Clean-Prefetcher: Look-Ahead Prefetching without Cache Pollution

Minshin Cho¹, Jae Young Hur², and Wooyoung Jang¹a

Abstract In this letter, we propose a novel prefetching technique that is free from cache pollution and thus achieves high performance for multicore processors. Unlike the conventional prefetchers that cause incorrect predictions, the proposed prefetcher reads instructions in advance only in determined paths and charges dynamic random access memory (DRAM) cells storing instructions in undetermined paths via refreshing DRAM cells. The DRAM cells highly charged will be quickly accessed. Since caches served by our prefetcher always store useful instructions, as a result, they are free from cache pollution that results in lower cache hit rate. In the case that SPEC CPU2006 benchmarks run on an 8-core processor, the proposed prefetcher consumes more 3.2% DRAM power, but achieves 12% higher performance on average.

key words: Prefetch, DRAM, refresh
Classification: Integrated circuits (memory, logic, analog, RF, sensor)

1. Introduction

As the number of cores integrated on a single die increases at faster rate than the improvement of memory performance [1, 2], a memory system becomes a critical performance bottleneck in state-of-the-art multicore processors. Many researchers have studied a variety of approaches for improving memory latency and capacity [3, 4, 5, 6, 7, 8, 9, 10]. Of them, a cache prefetcher that reads instructions from low-speed memories in advance can achieve the notable improvement of memory performance [11]. However, the prefetcher does not always improve memory performance. The reason is that the mispredictions of the prefetcher can evict instructions with high locality from a cache, and then fill useless ones in the cache, called cache pollution. In addition, the prefetcher can make urgent demand requests contend with mispredicted prefetch requests [12] and require meta-data unavailable in multicore processors for reading instructions in advance [13].

In this letter, we propose a novel clean-prefetcher that does not cause any cache pollution and thus improves the overall system performance. Our key observation is twofold. First, prefetching instructions in incorrect paths has critical negative impacts on cache hit rate. Even if a prefetcher is equipped with a reasonable branch predictor, the sum of performance gains from predicting correct paths can be less than that of performance losses from predicting incorrect paths. In particular, the losses can be more serious for multicore processors since cores share the cache polluted by useless instructions. Second, based on the branch prediction, reducing the latency of dynamic random access memories (DRAMs) can be more beneficial than prefetching instructions. Since DRAMs can be accessed much faster depending on the level of charges in the capacitor of DRAM cells [14, 15, 16], techniques that highly keep the charge of target DRAM cells prior to accessing the DRAM cells can be useful. Based on these observations, the major novelty and contribution of this letter are as follows: 1) We propose a novel prefetcher that generates not only demand and prefetch requests, but also row-refresh requests that recharge DRAM cells that are predicted to be accessed, and 2) we propose a memory controller that effectively performs demand, prefetch, and row-refresh requests, and 3) Our clean-prefetcher is evaluated in terms of performance and power consumption.

2. DRAM structure and operation

DRAM has three-dimensional structure, i.e., a bank, a row, and a column [17, 18]. Each bank is activated and deacti-vated by row activation (ACT) and precharge (PRE) commands, respectively, and read or write commands are executed to an activated bank. DRAM cells start leaking charge after they are disconnected from their bitline [19, 20, 21]. Thus, a refresh command is issued to avoid data loss within an allotted refresh period due to charge leakage from the capacitor of DRAM cells [22]. Each command is completed with the predefined interval time.

Timing parameters relevant to row activation can be further optimized [14, 15, 16] even though they have been determined by the DRAM standards [17]. Fig. 1 shows ACT and PRE operations depending on the level of charge in the capacitor of DRAM cells. When a row activation command is issued, the cell and bitline of DRAM share charges each other. Such charge sharing operation induces ΔV, and then ΔV is amplified by a sense amplifier. The initial level of charges has an effect on the response time of the sense amplifier. In Figs. 1(a) and 1(b), DRAM cells are partially and fully charged, respectively. After the charge shar-
ing operation, the sensing and restoring speed of the fully charged DRAM cell is faster than that of the partially charged DRAM cell. DRAM timing parameters relevant to sensing and restoring operations have been decided under the assumption that the sense amplifier is driven by the worst ΔV. Thus, the DRAM timing parameters can be reduced if DRAM cells that are soon accessed are detected and charged highly.

3. Prefetching without cache pollution

As instructions prefetched in a cache increase, cache hit rate can increase owing to improved memory utilization. However, in the case that the working set exceeds the limited cache capacity, other useful instructions with high temporal locality start to be evicted from a cache [11]. Furthermore, even if a prefetcher is equipped with an excellent conditional branch (CB) predictor, it does not always correctly prefetch instructions that will be executed after a CB instruction into a cache [28, 23]. The reason is that the latest applications such as deep learning [24], virtual/augmented reality [25], and high-quality multimedia [26] have extremely complex and irregular memory accesses. If the prefetcher often incorrectly predicts the next path for a CB instruction, instructions with high locality may be evicted from caches, and the caches can be filled with unnecessary instructions [27]. When a multicore processor simultaneously executes a variety of applications, especially, its shared cache that is polluted by useless instructions has more negative impacts on the overall system performance [12]. Therefore, a prefetching technique that does not result in cache pollution is required.

Fig. 2(a) is the structure of the proposed clean-prefetcher that is modified in the wrong-path instruction prefetcher proposed in [28]. The wrong-path instruction prefetcher reads instructions in both taken and not-taken paths of a CB instruction in advance. On the contrary, our prefetcher reads instructions in determined paths in advance via a prefetch request, and sends a row-refresh request to a memory controller to fully charge DRAM cells that store instructions in paths that depend on a CB instruction. The DRAM cells highly charged can rapidly provide instructions for a multicore processor as shown in Fig. 1(b) when they are actually accessed via demand and prefetch requests.

Our clean-prefetcher consists of a look-ahead program counter (LAPC), a decoder, a return address stack (RAS), a look-ahead controller (LAC), and several conditional branch trackers (CBTs). LAPC in the wrong-path instruction prefetcher reads instructions within fixed instruction distance from current program counter in advance even if the next path of a CB instruction is predicted. On the contrary, LAPC in our clean-prefetcher reads up to the coming CB instruction. A decoder checks if what LAPC reads is a branch instruction, and RAS stores a return address when a subroutine is called. A branch target buffer (BTB) stores the address, target address, and decision of the most recently executed branch instructions. LAC assigns taken or not-taken paths of a CB instruction to LAPC, and transfers prefetch and row-refresh requests to a memory controller. CBT newly added to our clean-prefetcher tracks a branch instruction in BTB for its assigned path.

Fig. 2(b) shows the operation of our clean-prefetcher with a simple example. We assume that a0 is being executed, and other instructions have not been in caches. LAPC starts fetching instructions such as a1, CBO, etc. in advance in a determined path via a prefetch request, and then stores them in caches. A decoder checks if there are some branch instructions prefetched by LAPC. In this example, it detects CBO, and informs LAC of the CBO detection. Next, LAC checks the recently executed CBO in BTB. If any history of CBO has not been in BTB, LAC sends row-refresh requests to a memory controller for recharging DRAM cells that store instructions in both taken and non-taken paths of CBO. The absence of the CBO history in BTB means that DRAM cells storing instructions executed after CBO have not been
recently used, and thus the DRAM cells may be discharged. In this example, previous CB0s were taken and not-taken as shown in the first and forth listings of BTB, respectively. In this case, LAC does not send row-refresh requests to a memory controller since DRAM cells storing instructions in both taken and not-taken paths of CB0 have been recently accessed. Then, LAC assigns the taken and not-taken paths of CB0 to CBTs 0 and 1, respectively. That is, LAC can send a row-refresh request to a memory controller for a CB path of which the history is not in BTB, and assign CBT to a CB path of which the history is in BTB.

CBTs 0 and 1 track their assigned paths to find the history of branch instructions in BTB. In this example, CBT 0 finds CB1. Since the taken path of CB1 has been included in the current path and CB1 previously executed was not taken as shown in the second listing of BTB, LAC does not send row-refresh request to a memory controller and keeps assigning the not-taken path of CB1 to CBT 0. Next, CBT 0 finds CB2 in the third listing of BTB. Since CB2 previously executed was just taken, LAC sends a row-refresh request to a memory controller for the not-taken path of CB2 and assigns the taken path of CB2 to CBT 0. Then, CBT 0 tracks the taken path including an d14, d15, and an instruction with an indirect address (indir). After Indir is executed, its next instructions can be determined. Since CBT 0 can no longer track CB instructions in BTB, thus, it is free from the assigned path. Meanwhile, CBT 1 tracks the not-taken path of CB0 in BTB. It sequentially encounters Call, Return, and Jump instructions converted to a CB instruction that is always taken. If CBT finds such a CB instruction in BTB, LAC should make CBT track its taken path and does not send row-refresh requests to a memory controller. Otherwise, CBT no longer tracks any paths. In this example, a Call instruction is in the fifth listing of BTB, and thus CBT 1 tracks a taken path starting at b6. Then, CBT 1 encounters a Return instruction in the sixth listing of BTB. Similarly, it tracks a taken path starting at a5. Next, CBT 1 finds a Jump instruction in the seventh listing of BTB and tries to track a taken path starting at c8. However, since the path has been already tracked by CBT 0, CBT 1 is free from its current path.

Whenever a CB instruction is executed and its path is determined, this procedure is repeated. When the path of CB0 is determined in Fig. 2(b), our clean-prefetcher writes a determined path in BTB and resumes either prefetching instructions in the determined path or row-refreshing DRAM cells storing instructions not recently executed in undetermined paths, referring to BTB. The depth of a path tracked in an instruction sequence depends on the number of CBT and the depth of BTB. Multiple CBTs and a deep BTB can make DRAM cells charged in more rows for fast accesses, but may increase design cost and power consumption.

4. Memory system supporting clean-prefetcher

4.1 Processing row-refresh requests

Our memory controller should perform a new row-refresh request which makes DRAM cells charged to support the proposed clean-prefetcher. The row-refresh request in DRAMs can be performed via a pair of ACT and PRE commands. In the case that a DRAM bank activated for refreshing a target row is not accessed by other demand or prefetch requests, our memory controller can skip a PRE command and operate in an open-page policy. Such an activated row can serve instructions for demand and prefetch requests just with column address strobe (CAS) latency since it saves time to deactivate a different row and activate the refreshed row again. Row-refresh requests should not interfere with not only demand requests, but also prefetch requests since instructions read in advance by our clean-prefetcher unlike conventional prefetchers are used soon. Thus, prefetch requests in our memory controller have a lower priority than demand requests, but a higher priority than others. On the contrary, row-refresh requests are served with a best-effort priority. Furthermore, in the case that any row to be refreshed by a row-refresh request is accessed by demand or prefetch requests, the priority of the row-refresh request is same as that of the demand or prefetch requests.

Our memory controller equips a small cache that stores row addresses recently refreshed (CRARR). The reason is that our clean-prefetcher can repeatedly generate the same row-refresh requests depending on the pattern of conditional branch instructions, and thus power consumption can be increased. For example, if the first listing of BTB is empty in Fig. 2(b), LAC will send a row-refresh request to a memory controller for the taken path of CB0. The row-refresh request makes all the DRAM cells in a single row charged. Since DRAM cells that store not only instructions c8, c9, CB1, c10, and CB2, but also instructions c11, c12, and c13 are in the same row, they are charged together. Let CB0 and CB1 taken and not-taken, respectively, and then, CB2 executed now. Since BTB does not have the history of previous CB2, LAC sends a row-refresh request for DRAM cells storing instructions c11, c12, and c13. However, the DRAM cells have been already charged via the previous row-refresh request. To prevent DRAM cells from being unnecessarily charged, our memory controller stores row addresses recently refreshed in CRARR and filters overlapped row-refresh requests.

4.2 Supporting multiple tRC and tRAS

Most of the DRAM timing parameters are programmed via mode register set (MRS) commands. The MRS commands take a long time to be executed, but they do not affect the system performance during power-up and initialization processes. However, in the case that some DRAM timing parameters are necessary to be changed, executing MRS commands while applications are running can result in the serious loss of the system performance. Thus, industrial memory systems usually rarely change DRAM timing parameters via MRS commands while applications are running.

On the contrary, activate-to-internal read or write delay time
predicted and thus fully charged in advance, they can serve demand and prefetch requests with 5.7ns shorter \( t_{\text{RCD}} \). Similarly, the charge of a DRAM cell capacitor is restored at different times according to the initial voltage of a DRAM cell. That is, a fully charged cell can achieve 11.2ns shorter \( t_{\text{RAS}} \) than a partially charged cell. Shortened \( t_{\text{RAS}} \) can make a bank that includes a refreshed row early deactivated by a PRE command and thus DRAM latency can be reduced when a different row of the bank are accessed.

We modify MARSSx86 [31] and DRAMSim2 [32] simulators for evaluating the proposed clean-prefetcher. MARSSx86 is a cycle-accurate full system simulator for multicore x86 CPUs, and DRAMSim2 is a cycle-accurate memory system simulator for DDR 2/3/4 DRAMs. They are configured as shown in Table I and run SPEC CPU2006 benchmarks [33].

Fig. 4 shows instruction per cycle (IPC), power consumption and cache miss rate of our clean-prefetcher normalized by the wrong-path instruction prefetcher [28]. The proposed prefetcher makes DRAM consume on average 2.4\% higher power for four benchmarks executed at the same time, as shown in Fig. 4(a). The reason is that all of the DRAM cells in a single row are fully charged thanks to some instructions that are expected to be accessed. CRARR in our memory controller minimizes this increase in power consumption by filtering overlapped row-refresh requests. On the contrary, our prefetcher achieves on average 10.5\% higher performance. This is because DRAM cells fully charged via a row-refresh request can serve demand and prefetch requests with shorter \( t_{\text{RCD}} \) and \( t_{\text{RAS}} \), and the cache miss rate resulting from our prefetcher decreases on average 2.5\%.

### Table I. Processor and memory configurations

<table>
<thead>
<tr>
<th>Processor</th>
<th>4-8 cores, 3 GHz, out-of-order, 4 issue per core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Last-level cache</td>
<td>2MB, 16-way associativity, 64B block size, write-back, pseudo pseudo least recently used (LRU) replacement policy</td>
</tr>
<tr>
<td>Memory controller</td>
<td>First-ready, first-come first-served, &quot;channel:row:rank: bank:column” mapping, 64-entry DRAM request buffer, open page policy</td>
</tr>
<tr>
<td>DRAM</td>
<td>1.6GHz data rate, 1 channel, 2 rank per channel, 8 banks per rank, 64B data width, 32K rows per bank, 2KB row, baseline 9b/cp/rras: 13.75/35ns, reduced tRCD/tRAS: 8.05/23.6ns</td>
</tr>
<tr>
<td>CRARR</td>
<td>128-entry, 2-way associativity, LRU replacement policy, 1ms-caching duration</td>
</tr>
</tbody>
</table>
In the case that the proposed prefetcher supports eight benchmarks simultaneously, it makes DRAM consume 3.2% more power on average, as shown in Fig. 4(b). On the contrary, the proposed prefetcher achieves on average 12% higher performance. This is because it is free from cache pollution in a last-level cache shared by many cores, instructions with high locality are hardly evicted, and thus the cache miss rate decreases on average 7.2%. As a result, our prefetcher can achieve higher performance in a processor equipping more cores and running more applications simultaneously.

6. Conclusion

Cache pollution resulting from the misprediction of a prefetcher has a significant impact on the performance of multicore processors. The reason is that the mispredicted instruction evicts the high-locality instructions of other cores from a shared last-level cache. Thus, in the letter, we proposed a charge-level-aware prefetcher that does not cause any cache pollution. Rather than predicting and prefetching instructions that may be unneeded in the next paths of a conditional branch instruction, our prefetcher charges DRAM cells storing the uncertain instructions, and prefetches instructions to be definitely used. Such a fully charged DRAM cell can be accessed with shorter DRAM latency for urgent demand requests. As a result, the proposed prefetcher achieves higher performance than the conventional prefetcher with a conditional branch predictor that causes cache pollution. In addition, our memory controller minimizes the increase in power consumption by filtering overlapped row-refresh requests. In conclusion, our clean-prefetcher can provide more opportunities to support multicore processors with high performance and reasonable cost.

Acknowledgments

The present research was conducted by the research fund of Dankook University in 2019.

References


[18] B. Jacob, D Wang, and S. Ng: Memory systems: cache, DRAM, disk(Morgan Kaufmann, 2010).


