Reducing Memory Bandwidth for Chip-Multiprocessors using Cache Injection

Edgar A. León and Arthur B. Maccabe
University of New Mexico

Scalable Systems Lab

Introduction

- CMPs significantly increase memory bandwidth pressure.
- Increasing disparity between memory and processor speeds (memory wall).
- Large on-chip caches.

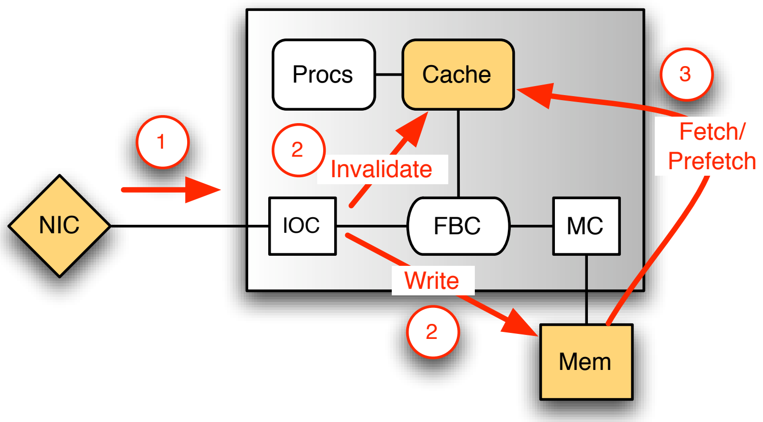
Decrease memory bandwidth utilization using cache injection of incoming network data.

- Evaluated cache injection on memory bandwidth.
- Compared cache injection with prefetching.

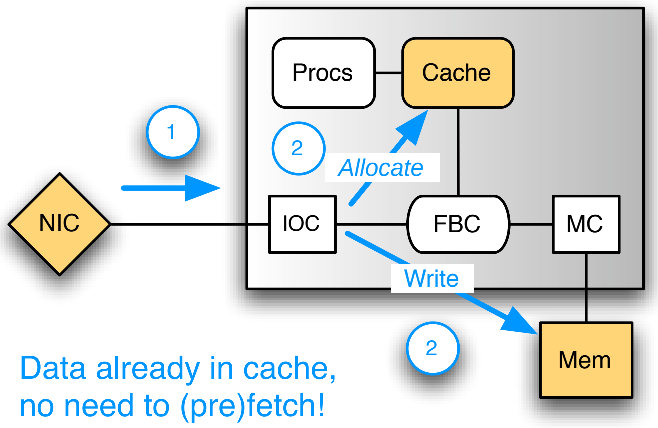
What is Cache Injection?

- Technique to reduce memory latency and memory bandwidth utilization on incoming network data.
- Data is moved directly from the NIC to a processor's cache.

NIC Memory Write Operation



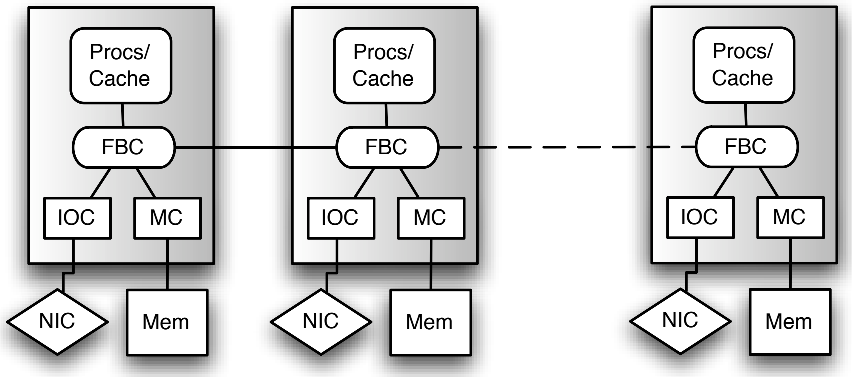
NIC *Cache Injection* Operation



Data already in cache, no need to (pre)fetch!

Architectural Assumptions

Based on IBM Power5 chip-multiprocessor architecture (procs/cache box is expected to become NUCA):



Experimental Evaluation

Using an OS-bypass, zero-copy communication system and a user level micro-benchmark:

- Quantify the effect of cache injection on memory bandwidth and execution time.
- Compare cache injection with prefetching.

Micro-benchmark reads sequentially each word of incoming network messages.

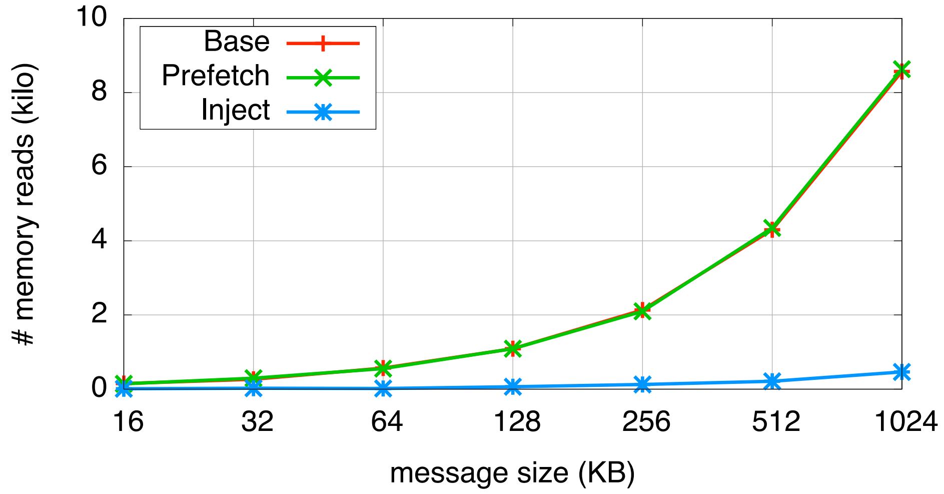
Simulation env.	Mambo: IBM PowerPC full-system simulator
OS	IBM K42 research OS.
Communication	OS-bypass, zero-copy UDP implementation
Processor freq.	1.65 GHz
L1 I/D cache	64KB/32KB, 2-way/4-way, 128B line
L2 cache	1.875MB, 3-slice, 10-way, 10 cycle latency
L3 cache	36MB, 3-slice, 12-way, 80 cycle latency
Main memory	512MB, 230 cycle latency

Acknowledgments

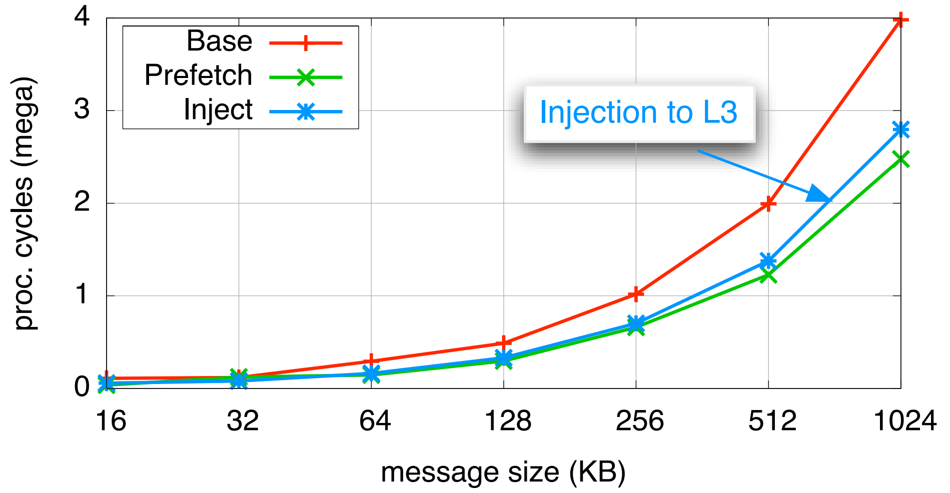
Scalable Systems Lab, UNM.
Novel Systems Architecture Group, IBM ARL.
Dr. Orran Krieger and the K42 team, IBM WRC.
This work was partially supported by an IBM grant.

Experimental Results

Memory Bandwidth



Run Time



Related Work

- Bohrer et al. Method and apparatus for accelerating I/O processing using cache injections. 2004. US Patent 6,711,650 B1.
- Huggahalli et al. Direct cache access for high bandwidth network I/O. ISCA 2005.

Conclusions and Future Work

- Cache injection reduces memory bandwidth dramatically on accesses to network data.
- Cache injection improves application performance and performs comparably with prefetching.
- Study cache injection for NUCA architectures and injection policies based on OS and compiler info.