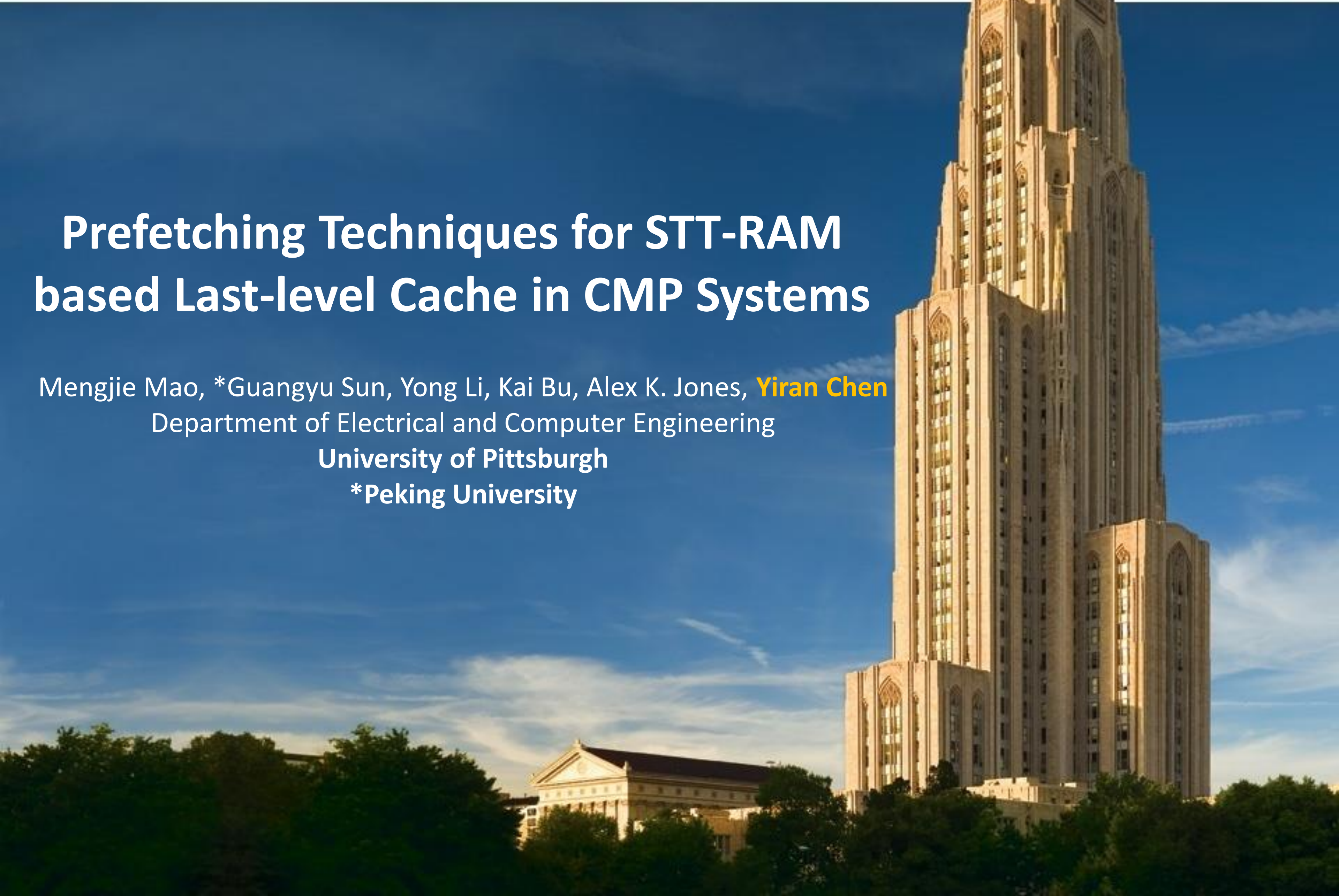
University of Pittsburgh

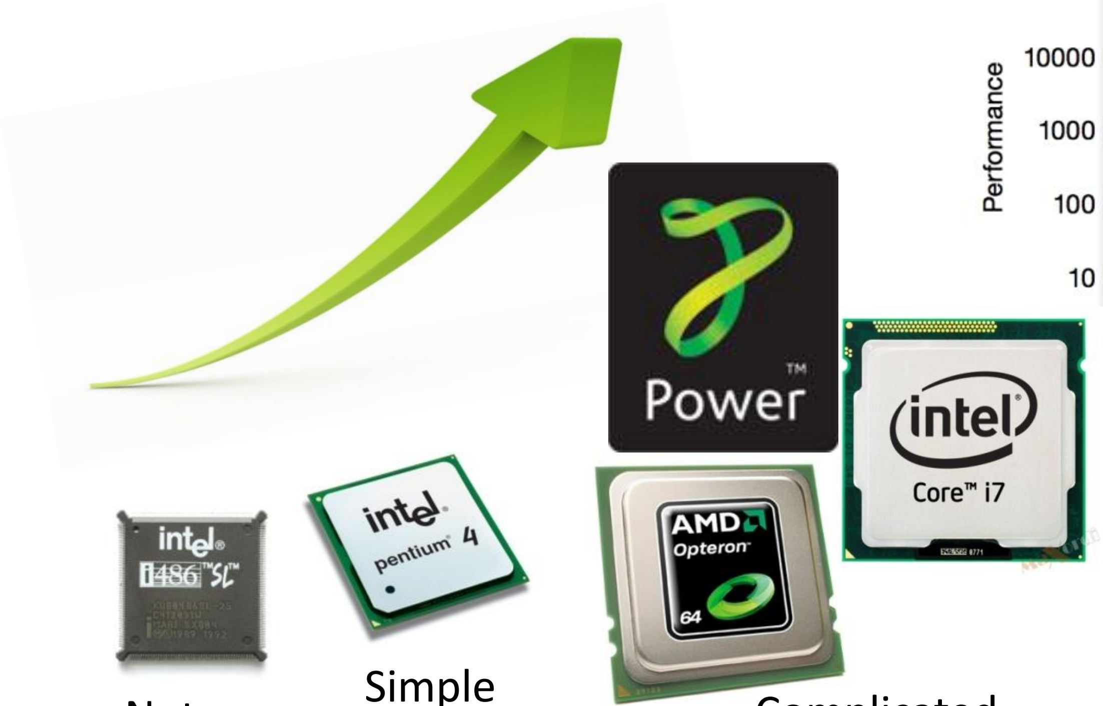


Prefetching Techniques for STT-RAM based Last-level Cache in CMP Systems

Mengjie Mao, *Guangyu Sun, Yong Li, Kai Bu, Alex K. Jones, **Yiran Chen**
Department of Electrical and Computer Engineering
University of Pittsburgh
***Peking University**

Introduction

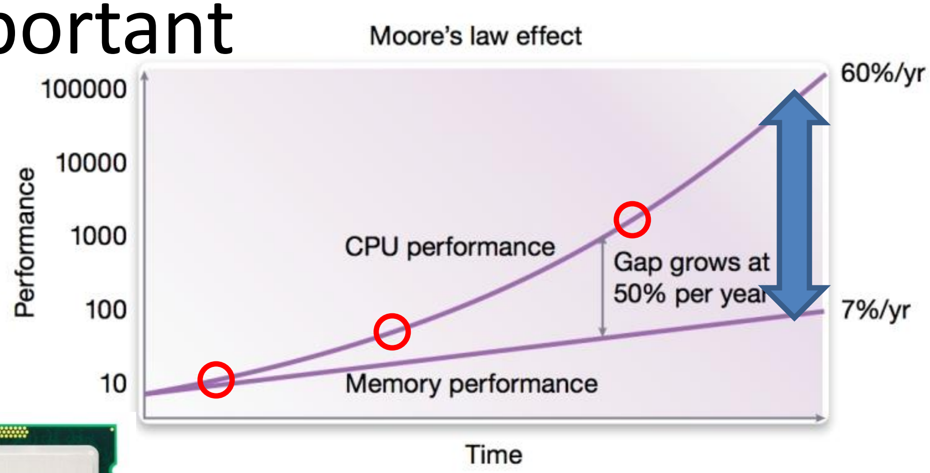
- Data prefetching is important



Not any

Simple
Software-aided

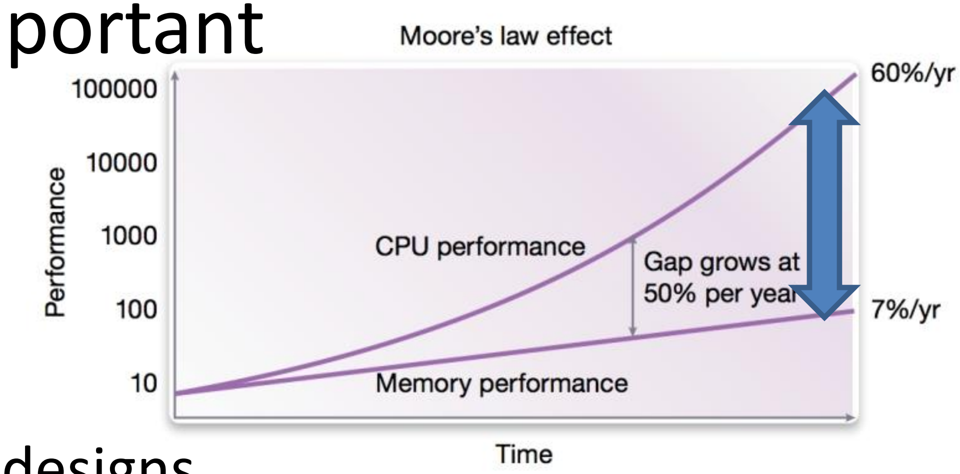
Complicated
Heterogeneous
Software/hardware



Prefetch data from memory to on-chip cache

Introduction

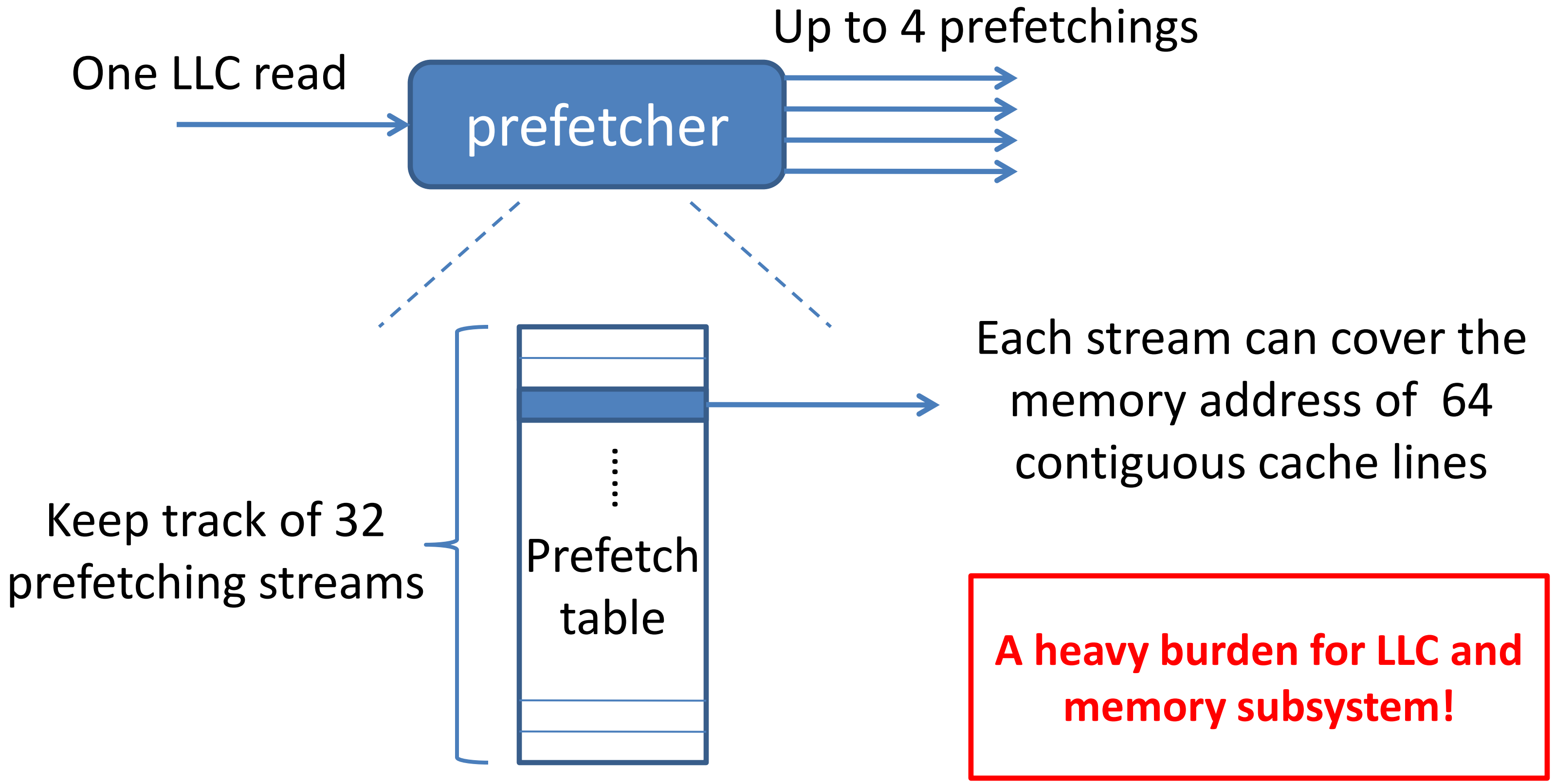
- Data prefetching is important



- But far from efficient
  - Average accuracy seldom exceeds 60% across most designs
  - 40% prefetching is useless
  - Wasted memory bandwidth
  - Access conflict and cache pollution
- Sensitive to write latency

Introduction

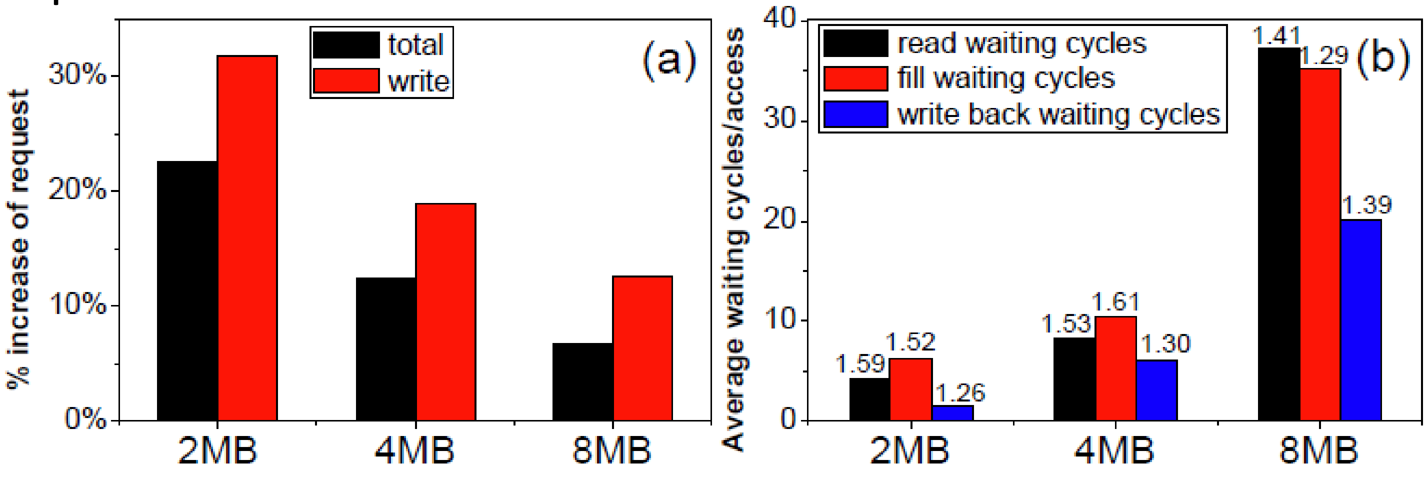
- A example—Stream prefetcher in IBM POWER4 microprocessor


Up to 4 prefetchings
One LLC read
prefetcher
Prefetch table
Keep track of 32 prefetching streams
Each stream can cover the memory address of 64 contiguous cache lines

A heavy burden for LLC and memory subsystem!

Introduction

- Spin-Transfer Torque Random Access Memory (STT-RAM) as Last-Level Cache (LLC)
  - High write cost of STT-RAM, ~ 10ns
  - Combine STT-RAM based LLC with data prefetching
  - High-density/small-size STT-RAM cell ->bank access conflicts
  - Low-density/large-size STT-RAM cells->prefetching-incurred cache pollution



Introduction

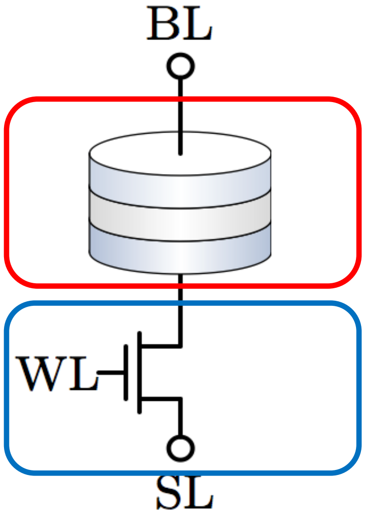
- Our work
  - Reduce the negative impact on system performance of CMPs with STT-RAM based LLC induced by data prefetching
  - Request Prioritization
  - Hybrid local-global prefetch control

Outline

- Motivation
- STT-RAM Basics
- Methodology
  - Request prioritization
  - Hybrid local-global prefetch control
- Result
- Conclusion

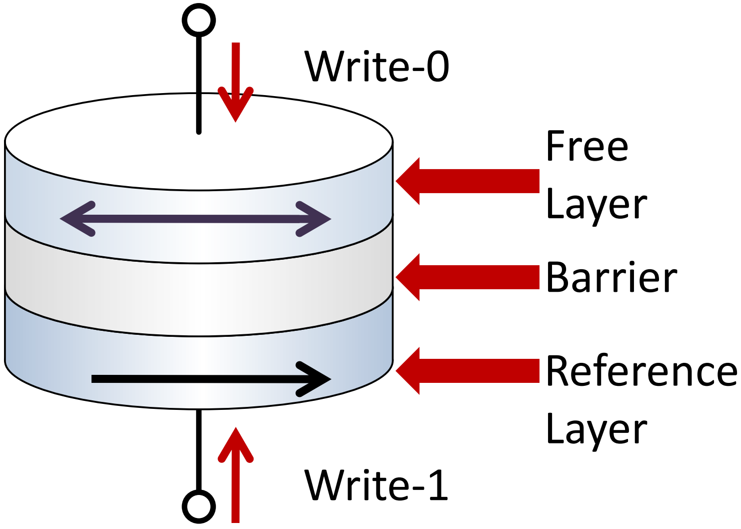
STT-RAM Basics

- STT-RAM Cell:
  - Transistor and MTJ (Magnetic Tunnel Junction);
  - Denoted as 1T-1J cell.
- MTJ:
  - Free Layer and Ref. Layer;
  - Read: Direction → Resistance;
  - Write: Current → Direction.



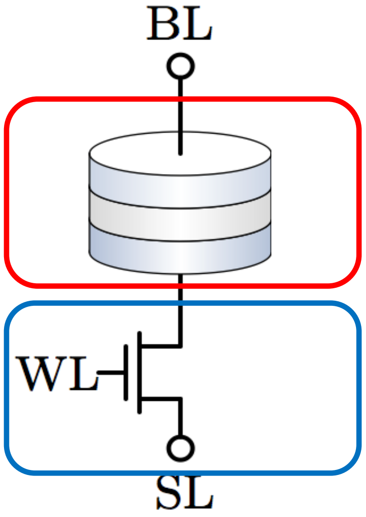
Parallel (R_{Low}), 0

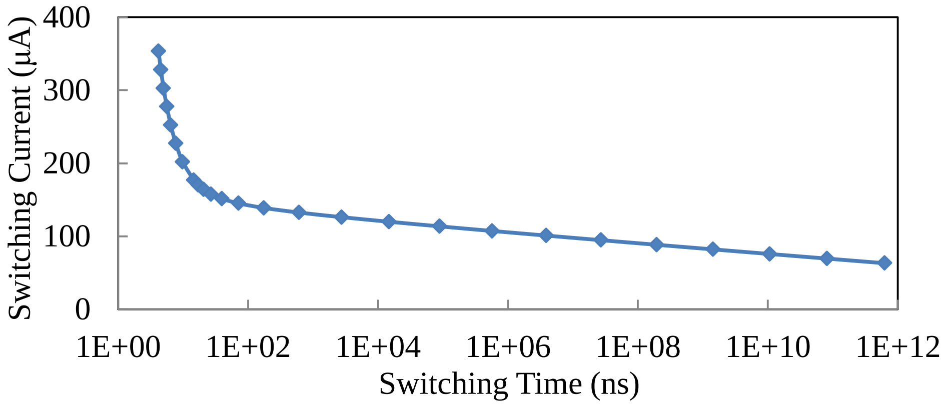
Anti-Parallel (R_{High}), 1

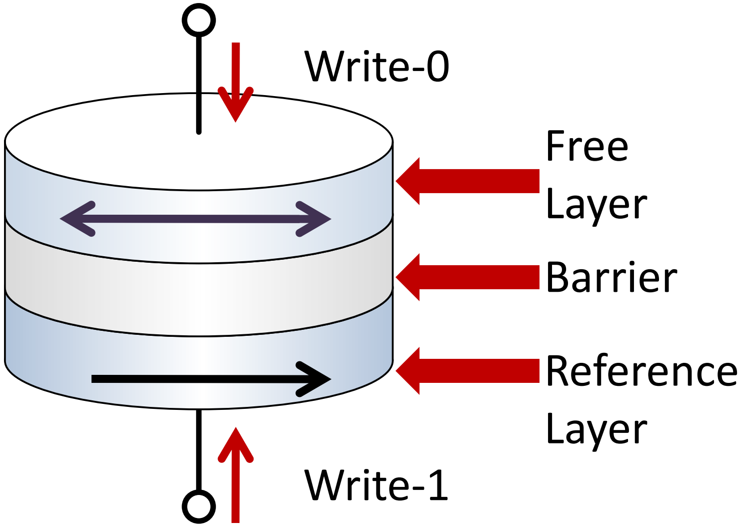


STT-RAM Basics

- STT-RAM Cell:
  - Transistor and MTJ (Magnetic Tunnel Junction);
  - Denoted as 1T-1J cell.
- MTJ:
  - Free Layer and Ref. Layer;
  - Read: Direction → Resistance;
  - Write: Current → Direction.
- Write latency VS. write current







Outline

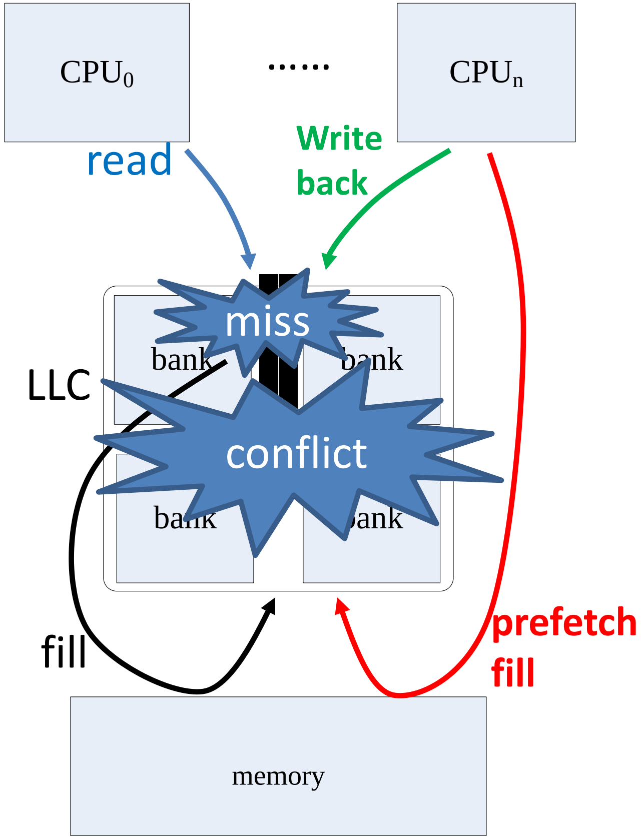
- Motivation
- STT-RAM Basics
- Methodology
  - Request Prioritization
  - Hybrid local-global prefetch control
- Result
- Conclusion

Request Prioritization(RP)

- Various LLC access requests
  - Access conflict
  - Further aggravated due to long write latency

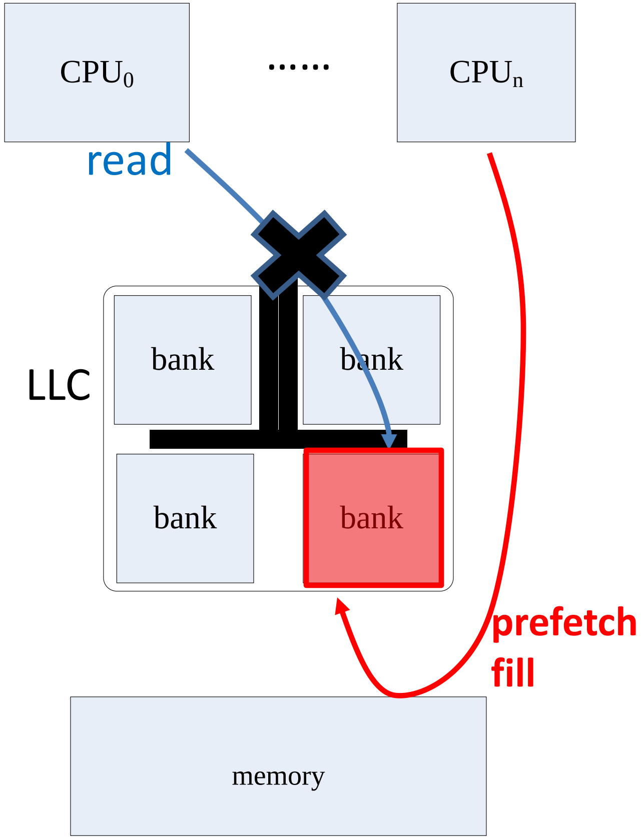
1. Read 2. Fill 3. Write back 4. Prefetch fill

- LLC access conflict--the source of performance degradation



Request Prioritization(RP)

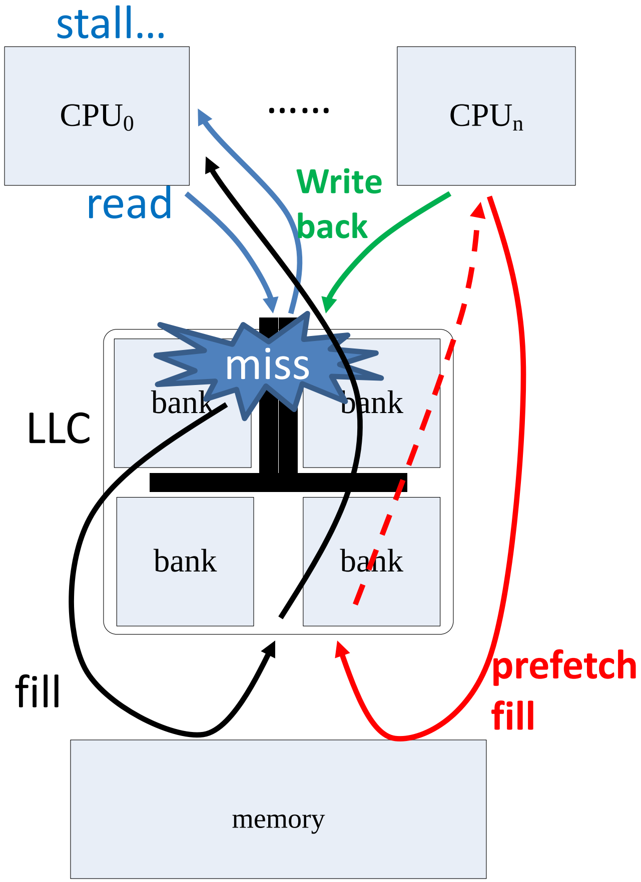
- Various LLC access requests
  - Access conflict
  - Further aggravated due to long write latency
- LLC access conflict--the source of performance degradation
  - Critical request is blocked by non-critical Request



Request Prioritization(RP)

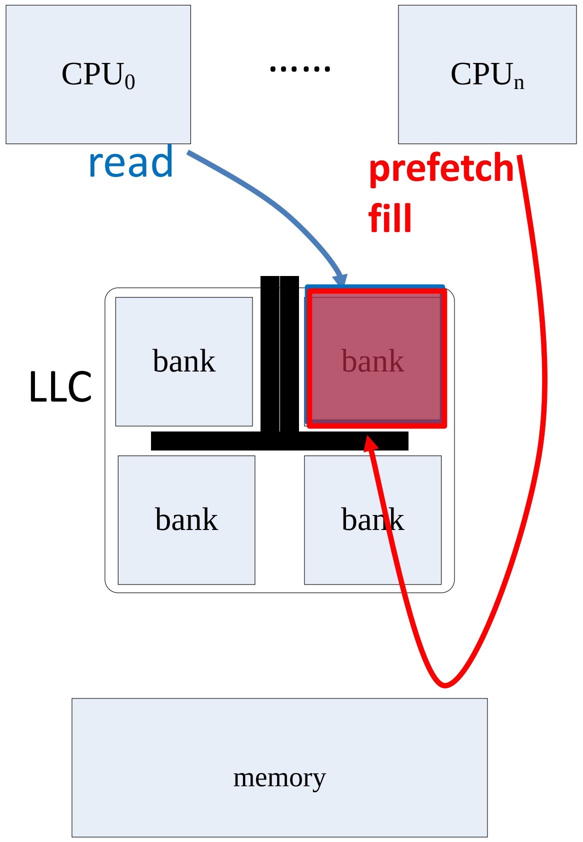
- Various LLC access requests
  - Access conflict
  - Further aggravated due to long write latency

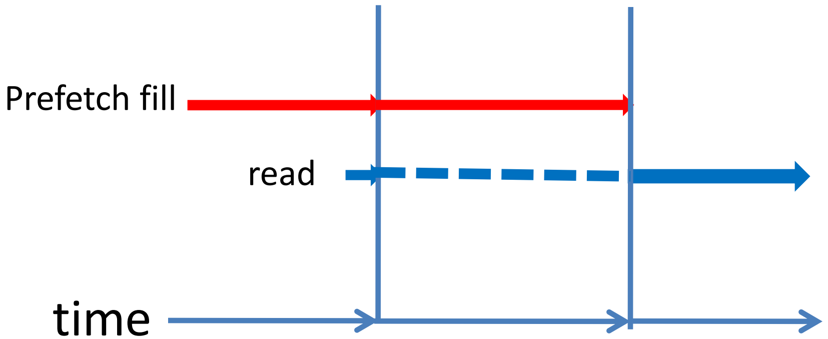
- LLC access conflict--the source of performance degradation
  - Critical request is blocked by non-critical Request
  - Read is the most critical
  - Fill is slightly less critical than read
  - Prefetch fill is less critical than read/fill
  - Write back is the least critical



Request Prioritization(RP)

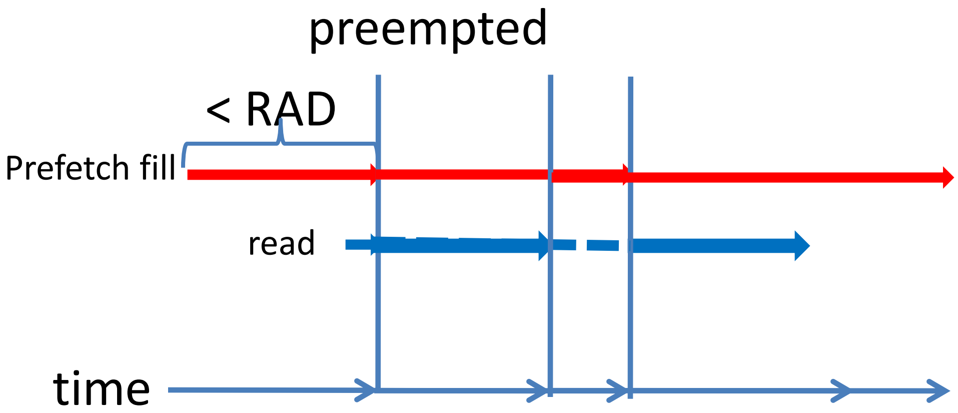
- Assign priority to individual request based on its criticality
  - Respond to the request based on its priority
- High-priority request is blocked by low priority-request

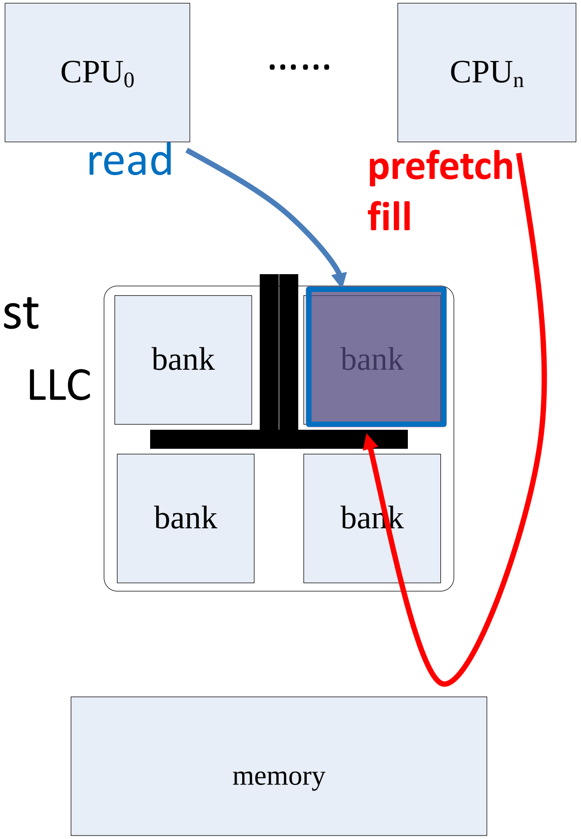




Request Prioritization(RP)

- Assign priority to individual request based on its criticality
  - Respond to the request based on its priority
- High-priority request is blocked by low priority-request
  - Retirement accomplishment degree (RAD)
  - Determines at what stage a low-priority request can be preempted by high-priority request





Outline

- Motivation
- STT-RAM Basics
- Methodology
  - Request prioritization
  - Hybrid local-global prefetch control
- Result
- Conclusion

Hybrid Local-global Prefetch Control (HLGPC)

- The prefetching efficiency is affected by the capacity (cell size) of STT-RAM based LLC
  - A large cell size alleviates bank conflict, by reducing the blocking time of write operations
  - Cache pollution incurred by prefetching also becomes severer due to the reduced total capacity
- Prefetch control considering LLC access contention
  - Dynamically control the aggressiveness of the prefetchers
  - Tune the prefetch distance/prefetch degree

Hybrid Local-global Prefetch Control (HLGPC)

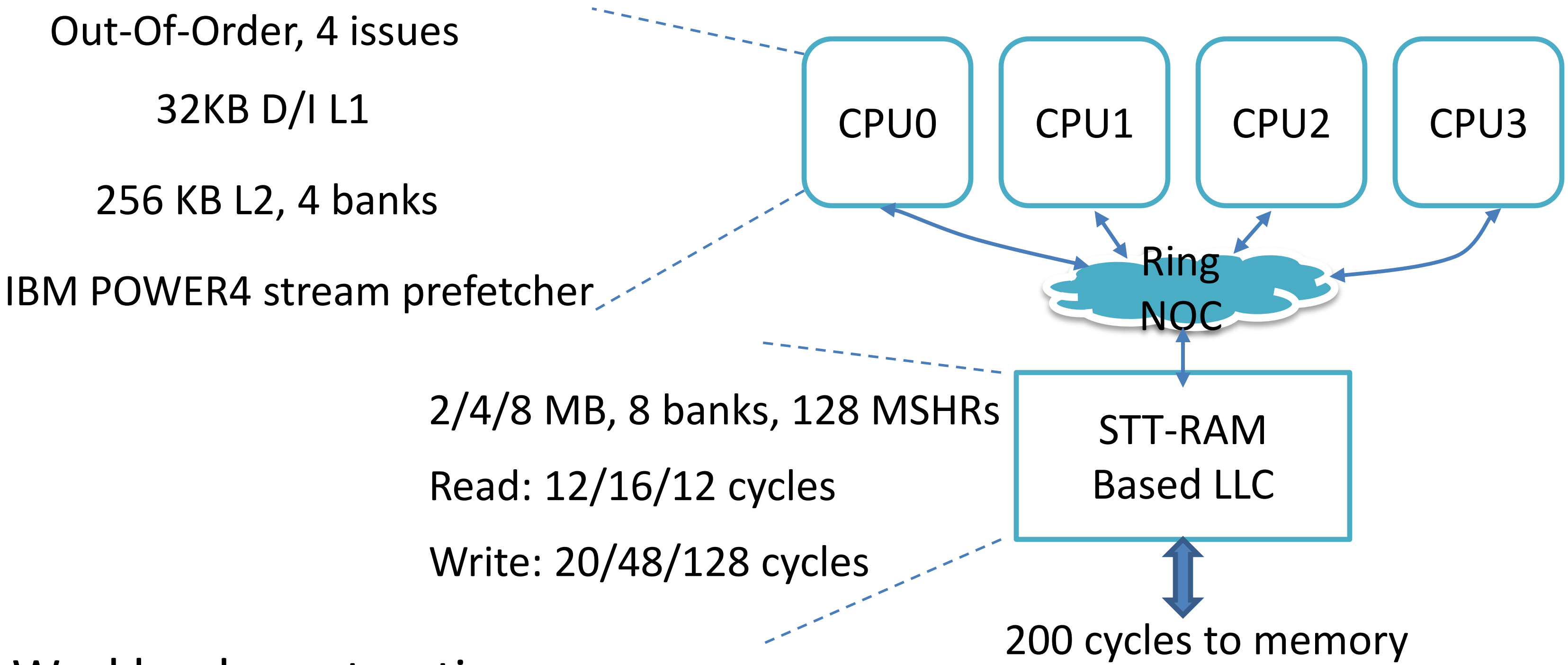
- Local (per core) prefetch control (LPC)
  - Focuses on maximizing the performance of each CPU core
- Hybrid Local-global Prefetch Control (HLGPC)
  - Achieve a balanced dynamic aggressiveness control
  - At core-level, feedback directed prefetching (FDP)* as the LPC
  - At chip-level, GPC may retain or override the decision of LPC based on the runtime information of LLC

Case	Core i's runtime information		LLC's info	Dicision
	Pref. Accuracy	Pref. Frequency	LLC acc. Freq.	
1	Low	High	High	Force scale down
2	High	High	High	Disable scale up
3	Low	High	Low	Disable scale up
4-8	-	-	-	Allow local decision

Outline

- Motivation
- STT-RAM Basics
- Methodology
  - Request prioritization
  - Hybrid local-global prefetch control
- Result
- Conclusion

Experiment Methodology



Out-Of-Order, 4 issues

32KB D/I L1

256 KB L2, 4 banks

IBM POWER4 stream prefetcher

2/4/8 MB, 8 banks, 128 MSHRs

Read: 12/16/12 cycles

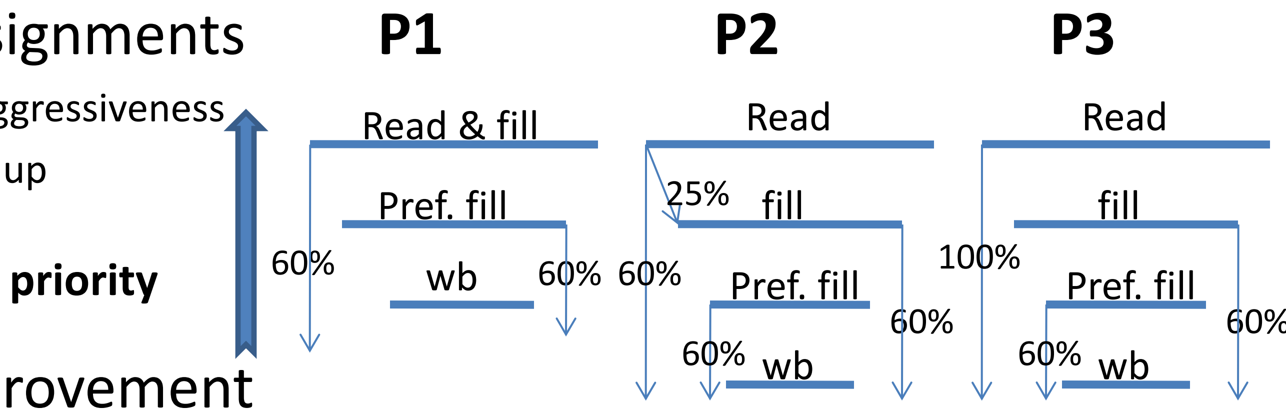
Write: 20/48/128 cycles

200 cycles to memory

- Workload construction
  - 2 SPEC CPU 2000 & 16 SPEC CPU 2006
  - 6 4-app workloads: at least 2 memory intensive, 1 prefetch intensive, 1 memory non-intensive

Enhancement by RP

- Three priority assignments
  - From P2 to P4, the aggressiveness of read request goes up



priority

P1

Read & fill

Pref. fill

wb

60%

60%

P2

Read

25%

fill

Pref. fill

wb

60%

60%

60%

P3

Read

fill

Pref. fill

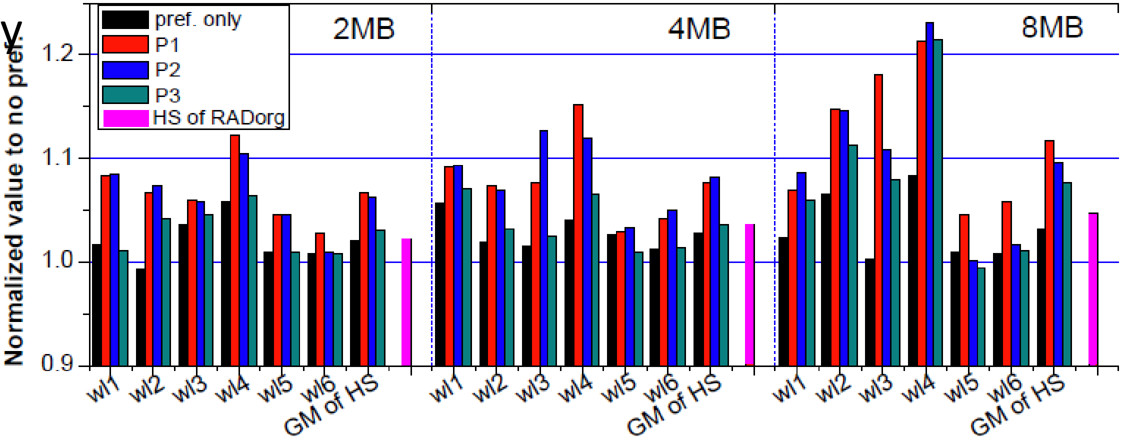
wb

100%

60%

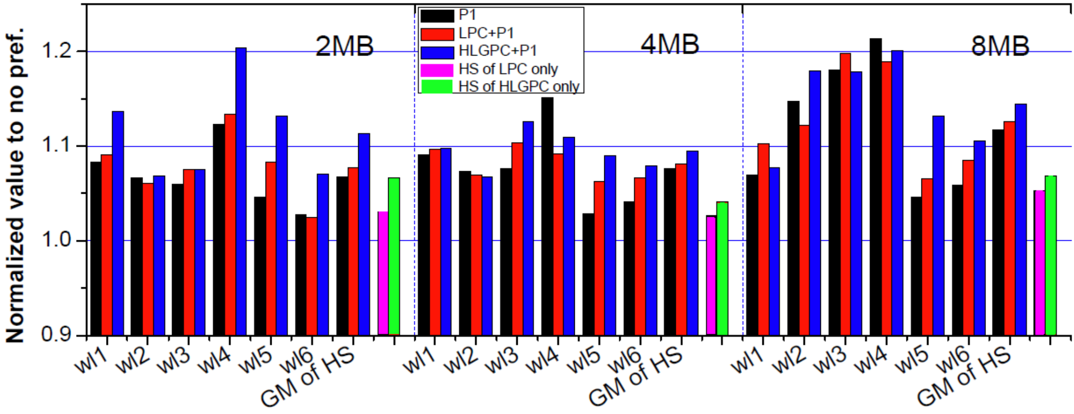
60%

- Performance improvement
  - Prioritizing LLC access requests always achieve system performance improvement
  - Achieves more substantial performance improvement for large LLC with long write access latency
  - The highest performance is achieved by P1

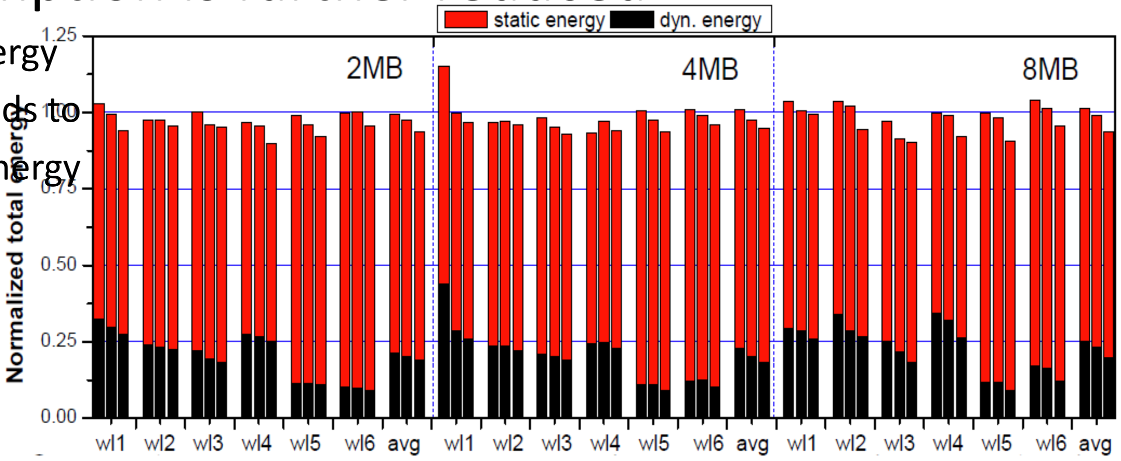


Effectiveness of HLGPC

- Performance improvement
  - *LPC alone* achieves little improvement
  - *HLGPC alone* is better
  - HLGPC+P1 achieves the highest improvement



- The total energy consumption is further reduced
  - Further decrease of dynamic energy
  - Much shorter execution time leads to continuously reduction of leakage energy
  - EDP is improved by 7.8%



1st: P1; 2nd: LPC+P1; 3rd: HLGPC+P1

Outline

- Motivation
- STT-RAM Basics
- Methodology
  - Request prioritization
  - Hybrid local-global prefetch control
- Result
- Conclusion

Conclusion

- In CMP systems with aggressive prefetching, STT-RAM based LLC suffers from increased LLC cache latency due to higher write pressure and cache pollution
- Request prioritization can significantly mitigate the negative impact induced by bank conflicts on large LLC
- Coupling GPC and LPC can alleviate the cache pollution on small LLC
- RP+HLGPC unveils the performance potential of prefetching in CMP systems
- System performance can be improved by 9.1%, 6.5%, and 11.0% for 2MB, 4MB, and 8MB STT-RAM LLCs; the corresponding LLC energy consumption is also saved by 7.3%, 4.8%, and 5.6%, respectively.

Q&A

- Thank you