CMPUT 313 Course Notes
(Part 3)
Telecommunications and Computer Networks

Janelle Harms
Updated Winter 2012

These notes are based on many sources including:

S. Tanenbaum, Computer Networks, Prentice Hall

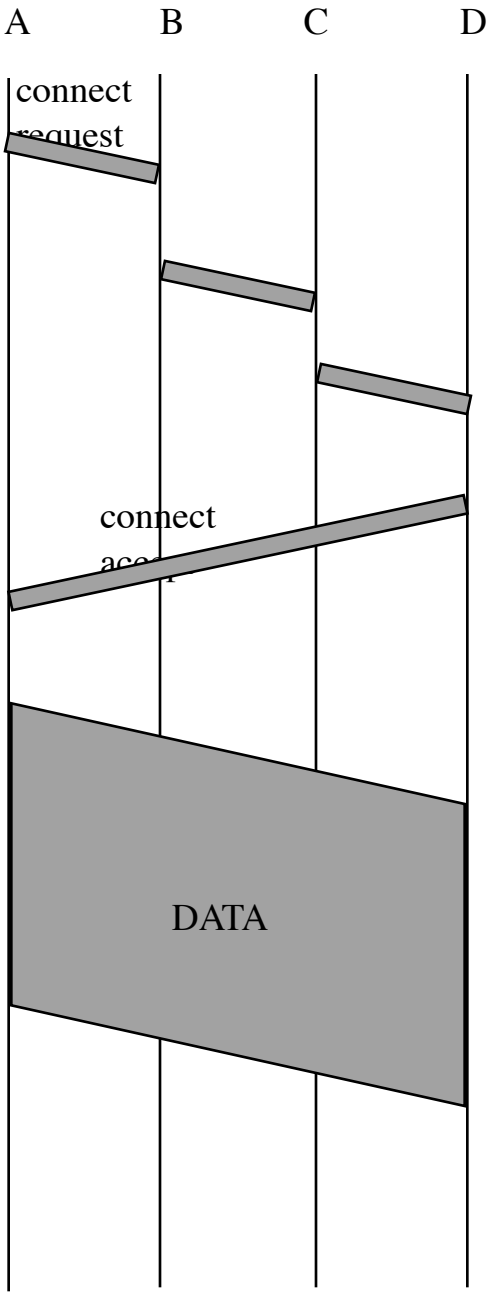
J. F. Kurose & K. W. Ross, Computer Networking, Addison Wesley

W. Stallings, Data and Computer Communications, Prentice Hall

Switching

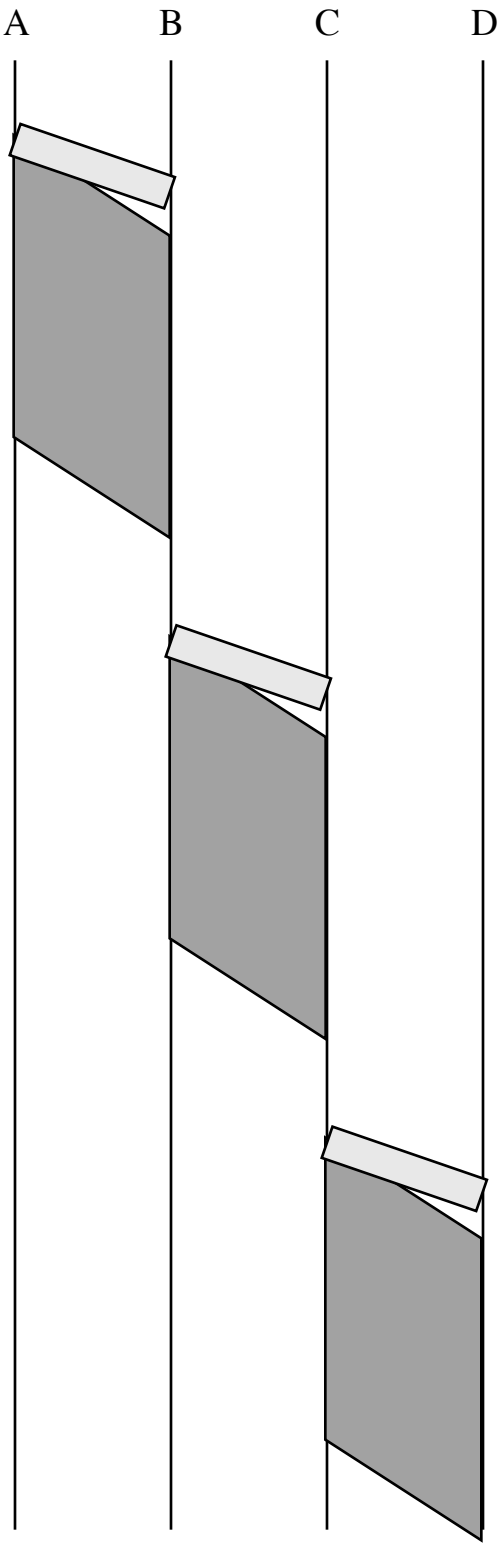
Circuit Switching

- establish a physical connection
- transmit data on line assigned to user
- disconnect



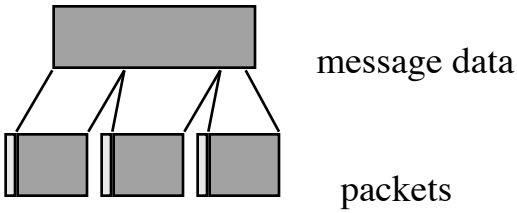
Message Switching

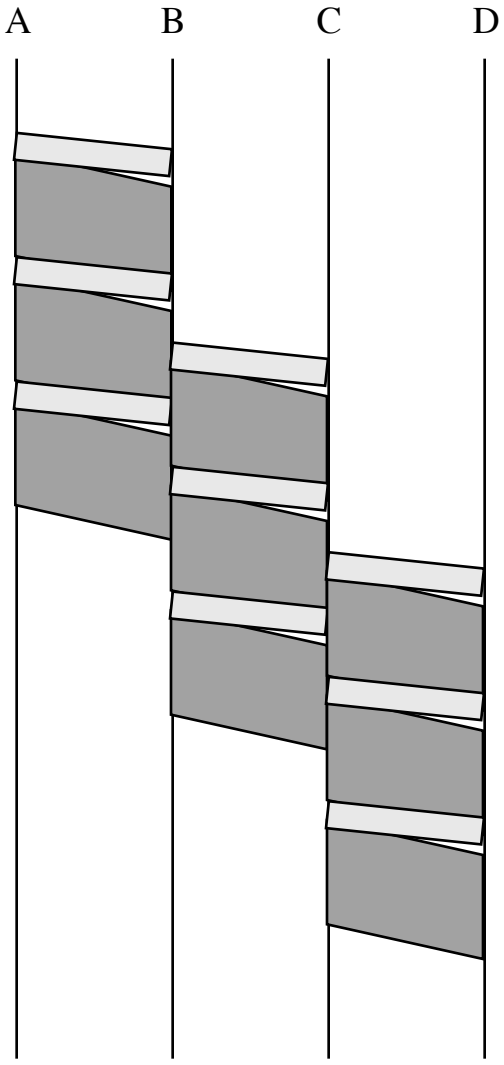
- message is a self-contained data unit
- store-and-forward



Packet Switching

- divide message into smaller units: packets





Network Layer

The **network layer** is responsible for packet delivery over the subnet.
Issues: routing, internetworking, congestion control

Services to the Transport Layer: connection-oriented vs connectionless.

Packet Transport Strategies

Datagram

- packet carries source and destination addresses
- packets are routed independently
- delivery is not guaranteed and packets may arrive out of order

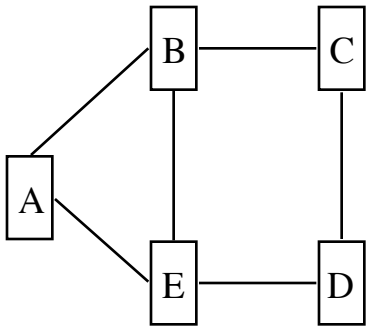
Virtual Circuit

- route from source to destination is chosen when a virtual circuit is set up
- each packet contains a VC identifier (VCI); each node has a table of VCs
- to avoid ambiguity, a VC may have different (but unique) id's between adjacent nodes

Routing Algorithms

Routing algorithms decide on which outgoing link an incoming packet should be transmitted.

Consider the following network of routers:



Many routing algorithms produce and maintain a routing table giving destination and preferred output link.

The preferred output link could be based on shortest path using

- distance metrics: hopcount, physical distance
- delay metrics: queuelength, estimated end-to-end delay

Routing Table for A

dst	output link
A	-----
B	B
C	B
D	E
E	E

Distance Vector Routing

- Each node has a routing table containing the following information for each destination:
  - preferred output link
  - estimated delay or distance to the destination (possible measures are hopcount, delay, number of packets queued on the path)
- Neighbouring nodes exchange delay information periodically and update both the delay and preferred output links

- Distance vector routing has problems with stability. It reacts quickly to good news (link comes up) but may react very slowly to bad news (link goes down: see Tane96 on "count to infinity" problem)

Routing Table for B

dst	link	delay
A	A	1
B	--	0
C	C	1
D	C	2
E	E	1

B receives the following distance vectors from:

A: 01221
C: 21012
E: 11120

Suppose the delays/distances represent queued packets and B has 5 packets queued to A, 3 " " to C, and 1 " " to E

New Routing Table for B

dst	link	delay
A	E	2
B	--	0
C	C	3
D	E	2
E	E	1

Link State Routing

- Each node periodically sends information about delay/distance to its neighbours to all other nodes.
- At each node, a Link State Packet is created with src, seq#, age and a list of the distance/delay to each neighbour of that node. For example:

src B	seq# 2	age 10	A 5 C 3 E 1

- Packets are distributed using flooding.
  - At each node, the latest LSP for each src is stored
  - If the incoming seq# > stored seq# for the source, the LSP is transmitted on all links except the link it came in on. Otherwise it is discarded.
  - The age field is decremented at every hop. If age is 0, the LSP is not forwarded.
- Each node receives LSPs from all other nodes and therefore has all th e information needed to find shortest paths between any nodes.

Broadcast Routing

Broadcast routing is required when a host needs to transmit messages to all other hosts. There are many ways of doing this:

- **Separately addressed packets:** sources sends a distinct packet to each destination

- **Flooding:** each node transmits the packet it receives on all links except the incoming link. Flooding needs a stopping mechanism (e.g. hopcount field, router memory)

- **Multidestination routing:** each packet contains information on desired destination. At a node, the packet is copied onto the links that provide the best routes to the destinations (based on the unicast routing table). The desired destination field is adjust accordingly

- **Spanning tree broadcasting:** a spanning tree is pre-specified (known to all nodes). An incoming broadcast packet is copied onto all output links that are part of the spanning tree (except the incoming link).

- **Reverse path forwarding**: if incoming link is the preferred link for sending packets back to the source, then forward the packet onto all links except the incoming; otherwise, discard the packet.

Multicast routing

Core-based tree

Source-based tree

Ad Hoc Wireless Networks

Ad hoc wireless networks have no fixed infrastructure

Characteristics

- Each mobile must forward traffic for other mobiles
  - Multi-hop routing
- They may be mobile
  - Topology changes continuously
- Mobiles have limited battery power
  - Must conserve energy
- Different mobile hosts may have different transmission ranges
  - Asymmetric paths

These characteristics create very challenging problems in providing basic connectivity, routing functionality and reliability.

Routing: on-demand versus pro-active

flooding

flood request for path, update tables or source route: DSR, AODV

position-based routing

Internetworking

Internetworking occurs at all layers:

- Layer 1: **repeaters** copy bits between LAN segments.
- Layer 2: **bridges** store and forward frames between LANs
- Layer 3: **routers** forward packets between different networks
- Higher Layers: **transport gateways**, **application gateways**

Bridges

- store and forward frames between LANs
- protocol differences:
  - frame formats
  - speeds
  - maximum frame lengths
  - other: priority, acknowledgment, etc.

Internetworking at the Network Layer

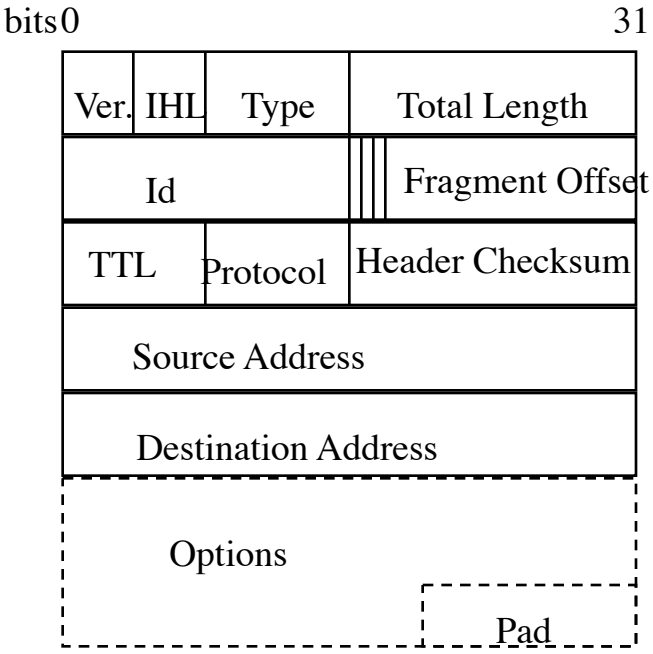
Issues:

- connection-oriented vs connectionless
- protocol conversion and address conversion
- packet size
  - fragmentation
    - transparent or non-transparent
    - identification of fragments
- support for multicast and broadcast
- effect of congestion, flow and error control schemes
- routing

Internet Protocol (IP)

- IP is a network layer protocol designed specifically for internetworking.
- It has simple functionality
  - an addressing scheme
  - best effort delivery

- **Header:**

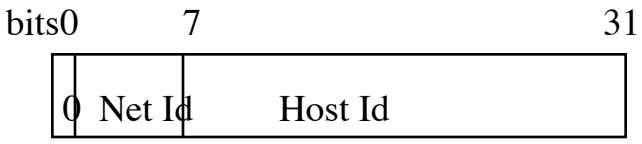


Version: 4
Header Length: measured in 32 bit words
Service Type:
Total Length: header plus data; measured in bytes (max 65,535 bytes)
Id: identifies packet
Flags: DF - do not fragment;
MF - more fragments
Fragment Offset: max 8192 fragments/packet
TTL: time to live
Protocol: transport layer protocol
Header checksum:
Source and Destination Addresses: (see later)
Options: for example, source routing, timestamps

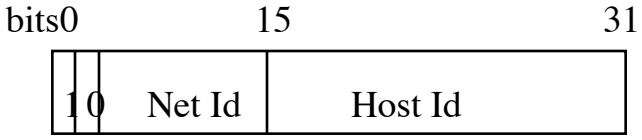
Internet Addressing

The Old Method (hierarchical)

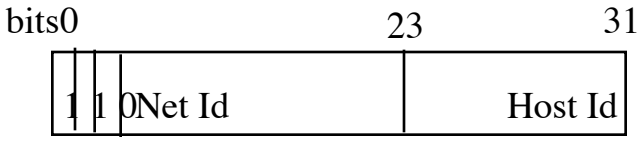
Class A Addresses



Class B Addresses



Class C Addresses



Class D Addresses

Internet addressing has evolved

Problems with the old scheme: poor utilisation of allocated address space and reaching the limits of the address space

Solutions

Subnetting

- Allows host part to be split into (subnet, host)
- Provides more organisation to local network.
- Uses a **mask** to remove the network part

Classless Internet Domain Routing (CIDR)

- Allocate variable-sized blocks (without regard to classes)
- More complex forwarding (routing tables need address plus 32 bit mask)
- Longest match is used to forward (allows grouping of addresses)

Development of new protocol version (IPv6)

- Has more bits to represent addresses
- Problem of phasing it in with legacy of old IPv4

IPv6: IP Version 6

- started designing early '90's; finished in late '90's.
- The transition has been difficult.

Design

- extended addresses (128 bits)
- streamlined header for faster processing
- ability to add QoS parameters

Header Contents: always 40 bytes

Version (4 bits)	6
Traffic Class (8 bits)	for QoS
Flow Label (20 bits)	for QoS

Payload length (16 bits) -- gives # bytes after 40 byte header

Next header (8 bits) -- transport layer protocol or pointer to options

Hop Limit (8 bits) -- decremented at each hop

DST Addr (128 bits)

SRC Addr (128 bits)

Note: --no fragmentation: if too small, router drops and sends error message

-- no header checksum

Internet Control Protocols

There are several other protocols used at the network layer of the Internet:

Internet Control Message Protocol (ICMP)

ICMP is used to report errors and to support testing. ICMP messages are encapsulated in IP packets. Some types of messages are:

- Destination unreachable
- Time exceeded
- Echo request, echo reply
- Timestamp request and reply

Address Resolution Protocol (ARP)

This protocol maps an IP address to a MAC address.

- The host that requires the mapping, broadcasts "who owns this IP address"; the owner responds with its Ethernet address.

There is also a Reverse Address Resolution Protocol (RARP) to go from a MAC address to an IP address.

Interior Gateway Routing Protocols

The Internet is constructed of several Autonomous Systems separated by gateway routers. Each Autonomous System has its own internal routing protocol.

Originally this was a distance vector protocol called RIP (Routing Internet Protocol). Now the recommended protocol is **Open Shortest Path First (OSPF)**

- It uses link state routing, supports different distance metrics, provides load balancing, and supports hierarchical routing.

Exterior Gateway Routing Protocol:

There must be a routing algorithm between Autonomous Systems

The Border Gateway Protocol (BGP)

- Uses distance vector routing (but communicates full paths)
- Routing information is exchanged between gateways using TCP.

Mobile IP

Provides location service for mobile users.

Design

- Each Mobile Host (MH) must be addressable using IP (home IP address)
- Fixed host software remains unchanged
- No overhead if the MH is at “home”

Mechanism

- A Mobile Host (MH)will register with a foreign agent when it leaves its home base.
- The Foreign Agent (FA)
  - Maintains information about MHs currently in its area.
  - Forwards traffic destined for the MHs in its area
- The Home Agent (HA)
  - Maintains information on the location of the MH (that is, it has the address of the FA).
  - Answers ARP requests for the MH IP address when the MH is away.
  - When a packet arrives for the MH, the HA tunnels the packet to the FA and tells the sender where to forward future packets.

Transport Layer

Transmission Control Protocol (TCP)

TCP is a transport layer protocol that uses IP at the network layer.

Design:

- End-to-end communication
- Connection-oriented
- Reliable delivery
- Byte stream
- Flexible -- accommodates heterogeneous end systems and adapts to changing network performance

At the sending host, after setting up a connection, the TCP protocol will:

- Accept data stream from the application
- Divide the data into pieces.
- Add a TCP header to each piece to form **segments**.
- Send segments using IP packets.

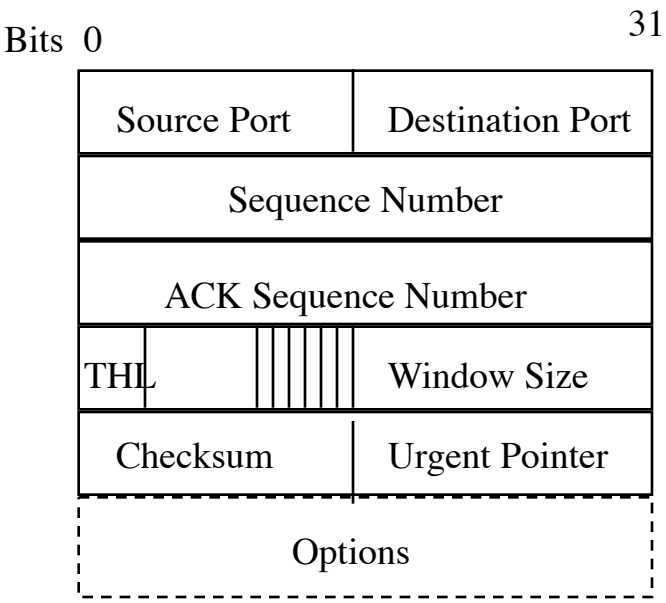
At the receiving host, the TCP protocol will:

- Reassemble the byte stream
- Deliver the data to the application

Connection Management:

- Connection establishment: 3-way handshake
- Connection release:

Header:



SRC, DST Ports: endpoints within the hosts
Sequence Number: first byte in data portion
Acknowledgement Number: next byte expected
TCP Header Length: number of 32 bit words in the header
Flags: URG: 1 if the urgent pointer is in use
ACK: 1 if ACK number is valid
PSH: 1 if must deliver data to application on arrival
RST: reject invalid segment, reset or refuse connection
SYN: used in connection establishment
FIN : release connection
Window Size: number of bytes that can be accepted
Checksum: checks header, pseudoheader (from IP) and data
Urgent Pointer: byte offset where urgent data is found
Options: for example, increase window size, Selective Repeate vs Go Back N, etc

TCP's Transmission Mechanism

TCP uses a sliding window protocol with variable maximum sender window size and variable timeout values.

- The receiver advertises the number of bytes it can accept (window size field) starting with the next byte expected (ACK field)
- The sender has at most window size number of bytes outstanding (unacked). The sender window size is the minimum of the receiver advertised window and the congestion window
- Flexibility: the sender can wait to fill buffer before transmitting; receiver can wait to receive more data before sending an acknowledgement.

- Some problems were identified and subsequent fixes to the protocol were made:
  - There is inefficiency if the sender and receiver send small amounts of data.
    * Receiver inefficiency can be fixed by delaying ACKs
    * Sender inefficiency is solved by Nagles' Algorithm (which does some accumulation at the sender side)
  - Silly window syndrome occurs when the receiver sends very small window updates. The solution is to force the receiver to wait until there is adequate space in its buffer.

TCP's Congestion Control Mechanism (Tahoe Version)

Congestion is detected by the loss of segments and controlled by reducing the sender window size

- Sender window = min (receiver advertised window, congestion window)

- The congestion control algorithm uses: the slow start algorithm, a threshold value to control window growth and variable timeout values.

Congestion Control Algorithm

CW = congestion window size (in bytes)
MSS = maximum segment size (in bytes)
T = threshold(in bytes)

For simplicity of discussion, assume that the sender always has data to send and, therefore, will send segments of size MSS when possible. Also assume that the receiver window is large.

If timeout occurs
T <-- CW / 2
CW <-- MSS
Execute the **slow start algorithm** (see below) until CW reaches T
After CW reaches T, for each "window" (CW bytes) acknowledged, increase CW by MSS

Note: if there is no timeout, the window grows to the receiver advertised window.

Slow Start (Jacobson)

This algorithm is used at the start of the connection and within the congestion control algorithm

CW <-- MSS
Sender sends 1 segment of size MSS and starts timer
If ACK arrives before timeout
CW <-- CW + MSS
sender sends 2 segments of size MSS
if both ACKed, the sender will send 4 segments of size MSS, etc

Effectively, as long as no timeout occurs and the receiver window is large, for each window (burst) ACKed, CW is doubled.

Timeout Values

The timeout value is dynamically set based on measurements of roundtrip time in the network.

E-RTT is the estimated round trip time
S-RTT is the sampled round trip time

E-RTT <-- α E-RTT + (1 - α) S-RTT
where α is the smoothing factor

Originally, the timer value was set to 2 E-RTT.

Updated version

The original version was found to be too rigid if there was large variation in S-RTT.
Therefore, variation is also calculated

$D = \delta D + (1 - \delta) \mid E\text{-}RTT - S\text{-}RTT \mid$

The Timer value is set to E-RTT + 4 D

Another Problem

Handling duplicate ACKs can be a problem.

Karns' Algorithm solution: if a packet is being retransmitted, suspend the estimation of RTT and double the timer on each successive failure.

TCP Variations

Reno Version: attempts to detect and react to loss sooner and in a less-drastic manner.

Fast Retransmit:

If the sender receives 3 duplicate ACKs for a segment (before it times out), then retransmit the segment and trigger Fast Recovery

Fast Recovery: (somewhat simplified)

T <-- CW / 2 and CW <-- T + 3 MSS

increase CW linearly with duplicate ACKS until ACK arrives for retransmitted packet (then it resets CW back to T)

(Note: slow start is only used at connection start up and when a timeout occurs)

New Reno Version

SACK

User Data Protocol (UDP)

This is the connectionless, "unreliable" transport layer protocol of the Internet.

It has an 8 byte header: Src Port, Dst Port, UDP length (header and data), UDP checksum (like TCP, the checksum also checks the pseudoheader)

The checksum could be set to 0 if not used. It if is used, the segment is dropped if an error is detected. There is no retransmission scheme for this protocol.

Application Layer Protocols

Domain Name Servers (DNS)

People prefer to deal with names and not numbers. The DNS will map name addresses to IP addresses.

Design:

- Distributed database with redundancy
- Domain-based: hierarchical structure
  - Name addresses are also hierarchical cs.ualberta.ca
  - High-level domains include: generic (for example .com,.edu,.gov, …) and country (for example .ca is Canada)
  - A local name server processes host queries; if it can't satisfy a request it sends the request up the hierarchy.
  - Requests are sent using UDP

File Transfer Protocol (FTP)

Application protocol to transfer a file from 1 host to another regardless of platform

FTP establishes 2 TCP connections: a control connection and a data connection

Simple Mail Transfer Protocol (SMTP)

The principle application protocol for electonric mail.

It uses TCP to transfer mail from the sender's mail server to the recipient's mail server.

Hypertext Transfer Protocol (HTTP)

This protocol defines how messages are passed between a browser and a web server.

It uses TCP to transport requests for objects and replies from the server.

- **Non-persistent** (http 1.0) a new connection is set up for each referenced object on the page.
- **Persistent** (http 1.1) the TCP connection is left open to allow more objects to be requested and transferred. It may allow **pipelining** were a new request can be issued before an old one is answered

Quality of Service (QoS)

Application or User View:

- Requires "service" from the network of a particular "quality".
- Different applications have different requirements: delay, loss, bandwidth

Network View

- Controls resources such as bandwidth and buffers
- Attempts to provide service to many different applications
- Problems to overcome: internetworking, possible congestion

Asynchronous Transfer Mode

Telecommunications (telephone companies) answer to providing broadband integrated services.

- Packet switching with fixed size packets called **cells**
- Independent of transmission speed
- Asynchronous transmission (statistical multiplexing)
- Connection-oriented
  - Uses virtual circuit routing
  - No acknowledgements (or retransmission)
- Designed to provide Quality of Service
  - Admission Control at connection set-up time
    - User specifies a traffic descriptor: peak, average and minimum cell rate
    - User also pecifies tolerance to cell loss ratio, cell transfer delay, cell delay variation
    - If the network can handle the new user, it reserves resources
  - Connections are "policed" using leaky bucket
  - Resource reservation at connection time (admission control)
  - Defines several classes of service
    - CBR: constant bit rate
    - VBR: variable bit rate
    - ABR: available bit rate
    - UBR: unspecified bit rate

Quality of Service over the Internet

The Pieces

- Admission control: limit the use of resources
- Classifier mechanism: sort customers into groups based on QoS requirements
- Packet Scheduler: control length of queues; differentially treat traffic of different classes
- Routing: balancing load through routing
- Signalling: passing of resource allocation information

Problem: legacy of existing networks; not everyone will change

Integrated Services (IntServ)

QoS provided by:

- A signalling protocol to reserve resources (RSVP)
- Admission control
- Classification and packet scheduling

Application provides traffic descriptor: peak rate, token bucket parameters, maximum packet size.

Network reserves resources on path to destination to support this traffic. If not enough resources, service is denied

Traffic that does not meet traffic specification is tagged and potentially discarded.

Classes of Service:

- Best effort
- Controlled load
- Guaranteed service

Problems with IntServ: poor scalability (flow-based state)

Differentiated Services (DiffServ)

Places complexity at the edge of the network instead of within the network

QoS provided based on:

- Policing and shaping at access end
- Packet forwarding scheme based on class of service bits in the IP header

User has a service level agreement (SLA) with the the Internet Service Provider(ISP) to receive differentiated services.

Premium Service

- Provides low delay and low jitter
- SLA specifies peak bandwidth
- Shaped at ingress and policed with excess dropped
- Premium queue: forward before other classes
- ISP limits the number of users in this class

Assured Service

- Better service than best effort
- Policed at ingress, marked as in or out
- Packets are placed in the Assured queue

Quality of Service Routing

- Select routes based on quality of service requirements of application
- Routes may be based on multiple constraints (bandwidth and delay, etc)