Chapter 15
Network Services and Performance

Defining the Services

- So far we have focused on the problem of getting data from one machine on a LAN to another machine on the same LAN
- Although this is an important service, networks typically provide lots of additional services
- In this chapter we will investigate what types of services network users would like from the network

Communication Services

- Applications often need more services from the network than simple packet delivery
- Not all networks offer all services
- Which services do applications need?
- Which services should a network implement?

The rest of this chapter describes useful services. The next chapter describes techniques for implementing these services.

Connection-Oriented Communication

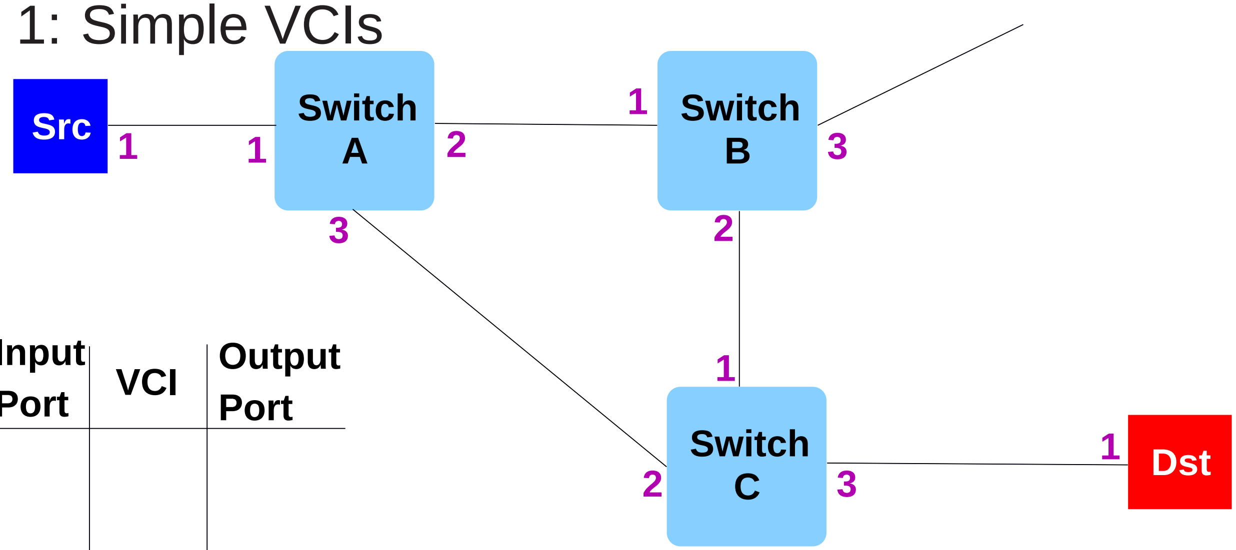
- Communication between exactly two machines
- Similar to telephone system; endpoints establish and maintain a connection as long as they have data to exchange
- One endpoint tries to start the connection
- The other endpoint agrees to the connection
- All further data exchange occurs over the pre-established connection
- Typically accessed via a stream interface
- Data transmission doesn't need to be continuous; like telephones, the connection remains in place even when no data is being transmitted
- So how do you stop a connection?

What exactly is a "Connection"?

- ATM provides connection-oriented service (we will use it as an example)
- recall that an ATM cell is ____ octets long with ____ octets of header and ____ octets of payload
- recall that an NSAP address is ____ octets long
- Question: How are ATM cells addressed? The header is not large enough to hold the destination NSAP address.
- Answer:

Virtual Circuits

- Example 1: Simple VCIs



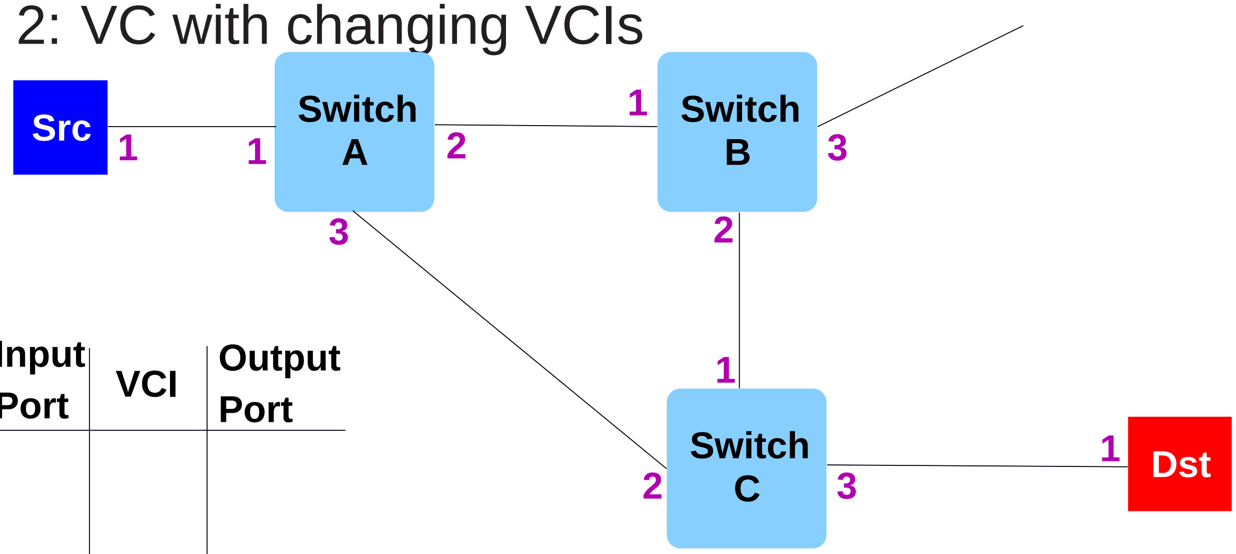
Input Port	VCI	Output Port

Input Port	VCI	Output Port

Input Port	VCI	Output Port

Virtual Circuits: (continued)

- Example 2: VC with changing VCIs



Input Port	VCI	Output Port

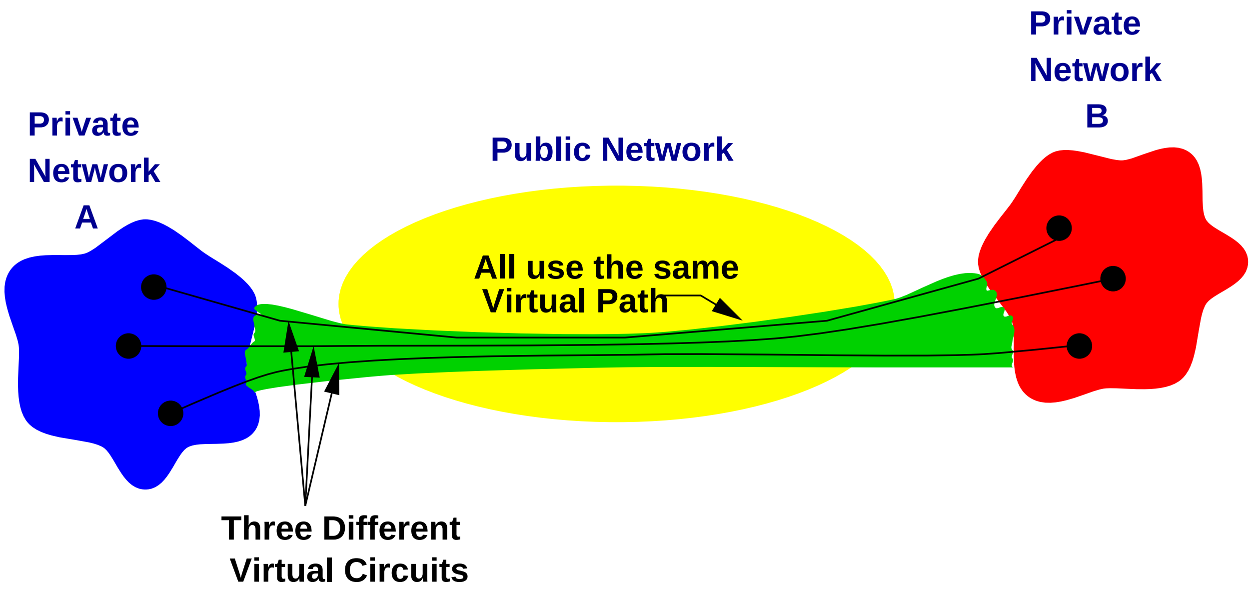
Input Port	VCI	Output Port

Input Port	VCI	Output Port

Virtual Circuits: (continued)

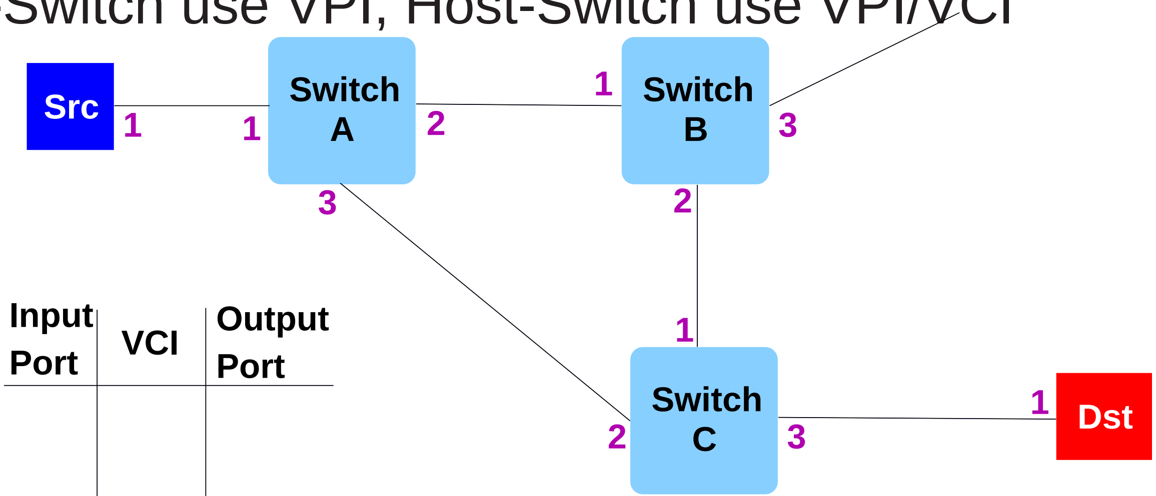
- Example 3: Hierarchical VCIs
  - Each VC identifier broken into two parts:
    1. VPI (Virtual Path Identifier)
    2. VCI (Virtual Circuit Identifier)
  - VPI used for similar destination traffic
  - VCI used with VPI to uniquely identify a VC

Virtual Circuits: (continued)



Virtual Circuits: (continued)

- Example 3 (continued): ATM Hierarchical VCs
  - Switch-Switch use VPI, Host-Switch use VPI/VCI



Input Port	VCI	Output Port

Input Port	VCI	Output Port

Input Port	VCI	Output Port

Connectionless Communication

- No connection necessary
- Source of data adds destination information to data and gives it to the network to deliver
- Network delivers each packet individually

- Packets can be delivered to multiple recipients

Communication Direction

- communication may be uni-directional
- communication may be bi-directional

Reliable Communication

- some apps may require reliable delivery
- other apps may be happy with best-effort delivery
- some apps may require that packets be delivered in the order they were sent (ordered delivery)
- other apps may not care about the ordering (out-of-order delivery)
- How reliable should reliable delivery be?

- We will revisit the issue of reliable delivery in the next chapter

Miscellaneous Services

- Request/Reply style communication
- Secure packet delivery
- Compressed/Encoded packet delivery
- Guaranteed bandwidth (what does this mean?)

- Guaranteed max delay
- Jitter bounded delivery
- ... etc ...

Performance

- users will want the network to delivery certain performance
- how do we define performance?
- what are the factors influencing performance?

Bandwidth/Throughput

- Bandwidth describes the number of bits that can be transmitted over the network in some period of time. Usually measured in Millions of bits/sec (Mbps).
- Really a measure of how fast we can send data out of a machine
- Thus you could think of it as the amount of time it takes to send out one bit
- In the analog world it is typically measured in Hertz (Hz)
- “b” vs. “B”

Latency/Delay

- Latency describes how long it takes a single bit to propagate from one machine to another machine.
- Measured in terms of time
- We can also use computer instructions to measure it (which is important since it tells how much work the computer can get done during the latency).
- Sometimes we are really more interested in Round-Trip-Time (RTT) as an alternate way to measure latency.

- Why are we interested in RTT?

Types of Latency

- Propagation Latency: The time it takes for 1 bit to travel across the wire.
  - cannot go faster than the speed of light
  - the speed of light depends on the media; The speed of light is:
  - the speed of light can even vary for the same basic type of media

Types of Latency: (continued)

- Transmission Latency: The time it takes to send a packet out of the machine.
- This depends on
  - the bandwidth of the network, and
  - the amount of data you want to send (i.e., packet size).
- may depend on operating system efficiency too

Types of Latency: (continued)

- Store-n-forward packet switches store packets for a (hopefully) brief amount of time before forwarding the packet.
- Usually the packet is stored in a *queue of packets* all waiting to be forwarded.
- Thus each bridge/switch/router between the source and destination adds a Queuing Delay to the overall latency.
- some people define a related term, Access Latency , as the time required to get access to the media. Others lump access latency in with the queueing delays.

Types of Latency: (continued)

Summary

- Latency = Propagation Delay + Transmission Delay + Queuing Delay
- Propagation Delay = Wire Distance / Speed-of-Light-on-wire
- Transmission Delay = Packet Size / Link Bandwidth
- Queuing Delay = (depends but) Avg Switch/Router Delay $\times$ Number of Routers on Path
- USA Latency (min over 3000 miles) = 24 ms

Types of Latency: (continued)

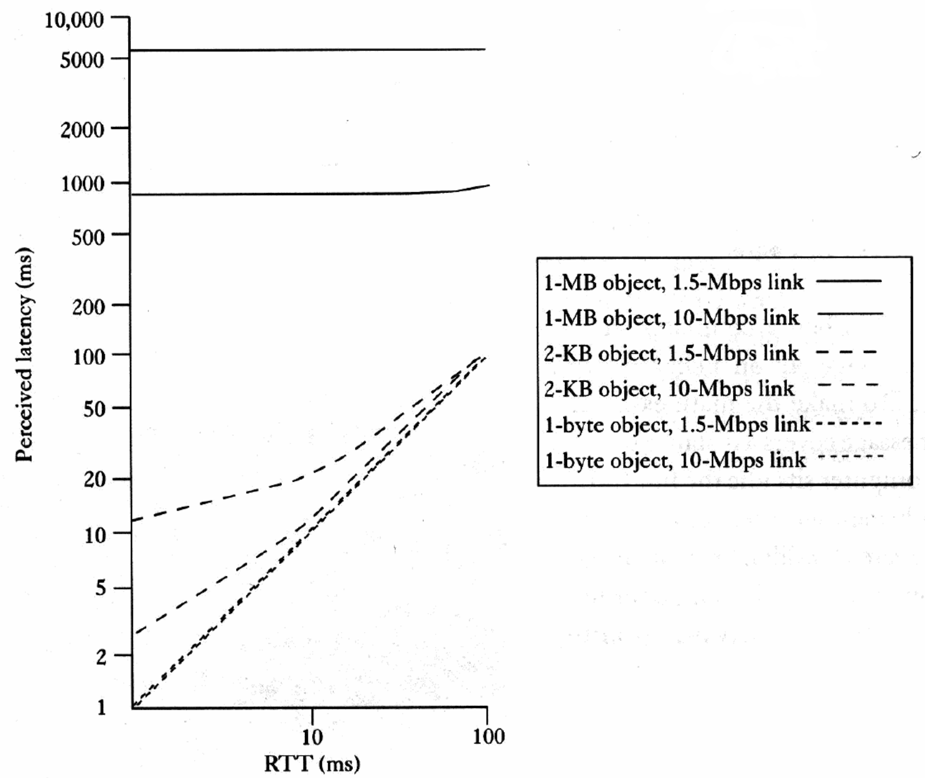


Figure 1: Perceived Latency (response time) versus Round Trip Time for various object sizes and link speeds.

Delay x Bandwidth Product

How much data can be in transit at any given time? The **Delay** $\times$ **Bandwidth Product** gives the volume of the network "pipe".

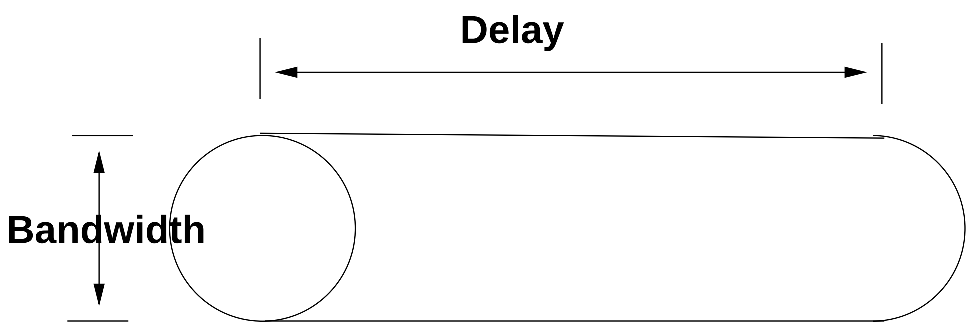


Figure 2: DelayxBandwidth Product

Delay x Bandwidth Examples: (continued)

Ethernet: Delay = 0.5 ms
Bandwidth = 10 Mbps
Holds = (0.5 x 10^{-3} sec) x (10 x 10^{6} bits/sec) = 5 x 10^{3} = 5 Kbps = 0.625 KBps

Modem Line: Delay = 85 ms
Bandwidth = 28.8 Kbps
Holds = (85 x 10^{-3} sec) x (28.8 x 10^{3} bits/sec) = 2.448 x 10^{3} = 2.5 Kbps = 0.31 KBps

Cross Country DS3 Line: Delay = 50 ms
Bandwidth = 45 Mbps
Holds = (50 x 10^{-3} sec) x (45 x 10^{6} bits/sec) = 2.25 x 10^{6} = 2.25 Mbps = 0.28 MBps

Delay x Bandwidth: (continued)

- The Delay x Bandwidth product is important because it is the amount of data the source will transmit before the first bit arrives at the destination!
- If we consider the RTT , a sender will transmit 2 x Delay x Bandwidth amount of data before hearing anything back from the destination.
- Thus, A sender might transmit a lot of data before it finds out if an error occurred .