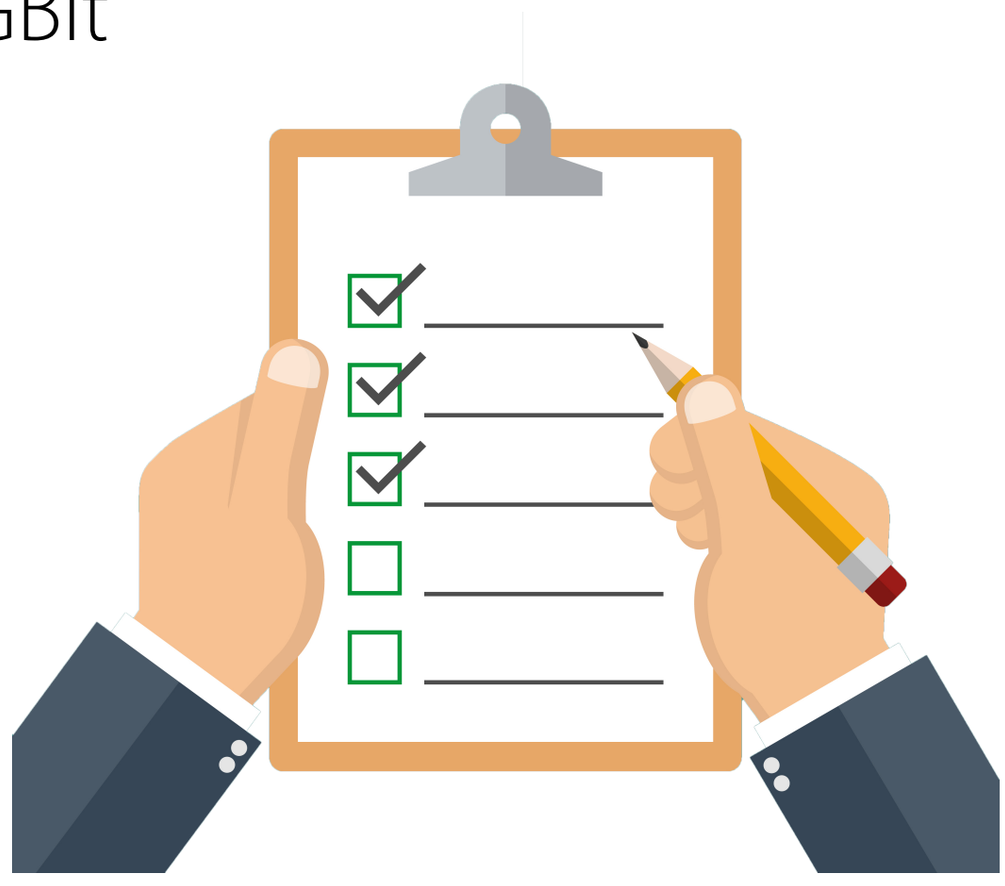
A Deep Dive into the Ethernet

Outline

- Cabling
- Modulation schemes: 10MBit - 100GBit
- Autonegotiation
- Energy-efficient Ethernet
- Power over Ethernet
- Flow control

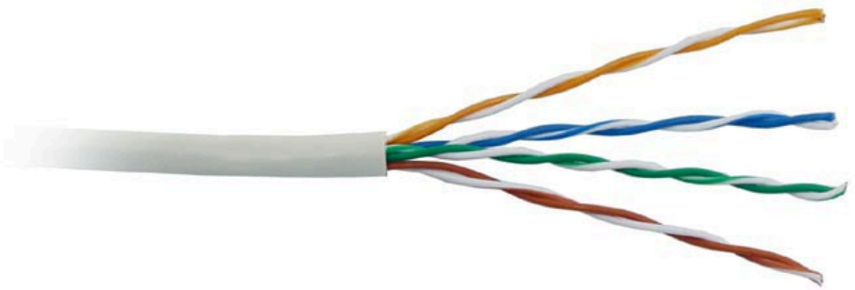


IEEE802.3

1980	1995	1999	2006
10Mb →	100Mb →	Gigabit →	10Gb

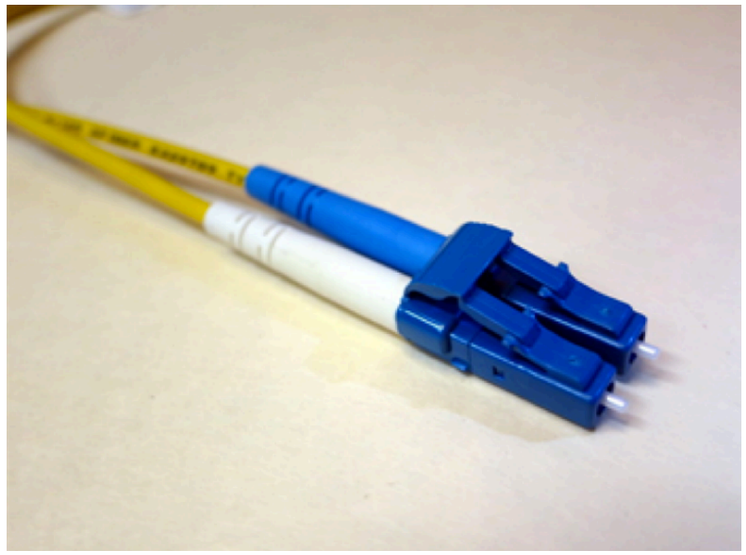
2010	2010	2017	> 2020
40Gb →	100Gb →	200/400 Gb →	Terabit

Cabling



✓

Twisted-pair



✓

Fiber

✗

Twinaxial

Twisted-pair copper

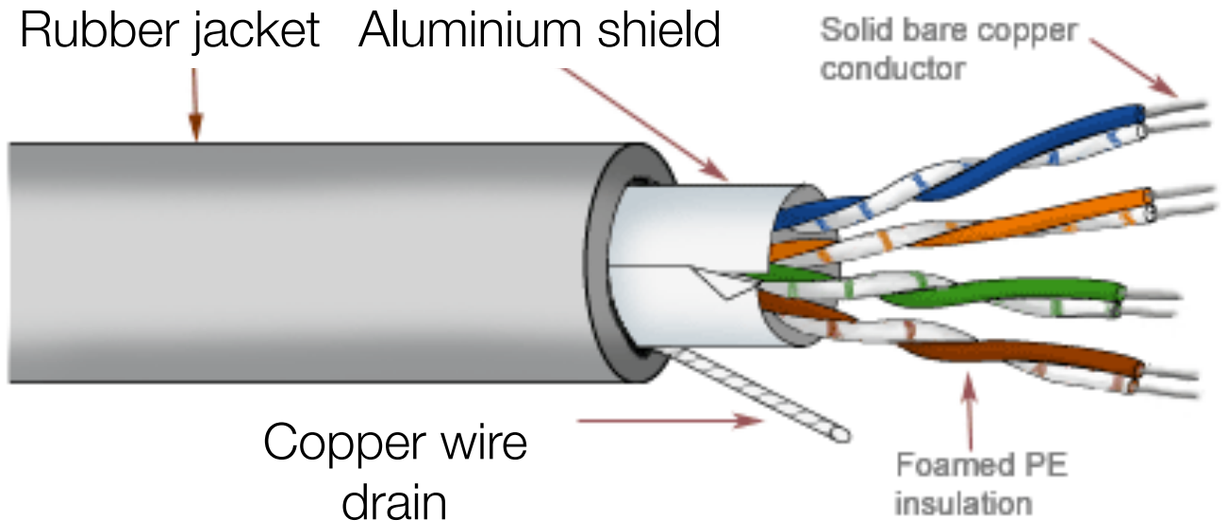
Category	Max	Bandwidth	Length
Cat 3	10Mb	16MHz	100m
Cat5/5e	1Gb	100MHz	100m
Cat6	10Gb	250MHz	55m
Cat6a	10Gb	500MHz	100m
Cat8	40Gb	2GHz	36m

2017-2018

No plans for twisted-pair support at 100Gb

100Gb expected to use 20GBaud→ Use fiber

Data rate	Bandwidth	Cable	
10MBit	10MHz	16MHz	
100MBit	31.25MHz	100MHz	
1GBit	62.5MHz	100MHz	
2.5GBit	100MHz	100MHz	
5GBit	200MHz	250MHz	
10GBit	400MHz	500MHz	



10/100Mb uses 2 pairs

Green + white	+TX
Green	-TX
White + orange	+RX
Orange	-RX

100MBit onwards use full-duplex

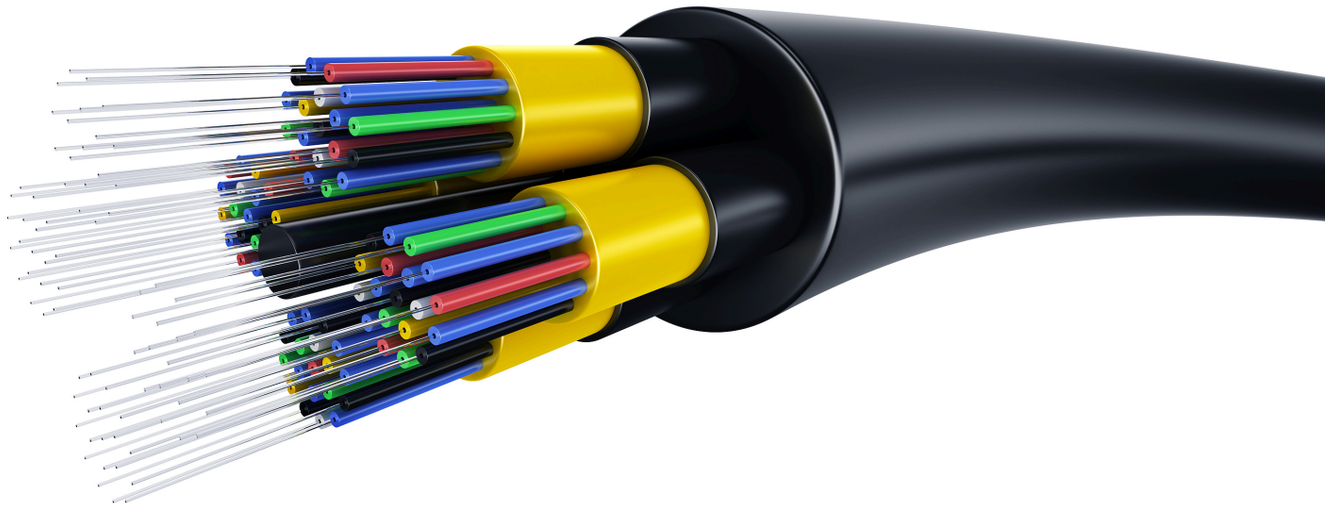
Gigabit onwards uses 4 pairs

Green + white	+A
Green	-A
Orange + white	+B
Orange	-B
Blue + white	+C
Blue	-C
Brown + white	+D
Brown	-D

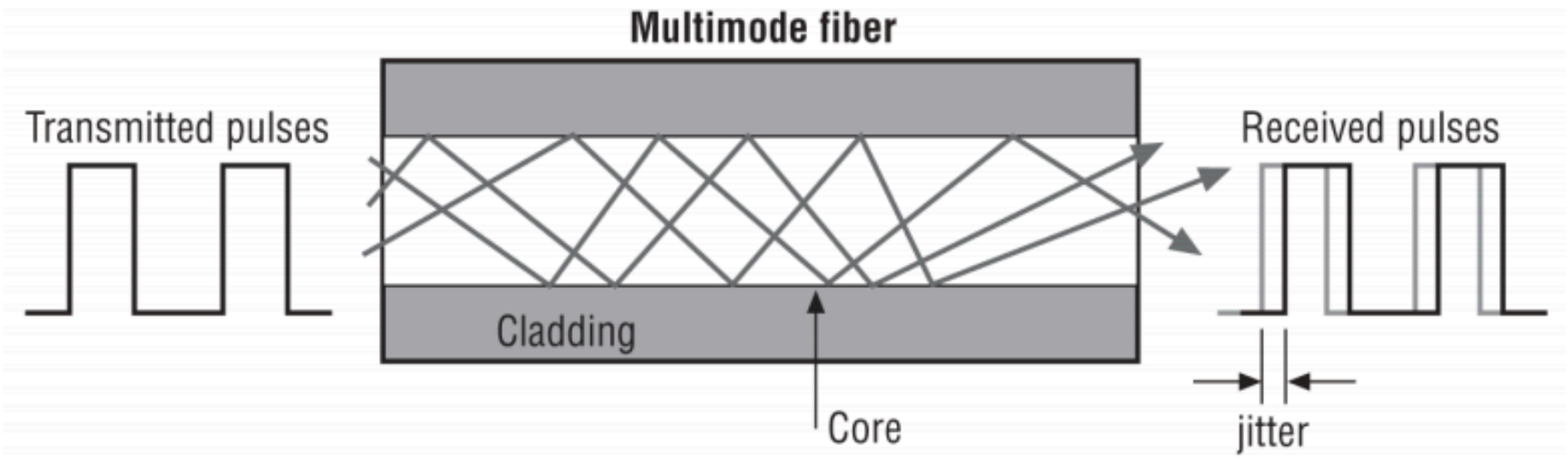
Fiber

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

Data rate	Range
100MBit	2km
1GBit	1km
10GBit	400m
40/100GBit	150m



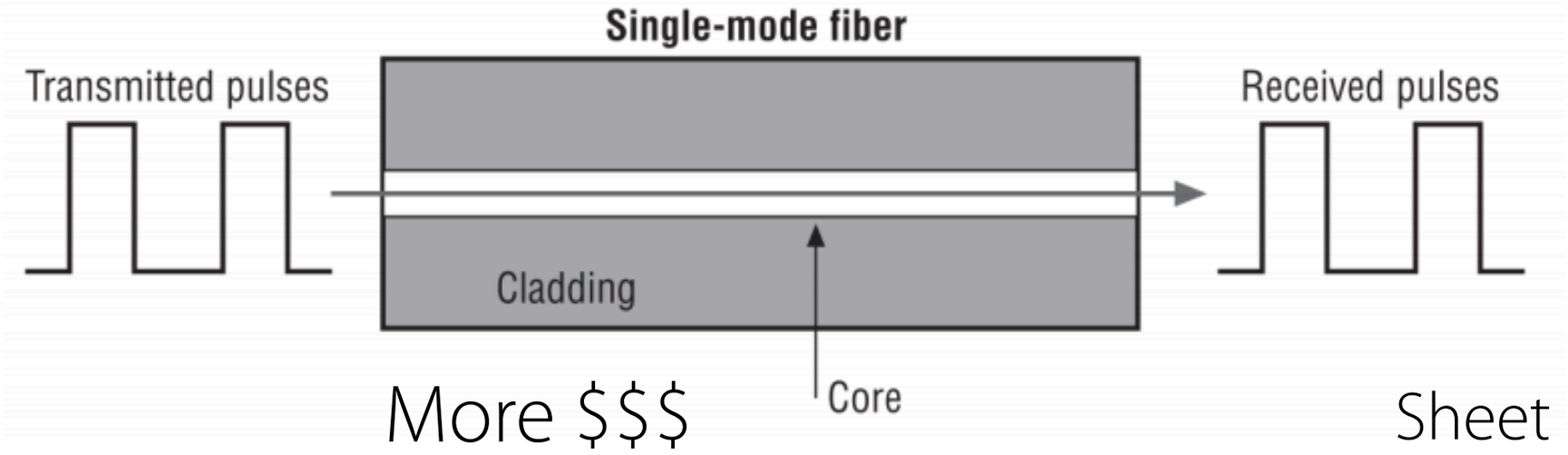
Shorter range due to modal dispersion
Signal spreads out in time



850 - 1300nm

50µm core

125µm cladding



1310 - 1550nm

8-10µm core

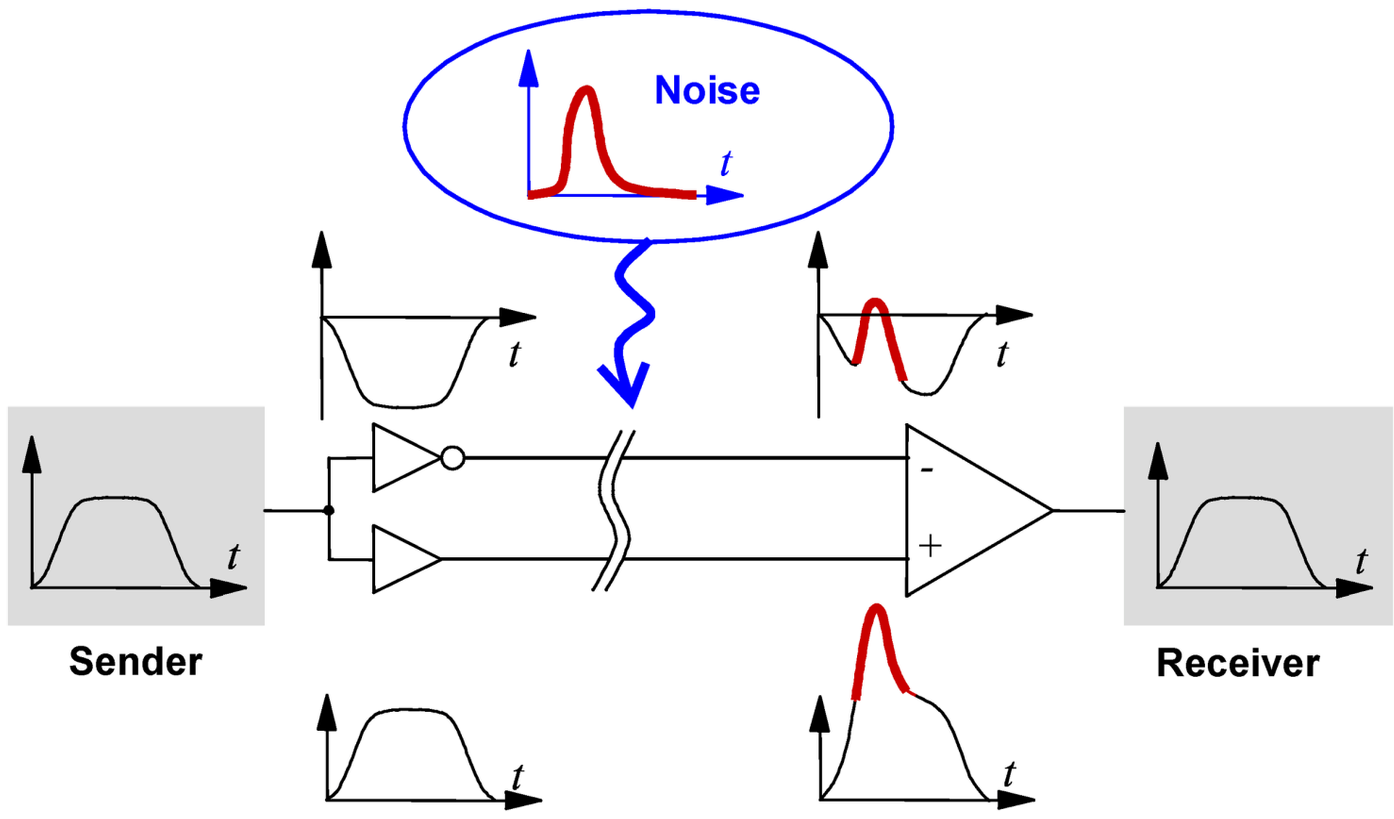
125µm cladding

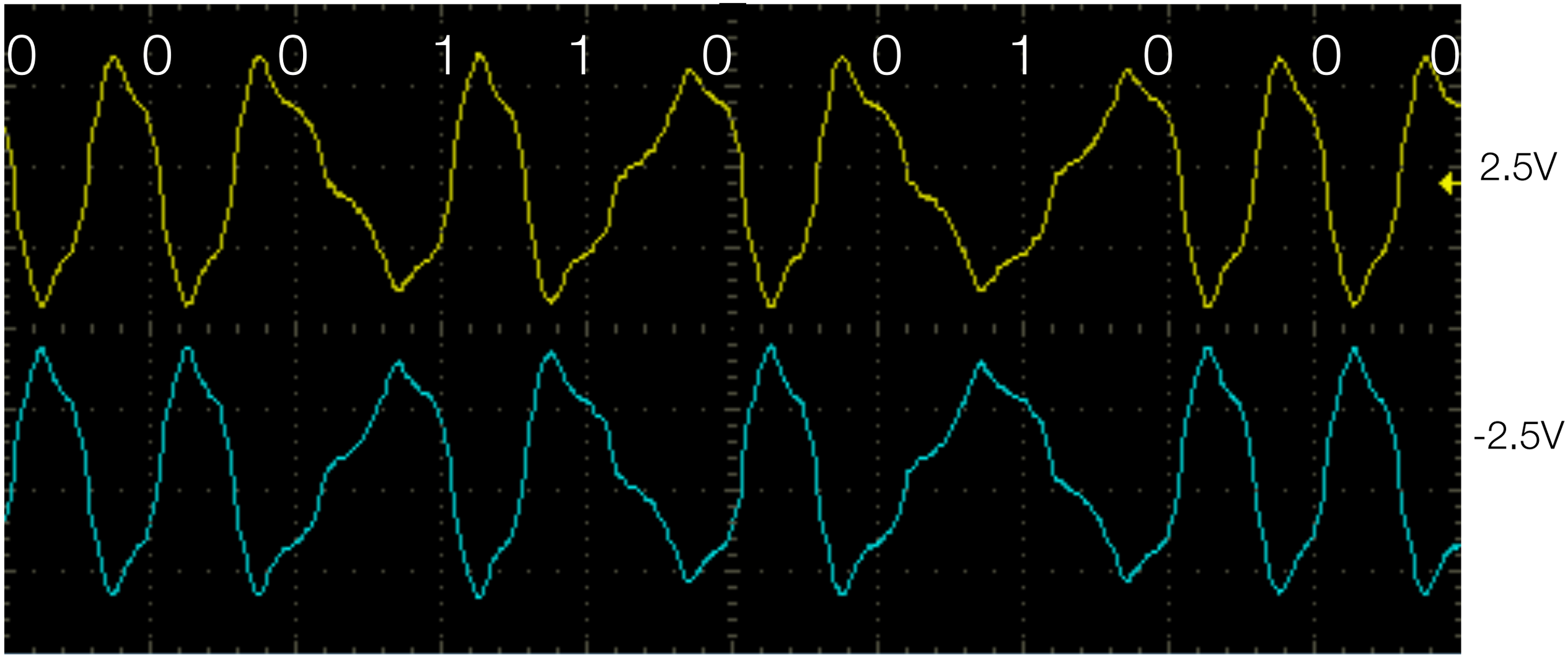
More $$$

Sheet of paper is 100µm thick

10MBit

- 10MHz clock
- 1 bit every 100 nanoseconds
- Manchester phase encoding
- Polarity reversal




0 0 0 1 1 0 0 1 0 0 0
2.5V
-2.5V
100ns

100MBit (Fast Ethernet)

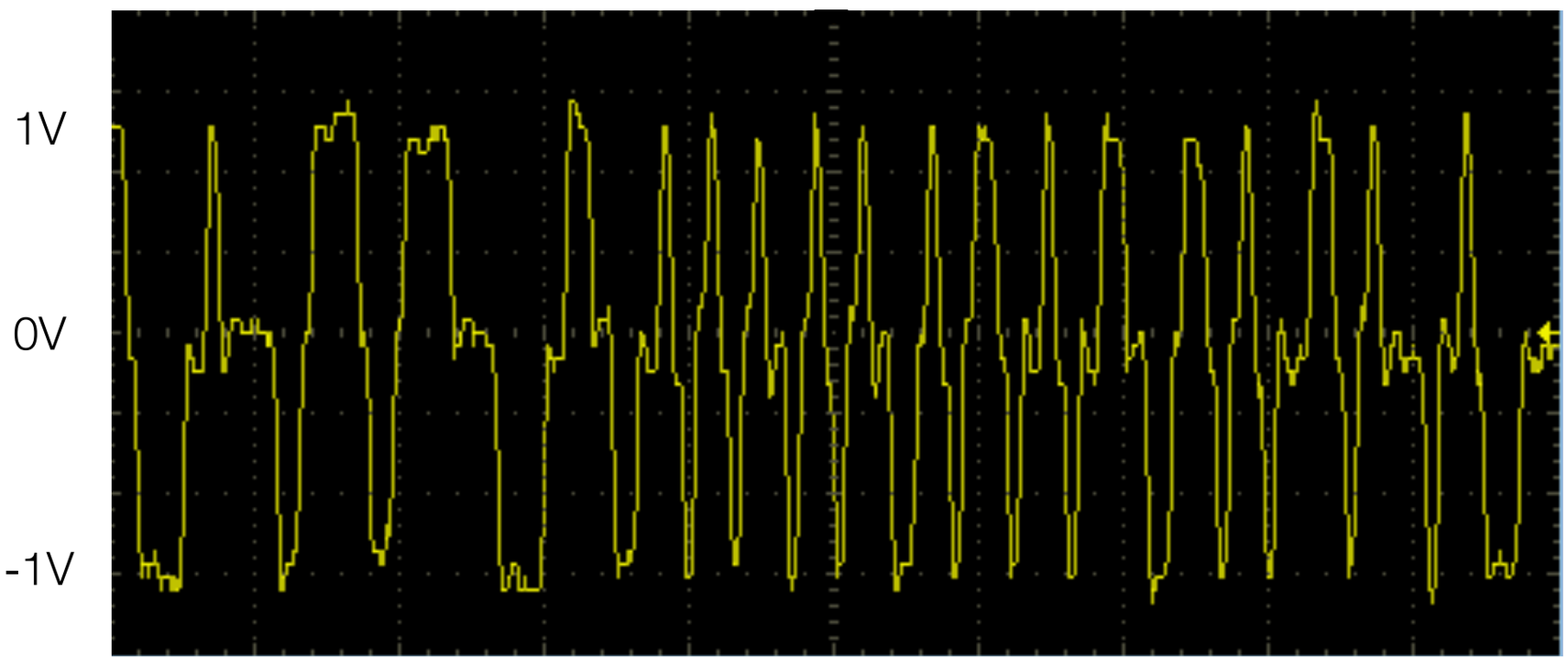
4 bits → 4B/5B → Scramble → MLT-3

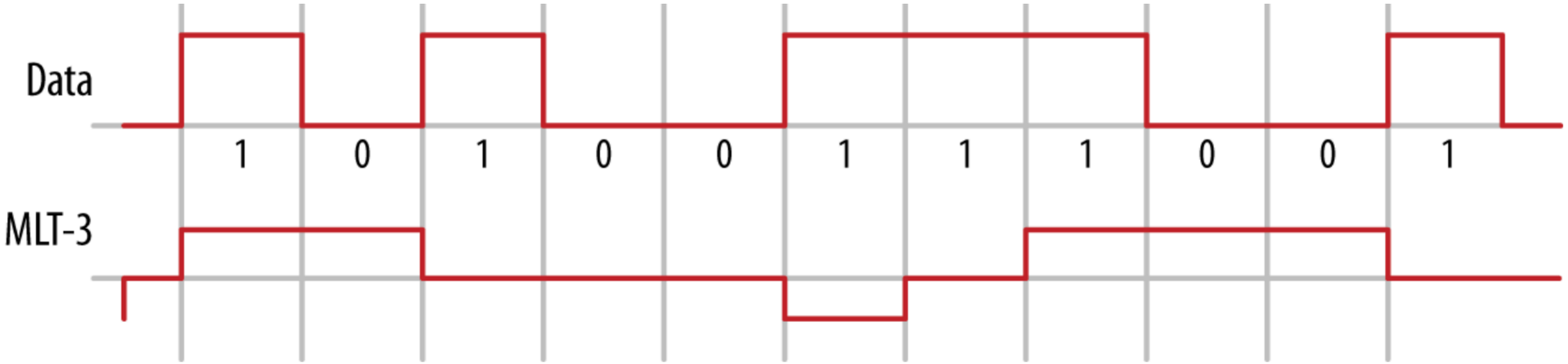
16 hex symbols
16 control symbols

At least 2 bit transitions per block for clock recovery

Spread across multiple frequencies to reduce interference

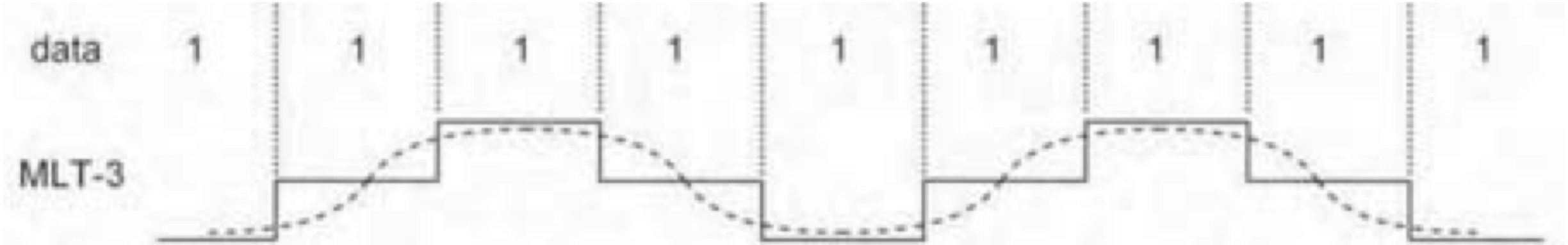
Reduce bandwidth





0: no transition, 1: transition

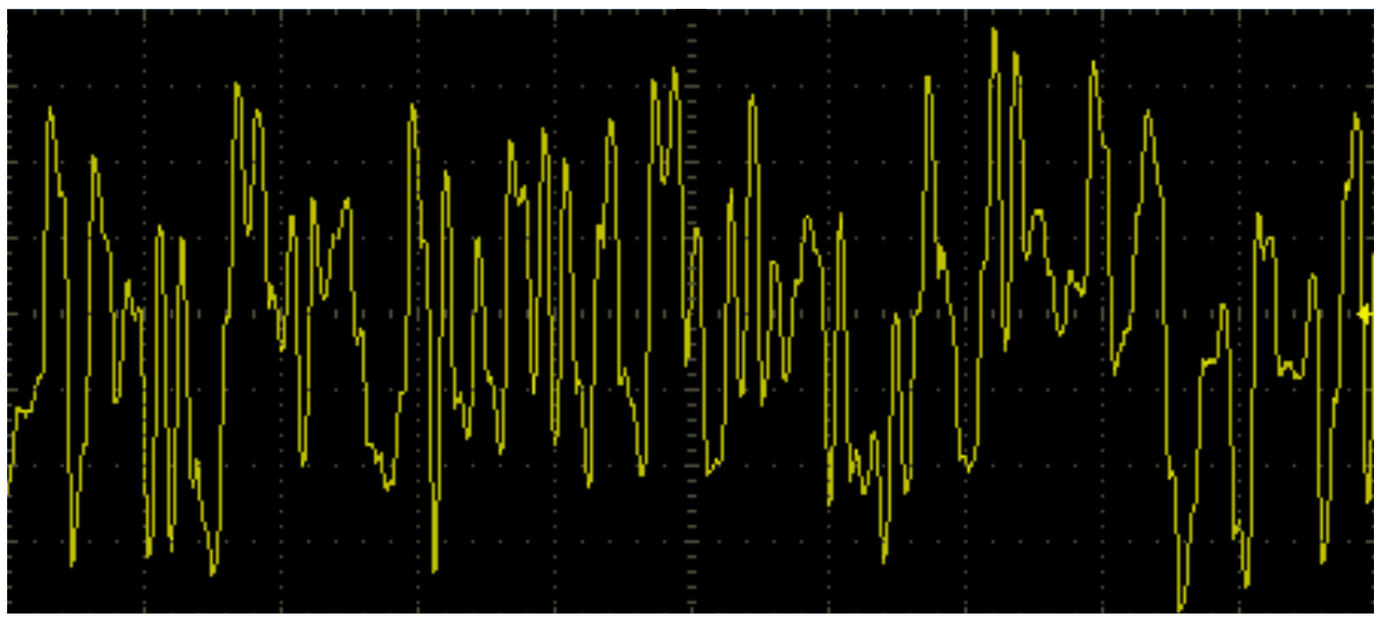
1 → 0 → -1 → 0 → 1

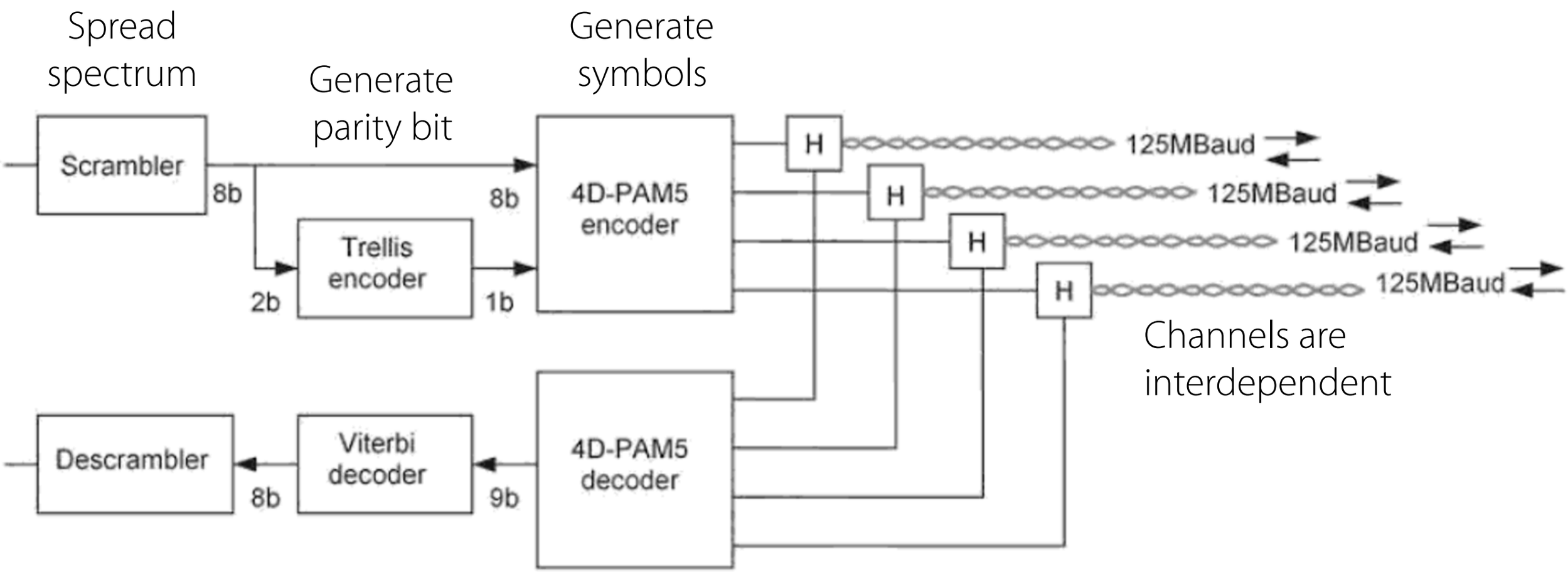


4B/5B creates 125MBaud signal
MLT-3 reduces fundamental frequency to 31.25MHz

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



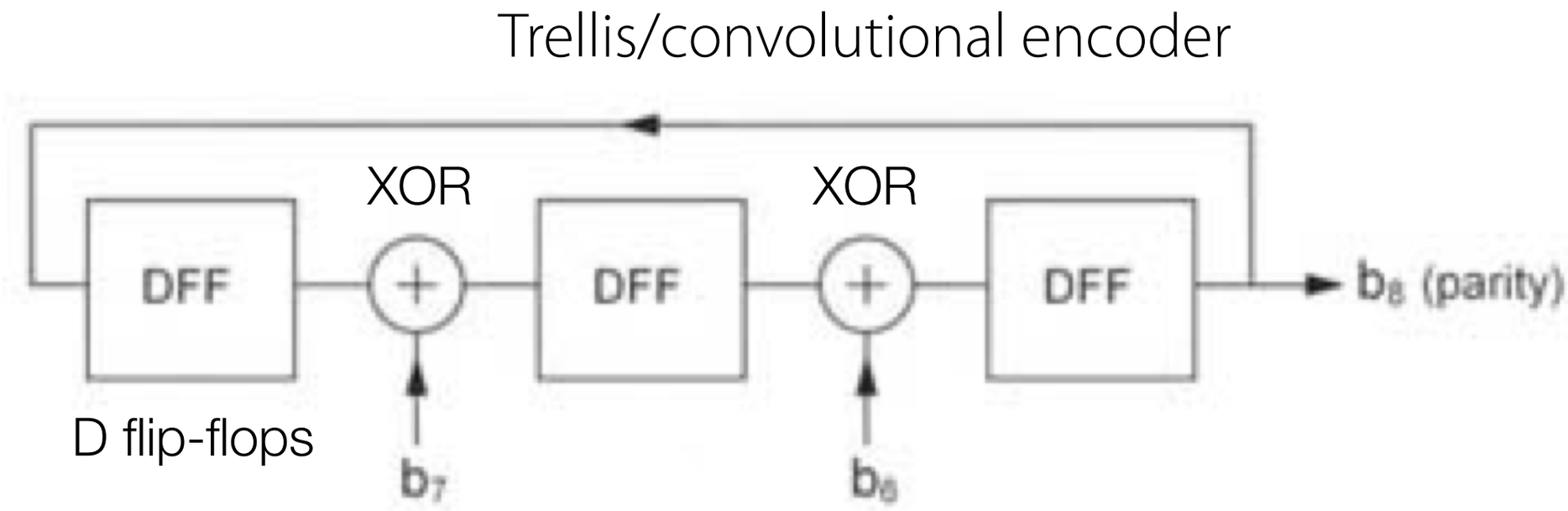


5^4=625 symbols in total

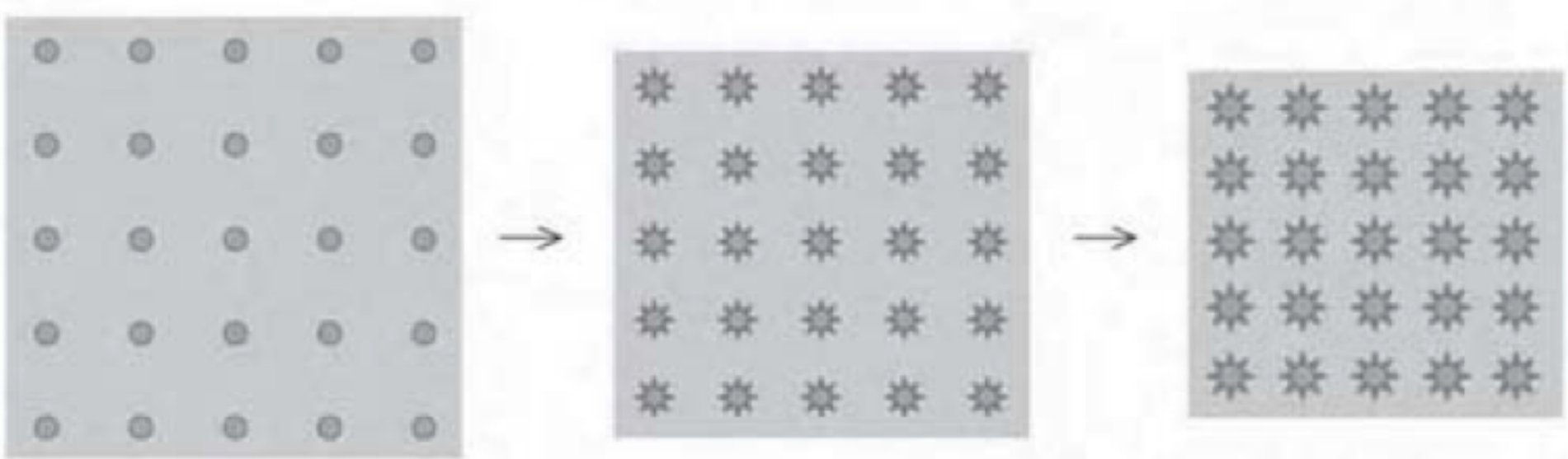
Encode 9 bit word: 8 data bits + 1 parity bit.
2^9=512 possible bitstrings

Remaining 113 symbols used for control or discarded

Parity bit generation



2D-PAM5



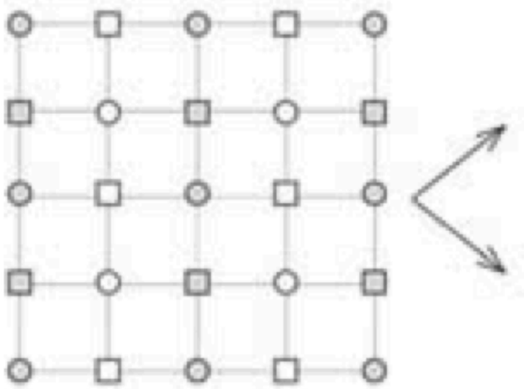
Attenuation - points move closer

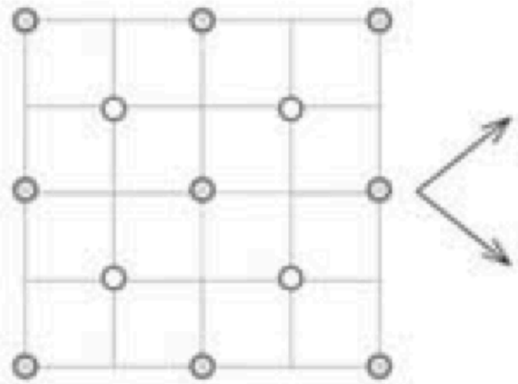
Noise - points blur

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

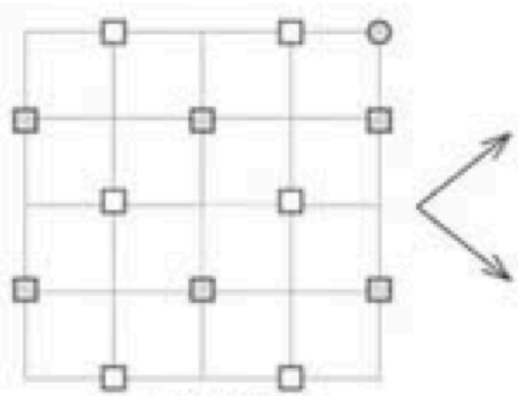
Decoder chooses most likely sequences using Hamming distance as distance metric

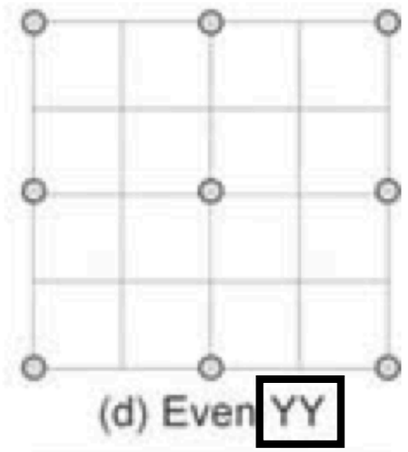
Symbols in a subset have distance 2

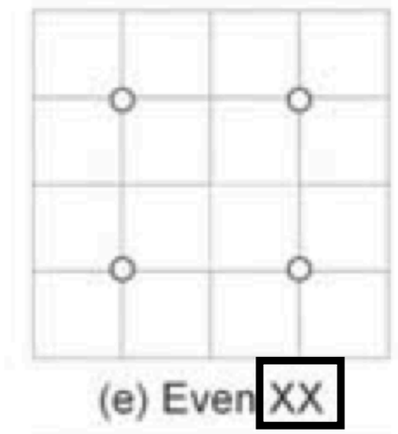


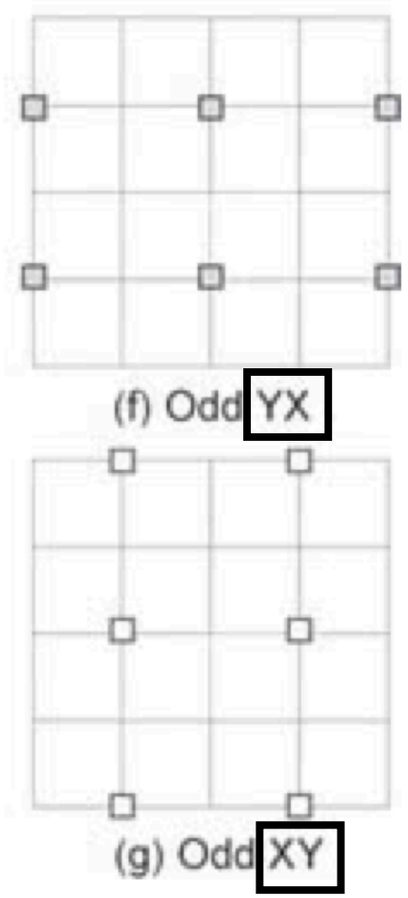


# of coordinates









= YY
= XX
= YX
= XY

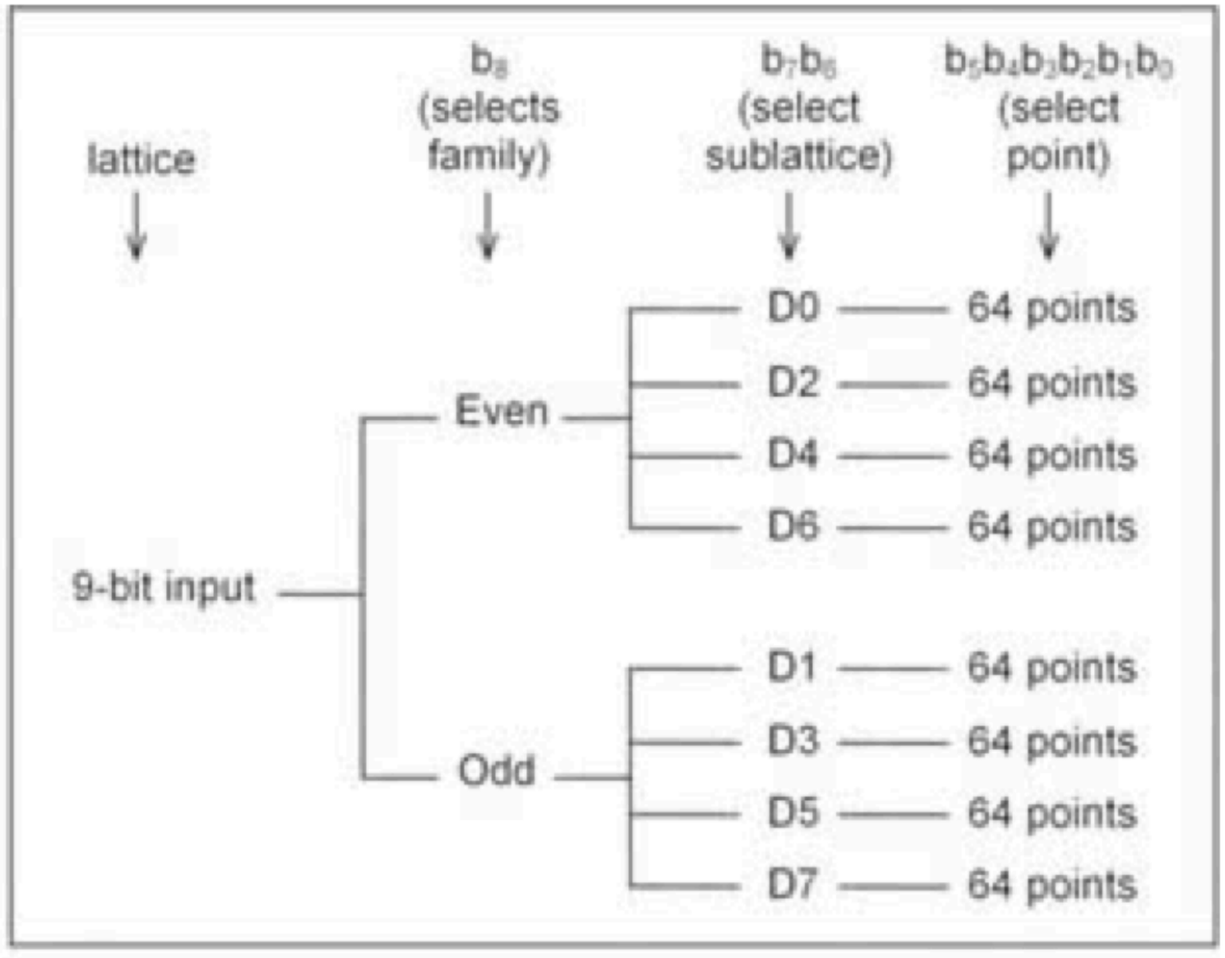
In 4D-PAM5 we have 16 subsets.
(XXXX, XXXY,…,YYYY)

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 >= 4

Sub-lattice	Contents	Number of points	Selected points
D0	XXXX + YYYY	$2^4+3^4=97$	64
D1	XXXY + YYYX	$(2^3)3+(3^3)2=78$	64
D2	XXYY + YYXX	72	64
D3	XXYX + YYXY	78	64
D4	XYYX + YXXY	72	64
D5	XYYY + YXXX	78	64
D6	XYXY + YXYX	72	64
D7	XYXX + YXYY	78	64
	Total	625	512

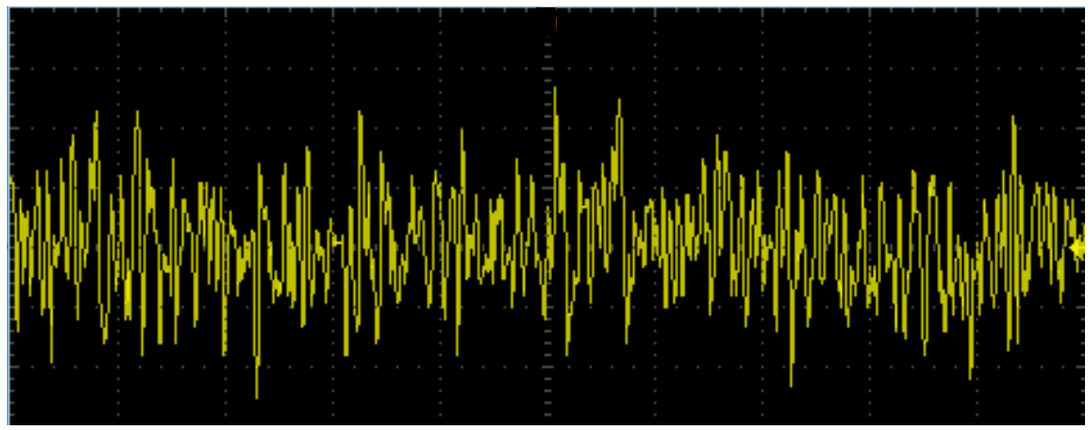


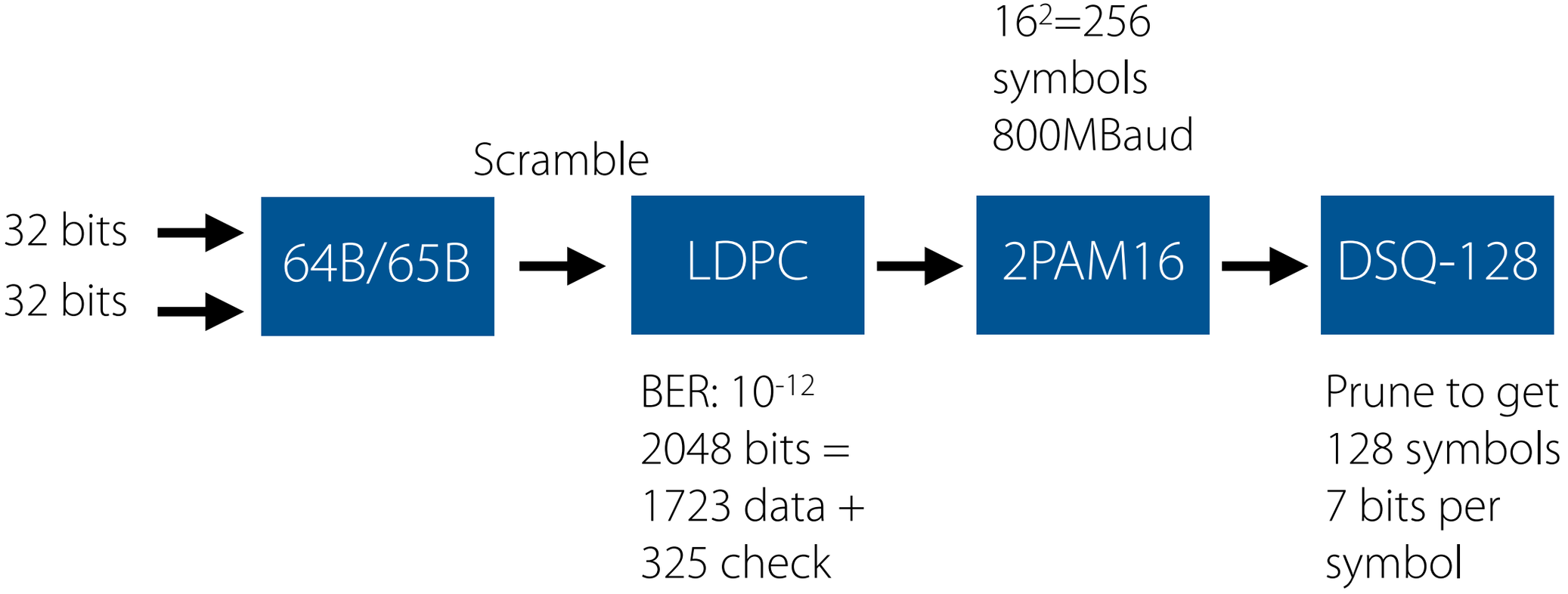
9 bit word → 1 of 512 symbols → voltage levels

10GBit

Each twisted pair takes care of 2.5Gbps both ways

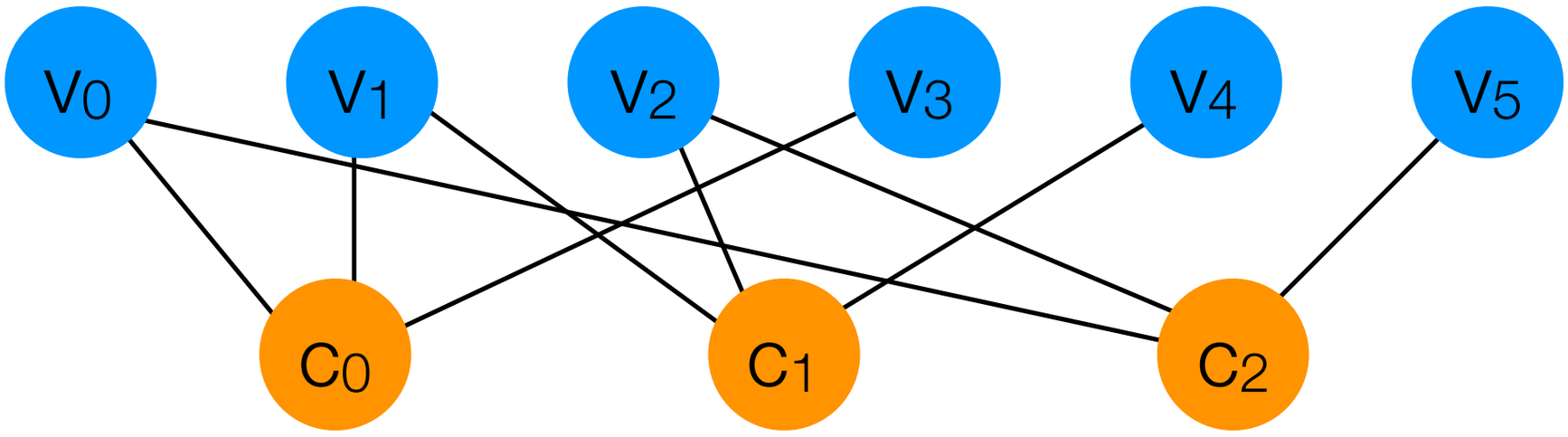
About 400MHz bandwidth





LDPC example

1	0	1	**0**	1	0	Received bits
0,1	1,0	1,1	1	1	0	Checked bits
1	0	1	**1**	1	0	Corrected bits



Variable nodes (2048)

Check nodes (325)

V_0: $V_1 \oplus V_3 \rightarrow 0$

V_1: $V_0 \oplus V_3 \rightarrow 1$

V_3: $V_0 \oplus V_1 \rightarrow 1$

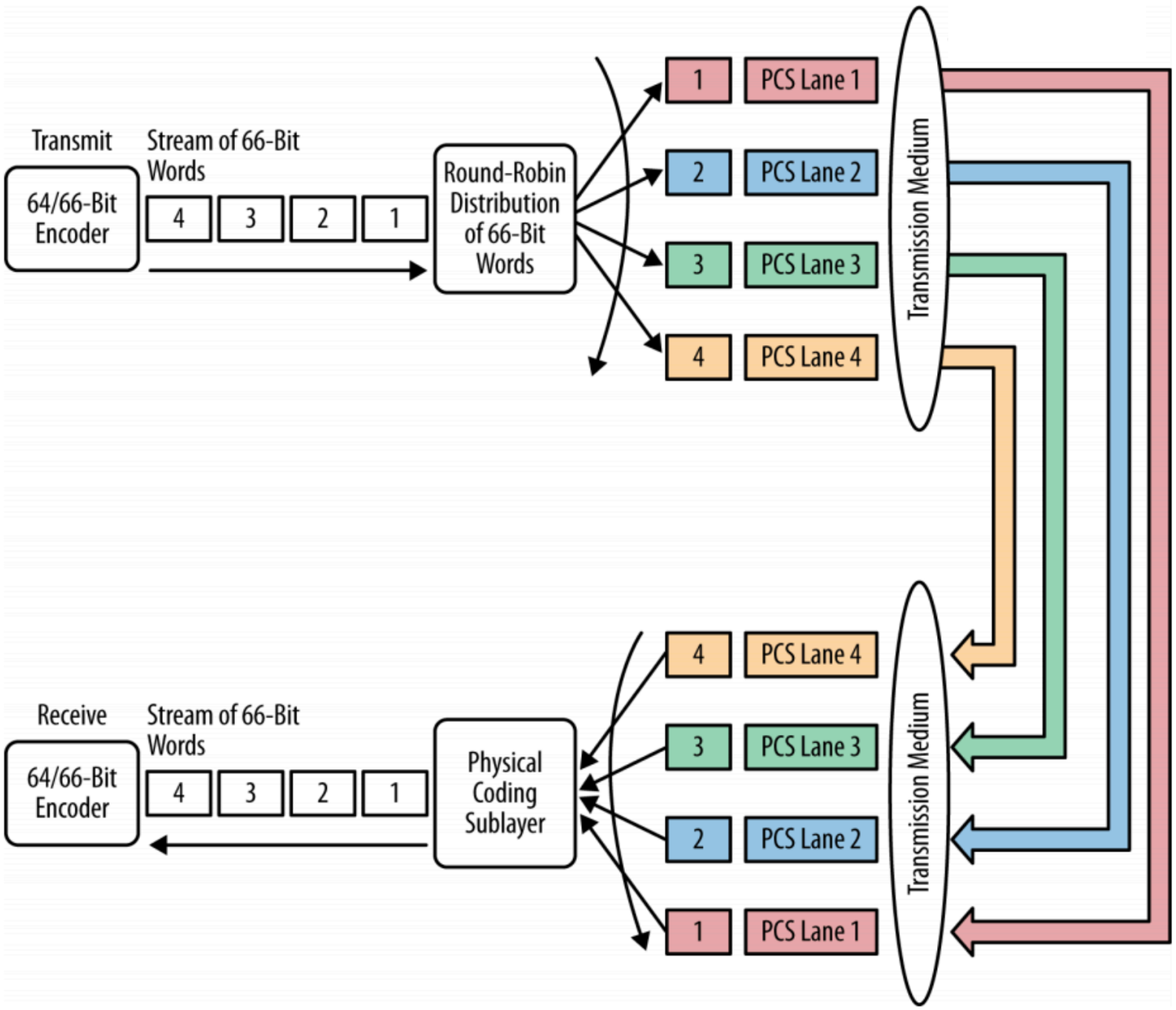
V_1 - 0
V_2 - 1
V_4 - 1

V_0 - 1
V_2 - 1
V_5 - 0

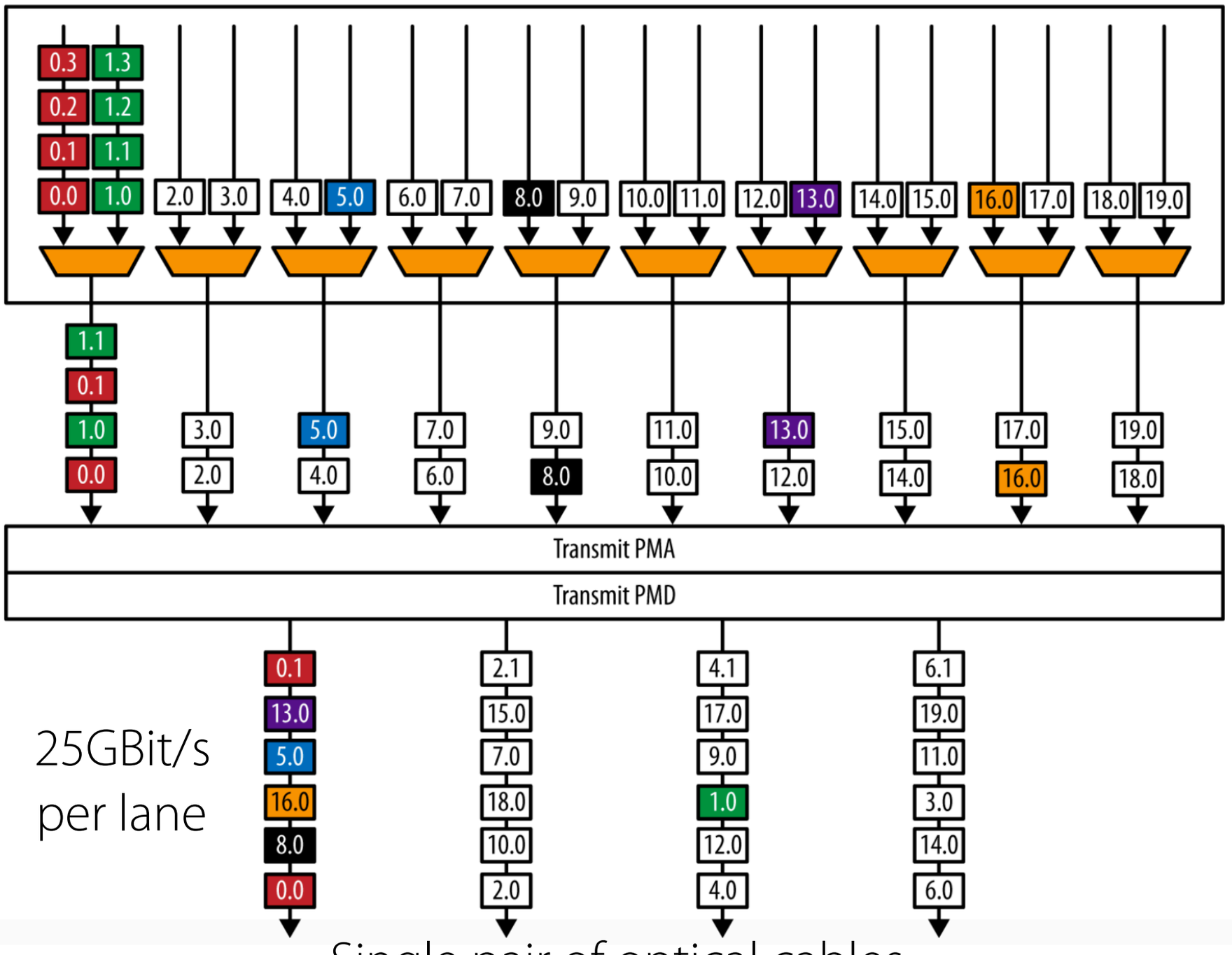
1. v nodes send bits to c nodes
2. c nodes send a v node the ⊕ of other received values
e.g. c_0 sends v_0 the value $v_1 \oplus v_4$
3. v nodes do a majority vote

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)

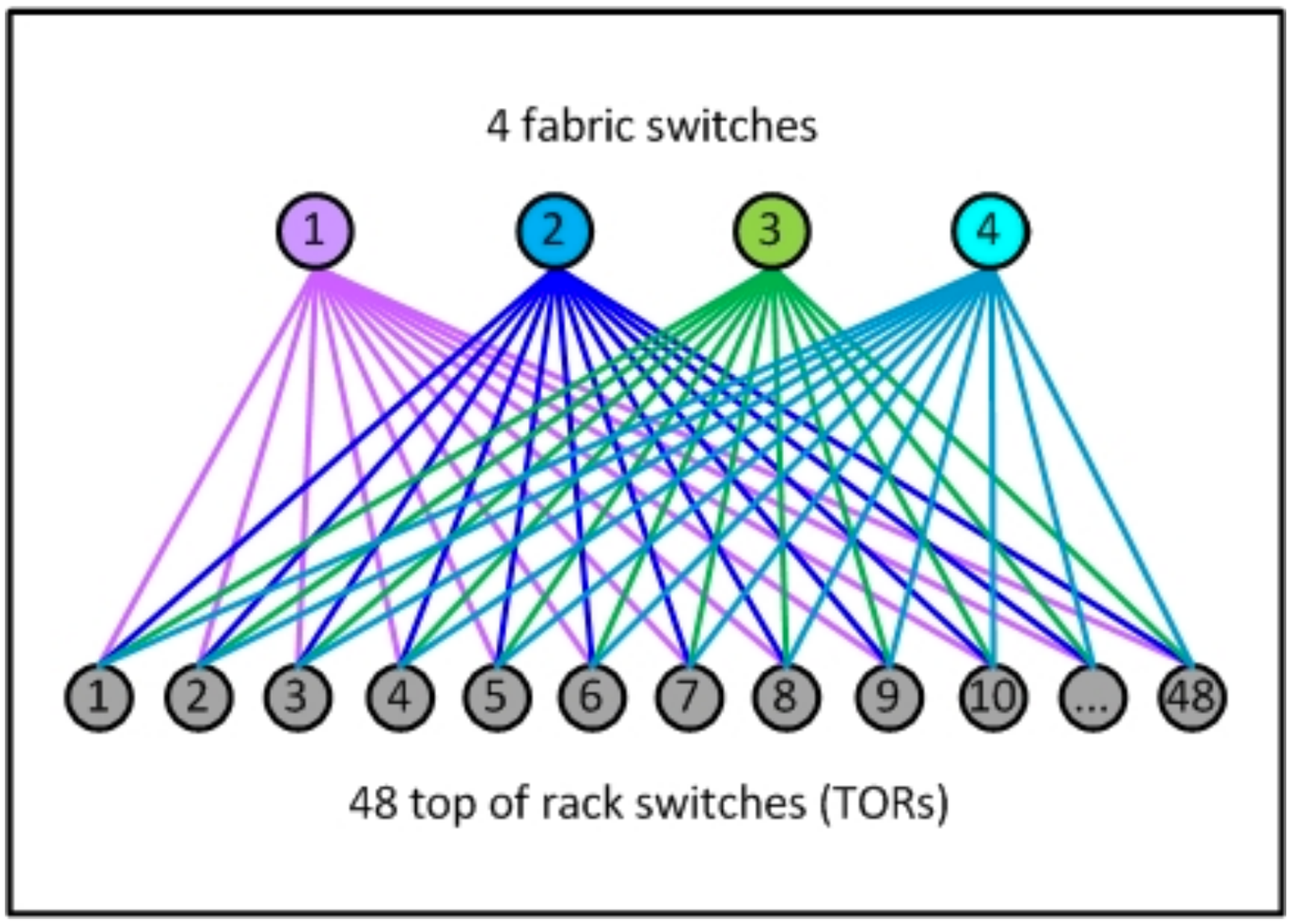
Single pair of optical cables

Datacenters

- Facebook (2014)
  - 4x40Gbps

HPC wire

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

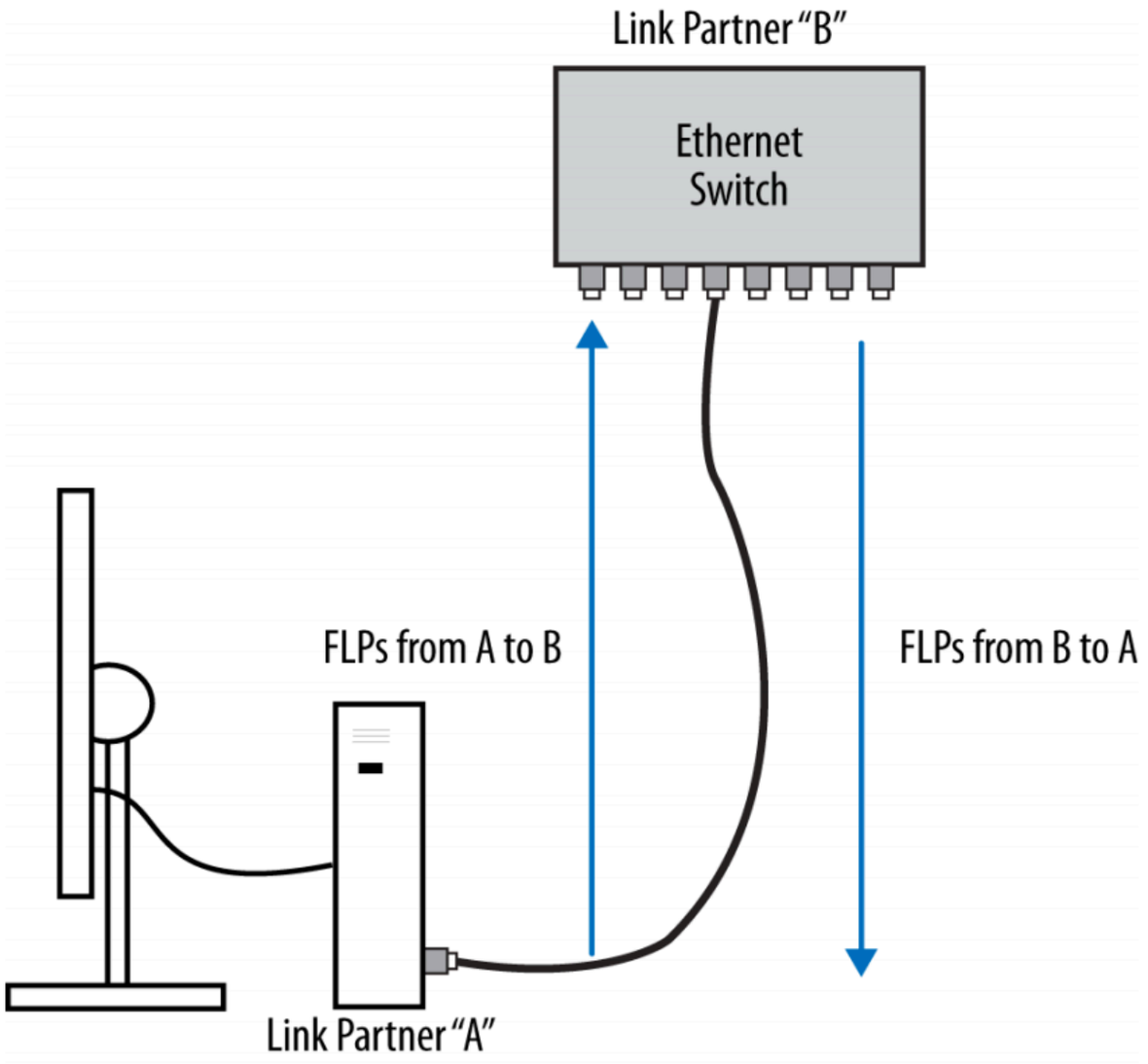


Autonegotiation

- 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

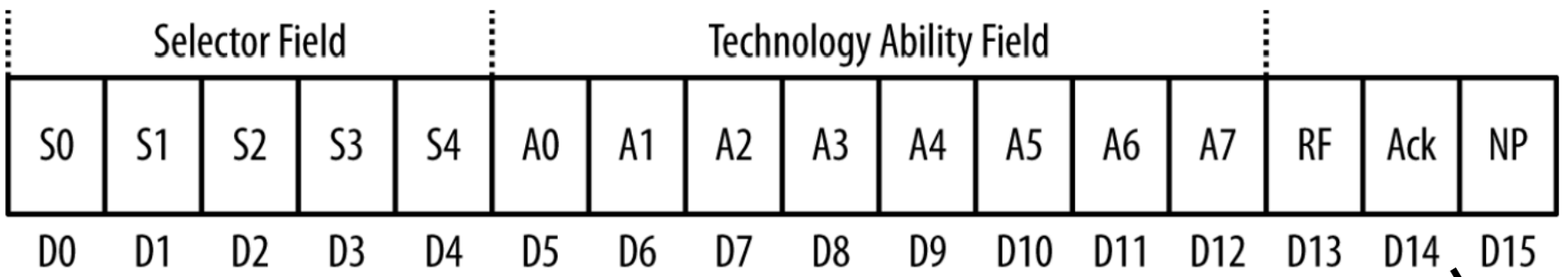
Autonegotiation

- 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

Next page bit to say if there's another message containing additional vendor info



LAN type: Ethernet

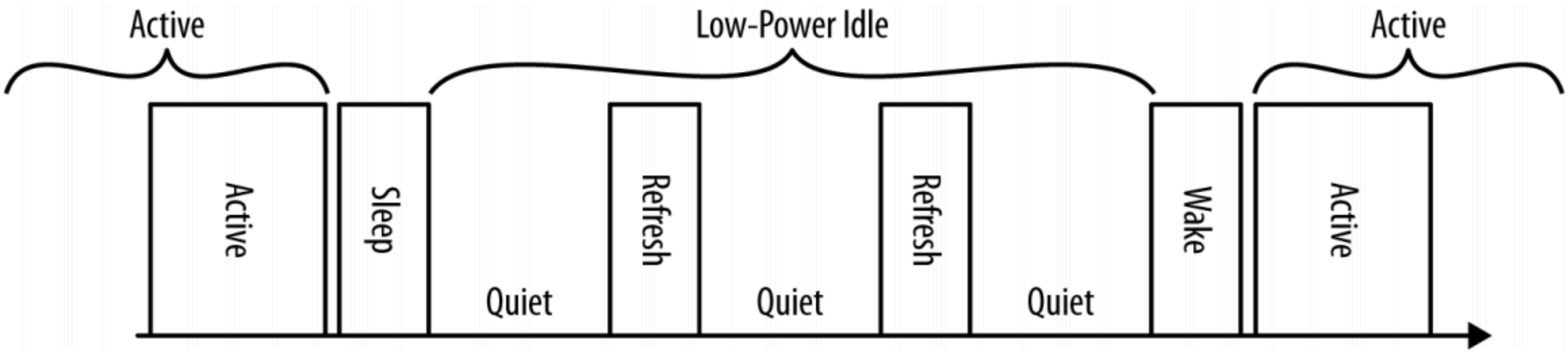
Speed, duplex, pause ability

Remote fault bit set if receiver senses failure

Set to 1 after receiving three messages

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%



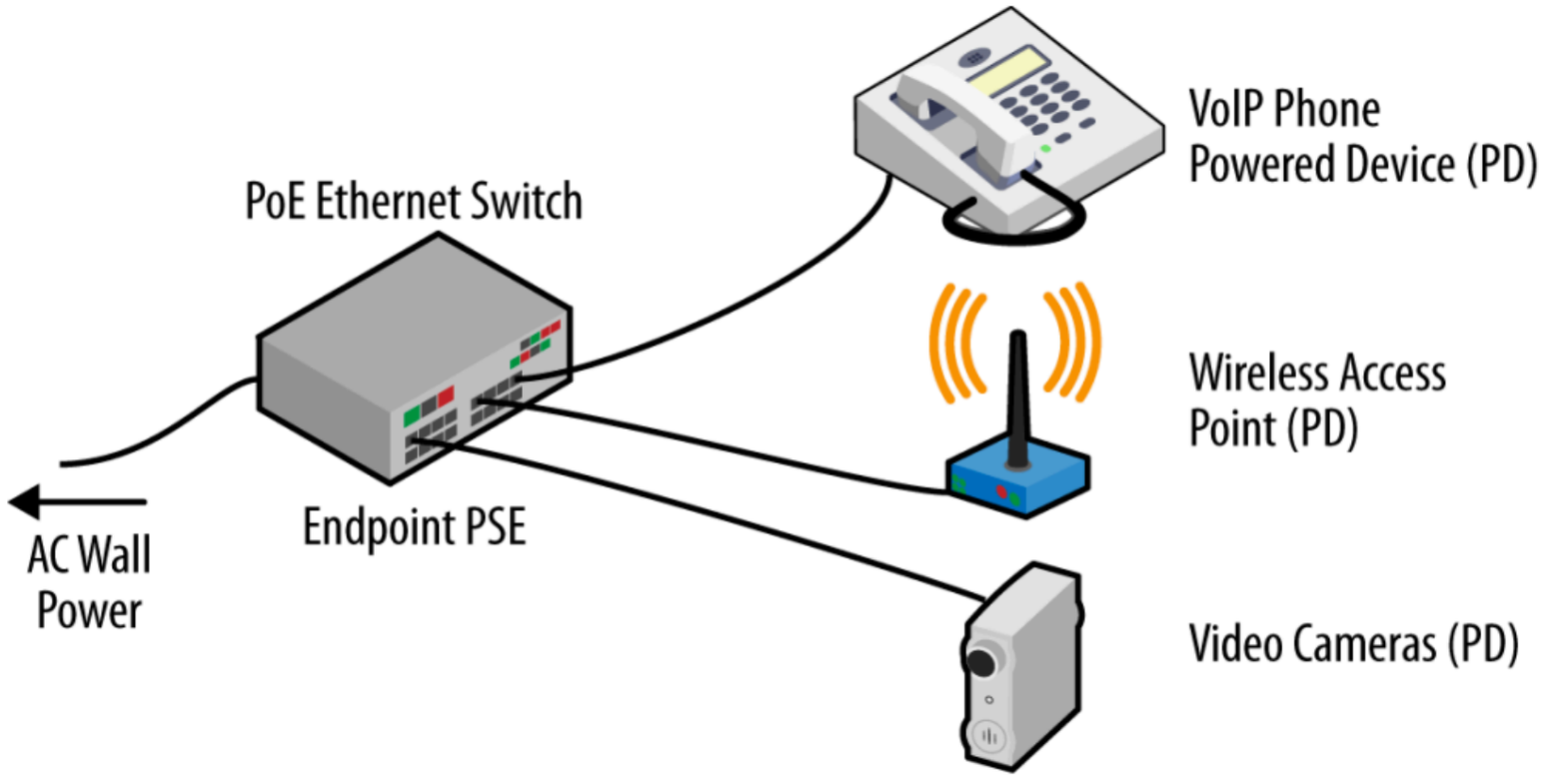
Sleep
Send LPI for *Time To Sleep*
If received LPI
Go to sleep

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

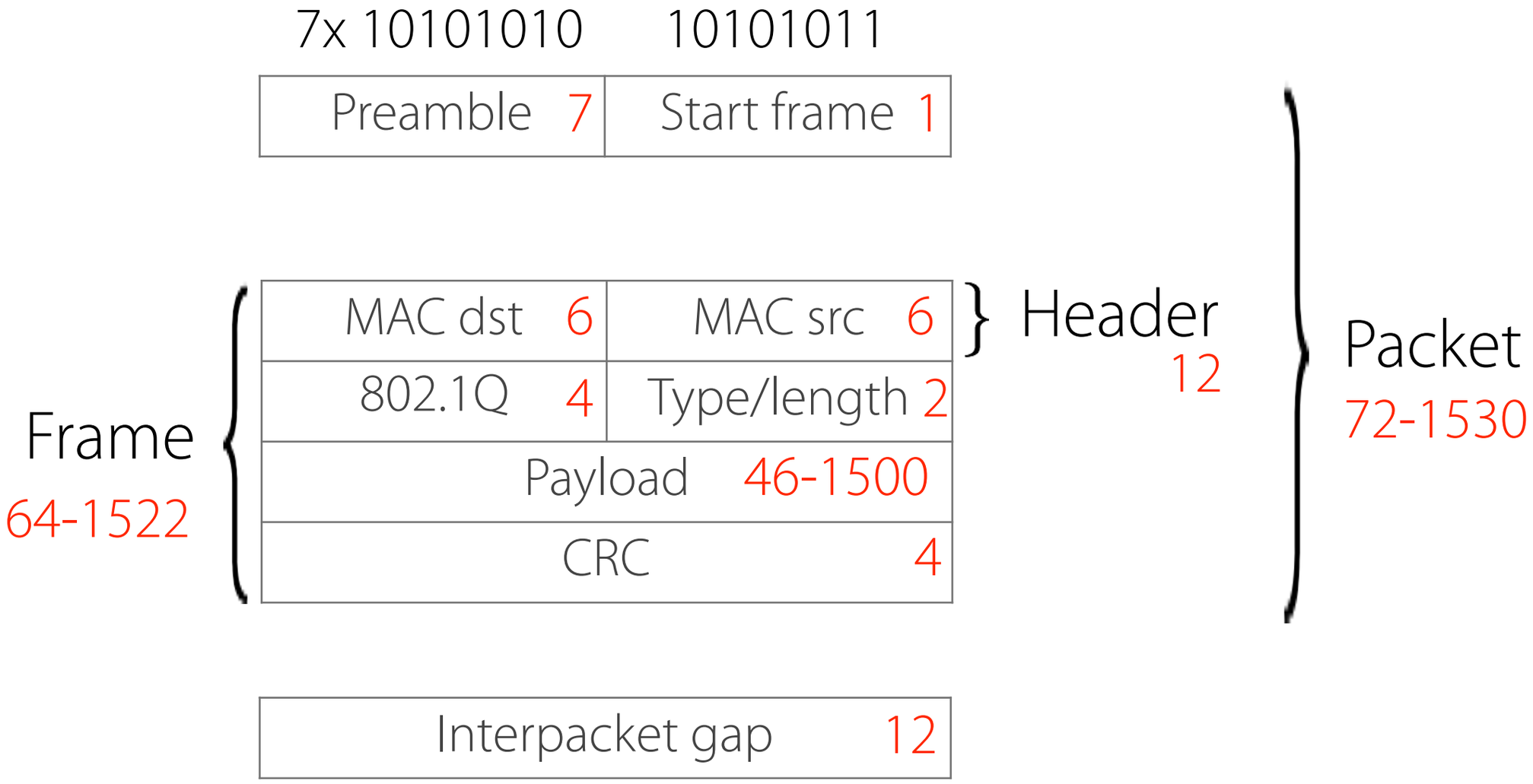
I (mA)	Pout (W)	Precv (W)
0-4	15.4	0.44-12.95
9-12	4	0.44-3.84
17-20	7	3.84-6.49
26-30	15.4	6.49-12.95
36-44	36	12.95-25.5

Flow Control

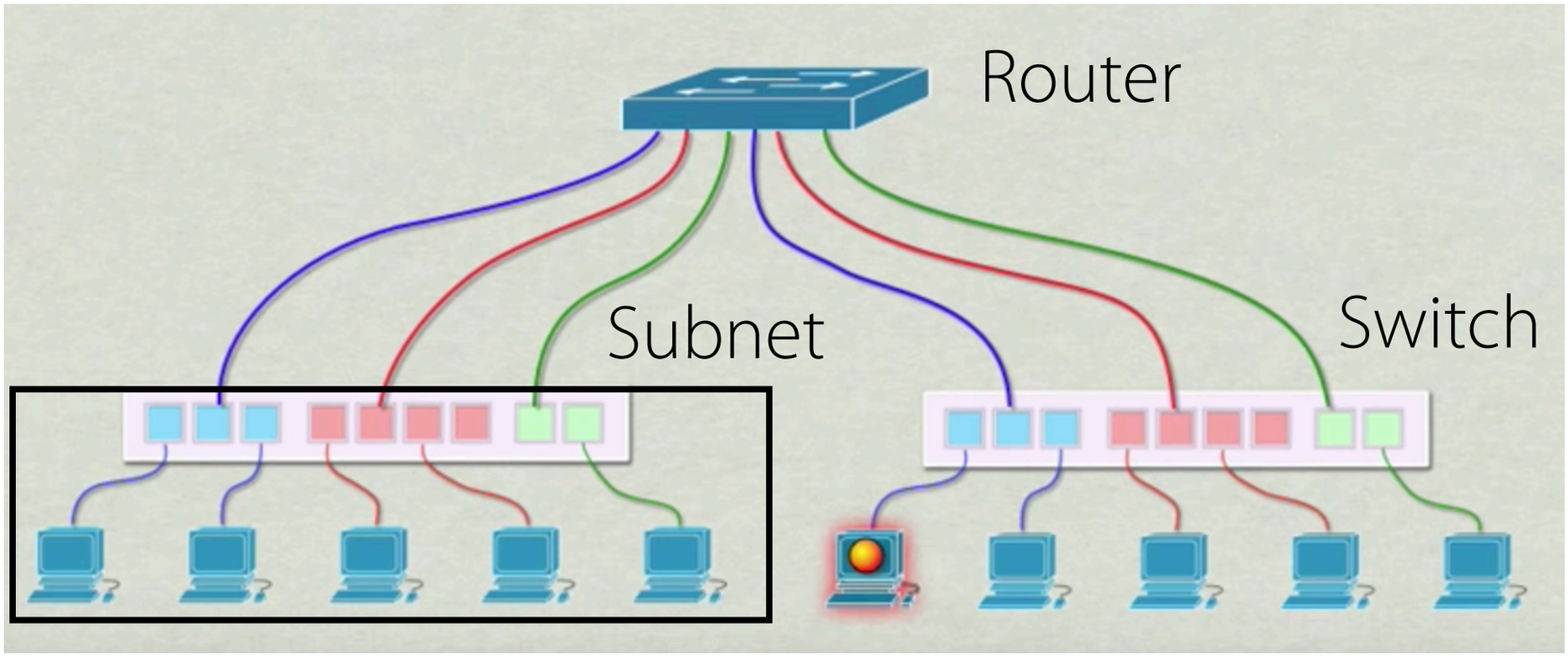
- Pause frame: tell sender to stop transmitting for X time
  - Used to handle NICs with small buffers
  - Send pause frame to 01-80-C2-00-00-01

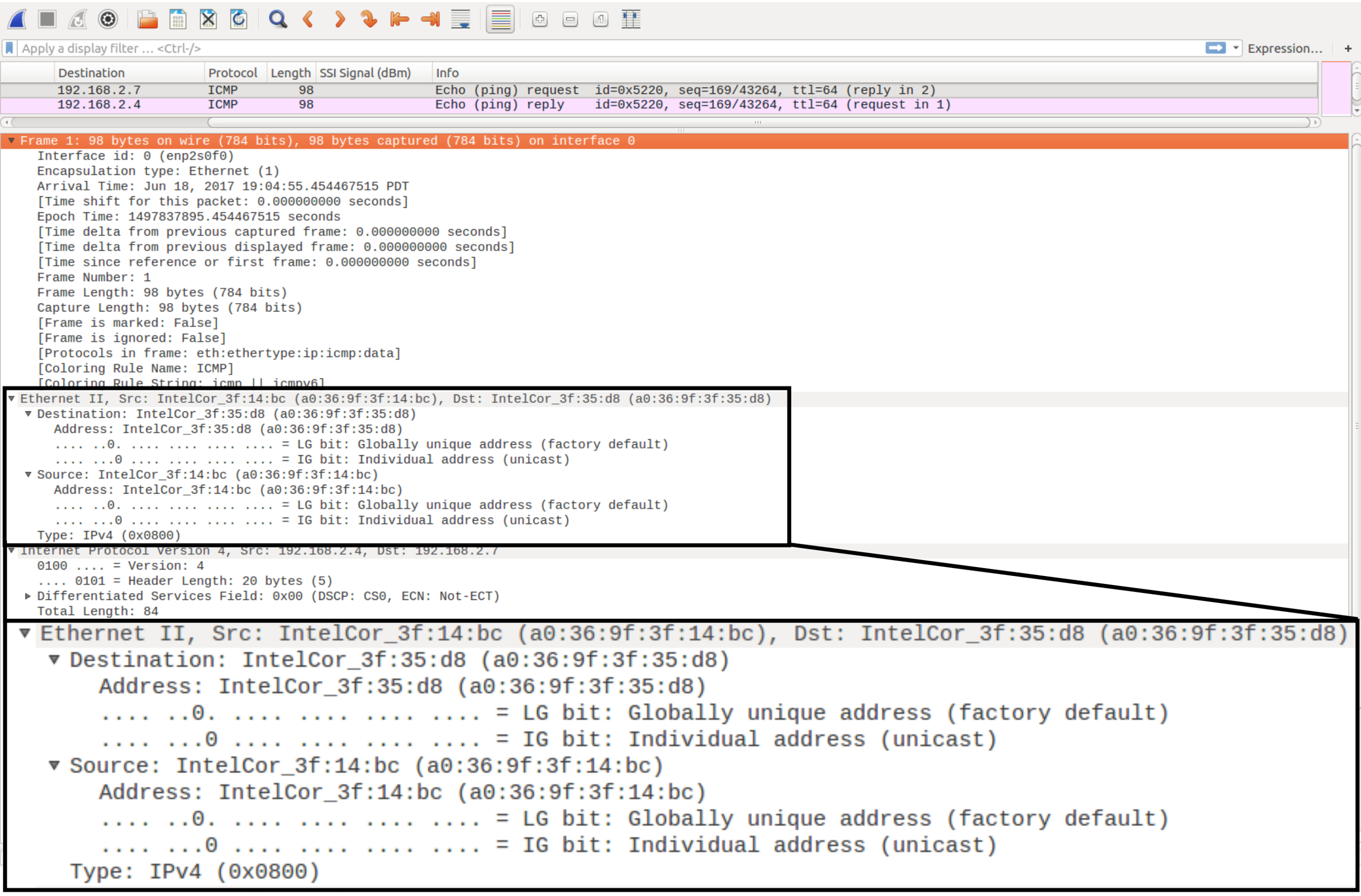
A Deep Dive into the Ethernet

Ethernet Packet



Q-tag: VLAN priority indicator
VLAN: LAN within a switch





Ethernet II, Src: IntelCor_3f:14:bc (a0:36:9f:3f:14:bc), Dst: IntelCor_3f:35:d8 (a0:36:9f:3f:35:d8)
Destination: IntelCor_3f:35:d8 (a0:36:9f:3f:35:d8)
Address: IntelCor_3f:35:d8 (a0:36:9f:3f:35:d8)
.... ..0. .... = LG bit: Globally unique address (factory default)
....0 .... = IG bit: Individual address (unicast)
Source: IntelCor_3f:14:bc (a0:36:9f:3f:14:bc)
Address: IntelCor_3f:14:bc (a0:36:9f:3f:14:bc)
.... ..0. .... = LG bit: Globally unique address (factory default)
....0 .... = IG bit: Individual address (unicast)
Type: IPv4 (0x0800)

Type={IPv4, IPv6, ARP, 802.1Q}