Ethernet

Parikshit Godbole
High Performance Computing - Technologies
C-DAC, Pune

Scope

- Networking Basics
- Ethernet:
  - Description
  - Layering
  - MAC detailing
  - Differences between 10/100/1000 operations
  - Other: Cabling, auto-negotiation
  - Designing with MII
- Ethernet equipment design
  - Hubs
  - Switches
  - Network Interface Card
- Future architectures – 10G, 40/100G Ethernet

Networking Basics

- LAN and WAN
  - Local area network consists of few tens to few hundreds of hosts connected together using similar networking technology and topology (e.g. Ethernet)
  - Wide area network consists of a very high number of hosts connected together using different networks. They are connected together using WAN links working at low to very high speeds (e.g. Internet)
  - WAN consists of hierarchical network components required to form a seamless interconnect across these diverse network technologies/topologies

Network Switching

- Circuit switching

  In a circuit-switched network, a dedicated physical path is established through the network and is held for as long as communication is necessary.

  e.g. A telephone network

- Packet Switching

  A packet-switched network, on the other hand, routes data in small pieces called packets, each of which proceeds independently through the network.

  e.g. LAN

Currently which mode are we communicating?

Local Area Network

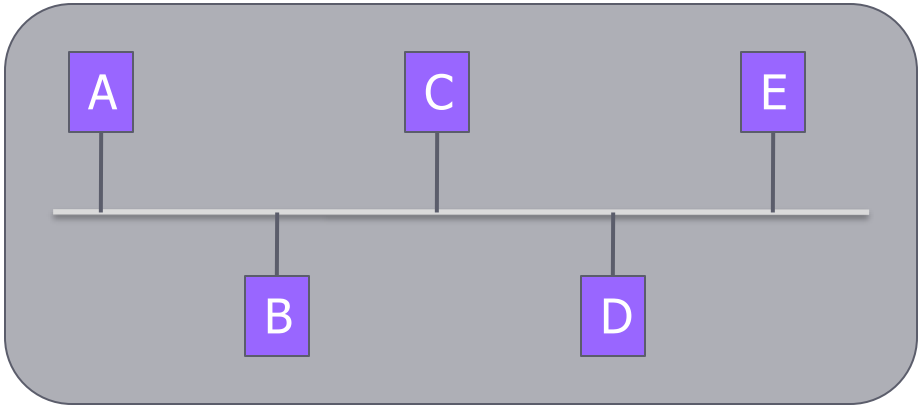
- LAN was devised as alternative to expensive, dedicated point-to-point connection.
- Initially,each LAN consists of a single shared medium, usually a cable, to which many hosts attach.
- Hosts take turns using medium to send packets.
- Many LAN technologies have been developed. To help understand similarities, each network is classified into a category according to its topology or general shape.

Network Topology

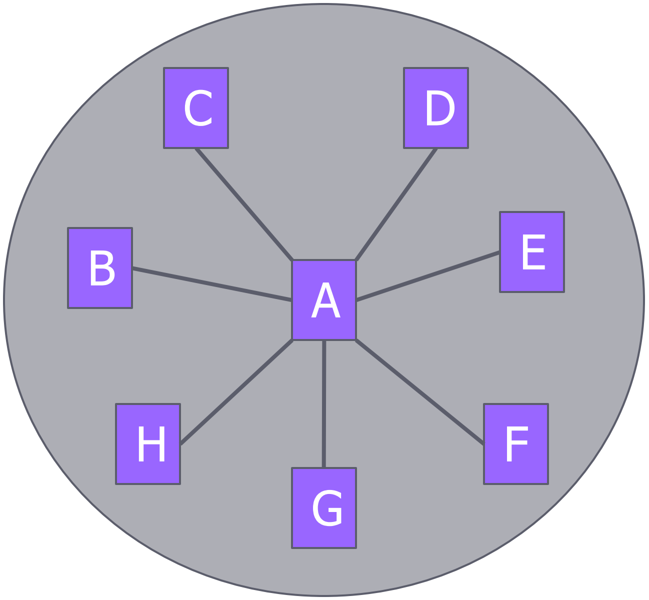
- It is the shape or physical connectivity of the network.
- Major goals when establishing topology
  - Route the traffic across the least cost path within the network.
  - Give the end user the best possible response time and throughput.
    - Latency and Bandwidth are two figures of merit.
- Each topology has advantages & disadvantages.

Common Topologies

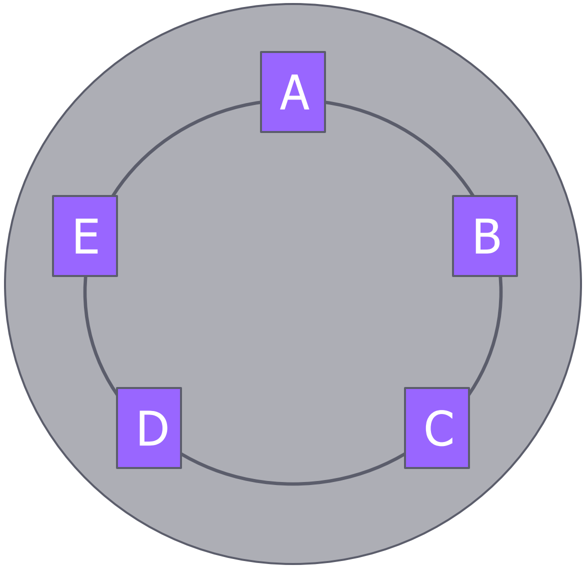
Bus



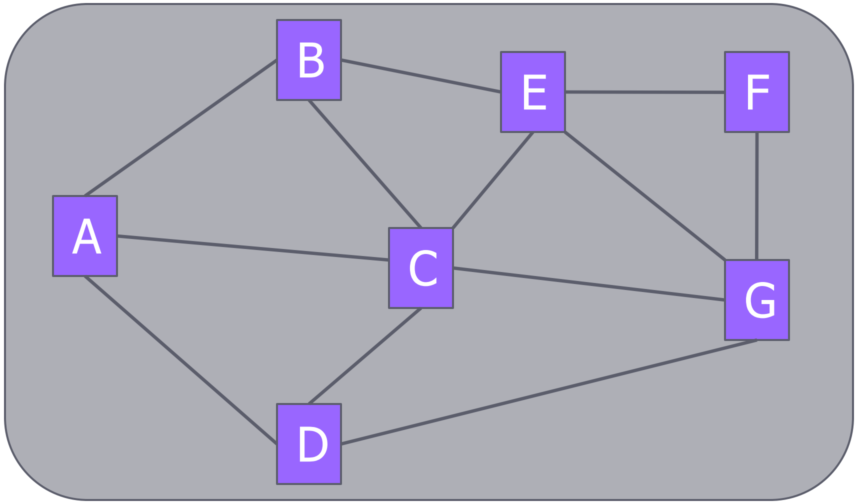
Star



Ring

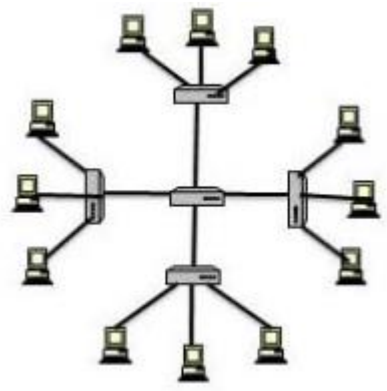


Mesh

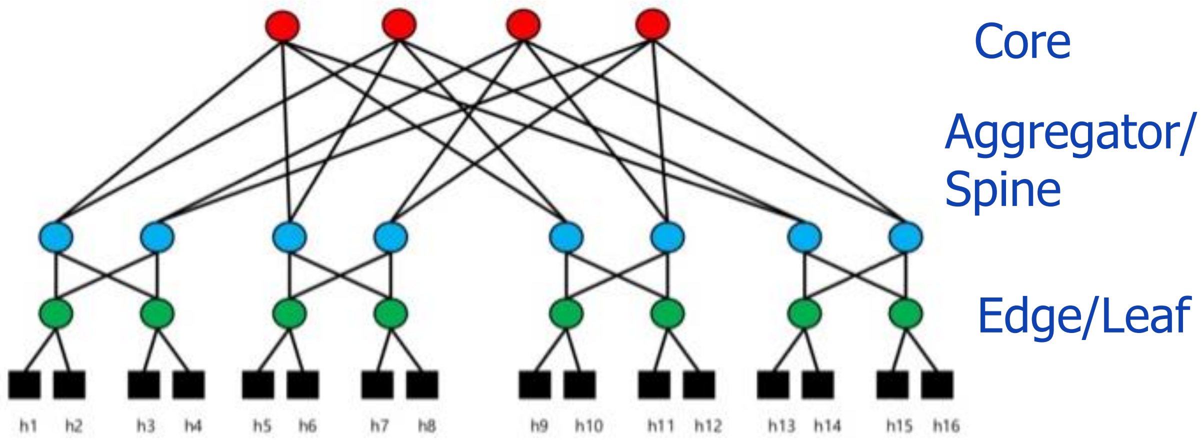


Common Topologies

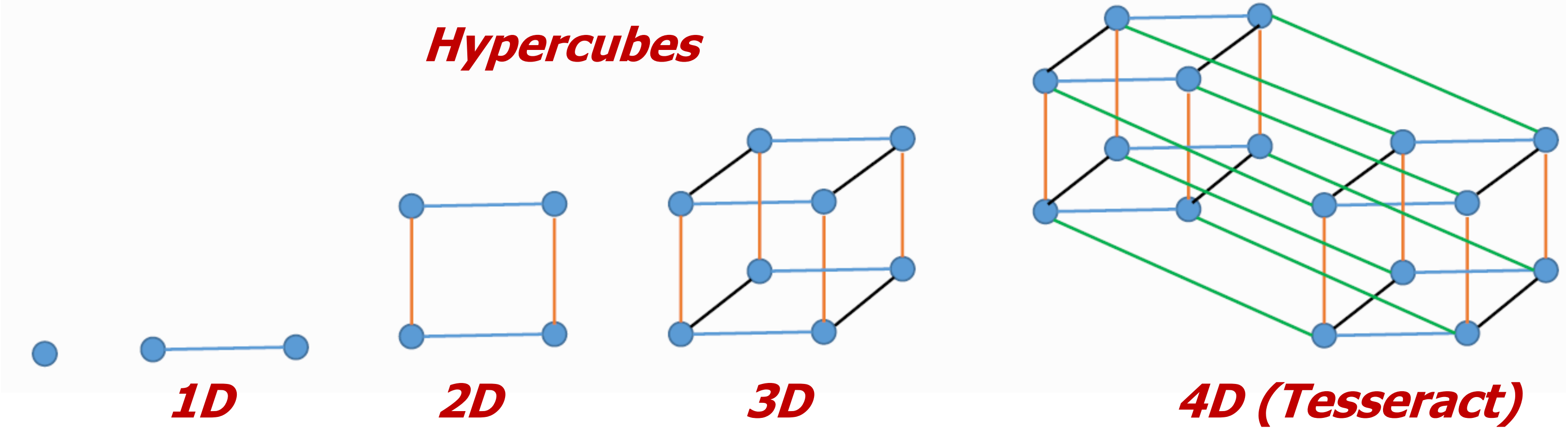
Extended Star



FAT Tree/ CLOS

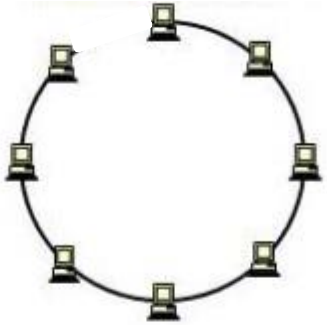


Hypercubes

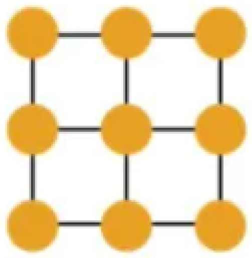


Common Topologies

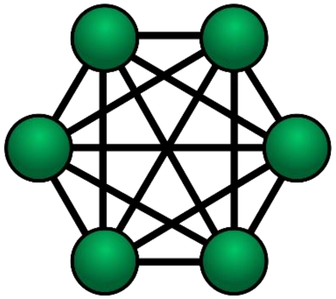
1-D Mesh



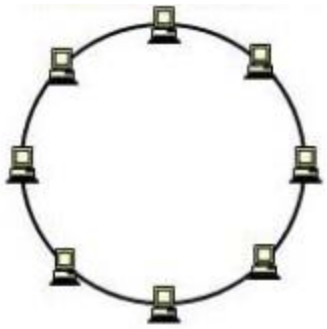
2-D Mesh



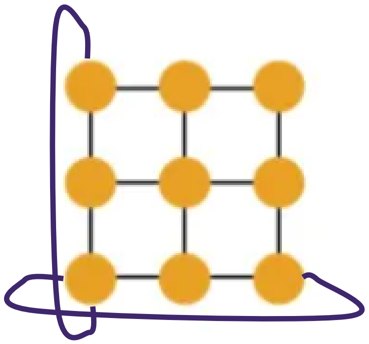
Fully Connected



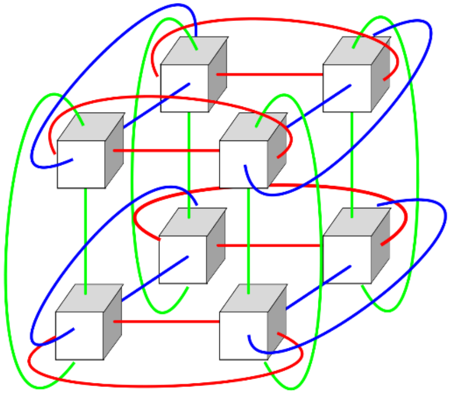
1-D Torus Ring



2-D Torus



3-D Torus

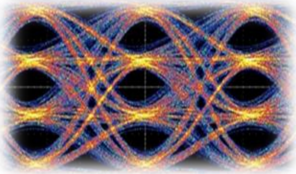


Why Ethernet?

- Several LAN technologies have been developed over last decades. However, Ethernet has surpassed them all to become a de-facto standard of LAN today
  - Over 85% of today’s LAN use Ethernet
- Factors behind popularity of Ethernet are
  - Seamless performance (10/100/1G/..)
  - Price/performance
  - Inter-operability
  - Scalability

What is Ethernet?

- The term ‘Ethernet’ today refers to family of LAN products covered by IEEE 802.3 standard which defines class of networks defined by CSMA/CD protocol (Carrier Sense multiple access/Collision detect)
- Ethernet is commonly known by the raw data speeds and is divided into following
  - 10 Mbps (10BaseT) : IEEE 802.3 (1985)
  - 100 Mbps (Fast Ethernet) : IEEE 802.3u (1995)
  - 1 Gbps (Gigabit Ethernet) : IEEE 802.3z (1998)
  - 10 Gbps (10G Ethernet) : IEEE 802.3ae (2004)
  - 40/100 Gbps : IEEE 802.3ba (2010 and 2014)
  - 200/400 Gbps : IEEE 802.3bs (2017)
  - Now 802.3bs (Electrical) and 802.3cd(Optical) covers 50Gbps, 100Gbps, 200Gbps and 400Gbps using PAM4

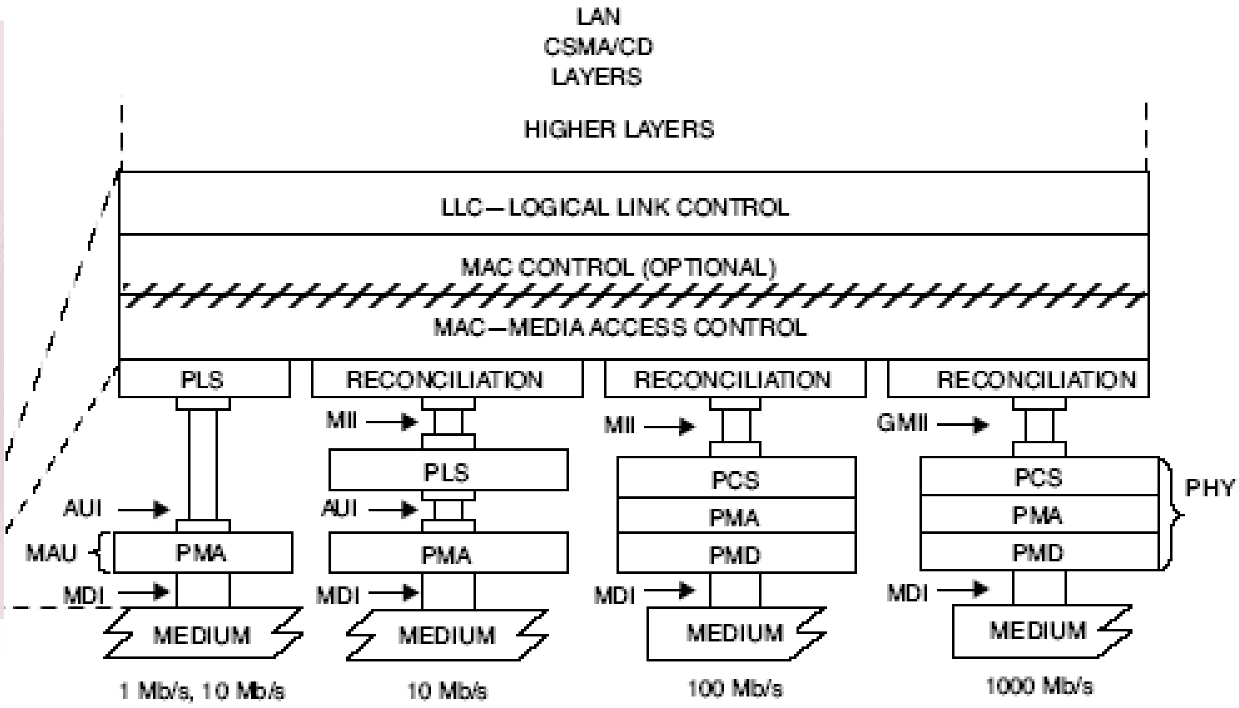


Brief History

- Inventor of Ethernet is Robert Metcalfe (Xerox) labs. The initial network consisted of a thick cable supporting ~3mbps rates. The network was called “Aloha” n/w and later renamed to “Ethernet” to signify that it could exist anywhere
- In 1980, three companies (Xerox, Digital and Intel) jointly developed 10mbps Ethernet specification V1.0
- With minor changes, IEEE adopted this standard as draft in 1983 and came out with ratified standard (802.3) in 1985
- This std continues to evolve to support new technologies, enhance speeds and support various media

IEEE 802.3

© art.com

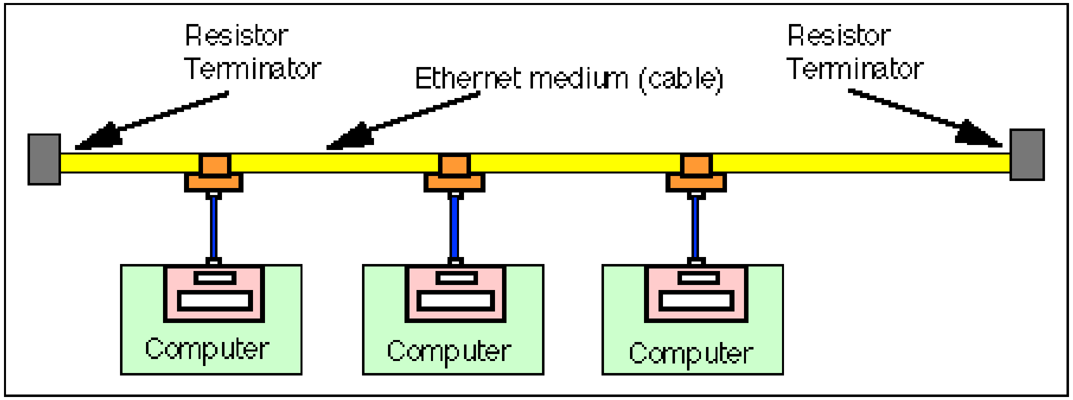


AUI = ATTACHMENT UNIT INTERFACE
MDI = MEDIUM DEPENDENT INTERFACE
MII = MEDIA INDEPENDENT INTERFACE
GMII = GIGABIT MEDIA INDEPENDENT INTERFACE
MAU = MEDIUM ATTACHMENT UNIT

PLS = PHYSICAL LAYER SIGNALING
PCS = PHYSICAL CODING SUBLAYER
PMA = PHYSICAL MEDIUM ATTACHMENT
PHY = PHYSICAL LAYER DEVICE
PMD = PHYSICAL MEDIUM DEPENDENT

Topology used for 802.3

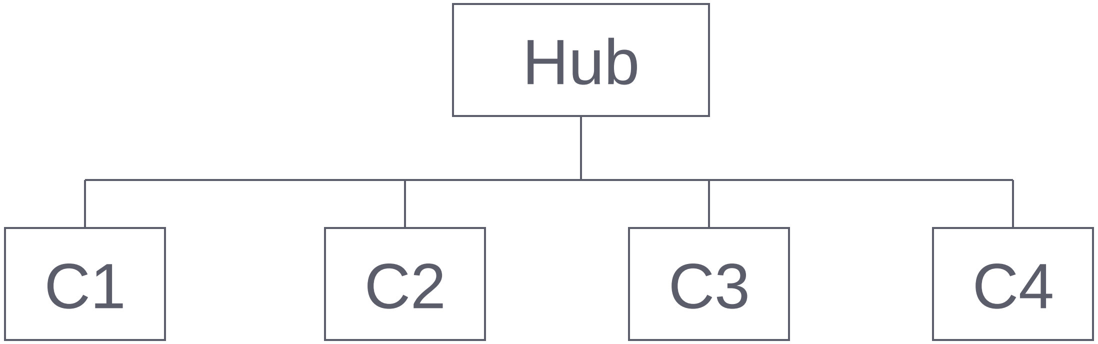
- Bus Topology



- A single back-bone cable.
- End-to-End length of cable is called “Ethernet segment”.
- A break in any segment will disable the entire network.

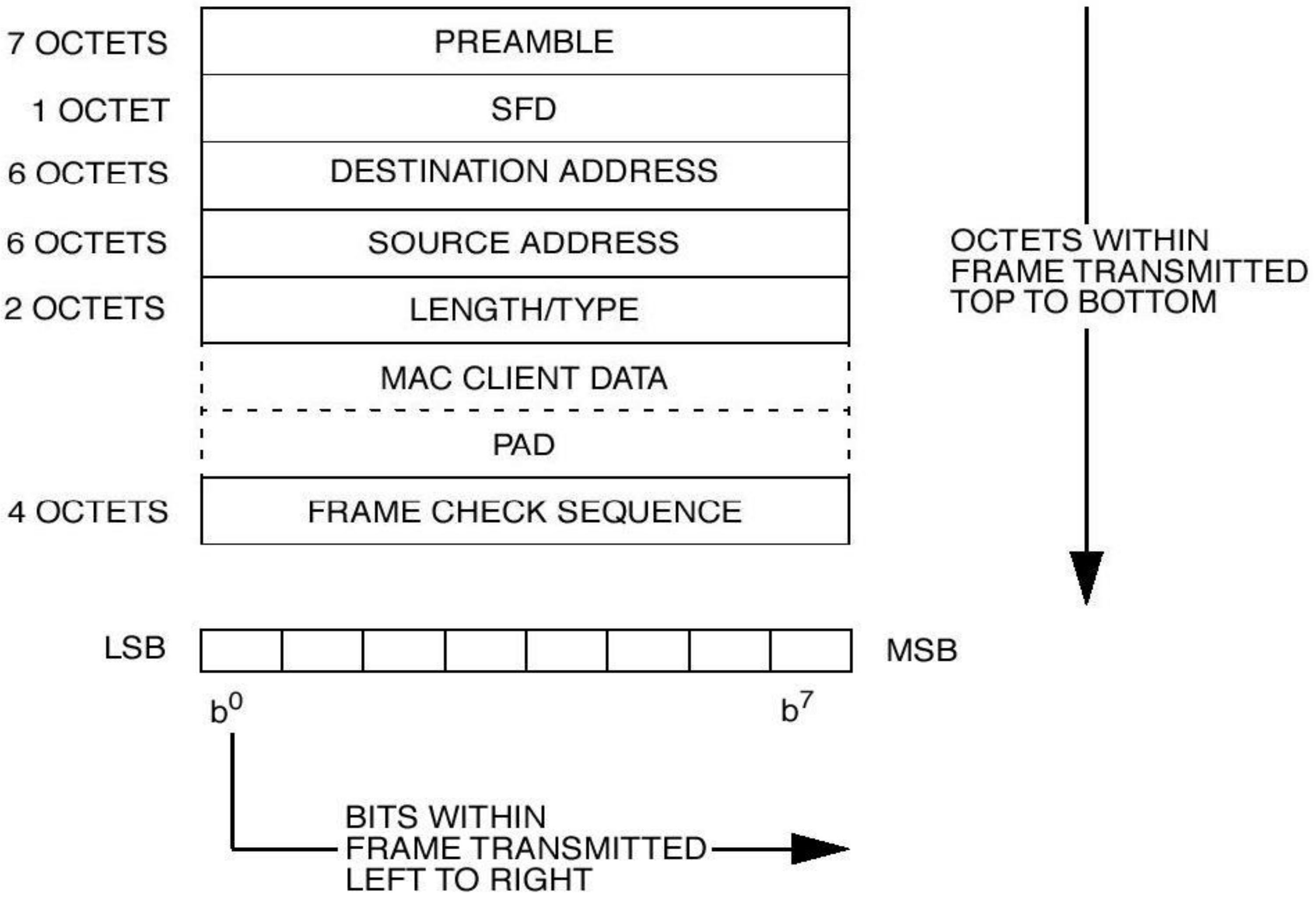
Topology used for 802.3

- Star Topology



- All nodes are connected to central HUB.
- A break in any segment will disable the node connected to that segment.
- More amount of cabling is required.
- Hub is equipment emulating a shared media

Ethernet Frame



Frame Format

- Preamble (7 Bytes)
  This field allows the physical layer device to synchronize itself to the arriving data packet.(10101010)
- Start Frame Delimiter (1 Byte)
  This character indicates start of the frame.(10101011)
- Destination Address Field (6 Bytes)
  The MAC address of the receiving node.
- Source Address Field (6 Bytes)
  The MAC address of the transmitting node.

Frame Format

- Length/Type Field (2 Bytes)
  The number of data bytes or type of packet.

- Data Field (0 to 1500 Bytes)
  The payload

- Padding
  For a proper operation of the CSMA/CD protocol a minimum frame size is prescribed. If the frame size is less than this figure then extra bytes are padded to the data .

- Frame Check Sequence (4 bytes)
  A 32 bit cyclic redundancy check value of all fields except preamble, SFD and FCS .

MAC Address

- MAC address is a unique 48 bit address assigned for each network device on the Ethernet represented in hex

  e.g.: 12:34:56:78:90:AB

  - Each manufacturer can apply to IEEE and get a unique set of MAC addresses for his products. IEEE provides the first 24 bits of the address. (Organizationally unique identifiers: OUIs)
    - E.g. CDAC 00:A0:22:XX:XX:XX
- The Broadcast address has all of its bits as ‘1’s
- A multicast address is associated with a logical grouping of nodes. This needs to be implemented at a higher level.

MAC

- The MAC sublayer defines a medium-independent facility, built on the medium-dependent physical facility provided by the Physical Layer, and under the access-layer-independent LAN LLC sublayer (or other MAC client).
- Functions generally associated with MAC
  - Data encapsulation (transmit and receive)
    - 1) Framing (frame boundary delimitation, frame synchronization)
    - 2) Addressing (handling of source and destination addresses)
    - 3) Error detection (detection of physical medium transmission errors)
  - Media Access Management
    - 1) Medium allocation (collision avoidance)
    - 2) Contention resolution (collision handling)

MAC

- 802.3 provides for two modes of operation of the MAC sublayer:
  - In *half duplex* mode, stations contend for the use of the physical medium, using the CSMA/CD algorithms specified.
  - The *full duplex* mode.

MAC in half-duplex mode

- All devices on the network have equal-priority access to the medium.
- Multiple nodes may simultaneously receive data from the medium but only one node can transmit at a time.
- The technique used to arbitrate is called Carrier Sense Multiple Access with collision detection (CSMA/CD).
- A node wishing to send data, first “listens” the medium.
- If some activity is going on then the node will defer its transmission until the activity ceases and a predetermined period of silence passes.
- This period of inactivity is known as IPG (inter packet gap)

MAC in half-duplex mode

- The IPG delineates each packet and allows all stations to detect carrier sense as inactive (IPG value is 96 bit times)
- If two or more nodes simultaneously starts transmitting then a Collision occurs.
- Each transmitting node monitors for "collision" and if detects one, stops immediately and sends a 32 bit jamming sequence.
- Jam period guarantees that stations at the extremes of the network are able to detect collision condition.
- If the collision is detected during preamble , Preamble/SFD sequence is completed and then Jam sequence is sent.

MAC in half-duplex mode

- After collision the MAC retries until either it is successful or a maximum number of attempts have been made and all have terminated due to collisions.
- The scheduling of the retransmissions is determined by a controlled randomization process called “truncated binary exponential backoff”.
- The delay is integer multiple of slot time.
- The number of slot times to delay before the nth (n= k) retransmission attempt is chosen as a uniformly distributed random integer r in the range:

$0 <= r < 2^k$

where $k = \min(n, 10)$

MAC in half-duplex mode

- If all attempts fail, this event is reported as an error.
- Algorithms used to generate the integer r should be designed to minimize the correlation between the numbers generated by any two stations at any given time.
- A round-trip delay called Slot time determines how long it takes to detect a collision.
- Slot time is fixed as 512 bit times for Ethernet.
- If a collision is detected after slot time then its called as Late collision and transmission is aborted immediately.

CSMA/CD MAC Layer Functionality

- Frame transmission
  - Accepts data from the LLC layer and constructs a frame
  - Presents a bit stream to the physical layer for transmission
- Frame Reception
  - Receives a bit stream from the physical layer
  - Presents the “data” in the frame to LLC.
  - Discards invalid MAC frames and frames not addressed to it
- Defers the transmission when medium is busy.

CSMA/CD MAC Layer Functionality (Tx)

- Appends preamble, SFD, DA, SA, length and FCS to all frames and inserts pad field for frames whose length is less than the minimum value.
- Delays transmission for specified interframe gap.
- Halts transmission when collision is detected.
- Enforces collision to ensure propagation throughout the network by sending a jamming message.
- Schedules retransmission after a collision until a specified retry limit is reached.

CSMA/CD MAC Layer Functionality (Rx)

- Verifies full octet (byte) boundary alignment.
- Discards invalid MAC frames received
  - frame length is inconsistent with length field
  - frame is not an integral number of octets in length.
  - CRC error
  - Runt frames(frames having length less than the minimum frame length)
- Removes preamble, SFD, DA, SA, length, FCS and padding before presenting it to the LLC.

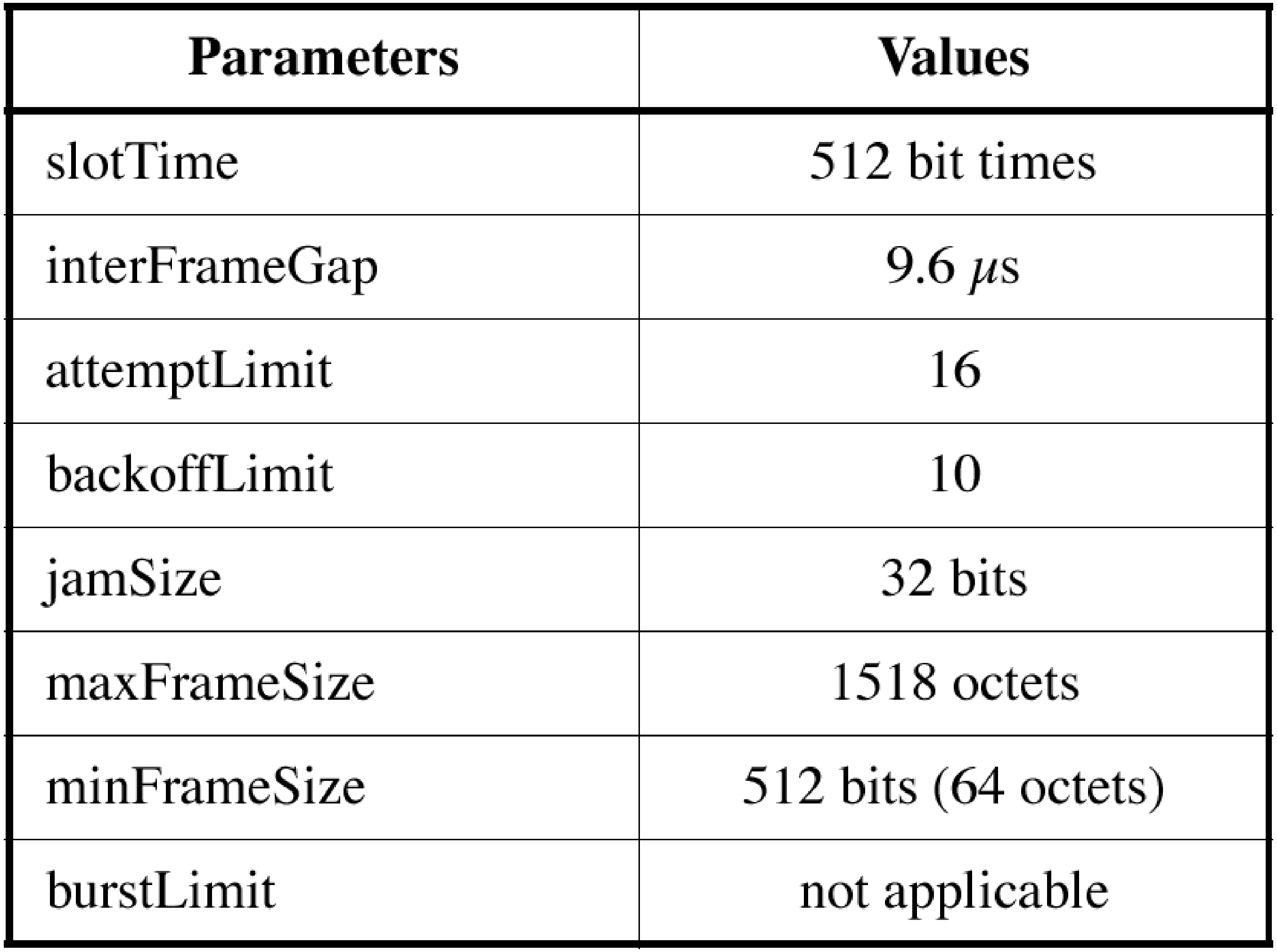
Physical Layer Functionality

- Accepts data from the MAC and decodes it.
- Gives the MAC status of link e.g. Carrier Detect, collision detect etc.
- Clock recovery
- Media dependent driver circuit.

Parameters Description

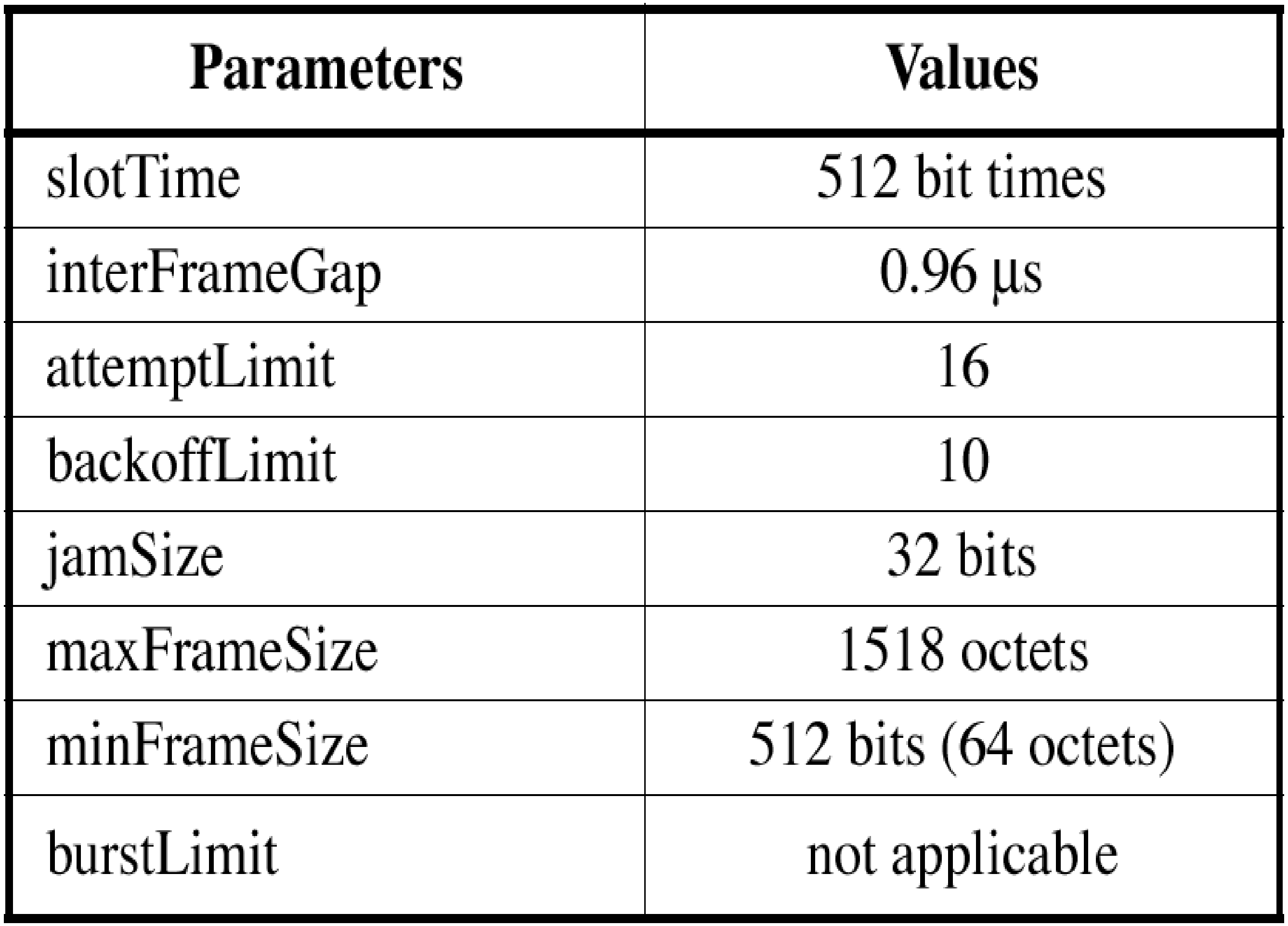
- Slot time is decided based upon the round trip delay.
- Minimum frame is also decided by the round trip delay.
  - For 10 and 100 mbps above two parameters are equal.
- Round trip delay determines the physical span of the network
- Attempt limit specifies the maximum number of retries.
- Back off limit specifies the maximum amount of time the MAC will wait before it retries.
- Maximum frame size is from Destination address field to CRC.

Parameters for 10Mbps



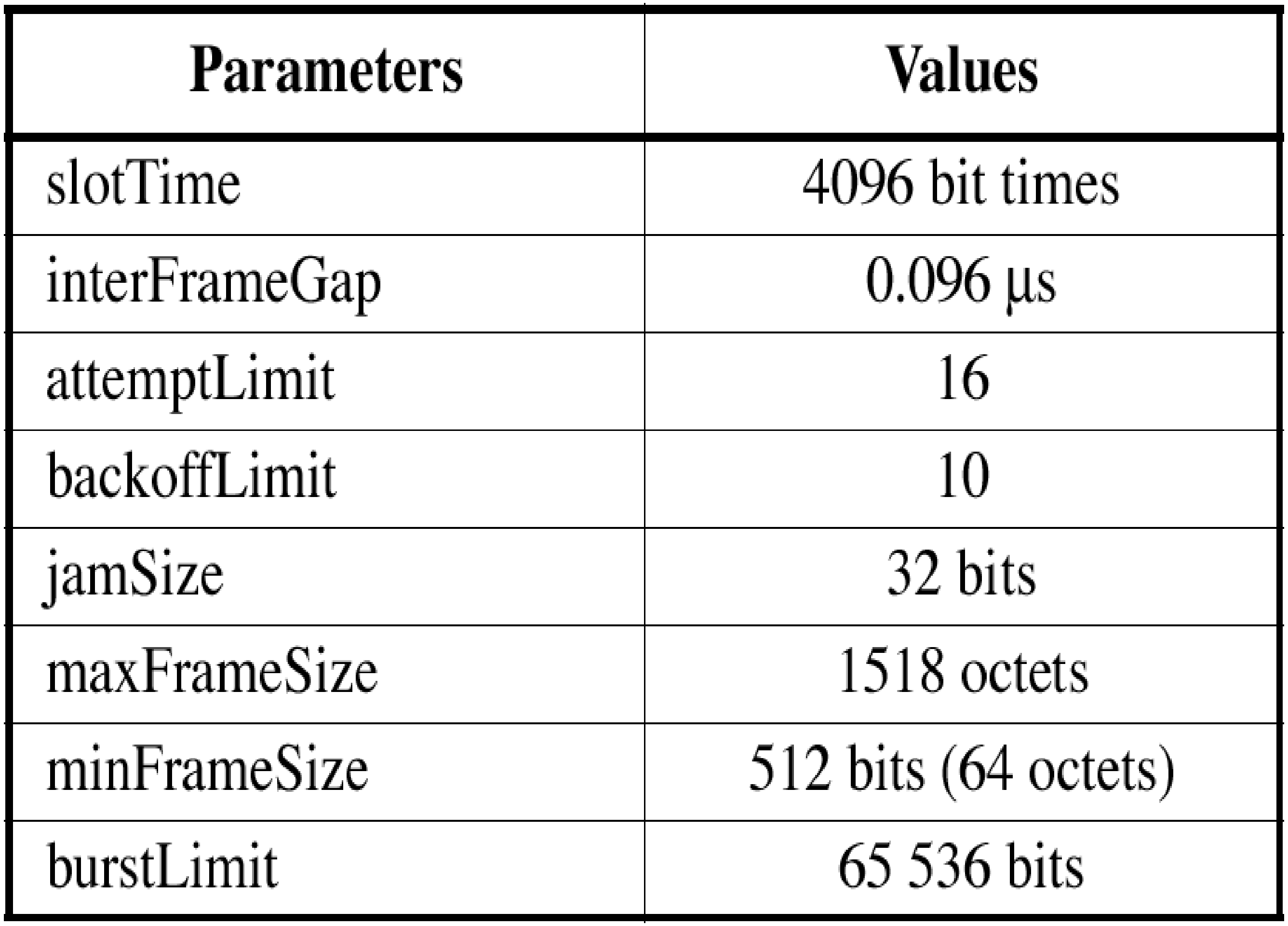
Parameters	Values
slotTime	512 bit times
interFrameGap	9.6 μs
attemptLimit	16
backoffLimit	10
jamSize	32 bits
maxFrameSize	1518 octets
minFrameSize	512 bits (64 octets)
burstLimit	not applicable

Parameters for 100 Mbps



Parameters	Values
slotTime	512 bit times
interFrameGap	0.96 µs
attemptLimit	16
backoffLimit	10
jamSize	32 bits
maxFrameSize	1518 octets
minFrameSize	512 bits (64 octets)
burstLimit	not applicable

Parameters for 1000Mbps



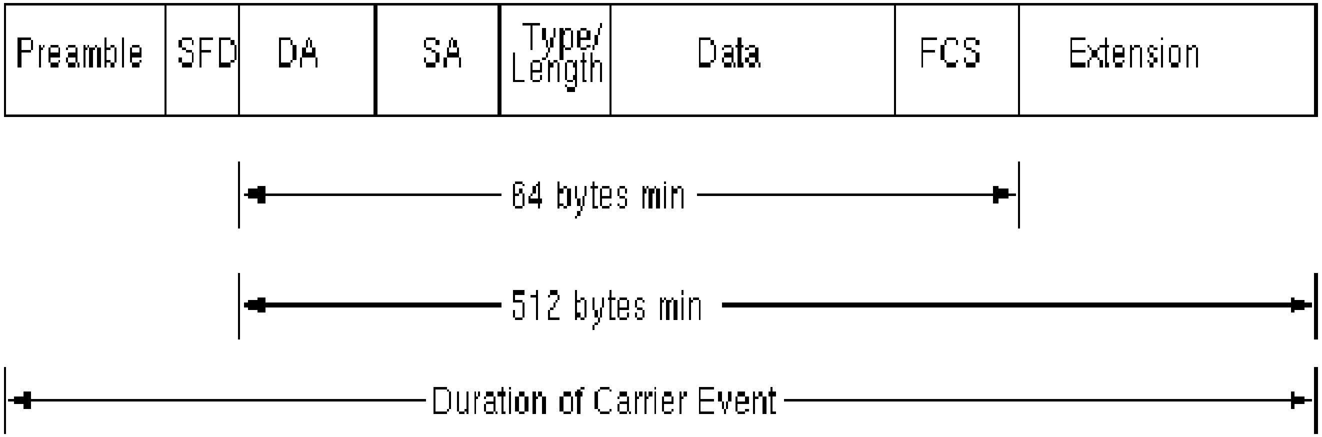
Parameters	Values
slotTime	4096 bit times
interFrameGap	0.096 µs
attemptLimit	16
backoffLimit	10
jamSize	32 bits
maxFrameSize	1518 octets
minFrameSize	512 bits (64 octets)
burstLimit	65 536 bits

Gigabit Ethernet

- Going from 10Mbps to 100Mbps the network diameter is reduced from 2Km to 200 m.
- If this were continued then for gigabit the length will come down to 20 m.
- So the work around is to fix the distance as 200 m and slot time as 512 bytes (4096 bits).
- Minimum packet length is still 64 bytes to have backward compatibility with lower-speed networks.
- Here the physical layer transmits a special signal carrier extension if frame size is less than the slot time.
- These special symbols are transmitted after the FCS.
- Special symbols are not considered part of frame and are handled in a special way at receiver.

Gigabit Ethernet

- Gigabit frame



Preamble | SFD | DA | SA | Type/ Length | Data | FCS | Extension

64 bytes min

512 bytes min

Duration of Carrier Event

SFD : Start of Frame Delimiter
DA : Destination Address
SA : Source Address
FCS : Frame Check Sequence

Gigabit Ethernet

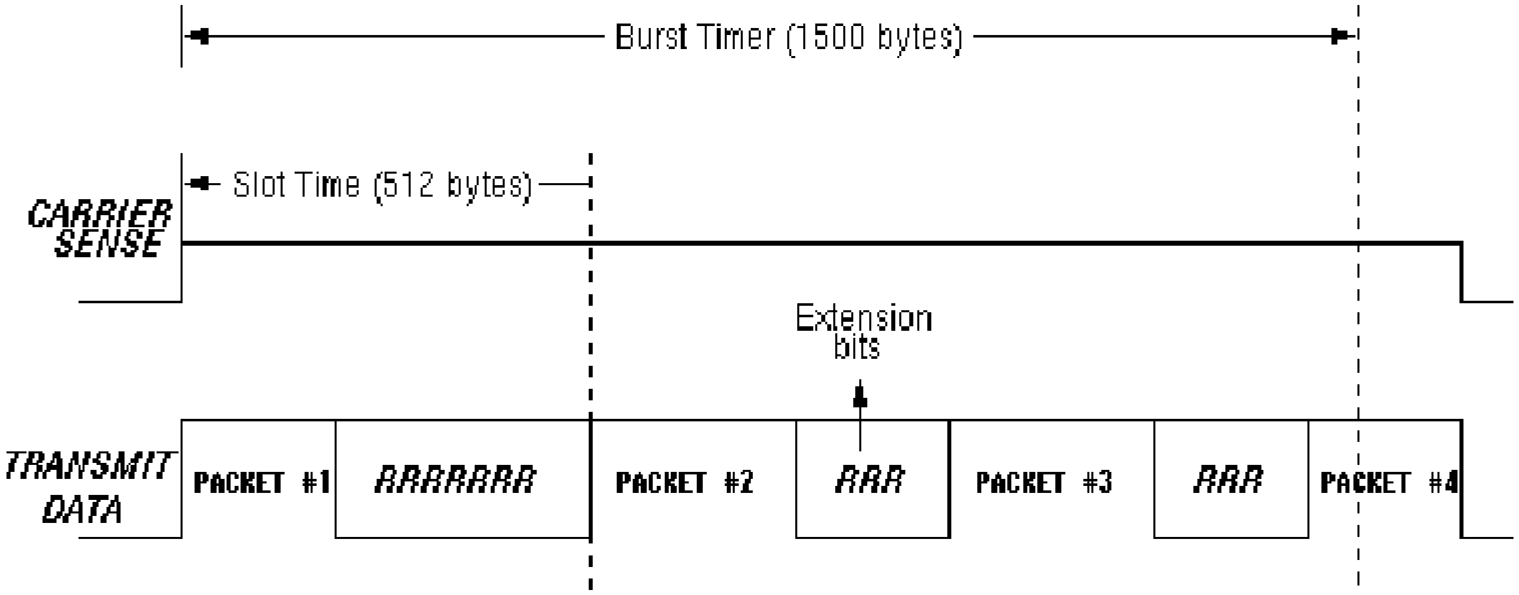
- Carrier extension is simple solution but it wastes the bandwidth.
- Up to 448 bytes are send as padding for small packets.
- Infact for large no of small packets the throughput is slightly better than Fast Ethernet.

Packet bursting:

- When a station has a number of packets to transmit, the first packet is padded to the slot time if necessary using carrier extension.
- Subsequent packets are transmitted back to back, with the minimum Inter-packet gap (IPG) until a burst timer (of 1500 bytes) expires.

Gigabit Ethernet

- Packet bursting



Full Duplex Mode

- This is a optional mode of operation allowing simultaneous communication between a pair of devices using point-to-point media segments that provides independent transmit and receive path.
- The aggregate capacity of full-duplex link is double.
- Segment length is no longer limited by timing requirements of shared channel half-duplex Ethernet.
- No need of CSMA/CD since there are exactly two stations connected with full duplex point-to-point link.
  - Flow control becomes essential to ensure optimal flow of data
- So this is supported between two hosts or between a host and a Ethernet Switch.

Full Duplex Mode

- Media systems that support full duplex mode:
  10BASE-T, 10BASE-FL, 100BASE-TX, 100BASE-FX, 100BASE-T2, 100BASE-X.
- Media systems that do not support full duplex mode:
  10BASE5, 10BASE2, 10BASE-FP, 10BASE-FB, 100BASE-T4.

Ethernet Physical Media

- **10BASE2**

  10 Mb/s, RG 58 coaxial cable.

  185m and 30 nodes.
- **10BASE5**

  10 Mb/s, coaxial cable (thicknet).

  500m and 100 nodes.
- **10BASE-F:**

  10 Mb/s, fiber optic cable.

  2000m.
- **10BASE-T**

  10 Mb/s, two pairs of twisted-pair telephone wire.

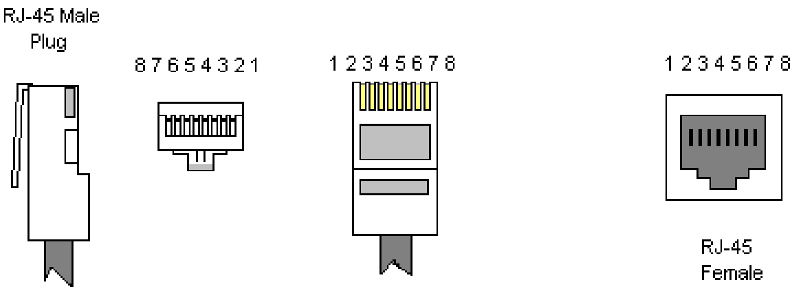
Media

- **100BASE-FX:** 100 Mb/s, 2 optical fibers,2000m.
- **100BASE-T2:** 100 Mb/s, 2 pairs of Category 3 or better balanced cabling, 100m.
- **100BASE-T4:** 100 Mb/s 4 pairs of Category 3, 4, and 5 unshielded twisted-pair (UTP) wire,100m.
- **100BASE-TX:** 100 Mb/s 2 pairs of Category 5 unshielded twisted-pair (UTP),100m.
- **100BASE-T:** 1000 Mb/s 4 pairs of Category 5 unshielded twisted-pair (UTP),100m.
- **1000BASE-CX:** 1000BASE-X over specialty shielded balanced copper jumper cable assemblies.

Media

- **1000BASE-LX:** 1000BASE-X using long wavelength laser devices over multimode and single-mode fiber.
- **1000BASE-SX:** 1000BASE-X using short wavelength laser devices over multimode fiber.

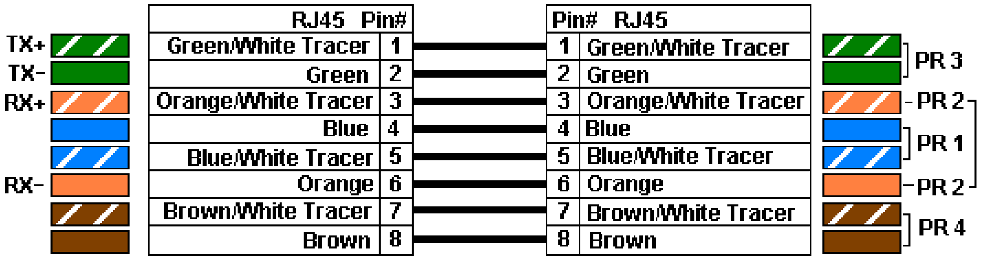
10/100/1000 Connectors



RJ45

Color Standard
EIA/TIA T568A

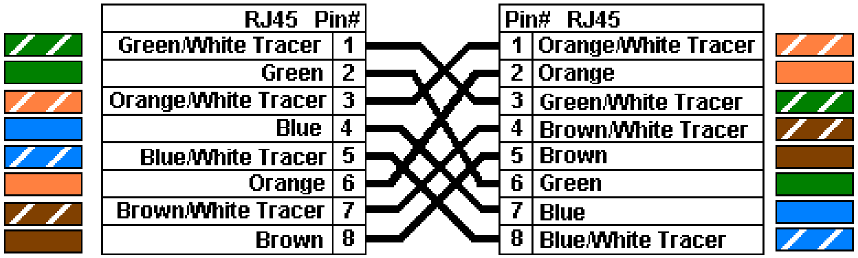
Ethernet Patch Cable



		RJ45	Pin#	Pin#	RJ45		
TX+	▨	Green/White Tracer	1	1	Green/White Tracer	▨	PR 3
TX-	🟩	Green	2	2	Green	🟩	
RX+	▨	Orange/White Tracer	3	3	Orange/White Tracer	▨	PR 2
	🟦	Blue	4	4	Blue	🟦	PR 1
	▨	Blue/White Tracer	5	5	Blue/White Tracer	▨	
RX-	🟧	Orange	6	6	Orange	🟧	PR 2
	▨	Brown/White Tracer	7	7	Brown/White Tracer	▨	PR 4
	🟫	Brown	8	8	Brown	🟫	

Color Standard
EIA/TIA T568A

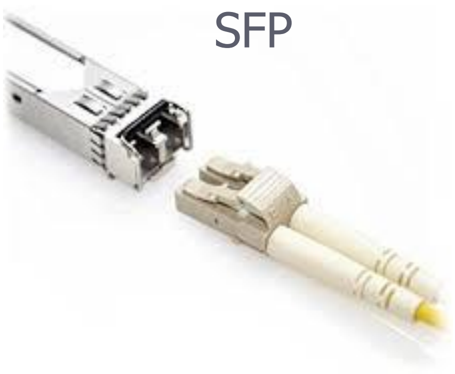
Ethernet Crossover Cable



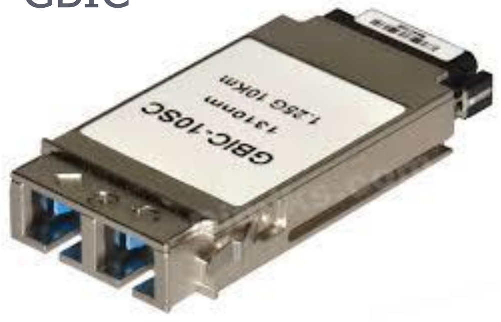
	RJ45	Pin#	Pin#	RJ45	
▨	Green/White Tracer	1	1	Orange/White Tracer	▨
🟩	Green	2	2	Orange	🟧
▨	Orange/White Tracer	3	3	Green/White Tracer	▨
🟦	Blue	4	4	Brown/White Tracer	▨
▨	Blue/White Tracer	5	5	Brown	🟫
🟧	Orange	6	6	Green	🟩
▨	Brown/White Tracer	7	7	Blue	🟦
🟫	Brown	8	8	Blue/White Tracer	▨

"A" is earlier

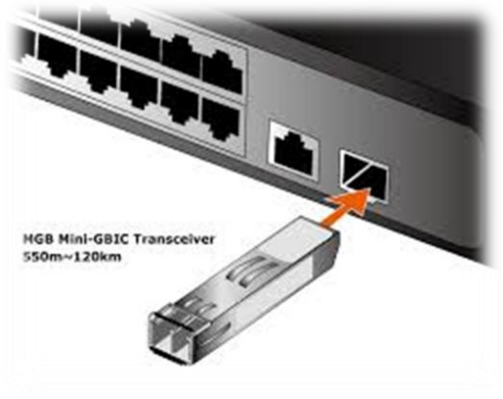
Gigabit connectors

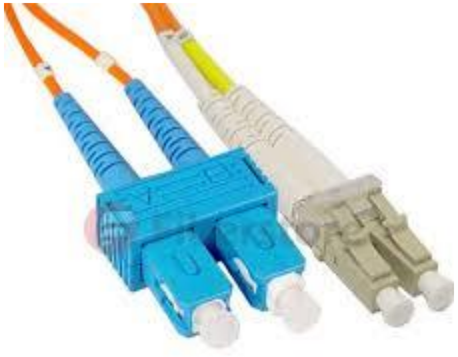


SFP



GBIC





SC LC

10Gbps options

SFP+

SFP+

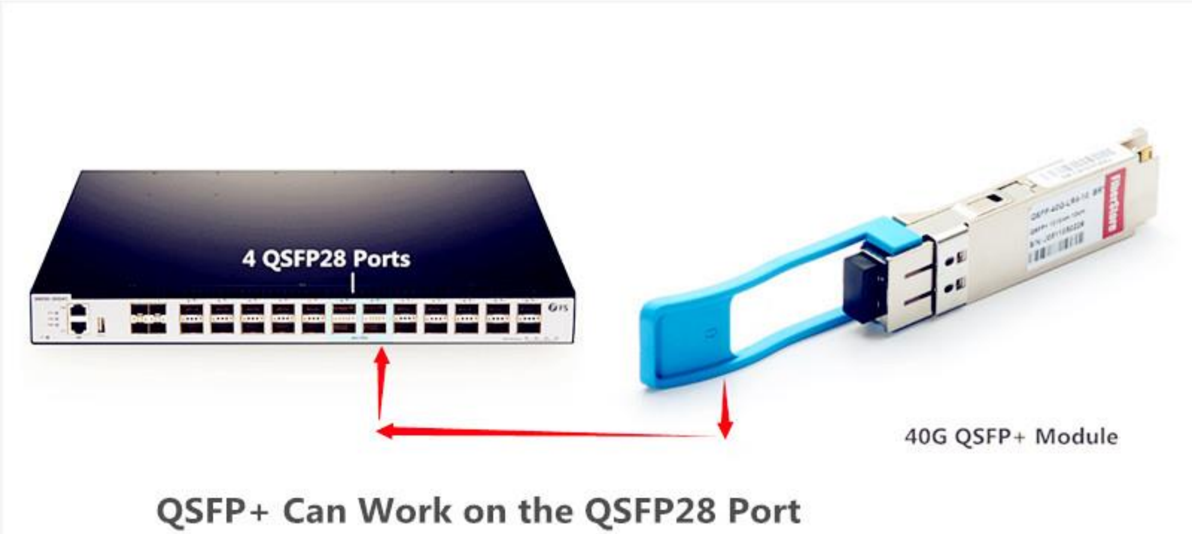
XFP

SFP/SFP+ to RJ45

XFP to CX4

CX4

40/100/200Gbps options



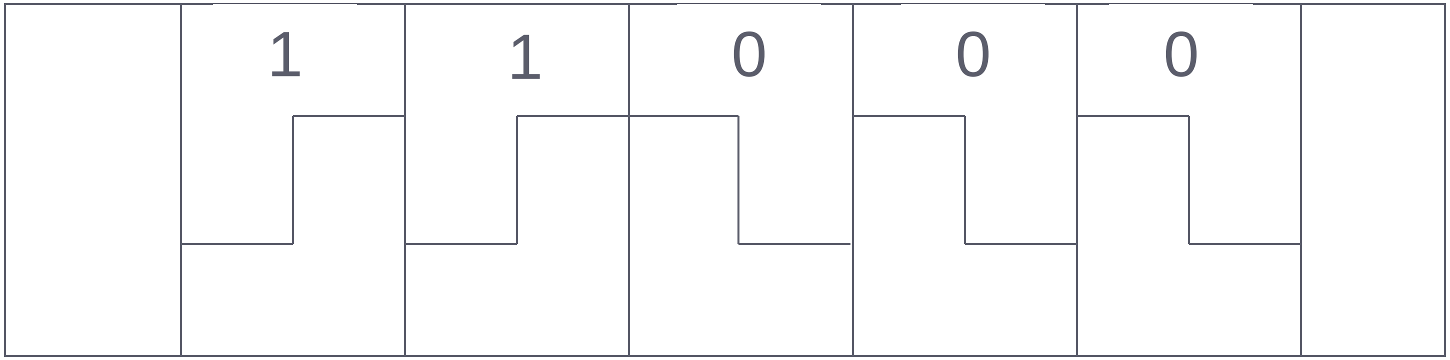
Molex QSFP stack 2x2

Coding

- Need of coding
  - Maintain DC balance for electrically isolated media (e.g. isolation transformers, AC coupling capacitors etc…)
  - Data/Control division allows for in-band control (e.g. flow control, detection of IDLE etc…)
  - Clock recovery
- Ethernet coding
  - Manchester coding (10mbps)
  - 4b/5b (100 Mbps)
  - 8b/10b (1Gbps)
  - 64b/66b (10 Gbps)

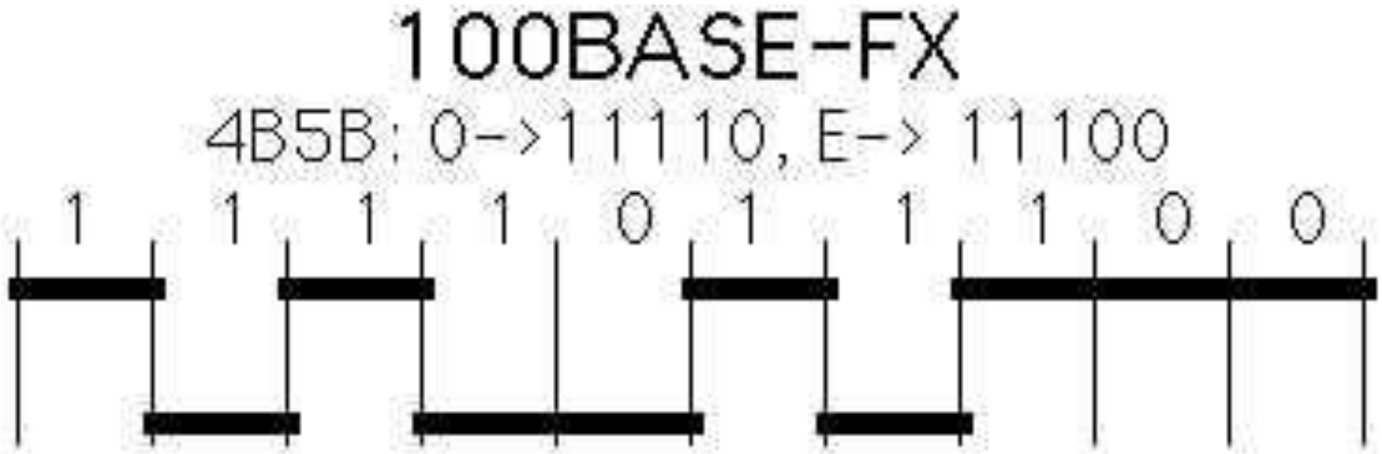
Manchester Encoding

- 10 Mbps cards used Manchester encoding and decoding is used.
- This decoding guarantee a transition every bit time.
- Due to the a transition every bit time, it requires more bandwidth.



Coding in 100BaseX

- Encoding used is 4B/5B, where a data nibble is converted to 5 bit code group.
- It uses NRZI for transmission.
- Effective data rate is 125Mbps.
- When nothing to transmit the NIC transmits IDLE character.



Transition to ‘1’ indicated by bit toggling

A 4B/5B NRZI bit pattern for input 0E (H)

Auto-Negotiation

- Need for a mechanism to accommodate multi-speed network devices.
- Auto-Negotiation detects the various modes that exist in the device on the other end of the wire, the Link Partner, and advertises it own abilities to automatically configure the highest performance mode of interoperation.
- As a standard technology, this allows simple, automatic connection of devices that support a variety of modes from a variety of manufacturers
- Auto-Negotiation acts like a rotary switch that automatically switches to the correct technology, such as 10BASE-T, 100BASE-TX, 100BASE-T4, or a corresponding Full duplex mode.

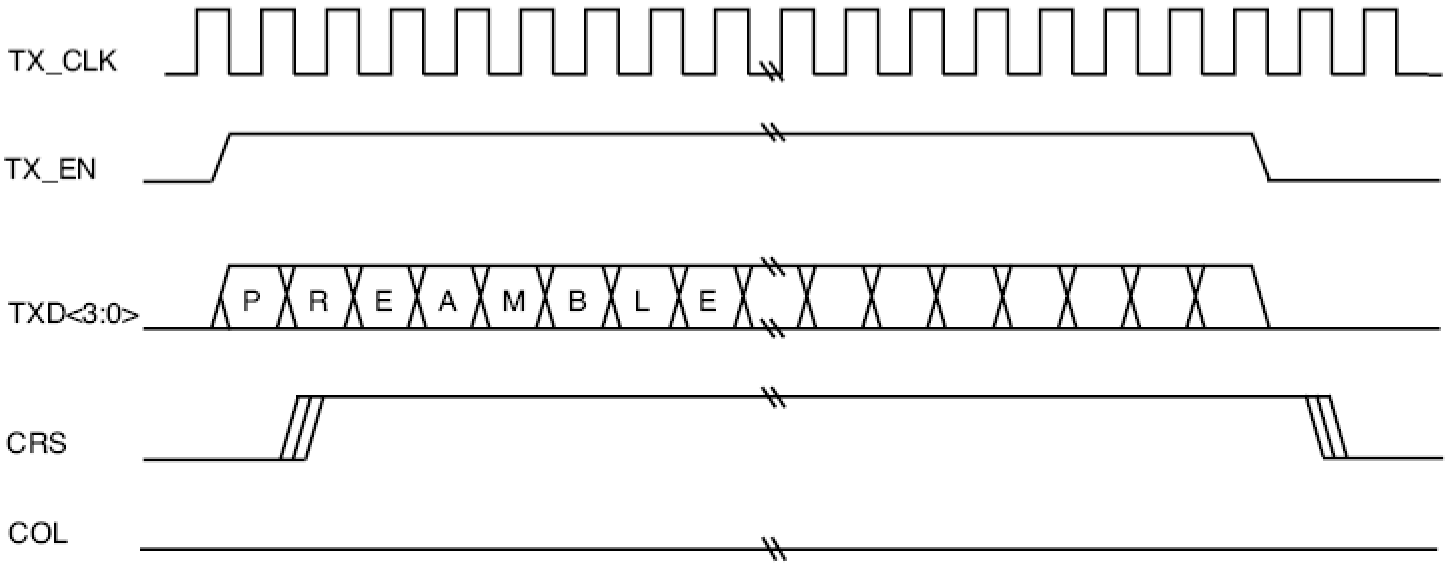
MII interface

- Media Independent Interface (MII), Gigabit-MII (GMII), Ten Gigabit-MII (XGMII) are popular interfaces to connect MAC and PHY equipment together
- MII is a simple nibble-wide interface for data along with a two-wire interface (MDIO) for control. Depending upon the data rate (10/100 mbps), it works either at 2.5 or 25 MHz
- MII is a full-duplex interface.
- MII can be used on PCB for connecting two chips together (e.g. MAC chip and PHY chip) or is available as connector also

Interface

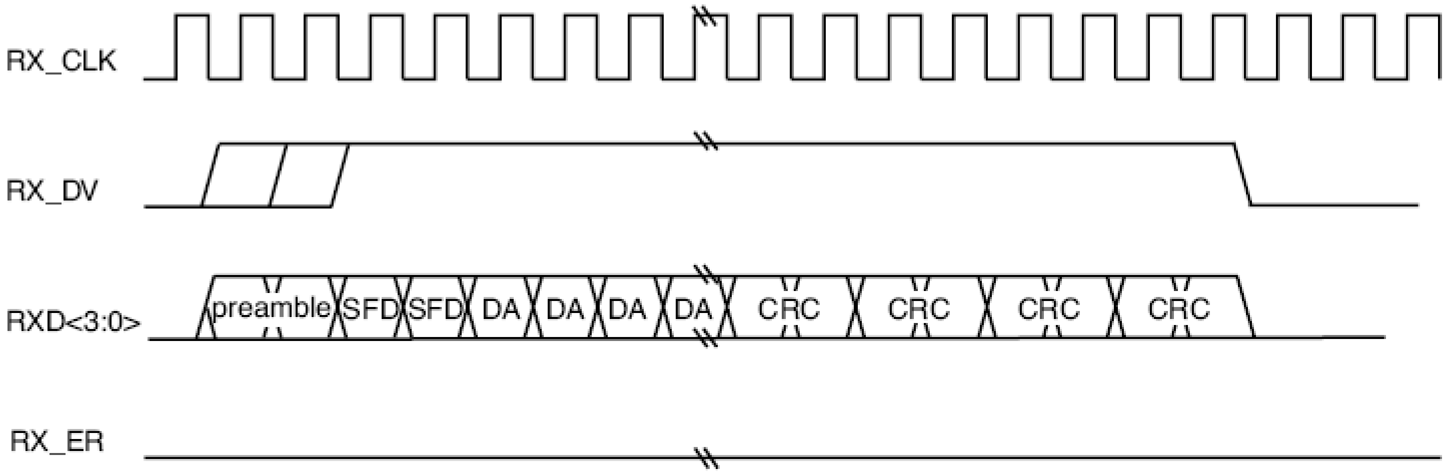
- Transmit
  - Txd<3:0>: tx data
  - Tx_en: transmit enable
  - Tx_er: error propagation
- Receive
  - Rxd<3:0>: rx data
  - Rx_dv: rx data valid
  - Rx_er: rx error
- Clocks
  - Rx_clk, tx_clk: rx and tx clocks
    - Generated by PHY, either 2.5 or 25 MHz
- Control
  - Crs: carrier sense
  - Col: collision detect
- Management
  - Mdc: clock
  - Mdio: data

MII



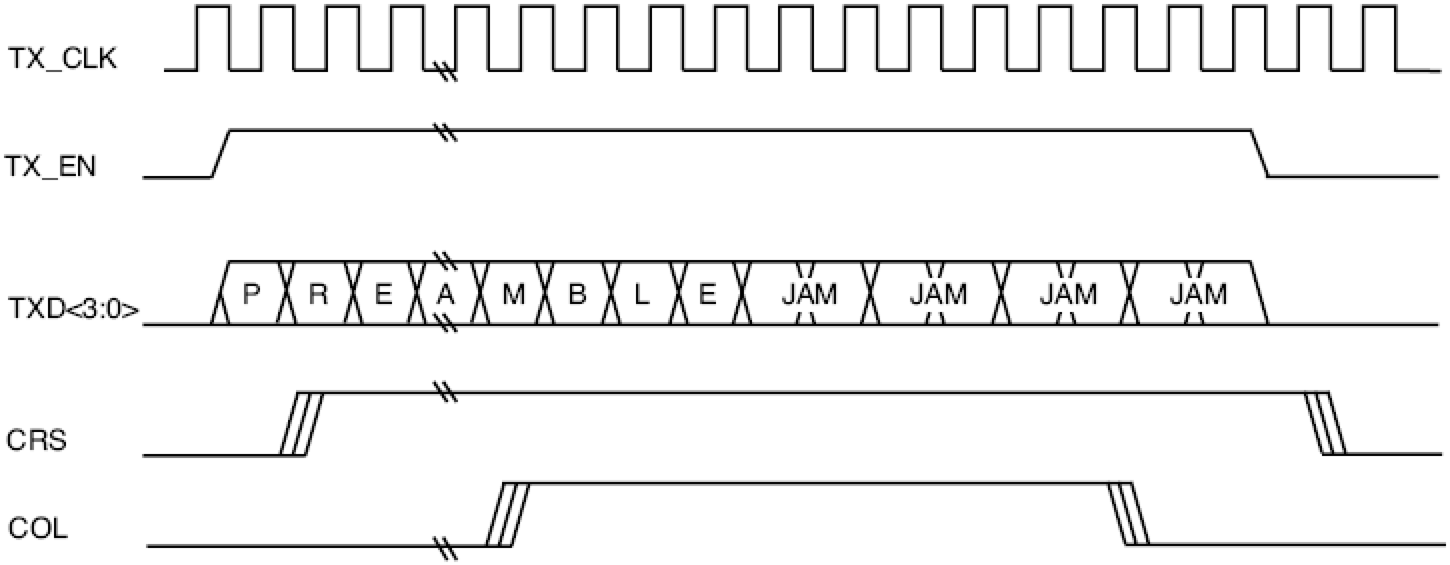
Normal frame transmission (no collisions)

MII



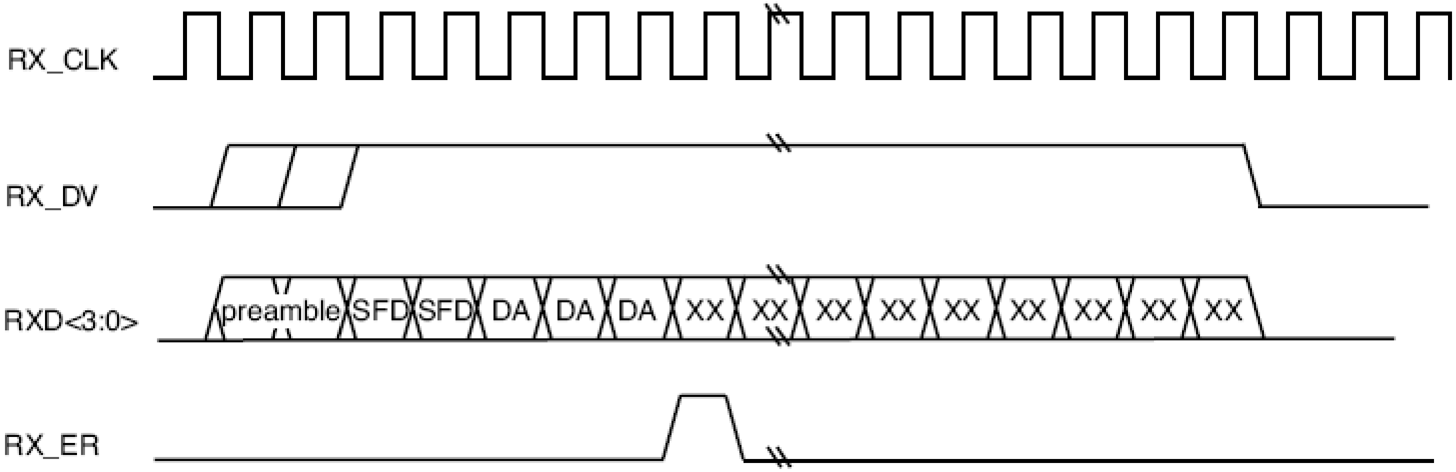
Normal frame reception with no errors

MII



Transmission with collision

MII

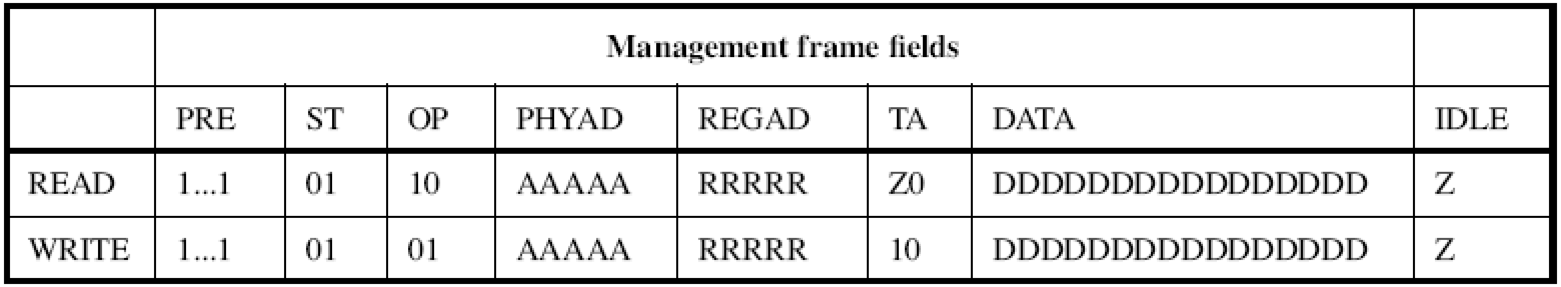


Frame reception with errors

MII

- Interface with LLC layer is left to the designer
- Simple MAC interface is capable of sending and receiving a single packet at a time. However modern MAC implementation enhance upon this basic capability by incorporating more intelligence in send/receive paths
- MDC/MDIO can be implemented by bit-toggling under software control (since performance isn’t an issue)

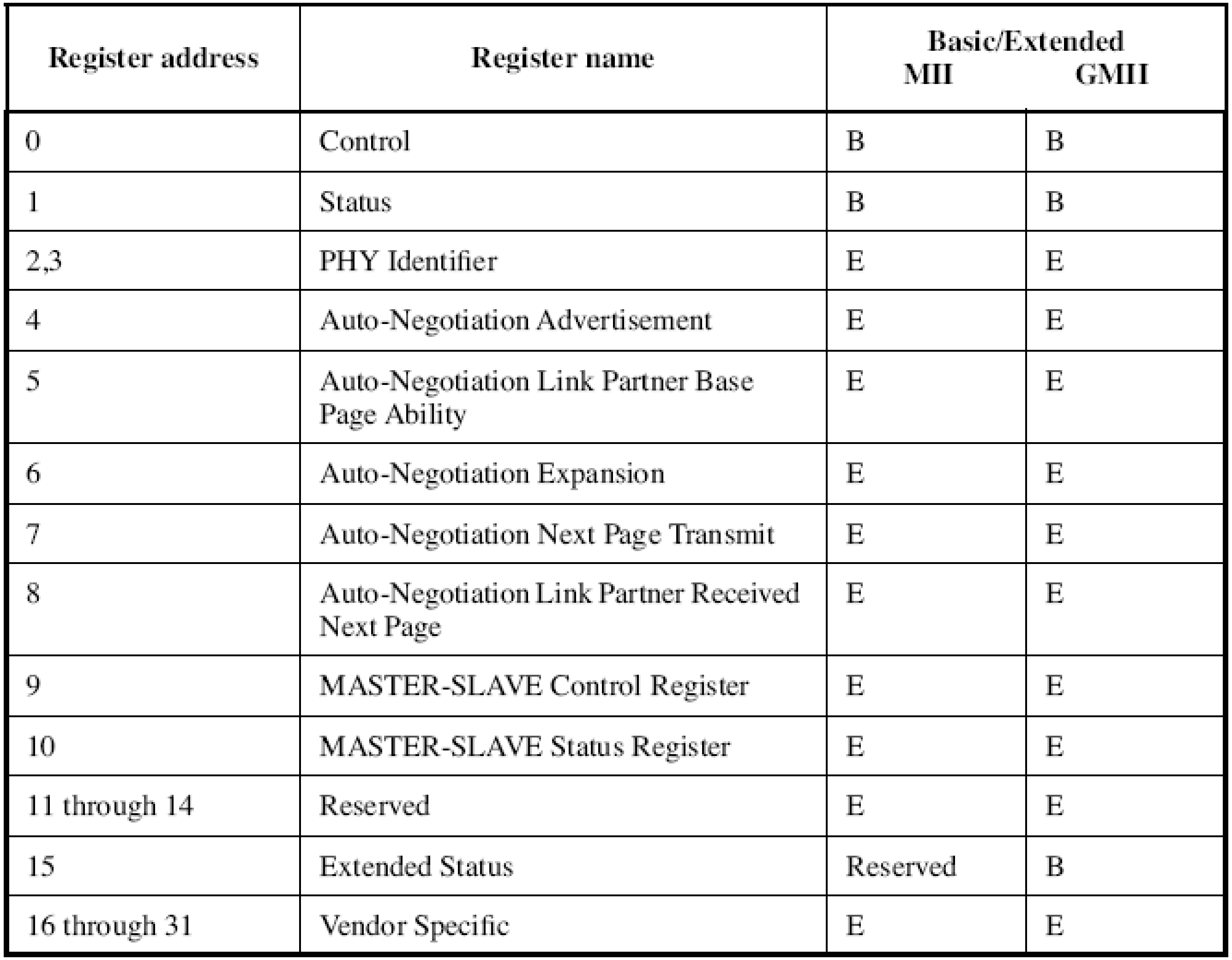
MDC/MDIO



	Management frame fields							
	PRE	ST	OP	PHYAD	REGAD	TA	DATA	IDLE
READ	1...1	01	10	AAAAA	RRRRR	Z0	DDDDDDDDDDDDDDDD	Z
WRITE	1...1	01	01	AAAAA	RRRRR	10	DDDDDDDDDDDDDDDD	Z

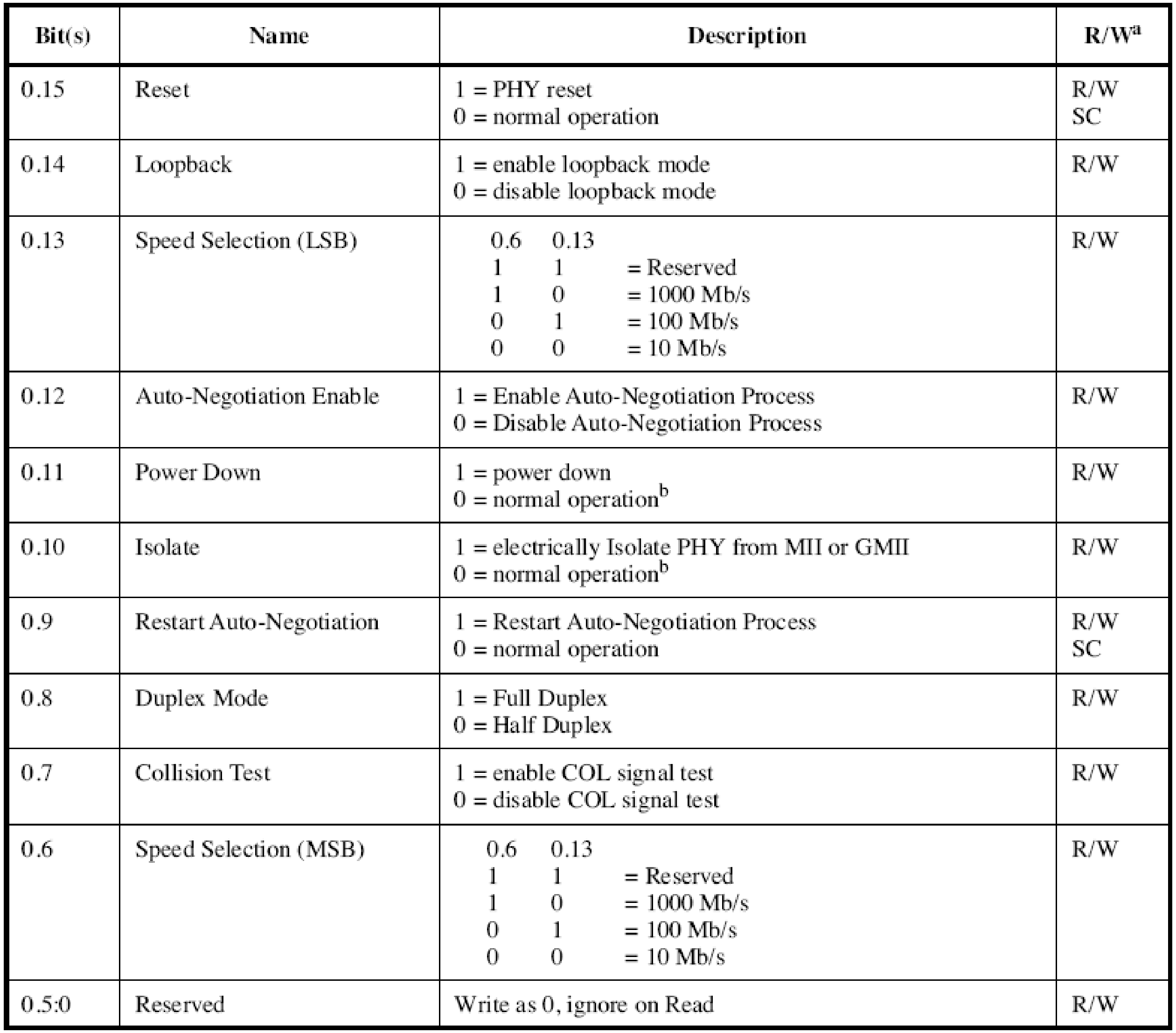
- Management frame format
  - IDLE
  - PRE (preamble)
  - ST (start of frame)
  - OP (operation code)
  - PHYAD (PHY address)
  - REGAD (register address)
  - TA (turn-around) (for rd cycles)
  - DATA

PHY registers



Register address	Register name	Basic/Extended	
		MII	GMII
0	Control	B	B
1	Status	B	B
2,3	PHY Identifier	E	E
4	Auto-Negotiation Advertisement	E	E
5	Auto-Negotiation Link Partner Base Page Ability	E	E
6	Auto-Negotiation Expansion	E	E
7	Auto-Negotiation Next Page Transmit	E	E
8	Auto-Negotiation Link Partner Received Next Page	E	E
9	MASTER-SLAVE Control Register	E	E
10	MASTER-SLAVE Status Register	E	E
11 through 14	Reserved	E	E
15	Extended Status	Reserved	B
16 through 31	Vendor Specific	E	E

Control register



Bit(s)	Name	Description	R/W[a]
0.15	Reset	1 = PHY reset 0 = normal operation	R/W SC
0.14	Loopback	1 = enable loopback mode 0 = disable loopback mode	R/W
0.13	Speed Selection (LSB)	0.6 0.13 1 1 = Reserved 1 0 = 1000 Mb/s 0 1 = 100 Mb/s 0 0 = 10 Mb/s	R/W
0.12	Auto-Negotiation Enable	1 = Enable Auto-Negotiation Process 0 = Disable Auto-Negotiation Process	R/W
0.11	Power Down	1 = power down 0 = normal operation[b]	R/W
0.10	Isolate	1 = electrically Isolate PHY from MII or GMII 0 = normal operation[b]	R/W
0.9	Restart Auto-Negotiation	1 = Restart Auto-Negotiation Process 0 = normal operation	R/W SC
0.8	Duplex Mode	1 = Full Duplex 0 = Half Duplex	R/W
0.7	Collision Test	1 = enable COL signal test 0 = disable COL signal test	R/W
0.6	Speed Selection (MSB)	0.6 0.13 1 1 = Reserved 1 0 = 1000 Mb/s 0 1 = 100 Mb/s 0 0 = 10 Mb/s	R/W
0.5:0	Reserved	Write as 0, ignore on Read	R/W

Status register

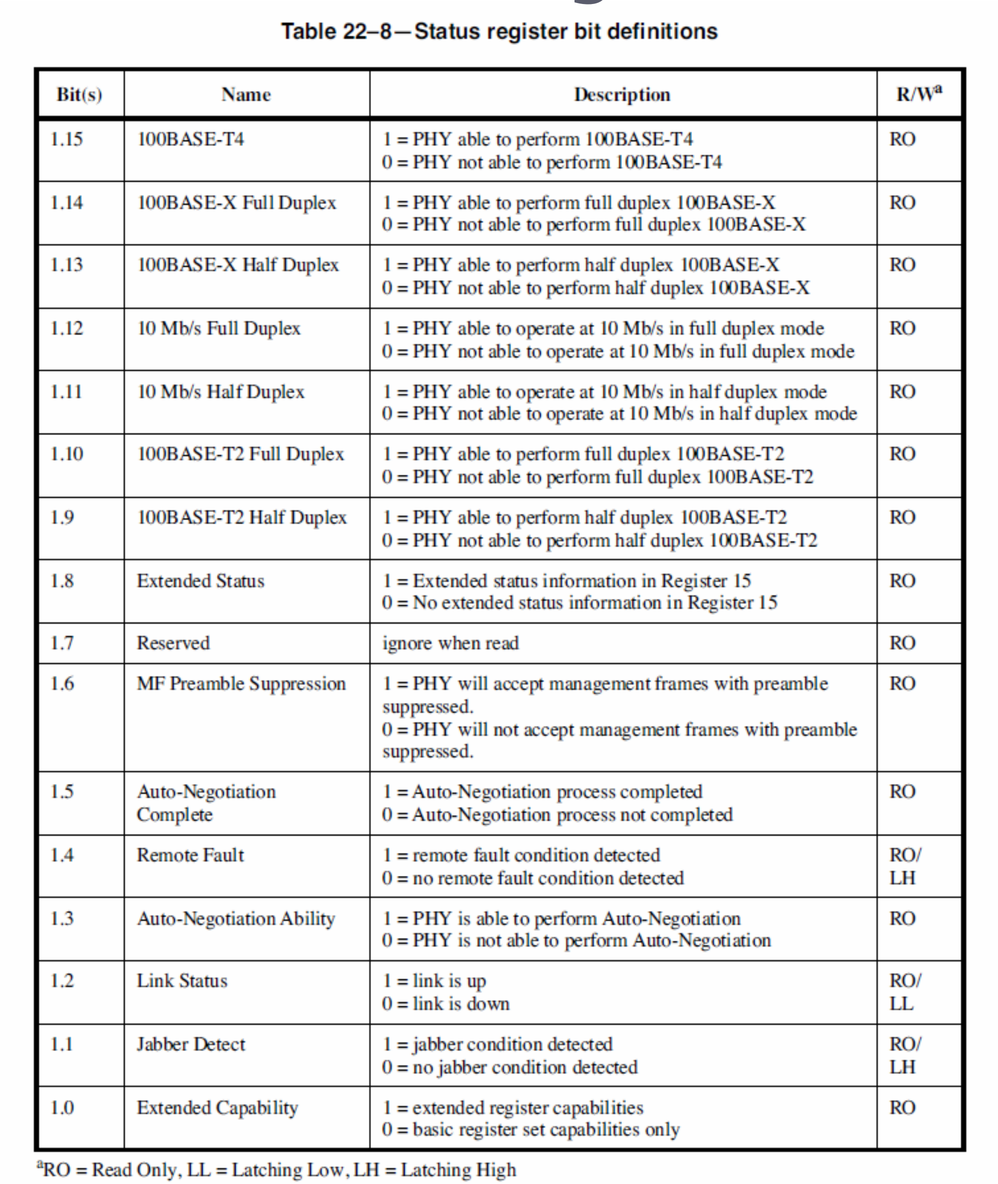


Table 22–8—Status register bit definitions

Bit(s)	Name	Description	R/W[a]
1.15	100BASE-T4	1 = PHY able to perform 100BASE-T4 0 = PHY not able to perform 100BASE-T4	RO
1.14	100BASE-X Full Duplex	1 = PHY able to perform full duplex 100BASE-X 0 = PHY not able to perform full duplex 100BASE-X	RO
1.13	100BASE-X Half Duplex	1 = PHY able to perform half duplex 100BASE-X 0 = PHY not able to perform half duplex 100BASE-X	RO
1.12	10 Mb/s Full Duplex	1 = PHY able to operate at 10 Mb/s in full duplex mode 0 = PHY not able to operate at 10 Mb/s in full duplex mode	RO
1.11	10 Mb/s Half Duplex	1 = PHY able to operate at 10 Mb/s in half duplex mode 0 = PHY not able to operate at 10 Mb/s in half duplex mode	RO
1.10	100BASE-T2 Full Duplex	1 = PHY able to perform full duplex 100BASE-T2 0 = PHY not able to perform full duplex 100BASE-T2	RO
1.9	100BASE-T2 Half Duplex	1 = PHY able to perform half duplex 100BASE-T2 0 = PHY not able to perform half duplex 100BASE-T2	RO
1.8	Extended Status	1 = Extended status information in Register 15 0 = No extended status information in Register 15	RO
1.7	Reserved	ignore when read	RO
1.6	MF Preamble Suppression	1 = PHY will accept management frames with preamble suppressed. 0 = PHY will not accept management frames with preamble suppressed.	RO
1.5	Auto-Negotiation Complete	1 = Auto-Negotiation process completed 0 = Auto-Negotiation process not completed	RO
1.4	Remote Fault	1 = remote fault condition detected 0 = no remote fault condition detected	RO/ LH
1.3	Auto-Negotiation Ability	1 = PHY is able to perform Auto-Negotiation 0 = PHY is not able to perform Auto-Negotiation	RO
1.2	Link Status	1 = link is up 0 = link is down	RO/ LL
1.1	Jabber Detect	1 = jabber condition detected 0 = no jabber condition detected	RO/ LH
1.0	Extended Capability	1 = extended register capabilities 0 = basic register set capabilities only	RO

[a] RO = Read Only, LL = Latching Low, LH = Latching High

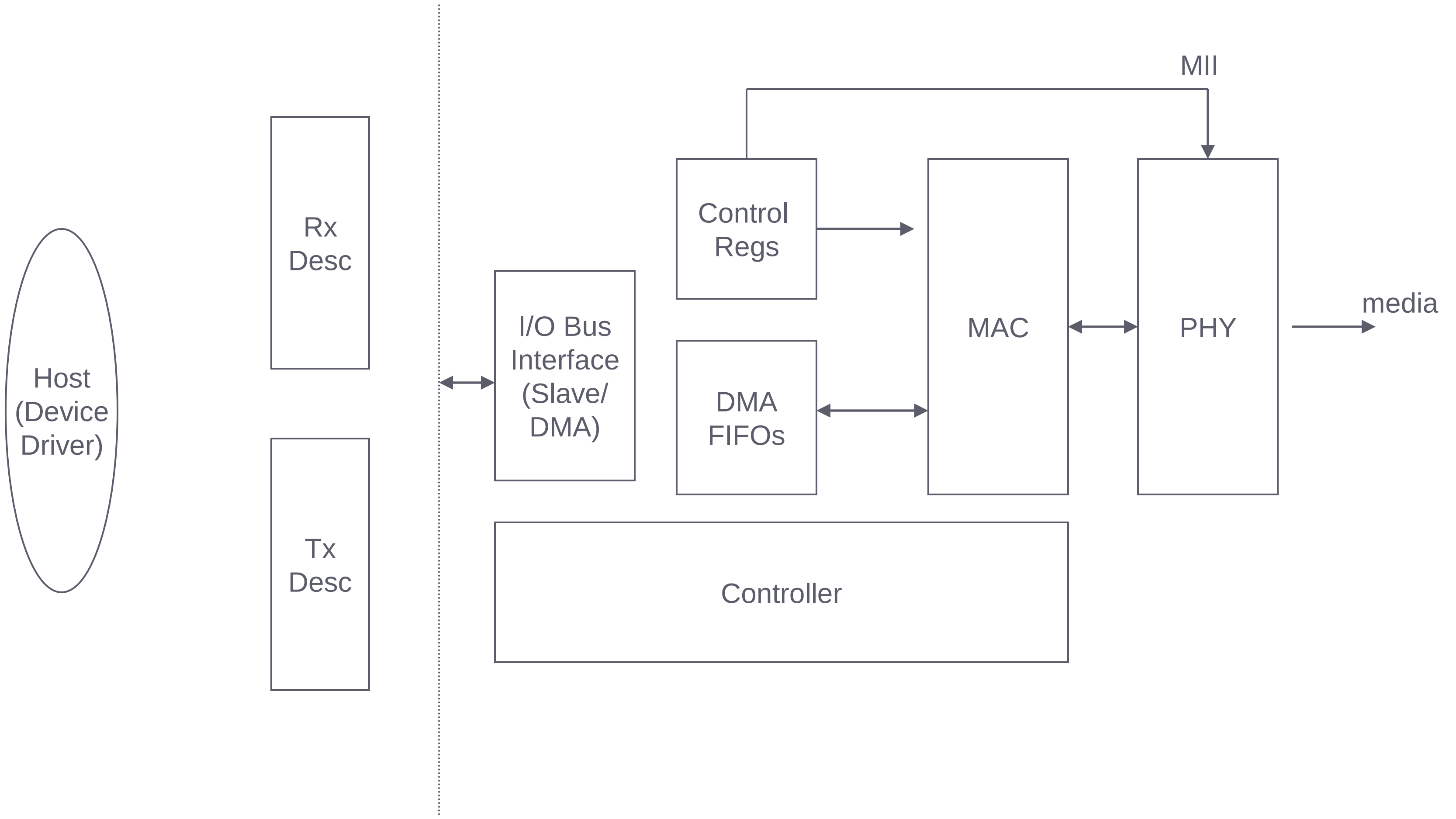
Ethernet Equipment

- Host: Network Interface Cards (NICs)
- Ethernet Hubs
- Ethernet Switches: L1, L3, ..

Network Interface Card

- Host Interface (control and data)
  - Host communicates with NIC using register interface for control and DMA for data. Typically, Ethernet data packets (single or linked list) are setup by the OS and handed over to host using some handshake mechanism
  - Similarly, rx packets are transferred to host memory using DMA. Preliminary checks such as host MAC addressing, CRC etc done by the hardware
- Hardware
  - Has control interface, DMA engine and control hardware
  - Control hardware communicates with configurable MAC
  - MAC and PHY may be built-in, else PHY chip will be external
  - MAC has proprietary control interface, PHY accessible thru MII

NIC Block Diagram



NIC operation

- Host interacts with NIC using registers (for control and status) and descriptors (for data)
- A descriptor is a structure in host memory which is shared between host and controller using a ownership scheme
- One or more descriptors are set up by the host s/w and indication given to controller. Controller services these and notifies the host (using polling or interrupt)
- A descriptor has pointer to data structures containing data such as packet payload (one or more segments) and other control info (e.g. destination MAC address)
- Control registers are for controlling operation of MAC (e.g. full duplex/half duplex), setting up station MAC address, etc
- Status regs are used to get status of phy link (down/10mbps/100mbps..), MAC stat counters etc.
- MII is used to access PHY layer. One or more devices can be accessed.

Ethernet HUB

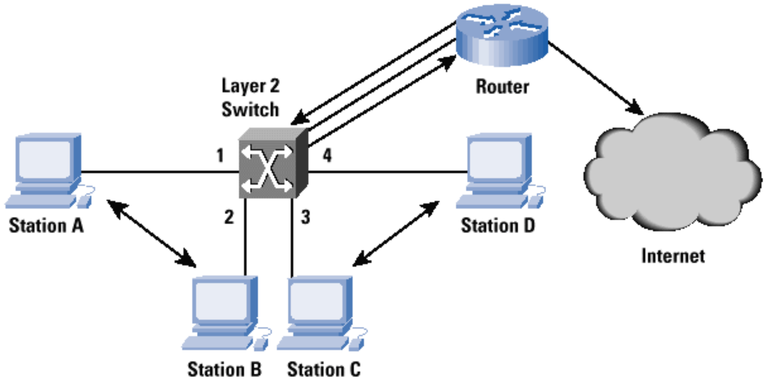
- Ethernet HUB is a form of repeater equipment. Functionally, it is identical to a network domain using shared media. It however allows structured cabling to be used in STAR topology.
- Hubs provide data communication between two endpoints at a given time. It prevents other ports to use the media during this time by suitably emulation line conditions on these ports (e.g. CRS asserted)
- Hubs are also required to repeat other events such as
  - Reporting collisions to all ports
  - Repeating runt frames
- Generally hubs do not allow for mixing of diff phy speeds. However newer implementations allow for this (and also for a faster 'uplink' for cascading purposes)
- Hub equipment is sometimes 'stackable'. This means that multiple instruments can be cascaded effortlessly to expand the network segment to a larger number of nodes beyond capacity of a single hub hardware. Stackable hubs are also easily managed using a single point of control.

Ethernet switch

- Ethernet switch is a device used to emulate a set of virtual routing channels between diff source and destinations. Multiple channels are emulated at the same time.
- Net effect is that using the same cabling infrastructure, throughput of a Ethernet network can be improved by substituting switches in place of hubs
- Most of the Ethernet switches do routing at the physical layer. However routing at higher layers (e.g IP addresses: L3, TCP sockets: L4 or TCP sessions:L5) can also be done
- Switches today are manageable. A switch also is a valid host on network and can be accessed remotely for status/control purposes.

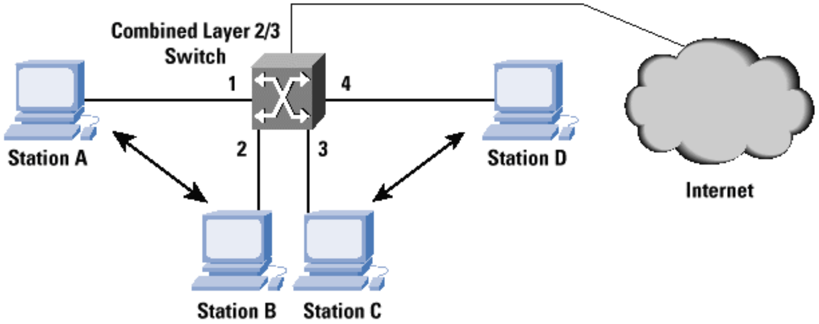
Bridges, Routers and L3 Switches

- Bridges are typically used to separate parts of a network that do not need to communicate regularly, but still need to be connected. Typically has one i/p and one o/p port
- Router is similar to L3 switches considering the switching aspect
- Router has WAN connectivity which L3 switches generally lack

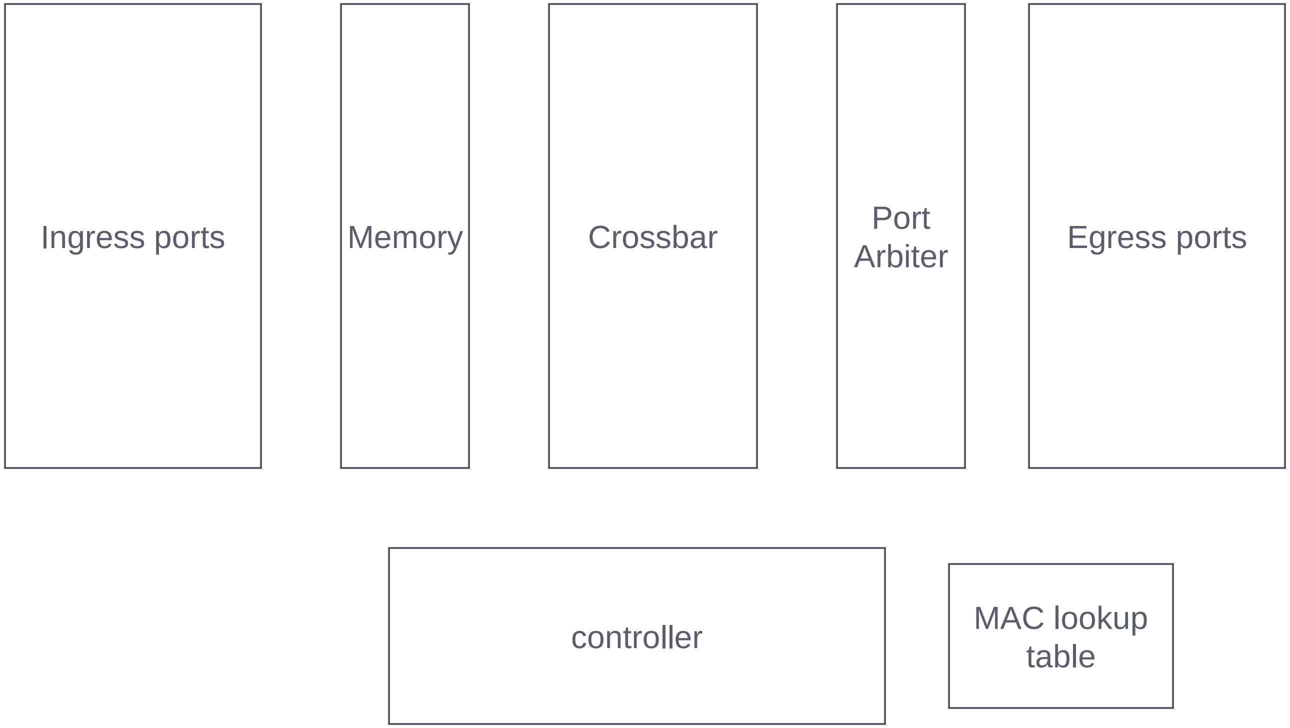


L2 switched LAN, router supports routing across multiple subnets

L2/L3 switched LAN supporting multiple subnets



Ethernet Switch Architecture

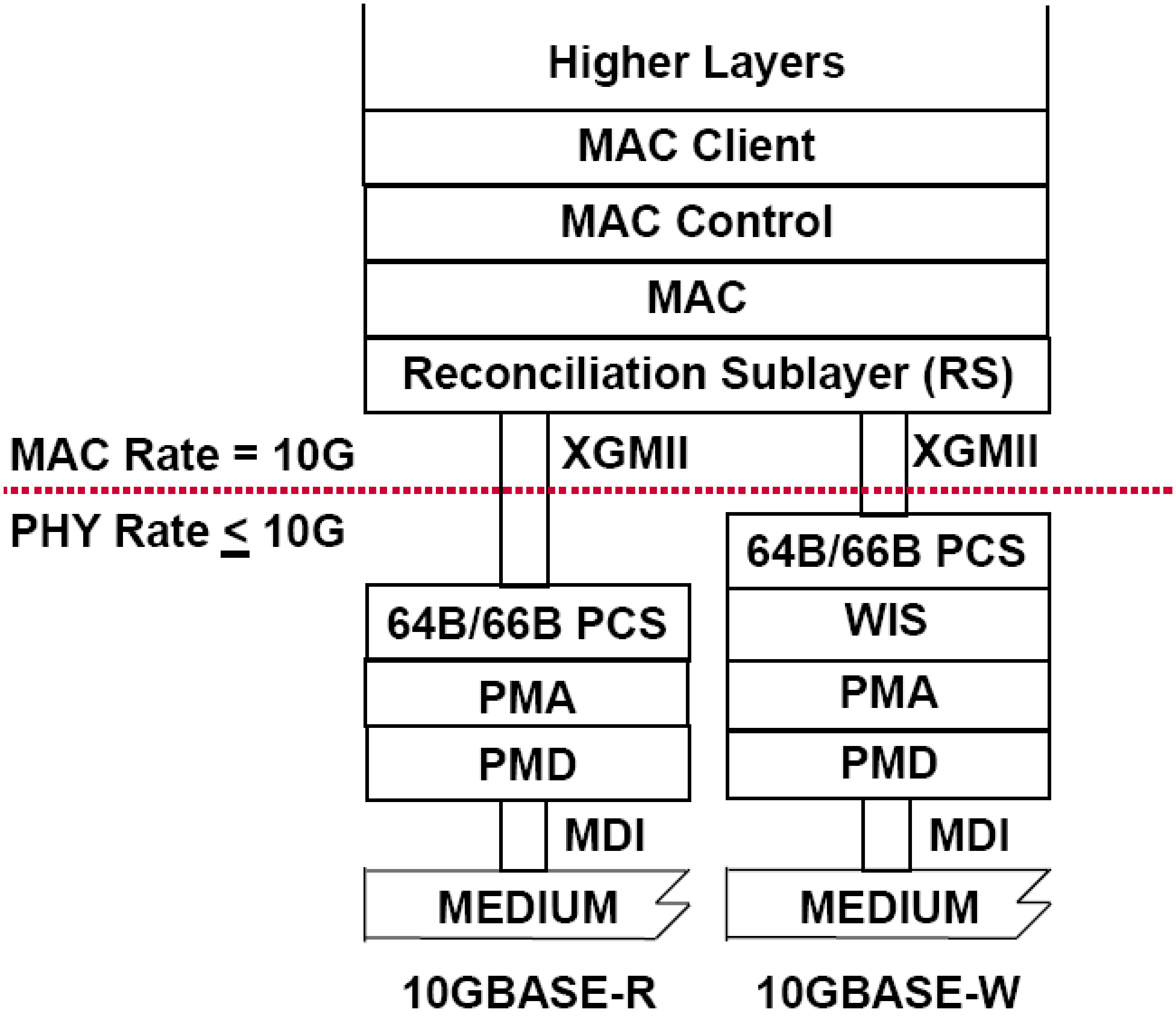


Ethernet Switch Architecture

- Basic elements: A crossbar which does M X N switching, Ingress and Egress ports, local memory and output port arbiter
- MAC address lookup table is implemented for fast resolution of MAC address → o/p port mapping
- Local memory is used for store and forward mode. Not necessary when cut-through routing is used (except when o/p port contention occurs)
- A fast network processor can be used as a controller
- One chip solutions for 10/100/1G switches are available today.

Future Architectures

- 10Gbps and future Ethernet: Only used as full duplex switched architecture
- Full hardware flow control support
- VLAN protocol is used to isolate and create virtual clusters without interfering. This also facilitates remote operation. (work from home etc…)
- Link aggregation protocol (802.3ad LAG) if implemented can support logical aggregation to create a larger data pipe or uplink. This is in particularly useful in data centers for backbone networking
- Although new high speed standard has been ratified recently, 10Gbps is still hard to reach in LANs as a default network due to cost of rebuilding

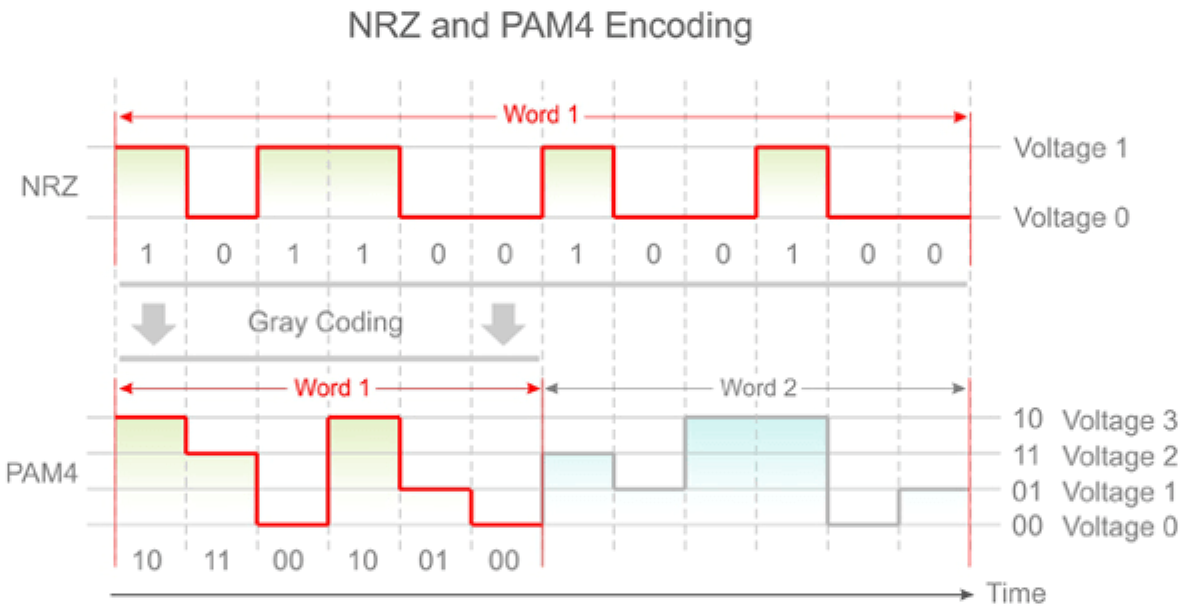

Higher Layers
MAC Client
MAC Control
MAC
Reconciliation Sublayer (RS)
MAC Rate = 10G
XGMII
XGMII
PHY Rate ≤ 10G
64B/66B PCS
64B/66B PCS
WIS
PMA
PMA
PMD
PMD
MDI
MDI
MEDIUM
MEDIUM
10GBASE-R
10GBASE-W

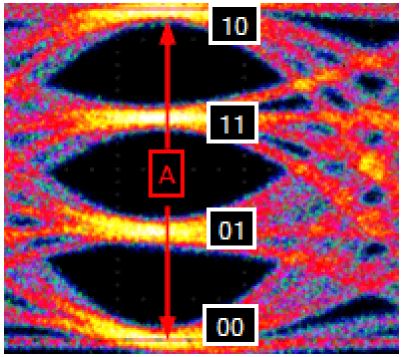
Future Architectures contd…

- 200/400Gbps Ethernet: Recently ratified as 802.3bs
- Finds its suitability in HPC, data centers, backbone networking
- Beyond 10Gbps rates, these standards support RDMA protocol at upper layer to speed up data communication to application layer
- Currently very few companies are sampling early products in this domain viz. Cisco, Juniper etc…
- This requires very sophisticated PHY layer implementation which is difficult to achieve
- 40G is essentially 10G x 4 lanes
- 100G is currently implemented as 10G x 10 lanes (2010), however later converted to 25G x 4 lanes (2014)

Future Architectures contd…

- Pulse Amplitude Modulation 4-level (PAM4) signalling introduced for 200/400 Gbps speeds in place of NRZ
- PAM4 uses 2-bits per symbol





Samtec.com

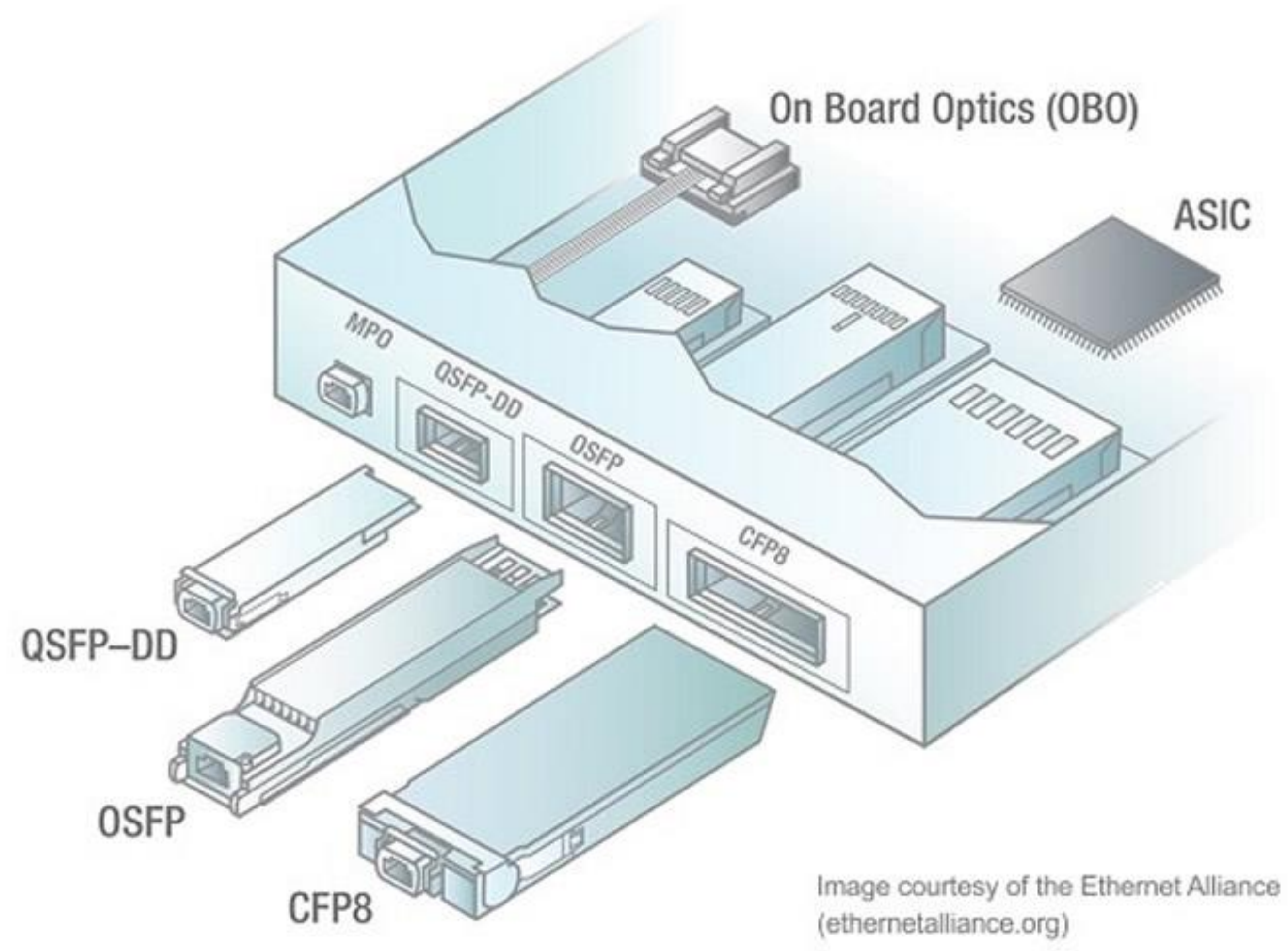
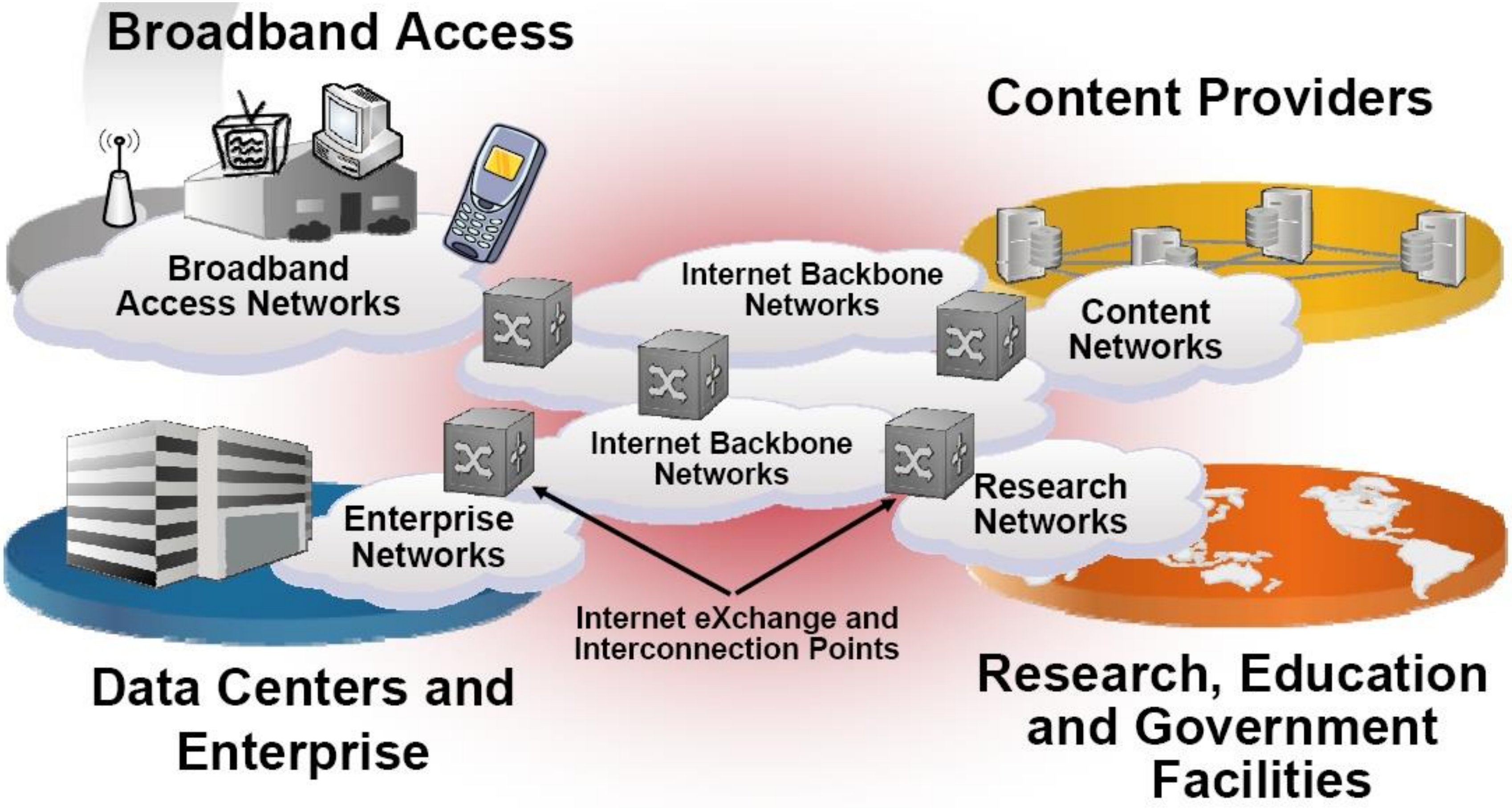


Image courtesy of the Ethernet Alliance (ethernetalliance.org)


Broadband Access
Content Providers
Broadband Access Networks
Internet Backbone Networks
Content Networks
Internet Backbone Networks
Research Networks
Enterprise Networks
Internet eXchange and Interconnection Points
Data Centers and Enterprise
Research, Education and Government Facilities

Thank You