementing KaffeOS. Many implementation decisions were driven by the goal of modifying Java as little as possible.

The primary purpose of KaffeOS is to run Java applications, which expect a well-defined environment of run-time services and libraries. We provide the standard Java API within KaffeOS.

We make use of various features of Java to support KaffeOS processes: Java class loaders, in particular, deserve some discussion. We also discuss some of the changes to the Java runtime. Finally, we discuss some aspects of the Kaffe implementation that affect the performance that we can achieve with our KaffeOS prototype.

3.1 Namespaces

Separate namespaces are provided in Java through the use of class loaders [26]. A class loader is an object that acts as a name server for classes. We use the Java class loading mechanism to provide KaffeOS processes with different namespaces. This use of Java class loaders is not novel, but is important because we have tried to make use of existing Java mechanisms when possible. When we use standard Java mechanisms, we can easily ensure that we do not violate the language semantics.

KaffeOS also uses class loaders to manage the namespace of shared objects. Process loaders delegate the loading of shared class to a shared loader, which means that all shared objects have well-understood types for all user processes. If we were not able to use delegation, KaffeOS would need to support a much more complicated type system. On the downside, the shared namespace becomes a global resource, which is harder to account for precisely.

3.2 Java Class Libraries

We examined each class in the Java standard libraries [8] to see how it interacted under the semantics of class loading. A class's members and their associated code are described by a sequence of bytes in a *class file*. Classes from identical class files that are loaded [26] by different class loaders are defined to be different in Java, although they have may identical behavior relative to the namespace defined by the loader that loaded them. We refer to such classes as *reloaded* classes. Reloading a class gives each instance its own copies of static fields. In KaffeOS, Java classes could be reloaded; they could be modified to be shared across processes; or they could be used unchanged. For each class, we decided which alternative to choose, subject to two goals: to share as many classes as possible, but to make as few code changes as necessary.

Certain classes must be shared between processes to support the use of the shared heaps or the kernel heap. By "shared class," we mean that the class is the same in both processes, and that the text is shared. If the class uses statics, these statics must be replaced with process-aware implementations if they cannot be eliminated. For example, the `java.lang.Object` class, which is the superclass of all object types, must be shared. If this type were not shared, it would not be possible for different processes to share generic objects! Because of some details of Java class loading, globally shared classes cannot directly refer to reloaded classes.

Non-shared classes should always be reloaded, so that each process gets its own instance. Reloaded classes do not share text in our current implementation, although they could. Because of some unfortunate decisions in the Java API design, some classes export static members as part of their public interface, which forces those classes to be reloaded. For example, `java.io.FileDescriptor` must be reloaded, because it exports the public static variables `in`, `out`, and `err` (stdin, stdout, and stderr, respectively). Other, possibly more efficient, ways to accomplish the same thing as reloading exist [14], but their impact on type safety are not fully understood. Out of roughly 600 classes in the core Java libraries, we are able to safely share about 430 (72%) of them. The rest of the classes are reloaded.

3.3 Java Language Issues

A few language compatability issues arose when building KaffeOS. For example, the Java language description assumes that all string literals are interned, and that equality can therefore be checked with a pointer comparison (the == operator). Unfortunately, to maintain such semantics, the interned string table would have to be a global (kernel) data structure—and user processes could allocate strings in an effort to make the kernel run out of memory. To deal with this problem, we chose to make separately intern strings for each process. As a result, the Java language use of pointer comparison to check string equality does not work for strings that were created in different heaps, and the `equals` method must be used instead. It is impractical for the JVM to hide this semantic change from applications. However, this issue arises only in rare situations, and then only in KaffeOS-aware applications that directly use KaffeOS features.

3.4 Kaffe limitations

Kaffe has relatively poor performance compared to commercial JVMs, for several reasons. First, its garbage collector is relatively primitive: it is a mark-and-sweep collector that is neither generational nor incremental. Second, it has a simple just-in-time bytecode compiler that translates each instruction individually. As a result, many unnecessary register spills and reloads are generated, and the native code that it produces is relatively poor. In addition, the version upon which KaffeOS is based does not use inlined synchronization primitives. To improve stability and performance, we have made some necessary additions to Kaffe, such as preemptive threading, class unloading, and improved reflection support. We have also optimized exception dispatch and improved the allocator.

4 Results

KaffeOS currently runs under Linux on the x86. We plan on porting it to the Itsy pocket computer from Compaq WRL; we have already ported Kaffe to the Itsy. To demonstrate the effectiveness of KaffeOS, we ran the following experiments:

- We measured three implementations of the write barrier. We ran the SPEC JVM98 benchmarks [33] on different configurations of KaffeOS, several versions of Kaffe, and the IBM JVM, which uses one of the fastest commercial JIT compilers [34] available. We must note that our results are not comparable with any published SPEC JVM98 metrics, as the measurements are not compliant with all of SPEC's run rules.

- We ran a servlet engine on KaffeOS to demonstrate that KaffeOS can prevent denial-of-service servlets from crashing a server. We also compared how the number of KaffeOS processes scales with how the number of OS processes scales.

Our measurements were all taken on a 500MHz "Katmai" Pentium III, with 256 Mbytes of SDRAM and a 100 MHz PCI bus, running Red Hat Linux 6.2. The processor has a split 32K L1 cache, and combined 512K L2 cache.

4.1 Write Barrier Implementations

To measure the cost of write barriers in KaffeOS, we implemented several versions:

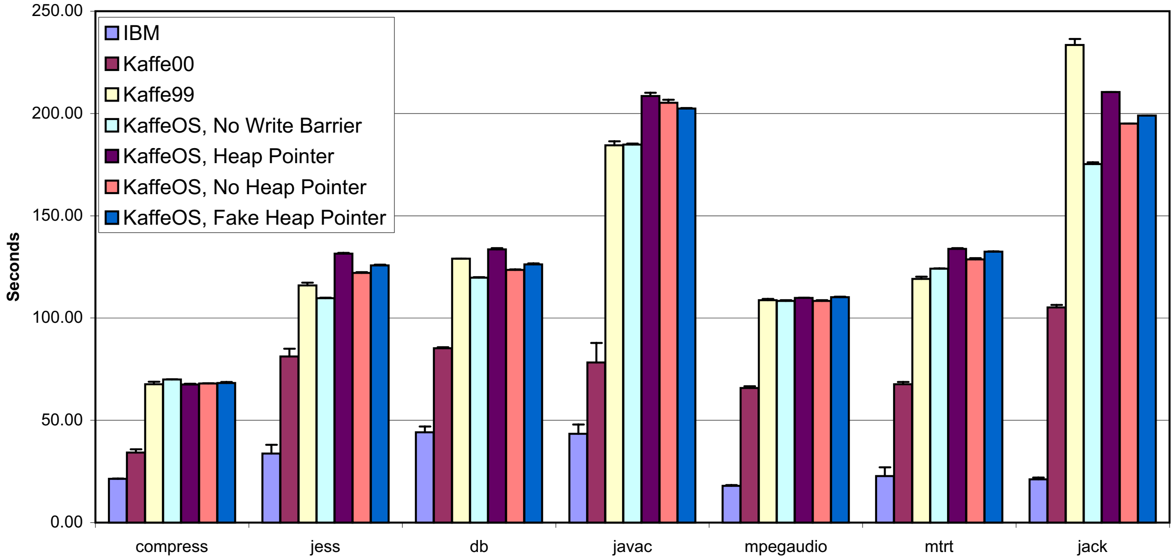

Figure 3: SPEC JVM98 run on various Java platforms. The error bars represent 95% confidence intervals. Each measurement is the result of three runs, except for jack on Kaffe99, which crashes after one run.

- *No Write Barrier*. We execute without a write barrier, and run everything on the kernel heap.
- *Heap Pointer*. At each heap pointer write, the write barrier consists of a call to a routine that finds an object's heap ID in the object header and performs the barrier checks. This implementation takes only 25 cycles with a hot cache, but adds 4 bytes per object.
- *No Heap Pointer*. At each heap pointer write, the write barrier consists of a call to a routine that finds an object's heap ID by looking at the page on which the object lies and performs the barrier checks. This implementation takes 41 cycles with a hot cache.
- *Fake Heap Pointer*. To measure the impact of the 4 bytes of padding in the *Heap Pointer* implementation, we use the third barrier implementation but add 4 bytes to each object.

The KaffeOS JIT compiler does not yet inline the write barrier routine. Inlining the write barrier would not necessarily improve performance, as it would lead to substantial code expansion.

We ran the SPEC JVM98 benchmark suite on two versions of Kaffe, and KaffeOS with different implementations of the write barrier. Kaffe99 is essentially version 1.0b4, which is the code base that KaffeOS was built on. This version is from May 1999. Kaffe00 is the current version of Kaffe, as of April 2000. The major performance differences between the two versions result from a better JIT, faster exception dispatch, and lightweight locking. The fast exception dispatch has been integrated into KaffeOS.

Figure 3 compares the results of our experiments. IBM's JVM is between 2–5 times faster than Kaffe00; Kaffe00 is about twice as fast as Kaffe99. The difference between Kaffe99 and KaffeOS without write barriers is due to the integration of some of the newer features in Kaffe00 into KaffeOS. The benefits of adding faster exception handling shows up strongly in `jack` because that benchmark raises many exceptions. The other KaffeOS configurations may gain some performance benefit because the kernel heap is collected separately from the user heap, which approximates the behavior of a generational garbage collector.

Benchmark	Barriers	Time	Percent
compress	0.017M	0.001s	0.00%
jess	7.9M	0.65s	0.59%
db	33.0M	2.70s	2.26%
javac	15.5M	1.27s	0.69%
mpegaudio	5.5M	0.45s	0.41%
mtrt	3.0M	0.25s	0.20%
jack	11.6M	0.95s	0.54%

Table 1: Number of write barriers executed for each SPEC JVM98 benchmark. "Time" is the total CPU cycle cost for the write barrier instructions, assuming the *No Heap Pointer* cost of 41 cycles; "percent" is the fraction of the *No Write Barrier* execution time.

Kaffe99 does not support profiling, so we can only make educated guesses as to the causes of write barrier overhead. If we compare the write barrier implementations to *No Write Barrier*, the total cost of the write barrier is about 11%. There is not a significant performance difference between the various implementations. The *Heap Pointer* implementation is slightly slower than the *No Heap Pointer* implementation, despite the fact that it takes fewer CPU cycles per barrier. The *Fake Heap Pointer* implementation shows that the *Heap Pointer* padding is part of the problem and that other ways of minimizing the write barrier penalty should be explored.

Table 1 gives the number of write barriers that are executed in each of the SPEC benchmarks. If we compute the time to execute the write barriers by using the cycle counts for the barriers, we see that it is a small percentage of the time for each benchmark. Therefore, almost all of the performance difference between KaffeOS and *No Write Barrier* is most likely due to "secondary" effects: cache pollution, cache conflicts, or even changed garbage collection behavior.

On a better system with a more effective JIT, the relative cost of using write barriers would increase. On the other hand, a good JIT compiler could perform several kinds of optimizations to remove write barriers. A compiler should be able to remove redundant write barriers, along the lines of array bounds checking elimination. It could even perform method splitting to specialize methods, so as to remove useless barriers along frequently used call paths. Again, we can only speculate as to what the performance penalty for implementing KaffeOS on the IBM JVM would be. Nevertheless, the performance of KaffeOS is much better than that of the IBM JVM in the presence of uncooperative applications, despite the raw performance difference between them.

4.2 Servlet Engine

A Java servlet engine provides an environment for running Java programs (servlets) at a server. Their functionality subsumes that of CGI scripts at Web servers: for example, servlets may create dynamic content or run database queries. We use a **MemHog** servlet to measure the effects of a denial-of-service attack. MemHog sits in a loop, repeatedly allocates memory, and keeps it from being garbage-collected.

We compared KaffeOS's ability to prevent the MemHog servlet from denying service with that of IBM's JVM. We used Apache 1.3.12, JServ 1.1 (Apache's servlet engine), and a free version of JSDK 2.0 to run our tests, without modification. JServ runs servlets in *servlet zones*, which are virtual servers. A single JServ instance can host one or more servlet zones. We ran each JServ in its own KaffeOS process. We compared KaffeOS against IBM's JVM, in two configurations: one servlet zone per JVM (IBM/1), and multiple servlet zones in one JVM (IBM/n).

When simulating this denial-of-service attack, we did what a system administrator concerned with availibility of his services would do: we restarted the JVM(s) and the KaffeOS process, respectively, whenever it crashed because of the effects caused by MemHog. In KaffeOS, MemHog will cause a single JServ to exit without affecting other JServs. If each JServ is started in its own IBM JVM, the whole JVM will eventually crash and be restarted. If all servlets are run in a single JServ on a single IBM JVM, the system runs out of memory in seemingly random places. This behavior resulted in exceptions that corrupted data structures that were shared between servlets in the surrounding JServ environment. This corruption eventually

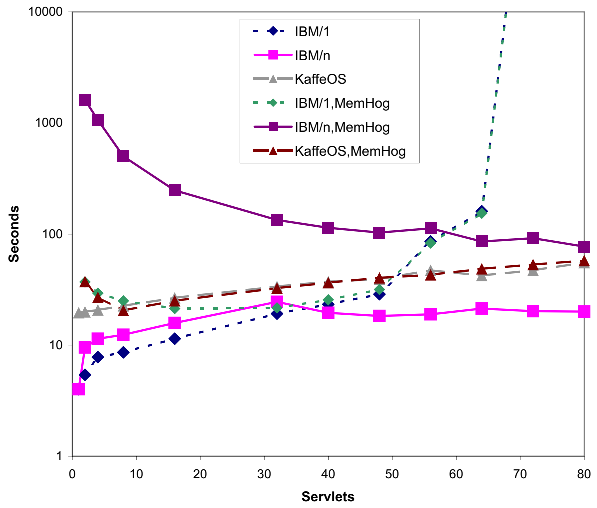

Figure 4: Scaling behavior of JVMs as the number of servlets increases. "IBM/1" means one IBM JVM per servlet; "IBM/n" means n servlets in one JVM. The "MemHog" measurements replace one of the good servlets with a MemHog. The y axis is the amount of time for the non-MemHog servlets to correctly respond to 1000 client requests.

led to a crash of the JVM, because of the lack of a user/kernel boundary.

Figure 4 illustrates the results of our experiments; note that the y axis uses a logarithmic scale. Running a separate KaffeOS process for each servlet has consistent performance, either with a MemHog running or without. This graph illustrates the most important feature of KaffeOS: that it can deliver consistent performance, even in the presence of uncooperative or malicious programs.

The graph shows that running each of the servlets in a single IBM JVM does not scale. This failure occurs because starting multiple JVMs eventually causes the machine to thrash. We estimate that each IBM JVM process takes about 2MB of virtual memory upon startup. We limited each JVM's heap size to 8MB in this configuration. An attempt to start 100 IBM JVMs rendered the machine inoperable.

If there are no uncooperative servlets running, using a single IBM JVM has the best performance. If there is a MemHog servlet running, such a configuration has worse performance than KaffeOS—despite the fact that KaffeOS is several times slower for individual servlets! This degradation is caused by a lack of isolation between servlets. However, as the ratio of well-behaved servlets to malicious servlets increases, the scheduler will yield less often to the malicious servlet. Consequently, the service of IBM/n,MemHog improves as the number of servlets increases. This effect is an artifact of our experimental setup and cannot be reasonably used to defend against denial-of-service attacks.

Finally, we observe a slight service degradation as the number of KaffeOS processes increases. This degradation is likely due to inefficiencies in the user-mode threading system and scheduler.

5 Related Work

We classify the related work into three broad categories: extensible operating systems, resource management in operating systems, and Java extensions for resource management.

5.1 Extensible Operating Systems

Extensible operating systems have existed for many years. Most of them were not designed to protect against malicious users, although a number of them support strong security features. None of them, however, provides strong resource controls. Pilot [30] and Cedar [36] were two of the earliest language-based systems. Their development at Xerox PARC predates a flurry of research in the 1990's on such systems. These systems include Oberon [41] and Juice [18], which are based on the Oberon language; SPIN [5], which is based on Modula-3; and Inferno [15], which is based on a language called Dis. Such systems can be viewed as single-address-space operating systems (see Opal [9]) that use type safety for protection.

VINO is a software-based (but not language-based) extensible system [32] that addresses resource management by wrapping kernel extensions within transactions. When an extension exceeds its resource limits, it can be safely aborted (even if it holds kernel locks) and its resources can be recovered. Transactions are a very effective mechanism, but they are also relatively heavyweight.

5.2 Resource Management

Several operating systems projects have focused on quality-of-service issues and real-time performance guarantees. Nemesis [25] is a single-address-space OS that focuses on quality-of-service for multimedia applications. Eclipse [6] introduced the concept of a *reservation domain*, which is a pool of guaranteed resources. Eclipse provides a guarantee of cumulative service, which means that processes execute at a predictable rate. It manages CPU, disk, and physical memory. Our work is orthogonal, because we examine the software mechanisms that are necessary to manage computational resources.

Recent work on resource management has examined different forms of abstractions for computational resources. Banga et al. [3] describe an abstraction called *resource containers*, which are effectively accounts from which resource usage can be debited. Resource containers are orthogonal to a process' protection domain: a process can contain multiple resource containers, and processes can share resource containers. In KaffeOS we have concentrated on the mechanisms to simply allow resource management; resource-container-like mechanisms could be added in the future.

5.3 Java Extensions

Besides KaffeOS, a number of other research systems have explored (or are currently exploring) the problem of supporting processes in Java.

The J-Kernel [21] and JRes [11, 12] projects at Cornell explore resource control issues without making changes to the Java virtual machine. The J-Kernel extends Java by supporting capabilities between processes. These capabilities are indirection objects that can be used to isolate processes from each other. JRes extends the J-Kernel with a resource management interface whose implementation is portable across JVMs. The disadvantage of JRes (as compared to KaffeOS) is that Jres is a layer on top of a JVM; therefore, it cannot account for JVM resources consumed on the behalf of applications. Cornell is also exploring type systems that can support revocation directly [22].

Alta [37] is a Java virtual machine that enforces resource controls based on a nested process model. The nested process model in Alta allows processes to control the resources and environment of other processes, including the class namespace. Additionally, Alta supports a more flexible sharing model that allows processes to directly share more than just objects of primitive types. Like KaffeOS, Alta is based on Kaffe, and, like KaffeOS, Alta provides support within the JVM for comprehensive memory accounting. However, Alta only provides a single, global garbage collector, so separation of garbage collection costs is not possible.

Balfanz and Gong [2] describe a multiprocessing JVM developed to explore the security architecture ramifications of protecting applications from each other, as opposed to just protecting the system from applications. They identify several areas of the JDK that assume a single-application model, and propose extensions to the JDK to allow multiple applications and to provide inter-application security. The focus of their multiprocessing JVM is to explore the applicability of the JDK security model to multi-processing, and they rely on the existing, limited JDK infrastructure for resource control.

Sun's original JavaOS [35] was a standalone OS written almost entirely in Java. It is described as a first-class OS for Java applications, but appears to provide a single JVM with little separation between applications. It was to be replaced by a new implementation termed "JavaOS for Business" that also ran only Java applications. "JavaOS for Consumers" is built on the Chorus microkernel OS [31] to achieve real-time properties needed in embedded systems. Both of these systems apparently require a separate JVM for each Java application, and all run in supervisor mode.

Joust [20], a JVM integrated into the Scout operating system [28], provides control over CPU time and network bandwidth. To do so, it uses Scout's path abstraction. However, Joust does not support memory limits on applications.

The Open Group's Conversant system [4] is another project that modifies a JVM to provide processes. It provides each process with a separate address range (within a single Mach task), a separate heap, and a separate garbage collection thread. Conversant does not support sharing between processes, unlike KaffeOS, Alta, and the J-Kernel.

6 Conclusions

We have described the design and implementation of KaffeOS, a Java virtual machine that supports the operating system abstraction of *process*. KaffeOS enables processes to be isolated from each other, to have their resources controlled, and still share objects directly. Processes enable the following important features:

- The resource demands of Java processes can be accounted for separately, including memory consumption and GC time.
- Java processes can be terminated if their resource demands are too high, without damaging the system.
- Termination reclaims the resources of the terminated Java process.

These features enable KaffeOS to run untrusted code safely, because it can prevent simple denial-of-service attacks that would disable standard JVMs. The cost of these features, relative to Kaffe, is reasonable. Because Kaffe's performance is poor compared to commercial JVMs, it is difficult to estimate the cost of adding such features to a commercial JVM—but we believe that the overhead should not be excessive. Finally, even though KaffeOS is substantially slower than commercial JVMs, it exhibits much better performance scaling in the presence of uncooperative code.

Acknowledgements

We thank many members of the Flux group, especially Patrick Tullmann, for enlightening discussion, comments on earlier drafts, and assistance in running some of the experiments.

References

[1] G. V. Back and W. C. Hsieh. Drawing the red line in Java. In *Proc. of the 7th HotOS*, Rio Rico, AZ, Mar. 1999. IEEE Computer Society.

[2] D. Balfanz and L. Gong. Experience with secure multi-processing in Java. In *Proc. of the Eighteenth ICDCS*, May 1998.

[3] G. Banga, P. Druschel, and J. C. Mogul. Resource containers: A new facility for resource management in server systems. In *Proc. of the Third OSDI*, pp. 45–58, New Orleans, LA, Feb. 1999.

[4] P. Bernadat, L. Feeney, D. Lambright, and F. Travostino. Java sandboxes meet service guarantees: Secure partitioning of CPU and memory. TR TOGRI-TR9805, The Open Group Research Institute, June 1998.

[5] B. Bershad, S. Savage, P. Pardyak, E. Sirer, M. Fiuczynski, D. Becker, S. Eggers, and C. Chambers. Extensibility, safety and performance in the SPIN operating system. In *Proc. of the 15th SOSP*, pp. 267–284, Copper Mountain, CO, Dec. 3–6, 1995.

[6] J. Bruno, E. Gabber, B. Ozden, and A. Silberschatz. The Eclipse operating system: Providing quality of service via reservation domains. In *Proc. of USENIX '98*, pp. 235–246, New Orleans, LA, June 1998.

[7] P. Chan and R. Lee. *The Java Class Libraries: Volume 2*. The Java Series. Addison-Wesley, second edition, November 1998.

[8] P. Chan, R. Lee, and D. Kramer. *The Java Class Libraries: Volume 1*. The Java Series. Addison-Wesley, second edition, November 1998.

[9] J. S. Chase, H. M. Levy, M. J. Feeley, and E. D. Lazowska. Sharing and protection in a single-address-space operating system. *ACM TOCS*, 12(4):271–307, 1994.

[10] P. Cheng, R. Harper, and P. Lee. Generational stack collection and profile-driven pretenuring. In *Proc. of PLDI '98*, pp. 162–173, Montreal, Canada, June 1998.

[11] G. Czajkowski, C.-C. Chang, C. Hawblitzel, D. Hu, and T. von Eicken. Resource management for extensible internet servers. In *Proc. of the 8th ACM SIGOPS European Workshop*, Sintra, Portugal, Sept. 1998.

[12] G. Czajkowski and T. von Eicken. JRes: a resource accounting interface for Java. In *Proc. of OOPSLA*, Vancouver, Canada, Oct. 1998. ACM.

[13] Delivering the promise of Internet computing: Integrating Java with Oracle8*i*. http://www.oracle.com/database/documents/-delivering_the_promise_twp.pdf, Apr. 1999.

[14] D. Dillenberger, R. Bordawekar, C. W. Clark, D. Durand, D. Emmes, O. Gohda, S. Howard, M. F. Oliver, F. Samuel, and R. W. S. John. Building a Java virtual machine for server applications: The JVM on OS/390. *IBM Systems Journal*, 39(1):194–210, 2000.

[15] S. Dorward, R. Pike, D. L. Presotto, D. Ritchie, H. Trickey, and P. Winterbottom. Inferno. In *Proc. of the 42nd IEEE COMPCON*, San Jose, CA, February 1997.

[16] D. R. Engler, M. F. Kaashoek, and J. O'Toole Jr. Exokernel: An operating system architecture for application-level resource management. In *Proc. of the 15th SOSP*, pp. 251–266, Dec. 1995.

[17] K. Farkas, J. Flinn, G. Back, D. Grunwald, and J. Anderson. Quantifying the energy consumption of a pocket computer and a Java virtual machine. In *Proc. of SIGMETRICS '00*, page to appear, Santa Clara, CA, June 2000.

[18] M. Franz. Beyond Java: An infrastructure for high-performance mobile code on the World Wide Web. In S. Lobodzinski and I. Tomek, editors, *Proc. of WebNet '97*, pp. 33–38, October 1997.

[19] L. Gorrie. Echidna — a free multiprocess system in Java. http://www.javagroup.org/echidna/.

[20] J. H. Hartman et al. Joust: A platform for communication-oriented liquid software. TR 97–16, Univ. of Arizona, CS Dept., Dec. 1997.

[21] C. Hawblitzel, C.-C. Chang, G. Czajkowski, D. Hu, and T. von Eicken. Implementing multiple protection domains in Java. In *Proc. of USENIX '98*, pp. 259–270, New Orleans, LA, June 1998.

[22] C. Hawblitzel and T. von Eicken. Type system support for dynamic revocation. In *2nd WCSSS*, Atlanta, GA, May 1999.

[23] T. Jaeger, J. Liedtke, and N. Islam. Operating system protection for fine-grained programs. In *Proc. of the 7th USENIX Security Symposium*, pp. 143–157, Jan. 1998.

[24] E. Jul, H. Levy, N. Hutchison, and A. Black. Fine-grained mobility in the Emerald system. *ACM TOCS*, 6(1):109–133, February 1988.

[25] I. M. Leslie, D. McAuley, R. J. Black, T. Roscoe, P. R. Barham, D. M. Evers, R. Fairbairns, and E. A. Hyden. The design and implementation of an operating system to support distributed multimedia applications. *IEEE JSAC*, 14(7):1280–1297, Sept. 1996.

[26] S. Liang and G. Bracha. Dynamic class loading in the Java virtual machine. In *Proc. of OOPSLA '98*, pp. 36–44, Vancouver, Canada, Oct. 1998.

[27] D. Malkhi, M. K. Reiter, and A. D. Rubin. Secure execution of Java applets using a remote playground. In *Proc. of the 1998 IEEE Symposium on Security and Privacy*, pp. 40–51, Oakland, CA, May 1998.

[28] D. Mosberger and L. L. Peterson. Making paths explicit in the Scout operating system. In *Proc. of the 2nd OSDI*, pp. 153–167, Seattle, WA, Oct. 1996.

[29] D. Plainfossé and M. Shapiro. A survey of distributed garbage collection techniques. In *Proc. of the 1995 IWMM*, Kinross, Scotland, Sept. 1995.

[30] D. D. Redell, Y. K. Dalal, T. R. Horsley, H. C. Lauer, W. C. Lynch, P. R. McJones, H. G. Murray, and S. C. Purcell. Pilot: An operating system for a personal computer. *CACM*, 23(2):81–92, 1980.

[31] M. Rozier, V. Abrossimov, F. Armand, I. Boule, M. Gien, M. Guillemont, F. Herrmann, C. Kaiser, S. Langlois, P. Léonard, and W. Neuhauser. The Chorus distributed operating system. *Computing Systems*, 1(4):287–338, Dec. 1989.

[32] M. I. Seltzer, Y. Endo, C. Small, and K. A. Smith. Dealing with disaster: Surviving misbehaved kernel extensions. In *Proc. of the 2nd OSDI*, pp. 213–227, Seattle, WA, Oct. 1996.

[33] SPEC JVM98 benchmarks. http://www.spec.org/-osg/jvm98/.

[34] T. Suganuma, T. Ogasawara, M. Takeuchi, T. Yasue, M. Kawahito, K. Ishizaki, H. Komatsu, and T. Nakatani. Overview of the IBM Java just-in-time compiler. *IBM Systems Journal*, 39(1):175–193, 2000.

[35] Sun Microsystems, Inc. JavaOS: A standalone Java environment, Feb. 1997. http:-//www.javasoft.com/products/javaos/javaos.-white.html.

[36] D. C. Swinehart, P. T. Zellweger, R. J. Beach, and R. B. Hagmann. A structural view of the Cedar programming environment. *ACM TOPLAS*, 8(4):419–490, Oct. 1986.

[37] P. Tullmann and J. Lepreau. Nested Java processes: OS structure for mobile code. In *Proc. of the Eighth ACM SIGOPS European Workshop*, pp. 111–117, Sintra, Portugal, Sept. 1998.

[38] D. J. Wetherall, J. Guttag, and D. L. Tennenhouse. ANTS: A toolkit for building and dynamically deploying network protocols. In *Proc. of IEEE OPENARCH '98*, San Francisco, CA, April 1998.

[39] T. Wilkinson. Kaffe – a Java virtual machine. http://www.transvirtual.com.

[40] P. R. Wilson. Uniprocessor garbage collection techniques. In *Proc. of the 1992 IWMM*, St. Malo, France, Sept. 1992. Expanded version, submitted to Computing Surveys.

[41] N. Wirth and J. Gutknecht. *Project Oberon*. ACM Press, New York, NY, 1992.