Journal of Algorithms **31,** 66–104 (1999)
Article ID jagm.1998.0985, available online at http://www.idealibrary.com on IDEAL®

The Influence of Caches on the Performance of Sorting[*]

Anthony LaMarca[†]

Xerox PARC, 3333 Coyote Hill Road, Palo Alto, California 94304

and

Richard E. Ladner[‡]

Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195

Received March 23, 1998; revised August 15, 1998

We investigate the effect that caches have on the performance of sorting algorithms both experimentally and analytically. To address the performance problems that high cache miss penalties introduce we restructure mergesort, quicksort, and heapsort in order to improve their cache locality. For all three algorithms the improvement in cache performance leads to a reduction in total execution time. We also investigate the performance of radix sort. Despite the extremely low instruction count incurred by this linear time sorting algorithm, its relatively poor cache performance results in worse overall performance than the efficient comparison based sorting algorithms. For each algorithm we provide an analysis that closely predicts the number of cache misses incurred by the algorithm.
© 1999 Academic Press

1. INTRODUCTION

Since the introduction of caches, main memory has continued to grow slower relative to processor cycle times. The time to service a cache miss to memory has grown from 6 cycles for the Vax 11/780 to 120 for the

[*]A preliminary version of this paper appears in Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, 1997, pp. 370–379.
[†]E-mail: lamarca@parc.xerox.com.
[‡]E-mail: ladner@cs.washington.edu.

AlphaServer 8400 [16, 21]. Cache miss penalties have grown to the point where good overall performance cannot be achieved without good cache performance. As a consequence of this change in computer architectures, algorithms that have been designed to minimize instruction count may not achieve the performance of algorithms that take into account both instruction count and cache performance.

One of the most common tasks computers perform is sorting a set of unordered keys. Sorting is a fundamental task and hundreds of sorting algorithms have been developed. In this paper we explore the potential performance gains that cache-conscious design offers in understanding and improving the performance of four popular sorting algorithms: mergesort,[1] quicksort [30], heapsort [56], and radix sort.[2] Mergesort, quicksort, and heapsort are all comparison based sorting algorithms while radix sort is not.

For each of the four sorting algorithms we choose an implementation variant with potential for good overall performance and then we heavily optimize this variant using traditional techniques to minimize the number of instructions executed. These heavily optimized algorithms form the baseline for comparison. For each of the comparison sort baseline algorithms we develop and apply memory optimizations in order to improve cache performance and, hopefully, overall performance. For radix sort we optimize cache performance by varying the radix.

In the process we develop some simple analytic techniques that enable us to predict the memory performance of these algorithms in terms of cache misses. Cache misses cannot be analyzed precisely due to a number of factors such as variations in process scheduling and the operating system's virtual to physical page-mapping policy. In addition, the memory behavior of an algorithm may be too complex to analyze completely. For these reasons the analyses we present are only approximate and must be validated empirically.

For comparison purposes we focus on sorting an array (4000 to 4,096,000 keys) of 64 bit integers chosen uniformly at random. Our main study uses trace-driven simulations and actual executions to measure the impact that our memory optimizations have on performance. We concentrate on three performance measures: instruction count, cache misses, and overall performance (time) on machines with modern memory systems. Our results can be summarized as follows:

1. For the three comparison based sorting algorithms, memory optimizations improve both cache and overall performance. The improvements

[1]Knuth [37] traces mergesort back to card sorting machines of the 1930s.

[2]Knuth [37] traces the radix sorting method to the Hollerith sorting machine that was first used to assist the 1890 U.S. census.

in overall performance for heapsort and mergesort are significant, while the improvement for quicksort is modest. Interestingly, memory optimizations to heapsort also reduce its instruction count. For radix sort the radix that minimizes cache misses also minimizes instruction count.

2. For large arrays, radix sort has the lowest instruction count, but because of its relatively poor cache performance, its overall performance is worse than the memory optimized versions of mergesort and quicksort.

3. Although our study was done on one particular architecture, we demonstrate the robustness of the results by showing that comparable speedups due to improved cache performance can be achieved on several other machines.

4. There are effective approximate analytic approaches to predicting the number of cache misses these sorting algorithms incur. In many cases the analysis is not difficult, yet is highly predictive of actual performance.

The main general lesson to be learned from this study is that because cache miss penalties are large, and growing larger with each new generation of processor, selecting the fastest algorithm to solve a problem entails understanding cache performance. Improving an algorithm's overall performance may require increasing the number of instructions executed while, at the same time, reducing the number of cache misses. Consequently, cache-conscious design of algorithms is required to achieve the best performance.

2. RELATED WORK

Algorithms in the context of a memory hierarchy have been studied for a long time. External sorting algorithms were studied in the context of main memory and magnetic tape or disk storage [37, 53−55]. Aggarwal and Vitter designed a particularly nice model for studying external memory algorithms and applied the model to external sorting [4]. Several of their algorithmic ideas in the external memory model transfer well to the cache-main memory model. In particular, they employ multimerging and multipartitioning to produce excellent external memory sorting algorithms. We also employ these techniques in the cache-main memory setting.

Although the similarities are strong, there are a several differences between the cache-main memory model and the external memory model that give them different characters. One important difference is that in the external memory model the algorithm has control over what data should be in main memory and what should be on disk. With hardware caches, however, when a cache miss occurs the algorithm has no choice as to where that new cache block is loaded or what block is evicted. Another difference is that in the external memory model computation is free and only disk I/Os are counted. This makes perfect sense because a disk I/O is many orders of magnitude more expensive than a computation step. In the cache-main memory model cache misses are not nearly as costly. Hence, care must be taken not to expend too many instructions when attempting to reduce the number of cache misses. Finally, there is a difference in scale. The caches closest to main memory are generally quite large relative to the size of memory. For example, it is common to see a 1 MB cache with a memory of 32 MB, making the ratio of memory to cache 32 : 1. In the external memory model the capacity of the disk is considered to be very large in comparison to the size of memory. Although the external memory model and cache-main memory model are different in character, algorithmic ideas such as multimerging and multipartitioning, are effective in both settings. Other algorithmic techniques from the external memory algorithms literature may also be effective in improving the cache performance and overall performance of memory resident algorithms [6–8, 14, 18, 38, 52].

In addition to the external memory model of Aggarwal and Vitter [4] there have been several other models of hierarchical memory that have been proposed [2, 3, 5]. In all these models the algorithm has control over where in the hierarchy the data is to reside.

Other related work includes research in compiler optimizations to improve cache performance [9, 13, 26, 35, 57, 58, 15], systems research to improve the cache performance while a program is running [10, 19, 28, 36, 49], trace modeling to better predict cache misses [1, 47, 50, 51], and theoretical research on optimal placement of data for direct mapped caches [24, 25, 32].

3. CACHES

In order to speed up memory accesses, small high speed memories called *caches* are placed between the processor and the main memory. Accessing the cache is typically much faster than accessing main memory. Unfortunately, because caches are smaller than main memory they can hold only a subset of its contents. Memory accesses first consult the cache to see if it contains the desired data. If the data is found in the cache, the main memory need not be consulted and the access is considered to be a cache *hit*. If the data is not in the cache it is considered a *miss*, and the data must be loaded from main memory. On a miss, the block containing the accessed data is loaded into the cache in the hope that data in the same block will be accessed again in the future. The *hit ratio* is a measure of cache performance and is the total number of hits divided by the total number of accesses.

The major design parameters of caches are:

- *Capacity*, which is the total number of bytes that the cache can hold.
- *Block size*, which is the number of bytes that are loaded from and written to memory at a time.
- *Associativity*, which indicates the number of different locations in the cache where a particular block can be loaded. In an *N-way set-associative* cache, a particular block can be loaded in N different cache locations. *Direct-mapped* caches have an associativity of 1, and can load a particular block only in a single location. *Fully associative* caches are at the other extreme and can load blocks anywhere in the cache.
- *Replacement policy*, which indicates the policy of which block to remove from the cache when a new block is loaded. For the direct-mapped cache the replacement policy is simply to remove the block currently residing in the cache.

In most modern machines, more than one cache is placed between the processor and the main memory. These hierarchies of caches are configured with the smallest, fastest cache next to the processor and the largest, slowest cache next to main memory. The largest miss penalty is typically incurred with the cache closest to main memory and this cache is often direct-mapped. Consequently, our design and analysis techniques will focus on improving the performance of direct-mapped caches. We assume that the cache parameters, block size, and capacity, are known to the programmer.

High cache hit ratios depend on a program's stream of memory references exhibiting locality. A program exhibits *temporal locality* if there is a good chance that an accessed data item will be accessed again in the near future. A program exhibits *spatial locality* if there is good chance that subsequently accessed data items are located near each other in memory. Most programs tend to exhibit both kinds of locality and typical hit ratios are greater than 90% [41]. With a 90% hit ratio, cutting the number of cache misses in half has the effect of raising the hit ratio to 95%. This may not seem like a big improvement, but with miss penalties on the order of 100 cycles, normal programs will exhibit speedups approaching 2 : 1 in execution time. Accordingly, our design techniques will attempt to improve both the temporal and spatial locality of the sorting algorithms.

Cache misses are often categorized into *compulsory*, *capacity*, and *conflict* misses [29]. Compulsory misses are those that occur when a block is first accessed and brought into the cache. Capacity misses are those caused by the fact that more blocks are accessed than can fit all at one time in the cache. Conflict misses are those that occur because two or more blocks that map to the same location in the cache are accessed. In this paper we address techniques to reduce the number of both capacity and conflict misses for sorting algorithms.

4. DESIGN AND EVALUATION METHODOLOGY

Cache locality is a good thing. When spatial and temporal locality can be improved at no cost it should always be done. In this paper, however, we propose techniques for improving locality even when it results in an increase in the total number of executed instructions. This represents a significant departure from traditional algorithm design and optimization methodology. We take this approach in order to show how large an impact cache performance can have on overall performance. Interestingly, many of the design techniques are not particularly new. Some have already been used in external memory algorithms, optimizing compilers, and in parallel algorithms. Similar techniques have also been used successfully in the development of the cache-efficient Alphasort algorithm [43].

4.1. *Performance Measures*

As mentioned earlier we focus on three measures of performance: instruction count, cache misses, and overall performance in terms of execution time. We now explain how each of these quantities were measured.

The dynamic instruction counts for our algorithms were measured using ATOM [48], a toolkit developed by DEC for instrumenting program executables on Alpha workstations. Using ATOM, we instrumented each algorithm to increment a counter after executing every instruction. In this way, we could run the instrumented algorithm and we could obtain an exact count of the number of instructions executed in a run.

To measure cache performance, we wrote a small library to emulate the behavior of a direct mapped cache. The emulated cache has a routine which is passed an address and returns whether the access was a hit or a miss. The emulated cache uses the stream of accesses to both maintain the state of the cache as well as to keep long term statistics on cache performance. The algorithms we wanted to measure were then instrumented with ATOM to make accesses to our emulated cache on every read and write. In effect, we are measuring cache performance with a trace-driven simulation. The only slight irregularity is that we are actually consuming the trace as it is being generated. This, however, in no way effects the accuracy of the measurement. Rather, this approach yields a cache simulation which perfectly measures the performance of a virtually indexed cache in the absence of context switches. We did not try to model the effects of context switches on cache performance, as they have very little impact when running with today's large quanta on lightly loaded machines. In all cases we configured the emulated cache's block size and capacity to be the same as the second level cache of the DEC Alphastation 250 (32 byte blocks with a $2{,}097{,}152 = 2^{21}$ byte capacity).

The final performance metric we measured was execution time, and these were captured using the hardware cycle timer on the DEC Alphastation 250. By simply reading the cycle counter before and after the execution of an uninstrumented algorithm and subtracting, an extremely accurate measure of elapsed cycles is obtained. Dividing this count by cycles per second (around 280,000,000) yields the execution time in seconds. Throughout this study, execution times represent the median of 15 trials.

4.2. *Analysis*

Finally, we provide analytic methods to predict cache performance in terms of cache misses. In some cases the analysis is quite simple. For example, traditional mergesort has a fairly oblivious pattern of access to memory, thereby making its analysis quite straightforward. However, the memory access patterns of the other algorithms, such as heapsort, are less oblivious requiring more sophisticated techniques and approximations to accomplish the analysis.

We choose a very simple model for analyzing cache performance. We assume there is a large memory that is divided up into *blocks* and a smaller cache that is divided up into C blocks. There are n keys and B is the number of keys that fit in a cache block. The keys are stored in a contiguous array of size n/B memory blocks.

In a direct-mapped cache each block of memory maps to exactly one block in the cache. We assume the simple mapping where memory block x is mapped to cache block $x \bmod C$. At any moment of time each block y in the cache is associated with exactly one block of memory x such that $y = x \bmod C$. In this case we say that block x of memory is in the cache. An access to memory block x is a *hit* if x is in the cache and is a *miss*, otherwise. As a consequence of a miss the accessed block x is brought into the cache and the previous block residing at cache location $x \bmod C$ is evicted.

We will model an algorithm simply as a sequence of accesses to blocks in memory. We assume that initially, none of the blocks to be accessed are in the cache. We do not distinguish between reads and writes because we assume a *copy back* architecture with a *write buffer* for the cache [29]. In the copy back architecture writes to the cache are not immediately passed to the memory. A write to cache block $y = x \bmod C$ is written to memory block x when a miss occurs, that is, when block z is accessed where $y = z \bmod C$ and $z \neq x$. The write buffer allows the writes to location x to almost always propagate to memory asynchronously.

This simple model does not exactly model the accesses a program might make in a real implementation. For example, the arrays declared in a program might not be allocated to contiguous segments of memory. In addition, the analysis ignores accesses to control variables and other small data structures because almost all of these accesses are hits and they cause few conflicts with accesses to sorting data. Finally, in some of our analyses we make simplifying assumptions in order to make the analysis tractable. Nonetheless, the trace-driven cache simulation results suggest that this simple model of a direct mapped cache yields analyses that accurately predict the number of cache misses incurred in the trace-driven cache simulations done with ATOM.

As a simple case for our analysis let us consider the number of cache misses incurred by a traversal of an array of keys like those employed in iterative mergesort, quicksort, and radix sort. If there are B keys per cache block, then in every B accesses to the array there is one cache miss. Hence, the number of cache misses per key is $1/B$.

4.3. *Data*

We evaluate our analyses on the DEC Alphastation 250 sorting eight byte keys. In our execution trials $C = 2^{16}$, $B = 4$ and n ranged from 4000 to 4,096,000 keys (4,096,000 was the largest set size we could use before disk paging became a factor). We chose our data uniformly at random. We believe that the choice of randomly chosen data yields a fair comparison among the algorithms. None of the algorithms that we study has a clear advantage or disadvantage over the others because of the choice of data set. It might seem that quicksort has an advantage because the worst case $\Omega(n^2)$ time can be avoided by using random data. However, quicksort can easily avoid worst case behavior by either choosing the pivot as a median of 3 or by choosing the pivot randomly.

5. MERGESORT

Two sorted lists can be merged into a single sorted list by traversing the two lists at the same time in sorted order, repeatedly adding the smaller key to the single sorted list. By treating a set of unordered keys as a set of sorted lists of length 1, the keys can be repeatedly merged together until a single sorted set of keys remains. Algorithms which sort in this manner are known as mergesort algorithms, and there are both recursive and iterative variants [31, 37].

5.1. *Base Mergesort Algorithm*

For a base algorithm, we chose an iterative mergesort [37] because it is both easy to implement and is very amenable to traditional optimization techniques. The algorithm uses two arrays, an input array and an auxiliary array. The standard iterative mergesort makes $\lceil \log_2 n \rceil$ passes over the input array, where the ith pass merges sorted subarrays of length 2^{i-1} into sorted subarrays of length 2^i on the auxiliary array. An important optimization that we employ is to alternate the merging process from one array to the other to avoid copying. If the number of passes is odd, then there is one final copy back from the auxiliary array to the input array. We make a number of other optimizations that reduce the number of instructions executed by the algorithm. These optimizations include: keeping the subarrays to be merged in opposite order to avoid unnecessary checking for end conditions, sorting subarrays of size 4 with a fast in-line sorting method, and loop unrolling. Thus, the number of merge passes is $\lceil \log_2(n/4) \rceil$. If $\lceil \log_2(n/4) \rceil$ is odd then an additional copy pass is needed to move the sorted array to the input array. Our base mergesort algorithm has a very low instruction count, executing the fewest instructions of any of our comparison based sorting algorithms.

5.2. *Memory Optimizations for Mergesort*

While the base mergesort executes few instructions, it has the potential for terrible cache performance. The base mergesort uses each data item only once per pass, and if the input array exceeds the capacity of the cache, keys are ejected before they are used again. If the set of keys is of size $BC/2$ or smaller, the entire sort can be performed in the cache and only compulsory misses are incurred. When the set size is larger than $BC/2$, however, temporal reuse drops off sharply and when the set size is larger than the cache, no temporal reuse occurs at all. To improve this inefficiency, we apply two memory optimizations to the base mergesort. Applying the first of these optimizations yields *tiled mergesort*, and applying both yields *multimergesort*.

Tiled mergesort employs an idea, called *tiling*, that is also used in some optimizing compilers [58]. Tiled mergesort has two phases. To improve temporal locality in the first phase, subarrays of length $BC/2$ are sorted using mergesort. The second phase returns to the base mergesort to complete the sorting of the entire array. In order to avoid the final copy if $\lceil \log_2(n/4) \rceil$ is odd, subarrays of size 2 are sorted in-line instead of size 4.

Tiling the base mergesort drastically reduces the misses it incurs, and the added loop overhead increases instruction counts very little.

While tiling improves the cache performance of the first phase, the second phase still suffers from the same problem as the base mergesort. Each pass through the source array in the second phase needs to fault in all of the blocks, and no reuse is achieved across passes if the set size is larger than the cache. To fix this inefficiency in the second phase, we employ a multiway merge similar to those used in external sorting (Knuth devotes a section of his book to techniques for multimerging [37, Section 5.4.1]). In multimergesort we replace the final $\lceil \log_2(n/(BC/2)) \rceil$ merge passes of tiled mergesort with a single pass that merges all of the pieces together at once. This single pass makes use of a memory-optimized heap to hold the heads of the lists being multimerged [39]. In order to avoid unnecessary cache misses when the heads of lists map to the same cache block, we load an entire block of keys into the heap from any list that requires an additional key in the heap. Thus, the heap has maximum size kB where k is the number of lists to be merged. We do not need the more sophisticated external mergesort algorithm employed by Aggarwal and Vitter [4] because kB is very small compared to the size of the cache. In order to avoid any unnecessary copying we make sure that in the first phase the result is found in the auxiliary array. To do this we in-line sort subarrays of size 2 instead of 4 if $\lceil \log_2(BC/8) \rceil$ is odd. Thus, the final multimerge phase leaves the sorted array in the input array as desired. The multimerge introduces several complications to the algorithm and significantly increases the dynamic instruction count. However, the resulting algorithm has excellent cache performance, incurring roughly a constant number of cache misses per key in our executions.

5.3. *Performance of Mergesort Algorithms*

As mentioned earlier, the performance of the three mergesort algorithms is measured by sorting sets of uniformly distributed 64 bit integers. Figure 1 shows the number of instructions executed per key, the cache missed incurred and the execution times for each of the mergesort variants. The instruction count graph shows that as expected, the base mergesort and the tiled mergesort execute almost the same number of instructions. The wobble in the instruction count curve for the base mergesort is due to the final copy that may need to take place depending on whether the final merge wrote into the source array or the auxiliary array [37]. When the set size is smaller than the cache, the multimergesort executes the same number of instructions as the tiled mergesort. Beyond that size, the multimerge is performed and this graph shows the increase it causes in

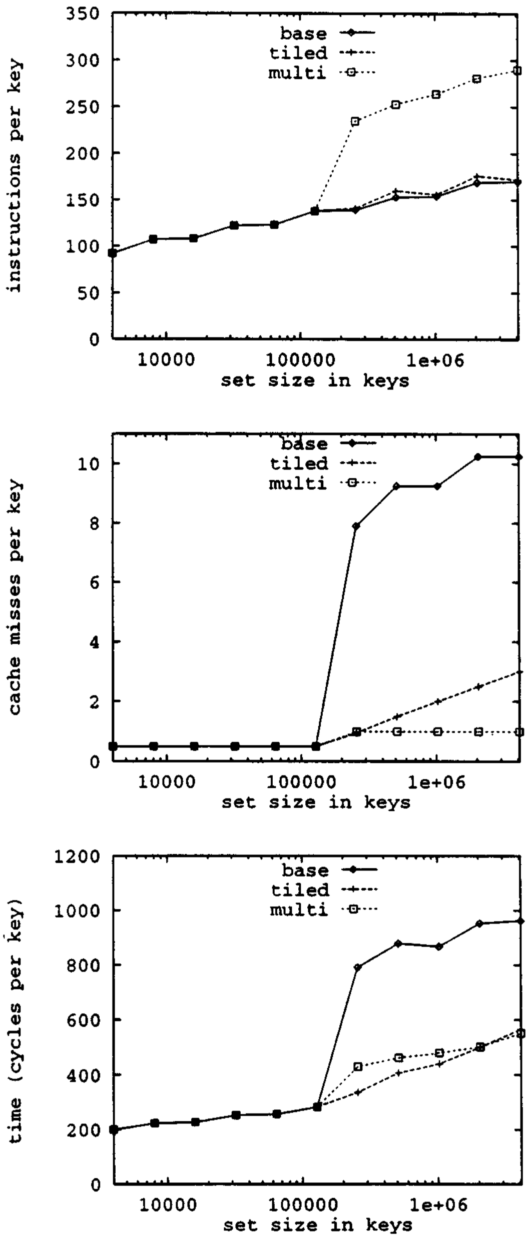

FIG. 1. Performance of mergesort on sets of 4000 to 4,096,000 keys. From top to bottom, the graphs show instruction counts per key, cache misses per key, and the execution times per key. Executions were run on a DEC Alphastation 250 and simulated cache capacity is 2 megabytes with a 32 byte block size.

the instruction count. For 4,096,000 keys, the multimerge executes 70% more instructions than the other mergesorts.

The most striking feature of the cache performance graph is the sudden increase in cache misses for the base mergesort when the set size grows larger than the cache. This graph shows the large impact that tiling the base mergesort has on cache misses; for 4,096,000 keys, the tiled mergesort incurs 66% fewer cache misses that the base mergesort. This graph also shows that the multimergesort is a clear success from a cache miss perspective, incurring slightly more than 1 miss per key regardless of the input size.

The graph of execution times shows that up to the size of the second level cache, all of these algorithms perform the same. Beyond that size, the base mergesort performs the worst due to the large number of cache misses it incurs. The tiled mergesort executes up to 55% faster than the base mergesort, showing the significant impact the cache misses in the first phase have on execution time. When the multiway merge is first performed, the multimergesort performs worse than the tiled mergesort due to the increase in instruction count. Due to lower cache misses, however, the multimergesort scales better and performs as well as the tiled mergesort for the largest set sizes. Because mergesort has a relatively oblivious pattern of memory accesses the speedups should be also achieved for arbitrary data sets.

5.4. *Cache Analysis of Mergesort Algorithms*

It is fairly easy to approximate the number of cache misses incurred by our mergesort variants as the memory reference patterns of these algorithms are fairly oblivious. We begin with our base algorithm. For $n \leq BC/2$ the number of misses per key is simply $2/B$, the compulsory misses. Because the other two algorithms employ the base algorithm for $n \leq BC/2$ then they all have the same number of misses per key in this range.

For $n > BC/2$, the number of misses per key in the iterative mergesort algorithm is approximately,

$$\frac{2}{B}\left\lceil \log_2 \frac{n}{4} \right\rceil + \frac{1}{B} + \frac{2}{B}\left(\left\lceil \log_2 \frac{n}{4} \right\rceil \bmod 2\right). \qquad (1)$$

The first term of expression (1) comes from the capacity misses incurred during the $\lceil \log_2 n/4 \rceil$ merge passes. In each pass, each key is moved from a source array to a destination array. Every Bth key visited in the source array results in one cache miss and every Bth key written into the destination array results in one cache miss. Thus, there are $2/B$ cache misses per key per pass. The second term, $1/B$, is the compulsory misses

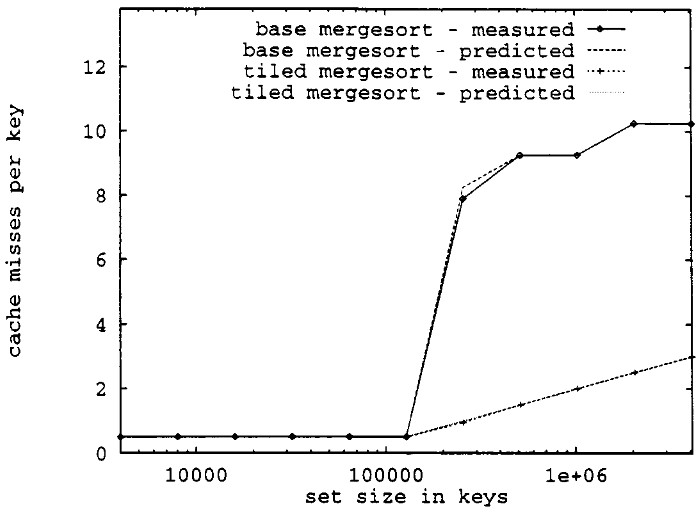

FIG. 2. Cache misses incurred by mergesort, measured versus predicted.

per key incurred in the initial pass of sorting into groups of four. The final term is the number of misses per key caused by the final copy, if there is one.

For $n > BC/2$ the number of misses per key in tiled mergesort is approximately,

$$\frac{2}{B}\left\lceil \log_2 \frac{2n}{BC} \right\rceil + \frac{2}{B}. \tag{2}$$

The first term of expression (2) is the number of misses per key for the final $\lceil \log_2 n/(BC/2) \rceil$ merge passes. The second term is the number of misses per key in sorting into $BC/2$ size pieces. The number of passes is forced to be even so there are no additional cache misses caused by a copy. Figure 2 shows how well the analysis of base mergesort and tiled mergesort predict their actual performance.

Finally, for $n > BC/2$, the number of misses per key in multimergesort is approximately,

$$\frac{4}{B}. \tag{3}$$

For $B = 4$, this closely matches the approximately 1 miss per key shown in cache misses graph of Fig. 1. The first phase of multimergesort is tiled mergesort which incurs $2/B$ misses per key. In the algorithm we make sure that the number of passes in the first phase is odd so that the second phase multimerges the auxiliary array into the input array. If we ignore the cache misses caused by accesses to the heap, the second phase multimerges the auxiliary array back into the input array causing another $2/B$ misses per key. The multimerge employs a heap array containing no more than $m = B\lceil 2n/BC \rceil$ keys. For practical purposes m is very small compared to the size of the cache. Thus, accesses to the heap are for the most part hits. Periodically but seldom the traversals of the input and auxiliary arrays capture the part of the cache where the heap maps. Thus, periodically but seldom the accesses to the heap are misses instead of hits. Thus, the number of cache misses per key in Fig. 1 for multimergesort is slightly more than 1.

6. QUICKSORT

Quicksort is an in-place divide-and-conquer sorting algorithm considered by most to be the fastest comparison-based sorting algorithm when the set of keys fit in memory [30]. In quicksort, a key from the set is chosen as the *pivot*, and all other keys in the set are partitioned around this pivot. This is usually accomplished by walking through an array of keys from the outside in, swapping keys on the left that are greater than the pivot with keys on the right that are less than the pivot. At the end of the pass, the set of keys is partitioned around the pivot and the pivot is guaranteed to be in its final position. The quicksort algorithm then recurses on the region to the left of the pivot and the region to the right. The simple recursive quicksort is a simple, elegant algorithm and can be expressed in less than 20 lines of code.

6.1. *Base Quicksort Algorithm*

An excellent study of fast implementations of quicksort was conducted by Sedgewick, and we use the optimized quicksort he develops as our base algorithm [45]. There are three main optimizations that he recommends. First, we use a stack instead of using recursion. Second, we use the median of 3 to select the pivot. Third, as Sedgewick suggests, our base algorithm does not use quicksort to sort small subproblems. Instead, it leaves all the subproblems below a certain threshold in size unsorted. Then in a final pass the array is sorted using an optimized insertion sort. We employ all the optimizations recommended by Sedgewick in our base quicksort.

6.2. *Memory Optimizations for Quicksort*

In practice, quicksort generally exhibits excellent cache performance. Because the algorithm makes sequential passes through the source array, all keys in a block are always used and spatial locality is excellent. Quicksort's divide-and-conquer structure also gives it excellent temporal locality. If a subset to be sorted is small enough to fit in the cache, quicksort incurs at most one cache miss per block before the subset is fully sorted. Despite this, improvements can be made and we develop two memory optimized versions of quicksort, *memory-tuned quicksort* and *multiquicksort*.

Our memory-tuned quicksort simply removes Sedgewick's elegant insertion sort at the end, and instead sorts small subarrays when they are first encountered using an unoptimized insertion sort. When a small subarray is encountered, it has just been part of a partitioning and this is an ideal time to sort it, because all of its keys should be in the cache. Although saving small subarrays until the end makes sense from an instruction count perspective, it is exactly the wrong thing to do from a cache performance perspective.

Multiquicksort employs a second memory optimization in similar spirit to that used in multimergesort. Although quicksort incurs only one cache miss per block when the set is cache-sized or smaller, larger sets incur a substantial number of misses. To fix this inefficiency, a single multipartition pass can be used to divide the full set into a number of subsets which are likely to be cache sized or smaller.

Multipartitioning is used in parallel sorting algorithms to divide a set into subsets for multiple processors [11, 33, 44] in order to quickly balance the load. In addition, multipartitioning has been used for external sorting to divide a set into subsets each of which fits in main memory [4]. We choose the number of pivots so that the number of subsets larger than the cache is small with sufficiently high probability. It is known that if k points are placed randomly in a range of length 1, the chance of a resulting subrange being of size x or greater is exactly $(1-x)^k$ [20, Vol. 2, p. 22]. In multiquicksort we partition the input array into $3n/(BC)$ pieces, requiring $(3n/(BC)) - 1$ pivots. Hence after the multipartition, the chance that a subset is larger than BC is $(1 - BC/n)^{(3n/(BC))-1}$. In the limit as n grows large, the percentage of subsets that are larger than the cache is e^{-3}, less than 5%.

Multiquicksort requires a number of auxiliary data structures. Unlike binary partitioning, k-way partitioning cannot be performed efficiently in-place. Instead a temporary list is allocated for each of the k subsets. It would be a very inefficient use of storage to allocate an array to each of these lists because we do not know their sizes ahead of time. Instead we use a linked list of blocks of keys for each list. Varying the number of keys per block from 100 and 5000 has little effect on performance, so that choosing blocks of 100 keys minimizes wasted storage without sacrificing performance. The k randomly chosen pivots are stored in an array. As each element of the source array is read, a binary search is performed to determine to which list the element belongs. Once all the keys in the source array are partitioned, each list is copied back to the source array and sorted using the base quicksort algorithm.

6.3. *Performance of Quicksort Algorithms*

Figure 3 shows the performance of the three quicksort algorithms sorting 64 bit uniformly distributed integers. The instruction count graph shows that the base quicksort executes the fewest instructions with the memory-tuned quicksort executing a constant number of additional instructions per key. This difference is due to the inefficiency of sorting the small subsets individually rather than at the end as suggested by Sedgewick. For large set sizes the multiquicksort performs the multipartition which results in a significant increase in the number of instructions executed.

The cache performance graph shows that all of the quicksort algorithms incur very few cache misses. The base quicksort incurs fewer than two misses per key for 4,096,000 keys, lower than all of the other algorithms up to this point with the exception of the multimergesort. The cache miss curve for the memory-tuned quicksort shows that removing the instruction count optimization in the base quicksort improves cache performance by approximately 0.25 cache misses per key. The cache miss graph also shows that the multiway partition produces a flat cache miss curve much the same as the curve for the multimergesort. The maximum number of misses incurred per key for the multiquicksort is slightly larger than one miss per key, validating the conjecture that it is uncommon for the multipartition to produce subsets larger than the size of the cache.

The graph of execution times shows the execution times of the three quicksort algorithms. All three of these algorithms perform similarly on our DEC Alphastation 250. This graph shows that sorting small subsets early is a benefit, and the reduction in cache misses outweighs the increase in instruction cost. The multipartition initially hurts the performance of the multiquicksort due to the increase in instruction cost, but the low number of cache misses makes it more competitive as the set size is increased. This graph suggests that if more memory were available and

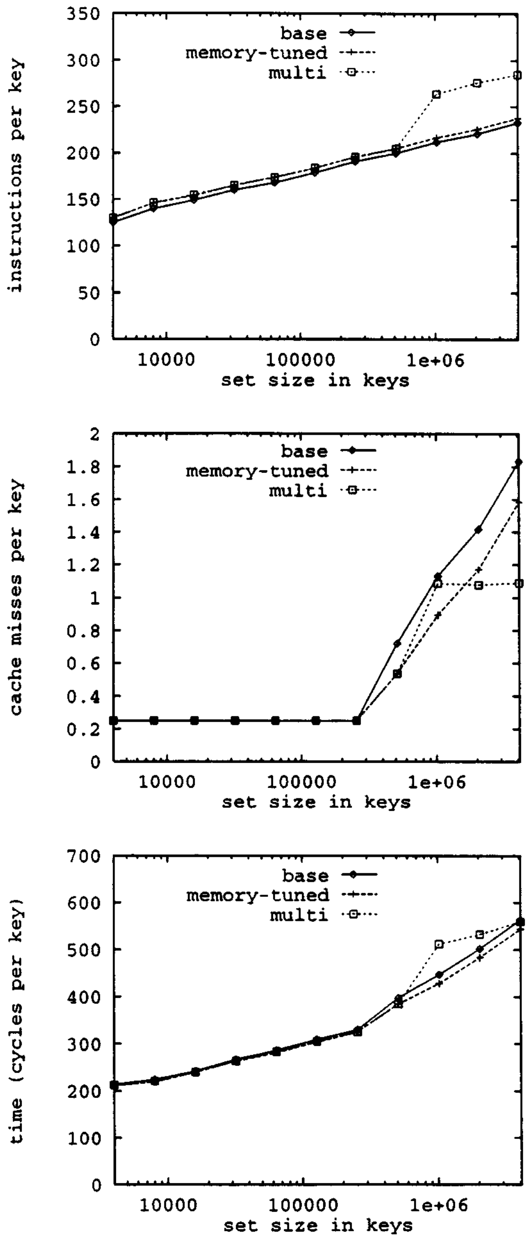

FIG. 3. Performance of quicksort on sets of 4000 to 4,096,000 keys. From top to bottom, the graphs show instruction counts per key, cache misses per key, and the execution times per key. Executions were run on a DEC Alphastation 250 and simulated cache capacity is 2 megabytes with a 32 byte block size.

larger sets were sorted, the multiquicksort would outperform both the base quicksort and the memory-tuned quicksort. If the pivot(s) in the quicksort algorithms are chosen so that the subproblems are very unequally balanced then any of the variants will have poor performance because both the number of instructions and the number of cache misses will be excessive. Otherwise, the performance of the quicksort variants should not vary much on arbitrary data from their performances on uniformly chosen data.

6.4. *Cache Analysis of Quicksort Algorithms*

Rather than using a complicated analysis that takes into account the median of three pivot choosing strategy, we take a simpler approach of assuming the pivot to be of any rank with equal probability. Because very few pivots are chosen before a subproblem is smaller than the cache the difference in cache performance between choosing the median of 3 or just choosing 1 as the pivot is only slight.

We begin by analyzing memory-tuned quicksort which has slightly simpler memory behavior than base quicksort. If $n \leq BC$ then memory-tuned quicksort incurs $1/B$ misses per key for the compulsory misses. Because base quicksort makes an extra pass at the end to perform the insertion sort, the number of cache misses per key incurred by base quicksort is $1/B$ plus the number of cache misses per key incurred by memory-tuned quicksort. The cache miss graph of Fig. 3 clearly shows the additional $1/B = 0.25$ misses per key for base quicksort in our simulation.

For $n > BC$ the expected number of misses per key in memory-tuned quicksort is approximately,

$$\frac{2}{B}\ln\left(\frac{n}{BC}\right) + \frac{5}{8B} + \frac{3C}{8n}. \tag{4}$$

We analyze the algorithm in two parts. In the first part we assume that partitioning an array of size m costs m/B misses if $m > BC$ and 0 misses otherwise. In the second part we correct this by estimating undercounted and overcounted cache misses. Let $M(n)$ be the expected number of misses incurred in quicksort under our assumption of the first part of the analysis. We then have the recurrence,

$$M(n) = \frac{n}{B} + \frac{1}{n}\sum_{i=0}^{n-1}\left(M(i) + M(n-i-1)\right), \quad \text{if } n > BC,$$

and $M(n) = 0$ if $n \le BC$. Using standard techniques [37] this recurrence solves to

$$M(n) = \frac{2(n+1)}{B} \ln\left(\frac{n+1}{BC+2}\right) + O\left(\frac{1}{n}\right)$$

for $n > BC$.

The first correction we make is undercounting the misses that are incurred when the subproblem first reaches size $\le BC$. In the preceding analysis we count this as zero misses, when in fact this subproblem may have no part in the cache. To account for this we add in n/B more misses because there are approximately n keys in all the subproblems that first reach size $\le BC$.

In the very first partitioning in quicksort, there are n/B cache misses, but not necessarily for subsequent partitionings. At the end of partitioning some of the array in the left subproblem is still in the cache. Hence there are hits that we are counting as misses in the earlier analysis. Note that the right subproblem does not have these hits because by the time the algorithm reaches the right subproblem its data has been removed from the cache.

We first analyze the expected number of subproblems of size $> BC$. This is given by the recurrence,

$$N(n) = 1 + \frac{1}{n} \sum_{i=0}^{n-1} \left(N(i) + N(n-i-1)\right), \quad \text{if } n > BC,$$

and $N(n) = 0$ if $n \le BC$. This recurrence solves exactly to

$$N(n) = \frac{n+1}{BC+2} - 1,$$

for $n > BC$. For each of these subproblems there is a left subproblem. On average, $BC/2$ of the keys in this left subproblem are in the cache. Not all the accesses to these keys are hits. While the right pointer in the partitioning enjoys the benefit of access to these keys, the left pointer eventually accesses blocks that map to these blocks in the cache, thereby replacing them. For purposes of this analysis, we assume that exactly $C/2$ blocks of the left subproblem are in the cache and that the right pointer starts on the last block in the cache. Assume that the left pointer starts on a block that maps to block i in the cache. If $i < C/2$ then as the two pointers move together the right pointer experiences cache hits while the left pointer experiences one miss per every B accesses. This continues until the two pointers eventually map to the same block in the cache. On average, the number of these additional hits by the right pointer is $(C - i)/2$. If $i \geq C/2$ then again the right pointer experiences hits until it and the left pointer maps to the same cache block. This yields another $(C - i)/2$ cache hits. But there are more hits in this case. Because $i \geq C/2$ then $i - C/2$ cache blocks lie to the left of the initial position of the left pointer. With high likelihood these cache blocks are eventually accessed by the right pointer before the left pointer reaches blocks that map to them. Thus, we have an additional $i - C/2$ cache hits. If we assume that i is equally likely to be any value between 1 and C, the expected number of additional cache hits per left subproblem is

$$\frac{1}{C}\left[\sum_{i=1}^{C/2} \frac{C-i}{2} + \sum_{i=C/2+1}^{C}\left(\frac{C-i}{2} + i - \frac{C}{2}\right)\right] = \frac{3C}{8}.$$

Hence the expected number of hits not accounted for in the computation of $M(n)$ is approximately,

$$\frac{3C}{8}N(n).$$

Adding up all the pieces, for $n > BC$, the expected number of misses per key is approximately,

$$\left(M(n) + \frac{n}{B} - \frac{3C}{8}N(n)\right)\Big/ n,$$

which is approximated closely by expression (4). Figure 4 shows how well this approximation of cache misses predicts the actual performance of memory-tuned quicksort. In addition, Figure 4 shows approximation of $1/B$ more misses per key for base quicksort is predicted well.

For multiquicksort the approximate analysis is quite simple. If $n \leq BC$ multiquicksort the number of misses per key is simply $1/B$, the compulsory misses. For $n > BC$ then the algorithm partitions the input into $k = 3n/(BC)$ pieces, the vast majority of which are smaller than the cache. For the purposes of our analysis we assume they are all smaller than the cache. In this multipartition phase we move the partitioned keys into k linked lists one for each partition. Each node in a linked list has room for 100 keys. We approximate the number of misses per key in the first phase as $2/B$, one miss per block in the input array and one miss per block in the

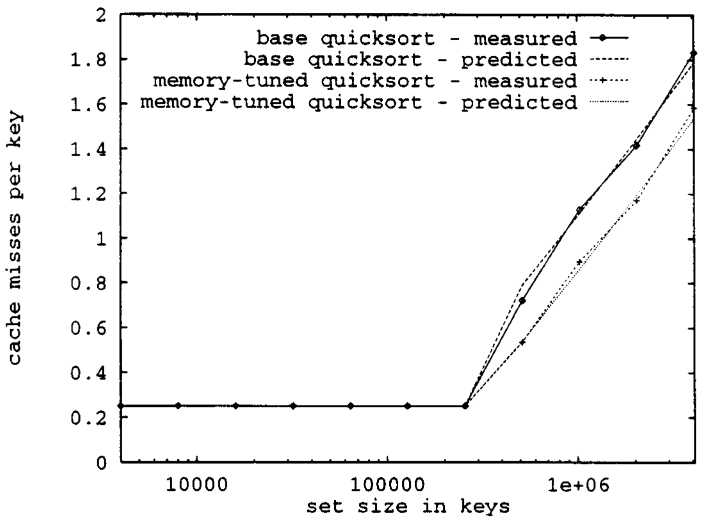

FIG. 4. Cache misses incurred by quicksort, measured versus predicted.

linked list. In the second phase each partition is returned to the input array and sorted in place. This costs approximately another $2/B$ misses per key. The cache miss overhead of maintaining the additional pivots is quite small for practical values of n, B, and C. The total is approximately

$$\frac{4}{B} \tag{5}$$

misses per key which closely matches the misses reported in Fig. 3.

7. HEAPSORT

While heaps are used for a variety of purposes, they were first proposed by Williams as part of the heapsort algorithm [56]. The heapsort algorithm sorts by first building a heap containing all of the keys and then removing them one at a time in sorted order. Using an array implementation of a heap results in a straightforward in-place sorting algorithm. On a set of n keys, Williams' algorithm takes $O(n \log n)$ steps to build the heap and $O(n \log n)$ steps to remove the keys in sorted order. In 1965 Floyd proposed an improved technique for building a heap with better average case performance and a worst case of $O(n)$ steps [22]. Williams' base algorithm with Floyd's improvement is still the most prevalent heapsort variant in use.

7.1. *Base Heapsort Algorithm*

As a base heapsort algorithm, we follow the recommendations of algorithm textbooks and we use an array implementation of a binary heap constructed using Floyd's method. In addition, we employ a standard optimization of using a sentinel at the end of the heap. The sentinel at the end of the heap ensures that every nonleaf node has exactly two children, thereby reducing the number of instructions needed during each remove-max operation. The literature contains a number of other optimizations that reduce the number of comparisons performed for both adds and removes [12, 17, 27], but in practice these increase the total number of instructions executed and do not improve performance. For this reason, we do not include them in our base heapsort.

7.2. *Memory Optimizations for Heapsort*

To this base heapsort algorithm, we now apply memory optimizations in order to further improve performance. In our paper [40] we showed that several memory optimizations improved the cache performance and overall performance of the heap used in the *hold model*. In the hold model, the number of elements in the heap is held constant as elements are alternately removed from and added to the heap. By adding some additional work between the removes and the adds, the hold model becomes a good approximation of a discrete event simulation with a static number of active events. The same optimizations also improve the performance of heapsort. The first optimization is to replace the traditional binary heap with a B-heap [34] where B is the number of keys that fit in a block. Generally, in a d-heap as described by Johnson [34], each nonleaf node has d children instead of the usual two. The use of d-heaps as a priority queue in an external memory environment has been studied [6, 38, 42]. If B is relatively small, say 4 or 8, there is an added advantage that the number of instructions executed for both add and remove-max is also reduced. The second optimization is to align the heap array in memory so that all B children lie on the same cache block. This optimization reduces what Lebeck and Wood refer to as *alignment misses* [41]. Also in our previous paper [40] we show that if a heap is larger than the size of the cache then Williams' original repeated-adds algorithm for building a heap incurs far fewer cache misses and performs better than Floyd's method. Thus, our memory-optimized heapsort dynamically chooses between Williams' repeated-adds method and Floyd's method for building a heap. If the heap is larger than the cache then Williams' method is chosen, otherwise Floyd's method is chosen. We call the base algorithm with these memory optimizations applied *memory-tuned heapsort*.

7.3. *Performance of Heapsort Algorithms*

Because the DEC Alphastation 250 has a 32 byte block size and because we are sorting eight byte keys, four keys fit in a block. As a consequence we choose the 4-heap in our memory-tuned heapsort. Figure 5 shows the performance of both the base heapsort and the memory-tuned heapsort. The instruction count curves show that memory-tuned heapsort executes fewer instructions than base heapsort.

The graph of cache performance shows that when the set to be sorted fits in the cache, the minimum 8 bytes/32 bytes = 0.25 compulsory misses are incurred per key for both algorithms. For larger sets the number of cache misses incurred by memory-tuned heapsort is less than half the misses incurred by base heapsort.

The execution time graph shows that the memory-tuned heapsort outperforms the base heapsort for all set sizes. The memory-tuned heapsort initially outperforms the base heapsort due to lower instruction counts. When the set size reaches the cache size, the gap widens due to differences in the number of cache misses incurred. For 4,096,000 keys, the memory-tuned heapsort sorts 81% faster than the base heapsort.

7.4. *Cache Analysis of Heapsort*

For $n \leq BC$ heapsort takes $1/B$ misses per key because it is an in-place algorithm. For $n > BC$ we directly apply the analysis technique, *collective analysis* that we used in a previous paper [40]. Collective analysis is an analytical framework for predicting the cache performance of algorithms when the algorithm's memory access behavior can be approximated using independent stochastic processes. As part of the analysis, the cache is divided into regions that are assumed to be accessed uniformly. By stating the way in which each process accesses each region, a simple formula can be used to predict cache performance. While collective analysis makes a number of simplifications which limit the class of algorithms that can be accurately analyzed, it serves well for algorithms whose behavior is understood, but whose exact reference pattern varies.

Our heapsort algorithm goes through two phases: the build-heap phase and the remove phase. For the build-heap phase, recall that William's method for building the heap simply puts each key at the bottom of the

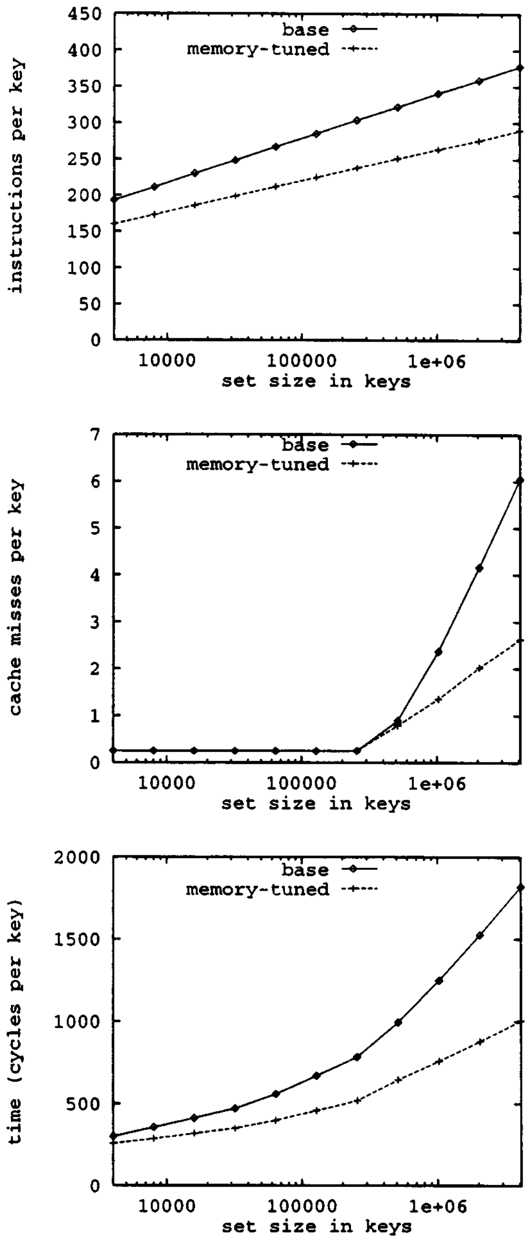

FIG. 5. Performance of heapsort on sets of 4000 to 4,096,000 keys. From top to bottom, the graphs show instruction counts per key, cache misses per key, and the execution times per key. Executions were run on a DEC Alphastation 250 and simulated cache capacity is 2 megabytes with a 32 byte block size.

heap and percolates it up to its proper place [56]. We pessimistically assume that all of these adds percolate to the root and that only the leaf-to-root path is in the cache. In this case, the probability that the next leaf is not in the cache is $1/B$, the chance that the parent of that leaf is not in the cache is $1/B^2$, the chance that the grandparent of that leaf is not in the cache is $1/B^3$, and so on. This gives the simple approximation of the number of misses incurred per key during the build phase of $\sum_{i=0}^{\infty} 1/B^{i+1} = 1/(B-1)$. Thus, we estimate the expected number of misses for the build-heap phase at $n/(B-1)$. It is interesting to note that this is within a small constant factor of the n/B compulsory misses that must be incurred by any build-heap algorithm.

The second phase of heapsort is where we apply collective analysis. As part of our collective analysis work, we analyzed the cache performance of d-heaps in the hold model. That analysis is summarized in our paper [40, Eq. (6)] where an interested reader can examine its derivation. We do not repeat that derivation here, instead we simply apply that equation to heapsort instead of the hold model. We divide the remove phase of heapsort into $n/(BC)$ subphases with each subphase removing BC keys. For $0 \leq i < n/(BC) - 1$ we model the removal of keys $BCi + 1$ to $BC(i+1)$ as BC steps on an array of size $n - BCi$ in the hold model. In essence, we are saying that the removal of BC keys from the B-heap is approximated by BC repeated remove-maxs and adds in approximately the same size B-heap. Admittedly, this approximation was one of convenience because we already had a complete approximate analysis of the B-heap in the hold model. Combining this remove phase estimate with our build-heap predictions yields Fig. 6. This graph shows our cache predictions for heapsort using both the traditional binary heap (base heapsort) and the 4-heap (memory-tuned heapsort). The predictions match the simulation results surprisingly well considering the simplifying assumptions made in the analysis.

8. RADIX SORT

Radix sort is the most important noncomparison based sorting algorithm used today. Knuth [37] traces the radix sort suitable for sorting in the main memory of a computer to a Master's thesis of Seward, 1954 [46]. Radix sort uses the principle that a b bit key can be thought of as a number in base 2^r of length $\lceil b/r \rceil$. The number r is called a *radix*. Seward pointed out that radix sort of n keys can be accomplished using an n key input array, an n key auxiliary array, and a count array of size 2^r which can hold integers up

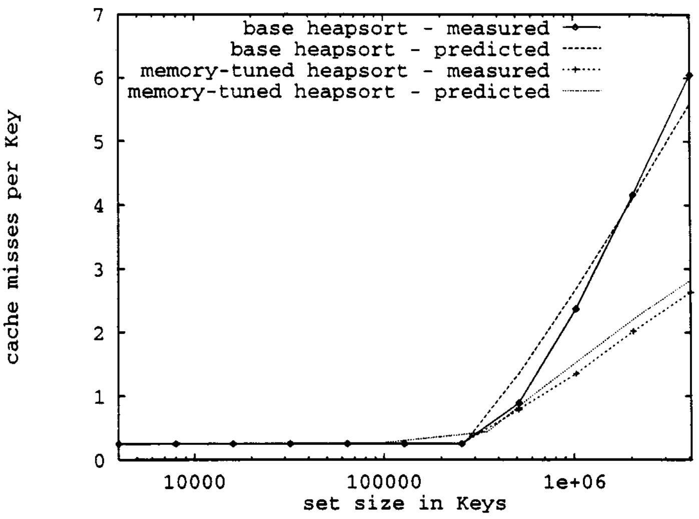

FIG. 6. Cache misses incurred by heapsort, measured versus predicted.

to size n. Knuth [37, p. 172] gives an excellent description of Seward's algorithm. Briefly summarizing, Seward's method does $\lceil b/r \rceil$ iterations each with two passes over the source array. The first pass accumulates counts of the number of keys with each radix. The counts are used to determine the offsets for buckets in the destination array. The second pass moves the source array into the buckets in the destination array. Similar to mergesort, the two n key arrays can alternate roles as source and destination array to avoid unnecessary copying. Seward's method is still a standard method found in radix sorting programs. Radix sort is often called a "linear time" sort because for keys of fixed length and for a fixed radix a constant number of passes over the data is sufficient to accomplish the sort, independent of the number of keys. Friend [23] suggested an improvement to reduce the number of passes over the source array, by accumulating the counts for the $(i + 1)$st iteration while concurrently moving keys to the destination array for the ith iteration. This requires a second count array of size 2^r. The first count array is used to keep track of the current offsets, while the second count array is used to accumulate counts for the next iteration. Friend's modification reduces both the instruction count and the cache misses of radix sort. Our radix sort is a highly optimized version of Seward's algorithm with Friend's improvement. The final task is to pick the radix which minimizes instruction count. This is done empirically because there is no universally best r which minimizes instruction count.

There is no obvious memory optimization for radix sort that is similar to those that we used for our comparison sorts. A simple memory optimization is to choose the radix which minimizes cache misses. As it happens, for our implementation of radix sort, a radix of 16 bits minimizes both cache misses and instruction count on an Alphastation 250. In the analysis subsection in the following text we take a closer look at the cache misses incurred by radix sort.

8.1. *Performance of Radix Sort*

For this study the keys are 64 bit integers and the counts can be restricted to 32 bit integers. With a 16 bit radix the two count arrays together are $\frac{1}{4}$ the size of the 2 Megabyte cache. Figure 7 shows the resulting performance. The instruction count graph shows radix sort's linear time behavior rather stunningly. The cache miss graph shows that when the size of the input array reaches the cache capacity the number of cache misses rapidly grows to a constant slightly more than three misses per key. The execution time graph clearly shows the effect that cache misses can have on overall performance. The execution time curve looks much like the instruction count curve until the input array exceeds the cache size at which time cycles per key increase according to the cache miss curve.

8.2. *Cache Analysis of Radix Sort*

The approximate cache miss analysis of radix sort is more complicated than our previous analyses for a number of reasons. First, there are a multitude of parameters to consider in this analysis: n, the number of keys; b, the number of bits per key; r, the radix; B, the number of keys per block; A, the number of counts per block; C, the capacity of the cache in blocks. Second, there are several cache effects to consider. We focus on what we feel are the three most significant cache effects, the capacity misses in traversals over the source and destination arrays, the conflict misses between the traversal of the source array and accesses to the two count arrays, and the conflict misses between the traversal of the source array and accesses to the destination array.

We will focus on analyzing the number of cache misses for a fixed large number of keys, while varying the radix. In the case of radix 16 this analysis attempts to predict the flat part of the cache curve in Figure 7. We assume that the size of the two count arrays is less than the cache capacity, $2^{r+1} \leq AC$. In addition, we assume that $r \leq b$. For $n > BC$, the expected number of cache misses per key in radix sort is approximated by sum of

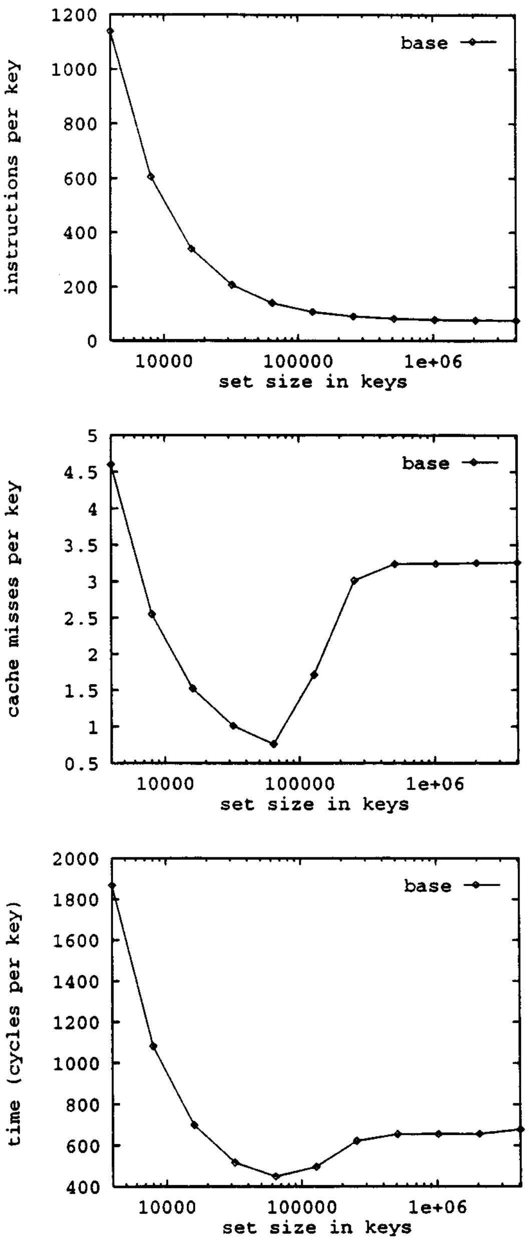

FIG. 7. Performance of radix sort on sets of 4000 to 4,096,000 keys. From top to bottom, the graphs show instruction counts per key, cache misses per key, and the execution times per key. Executions were run on a DEC Alphastation 250 and simulated cache capacity is 2 megabytes with a 32 byte block size.

three terms $M_{\text{trav}} + M_{\text{count}} + M_{\text{dest}}$ where M_{trav} is the number of misses per key incurred in traversing the source and destination arrays, M_{count} is the expected number of misses per key incurred in accesses to the count arrays, and M_{dest} is the expected number of misses per key in the destination array caused by conflicts with the traversal of the source array. We approximate M_{trav}, M_{count}, and M_{dest} by the following,

$$M_{\text{trav}} = \frac{1}{B}\left(2\left\lceil\frac{b}{r}\right\rceil + 1\right) + \frac{2}{B}\left(\left\lceil\frac{b}{r}\right\rceil \bmod 2\right), \tag{6}$$

$$\begin{aligned} M_{\text{count}} = {} & \frac{2^{r+1}}{ABC}\left(1 - \left(1 - \min\left(1, \frac{A}{2^r}\right)\right)^{BC}\right)\left\lfloor\frac{b}{r}\right\rfloor \\ & + \frac{2^{(b \bmod r)+1}}{ABC}\left(1 - \left(1 - \min\left(1, \frac{A}{2^{b \bmod r}}\right)\right)^{BC}\right), \end{aligned} \tag{7}$$

$$\begin{aligned} M_{\text{dest}} = {} & \frac{(B-1)2^r}{B^2C}\left(1 - \left(1 - \frac{1}{2^r}\right)^{BC}\right)\left\lfloor\frac{b}{r}\right\rfloor \\ & + \frac{(B-1)2^{b \bmod r}}{B^2C}\left(1 - \left(1 - \frac{1}{2^{b \bmod r}}\right)^{BC}\right). \end{aligned} \tag{8}$$

We start with the derivation of M_{trav}, Eq. (6). To implement Friend's improvement, there is an initial pass over the input array during which counts are accumulated in the first count array. Then there are $\lceil b/r \rceil$ iterations, where a source array is moved to the destination array according to the offsets established by one of the count arrays in the previous iteration. If $\lceil b/r \rceil$ is odd then the final sorted array must be copied back into the input array. In total, there are $2\lceil b/r \rceil + 1 + 2(\lceil b/r \rceil \bmod 2)$ passes over either a source or destination array. Ignoring any conflicts between the source and destination array, the number of misses per key is $1/B$ times the number of passes.

The quantity M_{count} accounts for the misses in random accesses to the two count arrays caused by the traversal over the source array. There are a total of $\lceil b/r \rceil + 1$ traversals over the source array that conflict with random accesses to at least one count array. If r divides b then the first and last traversal conflict with random accesses to one count array of size 2^r and all the other traversals conflict with random accesses to two count arrays of size 2^r. If r does not divide b then the first traversal conflicts with random accesses to one count array of size 2^r, the last traversal conflicts with random accesses to one count array of size $2^{b \bmod r}$, the second to last traversal conflicts with random access to one count array of size 2^r and one count array of size $2^{b \bmod r}$, and the remaining traversals, if any, conflict with random accesses to two count arrays of size 2^r. Generally, let us consider a pass over a source array that is in conflict with a count array of size 2^m. Consider a specific block x in one of the count arrays. Every BC steps of the algorithm the traversal of the source array ejects the block x from the cache. If the block is accessed in the count array before another BC steps of the algorithm then a miss is incurred in that access. The probability that block x is accessed in BC steps is $1 - (1 - \min(1, A/2^m))^{BC}$. The total expected number of misses is approximated by this quantity times the number of blocks in the array $(2^m/A)$ times the number of times x is visited in the traversal $(n/(BC))$. Thus, the expected number of misses per key incurred by a traversal over the source array in the count array of size 2^m is approximately,

$$\frac{2^m}{ABC}\left(1 - \left(1 - \min\left(1, \frac{A}{2^m}\right)\right)^{BC}\right). \tag{9}$$

If r divides b then the effect is $2b/r$ passes that conflict with a single count array of size 2^r. Expression (9) yields the first term of Eq. (7) (and the second term is negligible). If r does not divide b then the effect is $2\lfloor b/r \rfloor$ passes that conflict with a count array of size 2^r and two passes that conflict with a count array of size $2^{b \bmod r}$. This yields Eq. (7).

The quantity M_{dest} accounts for the conflict in the cache between the traversal over the source array and accesses into the destination array. Consider a specific block x in the destination array. In each iteration of the algorithm this block is visited exactly B times. The first visit is a cache miss that was accounted for in M_{trav} previously. Each of the remaining $B - 1$ visits is a cache hit unless the traversal in the source array visits the cache block of x before the next visit to x by the destination array. Assume that the number of pointers into the destination array is 2^m. As an approximation we assume that the $B - 1$ remaining accesses to block x each occur independently with probability $1/2^m$. Under this assumption, the probability that the traversal will visit the cache block of x before the next visit to x by the destination array is

$$\frac{2^m}{BC}\left(1 - \left(1 - \frac{1}{2^m}\right)^{BC}\right). \tag{10}$$

To see this, suppose that in exactly $i + 1$ steps the traversal will visit the cache block of x. The probability that x is accessed in the destination array before the traversal captures the cache block of x is approximately $(1 - 1/2^m)^i$. With probability $1/BC$ the traversal is $i + 1$ steps from capturing the cache block of x. Thus, the probability that the traversal captures the cache block of x before the access to block x is

$$\frac{1}{BC} \sum_{i=0}^{BC-1} \left(1 - \frac{1}{2^m}\right)^i,$$

which is equal to expression (10). To complete the derivation of Eq. (8) there are two cases to consider, depending on whether r divides b or not. If r divides b then there are b/r traversals of the source array that conflict with a traversal of the destination array. There are 2^r pointers into the destination array. The expected number of misses per key is then the number of passes, b/r, times $2^r/BC(1 - (1 - 1/2^r)^{BC})$ times $(B - 1)/B$, which totals to the first term of M_{dest}. In this case the second term of M_{dest} is negligible. If r does not divide b then there are $\lfloor b/r \rfloor$ traversals with 2^r pointers and one traversal with $2^{b \bmod r}$ pointers. Using an analysis similar to that in the preceding text yields the equation for M_{dest}. We should note that a more accurate formula for M_{dest}, one that does not treat the $B - 1$ accesses to block x as independent, can be derived. However, the more accurate formula is actually a complex recurrence that does not yield itself to a simple closed form solution nor to a tractable numerical solution. Our independence assumption seems to yield fairly accurate results.

An interesting question is how well our expression predicts cache performance as a function of the radix for large inputs sets. Figure 8 compares this approximation with the actual number of cache misses incurred per key, as a function of r, by radix sort for $n = 4{,}096{,}000$, $b = 64$, $A = 8$, $B = 4$, and $C = 2^{16}$. Figure 8 only goes up to radix 18 because beyond radix 18 the two count arrays are together larger than the cache. The theoretical prediction and the simulation agree that for these parameters the optimal radix is 16.

9. DISCUSSION

In this section we compare the performance of the fastest variant of each of the four sorting algorithms we examined in our study. In addition, we examine the performance of our algorithms on four additional architectures, in order to explore the robustness of our memory optimizations.

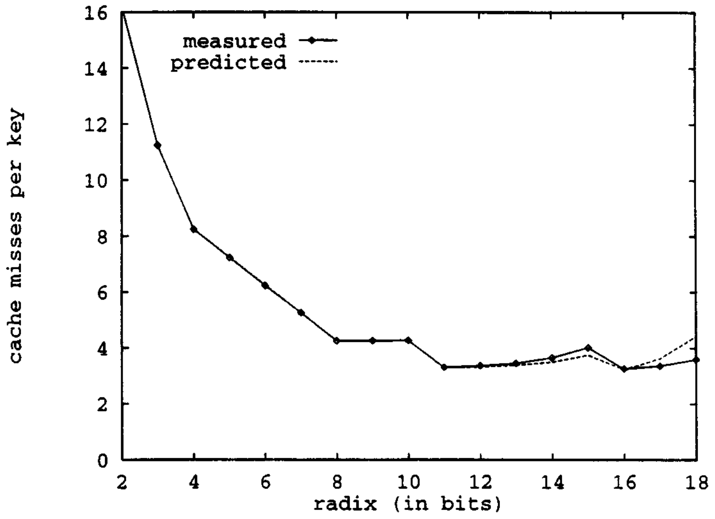

FIG. 8. Cache misses incurred by radix sort, measured versus predicted.

9.1. *Performance Comparison*

To compare performance across sorting algorithms, Figure 9 shows the instruction count, cache misses, and cycles executed per key for the fastest variants of heapsort, mergesort, quicksort, and radix sort.

The instruction count graph shows that the memory-tuned heapsort executes the most instructions, while radix sort executes the least. The cache miss graph shows that radix sort has the most cache misses, incurring more than twice as many misses as memory-tuned quicksort. The execution time graph strikingly shows the effect of cache performance on overall performance. For the largest data set memory-tuned quicksort ran 24% faster than radix sort even though memory-tuned quicksort performed more than three times as many instructions. This graph also shows that the small differences in instruction count and cache misses between the memory-tuned mergesort and memory-tuned quicksort offset to yield two algorithms with very comparable execution times.

9.2. *Robustness*

In order to determine if our experimental results generalize beyond the DEC Alphastation 250, we ran our programs on four other platforms: an IBM Power PC, a Sparc 10, a DEC Alpha 3000/400 and a Pentium base PC. Figure 10 shows the speedup that the memory-tuned heapsort achieves

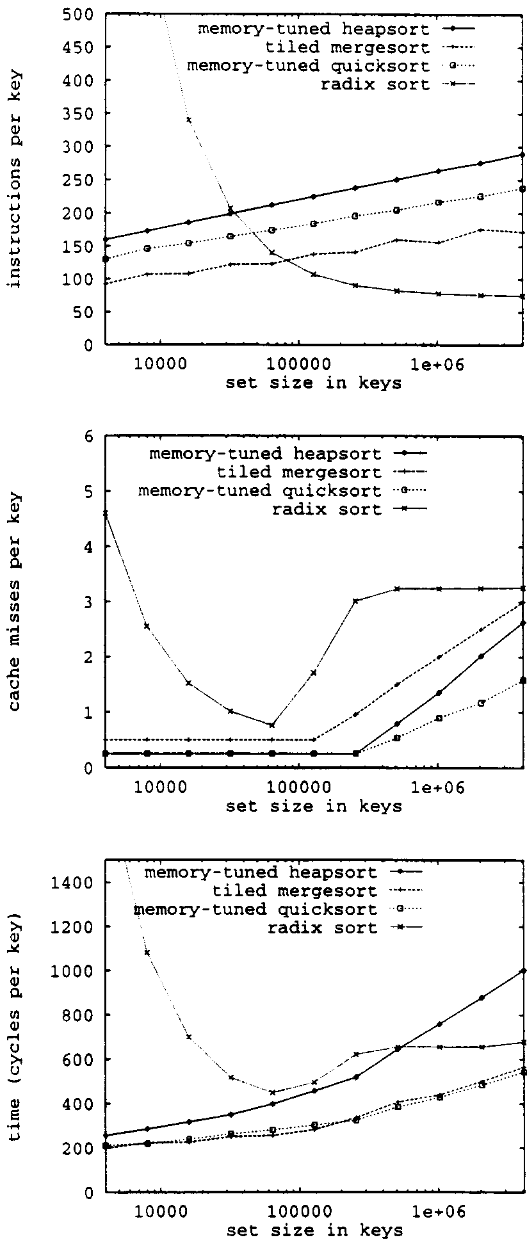

FIG. 9. Instruction count, cache misses, and execution time per key for the best heapsort, mergesort, quicksort, and radix sort on a DEC Alphastation 250. Simulated cache capacity is 2 megabytes, block size is 32 bytes.

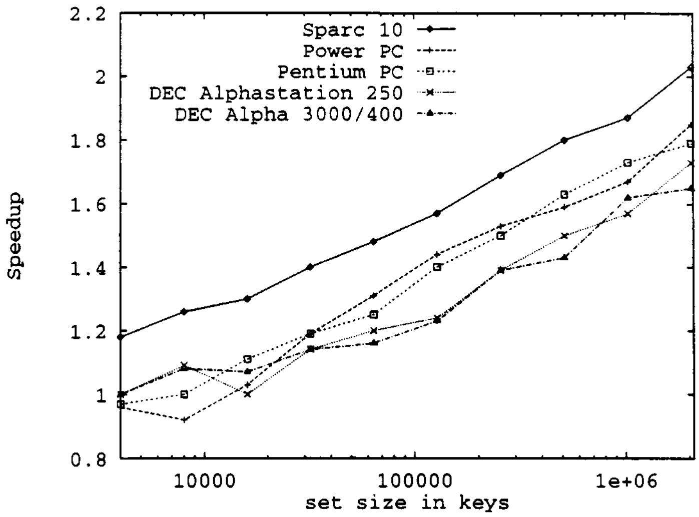

FIG. 10. Speedup of memory-tuned heapsort over base heapsort on five architectures.

over the base heapsort for the Alphastation 250 as well as these four additional machines. Despite the differences in architecture, all the platforms show similar speedups. It is interesting to note that all the speed-up curves show steady logarithmic growth.

Figure 11 shows the speedup of our tiled mergesort over the traditional mergesort for the same five machines. We chose tiled mergesort because it is a simple variant of the standard iterative mergesort that achieved good speedup in our base study on the DEC Alphastation 250. We wondered if similar speedups would be found on other machines. Unlike the heapsort case, the speedups for mergesort differ substantially. One clear difference is caused by the varying cache sizes on the machines. Machines with the smallest caches (Pentium, DEC 3000) had their speed-up curves rise sooner than machines with larger caches (Sparc, Alphastation). Another cause for the variation is the page mapping policies of the different operating systems. Throughout this study we have assumed that a block of contiguous pages in the virtual address space map to a block of contiguous pages in the cache. This is only guaranteed to be true when caches are virtually indexed rather than physically indexed [29]. Unfortunately, the caches on all five of our test machines are physically indexed. Fortunately, some operating systems, such as Digital Unix, have virtual to physical page mapping polices that attempt to map pages so that blocks of memory nearby in the virtual address space do not conflict in the cache [49]. Unlike

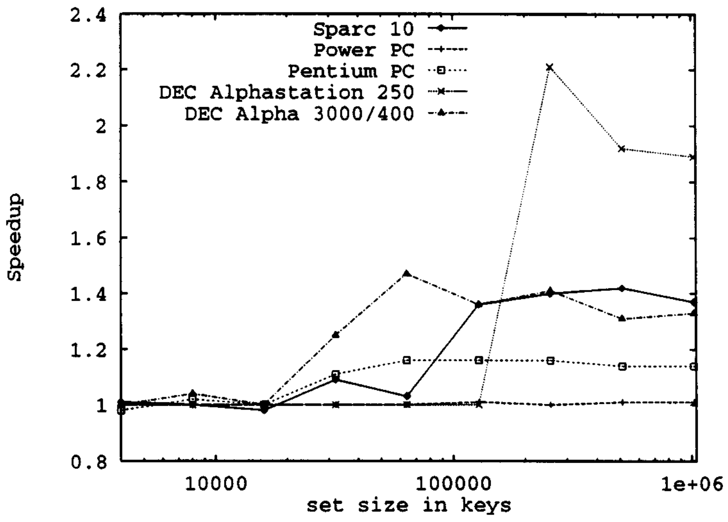

FIG. 11. Speedup of tiled mergesort over base mergesort on five architectures.

the heapsort algorithms, tiled mergesort relies heavily on the assumption that a cache-sized block of virtual pages does not conflict in the cache. As a result, the speedup of tiled mergesort relies heavily on the quality of the operating system's page mapping decisions. While the operating systems for the Sparc and the Alphas (Solaris and Digital Unix) make cache-conscious decisions about page placement, the operating systems for the Power PC and the Pentium (AIX and Linux) appear not to be as careful. Another interesting feature of the speed-up curves is how they all more or less flatten out eventually. This could be predicted because the two expressions (1) and (2) for the number of cache misses of base mergesort and tiled mergesort, respectively, differ by a constant.

10. CONCLUSIONS

We have explored the potential performance gains that cache-conscious design and analysis offers to classic sorting algorithms. The main conclusion of this work is that the effects of caching are extremely important and need to be considered if good performance is a goal. Despite its very low instruction count, radix sort is outperformed by both mergesort and quicksort due to its relatively poor locality. Despite the fact that the multimergesort executed 75% more instructions than the base mergesort, it sorts up to 70% faster. Neither the multimergesort nor the multiquicksort are in-place or stable. Nevertheless, these two algorithms offer something that none of the others do. They both incur very few cache misses, which renders their overall performance far less sensitive to cache miss penalties than the others. As a result, these algorithms can be expected to outperform the others as relative cache miss penalties continue to increase.

In this paper, a number of memory optimizations are applied to algorithms in order to improve their overall performance. What follows is a summary of the design principles applied in this work:

- Improving cache performance even at the cost of an increased instruction count can improve overall performance.
- Knowing and using architectural constants such as cache size and block size can improve an algorithm's memory system performance beyond that of a generic algorithm.
- Spatial locality can be improved by adjusting an algorithm's structure to fully utilize cache blocks.
- Temporal locality can be improved by padding and adjusting data layout so that structures are aligned within cache blocks.
- Capacity misses can be reduced by processing large data sets in cache-sized pieces.
- Conflict misses can be reduced by processing data in cache-block-sized pieces.

This paper also shows that despite the complexities of caching, the cache performance of algorithms can be reasonably approximated with a modest amount of work. Figures 2, 4, 6, and 8 show that our approximate analysis gives good information about cache performance.

ACKNOWLEDGMENT

We thank the anonymous referees for many helpful suggestions for improvements. We also thank Jim Fix and Mike Salisbury for carefully reading several versions of this paper and for making many useful suggestions for improvements.

REFERENCES

1. A. Agarwal, M. Horowitz, and J. Hennessy, An analytical cache model, *ACM Trans. Comput. Systems* **7**(2) (1989), 184–215.
2. A. Aggarwal, K. Chandra, and M. Snir, Hierarchical memory with block transfer, *in* ''The Twenty-Eighth Annual IEEE Symposium on Foundations of Computer Science,'' 1987, pp. 204–216.
3. A. Aggarwal, K. Chandra, and M. Snir, A model for hierarchical memory, *in* ''The Nineteenth Annual ACM Symposium on Theory of Computing,'' 1987, pp. 305–314.
4. A. Aggarwal and J. Vitter, The input/output complexity of sorting and related problems, *Comm. ACM* **31**(9) (1988), 1116–1127.
5. B. Alpern, L. Carter, E. Feig, and T. Selker, The uniform memory hierarchy model of computation, *Algorithmica* **12**(2–3) (1994), 72–109.
6. L. Arge, The buffer tree: A new technique for optimal I/O-algorithms, *in* ''Proceedings of the WADS'95,'' Lecture Notes in Computer Science, Vol. 955, pp. 334–345, Springer-Verlag, Berlin/New York, 1995.
7. L. Arge, P. Ferragina, R. Grossi, and J. Vitter, On sorting strings in external memory, *in* ''Proceedings of the STOC '97,'' 1997, pp. 540–548.
8. L. Arge, O. Procopiuc, S. Ramaswamy, T. Suel, and J. S. Vitter, Theory and practice of I/O-efficient algorithms for multidimensional batched searching problems, *in* ''Proceedings of the SODA '98,'' 1998.
9. T. M. Austin, ''Hardware and Software Mechanisms for Reducing Load Latency,'' Ph.D. dissertation, University of Wisconsin, Computer Sciences, 1996.
10. B. N. Bershad, D. Lee, T. Romer, and J. B. Chen, Avoiding conflict masses dynamically in large direct-mapped caches, *in* ''Sixth International Conference on Architectural Support for Programming Languages and Operating Systems,'' 1994, pp. 158–170.
11. G. Blelloch, C. Plaxton, C. Leiserson, S. Smith, B. Maggs, and M. Zagha, A comparison of sorting algorithms for the connection machine, *in* ''Proceedings of the Third ACM Symposium on Parallel Algorithms and Architecture,'' July 1991, pp. 3–16.
12. S. Carlsson, An optimal algorithm for deleting the root of a heap, *Inform. Process. Lett.* **37**(2) (1991), 117–120.
13. S. Carr, K. McKinley, and C. W. Tseng, Compiler optimizations for improving data locality, *in* ''Sixth International Conference on Architectural Support for Programming Languages and Operating Systems,'' 1994, pp. 252–262.
14. Y.-J. Chiang, Experiments on the practical I/O efficiency of geometric algorithms: Distribution sweep vs. plane sweep, *in* ''Proceedings of the WADS '95,'' Lecture Notes in Computer Science, Vol. 955, pp. 346–357, Springer-Verlag, Berlin/New York, 1995.
15. M. Cierniak and Wei Li, Unifying data and control transformations for distributed shared-memory machines, *in* ''Proceedings of the 1995 ACM Symposium on Programming Languages Design and Implementation,'' Assoc. Comput. Mach., New York, 1995, pp. 205–217.
16. D. Clark, Cache performance of the VAX-11/780, *ACM Trans. Comput. Systems* **1**(1) (1983), 24–37.
17. J. De Graffe and W. Kosters, Expected heights in heaps, *BIT* **32**(4) (1992), 570–579.
18. F. Dehne, W. Dittrich, and D. Hutchinson, Efficient external memory algorithms by simulating coarse-grained parallel algorithms, *in* ''Proceedings of the Ninth Annual ACM Symposium on Parallel Algorithms and Architectures,'' 1997, pp. 106–115.

19. A. Diwan, D. Tarditi, and E. Moss, Memory subsystem performance of programs using copying garbage collection, *in* ''Proceedings of the Twenty-First Annual ACM Symposium on Principles of Programming Languages,'' 1994, pp. 1–14.
20. W. Feller, ''An Introduction to Probability Theory and Its Applications,'' Wiley, New York, 1971.
21. D. Fenwick, D. Foley, W. Gist, S. VanDoren, and D. Wissell, The AlphaServer 8000 series: High-end server platform development, *Digital Tech. J.* **7**(1) (1995), 43–65.
22. R. W. Floyd, Treesort 3, *Comm. ACM* **7**(12) (1964), 701.
23. E. H. Friend, *J. ACM* **3** (1956), 152.
24. S. Gal, Y. Hollander, and A. Itai, Optimal mapping in direct mapped cache environments, *Math. Programming* **63** (1994), 371–387.
25. S. Gal and B. Klots, Optimal partitioning which maximizes the sum of the weighted averages, *Oper. Res.* **43** (1995), 500–508.
26. D. Gannon, W. Jalby, and K. Gallivan, Strategies for cache and local memory management by global program transformation, *J. Parallel Distributed Comput.* **5**(5) (Oct. 1988), 587–616.
27. G. Gonnet and J. Munro, Heaps on heaps, *SIAM J. Comput.* **15**(4) (1986), 964–971.
28. D. Grunwald, B. Zorn, and R. Henderson, Improving the cache locality of memory allocation, *in* ''Proceedings of the 1993 ACM Symposium on Programming Languages Design and Implementation,'' Assoc. Comput. Mach., New York, 1993, pp. 177–186.
29. J. Hennesey and D. Patterson, ''Computer Architecture A Quantitative Approach,'' 2nd ed., Morgan Kaufman, San Mateo, CA, 1996.
30. C. A. R. Hoare, Quicksort, *Comput. J.* **5** (1962), 10–15.
31. F. E. Holberton, *in* ''Symposium on Automatic Programming,'' 1952, pp. 34–39.
32. Y. Hollander and A. Itai, ''On the Complexity of Direct Caching,'' Technical Report CS0794, Technion, Computer Science Department, Jan. 1994.
33. L. Hui and K. C. Sevcik, Parallel sorting by overpartitioning, *in* ''Proceedings of the Sixth ACM Symposium on Parallel Algorithms and Architecture,'' June 1994, pp. 46–56.
34. D. B. Johnson, Priority queues with update and finding minimum spanning trees, *Inform. Process. Lett.* **4** (1975).
35. K. Kennedy and K. McKinley, Optimizing for parallelism and data locality, *in* ''Proceedings of the 1992 International Conference on Supercomputing,'' 1992, pp. 323–334.
36. R. Kessler and M. Hill, Page placement algorithms for large real-indexed caches, *ACM Trans. Comput. Systems* **10**(4) (Nov. 1992), 338–359.
37. D. E. Knuth, ''The Art of Computer Programming, Vol. III.—Sorting and Searching,'' Addison-Wesley, Reading, MA, 1973.
38. V. Kumar and E. Schwabe, Improved algorithms and data structures for solving graph problems in external memory, *in* ''Proceedings of the IEEE Symposium on Parallel and Distributed Processing,'' 1996, pp. 169–177.
39. A. LaMarca, ''Caches and Algorithms,'' Ph.D. dissertation, University of Washington, May 1996.
40. A. LaMarca and R. E. Ladner, The influence of caches on the performance of heaps, *J. Experimental Algorithmics* **1**(4) (1996).
41. A. Lebeck and D. Wood, Cache profiling and the spec benchmarks: a case study, *Computer* **27**(10) (Oct. 1994), 15–26.
42. D. Naor, C. Martel, and N. Matloff, Performance of priority queue structures in a virtual memory environment, *Comput. J.* **34**(5) (Oct. 1991), 428–437.
43. C. Nyberg, T. Barclay, Z. Cvetanovic, J. Gray, and D. Lomet, Alphasort: a RISC machine sort, *in* ''1994 ACM SIGMOD International Conference on Management of Data,'' May 1994, pp. 233–242.

44. R. Reischuk, Probabilistic parallel algorithms for sorting and selection, *SIAM J. Comput.* **14** (1985), 396–409.
45. R. Sedgewick, Implementing quicksort programs, *Comm. ACM* **21**(10) (Oct. 1978), 847–857.
46. H. H. Seward, Masters thesis, M.I.T. Digital Computer Laboratory Report R-232, 1954.
47. J. P. Singh, H. S. Stone, and D. F. Thiebaut, A model of workloads and its use in miss-rate prediction for fully associative caches, *IEEE Trans. Comput.* **41**(7) (1992), 811–825.
48. A. Srivastava and A. Eustace, ATOM: A system for building customized program analysis tools, *in* "Proceedings of the 1994 ACM Symposium on Programming Languages Design and Implementation," Assoc. Comput. Mach., New York, 1994, pp. 196–205.
49. G. Taylor, P. Davies, and M. Farmwald, The TBL slice: a low-cost high-speed address translation mechanism, *in* "Proceedings of the Seventeenth Annual International Symposium on Computer Architecture," 1990, 355–363.
50. O. Temam, C. Fricker, and W. Jalby, Cache interference phenomena, *in* "Proceedings of the 1994 ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems," 1994, 261–271.
51. O. Temam, C. Fricker, and W. Jalby, Influence of cross-interferences on blocked loops: A case study with matrix-vector multiply, *ACM Trans. Programming Languages Systems* **17**(4) (1995), 561–575.
52. D. E. Vengroff and J. S. Vitter, I/O-efficient scientific computation using TPIE, *in* "Proceedings of the Goddard Conference on Mass Storage Systems and Technologies," NASA Conference Publication 3340, Vol. II, 1996, pp. 553–570.
53. A. Verkamo, External quicksort, *Performance Evaluation* **8**(4) (Aug. 1988), 271–288.
54. A. Verkamo, Performance comparison of distributive and mergesort as external sorting algorithms, *J. Systems Software* **10**(3) (Oct. 1989), 187–200.
55. L. Wegner and J. Teuhola, The external heapsort, *J. Systems Software* **15**(7) (July 1989), 917–925.
56. J. W. Williams, Heapsort, *Comm. ACM* **7**(6) (1964), 347–348.
57. M. Wolf and M. Lam, A data locality optimizing algorithm, *in* "Proceedings of the 1991 ACM Symposium on Programming Languages Design and Implementation," Assoc. Comput. Mach., New York, 1991, pp. 30–44.
58. M. Wolfe, More iteration space tiling, *in* "Proceedings of Supercomputing '89," 1989, pp. 655–664.