Chapter 7 – Networking Support
Contents

- Packet-switched networks.
- The Internet.
- Web access and TCP congestion control.
- Network management; class-based queuing.
- Cloud interconnection networks.
- Storage area networks.
- Content delivery networks.
- Overlay networks.
Packet-switched networks

- A packet-switched network transports data units called packets through a maze of switches where packets are queued and routed towards their destination.
- A packet-switched network consists of:
  - **Network core** made up from routers and control systems interconnected by very high bandwidth communication channels.
  - **Network edge** where the end-user systems/hosts reside.
- **Packet** consists of a header which contains control information necessary for its transport through the network and a payload or data.
- Packets are subject to a variable delay, errors, and loss.
- A **network architecture** describes the protocol stack.
- **Protocol** a discipline for communication, it specifies the actions taken by the sender and the receiver of a data unit.
- **Host** a system located at the network edge capable to initiate and to receive communication, e.g., computer, mobile device, sensor.
The Internet

- Collection of separate and distinct networks.
- All networks operate under a common framework consisting of:
  - globally unique IP addressing.
  - IP routing.
  - global Border Gateway Routing (BGP) protocols.
- IP only provides best effort delivery - any router on the path from the source to the destination may drop a packet if it is overloaded.
- The Internet uses two transport protocols
  - UDP (User Datagram Protocol) - a connectionless datagram protocol. The UDP transport protocol assumes that error checking and error correction are either not necessary or performed by the application. Datagrams may arrive out of order, duplicated, or may not arrive at all.
  - TCP (Transport Control Protocol) - a connection-oriented protocol. TCP provides reliable, ordered delivery of a stream of bytes from an application on one system to its peer on the destination system.
The Internet protocol stack

Application Layer
- Teleconferencing
- Videoconferencing
- RealAudio
- WWW
- Email
- FTP

Transport Layer
- TCP
- UDP

Network Layer
- IP

Physical and Data Link Layers
- ATM
- Dial-up Modems
- LANs
- Wireless
- Direct Broadcast Satellite
- Cable
- Frame Relay
The Internet protocol stack (cont’d)

Streams of bits encoded as electrical, optical, or electromagnetic signals
IPv4 vs IPv6

- IPv4 has an addressing capability of $2^{32}$, or approximately 4.3 billion addresses, a number that proved to be insufficient.

- IPv6 has an addressing capability of $2^{128}$, or $3.4 \times 10^{38}$ addresses.

- Other major differences between IPv4 and IPv6:
  - IPv6 supports new multicast solutions and but not traditional IP broadcast.
  - IPv6 hosts can configure themselves automatically when connected to a routed IPv6 network using the Internet Control Message Protocol version 6.
  - Mandatory support for network security. Internet Network Security (IPsec) is an integral part of the base protocol suite in IPv6.

- Migration from IPv4 to IPv6 is a very challenging and costly proposition.
IP and MAC addresses, ports and sockets

- **IP address** → logical address assigned dynamically by a DHCP server. A host may have multiple IP addresses as it may be connected to more than one network.

- **MAC address** → unique physical address of each network interface.

- **Network interface** → hardware connecting a host with a network.

- **Port** → software abstraction for message delivery to an application.

- **Sockets** → software abstraction allowing an application to send and receive messages at a given port; implemented as two queues, one for incoming and the other for outgoing messages.
Sockets and ports

IP address: NetworkId+HostId
The relations between Internet networks

- Three type of relations:
  - Peering - two networks exchange traffic between each other's customers freely.
  - Transit - a network pays to another one to access the Internet.
  - Customer - a network is paid to allow Internet access.

- The networks are commonly classified as:
  - Tier 1 - can reach every other network on the Internet without purchasing IP transit or paying settlements.
  - Tier 2 - an Internet service provider who engages in the practice of peering with other networks, but who still purchases IP transit to reach some portion of the Internet; the common providers on the Internet.
  - Tier 3 - purchases transit rights from other networks (typically Tier 2 networks) to reach the Internet.
The relation of Internet networks based on the transit and paying settlements. There are three classes of networks, Tier 1, 2, and 3; an IXP is a physical infrastructure allowing ISPs to exchange Internet traffic.
The transformation of the Internet

- Web applications, cloud computing, and content-delivery networks are reshaping the definition of a network.
- Data streaming consumes an increasingly larger fraction of the available bandwidth as high definition TV sets become less expensive and content providers, such as Netflix and Hulu, offer customers services that require a significant increase of the network bandwidth.
- The “last mile” - the link connecting the home to the Internet Service Provider (ISP) network is the bottleneck.
- Google has initiated the Google Fiber Project which aims to provide 1Gb/s access speed to individual households through FTTH.
The transformation of the Internet. The traffic carried by Tier 3 networks increased from 5.8% in 2007 to 9.4% in 2009; Google applications accounted for 5.2% of the traffic in 2009.
The average download speed for broadband access advertised by several countries

Dan C. Marinescu
Chapter 7
14
Web access and TCP

- HTTP - the application protocol for Web access uses the TCP transport protocol.
- TCP supports mechanisms to avoid congestion and limit the amount of data transported over the Internet.
- Web access requires the transfer of large amounts of data as we can see in measurements reported by Google.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sample pages analyzed</td>
<td>$4.2 \times 10^9$</td>
</tr>
<tr>
<td>Average number of resources per page</td>
<td>44</td>
</tr>
<tr>
<td>Average number of GETs per page</td>
<td>44.5</td>
</tr>
<tr>
<td>Average number of unique host names encountered per page</td>
<td>7</td>
</tr>
<tr>
<td>Average size transferred over the network per page, including HTTP headers</td>
<td>320 KB</td>
</tr>
<tr>
<td>Average number of unique images per page</td>
<td>29</td>
</tr>
<tr>
<td>Average size of the images per page</td>
<td>206 KB</td>
</tr>
<tr>
<td>Average number of external scripts per page</td>
<td>7</td>
</tr>
<tr>
<td>Number of sample SSL (HTTPS) pages analyzed</td>
<td>$17 \times 10^6$</td>
</tr>
</tbody>
</table>
Congestion control in TCP

- Algorithms to control congestion include Tahoe, an algorithm based on: (1) slow start, (2) congestion avoidance, and (3) fast retransmit.

- Slow start means that:
  - (a) the sender starts with a window of two times MSS (Maximum Segment Size).
  - (b) for every packet acknowledged, the congestion window increases by 1 MSS so that the congestion window effectively doubles for every RTT (Round Trip Time).

- To overcome the limitations of the slow start application, strategies have been developed to reduce the time to download data over the Internet. For example,
  - Firefox 3 and Google Chrome open up to six TCP connections per domain.
  - Internet Explorer 8 opens 180 connections.
Congestion control in TCP (cont’d)

- The strategies used by the browsers to avoid the congestion control mechanisms circumvent the mechanisms for congestion control and incur a considerable overhead.
- The TCP latency is dominated by the number of RTTs during the slow start phase.
- Given that the average page size is 384 KB, a single TCP connection requires multiple RTTs to download a single page.
- It is argued that a better solution is to increase the initial congestion window of TCP. The effects of this solution:
  - It ensures fairness between short-lived transactions which are a majority of Internet transfers and the long-lived transactions which transfer very large amounts of data, e.g., audio and video streaming.
  - It allows faster recovery after losses through Fast Retransmission.
Class-Based Queuing (CBQ)

- The objectives of CBQ are to support:
  - Flexible link sharing for applications which require bandwidth guarantees such as VoIP, video-streaming, and audio-streaming.
  - Some balance between short-lived network flows, such as web searches, and long-lived ones, such as video-streaming or file transfers.

- CBQ aggregates the connections and constructs a tree-like hierarchy of classes with different priorities and throughput allocations. CBQ uses several functional units:
  - A classifier which uses the information in the packet header to assign arriving packets to classes.
  - An estimator of the short-term bandwidth for the class.
  - A selector/scheduler which identifies the highest priority class to send next and, if multiple classes have the same priority, to schedule them on a round-robin base.
  - A delayer to compute the next time when a class that has exceeded its link allocation is allowed to send.
Class-Based Queuing (CBQ) - packets are first classified into flows and then assigned to a queue dedicated to the flow; queues are serviced one packet at a time in round-robin order and empty queues are skipped.
CBQ link sharing for two groups: A→short-lived and B→long-lived traffic, allocated 25% and 75% of the link capacity. There are three classes with priorities 1, 2, and 3: (i) Real-time (RT) and the video streaming have priority 1 and are allocated 3% and 60%, respectively, (ii) Web transactions and audio streaming have priority 2 and are allocated 20% and 10%, respectively; (iii) In interactive (Intr) and file transfer (FTP) applications have priority 3 and are allocated 2% and 5%, respectively.
Class-Based Queuing (CBQ)

- A class is
  - overlimit → if over a certain recent period it has used more than its bandwidth allocation (in bytes per second).
  - underlimit → if it has used less.
  - atlimit → if it has used exactly its allocation.

- A leaf class is
  - satisfied if it is underlimit and has a persistent backlog.
  - unsatisfied otherwise.

- A non-leaf class is unsatisfied if it is underlimit and has some descendent class with a persistent backlog.
There are two groups A and B and three types of traffic, e.g., web, real-time, and interactive, denoted as 1, 2, and 3. (a) Group A and class A.3 traffic are underlimit and unsatisfied; classes A.1, A.2 and B.1 are overlimit, unsatisfied and with persistent backlog and have to be regulated; (b) Group A is underlimit and unsatisfied; Group B is overlimit and needs to be regulated; class A.1 traffic is underlimit; class A.2 is overlimit and with persistent backlog; class B.1 traffic is overlimit and with persistent backlog and needs to be regulated.
Hierarchical Token Buckets (HTB)

- Hierarchical Token Buckets (HTB) is a link sharing algorithm inspired by CBQ.

- The Linux kernel implements HTB.

- Each class has
  - An assured rate (AR).
  - A ceil rate (CR).

- HTB supports borrowing → If a class C needs a rate above its AR it tries to borrow from its parent; then the parent examines its children and, if there are classes running at a rate lower that their AR, the parent can borrow from them and reallocate it to class C.
Hierarchical Token Buckets (HTB)

HTB packet scheduling uses for every node a ceil rate in addition to the assured rate.
Cloud interconnection networks

- While processor and memory technology have followed Moore's law, the interconnection networks have evolved at a slower pace and have become a major factor in determining the overall performance and cost of the system.

- The networking infrastructure is organized hierarchically: servers are packed into racks and interconnected by a top of the rack router; the rack routers are connected to cluster routers which in turn are interconnected by a local communication fabric.

- The networking infrastructure of a cloud must satisfy several requirements:
  - Scalability.
  - Low cost.
  - Low-latency.
  - High bandwidth.
  - Provide location transparent communication between servers.
Location transparent communication

- Every server should be able to communicate with every other server with similar speed and latency.
- Applications need not be location aware.
- It also reduces the complexity of the system management.
- In a hierarchical organization *true location transparency is not feasible* and cost considerations ultimately decide the actual organization and performance of the communication fabric.
Interconnection networks - InfiniBand

- Interconnection network used by supercomputers and computer clouds.
  - Has a switched fabric topology designed to be scalable.
  - Supports several signaling rates.
  - The energy consumption depends on the throughput.
  - Links can be bonded together for additional throughput.

- The data rates.
  - single data rate (SDR) - 2.5 Gbps in each direction per connection.
  - double data rate (DDR) - 5 Gbps.
  - quad data rate (QDR) – 10 Gbps.
  - fourteen data rate (FDR) – 14.0625 Gbps.
  - enhanced data rated (EDR) – 25.78125 Gbps.

- Advantages.
  - high throughput, low latency.
  - supports quality of service guarantees and failover - the capability to switch to a redundant or standby system.
Routers and switches

- The cost of routers and the number of cables interconnecting the routers are major components of the cost of interconnection network.
- Better performance and lower costs can only be achieved with innovative router architecture → wire density has scaled up at a slower rate than processor speed and wire delay has remained constant.
- Router – switch interconnecting several networks.
  - low-radix routers – have a small number of ports; divide the bandwidth into a smaller number of wide ports.
  - high-radix routers - have a large number of ports; divide the bandwidth into larger number of narrow ports
- The number of intermediate routers in high-radix networks is reduced; lower latency and reduced power consumption.
- The pin bandwidth of the chips used for switching has increased by approximately an order of magnitude every 5 years during the past two decades.
Network characterization

- The diameter of a network is the average distance between all pairs of nodes; if a network is fully-connected its diameter is equal one.

- When a network is partitioned into two networks of the same size, the bisection bandwidth measures the communication bandwidth between the two.

- The cost.

- The power consumption.
Clos networks

- Butterfly network → the name comes from the pattern of inverted triangles created by the interconnections, which look like butterfly wings.
  - Transfers the data using the most efficient route, but it is blocking, it cannot handle a conflict between two packets attempting to reach the same port at the same time.
- Clos → Multistage nonblocking network with an odd number of stages.
  - Consists of two butterfly networks. The last stage of the input is fused with the first stage of the output.
  - All packets overshoot their destination and then hop back to it; most of the time, the overshoot is not necessary and increases the latency, a packet takes twice as many hops as it really needs.
- Folded Clos topology → the input and the output networks share switch modules. Such networks are called fat tree.
  - Myrinet, InfiniBand, and Quadrics implement a fat-tree topology.
- A 5-stage Clos network with radix-2 routers and unidirectional channels; the network is equivalent to two back-to-back butterfly networks.
- The corresponding folded-Clos network with bidirectional channels; the input and the output networks share switch modules.
(a) A 2-ary 4-fly butterfly with unidirectional links.

(b) The corresponding 2-ary 4-flat flattened butterfly is obtained by combining the four switches $S_0$, $S_1$, $S_2$, and $S_3$, in the first row of the traditional butterfly into a single switch $S_0'$, and by adding additional connections between switches.
Storage area networks

- Specialized, high-speed network for data block transfers between computer systems and storage elements.
- Consists of a communication infrastructure and a management layer.
- The Fiber Channel (FC) is the dominant architecture of SANs.
- FC it is a layered protocol.
A storage area network interconnects servers to servers, servers to storage devices, and storage devices to storage devices.
FC protocol layers

- Three lower-layer protocols: FC-0, the physical interface; FC-1, the transmission protocol responsible for encoding/decoding; and FC-2, the signaling protocol responsible for framing and flow control.
  - FC-0 uses laser diodes as the optical source and manages the point-to-point fiber connections.
  - FC-1 controls the serial transmission and integrates data with clock information.
  - FC-2 handles the topologies, the communication models, the classes of service, sequence and exchange identifiers, and segmentation and reassembly.

- Two upper-layer protocols:
  - FC-3 is common services layer.
  - FC-4 is the protocol mapping layer.
FC (Fiber Channel) protocol layers

- **FC-0**: Physical Interface
- **FC-1**: Transmission Code
- **FC-2**: Signaling Protocol
- **FC-3**: Common Services
- **FC-4**: SCSI, IP, ATM
FC classes of service

- **Class 1**: rarely used blocking connection-oriented service.
- **Class 2**: acknowledgments ensure that the frames are received; allows the fabric to multiplex several messages on a frame-by-frame basis; does not guarantee in-order delivery.
- **Class 3**: datagram connection; no acknowledgments.
- **Class 4**: connection-oriented service for multimedia applications; virtual circuits (VCs) established between ports, in-order delivery, acknowledgment of delivered frames; the fabric is responsible for multiplexing frames of different VCs. Guaranteed QoS, bandwidth and latency.
- **Class 5**: isochronous service for applications requiring immediate delivery, without buffering.
- **Class 6**: supports dedicated connections for a reliable multicast.
- **Class 7**: similar to Class 2, used for the control and management of the fabric; connectionless service with notification of non-delivery.
### The format of a FC frame

<table>
<thead>
<tr>
<th>Word 0</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 3-6</th>
<th>(0-2112 bytes) Payload</th>
<th>CRC</th>
<th>EOF (End Of Frame)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 bytes</td>
<td>3 bytes</td>
<td>3 bytes</td>
<td>18 bytes</td>
<td>0-2112 bytes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SOF (Start Of Frame)</td>
<td>Destination port address</td>
<td>Source port address</td>
<td>Control information</td>
<td>Payload</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FC networks

- An FC device has a unique id called the WWN (World Wide Name), a 64 bit address, the equivalent of the MAC address.
- Each port in the switched fabric has its own unique 24-bit address consisting of: the domain (bits 23 - 16), the area (bits 15 - 08), and the port physical address (bits 07-00).
- A switch assigns dynamically and maintains the port addresses.
- When a device with a WWN logs into the switch on a port, the switch assigns the port address to that device and maintains the correlation between that port address and the WWN address of the device using a Name Server.
- The Name Server is a component of the fabric operating system, running on the switch.
Content delivery networks (CDNs)

- CDNs are designed to support scalability, to increase reliability and performance, and to provide better security. In 2013, Internet video is expected to generate over 18 exabytes of data per month.

- A CDN receives the content from an **origin server**, then replicates it to its **edge cache servers**; the content is delivered to an end-user from the “closest” edge server.

- A CDN can deliver static content and/or live or on-demand streaming media.
  - Static content - media that can be maintained using traditional caching technologies as changes are infrequent. Examples: HTML pages, images, documents, software patches, audio and video files.
  - Live media - live events when the content is delivered in real time from the encoder to the media server.

- Protocols used by CDNs: Network Element Control Protocol (NECP), Web Cache Coordination Protocol (WCCP), SOCKS, Cache Array Routing Protocol (CARP), Internet Cache Protocol (ICP), Hypertext Caching Protocol (HTCP), and Cache Digest.
CDN design and performance

- Design and policy decisions for a CDNs.
  - The placement of the edge servers.
  - The content selection and delivery.
  - The content management.
  - The request routing policies.

- Critical metrics for CDN performance
  - Cache hit ratio - the ratio of the number of cached objects versus total number of objects requested.
  - Reserved bandwidth for the origin server.
  - Latency - based on the perceived response time by the end users.
  - Edge server utilization.
  - Reliability - based on packet-loss measurements.
Overlay networks

- An overlay network, or a virtual network, is a network built on top of a physical network.
  - The nodes of an overlay network are connected by virtual links which could traverse multiple physical links.
  - Overlay networks are widely used in many distributed systems such as peer-to-peer systems, content-delivery systems, and client-server systems; in all these cases, the distributed systems communicate through the Internet.
Scale-free networks

- The degree distribution of scale-free networks follows a power law.

- Many physical and social systems are interconnected by a scale-free network. Empirical data available for power grids, the web, the citation of scientific papers, or social networks confirm this trend.

- The majority of the vertices of a scale-free network:
  - Are directly connected with the vertices with the highest degree.
  - Have a low degree and only a few vertices are connected to a large number of edges.
A scale-free network is non-homogeneous; the majority of vertices have a low degree and only a few vertices are connected to a large number of edges; the majority of the vertices are directly connected with the highest degree ones.
Epidemic algorithms

- Epidemic algorithm mimic the transmission of infectious diseases and are often used in distributed systems to accomplish tasks such as:
  - disseminate information, e.g., topology information.
  - compute aggregates, e.g., arrange the nodes in a gossip overlay into a list sorted by some attributes in logarithmic time.
  - manage data replication in a distributed system.

- *Game of life* is a popular epidemic algorithm invented by John Conway.

- Several classes of epidemic algorithms exist. The concepts used to classify these algorithms
  - Susceptible (S),
  - Infective (I),
  - Recovered (R)
  refer to the state of the population subject to infectious disease and, by extension, to the recipient of information in a distributed system.
Types of epidemic algorithms

- Susceptible-Infective (SI) algorithms
  - apply when the entire population is susceptible to be infected; once an individual becomes infected it remains in that state until the entire population is infected.

- Susceptible-Infectious-Recover (SIR)
  - based on the model developed by Kermack and McKendrik which assumes
    - the following transition from one state to another: \( S \rightarrow I \rightarrow R \);
    - that the size of the population is fixed: \( S(t) + I(t) + R(t) = N \).

- Susceptible-Infective-Susceptible (SIS)
  - are particular cases of SIR models when individuals recover from the disease without immunity. If \( p = R(r)/I(r) \), then the number of newly infected grows until \( (1-p)/2 \) are infected and then decreases exponentially to \( (1-p) \).