ECE/CS 752: Midterm 2 Review

ECE/CS 752 Fall 2017

Prof. Mikko H. Lipasti
University of Wisconsin-Madison

Lecture notes based on notes by John P. Shen
Updated by Mikko Lipasti

Midterm 2 Review Topics

- Advanced Caches (Lect 11)
- Main Memory (Lect 12)
- Advanced Microarchitecture (Lect 13)
- Multiple Threads (Lect 14)

Readings-Advanced Caches

- Read on your own:
  - Review: Shen & Lipasti Chapter 3
  - W.-H. Wang, J.-L. Baer, and H. M. Levy. "Organization of a two-level virtual-real cache hierarchy," Proc. 16th ISCA, pp. 140-148, June 1989 (B6) Online PDF
  - Read Sec. 1, skim Sec. 2, read Sec. 3: Bruce Jacob, "The Memory System: You Can't Avoid It, You Can't Ignore It, You Can't Fake It," Synthesis Lectures on Computer Architecture 2009 4:1, 1-77. Online PDF
- To be discussed in class:
  - Review #1 due 11/1/2017: Andreas Sembrant, Erik Hagersten, David Black-Schaffer, "The Direct-to-Data (D2D) cache: navigating the cache hierarchy with a single lookup," Proc. ISCA 2014, June 2014.. Online PDF
  - Review #2 due 11/3/2017: Jishen Zhao, Sheng Li, Doe Hyun Yoon, Yuan Xie, and Norman P. Jouppi. 2013. Kiln: closing the performance gap between systems with and without persistence support. In Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-46). ACM, New York, NY, USA, 421-432. Online PDF
  - Review #3 due 11/6/2017: T. Shaw, M. Martin, A. Roth, "NoSQ: Store-Load Communication without a Store Queue," in Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, 2006. Online PDF

Advanced Memory Hierarchy

- Coherent Memory Interface
- Evaluation methods
  - Analytical, trace-driven, execution-driven. Simpoint selection
- Better miss rate: skewed associative caches, victim caches
- Reducing miss costs through software restructuring
- Beyond simple blocks
- Two level caches

Readings-Main Memory

- Read on your own:
  - Review: Shen & Lipasti Chapter 3
  - W.-H. Wang, J.-L. Baer, and H. M. Levy. "Organization of a two-level virtual-real cache hierarchy," Proc. 16th ISCA, pp. 140-148, June 1989 (B6) Online PDF
  - Read Sec. 1, skim Sec. 2, read Sec. 3: Bruce Jacob, "The Memory System: You Can't Avoid It, You Can't Ignore It, You Can't Fake It," Synthesis Lectures on Computer Architecture 2009 4:1, 1-77. Online PDF
- To be discussed in class:
  - Review #1 due 11/1/2017: Andreas Sembrant, Erik Hagersten, David Black-Schaffer, "The Direct-to-Data (D2D) cache: navigating the cache hierarchy with a single lookup," Proc. ISCA 2014, June 2014.. Online PDF
  - Review #2 due 11/3/2017: Jishen Zhao, Sheng Li, Doe Hyun Yoon, Yuan Xie, and Norman P. Jouppi. 2013. Kiln: closing the performance gap between systems with and without persistence support. In Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-46). ACM, New York, NY, USA, 421-432. Online PDF
  - Review #3 due 11/6/2017: T. Shaw, M. Martin, A. Roth, "NoSQ: Store-Load Communication without a Store Queue," in Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, 2006. Online PDF

Outline: Main Memory

- DRAM chips
- Memory organization
  - Interleaving
  - Banking
- Memory controller design
- Hybrid Memory Cube
- Persistent Memory
- Virtual memory
- TLBs
- Interaction of caches and virtual memory (Wang et al.)
- Large pages, virtualization

Readings-Adv. Microarch.

- Read on your own:
  - I. Kim and M. Lipasti, "Understanding Scheduling Replay Schemes," in Proceedings of the 10th International Symposium on High-performance Computer Architecture (HPCA-10), February 2004.
  - Srikanth Srinivasan, Ravi Rajwar, Haitham Akkary, Amit Gandhi, and Mike Upton, "Continual Flow Pipelines", in Proceedings of ASPLOS 2004, October 2004.
  - Ahmed S. Al-Zawawi, Vimal K. Reddy, Eric Rotenberg, Haitham H. Akkary, "Transparent Control Independence," in Proceedings of ISCA-34, 2007.
- To be discussed in class:
  - Review by 11/6/2017: T. Shaw, M. Martin, A. Roth, "NoSQ: Store-Load Communication without a Store Queue, " in Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, 2006.
  - Review by 11/8/2017: Andreas Sembrant et al., "Long term parking (LTP): criticality-aware resource allocation in OOO processors," Proc of MICRO-48, December 2015.
  - Review by 11/10/2017: Arthur Perais and André Seznec. 2014. EOLE: paving the way for an effective implementation of value prediction. In Proceeding of the 41st Annual International Symposium on Computer Architecture (ISCA '14). IEEE Press, Piscataway, NJ, USA, 481-492. Online PDF

Advanced Microarchitecture

- Memory Data Flow
  - Scalable Load/Store Queues
  - Memory-level parallelism (MLP)
- Register Data Flow
  - Instruction scheduling overview
    - Scheduling atomicity
    - Speculative scheduling
    - Scheduling recovery
  - EOLE: Effective Implementation of Value Prediction
- Instruction Flow
  - Revolver: efficient loop execution
  - Transparent Control Independence

Readings-Multiple Threads

- Read on your own:
  - Shen & Lipasti Chapter 11
  - G. S. Sohi, S. E. Breach and T.N. Vijaykumar. Multiscalar Processors, Proc. 22nd Annual International Symposium on Computer Architecture, June 1995.
  - Dean M. Tullsen, Susan J. Eggers, Joel S. Emer, Henry M. Levy, Jack L. Lo, and Rebecca L. Stamm. Exploiting Choice: Instruction Fetch and Issue on an Implementable Simultaneous Multithreading Processor, Proc. 23rd Annual International Symposium on Computer Architecture, May 1996 (B5)
- To be discussed in class:
  - Review #6 due 11/17/2017: Y.-H. Chen, J. Emer, V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," International Symposium on Computer Architecture (ISCA), pp. 367-379, June 2016. Online PDF

Executing Multiple Threads

- Thread-level parallelism
- Synchronization
- Multiprocessors: coherence, consistency
- Explicit multithreading
- Data parallel architectures
- Multicore interconnects
- Implicit multithreading: Multiscalar
- *Niagara case study*