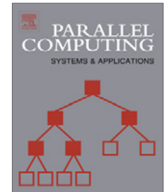
Parallel Computing 38 (2012) 533–551

Contents lists available at SciVerse ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco



Algorithm-level Feedback-controlled Adaptive data prefetcher: Accelerating data access for high-performance processors

Yong Chen [a,*], Huaiyu Zhu [b], Hui Jin [c], Xian-He Sun [c]

[a] Department of Computer Science, Texas Tech University, United States
[b] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, United States
[c] Department of Computer Science, Illinois Institute of Technology, United States

ARTICLE INFO

Article history:
Received 21 August 2010
Received in revised form 10 April 2012
Accepted 18 June 2012
Available online 29 June 2012

Keywords:
Data prefetching
Adaptive prefetching
Data access acceleration
Application accelerator
High-performance processors
Memory hierarchy
Memory wall

ABSTRACT

The rapid advance of processor architectures such as the emerged multicore architectures and the substantially increased computing capability on chip have put more pressure on the sluggish memory systems than ever. In the meantime, many applications become more and more data intensive. Data-access delay, not the processor speed, becomes the leading performance bottleneck of high-performance computing. Data prefetching is an effective solution to accelerating applications' data access and bridging the growing gap between computing speed and data-access speed. Existing works of prefetching, however, are very conservative in general, due to the computing power consumption concern of the past. They suffer low effectiveness especially when applications' access pattern changes. In this study, we propose an Algorithm-level Feedback-controlled Adaptive (AFA) data prefetcher to address these issues. The AFA prefetcher is based on the Data-Access History Cache, a hardware structure that is specifically designed for data access acceleration. It provides an algorithm-level adaptation and is capable of dynamically adapting to appropriate prefetching algorithms at runtime. We have conducted extensive simulation testing with the SimpleScalar simulator to validate the design and to analyze the performance gain. The simulation results show that the AFA prefetcher is effective and achieves considerable IPC (Instructions Per Cycle) improvement for 21 representative SPEC-CPU benchmarks.
© 2012 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid advance of semiconductor process technology and the evolvement of micro-architecture, the processor cycle times have been significantly reduced in the past decades. However, compared to the processor performance improvement, especially the aggregated processor performance of multicore/manycore architectures, data-access performance (latency and bandwidth) improvement has been at snail's pace. The memory speed has only increased by roughly 9% each year over the past two decades, which is significantly lower than the improvement speed of nearly 50% per year for processor performance [18]. This performance disparity between processor and memory is predicted to continually expand in next decades [18]. The unbalanced performance improvement leads to one of the significant performance bottlenecks in high-performance computing known as *memory-wall* problem [24,39]. The reason behind this huge processor-memory performance disparity is several folds. First, most advanced architectural and organizational efforts are focused on processor technology, instead of memory storage devices. Second, drastically improving semiconductor technology results in much smaller and more transistors to be built on chip for processing units and thus can achieve a high computational capability.

* Corresponding author.
E-mail addresses: yong.chen@ttu.edu (Y. Chen), hzhu10@illinois.edu (H. Zhu), hjin6@iit.edu (H. Jin), sun@iit.edu (X.-H. Sun).

0167-8191/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.parco.2012.06.002

Third, the primary technology improvement for memory device focuses on higher density, which results in much larger memory capacity, but the bandwidth and the latency are improved slowly. Multi-level cache hierarchy architectures have been the primary solution to avoiding large performance loss due to long memory-access delays. However, cache memories are designed based on data access locality. When applications lack data access locality due to non-contiguous data accesses, multi-level cache memory hierarchy does not work well.

Data prefetching is a common technique to accelerate data access and reduce the processor stall time on data accesses, and has been widely recognized as a critical companion technique of memory hierarchy solution to overcome the data access bottleneck issue [24,39,18]. Several series of commercial high-performance processors have adopted data prefetching techniques to hide long data-access latency [13,18,23]. As the term indicates, the essential idea of data prefetching is to observe data referencing patterns, then speculate future references and fetch the predicted reference data closer to processor before processor demands them. By overlapping computation and data accesses, data prefetching can overcome the limitations of cache memories, reduce long memory access latency, and speed up data-access performance. Numerous studies have been conducted and many strategies have been proposed for data prefetching [5,6,10,11,13,21,23,25,31,35,37].

Data prefetcher, as a data-access accelerator, has been adopted in production and found to be very effective for applications with regular data-access patterns, such as data streaming [23,13]. But in general, current prefetching techniques are very limited. Software solutions are too slow for cache level prefetching; whereas hardware prefetchers are static in nature and cannot change with the data-access patterns of the applications. Previous studies show that there is no enough effort to support hardware dynamic adaptation among different strategies depending on the runtime application characteristics [4]. The fact is that the effectiveness of data prefetching is application and data-access pattern dependent. There is no single universal prefetching algorithm suitable for all applications at runtime. A general and effective data prefetcher and accelerator must be dynamic in nature.

In this research, we propose an *Algorithm-level Feedback-controlled Adaptive* data prefetcher (*AFA prefetcher* in short) to provide a dynamic and adaptive prefetching methodology based on the recently proposed generic prefetching structure, Data-Access History Cache (DAHC) [6], and runtime feedback collected by hardware counters. DAHC is a new cache structure designed for data prefetching and data-access acceleration. It is capable of effectively tracking data-access history and maintaining correlations of both data-access address stream and instruction address stream. It can be used for efficient implementation of many data prefetching algorithms [6]. Hardware counters gain considerable attention in recent years and are becoming more and more important for improving the performance of contemporary processor architectures, operating system and applications [30,36,40]. This study explores this trend and assumes hardware counters are available within a data prefetcher that resides at the lowest cache level. Based on DAHC and available hardware counters, the AFA data prefetcher is able to recognize distinct data-access patterns and adapts to the corresponding appropriate prefetching algorithms at runtime. This adaptation methodology is significantly better than conventional static prefetching strategies. It improves the prefetching effectiveness, which in turn improves the overall performance. The AFA prefetcher does consume some transistors. However, the hardware chip space and the number of transistors integrated on chip are not limitations for current processor architectures. Trading chip space for lower data-access latency is a current trend [24,39]. The AFA prefetcher and DAHC follow the trend. A preliminary version of this research was published in [7], with extended background studies, design details, simulation details, evaluation tests, and related work discussions in this paper.

The contribution of this work is several folds:

First, we demonstrate that different prefetching algorithms exhibit substantial variation in performance improvement for diverse applications. The performance variance is largely due to distinct access patterns that different applications exhibit.
Second, we argue that the rapid advance of semiconductor technology and the trend of integrating considerable amount of hardware counters on chip provide us an opportunity to explore dynamic and adaptive strategy that can produce better performance improvement for various applications on average.
Third, we propose an algorithm-level adaptive prefetcher, named AFA prefetcher, that is able to dynamically adapt to well-performing algorithms at runtime based on a three-tuple evaluation metric. These three tuples are orthogonal and complementary to each other. We adopt innovative mechanisms to keep the hardware cost of the proposed mechanism low.
Fourth, we carry out extensive simulation testing to verify the proposed design and to evaluate the performance improvement. We also vary different simulation configurations and conduct sensitivity analysis of the proposed AFA prefetcher.
Last, to our knowledge, this study is the first work exploring algorithm-level dynamic adaptation to accelerate data access depending on applications' access pattern. Such an approach is promising and has a great potential of accelerating data accesses for high-performance processors such as multicore processors. We hope this study can bring dynamic data-access acceleration into community's attention and inspire more research efforts on accelerating applications' data-access performance.

The rest of this paper is organized as follows. Section 2 briefly reviews the DAHC structure. Section 3 presents the proposed AFA prefetcher design and discusses implementation related issues. Section 4 presents the simulation environment and simulation results. Section 5 discusses related work, and finally Section 6 concludes this study.

2. Data-Access History Cache

To fully exploit the benefits of data prefetching and to focus on accelerating data-access performance to achieve a high *sustained performance* instead of building extensive compute units to achieve a high *peak performance*, we proposed a generic prefetching-dedicated cache structure, named *Data-Access History Cache* (*DAHC*) [6]. The DAHC serves as a fundamental structure dedicated to data prefetching and data-access acceleration. The DAHC behaves as a cache for recent reference information instead of as a traditional cache for either instructions or data. Theoretically, it is capable of supporting any history-based prefetching algorithms.

The design rationale of DAHC is that history-based prefetching algorithms must rely on correlations among either instruction address stream or data address stream, or both. Thus, DAHC is designed to have three hardware tables: one data-access history table (DAH table) and two index tables, program counter index table (PIT) and address index table (AIT). The DAH table accommodates history details, while PIT and AIT maintain correlations from instruction and data address stream viewpoints respectively. Various prefetching algorithms thus can access these two tables to obtain the required correlation as necessary [6]. Fig. 1 illustrates the general design of DAHC and a high-level view of how it can be applied to support various prefetching algorithms. For instance, the traditional stride prefetching [5,13] can be supported by retrieving access history via PIT and DAH tables, while the Markov prefetching [21] can be supported by accessing AIT and DAH tables.

Fig. 2 illustrates the detailed structure of DAHC through an example. The DAH table consists of PC (program counter), chain_PC, Addr, chain_Addr and State fields. PC and Addr fields store the instruction address and data address separately. The chain_PC and chain_Addr point to an entry where the last access from the same instruction or the last access of the same address is located. Therefore, chain_PC and chain_Addr connect all accesses from the instruction stream and data stream perspectives. This design offers the fundamental mechanism to detect potential correlations and access patterns. The State field maintains state machine status used in prefetching algorithms. The PIT table has two fields, PC and Index. The PC field represents the instruction address, which is a unique index in this table. The Index field records the entry of the latest data access in the DAH table from the instruction stored in the correspondent PC field. It is the connection between the PIT and the DAH tables. The address index table is similarly defined. For instance, in Fig. 2, the DAH table captures four data accesses,

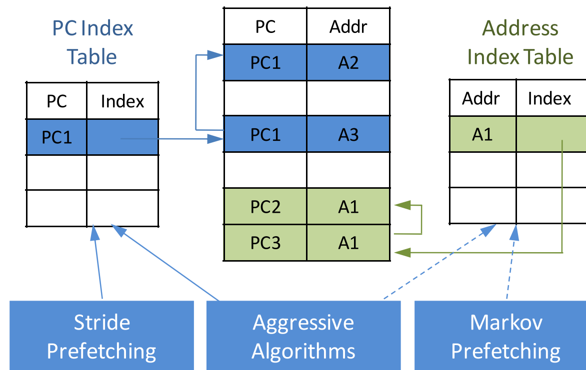


Fig. 1. DAHC general design and a high-level view. DAHC is designed to have three hardware tables: one data-access history table (DAH table) and two index tables, program counter index table (PIT) and address index table (AIT). The DAH table accommodates history details, while PIT and AIT maintain correlations from instruction and data address stream viewpoints respectively.

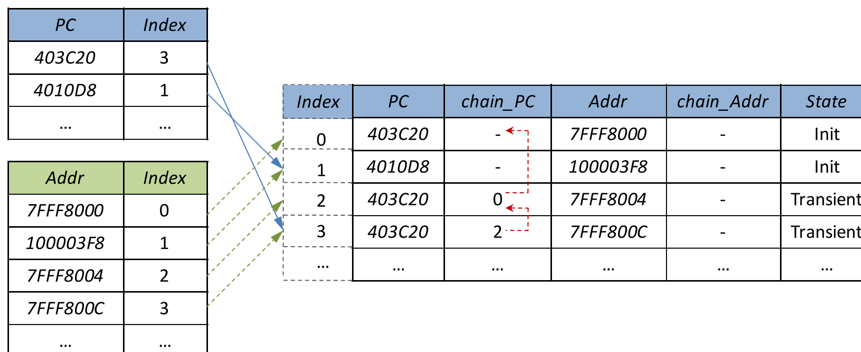


Fig. 2. DAHC design: PIT, AIT and DAH tables. The DAH table consists of PC (program counter), chain_PC, Addr, chain_Addr and State fields. The PIT table consists of PC and Index fields, and the AIT table consists of Addr and Index fields.

three of them issued by instruction 403C20 (stored in the PC field) and one by instruction 4010D8. The instruction 403C20 accesses data at addresses 7FFF8000, 7FFF8004 and 7FFF800C in sequence, which is shown through the Addr and chain_PC fields. The instruction 403C20 and 4010D8 are also stored in the PIT table, and the corresponding index field tracks the latest access from the DAH table, which are entry 3 and 1 respectively. The AIT table keeps each accessed address and the latest entry, as shown in the bottom left of the figure, thus connecting all the data accesses on the basis of the address stream.

We take the well-known stride prefetching [5,13] as an example to illustrate how DAHC can support various prefetching algorithms. Stride prefetching predicts future accesses based on strides of recent references [5]. This approach monitors data accesses and detects constant stride access patterns. It maintains a state machine with four states, initialization, transient, prediction and no-prediction, to track the detection of stride patterns from a specific instruction address and trigger prefetching when the state machine enters the prediction state. The conventional stride prefetching can be implemented with the DAHC as follows. First, when a miss occurs, the instruction address of this miss is searched in the PIT. If this instruction address does not match any entry in the PIT, which means it is the first time that we encounter this instruction address in current working window, no prefetching action is triggered. If the instruction address matches one entry (and only one entry as entries are unique), we follow the index chain to check whether a constant stride pattern is present. If such a pattern exists, one or more data blocks are prefetched depending on the configuration of prefetching degree and prefetching distance [6]. Note that since DAHC tracks a much larger window than the conventional stride prefetching implementation Reference Prediction Table (RPT) does, we are able to detect complex stride patterns such as structured patterns [6]. The detailed methodology of supporting stride prefetching and other prefetching algorithms through DAHC can be found in [6].

The DAHC provides a prototype design of a prefetching-dedicated structure. It works as a cache for data access information compared with conventional cache for either instructions or data. The DAHC can be placed at different memory hierarchy levels for various desired data prefetching. For instance, it can be used to track all accesses to first level cache and to serve as an L1 cache prefetcher. It can also be placed at the second level cache and to serves as an L2 cache prefetcher only. The straightforward design makes the implementation uncomplicated. The implementation of the DAHC is a specialized physical cache, like victim cache [20] or trace cache [28]. The PIT and AIT tables can be implemented with any associativity such as 2-way or 4-way. Since the index tables usually have less valid entries than the DAH table, it is unlikely that some entry is replaced due to a conflict miss [6]. Even if a conflict miss occurs, it does not affect the correctness except discarding certain access history. The DAH table can be implemented with a special structure where each entry can be located by using its index. The logic to fill/update the DAHC comes from the cache controller. The cache controller traps data accesses at the monitored level and keeps a copy of the access information in the DAHC. Note that DAHC design is general and it does not imply any restriction to the system environment. It works in a Chip Multiprocessor (CMP) or simultaneous multi-threading (SMT) environment, as well as in an environment where multiple applications are running concurrently.

3. Algorithm-level Feedback-controlled Adaptive data prefetcher

In this section, we present the design of an Algorithm-level Feedback-controlled Adaptive data prefetcher. This proposed AFA prefetcher leverages the powerful functionality provided by DAHC, supports multiple prefetching algorithms and dynamically adapts to those algorithms that perform well at runtime. The essential idea is using runtime feedback and evaluation to direct the dynamic adaptation. We first present the motivation of this work, and then discuss the evaluation metrics, the implementation and the methodology of directing adaptation.

3.1. Motivation

We have performed a simulation testing with five representative SPEC-CPU2000 benchmarks [42] including 179.art, 256.bzip2, 254.gap, 181.mcf and 171.swim with four different prefetching algorithms including stream [23], strided [5], Markov [21] and MLDT prefetching [35]. The simulation details are presented in Section 4. The reported Instructions Per Cycle (IPC) results are plotted in Fig. 3. The performance improvement of different prefetching algorithms exhibits significant variation for different benchmarks/applications as the results demonstrate. For instance, the strided prefetching achieves 29.9% IPC speedup for the art benchmark, while the Markov prefetching only achieves 0.92% improvement for the same benchmark. However, the Markov prefetching achieves 14.2% IPC speedup for the bzip2 benchmark, while the strided prefetching merely achieves 8.13% in this case. For another instance, the stream prefetching achieves over 18% IPC improvement for the gap benchmark, while the Markov prefetching even generates negative performance impact. The observations of these simulation testing clearly demonstrate that different algorithms are suitable for different workloads. There is a great need to support algorithm-level adaptation at runtime to direct the prefetcher to choose proper algorithms. Motivated by these observations, we propose an AFA prefetcher to adapt prefetching algorithms. The remaining subsections in this section present the design and implementation of the AFA prefetcher based on DAHC and feedbacks collected at runtime.

3.2. Evaluation metrics

While extensive studies exist in prefetching, few studies present a formalized metric to evaluate the effectiveness of prefetching algorithms. We analyze and sort out the essential and most critical criteria to model prefetching evaluation and

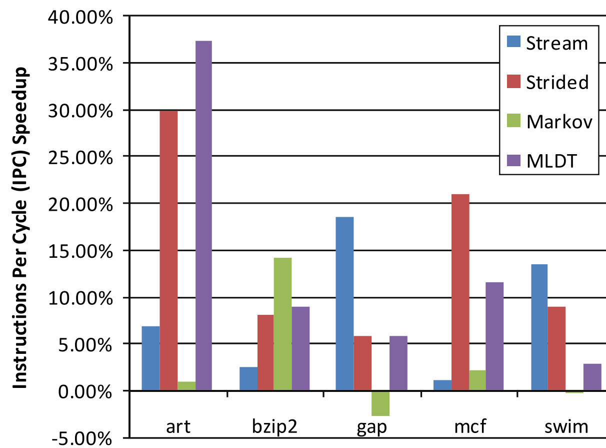


Fig. 3. IPC results of art, bzip2, gap, mcf and swim benchmarks with data prefetching. The performance improvement of different prefetching algorithms exhibits significant variation for different benchmarks. For instance, the strided prefetching achieves 29.9% IPC speedup for the art benchmark, while the Markov prefetching only achieves 0.92% improvement for the same benchmark. However, the Markov prefetching achieves 14.2% IPC speedup for the bzip2 benchmark, while the strided prefetching merely achieves 8.13% in this case.

present a formal definition in this study. These metrics provide a comprehensive evaluation of hardware prefetching methodologies. These metrics can be independently used to evaluate a prefetcher in addition to commonly used metrics, such as an IPC speedup metric.

3.2.1. Prefetch precision

The first and widely-adopted metric is termed prefetching precision or prefetching accuracy. This metric characterizes the percentage of prefetches that are actually accessed by demand requests, thus reflecting how accurate the prefetch requests and the prefetching algorithms are. We present a formal definition of prefetching precision as follows.

Definition 1. Prefetching precision is defined as the ratio between the number of distinct prefetched cache lines that are accessed by at least one demand request after being prefetched in and before being replaced out over the number of total prefetched cache lines.

By this definition, the prefetching precision models the *accuracy* of the prefetcher, i.e. the percent of *useful* cache lines in the *overall* cache lines prefetched. Note that we define a useful prefetch as being accessed at least once when the prefetched cache line resides in the prefetch destination. Therefore, a repeated access to that cache line in the current lifecycle does not account into the total number of useful prefetches. However, if a cache line is prefetched again into the destination after being displaced, and gets hit, this scenario will contribute to one useful prefetch. We refer these cache lines brought in by prefetch and accessed by demand requests as *prefetch hits*, in contrast with *demand hits*, those lines fetched by demand and hit again by other requests.

The prefetch precision should be considered as the most critical metric to evaluate or direct prefetch adaptation, as it describes the cost-efficiency of a prefetcher well. Taking prefetch precision into consideration, an aggressive but not accurate enough prefetcher should be largely avoided because it might produce a large number of useless prefetches, which significantly wastes resources, such as power and cache line slots. Instead, this metric favors a prefetcher with high confidence. The prefetch precision metric suggests that the prefetcher should focus on identifying the correct access pattern and make a highly-accurate prediction. Such an approach maximizes the hardware investment on the prefetcher and achieves a high cost-efficiency. An ideal prefetcher will produce a prefetching precision with value 1. In practice, the prefetching precision has a range from 0 to 1.

Though the prefetch precision is critical and straightforward, it merely describes one aspect of the problem under study – the prefetch precision does not quantify how effective the prefetcher is, i.e. how many misses among the overall misses are hided. The next metric we formalize addresses this limitation.

3.2.2. Prefetch coverage

The prefetch coverage metric is introduced to complement prefetch precision and quantify the other aspect of how well a prefetcher works. We formalize the prefetch coverage definition as follows.

Definition 2. Prefetch coverage is defined as the ratio of the number of misses reduced due to prefetches over the total number of misses that will occur without prefetching.

As the definition states, the prefetch coverage focuses on quantifying the ratio of the misses reduced, i.e. how wide a prefetcher covers the demand misses that are supposed to occur without the assistance of prefetching. A highly-accurate prefetcher does not necessarily provide a wide coverage. This is because such a prefetcher could be very conservative and takes

action only when the prefetcher has a high-confidence prediction. This conservativeness results in a high precision but the effectiveness in terms of miss reduction ratio is low. The vice versa holds as well, i.e. a prefetcher with wide coverage is not necessarily highly accurate since the prefetcher could be very aggressive (such as with a large prefetch degree) to improve the coverage while sacrificing the precision. In essence, the prefetch coverage and prefetch precision are complementary to each other, and together they quantify the effectiveness of a prefetcher from two aspects.

3.2.3. Prefetch pollution

While prefetch precision and prefetch coverage can reflect the prefetch effectiveness, or the positive side, of a prefetching algorithm well, they do not characterize the negative side of an algorithm. Cache pollution [37] is considered a critical down side of prefetching. When a cache line that is replaced by a prefetched line is later accessed by a demand request, cache pollution occurs. Such a cache miss will not happen without the interference of prefetching. This scenario, cache pollution, is referred as a negative side-effect of prefetching. We present a formal definition to describe prefetch pollution.

Definition 3. Prefetch pollution is defined as a ratio of the number of additional demand misses caused by prefetching that will not occur without prefetch interference over the number of misses that will occur without prefetching.

According to this definition, prefetch pollution quantifies the percent of extra demand misses due to prefetches, which means that those demand misses will not occur if prefetching is not adopted. The occurrence of these misses is due to the limited cache size and the replacement of useful cache lines by prefetched cache lines. Together with prefetch precision and prefetch coverage, the prefetch pollution completes a three-tuple (*precision*, *coverage*, *pollution*) or *PCP* in short, to evaluate a prefetching algorithm. These three metrics are complementary to each other, and assess an algorithm from both positive and negative aspects. Some literatures separate other metrics, such as lateness [34]. We observed that these metrics are well covered in the 3-tuple PCP metric. A separation of these additional metrics might be helpful, but might also cause confusion. In this study, the proposed dynamic adaptation is based on the 3-tuple PCP metric.

Note that these prefetch metrics not only consider whether a prefetched cache line will be used in the future or not, but also whether it will be used in a timely fashion or not. For instance, if a prefetched cache line is displaced before it is used either by a demand request or a more recently prefetch request, it is not a useful prefetch and is not counted as an accurate prefetch. The metrics introduced evaluate the effectiveness of prefetches from both important aspects, what to prefetch and when to prefetch, for a given prefetching algorithm.

3.3. Evaluation metrics: on-the-road

We have presented the formal definitions of a PCP metric to evaluate and direct a prefetcher in the previous section. We discuss the hardware design and realization of these metrics in the proposed AFA prefetcher in this section.

3.3.1. Realizing prefetch precision metric

To realize the prefetch precision metric, the AFA prefetcher utilizes two statistics counters for each evaluated algorithm, one counter for the prefetch hits, and one counter for overall prefetches. We refer these two counters as prefetch_hits and prefetch_total. In addition, to collect the statistic of prefetch hits, we need to distinguish the cache lines prefetched from demanded. This requirement results in a major hardware storage budget. We assume each cache line in the prefetch destination (L2 cache in this study) has one extra prefetch bit to represent whether this line is prefetched or fetched for each evaluated algorithm. When a cache line is prefetched into destination, this prefetch bit is set. If this cache line is ever accessed during its lifetime in the cache (after being prefetched and before being displaced), the prefetch_hits counter is increased and the prefetch bit is reset. By this way, the prefetched cache line is not counted as multiple hits even when it is accessed multiple times, which is consistent with the definition. A simple reason behind this decision is that the first hit acts like a regular demand request and fetches in data, and the future accesses will hit in cache. The actual saving of the prefetching is the first access. If a cache line is brought in by a normal demand request, the prefetch bit is not set. The combinatorial logic to maintain this prefetch bit, set and reset, is trivival, and the hardware implementation of this logic is not complicated. Note that if a cache line is prefetched by multiple algorithms, the corresponding prefetch bits will be set and a hit of this cache line will attribute to the statistics of each corresponding prefetcher.

3.3.2. Realizing prefetch coverage metric

The prefetch_hits counter discussed above can also be used in calculating the prefetch coverage. This is because that the statistics the prefetch_hits counter collects is the number of misses reduced due to prefetches. To compute the prefetch coverage, the AFA prefetcher needs another counter – the number of overall misses that will occur without prefetching. The AFA prefetcher utilizes another acounter, demand_misses counter, to collect the number of misses that occurs even with prefetching. The prefetch_hits counter represents the number of misses saved by prefetch, and the demand_misses counter represents the number of misses that still occur. The prefetch coverage is computed as:

$$prefetch_{coverpage} = \frac{prefetch_{hits}}{prefetch_{hits} + demand_{misses}}$$

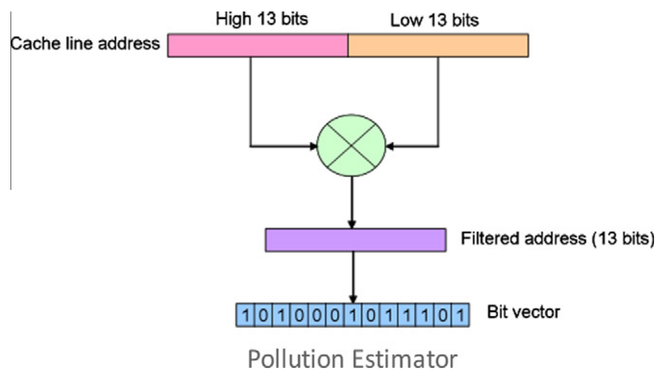


Fig. 4. Pollution estimator. The pollution estimator splits 26 bits of cache line address into two parts, high-order 13 bits and low-order 13 bits. These two parts are fed into an XOR logical unit, and a filtered address with 13 bits, is the output. This filtered address is used to index a bit vector and set the corresponding bit in the vector as indicating this cache line is replaced out due to a prefetch.

3.3.3. Realizing prefetch pollution metric

It is more challenging to collect the prefetch pollution statistics than to collect prefetch precision and coverage. The reason is that we never know whether the replaced cache line due to a prefetch will be used in future or not. An optimal solution to collecting the prefetch pollution metric is tracing down all these cache lines and detecting whether any future requests will access these lines. If such a cache line is detected, which means that this cache line is replaced out due to a prefetch but is needed by a demand request, a case of cache pollution is detected. However, such an optimal solution will require infinite-size storage to keep all past cache lines replaced out due to prefetches. This approach is not feasible in practice. Motivated from existing studies [1,26,34], the AFA prefetcher utilizes a Bloom filter [1] to estimate the percentage of cache pollution as shown in Fig. 4.

Suppose the cache line size is 64 B, and a cache block address is 26 bits. The pollution filter splits 26 bits into two parts, high-order 13 bits and low-order 13 bits. These two parts are fed into an XOR logical unit, and a filtered address, with 13 bits, is the output. This filtered address is used to index a bit vector and set the corresponding bit in the vector. The AFA prefetcher uses this filter to estimate the pollution. It tracks each cache line that is replaced out due to a prefetch and feeds this cache line address into the filter. A corresponding bit of the bit vector is set. It also feeds cache miss addresses into the filter, and if the corresponding bit is set, the AFA prefetcher estimates this cache line was in the cache but was replaced out due to a prefetch. After a cache pollution is detected, the corresponding bit is reset, as the cache line is fetched back into cache. The AFA prefetcher uses a pollution counter for each evaluated algorithm to accumulate the prefetch pollution statistics. The hardware cost of Bloom filters is discussed in Section III-F.

Note that the Bloom filter based pollution estimator provides a tradeoff between the accuracy of detecting prefetch pollution and the space requirement. An ideal solution of detecting every prefetch pollution will require infinite storage capacity; whereas the pollution estimator only requires limited and a small amount of storage (2^{13} bits of storage for each supported algorithm as discussed in Section 3.6). The downside of the pollution estimator is that it does not always guarantee an accurate detection. The previous studies, however, have shown that the Bloom filter based estimator provides a satisfactory accuracy for estimation [26,34]. In this study, it is adopted as an important mechanism to control the hardware cost of collecting prefetch pollution statistics. It has been verified effective through simulation tests.

3.4. Metrics collection

The AFA prefetcher periodically collects the PCP metric discussed above in order to make adaptation decision. The adaptive prefetching mechanism is designed to have two phases, *metrics collecting phase* and *stably prefetching phase*. In the collecting phase, all supported prefetching algorithms are enabled, and the statistics of each prefetching algorithm are tracked and collected. In the end of collecting phase, the PCP metric is computed for prefetching algorithm evaluation. In the stably prefetching phase, only the adaptively selected algorithm will be running, and all counters and pollution estimator are cleared and turned off. The decision to choose the working algorithm is discussed in the following subsection.

The switch between these two phases is controlled by a *phase timer*. There are many potential ways for measuring the time and providing the phase timer, such as utilizing CPU cycles, issued instructions, issued load/store instructions or load/store misses. We choose the approach that views each cache miss as one time tick and accumulates to measure the time and provides the phase control. This is a feasible design choice to control the adaptive prefetcher behavior because the number of cache misses can fairly represent how the prefetcher should react. In addition, this design is considered better than utilizing cycles or issued instructions, as those numbers could be huge and increase rapidly. Utilizing cache misses as time ticks is a much simpler way for the AFA prefetcher. The AFA prefetcher is empirically configured with a collecting phase as 1/8th of the stably prefetching phase, which means it collects statistics and makes adaptive selection decision in one unit of time, and prefetch with selected algorithms for eight units of time.

3.5. Adaptive selection

After collecting statistics and computing the metrics, the AFA prefetcher makes the decision to adaptively select the suitable prefetching algorithms. The decision is based on the evaluation of the performance of each prefetching algorithm, indicated by the 3-tuple metric, the precision, coverage and pollution. This evaluation is not complicated – simply done by comparing the runtime statistics against a preset threshold and classifying them as either high or low. If it is above the threshold, the prefetcher classifies the statistics as high. In contrast, if it is below the threshold, the prefetcher classifies it as low. In our simulation experiments, the threshold to distinguish a high/low prefetch precision, coverage and pollution are preset as 0.7, 0.3, and 0.2 respectively based on empirical experience. In practice, these thresholds can be measured and determined in advance for any specific architecture. It can also be tuned dynamically at runtime. Fig. 5 illustrates a table of eight levels of prefetching algorithm performance that can be recognized by the prefetcher.

The rationale of the arrangement of different levels in the performance table is rooted from the desired effectiveness of prefetchers, i.e. prefetching algorithms should remain low intrusion to the system, and strive for accuracy and coverage. In other words, the placement of these different levels is essentially driven by the desired QoS (Quality-of-Service) of prefetching algorithms. When a certain level of pollution is controlled, the AFA prefetcher favors accuracy over coverage because accurate prefetches are more effective than aggressive prefetches that provide wide coverage. Note that the thresholds that control the selections can be dynamically tuned to provide a fine control.

Based on the prefetching algorithm evaluation performance table, the AFA prefetcher is able to identify and choose best prefetching algorithms dynamically. In our current study, we propose and analyze two different mechanisms, best-strategy adaptive selection and multi-strategy adaptive selection, to select the algorithms adaptively.

The best-strategy adaptive selection always outputs the one performing best in the statistics collecting phase. This decision is made based on the algorithm evaluation and the performance table – the lowest level is assigned to have the highest priority. Within the same level, the precision is assigned to have the highest preference, while the coverage has the medium and the pollution metric has the lowest preference. This means that if the AFA prefetcher sees multiple algorithms falling into multiple levels, the prefetcher chooses the lowest level algorithms as the candidate. If the prefetcher detects multiple candidates sharing the same lowest level, it favors the one with the highest precision.

The best-strategy adaptation works by choosing the best strategy according to the defined policy (performance table and preference assignment) out of all supported algorithms. However, a potential limitation is that it only chooses one algorithm even if multiple strategies are performing well and can sometimes complement each other. In addition, this strategy always outputs one "relatively best" strategy, even though the best strategy might not work well enough at certain circumstances. Based on these observations, we introduce another adaptation strategy, multi-strategy adaptive selection, which chooses multiple optimal strategies according to the evaluation and performance table. This adaptation strategy uses the level as the selection criteria. For instance, if level 0 and level 1 are configured as the adaptation criteria, then the AFA prefetcher dynamically chooses all of these algorithms that fall into these levels, and use them in the stably prefetching phase. This adaptive strategy can also control the quality of the selection. If none of the algorithms satisfies the specified criteria, the prefetcher does not have any algorithm performing in the prefetching phase, until the algorithms are evaluated again in the next collecting phase. The selection criteria are preset, for instance, as level 0, 1 and 2 in our current simulation experiments.

3.6. Hardware cost

As discussed in previous subsections, the AFA prefetcher needs three counters for each evaluated algorithm and two counters for all algorithms to collect the required statistics in order to direct the adaptation. Each counter can be imple-

	Precision	Coverage	Pollution
Level 0	H	H	L
Level 1	H	L	L
Level 2	L	H	L
Level 3	L	L	L
Level 4	H	H	H
Level 5	L	H	H
Level 6	H	L	H
Level 7	L	L	H

Fig. 5. Prefetching algorithm performance table. The AFA prefetcher compares the runtime statistics against a preset threshold and classifies each of the 3-tuple metric, precision, coverage and pollution as either high or low. Based on the evaluation of the 3-tuple metric, each prefetching algorithm is classified into one of eight levels defined in this table.

mented with a 32-bit register. The overall required storage for counters would be 56 bytes, if assuming to support four distinct algorithms simultaneously. In addition to the counters, the AFA prefetcher needs one cache pollution estimator for each supported algorithm. The pollution estimator requires 2^{13} bits of storage for the bit vector. Therefore, the estimator consumes 1 kB for each supported algorithm. The cache structure needs a slight modification to support adaptive selection of prefetching algorithm as discussed previously. The modification is one bit for each supported algorithm. For a typical 1 MB L2 cache with 64 bytes cache line, it has 16,384 cache lines. To support one prefetching algorithm, the additional hardware cost will be 16,384 bits or 2 kB. Therefore, as a normal case, to support adaptation among four algorithms, the overall hardware cost will be around 12 kB. This hardware budget is trivial – as only around 1% compared to a regular 1 MB L2 cache. However, as the simulation verifies, the adaptive prefetching can considerably reduce cache misses and improve the overall system performance.

As discussed partially in the previous section, the combinatorial logic to realize the proposed adaptive prefetching is not complicated as well. The major required combinatorial logic resides in maintaining the prefetch bits within the cache line, maintaining statistics counters, filtering through prefetch pollution estimator, and adapting prefetching algorithms via the performance table. Maintaining prefetch bits is straightforward because it merely requires set/unset the corresponding bit according to whether a cache line is brought in due to a specific algorithm. Maintaining counters is a straightforward logic too. Filtering is slightly complicated, but as we show with the estimator, the hardware state machine can be described effortlessly. Adapting the algorithm mainly needs comparison logic, which can be implemented easily as well.

4. Simulation and performance analysis

We have carried out simulation experiments to study the feasibility of the proposed AFA prefetcher and analyze the potential performance impact. Stream prefetching [11,23], stride prefetching [5,13], Markov prefetching [21] and MLDT prefetching [35] algorithms were selected for simulation. The adaptive strategy of the AFA prefetcher is independent of the underlying prefetching algorithms. In theory, any prefetching algorithm can be supported by the AFA prefetcher. The stream, stride, Markov, and MLDT prefetching algorithms are chosen to evaluate the adaptive mechanism because these algorithms are classic, well known, and also widely used. The evaluation with other algorithms such as the PC/DC prefetching algorithm [25] and stream chaining algorithm [12] is a planned further study. This section discusses simulation details and presents the analytical results.

4.1. Simulation methodology

We have enhanced SimpleScalar simulator [3] with the DAHC [6] and the AFA prefetcher for the simulation verification and analysis. SimpleScalar tool set provides a detailed and high-performance simulation of modern processors. It takes binaries compiled for SimpleScalar architecture as input and simulates their execution on provided processor simulators. It has several different execution-driven processor simulators, ranging from extremely fast functional simulator to a detailed and out-of-order issue simulator [3].

We have chosen the most detailed, *sim-outorder simulator*, for our experiments. Fig. 6 shows the enhanced SimpleScalar simulator architecture. We have added two primary modules, DAHC module and AFA prefetcher module. The DAHC module simulates the functionality of the DAHC, as explained in [6]. The AFA prefetcher module implements the adaptive prefetching logic and four supported prefetching algorithms, stream, stride, Markov and MLDT. We have modified the cache line structure for identifying whether a cache line is brought in due to a prefetch, and which algorithm brings it in. We have created the required counters and the pollution estimator to simulate the evaluation as well. The combinatorial logic was simulated to compare the evaluation of each algorithm, and outputs the dynamically chosen suitable algorithms, with both best-strategy and multi-strategy selections. The multi-strategy selection was configured as level 0, level 1 and level 2 algorithms.

The AFA prefetcher was simulated to have an additional prefetch queue to store the prefetch requests. When the load/store issuing bandwidth is available after issuing demand requests, the prefetcher starts issuing the requests from the prefetch queue. If a newly generated prefetch request is already in the prefetch queue, the new request is simply dropped. If the prefetch queue is full, the new requests replace the old requests. The handling of prefetch requests is similar to the handling of regular load requests, with a slight difference that the effective address is computed based on prefetching algorithm, and any exceptions/faults generated by prefetches are discarded and the previous states are restored. The AFA prefetcher uses two groups of miss status handling registers (MSHRs) to track all on-the-fly memory access requests. One of them is used to filter the redundant misses from the stream and to hold outgoing requests issued. The other one holds outgoing requests issued by the AFA prefetcher.

4.2. Simulation setup

We used the Alpha-ISA and configured the simulator as a 4-way issue and 256-entry RUU processor. The instruction cache and data cache were split. L1 data cache was configured as 32 kB 2-way with 64 B cache line size. The latency was 2 cycles. L2 data cache was configured as 1 MB 4-way with 64 B cache line size. The latency of L2 cache was 12 CPU cycles. DAHC was configured with 1024 entries. We assume each DAHC access costs one CPU cycle. This is a reasonable assumption for a fairly

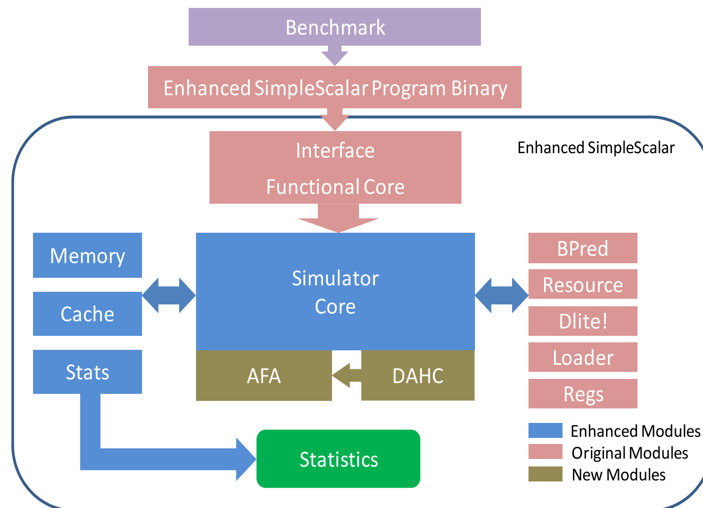


Fig. 6. Simulation architecture of AFA prefetcher. Two primary modules, DAHC module and AFA prefetcher module, were added into the existing SimpleScalar simulator. The DAHC module simulates the functionality of the DAHC, and the AFA prefetcher module implements the adaptive prefetching logic and four supported prefetching algorithms, stream, stride, Markov and MLDT.

small cache and is also commonly used in estimating the access latency of a hardware structure in existing studies [25,27]. The AFA prefetch queue was configured with 512 entries. The stream prefetching algorithm was configured to support four streams. The prefetch degree for all prefetching algorithms was configured as eight. The prefetch distance for strided, Markov and MLDT was configured as four. The memory access latency and bandwidth are assumed to be 120 cycles and 10 cycles/access respectively. The metrics collecting phase was configured as 32K cycles, the stably prefetching phase was configured as 256K cycles. Table 1 lists the configuration of the simulator in our simulation tests.

4.3. Simulation results

We have performed a series of simulation for performance evaluation with SPEC-CPU2000 benchmarks [42]. We fast forwarded the first 100M instructions and simulated the following 200M instructions to analyze the result. This evaluation setting (or a slight variation thereof) is widely used for evaluating architectural enhancements. Twenty-one out of total 26 benchmarks were tested successfully in our experiments. We have excluded the other five benchmarks (apsi, facerec, fma3d, perlbmk and wupwise) that had problems and did not finish the test.

4.3.1. Cache miss rate reduction

We first study the cache miss rate reduction with various prefetching algorithms and the AFA prefetcher. Fig. 7 plots the L2 cache miss rate reported by the simulator for the entire twenty-one benchmarks. This series of tests were conducted under eight cases, including the base case (without data prefetching), the cases with individual stream, strided, Markov and MLDT data prefetching, the cases with best-strategy and multi-strategy adaptation and the case with all supported prefetching algorithms running simultaneously.

As clearly shown from the results, different applications exhibit distinct access patterns, and thus the cache miss rate reduction of various prefetching algorithms have large variations. For instance, stream prefetching significantly reduced the misses for equake, gap, mgrid and swim benchmarks, while not for others like ammp, art, galgel and gcc. Instead, the

Table 1
Simulator configuration.

Issue width	4
Load store queue	64 entries
RUU size	256 entries
L1 D-cache	32 kB, 2-way set associative, 64 byte line, 2 cycle hit time
L1 I-cache	32 kB, 2-way set associative, 64 byte line, 1 cycle hit time
L2 Unified-cache	1 MB, 4-way set associative, 64 byte line, 12 cycle hit time
Block size	64 B
DAHC	1024 entries
AFA queue	512 entries
Memory latency	120 cycles
Memory bandwidth	10 cycles/access

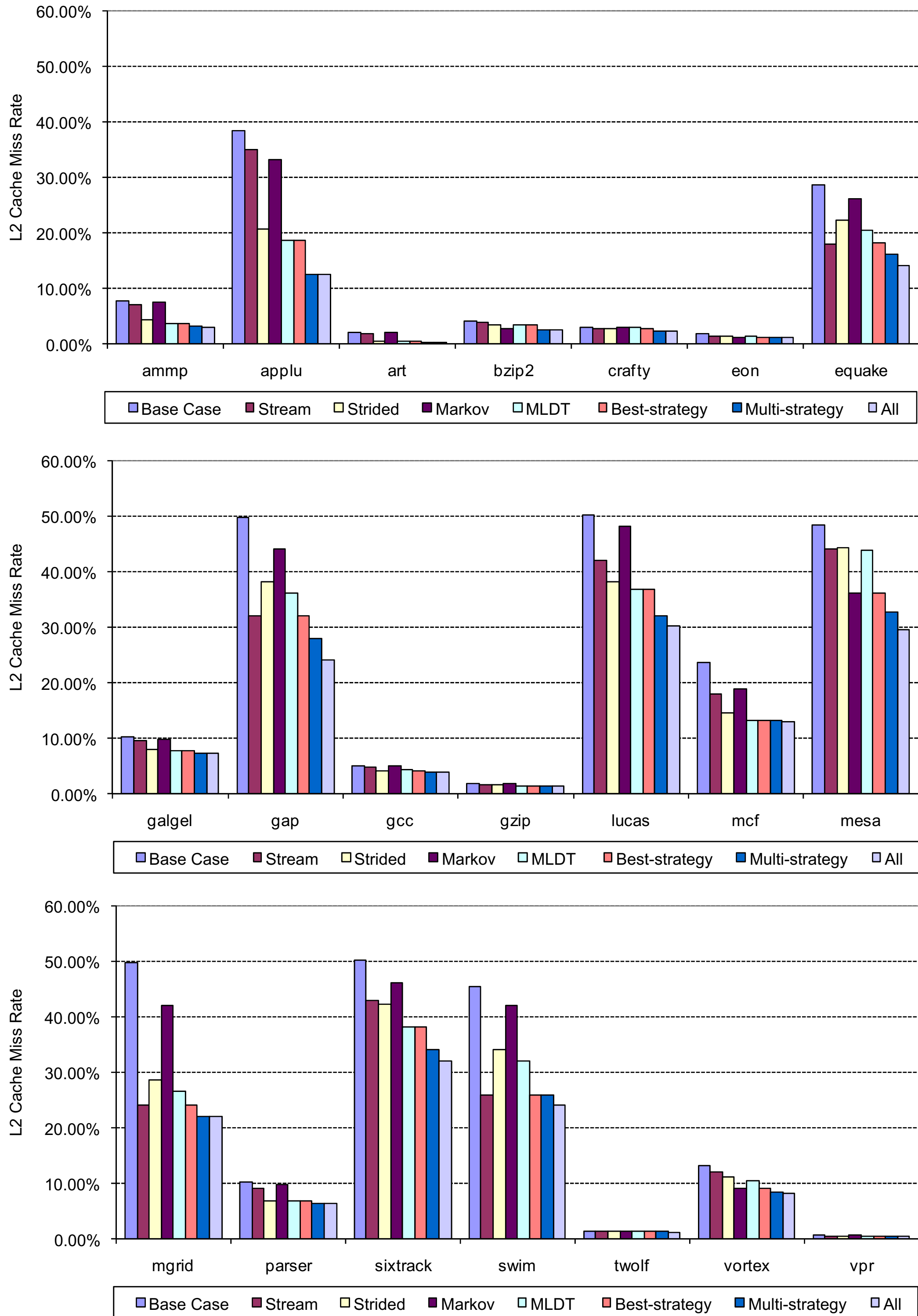


Fig. 7. Cache miss rate reduction of SPEC-CPU2000 benchmarks. Different applications exhibited distinct access patterns, and the cache miss rate reduction of various prefetching algorithms have large variations. The adaptation strategy can effectively identify the suitable prefetching algorithm and achieved better overall miss rate reduction.

strided prefetching performed extremely well for ammp, applu, lucas, mcf, etc., and Markov prefetching had considerable miss reduction for bzip2, eon and vortex benchmarks. The MLDT prefetching usually achieved a better miss rate reduction than strided prefetching, but still not for all benchmarks. These observations confirm that an adaptive prefetcher is desired to

be able to adjust to different application features at runtime to achieve a better overall prefetching performance. Simply adopting a fixed prefetching strategy cannot be the optimal solution.

The AFA prefetcher has demonstrated its strength through the simulation. From the reported miss rate results, we can tell that the best-strategy can almost achieve the largest miss rate reduction compared with each individual supported algorithm. This fact confirms that this adaptation strategy can effectively identify the suitable prefetching algorithm for current application access pattern. The multi-strategy adaptation can sometimes achieve an even better miss rate reduction, such as in the applu, gap, lucas, mesa and sixtrack benchmarks. The investigation shows that this further reduction is due to two or three well-performing prefetching strategies the multi-strategy adaptation identified and selected at different stages for these benchmarks. These optimal strategies are all able to generate effective prefetches at a specific stage. Table 2 lists the primary prefetching algorithms identified and selected by the AFA prefetcher for different benchmarks at runtime.

The last bar within each set of tests represents the miss rate reduction with all four prefetching algorithms working concurrently. The experimental results show that this case has achieved about the same reduction as multi-strategy adaptive prefetching, while better than adaptive strategies sometimes. However, as shown from the reported IPC results presented in the following subsection, the best miss reduction does not translate to best overall performance improvement because this strategy generates extensive replacements to the prefetch destination. In addition, the all-prefetching strategy consumes more resources than an adaptive strategy like the AFA prefetcher because the latter can identify the optimal ones and shut off low-efficiency prefetchers.

4.3.2. IPC improvement

Fig. 8 demonstrates the performance measurement in terms of IPC reported by SimpleScalar simulator. The results shown in the figure include twenty-one benchmarks under eight cases, similarly as discussed in the previous subsection. Fig. 9 reports the performance speedup (or slowdown) with the IPC speedup under all cases. Both the IPC measurement and the IPC speedup confirm that different prefetching algorithms benefit distinct benchmarks with different patterns. Any specific algorithm did not achieve the best IPC speedup for all benchmarks. Instead, these four supported prefetching algorithms have large variations in terms of the performance gain measured in IPC.

As shown from the reported results, the AFA prefetcher does have the capability to distinguish well-performing algorithms from others and adapts to these selected algorithms to achieve an overall better performance gain. For instance, the best-strategy adaptation has successfully identified strided prefetching suitable for ammp, applu, crafty, gcc, gzip, lucas, mcf, parser, twolf and vpr, while stream prefetching suitable for equake, gap, mesa, mgrid and swim. Both Markov and MLDT prefetching were also identified as better strategies at some cases, like Markov for the benchmark bzip2, eon and vortex, and MLDT for the benchmark art, galgel and sixtrack. Notice that better cache misses rate reduction does not necessarily result in a better IPC improvement. Take the applu benchmark as an example. The strided prefetching reduced less misses than MLDT did, but it produced better IPC improvement. This is because that MLDT involves more prediction overhead.

It is interesting to notice that multi-strategy adaptation usually generates better performance improvement than best-strategy does. This is because the multi-strategy adaptation is able to recognize multiple well-performing algorithms, and can benefit and complement each other while avoiding low-effective algorithms. Adopting all supported prefetching algorithms does not produce the best performance speedup. This observation reveals that adopting a low-accurate, low-coverage or high-pollution algorithm can even substantially worsen the performance. This fact has also been confirmed from most cases in the experiments.

The average performance improvement of all twenty-one benchmarks with each strategy is shown in Fig. 10. The best-strategy and multi-strategy adaptation mechanisms provided by the AFA prefetcher achieved 15.14% and 17.90% average improvement respectively, which is clearly better than all other cases. In summary, as verified from the simulation testing, the AFA prefetcher is able to dynamically choose proper algorithms for different applications and to achieve an overall better performance improvement of prefetching.

4.3.3. AFA behavior analysis

It is interesting to analyze the behavior of the AFA prefetcher in terms of the frequency of algorithm switches and the distribution of algorithms selected for better understandings.

The frequency of algorithm switches of both the best-strategy adaptation and multi-strategy adaptation is shown in Figs. 11 and 12 respectively. The frequency is counted as the ratio between the number of algorithm changes over the total

Table 2
Primary prefetching algorithms selected adaptively at runtime (B-S: best-strategy, M-S: multi-strategy; SM: stream, ST: strided, MK: Markov, MT: MLDT).

	ammp	applu	art	bzip2	crafty	Eon	equake	galgel	gap	gcc
B-S	ST	ST	MT	MK	ST	MK	SM	MT	SM	ST
M-S	ST	ST, MT	MT	MT, MK	ST, SM	MK	SM	ST, MT	SM	ST

gzip	lucas	mcf	mesa	mgrid	parser	sixtrack	swim	twolf	vortex	vpr
ST	ST	ST	SM	SM	ST	MT	SM	ST	MK	ST
ST, MT	SM, ST	ST	SM, MT	SM, MT	MT, ST	ST, MT, MK	SM	MT, ST	MK	MT, ST

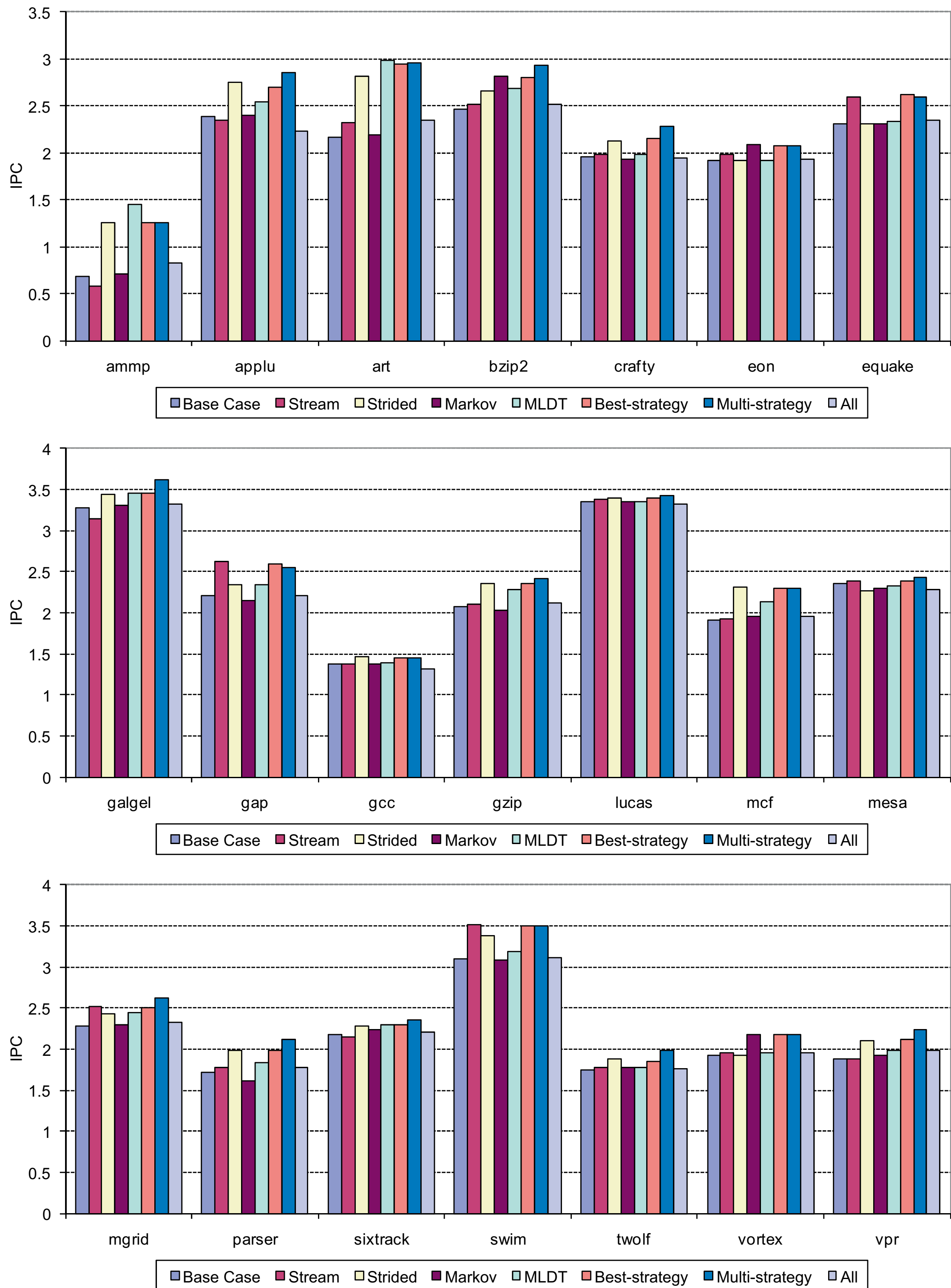


Fig. 8. Instructions Per Cycle (IPC) of SPEC-CPU2000 benchmarks with different data prefetching strategies. These results confirm that different applications exhibited distinct access patterns, and the IPC improvement of various prefetching algorithms have variations. The AFA prefetcher has the capability to distinguish well-performing algorithms from others and adapts to these selected algorithms to achieve an overall better performance gain.

number of metrics collection phases. For the best-strategy adaptation, if the algorithm selected for the stably prefetching phase is different in two consecutive phases, it is counted as one algorithm change or algorithm switch. For the

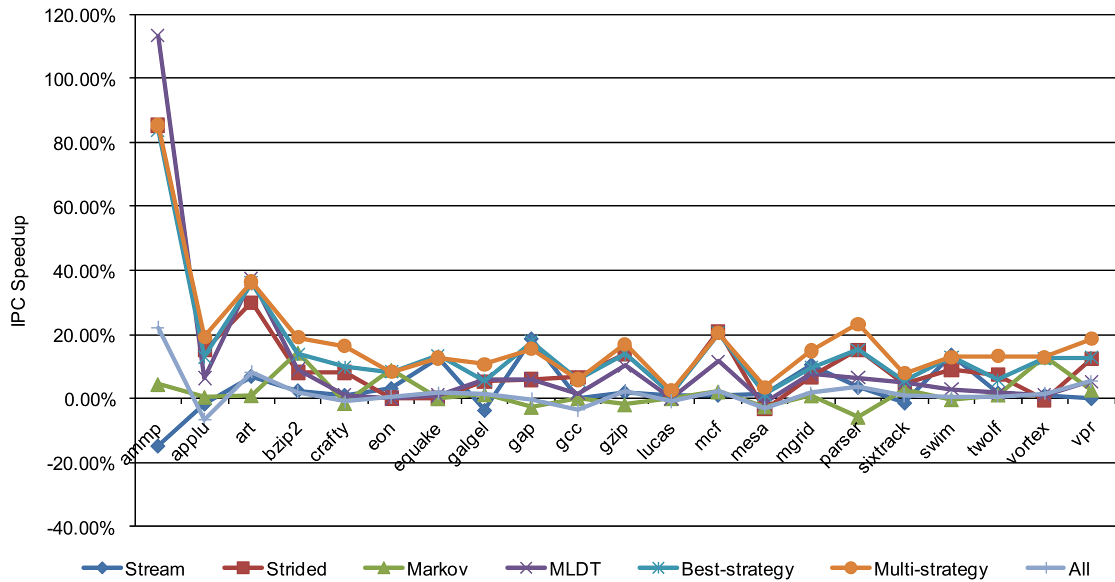


Fig. 9. IPC speedup with different data prefetching strategies. This figure combines all cases together and directly compares the performance improvement, IPC speedup, of different prefetching strategies.

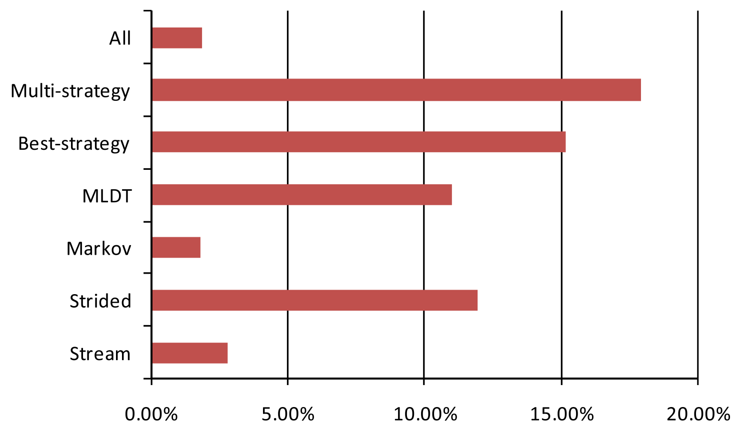


Fig. 10. Average IPC speedup. The best-strategy and multi-strategy adaptation mechanisms provided by the AFA prefetcher achieved 15.14% and 17.90% average improvement respectively, which is clearly better than all other cases.

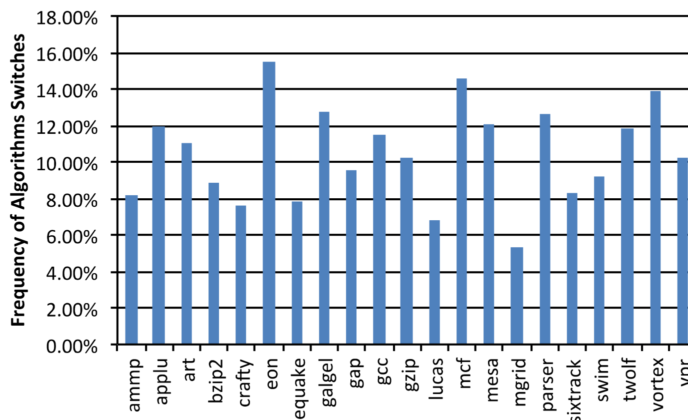


Fig. 11. Frequency of algorithms switches for best-strategy adaptation.

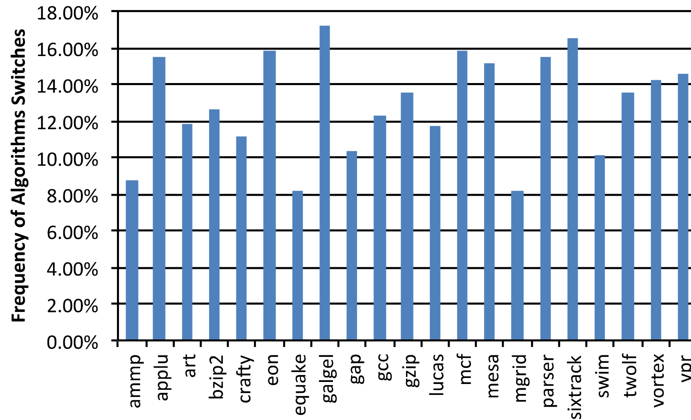


Fig. 12. Frequency of algorithms switches for multi-strategy adaptation.

multi-strategy adaptive selection, the algorithm change in two consecutive phases contains three scenarios: a new algorithm selected, an old algorithm removed, or both. It can be observed that the frequency of algorithm switches varies among applications. For the best-strategy adaptation, the arithmetic mean of the frequency is 10.13%. For the multi-strategy adaptation, the arithmetic mean is 12.68%, slightly higher than that in the best-strategy case. The driving force of the algorithm switches is the dynamic access pattern of applications. The ability of adapting the prefetching algorithms to different access patterns is desired for an effective prefetcher, and is a motivation of this study.

Fig. 13 shows the distribution of selected algorithms for the best-strategy adaption. This figure illustrates the detailed analysis of the AFA prefetcher behavior. Several observations can be made from this analysis. First, the detailed frequency distribution of selected algorithms confirms the need of adaption again. Different algorithms exhibit advantages over other algorithms for different applications. Second, in general, the strided prefetching and stream prefetching algorithms are selected most often. Third, certain algorithm such as the Markov prefetching performs weakly in general, but surprisingly well for some applications such as *vortex* and *bzip2*. An algorithm-level adaption is essential to address this issue and to identify the well performing algorithms depending on specific applications access features.

4.3.4. Sensitivity analysis

Cache pollution is an undesirable side effect of data prefetching techniques. In this subsection, we present an analysis of the performance sensitivity and cache pollution of the AFA prefetcher under different cache sizes.

All the prior simulation evaluations were carried out with 1 MB L2 cache. To observe the AFA prefetcher's sensitivity to different cache sizes and the impact of potential cache pollution, we varied L2 cache sizes as 512 kB and 2 MB to compare its performance with the case of 1 MB. The results show that a larger cache size helps the AFA prefetcher reduce more misses. A larger cache size lowered the extra misses due to the cache pollution, and therefore reduced the number of misses indirectly.

Figs. 14 and 15 show the IPC speedup results of the AFA best-strategy and multi-strategy, respectively, with three different L2 caches sizes. When the L2 cache size was increased, the IPC speedup improved for gap, lucas, mesa and sixtrack benchmarks. These benchmarks have high miss rates. Larger cache size hides the performance loss caused by cache pollution for these benchmarks. For the majority of benchmarks, the AFA prefetcher achieved stable IPC speedup. These simulation results demonstrate that the AFA prefetcher can have steady performance improvement with different cache sizes and is

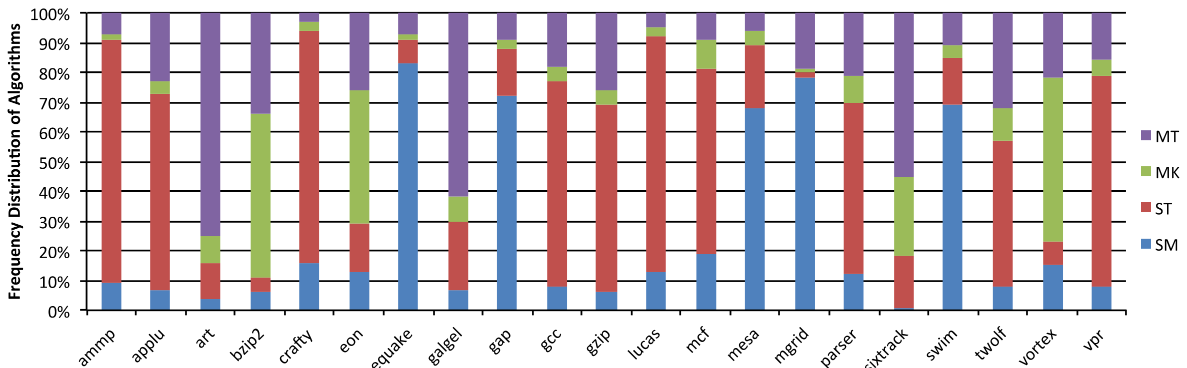


Fig. 13. Distribution of selected algorithms.

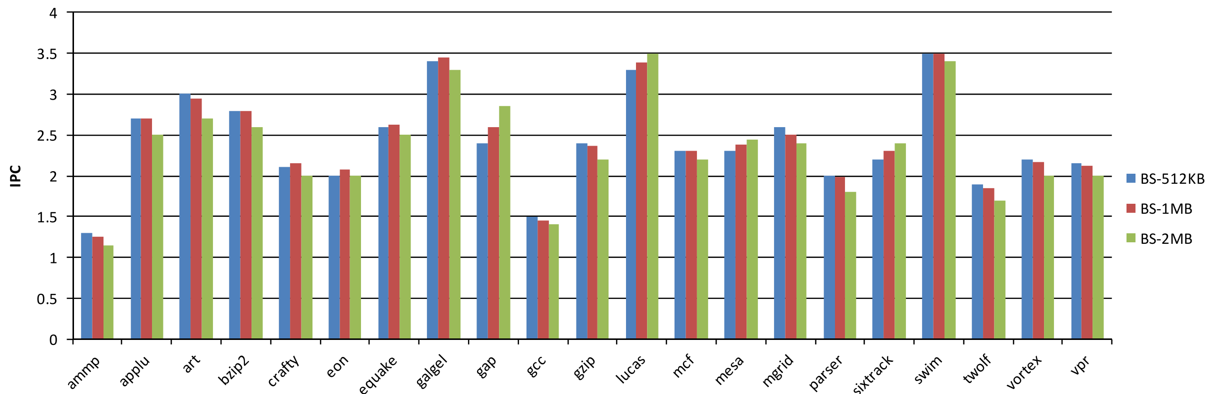


Fig. 14. IPC speedup of AFA best-strategy with 512 kB, 1 MB and 2 MB l2 caches. For the majority of benchmarks, the AFA prefetcher achieved stable Ipc speedup with different cache sizes and was not highly sensitive to different cache size configurations.

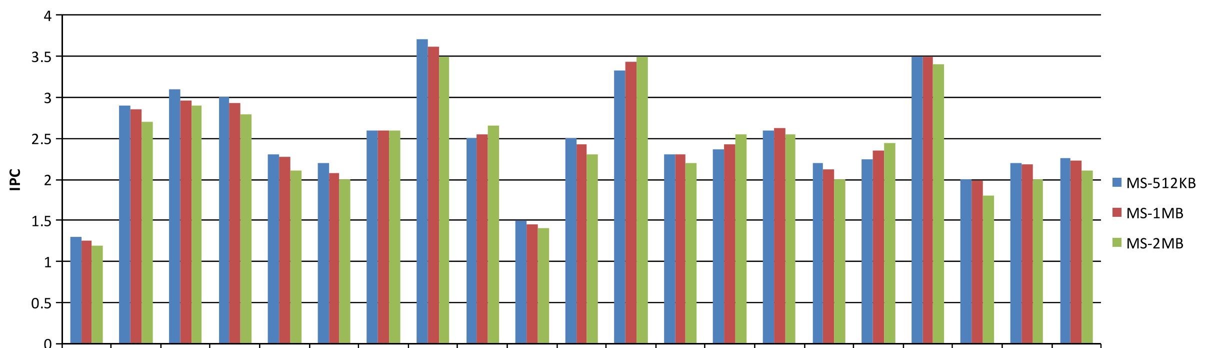


Fig. 15. IPC speedup of AFA multi-strategy with 512 kB, 1 MB and 2 MB L2 caches. For the majority of benchmarks, the AFA prefetcher achieved stable IPC speedup with different cache sizes and was not highly sensitive to different cache size configurations.

not highly sensitive to different cache size configurations. Although larger cache size helps reduce the cache pollution, the AFA prefetcher's dynamic and adaptive design makes it control the cache pollution well.

5. Related work

Data prefetching, as the name indicates, is a technique to fetch data before requested. A similar technique is instruction prefetching, which tries to speculate the future instructions and fetch them from memory in advance [15]. Data prefetching is usually classified as software prefetching and hardware prefetching [37]. Software prefetching is a technique to instrument prefetch instructions to the source code either by a programmer or by a complier during optimization phase. Hardware prefetching does not require modifications to binary or source code, and can benefit directly for existing executables. Existing hardware prefetching studies can be roughly classified into several categories: locality-based prefetching; correlation-based prefetching; and context-based prefetching. After describing previous work in each of these categories, we discuss how our AFA prefetcher is related to the existing work.

5.1. Locality-based prefetching

The locality-based prefetching category refers to a set of prefetching strategies that take advantage of data access locality in generating prefetches. The sequential prefetching, stream prefetching, and strided prefetching are representative strategies in this category.

Sequential prefetching is a basic prefetching strategy. It prefetches one or more blocks that follow the current missing block [10,11]. This prefetching mechanism takes advantage of spatial locality and assumes the applications usually request consecutive memory blocks. The one-block-lookahead (OBL) approach automatically prefetches the next block when an access of a block is initiated [11]. However, the limitation of this approach is that the prefetch may not be initiated early enough prior to processor's demand for the data to avoid a processor stall. To solve this issue, a variation of OBL prefetching, which fetches *k* blocks (called prefetching degree) instead of one block, is proposed [10]. Another variation is called adaptive

sequential prefetching, which varies prefetching degree k based on the prefetching efficiency [10]. The prefetching efficiency is a metric defined to characterize a program's spatial locality at runtime. Stream prefetching is a generalized sequential prefetching that supports the detection of multiple streams [23]. Strided prefetching approach [5,13] observes the pattern among strides of past accesses and thus predicts future references. It builds a state machine to track strides of accesses and generates prefetches when the state machine arrives at a stable state. The strided prefetching is generally implemented with a reference prediction table [5], or the recently proposed Data-Access History Cache [6].

5.2. Correlation-based prefetching

The correlation-based prefetching category refers to a set of strategies that exploit general correlation, not necessary the locality, among accesses. The representative strategies in this category include Markov prefetching, GHB-based correlation prefetching, DAHC-based correlation prefetching, epoch-based correlation prefetching, etc.

Markov prefetching was proposed by Joseph and Grunwald to capture the correlation between cache misses and prefetch data based on a state transition diagram [21]. The state transition diagram is built with states denoting an accessed data block. Probability of each state transition is maintained, so that most probable predicted data are prefetched in advance and the least probable predicted data references could be dropped from prefetching. Nesbit and Smith proposed the Global History Buffer (GHB) to maintain the recent history of data access [25]. It is an efficient structure for supporting different prefetching algorithms and has been adopted widely. Similar to GHB, Data Access History Cache (DAHC) [6] was proposed to facilitate data prefetching with a single structure and to support various algorithms simultaneously at runtime. More recently, a stream chaining method was proposed by Diaz and Cintra to link various localized streams into predictable chains such that multiple levels of correlation can be exploited by the prefetcher [12]. Since data-access patterns may change at runtime, adaptive mechanisms, such as feedback directed prefetching proposed by Srinath et al. [34], are often used to control the prefetch degree and the prefetcher's aggressiveness dynamically.

In addition, an epoch-based correlation prefetching was proposed by Chou to reduce late prefetches for correlation-based prefetching [9]. The idea is to localize misses within an to epoch (a fixed-length of time), and conduct prefetches based on epochs to prevent late prefetches. Lai et al. proposed a dead-block predictor (DBP) to improve correlation-based prefetching timeless by triggering prefetches and making replacement decision on time [22]. It predicts when a cache block is dead and can be evicted by tracking the time duration between when it comes into the cache and when it is evicted. Hu et al. also proposed time-keeping mechanisms to improve timely prefetching [19]. Bhattacharjee and Martonosi proposed two Inter-Core Cooperative (ICC) TLB prefetching mechanisms to exploit the correlation in TLB (Translation Lookaside Buffers) miss patterns across cores in chip multiprocessors (CMPs), which significantly improves data TLB (D-TLB) access performance [2]. Ebrahimi et al. proposed a hierarchy of prefetcher aggressiveness control structures to control prefetcher-caused inter-core interference by dynamically adjusting and coordinating the aggressiveness of multiple prefetchers in CMPs [14]. Their proposed structures improve the system performance and reduce bus traffic considerably on a multi-core system. Somogyi et al. proposed temporal, spatial, and spatio-temporal memory streaming to detect repeated patterns in specific memory regions and to boost memory access performance [32,38]. These approaches achieve remarkable performance for applications with regional repeated patterns.

5.3. Context-based prefetching

Context-based prefetching is a class of prefetching strategies that model data access as context and utilizes the relationship between current context (the miss access information) and the historical context to make predictions for data prefetching. A context-based prefetching method builds a state transition diagram with the access address strides (deltas) as states, and characterizes the correlation among miss address streams. Context-based data prefetching is similar to the context-based value predictor [29], but is used as a data prefetcher on the cache level. The Finite Context Method (FCM) is a representative context-based predictor that predicts the next value based on a finite number (order) of preceding values [29]. A variation of FCM technique, called Differential Finite Context Method (DFCM), was proposed in [17]. In this model, the context is formed by the differences between values instead of values themselves. DFCM can find more repeating patterns than FCM does, and in the meantime, it reduces the number of value prediction table entries needed in the original FCM [29]. P-DFCM [27] is a recently proposed data prefetcher based on DFCM. There are two major differences between P-DFCM and DFCM. First, P-DFCM prefetches on L2 load misses only, while DFCM prefetches on both load and store misses. Second, P-DFCM can prefetch when the PC (Program Counter) is known, while DFCM prefetches when both PC and its requesting data address are known. Recently, a multi-order context analysis approach has been proposed for context-based prefetching [8], where the order refers the length of the context. A multi-order analysis based context prefetching can improve the prefetching coverage, and thus improve the overall prefetching effectiveness.

5.4. Speculative-execution based prefetching

The speculative-execution based prefetching category refers a set of strategies that are completely different from the prior three categories. This set of strategies make the prediction of future data accesses by speculatively executing a fragment of code, whereas the prior three categories are all based on collecting the historical information and making the

prediction based on history. The speculative-execution based prefetching can be beneficial because it can steal cycles for speculative execution that are otherwise wasted due to data stall. It can also make use of the emerged multicore architecture and efficiently use one core to speculatively execute and warm up data cache for the process running on another core.

The representative speculative-execution based prefetching includes Zhou's dual-core execution (DCE) approach [41], Ganusov et al.'s future execution (FE) approach [16], Sun et al.'s data push server architecture [35] and Solihin et al.'s memory-side prefetching [33]. DCE and FE approach target for multi-core architecture. They use idle core to pre-execute future loop iterations to warm up cache (bring data to cache in advance). The data push server architecture utilizes a separate processing unit such as a separate core to conduct proactive push-based prefetching. The memory-side prefetching approach uses a memory processor residing within main memory to prefetch data proactively. The latter two approaches are also usually distinguished as push based prefetching from traditional pull based prefetching.

5.5. Relation to this work

Without the benefit of programmer or compiler hints, the effectiveness of hardware prefetching largely relies on the accuracy of prediction strategies. Incorrect prediction brings useless blocks into cache, consumes memory bandwidth and might cause cache pollution. To increase prefetching accuracy and coverage, hardware prefetching strategies should be able to make dynamic adaptation at runtime for different access patterns. This research targets to provide an algorithm-level hardware adaptive data prefetching to resolve this issue. Although a large body of data access acceleration and prefetching studies exist in locality-based, correlation-based, context-based and speculative-execution based prefetching, there are very limited studies exploiting dynamic and adaptive support for data prefetching strategies.

A few existing literature provide some form of adaptation, however [10,34], these strategies focus on adapting the prefetch degree and prefetch distance only. Our idea is motivated from the fact that no single prediction algorithm can work universally well for all applications. The adaptation at an algorithm-level is a necessity. This research provides such a solution and conducts simulation analysis with supporting four representative prefetching algorithms.

To the best of our knowledge, this research is the first work bringing algorithm-level adaptation into attention and exploiting a feedback-controlled dynamic adaptation mechanism. Nevertheless, the existing studies on adapting prefetch degree and distance are complementary to this study. They can be combined with this research to provide degree, distance and algorithm-level adaptation.

6. Conclusion

Advances in processor architectures such as multicore architectures have put more pressure than ever on reducing data-access latency for high-performance computing systems. Data-access latency has been recognized by many as the leading factor preventing high-sustained performance of applications. Data prefetching is an effective solution to accelerating data-access performance and to mitigating the fast growing processor-memory performance gap. Many hardware prefetching techniques have been widely used in contemporary processor architecture. They are successful for applications with simple data-access patterns, but notorious in certain cases for generating pollution and other overhead due to their low effectiveness. Previous study shows this low effectiveness is due to the lack of adaptation of existing hardware prefetchers. This study proposes an Algorithm-level Feedback-controlled Adaptive (AFA) prefetcher to support algorithm-level adaptation depending on application runtime data-access behavior. While existing adaptive prefetchers only adapt within a given prefetching algorithm, AFA can change prefetching algorithms at runtime. It provides more flexibility and, therefore, better performance. We have conducted extensive simulations with an enhanced SimpleScalar simulator to verify and evaluate the design. Simulation results have demonstrated a clear performance improvement over existing strategies. The AFA prefetcher will have an impact on accelerating applications' data-access performance.

Acknowledgements

This research is sponsored in part by the National Science Foundation under NSF Grant CCF-0621435 and CCF-0937877, and by ACM/IEEE High-Performance Computing Ph.D. Fellowship and Illinois Institute of Technology Fieldhouse Research Fellowship.

References

[1] B. Bloom, Space/time trade-offs in hash coding with allowable errors, Communications of the ACM 13 (7) (1970) 422–426.
[2] A. Bhattacharjee, M. Martonosi, Inter-core cooperative TLB for chip multiprocessors, in: ASPLOS, 2010, pp. 359–370.
[3] D.C. Burger, T.M. Austin, S. Bennett, Evaluating Future Microprocessors: The SimpleScalar Tool Set, University of Wisconsin-Madison Computer Sciences Technical, Report 1308, 1996.
[4] S. Byna, Y. Chen, X.-H. Sun, Taxonomy of data prefetching for multicore processors, Journal of Computer Science and Technology (JCST) 24 (3) (2009) 405–417.
[5] T.F. Chen, J.L. Baer, Effective hardware-based data prefetching for high performance processors, IEEE Transactions on Computers (1995) 609–623.
[6] Y. Chen, S. Byna, X.-H. Sun, Data access history cache and associated data prefetching mechanisms, in: Proceedings of the 2007 ACM/IEEE Supercomputing Conference, 2007.

[7] Y. Chen, H. Zhu, X.-H. Sun, An adaptive data prefetcher for high-performance processors, in: Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid, Computing (CCGrid'10), 2010.
[8] Y. Chen, H. Zhu, H. Jin, X.-H. Sun, Improving the effectiveness of context-based prefetching with multi-order analysis, in: 3rd International Workshop on Parallel Programming Models and Systems Software for High-End, Computing (P2S2), 2010.
[9] Y. Chou, Low-cost epoch-based correlation prefetching for commercial applications, in: MICRO, 2007.
[10] F. Dahlgren, M. Dubois, P. Stenstrom, Fixed and adaptive sequential prefetching in shared-memory multiprocessors, in: Proceedings of the 993 International Conference on Parallel Processing, 1993, pp. I56–I63.
[11] F. Dahlgren, M. Dubois, P. Stenstrom, Sequential hardware prefetching in shared-memory multiprocessors, IEEE Transactions on Parallel and Distributed Systems 6 (7) (1995) 733–746.
[12] P. Diaz, M. Cintra, Stream chaining: exploiting multiple levels of correlation in data prefetching, in: Proceedings of International Symposium on Computer Architecture, 2009, pp. 81–92.
[13] J. Doweck, Inside Intel Core Micro-architecture and Smart Memory Access, Intel White Paper, 2006.
[14] E. Ebrahimi, O. Mutlu, C.J. Lee, Y.N. Patt, Coordinated control of multiple prefetchers in multi-core systems, in: MICRO, 2009, pp. 316–326.
[15] A. Falcon, A. Ramirez, M. Valero, Effective instruction prefetching via fetch prestaging, in: Proceedings of the 19th International Parallel and Distributed Processing, Symposium, 2005.
[16] I. Ganusov, M. Burtscher, Future execution: a hardware prefetching technique for chip multiprocessors, in: Proceedings of the 14th Annual International Conference on Parallel Architectures and Compilation, Techniques, 2005.
[17] B. Goeman, H. Vandierendonck, K. Bosschere, Differential FCM: increasing value prediction accuracy by improving table usage efficiency, in: Proceedings of 7th International Symposium on High performance Computer, Architecture, 2001.
[18] J. Hennessy, D. Patterson, Computer Architecture: A Quantitative Approach, fourth ed., Morgan Kaufmann, 2006.
[19] Z. Hu, M. Martonosi, S. Kaxiras, Timekeeping in the memory system: predicting and optimizing memory behavior, in: ISCA, 2002, pp. 209–220.
[20] N.P. Jouppi, Improving direct-mapped cache performance by the addition of a small fully-associative cache and prefetch buffers, in: Proceedings of the 17th International Symposium on Computer, Architecture, 1990.
[21] D. Joseph, D. Grunwald, Prefetching using Markov predictors, in: Proceedings of the 24th Annual Symposium on Computer, Architecture, 1997.
[22] A.-C. Lai, C. Fide, B. Falsafi, Dead-block prediction & dead-block correlating prefetchers, in: ISCA, 2001, pp. 144–154.
[23] H.Q. Le, W.J. Starke, J.S. Fields, F.P. O'Connell, D.Q. Nguyen, B.J. Ronchetti, W.M. Sauer, E.M. Schwarz, M.T. Vaden, IBM Power6 microarchitecture, IBM Journal of Research and Development 51 (6) (2007) 639–662.
[24] S.A. McKee, Reflections on the memory wall, in: Proceedings of Computing, Frontiers (CF'04), 2004.
[25] K.J. Nesbit, J.E. Smith, Prefetching using a global history buffer, in: Proceedings of the 10th Annual International Symposium on High Performance Computer, Architecture, 2004.
[26] J. Peir, S. Lai, S. Lu, J. Stark, K. Lai, Bloom filtering cache misses for accurate data speculation and prefetching, in: Proceedings of the 16th International Conference on Supercomputing, 2002.
[27] L. Ramos, J. Briz, P. Ibañez, V. Viñals, Data prefetching in a cache hierarchy with high bandwidth and capacity, in: ACM Computer Architecture News, 2007, pp. 37–44.
[28] E. Rotenberg, S. Bennett, J.E. Smith, Trace cache: a low latency approach to high bandwidth instruction fetching, in: Proceedings of the 29th Annual IEEE/ACM International Symposium on Microarchitecture, 1996.
[29] Y. Sazeides, J. E. Smith, The predictability of data values, in: Proceedings of the 30th International Symposium on Microarchitecture, 1997.
[30] K. Shen, M. Zhong, S. Dwarkadas, C. Li, C. Stewart, X. Zhang, Hardware counter driven on-the-fly request signatures, in: Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems, 2008.
[31] J. Skeppstedt, M. Dubois. Hybrid compiler/hardware prefetching for multiprocessors using low-overhead cache miss traps, in: Proceedings of the International Conference on Parallel Processing, 1997.
[32] S. Somogyi, T.F. Wenisch, A. Ailamaki, B. Falsafi, Spatio-temporal memory streaming, in: International Symposium on Computer Architecture, 2009, pp. 69–80.
[33] Y. Solihin, J. Lee, J. Torrellas, Using a user-level memory thread for correlation prefetching, in: Proceedings of the 8th International Symposium on Computer, Architecture, 2002.
[34] S. Srinath, O. Mutlu, H. Kim, Y. Patt, Feedback directed prefetching: improving the performance and bandwidth-efficiency of hardware prefetchers, in: Proceedings of the 13th International Symposium on High Performance Computer, Architecture, 2007.
[35] X.H. Sun, S. Byna, Y. Chen, Improving data access performance with server push architecture, in: Proceedings of the NSF Next Generation Software Program Workshop in IPDPS'07, 2007.
[36] P.F. Sweeney, M. Hauswirth, B. Cahoon, P. Cheng, A. Diwan, D. Grove, M. Hind, Using hardware performance monitors to understand the behaviors of Java applications, in: Proceedings of the 3rd USENIX Virtual Machine Research and Technology Symposium, 2004.
[37] S.P. Vander Wiel, D.J. Lilja, When caches aren't enough: data prefetching techniques, IEEE Computer 30 (7) (1997) 23–30.
[38] T.F. Wenisch, S. Somogyi, N. Hardavellas, J. Kim, A. Ailamaki, B. Falsafi, Temporal streaming of shared memory, in: ISCA, 2005, pp. 222–233.
[39] W.A. Wulf, S.A. McKee, Hitting the memory wall: implications of the obvious, Computer Architecture News 23 (1) (1995) 20–24.
[40] X. Zhang, S. Dwarkadas, G. Folkmanis, K. Shen, Processor hardware counter statistics as a first-class system resource, in: Proceedings of the 11th Workshop on Hot Topics in Operating Systems, 2007.
[41] H. Zhou, Dual-core execution: building a highly scalable single-thread instruction window, in: Proceedings of the 14th International Conference on Parallel Architectures and Compilation, Techniques, 2005.
[42] Standard Performance Evaluation Corporation, SPEC Benchmarks. http://www.spec.org/.