

CS 6504: Advanced Networking for High-Performance Computing

Instructor:	Wuchun FENG Knowledge Works II, Room 2209 feng@cs.vt.edu
Class Time:	Tuesdays and Thursdays at 9:30 – 10:45 AM
Office Hours:	Thursdays 10:45AM – 12:30PM and by appointment
Required Text:	High Performance TCP/IP Networking: Concepts, Issues, and Solutions by Mahbub Hassan and Raj Jain
Prerequisites:	CS 5565 or equivalent CS 4504 or equivalent

Course Topics:

1. System-Area Networks
2. Wide-Area Networks
3. Programming Models for Networking
4. Performance Modeling of Networks

Learning Objectives:

1. Become conversant on approaches to high-performance networking, both in the system area and wide area
2. Learn to critically evaluate current literature in the field.
3. Work with a high-performance networking technology (or technologies), whether in theory, simulation, or implementation.
4. Identify new directions for research in high-performance networking.

Course Requirements:

1. Present and critique selected literature.
2. Read and provide a written critique (prior to presentation) and then participate in oral discussion of selected literature.
3. Propose a course project for in-depth study.
4. Conduct work for the course project, resulting in a 12-15 page technical report (IEEE or ACM camera-ready format).
5. Present course project to class and write-up course project.

Course Pointers on Communication:

1. Presenting
 - a. Oral Presentation Advice: <http://www.cs.wisc.edu/~markhill/conference-talk.html>
 - b. The Short Talk: <http://www.cs.cornell.edu/cv/ShortTalk.htm>
2. Writing
 - a. Writing Advice: <http://www.cs.berkeley.edu/~pattsrn/talks/writingtips.html>
 - b. Writing Technical Articles: <http://www.cs.columbia.edu/~hgs/etc/writing-style.html>

Tentative Topical Outline:

08/21	Administrivia
08/23 – 08/28	Overview of High-Performance Networking & Computing
08/30 – 09/04	What is the Future of High-Performance Networking? Discussion of Topical Interest
09/06 – 10/04	Wide-Area Networks (or System-Area Networks)
10/09 – 11/08	System-Area Networks (or Wide-Area Networks)
11/13 – 11/15	Project Reports
11/27 – 12/06	Potpourri

Grading: (Subject to Change)

10%	Attendance & Participation
10%	Homeworks and Quizzes
15%	Paper Presentations
15%	Written Critiques
10%	Project Proposal: Oral & Written
10%	Project Presentation (30 minutes each)
30%	Project Report

92 - 100	A	(Fractional percentages will be rounded to the nearest decimal place.)
90 - 91	A-	
88 - 89	B+	
82 - 87	B	
80 - 81	B-	
78 - 79	C+	
72 - 77	C	
70 - 71	C-	
68 - 69	D+	
62 - 67	D	
60 - 61	D-	
< 60	F	

Important Dates: (Subject to Change)

10/22/07	Project proposals due, e.g., a white paper with proposed preliminary work and scope. The paper should include the following sections: Motivation, Related Work, Proposed Project & Approach.
11/29/07 +	Project presentations.
12/06/07	Full paper due in IEEE-style format. See http://www.ieee.org/organizations/pubs/confpub/auxfiles/sample_manuscript.pdf

Graduate Honor System (<http://ghs.grads.vt.edu/>)

“The Graduate Honor Code establishes a standard of academic integrity. The code demands a firm adherence to a set of values and is founded on the concept of honesty with respect to the intellectual efforts of oneself and others.

Compliance with the Graduate Honor Code requires that all graduate students exercise honesty and ethical behavior in all their academic pursuits here at Virginia Tech, whether these undertakings pertain to study, course work, research, extension, or teaching.”

In addition to the above, you should be aware of Virginia Tech's University Policy of Class Attendance and the ACM and IEEE Code of Ethics. The tenets of the Graduate Honor Code (<http://ghs.grads.vt.edu>) will be strictly enforced in this course. With respect to collaboration, the following policies apply:

- All graded work is expected to be the original work of the individual student, unless otherwise directed by the instructor.
 - When working on homework assignments, both discussion and cooperative learning are allowed. However, copying is an honor code violation.
 - Projects are to be the work of an individual student or a team of at most two students, unless otherwise approved by the instructor. When working on projects, discussion of general concepts between teams, such as systems calls and software libraries, is permitted. However, discussing solutions to projects, specific code, or textual content in a report is an honor code violation. All source material used in any programming assignment must be properly cited.

Tentative Reading List

Many of the reading assignments will be taken directly from the required textbook – “High Performance TCP/IP Networking” – by Mahbub Hassan and Raj Jain and supplemented by the following reading material.

System-Area Networking

Ethernet vs. Ethernet Technologies

1. Boden et al., “Myrinet: A Gigabit-per-second Local Area Network,” *IEEE Micro*, Jan.-Feb. 1995.
<http://www.myri.com/research/publications/Hot.ps>
2. Petrini et al., “The Quadrics Network (QsNet): High Performance Clustering Technology,” *IEEE Micro*, Jan.-Feb. 2002.
http://www.lanl.gov/radiant/pubs/quadrics/qsnet_ieee-micro.pdf
3. Mellanox Technologies, “Introduction to Infiniband,” *White Paper*.
http://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf
4. Mellanox Technologies, “InfiniBand – Industry Standard Data Center Fabric is Ready for Prime Time,” *White Paper*.
http://www.mellanox.com/pdf/whitepapers/InfiniBand_DataCenter_Ready4PrimeTime_1_0.pdf
5. Hurwitz et al., “End-to-End Performance of 10-Gigabit Ethernet on Commodity Systems,” *IEEE Micro*, Jan.-Feb. 2004.
<http://www.lanl.gov/radiant/pubs/10GigE/ieeemicro-2004.pdf>
6. Intel, “10-Gigabit Ethernet Expands Network Bandwidth and Shrinks Latency,” *White Paper*.
http://www.intel.com/network/connectivity/resources/doc_library/white_papers/310526.pdf
7. Feng et al., “Performance Characterization of a 10-Gigabit Ethernet TOE,” *13th IEEE Int’l Symp. on High-Performance Interconnects (Hot Interconnects)*, August 2005.
http://www.lanl.gov/radiant/pubs/10GigE/hoti2005_10gige_toe_camera_ready.pdf
8. Chelsio Communications, “Time for TOE: The Benefits of 10Gbps TCP Offload,” *White Paper*.
http://www.chelsio.com/solutions/pdf/Chelsio_TOE_Value_Prop.pdf

Network Protocols for High-Speed Interconnects

1. von Eicken et al., “U-Net: A User-Level Network Interface for Parallel and Distributed Computing,” *15th ACM Symposium of Operating Systems Principles*, Dec. 1995.
<http://www.cs.cornell.edu/tve/u-net/papers/sosp.pdf>
2. Shivam et al., “EMP: Zero-Copy OS-Bypass NIC-Driven Gigabit Ethernet Message Passing,” *ACM/IEEE SC2001*, Nov. 2001.
<http://www.cs.duke.edu/~shivam/emp.pdf>
3. Bhoedjang et al., “User-Level Network Interface Protocols,” *IEEE Computer*, Nov. 1998.
<http://ieeexplore.ieee.org/iel4/2/15756/00730737.pdf?arnumber=730737>
4. Myricom Corporation, “Myrinet Express (MX): A High-Performance Portable Implementation of the MPI Message-Passing Interface Standard,” *White Paper*.
<http://www.myri.com/scs/MX/doc/mx.pdf>
5. Jin et al., “Performance Evaluation of RDMA over IP: A Case Study with Ammasso Gigabit Ethernet NIC,” *IEEE Workshop on High-Performance Interconnects for Distributed Computing*, Jul. 2004.
http://www-unix.mcs.anl.gov/~balaji/pubs/2005/hpidc/hpidc05.ammasso_eval.pdf
6. Balaji et al., “Analyzing the Impact of Supporting Out-of-Order Communication on In-Order Performance with iWARP,” *ACM/IEEE SC’07*, Nov. 2007.
<http://www-unix.mcs.anl.gov/~balaji/pubs/2007/sc/sc07.iwarp.pdf>

Parallel Programming Models and Environments

1. Gropp et al., "A High-Performance, Portable Implementation of the MPI Message-Passing Interface Standard," *IEEE Journal of Parallel Computing*, 1996. To be made available.
 - MPI Reference: <http://www-unix.mcs.anl.gov/mpi/> and <http://www-unix.mcs.anl.gov/mpi/mpl-standard/mpl-report-1.0.ps>
2. Geist et al., "MPI-2: Extending the Message-Passing Interface," *EuroPar*, February 1996. <http://citeseer.ist.psu.edu/69768.html>
3. Dagum et al., "OpenMP: An Industry-Standard API for Shared-Memory Programming". *IEEE Computational Science & Engineering*, Jan.-Mar. 1998. <http://www.idi.ntnu.no/~banino/teaching/TDT24/files/c1046bw-1.pdf>
 - OpenMP Reference: <http://www.openmp.org>
4. Carlson et al., "Introduction to UPC and Language Introduction," *Technical Report*, 1999. <http://upc.lbl.gov/publications/upctr.pdf>
 - UPC Reference: http://upc.lbl.gov/docs/user/upc_spec_1.2.pdf
5. SourceForge, "Architectural Specification – Official Sockets Framework and Sockets Direct Protocol," *White Paper*. http://infiniband.sourceforge.net/archive/LinuxSAS_SDP.pdf http://infiniband.sourceforge.net/archive/OSF_SDP_HLD.pdf
6. Protic et al., "Distributed Shared Memory: Concepts and Systems," *IEEE Parallel and Distributed Technology*, 1996. To be made available.
7. Amza et al., "TreadMarks: Shared Memory Computing on Networks of Workstations," *IEEE Computer*, 1996. <http://www.cs.rice.edu/~willy/papers/computer96.ps.gz> (<http://citeseer.ist.psu.edu/cache/papers/cs/2972/http:zSzzSzwww.cs.rice.edu:zSzzSzSystemsSzpapersSzpapersSzsovervie w94.pdf/amza96treadmarks.pdf>)
8. Nieplocha et al., "Global Arrays: A Portable 'Shared Memory' Programming Model for Distributed Memory Computers," *ACM/iEEE SC94*, Nov. 1994. <http://www.emsl.pnl.gov/docs/global/papers/super94.pdf>

Wide-Area Networks

Packet-Switched Networking

1. Semke et al., "Automatic TCP Buffer Tuning," *ACM/SIGCOMM*, Oct. 1998. http://www.psc.edu/networking/ftp/papers/autotune_sigcomm98.ps
2. Weigle et al., "A Comparison of TCP Automatic Tuning Techniques for Distributed Computing," *10th IEEE Int'l Symp. on High-Performance Distributed Computing*, July 2002. <http://www.lanl.gov/radiant/pubs/hptcp/hpdc02-drs.pdf>
3. Hellal et al., "Analysis of TCP Vegas and TCP Reno," *Telecommunication Systems*, 2000. <http://citeseer.ist.psu.edu/303040.html>
4. Mo et al., "Analysis and Comparison on TCP Reno and Vegas," *IEEE INFOCOM*, Mar. 1999. <http://fleece.ucsd.edu/~tjavid/ECE257A/Papers2/VegasReno.pdf>
5. Kurata et al., "Fairness Comparisons Between TCP Reno and TCP Vegas ...," *INET 2000*, July 2000. <http://www.anarg.jp/~murata/papers/k-kurata00inet-ComparisonsRenoVegas.pdf>
6. Weigle et al., "A Case for TCP Vegas in High-Performance Computation Grids," *9th IEEE Symposium on High-Performance Distributed Computing*, Aug. 2001. <http://www.lanl.gov/radiant/pubs/hptcp/hpdc01-renovegas.pdf>
7. Hengartner et al., "TCP Vegas Revisited," *IEEE INFOCOM*, Mar. 2000. <http://netlab.caltech.edu/FAST/references/infocom00.pdf>
8. Feng et al., "Enabling Compatibility Between TCP Reno and TCP Vegas," *IEEE Symposium on Applications and the Internet*, Jan. 2003. <http://www.lanl.gov/radiant/pubs/hptcp/compatibility.ps>

9. Choe et al., "Stabilized Vegas," *IEEE INFOCOM*, Mar. 2003.
http://www.ieee-infocom.org/2003/papers/56_01.PDF
10. Mo et al., "Fair End-to-End Window Based Congestion Control," *IEEE Transactions on Networking*, Oct. 2000.
11. Jin et al., "FAST TCP: From Theory to Experiments," *IEEE Network*, Jan.-Feb. 2005.
<http://www.lanl.gov/radiant/pubs/hptcp/IEEE-Network-FAST-2005-Camera.pdf>

Circuit-Switched & Hybrid Networking

1. Veeraraghavan et al., "CHEETAH: Circuit-Switched High-Speed End-to-End Transport Architecture," *SPIE/IEEE Optical Networking and Computer Communications Conference (OptiComm)*, Oct. 2003.
<http://www.lanl.gov/radiant/pubs/cheetah/opticom2003.pdf>
2. Veeraraghavan et al., "Scheduling and Transport for File Transfers on High-Speed Optical Circuits," *Journal of Grid Computing*, June 2004.
<http://www.lanl.gov/radiant/pubs/drs/jgc2004.pdf>
3. He et al., "Reliable Blast UDP: Predictable High-Performance Bulk-Data Transfer," *IEEE International Conference on Cluster Computing*, Sept. 2002.
<http://www.evl.uic.edu/cavern/papers/cluster2002.pdf>
4. Zheng et al., "FRTP: Fixed-Rate Transport Protocol ...," *IEEE ACM International Conference on Broadband Communications, Networks, and Systems*, Oct. 2004.
http://www.broadnets.org/2004/workshop-papers/Pathnets/01_FRTP-XuanZheng.pdf
5. Sivakumar et al., "Simple Available Bandwidth Utilization Library for High-Speed Wide-Area Networks," *Journal of Supercomputing*, 2003.
<http://www.dataspaceweb.net/papers/sabul-jsc-03.pdf>
6. Xiong et al., "LambdaStream – A Data Transport Protocol for Streaming Network-Interactive Applications over Photonic Networks," *3rd International Workshop on Protocols for Fast Long-Distance Networks*, Feb. 2005.
<http://snorky.evl.uic.edu/files/pdf/lambdaStream.pdf>
7. Banerjee et al., "RAPID:: An End-System Aware Protocol for Intelligent Data-Transfer over LambdaGrids," *IEEE International Parallel & Distributed Processing Symposium*, Apr. 2006.
<http://www.cs.ucdavis.edu/~ghosal/Research/publications/ipdps-2006.pdf>
8. Datta et al., "End-System Aware, Rate-Adaptive Protocol for Network Transport in LambdaGrid Environments," SC'06, 2006.
<http://sc06.supercomputing.org/schedule/pdf/pap229.pdf>

Networking in Datacenters (SDP, iWARP, etc.)

1. Balaji et al., "Sockets Direct Protocol over InfiniBand in Clusters: Is it Beneficial?," *IEEE International Symposium on Performance Analysis of Systems and Software*, Mar. 2004.
2. Balaji et al., "Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol (SDP) over InfiniBand," *Communication Architecture for Clusters (CAC)*, Apr. 2006.
3. Jin et al., "Performance Evaluation of RDMA over IP: A Case Study with Ammasso Gigabit Ethernet NIC," *IEEE Workshop on High-Performance Interconnects for Distributed Computing*, Jul. 2004. (Same paper as above.)
4. Balaji et al., "Supporting iWARP Compatibility and Features for Regular Network Adapters," *Workshop on Remote Direct Memory Access (RDMA): Applications, Implementations, and Technologies (RAIT 2005)*, Sept. 2005.
5. Narravula et al., "Architecture for Caching Responses with Multiple Dynamic Dependencies in Multi-Tier Data-Centers over InfiniBand," *IEEE/ACM International Symposium on Cluster Computing and the Grid*, May 2005.

6. Balaji et al., "On the Provision of Prioritization and Soft QoS in Dynamically Reconfigurable Shared Datacenters over InfiniBand," *IEEE Int'l Symp. on Performance Analysis of Systems and Software*, Mar. 2005.

I/O, File Systems, and Storage Systems

1. Wu et al., "PVFS over InfiniBand: Design and Performance Evaluation," *32nd International Conference on Parallel Processing*, Oct. 2003.
2. Yu et al., "High Performance Support of Parallel Virtual File System (PVFS2) over Quadrics," *19th ACM International Conference on Supercomputing (ICS '05)*, June 2005.
3. Liu et al., "Evaluating the Impact of RDMA on Storage I/O over InfiniBand . SAN-03 Workshop (in conjunction with HPCA), Feb. 2004.
4. Callaghan et al., "NFS over RDMA," *ACM SIGCOMM Workshops*, 2003.
5. DeBergalis et al., "The Direct Access File System," *FAST*, 2003.
6. Eisler, "Data ONTAP GX: A Scalable Storage Cluster," *FAST*, 2007.

Potpourri: Performance Modeling

1. Culler et al., "LogP: Towards a Realistic Model of Parallel Computation," <http://www.cs.berkeley.edu/~culler/papers/logp.ps>
2. Shivam et al., "On the Elusive Benefits of Protocol Offload," *ACM SIGCOMM Workshop on Network-I/O Convergence: Experiences, Lessons and Implications (NICELI)*, Aug. 2003. <http://issg.cs.duke.edu/publications/niceli03.pdf>
3. Hochstein et al., "Parallel Programmer Productivity: A Case Study of Novice Parallel Programmers," *ACM/IEEE SC2005*, Nov. 2005. <http://www.cs.umd.edu/~lorin/pubs/sc05.pdf>

Examples of Project Ideas

1. Process-to-Core Mapping in Multicore Environments
2. Energy-Aware Networking with 10G Switches
3. TCP/IP Backend to Conceptual (<http://www.ccs3.lanl.gov/~pakin/software/conceptual/>)
4. RAPID++: Rate-Based Protocol for LambdaGrids
5. UOE Offload Engine: Chelsio T110 or Myri-10G
6. NetEffect 10G Adapter / OpenFabrics / OFED
7. iWARP over Reliable UDP Protocol
8. Eventless Monitoring and Self-Tuning of Network Protocols in the Presence of Offload Engines
9. Detailed and Updated Examination of TCP Autotuning Techniques
10. Networking in Linux vs. Plan 9
11. Dynamic Flow-Control Adaptation: Linux, Plan 9, others?
12. WAN compatibility for Myri-10G adapters → partial offload of TCP/IP
13. A Virtual Computing Laboratory Running Atop a LAMPS Environment
14. End-to-End TCP via Composition

And on the "not really networking" front ...

15. Transform mpiBLAST into SocketsBLAST or GridBLAST/GridRuby
16. Port mpiBLAST to the MPI-2 interface