Interconnection Networks for High-Performance Systems

Prerequisite(s): ECE 6100 / CS6290, or equivalent
Instructor: Tushar Krishna

Course Objectives

Interconnection Networks refer to the communication fabric *within* a computer system. They occur at various scales – from on-chip networks (OCN)/Networks-on-Chip (NoCs) in billion-transistor many-core chips, to custom high-speed wired networks in supercomputers, to optical fiber networks within datacenters**.** The growing emphasis on parallelism, distributed computing, and energy-efficiency across all these systems makes the design of the communication fabric critical to both high-performance and low power consumption.

This course examines the architecture, design methodology, and trade-offs of interconnection networks. In the first half of the course, lectures will cover the fundamentals of interconnection networks – topology, routing, flow-control, microarchitecture, network interfaces, and system interactions - getting to the research frontier at each level. The second half of the course will focus on state-of-the-art research and case studies, using a mix of lectures, student presentations, paper readings, discussions, and debates on contrasting approaches. Towards the end, the important role of interconnection networks in emerging domains such as Deep Learning, Edge Computing and the Internet-of-Things will be explored.

A semester long programming-heavy project will focus on solving a research problem in the domain of interconnection networks, and can lead to publications in reputed conferences/journals. Projects aligned with the students' own MS/PhD research will also be encouraged. Projects from past iterations of the course have led to student publications in top-tier conferences and journals.

The material covered in this course bridges the gap between disciplines/courses such as VLSI interconnects, digital communication, computer architecture, distributed systems, and computer networks. At the end of this course, students will be able to appreciate both the architectural and the physical design nuances/trade-offs involved in interconnection network design. They will acquire the skill sets to design state-of-the-art on-chip networks for multicore processors from industry leaders like Intel, AMD, IBM, Qualcomm, and so on. They will also be able to design and evaluate large-scale networks inside datacenters maintained by Google, Facebook, Amazon, and Microsoft. Students will also be trained to review and critique state-of-the-research papers on Interconnections Networks / Networks-on-Chip from leading conferences and journals.

Course Text

The material for this course will be derived from the following texts:

1. N. E. Jerger, T. Krishna, and L.-S Peh, "On-Chip Networks, Second Edition" Morgan Claypool Publishers, 2017. *[required]*
2. W. Dally and B. Towles, "Principles and Practices of Interconnection Networks," Morgan Kauffman Publishers, 2004. *[optional]*
3. J. Duato, S. Yalamanchili, L.Ni, "Interconnection Networks: An Engineering Approach," Morgan Kauffman Publishers, 2002. *[optional]*
4. Papers from recent conferences: **ISCA, MICRO, HPCA, ASPLOS, NOCS, DATE, DAC, ICCAD, ICCD, ISSCC**

Syllabus and Outline

1. **Introduction to Interconnection Networks**
   - Introduction
   - Types of Networks
   - Evaluation Metrics

2. **Topology**
   - Metrics for comparing topologies
   - Direct Topologies
   - Indirect Topologies
   - Hierarchical Topologies

3. **Routing**
   - Deterministic Routing
   - Oblivious Routing
   - Adaptive Routing

4. **Flow-Control**
   - Message-based Flow Control
   - Packet-based Flow Control
   - Flit-based Flow Control
   - Virtual Channels

5. **Deadlocks**
   - Channel Dependency Graph
   - Turn Model
   - Up*/Down* Routing
   - Escape Virtual Channels
   - Deadlock Recovery

6. **Microarchitecture**
   - Router Organization
   - Pipeline
   - Optimizations
   - Buffer Management
   - Crossbar Design
   - Allocators and Arbiters

7. **System Interface**
   - Shared Memory Multiprocessors
     - Cache Coherence
     - Deadlocks
   - Message Passing

8. **Implementation: RTL and Circuits**
   - Wire Delay
   - Router Pipelines
   - Power Consumption
   - Area Overheads

9. Advanced Topics

- Physical and Virtual Express Topologies
- Single-cycle Multi-hop Networks
- Multicast Communication
- Silicon Photonics
- Reliability and Faults
- GPU Networks
- FPGA Networks

10. System-level Networks

- Supercomputer Networks
- Datacenter Networks

11. Case Studies with Real Chips

- Supercomputers
  - D E Shaw Research Anton 2
  - IBM BlueGene Q
- Multicore
  - Intel SCC
  - ST Spidergon
  - Tilera TILE64
  - UT Austin TRIPS
  - University of Michigan Swizzle Switch
- Accelerators
  - IBM TrueNorth
  - Google TPU

12. Emerging Trends

- Heterogeneous Systems
- Spatial Accelerators (Deep Learning, Graph Processing)
- Internet-of-Things and Edge-Computing

Course Grading

Lab Assignments	30%
Midterm	15%
Paper Critiques	10%
Presentation on Paper/Case Study	10%
Project	35% (Report 25%, Presentation 10%)

The early part of the course will cover fundamentals of interconnection networks (topology, routing, flow-control, and microarchitecture). The midterm will test knowledge of this theory. The remainder of the course will follow a seminar format involving paper readings, discussions and presentations. The lab assignments will introduce the students to the Garnet2.0 Network-on-Chip simulator (distributed within the gem5 (www.gem5.org) open source full-system multi-core simulator written in C++). The research project will cover problems in modern network design. Students will study relevant papers, propose a solution, implement the solution (via simulation) document the project (short paper) and present the paper in a conference format (20 minutes).

Absence and Re-Examination Policy

In case students miss a deadline or an exam, the course will abide by the institute policy on student absences (http://www.catalog.gatech.edu/rules/4/)

Learning Accommodations

If needed, the course will make classroom accommodations for students with disabilities. These accommodations should be arranged in advance and in accordance with the office of Disability Services (http://www.adapts.gatech.edu)

Academic Integrity

Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. For information on Georgia Tech's Academic Honor Code, please visit
http://www.catalog.gatech.edu/policies/honor-code/ or http://www.catalog.gatech.edu/rules/18/

Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.