ECE 435 – Network Engineering Lecture 18

Vince Weaver

http://web.eece.maine.edu/~vweaver

vincent.weaver@maine.edu

7 November 2017

Announcements

- HW#8 was posted
- Project topics
- Lab next week
- News: someone managed to mess up BGP and blackhole a bunch of the internet on Monday (Level3 messed up route, things like Comcast down)

Questions From Last Time

- Broadcast MAC addresses? Bit vs ff:ff:ff:ff:ff:ff
  What about the “Group Address” multicast bit? Shown as the high bit? Remember that Ethernet sends LSB first. So it’s addresses like: 01-80-C2-00-00-00

- Why IEEE standards start with 802
  Next available? Also co-incidentally first meeting was Feb. 1980
  For example, IEEE floating point is IEEE 754 but first meeting was not April 1975

Ethernet Transmission

- MAC puts data into frame
- In half-duplex CSMDA/CD senses carrier. Waits until channel clear
- Wait for an inter-frame-gap (IFG) 96 bit times. Allows time for receiver to finish processing
- Start transmitting frame
- In half-duplex, transmitter should check for collision.

  Co-ax, higher voltage than normal

  For twisted pair, noticing signal on the receive while transmitting

- If no collision, then done
- If collision detected, a *jam* signal is sent for 32-bits to ensure everyone knows. Pattern is unspecified (can continue w data, or send alternating 1s and 0s)
- Abort the transmission
- Try 16 times. If can't, give up
- Exponential backoff. Randomly choose time from 0 to $2^k - 1$ where k is number of tries (capping at 10). Time slot is 512 bits for 10/100, 4096 for 1Gbs
- Wait the backoff time then retry

Ethernet Receiving

- Physical layer receives it, recording bits until signal done. Truncated to nearest byte.
- If too short (less than 512 bits) treated as collision
- If destination is not the receiver, drop it
- If frame too long, dropped and error recorded
- If incorrect FCS, dropped and error recorded
- If frame not an integer number of octets dropped and error recorded
- If everything OK, de-capsulated and passed up

- Frame passed up (minus preamble, SFD, and often crc)
- Promiscuous mode?

Maximum Frame Rate

- 7+1 byte preamble 64-byte frame, IFG of 12 bytes between transmissions. equals 672 bits. In 100Mbps system 148,800 frames/second

Full Duplex MAC

- Early Ethernet was coaxial in a bus
- Twisted pair has replaced this, usually in a hub/or switch star topology
- 10BASE-T and 100BASE-TX pair for transmit or receive
- inefficient. Since point to point, why do you need arbitration?
- Full-duplex introduced in 1997. Must be able to

transmit/receive w/o interference, and be point to point.

- Full duplex effectively doubles how much bandwidth between. Also it lifts the distance limit imposed by collision detection

Ethernet Flow Control

- Flow control is optional
- In half duplex a receiver can transmit a “false carrier” of 1010..10 until it can take more.
- Congested receiver can also force a collision, causing a backoff and resend. Sometimes called force collision
- Above schemes called “back pressure”
- For full duplex can send a PAUSE frame that specifies how much time to wait.

Fast Ethernet (100MB)

- 10MB not fast enough! What can we do?
  - FDDI and Fibrechannel (fast optic-ring), too expensive
  - Can we just multiply all speeds by 10? Or else come up with some completely new better thing?
  - IEEE decided to just keep everything same, just faster
  - The other group went off and made 802.12 100BaseVG (which failed)
- 802.3u 1995
- 10BASE-TX most common

- Bit time from 100nsec to 10nsec
- Uses twisted pair/switches, no coax
  - To use cat3 100BASE-T4 wiring needed 4 twisted pair and complex encoding, no Manchester, ternary
  - To use cat5 wiring 100BASE-TX. Two twisted pair, one to hub, one from.
- Often split between MAC (media access controller) and PHY (physical interface). Card configures the PHY via the MII (media independent interface)

  Goal was you could have same card but interchangeable PHY (twisted pair, fiber, etc). 4bit bus

Interface requires 18 signals, only two can be shared if multiple PHY

So RMII (reduced) was designed. Clock doubled, only 2-bit bus. Fewer signal wires.

- 100BASE-TX:
  - 2 pairs. One pair 100MB each direction, full duplex, 100m distance
  - Raw bits (4 bits wide at 25MHz at MII) go through 4B/5B encoding clocked at 125MHz. (DC equalization and spectrum shaping)
  - Then NRZI encoding (transition at clock if 1, none if

0).

- TX then goes through MLT-3 encoding (-1,0,+1,0. Transition means 1, no transition means 0) 31.25MHz, much like FDDI

Router vs Hub vs switch

- Hub all frames are broadcast to all others
  Bandwidth is shared (only say 100MB for all)
- Switch – direct connection, no broadcast. Has to be intelligent. Each point to point connection full bandwidth.

  no collisions. Internally either own network to handle collisions, or else just buffer RAM that can hold onto frames until the coast is clear.
- Multi-speed hubs?

When 10/100MB first came out, cheap hubs could only run at 10MB or 100MB. But switches *really* expensive. They had a compromise 10/100MB hub that internally had a hub for both then a mini-switch to bridge the gap.

- Direct Ethernet connection. Need a special loopback cable?

  Modern cards can detect direct connect and swap the wires for you
- Router will move frames from one network to another
- Lights. How many ports? Uplink ports?
- Power over ethernet

- Method B: In 10/100 Base T, only of the 4 pairs in Cat5 used. So send voltage down spare pairs
- Method A: send DC voltage down with the signals floating on top
- Original 44 VDC, 15.4W
- POE+ 25W
- Need special switch to send power, and device on other end has to support it.

Gigabit Ethernet

- Two task forces working in 1998/1999
- 802.3z 1998 (fiber), 802.3ab 1999 (copper)
- Could still use hub, problem was the CSMA/CD restriction.
  - ○ About 200m for 100Mbps.
  - ○ For Gb would have been 20m which is not very far.
  - ○ Carrier extension: hardware transparently pads frames to 512 bytes

    Wasteful, 512 bytes to send 64 bytes of data

  - Frame bursting: allow sender to sends sequence of multiple frames grouped together
- Better solution is just use full duplex
- 1000Base-SX (fiber)/LX (fiber)/CX (shielded)/T (cat 5), more
- Fiber
  - No Manchester, 8B/10B encoding. chosen so no more than four identical bits in row, no more than six 0s or six 1s
need transitions to keep in sync
try to balance 0s and 1s? keep DC component low so

can pass through transformers?

- 1000BASE-T
  - 5 voltage levels, 00, 01, 10, 11, or control. So 8 bits per clock cycle per pair, 4 pairs running at 125MHz, so 1GBps
  - simultaneous transmission in both directions with adaptive equalization (using DSPs), 5-level pulse-level modulation (PAM-5) [technically 100BASE-TX is PAM-3]. Diagram? looks sort of like a sine wave as cycle through the voltages.
  - four-dimensional trellis coded modulation (TCM) 6dB

coding gain across the four pairs

  - ○ Autonegotiation of speed. Only uses two pairs for this, can be trouble if pairs missing.

- Fast enough that computers at time had trouble saturating such a connection
- Jumbo Frames? 9000 byte?

Even Faster Ethernet

http://www.theregister.co.uk/2017/02/06/decoding_25gb_ethernet_and_beyond/

- Misquote: Not sure what the network will be like in 30 years, but they will call it Ethernet.

- 2.5Gb: 802.3bz (Sep 2016?)
  Like 10Gb but slower. Can't run 10Gb over Cat5e
  Power over Ethernet (for using on wireless access points)
  Power with signal overlaid on top.
  2.5Gb on Cat 5e, 5Gb on Cat6

- 10Gb: 802.3ae-2002. Full duplex, switches only
  Need Cat6a or Cat7 for links up to 100m
  Expensive. Lots of kind. 10GBASE-T, 802.3an-2006
  100m over cat6a, 55m Cat6
  additional encoding overhead, higher latency
  Tomlinson-Harashin precoding (THP), PAM15 in two-dimensional checkerboard DSQ128 at 800Msymbol/s

- 25Gb, 802.3by. 25GBASE-T, 50GBASE-T. Available, if copper only a few meters

- 40GB, 100GB. 802.3ba-2010, 802.3bg-2011, 802.3bj-

2014, 802.3bm-2015
40GBASE-T twisted pair 40GBit/s 30m. QFSP+ connectors, like infiniband

- Terabit? still under discussion

Autonegotiation

- How figure out line speed and duplex

What does your machine have

- skylake machine:

```
[   18.240021] e1000e: eth0 NIC Link is Up 1000
```

- Raspberry Pi:

```
[   77.110505] smsc95xx 1-1.1:1.0 eth0: link up
```

- Haswell machine:

```
[    3.907651] tg3 0000:03:00.0 eth0: Tigon3 [p
[    3.919115] tg3 0000:03:00.0 eth0: attached
[    3.929838] tg3 0000:03:00.0 eth0: RXcsums[1
[    3.938174] tg3 0000:03:00.0 eth0: dma_rwctr
[   13.758613] IPv6: ADDRCONF(NETDEV_UP): eth0:
[   15.404905] tg3 0000:03:00.0 eth0: Link is u
[   15.411479] tg3 0000:03:00.0 eth0: Flow cont
```

Linux OS Support

- When frame comes in, interrupt comes in
- Allocates `sk_buff` copies in
- Old: `net_if_rx()` interrupt, `net_rx_action()` interrupt/polling
- `net_if_receive_skb()`
- passes it to proper net level (`ip_recv()`, `ip_ipv6_recv()`, `arp_recv()`)

- for send

- `net_tx_action()`
  `dev_queue_xmit()` and then deallocate `sk_buff`

- `qdisc_run()` selects next frame to transmit and calls `dequeue_skb()`

Why might you want to split up LANs

- Bandwidth concerns
- Different groups, privacy/security
- Equipment costs
- Distance
- Reliability (equipment failure)

Bridging

- How do you connect together multiple groups of machines into one big LAN?
- An interconnection at the link layer is called a MAC bridge, or bridge. Also a Layer-2 switch
- IEEE 802.1D
- Transparent bridge, as users are not aware of them
- Bridge acts in promiscuous mode (receives every frame

on the LAN) so it can find ones that need to forward on across the bridge

- How does bridge learn the MAC addresses? self-learning. It watches for frames coming in and their source address. Puts in table. How does it learn where destination is? It broadcasts to all. Once the destination also sends a frame (so its source is known) then the switch updates its table and no longer broadcasts.

- How do you handle machines that are moved? Aging mechanism. If not heard from for a while, expire the table

- Multicast or Broadcast, can follow GMRP or GARP to limit how far it is broadcast

Bridge vs Switch

- Before 1991 a switch was a bridge (in the standard)
- In 1991 Kalpana made a “switch” and differentiated it by cut-through instead of store and forward
- Store and forward – whole frame received before resent larger latency, no problem with broadcast, can check FCS
- cut-through – can start transmitting before receiving completely (destination MAC at beginning). Slightly

better latency, broadcast not possible, too late to check FCS

- These day most are store and forward
- Differences
  - repeater – purely electronic, resends voltages (original Ethernet allowed four)
  - hubs – frames coming in one port sent to all others creates a collision domain
  - bridge – connects two or more LAs. Each line own collision domain

can maybe bridge different types of networks (ethernet/token, wired/wireless)

- switch –
- router – actually strips off headers and looks at packets

Spanning Tree Protocol

- Invented by Radia Perlman at DEC

- Can have problems if cause a loop in the topology. Frames can circulate loop forever

- 802.1D
  - Each switch and port assigned an ID with priority
  - Each link assigned a cost, inversely proportional to link speed

- The lowest ID gets to act as root (there is a protocol on how to elect the root)
- Each LAN connected to upstream port in active topology, called the dedicated port. Receives from root port
- Config info comes from root as bridge protocol data unit (BPDU) on reserved multicast address 01:80:c2:00:00:00
- Switch may configure itself based on BPDU.

Bridging 802.11 to 802.3

- Need to strip off one header, put new one on
- Need to put fields in as needed, recalc checksum, etc
- What if bridging faster net to slower one
- What if maximum frame size different on different LANs? Can't always fragment
- What if one has encryption and one doesn't
- What of quality of service?

VLAN

- How to switch machines between networks? Request? Someone in wiring closet?
- Physical LAN
- What if want to partition a switch so some nodes are on one and one on another (virtual LANs)
- IEEE 802.1Q
- can have priority

- link aggregation, combine two links for higher bandwidth

- why split up?
  Security (someone in promisc mode not see everything)
  Load – two groups, one not happy if other group takes up all bandwidth
  Broadcasting – when asks for a connection, broadcasts to all
  broadcast storms – entire LAN brought down with all machines broadcasting

- how to bridge VLANs? special VLAN field in Ethernet frame

priority, CDI (makes connectionless interface have some manner of connection)