A Deep Dive into the Ethernet
Outline

• Cabling

• Modulation schemes: 10MBit - 100GBit

• Autonegotiation

• Energy-efficient Ethernet

• Power over Ethernet

• Flow control
IEEE 802.3

- 2010: 40Gb → 2010: 100Gb → 2017: 200/400 Gb → Terabit

Future developments: Termabit (COMING SOON) > 2020
Cabling

- Twisted-pair
- Fiber
- Twinaxial
Twisted-pair copper

<table>
<thead>
<tr>
<th>Category</th>
<th>Max</th>
<th>Bandwidth</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat 3</td>
<td>10Mb</td>
<td>16MHz</td>
<td>100m</td>
</tr>
<tr>
<td>Cat5/5e</td>
<td>1Gb</td>
<td>100MHz</td>
<td>100m</td>
</tr>
<tr>
<td>Cat6</td>
<td>10Gb</td>
<td>250MHz</td>
<td>55m</td>
</tr>
<tr>
<td>Cat6a</td>
<td>10Gb</td>
<td>500MHz</td>
<td>100m</td>
</tr>
<tr>
<td>Cat8</td>
<td>40Gb</td>
<td>2GHz</td>
<td>36m</td>
</tr>
</tbody>
</table>

No plans for twisted-pair support at 100Gb
100Gb expected to use 20GBaud → Use fiber
<table>
<thead>
<tr>
<th>Data rate</th>
<th>Bandwidth</th>
<th>Cable</th>
</tr>
</thead>
<tbody>
<tr>
<td>10MBit</td>
<td>10MHz</td>
<td>16MHz</td>
</tr>
<tr>
<td>100MBit</td>
<td>31.25MHz</td>
<td>100MHz</td>
</tr>
<tr>
<td>1GBit</td>
<td>62.5MHz</td>
<td>100MHz</td>
</tr>
<tr>
<td>2.5GBit</td>
<td>100MHz</td>
<td>100MHz</td>
</tr>
<tr>
<td>5GBit</td>
<td>200MHz</td>
<td>250MHz</td>
</tr>
<tr>
<td>10GBit</td>
<td>400MHz</td>
<td>500MHz</td>
</tr>
</tbody>
</table>
10/100Mb uses 2 pairs

<table>
<thead>
<tr>
<th>Color Combination</th>
<th>+TX</th>
<th>-TX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green + white</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td>-TX</td>
<td></td>
</tr>
<tr>
<td>White + orange</td>
<td></td>
<td>+RX</td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td>-RX</td>
</tr>
</tbody>
</table>

100MBit onwards use full-duplex

Gigabit onwards uses 4 pairs

<table>
<thead>
<tr>
<th>Color Combination</th>
<th>+A</th>
<th>-A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green + white</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Green</td>
<td>-A</td>
<td></td>
</tr>
<tr>
<td>Orange + white</td>
<td>+B</td>
<td></td>
</tr>
<tr>
<td>Orange</td>
<td></td>
<td>-B</td>
</tr>
<tr>
<td>Blue + white</td>
<td>+C</td>
<td></td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>-C</td>
</tr>
<tr>
<td>Brown + white</td>
<td>+D</td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td></td>
<td>-D</td>
</tr>
</tbody>
</table>
Fiber

- Isolated from external electrical noise
- Longer range
- 40Gb - 12 fibers
- 100Gb - 24 fibers
- LED at 850nm for 100MBit
- Laser for 1GBit and above

<table>
<thead>
<tr>
<th>Data rate</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>100MBit</td>
<td>2km</td>
</tr>
<tr>
<td>1GBit</td>
<td>1km</td>
</tr>
<tr>
<td>10GBit</td>
<td>400m</td>
</tr>
<tr>
<td>40/100GBit</td>
<td>150m</td>
</tr>
</tbody>
</table>
Shorter range due to modal dispersion
Signal spreads out in time

**Multimode fiber**
- Transmitted pulses
- Cladding
- Core
- Received pulses
- Jitter

850 - 1300nm
50μm core
125μm cladding

**Single-mode fiber**
- Transmitted pulses
- Cladding
- Core
- Received pulses

1310 - 1550nm
8-10μm core
125μm cladding

More $$$
Sheet of paper is 100μm thick
10MBit

- 10MHz clock
- 1 bit every 100 nanoseconds
- Manchester phase encoding
- Polarity reversal
100MBit (Fast Ethernet)

4 bits → 4B/5B → Scramble → MLT-3

- At least 2 bit transitions per block for clock recovery
- Spread across multiple frequencies to reduce interference
- Reduce bandwidth

16 hex symbols
16 control symbols
0: no transition, 1: transition

1 \rightarrow 0 \rightarrow -1 \rightarrow 0 \rightarrow 1
4B/5B creates 125MBaud signal
MLT-3 reduces fundamental frequency to 31.25MHz
Gigabit Ethernet

- 125MBaud per twisted pair.
- 500MBaud in total (2 bits per symbol)
- 4D-PAM5 (5 voltage levels, spread across 4 channels)
- Trellis modulation for parity bit
Spread spectrum
Generate parity bit
Generate symbols

Channels are interdependent

$5^4 = 625$ symbols in total

Encode 9 bit word: 8 data bits + 1 parity bit. $2^9 = 512$ possible bitstrings

Remaining 113 symbols used for control or discarded
Parity bit generation

Trellis/convolutional encoder

D flip-flops

17
2D-PAM5

Attenuation - points move closer

Noise - points blur

Subsequent symbols must have a Euclidean distance of at least X
Only specific sequences are allowed.

Decoder chooses most likely sequences using Hamming distance as distance metric
Symbols in a subset have distance 2

In 4D-PAM5 we have 16 subsets. (XXXX, XXXY, ..., YYYY)

Symbols have distance 4
Group 16 subsets into 8 sub-lattices:
Each pair of subsets are complements of each other

Distance between any two symbols within sub lattice $\geq 4$

9 bit word $\rightarrow$ 1 of 512 symbols $\rightarrow$ voltage levels
10GBit

Each twisted pair takes care of 2.5Gbps both ways

About 400MHz bandwidth

32 bits 32 bits → 64B/65B → LDPC → 2PAM16 → DSQ-128

Scramble

BER: $10^{-12}$
2048 bits = 1723 data + 325 check

$16^2 = 256$ symbols
800MBaud

Prune to get 128 symbols
7 bits per symbol
**LDPC example**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>1</th>
<th>0</th>
<th>Received bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Checked bits</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Corrected bits</td>
</tr>
</tbody>
</table>

Variable nodes (2048)

Check nodes (325)

1. $v$ nodes send bits to $c$ nodes
2. $c$ nodes send a $v$ node the $\oplus$ of other received values
   e.g. $c_0$ sends $v_0$ the value $v_1 \oplus v_4$
3. $v$ nodes do a majority vote
40GBit

- 64/66 bit encoder from 10GBit
- 10.3125GBaud per lane
- Round robin across 4 PCS (physical coding sublayer)
- PCS does coding and negotiation
- Lanes can be optical or twisted cable
100GBit

- 20 PCS lanes
- Multiplexers
- 10 electrical lines
- Framing, octet sync, de/scramble
- Convert to bits to signal
- 25GBit/s per lane
- 4 optical lanes (4 different wavelengths)
- Single pair of optical cables
Datacenters

- Facebook (2014)
  - 4x40Gbps

Facebook Dreams of Terabit Ethernet
By Michael Feldman
February 3, 2010

4 fabric switches

48 top of rack switches (TORs)
Autonegotiation

- Negotiates highest speed, duplex + other capabilities
- Only requires 2 twisted pairs
- Occurs whenever link is re-connected
- Time
  - 2-3 seconds for 10/100MBit
  - 5-6 seconds for Gigabit onwards
- Doesn’t check if cable is correct
  - Won’t notice if you use Cat3 with 1GBASE
- Auto-downgrades if a speed can’t be established
Autonegotiation

- Sends a single Fast Link Pulse (similar to heartbeat)
  - 33 bursts, 10ns long each
  - 17 odd pulses for clock
  - 16 even pulses for data:
    - presence/absence => 1/0
  - 16-bit link code words
Base page message

LAN type: Ethernet
Speed, duplex, pause ability
Remote fault bit set if receiver senses failure
Set to 1 after receiving three messages

Next page bit to say if there's another message containing additional vendor info
Energy efficient ethernet (100MB-10GB)

- By default Ethernet sends IDLE symbols as a heartbeat.
- Channel is always occupied
- Low Power Idle mode
  - If there are no frames: shut down
- Savings - $450 million across the US per year
  - Idle link: 91% for 1G and 74% for 100MBit
  - 191 links with normal bursty traffic: 15%
Sleep
Send LPI for *Time To Sleep*
If received LPI
Go to sleep

Refresh
Check for disconnection

Wake
Send IDLE for *Time to Wake*
Power Over Ethernet (10MB - 1GB)

- Alt-A (common-mode): power and data share same wire
- Alt-B (spare-pair): power and data on separate wires
- Up to 25.5W per device
Power Over Ethernet

- Detection:
  - Send 2.7-10.1V voltage check for 25kΩ resistance

- Classification
  - Send 15.5-20.5 voltage
  - Measure current draw
  - Dynamic adjustment. 10 second wait
  - Checks if link is connected
  - Regulates voltage and current

<table>
<thead>
<tr>
<th>I (mA)</th>
<th>Pout (W)</th>
<th>Precv (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>15.4</td>
<td>0.44-12.95</td>
</tr>
<tr>
<td>9-12</td>
<td>4</td>
<td>0.44-3.84</td>
</tr>
<tr>
<td>17-20</td>
<td>7</td>
<td>3.84-6.49</td>
</tr>
<tr>
<td>26-30</td>
<td>15.4</td>
<td>6.49-12.95</td>
</tr>
<tr>
<td>36-44</td>
<td>36</td>
<td>12.95-25.5</td>
</tr>
</tbody>
</table>
Flow Control

- Pause frame: tell sender to stop transmitting for X time
  - Used to handle NICs with small buffers
  - Send pause frame to 01-80-C2-00-00-01
A Deep Dive into the Ethernet
Ethernet Packet

Frame 64-1522

7x 10101010 10101011
Preamble 7 Start frame 1

MAC dst 6 MAC src 6
802.1Q 4 Type/length 2
Payload 46-1500
CRC 4

Interpacket gap 12

Packet 72-1530

Header 12
Q-tag: VLAN priority indicator
VLAN: LAN within a switch
Type={IPv4, IPv6, ARP, 802.1Q}