SableVM: A Research Framework for the Efficient Execution of Java Bytecode[1]

Sable Technical Report No. 2000-3

Etienne M. Gagnon and Laurie J. Hendren
Sable Research Group
School of Computer Science
McGill University
[gagnon,hendren]@sable.mcgill.ca

November 2000

Abstract

SableVM is an open-source virtual machine for Java[2], intended as a research framework for efficient execution of Java bytecode. The framework is essentially composed of an extensible bytecode interpreter using state-of-the-art and innovative techniques. Written in the C programming language, and assuming minimal system dependencies, the interpreter emphasizes high-level techniques to support efficient execution. In particular, we introduce new data layouts for classes, virtual tables and object instances that reduce the cost of interface method calls to that of normal virtual calls, allow efficient garbage collection and light synchronization, and make effective use of memory space.

1 Introduction & Motivation

Over the last few years, Java[21] has rapidly become one of the most popular general purpose object-oriented (OO) programming languages. Java programs are compiled into *class files* which include type information and platform independent bytecode instructions. On a specific platform, a runtime system called a *virtual machine*[24] loads and links class files then executes bytecode instructions. The virtual machine collaborates with the standard class libraries to provide key services to Java programs, including threads and synchronization, automatic memory management (garbage collection), safety features (array bound checks, null pointer detection, code verification), reflection, dynamic class loading, and more.[3]

Early Java virtual machines were simple bytecode interpreters. Soon, the quest for efficiency led to the addition of *Just-In-Time compilers* (JIT) to virtual machines, an idea formerly developed for other OO runtime systems like Smalltalk-80[17] and Self-91[15]. In a few words, a just-in-time compiler works by compiling bytecodes to machine specific code on the first invocation of a method. JITs range from the very naive, that use templates to replace each bytecode with a fixed sequence of native code instructions (early versions of Kaffe[5] did this), to the very sophisticated that perform register allocation, instruction scheduling and other scalar optimizations (e.g. [8, 23, 29, 33]).

JITs face two major problems. First, they strive to generate good code in very little time, as compile time is lost to the running application. Second, the code of compiled method resides in memory; this augments the pressure on the memory manager and garbage collector. Recent virtual machines try to overcome these problems. The main trend is to use dynamic strategies to find *hot* execution paths, and only optimize these areas (e.g. [4, 10, 16]). HotSpot[4], for example, is a mixed interpreter and compiler environment. It only compiles and optimizes *hot spots*. Jalapeno[10, 11], on the other hand, always compiles methods (naively at first), then uses adaptive online feedback to recompile and optimize hot methods. These techniques are particularly suited to virtual machines executing long running programs in server environments. The optimizer can be relatively slow and consist of a fully fledged optimizing compiler using intermediate representations and performing *costly* aggressive optimizations, as compile time will be amortized on the overall running time.

Our research complements these approaches by exploring opportunities for making the virtual machine execute efficiently. Rather than looking at fine grain techniques, like register allocation and instruction scheduling, we address the fundamental problem of data layout in a dynamic Java environment. While Java shares many properties with other object-oriented languages, the set of runtime constraints enforced by the verifier and the basic services provided for each object (hash code, locking) are unique. This leads us to revisit traditional data structures used in object-oriented runtime environments, and adapt them to fully take advantage of the properties of the Java runtime environment.

[2] In this document, the term *Java* means: *the Java programming language*.

[3] There exist static compilers that directly compile Java programs to machine code (e.g. [2, 3, 7]). The constraints of static and dynamic environments are quite different. Our research focuses on dynamic Java execution environments.

As a testbed for evaluating our proposed data structures and algorithms, we are designing and implementing SableVM, a standards conforming open-source virtual machine. Written in the C programming language, and depending on the POSIX application programming interface (API), it is meant as a small and portable interpreter[4]. It can be used as an experimental framework for extending the bytecode language. It can also be used as an efficient virtual machine for embedded systems, or as a profiling interpreter in a hybrid interpreter/just-in-time optimizing-compiler environment.

The remaining part of this document is structured as follows. Section 2, we state the contributions of this paper. In section 3, we give an overview of the SableVM framework. In section 4, we describe SableVM's threaded interpreter. In section 5, we introduce our classification of virtual machine memory. In section 6, we explore new techniques for efficient garbage collection, interface method invocation and synchronization. In section 7, we discuss our proposed experiments. Finally, in section 8, we present our conclusions.

2 Contribution

The specific contributions of this research are as follows.

- Categorization of virtual machine memory into separate areas exhibiting different management needs.
- Introduction of an object instance data layout grouping reference fields sequentially, enabling faster garbage collection tracing.
- Introduction of a virtual method lookup table layout enabling constant time interface method lookup in presence of dynamic loading.
- Improvement of the bimodal field based thin lock algorithm[13, 26] (without busy-wait) to eliminate the one bit overhead per object instance.
- Introduction of a simple strategy to provide dynamically growing stacks.
- Design and implementation of SableVM, an experimental framework for testing high-level techniques for efficient execution and extensions to the Java virtual machine.
- Study of the efficiency achievable by a portable Java bytecode interpreter written in C.

3 Framework Overview

As shown in Figure 1, the SableVM experimental framework is a virtual machine composed of five main components: interpreter, memory manager, verifier, class loader, and native interface. In addition, the virtual machine implements various services required by the class library (e.g.: synchronization and threads).

SableVM is entirely[5] written in portable C. Thus, its source code is readable and simple to modify. This makes an ideal framework for testing new high-level implementation features or bytecode language extensions. For example, adding a new arithmetic bytecode instruction entails making a minor modification to the class loader, adding a few rules to the verifier, and finally adding the related interpreter code. This is pretty easy to do in SableVM, as compared to a virtual machine written in assembly language. It is also possible to add a JIT compiler and feedback mechanisms to SableVM.

The current implementation of SableVM targets the Linux operating System on Intel x86 processors. It uses the GNU libc implementation of POSIX threads to provide preemptive operating system level threads. SableVM has a very clean implementation the Java Native

[4]SableVM depends on the open-source *GNU Classpath[1]* class library for providing standard library services.

[5]Exceptions: We assume a POSIX function library, we use the *label as values* GNU C extension (see Figure 2(b)), and there is a single line of assembly code (*compare-and-swap*), for implementing light locks.

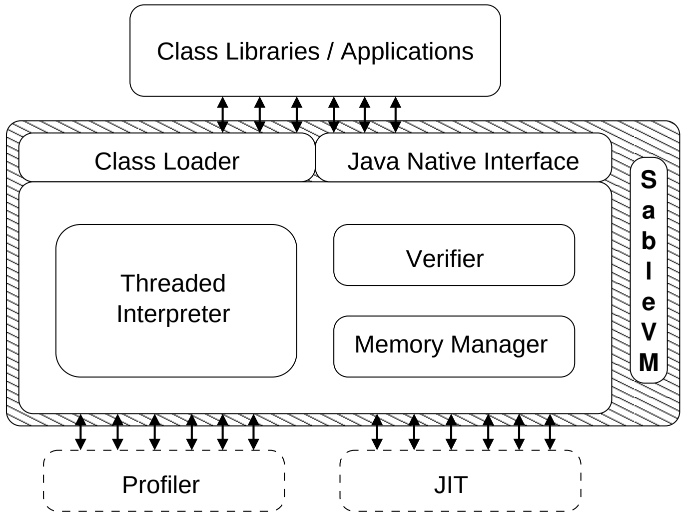

Figure 1: The SableVM experimental framework

Interface (version 1.2). For example, it can run multiple virtual machine instances on separate threads in a single process.

4 Threaded Interpreter

SableVM's interpreter is a threaded interpreter. Pure bytecode interpreters suffer from expensive dispatch costs: on every iteration, the dispatch loop fetches the next bytecode, looks up the associated implementation address in a table (explicitly, or through a *switch* statement), then transfers the control to that address. *Direct threading*[20] reduces this overhead: in the executable code stream, each bytecode is replaced by the address of its associated implementation. In addition, each bytecode implementation ends with the code required to dispatch the next opcode. This is illustrated in figure 2. This technique eliminates the table lookup and the central dispatch loop (thus eliminating a branch instruction to the head of the loop). As these operations are expensive on modern processors, this technique has been shown to be quite effective[20, 27].

```
/* code */

char code[] = {
  ICONST_2, ICONST_2,
  ICONST_1, IADD, ...
}
char *pc = code;

/* dispatch loop */

while(true) {
  switch(*pc++) {
  case ICONST_1: *++sp = 1; break;
  case ICONST_2: *++sp = 2; break;
  case IADD:
    sp[-1] += *sp; --sp; break;
  ...
}}
```

(a) Pure bytecode interpreter

```
/* code */

void *code[] = {
  &&ICONST_2, &&ICONST_2,
  &&ICONST_1, &&IADD, ...
}
void **pc = code;

/* implementations */

goto **(pc++);

ICONST_1: *++sp = 1; goto **(pc++);
ICONST_2: *++sp = 2; goto **(pc++);
IADD:
  sp[-1] += *sp; --sp; goto **(pc++);
...
```

(b) Threaded Interpreter

Figure 2: Pure and threaded interpreters

One complication of translating bytecodes to direct threaded code is that bytecodes use a single byte of storage (plus operands), while the address of an implementation requires `sizeof(void *)` bytes. To simplify the translation process, SableVM translates a method body of N bytes to a buffer of N * `sizeof(void *)` bytes[6], replacing each *bytecode + operands* of n bytes by an *address + operands* of n * `sizeof(void *)` bytes[7]. This greatly simplifies the resolution of branch destination in the threaded code buffer. It also enables a one pass translation of bytecodes to threaded code.

Method bodies are translated to threaded code on their first invocation. We take advantage of this translation to do some optimizations. For example, we precompute absolute branch destinations, we translate overloaded bytecodes like the `GET_FIELD` instruction to separate implementation addresses (e.g. `GET_FIELD_INT`, `GET_FIELD_FLOAT`, ...), and we inline *constant pool* references to direct operand values.

This one pass translation is much simpler than the translation done by even the most naive just-in-time compiler, as each bytecode maps to an address, not a variable sized implementation. However, unlike a JIT, the threaded interpreter still pays the cost of an instruction dispatch for each bytecode. Piumarta[27] has shown a technique to reduce this overhead, using selective inlining in a portable manner at the cost of additional memory[8]. SableVM implements this technique optionally through a runtime flag, as it might not be appropriate for systems with little memory.

5 Memory Management

Memory management is a central issue in the design of SableVM. Most of the high-level performance enhancements introduced in this research are related to memory management.

In this section, we classify the different sections of memory of the Java virtual machine according to two criteria: allocation and release behavior, and management control. We define the four resulting categories (shared, thread specific, class loader specific, and system memory), and discuss how the SableVM implementation takes advantage of the specific properties of each category.

5.1 Shared Memory

Shared memory is managed by the virtual machine and potentially allocated and modified by many threads executing methods of various class loaders.

This memory consists primarily of the Java heap (which is garbage collected), and global JNI references. The allocation and release behavior of such memory is highly application dependent.

5.1.1 SableVM Implementation

SableVM manages this memory using the `mmap()` and `munmap()` POSIX function calls, which interact with the operating system virtual memory system. SableVM maps the special Linux device `/dev/zero` to obtain zeroed memory pages[9]. The advantage of unmapping memory, relative to simply freeing memory (using the `malloc()/free()` ANSI C functions) is that unmapped pages need not be dumped to disk by the virtual memory system, when they are to be to be ejected from RAM. This is particularly useful in the case of copying collectors. SableVM's

[6]The unused space is recycled by the class loader memory manager (see section 5.3).

[7]To be accurate, we use the maximum of `sizeof(void *)` and `sizeof(jint)`.

[8]On some processors, this technique requires one line of assembly code to synchronize the instruction and data caches.

[9]This can be replaced by explicit initialization to zero on operating systems that do not provide the equivalent of /dev/zero.

default garbage collector is a simple precise copying collector. After collection, SableVM unmaps the old semi-space and maps a new zeroed semi-space. On cooperative operating systems, this has the potential of reducing the typical thrashing of copying collectors.

5.2 Thread Specific Memory

Thread specific memory is also managed by the virtual machine, but is allocated for internal management of each Java thread.

This memory consists primarily of Java stacks, JNI local reference frames for each stack, and internal structures storing thread specific data like stack information, JNI virtual table, and exception status.

This memory exhibits precise allocation and release patterns. Thread specific structures have a life time similar to their related thread. So, this memory can be allocated and freed (or recycled) at the time of respective creation and death of the underlying thread. Stacks do not conceptually suffer from fragmentation, as they grow and shrink on one side only. This property is shared by JNI local reference frames.

5.2.1 SableVM Implementation

The SableVM implementation takes advantage of the stack allocation behavior to manage thread specific memory independently from shared memory. SableVM allocates thread structures on thread creation but does not release them at thread death. Instead, it manages a free list to recycle this memory on future thread creation. Java stack memory is managed differently: an initial stack is mapped (using `mmap()`) on thread creation. Then this stack is expanded (using Linux's `mremap()`) as required by the computation. The `mremap()` function is similar to the ANSI C `realloc()` function, but it interacts with the virtual memory system, saving many copying operations by remapping pages of memory when the memory block needs to be relocated elsewhere to grow. SableVM never shrinks the size allocated for a Java stack. It simply unmaps it on thread death. On systems that do not have `mremap()`), thread allocation and growth can be managed using `malloc()` and `realloc()` instead (at a higher reallocation cost). Initial allocation size and growth are controlled through runtime parameters. Note that all Java stack frame information is stored relatively (e.g. offset to previous/next frame, instead of direct address), which enables stacks to be moved.

The advantage of a stack whose maximum size can only grow, is simpler stack memory management at method call boundaries. It also simplifies the virtual machine usage, as the user needs not specify a maximum stack size, which can be difficult to do for each thread of a multi-threaded application. There's no need to manage the potential fragmentation of non-movable stacks. It would be possible to allocate Java stack frames on the shared garbage collected Java heap but this would put unnecessary additional pressure on the garbage collector, and it would not take advantage of the highly regular allocation and release behavior of this memory.

5.3 Class Loader Specific Memory

Class loader specific memory is managed by the virtual machine and is allocated for internal management of each class loader.

This memory consists primarily of the internal data structures used to store class loaders, classes and methods, and related information. This includes method bodies in their various forms like bytecode, direct threaded code, inlined threaded code, and compiled code (in the presence of a JIT). It also includes virtual tables.

This memory exhibits precise allocation and release patterns. This memory is allocated at class loading time, and at various preparation, verification, and resolution execution points. This memory differs significantly from stacks and the shared garbage collected heap in that once it is allocated, it must stay at a fixed location, and it is unlikely to be released soon. The Java virtual machine specification allows for potential unloading of all classes of a class loader as a group, if no direct or indirect references to the class loader, its classes, and theirs instances remain. In such a case, and if a virtual machine supports class unloading, all memory used by a class loader and its classes can be released at once.

5.3.1 SableVM Implementation

SableVM manages this memory on a class loader basis. It allocates chunks of memory using `mmap()` then provides its own allocator for distributing smaller memory fragments. This has many advantages.

It allows the allocation of many small memory blocks without the usual cost of memory space overhead. ANSI C's `malloc()` has to use additional memory space to store the size of each allocated block in prevision of future `free()` calls, which our allocator needs not worry about. Memory chunks are unmapped on class unloading without regard for internal sub-allocation.

A corollary to the previous point is that class unloading is more efficient than using regular `malloc()` and `free()` calls, as there is no need to incrementally aggregate small memory areas, as would happen with a sequence of `free()` calls.

It also allows class parsing and decoding in one pass without memory overhead, by allocating many small memory blocks. This is usually not possible, as it is not possible to estimate the memory requirement for storing internal class information before the end of the first pass. The usual alternatives are to either pay a memory overhead for allocating small blocks, or do 2 passes over the class file; one pass to compute memory requirements, and the second to extract the information and store it in the allocated memory. But even then, it does not solve the problem of many small allocations which are required to store threaded or compiled code, and any other linking information computed lazily throughout execution.

Finally, and importantly, it allows for irregular memory management strategies: it makes it possible to return sub-areas of an allocated block to the allocator, if these areas are known not to be used. We take advantage of this to improve the representation of interface method lookup tables[10].

5.4 System Memory

System memory is the remaining memory, on which we have essentially no control. It consists of the memory used for virtual machine code, native stacks, ANSI C's memory manager (`malloc()` and `free()`), and dynamically linked native libraries, and any other uncontrollable memory.

5.4.1 SableVM Implementation

As SableVM is written in portable C, it has no say on the management of this memory. Assembly and system specific virtual machines can (and should probably) manage this memory, in which case it could be classified among the previous three categories. For example, dynamically linked native libraries would be classified as class loader specific memory.

The only attempt to control this memory is the following. SableVM manages its own Java stacks (one per thread), and does not use a C function call to implement Java method calls. Thus, it makes a minimal use of the native C stack. Recursive function calls are only possible through native JNI calls. (A native JNI method is allowed to invoke the virtual machine interpreter which may in turn call back the JNI native method.)

6 Performance Enhancements

In this section, we present a set of new high-level performance enhancements. These enhancements concern the areas of garbage collection, interface method invocation, and locking.

[10]See section 6.2.

6.1 Efficient Garbage Collection Traces

The Java heap is by definition a garbage collected area. A Java programmer has no control on the deallocation of an object. Garbage collectors can be divided into two major classes: tracing and non-tracing collectors. Non-tracing collectors (mainly *reference counting*) cannot reclaim cyclic data structures, are a poor fit for concurrent programming models, and have a high reference count maintenance overhead. For this reason, Java virtual machine designers usually opt for a tracing collector.

There exist many tracing collectors[22]. The simplest models are mark-and-sweep, copying and mark-compact. The common point to all tracing collectors (including advanced generational, conservative and incremental techniques) is that they must trace a subset of the heap, starting from a root set, looking for *reachable* objects. Tracing is often one of the most expensive steps of garbage collection[22]. For every root, the *garbage collector* (gc) looks up the type of the object to find the offset of its reference fields, then it recursively visits the objects referenced by these fields.

We propose a new object layout that optimizes the placement of reference fields to allow efficient gc tracing. First, we explain the problem. To provide efficient field access, it is desirable to place fields at a constant offset from the object header, regardless of inheritance. This is easily achieved in Java as instance fields can only be declared in classes (not in interfaces), and classes are restricted to single inheritance. Fields are laid out consecutively after the object header, starting with super class fields then subclass fields, as shown in Figure 3(a). When tracing such an object, the garbage collector must access the object's class information to discover the offset of its reference fields, then access the superclass information to obtain the offset of its reference fields, and so on. As this process must be repeated for each traced object, it is quite expensive.

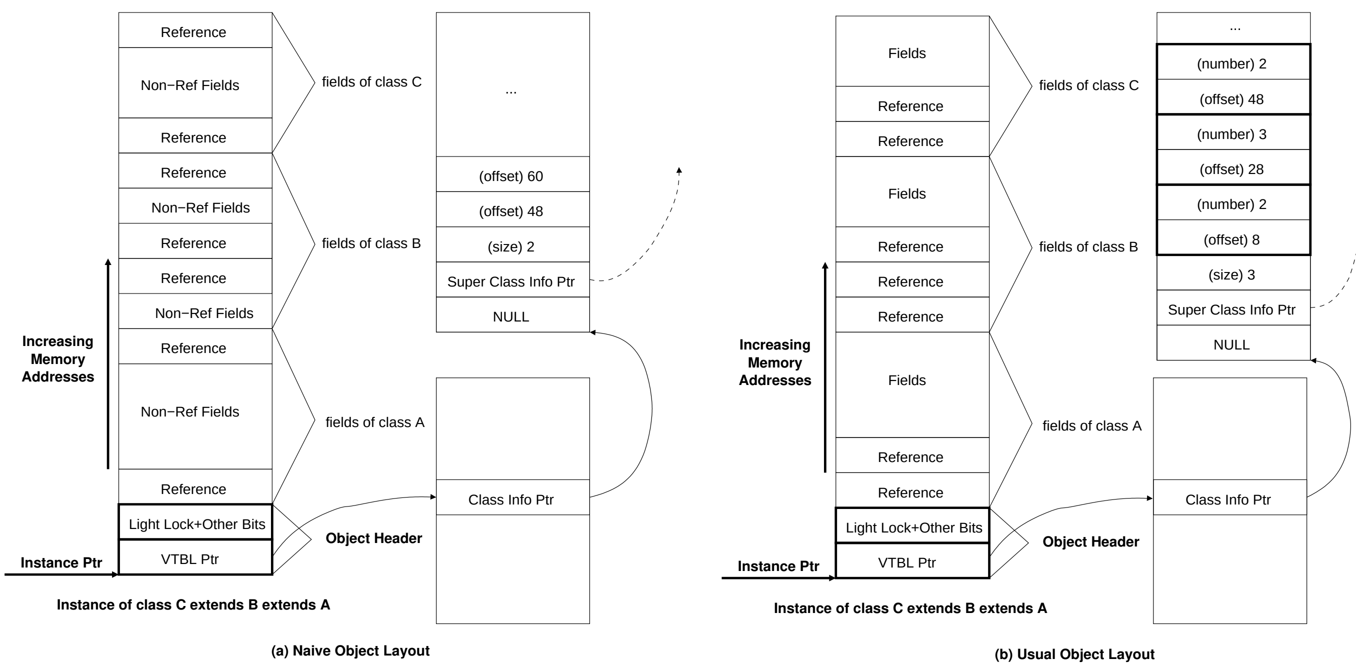

Figure 3: Virtual table layout

There are two improvements that are usually applied to this naive representation. First, reference fields can be grouped together in the layout of each class. Secondly, each class can store an array of offsets and counts of reference fields for itself and all its super classes. This is shown in Figure 3(b). The number of memory accesses needed to trace an object, in this case, is n (the number of references) + 3 (virtual table pointer, class pointer, array size) + 2 * array size (each array element has two values: base offset and reference number). Two nested loops (and loop variables) are required: one to traverse the array, and one for each array element (accessing the related number of references).

We propose a new object layout that further reduces the number of memory accesses required to trace an object. Our solution is to group all reference fields consecutively. To maintain the constant offset property, we simply grow objects in both directions, placing non-reference fields after the object header, and reference fields in front of it. The left part of Figure 4 illustrates the layout of object instances. In the object instance layout, the instance starting point is *possibly* a reference field. The instance grows both ways from the object header, which is located in the middle of the instance. References are placed before the header, and other fields are placed after it. The right part of Figure 4 illustrates the layout of array instances. Array element are placed in front or after the array instance header, depending on whether the element type is a reference or a non-reference type, respectively.

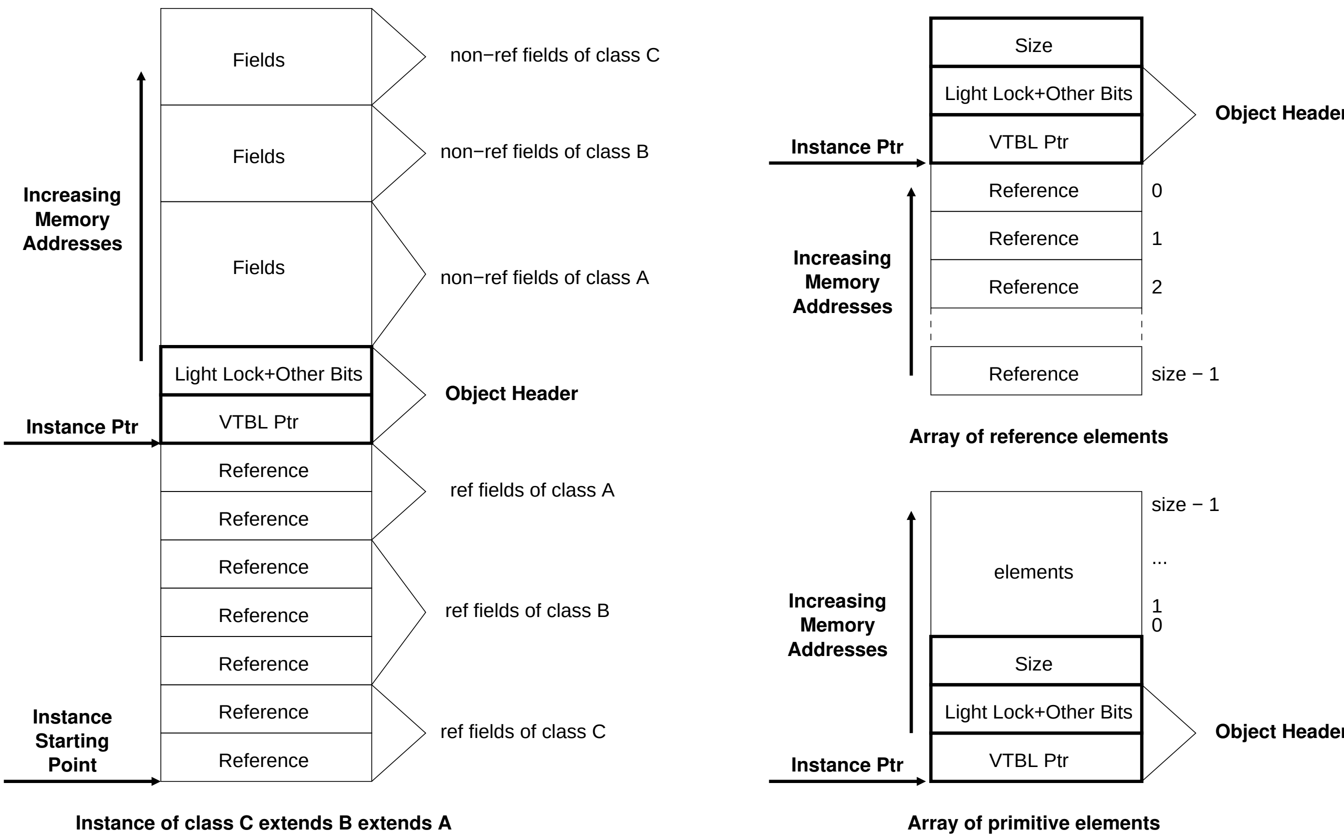

Figure 4: Object and array instance layout

At this point, we must distinguish the two ways in which an object instance can be reached by a tracing collector. The first way is through an object reference that points to the object header (which is in the middle of the object). The second way is through its starting point. In the first case, our representation allows a very efficient and simple tracing algorithm. The gc accesses the class information to get the number of references n, and stores it in a local variable (or register). Then, it simply traces n references in front of the object header. This is much simpler than tracing the traditional object layout.

In the second case, the object instance is reached is from its starting point in memory, which might be either a reference field or the object header (if there are no reference fields in this instance). This happens when a copying collectors scans the *tospace*, starting from its beginning. This is a problem with our representation, as the gc does not yet know the type of the object it is scanning. In other words, it does not know whether the initial word is a reference or a virtual table pointer.

The following elegant and efficient algorithm solves the problem. The gc keeps the start and end addresses of the *fromspace* in local variables (or registers). When it starts scanning, it compares the value of the first word to NULL. If it is NULL, it is not a virtual table pointer, and the gc repeats the process with the next word. If it is not NULL, the value is compared to the *fromspace* boundaries. If it is within these boundaries, the reference is changed to the forwarding *tospace* value of the referenced *fromspace* object (after it is copied, if necessary). If it is not within boundaries, then it is a virtual table pointer. In this case, the gc uses it to access class information and retrieves the offset of the next object to scan. The advantage of this approach is that every reference field is read only once and quickly compared with boundary values. Finally, when the virtual table pointer is accessed, it is immediately used and need not be remembered. If boundary values are kept in registers, the total number of memory operations will be limited to n (the number of reference fields) + 3 (virtual table pointer, class pointer, offset value). If boundary values are kept in local variables, they are likely available in the data cache. This overhead is again lower than the overhead of tracing the traditional object layout.

Objects are also scanned from their starting point in the sweep phase of a mark-and-sweep collector. Our technique is easily adapted to generational collection by choosing the appropriate range. The only requirement is not to have any virtual table within the maximum and minimum boundaries of all Java heap segments. In fact, the *reference-or-vtable-ptr* test could be simplified by keeping all virtual tables either above or below heap segments, but this would be difficult to achieve for SableVM's class loader memory manager.

We also provide an alternative *reference-or-vtable-ptr* test which does not depend on bound checks. Figure 5 illustrates this alternative test. The ambiguous value is compared to NULL. If it is NULL, then it is a reference. If it is not NULL, then it is safe to dereference this value twice, as shown in Figure 5, because it is either a reference which necessarily points to a virtual table pointer, or it is a virtual table pointer which necessarily points to a class information pointer. The value retrieved by the double indirection is then compared to NULL. If it is NULL, then it is the first word of a class information structure, so the initial ambiguous value is a virtual table pointer. If it is not NULL, then it is a class information pointer, and the initial ambiguous value is a reference.

This alternative test is a little more expensive, because of the double indirection "`**ref`" which causes two additional memory accesses per reference. This test would be useful for memory managers that seldom scan objects from their starting point, when it would be difficult to provide a single range of memory including all the Java heap, but no virtual tables. The only requirement for this alternative test to work is to put NULL at the start of the class information structure.

The idea of using a bidirectional object layout has been suggested[25, 28] as a mean to provide efficient access to instance data and dispatch information in languages supporting multiple inheritance (like C++). In [14], Bartlett proposed a garbage collector which requires grouping references at the head of objects.

6.2 Constant Time Interface Method Lookup

Our next enhancement addresses a problem raised by multiple inheritance of interfaces in Java. The virtual machine instruction set contains an *invokeinterface* instruction, used to invoke interface methods. A common technique to implement this instruction efficiently is to prepare multiple virtual tables for each class: a main virtual table used for normal virtual method invocation, and one additional virtual table for each interface directly or indirectly implemented by the class[5]. Each method declared in an interface is given an index within its virtual table. After preparation, each *invokeinterface* has two arguments: an interface number, and a method index. On execution, the *invokeinterface* instruction operates its method lookup in two steps.

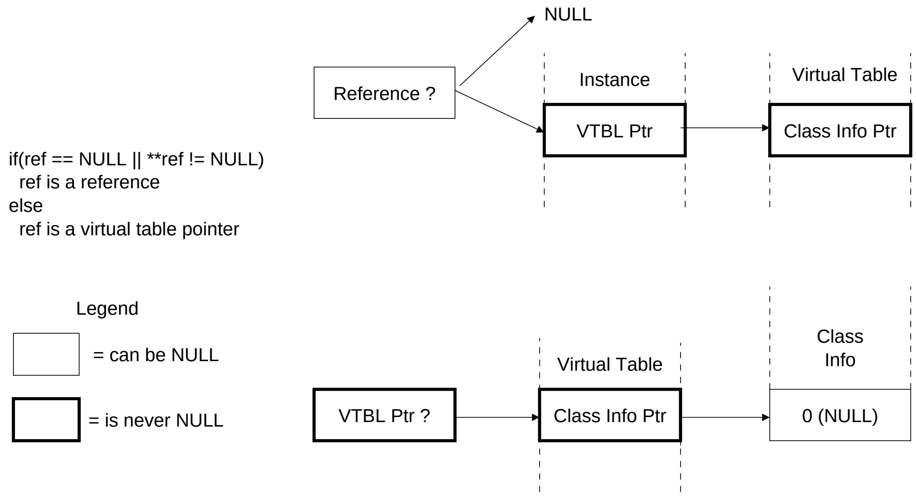

Figure 5: Alternative reference or virtual table test

It first lookups up the appropriate virtual table (using linear, binary or hashed search), then it retrieves the method pointer in a single operation from the virtual table entry located at the given method index. This interface lookup procedure has the following overhead over normal virtual method lookup: it needs to do a search to find the appropriate virtual table. It would be possible to implement an constant time lookup using *compact encoding*[32], but unfortunately, dynamic class loading requires updating this information dynamically, which is difficult to do in a multi-threaded Java environment. Our approach is simple, and does not require dynamic recomputation of tables or code rewrite.

The idea of maintaining multiple virtual tables in case of multiple inheritance is reminiscent of C++ implementations[19]. But, Java's multiple inheritance has a major semantic difference: it only applies to interfaces which may only declare method signatures without providing an implementation. Furthermore, if a Java class implements two distinct interfaces which declare the same method signature, this class satisfies both interfaces by providing a single implementation of this method. C++ allows the inheritance of distinct implementations of the same method signature.

We take advantage of this important difference to rethink the appropriate data structure needed for efficient interface method lookup. Our ideas originate from previous work on efficient method lookup in dynamically typed OO languages using of *selector-indexed dispatch tables*[12, 18, 31]. We assign a globally unique increasing index[11] to each method signature declared in an interface. A method signature declared in multiple interfaces has a single index. When the virtual table of a class is created, we also create an *interface virtual table* that grows down from the normal virtual table. This interface virtual table has a size equal to the highest index of all methods declared in the direct and indirect super interfaces of the class. For every declared super interface method, the entry at its index is filled with the address of its implementation. Interface invocation is then encoded with the *invokeinterface* instruction, and a single interface method index. The execution of *invokeinterface* can then proceed at the *exact same* cost as an *invokevirtual*. In fact, in SableVM's threaded interpreter, both instructions share the same implementation!

The interface virtual table is a sparse array of method pointers. As more interfaces are loaded, with many interface method signatures, the amount of free space in interface virtual tables grows. The traditional approach has been to use table compression techniques to reduce the amount of free space. However, these techniques are poorly adapted to concurrent and

[11] In reality, we use a decreasing index, starting at at -1, to allow direct indexing in the *interface virtual table*.

dynamic class loading environments like the Java virtual machine, as they require dynamic recompilation.

Our approach differs. Instead of compressing interface virtual tables, we simply return the free space in them to the related class loader memory manager (see section 5.3). This memory is then used to store all kinds of class loader related memory. In other words, we simply recycle the free space of sparse interface virtual tables within a class loader. The layout of interface virtual tables is illustrated in Figure 6.

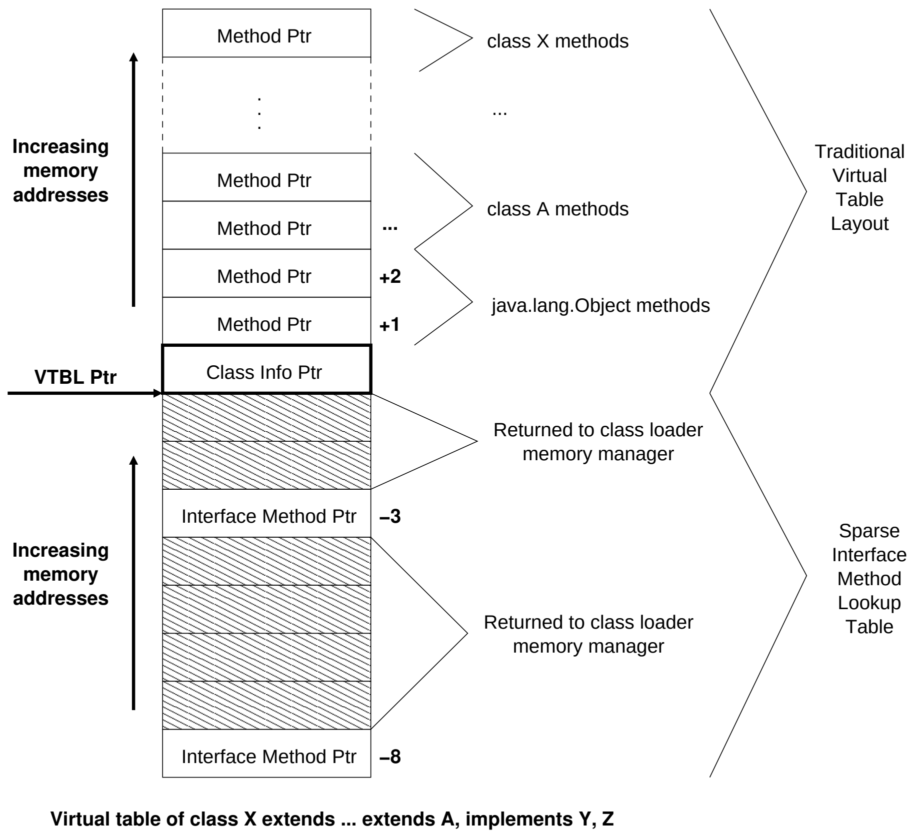

Figure 6: Virtual table layout

As interface usage in most Java programs range from very low to moderate, we could argue that it is unlikely that the free space returned by interface virtual tables will grow faster than the rate at which it is recycled. However, in order to handle pathological cases, we also provide a very simple technique, which incurs no runtime overhead, to limit the maximal growth of interface virtual tables. To limit this growth to N entries, we stop allocating new interface method indices as soon as index N is given. Then, new interface method signatures are encoded using traditional techniques. The trick to make this possible is to encode interface calls differently, based on whether the invoked method signature has been given an index or not. The traditional technique used to handle overflow can safely ignore all interface methods which have already been given an index.

6.3 Improved Thin Locks

Our final enhancement improves upon Onodera's bimodal field locking algorithm[26], a modified version of Bacon's thin lock algorithm[13], but without busy-wait transition from light to heavy mode.

Bacon's thin lock algorithm can be summarized as follows. Each object instance has a one lock word in its header[12]. To acquire the lock of an object, a thread uses the *compare-and-swap* atomic operation to compare the current lock value to zero, and replace it with its thread identifier. If the lock value isn't zero, this means that either the lock is already inflated, in which case a normal locking procedure is applied, or the lock is thin and is already acquired by some thread. In the latter case, if the owning thread is the current one, a nesting count (in the lock word) is increased. If the owning thread is not the current one, then there is contention, and Bacon's version of the algorithm busy-waits, spinning until it acquires the lock. When it is finally acquired, it is inflated. Unlocking non-inflated locks is simple. On each unlock operation, the nesting count is decreased. When it reaches 0, the lock byte is replaced by zero, releasing the lock.

The advantages of this algorithm are that a single atomic operation is needed to acquire a thin lock in absence of contention, and more importantly, no atomic operation is required to unlock an object[13].

Onodera eliminates the busy wait in case of contention on a thin lock, using a single additional bit in each object instance. The role of this *contention bit* is to indicate that some other thread is waiting to acquire the current thin lock. Onodera's algorithm differs from the previous algorithm at two points. First, when a thread fails to acquire a thin lock (because of contention), it acquires a fat monitor for the object, sets the contention bit, checks that the thin lock was not released, then puts itself in a waiting state. Second, when a thin lock is released (e.g. lock word is replaced by zero), the releasing thread checks the contention bit. If it is set, it inflates the lock, and notifies all waiting threads[14].

The overhead of Onodera's algorithm over Bacon's is the contention bit test on unlocking, a fairly simple non-atomic operation, and the one bit per object. This bit has the following restriction: it must not reside in the lock word. This is a problem. It is important to keep the per object space overhead as low as possible, as Java programs tend allocate many small objects. It is now common practice to use 2 word headers in object instances; one word for the virtual pointer, and the second for the lock and other information. The contention bit cannot reside in either of these two words (putting it in the virtual table pointer word would add execution overhead to method invocation, field access, and any other operation dereferencing this pointer). As objects need to be aligned on a word multiple (for the atomic operation to work), this one bit overhead might well translate into a whole word overhead for small objects. Also, it is likely that the placement of this bit will be highly type dependent, which complicates the unlocking test.

Our solution to this problem is to put the *contention bit* in the thread structure, instead of in the object instance. This simple modification has the advantage of eliminating the per object overhead while maintaining the key properties of the algorithm, namely, fast thin lock acquisition with a single atomic operation, fast thin lock unlocking without atomic operations, and no busy-wait in case of contention.

We modify Onodera's algorithm as follows. In SableVM, each thread has a related data structure containing various information, like stack information and exception status. In this structure, we add the contention bit, a *contention lock*[15], and a linked list of (waiting thread, object) tuples. Then we modify the lock and unlock operation as described in the following two subsections.

6.3.1 Modifications to the lock operation

The lock operation is only modified in the case of contention on a thin lock.

[12] Only 24 bits of that word are used for locking on 32 bit systems. 8 bits remain free for other uses.

[13] Unlike Agesen's recent *meta-lock* algorithm[9] which requires an atomic operation for unlocking objects.

[14] This is a simplified description. Please refer to the original paper[26] for details.

[15] The contention lock is a simple non-recursive mutex.

When a thread x_t fails to acquire a thin lock on object z_o due to contention (because thread y_t already owns the thin lock), then (1) thread x_t acquires the contention lock of the owning thread (y_t), and (2) sets the contention bit of thread y_t, then (3) checks that the lock of object z_o is still thin and owned by thread y_t. If the check fails, (4a) the contention bit is restored to its initial value, the contention lock is released and the lock operation is repeated. If the check succeeds, (4b) the tuple (x_t, z_o) is added to the linked list of thread y_t, then thread x_t is put in the waiting state (temporarily releasing the contention lock of thread y_t, while it sleeps). Later, when thread x_t wakes up (because it was signaled), it releases the re-acquired contention lock and repeats the lock operation.

6.3.2 Modifications to the unlock operation

The unlock operation is modified to check the contention bit of the currently executing thread. This check is only done when a lock is actually released (as locks are recursive), after releasing the lock.

When the lock of object b_o is released by thread y_t, and if the contention bit of thread y_t is set, then (1) thread y_t acquires its own contention lock, and (2) iterates over all the elements of its tuple linked list. For each tuple (x_t, z_o), if ($z_0 = b_o$), thread x_t is simply signaled. If ($z_o \neq b_o$), the lock of object z_o is inflated[16] (if it is thin), then thread x_t is signaled. Finally, (3) thread y_t empties its tuple linked list, clears its contention bit, and releases its contention lock.

7 Implementation Status and Performance Evaluation

SableVM is under ongoing development[6]. It can currently run single threaded, non-memory intensive applications. Within the next month, thread support and garbage collection will be added. Most of the data layouts introduced in this paper are already in place. In the final version of this paper, we will provide a section on performance measurements, comparing the speed of SableVM to other interpreters and JITs on the Linux platform, and measuring the effectiveness of our memory recycling strategy. We will also measure the dynamic average and maximal size of Java stacks. Finally, we will discuss our results.

8 Conclusion and Future Work

In this paper, we have presented SableVM, a framework for testing high-level performance enhancements and extensions to the Java virtual machine. SableVM is written in portable C and is currently under development for Intel x86 processors running the Linux operating system.

The main goal of the SableVM was the design and implementation of a virtual machine suitable for research which is easy to modify, can simply handle language/bytecode extensions, and also provides a testbed for various implementation strategies, particularly those relating to how memory is handled.

The core of SableVM is a threaded interpreter which has been designed to reduce the interpretive overhead of the instruction dispatch loop. Having an efficient interpreter allows such a system to be used in embedded systems where restricting the size of the virtual machine implementation is important. Further, a fast interpreter can be part of a mixed-mode virtual machine, where some code is interpreted and some code is compiled with a naive or optimizing JIT compiler. One goal of our work is to examine how efficient the interpreter can be, while still keeping the implementation in portable C.

In the design of SableVM, a lot of attention was paid to the layout and management of memory. Our view is that a good approach to memory will benefit both the SableVM interpreter, as well as any JIT that may be added in the future. We have provided our view of how memory

[16]Notice that thread y_t necessarily owns the lock of object z_o, as a only one lock (on object b_o) has been released by thread y_t since it last cleared its contention bit and emptied its tuple list.

can be divided into four key areas based on two criteria: its allocation and release behavior, and what part of the implementation controls its management. An underlying theme is that we want to take best advantage of particular properties of memory behavior and also take advantage of low-level memory primitives to reduce overhead.

We have also given a detailed description of how objects can be laid out in order to optimize garbage collection traces. We showed how objects can grow in both directions from the object header, thus allowing all reference fields to be placed consecutively. Further, we showed how this layout makes tracing objects simple and efficient.

Another feature of our memory management is our approach to implementing invokeinterface instructions. We showed how a sparse interface virtual table could reduce the cost of an invokeinterface instruction to that of an invokevirtual. We also demonstrated that the sparse representation need not waste memory because unused holes in the interface table could be recycled and used by the class loader memory manager.

Our last improvement was an improvement on thin locks that is also related to memory. The key point is that a cheap implementation of a thin lock without busy-wait needs an extra bit of memory, but putting this bit in the object instances means that small objects might need an extra word of memory to store this single bit. Our solution is to provide an approach where the contention bit is stored in the thread structure, rather than in the object instances. This allows the inexpensive locking mechanism, while at the same time allowing for two word object headers (with no extra overhead).

We are very excited about our work on SableVM, and we are now at the point where we can start to perform experimental studies. We plan to have experimental results examining the speed of our VM and the behavior of our suggested memory strategies for the final version of the paper. In the future we would like to push the experiments further and examine connections to different garbage collection schemes and the use of class file attributes [30] for escape analysis to improve memory allocation and garbage collection.

Other groups have expressed an interest in adding other components to the VM, including a JIT compiler. We encourage such collaboration and the SableVM source is publicly-available at `http://www.sablevm.org/`.

References

[1] Classpath. http://www.classpath.org/.

[2] GCJ. http://sources.redhat.com/java/.

[3] Harissa. http://www.irisa.fr/compose/harissa/harissa.html.

[4] HotSpot. http://java.sun.com/products/hotspot/whitepaper.html.

[5] Kaffe. http://www.kaffe.org/.

[6] SableVM. http://www.sablevm.org/.

[7] Toba. http://www.cs.arizona.edu/sumatra/toba/.

[8] Ali-Reza Adl-Tabatabai, Michał Cierniak, Guei-Yuan Lueh, Vishesh M. Parikh, and James M. Stichnoth. Fast, effective code generation in a Just-in-Time Java compiler. In *Proceedings of the ACM SIGPLAN '98 Conference on Programming Language Design and Implementation*, pages 280–290. ACM Press, 1998.

[9] Ole Agesen, David Detlefs, Alex Garthwaite, Ross Knippel, Y. S. Ramakrishna, and Derek White. An efficient meta-lock for implementing ubiquitous synchronization. In *Proceedings of the Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 207–222. ACM Press, November 1999.

[10] B. Alpern, C. R. Attanasio, J. J. Barton, M. G. Burke, P. Cheng, J.-D. Choi, A. Cocchi, S. J. Fink, D. Grove, M. Hind, S. F. Hummel, D. Lieber, V. Litvinov, M. F. Mergen, T. Ngo, J. R. Russell, V. Sarkar, M. J. Serrano, J. C. Shepherd, S. E. Smith, V. C.

Sreedhar, H. Srinivasan, and J. Whaley. The Jalapeño virtual machine. *IBM Systems Journal*, 39(1):211–238, October 2000.

[11] Bowen Alpern, Dick Attanasio, Anthony Cocchi, Derek Lieber, Stephen Smith, Ton Ngo, and John J. Barton. Implementing Jalapeno in Java. In *Proceeings of the 1999 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages & Applications (OOPSLA'99)*, volume 34.10 of *ACM Sigplan Notices*, pages 314–324. ACM Press, November 1999.

[12] Cox B. *Object-Oriented Programming: An evolutionary Approach.* Addison-Wesley, 1987.

[13] David F. Bacon, Ravi Konuru, Chet Murthy, and Mauricio Serrano. Thin locks: Featherweight synchronization for Java. In *Proceedings of the ACM SIGPLAN'98 Conference on Programming Language Design and Implementation (PLDI)*, pages 258–268. ACM Press, June 1998.

[14] Joel F. Bartlett. Compacting garbage collection with ambiguous roots. Technical Report 88.2, Digital – Western Research Laboratory, 1988.

[15] C. Chambers, D. Ungar, and E. Lee. An efficient implementation of SELF a dynamically-typed object-oriented language based on prototypes. In *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA)*, volume 24, pages 49–70. ACM Press, October 1989.

[16] Michał Cierniak, Guei-Yuan Lueh, and James N. Stichnoth. Practicing JUDO: Java under dynamic optimizations. In *Proceedings of the ACM SIGPLAN '00 Conference on Programming Language Design and Implementation*, pages 13–26, Vancouver, British Columbia, June 2000. ACM Press.

[17] Peter Deutsch and Alan M. Schiffman. Efficient implementation of the Smalltalk-80 system. In *Conference Record of the Eleventh Annual ACM Symposium on Principles of Programming Languages*, pages 297–302. ACM Press, January 1984.

[18] Karel Driesen. Selector table indexing & sparse arrays. *SIGPLAN Notices: Proc. 8th Annual Conf. Object-Oriented Programming Systems, Languages, and Applications, OOPSLA*, 28(10):259–270, September 1993.

[19] Margaret A. Ellis and Bjarne Stroustrup. *The Annotated C++ Reference Manual.* Addison-Wesley, December 1990.

[20] Anton M. Ertl. A portable Forth engine. http://www.complang.tuwien.ac.at/forth/threaded-code.html.

[21] James Gosling, Bill Joy, Guy Steele, and Gilad Bracha. *The Java Language Specification Second Edition.* Addison-Wesley, 2000.

[22] Richard Jones and Rafael Lins. *Garbage collection: algorithms for automatic dynamic memory management.* Wiley, 1996.

[23] Andreas Krall. Efficient JavaVM Just-in-Time compilation. In *Proceedings of the 1998 International Conference on Parallel Architectures and Compilation Techniques (PACT '98)*, pages 205–212. IEEE Computer Society Press, October 1998.

[24] Tim Lindholm and Frank Yellin. *The Java Virtual Machine Specification.* Addison-Wesley, second edition, 1999.

[25] Andrew C. Myers. Bidirectional object layout for separate compilation. In *OOPSLA '95 Conference Proceedings: Object-Oriented Programming Systems, Languages, and Applications*, pages 124–139. ACM Press, October 1995.

[26] Tamiya Onodera and Kiyokuni Kawachiya. A study of locking objects with bimodal fields. In *Proceeings of the 1999 ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages & Applications (OOPSLA'99)*, volume 34.10 of *ACM Sigplan Notices*, pages 223–237. ACM Press, November 1999.

[27] Ian Piumarta and Fabio Riccardi. Optimizing direct threaded code by selective inlining. In *SIGPLAN '98 Conference on Programming Language Design and Implementation*, pages 291–300. ACM Press, June 1998.

[28] William Pugh and Grant Weddell. Two-directional record layout for multiple inheritance. *ACM SIGPLAN Notices*, 25(6):85–91, June 1990.

[29] 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.

[30] Raja Vallée-Rai, Phong Co, Etienne Gagnon, Laurie Hendren, Patrick Lam, and Vijay Sundaresan. Soot - a Java bytecode optimization framework. In *Proceedings of CASCON '99*, pages 125–135, 1999.

[31] Jan Vitek and R. Nigel Horspool. Taming message passing: Efficient method look-up for dynamically typed languages. In *Object-Oriented Programming, Proceedings of the 8th European Conference ECOOP'94*, volume 821 of *Lecture Notes in Computer Science*, pages 432–449. Springer, July 1994.

[32] Jan Vitek, R. Nigel Horspool, and Andreas Krall. Efficient type inclusion tests. In *OOPSLA '97 Conference Proceedings: Object-Oriented Programming Systems, Languages, and Applications*, pages 142–157. ACM Press, October 1997.

[33] Byung-Sun Yang, Soo-Mook Moon, Seongbae Park, Junpyo Lee, SeungIl Lee, Jinpyo Park, Yoo C. Chung, Suhyun Kim, Kemal Ebcioğlu, and Erik Altman. LaTTe: A Java VM Just-in-Time compiler with fast and efficient register allocation. In *Proceedings of the 1999 International Conference on Parallel Architectures and Compilation Techniques (PACT '99)*, pages 128–138. IEEE Computer Society Press, October 1999.