CIT 561 Advanced Parallel Data Systems

Catalog Description: **This course provides an introduction to the techniques and technologies used in high performance and cloud computing. Topics covered in this course will focus on aspects of the design, implementation, and use of high performance systems progressively from the hardware layer through the operating system up to the application level. Topics will include: commodity hardware and novel architectural storage components; the architecture and use of parallel file systems, including PVFS2 and Lustre; reliability and scheduling; virtualization and fault tolerant strategies for Petascale computing; system architectures for data intensive computing and workflows; parallel I/O systems; and grid and cloud computing architectures. The driving outcome for this course is for students to understand and apply advanced high performance computing concepts, architectures, and software components to develop and operate a high performance computing environment.**

Organization and Schedule Credit 3

Semester	*Meeting Type*	*Days*	*Time*	*Location*
Spring '14	Lecture 01	T/TH	10:30-11:45am	Knoy 205

Prerequisites *CIT499M High Performance Computing Systems, or Consent of Instructor Required*

Course Description and/or Theme

The objective of this course is to provide students an introduction to the theory, design, deployment, and use of parallel data systems that support data intensive computing, focused on use for high performance computing and cyberinfrastructure systems. The course is structured to provide an introduction to system architectures used for high performance computing and parallel storage systems, performance considerations of systems and applications, the trends and forces within the computer industry that are driving the development of cost effective storage technologies, the adoption of cluster computing, the design of high performance storage systems from commodity components, and the operation and management of high performance storage systems. As part of the course, students will build a small-scale parallel storage system from commodity components, benchmark their system using microbenchmarks and application benchmarks, and compare the performance of their system with other storage systems. The optional textbook used for this course will be Parallel I/O for High Performance Computing by John M. May, as well as various papers on recent topics. If you need additional help with Linux Systems Administration, you should pick up a text on Linux SysAdmin. The Linux Administration Handbook (2nd edition) by Evi Nemeth would be a good choice. The readings for the course will be from journal and conference papers as well as selected sections from textbooks.

Information Technology Used In This Course

- Linux, specifically Fedora Linux
- Dell Desktops
- Purdue Condor BoilerGrid
- NSF Teragrid Resources (if available)
- Amazon Web Service, Microsoft Azure and OpenStack
- Gigabit Ethernet networking equipment, specifically Force-10 and Netgear
- Inifiniband, Myrinet-2000, and Myrinet-10g networking equipment
- Microsoft *Word, Excel* (for homework and laboratory assignments)
- Virtualization technology

Course Instructors

Name	*Office*	*Phone*
Ioannis Papapanagiotou	Knoy 213	494-4677
Thomas J. Hacker	Knoy 211	494-4465

Recommended Textbooks, Lab Manuals, and Supplies

- May, J., Parallel I/O for High Performance Computing, Morgan Kaufmann Press, 2000, ISBN 1558606645
- Linux Administration Handbook (2nd Edition) by Evi Nemeth, Prentice Hall, 2006, ISBN 0131480049. *This is a reference book on Linux Systems Administration.*
- Distributed and Cloud Computing: From Parallel Processing to the Internet of Things, By K. Hwang et al., Elsevier 2012.

The student who successfully completes this course must:

1. Understand the factors that motivate the use and development of high performance parallel storage systems, the importance of fast data rates to the efficient use of HPC systems and processors, and the effects of poor I/O on application performance.

2. Demonstrate skill in developing high performance storage systems by building, benchmarking, and optimizing a small commodity-based storage system based on the Linux operating system and other open source software packages.

3. Demonstrate understanding and knowledge of the core concepts of high performance storage systems in the context of high performance computing, which include:
    - The storage hierarchy, and the speeds, feeds, and costs at each level of the hierarchy.
    - The technological characteristics of components at each level of the storage hierarchy, and the effects of these technologies on reliability and performance.
    - The range of interconnection technologies available for parallel data systems, and the effects of these technologies on storage performance. The interconnection technologies include: IB, Myrinet, GigE.
    - Historical trends and forces in storage broadly as well as the range of storage technologies available today. This list includes tape technology, magnetic disk, non-volatile RAM, and other types of memory devices.
    - Designing high performance storage systems that are cost effective and reliable.

Course Outline (Subject to change)

Week 1	Introduction and Motivation for Parallel Data Systems	Week 9	Designing High Performance Storage Systems
Week 2	HPC Application I/O Requirements and Data Storage Technology Overview	Week 10	Break
Week 3	Disk and Tape Technologies	Week 11	Introduction to Cloud Computing
Week 4	Memory Devices and Communications	Week 12	Cloud Computing: Basic Architecture
Week 5	Communications and Storage Systems for High Performance Computing	Week 13	Storage on the Cloud
Week 6	Parallel File Systems I	Week 14	Distributed File Systems
Week 7	Parallel File Systems II	Week 15	Database Systems
Week 8	Using Parallel Storage Systems and Reliability and Fault Tolerance	Week 16	Scientific Computing and Workflow Systems