Interconnection Networks for High-Performance Systems
ECE 8823 A / CS 8803 – ICN (Spring 2018)

Instructor: Tushar Krishna (tushar@ece.gatech.edu)
When: MW 3:00 – 4:15 pm
Where: CoC 102
Web: http://tusharkrishna.ece.gatech.edu/teaching/icn_s18/

Course Objectives

Interconnection Networks form the backbone of all computer systems today. They occur at various scales across all high-performance systems - systolic-arrays within Google's Deep Learning **TPU**, high-bandwidth crossbars inside modern **GPU**s, soft transport macros on **FPGAs**, mesh networks-on-chip (**NoC**) in many-core processors, **interposer** fabrics on package, **QPI** in multi-socket servers, **Infiniband** in supercomputers/clusters, and Fat-Tree **datacenter** networks in the cloud. The growing emphasis on parallelism, scalability, 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 will examine the similarities, differences, and trade-offs in the architecture and implementation of interconnection fabrics across all these systems. Given the breadth of topic areas (computer architecture, VLSI interconnects, computer networks, and distributed systems), students will get a glimpse into designing systems and optimizing for data movement at various scales – from on-chip to cloud-scale.

This year, there will be a particular focus on two emerging domains where interconnection networks are playing an increasing role: ***Deep Learning Accelerators*** and Fog-Platforms for running ***Internet-of-Things (IoT)*** applications.

Course Structure: This is an advanced graduate course, structured around a mix of lectures on the fundamentals of interconnection networks (topology, routing, flow-control, microarchitecture, network and system interfaces), student presentations, and paper critiques. A series of programming-heavy labs will bring everyone up to speed with an interconnection networks simulator Garnet2.0, that is distributed within the gem5 (www.gem5.org) open source full-system multi-core simulator.

A half-semester long research project will focus on solving open-ended research problems on interconnection networks across any domain of computing. Projects aligned with students' own graduate research (MS/PhD) will be encouraged if they have an exciting networks component. In the past two iterations of this course, projects directly from the course have led to publications in top conferences such as HPCA, ASPLOS, ICCAD, ISPASS and NOCS.

Minor in ECE for CS Students: To get credits for a minor in ECE, register for the ECE section.

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, 2nd Edition" Morgan Claypool Publishers, 2017.
2. W. Dally and B. Towles, "Principles and Practices of Interconnection Networks," Morgan Kauffman Publishers, 2004.
3. Papers from recent conferences: **ISCA, MICRO, HPCA, ASPLOS, SIGCOMM, NSDI, NOCS, DATE, DAC, 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. Emerging Technologies
- Silicon Photonics
- Reliability and Faults

10. Interconnection Networks across High-Performance Systems

Many-core Processors
GPUs
FPGA
Deep Learning Accelerators
On-Package Interconnects
2.5D/3D Systems
Supercomputers
Datacenters
IoT Systems

Course Grading

Lab Assignments	25%
Paper Critiques	10%
Paper Presentation	10%
Peer Reviews	5%
Midterm	10%
Project	40% (Report 25%, Presentation 15%)

Course Policies

If you have a documented emergency or a university mandated reason because of which you have to miss an exam, get in touch with the instructor before (preferable) or latest by the day of the exam.

Learning Accommodations

If needed, we 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)