

Network performance

Networks: speed at a premium

- if slow typically not used in practice
- e.g., cryptographic protocols tend to be turned off at routers due to overhead

Network design approach:

- emphasis on lightweight network core
- push heavyweight stuff toward the edge (i.e., host/server)
- called end-to-end paradigm
- historically: guided Internet design and evolution
- other approaches have been tried and failed

Performance yardsticks:

- bandwidth in bps (bits-per-second)
 - from bandwidth of physical media (Hz) to bps
 - nominal bandwidth: ignore slow-down due to software and control overhead
- throughput (bps): includes protocol overhead
 - firmware in NIC and device driver in OS
 - in addition: app/user space overhead (not necessarily bad)
- latency and delay: unit msec (millisecond)
 - latency: signal propagation speed (bounded by SOL)
 - processing and buffering delay (queueing)
- jitter: delay variation
 - average delay small but max delay large
 - bad for multimedia

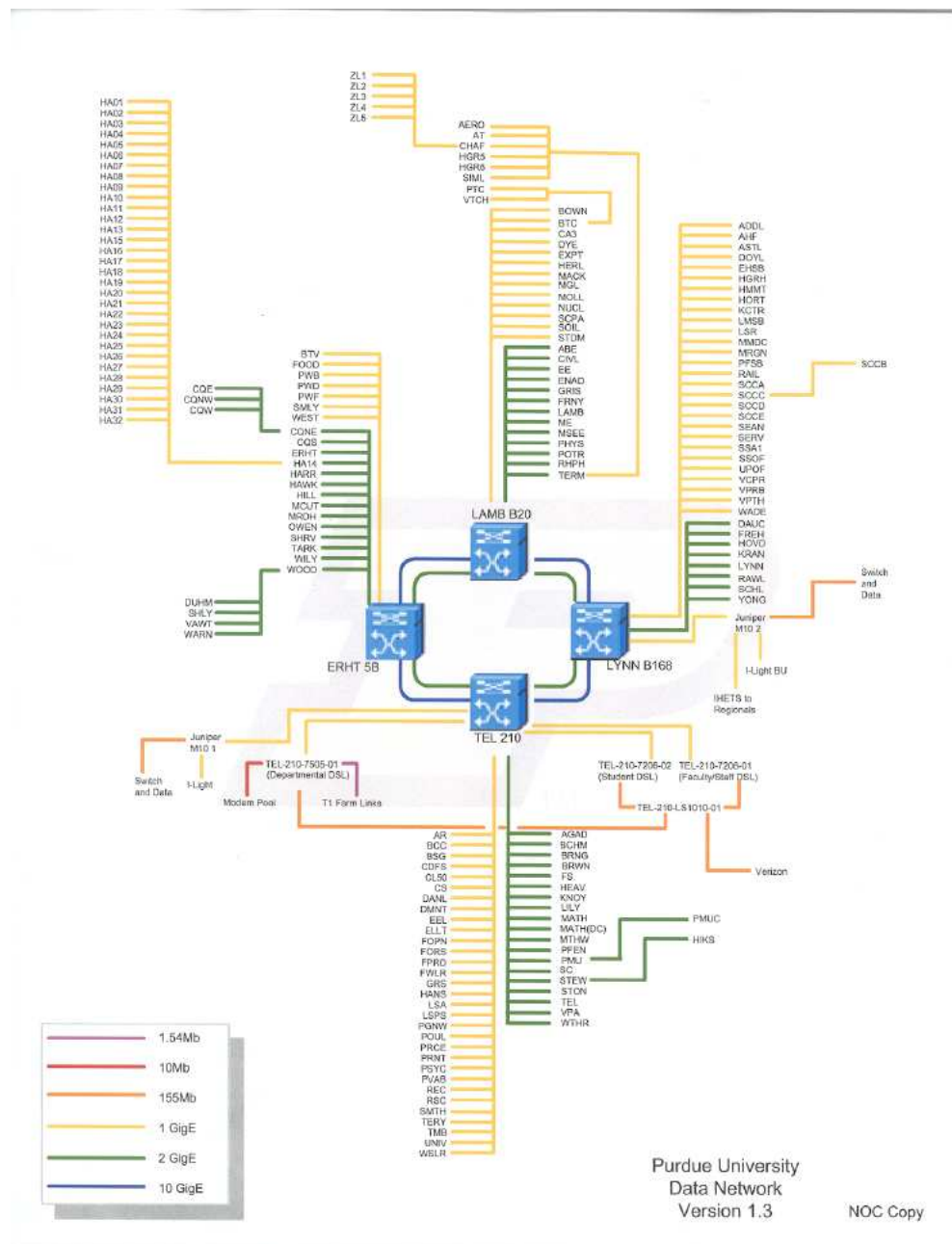
Meaning of “high-speed” networks:

- signal propagation speed is bounded by SOL (speed-of-light)
 - $\sim 186\text{K miles/s}$ ($\sim 300\text{K km/s}$)
 - optical fiber, copper: slower than SOL
- Ex.: latency from Purdue to West Coast
 - for 2000 miles: $\sim 10\text{ msec}$ ($= 2000/186000$)
 - lower bound
- Ex.: geostationary satellites at $\sim 22.2\text{K miles}$
 - latency: $\sim 120\text{ msec}$
 - end-to-end (one-way): $\sim 240\text{ msec}$
 - round-trip (two-way): $\sim 480\text{ msec}$
 - roughly: half a second
 - fundamental limit faced by apps

Meaning of high-speed:

- a single bit cannot go faster
 - can only increase bandwidth (bps): bits packed into 1 second
 - cramming bits sequentially in time
 - sending bits in parallel across multiple channels
 - analogous to widening highway, i.e., more lanes
 - also called broadband
- focus: high-speed \Leftrightarrow many lanes
 - effect: completion time of large files shorter
 - in this sense, “higher” speed
 - for small files: marginal benefit
 - Internet workload: most files are small, minority is very large
 - but minority consumes bulk of network bandwidth

Example network pics: Purdue's backbone network



What is traveling on the wires?

Mixture of:

Bulk data (data, image, video, audio files), voice, streaming video/audio, real-time interactive data (e.g., games, social media, etc.), AI related traffic.

→ bulk of Internet traffic: mediated by TCP protocol (which runs on top of IP)

→ primarily a giant client/server system

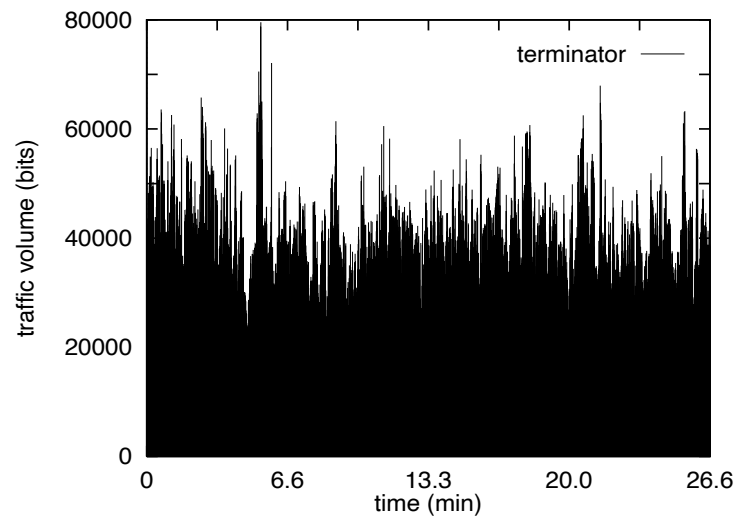
Multimedia (video/audio) streaming: rapid rise

→ streaming video: e.g., YouTube, Netflix

→ real-time: e.g., VoIP, video conferencing, games

→ may be target of traffic shaping (e.g., fine print of data plans)

Example: traffic is “bursty”: MPEG compressed real-time video



Reason:

- video compression
 - utilize inter-frame compression
- burstiness is not good for networks

Example: 90/10 (or 80/20) property of files

→ called “mice and elephants”

→ spikes caused by elephants

→ target of active traffic control (e.g., TCP)

→ most flows are mice

How to make sense of all this?

We will investigate three aspects:

- architecture
 - system design, real-world manifestation
- algorithms
 - how do the components work
- implementation
 - how are they actually implemented
 - additional complications

Key concern and common thread: performance

- networks are real-world systems
- slow means likely not used in practice
- performance concerns heavily influence architecture, algorithm, implementation