EXPLORING THE IEEE 802 ETHERNET ECOSYSTEM

June 2017

George Sullivan, Senior IEEE Member

Adjunct Professor NYU Tandon School of Engineering

Northrop Grumman Technical Fellow (retired)

BEFORE WE SHARE OUR OPINIONS……

- "At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position, explanation, or interpretation of the IEEE."
- IEEE-SA Standards Board Operation Manual (subclause 5.9.3)

IEEE-SA STANDARDS DEVELOPMENT

Open, consensus-based process

- Open – anybody can participate (payment of meeting fees may be needed)
- Individual standards development
  - Each individual has one vote
- Corporate standards development
  - One company/one vote
- The IEEE, IETF, and ITU-T coordinate their standards development activities through formal liaison letters and informal discussions.
- Results frequently adopted by national, regional, and international standards bodies
- IEEE Standard published after approval
- Standard is valid for 10 years after approval
- After 10 years, must be revised or withdrawn

SI TERMINOLOGY FOR ORDERS OF MAGNITUDE

Magnitude	Binary	Prefix	Symbol
10^{24}	2^{80}	Yotta	Y
10^{21}	2^{70}	Zetta	Z
10^{18}	2^{60}	Exa	E
10^{15}	2^{50}	Peta	P
10^{12}	2^{40}	Tera	T
10^{9}	2^{30}	Giga	G
10^{6}	2^{20}	Mega	M
10^{3}	2^{10}	kilo	k
10^{0}	2^{0}	uni	1
10^{-3}	2^{-10}	milli	m
10^{-6}	2^{-20}	micro	u
10^{-9}	2^{-30}	nano	n
10^{-12}	2^{-40}	pico	p
10^{-15}	2^{-50}	femto	f
10^{-18}	2^{-60}	atta	a
10^{-21}	2^{-70}	zepto	z
10^{-24}	2^{-80}	yocto	y

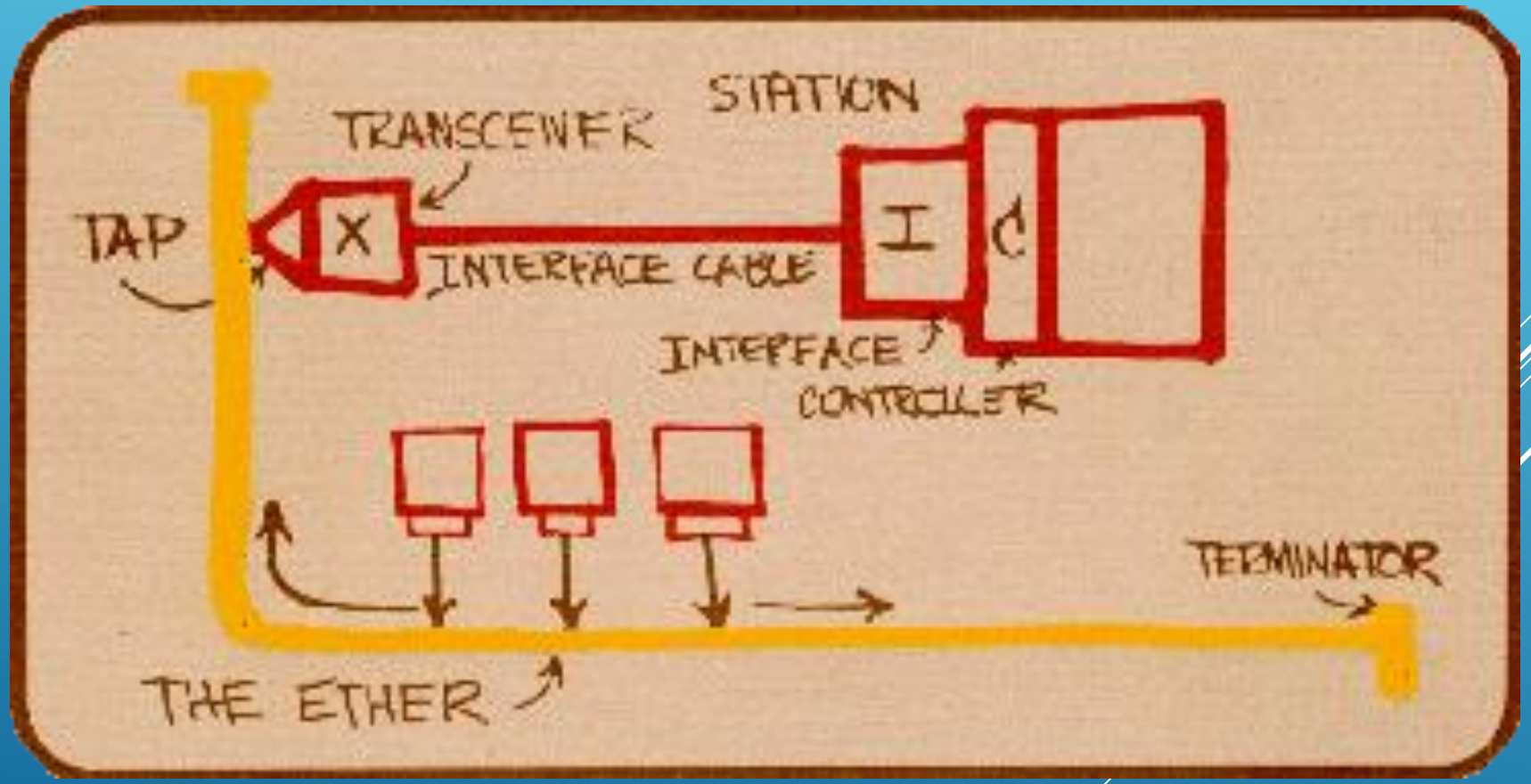
←"one"

Be sure to properly distinguish symbols with upper / lower case

AGENDA

- IEEE 802 Networks
  - Who
  - What
  - Why
  - When
  - Where
  - How
  - How Much
  - References

1st Sketch of Ethernet - 1973

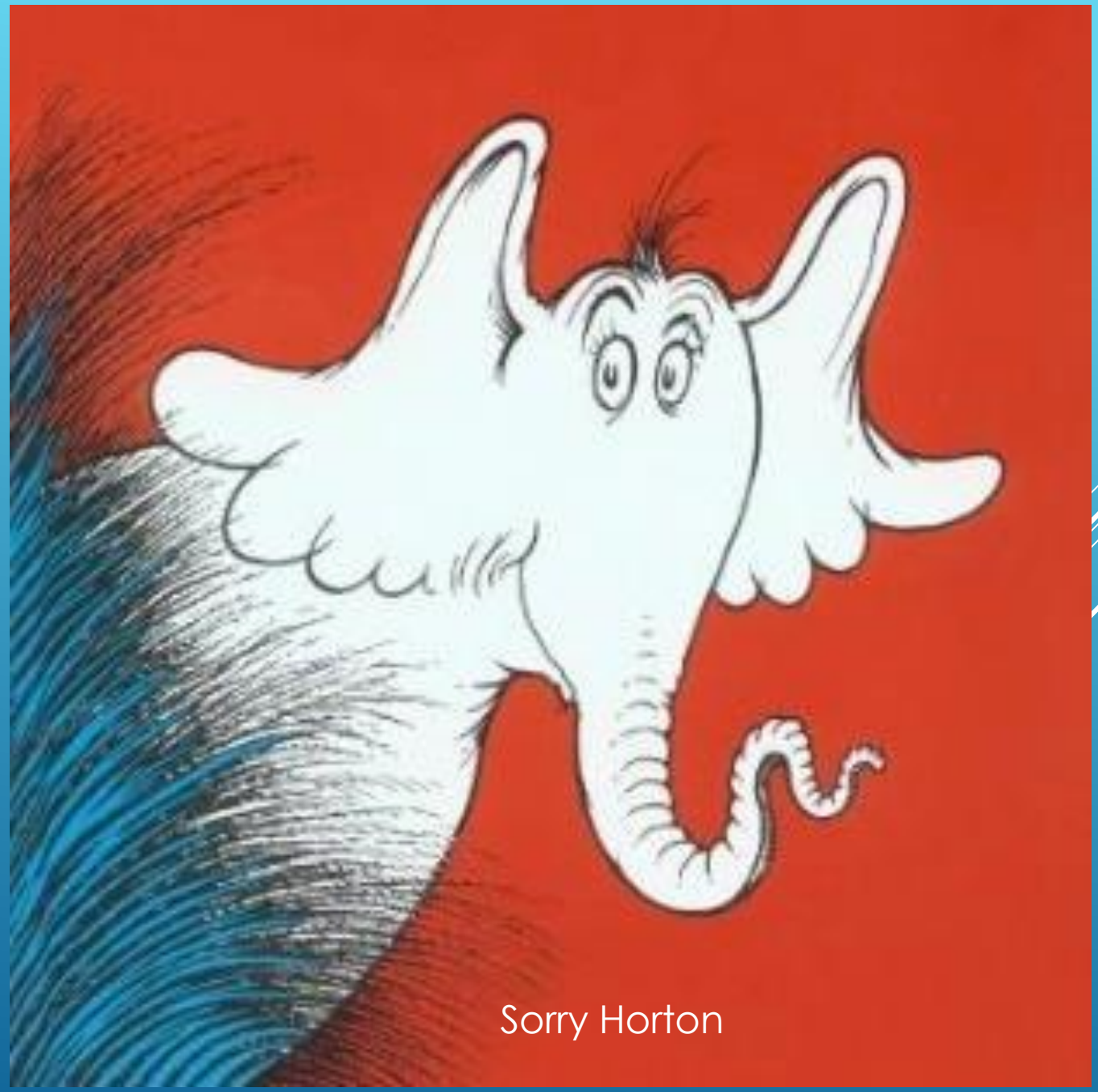


WHO

IEEE 802 Networks Have NO Who

802 addresses do not designate people

IEEE 802 addresses are assigned to physical interfaces



IEEE 802 IS A STANDARDS COMMITTEE DEVELOPING NETWORK STANDARDS

- IEEE 802 standards specify frame based networks with source and destination addressing, and asynchronous timing.
- In a **frame-based** system, the data structure is a variable-length sequence of data octets.

- By contrast, **cell-based** communication transmits data in small fixed-length units within specified time intervals
- **Isochronous** communication transmits data as a steady stream of octets, or groups of octets, at equal time intervals over a circuit switched architecture substrate

IEEE 802 STANDARDS SUMMARY

Standard	Name	Standard	Name
802.1	Bridging & Management	802.13	unassigned
802.2	Logical Link Control	802.14	Cable Modem
802.3	Ethernet	802.15	Wireless PANs
802.4	Token Bus	802.16	Broadband Wireless MANs
802.5	Token Ring	802.17	Resilient Packet Ring
802.6	DQDB Metropolitan Area Network	802.18	Radio Regulatory TAG
802.7	Broadband TAG	802.19	TV Whitespace Coexistence
802.8	Fiber Optic TAG	802.20	Mobile Broadband Wireless Access (MBWA)
802.9	Integrated Services LAN (ISLAN)	802.21	Media Independent Handover
802.10	Interoperable LAN Security (SILS)	802.22	Wireless Regional Area Networks
802.11	Wireless LANs	802.23	Emergency Services
802.12	Demand Priority	802.24	Vertical Applications TAG

WHEN - BRIEF ETHERNET HISTORY

- Ethernet was developed at Xerox PARC between 1973 and 1974. It was inspired by ALOHAnet, which Robert Metcalfe had studied. It was first documented on May 22, 1973, where he named it after the disproven luminiferous ether as an "omnipresent, completely-passive medium for the propagation of electromagnetic waves". In 1975, Xerox filed a patent application listing Metcalfe, David Boggs, Chuck Thacker, and Butler Lampson as inventors. In 1976, after the system was deployed at PARC, Metcalfe and Boggs published a seminal paper.
- Metcalfe left Xerox in June 1979 to form 3Com. He convinced Digital Equipment Corporation (DEC), Intel, and Xerox to work together to promote Ethernet as a standard. The so-called "DIX" standard, for "Digital/Intel/Xerox", specified 10 Mbit/s Ethernet, with 48-bit destination and source addresses and a global 16-bit Ethertype field. It was published on September 30, 1980 as "The Ethernet, A Local Area Network. Data Link Layer and Physical Layer Specifications". Version 2 was published in November, 1982 and defines Ethernet II. Formal standardization efforts and resulted in the draft publication of IEEE 802.3 on June 23, 1983.
- 3Com shipped its first 10 Mbit/s Ethernet 3C100 NIC in March 1981, and that year started selling adapters for PDP-11s and VAXes, as well as Multibus-based Intel and Sun Microsystems computers. This was followed quickly by DEC's Unibus to Ethernet adapter, which DEC sold and used internally to build its own corporate network, which reached over 10,000 nodes by 1986, making it one of the largest computer networks in the world.
- An Ethernet adapter card for the IBM PC was released in 1982, and, by 1985, 3Com had sold 100,000. By the early 1990s, Ethernet became a must-have feature for modern computers. Ethernet ports began to appear on some PCs and most workstations. This process was greatly sped up with the introduction of 10BASE-T at which point Ethernet ports appeared even on low-end motherboards.

802.3 EVOLUTION STANDARDS LIST (BEFORE 1990)

Name	Year	Description
Experimental Ethernet	1973	2.94 Mbit/s over coaxial cable (coax) bus
Ethernet II (DIX v2.0)	1982	10 Mbit/s over thick coax. Frames have a Type field. This frame format is used on all forms of Ethernet by protocols in the Internet protocol suite.
IEEE 802.3 standard	1983	10BASE5 10 Mbit/s over thick coax. Same as Ethernet II (above) except Type field is replaced by Length, and an 802.2 LLC header follows the 802.3 header. Based on the CSMA/CD Process.
802.3a	1985	10BASE2 10 Mbit/s over thin Coax (a.k.a. thinnet or cheapernet)
802.3b	1985	10BROAD36 Use of CATV RF Modems for 10 Mbit/s Operation
802.3c	1985	10 Mbit/s repeater specs
802.3d	1987	Fiber-optic inter-repeater link (FOIRL)
802.3e	1987	1 Mbit/s 1BASE5 or StarLAN

802.3 STANDARDS LIST (1990 TO 1999)

Name	Year	Description
802.3i	1990	10BASE-T 10 Mbit/s over twisted pair
802.3j	1993	10BASE-F 10 Mbit/s over Fiber-Optic
802.3u	1995	100BASE-TX, 100BASE-T4, 100BASE-FX Fast Ethernet at 100 Mbit/s w/autonegotiation
802.3x	1997	Full Duplex and flow control; also incorporates DIX framing, so there's no longer a DIX/802.3 split
802.3-1998	1998	A revision of base standard incorporating the above amendments and errata
802.3ac	1998	Max frame size extended to 1522 bytes (to allow "Q-tag") The Q-tag includes 802.1Q VLAN information and 802.1p priority information.
802.3y	1998	100BASE-T2 100 Mbit/s over low quality twisted pair
802.3z	1998	1000BASE-X Ethernet over Fiber-Optic at 1 Gbit/s
802.3ab	1999	1000BASE-T Ethernet over twisted pair at 1 Gbit/s

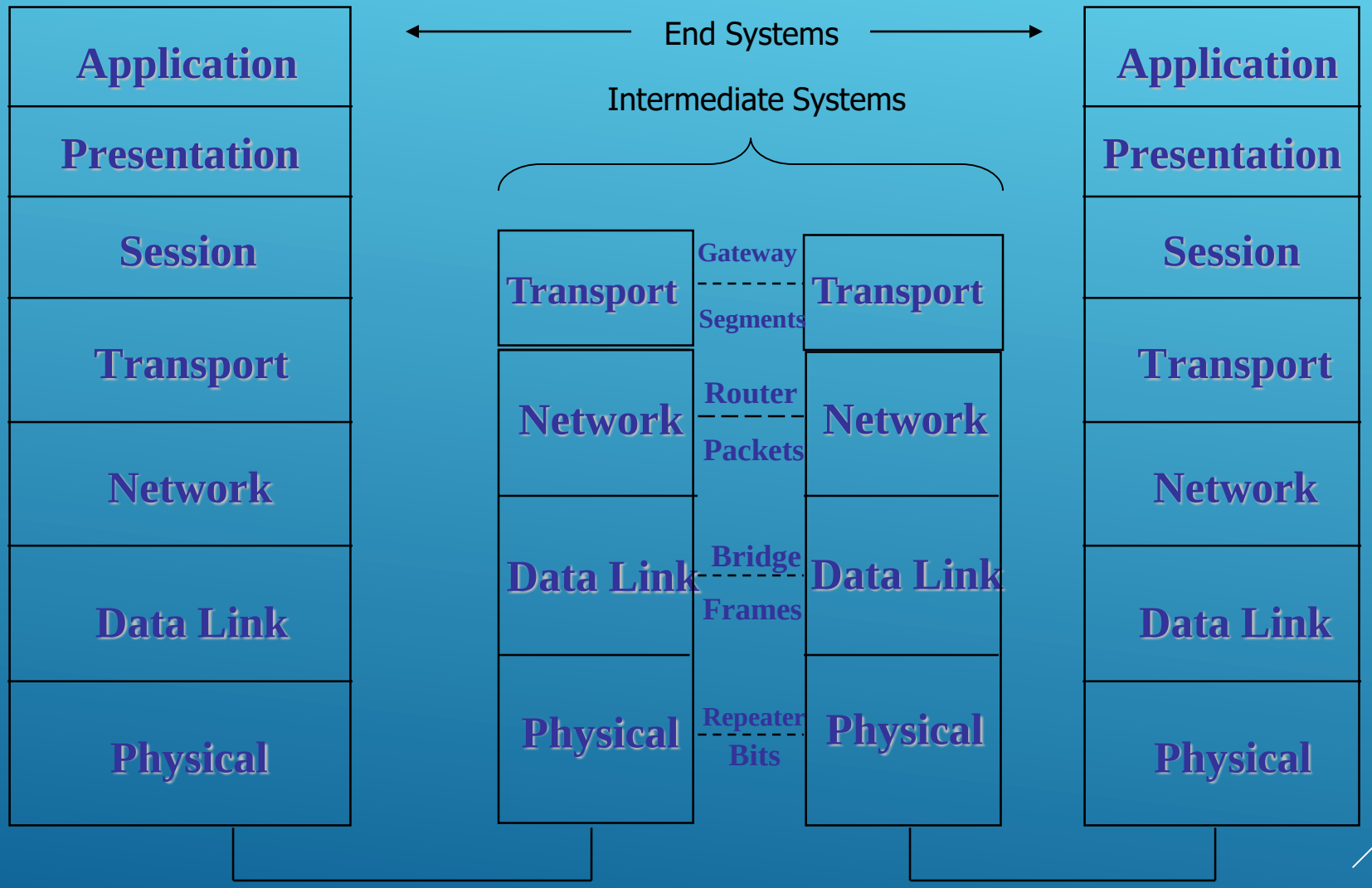
802.3 STANDARDS LIST (2000 TO 2010)

Name	Year	Description
802.3ad	2000	Link aggregation for parallel links, since moved to IEEE 802.1AX
802.3-2002	2002	A revision of base standard incorporating the three prior amendments and errata
802.3ae	2002	10 Gigabit Ethernet over fiber; 10GBASE-SR, 10GBASE-LR, 10GBASE-ER, 10GBASE-SW, 10GBASE-LW, 10GBASE-EW
802.3af	2003	Power over Ethernet (15.4 W)
802.3ah	2004	Ethernet in the First Mile
802.3ak	2004	10GBASE-CX4 10 Gbit/s Ethernet over twinaxial cables
802.3-2005	2005	A revision of base standard incorporating the four prior amendments and errata.
802.3an	2006	10GBASE-T 10 Gbit/s Ethernet over unshielded twisted pair (UTP)
802.3aq	2006	10GBASE-LRM 10 Gbit/s Ethernet over multimode fiber
802.3as	2006	Frame expansion
802.3au	2006	Isolation requirements for Power over Ethernet (802.3-2005/Cor 1)
802.3ap	2007	Backplane Ethernet (1 and 10 Gbit/s over printed circuit boards)
802.3aw	2007	Fixed an equation in the publication of 10GBASE-T (released as 802.3-2005/Cor 2)
802.3-2008	2008	A revision incorporating the 802.3an/ap/aq/as amendments, two corrigenda and errata. Link aggregation was moved to 802.1AX.
802.3-2008/Cor 1	2009	Increase Pause Reaction Delay timings which are insufficient for 10 Gbit/s (workgroup name was 802.3bb)
802.3at	2009	Power over Ethernet enhancements (25.5 W)
802.3av	2009	10 Gbit/s EPON
802.3bc	2009	Move and update Ethernet related TLVs (type, length, values), previously specified in Annex F of IEEE 802.1AB (LLDP) to 802.3.
802.3az	2010	Energy-efficient Ethernet
802.3ba	2010	40 Gbit/s and 100 Gbit/s Ethernet. 40 Gbit/s over 1 m backplane, 10 m Cu cable assembly (4×25 Gbit or 10×10 Gbit lanes) and 100 m of MMF and 100 Gbit/s up to 10 m of Cu cable assembly, 100 m of MMF or 40 km of SMF respectively
802.3bd	2010	Priority-based Flow Control. An amendment by the IEEE 802.1 Data Center Bridging Task Group (802.1Qbb) to develop an amendment to IEEE Std 802.3 to add a MAC Control Frame to support IEEE 802.1Qbb Priority-based Flow Control.

802.3 STANDARDS LIST (2011 TO PRESENT)

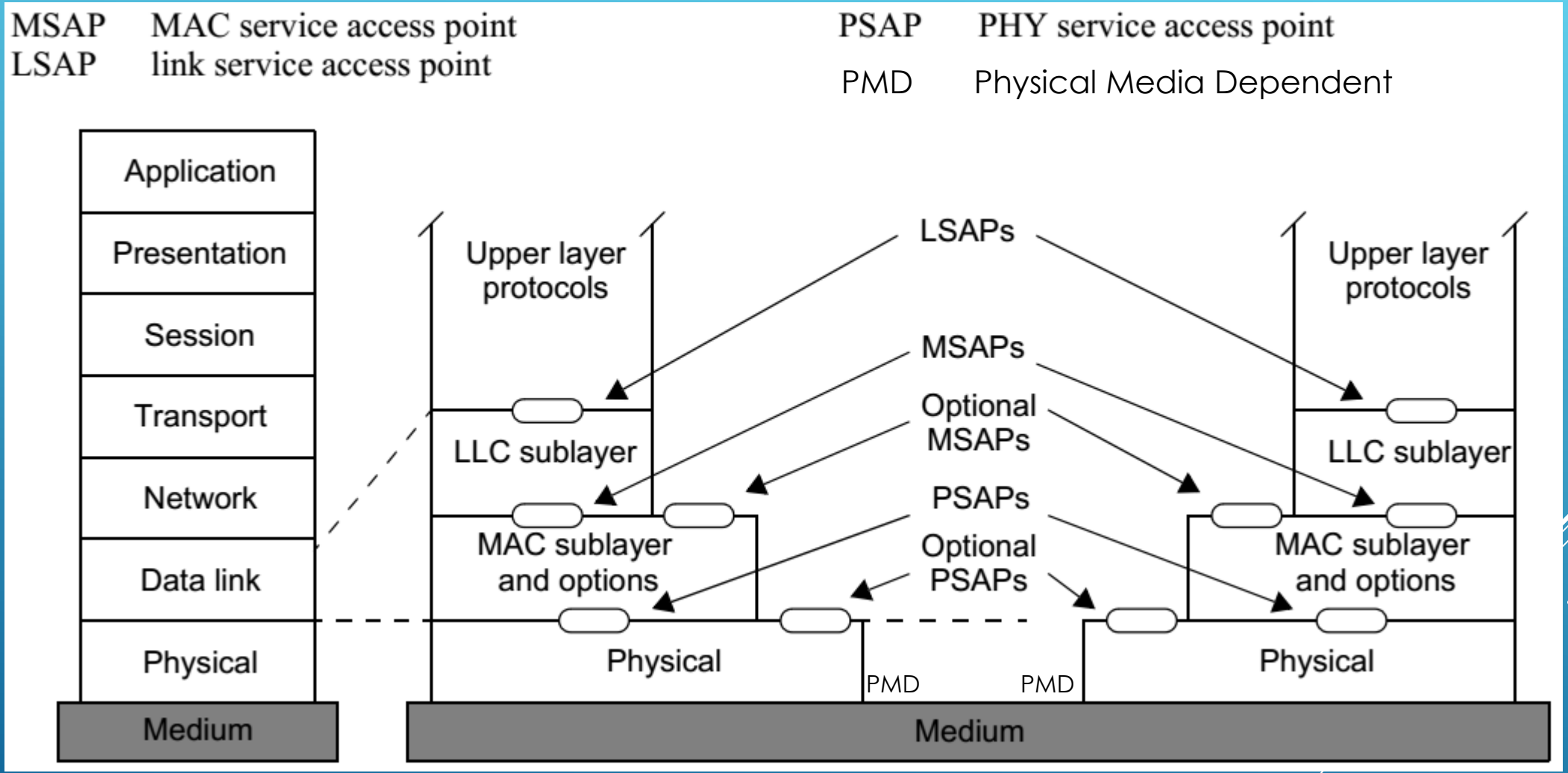
Name	Year	Description
802.3.1	2011	MIB definitions for Ethernet. It consolidates the Ethernet related MIBs present in Annex 30A&B, various IETF RFCs, and 802.1AB annex F into one master document with a machine readable extract. (workgroup name was P802.3be)
802.3bf	2011	Provide an accurate indication of the transmission and reception initiation times of certain packets as required to support IEEE P802.1AS.
802.3bg	2011	Provide a 40 Gbit/s PMD which is optically compatible with existing carrier SMF 40 Gbit/s client interfaces (OTU3/STM-256/OC-768/40G POS).
802.3-2012	2012	A revision of base standard incorporating the 802.3at/av/az/ba/bc/bd/bf/bg amendments, a corrigenda and errata.
802.3bk	2013	This amendment to IEEE Std 802.3 defines the physical layer specifications and management parameters for EPON operation on point-to-multipoint passive optical networks supporting extended power budget classes of PX30, PX40, PRX40, and PR40 PMDs.
802.3bj	2014	Define a 4-lane 100 Gbit/s backplane PHY for operation over links consistent with copper traces on "improved FR-4" (as defined by IEEE P802.3ap or better materials to be defined by the Task Force) with lengths up to at least 1 m and a 4-lane 100 Gbit/s PHY for operation over links consistent with copper twinaxial cables with lengths up to at least 5 m.
802.3-2015	2015	802.3bx – a new consolidated revision of the 802.3 standard including amendments 802.3bk/bj/bm
802.3bm	2015	100G/40G Ethernet for optical fiber
802.3bw	2015	100BASE-T1 – 100 Mbit/s Ethernet over a single twisted pair for automotive applications
802.3bp	2016	1000BASE-T1 – Gigabit Ethernet over a single twisted pair, automotive & industrial environments
802.3bq	2016	25G/40GBASE-T for 4-pair balanced twisted-pair cabling with 2 connectors over 30 m distances
802.3by	2016	Optical fiber, twinax and backplane 25 Gigabit Ethernet
802.3bz	2016	2.5GBASE-T and 5GBASE-T – 2.5 Gigabit and 5 Gigabit Ethernet over Cat-5/Cat-6 twisted pair
802.3bs	2017 (Dec.) (TBD)	200GbE over single-mode fiber and 400GbE over optical physical media
802.3bt	2017 (TBD)	Power over Ethernet enhancements up to 100 W using all 4 pairs balanced twisted-pair cabling, lower standby power and specific enhancements to support IoT applications (e.g. Lighting, sensors, building automation).
802.3cc	2017 (TBD)	25 Gbit/s over Single Mode Fiber
802.3cd	2018 (TBD)	Media Access Control Parameters for 50 Gbit/s and Physical Layers and Management Parameters for 50 Gbit/s, 100 Gbit/s, and 200 Gbit/s Operation
802.3ca	2019 (TBD)	100G-EPON – 25 Gbit/s, 50 Gbit/s, and 100 Gbit/s over Ethernet Passive Optical Networks

OSI NETWORK REFERENCE MODEL



IEEE 802 networks are at layers 1 & 2

IEEE 802 & OSI REFERENCE MODEL



IEEE 802 RM for end stations

ETHERNET DEVICES

Gibson Ethernet Electric Guitar

Pipe Organs

Refrigerator

Cisco Unified IP Phone 7975G

High-fidelity wideband audio & improved navigation options; Internet Low Bitrate Codec support for use in lossy networks Gigabit Ethernet connectivity

Pioneer Elite Receiver

Ethernet Office Lighting

Making sense: Sensors to the right of these LED lights keep an eye on motion, temperature and light levels, and include a backup infrared control. The red light is not a sensor, it's a status indicator

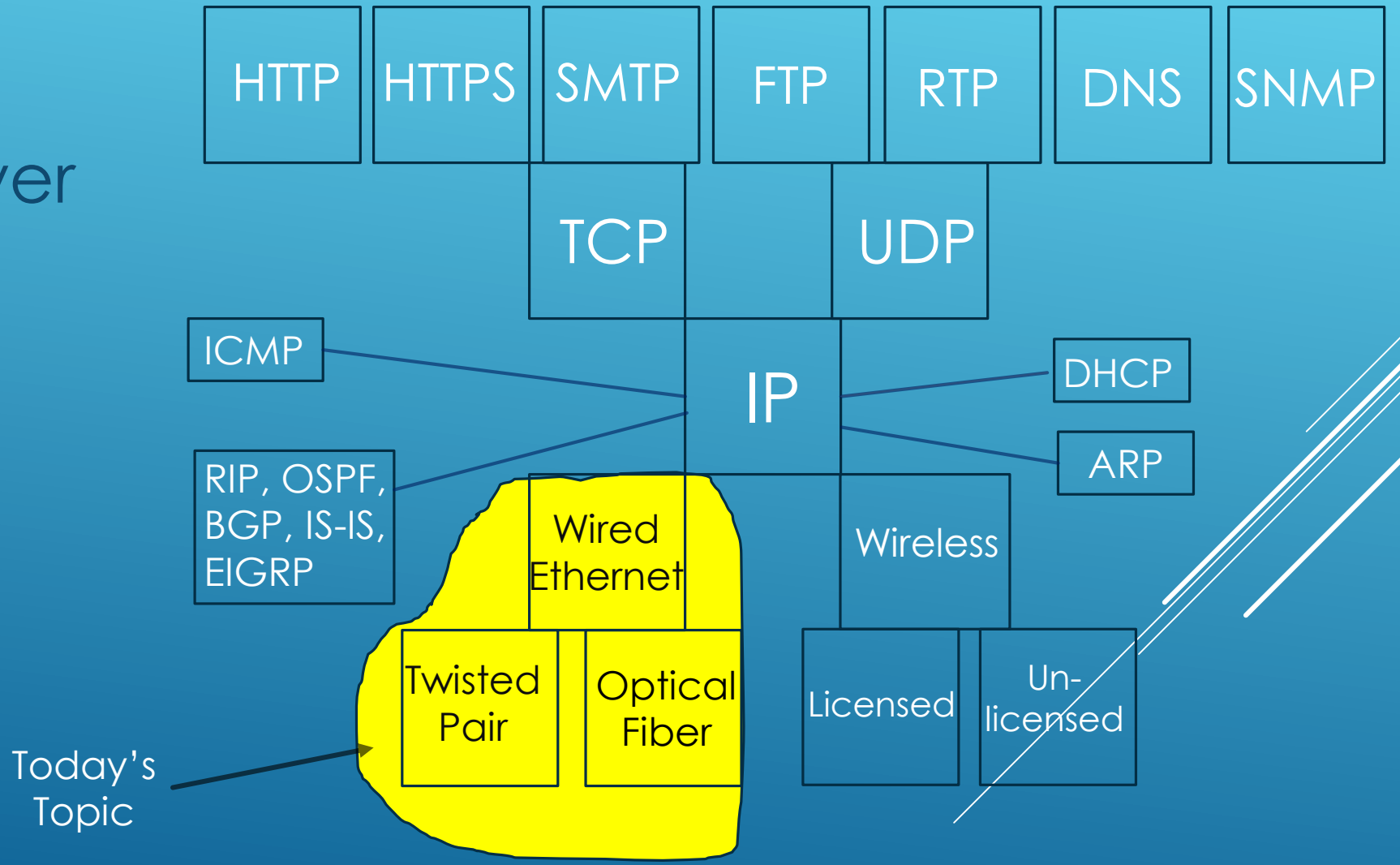
Video Surveillance

WHERE

- Used in devices intended to communicate over a wired standardized communication network
- IEEE 802 networks corresponds to Layers 1 & 2 of the OSI Network Protocol Reference Model
- Global (ubiquitous everywhere)
  - IEEE 802.3 Ethernet standard has been adopted internationally

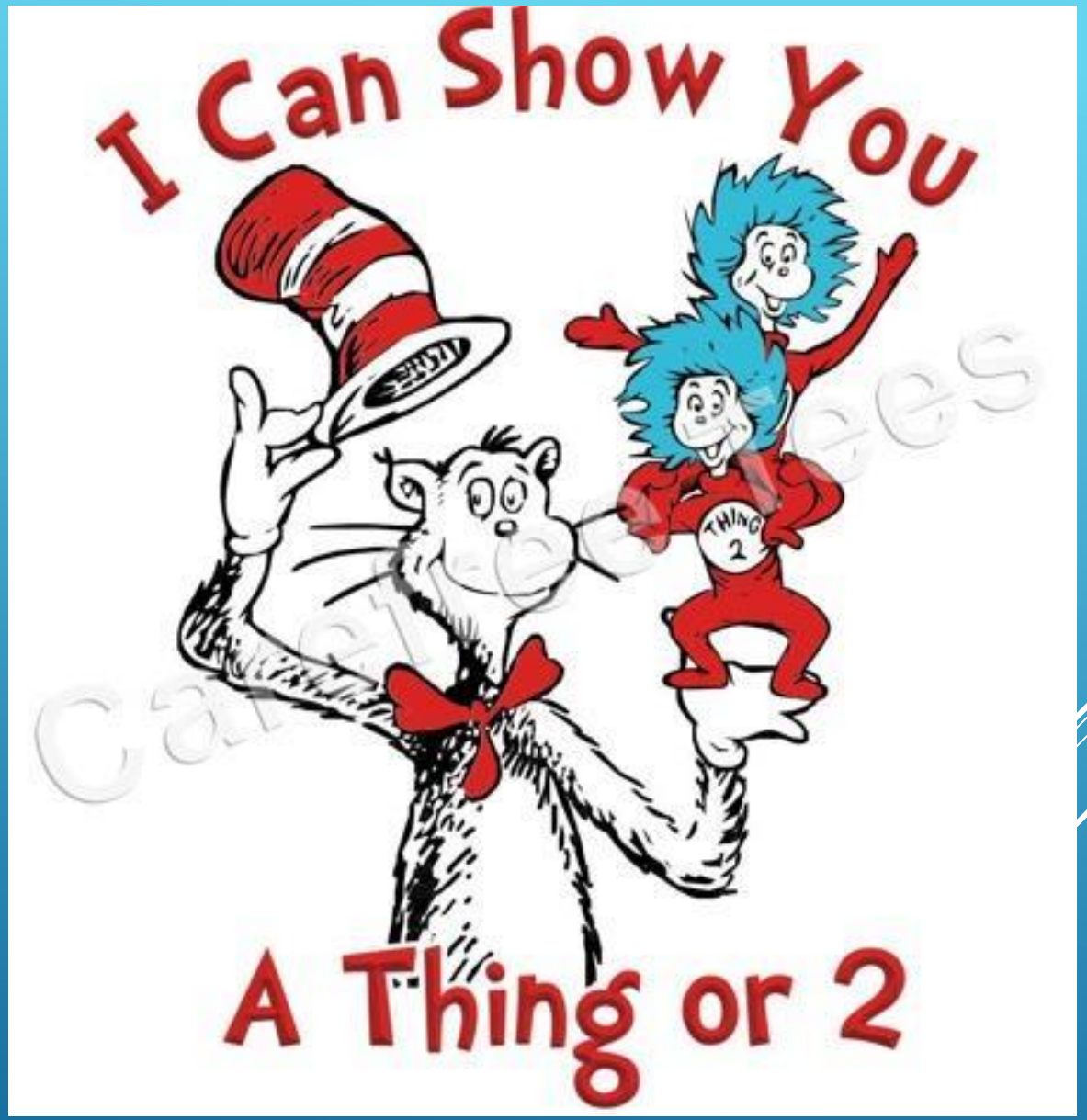
INTERNET LAYERED ARCHITECTURE

- Application Layer
- Transport Layer
- Network Layer
- Data Link Layer
- Physical Layer



HOW

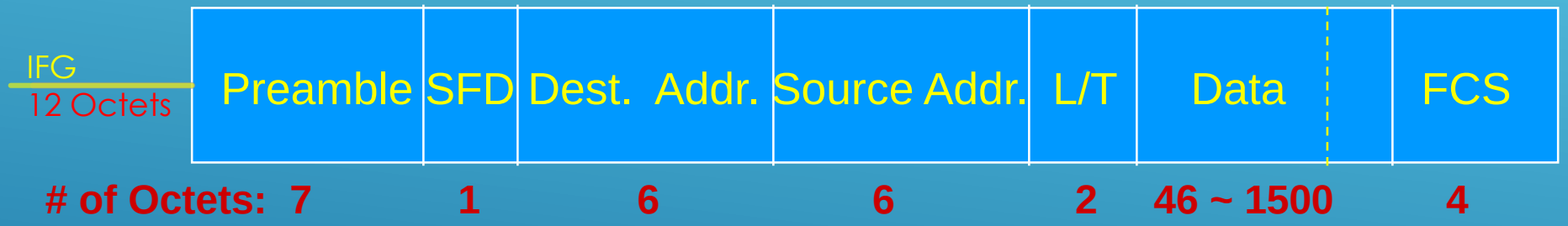
- Most of the rest of this presentation describes how IEEE 802.3 Ethernet networks work



BASIC 802.3 HEADER FIELDS

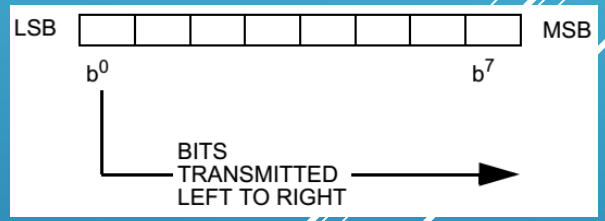
L/T - Length or Type Field
If ≤ 1500 = Length, 802.3
If ≥ 1536 = EtherType, Ethernet V.2

Padding (if required)



Value	Ethertype
0x0800	IPv4
0x0806	ARP
0x8100	802.1Q
0x86DD	IPv6

Each octet of a MAC frame, except the FCS, is transmitted least significant bit first.



- Preamble is 7 octets of 10101010 to synchronize receiver
- Start Frame Delimiter (SFD) is 10101011 to begin the frame
- Data field must be integral number of octets
- If Data is <46 octets, padding must be added. Padding is undefined.
  - For operating system developers: it's considered to be a security threat to send uninitialized padding data!

802 ADDRESS STRUCTURE

- 48 Bit (6 Octet) Standard MAC or Hardware Address
  - 1st Bit denotes Individual (0) or Group (1)
  - The 1st bit must be 0 for a source address
  - Group identifies more than one or all nodes
  - 2nd Bit denotes Global (0) or Local (1)
    - *Note that for the group address, this bit is also a 1*
  - Bits 3 to 24 comprise an IEEE assigned OUI if 2nd Bit is Global
  - Bits 25 to 48 comprise network hardware serial number if 2nd bit = 0
  - Bits 3 to 48 must be set to a unique number per network if 2nd bit = 1
  - Each octet of each address field shall be transmitted least significant bit first.
- Usually Written in Hexadecimal Notation as 6 Ordered Pairs
  - For Example: 00:00:69:3F:BD:05
- OUI: Organizational Unit Identifier assigned by IEEE to Hardware Suppliers

MULTICAST ADDRESSES

- Ethernet frames with a value of 1 in the least-significant bit of the first octet of the destination address are treated as multicast frames and are flooded to all points on the network.
- Frames with all ones in the destination are sometimes referred to as broadcasts
- Interfaces may join more than one group
- May not ever be used as a source address

ETHERNET MULTICAST ADDRESS EXAMPLES

Ethernet multicast address	EtherType Field	Usage
01-80-C2-00-00-00		Spanning Tree Protocol (for bridges) IEEE 802.1D
01-80-C2-00-00-00 or 01-80-C2-00-00-03 or 01-80-C2-00-00-0E	0x88cc	Link Layer Discovery Protocol (LLDP)
01-80-C2-00-00-08	0x0802	Spanning Tree Protocol (for provider bridges) IEEE 802.1ad
01-80-C2-00-00-01	0x8808	Ethernet flow control (Pause frame) IEEE 802.3x
01-80-C2-00-00-02	0x8809	Ethernet OAM Protocol IEEE 802.3ah (A.K.A. "slow protocols") Operations, Administration, and Maintenance
01-80-C2-00-00-30 - 01-80-C2-00-00-3F	0x8902	Ethernet CFM Protocol IEEE 802.1ag Connectivity Fault Management
01-00-5E-00-00-00 - 01-00-5E-7F-FF-FF	0x0800	IPv4 Multicast, insert the low 23 Bits of the multicast IPv4 Address into the Ethernet Address (RFC 7042)
33-33-xx-xx-xx-xx	0x86DD	IPv6 Multicast (RFC 2464), insert the low 32 Bits of the multicast IPv6 Address into the Ethernet Address (RFC 7042)
01-1B-19-00-00-00, or 01-80-C2-00-00-0E	0x88F7	Precision Time Protocol (PTP) version 2 over Ethernet

OBTAINING IEEE 802 ADDRESSES

- OUI & Ethertype assignments may be purchased from the IEEE Registration Authority
  - Registration Authority, IEEE Standards Department,
  - P.O. Box 1331,
  - 445 Hoes Lane,
    Piscataway, NJ 08855-1331, USA;
  - +1 732 562 3813; fax +1 732 562 1571.
- URL: http://standards.ieee.org/develop/regauth/

- Make up your own by setting global/local bit to 1
  - Be sure to keep each locally administered address unique within the network

IEEE 802.3 FCS

- 32 bit field used to detect frame transmission errors
- Checks all fields starting after the frame delimiter inclusive of the last data bit or pad bit in the data field
- FCSs detect all single bit errors, any odd number of errors, any errors of 32 bits or less, and most other multiple bit errors
- Probability of undetected errors is greater in longer frames
- Calculating FCS values consume most processing time for a protocol
- Upon reception, the FCS value is recalculated and compared to the value actually received
- If the recalculated FCS value is equal to the FCS value received, then the frame is considered error-free and passed up the stack for further processing
- If the recalculated FCS value differs from the FCS value received, then the frame is discarded

ETHERNET FCS (CRC) (CLAUSE 3.2.9)

A cyclic redundancy check (CRC) is used by the transmit and receive algorithms to generate a 32 bit CRC value for the FCS field. This value is computed as a function of the contents of the protected fields of the MAC frame: the Destination Address, Source Address, Length/Type field, MAC Client Data, and Pad (that is, all fields except FCS). The encoding is defined by the following generating polynomial:

$$G(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^{8} + x^{7} + x^{5} + x^{4} + x^{2} + x + 1$$

The CRC value corresponding to a given MAC frame is defined by the following procedure:

1. The first 32 bits of the frame are complemented.
2. The n bits of the protected fields are then considered to be the coefficients of a polynomial $M(x)$ of degree $n - 1$. (The first bit of the Destination Address field corresponds to the $x^{(n-1)}$ term and the last bit of the MAC Client Data field (or Pad field if present) corresponds to the x^0 term.)
3. $M(x)$ is multiplied by x^{32} and divided by $G(x)$, producing a remainder $R(x)$ of degree ≤ 31.
4. The coefficients of $R(x)$ are considered to be a 32-bit sequence.
5. The bit sequence is complemented and the result is the CRC.
6. The 32 bits of the CRC value are placed in the FCS field so that the x^{31} term is the left-most bit of the first octet, and the x^0 term is the right most bit of the last octet.

- The bits of the CRC are thus transmitted in the order x^{31}, x^{30},…, x^1, x^0.

JUMBO FRAMES

- Non-standard vendor specific implementations
- Started with iSCSI and remote disks. Now used for remote storage area networking
- A single *Etheresque* frame contains an entire contiguous disk block
  - Maximize performance
  - Minimize overhead
- Typical maximum jumbo frame size is 9216 octets
- But decreases "Fairness" – deferring other stations from using the network
- FCS becomes less effective at detecting possible errors
  - Probability of an undetected errored frame increases
- Should only be used locally or only over non error-prone circuits
- Audio-Video Network Domains are considered incompatible with jumbo frames (802.1ba Section 6.3)

INVALID MAC FRAMES

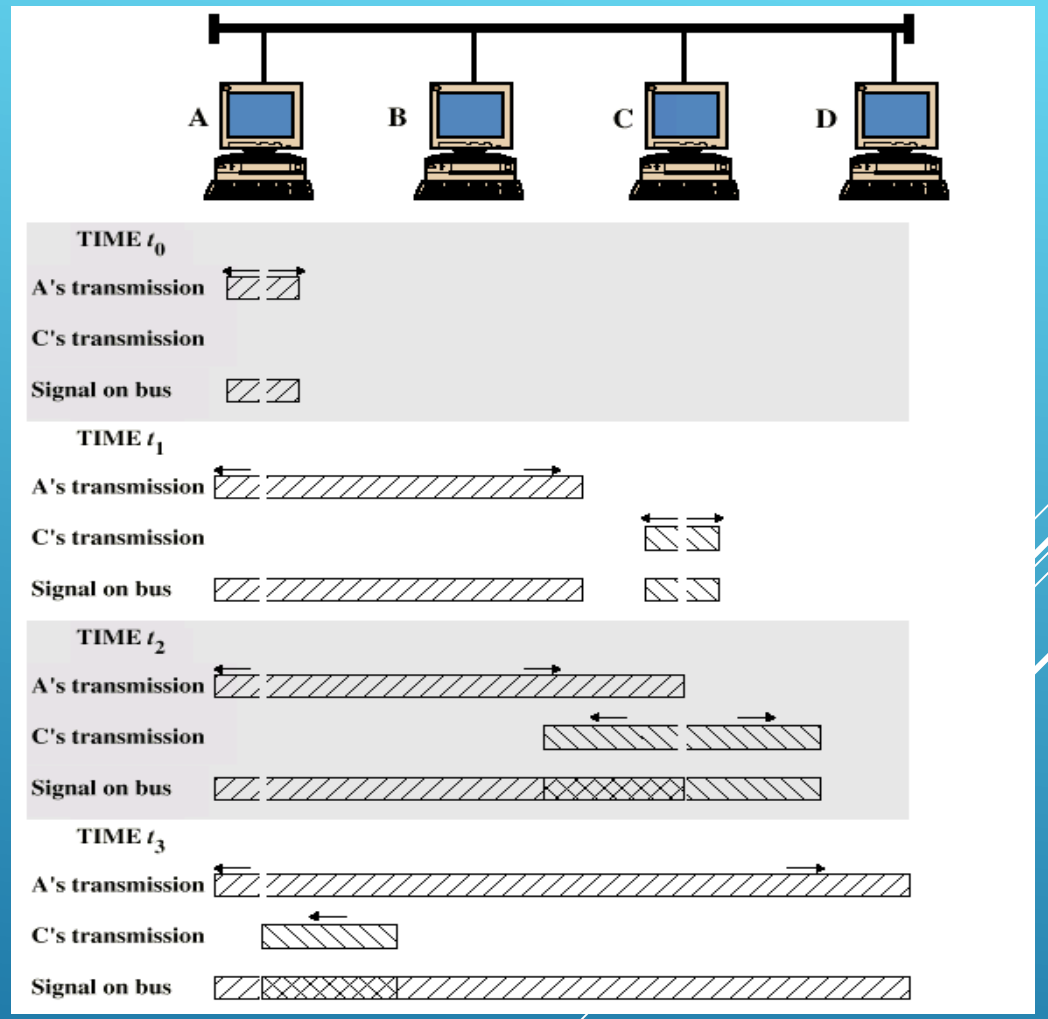
- An invalid MAC frame shall be defined as one that meets at least one of the following conditions:
  - The frame length is inconsistent with a length value if specified in the length/type field.
    - If the length/type field contains a type value, then the frame length is assumed to be consistent with this field and should not be considered an invalid frame on this basis.
  - It is not an integral number of octets in length. (frame alignment error)
  - The bits of the incoming frame (exclusive of the FCS field itself) do not generate a CRC value identical to the one received.
  - The contents of invalid MAC frames shall not be passed to the LLC or MAC Control sublayers.
  - The occurrence of invalid MAC frames may be communicated to network management.
- A runt frame is an Ethernet frame that is less than the IEEE 802.3's minimum length of 64 octets. Runt frames are most commonly caused by collisions; other possible causes are a malfunctioning network card, buffer underrun, or software issues

OPERATIONAL MODES

- In half duplex mode, stations contend for the use of the physical medium, using the CSMA/CD algorithms. Bidirectional communication is accomplished by rapid exchange of frames, rather than full duplex operation. Half duplex operation is possible on all supported media; it is required on those media that are incapable of supporting simultaneous transmission and reception without interference.
- The full duplex mode of operation can be used when all of the following are true:
  1. The physical medium is capable of supporting simultaneous transmission and reception without interference (e.g., 10BASE-T, 10BASE-FL, and 100BASE-T).
  2. There are exactly two stations on the LAN. This allows the physical medium to be treated as a full duplex point-to-point link between the stations. Since there is no contention for use of a shared medium, the multiple access (i.e., CSMA/CD) algorithms are unnecessary.
  3. Both stations on the LAN are capable of and have been configured to use full duplex operation.
- The most common configuration for full duplex operation consists of a central bridge (switch) with a dedicated LAN connecting each bridge port to a single device interface.

CSMA/CD OPERATION

- Propagation time is much less than transmission time
- All stations know that a transmission has started almost immediately
- First listen for clear medium (carrier sense)
- If medium idle, transmit
- If two stations start at the same instant, collision
- Wait reasonable time (round trip plus ACK contention)
- No ACK then retransmit
- Max utilization depends on propagation time (medium length) and frame length
  - Longer frame and shorter propagation gives higher utilization
  - Shorter frames promote "fairness"
    - "Fairness" is the number of transmit opportunities / time
- With CSMA, collision occupies medium for duration of transmission. With CSMA/CD:
- Stations listen whilst transmitting
- If medium idle, transmit
- If busy, listen for idle, then transmit
- If collision detected, jam then cease transmission
- After jam, wait random time then start again
  - Binary exponential back off



VLANS 802.1Q-9

- VLAN technology functions by logically segmenting the network into different *broadcast domains* so that packets are only switched between ports that are designated for the same VLAN.
- This approach also improves scalability, particularly in LAN environments that support broadcast- or multicast-intensive protocols and applications that flood packets throughout the network
- Regardless of physical location within a campus or of an interface type, managers can define workgroups based on logical function rather than physical location through simple port configuration. Using switches and routers that have embedded VLAN intelligence obviates the need for recabling to extend connectivity in switched LAN environments.
- VLAN tags include priority bits that indicate a specific class of service
- May be recursive (VLAN frame encapsulated within a VLAN frame)
- If frame is recursive, then maximum length may be extended to 2kB. User data field maximum length still maintained at 1.5kB as per (802.3as)

ETHERNET FRAME W/ VLAN TAG

Drop eligible indicator *(DEI)*: a 1-bit field. May be used separately or in conjunction with priority bits to indicate frames eligible to be dropped in the presence of congestion.

VLAN values of 0x000 and 0xFFF are reserved

VLAN 0x001 reserved by most vendors for management VLAN

Priority	Description	Target Time
0	Best Effort (default)	
1	Background	
2	Excellent Effort	
3	Critical Application	
4	Video	< 100 mS
5	Voice	< 10 mS
6	Internetwork Control	
7	Network Control	

These values may be used to prioritize different classes of traffic.

VLAN tags increase the maximum frame length to 1522 octets

VLAN Protocol ID 0x8100

VLAN tag fields

Tag Control Info

Padding (if required)

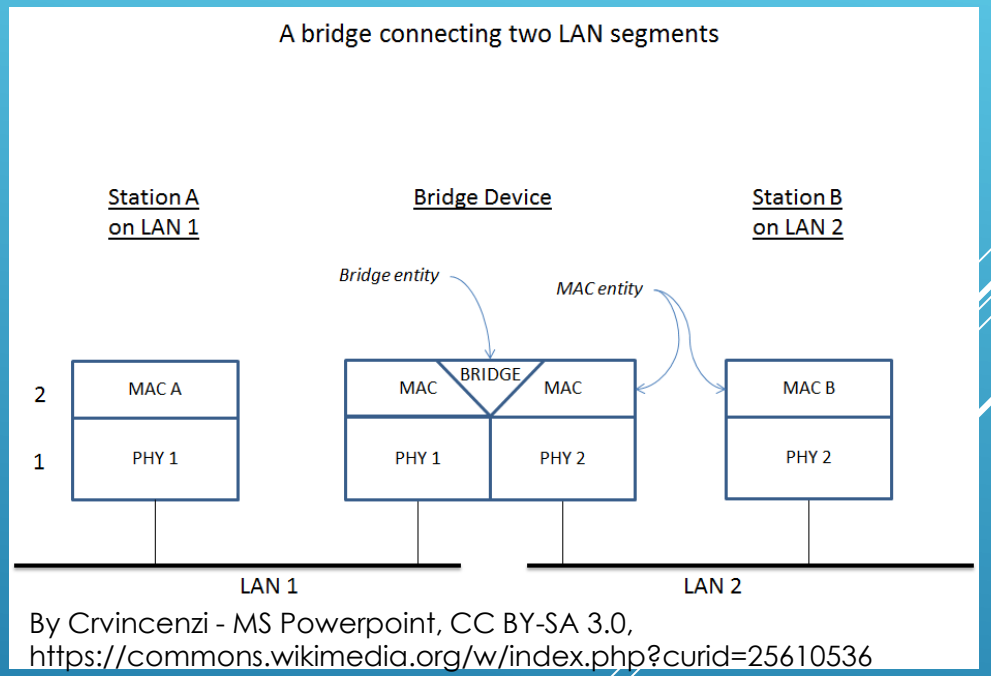
Start Frame Delimiter

Length or Type Field
If ≤ 1500 = Length, 802.3
If ≥ 1536 = Type, Ethernet V.2

Frame check sequence

BRIDGING 802.1Q-8

- A bridge connects two network segments, by deciding on a frame-by-frame basis whether or not to forward from one network to the other.
- A store and forward technique is used such that during forwarding the frame integrity is verified on the source network and CSMA/CD delays are accommodated on the destination network.
- Fast switching bridges forward frames immediately after receiving the destination address thus improving latency
- Contrary to repeaters that simply forward all frames, bridges only forward frames that are required to cross the bridge.
- Additionally, bridges reduce collisions by partitioning the collision domain.
- The multiport bridge function serves as the basis for network switches.



By Crvincenzi - MS Powerpoint, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=25610536

TRANSPARENT LEARNING BRIDGE

- A transparent bridge uses a forwarding database to send frames across network segments. The forwarding database starts empty - entries in the database are built as the bridge receives frames. If an address entry is not found in the forwarding database, the frame is flooded to all other ports of the bridge, flooding the frame to all segments except the one from which it was received. By means of these flooded frames, the destination network will respond and a forwarding database entry will be created.
- In the context of a two-port bridge, one can think of the forwarding database as a filtering database. A bridge reads a frame's destination address and decides to either forward or filter. If the bridge determines that the destination node is on another segment on the network, it forwards (transmits) the frame to that segment. If the destination address belongs to the same segment as the source address, the bridge filters (discards) the frame. As nodes transmit data through the bridge, the bridge establishes a filtering database of known MAC addresses and their locations on the network. The bridge uses its filtering database to determine whether a frame should be forwarded or filtered.
- Transparent bridging can also operate over devices with more than two ports. As an example, consider a bridge connected to three hosts, A, B, and C. The bridge has three ports. A is connected to bridge port 1, B is connected to bridge port 2, C is connected to bridge port 3. A sends a frame addressed to B to the bridge. The bridge examines the source address of the frame and creates an address and port number entry for A in its forwarding table. The bridge examines the destination address of the frame and does not find it in its forwarding table so it floods it to all other ports: 2 and 3. The frame is received by hosts B and C. Host C examines the destination address and ignores the frame. Host B recognizes a destination address match and generates a response to A. On the return path, the bridge adds an address and port number entry for B to its forwarding table. The bridge already has A's address in its forwarding table so it forwards the response only to port 1. Host C or any other hosts on port 3 are not burdened with the response. Two-way communication is now possible between A and B without any further flooding in network.
- Both source and destination addresses are used in this algorithm: source addresses are recorded in entries in the table, while destination addresses are looked up in the table and matched to the proper segment to send the frame to.
- Forwarding and filtering tables are usually ephemeral; they need to "age out" over time to allow for network changes. Aging typically occurs over several minutes of elapsed time. Administrative filters for security and management are not subject to aging.

SPANNING TREE PROTOCOL 802.1Q-13

- Spanning Tree Protocol (STP) is a network protocol that builds a logical loop-free topology for Ethernet networks.
- STP prevents bridge loops and the broadcast storms that results from them.
- Spanning tree allows network designs to include spare (redundant) links to provide automatic backup paths if an active link fails. This is done without bridge loops, or manual intervention.

HOW STP WORKS

- Lowest root bridge ID - Determines the root bridge
- Lowest cost path to the root bridge - Favors the upstream switch with the least cost to root
- Lowest sender bridge ID - Serves as a tie breaker if multiple upstream switches have equal cost to root
- Lowest sender port ID - Serves as a tie breaker if a switch has multiple links to a single upstream switch, where:
  - Bridge ID = priority (4 bits) + locally assigned system ID extension (12 bits) + ID [MAC address] (48 bits); the default bridge priority is 32768
  - Port ID = priority (4 bits) + ID (Interface number) (12 bits); the default port priority is 128.
- The access speeds of the links determine the path cost that STP assumes calculated by the formula: 20 Tbps/link_data_rate

STP BPDU

- BPDU – Bridge Port Data Unit exchange information about bridge IDs and root path costs
- Every 2 seconds a bridge sends a BPDU frame using the unique MAC address of the port itself as a source address, and a destination address of the STP multicast address 01:80:C2:00:00:00.
- STP switch port states:
- Blocking - A port that would cause a switching loop if it were active. No user data is sent or received over a blocking port, but it may go into forwarding mode if the other links in use fail and the spanning tree algorithm determines the port may transition to the forwarding state. BPDU data is still received in blocking state. Prevents the use of looped paths.
- Listening - The switch processes BPDUs and awaits possible new information that would cause it to return to the blocking state. It does not populate the MAC address table and it does not forward frames.
- Learning - While the port does not yet forward frames it does learn source addresses from frames received and adds them to the filtering database (switching database). It populates the MAC address table, but does not forward frames.
- Forwarding - A port receiving and sending data, normal operation. STP still monitors incoming BPDUs that would indicate it should return to the blocking state to prevent a loop.
- Disabled - Not strictly part of STP, a network administrator can manually disable a port

MULTIPLE REGISTRATION PROTOCOL 802.1Q-10

- Generic framework allowing bridges to register and de-register attribute values, like VLAN identifiers and multicast group membership. It defines the architecture, rules of operation, state machines and variables for the registration and de-registration of attribute values.
- Multiple MAC Registration Protocol is a data link layer protocol to register multicast MAC addresses on multiple switches. It is an MRP application, included in 802.1Q.
- The purpose of MMRP is to allow multicast traffic in bridged LANs to be confined (pruned) to areas of the network where it is required.
- Group membership information frame. This indicates the presence of MMRP participants that are members of a particular Group(s), and carries the group MAC address(es) associated with the Group(s). The exchange of Group membership information results in the creation or updates of MAC Address Registration Entries in the FDB to indicate the Port(s) and VID(s) of the VLAN(s) that members of the Group(s) have registered.

PAUSE & PRIORITY FLOW CONTROL (PFC) FRAMES

- The optional PAUSE operation is used to inhibit transmission of data frames for a specified period of time. (802.3 clause 31)
- A MAC Control client wishing to inhibit transmission of data frames from another station on the network generates a MA_CONTROL.request primitive specifying:
  1. The globally assigned 48-bit multicast address 01-80-C2-00-00-01,
  2. The PAUSE opcode,
  3. A request_operand indicating the length of time for which it wishes to inhibit data frame transmission in the form of two byte unsigned integer. This number is the requested duration of the pause. The pause time is measured in units of pause "quanta", where each unit is equal to 512 bit times.
- The PAUSE operation cannot be used to inhibit transmission of MAC Control frames.
- PAUSE frames shall only be sent by DTEs configured to the full duplex mode of operation.
- Bridges will not forward frames sent to the PAUSE multicast destination address
- PFC Pauses are applied separately for each class of service
- Pause operates only at the local link layer; it is Not an END-to-END service

ETHERNET MANAGEMENT (CLAUSES 30, 57) (802.3.1)

- Ethernet Standard 802.3.1 defines MIB Management Information Base (MIB) module specifications for IEEE Std 802.3.
- It includes the Structure of Management Information Version 2 (SMIv2) MIB module specifications formerly published by the Internet Engineering Task Force (IETF), and the managed object branch and leaf assignments provided in the MIB modules, as well as extensions resulting from recent amendments to IEEE Std 802.3.
- The SMIv2 MIB modules are intended for use with the Simple Network Management Protocol (SNMP), commonly used to manage networks.
- Ethernet Operation, Administration, and Maintenance (OAM) is composed of a core set of functions and a set of optional functional groups. OAM provides network operators the ability to monitor the health of the network and quickly determine the location of failing links or fault conditions.
- The core functions include discovery operations (determining if the other end of the link is OAM capable and what OAM functions it supports), state machine implementation, and some critical event flows including:
  - Remote fault indication
  - Link monitoring
  - Remote loopback
- IEEE 802.3 OAM does not include encryption or authentication mechanisms.

ETHERNET MANAGEMENT YANG MODELS P802.3CF

- This project will develop YANG models from Clause 30 objects in IEEE Std 802.3-2015 and published amendments to enable NETCONF (RFC6241) management of IEEE Std 802.3 Ethernet.
- The proposed standard will use the NETCONF protocol (RFC6241) and the YANG data modeling language (RFC6020, RFC7950).
- The NETCONF protocol uses a remote procedure call (RPC) paradigm. A client encodes an RPC in XML [W3C.REC-xml-20001006] and sends it to a server using a secure, connection-oriented session. The server responds with a reply encoded in XML. The contents of both the request and the response are fully described in XML DTDs or XML schemas, or both, allowing both parties to recognize the syntax constraints imposed on the exchange.
- The NETCONF protocol is a building block in a system of automated configuration. NETCONF can be used in concert with XML-based transformation technologies, such as XSLT [W3C.REC-xslt-19991116], to provide a system for automated generation of full and partial configurations. The system can query one or more databases for data about networking topologies, links, policies, customers, and services.
- The base protocol includes the following protocol operations:

1. get
2. get-config
3. edit-config
4. copy-config
5. delete-config
6. lock
7. unlock
8. close-session
9. kill-session

SELF-SIMILAR NATURE OF ETHERNET TRAFFIC

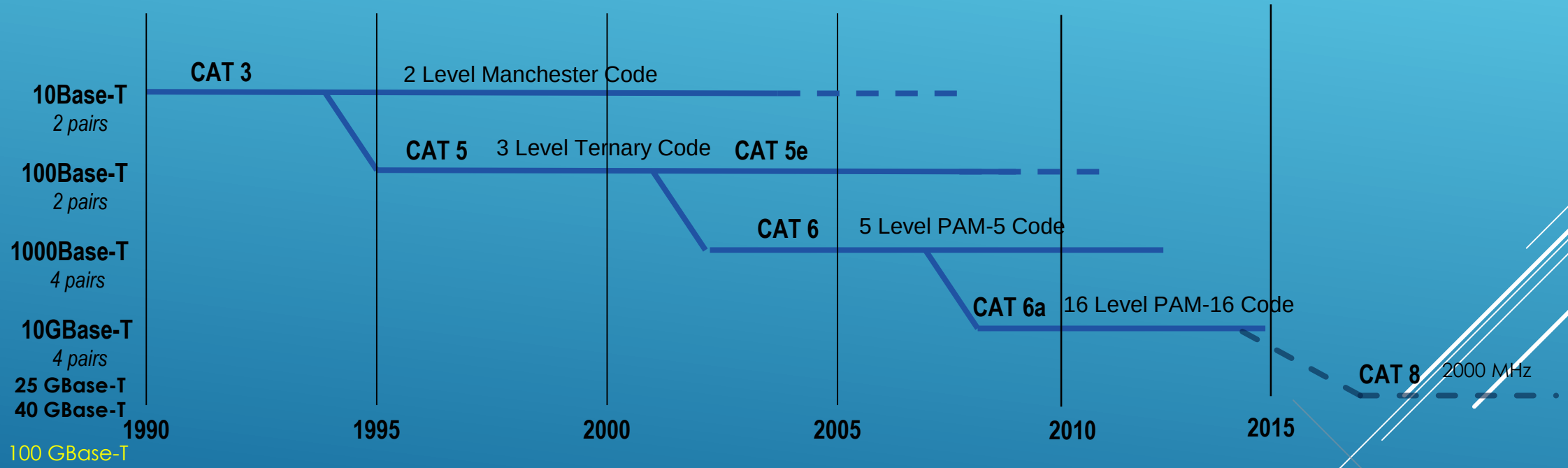
- Ethernet LAN traffic is statistically self-similar; none of the commonly used traffic models are able to capture its fractal-like behavior; that such behavior has serious implications for the design, control, and analysis of high-speed, cell-based networks; and that aggregating streams of such traffic intensifies the self-similarity ("burstiness") instead of smoothing it.
- This is supported by a rigorous statistical analysis of hundreds of millions of high quality Ethernet traffic measurements, coupled with a discussion of the underlying mathematical and statistical properties of self-similarity and their relationship with actual network behavior.
- The analysis of the Ethernet data shows that the generally accepted argument for the "Poisson-like" nature of aggregated traffic, namely, that it becomes smoother (less bursty) as the number of traffic sources increases, has very little to do with reality. Using the degree of self-similarity (which typically depends on the utilization level of the Ethernet and can be defined via the Hurst parameter) as a measure of "burstiness": the burstiness of LAN traffic typically intensifies as the number of active traffic sources increases, contrary to commonly held views.
- The impact of the self-similar nature of packet traffic for high-speed networks are already ample:
    1. Source models for individual users show extreme variability in terms of interarrival times of packets (Le., the infinite variance syndrome),
    2. Commonly used measures for "burstiness" such as the index of dispersion (for counts), the peak-to-mean-ratio, or the coefficient of variation (for interarrival times) are no longer meaningful for self-similar traffic but can be replaced by the Hurst parameter,
    3. The nature of congestion produced by self-similar network traffic models differs drastically from that predicted by standard formal Poisson models and displays a far more complicated picture than has been assumed in the past
    4. First analytic results show a clear distinction between predicted performance of certain queueing models with traditional input streams and the same queueing models with self-similar inputs.
- Finally, in light of the same fractal-like behavior recently observed in Variable Bit Rate video traffic and the more complicated nature of congestion due to the self-similar traffic behavior can be expected to persist even when we move toward a more heterogeneous environment. Thus, we believe based on our measured traffic data that the success or failure of, for example, a proposed congestion control scheme will depend on how well it performs under a self-similar rather than under one of the standard formal traffic scenarios.
- From Paper:
    - IEEWACM TRANSACTIONS ON NETWORKING, VOL. 2, NO. 1. FEBRUARY 1994
    - *On the Self-similar Nature of Ethernet Traffic (Extended Version)*
    - Will E. Leland, Member, IEEE, Murad S. Taqqu, Member, IEEE, Walter Willinger, and Daniel V. Wilson, Member, IEEE

ETHERNET MEDIA

- Media Dependent Layer
  - Data Rate
  - Encoding & Scrambling
  - Connectors
- Copper
  - UTP – Unshielded Twisted Pair
  - STP – Shielded Twisted Pair
  - Twinax
  - Ethernet in the First Mile (EFM)
  - Coaxial Cable (obsolete)
  - Electrical Backplanes
    - Backplane Ethernet is primarily intended to operate over differential, controlled impedance traces up to 1 m, including two connectors, on printed circuit boards residing in a backplane environment.
- Fiber
  - SMF - Single Mode Fiber
  - MMF – Multi Mode Fiber
  - WDM – Wavelength Division Multiplexing
  - CWDM – Coarse Wavelength Division Multiplexing
  - DWDM – Dense Wavelength Division Multiplexing

TWISTED PAIR AND ETHERNET TIMELINE

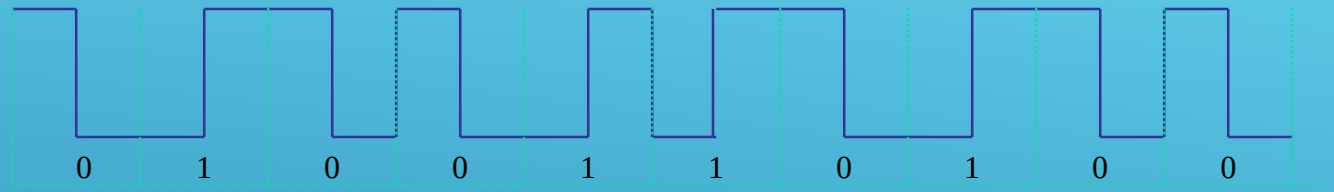
- Ethernet advances have driven the demand for higher quality cabling systems
- Frequencies on the cable have increased; signal level step sizes have decreased
- Higher speed operation is more susceptible to attenuation and noise



Before 1990, Ethernet used coaxial cable in a bus topology, half-duplex
After 1990, Ethernet used UTP in a star topology, full duplex

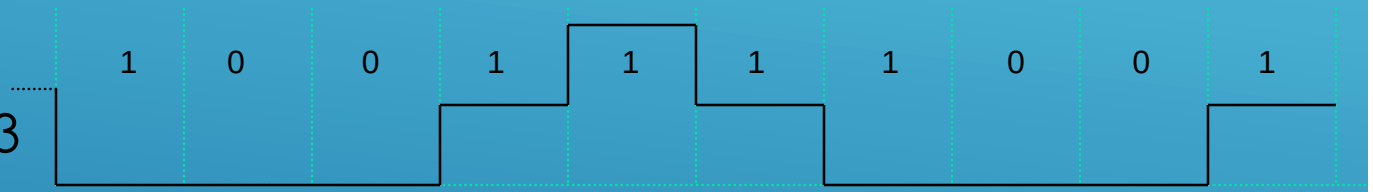
ETHERNET ENCODING FOR UTP PMD

- 10Base-T
  - 10 Mbps
  - Manchester

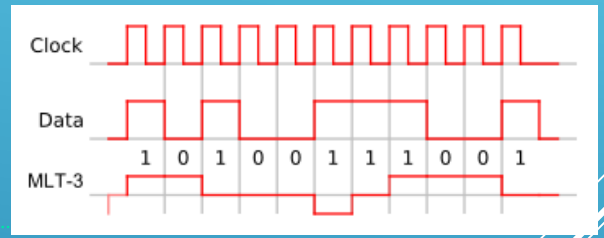


Manchester Code: 100 nS Bit Time Interval, Transitions MUST occur in middle of bit time intervals
Rising edge in middle = 1, Falling edge in middle = 0

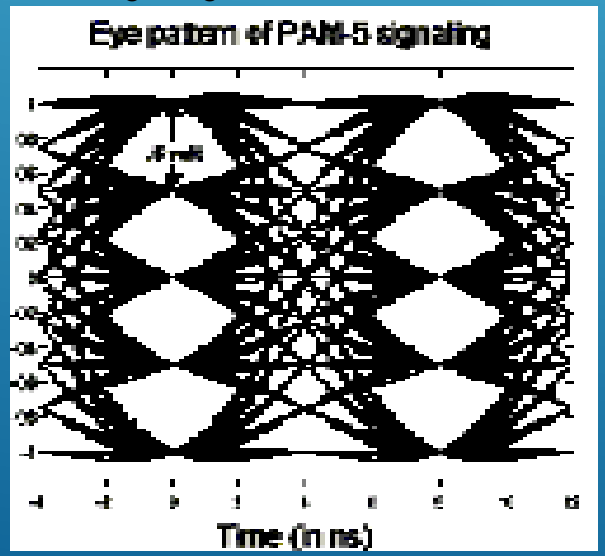
- 100Base-T
  - 100 Mbps
  - 4B/5B, MLT-3



MLT-3 Code: 8.0 nS Bit Time Interval, If next bit is 0, then no Δ in output level; if next bit is 1, then transition to 0v if previous level was either ±V; if proceeding level was 0v, then go to the level of ±V opposite that of the previous ±V level, transitions MUST occur at beginning of bit time intervals



- 1000Base-T
  - 1000 Mbps
  - 4D-PAM-5 Encoding (5 level, quinary)



5 voltage levels give $5^4 = 625$ possible code words. 256 are data words, an additional 256 code words represent the same data words (providing 100% redundancy). The remaining 113 words are used for control and idle signals. Each 8-bit word is coded as a four- dimensional vector of quinary symbols spaced by a time interval of 8 nanoseconds (ns). These symbols are selected from the set {-2, -1, 0, +1, +2}.

- 10GBase-T
  - 10,000 Mbps
  - 64B/65B, PAM-16 Encoding, (16 level)

4B5B ENCODING SYMBOL TABLE

Data	Symbol	Data	Symbol	Control	Symbol
0000	11110	1000	10010	Quiet (Q)	00000
0001	01001	1001	10011	Idle (I)	11111
0010	10100	1010	10110	Halt (H)	00100
0011	10101	1011	10111	Start (1) Delimiter (J)	11000
0100	01010	1100	11010	Start (2) Delimiter (K)	10001
0101	01011	1101	11011	End (T) Delimiter	01101
0110	01110	1110	11100	Reset (R)	00111
0111	01111	1111	11101	Set (S)	11001

8B10B ENCODING

- 8B10B transmission code provides the following functions:
  - Improves transmission characteristics
  - Enables bit- level clock recovery
  - Improves error detection
  - Separates data symbols from control symbols
  - Derives bit and word synchronization

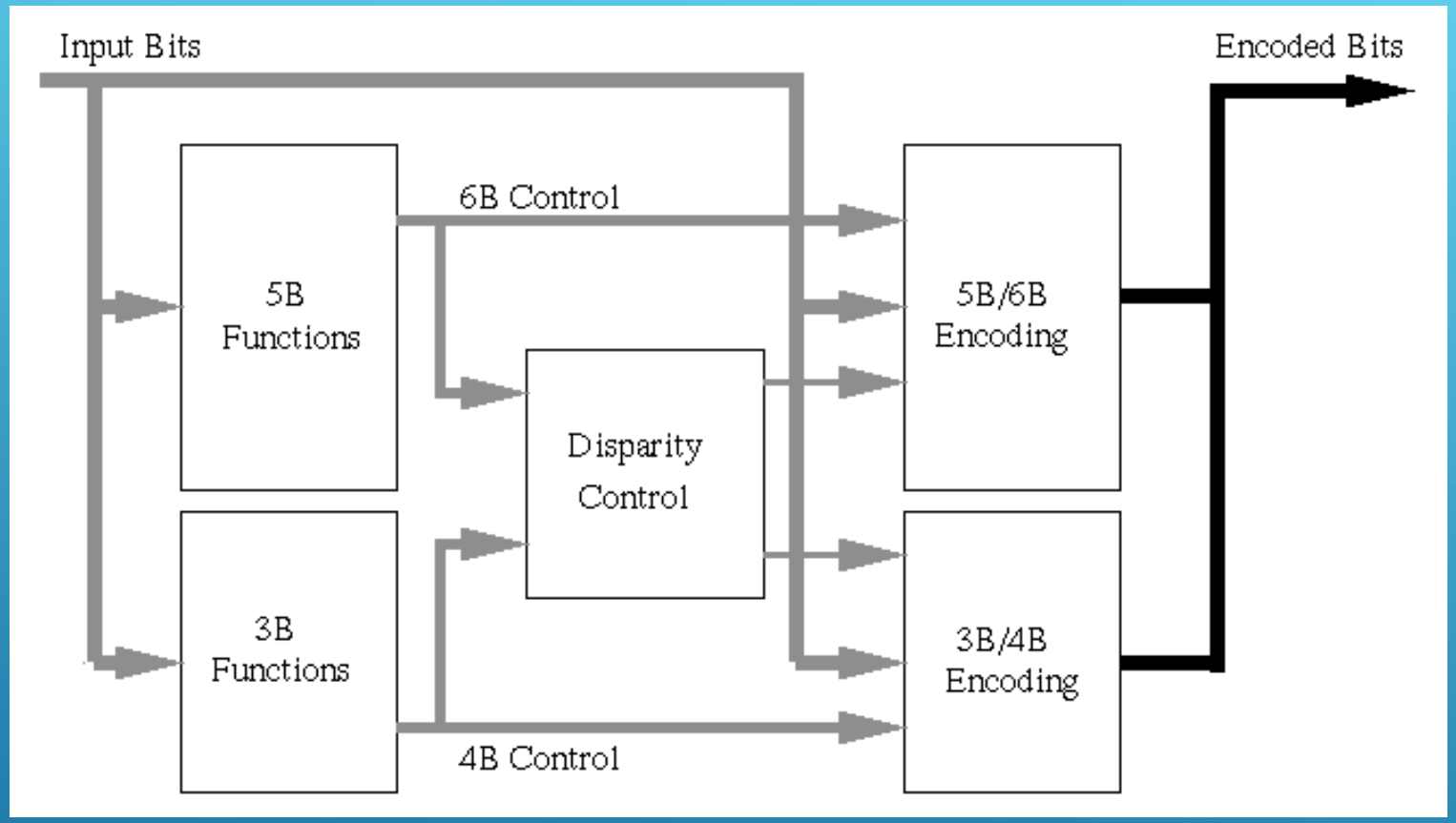
Gigabit Ethernet uses 8B10B encoding

- The data bytes are encoded into 10-bit data characters resulting into 1024 possible characters. 2x256=512 are reserved for the data byte transfers. One character representative has more 1's, the other has more 0's and are selected according to the current disparity *(see below)*. 12 special characters are defined for special signaling. The rest of the 1024-512-12 are not allowed for transmission and indicate transmission errors or unsynchronized status once they are received at the destination. Ordered sets are flexible building blocks which may be used for in-band and or out-of-band protocol functions.
- 8B10B code recognizes the idea of a Running Disparity (the difference between the number of 1's and 0's transmitted). The sender keeps the running disparity around zero, the receiver checks the data stream according to this rules and is thus able to detect some transmission errors and maintain DC balance.

8B10B DETAILS

- An unencoded information byte is composed of eight information bits A, B, C, D, E, F, G, H and the control variable Z. This information is encoded into the bits a, b, c, d, e, f, g, h, I, j of a 10-bit Transmission Character. The control variable has either the value D (D-type) for Data characters or the value K (K-type) for special characters. Each valid Transmission Character has been given a name using the following convention: Zxx.y, where Z is the control variable of the unencoded information byte, xx is the decimal value of the binary number composed of the bits E, D, C, B, and A, and y is the decimal value of the binary number composed of the bits H, G of the unencoded information byte in that order. For example the name of the Transmission Character composed of the hexadecimal "BC" special (K-type) code is K28.5.
- The information received is recovered 10 bits at a time and those Transmission Characters used for data (D-type) are decoded into the one of the 256 8-bit combinations. Some of the remaining Transmission Characters (K-type) referred to as special characters, are used for protocol management functions. Codes detected at the receiver that are not D- or K- type are signaled as code violation errors.
- Each data byte or special character has two (not necessarily different) transmission codes. The data bytes and special characters are encoded into these codes respectively, depending on the initial Running Disparity (RD). The RD is a binary parameter, which is calculated upon the balance of ones and zeros in the sub-blocks (the first six bits and the last four bits) of a transmission character. A new RD is calculated from the transmitted character at both the transmitter and the receiver. If the detected character has opposite RD the transmitter should have sent, (depending on the RD of the previous bit stream) the receiver indicates a disparity violation condition. A Transmission Word is composed of four contiguous transmission characters.

8B10B ENCODING PROCESS



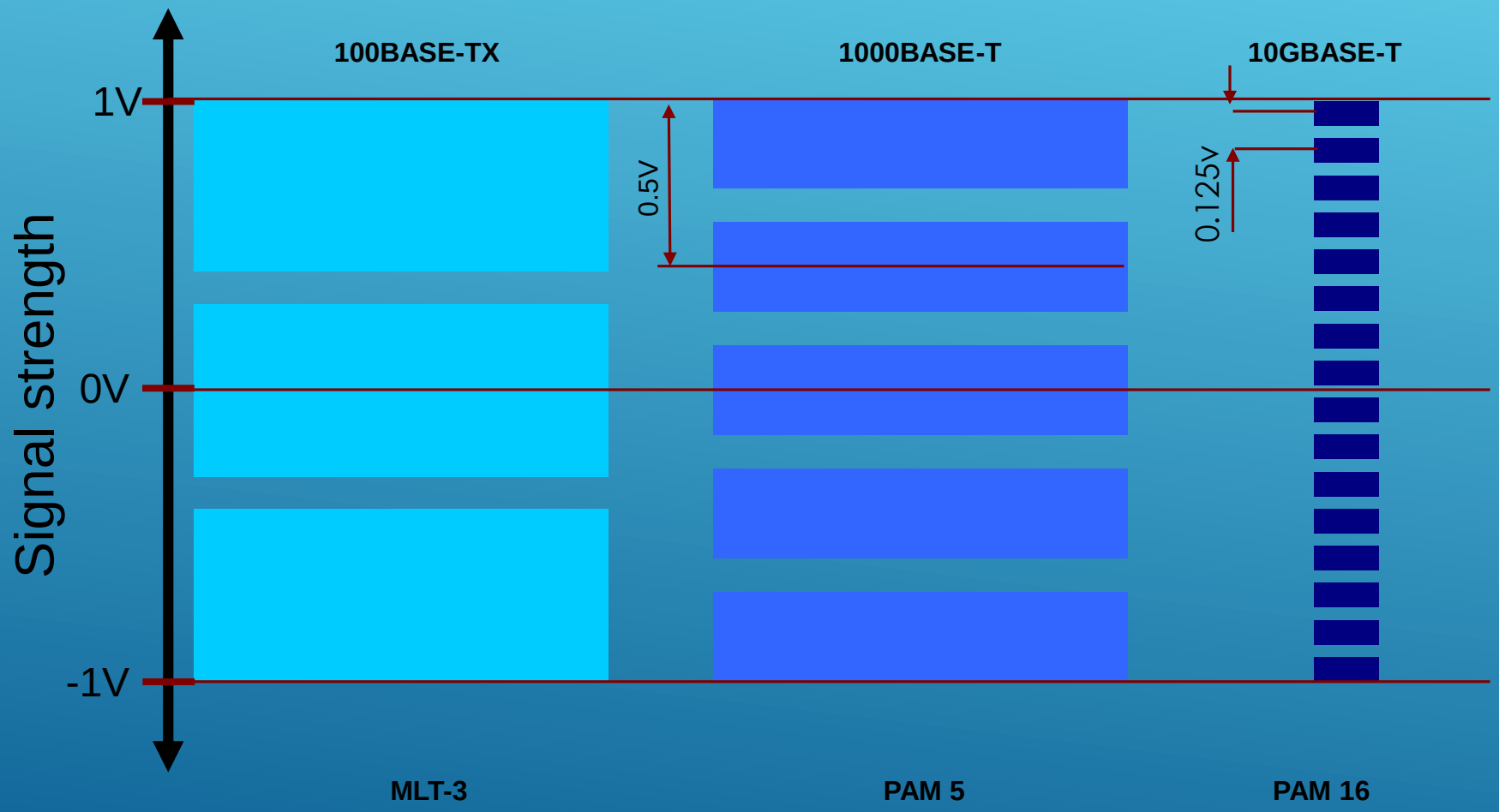
Note that 8b/10b is the encoding scheme, not a specific code. While many applications do use the same code, there exist some incompatible implementations; for example, Transition Minimized Differential Signaling, which also expands 8 bits to 10 bits, but it uses a completely different method to do so.

- 8B10B guarantees a maximum run length of 5 bits
- The lowest transition density that can be indefinitely maintained under the encoding scheme is 30 transitions per 100 bits
- It can detect all single-bit and many other errors
- It contains three different comma characters.
- DC-balanced property: it generates a bit stream with a balanced number of '1' and '0' bits.

10G ETHERNET ENCODING

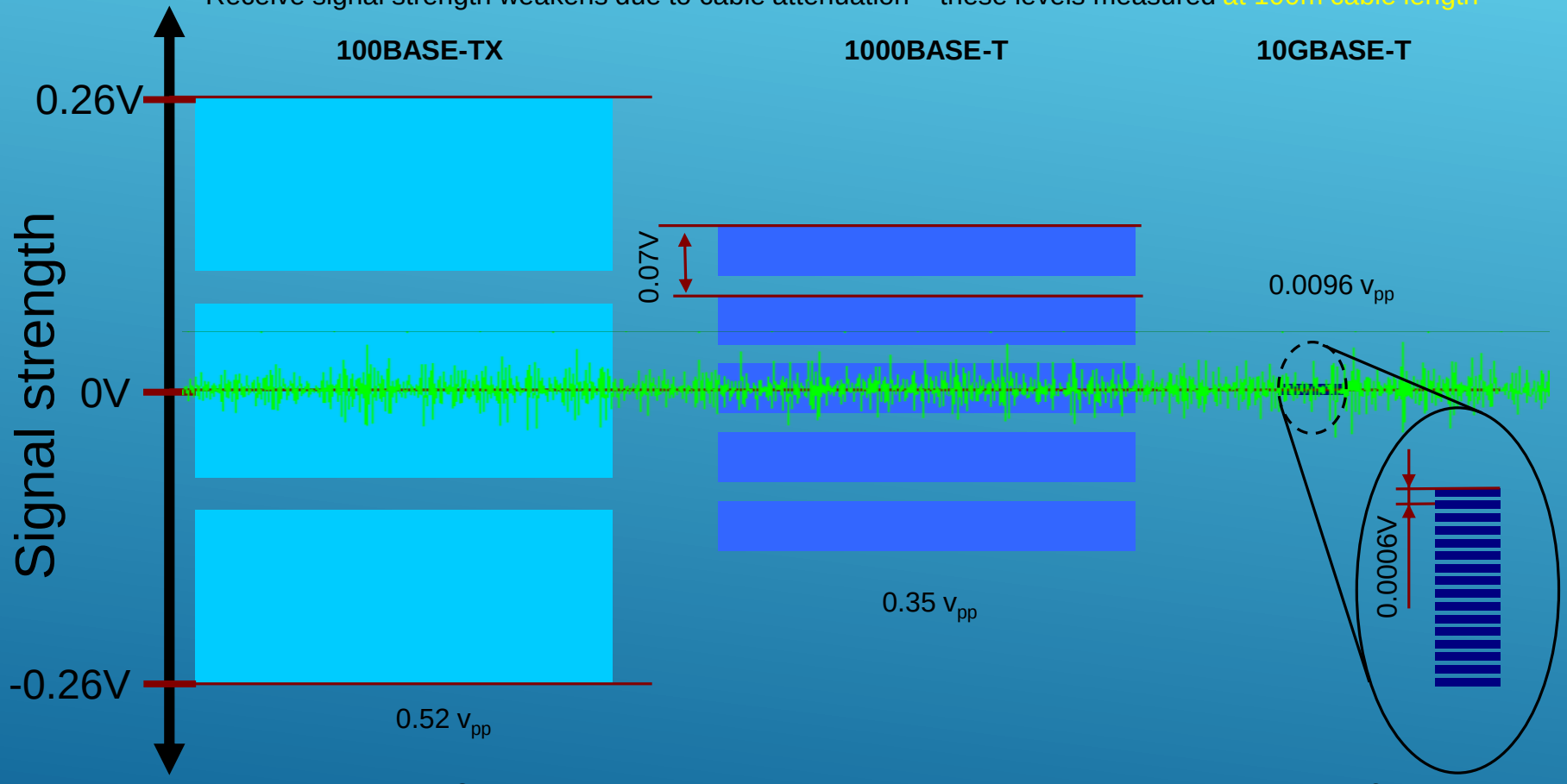
- The aggregate data rate of 10 Gb/s is achieved by transmitting 2500 Mb/s in each direction simultaneously on each wire pair. Baseband 16-level PAM signaling with a modulation rate of 800 Megasymbol per second is used on each of the wire pairs.
- Ethernet data and control characters are encoded at a rate of 3.125 information bits per PAM16 symbol, along with auxiliary channel bits in a 64b65b configuration. Two consecutively transmitted PAM16 symbols are considered as one two-dimensional (2D) symbol. The DSQ128 symbols are obtained by concatenating two time-adjacent 1D PAM16 symbols and retaining among the 256 possible Cartesian product combinations, 128 maximally spaced 2D symbols. The resulting checkerboard constellation is based on a lattice called RZ
- The 2D symbols are selected from a constrained constellation of 128 maximally spaced 2D symbols, called DSQ128 (double square 128). After link startup, PHY frames consisting of 512 DSQ128 symbols are continuously transmitted. The DSQ128 symbols are determined by 7-bit labels, each comprising 3 uncoded bits and 4 LDPC-encoded bits. The 512 DSQ128 symbols of one PHY frame are transmitted as 4 × 256 PAM16 symbols over the four wire pairs.
- Data and Control symbols are embedded in a framing scheme that runs continuously after startup of the link. The modulation symbol rate of 800Msymbols/s results in a symbol period of 1.25 ns.
- 25 and 40 GBase-T use similar encoding schemes
- 64b/66b encoding, introduced for 10 Gigabit Ethernet's 10GBASE-R Physical Medium Dependent (PMD) interfaces, is a lower-overhead alternative to 8b/10b encoding, having a two-bit overhead per 64 bits (instead of eight bits) of encoded data. This scheme does not explicitly guarantee DC balance, short run length, and transition density (these features are achieved statistically via scrambling).
- 2.5GBASE-T and 5GBASE-T use the same encoding as 10GBASE-T slowed by factor four or two, respectively

ETHERNET TRANSMIT SIGNAL LEVELS



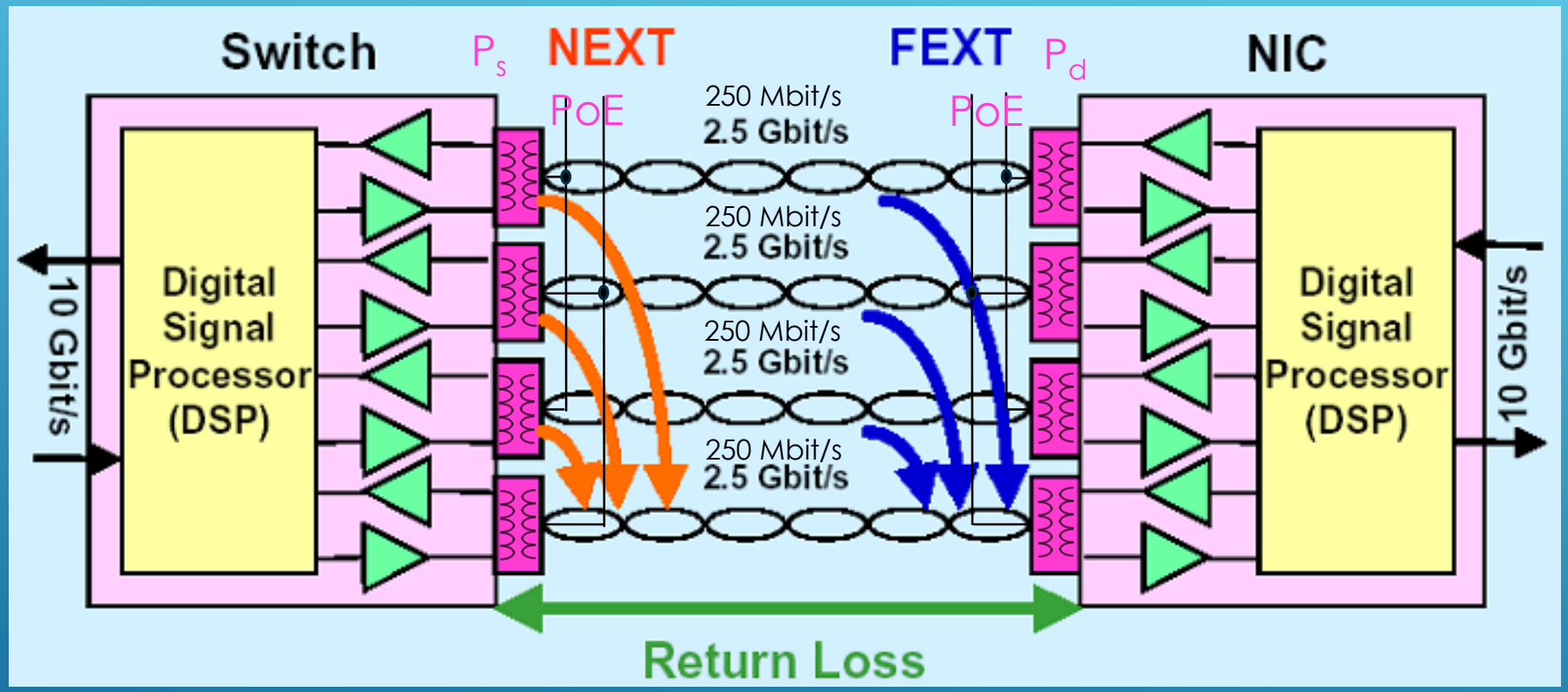
ETHERNET RECEIVE SIGNAL LEVELS

Receive signal strength weakens due to cable attenuation – these levels measured at 100m cable length



Green signal represents noise picked up by the cables

1 AND 10GBASE-T INTERFACES



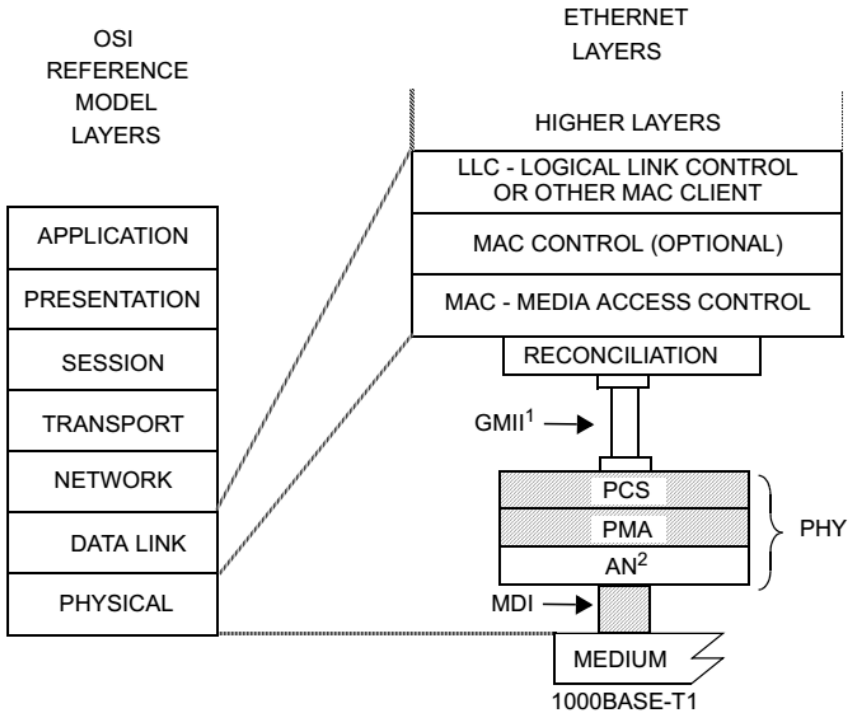
802.3 SINGLE TP (10, 100, 1000 MBPS.)

IEEE 802.3cg 10Mb/s Single Twisted Pair Ethernet

Objectives:

- Support 10Mb/s operation in automotive & industrial environments over single balanced twisted-pair cabling
  - Support for optional auto-negotiation over single-pair
- Link Segment: up to 4 connectors for up to at least 15m
- Define link segment & PHY to support point-to-point operation with 10 inline connectors using balanced cabling (1-pair) for up to at least 1km

- Two amendments published already
  - 802.3bw-2015: Amendment 1: 100Mb/s over single TP cable
  - 802.3bp-2016: Amendment 4: 1Gb/s over single TP cable



MDI = MEDIUM DEPENDENT INTERFACE
GMII = TEN GIGABIT MEDIA INDEPENDENT INTERFACE

NOTE 1—GMII is optional
NOTE 2—Auto-Negotiation is optional

PCS = PHYSICAL CODING SUBLAYER
PMA = PHYSICAL MEDIUM ATTACHMENT
PHY = PHYSICAL LAYER DEVICE
AN = AUTO-NEGOTIATION

SINGLE TP OPERATION

- The 1000BASE-T1 PHY operates using full-duplex communications over a single twisted-pair copper cable with an effective rate of 1 Gb/s in each direction simultaneously.
- The 1000BASE-T1 PHY utilizes 3 level Pulse Amplitude Modulation (PAM3) transmitted at a 750 MBd rate.
- The PCS operates in two modes: data & training. In the data mode, the PCS Transmit function data path starts with input data to the PCS every 8 ns. Data and control from ten GTX_CLK cycles are 80B/81B encoded into an 81-bit "81B block" that encodes every possible combination of data and control (control signals include transmit error propagation, receive error, assert low power idle, and inter-frame signaling. Each set of 45 80B/81B blocks along with 9 bits of 1000BASE-T1 OAM data is processed by a Reed Solomon-FEC encoder. The RS-FEC encoder adds 396 RS-FEC parity bits and the resulting 4050 bits (45 80B/81B blocks = 3645 bits, 9 bits of 1000BASE-T1 OAM, and 396 bits of FEC parity bits) are scrambled using a 15-bit side-stream scrambler. The 4050 bits are referred to interchangeably as a PHY frame or as a Reed-Solomon frame. Each group of 3 bits of the scrambled data is converted to 2 PAM3 symbols by the 3B2T mapper (the 4050 bits in the PHY frame become 2700 PAM3 symbols) and passed to the PMA.
- In the training mode, the PCS generates only a PAM2 pattern with periodic embedded data that enables the receiver at the other end to train and synchronize timing, scrambler seeds, and capabilities.

AUTONEGOTIATION (CLAUSE 28)

- **AutoNegotiation** is a procedure by which two connected devices (that use an eight-pin modular connector) select common parameters, such as speed, duplex mode, & flow control.
- The connected devices first share their capabilities regarding these parameters and then choose the highest performance transmission mode they both support.
  - Data rate
  - half duplex capability
  - whether the device is single port or multiport
  - whether master/slave is manually configured or not
  - whether the device is manually configured as master or slave
  - Energy Efficient Ethernet (EEE) operation
- AutoNegotiation is designed in a way that allows it to be easily expanded as new technologies are developed. When a new technology is developed, the following is required:
  1. The appropriate Selector Field value to contain the new technology must be selected and allocated.
  2. A Technology bit must be allocated for the new technology within the chosen Selector Field value.
  3. The new technology's relative priority within the technologies supported within a Selector Field value must be established.

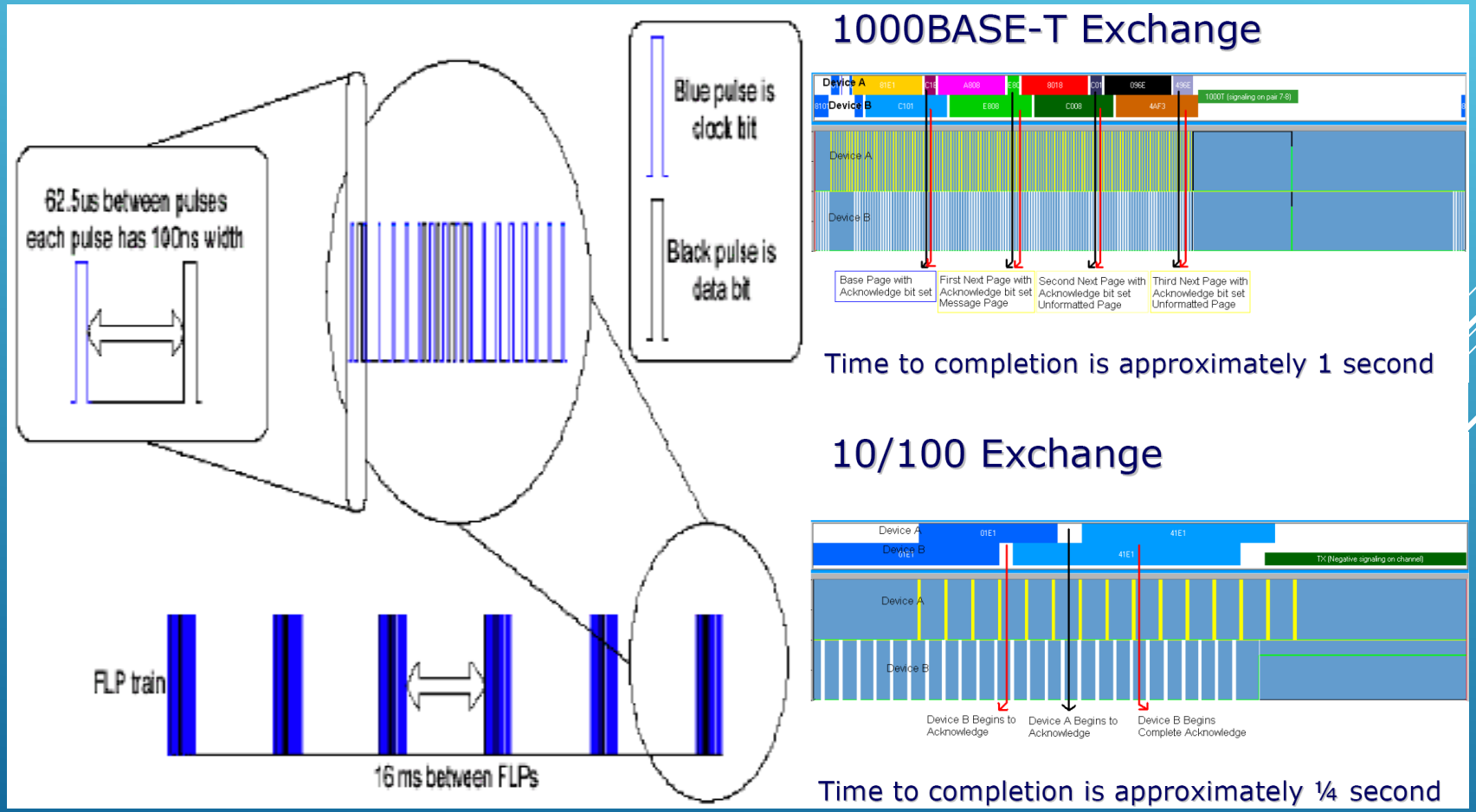
AUTONEGOTIATION

- Auto-negotiation is based on pulses that detect the presence of a connection to another device. These connection present pulses are sent by Ethernet devices when they are not sending or receiving any frames. They are unipolar positive-only electrical pulses of a nominal duration of 100 ns, with a maximum pulse width of 200 ns, generated at a 16 ms time interval (with a timing variation tolerance of 8 ms). These pulses are called normal link pulses (NLP) in the auto-negotiation specification.
- Three bursts of Fast Link Pulses, are used by autonegotiating devices to declare their capabilities.
- Auto-negotiation uses pulses labeled as NLP. NLP are unipolar, positive-only, and of the nominal duration of 100 ns; but each pulse burst consists of 17 to 33 pulses sent 125 µs apart. Each such pulse burst is called a fast link pulse (FLP) burst. The time interval between the start of each FLP burst is the same 16 milliseconds as between normal link pulses (variation tolerance of 8 ms).
- The FLP burst consists of 17 NLP at a 125 µs time interval (with a tolerance of 14 µs). Between each pair of two consecutive NLP (i.e. at 62.5 µs after first NLP of the pulse pair) an additional positive pulse may be present. The presence of this additional pulse indicates a logical 1, its absence a logical 0. As a result, every FLP contains a data word of 16 bits. This data word is called a link code word (LCW). The bits of the link code word are numbered from 0 to 15, where bit 0 corresponds to the first possible pulse in time and bit 15 to the last.
- Every fast link pulse burst transmits a word of 16 bits known as a link code word. The first such word is known as a base link code word, and its bits are used as follows:
  - 0–4: selector field: it indicates which standard is used between IEEE 802.3 and IEEE 802.9
  - 5–12: technology ability field: this is a sequence of bits that encode the possible modes of operations
  - 13: remote fault: this is set to one when the device is detecting a link failure
  - 14: acknowledgement: the device sets this to one to indicate the correct reception of the base link code word from the other party; this is detected by the reception of at least three identical base code words
  - 15: next page: this bit is used to indicate the intention of sending other link code words after the base link code word;

AUTONEGOTIATION

- Upon receipt of the technology abilities of the other device, both devices decide the best possible mode of operation supported by both devices. The priority among modes is as follows:

1. 40GBASE-T full duplex
2. 25GBASE-T full duplex
3. 10GBASE-T full duplex
4. 5GBASE-T full duplex
5. 2.5GBase-T full duplex
6. 1000BASE-T full duplex
7. 1000BASE-T half duplex
8. 100BASE-T2 full duplex
9. 100BASE-TX full duplex
10. 100BASE-T2 half duplex
11. 100BASE-T4 half duplex
12. 100BASE-TX half duplex
13. 10BASE-T full duplex
14. 10BASE-T half duplex



POWER OVER ETHERNET (POE) (CLAUSE 33)

Two modes, A and B, are available. Nominal voltage is 56 Volts measured at the PSE

- Mode A delivers power on the data pairs of 100BASE-TX or 10BASE-T. Mode B delivers power on the spare pairs. For 1000Base-T and 10GBase-T Ethernet which have no spare pairs all power is delivered using the phantom technique. Mode A has two alternate configurations (MDI and MDI-X), using the same pairs but with different polarities. In mode A, pins 1 and 2 form one side of the 48 V DC, and pins 3 and 6 form the other side. These are the same two pairs used for data transmission in 10BASE-T and 100BASE-TX, allowing the provision of both power and data over only two pairs in such networks. The free polarity allows PoE to accommodate for crossover cables, patch cables and Auto MDI-X.
- Mode B, uses the "spare" pairs in 10BASE-T and 100BASE-TX. Mode B, therefore, requires a 4-pair cable.
- The PSE (*power sourcing equipment*), not the PD (*powered device*), decides whether power mode A or B shall be used. PDs that implement only Mode A or Mode B are disallowed by the standard. The PSE can implement mode A or B or both.
- A PD indicates that it is standards-compliant by placing a 25 kΩ resistor between the powered pairs. If the PSE detects a resistance that is too high or too low (including a short circuit), no power is applied. This protects devices that do not support PoE.
- An optional "power class" feature allows the PD to indicate its power requirements.
- To stay powered, the PD must continuously use 5–10 mA for at least 60 ms with no more than 400 ms since last use or else it will be unpowered by the PSE.
- In order to prevent potential oscillations between the PSE and PD, the sum of the PSE port output impedance, the cable impedance, the PD input port circuitry impedance and the PD EMI output filter impedance should be lower than the PD power supply input impedance.

FIBER OPTIC OPTIONS

- Wide Range of Data Rates
  - 10 Mbps to 400 Gbps
- Distance to 40 Km
- Both Single (SMF) and Multi Mode (MMF) Fiber
- Plastic Optical Fiber (50m, 1Gbps, P802.3bv)
- Internal Wavelength Division Multiplexing
- From a single strand (EPON) to 32 Strands per interface

FIBER OPTIC DATA CENTERAPPLICATIONS

- The majority of data center optics use multimode fiber (MMF) technology because of low cost. The maximum specified reach is 300 m, although most deployed links are below 100 m. The 10GESR optics use a single vertical cavity surface emitting laser (VCSEL) transmitter and are packaged in the small form-factor pluggable (SFP+) module, establishing the baseline bit-per-second cost vs. volume reference curve for higher-data-rate structured MMF optics.
- The defining characteristics of SMF general data center optics are a minimum loss budget of 6 dB, and link budget penalties supporting up to 10 km duplex SMF reach. The 10GE-LR duplex SMF optics specifications are defined in the 802.3ae-2002 standard, use a single DFB laser transmitter, and are packaged in the 10G form-factor pluggable (XFP) or SFP+ module. The SFP+ implementation establishes the baseline bit-per-second cost vs. volume reference curve for higher-data-rate general data center SMF optics.

IEEE 802.3BA

PMD	Link Distance	Fiber Count and Media Type	Technology
40GBASE-SR4	100 m OM3 150 m OM4	8-f MMF (12-f MPO)	4x10G parallel NRZ 850nm
40GBASE-LR4	10 km	2-f SMF	4x10G CWDM NRZ 4 wavelengths around 1300nm
100GBASE-SR10	100 m OM3 150 m OM4	20-f MMF (24-f MPO)	10x10G parallel NRZ 850 nm
100GBASE-LR4	10 km	2-f SMF	4x25G CWDM NRZ 4 wavelengths around 1300nm
100GBASE-ER4	40 km	2-f SMF	4x25G CWDM NRZ 4 wavelengths around 1300nm

802.3BM

PMD	Link Distance	Fiber Count and Media Type	Technology
40GBASE-ER4	30 km (40 km engineered link)	2-f SMF	4x10G CWDM NRZ 4 wavelengths around 1300nm
100GBASE-SR4	70 m OM3 100 m OM4	8-f MMF (12-f MPO)	4x25G parallel NRZ 850 nm

802.3BS & 802.3CD

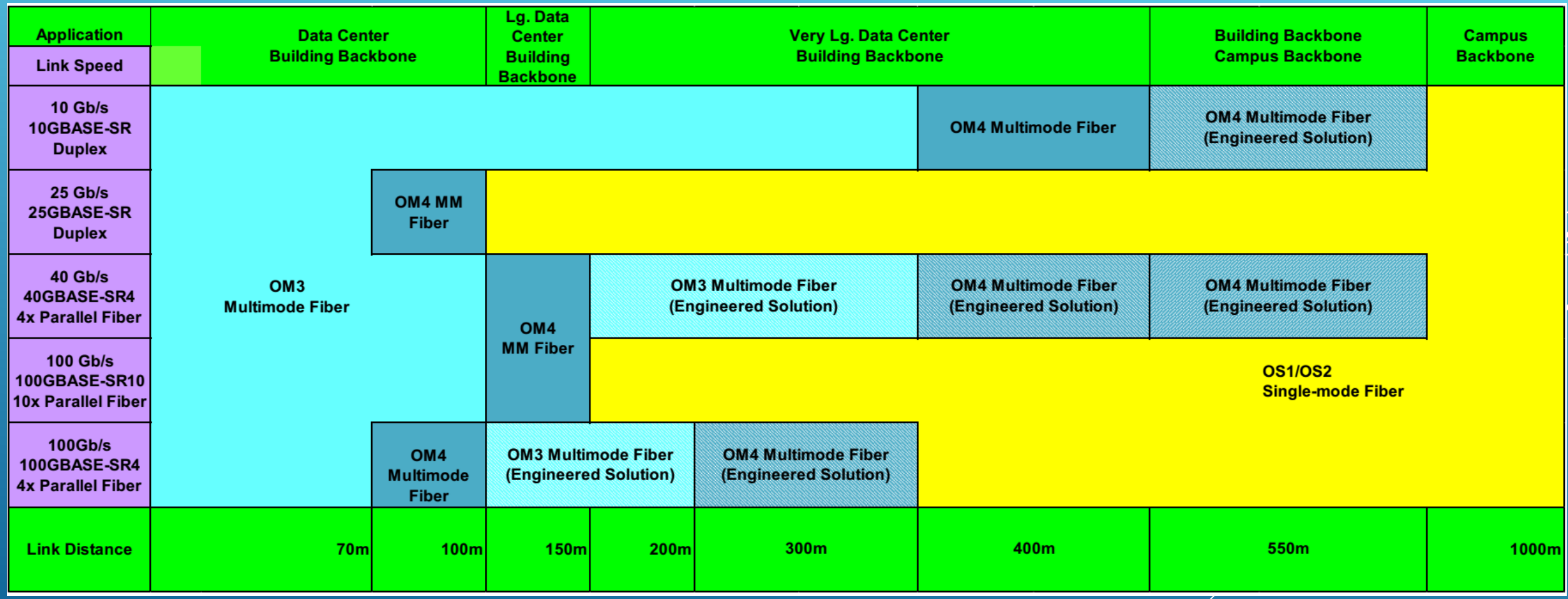
Table 2: IEEE 802.3bs Variants						
Rate	Encoding	Symbol Rate	Fiber(s) per Direction	Transceiver Interface	Wavelengths	Reach
200GBASE-DR4	PAM 4	26.5 GBd	4 Lanes SMF	MPO-12	1	500m
200GBASE-FR4	PAM 4	26.5 GBd	1 Lane SMF	LC	4	2km
200GBASE-LR4	PAM 4	26.5 GBd	1 Lane SMF	LC	4	10km
400GBASE-SR16	NRZ	26.5 GBd	16 Lanes MMF	MPO-32	1	100m
400GBASE-DR4	PAM 4	53.0 GBd	4 Lanes SMF	MPO-12	1	500m
400GBASE-FR8	PAM 4	26.5 GBd	1 Lane SMF	LC	8	2km
400GBASE-LR8	PAM 4	26.5 GBd	1 Lane SMF	LC	8	10Km

Table 3: Proposed IEEE 802.3cd Variants						
Rate	Encoding	Symbol Rate	Fiber(s) per Direction	Transceiver Interface	Wavelengths	Reach
50GBASE-SR	PAM 4	26.5 GBd	1 Lane MMF	LC	1	100m
50GBASE-FR	PAM 4	26.5 GBd	1 Lane SMF	LC	1	2km
50GBASE-LR	PAM 4	26.5 GBd	1 Lane SMF	LC	1	10km
100GBASE-SR2	PAM 4	26.5 GBd	2 Lanes MMF	MPO-12	1	100m
100GBASE-DR	PAM 4	53.0 GBd	1 Lane SMF	LC	1	500m
200GBASE-SR4	PAM 4	26.5 GBd	4 Lanes MMF	MPO-12	1	100m

802.3BY

PMD	Link Distance	Fiber Count and Media Type	Technology
25GBASE-SR	100 m OM4	2-f MMF	1x25G NRZ

STANDARD 802.3 FIBER LINK DISTANCES



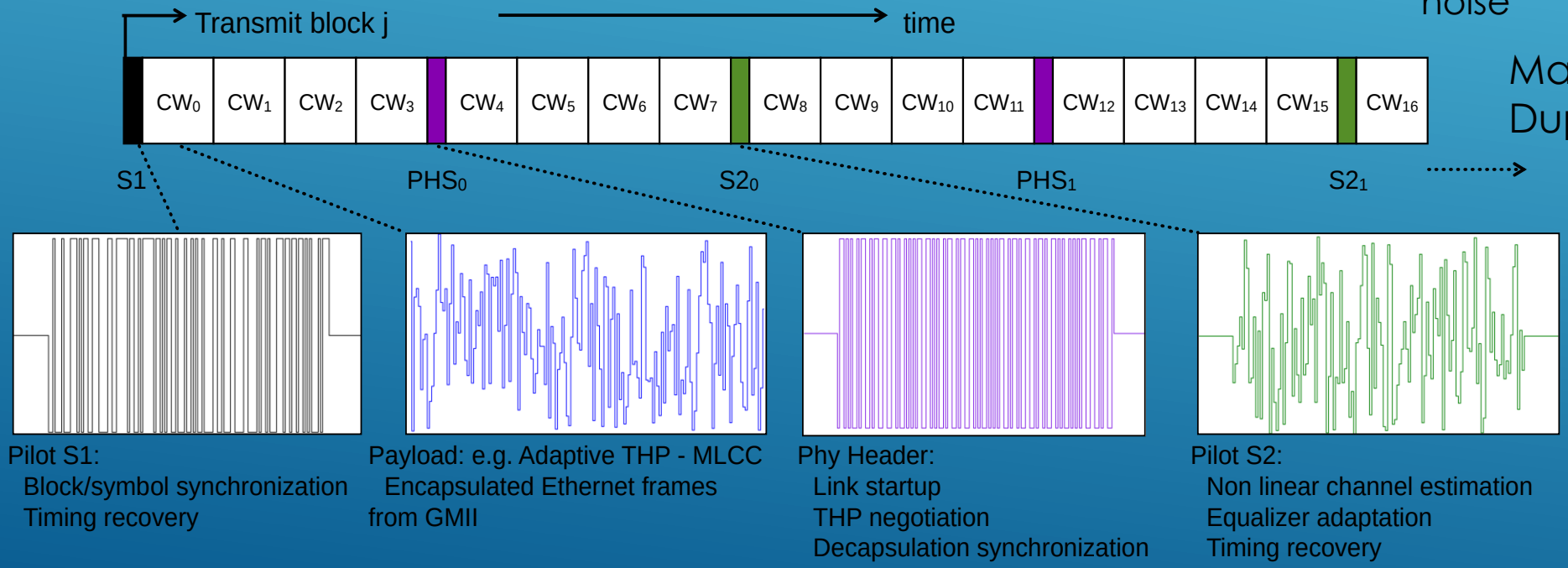
PLASTIC OPTICAL FIBER (802.3BV)

DATA ENCAPSULATION

- Periodic transmission structure
  - To have a fast link establishment (Less than 50 ms)
  - To have big tolerance to clock frequency mismatch (>+-200 ppm)
  - To have a fast negotiation of THP TX coefficients
  - To track and equalize channel changes with temperature, bending and vibration
  - To implement Low Power Idle (LPI) mode for Energy Efficient Ethernet (EEE)

Encoding

- PAM-16 with THP (312.5 Mbaud) Optimum solution for current LED, Fiber and PD and feasible TIA
- THP is used to: Solve Inter Symbol Interference (ISI) in combination with high spectral efficiency coded modulation
- Allows whitening of TIA non-white noise



Maximum 50 Meter Distance Duplex Fiber

Pilot S1:
Block/symbol synchronization
Timing recovery

Payload: e.g. Adaptive THP - MLCC
Encapsulated Ethernet frames from GMII

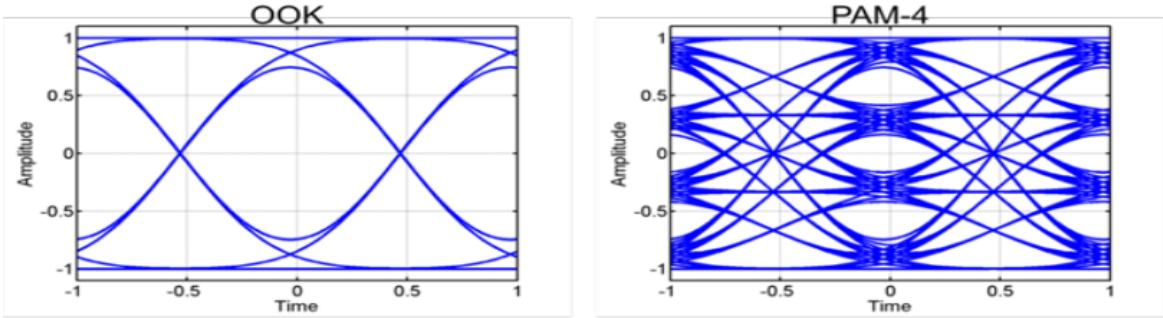
Phy Header:
Link startup
THP negotiation
Decapsulation synchronization

Pilot S2:
Non linear channel estimation
Equalizer adaptation
Timing recovery

PROPOSED MULTILEVEL OPTICAL SIGNALING

PAM-4

- Increases the bit rate 2x



- **Currently under discussion in IEEE and FC for next generation solutions**
  - Could leverage CWDM efforts to further expand fiber capacity
  - Discussion of possible 50Gb/s/lane rates
- **Advanced modulation formats require higher receiver sensitivity than OOK**
  - Have to accommodate “multiple eyes” within same vertical interval
- **Receiver sensitivity requirements can be reduced via Equalization and/or FEC**

25 GBPS OPTICAL ETHERNET 802.3CC

PMD	Link Distance	Fiber Count and Media Type	Technology
25GBASE-LR	10 km SMF	2-f SMF	1x25G NRZ
25GBASE-ER	40 km SMF	2-f SMF	1x25G NRZ

802.3CD

PMD	Link Distance	Fiber Count and Media Type	Technology
50GBASE-SR	100 m OM4	2-f MMF	1x50G PAM-4 850nm
50GBASE-FR	2 km	2-f SMF	1x50G PAM-4 1300nm
50GBASE-LR	10 km	2-f SMF	1x50G PAM-4 1300nm
100GBASE-SR2	100 m	4-f MMF	2x50G PAM-4 850nm
100GBASE-DR	500 m	2-f SMF	1x100G PAM-4 1300nm
200GBASE-SR4	100 m	8-f MMF	4x50G parallel PAM-4 850nm

OPTICAL ETHERNET FUTURE P802.3BS

PMD	Link Distance	Fiber Count and Media Type	Technology
400GBASE-SR16	100 m OM4 (32-f MPO)	32-f MMF	16x25G parallel NRZ 850nm
400GBASE-DR4	500 m	8-f SMF	4x100G parallel PAM4 1300nm
400GBASE-FR8	2 km	2-f SMF	8x50G CWDM PAM4 8 wavelengths around 1300nm
400GBASE-LR8	10 km	2-f SMF	8x50G CWDM PAM4 8 wavelengths around 1300nm
200GBASE-DR4	500 m	8-f SMF	4x50G Parallel PAM4 1300nm
200GBASE-FR4	2 km	2-f SMF	4x50G CWDM PAM4 4 wavelengths around 1300nm
200GBASE-LR4	10 km	2-f SMF	4x50G CWDM PAM4 4 wavelengths around 1300nm

EPON, GPON 802.3AH, 802.3AV CLAUSE 75

- EPON uses standard 802.3 Ethernet frames with symmetric 1 gigabit per second upstream and downstream rates
- 10 Gbit/s EPON or 10G-EPON is defined by IEEE 802.3av. The downstream wavelength plan support simultaneous operation of 10 Gbit/s on one wavelength and 1 Gbit/s on a separate wavelength for operation of IEEE 802.3av and IEEE 802.3ah on the same PON concurrently.
- The upstream channel can support simultaneous operation of IEEE 802.3av and 1 Gbit/s 802.3ah simultaneously on a single shared (1310 nm) channel.
- A PON takes advantage of wavelength division multiplexing (WDM), using one wavelength for downstream traffic and another for upstream traffic on ONE single mode fiber. EPON, and GPON have the same basic wavelength plan and use:
  - 1490 nanometer (nm) wavelength for downstream traffic
  - 1310 nm wavelength for upstream traffic
  - 1550 nm for optional downstream overlay services, typically RF (analog) video
- The standards describe optical budgets; most common is 28 dB of loss budget for both EPON and GPON. 28 dB allows about 20 km with a 32-way split. Forward error correction (FEC) provides another 2–3 dB of loss budget on GPON systems. Both the GPON and EPON protocols permit split ratios up to 128 subscribers for GPON, up to 32,768 for EPON. In practice most PONs are deployed with a split ratio of 1:32 or smaller.
- A PON consists of a central node, called an optical line terminal (OLT), one or more user nodes, called optical network units (ONUs) or optical network terminals (ONTs), and the fibers and splitters between them, called the optical distribution network (ODN). Some ONUs implement a separate subscriber unit to provide services such as telephony, Ethernet data, or video.
- An OLT provides the interface between a PON and a service provider's core network. These typically include:
  - IP traffic over Fast Ethernet, gigabit Ethernet, or 10 Gigabit Ethernet;
  - Standard TDM interfaces such as SDH/SONET;
  - ATM UNI at 155–622 Mbit/s.
- The ONT or ONU terminates the PON and presents the native service interfaces to the user. These services can include voice (plain old telephone service (POTS) or voice over IP (VoIP)), data (typically Ethernet or V.35), video, and/or telemetry (TTL, ECL, RS530, etc.) Often the ONU functions are separated into two parts:
  1. The ONU, which terminates the PON and presents a converged interface—such as DSL, coaxial cable, or multiservice Ethernet—toward the user;
  2. Network termination equipment (NTE), which inputs the converged interface and outputs native service interfaces to the user, such as Ethernet and POTS.
- A PON is a shared network, in that the OLT sends a single stream of downstream traffic that is seen by all ONUs. Each ONU only reads the content of those packets that are addressed to it.
- Encryption is used to prevent eavesdropping on downstream traffic.

GPON ARCHITECTURE OVERVIEW

- In the downstream direction (from the OLT to an ONU), signals transmitted by the OLT pass through a 1:N passive splitter or a cascade of splitters and reach each ONU.
- In the upstream direction (from the ONUs to the OLT), the signal transmitted by an ONU would only reach the OLT, but not other ONUs. To avoid data collisions and increase efficiency ONU's transmissions are arbitrated.
- This arbitration is achieved by allocating a transmission window (grant) to each ONU. An ONU defers transmission until its grant arrives. When the grant arrives, the ONU transmits frames at wire speed during its assigned time slot.

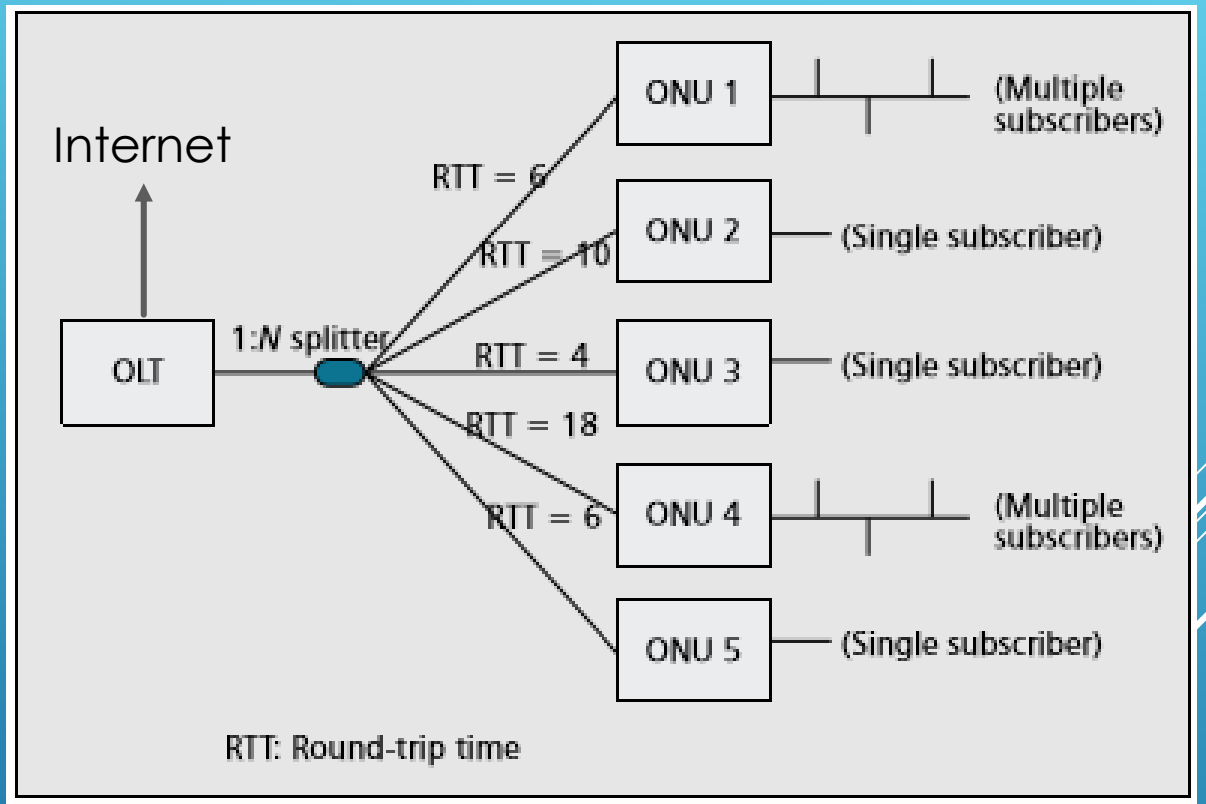


FIGURE 1. Network architecture of an EPON with one optical line terminal (OLT) and N = 5 optical network units (ONUs), each with a different round-trip time (RTT).

LINK LAYER DISCOVERY PROTOCOL (LLDP) 802.1AB

- Advertises connectivity and management information about the local station to adjacent stations on the same IEEE 802 LAN.
- Receives network management information from adjacent stations on the same IEEE 802 LAN.
- Operates with all IEEE 802 access protocols and network media.
- Establishes a network management information schema and object definitions that are suitable for storing connection information about adjacent stations.
- Provides compatibility with the IETF *PTOPO* MIB (IETF RFC 2922)
  - physical topology Management Information Base
- Information that may be retrieved include:
  - System name and description
  - Port name and description
  - VLAN number
  - IP management address
  - System capabilities (switching, routing, etc.)
  - MAC/PHY information
  - PoE MDI power
  - Link aggregation

Media Endpoint Discovery is an enhancement of LLDP, known as LLDP-MED, that provides the following facilities:

- Auto-discovery of LAN policies (such as VLAN, Layer 2 Priority and Differentiated services (Diffserv) settings) enabling plug and play networking.
- Device location discovery to allow creation of location databases and, in the case of Voice over Internet Protocol (VoIP), Enhanced 911 services.
- Extended and automated power management of Power over Ethernet (PoE) end points.
- Inventory management, allowing network administrators to track their network devices, and determine their characteristics (manufacturer, software and hardware versions, serial or asset number).

802.1AB FRAME STRUCTURE

LLDP Ethernet frame structure

Preamble	Destination MAC	Source MAC	Ethertype	Chassis ID TLV	Port ID TLV	Time to live TLV	Optional TLVs	End of LLDPDU TLV	Frame check sequence
	01:80:c2:00:00:0e, or 01:80:c2:00:00:03, or 01:80:c2:00:00:00	Station's address	0x88CC	Type=1	Type=2	Type=3	Zero or more complete TLVs	Type=0, Length=0	

Each of the TLV components has the following basic structure:

TLV structure

Type	Length	Value
7 bits	9 bits	0-511 octets

Custom TLVs[note 2] are supported via a TLV type 127. The value of a custom TLV starts with a 24-bit organizationally unique identifier and a 1 byte organizationally specific subtype followed by data. The basic format for an organizationally specific TLV is shown below:

Organizationally specific TLV

Type	Length	Organizationally unique identifier (OUI)	Organizationally defined subtype	Organizationally defined information string
7 bits—127	9 bits	24 bits	8 bits	0-507 octets

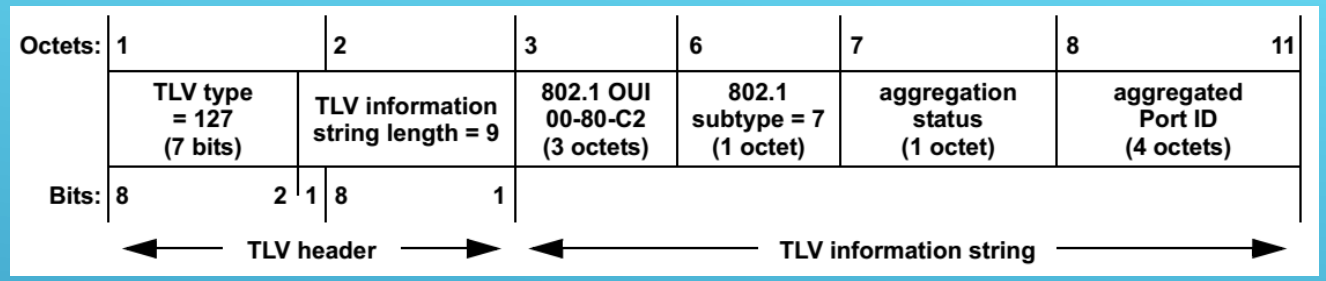
According to IEEE Std 802.1AB, §9.6.1.3, "The Organizationally Unique Identifier shall contain the organization's OUI as defined in IEEE Std 802-2001." Each organization is responsible for managing their subtypes.

- LLDP information is sent by devices from each of their interfaces at a fixed interval, in the form of an Ethernet frame. Each frame contains one LLDP Data Unit (LLDPDU). Each LLDPDU is a sequence of type-length-value (TLV) structures.
- The Ethernet frame used in LLDP has its destination MAC address typically set to a special multicast address that 802.1D-compliant bridges do not forward. Other multicast and unicast destination addresses are permitted. The EtherType field is set to 0x88cc.
- Each LLDP frame starts with the following mandatory TLVs: Chassis ID, Port ID, and Time-to-Live. The mandatory TLVs are followed by any number of optional TLVs. The frame ends with a special TLV, named end of LLDPDU in which both the type and length fields are 0

LINK AGGREGATION 802.1AX

- Link Aggregation allows one or more links to be aggregated together to form a Link Aggregation Group (LAG), treated as if it were a single link.
- Link Aggregation specifies the establishment of logical links, consisting of N parallel instances of full-duplex point-to-point links.
- All aggregated links must have the same port speed
- Link aggregation does NOT support half duplex nor multipoint configurations
- Failover occurs automatically
- Management attributes are defined for link aggregation

LINK AGGREGATION CONTROL PROTOCOL



- The Link Aggregation Control Protocol (LACP) controls the bundling of several physical ports together to form a single logical channel. LACP negotiates an automatic bundling of links by sending LACP packets to the peer.
- Usually up to 8 bundled ports are allowed in a port channel.
- LACP packets are sent with multicast group MAC address 01:80:c2:00:00:02
- LACP packets are transmitted every second during LACP detection period
- Keep alive mechanism for link member: (default: slow = 30s, fast=1s)
- LACP can have the port-channel load-balance mode :
- link (link-id) Integer that identifies the member link for load balancing. The range is from 1 to 8.
- LACP mode :
  - active : Enables LACP unconditionally.
  - passive : Enables LACP only when an LACP device is detected. (This is the default state)

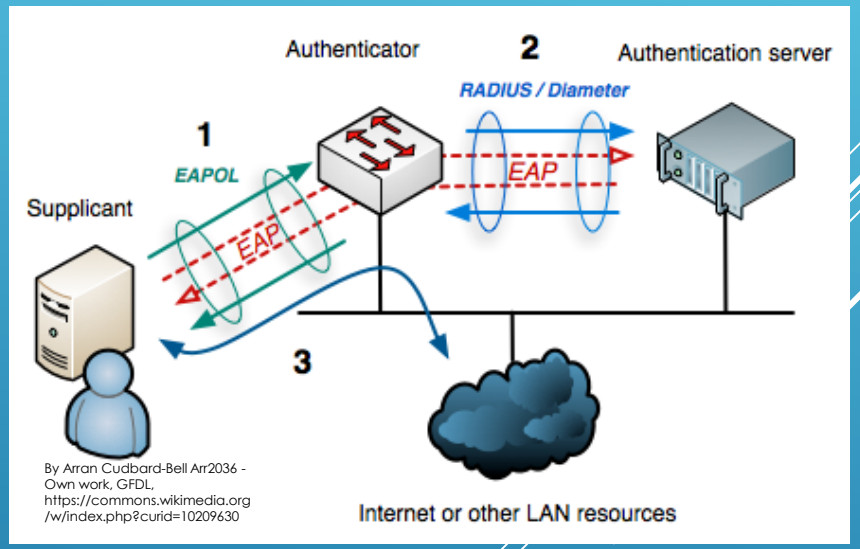
PORT-BASED NETWORK ACCESS CONTROL 802.1X

- Std. 802.1X specifies an architecture, functional elements, and protocols to support mutual authentication between the clients of ports attached to the same LAN and secure communication between the ports.
- Use the service provided by the LAN MAC, at a common service access point, to support a Controlled Port that provides secure access-controlled communication and an Uncontrolled Port that supports protocols that initiate the secure communication or do not require protection.
- Support mutual authentication between a Port Access Entity (PAE) associated with a Controlled Port, and a peer PAE associated with a peer port in a LAN attached station that desires to communicate through the Controlled Port.
- Secure the communication between the Controlled Port and the authenticated peer port, excluding other devices attached to or eavesdropping on the LAN.
- Provide the Controlled Port with attributes that specify access controls appropriate to the authorization accorded to the peer station or its user.
- This standard specifies the use of EAP, the Extensible Authentication Protocol (IETF RFC 3748) to support authentication using a centrally administered Authentication Server and defines EAP encapsulation over LANs to convey the necessary exchanges between peer PAEs attached to a LAN.

PORT BASED NETWORK ACCESS CONTROL

802.1X authentication involves 3 parties: a supplicant, an authenticator, and an authentication server.

- The supplicant is a client device (such as a laptop) that wishes to attach to the LAN/WLAN.
- The authenticator is a network device, such as an Ethernet switch or wireless access point.
- The authentication server is a host running software supporting the RADIUS and EAP protocols.
  - In some cases, the authentication server software may run on the authenticator hardware.
- The authenticator acts like a security guard to a protected network. The supplicant is not allowed access through the authenticator to the protected side of the network until the supplicant's identity has been validated and authorized.
- With port authentication, the supplicant provides credentials, such as user name/password or digital certificate, to the authenticator, and the authenticator forwards the credentials to the authentication server for verification.
- If the authentication server determines the credentials are valid, the supplicant is allowed to access resources located on the protected side of the network

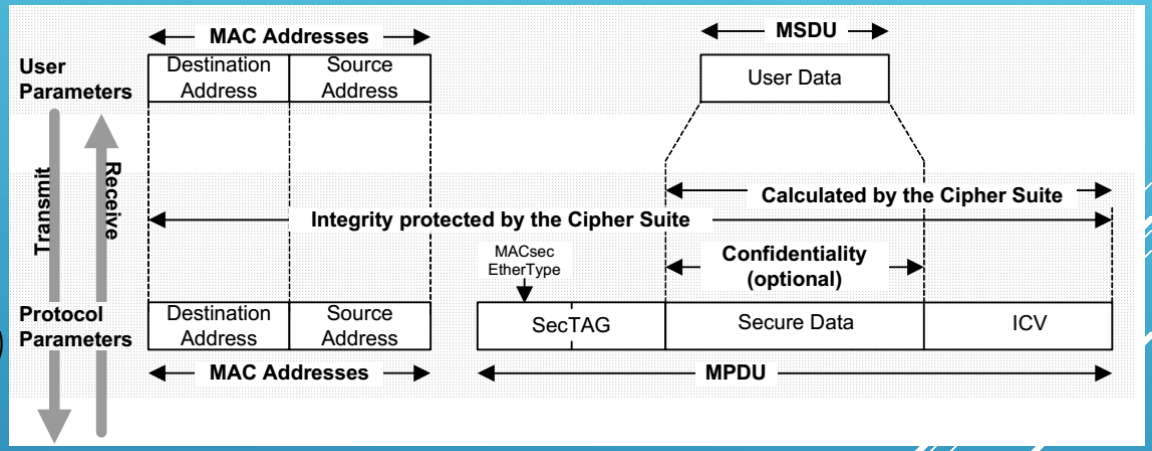


802.1X PROCEDURES

1. Initialization On detection of a new supplicant, the port on the switch (authenticator) is enabled and set to the "unauthorized" state. In this state, only 802.1X traffic is allowed; other traffic, such as the Internet Protocol (and with that TCP and UDP), is dropped.
2. Initiation To initiate authentication the authenticator will periodically transmit EAP-Request Identity frames to a specific Layer 2 address on the local network segment. The supplicant listens on this address, and on receipt of the EAP-Request Identity frame it responds with an EAP-Response Identity frame containing an identifier for the supplicant such as a User ID. The authenticator then encapsulates this Identity response in a RADIUS Access-Request packet and forwards it on to the authentication server. The supplicant may also initiate or restart authentication by sending an EAPOL-Start frame to the authenticator, which will then reply with an EAP-Request Identity frame.
3. Negotiation The authentication server sends a reply (encapsulated in a RADIUS Access-Challenge packet) to the authenticator, containing an EAP Request specifying the EAP Method (The type of EAP based authentication it wishes the supplicant to perform). The authenticator encapsulates the EAP Request in an EAPOL frame and transmits it to the supplicant. At this point the supplicant can start using the requested EAP Method, or do an NAK and respond with the EAP Methods it is willing to perform.
4. Authentication If the authentication server and supplicant agree on an EAP Method, EAP Requests and Responses are sent between the supplicant and the authentication server (translated by the authenticator) until the authentication server responds with either an EAP-Success message (encapsulated in a RADIUS Access-Accept packet), or an EAP-Failure message (encapsulated in a RADIUS Access-Reject packet). If authentication is successful, the authenticator sets the port to the "authorized" state and normal traffic is allowed, if it is unsuccessful the port remains in the "unauthorized" state. When the supplicant logs off, it sends an EAPOL-logoff message to the authenticator, the authenticator then sets the port to the "unauthorized" state, once again blocking all non-EAP traffic.

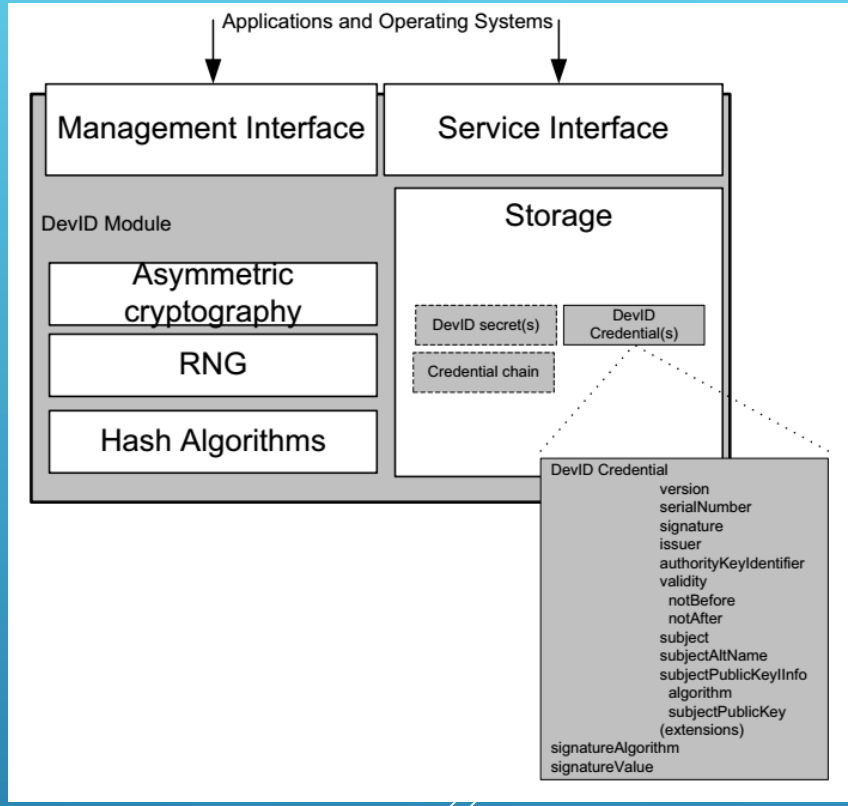
MEDIA ACCESS CONTROL SECURITY 802.1AE

- The IEEE 802.1AE (MACsec) standard specifies a set of protocols to meet the security requirements for protecting data traversing Ethernet LANs.
- It specifies the implementation of a MAC Security Entities (SecY) that can be thought of as part of the stations attached to the same LAN, providing secure MAC service to the client. The standard defines:
  - MACsec frame format, which is similar to the Ethernet frame, but includes additional fields:
  - Security Tag, which is an extension of the EtherType
  - Message authentication code (ICV)
  - Secure Connectivity Associations that represent groups of stations connected via unidirectional Secure Channels
  - Security Associations within each secure channel. Each association uses its own key (SAK). More than one association is permitted within the channel for the purpose of key change without traffic interruption
    - (802.1ae standard requires devices to support at least two)
  - A default cipher suite of either GCM-AES-128 (Galois/Counter Mode with 128-bit key) or GCM-AES-256 using a 256 bit key
- Security tag inside each frame in addition to EtherType includes:
  - association number within the channel
  - packet number to provide unique initialization vector for encryption and authentication algorithms as well as protection against replay attack
  - optional LAN-wide secure channel identifier (not required on point-to-point links).
- MACsec allows unauthorized LAN connections to be identified and excluded from communication within the network. In common with IPsec and SSL, MACsec defines a security infrastructure to provide data confidentiality, data integrity and data origin authentication.
- By assuring that a frame comes from the station that claimed to send it, MACSec can mitigate attacks on Layer 2 protocols.



SECURE DEVICE IDENTITY 802.1AR

- Devices that compose a network are designed for unattended autonomous operation and might not support user authentication.
- This standard specifies unique per-device identifiers (DevID) and the management and cryptographic binding of a device to its identifiers, the relationship between an initially installed identity and subsequent locally significant identities, and interfaces and methods for use of DevIDs with existing and new provisioning and authentication protocols.
- DevID capability incorporates a globally unique manufacturer provided Initial Device Identifier (IDevID), stored in a way that protects it from modification.
- Each LDevID is bound to the device in a way that makes it infeasible for it to be forged or transferred to a device with a different IDevID without knowledge of the private key used to effect the cryptographic binding.
- Especially relevant for Internet of Things (IoT) devices



PRECISION CLOCK SYNCHRONIZATION PROTOCOL 802.1AS

- 802.1AS specifies the protocol and procedures used to ensure that synchronization requirements are met for time-sensitive applications, across bridged and virtual bridged local area networks consisting of LAN media where the transmission delays are fixed and symmetrical; for example, IEEE 802.3 full-duplex links. This includes the maintenance of synchronized time during normal operation and following addition, removal, reconfiguration, or failure of network components.
- It specifies the use of IEEE 1588 specifications where applicable in the context of IEEE Std. 802.1D-2004 and IEEE Std. 802.1Q-2005.
- Synchronization to an externally provided timing signal (e.g., a recognized timing standard such as UTC or TAI) is not part of this standard but is not precluded.
- This standard enables stations attached to bridged LANs to meet the respective jitter, wander, and time synchronization requirements for time-sensitive applications.
- A port attached to a full-duplex, point-to-point link uses the PTP peer delay protocol to measure propagation delay across the link.
- Synchronization information is transported using the PTP messages Sync and Follow_Up.

ETHERNET SYNC TECHNIQUES

- The Ethernet family of computer networks do not carry clock synchronization information. Several means are defined to address this issue. IETF's Network Time Protocol, & IEEE's 1588-2008 Precision Time Protocol are some of them.
- SyncE was standardized by the ITU-T, in cooperation with IEEE, as three recommendations:
  1. ITU-T Rec. G.8261 that defines aspects about the architecture and the wander performance of SyncE networks
  2. ITU-T Rec. G.8262 that specifies Synchronous Ethernet clocks for SyncE
  3. ITU-T Rec. G.8264 that describes the specification of Ethernet Synchronization Messaging Channel (ESMC)
- SyncE architecture minimally requires replacement of the internal clock of the Ethernet card by a phase locked loop in order to feed the Ethernet PHY.
- Clocks, are defined in terms of accuracy, noise transfer, holdover performance, noise tolerance and noise generation. The basic IEEE 802.3 standard specifies Ethernet clocks to be within ±100 ppm. EECs accuracy must be within ±4.6 ppm. In addition, by timing the Ethernet clock, it is possible to achieve Primary Reference Clock (PRC) traceability at the interfaces.
- Precision Time Protocol (PTP) uses the following message types.
  1. Sync, Delay_Req, Follow_Up and Delay_Resp messages are used by ordinary and boundary clocks and communicate time-related information used to synchronize clocks across the network.
  2. Pdelay_Req, Pdelay_Resp and Pdelay_Resp_Follow_Up are used by transparent clocks to measure delays across the communications medium so that they can be compensated for by the system.
  3. Announce messages are used by the best master clock algorithm in IEEE 1588-2008 to build a clock hierarchy and select the grandmaster.
  4. Management messages are used by network management to monitor, configure and maintain a PTP system.
  5. Signaling messages are used for non-time-critical communications between clocks.

NTP COMPARED TO ETHERNET PTP

Criteria	NTP	PTP (IEEE 1588)
Peak time transfer error	> 1ms (10^{-3}s)	> 100 ns (10^{-7}s)
Primary error source	Routers	Routers, switches, port contention, o/s stack delay, network etc.
Implementation	Hardware or software servers; software clients	Hardware masters; hardware or software clients (slaves)
Mode of operation	Clients pull time from server	Masters push time to slaves (clients)
On path support	Non existent and not possible	Not required, but possible through transparent clocks and boundary clocks (enhances performance)
Relative cost of solution	Inexpensive	More expensive (higher precision solutions cost more)
Metrics, monitoring & management	Exists, but minimal	Extensive in band metrics for monitoring and management

(from Spectracom)

WAKE ON LAN (NOTE: WOL IS NOT AN IEEE 802 SERIES STANDARD)

- Wake-on-LAN ("WoL") is implemented using a specially designed packet called a magic packet, which is sent to all computers in a network, including the computer to be awakened. The magic packet contains the MAC address of the destination computer, an identifying number built into each network interface card ("NIC") that enables it to be uniquely recognized on a network.
- Powered-down computers capable of Wake-on-LAN contain network devices able to "listen" to incoming packets in low-power mode while the system is powered down. If a magic packet is received that is directed to the device's MAC address, the NIC signals the computer's power supply or motherboard to initiate system wake-up, and turn on the power.
- In order for Wake-on-LAN to work, parts of the network interface need to stay on. This consumes a small amount of standby power, that is typically much less than normal operating power.
- The magic packet is a broadcast frame containing within its payload 6 bytes of FF FF FF FF FF FF in hexadecimal, followed by sixteen repetitions of the target computer's 48-bit MAC address, for a total of 102 bytes.
- WoL may be sent as any network- and transport-layer protocol, although it is typically sent as a UDP datagram to port 0, 7 or 9, or directly over Ethernet as EtherType 0x0842.
- A standard magic packet has the following limitations:
  - Requires destination computer MAC address (also may require a SecureOn password)
  - Does not provide a delivery confirmation
  - May not work outside of the local network
  - Requires hardware support of Wake-On-LAN on destination computer

ETHERNET TRANSCEIVERS

- Transceiver Basics - Definitions and Acronyms
  - PMD - Physical Media Dependent
  - MSA - Multi Source Agreement
  - OSA - Optical sub-assembly (TOSA & ROSA)
  - MDI - Media Device Interface
  - QSFP+ - Quad Small Form-Factor Pluggable
  - SFF - Small Form Factor
  - SFP - Small Form-Factor Pluggable
  - CFP - 100G Form-Factor Pluggable
  - CSFP - Compact Small Form-Factor Pluggable
  - VCSEL – Vertical Cavity Surface Emitting Laser

TRANSCEIVER MODULES

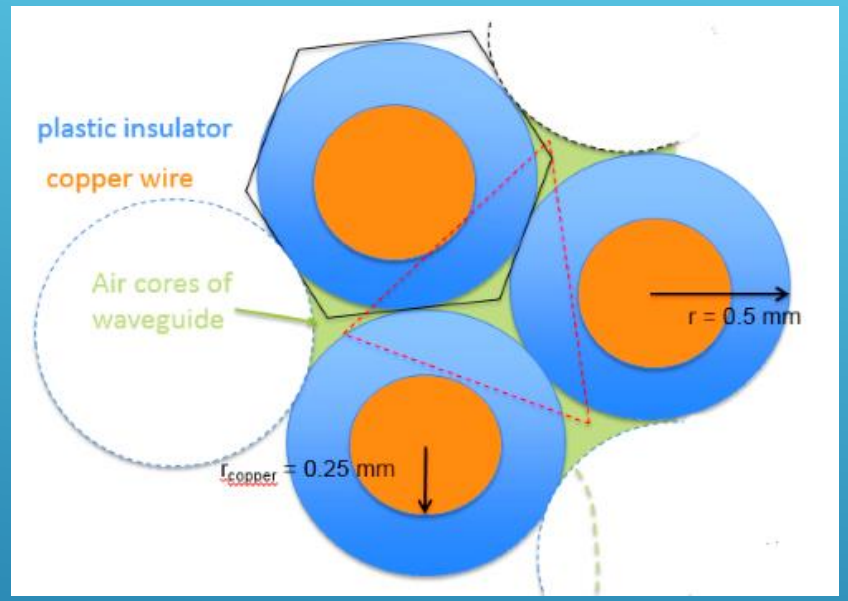
- SFP+ the enhanced small form-factor pluggable transceiver, developed by the ANSI T11 fibre channel group is small and low power. SFP+ has become the most popular socket on 10GE systems. SFP+ modules do only optical to electrical conversion, no clock and data recovery, putting a higher burden on the host's channel equalization. SFP+ modules share a common physical form factor allowing higher port density and the re-use of existing designs for 24 or 48 ports in a 19" rack width blade.
- Optical modules are connected to a host by SFI interface that uses a single lane data channel and 64b/66b encoding.
- Shortwave WDM (SWDM): The ability to multiplex multiple lanes onto a single fiber to reduce the fiber count and enable duplex-LC interfaces for 40GbE, 100GbE, and beyond
- Lane rates > 25 Gb/s: Developing and standardizing technology that enables multimode VCSELs to operate at 50 Gb/s and beyond, to enable future generations of both single-lane and multiple lane optical interfaces
- Wideband MMF: Support for the definition and standardization of wideband multimode fiber to enable WDM transmission over links that are greater than 300m

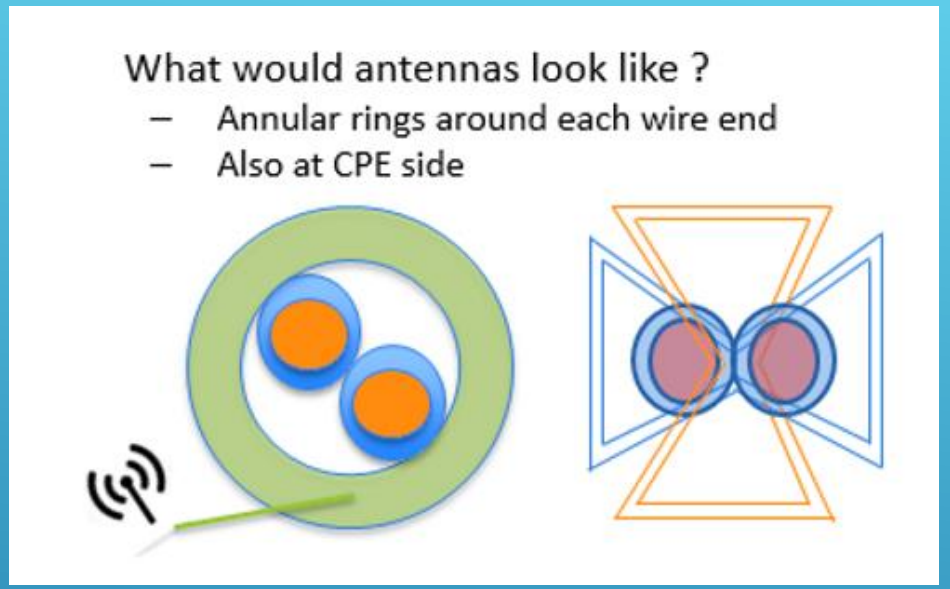
FIBER OPTIC FUTURE ENHANCEMENTS

- MMF/VCSEL solution will continue seeing pressure from competing SMF/Long wavelength laser solutions
- VCSEL/MMF solutions offer several advantages worth preserving
  - Lowest energy consumption
  - Dust/debris immunity at connections (robust operation, MACs)
  - Support break-out via parallel optics (vs. WDM breakout)
  - Large installed base (>80% of DC fiber media)
- Benefits preserved for the bulk of 100G/128GFC/400G/512G channels by supporting 100m reach (even in break-out implementations)
  - using 4x parallel solutions

TERABIT DSL (TDSL)

Use of a copper pair's sub-millimeter Waveguide modes





John Cioffi unveiled ideas behind Terabit DSL (TDSL). They include carrying 50-600 GHz wireless signals through the tiny spaces between individual twisted pairs, a terabit/second over 100 meters, 100 Gbits/s at 300 meters and 10 Gbits/s at 500 meters. Concepts in TDSL had their start two years ago when he was reviewing 60 GHz MIMO work at a wireless lab at NYU led by Ted Rappaport. "I started thinking whether there was an analog for this work in cables" Cioffi said. TDSL is a combination of millimeter waveguides and vectoring signal processing. Instead of injecting current in the wires, he foresees putting almost microscopic antennas like a donut or bowtie around the wire--not touching it. "Shooting an IR laser at those elements creates a millimeter wave similar to what's done in coherent optics, but we use a different frequency and an array of signals instead of just one used in optics."

From EETimes, 5-9-2017

HOW MUCH

- Difficult to pin down costs
  - Cabling Infrastructure
  - Code
  - Managing and Maintaining
  - Security

Reference	URL
IETF RFCs	www.rfc-editor.org
NIST	www.nist.gov/
NANOG	www.nanog.org/
Internet Protocol Journal	ipj.dreamhosters.com/
Telecommunications Industry Association	www.tiaonline.org
Wikipedia PoE	https://en.wikipedia.org/wiki/Power_over_Ethernet
Guidelines for Use Organizationally Unique Identifier	http://standards.ieee.org/develop/regauth/tut/eui.pdf
IEEE RA Assigned Ethertype List	http://standards-oui.ieee.org/ethertype/eth.txt
IEEE 802.1Q	https://en.wikipedia.org/wiki/IEEE_802.1Q
Ethernet Guitar	http://www.gibson.com/Products/Electric-Guitars/Les-Paul/Gibson-USA/HD-6X-Pro-Digital-Guitar/Features.aspx
Ethernet Office Lighting	http://luxreview.com/article/2015/07/here-s-what-will-really-make-led-office-lighting-take-off
PTP and NTP comparison	http://www.en4tel.com/pdfs/NTPandPTP-A-Brief-Comparison.pdf
The Switch Book by Rich Siefert	www.wiley.com/compbooks/seifert ISBN#: 0-471-34586-5
Gigabit Ethernet by Rich Siefert	ISBN#: 0-201-18553-9
IEEE Get Program 802 Standards	http://standards.ieee.org/about/get/

JUST IN CASE

WHEN – LAN STANDARDIZATION

- In February 1980, the IEEE started project 802 to standardize LANs. The "DIX-group" with Gary Robinson (DEC), Phil Arst (Intel), and Bob Printis (Xerox) submitted the so-called "Blue Book" CSMA/CD specification as a candidate for the LAN specification.
- In addition to CSMA/CD, Token Ring (IBM) and Token Bus (General Motors & Boeing) were also considered as candidates for a LAN standard. Competing led to strong disagreement over which technology to standardize. In December 1980, the group split into three subgroups, and standardization proceeded separately for each proposal.
- Delays in the standards process put the market introduction of the Xerox Star workstation and 3Com's Ethernet LAN products at risk. Considering the business implications David Liddle (General Manager, Xerox Office Systems) and Metcalfe (3Com) strongly supported a proposal of Fritz Röscheisen (Siemens) for an alliance in the emerging office communication market, including Siemens' support for the international standardization of Ethernet (April 10, 1981).
- Ingrid Fromm, Siemens' representative to IEEE 802, achieved broader support for Ethernet by the establishment of a competing Task Group "Local Networks" within the European standards body ECMA TC24. On March 1982, ECMA TC24 reached an agreement on a standard for CSMA/CD based on the IEEE 802 draft. Because the DIX proposal was most technically complete and because of the speedy action taken by ECMA which decisively contributed to the conciliation of opinions within IEEE, the IEEE 802.3 CSMA/CD standard was approved in December 1982. IEEE published the 802.3 standard as a draft in 1983 and as a standard in 1985.
- Approval of Ethernet on the international level was achieved by a similar, action with Fromm as the liaison officer working to integrate with International Electrotechnical Commission (IEC) Technical Committee 83 (TC83) and International Organization for Standardization (ISO) Technical Committee 97 Sub Committee 6 (TC97SC6). The ISO 8802-3 standard was published in 1989.

CLASSIFYING POWER DEMAND

Stages of powering up a PoE link

Stage	Action	Volts specified [V]	
		802.3af	802.3at
Detection	PSE detects if the PD has the correct signature resistance of 19–26.5 kΩ	2.7–10.1	
Classification	PSE detects resistor indicating power range (see below)	14.5–20.5	
Mark 1	Signals PSE is 802.3at capable. PD presents a 0.25–4 mA load.	—	7–10
Class 2	PSE outputs classification voltage again to indicate 802.3at capability	—	14.5–20.5
Mark 2	Signals PSE is 802.3at capable. PD presents a 0.25–4 mA load.	—	7–10
Startup	Startup voltage	> 42	> 42
Normal operation	Supply power to device	37–57	42.5–57

CLASSIFYING POWER DEMAND

Power levels available

Class	Usage	Classification current [mA]	Power range at PD [W]	Power from PSE [W]	Class description
0	Default	0–4	0.44–12.94	15.4	Classification unimplemented
1	Optional	9–12	0.44–3.84	4.00	Very Low power
2	Optional	17–20	3.84–6.49	7.00	Low power
3	Optional	26–30	6.49–12.95	15.4	Mid power
4	Valid for Type 2 (802.3at) devices, not allowed for 802.3af devices	36–44	12.95–25.50	30	High power
5	Valid for Type 3 (802.3bt) devices		–40	45	
6			–51	60	
7	Valid for Type 4 (802.3bt) devices		–62	75	
8			–71.3	99	
8+			–99.9		

CLASSIFYING POWER DEMAND

LLDP- MED Advanced Power Management

TLV Header		MED Header		Extended power via MDI			
Type (7 bits)	**Length (9 bits)**	**TIA OUI (3 octets)**	**Extended power via MDI subtype (1 octet)**	**Power type (2 bits)**	**Power source (2 bits)**	**Power priority (4 bits)**	**Power value (2 octets)**
127	7	00-12-BB	4	PSE or PD	Normal or Backup conservation	Critical, High, Low	0–102.3 W in 0.1 W steps

MED - Media Endpoint Discovery
MDI - Media Dependent Interface

The setup phases are as follows:

- PSE (provider) tests PD (consumer) physically using 802.3af phase class 3.
  - PSE powers up PD.
- PD sends to PSE: I'm a PD, max power = X, max power requested = X.
- PSE sends to PD: I'm a PSE, max power allowed = X.
  - PD may now use the amount of power as specified by the PSE.

The rules for this power negotiation are:

- PD shall never request more power than physical 802.3af class
- PD shall never draw more than max power advertised by PSE
- PSE may deny any PD drawing more power than max allowed by PSE
- PSE shall not reduce power allocated to PD, that is in use
- PSE may *request* reduced power, via conservation mode

UPCOMING POE ENHANCEMENTS

IEEE 802.3bt DTE Power via MDI over 4-pair

- As of February, 2017
- Split Type 3 & Type 4 to new clause (clause 145)
  - Support Operation of 10GBASE-T
  - Comply with SELV (Safety Extra Low Voltage) requirements as define in ISO/IEC 60950
  - Send 60w from PSE for 49w @ PD (Type 3)
  - Send 100w from PSE to get 70w @ PD (Type 4)
  - Uses all 4 pairs
  - Expected Current levels 850mA to 1000mA per pair
- Expect standard in early 2018

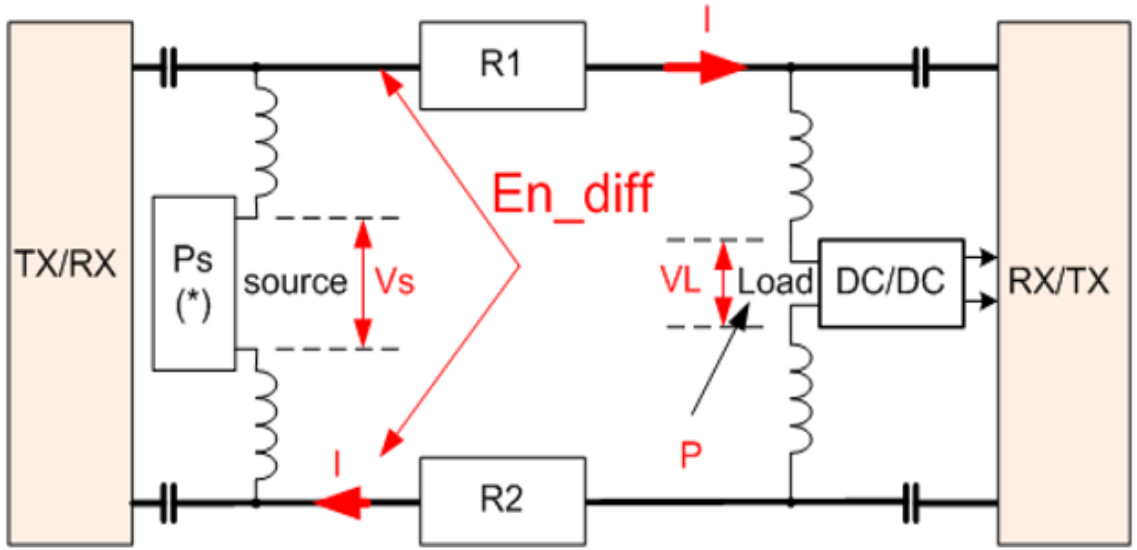
PROPOSED POE ENHANCEMENTS

IEEE 802.3bt DTE Power via MDI over 4-pair

Standard	IEEE 802.3af	IEEE 802.3at	IEEE 802.3bt	
	PoE	PoE+	4-pairs PoE or 4PPoE	
Type	1	2	3	4
Status	Released	Released	Expected early 2018	
Maximum number of energized pairs	2	2	4	4
Maximum DC current per pair	350 mA	600 mA	960 mA	960 mA
Maximum power delivered by the Power Sourcing Equipment (PSE)	15.4 watt	30.0 Watt	60.0 Watt	99.9 Watt
Minimum required power at the Powered Device (PD)	12.95 Watt	25.5 Watt	51.0 Watt	71.0 Watt

PROPOSED IEEE 802.3 BU POE SINGLE TP

1 Pair Power Over Data Implementation **Example**.



- The above power/data distribution technique and power/data interface is well known from XDSL applications and other telephony system for a long time. It works.
- Normally R1~=R2 for low CM noise
- DC Loop resistance=R=R1+R2
- *(*)Ps may be DC/DC converter fed by car battery after automotive protection circuitry block or connected to car battery w/o additional DC/DC block pending load power needs, cable loop resistance, Vs_min required to deliver max load power requirements and noise performance requirements*

HIGHEST SPEED PROPOSED STANDARD

IEEE 802.3bs 200 Gb/s & 400 Gb/s Ethernet Task Force

Task Force added 200 Gb/s capability to this project in May, 2016 to support a cost and performance optimized migration path to 400 Gb/s.

- Provide physical layer specifications, which support 200 Gb/s operation over:
  - At least 500 m of 4-lane parallel SMF
  - At least 2 km of SMF (4λ WDM duplex fiber)
  - At least 10 km of SMF (4λ WDM duplex fiber)

- Provide physical layer specifications, which support 400 Gb/s operation over:
  - At least 100 m of MMF (sixteen lanes/ 32 fibers total)
  - At least 500 m of SMF (four lanes / 8 fibers total)
  - At least 2 km of SMF (8λ WDM duplex fiber)
  - At least 10 km of SMF (8λ WDM duplex fiber)
  - Expect standard at end of 2017

HIGHER SPEED PROPOSED STANDARD

IEEE P802.3cd 50 Gb/s, 100 Gb/s, and 200 Gb/s Ethernet

NGOATH = Next Generation One And Two Hundred

50 Gb/s Ethernet PHYs

- Define single-lane 50 Gb/s PHYs for operation over
  - copper twin-axial cables with lengths up to at least 3m.
  - printed circuit board backplane with a total channel insertion loss of <= 30dB at 13.28125 GHz.
  - MMF with lengths up to at least 100m
  - SMF with lengths up to at least 2km
  - SMF with lengths up to at least 10km

100 Gb/s Ethernet PHYs

- Define a two-lane 100 Gb/s PHY for operation over
  - copper twin-axial cables with lengths up to at least 3m.
  - printed circuit board backplane with a total channel insertion loss of <= 30dB at 13.28125 GHz.
  - MMF with lengths up to at least 100m
  - SMF with lengths up to at least 500m
- Define a 100 Gb/s PHY for operation over SMF with lengths up to at least 2 km

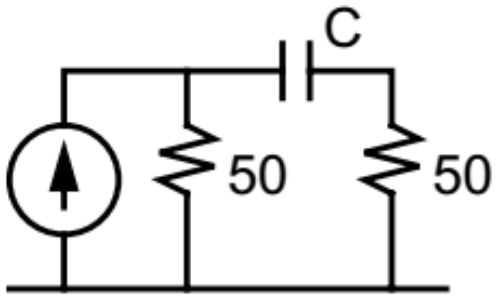
200 Gb/s Ethernet PHYs

- Define four-lane 200 Gb/s PHYs for operation over
  - copper twin-axial cables with lengths up to at least 3m.
  - printed circuit board backplane with a total channel insertion loss of <= 30dB at 13.28125 GHz.
- Define 200 Gb/s PHYs for operation over MMF with lengths up to at least 100m

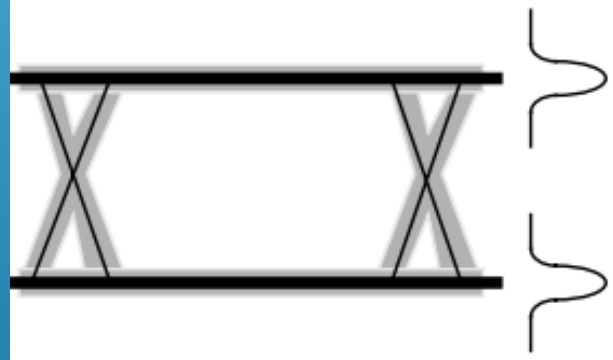
LINK AGGREGATION OBJECTIVES

Objective	Description
Increased bandwidth	capacity of multiple links combined into one logical link
Linearly incremental bandwidth	Bandwidth can be increased in unit multiples instead of the order-of-magnitude increase available through Physical Layer options
Increased availability	failure of a single link within a LAG need not cause failure from the perspective of an Aggregator Client
Load sharing	Aggregated traffic is distributed across multiple links
Automatic configuration	In the absence of manual overrides, a set of LAGs is automatically configured, & individual links are allocated to those groups.
Rapid configuration and reconfiguration	For changes in physical connectivity, Link Aggregation will converge to a new configuration, typically in milliseconds for link down events and 1 s or less for link up events.
Deterministic behavior	aggregation can be independent of the order of events
Low risk of duplication or misordering	For both steady-state operation and link (re)configuration
Support of existing MAC Clients	No changes required to higher-layer protocols or applications
Backwards compatibility	Links that cannot take part in Link Aggregation operate as normal, individual links.
Accommodation of differing capabilities and constraints	Devices with differing hardware and software constraints are accommodated to the extent possible.
No change to frame formats	Link aggregation neither adds to, nor changes the contents of frames exchanged between Aggregator Clients
Network management support	standard specifies management objects for configuration, monitoring, and control
Dissimilar MACs	Link Aggregation can be used over any medium offering the Internal Sublayer Service.

Baseline Wander



A typical laser system will require a coupling capacitor at both TX and RX



Eye diagram shows Gaussian wander due to statistical fluctuation of 1's and 0's charging coupling capacitor C

The average value of N random bits is 1/2 with a sigma of 1/sqrt(N)

A coupling capacitor can be approximately considered to be performing a moving average over N bits, where N ~= 2*PI*RC/(Tbit)

Therefore baseline wander is well approximated by a gaussian distribution with a sigma proportional to 1/sqrt(RC)

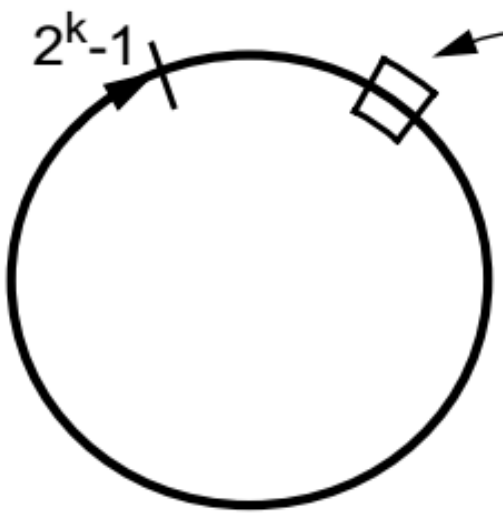
Scrambler selection

jamming probability analysis

Expected time for 64-bit run-length due to random data:

$$MTTF(random) = \frac{2^{64}}{2 \cdot 10.3\,GHz} \approx 29 years$$

Expected time for 64-bit run-length due to jamming:



1500byte packet (26 byte overhead)
gives 94 chances to "hit" scrambler

$$MTTF(jamming) = \frac{(2^{k}-1) \bullet 1526 \bullet 8}{94 \bullet (2) \bullet 10.3\,GHz}$$

k=57 gives MTTF of 29 years