CS594 Special Topics:
Large-Scale Computer Systems

Zhiling Lan

December 2023

1 Course Goals

Large-scale computer systems play a pivotal role in both high-performance computing and cloud computing. This class is designed to enable students to keep up with the latest developments in modern computing platforms, with hardware and software working in concert to deliver good levels of performance and efficiency. The lectures cover a broad array of topics including heterogeneous architecture, networking, storage, power management, availability and reliability, resource management, and emerging topics such as machine learning for systems. These concepts are reinforced with research paper readings and hands-on projects that involve computer system design and analysis. By the end of this course, students will:

- Gain a deep understanding of the core principles and technologies behind large-scale computer systems that power high-performance computing and cloud computing.
- Develop practical skills in concerted hardware and software development for massive computing.
- Dive into emerging areas of interest such as machine learning applications for system optimization.
- Apply the knowledge to solve real-world problems through hands-on projects and case studies.

Upon completion of this course, students will be well-prepared to tackle the challenges and opportunities presented by modern computer systems in the context of cloud computing and high-performance computing.

2 Course Materials

A recommended textbook is “Datacenter as a Computer: An Introduction to the Design of Warehouse–scale Machines” by L. Barroso, Jimmy Clidaras, U.

Hoelzle (BCH), 3rd edition, Morgan & Claypool Publishers, 2019. A number of research papers taken from premier cloud computing conferences (e.g., USENIX conferences) and high-performance computing conferences (e.g., SC, HPDC, IPDPS, Cluster) will be used for class reading.

All the reading material will be provided to students as either PDF files or pointers to online resources.

3 Course Outline (Tentative)

Week	Topic	Readings
Core concepts		
Week 1	Software infrastructure	BCH ch. 1,2
Week 2	Hardware architectures	BCH ch. 3,6
Week 3	Power management	BCH ch. 4,5
Week 4	Availability and reliability	BCH ch. 7
Week 5	Storage and file systems	Research papers
Week 6	Networking	Research papers
Week 7	Resource management	Research papers
Week 8	Monitoring and analysis	Research papers
Emerging topics		
Week 9	Emerging architectures	Research papers
Week 10	Predictive analysis	Research papers
Week 11	ML for resource management	Research papers
Week 12	ML for memory/storage	Research papers
Week 13	ML for compiler	Research papers
Week 14	ML for crosscutting	Research papers
Week 15	LLMs for systems	Research papers

4 Course Work

The course is a combination of lectures and paper reading.

- Reading, presenting, discussing research papers.
- A midterm exam.
- A semester-long research project.

5 Prerequisite

Students are expected to have taken and received an “C” or better in “Introduction to High Performance Computing”. We also expect students to have basic understanding of computer systems through a course like CS361 “Systems Programming” and CS461 “Operating Systems”.