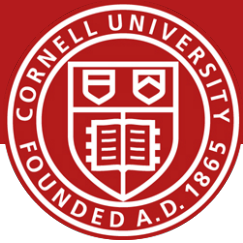


Data Center Networks and Basic Switching Technologies

Hakim Weatherspoon

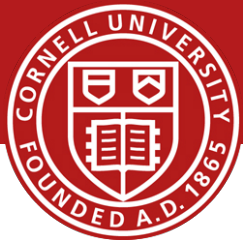
Assistant Professor, Dept of Computer Science

CS 5413: High Performance Systems and Networking

September 15, 2014

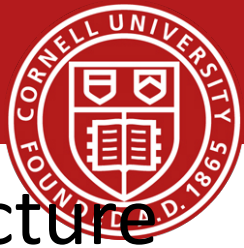
Slides used and adapted judiciously from Computer Networking, A Top-Down Approach

Where are we in the semester?



- Overview and Basics
- Data Center Networks
  - Basic switching technologies (*today*)
  - Data Center Network Topologies
  - Software Routers (eg. Click, Routebricks, NetMap, Netslice)
  - Alternative Switching Technologies
  - Data Center Transport
- Data Center Software Networking
  - Software Defined networking (overview, control plane, data plane, NetFGPA)
  - Data Center Traffic and Measurements
  - Virtualizing Networks
  - Middleboxes
- Advanced Topics

Goals for Today

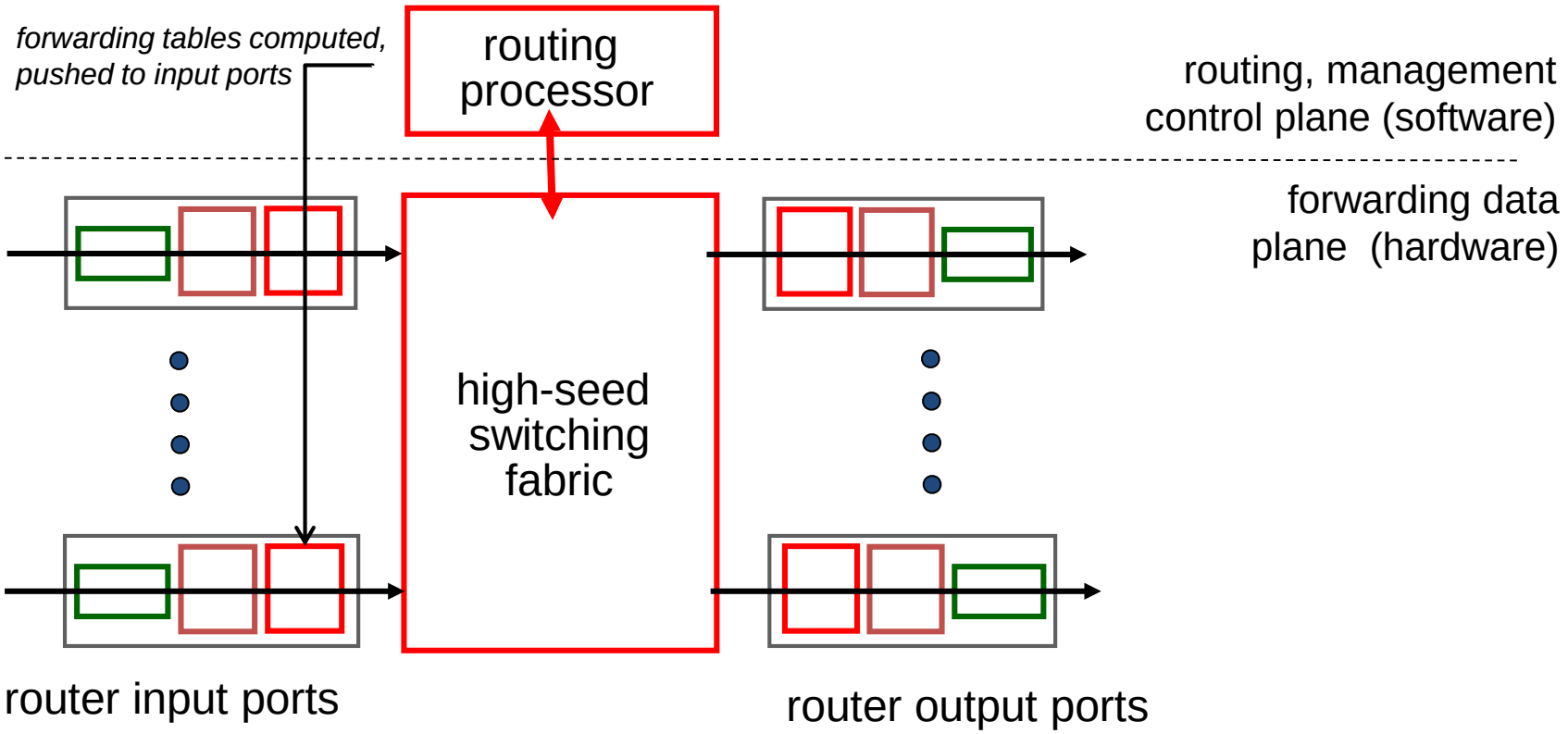


- Basic Switching Technologies/Router Architecture Overview
  - See Section 4.3 in book
- A 50-Gb/s IP Router
  - Craig Partridge , Senior Member , Philip P. Carvey , Isidro Castineyra , Tom Clarke , John Rokosz , Joshua Seeger , Michael Sollins , Steve Starch , Benjamin Tober , Gregory D. Troxel , David Waitzman , Scott Winterble. *IEEE/ACM Transactions on Networking (ToN)*, Volume 6, Issue 3 (June 1998), pages 237-248.

Router Architecture Overview

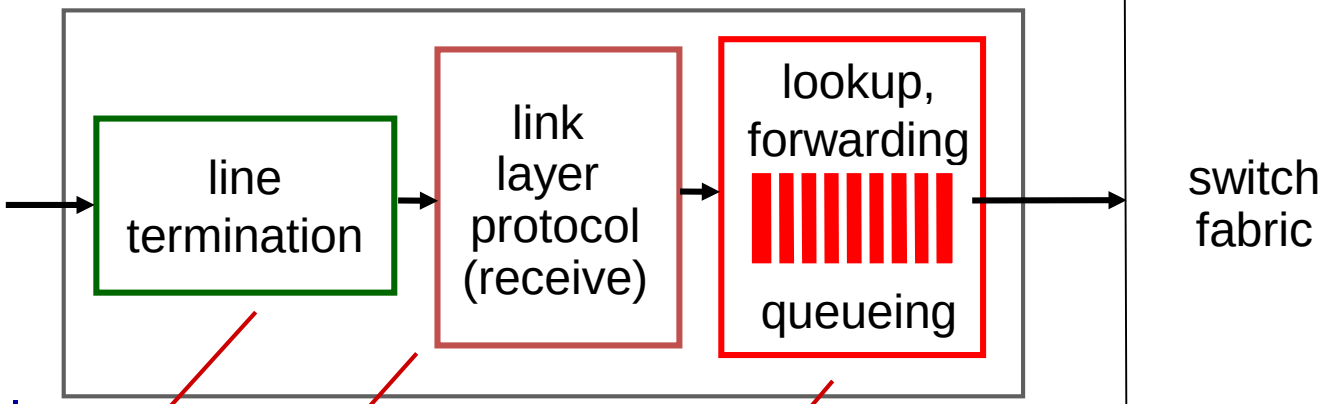
two key router functions:

- run routing algorithms/protocol (e.g. RIP, OSPF, BGP)
- *forwarding* datagrams from incoming to outgoing link



Router Architecture Overview

Input Port Functions



physical layer:
bit-level reception

data link layer:
e.g., Ethernet
see chapter 5

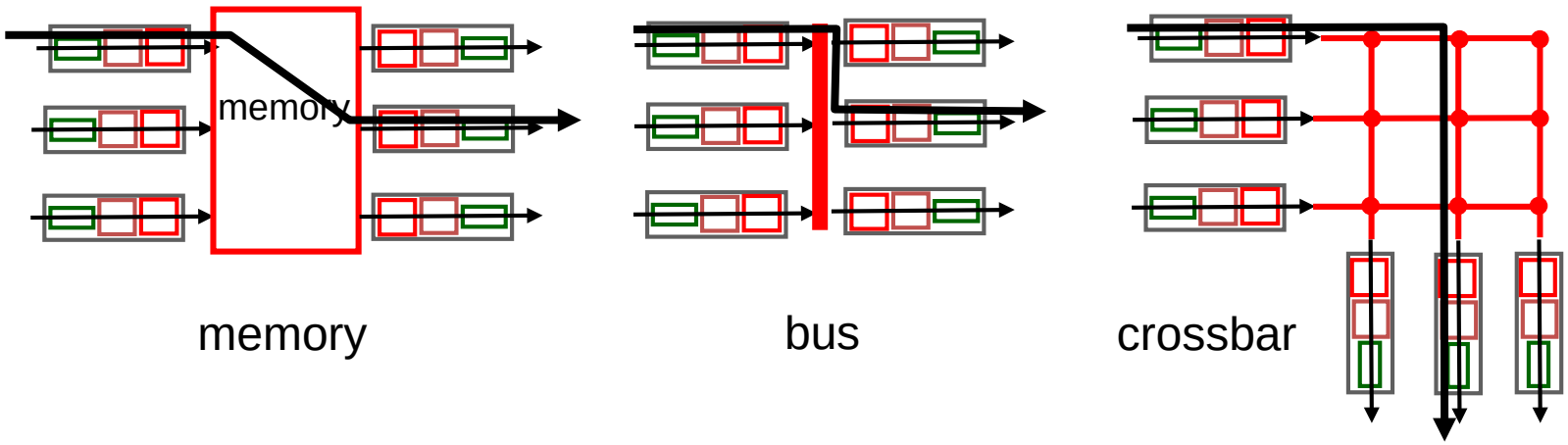
decentralized switching:

- given datagram dest., lookup output port using forwarding table in input port memory *("match plus action")*
- ***goal: complete input port processing at 'line speed'***
- ***queuing***: if datagrams arrive faster than forwarding rate into switch fabric

Router Architecture Overview

Switching Fabrics

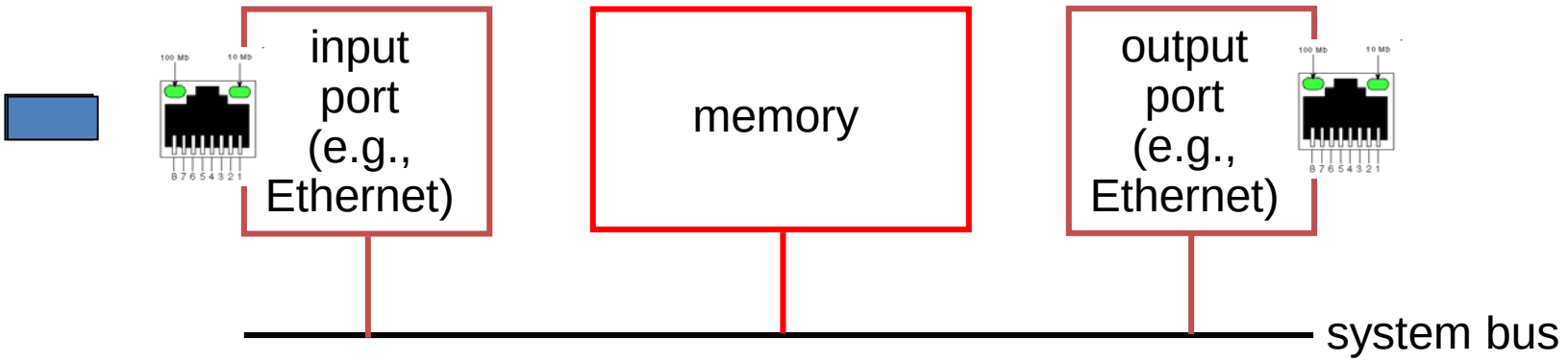
- transfer packet from input buffer to appropriate output buffer
- switching rate: rate at which packets can be transfered from inputs to outputs
  - often measured as multiple of input/output line rate
  - N inputs: switching rate N times line rate desirable
- three types of switching fabrics



Router Architecture Overview

Switching via Memory: First Generation Routers

- traditional computers with switching under direct control of CPU
- packet copied to system's memory
- speed limited by memory bandwidth (2 bus crossings per datagram)

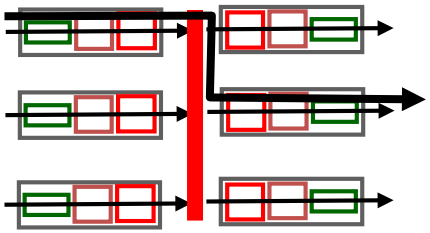


Router Architecture Overview

Switching via a bus

- datagram from input port memory

  to output port memory via a shared bus
- *bus contention:* switching speed limited by bus bandwidth
- 32 Gbps bus, Cisco 5600: sufficient speed for access and enterprise routers

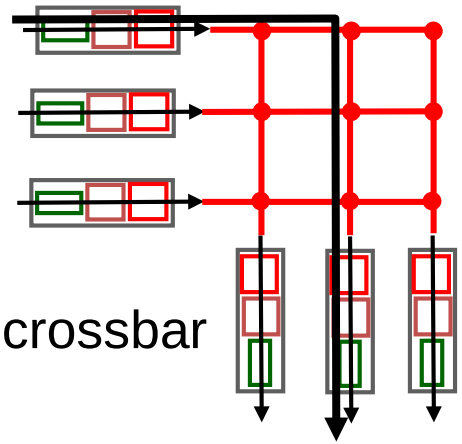


bus

Router Architecture Overview

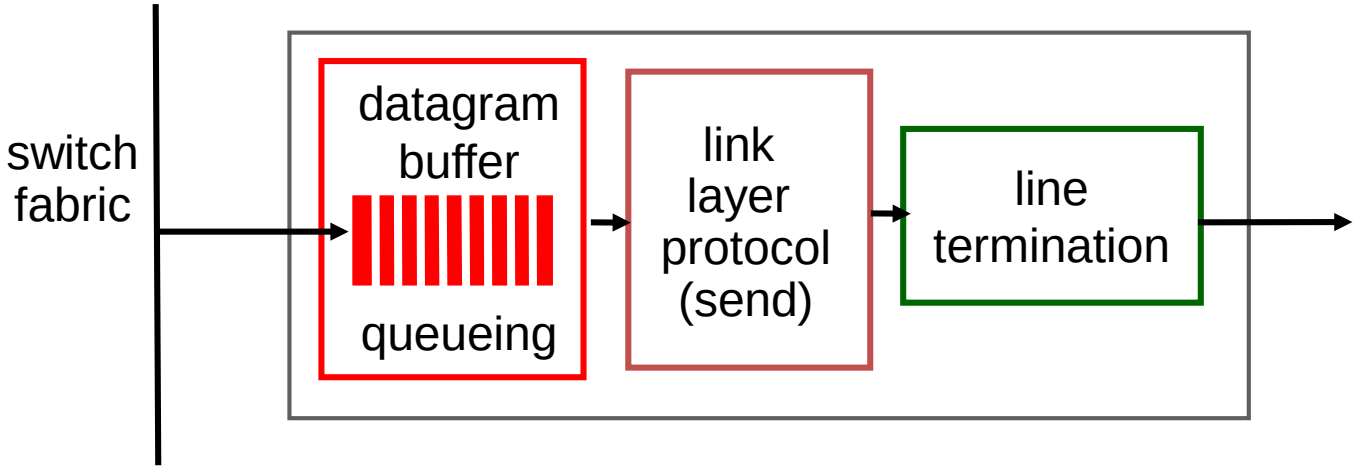
Switching via interconnection network

- overcome bus bandwidth limitations
- banyan networks, crossbar, other interconnection nets initially developed to connect processors in multiprocessor
- advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- Cisco 12000: switches 60 Gbps through the interconnection network



Router Architecture Overview

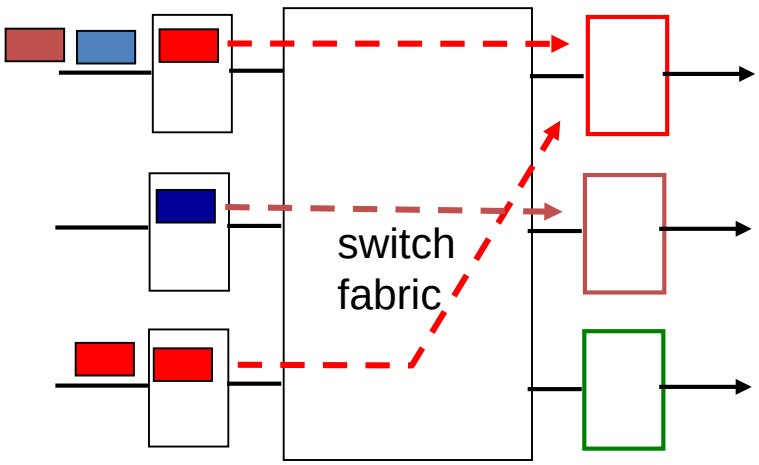
Output Ports



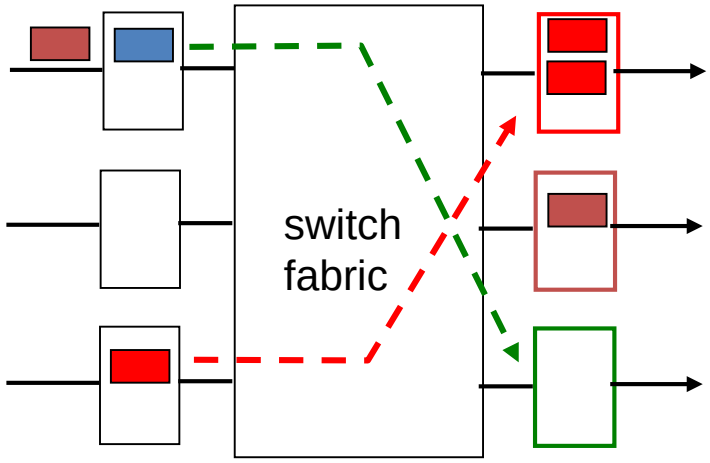
- *buffering* required when datagrams arrive from fabric faster than the transmission rate
- *scheduling discipline* chooses among queued datagrams for transmission

Router Architecture Overview

Output Port Queuing



at *t*, packets more from input to output



one packet time later

- buffering when arrival rate via switch exceeds output line speed
- *queueing (delay) and loss due to output port buffer overflow!*

Router Architecture Overview

How much buffering?

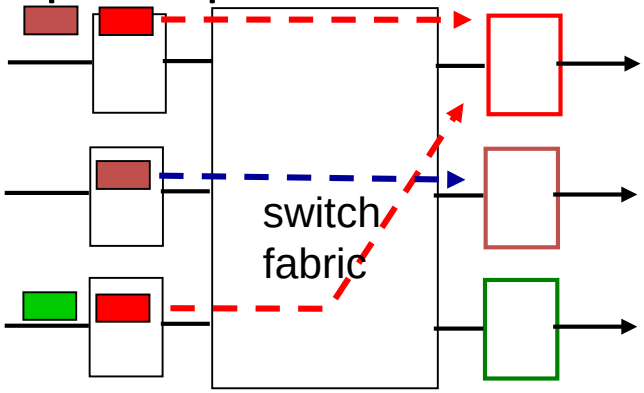
- RFC 3439 rule of thumb: average buffering equal to "typical" RTT (say 250 msec) times link capacity C
  - e.g., C = 10 Gpbs link: 2.5 Gbit buffer
- recent recommendation: with *N* flows, buffering equal to

$$\frac{RTT \cdot C}{\sqrt{N}}$$

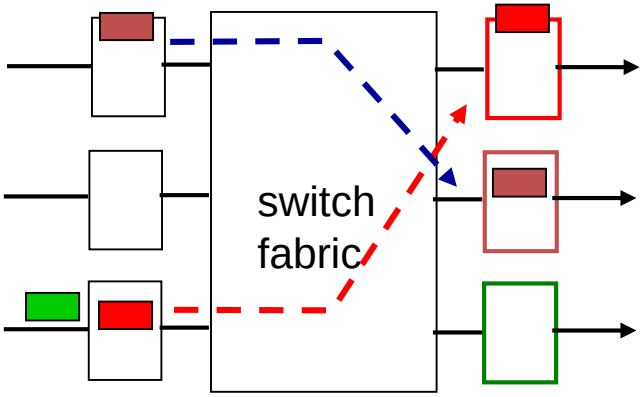
Router Architecture Overview

Input Port Queuing

- fabric slower than input ports combined -> queueing may occur at input queues
  - *queueing delay and loss due to input buffer overflow!*
- Head-of-the-Line (HOL) blocking: queued datagram at front of queue prevents others in queue from moving forward

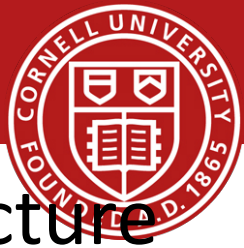


output port contention:
only one red datagram can be transferred.
lower red packet is blocked



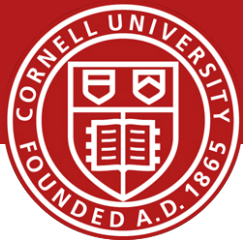
one packet time later:
green packet
experiences HOL
blocking

Goals for Today



- Basic Switching Technologies/Router Architecture Overview
  - See Section 4.3 in book
- A 50-Gb/s IP Router
  - Craig Partridge , Senior Member , Philip P. Carvey , Isidro Castineyra , Tom Clarke , John Rokosz , Joshua Seeger , Michael Sollins , Steve Starch , Benjamin Tober , Gregory D. Troxel , David Waitzman , Scott Winterble. *IEEE/ACM Transactions on Networking (ToN)*, Volume 6, Issue 3 (June 1998), pages 237-248.

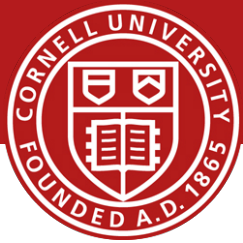
Multigigabit Router (MGR)



Architecture

- Network interfaces (Line cards)
- Forwarding Engine
- Network Processor
- Switching Fabric
  - .

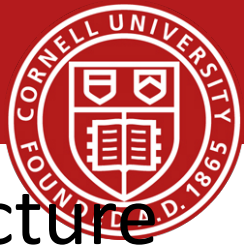
Multigigabit Router (MGR)



Contributions

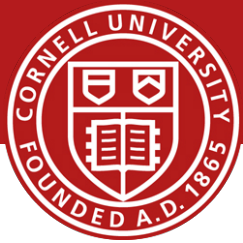
- Network interfaces (Line cards)
  - Forwarding Engine distinct from line cards
- Forwarding Engine
  - Complete set of forwarding tables, fast path
  - QoS
- Network Processor
  - Updates Routing Table
  - Separates and handles slow path
- Switching Fabric
  - Switched backplane

Goals for Today



- Basic Switching Technologies/Router Architecture Overview
  - See Section 4.3 in book
- A 50-Gb/s IP Router
  - Craig Partridge , Senior Member , Philip P. Carvey , Isidro Castineyra , Tom Clarke , John Rokosz , Joshua Seeger , Michael Sollins , Steve Starch , Benjamin Tober , Gregory D. Troxel , David Waitzman , Scott Winterble. *IEEE/ACM Transactions on Networking (ToN)*, Volume 6, Issue 3 (June 1998), pages 237-248.

Before Next time



- Project Proposal
  - **due this Friday, Sept 19**
  - Meet with groups, TA, and professor
- Lab2
  - Multi threaded TCP proxy
  - **Due this Friday, Sept 19**

- ***Required review and reading***
  - "A Guided Tour Through Datacenter Networking," D. Abts and B. Felderman. *Communications of the ACM (CACM)*, Volume 55, Issue 6 (June 2012), pages 44-51.
  - http://dl.acm.org/citation.cfm?id=2184335
  - http://wwwnew.cs.princeton.edu/courses/archive/spring13/cos598C/google-network.pdf

- Check piazza: http://piazza.com/cornell/fall2014/cs5413
- Check website for updated schedule