CS 838 - CMP Prefetching

Kyle Nesbit
Nick Lindberg
Outline

- Overview
- CZone Prefetching
- Simulator Modifications
- Results
- Ongoing Work
Overview

- As with most recent prefetching research, this work focuses on L2 cache misses.
CZone Prefetching

- CZone prefetching divides memory into Concentration Zones (CZones) and detects constant stride accesses within each zone [Palacharla and Kessler].

- Uses the miss address stream. Can prefetch lower in the cache hierarchy without modifying upper levels (i.e. doesn’t use the PC).

- Is nearly as affective as PC prefetching.

- Does not suffer from PC conflicts that can occur in a CMP environment.
Simulator Setup

- Added new cache state for prefetches.
- If an access hits a prefetched line, it is sent to the prefetcher as a cache miss (keeps miss stream intact)
- Prefetches go into L1 -> L2 request queue as GETS.
- If the prefetch hits L2, it is dropped (i.e. no L1 response message is sent).
- Otherwise prefetch is sent to the appropriate directory.
- When the prefetch returns from the network (it must return) it is entered into the cache, once again there is no L1 response.
Results

- Ran OLTP, Apache, and Zeus workloads with default configuration file except for the number processors and number of processors per chip.
- Two baseline systems, a MP system with four chips and one processor per chip, and a CMP system with four chips and four processors per chip.
- Simulated varying prefetch degrees (aggressiveness).
- We have not been able to run all workloads and configurations with the same simulator code, so we can only present qualitative results.
- Our next plot illustrates the most prevalent trends we have seen across workloads.
Results

CMP (4 x 4) vs. MP (4 x 1) Performance

Speedup

Prefetch Degree

baseline
4 x 4 CMP
4 x 1 MP
Ongoing Work

• Divide prefetches and demand fetch resources.
• Put prefetches in separate message buffers.
• Add support to drop prefetches.
• Prefetch buffers are a finite sized circular queue, allowing newer prefetch requests to overwrite stale prefetch requests.
• Do not allocate TBE entries for prefetches.
• Do cache replacement once prefetched data has returned.
• Incorporate state of cache lines into prefetch algorithm.
Ongoing Work

- Add Global History Buffer CZone / Delta Correlation (GHB C/DC) prefetching [Nesbit, Dhodapkar and Smith].
- GHB C/DC can prefetch more abstract access patterns and has better prefetch accuracy.
Questions?