CS7290 Advanced Microarchitecture

Fall 2016

[Home][Course overview][Schedule][Reading]

Instructor: Prof. Hyesoon Kim
Email: hyesoon at cc dot gatech dot edu
Office hours: M/W 4:00-4:30 (or by appointment)
Office: KACB 2344
Class time: 4:35-5:55 MW
Class room: KACB 2456

CS7290 Advanced Microarchitecture

Fall 2016

[Home][Course overview][Schedule][Reading]

Course Description
This is a graduate-level advanced computer architecture course In this course, we will extend our knowledge about various topics in computer architecture. This course will cover latest computer architecture papers and advanced topics in microarchitecture, memory systems, GPU architectures, performance/power modeling, and emerging architectures. This course is an open-ended project-based course in which you will choose your topic. The project topics can be research oriented or implementation oriented. It can be optimizing architecture with a timing simulator, developing a timing simulator, implementing an architecture on FPGA boards, developing performance/power models, etc.

Topics:
- Micro-architecture
- Memory systems
- GPU architectures
- Performance, power modeling
- Emerging architectures

Pre-requisite:
CS6290/CS4290/ECE6100/ECE4100 or an equivalent course **A**

Textbooks: There is no required textbook for the course.

Course Home Pages: http://www.cc.gatech.edu/~hyesoon/fall16/cs7290/index.html

Grading Scheme

Lecture Scribing (5%)
In-class quiz/participation (15%)
Project (60%)
Mid-term exam (20%)

Final grade algorithm:
90 ~ 100 (out of 100) or Outstanding project: A
80 ~ 90 or good project : B
70 ~ 80 and completed project: C
Otherwise F

- **Lecture scribing** Each student is responsible for scribing the lecture notes.

- **Class participation** Class discussion is one of the important part of the class. Some discussions will be lead by students.

- **Student Honor Code:** We encourage students to discuss and collaborate together. However, Zero tolerance toward a violation of the student honor code. If you have any questions, please ask to the instructor.

- **Office Hours:** Please respect the office hours of the instructor by planning ahead. Other times are possible by appointments.

- Student Information : Please complete this form by Aug 26th Friday 6 pm .

CS7290 Advanced Microarchitecture

Fall 2016

[Home][Course overview][Schedule][Reading]

Tentative Schedule

Week	Dates	Topics	Readings	MISC
1	8/22, 8/24	Overview and classical machines		
2	8/29, 8/31	front-end optimizations	[TRACE1, BR1]	
3	9/5, 9/7	front-end optimizations		No class (9/5)
4	9/12, 9/14	Scheduling		
5	9/19, 9/21	Microarchitecture in pipeline		
6	9/26, 9/28	First project presentations/Optimizing memory hierarchy	[RDI][PIF] [UCP][AIP] [TAP] [COM] [COM2] [TLB1]	
7	10/3,10/5	Coherence/consistency	[WPW] [COH1] [CSB][PH]	guest lecture
8	10/10, 10/12	Coherence/consistency		No lecture on 10/10 (fall recess)
9	10/17, 10/19	coherence/consistency, TLB optimization		
10	10/24, 10/26	GPU/vector architectures		
11	10/31, 11/2	GPU/vector architectures/Project 2nd presentations		
12	11/7, 11/9	3D stacking architecture		
13	11/14, 11/16	Performance & Power modeling	[MCP] [PGPU]	
14	11/21, 11/23	mid-term		No class on 11/23
15	11/28, 11/30	non-conventional architecture		
16	12/5, 12/7	Project presentations		

CS7290 Advanced Microarchitecture

Fall 2016

[Home][Course overview][Schedule][Reading]

Reading Papers

- Front-end and Branch Predictors
  - [TRACE1] Eric Rotenberg, Steve Bennett, Jim Smith. Trace Cache: a Low Latency Approach to High Bandwidth Instruction Fetching. April, 1996
  - [BR1] Marius Evers, Sanjay J. Patel, Robert S. Chappell, and Yale N. Patt. 1998. An analysis of correlation and predictability: what makes two-level branch predictors work. In Proceedings of the 25th annual international symposium on Computer architecture (ISCA '98),
  - Jimenez, D.A; Lin, C., "Dynamic branch prediction with perceptrons," High-Performance Computer Architecture, 2001. HPCA. The Seventh International Symposium on , vol., no., pp.197,206, 2001
  - [RDI] RDIP: Return-address-stack Directed Instruction Prefetching (MICRO13)
  - [PIF] Proactive Instruction Fetch (MICRO11)
- Cache Optimizations
  - [UCP] Utility-Based Cache Partitioning: A Low-Overhead, High-Performance, Runtime Mechanism to Partition Shared Caches. Moinuddin K. Qureshi and Yale N. Patt. MICRO'06.
  - [AIP] Moinuddin K. Qureshi, Aamer Jaleel, Yale N. Patt, Simon C. Steely, and Joel Emer. 2007. Adaptive insertion policies for high performance caching. In Proceedings of the 34th annual international symposium on Computer architecture (ISCA '07).
  - Aamer Jaleel, Kevin B. Theobald, Simon C. Steely, Jr., and Joel Emer. < 2010. High performance cache replacement using re-reference interval prediction (RRIP). In Proceedings of the 37th annual international symposium on Computer architecture (ISCA '10)
  - [TAP] Jaekyu Lee; Hyesoon Kim, "TAP: A TLP-aware cache management policy for a CPU-GPU heterogeneous architecture," High Performance Computer Architecture (HPCA), 2012
  - [ACC]Alaa R. Alameldeen and David A. Wood. 2004. Adaptive Cache Compression for High-Performance Processors. In Proceedings of the 31st annual international symposium on Computer architecture(ISCA '04)
  - [COM] Alaa R. Alameldeen and David A. Wood. 2004. Adaptive Cache Compression for High-Performance Processors. In Proceedings of the 31st annual international symposium on Computer architecture (ISCA '04)
  - [COM2] Magnus Ekman and Per Stenstrom. 2005. A Robust Main-Memory Compression Scheme. In Proceedings of the 32nd annual international symposium on Computer Architecture (ISCA '05)
- Prefetching
  - [WPW]Jaekyu Lee, Hyesoon Kim, and Richard Vuduc, When Prefetching Works, When It Doesn't, and Why , TACO 2012
- TLB

- [TLB1] Gokul B. Kandiraju and Anand Sivasubramaniam. 2002. Going the distance for TLB prefetching: an application-driven study. In Proceedings of the 29th annual international symposium on Computer architecture (ISCA '02). IEEE Computer Society, Washington,
- Abhishek Bhattacharjee and Margaret Martonosi. 2010. Inter-core cooperative TLB for chip multiprocessors. In Proceedings of the fifteenth edition of ASPLOS on Architectural support for programming languages and operating systems (ASPLOS XV).
- Shekhar Srikantaiah and Mahmut Kandemir. 2010. Synergistic TLBs for High Performance Address Translation in Chip Multiprocessors. In Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '43)
- Abhishek Bhattacharjee, Daniel Lustig, and Margaret Martonosi. 2011. Shared last-level TLBs for chip multiprocessors. In Proceedings of the 2011 IEEE 17th International Symposium on High Performance Computer Architecture (HPCA '11).
- Binh Pham, Viswanathan Vaidyanathan, Aamer Jaleel, and Abhishek Bhattacharjee. 2012. CoLT: Coalesced Large-Reach TLBs. In Proceedings of the 2012 45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-45)
- Arkaprava Basu, Jayneel Gandhi, Jichuan Chang, Mark D. Hill, and Michael M. Swift. 2013. Efficient virtual memory for big memory servers. In Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA '13).

- Coherence

  - [CSB]Culler and Singh, Parallel Computer Architecture Chapter 5.1 (pp 269 – 283), Chapter 5.3 (pp 291 – 305)
  - [PH] P&H, Computer Organization and Design Chap 5.8
  - [HP] H&P, Computer Architecture, Chap 5.4-5.6 [COH1] Andreas Moshovos. 2005. RegionScout: Exploiting Coarse Grain Sharing in Snoop-Based Coherence. In Proceedings of the 32nd annual international symposium on Computer Architecture (ISCA '05).
  - Daehoon Kim, Jeongseob Ahn, Jaehong Kim, and Jaehyuk Huh. 2010. Subspace snooping: filtering snoops with operating system support. In Proceedings of the 19th international conference on Parallel architectures and compilation techniques (PACT '10).

- Non-conventional architecture

  - [DFA1] Dennis and Misunas, "A Preliminary Architecture for a Basic Data Flow Processor," ISCA 1974.
  - [DFA2]Arvind and Nikhil, "Executing a Program on the MIT Tagged-Token Dataflow Architecture," IEEE TC 1990
  - [ATU] Micron Automata Processor

- GPU architectures

  [BGPU]"Performance analysis and tuning for GPGPUs." Synthesis Lectures on Computer Architecture, Morgan & Claypool

- Power

  [MCP] Sheng Li, Jung Ho Ahn, Richard D. Strong, Jay B. Brockman, Dean M. Tullsen, and Norman P. Jouppi. 2009. McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 42)
  - [PGPU] Sunpyo Hong and Hyesoon Kim. An integrated GPU power and performance model. In Proceedings of the 37th annual international symposium on Computer architecture (ISCA '10).
