Link Layer Layer and LANS

- This note is based on Chapter 6 of the textbook. You are recommended to read the textbook for more details.

- Outline of Chapter 6
  - Overview
  - Error detection and correction
  - Multiple access protocols
  - Switched Local Area Networks
  - Link virtualization
  - Data center networking
  - A day in the life of a web request

- Overview
  We first introduce some terminology. We use node for a host or router, use link for a communication channel connecting nodes. We use frame for a layer-2 data packet which encapsulates a datagram from a network layer protocol, for example, an IP datagram.

  Data link layer protocols deal with the communication in a same network. More specifically, a data-link layer protocol transfers frames from one node to a physically adjacent node over a link. Different links may use different protocols for data transfer, for example, IEEE802.3 for Ethernet, IEEE802.11 for wireless LAN. Different links may provide different services.

- Link Layer Services
  Major services of a data link layer protocol include framing and link access. Framing is to encapsulate a datagram into a data frame. Link access is to transmit data frame to a link. A key component in link access is channel access if the communication medium of the link is shared by multiple nodes. At data link layer, MAC (also called physical/Ethernet) address is usually used to identify a network interface of a node. Error detection and correction are defined at data link layer for reliable data delivery. These functions usually used in wireless links. Flow control is also defined in data link layer. A link can be point-to-point link or multi-point link. A point-to-point link is a link between two nodes. A multi-point link is a link which can be shared by multiple (more than two) nodes. A link can be simplex, half-duplex and full-duplex. A simplex link is the one that data transmission is in one direction, like cable TV. A half-duplex link is the one that the transmission is in either of two directions but only in one direction at a time. A full duplex link is the one that the transmission can be in both directions simultaneously.

- Connecting Node to Link
  A node is connected to a link via a network interface card (NIC) which is a device consisting of hardware, software and firmware (see Figure 1). An NIC at a source node receives a datagram from a upper (network) layer protocol, encapsulates the datagram in a frame (adds
frame header, info for error detection and/or correction, flow control etc) and transmits the frame to a link. An NIC at a destination node receives a frame from a link, checks and/or recovers errors, extracts the datagram from the frame and passes the datagram to upper layer protocol.

**Error Detection and Correction**

In digital systems, redundant bits are often used for error detection and correction. For example, a parity bit is used to detect one bit error. A two dimensional bit parity check is to arrange the bit-sequence of data into a two-dimensional array and perform parity check for each row and each column (see Figure 2). This scheme can correct a single bit error.

Commonly used bit error detection methods include parity check, cyclic redundancy check (CRC), and checksum. Parity check and CRC are well used at the data link layer, while checksum is usually used at higher layers.

**Parity Check**

A sender appends a parity bit to a block of data and a receiver checks the parity of 1’s in the data. If the pre-defined parity is not changed then assume no error, otherwise an error is detected. The following example shows the efficiency of parity check. Assume that the probability of a single bit error $P_b = 10^{-6}$ and there are $F = 10^3 = 1024$ bits in the data. Let $P_e$ be the probability that a receiver accepts the data with an error.

- If no error detection is used,
  $$P_e \approx F \times P_b \approx 10^{-3}$$

- If parity check is used,
  $$P_e = \sum_{k=1}^{F/2} \binom{F}{2k} (P_b)^{2k} (1 - P_b)^{F-2k}$$
  $$\approx \frac{10^{-6}}{2}$$
The approach is simple but can only detect errors caused by odd number of bit changes. It is not reliable for large block of data.

- **Checksum**

A sender views the data (a string of binary bits) as rows of binary bits and computes a checksum code by some operations on the rows, examples include: computing the sum of the rows by bitwise exclusive OR as the checksum code; viewing each row as an integer and computing the sum of the rows (integers) as the checksum code. The sender sends both the data and the checksum code to the receiver which computes a checksum code from the received data by the same method used by the sender and compares the two checksum codes (one received from sender and the other computed by receiver). If the two checksum codes are the same, no error is assumed, otherwise there is an error during the transmission. Checksum is more powerful than parity check and often used by higher layer protocols.

- **Cyclic Redundancy Check (CRC)**

CRC is a more powerful error detection scheme used at link layer and works as follows: View a $k$ bits data block $D$ as a binary number. Sender

- Choose a number $G$ of ($r + 1$) bits.
- Make a binary number of $D2^r$.
- Divide $D2^r$ by $G$ to get remainder $R$ (CRC bits).
- Sends $M = D2^r \oplus R$.

Receiver checks if $M$ is a multiple of $G$. If so, no error is assumed. Otherwise, an error is detected.
• **CRC Example**
Assume that $D = 101011$ ($k = 6$ bits) and $G = 1001$ ($r + 1 = 4$ bits). Then the CRC bits $R = 110$, $D2^r = 101011000$ and the data sent $M = 10101110$ (see Figure 3).

• **Mathematics Theory for CRC**
Sender,
\[
\frac{D2^r}{G} = Q + \frac{R}{G},
\]

$R$ has at most $r$ bits.
Receiver,
\[
\frac{M}{G} = \frac{D2^r + R}{G} = Q + \frac{R}{G} + \frac{R}{G} = Q,
\]
no remainder if there is no transfer error.

CRC can not detect an error which changes $M$ into another multiple of $G$. However, this can happen with extremely low probability for an appropriately chosen $G$.

• **Error correction**
A simple approach for error correction is to retransmit the message if an error is detected during the transmission. Using error correction code is another approach for error correction. In this approach, redundant bits are added to the transmitted data in such a way that the receiver can finds which bit has an error. For this purpose, the number of redundant bits is much larger (usually proportional to the length of the data) than that used in CRC or CKS.

The *block code technique* is a well used error correction code. In this technique, assume a data block of $k$ bits is transmitted. Each of the $2^k$ sequences of $k$ bits is mapped to a unique
codeword of \( n \) (\( n > k \)) bits and the codeword is transmitted. If the codeword is designed in such a way that the Hamming distance between any two codewords is at least \( 2i + 1 \), then the scheme can correct any error up to \( i \) bits. For example, for \( k = 2 \) and \( n = 5 \) we can have the following mapping from the data to the codewords:

<table>
<thead>
<tr>
<th>Data block</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>00000</td>
</tr>
<tr>
<td>01</td>
<td>00111</td>
</tr>
<tr>
<td>10</td>
<td>11001</td>
</tr>
<tr>
<td>11</td>
<td>11110</td>
</tr>
</tbody>
</table>

It is easy to check that the above block code scheme can correct any error of a single bit.

- **Multiple Access Protocols**

  Links can be classified into point-to-point links and multi-point links. A point-to-point link is used to connect two nodes. A multi-point link usually uses a shared communication medium to connect more than two nodes. Such a link is also called a broadcast link because when a node transmits data to the link, the data are broadcasted on the link and every node sharing the link can hear the broadcast. Usually, two or more simultaneous transmissions on the link is not allowed because of the signal interference. Key issues in accessing multi-point links is to assign nodes to different links (channels) and coordinate nodes assigned to a same broadcast link to access the link exclusively. Examples of broadcast links include radios, Ethernet, wireless LAN, cell phone networks, etc. Media Access Control (MAC) protocols are used to coordinate nodes’ access to broadcast links.

- **MAC Approaches**

  The communication medium used by a link may have a large bandwidth. Channel partition is an effective approach to use the rich bandwidth of the medium: the bandwidth of the medium is partitioned into multiple sub-bands and each sub-band is used to support a channel. Well used channel partition techniques include: Time Division Multiplexing (TDM), Frequency Division Multiplexing (FDM) and Code Division Multiplexing (CDM, usually used in wireless networks).

  Another approach for media access control is to use all bandwidth of a link for one channel and nodes connected to the link share the channel. The access of nodes to the channel can be a random one or coordinated one. In the random access, a node can access the channel randomly but need to compete for accessing the channel. Collisions happen if multiple nodes access the channel at the same time. Collision detection and recovery mechanisms are required. In the coordinated access, there is a central device which coordinates nodes’ access to the channel.

- **TDM**

  TDM usually refers to synchronous time division multiplexing which partitions a time unit into multiple time slots and assigns time slots of each time unit to channels in round-robin, one slot supports one channel. The entire bandwidth of a link is used for the channel in the assigned time slot.
• **FDM**

FDM partitions the bandwidth of a link into non-overlapping sub-bands, each band supports one channel. In optical networks, FDM is called wavelength division multiplexing (WDM). In FDM, the entire time is used for a channel.

• **MAC Protocols for Random Access**

In a broadcast link with random access, the entire bandwidth of the link is used for one channel. Nodes can access the link randomly. If two or more nodes access the link at the same time, collisions happen. MAC protocols for random access specify how to detect collisions and recover from collisions. Examples of such protocols include ALOHA, slotted ALOHA, CSMA, CSMA/CD and CSMA/CA.

• **ALOHA**

ALOHA is a MAC protocol for packet radio networks, originally designed for the network to connect Hawaii Univ campuses in different islands. The protocol is very simple. Sender transmits a frame, set-up a timer and waits for an ACK of the sent frame. If the ACK is received before the time-out, then the transmission is OK; otherwise the sender retransmits the frame. A receiver sends an ACK for a correctly received frame and ignores a damaged one. ALOHA is simple but has a low efficiency if the average number of accesses to the link is high. To see this, assume that the transmission and propagation time for one frame is 1, the frame sent at time $t_0$ collides with other frames sent in the time period of $[t_0 - 1, t_0 + 1]$.

• **Slotted ALOHA**

To improve the efficiency (reduce the probability of collisions) of ALOHA, slotted ALOHA divides time into time slots, one slot for transmitting one frame, and synchronizes nodes such that a node can start transmission only at beginning of a slot. Collisions happen if there are two or more transmissions at the same time slot. Collision probability of slotted ALOHA is about half of that of ALOHA.

• **Carrier Sense Multiple Access (CSMA)**

To further reduce the collision probability, one idea is that before a transmission, a node first listens to the link to see if the link is used by another node. If the link is idle then the node transmits a frame, otherwise the node waits for the link becoming idle (see Figure 4). This can further reduce the probability of collisions but collisions can still happen due to signal propagation delay. For example, two nodes are separated by some distance, one node starts transmission but the other node may not hear the transmission immediately and starts a transmission as well. Then a collision happens.

• **CSMA/CD (Collision Detection)**

In some communication media, collision detection technique can be used to improve the data rate of the link. CSMA/CD is designed for this purpose. In CSMA/CD, CSMA is used to decide if a node can start a transmission. While transmitting, the node keeps listening to the link. If there are more than one transmissions, the node detects a collision by observing a higher signal level than a normal transmission. If a collision is detected, the node stop
transmission and go back to CSMA. Figure 5 explains CSMA/CD. Compared with the simple CSMA, CSMA/CD has a shorter collision time. It is easy to detect collision by abnormal high signal level in wired links but difficult for wireless links. So, CSMA/CD is not used in wireless networks, instead CSMA/CA (Collision Avoidance) is used in some wireless networks such as wireless LAN.

- **Coordinated MAC**

  There are two approaches for coordinated the media access control, one is called polling and the other is token passing. In polling, nodes connected to a link are partitioned to a master node and slave nodes. The master node invites slave hosts to transmit in turn. In token passing, nodes are usually connected to a ring network and there is a control token passed from one node to the next on the ring. A node must seize a token to transmit a frame and passes the token to the next node after the transmission. Examples of polling protocols include Bluetooth and IEEE802.11 for wireless LAN (a mix of CSMA/CA and polling). Examples
The maximum throughput of a coordinated LAN of $N$ nodes. Assume that each of the $N$ nodes has a transmission rate $R$ and can transmit at most $m$ bits in one turn. Let $d$ be the delay between the end of one node’s transmission and the begin of the next node’s transmission. The time for one round of $N$ transmissions is $N(m/R + d)$. The maximum throughput is

$$\frac{Nm}{N(m/R + d)}.$$

• **Summary of MAC Protocols**

Channel partition is to divide the bandwidth of a link into multiple sub-bands, each sub-band is used to support a channel. TDM, FDM and CDM (not explained) are commonly used channel partition techniques.

ALOHA, slotted ALOHA, CSMA, CSMA/CD and CSMA/CA (not explained) are media access control protocols which allow multiple nodes to have random access to a broadcast link. Collision detection is easier in wired links but difficult in wireless links. CSMA/CD is used in Ethernet (IEEE802.3) and CSMA/CA (Collision Avoidance) is used in wireless LAN (IEEE802.11)

Polling from a central node and token passing are techniques for coordinated access to broadcast links. Examples of such protocols include Bluetooth (polling), IEEE802.11 (mix of polling and random access), FDDI (token passing) and IBM Token Ring (token passing).

• **Link Layer Addressing**

At link layer, a physical (hardware) address is used to identify a network interface. Commonly used physical addresses include the 48bits Ethernet (MAC) address and the IEEE 64 bits address.

An MAC address has 48 bits and is used in most LANs. Each network interface card (NIC) has an unique MAC unicast address. A manufacturer buys a block of MAC addresses from IEEE to guarantee the global uniqueness of the addresses. An MAC address is independent but an IP address is dependent on networks. So when a node is moved from one network to another network, the MAC address of the node does not need to be changed but the IP address may have to be changed. This has been considered as a disadvantage of IPv4 address.

An MAC address is expressed in hexadecimal notation, e.g., the address of all 1 for broadcast is expressed by $FF − FF − FF − FF − FF − FF$.

• **MAC Address and ARP**

MAC addresses are used for frame transfer in physical networks. To deliver a datagram to a next node (router or destination host), a router needs to find the MAC address of the next node. Address resolution protocol (ARP) is used for this purpose. Assume that Host $A$ wants to find Host $B$’s MAC address with $B$’s IP address.

- $A$ broadcasts an ARP request, containing $B$’s IP address.
Figure 6: Deliver a datagram from a network to another network (1).

- Every node on the network receives the ARP request.
- B sends A an ARP reply with B’s MAC address.

To reduce the ARP requests, a node keeps the known IP-MAC address pair in a cache.

- **Routing from Network to Network**

Assume that Host A sends a datagram to Host B in two different networks connected by Router R.

- A knows B’s IP address.
- A gets the 1st hop router R’s IP address from A’s routing table.
- A gets the 1st hop router R’s MAC address from ARP.
- A creates the IP datagram with IP source A and destination B.
- A creates a link layer frame with R’s MAC address as destination, the frame contains A-to-B datagram as data.
- A sends frame to R and R gets the datagram from the frame.
- R forwards A-to-B datagram based on routing table of R to B.
- R creates a link-layer frame with B’s MAC address as destination, the frame contains A-to-B datagram, and sends the frame to B.

Figures 6-9 give an explanation on the above operations.

- **Ethernet**

Ethernet is a widely used LAN and has bandwidth from 10Mbps to 10Gbps. Ethernet provides connectionless and best effort data delivery service. It uses CSMA/CD for media access control. Ethernet has a bus topology, nodes are attached to a shared coaxial cable, or a star topology, nodes are connected to a hub (or switch).

- **Ethernet Frame**

The Ethernet frame format has a preamble field which is 7 bytes of 10101010 followed by one byte of 10101011, a destination address field of 6 bytes, a source address field of 6 bytes, a type field of 2 bytes (for length of data in the frame or higher layer protocol type), a data portion of 46-1500 bytes, and a CRC field of 4 bytes for error detection at receiver (see Figure 10).
Figure 7: Deliver a datagram from a network to another network (2).

Figure 8: Deliver a datagram from a network to another network (3).
Figure 9: Deliver a datagram from a network to another network (4).

Figure 10: Ethernet frame format.
• Ethernet CSMA/CD

Ethernet uses CSMA/CD for media control. The details of the CSMA/CD are as follows: The NIC of a node senses the network. If it is idle then the NIC transmits frame $F$, otherwise waits until the network becomes idle and then transmits $F$. After the transmission the NIC keeps listening to the network. If the NIC transmits the entire $F$ without detecting another transmission, then the transmission of $F$ is done. Otherwise, NIC (detects a collision) aborts $F$ and sends a jam signal to tell everyone a collision is detected. After the aborting, the NIC does not retransmit the frame immediately to avoid another collision (if both nodes involved in a collision do the retransmissions immediately, there is a high probability of another collision) but waits for a randomly decided time period. This waiting is called an exponential back off:

- After $m$th collision, NIC chooses a $k$ at random from $\{0, 1, \ldots, 2^m - 1\}$
- NIC wait for a time decided by $k$ and then go to Step 1 for retransmission of $F$

A collision example with CSMA/CD. Assume that Nodes $A$ and $B$ on an Ethernet $N$ and the propagation delay between $A$ and $B$ is 301 bits.

- At time $t = 0$, $A$ transmits a frame of 300 bits to $N$.
- At time $t = 150\text{bits}$, $B$ transmits a frame of 150 bits to $N$.
- At time $t = 150 + 150 = 300\text{bits}$, $A$ and $B$ finish transmissions.
- At time $t = 301\text{bits}$, $A$’s first bit arrives at $B$.
- At time $t = 301 + 150 = 451\text{bits}$, $B$’s first bit arrives at $A$.
- There is a collision, but not detected by CSMA/CD.
- Solution, data frame has a minimum size such that the transmission time for the frame is greater than the largest propagation delay between any two nodes in Ethernet

There are different Ethernet standards, commonly used are Ethernet V2 and IEEE802.3. Ethernet has different bandwidth from 10Mbps to 10Gbps and can use different physical layer media, cable, optical fiber, etc.

• Link Layer Switches

A modern Ethernet has a star topology, nodes are connected to a hub or switch. A hub is a physical layer repeater providing one-to-all connection. In an Ethernet connected by a hub:

- Bits come in from one link go out to all other links.
- A host can collide with another host connected to the hub.
- There is no buffer at hub, no CSMA/CD at hub but a node detects collisions.

An Ethernet can also be formed by connecting nodes to a switch which can provides one-to-one, one-to-many and one-to-all connections. In an Ethernet connected by a switch:

- Multiple simultaneous transmissions are allowed.
- Switch perform store and forward switching based on switch table.
Switch tables are created by self-learning algorithms.

**Forward Frames at Switch**

Each switch (say $S$) has a switch table $T$. Each entry of $T$ has a MAC address and an interface to reach the MAC address. Initially, $T$ is manually set-up or empty. When a frame from Host $A$ is received from Interface $I$ at $S$, $S$ creates an entry in $T$ with $A$’s address and interface $I$. In forwarding a frame $F$, $S$ checks destination MAC address $D_F$ of $F$ and take the following actions:

- IF $S$ finds an entry for $D_F$ and the interface to reach $D_F$ is the same as that from which $F$ is received, $S$ discards $F$.
- IF $S$ finds an entry for $D_F$ and the interface to reach $D_F$ is different from that from which $F$ is received, $S$ forwards $F$ to $I$.
- Otherwise, $S$ forwards $F$ to all interfaces except the one from which $F$ is received.

**Interconnecting Switches**

Switches can be connected together to make a larger network. In this case, a switch treats another switch connected to it as a node. Each switch uses self learning to create switch table.

**Switches vs Routers**

Both switches and routers perform store-and-forward packets. Routers work at network layer (layer-3 switches) and forward datagrams based on IP address and routing table. Routers use routing algorithms to create and maintain routing tables. Routers can connect networks of different technologies.

Switches work at data-link layer (layer-2 switches) and forward frames based on MAC address and switch table. Switches use self-learning algorithm to create and maintain switch tables. Switches are used to connect networks of a same technology.

**Point to Point Data Link Control**

A point-to-point link provides a dedicated connection between two nodes, and thus, media access control is not needed. Usually, explicit MAC addresses are not needed for the nodes as well. Point-to-point links are widely used in wide area networks (WAN) to provide connections of large distance. Point-to-Point Protocol (PPP) is a High Level Data Link Control (HDLC) protocol for transferring data on point-to-point links.

**Point-to-Point Protocol**

Below are major functions of PPP:

- Packet framing, encapsulate any network layer (not only IP) datagram in data link layer frame, one frame carry one datagram.
- Multiplexing, allow datagrams from multiple network layer protocols transferred on a same channel.
- Bit transparency, no constraint on bit pattern of data.
• Provide error detection but no error correction.
• Monitor connection, detect and report signal link failure to network layer.
• Address negotiation, each endpoint can learn/configure the network (e.g., IP) address of the other.

• PPP Data Frame
A PPP data frame has the following fields:
  • Flag (01111110) used as delimiter.
  • Address (11111111) not used, can be omitted.
  • Control (00000011) not used, can be omitted.
  • Protocol, one or two bytes, used to indicate upper layer protocol types.
  • Info, data from upper layer.
  • Check, two or four bytes cyclic redundancy check for error detection.

• Byte Stuffing
PPP has no constraint on bit pattern in data and uses 01111110 as delimiter (start and end) of a frame. This will have the following problem: when a receiver see a 01111110 pattern, is it a delimiter or a part of data? One approach to solve this problem is that
  • Sender, add extra byte 01111110 after each 01111110 data byte.
  • Receiver, take two consecutive 01111110 bytes as data and single 01111110 byte as delimiter.

• Virtual LAN and Link Virtualization
LANs using the same addressing scheme can be interconnected by layer 2 switches to make a larger network. The interconnection can be configured as one virtual LAN (VLAN). Hosts in a VLAN communicate with each other as if they are in a same physical LAN.

A connection between two nodes in the Internet may be viewed as a virtual link: the connection may consists of multiple real networks of different technologies but there is a single standard for the format of data packets and forwarding the data packets over the connection, and thus the routers who forward (receive) data packets to (from) this connection treat the connection as a single link. Multiple Protocol Label Switching (MPLS) is an example of link virtualization. MPLS provides packet-switched virtual circuit connection service. Two nodes connected by an MPLS circuit can be viewed connected by a virtual link. MPLS appends an MPLS header to an IP datagram for the switching. The MPLS header has a label field which contains the label for a flow.

• Synthesis
Major protocols used for transferring data from one application process to another one in the Internet have been introduced. Next, we will see a global picture on how data are transferred by an example. In this example, we assume that a user attaches a lap-top to a campus network
and accesses a website, say www.google.com (see Figure 11). We will review and understand protocols (at all layers) involved in this web application.

Connect to network (see Figure 12).

- Laptop sends a DHCP request to get IP addresses for itself, the first-hop router and DNS server.
- The DHCP request is carried by a UDP segment, the UDP segment is carried by an IP datagram, the IP datagram is carried by an Ethernet frame (IEEE802.3).
- The Ethernet frame is broadcasted on LAN and received by a DHCP server.
- The DHCP server sends laptop a DHCP reply to provide the info requested.

Get IP address of the website (see Figures 13 and 14).

- Laptop first uses DNS to get IP address of www.google.com.
- The DNS query is carried by a UDP fragment, UDP is carried by an IP datagram, the IP datagram is carried by an Ethernet frame, and the Ethernet frame is sent to a DNS server via first-hop router.
- ARP is used to get MAC address of 1st-hop router.
- The DNS query is sent to 1st-hop router.
- The DNS query is sent to the DNS server via 1st-hop router of the lap-top, IP forwards the datagram carried the query based on routing tables created by routing protocols such as RIP, OSPF, BGP.
- The DNS server sends the laptop a DNS reply containing IP address of www.google.com.

Connect to www.google.com (see Figure 15).

- Laptop opens TCP connection to the server.
Figure 12: An example on the Internet communication, connecting to a network.

Figure 13: An example on the Internet communication, get destination IP address (1).
Figure 14: An example on the Internet communication, get destination IP address (2).

- Server accepts the request and creates a TCP connection.
- HTTP request is sent to the TCP socket and routed to the server.
- Server sends the laptop data.

The web page is displayed.
Figure 15: An example on the Internet communication, connecting to destination.