Introducing A New Cache Design into Vector Computers*

Qing Yang
Dept. of Electrical and Computer Engineering
University of Rhode Island, Kingston, RI, 02881
email qyang@ele.uri.edu

ABSTRACT

This paper introduces an innovative cache design for vector computers, called prime-mapped cache. By utilizing the special properties of a Mersenne prime, the new design does not increase the critical path length of a processor, nor does it increase the cache access time as compared to existing cache organizations. The prime-mapped cache minimizes cache miss ratio caused by line interferences that have been shown to be critical for numerical applications by previous investigators. With negligibly additional hardware cost, we observe significant performance gains by adding the proposed cache memory into an existing vector computer. We study the performance of the new design analytically based on a generic vector computation model. The analytical model is validated through extensive simulation experiments. Our performance analysis on various vector access patterns shows that the prime-mapped cache performs significantly better than conventional cache organizations in the vector processing environment. The performance gain will increase with the increase of the speed gap between processors and memories.

1 Introduction

While cache memories have been successfully used in general purpose computers to boost system performance [1], their effectiveness for vector processing has not been established. Most of existing supercomputer vector processors typically do not have cache memories because of perceived poor performance for vectorized numerical algorithms. This common believe results primarily from three considerations: First, numerical programs generally have data sets that are too large for the current cache sizes. Sweep accesses of a large vector may result in complete reloading of the cache before the processor reuses them. Secondly, address sequentiality which has been an important assumption in conventional caches may not be as good in vectorized numerical algorithms that usually access data

*This research is partially supported by National Science Foundation under grants No. CCR-8909672 and MIP-9208041. This paper is a revised version of "A novel cache design for vector processing" by Qing Yang and Liping W. Yang which appeared in Proceedings of the 19th International Symposium on Computer Architecture, Gold Coast, Australia, May 1992.

with certain *stride* which is the difference between addresses associated with consecutive vector elements. And third, register files and highly interleaved memories have been commonly used to achieve high memory bandwidth required by vector processing. It is not clear whether cache memories can significantly improve the performance of such systems. However, with the rapid advances in device technology and increased speed gap between processor and memory [2], it has become increasingly important to study the performance of cache memories for vector processors [3]

The first concern about possible poor performance of vector cache memories (we will call a cache for vector data a *vector cache*) has been studied by a number of researchers [4, 5, 6]. It is well known that the memory hierarchy can be better utilized if numerical algorithms are *blocked*. Blocking is a general program optimization technique that promotes data reuse in high speed memories. It has been shown that blocking is very effective for many algorithms in linear algebra [6]. Lam et al. have recently studied the cache performance of a blocked matrix multiplication algorithm on general purpose computers [7]. It is shown in their study that blocking factor (the size of inner loop) has significant impact on cache performance. In [4], So and Zecca presented an interesting performance study for vector caches by means of trace driven simulations [4]. They have shown that although the program locality of vector executions is significantly different from that of scalar executions the cache hit ratio is high enough to take advantage of a cache. Their study is based on traces of a set of fixed size programs that are either cache-optimized subroutines from the machine library or highly vectorized for vector machines. It is not clear how the cache behaves for generally blocked programs with different problem sizes.

Sequentiality of vector addresses depends on vector access stride that varies widely in numerical algorithms. Since the basic storage unit in a cache is a cache line that consists of a group of consecutive memory words, cache pollution may result if the access stride is not one. Large cache lines may exploit the spatial locality of vector accesses with small stride but may lead to poor cache performance for large strides [8]. Small cache lines, on the other hand, may increase the number of cache misses depending on the vector stride. Fu and Patel [8] have recently presented a comprehensive study on the effects of cache line sizes on the performance of vector caches. They proposed two prefetching schemes, sequential-prefetching and stride-prefetching, for vector caches to reduce the influence of long stride vector accesses. As shown in [8], certain performance improvements as result of the two prefetching schemes are obtained. However, the cache miss ratios for some applications considered in [8] are still very high (40% in some cases). This is because that not only does the poor sequentiality of vector data result in cache pollutions but also a large amount of *interference misses* [7].

In spite of the fact that most supercomputer vector machines use interleaved memory and large register files to speed up memory accesses, the increased speed gap between processor and memory may still favor a cache memory to bridge this gap. This is because not only is a register file relatively small that can hardly hold the working set of a program but also it requires extra efforts in software to manage it. Cache memory, on the other hands, is transparent to programmers. Highly interleaved memory may provide enough bandwidth to single stream vector accesses, but the memory speed has to be extremely fast and the number of interleaved memory modules has

to be excessively large in order to provide enough bandwidth for multiple stream vector accesses. Baily has shown in his study [9] that hundreds and even thousands of interleaved memory modules are needed to achieve a reasonable memory performance for multiple stream vector accesses. Furthermore, vector processing has become a mainstream form of computing ranging from superminis to workstations. Vector processors have also been incorporated into mainframes as built-in accelerators for computationally intensive applications. For these types of machines, a cache memory can be a cost-effective enhancement towards a smooth memory hierarchy [4, 10, 8]. Several recent vector computers feature a cache-based memory hierarchy such as IBM 3090 [4], Alliant FX/8 [10] and VAX 6000 [11].

Although cache memories have potential for improving the performance of future vector processors, there are practical reasons why such vector caches have not yet been satisfactorily efficient. A single miss in the vector cache results in a number of processor stall cycles equal to the entire memory access time, while the memory accesses of a vector processor without cache are fully pipelined. In order to benefit from a vector cache, the miss ratio must be kept extremely small. In general, cache misses can be classified into three categories [2]: compulsory, capacity, and conflicts. The compulsory misses are the misses in the initial loading of data, which can be properly pipelined in a vector computer. The capacity misses are due to the size limitation of a cache to hold data between references. If application algorithms are properly blocked as mentioned above, the capacity misses can be attributed to the compulsory misses for the initial loading of each block of data provided that the block size is less than cache size. The last category, conflict misses, plays a key role in the vector processing environment. Conflicts can occur when two or more elements of the same vector are mapped to the same cache line or elements from two different vectors compete for the same cache line. The former is called *self-interference* whereas the latter is called *cross-interference* [7]. The recent study on blocked matrix multiply algorithm [7] shows that the self-interference misses increase drastically and dominate cache misses after the fraction of a 16K-word cache being used exceeds 3%.

Since conflict misses that significantly degrade vector cache performance have a lot to do with vector access stride, one may wish to adjust the size of an application problem to make a good access stride for a given machine. However, not only does this approach give a programmer a burden of knowing architecture details of a machine but also infeasible for many applications. It is known that the number of conflicts is minimum if the stride of accessing a vector is relative prime to the number of sets which is a power of 2 in conventional caches. Note that the stride required to access the major diagonal of a matrix is one greater than the stride required to access a row of the matrix stored in a column-major. Therefore, it is not possible to make both row access and major diagonal access efficient because one stride or the other is not relative prime to the number of sets of any direct or set-associative cache.

In this paper, we propose a new cache design suitable for use in vector processing and in particular for vector computers. Our objective is to provide a vector cache memory system that is capable of efficiently handling numerical programs having data sets of various sizes and various access strides. Our new cache design, referred to as *prime-mapped cache*, minimizes cache miss ratio in vector processing. The primary feature of the design is that memory data are mapped into

cache lines according to a Mersenne prime [12]. The cache access logic of the new mapping scheme keeps virtually the same as the traditional cache organizations, resulting in no additional delay for cache accesses. At the meantime, the address generation takes no longer than the normal address calculation time due to the special properties of the Mersenne number. Moreover, generating addresses for cache access is done in parallel with normal address calculations resulting in no performance penalty. Thus, it has the advantages of both direct-mapped cache and fully associative cache if replacement is not considered.

In order to evaluate the performance potential of the new design, we will carry out performance analysis on various architecture approaches. Evaluating the performance of cache-based computer systems is a difficult task because of the complexity of program behaviors. One needs to consider the locality property of an application and reuse factors among other things. Traditional performance evaluation techniques can be broadly classified into three categories: trace-driven simulation[13], event-driven simulation[14] and analytical modeling [15, 16]. Trace-driven simulation is based on actual traces of programs running on a system. Therefore, it provides the most reliable and accurate performance estimates for given programs on a given system. Since the collection of traces is a time consuming process and requires special hardware or software supports, trace-driven simulation may not be effective in capturing the exact performance behavior of hypothetical architectures. Moreover, address traces are collected from representative programs of fixed sizes, which is not appropriate for evaluating vector cache performance because the performance of vector processing is very sensitive to problem sizes. It is shown in [7] that an algorithm with one problem size can run at twice the speed of the same algorithm with a different size. If one considers a wide range of sizes for each problem, on the other hand, the cost of collecting traces would be prohibitively large. We therefore present performance study by means of analytical modeling and discrete event-driven simulations. Analytical models are developed for two simple vector processor models: memory-register vector processor model and cache-based vector processor model. The advantage of analytical modeling is that it provides us with a quick and insightful performance estimate of a given design without the limitation of application problem sizes. Based on the analytical model, one can easily pinpoint where the potential performance bottleneck is. Furthermore, most numerical algorithms are highly structured, which makes analysis possible and fairly accurate. We also carry out discrete event-driven simulation experiments to validate our analysis. It is shown that our analytical models match very well with the simulation experiments.

As examples of applying our analytical models, we analyze caching behaviors of three typical vector access patterns, namely random multistride access, subblock access and FFT access. Numerical results have shown that the prime-mapped cache outperforms both the vector computer without cache and the vector computer with a conventional cache for all vector access patterns considered. The performance improvement ranges from 40% to a factor of 3 depending on the memory cycle time, blocking factor and access patterns.

The paper is organized as follows. Section 2 presents the cache design and discussions on hardware issues. In Section 3, we present an analytical performance model for two vector processors with and without a cache. We will carry out performance analyses and comparisons under different access patterns for different architectures in Section 4. Section 5 concludes the paper.

2 Minimizing Vector Cache Misses

In this section, we consider several alternatives to reduce vector cache misses. By vector cache, we mean, in this paper, the cache that holds vector data accessed by vector load or store instructions. We assume that scalar data have a separate cache [11].

2.1 Can Associativity Help?

Since cache misses in a vector cache result mainly from line conflicts as discussed in the introduction, one would naturally think of set-associative-mapped cache with higher associativity or fully associative cache. It is well known that higher associativity has the advantages of less line conflicts and the flexibility of implementing a replacement algorithm such as LRU. Lam et al [7] have shown that the average miss rates of a set-associative cache are relatively lower than that of direct-mapped cache. However, the fraction of cache used in case of set-associative cache still remains very small (in the range of 20% to 60%) though it is larger than that of the direct-mapped cache. Moreover, the standard deviation of miss rate increases steadily with the blocking factor implying that the execution time for some problem sizes can be significantly worse than others. For the same cache size, increasing associativity results in decreased number of sets that data can be mapped to although several cache lines can be mapped to the same set. As a result, we will not see significant reduction in terms of interference misses. As an example, consider 2 alternative designs of an eight-line cache memory: 2-way set-associative and direct-mapped. Suppose that a vector is loaded into the cache with a stride that is an even number. No matter which one of the two designs one wishes to consider, only four or less number of cache lines can be placed in the cache. In other words, the number of line interferences is the same for either case if the vector has more than 4 elements. While this example is simple, it demonstrates a potential problem that exists in any existing cache design except for a fully associative cache. Furthermore, increasing the associativity also increases the cache hit time [17] which has negative effects on system performance. As to the replacement algorithm, due to the nature of numerical algorithms, we may not be able to take this advantage as we do for most other applications since serial access to vectors dictates against LRU replacement [3]. Whether there exists a better replacement algorithm needs further study.

2.2 Effects of Line Size

Cache line size is one of the most important parameters in a cache design. Larger cache line sizes reduce compulsory misses if an application has a high spatial locality. At the meantime, larger cache lines also reduces the number of lines in the cache, giving rise to more conflicts. In addition, non-unit access strides may also result in cache pollutions since the loaded excess data may never be used before being replaced. Cache pollution degrades cache performance significantly because not only do the unused data use cache space but also memory bandwidth [8]. Fu and Patel [8] observed that cache line sizes have unpredictable impacts on vector cache performance. The best cache line size of one program may be the worst for another program in a given cache design. Their observations suggest that an optimal cache line size for vector processing be difficult to determine unless all

application programs have the same stride characteristics. There are also other considerations in deciding cache line size such as bus width, degree of memory interleaving etc.. Cache line size selection for the vector cache is more complicated than general purpose computers.

2.3 Prime-Mapped Cache

It is clear from the above discussions that any incremental work on the existing cache organization will not help significantly for vector processing. We propose a cache design with a novel mapping scheme called *prime-mapping* that attempts to minimize interference misses. The idea behind the new design is simple. Instead of having 2^c sets in the cache, we allow only a prime number of sets to reside in the cache. Using a prime number in a memory system to avoid memory access conflicts was attempted before. In the early 70's, Budnik and Kuck suggested the use of prime number of memory modules in the context of a parallel computer [18]. The idea was latter developed by Burroughs in the design of the BSP computer [19]. However, addressing for such prime-number memory systems is much more complex than for memory systems in which the number of memory modules is a power of 2. This complicated addressing is more critical in cache design since we can not afford to allow any additional delay for cache accesses. Any increase in cache hit time will affect not only the average access time but also the clock rate of the CPU. Therefore, it is of essential importance that any attempt in designing a new cache must ensure no increase in the length of the critical path of a processor.

Our approach here is to utilize the special properties of a class of prime numbers: the Mersenne primes. A Mersenne number is of the form $2^c - 1$ for some c that makes $2^c - 1$ a prime. Examples of such c are 2, 3, 5, 7, 13, 17, and 19 etc.. A cache line with address A_l is mapped into set number $A_l \ mod \ (2^c - 1)$ in the cache. Since Mersenne number is a prime, a vector access to the cache with this mapping scheme can be made conflict-free. Meanwhile, the address space is also well utilized since the number of logical set in the prime-mapped cache is just one less than a power of 2.

A detailed description of the *Prime-mapped* scheme is as follows: Each memory address, same as conventional cache-based computer system[17], is partitioned into three fields: $W = log_2(line\ size)$ bits of word address in a line (offset); $c = log_2(number\ of\ sets + 1)$ bits of index; and the remaining tag bits of tag. The access logic of the prime-mapped cache consists of three components: data memory, tag memory, and matching logic. Same as a set-associative cache, the data memory contains a set of address decoders and cached data; the tag memory stores tags corresponding to the cached lines; and the matching logic checks if the tag in an issued address matches the tag in the cache. The cache lookup process is exactly the same as the set-associative cache and hence takes the same amount of time as the set-associative cache. However, the index field used to access the data memory is not just a subfield of the original address word issued by the processor since the modulus for cache mapping is not a power of 2 any more. It is the residue of the line address modulo a Mersenne number. How efficient this modular operation can be done is essential to the cache performance. In the following, we show how the index conversion can be done without adding additional delay in the address mapping process.

Figure 1 shows the block diagram of the address computation logic. It consists of two parallel parts: one normal address calculation unit that generates addresses for accessing the main mem-

ory and the other address generator that calculates addresses for accessing the vector cache. For notational convenience, we call the address for accessing the main memory *as memory address* and the one for accessing the vector cache as *cache address*. The normal address calculation unit is a part of address control logic or functional unit that any existing vector computers should have. The memory address of a vector element is calculated based on a vector stride and the address of the previous element of the vector. For generating a cache address, the tag field and the word (offset) field are the same as that for memory address. It is only the index field (c bits) that needs to be calculated in a Mersenne number form through the additional address generator. Since a Mersenne number is defined as an integer modulo 2^c-1, arithmetic operations are greatly simplified by noting that $2^c \ mod \ (2^c-1) = 1$. Consider any two integers represented in c-bit binary form. If we add the two numbers we obtain a $(c+1)$-bit sum with the most significant bit being carry bit. But $2^c \equiv 1 \ mod \ (2^c-1)$. Thus, addition modulo a Mersenne number is performed very simply by using a conventional full binary adder of c bits and by folding the most significant carry bit output back into the least significant carry bit input. Therefore, a c-bit full adder is sufficient to perform the address generation. The index field of the cache address of a vector element is obtained by adding the stride to the index field of the cache address of the previous element of the vector. Note that both the stride and the previous index value should be in the Mersenne number form as explained shortly. Because the addition is performed on a c-bit field which is a portion of a memory word, this process should take no longer than the normal address calculation process. Two addresses are therefore generated concurrently: one for cache access and the other for memory access in case of a cache miss.

It remains to consider how to generate the cache address of the first element of a vector and the vector stride in a Mersenne number form efficiently. Let the starting line address of a vector be A_{start} that consists of two fields: tag_A and $index_A$ with the lengths of each field being tag and c, respectively. We ignore the offset field of an address since a line is the basic unit for cache mapping. Suppose the least significant c bits of tag_A are represented by tag_{A1} and the second c bits are represented by tag_{A2} and so forth. In order to map the starting line of the vector into the prime-mapped cache, we need to perform a modular operation $A_{start} \ mod \ (2^c-1)$. Since

$$A_{start} = index_A + 2^c tag_{A1} + 2^{2c} tag_{A2} + \cdots,$$

and $2^c \ mod \ (2^c-1) = 1$, we can write

$$A_{start} \ mod \ (2^c-1) = index_A + tag_{A1} + tag_{A2} + \cdots.$$

Therefore, to derive the index field of the starting line address of a vector, what we need to do is to perform a sequence of c bit additions. If the length of tag_A (tag) in bits is less than or comparable to the length of $index_A$ (c), then we just need one c-bit addition plus taking care of the carry bit, which is equivalent to a $2c$-bit addition. Note that this $2c$-bit addition should take less time than the normal address calculation which operates on whole word length. In practice, it is often the case that the tag field is comparable to the index field. For example, the cache size of the Alliant FX/8 is 128K bytes (16K double words) [10] giving rise to an index length of 14 bits if the line size is 8 bytes. If the address length is 32 bits, the remaining tag field consists of maximum of 15 bits.

We then need to perform a 14-bit addition and add back the most significant bit of the starting address and the carry bit into the least significant bit of the sum. Similarly, the VAX 6000 Model 400 vector processor has a cache of 1 Mbytes that can easily make the tag field comparable to the index field. If it is desired in a design that the tag field be much longer than the index field, we would need at most one or a few more level of additions of c bits in calculating the index field of the starting address of a vector.

The address conversion logic for the starting element of a vector is also shown in Figure 1 assuming $tag \leq c$. Two multiplexors are used to select addends for the $c-$bit adder. The multiplexors will select two fields (tag and index) of the memory address word to do addition if it is for the starting element. For all other elements of the vector, the multiplexors will select index value of the cache address of the previous element and the stride to do addition. The generation of the vector starting address for cache access can be done in parallel with the memory access of the vector element if the vector is accessed first time. It is safe to assume that a couple of stages of c bit additions can finish before the first element of a vector loaded from the main memory arrives in the vector cache. Thus, the loaded data can be written into the vector cache using the newly calculated index value. In this case, there is no performance penalty resulting from the address conversion. The converted cache address of the starting element can be stored in a special register for future reuses if the vector is going to be accessed again. Subsequent accesses to the elements in the same vector can be done using two possible addresses: cache addresses as discussed above or memory addresses if a miss occurs. The calculation of these addresses does not incur any additional delay compared to existing machines as evidenced previously.

Converting an integer stride into a Mersenne number can be done in the same way as for the starting address. That is, just additions are needed. This process can be done at the time when the vector stride is loaded into the vector stride register. However, a special register of c bits is needed to keep the converted stride as shown in Figure 1.

It should be noted that the number of physical lines in the cache is still 2^c though there are only $2^c - 1$ logical lines. The reason is that there are 2 different 0's in Mersenne number $(0, 0 \cdots, 0)$ and $(1, 1, \cdots, 1)$ that correspond to 2 different memory addresses. Whenever the index field becomes all 1's, we need to reset the index field. This can be easily implemented as shown in Figure 1.

The additional hardware cost as result of this new mapping scheme includes 2 multiplexors, a full adder and a few registers as shown in Figure 1. Even a small percentage of performance improvement can justify such insignificant hardware cost. As will be evidenced latter in this paper, we expect to double and even triple the performance of direct-mapped cache and the performance of the machines without cache. However, there is still a trade-off between performance and hardware cost. The registers storing the starting addresses of vectors, for example, add more cost to the machine. If several vectors are accessed concurrently, we not only need more registers but also the control logic to control the access of these registers. If we can afford to spend 1 or 2 more cycles at each vector start-up time, we may eliminate these registers. Each time a vector is accessed again, we recalculate its starting address even though the vector is already in the cache. This recalculation requires just additions of subwords, which takes small amount of time and may also be pipelined. On the other hand, if performance is absolutely important one can pay for the registers.

2.4 Cache Size Limitations

Readers may have already noticed that using Mersenne number to implement cache mapping poses cache size limitations. The values of c that make Mersenne number a prime are 2, 3, 5, 7, 13, 17, and 19 etc.. There is a gap between each 2 subsequent Mersenne primes. For example, the next larger cache size beyond $8K$ sets is $128K$ sets. Such size gaps can be avoided in actual designs in a number of ways. Let us say that the cache size is measured in terms of memory words. Then changing the line size from 1 word to 16 words results in consecutive cache sizes from $8K$ (2^{13}) to $128K$ (2^{17}) words. On the other hand, if an optimal line size is chosen, one can also change the degree of associativity to have different cache sizes. From a direct-mapped cache to 8-way set-associative, one can fill up the size gap between $8K$ and $128K$ lines. Therefore, there are 3 design dimensions: number of sets (modulus), line size, and associativity. A computer designer has enough space to exercise his/her design trade-offs.

3 An Analytical Model

In this section, we present an analytical performance model based on the simple vector performance model given in [2]. It should be noted that the analysis and the numerical results presented here are not meant to be predictors of performance of any realistic computer system. Rather, they are meant to give a quick insight into the simplified vector processor models and to show some potential problems that might exist in a vector cache memory. By varying different input parameters over a wide range, we use the analytical model to comparatively study the performance of different alternatives of cache designs.

3.1 A Vector-Computation Model

Two simplified vector processor models considered here are shown in Figures 2 and 3 which will be referred to as MM-model and CC-model, respectively. Both models have a vector processing unit, a set of vector registers with maximum vector length of V_L words, and a set of low order-bit interleaved memory modules (or banks) that are connected to the vector processor through three pipelined buses. Two of the three buses are read buses and the other is write bus. Each of these buses contains separate data bus and address bus. There are totally $M = 2^m$ interleaved memory banks each of which has the access time of t_m cycles. The second processor model (CC-model) shown in Figure 3 differs from the first one (MM-model) only in that it has a cache memory of size C sets between the processor and memories. We assume that this cache memory is used purely for vector data similar to the split vector cache of VAX 6000 [11]. A line of data can be transferred on one of the buses in one cycle.

Cache miss ratio has been used by many researchers as a performance measure to evaluate cache performance. However, it is not a very good performance measure in this context because we intend to compare the performance differences between the MM-model and the CC-model. The performance measures that we will use in this paper are the *total execution time* of an application program and the *speedup* resulting from adding cache memories into the MM-model vector computer.

The *speedup* is defined as the ratio of the execution time on the MM-model to the execution time on the CC-model. In order to estimate the total execution time, we present in the following a simple and generic vector computational model that attempts to cover a wide spectrum of numerical algorithms.

Our simplified computational model assumes that application programs are blocked into several segments. One of the vector data that an application program operates on is partitioned into several sub-blocks of size B each. We call this B a *blocking factor*. For example, if a matrix is blocked into several submatrices of size $b \times b$, then the blocking factor is b^2. A sub-block of data may be used several times by one program segment. We define a reuse factor, R, as the number of times a block of data is reused. During each vector operation, the processor may load two vectors from the memory simultaneously or load just one vector with the other operand available in a register. It is assumed that the probability that the processor accesses two vector streams simultaneously from the memory during one vector operation is P_{ds} (double stream). The probability that the processor accesses just a single vector stream with the other operand available in a register is P_{ss} $(= 1 - P_{ds})$. The system performs operations similar to the SAXPY (Single-precision $A * X$ Plus Y:$Y = a * X + Y$) operations. The processor first loads one or two vectors into its registers and then performs arithmetic operations on the two vectors or on one vector and a scalar in a scalar register.

We now determine the length of the second vector for double stream vector accesses. In order to simplify our analysis, we further assume that all single-stream vector accesses are evenly intercepted by one-double-stream accesses if $P_{ds} \neq 0$. In other words, every sequence of $\frac{BP_{ss}}{BP_{ds}} = \frac{P_{ss}}{P_{ds}}$ consecutive single-stream vector accesses is followed by one double stream vector access which is again followed by a statistically identical sequence of single stream accesses, and so forth. This assumption is backed by many realistic numerical algorithms [7, 20, 21]. Thus, one can imagine one of the vectors of length B as a two dimensional matrix with $P_{ss}/P_{ds}+1$ columns and BP_{ds} rows. The vector length of each column is BP_{ds}. After every P_{ss}/P_{ds} column accesses of the imagined matrix, the processor loads two vectors to perform vector operations on the first and the second vector. Therefore, the length of the second vector can be considered as BP_{ds}.

Although the assumed computation model is simplified, it is very close to many realistic applications. For example, the blocked matrix multiply algorithm in [7] has the blocking factor of b^2 since the matrices are blocked into submatrices of size $b \times b$. The reuse factor of each block is b and each sequence of $b-1$ single stream vector accesses is followed by a double stream access. Therefore, the fraction of time when the processor accesses a single vector stream is $(b-1)/b$ while the fraction of double stream accesses is $1/b$. The length of the first vector is then b^2 and the length of the second vector is b. Similarly, the blocked LU decomposition algorithm with a blocking factor of b^2 [20] has an average reuse factor of $3b/2$ and the blocked FFT algorithm [21] with a blocking factor of b has a reuse factor of $Log_2 b$.

The access strides of loading the two vectors are denoted by s_1 and s_2, respectively. Access strides vary widely depending on application programs. In general, unit stride occurs more frequently than other strides in most numerical algorithms [8]. We therefore use $P_{stride1}$ to denote the probability that an access stride is 1. If a stride is not 1, we assume that it takes any other integer values

equally likely. Since we are interested in memory bank conflicts and cache line conflicts, a stride is assumed to take an integer value in the range of $(1, 2, \cdots, M)$ for the MM-model and in the range of $(1, 2, \cdots, C)$ for the CC-model due to modular operations. We also assume that writing results into memory or cache will not delay the normal vector operations. This assumption is not a severe restriction because it can be approached in real machines by having write buffers, separate data bus for writing and separate write port for memories [2].

In summary, our simple vector computational model consists of the following seven-tuple:

$$VCM = [B, R, P_{ds}, s_1, s_2, P_{stride1}(s_1), P_{stride1}(s_2)],$$

where

B:	blocking factor or the size of each sub-vector;
R:	reuse factor, the number of times a sub-vector is reused;
P_{ds}:	the probability that a processor accesses two vector streams;
s_i:	stride of accessing ith vector;
$P_{stride1}(s_i)$:	the probability that $s_i = 1$.

By properly selecting these model parameters, the model can fit into a variety of numerical algorithms and various vector access patterns. For example, if we set $VCM = [b, r, 1, 1, P, 1, 1/C]$, we have double stream vector accesses to columns and rows of a $b \times b$ submatrix of a $P \times Q$ matrix stored in column-major. Each pair of column and row are used for r times. If $VCM = [b, r, 0, P + 1, *, 1/C, *]$ [1], then we get a single vector access stream to a major diagonal of a $b \times b$ submatrix of the same $P \times Q$ matrix. FFT access and sub-block access can also be easily modeled with some minor modifications as will be discussed later in the paper.

3.2 Execution Time of the MM-Model

Let us consider the MM-model of Figure 2 first. The vector performance can be characterized by the execution time for a sequence of operations on a vector of length B, T_B. This quantity depends on a number of factors including the overhead for computing the starting addresses and setting up vector controls, the overhead for executing scalar code for strip-mining (fine level of blocking), and the maximum vector register length, V_L. By assuming these parameters to be constant denoted by T_o, we have

$$T_B = T_o + \lceil \frac{B}{V_L} \rceil \cdot T^M_{start} + B \cdot T^M_{elemt}, \tag{1}$$

where T^M_{start} is the start-up time for each inner most loop and T^M_{elemt} [2] is the time for processing one element of the vector ignoring the start-up time. The start-up time depends on the memory access time and the type of arithmetic operations to be performed. It should not be difficult to determine its value for a given set of system parameters. In our discussions, we assume T_{start} to be constant for a given memory access time. If the system is fully pipelined with one functional unit, T^M_{elemt} should be one in an ideal case. In reality, T^M_{elemt} is often greater than one because of stalls resulting

[1] The symbol "*" means undefined

[2] For notational convenience, we use superscript M and C on a parameter to distinguish MM-model (Figure 2) and CC-model (Figure 3) vector processors, respectively.

from memory accesses, data dependencies and control dependencies. Since our primary interest here is memory system performance, we will concentrate on memory stalls only. The major source of memory stalls for vector processors is memory conflicts [22, 23, 9]. A number of storage schemes to minimize memory conflicts stands out in the literature. Some examples are [24, 22, 19, 18] to list a few. In our analysis, we consider only the low-order bit interleaving scheme although some conflict-free dynamic storage schemes can provide about 18% better performance than the simple interleaving [22]. In the following, we derive T^M_{elemt} by taking into account memory stalls.

Memory stalls can result from two types of bank conflicts: One is the conflicts between elements in the same vector (*self-interference*) denoted by I^M_s, and the other occurs between elements from two different vector access streams (*cross-interference*). Let us first consider the number of possible stall cycles of accessing one vector stream with stride s_1. When a vector access stream traverses across all the memory banks (one sweep), the number of memory banks visited by the access stream, the return number (N_r), is given by $M/gcd(M, s_1)$ [23], where $gcd(M, s_1)$ is the greatest common divisor of M and s_1. If $V_L > M/gcd(M, s_1)$, memory conflicts occur within the vector access stream. For a special case where $gcd(M, s_1) = 2^m = M$, all element requests go to a single memory bank. Since M is a power of 2, $gcd(M, s_1)$ is in the form of 2^i for some $i = 0, \cdots m-1$. For each 2^i, there may be a number of values of s_1 within M such that $gcd(M, s_1) = 2^i$. In order to find the distribution of return number N_r which determines average memory stalls, we need to find out the number of integers within M that share the same $gcd(M, s_1) = 2^i$. This number is clearly the same as the number of values of s_1 such that $gcd(\frac{M}{2^i}, \frac{s_1}{2^i}) = 1$, which is Euler's function [12] given by

$$\phi(\frac{M}{2^i}) = \frac{M}{2^{i+1}} = 2^{m-i-1}, \quad for\ i < m.$$

For each such s_1, the return number, N_r, is 2^{m-i}. Recall that the probability of $s_1 = 1$ is $P_{stride1}$ and the probability of s_1 taking any other value in $(2, \cdots, M)$ is $\frac{1-P_{stride1}}{M-1}$. Thus, the probability distribution of N_r is given by

$$Pr\{N_r = 2^i\} = \begin{cases} \frac{1-P_{stride1}}{M-1}, & i = 0, \\ (1 - P_{stride1})\frac{2^{i-1}}{M-1}, & i = 1,\ 2,\ \cdots v-1, \\ 1 - \sum_{N_r=1}^{V_L/2} Pr\{N_r\} = 1 - \frac{V_L(1-P_{stride1})}{2(M-1)}, & i = v. \end{cases} \tag{2}$$

Now consider a vector access stream. If the memory cycle time is greater than the return number, i.e. $t_m > M/gcd(M, s_1)$, memory stalls occur. Each one of the V_L/N_r sweeps are delayed by $t_m - N_r$ cycles. If $gcd(M, s_1) = 2^m = M$, then each of V_L elements will be delayed by $t_m - 1$ cycles. Therefore the total number of stall cycles resulting from accessing a vector of length V_L, I^M_s, is given by

$$I^M_s = \frac{1 - P_{stride1}}{M-1}[\sum_{i=\lceil log_2 \frac{M}{t_m} \rceil}^{m-1} (t_m - 2^{m-i})2^{m-i-1}\frac{V_L}{2^{m-i}} + V_L(t_m - 1)].$$

The lower limit of the summation in the above equation is determined based on the condition $t_m \geq M/2^i$. Notice that the above equation assumes that $t_m < M$ so that unit stride does not incur

any stalls. Simplifying the above expression we have

$$I_s^M = V_L \frac{1 - P_{stride1}}{(M-1)} \cdot [t_m + \frac{t_m}{2} \lfloor log_2 t_m \rfloor - 2^{\lfloor log_2 t_m \rfloor}]. \tag{3}$$

Next, let us consider memory stalls as result of cross-interferences between two independent vector access streams. Let the stride of accessing the second vector stream be s_2 which has the same distribution as s_1. Let the memory bank difference between the starting addresses of the two vectors be uniformly distributed between 1 and M. Conditioning on the return numbers of the two access streams, N_{r1} and N_{r2}, we have the probability that the two vector access streams overlap

$$P_{ov} = Pr\{overlap | N_{r1} = n_1, N_{r2} = n_2\} = \begin{cases} \frac{n_1+n_2-1}{M}, & if\ M \geq n_1 + n_2; \\ 1, & otherwise. \end{cases} \tag{4}$$

If two vector access streams do not overlap, then the number of memory stalls will depend on the vector access stream that has smaller return number. Therefore, the number of stalls (NS_{nov}) in this case given that $N_{r1} = n_1$ and $N_{r2} = n_2$ is given by

$$NS_{nov} = \delta(t_m - min(n_1, n_2)) \cdot \frac{V_L}{min(n_1, n_2)},$$

where $\delta(x) = x$ if $x \geq 0$ and $\delta(x) = 0$ otherwise. If the two vector access streams overlap, the overlapped elements of the second vector can not start memory access until after $\frac{V_L}{n_1} \cdot t_m - 1$ cycles which are the time needed to finish all memory accesses of the first vector in that memory bank. The reason of lessing 1 here is that the load instruction of the second vector starts memory access 1 cycle later than the first vector access. After we start memory accesses for the overlapped elements of the second vector, it will again take $t_m \cdot \frac{V_L}{n_2}$ cycles to finish up all memory accesses in the memory bank. Therefore, the number of stalls (NS_{ov}) in case of overlapping given that $N_{r1} = n_1$ and $N_{r2} = n_2$ is given by

$$NS_{ov} = \frac{V_L}{n_1} \cdot t_m - 1 + t_m \cdot \frac{V_L}{n_2}.$$

The expected number of memory stalls conditioning on the two return numbers is therefore given by

$$E[stalls | N_{r1} = n_1, N_{r2} = n_2] = P_{ov} \cdot NS_{ov} + (1 - P_{ov}) \cdot NS_{nov}.$$

Removing the conditions, we have the average number of memory stalls resulting from a double stream vector access. It is given by

$$E[stalls] = \sum_{n_1=1}^{V_L} \sum_{n_2=1}^{V_L} E[stalls | N_{r1} = n_1 N_{r2} = n_2] Pr\{N_{r1} = n_1\} Pr\{N_{r2} = n_2\}, \tag{5}$$

where the distribution of N_{r1} and N_{r2} is given by Equation (2).

T_{elemt}^M can now be expressed as

$$\begin{aligned} T_{elemt}^M &= 1 + average\ stall\ cycles\ per\ element \\ &= 1 + \frac{E[stalls] \cdot P_{ds} + I_s^M \cdot P_{ss}}{V_L}. \end{aligned} \tag{6}$$

Substituting Equation (6) into Equation (1) we get the execution time of one block. Assume that the total data size of a program is N which is blocked into several segments of length B. Taking into account the reuse factor, R, the total execution time T_N^M can be expressed as:

$$T_N^M = T_B \cdot R \cdot \lceil \frac{N}{B} \rceil,$$

where T_B is given by Equation (1).

3.3 Execution Time of Direct-Mapped CC-Model

Now let us consider the CC-model of Figure 3 that has a direct-mapped cache of size $C = 2^c$ lines. We assume in this section that a cache line contains one double precision word of data.

The processor initially loads one or two vectors into the cache while performing the designated vector operations. The process of the initial loading will take the amount of time given by Equation (1). After the vectors are loaded into the cache, the remaining operations on the same set of data are expected to be performed in the cache without accessing memories in an ideal situation. The total execution time of the CC-model (Figure 3) is given by

$$T_N^C = \{T_B + [T_o + \lceil \frac{B}{V_L} \rceil \cdot T_{start}^C + B \cdot T_{elemt}^C] \cdot (R-1)\} \cdot \lceil \frac{N}{B} \rceil, \tag{7}$$

where T_B covers the effects of compulsory and capacity misses as discussed above. Similar to the analysis of the MM-model, we use T_{elemt}^C to include memory stalls due to cache misses. We will consider in the following the effects of cache misses caused by line interferences on the system performance.

Suppose that the processor tries to load a vector of length B into the vector cache with a stride s_1 chosen from a random number having the distribution described in the beginning of this section. The number of cache lines occupied by the vector is given by $C/gcd(C, s_1)$. Thus, there will be $B - C/gcd(C, s_1)$ self-interferences if the vector length B is greater than or equal to $C/gcd(C, s_1)$. Again the number of s_1's such that $gcd(C, s_1) = 2^{(c-i)}$ is 2^{i-1} (divisor function) [12]. If $gcd(C, s_1) = C$, then there would be $B - 1$ conflicts. Assuming lock-up free cache, the average number of stalls due to self-interference misses in the B-element vector is given by

$$I_s^C(B) = \frac{1 - P_{stride1}}{C-1} [\sum_{i=1}^{c - \lceil log_2 \frac{C}{B} \rceil} (B - \frac{C}{2^{c-i}}) 2^{i-1} + B - 1] \cdot t_m,$$

where the upper limit of the summation is determined in such a way that $B - C/2^i$ is always positive. Notice that there is no self-interference if the access stride is one and the block size B is less than cache size. For every self-interference miss the processor stalls for t_m cycles. Simplifying the above expression we have

$$I_s^C(B) = \frac{1 - P_{stride1}}{C-1} \frac{1}{3} (3B2^{\lfloor log_2 B \rfloor} - 2 \cdot 2^{2\lfloor log_2 B \rfloor} - 1) \cdot t_m. \tag{8}$$

For B being a power of 2, we have

$$I_s^C(B) = \frac{1 - P_{stride1}}{3(C-1)}(B^2 - 1) \cdot t_m.$$

The cross interference misses can be calculated in a similar way as for the MM-model assuming that the two vectors are independent and their starting addresses differ by a random number. The probability distribution of the return number, the number of cache lines occupied when loading a vector, is obtained using Equation (2) by substituting M and v with C and $logB$ respectively. Similarly, the probability that the two vectors overlap is given by Equation (4) by replacing M by C. If the two vectors do not overlap in the cache, the expected number of stalls conditioning on the two return numbers n_1 and n_2 is given by

$$NS_{nov} = t_m(B - n_1) + t_m(B - n_2).$$

If the two vectors overlap in the cache, then the average overlap length is $\frac{n_1 n_2}{n_1+n_2-1}$. Therefore, the number of stalls resulting from cache misses given that the two vectors overlapped is given by

$$NS_{ov} = t_m(B - n_1) + t_m(B - n_2) + t_m\frac{n_1 n_2}{n_1 + n_2 - 1}.$$

The total stalls, $E^C[stalls]$, are calculated using Equation (5) with the upper limits for the summations being B and BP_{ds} respectively. Thus, the time for processing one vector element, T^C_{elemt}, is

$$T^C_{elemt} = 1 + P_{ss}\frac{I_s^C(B)}{B} + P_{ds}\frac{E^C[stalls]}{B}. \tag{9}$$

Substituting Equation (9) into Equation (7) we obtain the total execution time of the direct-mapped CC-model.

3.4 Execution time of Prime-Mapped Cache

Since the modulus in the prime-mapped cache is a prime number, interference misses are minimum. For the random stride access pattern, the self-interference misses occur only when the stride is an integral multiple of the cache size. Therefore, I_s^C is simply given by

$$I_s^C(B) = \frac{(1 - P_{stride1})(B-1)}{C-1}t_m, \tag{10}$$

assuming that the stride is uniformly distributed between 2 and C if it is not 1. Cross-interferences are computed in the same way as for the direct-mapped cache. Substituting Equation (10) into Equation (9) we obtain the time for processing one vector element for prime-mapped cache. The total execution time is calculated using Equation (7).

3.5 Validation of the Analytical Model

In order to verify the correctness of our analytical model developed in the last subsection, we have written a simulator. The simulator is synchronous (Time-driven Simulation) in the sense that all the system activities occur at discrete time intervals which are processor cycles. The simulator simulates a vector processor, a set of interleaved memory banks, and three data buses that connect the processor and memories. The vector processor in the simulator can be in one of two states: *active* doing useful computations or *idle* waiting for a memory access. At the beginning of each simulation cycle, we check if the processor is in active or idle state. If it is in active state, a vector access request is generated and the processor is made idle. Whether the request is for a single vector or two vectors is determined by P_{ss}. The starting memory bank number of the vector access(es) is selected from the M memory banks equally likely. At the mean time the access stride is also determined based on the random number following the assumed distribution. Once the starting address and the stride are determined, V_L element requests are then appended to the corresponding memory bank queues. Every two consecutive element requests of the vector are separated by the chosen constant stride. Before the cycle ends, we update all the memory queues, decrease the remaining memory time of the element request at the head of a queue, and return the request to the processor if the remaining time is 0. If the processor has collected all element requests it had initiated, it resumes active for the following cycle. The statistics that we collected from the simulator is T_{elemt} which is 1 plus the average number of stalls per vector element. We run the simulator for 1,000,000 cycles in each simulation run. All the simulation results reported in this paper are the average of 10 such runs.

Table 1. Comparison between analysis and simulation in terms of T_{elemt}. (V_L=64, $P_{stride1} = 0.25$, $P_{ss} = 1.0$)

tm	M	=64	M	=128	M	=256
	Sim	Ana	Sim	Ana	Sim	Ana
5	1.69	1.69	1.66	1.66	1.64	1.64
10	1.89	1.91	1.80	1.80	1.76	1.75
15	2.10	2.13	1.92	1.95	1.89	1.86
20	2.35	2.38	2.12	2.11	1.94	1.99
25	2.59	2.64	2.25	2.29	2.21	2.11
30	2.86	2.89	2.46	2.45	2.24	2.23
35	3.10	3.17	2.54	2.63	2.30	2.36
40	3.39	3.46	2.66	2.81	2.46	2.49

Table 2. Comparison between analysis and simulation in terms of T_{elemt}. (V_L=64, $P_{stride1} = 0.25$, $P_{ss} = 0.0$)

tm	M	=64	M	=128	M	=256
	Sim	Ana	Sim	Ana	Sim	Ana
5	2.36	2.28	2.23	2.17	2.01	1.91
10	2.74	2.77	2.50	2.47	2.23	2.09
15	3.19	3.26	2.77	2.77	2.45	2.29
20	3.76	3.76	3.11	3.08	2.69	2.49
25	4.06	4.25	3.45	3.40	2.86	2.68
30	4.66	4.74	3.73	3.72	3.09	2.88
35	5.18	5.25	4.07	4.04	3.30	3.08
40	5.80	5.74	4.38	4.36	3.50	3.30

We first verify our analysis by varying the memory cycle time, t_m. Table 1 shows the comparison between the simulation results and the analytical results with $P_{ss} = 1$ and $P_{stride1} = 0.25$. It is shown in the table that the two results are in an excellent agreement for all considered values of t_m. When we decrease (P_{ss}) from 1 to 0 implying that all vector accesses are double streams, similar agreement is also observed as shown in Table 2. In all cases, the maximum difference between the simulation and the analysis is about 6% indicating a high accuracy of our analytical model.

We have also varied the proportion of single stream vector accesses, P_{ss}, and the probability of unit stride, $P_{stride1}$, to compare the two results. Table 3 shows the numerical results for different values of P_{ss} and Table 4 shows the numerical results for different values of $P_{stride1}$. Similar agreements are also observed from these two tables.

Table 3. Comparison between analysis and simulation in terms of T_{elemt}. (V_L=64, $P_{stride1} = 0.25$, $tm = 20$)

P_{ss}	M	=64	M	=128
	Sim	Ana	Sim	Ana
0.0	3.69	3.76	3.12	3.09
0.2	3.46	3.48	2.99	2.89
0.4	3.24	3.21	2.72	2.70
0.6	2.98	2.93	2.56	2.51
0.8	2.74	2.66	2.37	2.31
1.0	2.38	2.38	2.12	2.12

Table 4. Comparison between analysis and simulation in terms of T_{elemt}. (V_L=64, $P_{ss} = 0.9$, $tm = 20$)

$P_{stride1}$	M	=64	M	=128
	Sim	Ana	Sim	Ana
0.0	2.80	2.71	2.32	2.31
0.2	2.69	2.59	2.26	2.25
0.4	2.55	2.46	2.18	2.18
0.6	2.32	2.33	2.11	2.12
0.8	2.09	2.21	2.02	2.06
1.0	1.89	2.08	1.89	1.99

4 Performance Analysis and Comparison

We now carry out performance analysis and comparison of different architectural alternatives based on the simple analytical model presented above. The degree of cache associativity is considered to be 1 since we are only interested in relative performance of the new cache design and the conventional cache design. One may increase the associativity of both designs to study the cache performance, which is not considered here to control the length of the paper. In the following figures, we fix the values of V_L, T_o, T^M_{start} and T^C_{start} at 64, 20, $15 + t_m$ and 15 [2], respectively. Unless otherwise specified, the probability of unit stride access, $P_{stride1}$, is fixed at 0.25 which is the average value of the experimental data reported in [8].

Random Multistride Accesses

Figure 4 shows the *speedup* as a function of memory access time in terms of processor cycles assuming random stride having the distribution given previously. The *speedups* are obtained by dividing the total execution time of the MM-model by the execution times of the corresponding CC-models. Two curves are drawn corresponding to the direct-mapped CC-model and the prime-mapped CC-model, respectively. If memory access time is small, it can be seen from the figure that adding a direct-mapped cache memory does not help. The MM-model performs even better than the CC-model that has a direct-mapped cache memory. This is primarily due to the fact that any one of

irregularly distributed cache misses results in a memory access that takes t_m cycles. The interleaved memory with pipelining may well satisfy the bandwidth requirement of the processor. However, as the speed gap between processor and memory increases, i.e. the number of cycles needed for accessing memory increases, the performance of the direct-mapped CC-model takes over. Still, the speedup of the direct-mapped CC-model is limited. The prime-mapped cache, on the other hand, shows significant speedup which is getting even larger as memory access time increases. The prime-mapped CC-model outperforms both the MM-model and the direct-mapped CC-model over entire range of memory times considered. We notice that the prime-mapped CC-model runs up to three times faster than both the direct-mapped CC-model and the MM-model.

It was noted in [7] that different blocking factors show different performances. To observe this effects, we plotted the performance vs blocking factors with other parameters being fixed at the values shown in Figure 5. As shown in the figure, speedup resulting from adding the direct-mapped cache quickly drops down below 1 after the blocking factor is about 1K, while the speedup of the prime-mapped cache remains flat. This figure indicates that improper choice of blocking factor can make the performance of the direct-mapped CC-model worse than that of the MM-model, whereas the prime-mapped cache is not sensitive to the blocking factor. This is another advantage of the prime-mapped cache.

In the above two figures, we have fixed the probability of accessing a vector with stride 1, $P_{stride1}$, at 0.25. To examine the effect of this parameter on system performance, we plotted speedups by varying $P_{stride1}$ as shown in Figure 6. High probability of unit stride will reduce both memory conflicts and cache line conflicts. As a result, the performance difference between the two mapping schemes comes close as this probability increases. When the probability of having unit stride is 1, the performance of the two mapping schemes is the same. Whenever this probability is nonunit, the prime-mapped cache performs better than the direct-mapped cache.

Figure 7 shows the effect of proportion of double access streams (P_{ds}) on the performance of the three models. When the proportion of double access streams is large, cross-interference between the two vectors in the main memory or in the cache becomes large. Notice that the cross-interference in the prime-mapped cache is more severe than that in direct-mapped cache because the vector footprint in the former is larger than the vector footprint in the latter. Notice also that cross-interference in the main memory is more severe than cross-interference in the cache for small footprints, giving rise to increased speedup of direct-mapped cache as P_{ds} increases. For prime-mapped cache, on the other hand, cross-interference is relatively larger because of the large sizes of the two vectors that can be placed in the cache. As a result, the speedup of the prime-mapped cache decreases with the increase of the proportion of double stream accesses. Nevertheless, the prime-mapped cache still performs better than the direct-mapped cache over all possible fractions of double stream accesses. This is mainly because of the large effect of self-interferences.

It should be pointed out here that cross interferences in the prime-mapped cache can be completely eliminated by increasing the cache associativity. A 2-way set associative cache with a prime number of sets can hold two vector streams without conflicts; a 4-way set associative cache can hold 4 vector streams and so forth. The same is not true for conventional set-associative caches due to a large amount of self-interferences.

Sub-block Accesses

Sub-block accesses have proven to be important in many vector processing applications. Blocked matrix multiply is an example that needs to access a submatrix of a large matrix. Let the original large matrix be dimensioned $P \times Q$, and the sub-block be dimensioned $b_1 \times b_2$. A sub-block access can be characterized as b_2 stride-1 accesses to column vectors of length b_1. The distance between the starting addresses of two successive column vectors is P.

Sub-block accesses can be easily made conflict free for any arbitrary two dimensional matrix by properly selecting blocking factors. Notice that this is either impossible or prohibitively costly for the MM-model since the modulus is a power of 2 [22]. Consider the above matrix with dimensions $P \times Q$ that is stored in a column-major. To make accesses to a sub-block of dimension $b_1 \times b_2$ conflict free in the prime-mapped cache, what we need to do is to select b_1 and b_2 in such a way that the following conditions are satisfied.

$$b_1 \leq min(P \ mode \ C, \ C - P \ mod \ C), \ and$$
$$b_2 \leq \lfloor C/min(P \ mode \ C, \ C - P \ mod \ C) \rfloor,$$

where C is cache size. A row-major storage can also be realized similarly by interchanging the corresponding dimensions. It can be easily shown that a sub-block with b_1 and b_2 satisfying the above conditions can be stored in the cache without line interference. This is because that b_1 elements of any column vector are stored in consecutive memory locations and hence no conflict can occur among them. Moreover, since $(P \ mode \ C) \geq b_1$ and $(C - P \ mod \ C) \geq b_1$, the first elements of any two column vectors are mapped into two cache locations that are at least b_1 lines apart as long as $b_2 \leq \lfloor C/min(P \ mode \ C, \ C - P \ mod \ C) \rfloor$. The fraction of the cache being used is $b_1 b_2 / C$. If we let $b_1 = min(P \ mode \ C, \ C - P \ mod \ C)$, and $b_2 = \lfloor \frac{C}{b_1} \rfloor$. then the cache utilization is close to 1. In other words, conflict free access is possible to the submatrix even with the cache utilization approaching 1.

Suppose that a subblock of $b_1 \times b_2$ that is stored in column-major is reused for R times in a program. We can use the analysis presented in the previous section to calculate the execution time by assuming that each subcolumn of a subblock can be contained in a cache set. Figure 8 shows the *speedup* as a function of memory access time in terms of processor cycles. With the blocking factor of 1K, the direct-mapped CC-model just manages to have some speedup. The speedup is about 1.6 for memory access time being 32 cycles. If the blocking factor is 2K, i.e. the fraction of cache being used is about a quarter, we can hardly see any speedup. The prime-mapped cache, however, shows significant speedup as long as the blocking factor is less than the cache size. It is clear from the figure that adding the newly proposed cache memory triples the vector performance after the memory time exceeds 28 cycles.

In Figure 9, we fix blocking factor b_2 which is the number of columns in each submatrix at 1K while changing blocking factor b_1 which is the length of each column in a submatrix and also the number of words in a set. Note that the blocking factor is changed in terms of $log_2 b_1$ in the figure rather than b_1. Since matrix Y is stored in column-major, elements along any column are in consecutive memory locations. As a result, changing subcolumn size does not change number of line conflicts as long as b_2 keeps the same, which makes the two curves remain relative flat. However,

when we fix b_1 while changing b_2, the performance of the direct-mapped CC-model drops drastically after b_2 exceeds 64 ($log_2 b_2 = 6$) as shown in Figure 10. The initial increase in speedup of direct-mapped CC-model is due to the increase of vector length (within V_L) of each vector operation. In both cases (Figures 9 and 10), the prime-mapped cache organization performs significantly better than the direct-mapped CC-model. Moreover, the speedup remains unchanged for all possible blocking factors because it is conflict-free as long as the subblock is smaller than cache size.

To observe how the reuse factor affect the performance of vector caches we plotted *speedup* as a function of reuse factor R while fixing the blocking factors as shown in Figure 11. As expected, the performance of the two cache organizations is the same when the reuse factor is 1. Recall that the initial loading of the cache is pipelined and is done concurrently with the vector operation. The performance difference between the two mapping schemes is getting larger as the reuse factor increases due to the fact that the prime-mapped CC-model has a much higher chance of finding in cache the reused data than the direct-mapped counter part does. Moreover, it can be seen in this figure that after the reuse factor exceeds certain value, the performance change is not significant.

FFT Accesses

The FFT accesses result from the Cooley and Tukey's Fast Fourier Transform (FFT) algorithm. The FFT is a basic tool in various scientific and engineering disciplines, ranging from oil exploration to artificial intelligence. Efficient FFT computation is desirable for any engineering/scientific computers. While the FFT algorithm performs well on scalar computers, it does not perform as well on vector computers because the access stride changes after each stage of computation. Moreover, other than the final stage all strides are powers of 2, which results in a lot of line conflicts in a direct-mapped cache. The FFT in its original form can only keep a very small amount of data in the cache if the data size that has to be a power of 2 is greater than the cache size. A tremendous amount of efforts has been made during the past 20 years to optimize the performance of FFT on a vector machine [21].

One way to implement the FFT algorithm in a memory hierarchical system is to map the input data into a two dimensional array [21]. Suppose an N-point data array to be transformed can be represented as $N = B_2 \times B_1$. Then the input data can be considered to be a matrix of B_1 columns and B_2 rows. The matrix is stored in a column-major. The algorithm proceeds by first doing B_1-point row FFTs for B_2 times each of which is expected to be done within the local cache. After all B_2 row FFTs are done, we multiply twiddle factors and perform column FFTs for B_1 times. Again it is expected that each of the column FFTs is done inside the cache. If B_2 is less than the cache size, each of the column FFTs can be done in the cache without misses except for the compulsory misses since the array is stored in a column-major. However, the row FFTs of the first step may not be guaranteed to be done inside the cache without misses. Depending on the value of B_2, conflict misses may occur. The number of interference misses is obviously given by $B_1 - C/gcd(B_2, C)$ for $B_1 > C/gcd(B_2, C)$.

The total execution time of this algorithm on the direct-mapped cache can be computed by using Equation (7) twice. First, we substitute B and R in (7) by B_1 and $log_2 B_1$, respectively. T^C_{elemt} should include the above self interferences for a given B_2. Notice that $P_{ds} = 0$ if we assume

that twiddle factors are available in the registers. Second, Equation (7) is used again with B and R being replaced by B_2 and $log_2 B_2$, respectively. T^C_{elemt} can be considered to be 1 assuming that $B_2 < C$. The final total execution time is the sum of the above two times. The execution time of prime-mapped cache can be done similarly by noting that there are no self-interference misses unless $B_2 = C$. Compulsory misses (T_B) should also be adjusted based on the FFT stride characteristics.

Figure 12 shows the performance comparison of the FFT algorithm on the two different cache organizations. The curves are the average clock cycles per point which is the total execution time divided by N. In the figure, we fix one dimension while varying the other. It is observed from these curves that the prime-mapped cache outperforms the direct-mapped cache by a factor of more than 2. The improvement is valid over all possible values of the blocking factor B_2. With the prime-mapped cache, the FFT algorithm can be implemented very easily. Optimization is guaranteed as long as the block size is less than the cache size.

5 Conclusions

In this paper, we have proposed an innovative cache mapping scheme for vector computers, called prime-mapped cache. The new cache organization minimizes cache misses caused by cache line interferences that have been shown to be critical in numerical applications. The cache lookup time of the new mapping scheme keeps the same as conventional caches. Generation of cache addresses for accessing the prime-mapped cache can be done in parallel with normal address calculations. This address generation takes shorter time than the normal address calculation due to the special properties of the Mersenne prime. Therefore, the new mapping scheme does not result in any performance penalty as far as the cache access time is concerned. The hardware cost introduced by the prime mapping scheme includes 2 multiplexors, a full adder and a few registers.

We have developed an analytical performance model for a generic vector computational model that covers a wide spectrum of numerical algorithms. The analytical model is simple and precise for a given set of vector access patterns. Simulation experiments have been carried out to validate our analysis. It has been shown that our analysis is in a good agreement with simulations. Three typical vector access patterns have been considered in our performance analysis: random multi-stride, submatrix and FFT accesses. Numerical results have shown that the prime-mapped cache outperforms both the vector computer without cache and the vector computer with a conventional cache for all vector access patterns considered. The performance improvement ranges from 40% to a factor of 3 depending on the memory cycle time, blocking factor and access patterns.

Our conclusion is that cache memory can improve the performance of vector processing provided that application programs can be blocked. With the new mapping scheme, the cache memory can provide significant performance improvement which will become larger as the speed gap between processor and memory increases.

Acknowledgement

The author is grateful to Mr. J. Huang for his help in simulations. The paper has also benefited from several discussions with Dr. Jim Cooley of URI and Dr. Dave Harper of UT Dallas. The

author would like to thank the anonymous referees for their valuable comments that greatly helped improving the quality of the paper. Special thanks are due to Ms. Liping W. Yang for her continued support as well as her helps in numerical computations.

References

[1] A. J. Smith, "Cache memories," *Computing Surveys*, vol. 14, pp. 472–530, Sept. 1982.

[2] J. L. Hennessy and D. A. Patterson, *Computer Architecture, A Quantitative Approach.* Morgan Kaufmann, 1990.

[3] H. S. Stone, *High Performance Computer Architecture.* Addison-Wesley, 1990.

[4] K. So and V. Zecca, "Cache performance of vector processors," in *Proc. 15th. Int'l Symp. on Comp. Arch.*, pp. 261–268, 1988.

[5] D. Gannon, W. Jalby, and K. Gallivan, "Strategies for cache and local memory management by global program transformation," in *Int'l Conf. on Supercomputing*, 1987.

[6] J. Dongarra and et al., "A set of level 3 basic linear algebra subprograms," *ACM Trans. on Math. Soft.*, vol. 16-1, pp. 1–17, March 1990.

[7] M. S. Lam, E. E. Rothberg, and M. E. Wolf, "The cache performance and optimizations of blocked algorithms," in *Proc. of Arch. Supp. for Prog. Lang. and Opr. Sys.*, pp. 63–74, April 1991.

[8] J. W. C. Fu and J. H. Patel, "Data prefetching in multiprocessor vector cache memories," in *Proc. 18th. Int'l Symp. on Comp. Arch.*, pp. 54–63, 1991.

[9] D. H. Bailey, "Vector computer memory bank contention," *IEEE Trans. on Computers*, vol. C-36, pp. 293–298, MARCH 1987.

[10] W. Abu-Sufah and A. D. Malony, "Vector processing on the Alliant FX/8 multiprocessor," in *Int. Conf. on Parallel Processing*, pp. 559–566, Aug. 1986.

[11] D. Bhandarkar and R. Brunner, "VAX vector architecture," in *Proc. 17th. Int'l Symp. on Comp. Arch.*, pp. 204–215, 1990.

[12] A. J. Pettofrezzo and D. R. Byrkit, *Elements of Number Theory.* Prentice-Hall, 1970.

[13] A. Agarwal, R. Simoni, J. Hennessy, and M. Horowitz, "An evaluation of directory schemes for cache coherences," in *The 15th Ann. Int. Symp. on Comp. Arch.*, pp. 280–289, June 1988.

[14] J. Archibald and J.-L. Baer, "Cache coherence protocols: Evaluation using multiprocessor simulation model," *ACM Tran. on Comput. Systems*, vol. 4, pp. 273–298, November 1986.

[15] Q. Yang, L. Bhuyan, and B.-C. Liu, "Analysis and comparison of cache coherence protocols for a packet-switched multiprocessor," *IEEE Trans. on Comput.*, vol. 38, pp. 1143–1153, August 1989. Special Issue on Distributed Computer Systems.

[16] M. Dubois and F. Briggs, "Effect of cache coherency in multiprocessors," *IEEE Tran. on Comput.*, vol. C-31, pp. 1083–1099, Nov. 1982.

[17] M. D. Hill, "A case for direct-mapped caches," *IEEE Computer*, pp. 25–40, Dec. 1988.

[18] P. Budnik and D. J. Kuck, "Organization and use of parallel memories," *IEEE Trans. on Computers*, pp. 1566–1569, Dec. 1971.

[19] D. H. Lawrie and C. R. Vora, "The prime memory system for array access," *IEEE Trans. on Computers*, vol. C-31, pp. 435–441, May 1982.

[20] J. Armstrong, "Algorithm and performance notes for blocked LU factorization," in *Int. Conf. on Parallel Processing*, pp. III–161–164, Aug. 1988.

[21] J. W. Cooley, *The Structure of FFT and Convolution Algorithms*. IBM T. J. Watson Research Center, 1990. Research Report.

[22] D. T. HarperIII, "Block, multistride vector, and FFT accesses in parallel memory systems," *IEEE Trans. on Parallel and Distributed Systems*, Jan. 1991.

[23] W. Oed and O. Lange, "On the effective bandwidth of interleaved memories in vector processor systems," *IEEE Trans. on Computers*, vol. C-34, pp. 949–957, Oct. 1985.

[24] R. Raghavan and J. P. Hayes, "On randomly interleaved memories," in *Proceedings of Supercomputing-90*, pp. 49–58, Nov. 1990.

[25] Q. Yang and L. W. Yang, "A novel cache design for vector processing," *the 19th Int'l Symp. on Computer Architecture*, May 1992. Gold Coast, Australia.