An Experimental Comparison of Cache-oblivious and Cache-aware Programs DRAFT: DO NOT DISTRIBUTE

Kamen Yotov
IBM T. J. Watson Research Center
kyotov@us.ibm.com

Tom Roeder, Keshav Pingali
Cornell University
{tmroeder,pingali}@cs.cornell.edu

John Gunnels,Fred Gustavson
IBM T. J. Watson Research Center
{gunnels,fg2}@us.ibm.com

Abstract

Cache-oblivious algorithms have been advanced as a way of circumventing some of the difficulties of optimizing applications to take advantage of the memory hierarchy of modern microprocessors. These algorithms are based on the divide-and-conquer paradigm – each division step creates sub-problems of smaller size, and when the working set of a sub-problem fits in some level of the memory hierarchy, the computations in that sub-problem can be executed without suffering capacity misses at that level. In this way, divide-and-conquer algorithms adapt automatically to all levels of the memory hierarchy; in fact, for problems like matrix multiplication, matrix transpose, and FFT, these recursive algorithms are optimal to within constant factors for some theoretical models of the memory hierarchy.

An important question is the following: how well do carefully tuned cache-oblivious programs perform compared to carefully tuned cache-conscious programs for the same problem? Is there a price for obliviousness, and if so, how much performance do we lose? Somewhat surprisingly, there are few studies in the literature that have addressed this question.

This paper reports the results of such a study in the domain of dense linear algebra. Our main finding is that in this domain, even highly optimized cache-oblivious programs perform significantly worse than corresponding cache-conscious programs. We provide insights into why this is so, and suggest research directions for making cache-oblivious algorithms more competitive with cache-conscious algorithms.

1. Introduction

The contributions of this paper are the following.

- We present detailed experiments on a number of high-performance platforms that show that even highly tuned recursive cache-oblivious programs may perform significantly worse than highly tuned cache-conscious programs for the same problem.
- We argue that the performance problem arises in part because the schedule of operations in recursive codes may be sub-optimal for exploiting processor pipelines. We show that the schedule of operations in iterative codes can make better use of processor pipelines.
- We argue that the rest of the performance problem arises from memory latency. Using analytical models, we point out that cache blocking serves two purposes: it can reduce the effective latency of memory requests and it can reduce the bandwidth required from memory. We argue quantitatively that I/O optimality [28] addresses bandwidth concerns but not memory latency necessarily; therefore, recursive cache-oblivious codes may be I/O optimal but their performance may still be hurt by memory latency. In highly tuned iterative cache-conscious codes, the effective latency of memory requests is reduced by pre-fetching. We believe that this is needed in cache-oblivious programs as well; however, pre-fetching appears to be more complicated for cache-oblivious programs because of their complex control structure.

1.1 Memory Hierarchy Problem

The performance of many programs on modern computers is limited by the performance of the memory system in two ways. First, the latency of memory accesses can be many hundreds of cycles, so the processor may be stalled most of the time, waiting for loads to complete. Second, the bandwidth from memory is usually far less than the rate at which the processor can consume data.

Both problems can be addressed by using caching – if most memory requests are satisfied by some cache level, the effective memory latency as well as the bandwidth required from memory are reduced. As is well known, the effectiveness of caching for a problem depends both on the algorithm used to solve the problem, and on the program used to express that algorithm (simply put, an algorithm defines only the dataflow of the computation, while a program for a given algorithm also specifies the schedule of operations and may perform storage allocation consistent with that schedule). One useful quantity in thinking about these issues is *algorithmic data reuse*, which is an abstract measure of the number of accesses made to a typical memory location by the algorithm. For example, the standard algorithm for multiplying matrices of size $n \times n$ performs $O(n^3)$ operations on $O(n^2)$ data, so it has excellent algorithmic data reuse since each data element is accessed ***O(n)*** times; in contrast, matrix transpose performs $O(n^2)$ operations on $O(n^2)$ data, so it has poor algorithmic data reuse. When an algorithm has substantial algorithmic data reuse, the challenge is to write the program so that the memory accesses made by that program exhibit both spatial and temporal locality. In contrast, programs that encode algorithms with poor algorithmic data reuse are concerned largely with exploiting spatial locality.

1.2 Programming styles

Two programming styles are common in the domain of dense linear algebra: *iterative* and *recursive*.

In the iterative programming style, computations are implemented as nested loops. It is well known that naïve programs written in this style exhibit poor temporal locality and do not exploit caches effectively. Temporal locality can be improved by tiling the loops either manually or with restructuring compilers [30, 37]. The resulting program can be viewed as a computation over block matrices; tile sizes must be chosen so that the working set of each block computation fits in cache [12, 33]. If there are multiple cache levels, it may be necessary to tile for each one. Tiling for registers requires loop unrolling [2]. Since tile sizes are a function of cache capacity, and loop unroll factors depend on the number of available registers and on instruction cache capacity, this style of coding is called *cache-conscious* programming because the code either explicitly or implicitly embod-

ies parameters whose optimal values depend on the architecture[1]. Simple architectural models or empirical search can be used to determine these optimal values [38, 35, 13]. Cache-conscious programs for dense linear algebra problems have been investigated extensively by the numerical linear algebra community and the restructuring compiler community. The Basic Linear Algebra Subroutine (BLAS [1]) libraries produced by most vendors are cache-conscious iterative programs, as are the matrix factorization routines in libraries like LAPACK [3].

In the recursive programming style, computations are implemented with divide-and-conquer. For example, to multiply two matrices A and B, we can divide one of the matrices (say A) into two sub-matrices A_1 and A_2, and multiply A_1 and A_2 by B; the base case of this recursion is reached when both A and B have a single element. Programs written in this divide-and-conquer style perform *approximate* blocking in the following sense. Each division step generates sub-problems of smaller size, and when the working set of some subproblem fits in a given level of cache, the computation can take place without suffering capacity misses at that level. The resulting blocking is only approximate since the size of the working set may be smaller than the capacity of the cache.

An important theoretical result about divide-and-conquer algorithms was obtained by Hong and Kung [28] who also introduced the *I/O model* to study the memory hierarchy performance of algorithms. This model considers a two level memory hierarchy consisting of a cache and main memory. The processor can compute only with values in the cache, so it is necessary to move data between cache and memory during program execution. The *I/O complexity* of a program is an asymptotic measure of the total volume of data movement when that program is executed. Hong and Kung showed that divide-and-conquer programs for matrix multiplication and FFT are optimal under this measure. This result was extended to matrix transposition by Frigo et al., who also coined the adjective *cache-oblivious* to describe these programs because cache parameters are not reflected in the code [20].

A natural question is the following: *in domains like dense linear algebra in which it is important to exploit caches, how well do highly-optimized recursive programs perform compared to highly-optimized iterative cache-conscious programs*? Is there a performance penalty, in other words, that cache-oblivious recursive programs pay for the ability to adapt automatically to the memory hierarchy? Somewhat surprisingly, we have not found any definitive studies of this question in the literature; the few comparative studies that we have found have compared the performance of optimized recursive programs with that of *unoptimized* iterative programs. For example, Figure 1 shows the results of one study that found that a recursive implementation of matrix multiplication on an Itanium-2 outperforms an iterative implementation [29]. However, careful examination of this graph shows that the cache-oblivious implementation runs at about 30 Mflops; in comparison, the cache-conscious iterative native BLAS on this machine runs at almost 6 GFlops, as we discuss later in this paper.

1.3 Organization of this paper

In this paper, we describe the results of a study of the relative performance of highly-optimized recursive and iterative programs for matrix multiplication and matrix transpose on four modern architectures: IBM Power 5, Sun UltraSPARC IIIi, Intel Itanium 2, and Intel Pentium 4 Xeon. The Power 5 is an out-of-order RISC processor, the UltraSPARC is an in-order RISC processor, the Itanium is a long instruction word processor, and the Pentium is a CISC processor. Between them, these machines cover the spectrum of current high-performance processors. Key parameters of these machines are shown in Table 1. The programs we evaluate are generated by a domain-specific compiler we are building called BRILA (Block Recursive Implementation of Linear Algebra). The compiler takes recursive descriptions of linear algebra problems, and produces optimized iterative or recursive programs as output. It also implements key optimizations like scalar replacement [10], register allocation and operation scheduling at the level of the C program; these optimizations can be turned on or off as desired. Wherever appropriate, we compared the code produced by BRILA with code in libraries like ATLAS [35].

[1] Strictly speaking, these codes are both processor-conscious and cache-conscious, but we will use standard terminology and just call them cache-conscious.

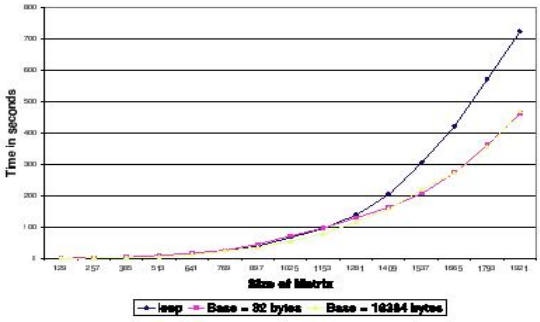


Figure 1. An empirical study of recursive and iterative matrix multiplication codes [29]

	Pentium 4 Xeon	Itanium 2	Power 5	UltraSPARC IIIi
Vendor CC	Intel C 9.0	Intel C 9.0	IBM XLC 7.0	Sun C 5.5
GCC	gcc 3.4.3	gcc 3.4.3	gcc 3.4.3	gcc 3.2.2
OS Version	Linux 2.6.9	Linux 2.6.9	IBM AIX 5.3	Sun Solaris 9
PAPI Version	3.0.8.1	3.0.8.1	3.0.8.1	3.0.8.1
BLAS Version	Intel MKL 8.0	Intel MKL 8.0	ESSL 4.2.0.2	Sun Studio 8
CPU Frequency	3.6 GHz	1.5 GHz	1.65 GHz	1.06 GHz
CPU Peak Rate	7.2 GFlops	6.0 GFlops	6.6 GFlops	2.12 GFlops
Has FMA	No	Yes	Yes	No
Has RegRelAddr	Yes	No	Yes	Yes
# of Registers	8	128	32	32
L1 Size	16 kB	16 kB	32 kB	64 kB
L1 Line Size	64 B	64 B	128 B	32 B
L2 Size	2 MB	256 kB	1.875 MB	1 MB
L2 Line Size	128 B	128 B	128 B	32 B
L3 Size	n/a	3 MB	36 MB	n/a
L3 Line Size	n/a	128 B	512 B	n/a
I-Cache Size	12 k micro-ops	16 kB	64 kB	32 kB

Table 1. Software and Hardware parameters

In Section 2, we motivate approximate blocking by giving a quantitative analysis of how blocking can reduce the required bandwidth from memory. This analysis provides a novel way of thinking about the I/O optimality of recursive algorithms for problems like matrix multiplication.

In Section 3, we discuss the performance of naïve iterative and recursive programs. These programs are *processor-oblivious* because they do not exploit registers and pipelines in the processor; they are also cache-oblivious. Therefore, neither program performs well on any architecture.

In Sections 4 and 5, we evaluate approaches for making the recursive and iterative programs processor-aware. The goal is to enable these programs to exploit registers and processor pipelines. This is accomplished by generating long basic blocks of instructions, using unrolling of loops and of recursive calls respectively, which are called from the main computations. These long basic blocks are called *microkernels* in this paper. Microkernels also serve to reduce loop and recursive call overhead. We discuss a number of algorithms for register allocation and scheduling of the microkernels, which we have implemented in the BRILA compiler. The main finding in this section is that we were unable to produce a microkernel for the recursive code that performed well, even after considerable effort. In contrast, microkernels from the iterative code obtain near peak performance. Therefore, in the rest of our studies, we only used microkernels obtained from the iterative code.

In Section 6, we study the impact of adding cache-awareness to the processor-aware code obtained in the previous section. We study the performance of programs obtained by wrapping recursive and cache-blocked iterative outer control structures around the iterative micro-kernels from the previous section. We also measure the performance obtained by using the native BLAS on these machines. The main finding in this section is that prefetching is important to obtain better performance. While prefetching is easy if the outer control structure is iterative, it is not clear how to accomplish this if the outer control structure is recursive.

Section 7 presents some preliminary findings about Matrix Transposition. Section 8 discusses related work, and Section 9 concludes with ideas for improving the performance of cache-oblivious algorithms.

2. Approximate Blocking

In this section, we give a quantitative estimate of the impact of blocking on effective memory latency as well as on the bandwidth required from memory. This analysis provides a novel way of looking at approximate blocking in cache-oblivious programs. As a running example, we use Matrix-Matrix Multiply (MMM) on the Intel Itanium 2 architecture. The Itanium 2 can execute 2 FMAs (fused multiply-adds) per cycle, so to multiply two $N \times N$ matrices, this platform would ideally spend $\frac{N^3}{2}$ cycles. However, any naïve version of matrix multiplication will take much longer because the processor spends most of its time waiting for memory loads to complete.

To examine the impact of blocking on the overhead from memory latency and bandwidth, we first consider a simple, two-level memory model consisting of one cache level and memory. The cache is of capacity C, with line size L_C, and has access latency l_C. The access latency of main memory is l_M. We consider blocked MMM, in which each block computation multiplies matrices of size $N_B \times N_B$. We assume that there is no data reuse between block computations.

2.1 Upper Bound on N_B

We derive an upper bound on N_B by requiring the size of the working set of the block computation to be less than the capacity of the cache, C. The working set depends on the schedule of operations, but it is bounded above by the size of the sub-problem. Therefore, the following inequality is a conservative approximation, although better approximations exist [38].

$$3N_B^2 \leq C \tag{1}$$

2.2 Effect of Blocking on Latency

The total number of memory accesses each block computation makes is $4N_B^3$. Each block computation brings $3N_B^2$ data into the cache, which results in $\frac{3N_B^2}{L_C}$ cold misses. If the block size is chosen so that the working set fits in the cache and there are no conflict misses, the cache miss ratio of the complete block computation is $\frac{3}{4N_B \times L_C}$. Assuming that memory accesses are not overlapped, the expected memory access latency is as follows.

$$l = \left(1 - \frac{3}{4N_B \times L_C}\right) \times l_C + \frac{3}{4N_B \times L_C} \times l_M \tag{2}$$

Equation 2 shows that the expected latency decreases with increasing N_B, so latency is minimized by choosing the largest N_B for which the working set fits in the cache. In practice, the expected memory latency computed from Equation 2 is somewhat pessimistic because loads can be overlapped with each other or with actual computations, reducing the effective values of l_C and l_M. These optimizations are extremely important in the generation of the micro-kernels, as we describe in Section 4. Furthermore, hardware and software prefetching can also be used to reduce effective latency, as discussed in Section 6.

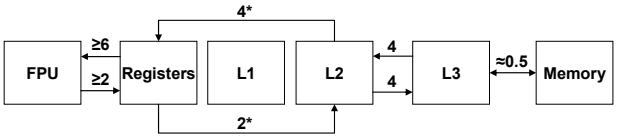


Figure 2. Bandwidth of the Itanium 2 memory hierarchy, measured in doubles/cycle. *Note: (1) Floating-point values are not cached at L1 in Itanium 2, they are transferred directly to / from L2 cache; (2) L2 cache can transfer 4 values to floating point registers and 2 values from floating point registers per cycle, but there is a maximum total of 4 memory operations.

2.3 Effect of Blocking on Bandwidth

In the restructuring compiler community, blocking is seen as a technique for reducing the effective latency of memory accesses. To understand the virtues of the cache-oblivious approach, it is better to view blocking as a technique for reducing the bandwidth required from memory.

Each FMA operation in MMM reads three values and writes one value. The required bandwidth to perform these reads and writes is $4N^3 \div \frac{N^3}{2} = 8$ doubles/cycle. Figure 2 shows the bandwidth between different levels of the memory hierarchy of the Itanium (floating-point values are not cached in the L1 cache on the Itanium). It can be seen that the register-file can sustain the required bandwidth but memory cannot.

To reduce the bandwidth required from memory, we can block the computation for the register-file. Since each block computation requires $4N_B^2$ data to be moved, our simple memory model implies that the total data movement is $\left(\frac{N}{N_B}\right)^3 \times 4N_B^2 = \frac{4N^3}{N_B}$. The ideal execution time of the computation is still $\frac{N^3}{2}$, so the bandwidth required from memory is $\frac{4N^3}{N_B} \div \frac{N^3}{2} = \frac{8}{N_B}$ doubles/cycle. Therefore, cache blocking by a factor of N_B reduces the bandwidth required from memory by the same factor.

We can now write the following lower bound on the value of N_B, where $B(L1, M)$ is the bandwidth between cache and memory.

$$\frac{8}{N_B} \leq B(L1, M) \tag{3}$$

Inequalities 1 and 3 imply the following inequality for N_B:

$$\frac{8}{B(L1, M)} \leq N_B \leq \sqrt{\frac{C}{3}} \tag{4}$$

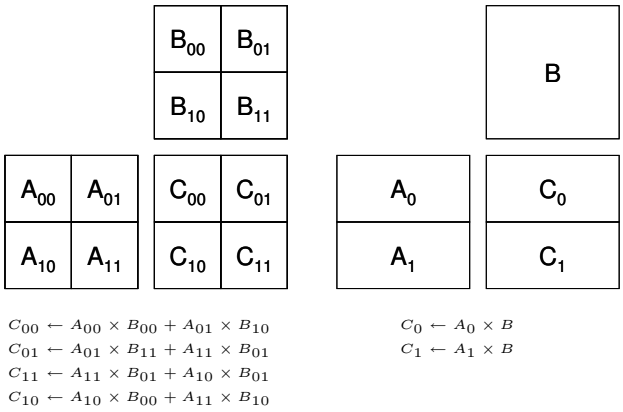
This argument generalizes to a multi-level memory hierarchy. If $B(L_i, L_{i+1})$ is the bandwidth between levels i and $i+1$ in the memory hierarchy, $N_B\,(i)$ is the block size for the i^{th} cache level, and C_i is the capacity of this cache, we obtain the following inequality:

$$\frac{8}{B(L_i, L_{i+1})} \leq N_B\,(i) \leq \sqrt{\frac{C_i}{3}} \tag{5}$$

In principle, there may be no values of $N_B\,(i)$ that satisfy the inequality. This can happen if the capacity of the cache as well as the bandwidth to the next level of the memory hierarchy are small. According to this model, the bandwidth problem for such problems cannot be solved by blocking[2].

For the Itanium 2, we have seen that register blocking is needed to prevent the bandwidth between registers and L2 cache from becoming the bottleneck. If $N_B\,(R)$ is the size of the register block, we see that $\frac{8}{4} \leq N_B\,(R) \leq \sqrt{\frac{126}{3}}$. Therefore, $N_B\,(R)$ values between 2 and 6 will suffice. If we use $N_B\,(R)$ in this range, the bandwidth required from L2 to registers is between 1.33 and 4 doubles per

[2] In practice, there may be other bottlenecks such as the inability of the processor to sustain a sufficient number of outstanding memory requests.



(a) All Dimensions (b) Largest Dimension

Figure 3. Two divide-and-conquer strategies for MMM

cycle. Note that this much bandwidth is also available between the L2 and L3 caches. Therefore, it is not necessary to block for the L2 cache to ameliorate bandwidth problems. Unfortunately, this bandwidth exceeds the bandwidth between L3 cache and memory. Therefore, we need to block for the L3 cache. The appropriate inequality is $\frac{8}{0.5} \leq N_B\left(L_3\right) \leq \sqrt{\frac{4\text{MB}}{3}}$. Therefore, $N_B\left(L_3\right)$ values between 16 and 418 will suffice.

Thus, for the Itanium 2, there is a range of block sizes that can be used. Since the upper bound in each range is more than twice the lower bound, the approximate blocking of a divide-and-conquer implementation of a cache-oblivious program will generate subproblems in these ranges, and therefore bandwidth is not a constraint. Of course, latency of memory accesses may still be a problem. In particular, since blocking by cache-oblivious programs is only approximate, the analysis of Section 2.2 suggests that reducing the impact of memory latency is more critical for cache-oblivious codes than it is for cache-conscious codes. We will revisit this point in more detail in Section 6.

3. Naïve codes

In this section, we discuss naïve recursive and iterative programs that are oblivious to both the processor and the memory hierarchy. There are two high-level design decisions to be made when writing either program: what control structure and what data structure to use.

Figure 3 shows two recursive control structures for implementing matrix multiplication. A well-known approach is to bisect both A and B along rows and columns, generating eight independent subproblems as shown in Figure 3(a). The recursion terminates when the matrices consist of single elements. For obvious reasons, we refer to this strategy as the *all-dimensions* (AD) strategy.

Bilardi et al. [9] have pointed out that it is possible to optimize memory hierarchy performance by using a Gray code order to schedule the eight sub-problems so that there is always one sub-matrix in common between successive sub-problems. One such order can be found in Figure 3(a) if the sub-problems are executed in left-to-right, top-to-bottom order. For example, the first two sub-problems have C_{00} in common, and the second and third have A_{01} in common.

An alternative control strategy is the *largest-dimension* (LD) strategy proposed by Frigo et al. [20], in which only the largest of the three dimensions is divided, as shown in Figure 3(b). Both the AD and LD strategies lead to programs that are optimal in the Hong and Kung I/O complexity model [28].

As a baseline for performance comparisons, we used the simple iterative version of matrix multiplication. The three loops in this program are fully permutable, so all six orders of the loop nest compute the same values. In our experiments, we used the *jki* order. For the experiments in this section, we chose row-major array order. Note that the *jki* loop order is the worst loop order for exploiting spatial locality if arrays stored in row-major order (as discussed in [14]). We chose this order to eliminate any advantage the iterative code might obtain from exploiting spatial locality.

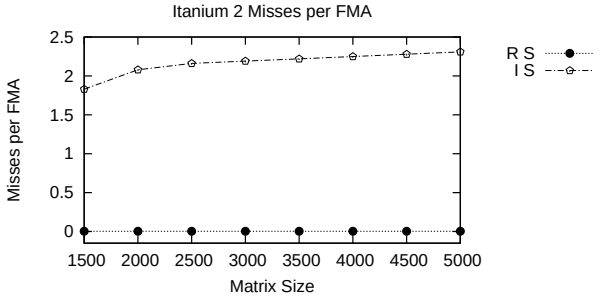


Figure 4. Data cache misses per FMA instruction in MMM

As an aside, we mention that we investigated Morton-Z storage order [27] as an alternative to row-major order. Accessing array elements is substantially more complex for Morton-Z order, especially for matrices whose dimensions are not a power of two. Even for matrices whose dimensions are a power of two, we rarely found any improvement in performance. This finding is consistent with previous studies that have concluded that space-filling storage orders like Morton-Z order pay off only when the computation is out of core [8].

Figure 10 shows the results of performing complete MMMs on the machines in our study (for lack of space, we have consolidated all the performance graphs for each machine into a single graph). Since we explore a large number of implementations in this paper, we use a tuple to distinguish them, the first part of which describes the outer control structure.

- R – using the Recursive AD control structure;
- I – using a triply-nested Iterative control structure;

The second part of the tuple describes the microkernel, and it will be explained as the microkernels are developed in Sections 4 and 5. However, when the outer control structure invokes a single statement to perform the computations, we use the symbol S (for Statement). For completeness, we include performance lines for MMM performed by the Vendor BLAS using standard row-major storage format for the arrays.

With this notation, note that the lines labelled R S in Figure 10 shows the performance of the AD cache-oblivious program, while the lines labelled I S shows the performance of the nested loop program. Both programs perform very poorly, obtaining roughly 1% of peak on all the machines. As a point of comparison, vendor BLAS on the Itanium 2 achieves close to peak performance. The performance of LD was close to that of AD on all machines, so we do not discuss it further.

3.1 Discussion

To get some insight into why these programs perform so poorly, we studied the assembly listings and the values of various hardware counters on the four machines. This study revealed three important reasons for the poor performance.

- As is well-known, the major problem with the recursive program is the overhead of recursion, since each division step in the divide-and-conquer process involves a procedure call. Our measurements on the Itanium showed that this overhead is roughly 100 cycles per FMA, while on the UltraSPARC, it is roughly 360 cycles. This integer overhead is much less for the iterative program.
- A second reason for poor performance is that the programs make poor use of registers. Compilers do not track register values across

procedure calls, so register blocking for the recursive code is difficult. In principle, compilers can perform register blocking for the iterative program, but none of the compilers were able to accomplish this.

- Finally, a remarkable fact emerges when we look at the number of L2 cache misses on the Itanium. Figure 4 shows the number of cache misses per FMA for the iterative and recursive programs. The iterative program suffers roughly two misses per FMA. This makes intuitive sense because for the *jki* loop order, the accesses to A_{ik} and C_{ij} miss in the cache since the A and C arrays are stored in row-major order but are accessed in column-major order. The element B_{kj} is invariant in the innermost loop, so it does not cause cache misses. Therefore, each iteration of the innermost loop performs one FMA and misses on two references, resulting in a miss ratio of 0.5. In short, poor memory hierarchy behavior limits the performance of the iterative code. Remarkably, the recursive program suffers only 0.002 misses per FMA, resulting in a miss ratio of 0.0005! This low miss ratio is a practical manifestation of the theoretical I/O optimality of the recursive program. Nevertheless, the poor performance of this code shows that I/O optimality alone does not guarantee good overall performance.

To improve performance, it is necessary to massage the recursive and iterative codes so that they become more processor-aware and exploit the processor pipeline and the register file. Section 4 describes how processor-awareness can be added to recursive codes. Section 5 describes how this can be done for iterative codes.

4. Processor-aware recursive codes

To make the recursive program processor-aware, we generate a long basic block of operations called a microkernel that is obtained by unrolling the recursive code completely for a problem of size $R_U \times R_U \times R_U$ [20]. The overall recursive program invoke the microkernel as its base case. There are two advantages to this approach. First, it is possible to perform register allocation and scheduling of operations in the microkernel, thereby exploiting registers and the processor pipeline. Second, the overhead of recursion is reduced.

We call the long basic block a *recursive* microkernel since the multiply-add operations are performed in the same order as they were in the original recursive code. The optimal value of R_U is determined empirically for values between 1 and 15.

Together with the control structure, one needs to worry about which data structure to use to represent the matrices. Virtually all high-performance BLAS libraries internally use a form of a blocked matrix, such as Row-Block-Row (RBR). An alternative is to use a recursive data layout, such as a space filling curve like Morton-Z [27]. We compared the MMM performance using both these choices and we rarely saw any performance improvement using Morton-Z order over RBR. Thus we use RBR in all experiments in this paper, and we chose the data block size to match our kernel block size[3].

We considered three different approaches to performing register allocation and scheduling for the microkernel.

4.1 R($R_U \times R_U \times R_U$, NN)

The first approach is to use the native compiler on each platform to compile the microkernel. We call this version R($R_U \times R_U \times R_U$, NN) because it is generated from Recursive inlining when data is R_U along the three dimensions; NN stands for Native-None, and it means that the native compiler is used to schedule the code and no other register allocation is performed.

Figure 5 shows the performance of different microkernels in isolation on the four architectures of interest. Intuitively, this is the performance obtained by a microkernel if all of its memory accesses are satisfied by the highest cache level (L2 on the Itanium and L1 on the other machines). This performance is measured by invoking the microkernel repeatedly on the same data (the RBR format ensures that there are no conflict misses at the highest cache level).

We focus on the UltraSPARC results. Figure 5(b) shows that the performance of the microkernel on the UltraSPARC in isolation is only about 11% of peak. This is also the performance of the complete MMM computation using this microkernel (190 MFlops out of 2.12 GFlops). The line labelled "ideal" corresponds to the highest performance one can achieve for a microkernel of given size, given the cost of ramping up and draining the computations of the microkernel. The line labelled "T(4x4x120,BC)" shows the performance of the best *iterative* microkernel, discussed in detail in Section 5.

Overall performance is better than that of the naïve recursive version discussed in Section 3 because the overhead of recursion is amortized over the computations in the microkernel. An examination of the assembly code, however, showed that the compilers were not able to register allocate array elements. This optimization requires the compiler to discover whether or not the matrices are aliased. Even in the best production compilers, this alias analysis is often insufficient.

Some compilers can be told to assume that there is no aliasing in the microkernel. We found that the Intel C compiler (version 9.0) on the Itanium 2 was able to produce code comparable in performance to that of our most advanced recursive microkernel (Section 4.3) if it is told that no aliasing occurs in the microkernel.

4.2 R($R_U \times R_U \times R_U$, BB)

At Frigo's suggestion [18], we addressed this problem by implementing modules in the BRILA compiler that (i) used Belady's algorithm [5] to perform register allocation on the unrolled microkernel code, and then (ii) performed scheduling on the resulting code. Our implementation of Belady's algorithm is along the lines of [26]. This code was then compiled using the native compiler on each platform. In these experiments, we ensured that the native compiler was used only as a portable assembler, and that it did not perform any optimizations that interfered with BRILA optimizations.

The key idea behind using Belady's algorithm is that when it is necessary to spill a register, the value that will be needed furthest in the future should be evicted. This value is easy to discover in the context of microkernels, since we have one large basic block. The Belady register allocation algorithm is guaranteed to produce an allocation that results in the minimum number of loads. Different architectures require slightly different versions of the allocator. For instance, on the Itanium 2, Belady register allocation is implemented in two passes – one to allocate floating-point registers and a subsequent one to allocate integer registers. This division is necessary because the Itanium 2 architecture does not have a register-relative addressing mode, so the address of each memory operation needs to be pre-computed into an integer register. To decide on an allocation for the integer registers, we need to know the order of floating-point memory operations, but this order is not known before the floating-point registers themselves are allocated.

The BRILA scheduler is a simplified version of a general instruction scheduler found in compilers, since it has to handle only a basic block of floating-point FMAs (or multiplies and adds when the architecture does not have an FMA instruction), floating-point loads and stores, and potentially integer adds (for pointer arithmetic on Itanium 2). It accepts a simple description of the architecture and uses it to schedule the instructions appropriately. A brief description of the scheduler is presented in Figure 6.

We call the resulting microkernel, generated by using the Belady register allocation algorithm and the BRILA scheduler, R($R_U \times R_U \times R_U$, BB), where BB stands for BRILA-Belady.

Figure 5(b) shows that on the UltraSPARC, the performance in isolation of this microkernel is above 40% of peak for $R_U > 3$. The performance of the complete MMM is only at about 640 MFlops, or just about 32% of the 2 GFlops peak rate. Note that on the Itanium 2,

[3] However, the native BLAS on all the machines use standard row-major order for the input arrays and copy these arrays internally into RBR format, so care should be taken in performance comparisons with the native BLAS.

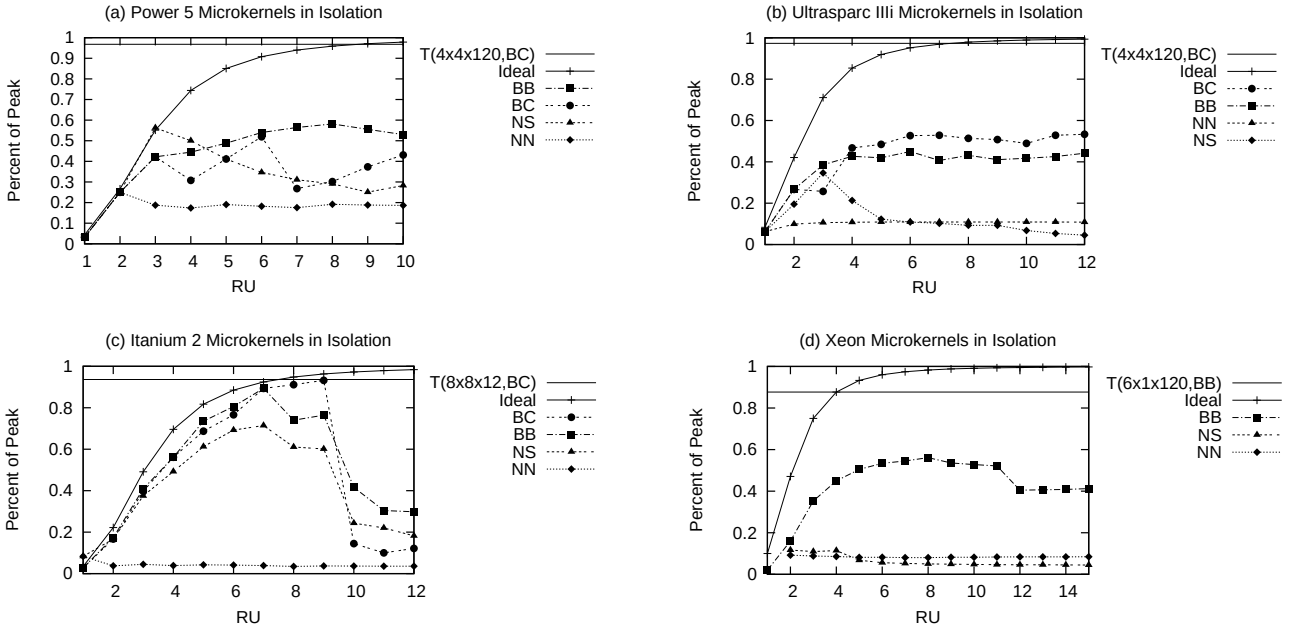


Figure 5. Microkernel performance in isolation

- The scheduler works on blocks of the following instruction types:
  - Floating-point FMA, multiply, and add;
  - Floating-point load and store;
  - Integer arithmetic for address computation.
- The scheduler is parameterized by a description of the target architecture, which consists of:
  - HasFMA : bool – specifies whether the architecture has a floating-point FMA instruction.
  - HasRegRelAddr : bool – specifies whether the architecture supports register relative addressing mode, or all addresses need to computed into an integer register in advance (e.g. Itanium 2).
  - Latency : instruction → bool – specifies the latency in cycles of all instructions of interest.
  - Set of possible instruction bundles, each of which can be dispatched in a single cycle. The way we describe this set is by first mapping each instruction of interest to an instruction *type*. Each instruction type can be dispatched to one or more different execution *ports* inside the processor. Finally, the processor can dispatch at most one instruction to each execution port, for a subset of execution ports per cycle. We enumerate the possible sets of execution ports that can be dispatched together.
- The scheduler produces instruction bundles at each step as follows:
  1. Considers all instructions which have not been scheduled yet;
  2. Without changing the relative order of the instructions, removes all instructions from the list which depend on instructions that have not been scheduled yet;
  3. Greedily selects the largest subset of instructions from the resulting list which matches one of the subsets of execution ports the processor supports. It ensures that:
     - Instructions of the same type execute in program order;
     - Instructions from different types are given different execute preferences. The scheduler prefers to dispatch computational instructions most, followed by loads and stores, followed by integer arithmetic (if necessary).

Figure 6. The BRILA instruction scheduler

register spills hurt the performance of this microkernel for $R_U > 7$. An even greater drop occurs for $R_U > 9$ because the microkernel overflows the I-Cache.

Interestingly, for the R($R_U \times R_U \times R_U$, NS) microkernel, Figure 10(a) shows that the IBM XLC Compiler at its highest optimization level is able to produce code which is slightly faster than the corresponding BB microkernel.

4.3 R($R_U \times R_U \times R_U$, BC)

Although Belady's algorithm minimizes the number of loads, it is not necessarily the best approach to register allocation on modern machines; for example, we can afford to execute more than the optimal number of loads from memory if they can be performed in parallel with each other or with computation[4]. Therefore, we decided to investigate an integrated approach to register allocation and scheduling [23, 7, 31]. Figure 7 briefly describes the algorithm we implemented in BRILA.

Both UltraSPARC IIIi and Itanium 2 are in-order architectures, and precise scheduling is extremely important for achieving high-performance. Figure 5(b) and 5(c) show that the BC strategy works better on these architectures than the other strategies discussed in this section. As we can see in Figure 5(b), the performance of this microkernel in isolation on the UltraSPARC is about 50% of peak for $R_U > 3$. The performance of the complete MMM is about 760 MFlops, or just about 38% of peak. On the Itanium 2 architecture, the performance of the microkernel in isolation is 93% of peak. Although this level of performance is not sustained in the complete MMM, Figure 10(c) shows that the complete MMM reaches about 3.8 GFlops or about 63% of peak.

The situation is more complex for the Power 5 and Pentium 4 Xeon since these are out-of-order architectures and the hardware reorders instructions during execution. Figures 10(a) and 5(a) show that the Belady register allocation strategy (BB) performs better on Power 5 than the integrated graph coloring and scheduling approach (BC).

[4] Belady invented his policy while investigating page replacement policies for virtual memory systems, and the algorithm is optimal in that context since page faults cannot be overlapped. Basic blocks, however, have both computation and memory accesses to schedule, and can overlap them to gain higher performance.

1. Generate the sequence of FMA operations in the same way we do this for $\mathsf{R}(R_U \times R_U \times R_U, \mathsf{NN})$
2. Generate an approximate schedule for this sequence:
   1. Consider the issue width of the processor for floating-point operations and schedule that many FMA instructions per cycle. Assume that an arbitrary number of other instructions (memory, integer arithmetic) can be executed in each cycle.
   2. If the processor has no FMA instruction, break each FMA into its two components, replace the FMA with its multiply part, and schedule its add part for $\mathsf{Latency}\,(\mathsf{multiply})$ cycles later;
   3. Assume an infinite virtual register file, and allocate each operand of each computational floating-point instruction (FMA, multiply, or add) into a different virtual register.
      1. Schedule a memory load into a register $\mathsf{Latency}\,(\mathsf{load})$ cycles before the FMA (or multiply) using the corresponding value.
      2. Schedule a memory store from a register $\mathsf{Latency}\,(\mathsf{FMA})$ (or $\mathsf{Latency}\,(\mathsf{add})$) cycles after the FMA (or add) that modifies that register.
      3. Whenever the life spans of two registers that hold the same physical matrix element overlap, we merge them into a single virtual register and eliminate unnecessary intermediate loads and stores. This step is required to preserve the semantics of the microkernel.
      4. Additionally, when the life spans of two registers that hold the same physical matrix element do not overlap, but are close (say at most δ cycles apart), we still merge them to take advantage of this reuse opportunity. This step is not required to preserve the correctness of the program, but can allow significant reuse of already loaded values. The parameter δ depends on architectural parameters.
5. Use graph coloring to generate a virtual to logical register mapping.
6. Use the BRILA scheduler, described in Figure 6, for the corresponding architecture to produce a final schedule for the microkernel.

Figure 7. Integrated register allocator and scheduler in BRILA

Intuitively, this occurs because the out-of-order hardware schedules around stalls caused by the Belady register allocation.

On the Pentium 4 Xeon, there are too few registers to perform complex scheduling. Our previous experience with the x86 architecture is that it is better to let the out-of-order execution hardware perform register renaming and instruction reordering [38]. Therefore, we used the Belady register allocation algorithm and scheduled dependent loads, multiplies, and adds *back-to-back*. Figure 5(d) shows that the microkernel in isolation gets roughly 50% of peak.

4.4 Discussion

Our work on adding processor-awareness to recursive MMM codes led us to the following conclusions.

- The microkernel is critical to overall performance. Producing a high-performance microkernel is a non-trivial job, and requires substantial programming effort.
- The performance of the program obtained by following the canonical recipe (recursive outer control structure and recursive microkernel) is substantially lower than the near-peak performance of highly optimized iterative codes produced by ATLAS or in vendor BLAS. The best we were able to obtain was 63% of peak on the Itanium 2; on the UltraSPARC, performance was only 38% of peak.
- For generating the microkernel code, using Belady's algorithm followed by scheduling may not be optimal. Belady's algorithm minimizes the number of loads, but minimizing loads does not necessarily maximize performance. An integrated register allocation and scheduling approach appears to perform better.
- Most compilers we used did not do a good job with register allocation and scheduling for long basic blocks. This problem has been investigated before [23, 7, 31]. The situation is more muddied when processors perform register renaming and out-of-order instruction scheduling. The compiler community needs to pay more attention to this problem.

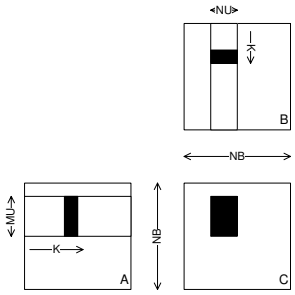


Figure 8. Iterative microkernel used in ATLAS

5. Processor-aware iterative codes

We now discuss how processor-awareness can be added to iterative codes.

5.1 Iterative microkernels

The ATLAS system and many other numerical linear algebra libraries use iterative microkernels whose structure is shown pictorially in Figure 8. Unlike the recursive microkernels described in Section 4 that have a single degree of freedom R_U, the iterative microkernels have three degrees of freedom called K_U, N_U, and M_U. The microkernel loads a block of the C matrix of size $M_U \times N_U$ into registers, and then accumulates the results of performing a sequence of size K_U of outer products between small column vectors of A and small row vectors of B.

Our iterative microkernels are generated by BRILA as follows.

1. Start with a simple kji triply-nested loop for performing an MMM with dimensions $\langle K_U, N_U, M_U \rangle$ and unroll it completely to produce a sequence of $M_U \times N_U \times K_U$ FMAs.
2. Use the algorithm described in Figure 7 for register allocation and scheduling, starting with the sequence of FMAs generated above. As in Section 4.3, we use Belady register allocation and schedule dependent instructions back-to-back on the Pentium 4 Xeon.

We examined the schedule of our microkernel and compared it to the structure of the ATLAS microkernel, which is shown in Figure 8. Both perform the computation instructions in the same order and keep the submatrix of C in registers at all times. Our compiler uses a description of the architecture to schedule the loads from A and B more precisely. ATLAS relies on the native compiler.

The iterative microkernel generated in this way has a number of advantages. For the kji loop order, the number of required registers does not depend on the K_U parameter [38]. Thus we can optimize the values of M_U and N_U to make the working set of the microkernel fit in the register file. Then, we can optimize the value of K_U to make the code of the microkernel fit in the instruction cache. In principle, we can generate recursive microkernels for non-square blocks, but their dimensions are not independent since each dimension affects both register allocation and instruction cache utilization.

Table 2 shows the performance of our iterative microkernels in isolation (also shown as a solid flat horizontal line in Figure 5(a-d)). We name the iterative microkernels with T for Tiled, the block size $M_U \times N_U \times K_U$ and the allocation - scheduling pair (BC or BB).

It can be seen that iterative microkernels perform substantially better than recursive microkernels on most architectures, obtaining close to peak performance on most of them.

Architecture	Micro-Kernel	Percent
Power 5	R($8 \times 8 \times 8$, BB)	58%
	T($4 \times 4 \times 120$, BC)	98%
UltraSPARC IIIi	R($12 \times 12 \times 12$, BC)	53%
	T($4 \times 4 \times 120$, BC)	98%
Itanium 2	R($9 \times 9 \times 9$, BC)	93%
	T($8 \times 8 \times 12$, BC)	94%
Pentium 4 Xeon	R($8 \times 8 \times 8$, BB)	56%
	T($6 \times 1 \times 120$, BB)	87%

Table 2. Performance of the best microkernels in isolation.

5.2 Overall MMM Performance

To perform complete MMMs, the iterative microkernel is wrapped in an outer control structure consisting of a triply-nested loop that invokes the iterative microkernel within its body. The resulting code is processor-aware but not cache-aware, and therefore has a working set of a matrix, a panel of another matrix and a block from the third matrix; because it uses a microkernel, it does provide register blocking. The experimental results are labelled with I, followed by the microkernel name from Table 2 in Figure 10(a-d).

On all four machines, the performance trends are similar. When the problem size is small, performance is great because the highly-tuned iterative microkernel obtains its inputs from the highest cache level. However, as the problem size increases, performance drops rapidly because there is no cache blocking. This can be seen most clearly on the Power 5. Performance of I T($4 \times 4 \times 120$, BC) is initially at 5.8 GFlops. When the working set of the iterative version becomes larger than the 1920KB L2 cache (for matrices of size $480 \times 480 \times 480$), performance drops to about 3.8 GFlops. Finally, when the working set of the iterative version becomes larger than the 36MB L3 cache (for matrices of size $2040 \times 2040 \times 2400$), performance drops further to about 2 GFlops, about 30% of peak.

5.3 Discussion

Table 2 shows that on a given architecture, iterative microkernels are larger in size than recursive microkernels. It is possible to produce larger iterative microkernels because of the decoupling of the problem dimensions: the size of the K_U dimension is limited only by the capacity of the instruction cache, and is practically unlimited if a software-pipelined loop is introduced along K_U.

In summary, iterative microkernels outperform recursive microkernels by a wide margin on three out of four architectures we studied; performance of the recursive microkernel was close to that of the iterative microkernel only on the Itanium. Since overall MMM performance is bounded above by the performance of the microkernel, these results suggest that use of recursive microkernels is not recommended. In the rest of this paper, we will therefore focus exclusively on iterative microkernels. However, the benefits of a highly optimized iterative microkernel are obtained only for small problem sizes. We address this problem next.

6. Incorporating cache blocking

Without cache blocking, the performance advantages of the highly optimized iterative microkernels described in Section 5 are obtained only for small problem sizes; once the working set of the problem is larger than the capacity of the highest cache level, performance drops off. To sustain performance, it is necessary to block for the memory hierarchy.

In this section, we describe two ways of accomplishing this. The first approach is to wrap the iterative microkernel in a recursive outer control structure to perform approximate blocking. The second approach is to use iterative outer control structures and perform explicit cache tiling.

6.1 Recursive outer control structure

Figure 10 presents the complete MMM performance of the iterative microkernels within a recursive outer structure. The corresponding lines are labelled with R followed by the name of the microkernel from Table 2. On all four machines, performance stays more or less constant independent of the problem size, demonstrating that the recursive outer control structure is able to block approximately for all cache levels. The achieved performance is between 60% (on the UltraSPARC IIIi) and 75% (on the Power 5) of peak. While this is good, overall performance is still substantially less than the performance of the native BLAS on these machines. For example, on the Itanium, this approach gives roughly 4 GFlops, whereas vendor BLAS obtains almost 6 GFlops; on the Ultrasparc III, this approach obtains roughly 1.2 GFlops, whereas vendor BLAS gets close to 1.6 GFlops.

6.2 Blocked iterative outer control structure

These experiments suggest that if an iterative outer control structure is used, one approach is to tile explicitly for all levels of the memory hierarchy. A different approach that leads to even better performance emerges if one studies the handcoded BLAS on various machines. On most machines, the handcoded BLAS libraries are proprietary, so we did not have access to them. However, the ATLAS distribution has hand-optimized codes for some of the machines in our study, so we used those in our experiments. The minikernels in these codes incorporate one level of cache tiling, and perform prefetching so that while one minikernel was executing, data for the next minikernel was prefetched from memory. The resulting performance was very close to that of vendor BLAS on all machines (we do not show these performance lines in Figure 10 to avoid cluttering the figure).

To mimic this structure, we used the BRILA compiler to generate a *minikernel* composed of a loop nest wrapped around the $M_U \times N_U \times K_U$ iterative microkernel; the minikernel performs a matrix multiplication of size $N_B \times N_B \times N_B$. This is essentially the structure of the minikernel used in the ATLAS system [35]. For our experiments, we set N_B to 120. As with the microkernels, this minikernel can then be used with either recursive or iterative outer control structures. The experimental results in Figure 10 show a number of interesting points. The recursive and iterative outer control structures achieve almost identical performance for most problem sizes. For instance, on the Power 5, R T(120,4x4x120,BC) reaches nearly 6 GFlops and maintains its performance through matrix sizes of 5000×5000. I T(120,4x4x120,BC) matches this performance until the matrices become too large to fit in the L2 cache; performance then falls off because there is no tiling for the L3 cache.

We have not yet implemented prefetching in BRILA, but for iterative minikernels, the memory access pattern is regular and predictable, so instructions that touch memory locations required for successive microkernels can be inserted into the computationally intensive code of earlier microkernels without performance penalty. However, it is not clear how one introduces prefetching into programs with a recursive outer control structure. Following the line of reasoning described in Section 2, we believe this is required to raise the level of performance of the recursive approach to that of the iterative approach. Whether prefetching can be done in some cache-oblivious manner remains to be seen.

6.3 Discussion

Our minikernel work led us to the following conclusions.

- Wrapping a recursive control structure around the iterative microkernel gives a program that performs reasonably well since it is able to block approximately for all levels of cache and block exactly for registers.
- If an iterative outer control structure is used, it is necessary to block for relevant levels of the memory hierarchy.
- To achieve performance competitive with hand-tuned kernels, minikernels need to do data prefetching. It is clear how to do

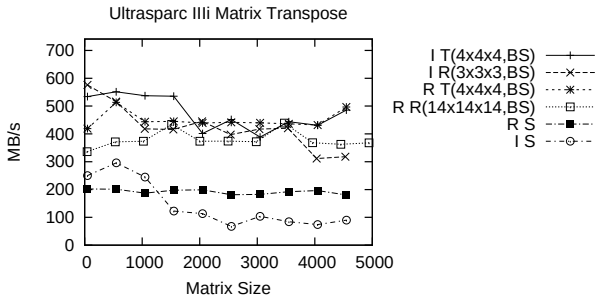


Figure 9. Out-of-place Matrix Transpose on the UltraSPARC IIIi.

this for an iterative outer control structure but it is not clear how to do this for a recursive outer control structure.

7. Matrix Transpose

We have just started our study of out-of-place Matrix Transposition (MT), a kernel that is very different in behavior than MMM. Unlike MMM, MT does $O(N^2)$ work on $O(N^2)$ data, so there is no algorithmic reuse, but it can benefit from exploiting spatial locality in data cache and data TLB. There are no multiply-add operations in the microkernel, but an important performance metric is the rate at which data is stored into B.

Figure 9 shows the results of running various Transpose algorithms on the UltraSPARC IIIi. We use the same notation for these algorithms as we did for MMM. For Transpose, all matrices are stored in Row-Block-Row order when they are blocked. The naïve algorithms operate on the standard row-major data structure.

We used the BRILA compiler to produce the microkernel as follows. We unroll the iterative or recursive code as for MMM, but instead of scheduling the loads and stores to perform the multiply-adds as early as possible, the scheduler tries to store into B as early as possible. The label BS corresponds to BRILA-Store, since our metric involves optimizing the stores in the microkernel.

The iterative transpose algorithm I S, consisting of a doubly nested loop, performs reasonably well for small matrices, but performance quickly falls below that of the recursive algorithm as the size of the matrix grows. This drop in performance corresponds to an increase in cache misses: the doubly nested loop walks one of the matrices in row major order but the other in column major. As less and less of the matrix fits into the cache, more and more cycles of the iterative algorithm are taken up in waiting for data from memory.

The recursive algorithm R S has more consistent performance, but it does not achieve the higher performance of the blocked kernels. As in MMM, its recursive structure provides enough locality to avoid the high cache miss penalty even for large matrices, but the overhead of recursing down to 1×1 is too high.

We investigated all combinations of outer and inner blocked control structures. Figure 9 shows that any amount of blocking provides better performance than the naïve R S and I S versions. The blocking provides spatial locality, and avoids the cache misses discussed for the iterative outer structure while reducing the recursive overhead enough to improve the performance of the recursive outer structure.

We plan to compare the performance of these generated versions with the performance of hand-written transpose programs, and we will report those numbers in the final paper.

8. Related Work

Four domains are closely related to our research. First, hand-tuned numerical libraries are critical to our work since they provide a high water mark for performance on different architectures. They also suggest optimization strategies that can often be incorporated into compilers. Second, the cache-oblivious program advocated originally in [20] has grown to encompass computations for many domains; these algorithms guide our recursive implementation and data structures. Third, the restructuring compiler community has developed many techniques for improving the memory behavior of programs. Finally, some studies have compared recursive and iterative code for matrix multiply as well as linear and recursive blocking for data structures.

8.1 Numerical linear algebra libraries

There is a vast literature on this subject. A magisterial survey of this field can be found in the classic text by Golub and van Loan [22].

The central routines in dense numerical linear algebra are the Basic Linear Algebra Subroutines (BLAS) [1]. Matrix multiplication is perhaps the most important routine in the BLAS. Most high-performance BLAS routines are produced by hand (for example Goto [24]).

The ATLAS system [35] allows a measure of automation in generating BLAS libraries. ATLAS is essentially a code generator that can generate BLAS routines, given the values of certain numerical parameters like the L1 cache tile size, the register tile sizes, etc. Optimal values of these parameters are determined by using empirical search over a sub-space of possible parameter values. The ATLAS distribution also has hand-tuned versions of BLAS routines that it uses to produce the library if they perform better than the code produced by the code generator. On most machines, these hand-tuned versions perform better than the code produced by using empirical search.

8.2 Restructuring compilers

There is a large body of existing work on compiler transformations for restructuring high-level algorithmic descriptions to improve memory hierarchy performance. These include linear loop transformations [4, 16, 30, 36], loop tiling [37] and loop unrolling [2]. Other work has focused on algorithms for estimating optimal values for parameters associated with these transformations, such as tile sizes [12, 33] and loop unroll factors [2].

8.3 Cache-Oblivious Algorithms

This area of work was inspired by the classic paper of Hong and Kung in which they introduced the I/O model and showed that divide-and-conquer versions of matrix multiplication and FFT are I/O optimal [28]. Frigo, Leiserson and co-workers generalized the results in this paper; they also coined the term "cache-oblivious algorithms" [20]. They also produced FFTW [19], the highly successful FFT library generator. Independent of this work, the scientific computing community has investigated divide-and-conquer algorithms [27, 11, 15].

A significant body of work (for example [34, 9, 6, 19]) has been inspired by the original work on cache-oblivious algorithms. Other work has investigated cache-oblivious data structures, starting from the idea of recursive storage order, and currently, efforts [21, 32] are underway to improve native compiler support for these orders. These recursive data structures allow the data structure for MMM to match its control structure. Their main problem is the need to compute the block offset; this computation is $O(1)$ for regular iterative blocking strategies, but is in general $O(logN)$ for recursive data structures.

The work on cache-oblivious algorithms and data structures is complementary to our own; we integrate results from this literature into the BRILA compiler. We rely on these results to provide high performance cache-oblivious formulations of common linear algebra problems. Our work has now turned up several directions for further research in cache-oblivious algorithms.

8.4 Comparison Studies

The work of Bilardi et al. [9] is closest to ours. They were interested in the issue of *performance portability* – if you are restricted to using a single C program on all architectures, what fraction of peak performance do you get for matrix multiplication on different architectures? Their study used a cache-oblivious program, and they compared its performance with the performance of the code produced by ATLAS.

Our work is complementary to this work since we permit the BRILA compiler to perform architecture dependent optimizations. In other words, we are interested in the portability of the compiler, and necessarily of the application code itself.

Frens and Wise have implemented cache-oblivious programs for MMM and other numerical routines[17]. It appears that this work uses recursive microkernels, and they have not studied the use of iterative kernels with different outer control structures as we have done.

9. Future Work

The results in this paper suggest several directions for future research.

- The performance of recursive microkernels is substantially worse than that of iterative microkernels on all architectures we studied. How can we produce better recursive microkernels?
- Better register allocation and scheduling techniques are needed for long basic blocks. Contrary to popular belief, using Belady's algorithm followed by scheduling is not necessarily optimal because minimizing the number of loads is not necessarily correlated to optimizing performance on architectures that support multiple outstanding loads and can overlap loads with computation.
- Wrapping a recursive outer structure around an iterative microkernel provides approximate blocking for all levels of cache, and performs better than wrapping a simple loop around the microkernel. However, pre-fetching is easier for the iterative outer structure, and boosts performance well above that of the recursive outer structure program. How do we integrate prefetching into cache-oblivious algorithms?
- The naïve recursive code described in Section 3 is I/O optimal, but delivers only 1% of peak on all architectures. Intuitively, the I/O complexity of a program describes only one dimension of its behavior, and focusing on I/O optimality alone may be misleading when it comes to overall performance. What models are appropriate for describing other dimensions of program behavior to obtain a comprehensive description of program performance?

Acknowledgements: We would like to thank Matteo Frigo and Gianfranco Bilardi for many useful discussions.

References

[1] Basic Linear Algebra Routines (BLAS). `http://www.netlib.org/blas`.

[2] R. Allan and K. Kennedy. *Optimizing Compilers for Modern Architectures*. Morgan Kaufmann Publishers, 2002.

[3] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, editors. *LAPACK Users' Guide. Second Edition.* SIAM, Philadelphia, 1995.

[4] Uptal Banerjee. Unimodular transformations of double loops. In *Languages and compilers for parallel computing*, pages 192–219, 1990.

[5] L. A. Belady. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal*, 5(2):78–101, 1966.

[6] Michael A. Bender, Jeremy T. Fineman, Seth Gilbert, and Bradly C. Kuszmaul. Concurrent cache-oblivious b-trees. In *Proc. of the 17th Annual ACM Symposium on Parallelism in Algorithms and Architectures*, pages 228–237, 2005.

[7] David A. Berson, Rajiv Gupta, and Mary Lou Soffa. Integrated instruction scheduling and register allocation techniques. In *LCPC '98*, pages 247–262, London, UK, 1999. Springer-Verlag.

[8] Gianfranco Bilardi. Personal communication, 2005.

[9] Gianfranco Bilardi, Paolo D'Alberto, and Alex Nicolau. Fractal matrix multiplication: A case study on portability of cache performance. In *Algorithm Engineering: 5th International Workshop, WAE*, 2001.

[10] David Callahan, Steve Carr, and Ken Kennedy. Improving register allocation for subscripted variables. In *PLDI*, pages 53–65, 1990.

[11] Siddhartha Chatterjee, Alvin R. Lebeck, Praveen K. Patnala, and Mithuna Thottethodi. Recursive array layouts and fast parallel matrix multiplication. In *ACM Symposium on Parallel Algorithms and Architectures*, pages 222–231, 1999.

[12] S. Coleman and K. S. McKinley. Tile size selection using cache organization and data layout. In *PLDI*, 1995.

[13] Keith D. Cooper, Devika Subramanian, and Linda Torczon. Adaptive optimizing compilers for the 21st century. *J. Supercomput.*, 23(1):7–22, 2002.

[14] J. J. Dongarra, F. Gustavson, and A. Karp. Implementing linear algebra algorithms for dense matrices on a vector pipeline machine. *SIAM Review*, 26(1):91–112, 1984.

[15] E. Elmroth, F. G. Gustavson, B. Kågström, and I. Jonsson. Recursive blocked algorithms and hybrid data structures for dense matrix library software. *SIAM Review*, 46(1):3–45, March 2004.

[16] Paul Feautrier. Some efficient solutions to the affine scheduling problem - part 1: one dimensional time. *International Journal of Parallel Programming*, October 1992.

[17] Jeremy D. Frens and David S. Wise. Auto-blocking matrix-multiplication, or tracking blas3 performance from source code. In *Proc. ACM Symp. on Principles and Practice of Parallel Programming*, pages 206–216, 1997.

[18] Matteo Frigo. Personal communication, 2005.

[19] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2), 2005.

[20] Matteo Frigo, Charles E. Leiserson, Harald Prokop, and Sridhar Ramachandran. Cache-oblivious algorithms. In *FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, page 285. IEEE Computer Society, 1999.

[21] Steven Gabriel and David Wise. The opie compiler: from row-major source to morton-ordered matrices. In *Proc. 3rd Workshop on Memory Performance Issues*, pages 136–144, 2004.

[22] Gene Golub and Charles Van Loan. *Matrix Computations*. John Hopkins University Press, 1993.

[23] J. R. Goodman and W.-C. Hsu. Code scheduling and register allocation in large basic blocks. In *ICS '88*, pages 442–452, New York, NY, USA, 1988. ACM Press.

[24] Kazushige Goto and Robert van de Geijn. On reducing TLB misses in matrix multiplication. FLAME working note #9. Technical report, The University of Texas at Austin, Department of Computer Science, Nov 2002.

[25] John Gunnels, Fred Gustavson, Greg Henry, and Robert van de Geijn. Matrix multiplication kernels: Synergy between theory and practice leads to superior performance. In *PARA*, 2004.

[26] Jia Guo, María Jesús Garzarán, and David Padua. The power of Belady's algorithm in register allocation for long basic blocks. In *Proc. 16th International Workshop in Languages and Parallel Computing*, pages 374–390, 2003.

[27] Fred Gustavson. Recursion leads to automatic variable blocking for dense linear-algebra algorithms. *IBM Journal of Research and Development*, 41(6):737–755, 1999.

[28] Jia-Wei Hong and H. T. Kung. I/O complexity: The red-blue pebble game. In *Proc. of the thirteenth annual ACM symposium on Theory of computing*, pages 326–333, 1981.

[29] Piyush Kumar. Cache-oblivious algorithms. In *Lecture Notes in Computer Science 2625*. Springer-Verlag, 1998.

[30] W. Li and K. Pingali. Access Normalization: Loop restructuring for NUMA compilers. *ACM Transactions on Computer Systems*, 1993.

[31] Cindy Norris and Lori L. Pollock. An experimental study of several cooperative register allocation and instruction scheduling strategies. In *MICRO 28*, pages 169–179, Los Alamitos, CA, USA, 1995. IEEE Computer Society Press.

[32] Rajeev Raman and David Wise. Converting to and from the dilated integers. http://www.cs.indiana.edu/ dswise/Arcee/castingDilated-comb.pdf, 2004.

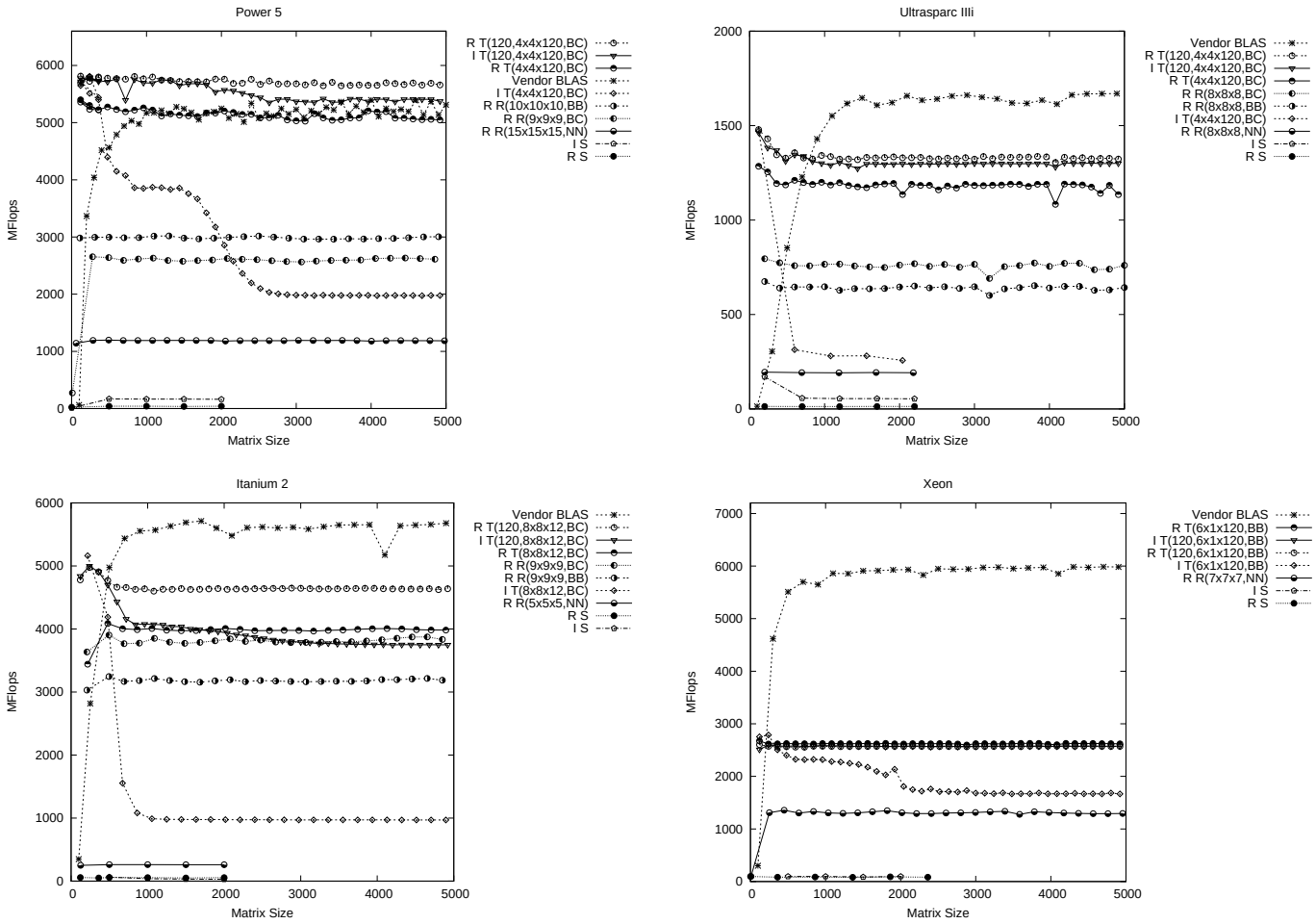


Figure 10. Complete MMM performance

[33] Robert Schreiber and Jack Dongarra. Automatic blocking of nested loops. Technical Report CS-90-108, Knoxville, TN 37996, USA, 1990.

[34] Sivan Toledo. A survey of out-of-core algorithms in numerical linear algebra. In *External memory algorithms*. American Mathematical Society, Boston, MA, 1999.

[35] R. Clint Whaley, Antoine Petitet, and Jack J. Dongarra. Automated empirical optimization of software and the ATLAS project. *Parallel Computing*, 27(1–2):3–35, 2001.

[36] Michael E. Wolf and Monica S. Lam. An algorithmic approach to compound loop transformations. In *Advances in Languages and Compilers for Parallel Computing*. Pitman Publisher, 1991.

[37] M. Wolfe. Iteration space tiling for memory hierarchies. In *Third SIAM Conference on Parallel Processing for Scientific Computing*, December 1987.

[38] Kamen Yotov, Xiaoming Li, Gang Ren, Maria Garzaran, David Padua, Keshav Pingali, and Paul Stodghill. Is search really necessary to generate high-performance BLAS? *Proceedings of the IEEE*, 93(2), 2005.