Cooperative Prefetching: Compiler and Hardware Support for Effective Instruction Prefetching in Modern Processors

Chi-Keung Luk

Department of Computer Science
University of Toronto
Toronto, Canada M5S 3G4
luk@eecg.toronto.edu

Todd C. Mowry

Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
tcm@cs.cmu.edu

Abstract

Instruction cache miss latency is becoming an increasingly important performance bottleneck, especially for commercial applications. Although instruction prefetching is an attractive technique for tolerating this latency, we find that existing prefetching schemes are insufficient for modern superscalar processors since they fail to issue prefetches early enough (particularly for non-sequential accesses). To overcome these limitations, we propose a new instruction prefetching technique whereby the hardware and software cooperate *to hide the latency as follows. The hardware performs aggressive sequential prefetching combined with a novel* prefetch filtering *mechanism to allow it to get far ahead without polluting the cache. To hide the latency of non-sequential accesses, we propose and implement a novel compiler algorithm which automatically inserts* instruction-prefetch instructions *into the executable to prefetch the targets of control transfers far enough in advance. Our experimental results demonstrate that this new approach results in speedups ranging from 9.4% to 18.5% (13.3% on average) over the original execution time on an out-of-order superscalar processor, which is more than double the average speedup of the best existing schemes (6.5%). This is accomplished by hiding an average of 71% of the original instruction stall time, compared with only 36% for the best existing schemes. We find that both the* prefetch filtering *and* compiler-inserted prefetching *components of our design are essential and complementary, that the compiler can limit the code expansion to less than 10% on average, and that our scheme is robust with respect to variations in miss latency and bandwidth.*

1. Introduction

The latency of fetching instructions is a key performance bottleneck in modern systems, and the problem is expected to get worse as the gap between processor and memory speeds continues to grow. While instruction caches are a crucial first step, they are not a complete solution. For example, a study conducted by Maynard *et al.* [7] demonstrates that many commercial applications suffer from relatively large instruction cache miss rates (e.g., over 20% in an 8KB cache) due to their large instruction footprints and poor instruction localities. To further tolerate this latency, one attractive technique is to automatically *prefetch* instructions into the cache before they are needed.

[0] To appear in *Proceedings of Micro-31*, Nov. 30–Dec. 2, 1998.

1.1. Previous Work on Instruction Prefetching

Several researchers have considered instruction prefetching in the past. We will begin by discussing and then quantitatively evaluating four of the most promising techniques that have been proposed to date, all of which are purely hardware-based: *next-N-line* prefetching [10, 11], *target-line* prefetching [12], *wrong-path* prefetching [8], and *Markov* prefetching [3].

Before we begin our discussion, we briefly introduce some prefetching terminology. The *coverage factor* is the fraction of original cache misses that are prefetched. A prefetch is *unnecessary* if the line is already in the cache (or is currently being fetched), and is *useless* if it brings a line into the cache which will not be used before it is displaced. An ideal prefetching scheme would provide a coverage factor of 100% and would generate no unnecessary or useless prefetches. In addition, the *timeliness* of prefetches is also crucial. The *prefetching distance* (i.e. the elapsed time between initiating and consuming the result of a prefetch) should be large enough to fully hide the miss latency, but not so large that the line is likely to be displaced by other accesses before it can be used (i.e. a useless prefetch).

The idea behind *next-N-line prefetching* [10, 11] is to prefetch the N sequential lines following the one currently being fetched by the CPU. A larger value of N tends to increase the prefetching distance, but also increases the likelihood of polluting the cache with useless prefetches. The optimal value of N depends on the line size, the cache size, and the behavior of the application itself. Next-N-line prefetching captures sequential execution as well as control transfers where the target falls within the next N lines. It is usually included as part of other more complex instruction prefetching schemes, and based on our experiments, it accounts for most of the performance benefit of these previously existing schemes.

To further expand the scope of prefetching to capture more control transfer targets, Smith and Hsu [12] proposed *target-line prefetching* which uses a prediction table to record the address of the line which most recently followed a given instruction line, thus enabling hardware to prefetch targets whenever an entry is found in this table. They observed that combining target-line prefetching with next-1-line prefetching produced significantly better results than either technique alone.

Rather than relying on history tables, Pierce and Mudge [8] proposed *wrong-path prefetching* which combines next-N-line prefetching with always prefetching the target of control transfers with static target addresses. Hence for conditional branches, both the target and fall-through lines will always be prefetched. However, since target addresses cannot be determined early, this scheme only outperforms next-N-line prefetching when a conditional branch is initially untaken but later taken (assuming that enough time has passed to hide the latency but not so much that the line has been displaced). Their results indicated that wrong-path prefetching performed slightly better than next-1-line prefetching on average.

Joseph and Grunwald [3] proposed *Markov prefetching*, which correlates consecutive miss addresses. These correlations are stored in a *miss-address prediction table* which is indexed using the current miss address, and which can return multiple predicted addresses. The Joseph and Grunwald study focused primarily on data cache misses, and did not compare Markov prefetching with techniques designed specifically for prefetching instructions.

Finally, we note that while Xia and Torrellas [13] considered instruction prefetching for codes where the layout has already been optimized using profiling information, we focus only on techniques which do not require changes to the instruction layout in this study.

1.1.1. Performance of Existing Instruction Prefetching Techniques

To quantify the performance benefits and limitations of the four prefetching techniques described above, we implemented each of them within a detailed, cycle-by-cycle simulator which models an out-of-order four-issue superscalar processor based on the MIPS R10000 [14]. We model a two-level cache hierarchy with split 32 KB, two-way set-associative primary instruction and data caches and a unified 1 MB, four-way set-associative secondary cache. Both levels use 32 byte lines. The penalty of a primary cache miss that hits in the secondary cache is at least 12 cycles, and the total penalty of a miss that goes all the way to memory is at least 75 cycles (plus any delays due to contention, which is modeled in detail). To provide better support for instruction prefetching, we further enhanced the primary instruction cache relative to the R10000 as follows: we divide it into four separate banks, and we add an eight-entry victim cache [4] and a 16-entry prefetch buffer [3]. Further details on our experiments will be presented later in Section 5.

Table 1. Parameters used in the evaluation of existing instruction prefetching techniques.

Technique	# of Lines Sequentially Prefetched	Target Prefetching Parameters: # of Targets	Table Size	Indexing Method
Next-N-Line	N = 2, 4, 8	0	0	*N/A*
Target-Line	2	1	64 entries	direct-mapped with tags
Wrong-Path	2	1	0	*N/A*
Markov	2	2	512 KB	direct-mapped with tags

Table 1 summarizes the prefetching parameters used throughout our experiments. These parameters were chosen through experimentation in an effort to maximize the performance of each scheme. All schemes effectively include next-2-line prefetching. (Although next-2-line prefetching was not in the original Markov prefetching design [3], we added it since we found that it improves performance.) When a target is to be prefetched, we prefetch two consecutive lines starting at the target address.

Figure 1 shows the performance impact of each prefetching scheme on a collection of seven non-numeric applications (which are discussed more in Section 5). We show three different versions of next-N-line prefetching (where N = 2, 4, and 8) in Figure 1, along with the original case without prefetching (**O**) and the case with a perfect instruction cache (**P**). Each bar represents execution time normalized to the case without prefetching, and is broken down into three categories corresponding to all potential graduation slots.[1] The bottom section (*Busy*) is the number of slots when instructions actually graduate, the top section (*I-Miss Stall*) is any non-graduating slots that would not occur with a perfect instruction cache, and the middle section (*Other Stall*) is all other slots where instructions do not graduate.

We observe from Figure 1 that despite significant differences in complexity and hardware cost, the various prefetching schemes offer remarkably similar performance, with no single scheme clearly dominating. Perhaps surprisingly, the best performance is achieved by either next-4-line or next-8-line prefetching in all cases except `perl`; even in `perl`, next-4-line prefetching is still within 1% of the best case. The reason for this is that the bulk of the benefit offered by each of these schemes is due to prefetching

[1] The number of graduation slots is the issue width (4 in this case) multiplied by the number of cycles. We focus on graduation rather than issue slots to avoid counting speculative operations that are squashed.

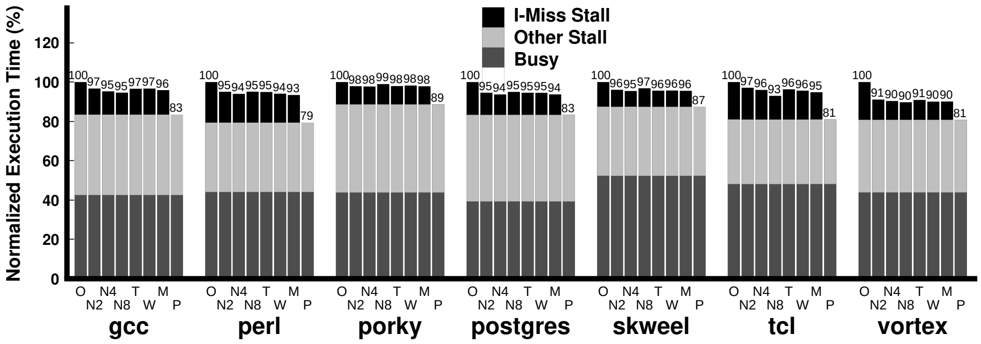

Figure 1. Performance of existing instruction prefetching techniques (O = original, Nx = next-x-line prefetching, T = target-line prefetching, W = wrong-path prefetching, M = Markov prefetching, P = perfect instruction cache).

sequential accesses.

Finally, we see in Figure 1 that these schemes are hiding no more than half of the stall time due to instruction cache misses. Through a detailed analysis of why these schemes are not more successful (further details are presented later in Section 6.1), we observe that although the coverage is generally quite high, the real problem is the *timeliness* of the prefetches—i.e. prefetches are not being launched early enough to hide the latency. Hence there is significant room for improvement over these existing schemes.

1.2. Our Solution

To hide instruction cache miss latency more effectively in modern microprocessors, we propose and evaluate a new fully-automatic instruction prefetching scheme whereby the compiler and the hardware cooperate to launch prefetches earlier (therefore hiding more latency) while at the same time maintaining high coverage and actually *reducing* the impact of useless prefetches relative to today's schemes. Our approach involves two novel components. First, to enable more aggressive sequential prefetching without polluting the cache with useless prefetches, we introduce a new *prefetch filtering* hardware mechanism. Second, to enable more effective prefetching of non-sequential accesses, we introduce a novel compiler algorithm which inserts explicit *instruction-prefetch instructions* into the executable to prefetch the targets of control transfers far enough in advance. Our experimental results demonstrate that our scheme provides significant performance improvements over existing schemes, eliminating roughly 50% or more of the latency that had remained with the best existing scheme.

This paper is organized as follows. We begin in Section 2 with an overview of our approach, and then present further details on the architectural and compiler support in Sections 3 and 4. Sections 5 and 6 present our experimental methodology and our experimental results, and finally we conclude in Section 7.

2. Cooperative Instruction Prefetching

We begin this section with a high-level overview of our prefetching scheme. To make our approach concrete, we also present an example illustrating prefetch insertion.

2.1. Overview of the Prefetching Algorithm

As we mentioned earlier, the key challenge in designing a better instruction prefetching scheme is to be able to launch prefetches earlier—i.e. to achieve a larger *prefetching distance*. Let us consider the sequential and non-sequential portions of instruction streams separately.

2.1.1. Prefetching Sequential Accesses

Since the addresses within sequential access patterns are trivial to predict, they are well-suited to a purely hardware-based mechanism such as next-N-line prefetching. To get far enough ahead to fully hide the latency, we would like to choose a fairly large value for N (e.g., $N = 8$ in our experiments). However, the problem with this is that larger values of N increase the probability of overshooting the end of the sequence and polluting the cache with useless prefetches. For example, next-8-line prefetching performs worse than next-4-line prefetching for four cases in Figure 1 (`perl`, `porky`, `postgres`, and `skweel`) due to this effect.

The ideal solution would be to prefetch ahead aggressively (i.e. with a large N) but to stop upon reaching the end of the sequence. Xia and Torrellas [13] proposed a mechanism for doing this which uses software to explicitly mark the likely end of a sequence with a special bit. In contrast, we achieve a similar effect using a more general *prefetch filtering* mechanism which automatically detects and discards useless prefetches before they can pollute the instruction cache. We will explain how the prefetch filter works in detail later in Section 3.3, but the basic idea is to use two-bit saturating counters stored in the secondary cache tags to dynamically detect cases where lines have been repeatedly prefetched into the primary instruction cache but were not accessed before they were displaced (i.e. *useless* prefetches). When prefetches for such lines subsequently arrive at the secondary cache, they are simply dropped. One advantage of our approach is that it adapts to the dynamic branching behavior of the program, rather than relying on static predictions of likely control flow paths. In addition, our filtering mechanism is equally applicable to *non-sequential* as well as sequential prefetches.

2.1.2. Prefetching Non-Sequential Accesses

In contrast with sequential access patterns, purely hardware-based prefetching schemes are far less successful at prefetching *non-sequential* instruction accesses early enough. Wrong-path prefetching does not attempt to predict the target address of a given branch early, but instead hopes that the same branch will be revisited sometime in the not-too-distant future with a different branch outcome. Both target-line and Markov prefetching rely on building up history tables to predict addresses to prefetch along control targets. However, if a control transfer is encountered for the first time or if its entry has been displaced from the finite history table, then its target will not be prefetched.[2] Perhaps more importantly, even if a valid entry is found in the history table, it is often too late to fully hide the latency of prefetching the target since the processor is already accessing the line containing the branch.

To overcome these limitations, we rely on *software* rather than hardware to launch non-sequential instruction prefetches early enough. To avoid placing any burden on the programmer, we use the compiler to insert these new instruction-prefetching instructions automatically. As we describe in further detail later in Section 4, our compiler algorithm moves prefetches back by a specified *prefetch-scheduling distance* while being careful not to insert prefetches that would be redundant with either next-N-line prefetching or other software instruction prefetches. Since many control transfers within procedures have targets within the N lines covered by our next-N-line prefetcher, the bulk of the instructions inserted by our compiler algorithm are for prefetching *across* procedure boundaries (as we show later in Section 5). Hence, although it is an oversimplification, one could think of our scheme as being primarily hardware-based for *intraprocedural* prefetching, and primarily software-based for *interprocedural* prefetching.

While direct control transfers (i.e. where the target address is statically known) are handled in a straightforward way by our algorithm, *indirect jumps* require some additional support for software to generate the target addresses early. We have proposed and evaluated a number of ways to

[2] Note that although our prefetch filtering mechanism can also potentially suffer from the limitations of learning within a finite table, we find that it is far more important to prefetch target addresses early enough rather than filtering out all useless prefetches.

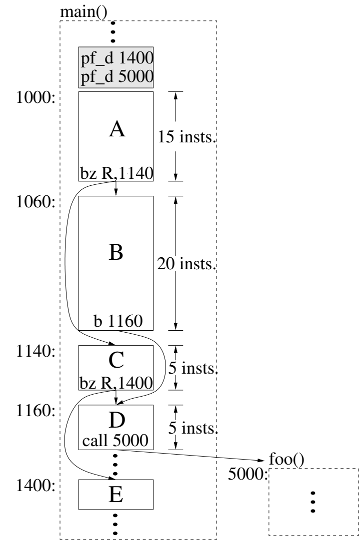

Figure 2. Example of prefetch insertion.

prefetch indirect jumps [6]. However, since our experimental results indicate that the marginal performance benefit of supporting indirect prefetches is quite small (less than 1% speedup), we do not consider them further in this study.

While the advantage of software-controlled instruction prefetching is that it gives us greater control over issuing prefetches early, the potential drawbacks are that it increases the code size and effectively reduces the instruction fetch bandwidth (since the prefetch instructions themselves consume part of the instruction stream). Fortunately, our experimental results demonstrate that this advantage outweighs any disadvantages.

2.2. Example of Prefetch Insertion

To make our discussion more concrete, Figure 2 shows an example of prefetch insertion. We assume the following: a cache line is 32 bytes long; an instruction is 4 bytes long (hence one cache line contains 8 instructions); hardware next-8-line prefetching is enabled; and the prefetch-scheduling distance is 20 instructions. This example shows two procedures, `main()` and `foo()`, where `main()` contains five basic blocks (labeled `A` through `E`). Two prefetches (`pf_d`) have been inserted at the beginning of basic block `A`: one targeting block `E`, and the other targeting procedure `foo()`. The compiler does not insert software prefetches for blocks `B`, `C` or `D` at `A` since they will already be handled by next-8-line prefetching. The prefetch targeting `E` is inserted in block `A` rather than in block `C` in order to guarantee a prefetching distance of at least 20 instructions. Although there are two possible paths from `A` to `foo()` (i.e. `A`→`B`→`D`→`foo()` and `A`→`C`→`D`→`foo()`), the compiler inserts only a single prefetch of `foo()` in `A` (rather than inserting one in A and one in B) because (i) A *dominates*[3] both paths, and (ii) the compiler determines that these prefetched instructions are not likely to be displaced by other instructions fetched along the path A→B→D→foo().

3. Architectural Support

Our prefetching scheme requires new architectural support. In this section, we describe our extensions to the instruction set, how these new instructions affect the pipeline, and the new hardware that we add to the memory system (including the prefetch filter).

3.1. Extensions to the Instruction Set Architecture

Without loss of generality, we assume a base instruction set architecture similar to the MIPS ISA [5]. Within a 32-bit MIPS instruction, the high-order six bits contain the opcode. For the jump-type instructions, the remaining 26 bits contain the low-order bits of the target word address. We will use this same instruction format as our starting point.

There are many ways to encode our new instruction-prefetch instructions, and Figure 3(a) shows just one of the possibilities. An opcode is designated to identify instruction-prefetch instructions. In contrast with the standard jump-type instruction format, we assume that 24 bits (bits 2 through 25) contain information for computing the prefetch address(es), bit 1 indicates the prefetch type, and bit 0 is currently not used. The prefetch type `pf_d` stores a single prefetch address in a format similar to a MIPS jump address. The only difference is that since the lower two bits are ignored, it effectively encodes a 16-byte-aligned address. (Since most machines have at least 16 byte instruction lines, this is not a limitation.) The `pf_c` type is a *compact* format which encodes two target addresses within the 24-bit field in the form of offsets between the target address lines and the prefetch instruction line itself (again, a single offset bit represents 16 bytes); each offset is 12 bits wide.

3.2. Impact on the Processor Pipeline

Many recent processors have implemented instructions for data prefetching [2, 9, 14]. With respect to pipelining, our *instruction* prefetches differ in two important ways from data prefetches: (i) the pipeline stage in which the prefetch address is known, and (ii) the computational resources consumed by the prefetches. Figure 3(b) contrasts the pipeline for data prefetches in the MIPS R10000 [14] with the pipeline for our instruction prefetches in an equivalent machine. As we see in Figure 3(b), the prefetch address of a `pf_d` instruction prefetch is known immediately after the *Decode* stage (`pf_c` type prefetches would require some additional time), while the address for a data prefetch is not

[3] Node d of a flow graph dominates node n if every path from the initial node of the flow graph to n goes through d [1].

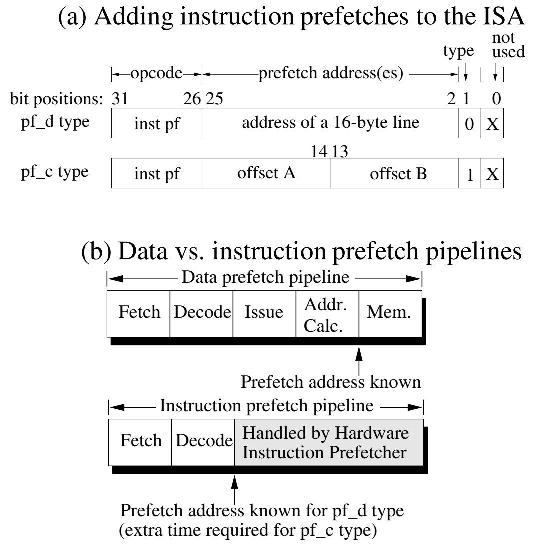

Figure 3. Possible extensions to the ISA and the CPU pipeline for instruction prefetches.

known until it is computed in the *Address Calculate* stage. Hence a `pf_d` instruction prefetch can be initiated two cycles earlier than a data prefetch. In addition, since instruction prefetches do not go through the latter three pipeline stages of a data prefetch (instead they are handled directly by the hardware instruction prefetcher after they are decoded), they do not contend for processor resources including functional units, the reorder buffer, register file, etc. In effect, the instruction prefetches are removed from the instruction stream as soon as they are decoded, thereby having minimal impact on most computational resources.

3.3. Extensions to the Memory Subsystem: Prefetch Filtering

We add two components to the memory subsystem to support instruction prefetching: the *I-prefetcher* and the *prefetch filter*. The I-prefetcher generates both hardware- and software-initiated prefetch addresses and first checks whether they are already present in the primary instruction cache (the "I-cache"); if not—and if a prefetch for the same line is not already outstanding—then the prefetch is forwarded on to the prefetch filter.

The *prefetch filter* sits between the I-prefetcher and the L2 cache to reduce the number of useless prefetches. In addition, a *prefetch bit* is associated with each line in the I-cache to remember whether the line was prefetched but not yet used, and a two-bit saturating counter value is associated with each line in the L2 cache to record the number of *consecutive* times that the line was prefetched but not used before it was replaced. The prefetch filtering mechanism works as follows. When a line is *fetched* from the L2 cache to the I-cache, both the prefetch bit and the saturating

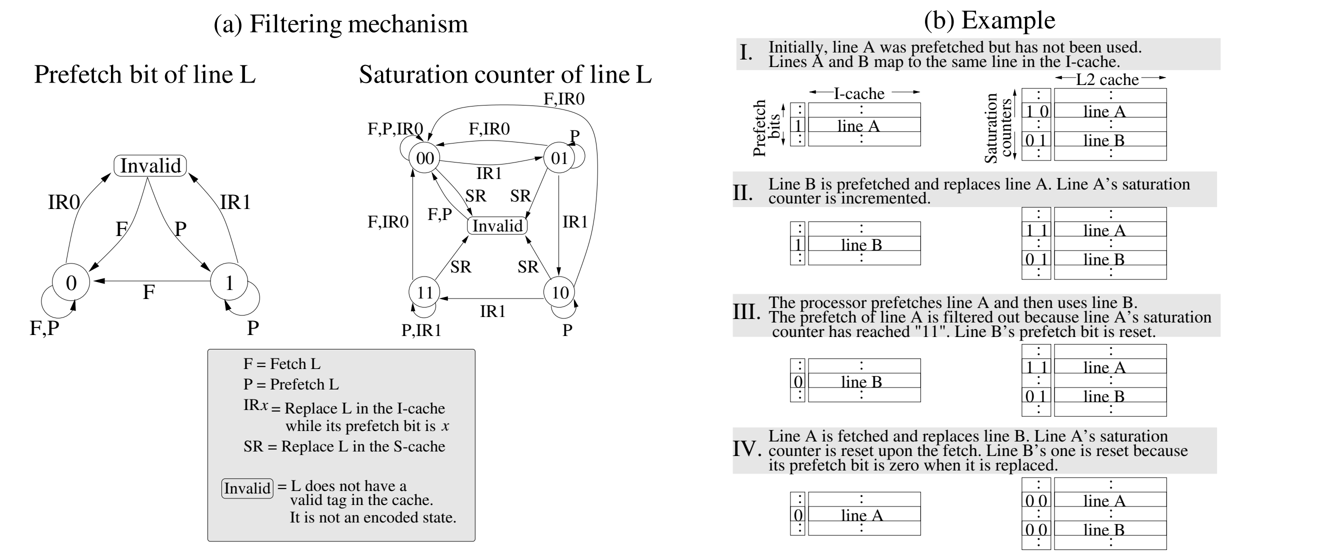

Figure 4. Prefetch filtering mechanism and example.

```
void schedule_prefetches(E) {
    foreach basic block B in the executable E do
        schedule(B, B, 0, {});
}

// Consider attaching a prefetch for T to B where:
//   B = current basic block, T = prefetch-target basic block,
//   D = the prefetching distance between B and T
//   S = set of basic blocks scheduled so far
//   SCHED_DIST = prefetch-scheduling distance
//   N = the N used in hardware next-N-line prefetching
void schedule(B, T, D, S) {
    if (B ∉ S) { // continue only if B hasn't been scheduled
        S = S ∪ {B};
        // Attach a prefetch if it is sufficiently early and
        // if it is necessary.
        if ((D ≥ SCHED_DIST)
            and not locality_likely(B, T)
            and not nextNline_prefetchable(B, T, N)
            and not prefetch_already_exists(B, T)) {
                attach_prefetch(B, T);
        }
        foreach basic block P which can reach B
            in a single direct control transfer do {
            // update prefetching distance conservatively
            D' = D + min_length(P);
            schedule(P, T, D', S);
        }
    }
}

boolean locality_likely(B, T) {
    // If B and T are in the same loop or recursive procedure
    // chain that accesses a very small volume of instructions
    // relative to the I-cache size, it is likely that T is already
    // in the I-cache when we are executing B. In this case,
    // we return TRUE; otherwise return FALSE.
}

boolean nextNline_prefetchable(B, T, N) {
    // Determine whether T is within N cache lines of B.
}

boolean prefetch_already_exists(B, T) {
    // Check whether a prefetch for T is already attached to B.
}

void attach_prefetch(B, T) {
    // Insert a prefetch of T before the first instruction in B.
}

int min_length(B) {
    // Return the number of instructions executed in basic block
    // B. If B doesn't end with a procedure call, this is simply the
    // number of instructions in B; otherwise, this is the number
    // of instructions in B plus the length of the shortest path
    // from the beginning to the end of the procedure called by B.
}
```

Figure 5. Pseudo-code representation of our prefetch scheduling compiler algorithm.

counter value are reset to zero. (Although another possibility is to *decrement* the saturation counter upon a normal fetch, our experiments showed that this performs worse than resetting the counter to zero.) When a line is *prefetched* from the L2 cache to the I-cache, its prefetch bit is *set* to one and its saturation counter does not change. When a prefetched line is actually used by a fetch, its prefetch bit is *reset* to zero. When a prefetched line l in the I-cache is replaced by another line, then if the prefetch bit of line l is set, its saturation counter is incremented (unless it has already saturated, of course); otherwise, the counter is reset to zero. When the prefetch filter receives a prefetch request for line

Table 2. Prefetch optimization passes.

Order	Purpose	Description
1	Combining prefetches at dominators	Boosts prefetches that have been attached to a basic block b in the prefetch scheduling phase to b's nearest dominator (other than b itself) if the boosting is not harmful (it is harmful when the boosted prefetches will displace other useful instructions from the cache before b is referenced). After this boosting process, the compiler could combine some prefetches at dominators. For example, Figure 6(b) shows the result of combining the two prefetches of line `y` into one after boosting prefetches from basic blocks `D`, `E`, and `F` into their dominator `C`.
2	Eliminating unnecessary prefetches	A prefetch targeting a line l is *unnecessary* if l resides in the I-cache on *all* possible paths reaching the prefetch instruction. To eliminate unnecessary prefetch instructions, the compiler estimates which lines reside in the I-cache at each prefetch instruction using an algorithm similar to the one for computing *available expressions* in classical code optimization [1]. In our case, the *gen* set of a basic block b is the set of lines fetched or prefetched by b while the *kill* set is the set of lines displaced by b. In our example, since line `z` will definitely be in the I-cache when we enter basic block `C` regardless of whether we came from `A` or `B`, the prefetch of line `z` in `C` is unnecessary and therefore is eliminated, as shown in Figure 6(c).
3	Compressing prefetches	The compiler checks whether multiple `pf_d` prefetches in the same basic block can be compressed into a single compact prefetch. For each basic block b, the compiler needs to compute the offsets between the starting address of b and the target addresses of all `pf_d` prefetches scheduled in b. It then attempts to fit these offsets into a minimum number of compact prefetch instructions. Our example assumes that the address offsets of both lines `x` and `y` are representable within 12 bits, and therefore the two `pf_d` prefetches in `C` are compressed into a single `pf_c` prefetch, as shown in Figure 6(d).
4	Hoisting prefetches	Finally, the compiler hoists prefetches scheduled inside a loop up to the nearest basic block that dominates but is not part of the loop, if the prefetches do not need to be re-executed at every iteration (which may not be the case if each iteration can access a large volume of instructions). In some cases, a *pre-header* block will be created for the loop to hold the hoisted prefetches. For example, in Figure 6(e), a pre-header `C'` is created to immediately precede the header (i.e. `C`) of the loop containing `C`, `D`, `E`, and `F` to hold the hoisted `pf_c` prefetch. While this optimization does not reduce the code size, it can reduce the number of *dynamic* prefetches.

l, it will either respond normally if the counter value is below a threshold T, or else it will drop the prefetch and send a "prefetch canceled" signal to the processor if the counter has reached T (in our experiments, $T = 3$). Figure 4(a) summarizes the filtering mechanism in two finite automata, and Figure 4(b) is an example of prefetch filtering.

4. Compiler Support

The compiler is responsible for automatically inserting prefetch instructions into the executable. Note that since prefetch insertion is most effective if it begins after the code is otherwise in its final form, this new pass occurs fairly late in the compilation: perhaps at link time, or in our case, we implemented it as a binary rewrite tool. The goal of the compiler is to schedule prefetches to achieve high coverage and satisfactory prefetching distances while at the same time minimizing the static and dynamic instruction overhead. Hence our compiler algorithm has two major phases: *prefetch scheduling* and *prefetch optimization*. Figure 5 shows a pseudo-code representation of our prefetch scheduling algorithm.

After generating an initial prefetch schedule, the compiler then performs the four optimization passes described in Table 2 and illustrated through the running example in Figure 6. We used a complete implementation of this algorithm in our experiments, and further details on the algorithm can be found in a technical report [6].

5. Experimental Framework

We performed our experiments on seven non-numeric applications which were chosen because their relatively

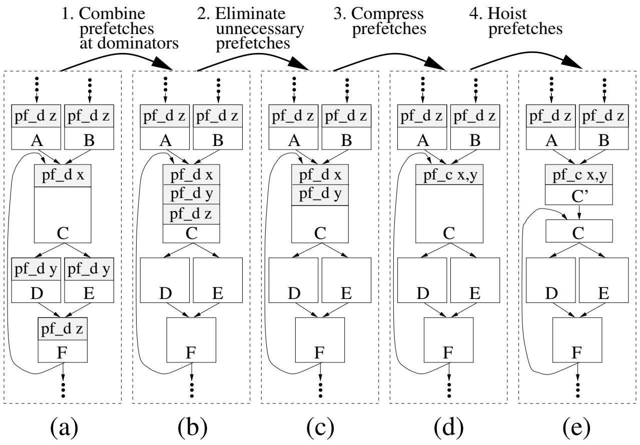

Figure 6. Example of prefetch optimization. `A` to `F` are basic blocks; `x`, `y` and `z` are cache line addresses. `C` is a dominator of `D`, `E`, `F`, and `C` itself. Part (a) is the initial schedule, and part (e) is the final optimized schedule.

large instruction footprints result in poor instruction cache performance. These applications are described Table 3, and all of them were run to completion. Table 3 also shows the number of software prefetches inserted into the executable, broken down into the interprocedural and intraprocedural cases. As we see in Table 3, the software component of our scheme mainly targets interprocedural prefetching.

We performed detailed cycle-by-cycle simulations of our applications on a dynamically-scheduled, superscalar processor similar to the MIPS R10000 [14]. Our simulator models the rich details of the processor including the pipeline, register renaming, the reorder buffer, branch pre-

Table 3. Benchmark characteristics. Note: the "combined" miss rate is the fraction of instruction fetches which suffer misses in both the 32KB I-cache *and* the 1MB L2 cache. The "static prefetch count" is the number of prefetches inserted, normalized to the size of the original executable. Prefetches are classified as either *interprocedural* or *intraprocedural*, depending on whether the prefetch target and the prefetch itself are in the same procedure.

Name	Description	Input Data Set	Instructions Graduated	Miss Rate		Static Prefetch Count (% of Original Executable Size)	
				I-Cache	Combined	Interprocedural	Intraprocedural
Gcc	The GNU C compiler drawn from SPEC92	The stmt.i in the reference input set	136.0M	2.63%	0.10%	6.3%	1.6%
Perl	The interpreter of the Perl language drawn from SPEC95	A Perl script called a2ps.pl which converts ascii to postscript	41.4M	5.03%	0.06%	8.3 %	1.7 %
Porky	A SUIF compiler pass for simplifying and rearranging codes	The compress95.c program in SPEC95 (default optimizations)	86.8M	2.38%	0.06%	7.1%	0.4%
Postgres	The PostgreSQL database management system [15]	A subset of queries in the Postgres Wisconsin benchmark	46.0M	3.76%	0.16%	8.3%	0.6%
Skweel	A SUIF compiler pass for loop parallelization	A program that computes Simplex (all optimizations)	68.1M	2.22%	0.06%	7.5 %	0.6%
Tcl	An interpreter of the script language Tcl version 7.6	Tcltags.tcl which makes Emacs-style TAGS file for Tcl source	37.5M	2.78%	0.02%	7.2%	1.0%
Vortex	The Vortex object-oriented database program drawn from SPEC95	A reduced SPEC95 input set	193.0M	6.48%	0.08%	10.3%	0.7%

Table 4. Simulation parameters for the baseline architecture.

Pipeline Parameters	
Fetch & Decode Width	8 aligned sequential instructions
Issue & Graduate Width	4
Functional Units	2 Integer, 2 FP, 2 Memory, 2 Branch
Reorder Buffer Size	32
Integer Multiply	12 cycles
Integer Divide	76 cycles
All Other Integer	1 cycle
FP Divide	15 cycles
FP Square Root	20 cycles
All Other FP	2 cycles
Branch Prediction Scheme	2-bit Counters

Memory Parameters	
Line Size	32B
I-Cache	32KB, 2-way set-associative, 4 banks
Inst. Prefetch Buffer	16 entries
D-Cache	32KB, 2-way set-associative, 4 banks
Victim Buffers	8 entries each for data and inst.
Miss Handlers (MSHRS)	32 each for data and inst.
Unified S-Cache	1MB, 4-way set-associative
Primary-to-Secondary Miss Latency	12 cycles (plus any delays due to contention)
Primary-to-Memory Miss Latency	75 cycles (plus any delays due to contention)
Primary-to-Secondary Bandwidth	32 bytes/cycle
Secondary-to-Memory Bandwidth	8 bytes/cycle

diction, branching penalties, speculative instruction fetching (including incorrect execution paths), the memory hierarchy (including tag, bank, and bus contention), etc. Table 4 shows the parameters used in our model for the bulk of our experiments (we vary the latency and bandwidth later in Section 6.5). As shown in Table 4, we enhanced the memory subsystem in a few ways relative to the R10000 to provide better support for instruction prefetching—e.g., we added an eight-entry victim cache [4] and a 16-entry prefetch buffer [3].

We compiled each application as a "nonshared" executable with `-O2` optimization using the standard MIPS C compilers under IRIX 5.3. We implemented our compiler algorithm as a standalone pass which reads in the MIPS executable and modifies the binary. However, since we did not have access to a complete set of binary rewrite utilities, we tightly integrated our compiler pass with our simulator so that rather than physically generating a new executable, we instead pass a logical representation of the new binary to the simulator which it can then model accurately. For example, the simulator fetches and executes all of the new instruction prefetches as though they were in a real binary, and it remaps all instruction layouts and addresses to correspond to what they would be in the modified binary. Hence we truly emulate the physical insertion of prefetches at the expense of decreased simulation speed.

6. Experimental Results

We now present results from our simulation studies. We start by evaluating the overall performance of our cooperative prefetching scheme. Next, we examine the relative importance of the two key components of our scheme: prefetch filtering and compiler-inserted prefetching. We also quantify the impact of our compiler's prefetch optimizations, and of varying the prefetch-scheduling distance parameter, on the code size and performance. We then explore the impact of varying cache latencies and bandwidths on the performance of our scheme. Finally, we evaluate whether cooperative prefetching is cost effective.

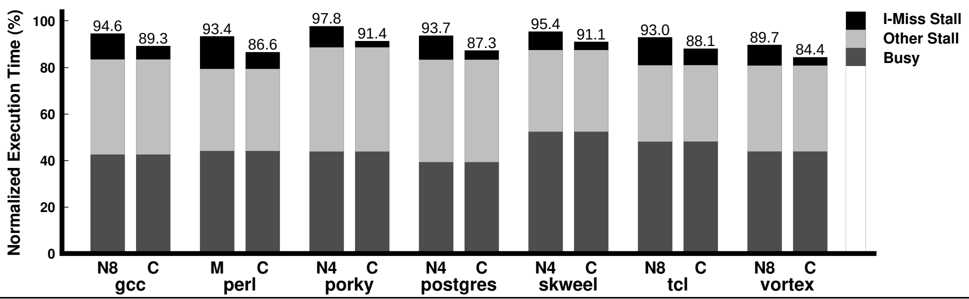

Figure 7. Performance comparison of cooperative prefetching and the best performing existing schemes of individual applications (Nx = next-x-line prefetching, M = Markov prefetching, C = cooperative prefetching).

6.1. Performance of Cooperative Prefetching

Our cooperative prefetching scheme includes compiler-inserted `pf_d` and `pf_c` prefetches, hardware-based next-8-line prefetching, and prefetch filtering. A prefetch-scheduling distance of 20 instructions is used for all applications. (We will discuss the impact of the prefetch-scheduling distance more later in Section 6.4.)

Figure 7 shows the performance impact of cooperative instruction prefetching. For each application, we show two cases: the bar on the left is the best previously-existing prefetching scheme (seen earlier in Figure 1), and the bar on the right is cooperative prefetching (**C**). Note that the number of instructions that actually graduate (i.e. the *busy* section) is equal in both cases because instruction prefetches are removed from the instruction stream once they are decoded (see Section 3.2). As we see in Figure 7, our cooperative prefetching scheme offers significant speedups over existing schemes (6.4% on average) by hiding a substantially larger fraction of the original instruction cache miss stall times (71% on average, as opposed to an average reduction of 36% for the best existing schemes).

To understand the performance results in greater depth, Figure 8 shows a metric which allows us to evaluate the coverage, timeliness, and usefulness of prefetches all on a single axis. This figure shows the total I-cache misses (including both fetch and prefetch misses) normalized to the original case (i.e. without prefetching) and broken down into the following four categories. The bottom section is the number of fetch misses that were not prefetched (this accounts for 100% of the misses in the original case, of course). The next section (*Late Prefetched Misses*) is where a miss has been prefetched, but the prefetched line has not returned in time to fully hide the miss (in which case the instruction fetcher stalls until the prefetched line returns, rather than generating a new miss request). The *Prefetched Hits* section is the most desirable case, where a prefetch fully hides the latency of what would normally have been a fetch miss, converting it into a hit. Finally, the top section is useless prefetches which bring lines into the cache that are not accessed before they are replaced.

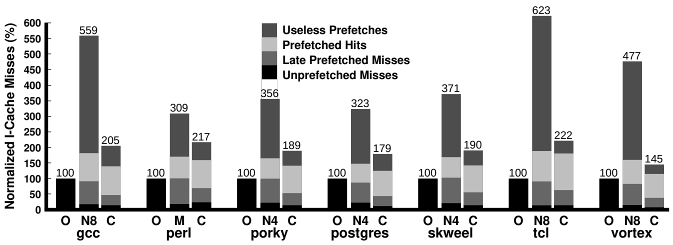

Figure 8. Breakdown of all I-cache misses. (O = original, Nx = next-x-line prefetching, M = Markov prefetching, C = cooperative prefetching).

Figure 8 shows that both cooperative prefetching and the best existing prefetching schemes achieve large coverage factors, as indicated by the small number of unprefetched misses. The main advantage of our scheme is that it is more effective at launching prefetches early enough. This is demonstrated in Figure 8 by the significant reduction in *late prefetched misses*, the bulk of which have been converted into *prefetched hits*. We also observe in Figure 8 that both cooperative prefetching and existing schemes experience a certain amount of *cache pollution* since the sum of the bottom three sections of the bars adds up to over 100%. However, the *prefetch filtering* mechanism used by cooperative prefetching helps to reduce this problem, thereby resulting in a smaller total for the bottom three sections than the best existing scheme in all of our applications. In addition, Figure 8 shows another benefit of prefetch filtering: it dramatically reduces the number of useless prefetches. The reduction in total useless prefetches ranges from 2.4 in `perl` to 10.6 in `tcl`—on average, cooperative prefetching has achieved a sixfold reduction in useless prefetching.

6.2. Importance of Prefetch Filtering and Software Prefetching

Two components of the cooperative prefetching design contribute to its performance advantages: prefetch filtering and compiler-inserted software prefetching. To isolate the contributions of each component, Figure 9 shows their performance individually as well as in combination. The

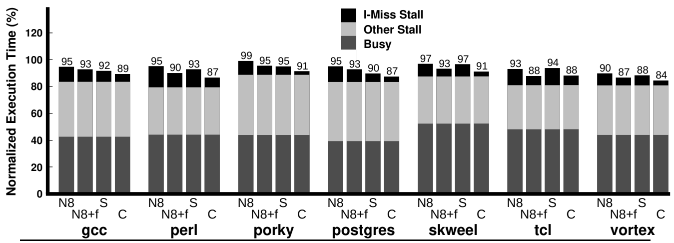

Figure 9. Performance of four different combinations of prefetch filtering and compiler-inserted prefetching (N8 = next-8-line prefetching alone, N8+f = next-8-line prefetching with prefetch filtering, S = compiler-inserted prefetching alone without prefetch filtering, C = cooperative prefetching).

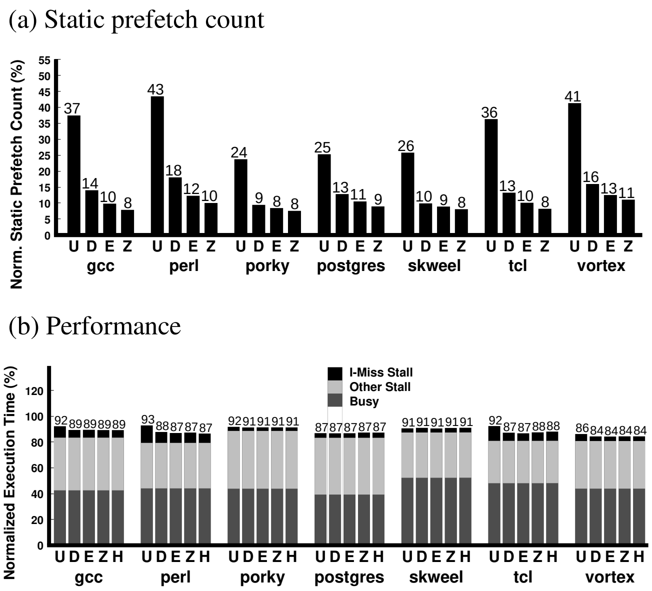

Figure 10. Impact of prefetch optimization on (a) the static prefetch count and (b) the performance of cooperative prefetching. (U = unoptimized, D = combining prefetches at dominators, E = case D plus eliminating unnecessary prefetches, Z = case E plus compressing prefetches, H = case Z plus hoisting prefetches.) The y-axis of (a) is normalized to the number of instructions in the original executable.

relative importance of prefetch filtering versus compiler-inserted prefetching varies across the applications: in `tcl`, prefetching filtering is more important, and in `postgres`, compiler-inserted prefetching is more important. In all cases, the best performance is achieved when both techniques are combined, and in all but one case this results in a significant speedup over either technique alone. Intuitively, the reason for this is that the benefits of prefetch filtering (i.e. avoiding cache pollution) and software prefetching (i.e. issuing non-sequential prefetches early enough) are *orthogonal*. Hence both components of our design are clearly important for performance and are complementary in nature.

6.3. Impact of Prefetching Optimizations

To evaluate the effectiveness of the compiler optimizations discussed earlier in Section 4 in reducing the number of prefetches, we measured their impact both on code size and performance. Figure 10(a) shows the number of static prefetches remaining as each optimization pass is applied incrementally, normalized to the original code size. Without any optimization (**U**), the code size can increase by over 40%. Combining prefetches at dominators (**D**) dramatically reduces the prefetch count by more than half in all applications except `postgres`. Eliminating unnecessary prefetches and compressing prefetches further reduces the prefetch count by a moderate amount. (Prefetch hoisting has no effect on the static prefetch count, and therefore is not shown in Figure 10(a).) Altogether, the prefetch optimizations limit the prefetch count to only 9% of the original code size on average.

Figure 10(b) shows the impact of these optimizations on performance. As we see in this figure, combining prefetches at dominators results in a noticeable performance improvement in several cases (e.g., `gcc`, `perl`, and `tcl`). The other optimizations have a negligible performance impact. In fact, prefetch compression and hoisting sometimes degrade performance by a very small amount by changing the order in which prefetches are launched.

6.4. Varying the Prefetch-Scheduling Distance

A key parameter in our prefetch scheduling compiler algorithm is the *prefetch-scheduling distance* (i.e. $SCHED\_DIST$ in Figure 5). When choosing a value for this parameter, we must consider the following trade-offs: we would like the parameter to be large enough to hide the expected miss latency, but setting the parameter too high can increase the code size (since more prefetches must be inserted to cover a larger number of unique incoming paths) and increase the likelihood of polluting the cache. In our experiments so far, we have used a prefetch-scheduling distance of 20 instructions, which is roughly equal to the product of the expected IPC ($\sim$1.6) and the primary-to-secondary miss latency ($\geq$ 12 cycles). To determine the sensitivity of cooperative prefetching to this parameter, we varied the prefetch-scheduling distance across a range of five values from 12 to 28 instructions, and measured the resulting impact on both code size and performance (shown in Figures 11(a) and 11(b), respectively).

As we observe in Figure 11(a), increasing the prefetch-scheduling distance can result in a noticeable increase in the code size. Fortunately, even with a prefetch-scheduling distance as large as 28 instructions, the compiler is still able to limit the code expansion to less that 11% on average, due

Table 5. Impact of latency and bandwidth variations on prefetching performance. Each table entry contains the average speedup relative to the original code, along with the average fraction of the original *I-Miss Stall* time that has been eliminated in parentheses.

Scheme	Baseline (Latency = 12 cycles, Bandwidth = 32B/cycle)	Latency Variations (Bandwidth = 32B/cycle)		Bandwidth Variations (Latency = 12 cycles)		
		6 cycles	24 cycles	8B/cycle	16B/cycle	Unlimited
Best Existing	6.5 % (36.0%)	1.7% (16.7%)	14.2% (42.8%)	5.1% (24.3 %)	6.0% (30.8 %)	7.3% (39.4%)
Cooperative	13.3% (71.0%)	5.8% (59.6%)	24.4% (69.6%)	11.9% (53.4 %)	12.5% (63.2%)	13.7% (72.7%)
Perfect I-cache	20.0% (100.0%)	10.2% (100.0%)	39.2% (100.0%)	25.1% (100.0%)	21.8% (100.0%)	20.1% (100.0%)

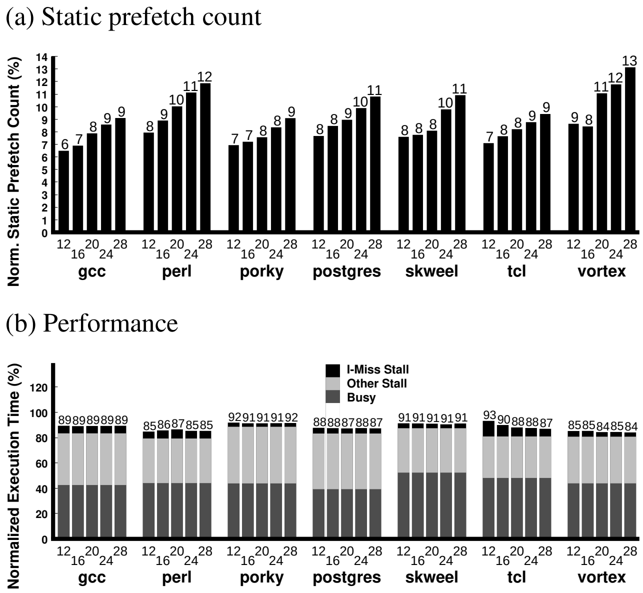

Figure 11. Impact of the prefetch-scheduling distance on (a) the static prefetch count and (b) the performance of cooperative prefetching. (x = a prefetch-scheduling distance of x instructions is used in the compiler scheduling; the case 20 is the default for our basic cooperative prefetching.) The y-axis of (a) is normalized to the number of instructions in the original executable.

to the optimizations discussed in the previous section. In contrast, the *performance* offered by cooperative prefetching is less sensitive to the prefetch-scheduling distance, as we see in Figure 11(b). While `tcl` enjoys a 6% speedup as we increase this parameter from 12 to 28 cycles, the other applications experience no more than a 2% fluctuation in performance across this range of values. Hence we observe that performance is not overly sensitive to this parameter.

6.5. Impact of Latency and Bandwidth Variations

We now consider the impact of varying miss latencies and available bandwidth between the primary and secondary caches on the performance of cooperative prefetching. Recall that in our experiments so far, the primary-to-secondary miss latency has been 12 cycles (plus any delays due to contention) and the primary-to-secondary cache bandwidth has been 32 bytes/cycle. Table 5 shows the impact of varying these parameters. Starting with the middle two columns in the table, we see the performance of the best performing existing schemes and cooperative prefetching when the primary-to-secondary latency is decreased to 6 cycles and increased to 24 cycles. (Note that the compiler's prefetch-scheduling distance was set to 12 and 28 instructions, respectively, for the 6-cycle and 24-cycle cases.) As we see in Table 5, cooperative prefetching still performs well under both latencies, and results in even larger improvements as the latency grows. In the 24-cycle case, cooperative prefetching results in an average speedup of 24.4%, which is significantly larger than the 14.2% speedup offered by the best existing scheme.

Turning our attention to bandwidth, the rightmost three columns in Table 5 show the impact of decreasing the primary-to-secondary cache bandwidth from 32 bytes/cycle to 8 and 16 bytes/cycle, and of increasing it to unlimited bandwidth. There are two things to note from these results. First, we see in Table 5 that while reducing the bandwidth does degrade the performance of cooperative prefetching somewhat—from an average speedup of 13.3% to 11.9%—the overall performance gain still remains high. Hence cooperative prefetching can achieve good performance with realistic amounts of bandwidth. (Note that this bandwidth includes servicing data cache misses as well.) Second, we observe in Table 5 that *increasing* the bandwidth beyond 32 bytes/cycle does not significantly improve the performance of cooperative prefetching (the average speedup only increases from 13.3% to 13.7%). Therefore cooperative prefetching is not bandwidth-limited, and it is more likely that it is limited by other factors (e.g., cache pollution, achieving a sufficient prefetching distance, etc.).

6.6. Cost Effectiveness

Having demonstrated the performance advantages of cooperative prefetching, we now focus on whether the additional hardware support is cost effective. One alternative to cooperative prefetching would be to simply increase the cache sizes by a comparable amount. (Note that this is overly simplistic since the primary cache sizes are often limited more by access time than the amount of silicon area available.) For our baseline architecture, the additional stor-

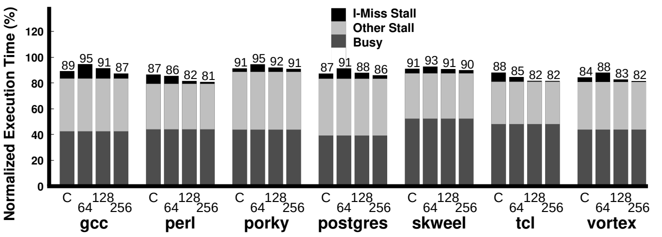

Figure 12. Performance comparison of cooperative prefetching and larger I-caches (C = a 32 KB I-cache with cooperative prefetching, x = an x KB I-cache without prefetching). The y-axis is normalized to the execution time of a 32 KB I-cache without prefetching.

age necessary to support basic cooperative prefetching is 640 bytes at the level of the primary I-cache (128 bytes for the prefetch bits used by prefetch filtering, and 512 bytes for the prefetch buffer), and 8 KB for the 2-bit saturating counters added to the L2 cache.

Figure 12 compares the performance of a 32 KB I-cache with cooperative prefetching with that of three larger I-caches, ranging from 64 KB to 256 KB, without prefetching. It is encouraging that the average speedup achieved by cooperative prefetching (13.3%) is greater than that obtained by doubling the cache size from 32 KB to 64 KB (10.8%) despite of the substantially higher hardware cost of the larger cache. In addition, cooperative prefetching outperforms the 128 KB I-cache in three of the seven applications, and is within 2% of the performance with a 256 KB I-cache in five cases. Overall, cooperative prefetching appears to be a more cost-effective method of improving performance than simply increasing the I-cache size.

7. Conclusions

To overcome the disappointing performance of existing instruction prefetching schemes on modern microprocessors, we have proposed and evaluated a new prefetching scheme whereby the hardware and software cooperate as follows: the hardware performs aggressive next-N-line prefetching combined with a novel *prefetch filtering* mechanism to get far ahead on sequential accesses without polluting the cache, and the compiler uses a novel algorithm to insert explicit *instruction-prefetch instructions* into the executable to prefetch non-sequential accesses. Our experimental results demonstrate that our scheme significantly outperforms existing schemes, eliminating 50% or more of the latency that had remained with the best existing scheme. This reduction in latency translates into a 13.3% average speedup over the original execution time on a state-of-the-art superscalar processor, which is more than double the 6.5% speedup achieved by the best existing scheme, and much closer to the maximum 20% speedup (for these applications and this architecture) in the ideal instruction prefetching case. These improvements are the result of launching prefetches earlier (thereby hiding more latency), while at the same time reducing the cache-polluting effects of useless prefetches dramatically. Given these encouraging results, we advocate that future microprocessors provide instruction-prefetch instructions along with the prefetch filtering mechanism.

8. Acknowledgments

We thank Earl Killian for his many helpful suggestions and comments throughout this work. Chi-Keung Luk is partially supported by a Canadian Commonwealth Fellowship. Todd C. Mowry is partially supported by a Faculty Development Award from IBM.

References

[1] A. V. Aho, R. Sethi, and J. D. Ullman. *Compilers: Principles, Techniques and Tools*. Addison Wesley, 1986.

[2] D. Bernstein, D. Cohen, A. Freund, and D. E. Maydan. Compiler techniques for data prefetching on the PowerPC. In *PACT'95*, June 1995.

[3] D. Joseph and D. Grunwald. Prefetching using markov predictors. In *ISCA'97*, pages 252–263, June 1997.

[4] N. P. Jouppi. Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers. In *ISCA'90*, pages 364–373, May 1990.

[5] G. Kane and J. Heinrich. *MIPS RISC Architecture*. Prentice Hall, 1992.

[6] C.-K. Luk and T. C. Mowry. Compiler and hardware support for automatic instruction prefetching: A cooperative approach. Technical Report CMU-CS-98-140, Carnegie Mellon University, June 1998.

[7] A. Maynard, C. Donnelly, and B. Olszewski. Contrasting characteristics and cache performance of technical and multi-user commerical workloads. In *ASPLOS-VI*, pages 145–156, October 1994.

[8] J. Pierce and T. Mudge. Wrong-path prefetching. In *MICRO-29*, pages 264–273, Dec. 1996.

[9] V. Santhanam, E. Gornish, and W.-C. Hsu. Data prefetching on the HP PA8000. In *ISCA'97*, June 1997.

[10] A. Smith. Sequential program prefetching in memory hierarchies. *IEEE Computer*, 11(2):7–21, 1978.

[11] A. Smith. Cache memories. *Computing Surveys*, 14(3):473–530, Sept. 1982.

[12] J. Smith and W.-C. Hsu. Prefetching in supercomputer instruction caches. In *Supercomputing'92*, pages 588–597, 1992.

[13] C. Xia and J. Torrellas. Instruction prefetching of system codes with layout optimized for reduced cache misses. In *ISCA'96*, pages 271–282, June 1996.

[14] K. Yeager. The MIPS R10000 superscalar microprocessor. *IEEE Micro*, April 1996.

[15] A. Yu and J. Chen. *The Postgres95 User Manuel v1.0*. University of California at Berkeley, Sept 1996.