Hardware Support for Effective Helper Threaded Data Prefetching for CMPs*

Min Cai, Zhimin Gu

Beijing Institute of Technology

1 Introduction

2 Hardware Metrics Reflecting the Effectiveness of the HT

2.1 Accuracy

Definition. A measure of how accurately the HT can predict the memory addresses that will be accessed by the MT.

Formula. HT request accuracy = # useful HT requests / # HT requests sent to memory, where # useful prefetches is # HT fetched cache blocks that are used by MT requests while they are in the L2 cache.

Hint. When the accuracy of HT requests is low, the HT may be harmful to the MT performance other than useful.

2.2 Lateness

Definition. A measure of how timely the HT requests are with respect to the MT accesses that need the HT fetched data.

Formula. HT request lateness = # late HT requests / # useful HT requests, where an HT request is defined to be late if the HT fetched data has not yet returned from the main memory by the time an MT load or store instruction requests the HT fetched data.

Hint. Even though the HT requests are accurate, the HT might not be able to improve the MT performance if the HT requests are very late.

*Project (No. 61070029) Supported by the National Natural Science Foundation of China.

2.3 Cache Pollution

Definition. A measure of the disturbance caused by HT fetched data in the L2 cache.

Formula. HT request generated cache pollution = # MT misses caused by HT / # MT misses, where an MT miss is defined to be caused by the HT if it would not have occurred had the HT not been present.

Hint. If the HT generated cache pollution is high, the MT performance can degrade because useful data in the cache could be evicted by HT fetched data.

3 Parameters for Tuning the Effectiveness of the HT

3.1 Aggressiveness of the HT

3.1.1 HT request distance (lookahead)

Definition. How many cache blocks ahead of the the MT access stream the HT can send requests.

3.1.2 HT request degree (stride)

Definition. How many cache blocks can be prefetched in each turn before synchronization

3.2 Insertion Position of the HT Fetched Data in the LRU Stack

4 Hardware Support for Effective Helper Threaded Data Prefetching for CMPs

4.1 Phase Based Sampling of the HT Effectiveness Hardware Metrics

4.2 Communicating the HT Effectiveness Hardware Metrics to the Program via Register Based Polling

A register based polling process is used based on the existing pseudo call handling mechanism implemented in the simulator, with the details as follows.

1. After the HT program has done one turn of data prefetching work, it puts the next expected MT memory address range (a candidate is just the last data block address that the MT requests) in some unused registers, and suspends its execution;

2. Processor reads the next expected MT memory address range in the unused registers and remembers when an MT access's address is within the expected address range set by the HT program;

3. The HT (the manager thread) periodically wakup and poll in some unused register to check whether an expected MT access has occurred;

4. If the HT (the manager thread) detects an expected MT access has occured, it will resumes and (the manager thread wakes up and notifies the HT to) begin the next turn of data prefetching work.

4.3 Dynamically Adjusting the Aggressiveness of the HT

5 Experimentation

5.1 Simulation Setup

5.2 Results

6 Related Work

7 Conclusion

Acknowledgments

This work was supported by the National Natural Science Foundation of China under the contract No. 61070029.