Layering

- Abstracting functionality to different parts (layers) of the network
- Each layer implements a protocol
- Each layer communicates with the layers above and below it

Protocol

- Defines a mechanism for communication within a layer
  - peer interface
- Defines a mechanism for communication with services in layers directly above this layer
  - service interface

Example: web transaction

- HTTP: browser-to-web server communication
- TCP: host-to-host communication
- IP: network-to-network communication
- Other protocols operating “under the hood”
  - more on this in next few lectures
Architecture

- Defined in terms of a protocol stack
- Two models:
  1. OSI (Open Systems Interconnect) --- 7 layers
  2. “Internet” --- 4 layers

What each layer does

- Application: Self-explanatory
- Presentation: Data formatting
- Session: Organization of multiple transport streams (e.g., audio + video + control)
- Transport: Message exchange, host-to-host connectivity (TCP, UDP)
- Network: Packet routing
- Data link: Bit framing
- Physical: Signalling (bit transmission)

OSI architecture

- Application
- Presentation
- Session
- Transport
- Network
- Data link
- Physical

End host

Network node

Internet architecture

- Layers 1-4 and 7 from OSI stack
- Network (aka link) layer combines layers 1 and 2 from OSI stack
  - network = hardware and software connections between nodes
- IP is layer 3 from OSI stack
- Application layer combines layers 5-7 from OSI stack
Each layer takes care of...

- How data is sent
- Error handling
- Encapsulating data from the layer above
- Adding appropriate headers (routing/handling information)

Note

- Layers are abstractions, not strict delineators!
- Some protocols operate “between” layers
- Can have multiple protocols operating in the same layer, on top of each other
- Examples:
  - RTP: transport layer, operates over UDP or TCP
  - MPLS: Layer 2/3, can operate over ATM, which operates over SONET

Measuring computer network performance

- Bandwidth
- Throughput
- Latency
- Packet loss
- ...

Bandwidth

- One of the most misused and misunderstood terms in networking
- Our definition: The maximum number of bits per second that a given link can transmit
  - e.g., Ethernet = 100 Mbps (100 million bits per second)
Bandwidth (cont.)

- Other definitions:
  - engineer’s: width of a frequency band (Hz)
  - application bandwidth: how many bits per second it requires from the network in order to operate at acceptable performance levels
  - overall bandwidth: how many bits per second a “logical channel” between two hosts can transmit
    - aka bottleneck bandwidth
  - available bandwidth: how many bits per second a link or channel can provide to a particular host or application
    - \( \text{BW}_{\text{available}} = \text{BW}_{\text{total}} - \text{BW}_{\text{currently in use}} \)

Throughput

- Definition: the number of bits per second actually transmitted by a link in practice
- Can refer to
  - the link overall
  - a specific application
  - a specific set of hosts
  - a logical channel

Latency

- The amount of time it takes a message (a collection of bits) to travel from one point to another, in seconds
  - link latency
  - channel latency
- Usually, we're interested in the round-trip time
  - how long it takes a message to get from one point to another and back

Calculating latency

- Latency is the sum of three components:
  - propagation delay = distance/speed
    - speed \( \sim 2 \times 10^8 \text{ m/s} \)
  - transmission delay = (message size)/bandwidth
  - queueing delay = measured or approximated
Measuring network performance

- Example: 2 messages to be sent between a client and a server
  - one message is 5 bytes
  - one message is 20 MB
- 2 possible channels on which to send the data
  - 1 Mbps channel
  - 10 Mbps channel
- Channel is between Northfield and San Diego (2400 km)
- Q: What is the latency (one way) for each channel and each message?

Measuring network performance

- Some applications are latency-bound
  - latency calculation is dominated by the propagation delay
  - typically applications that send small messages
- Some applications are bandwidth-bound
  - latency calculation is dominated by the transmission delay, which in turn is dominated by the channel bandwidth
  - typically applications that send large messages

Delay-bandwidth product

- “Volume” of a link or channel
- How many bits you can fit into a channel or link
- Also, how many bits a host must send before the first bit arrives at the receiver