IEEE P802.3ae 10Gb/s Ethernet Task Force

Blue Book

12th July 2000

Proposal for an Open Loop PHY Rate Control Mechanism

Shimon Muller
Ariel Hendel

Sun Microsystems Inc.
Computer Systems

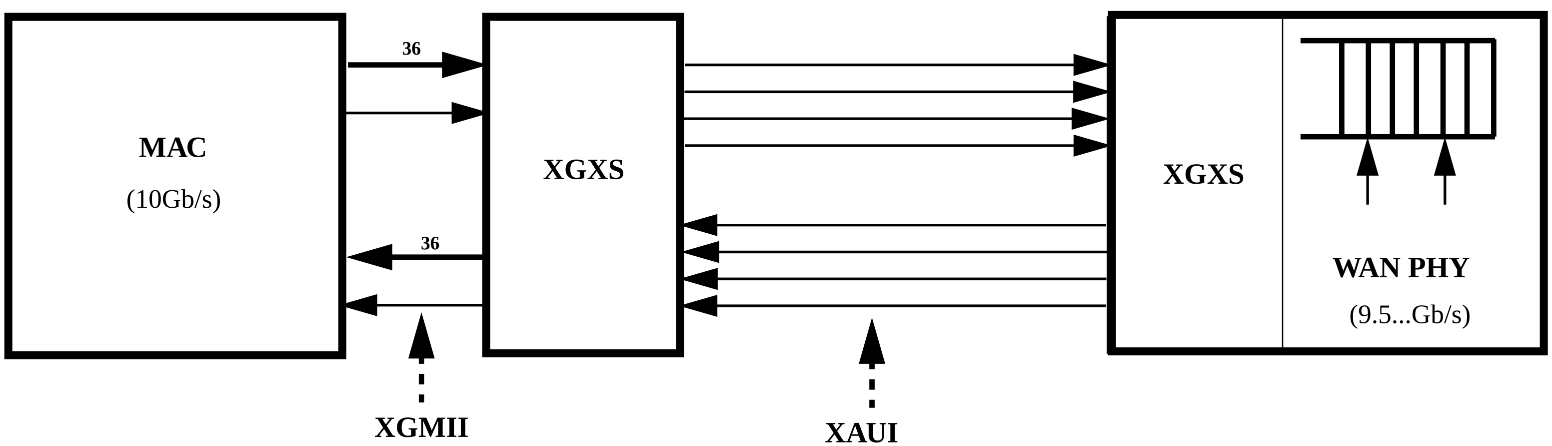
July 11, 2000

Outline

- Introduction
  - Why is a Rate Control Mechanism Necessary for 802.3ae
- MAC<->PHY Rate Control Alternatives
- MAC Self-Pacing Proposal
  - Concept
  - Implementation Implications
  - Standard Implications
- Summary

Introduction --- Why Rate Control for 802.3ae?

- **At the November 1999 meeting, the HSSG adopted the following objectives for 802.3ae:**
  - **Support a speed of 10.0000 Gb/s at the MAC/PLS service interface**
  - **Define two families of PHYs:**
    - **A LAN PHY, operating at a data rate of 10.0000 Gb/s**
    - **A WAN PHY, operating at a data rate compatible with the payload rate of OC-192c/SDH VC-4-64c**
  - **Define a mechanism to adapt the MAC/PLS data rate to the data rate of the WAN PHY**



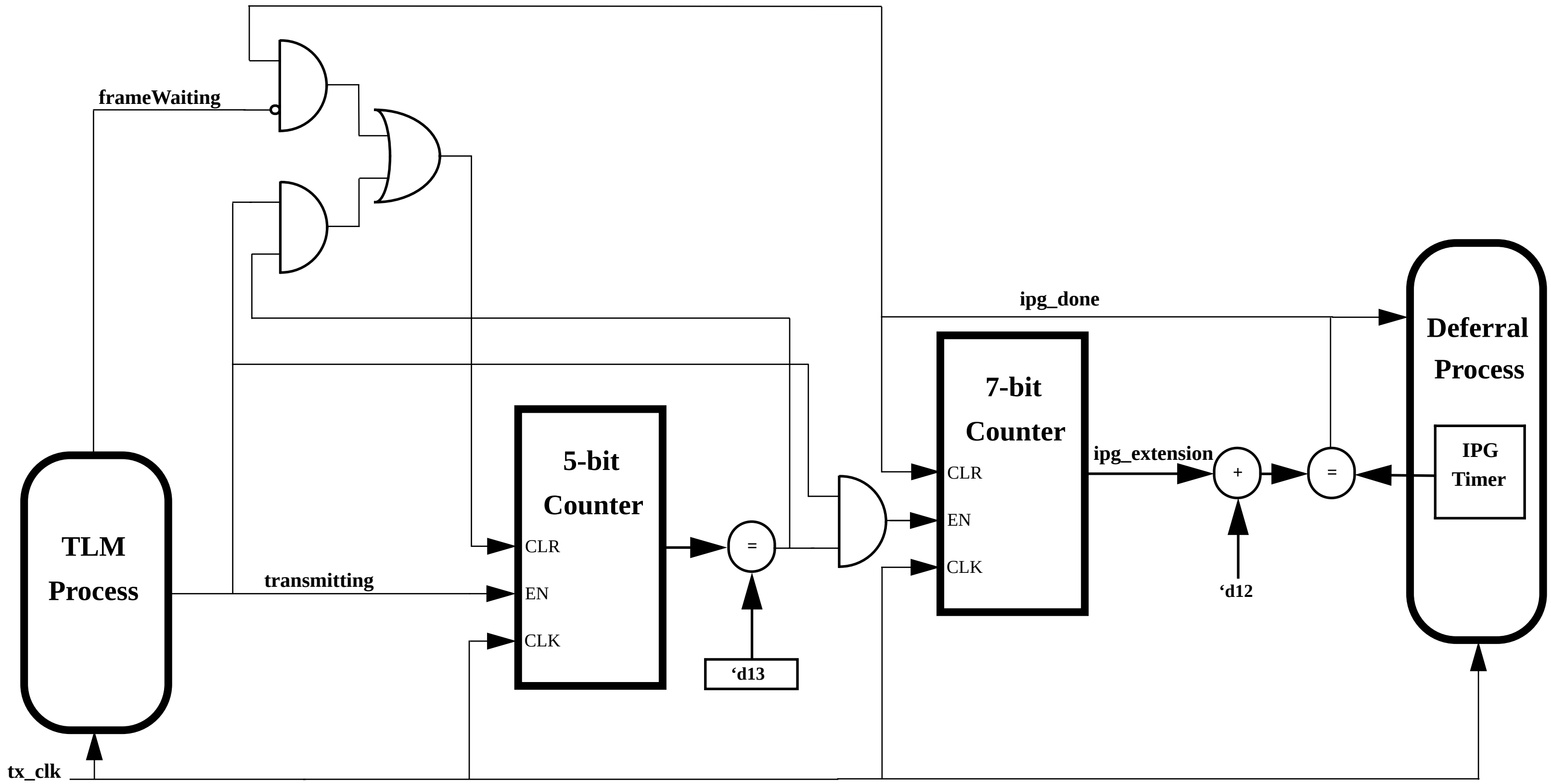
MAC<->PHY Rate Control Alternatives

- Fine granularity rate control
  - Word-by-Word hold signalling
  - Clock stretching
- Packet granularity rate control
  - Frame-based
  - Carrier Sense based
  - Busy Idle
  - Self pacing in the MAC

MAC Self-Pacing Proposal

- **Concept**
  - **The MAC “knows” that the PHY is slower and by how much**
  - **The MAC adapts its average data rate by extending the IPG after each frame transmission**
    - **This guarantees that the MAC never exceeds the average data rate in the PHY, with packet granularity**
  - **The IPG extension is “dynamic”**
    - **Depends on the size of the previously transmitted frames**
  - **The PHY is only required to sustain the transmission of one maximum size packet**
    - **Requires a rate adaptation fifo in the PHY of ~64 bytes (plus framer overhead)**

MAC Self-Pacing Proposal --- Implementation



Notes:

*** transmitting --- signal that frames the transmission of a frame in the MAC**

*** ipg_done --- signal that indicates the completion of IPG transmission**

*** frameWaiting --- signal that indicates that a frame was passed to the MAC for transmission**

MAC Self-Pacing Proposal --- Pascal Changes

- **Transmit state variables (4.2.7.2)**
  - ***const***
    ifsExtensionRatio = ...; {In bits, determines the number of bits in a frame that will require one octet of interFrameSpacing extension, see 4.4}
  - ***var***
    paceMode: Boolean; {Indicates the desired mode of operation, ... static variable}
    ifsExtensionCount: 0... {In bits, running counter that counts the number of bits during frame transmission that will be considered for minimum interFrameSpacing extension}
    ifsExtensionSize: 0... {In octets, running counter that counts the integer number of octets to be added to minimum interFrameSpacing}

- **State variable initialization (4.2.7.5)**
  - *procedure* Initialize;

```
begin
   .................
   paceMode := ...;
   ifsExtensionCount := 0;
   ifsExtensionSize := 0;
   while carrierSense or receiveDataValid do nothing
end;
```

MAC Self-Pacing Proposal --- Pascal Changes (cont.)

- **Frame transmission (4.2.8)**
  - *function* TransmitLinkMgmt: TransmitStatus;

```
begin
        ..........................................................
        ..........................................................
  begin {loop}
    if bursting then frameWaiting := true
    else
      begin
        if attempts > 0 then BackOff;
        ~~if halfDuplex then~~ frameWaiting := true;
        .........................................................
        .........................................................
      end;
    lateCollision := false;
    StartTransmit;
    frameWaiting := false;    (underlined)
    if halfDuplex then
      begin
        ~~frameWaiting := false;~~
      .........................................................
      .........................................................
      end {half duplex mode}
    else while transmitting do nothing
  end; {loop}
      .........................................................
      .........................................................
end; {TransmitLinkMgmt}
```

MAC Self-Pacing Proposal --- Pascal Changes (cont.)

```
process BitTransmitter;
  begin
    cycle {outer loop}
      if transmitting then
        begin {inner loop}
          extendError := false;
          PhysicalSignalEncap;
          while transmitting do
            begin
              if (currentTransmitBit > lastTransmitBit) then TransmitBit(extensionBit)
              else if extendError then TransmitBit(extensionErrorBit)
              else
                begin
                  TransmitBit(outgoingFrame[currentTransmitBit]);
                  ifsExtensionCount := ifsExtensionCount + 1;
                  if ((ifsExtensionCount mod 8) = 0) then
                    if ((ifsExtensionCount mod ifsExtensionRatio) = 0) then
                                                ifsExtensionSize := ifsExtensionSize + 1
                end;
              if newCollision then StartJam else NextBit
            end;
          if bursting then
            begin
              InterFrameSignal;
              if extendError then
                if transmitting then transmitting := false
                else IncLargeCounter(lateCollision);
                bursting := bursting and (frameWaiting or transmitting)
            end
        end {inner loop}
    end {outer loop}
  end; {BitTransmitter}
```

MAC Self-Pacing Proposal --- Pascal Changes (cont.)

```
process Deference;
  begin
    if halfDuplex then cycle {half duplex loop}
      ..........................................................
      ..........................................................
    end {half duplex loop}
    else cycle {full duplex loop}
      while not transmitting do nothing;
      deferring := true;
      while transmitting do nothing;
      StartRealTimeDelay;
      while RealTimeDelay(interFrameSpacing) do nothing;
      while paceMode and (ifsExtensionSize > 0) do
        begin
          Wait (8);
          ifsExtensionSize := ifsExtensionSize - 1
        end;
      if frameWaiting then ifsExtensionCount := ifsExtensionCount mod ifsExtensionRatio
      else ifsExtensionCount := 0;
      deferring := false
    end {full duplex loop}
  end; {Deference}
```

Summary

- The open loop rate control achieves rate adaptation by extending the IPG between frames, controlled by the MAC
- This method for rate adaptation, as proposed, has the following advantages:
  - Simple
  - Cheap
  - Very precise
    - Worst case imprecision is less than 0.05751%
  - Independent of the PHY and the MAC/PHY interconnect
  - The self contained nature of this mechanism provides a robust solution

IEEE P802.3ae 10Gb/s Ethernet MDC/MDIO Proposal

David Law, Edward Turner - 3Com
Howard Frazier - Cisco Systems
Rich Taborek, Don Alderrou- nSerial

Contribution from :

Alan Ames and Bob Noseworthy - UNH InterOperability Lab

Initial Issue

- Need register access to external XGXS interfaces as well as PHY internal registers

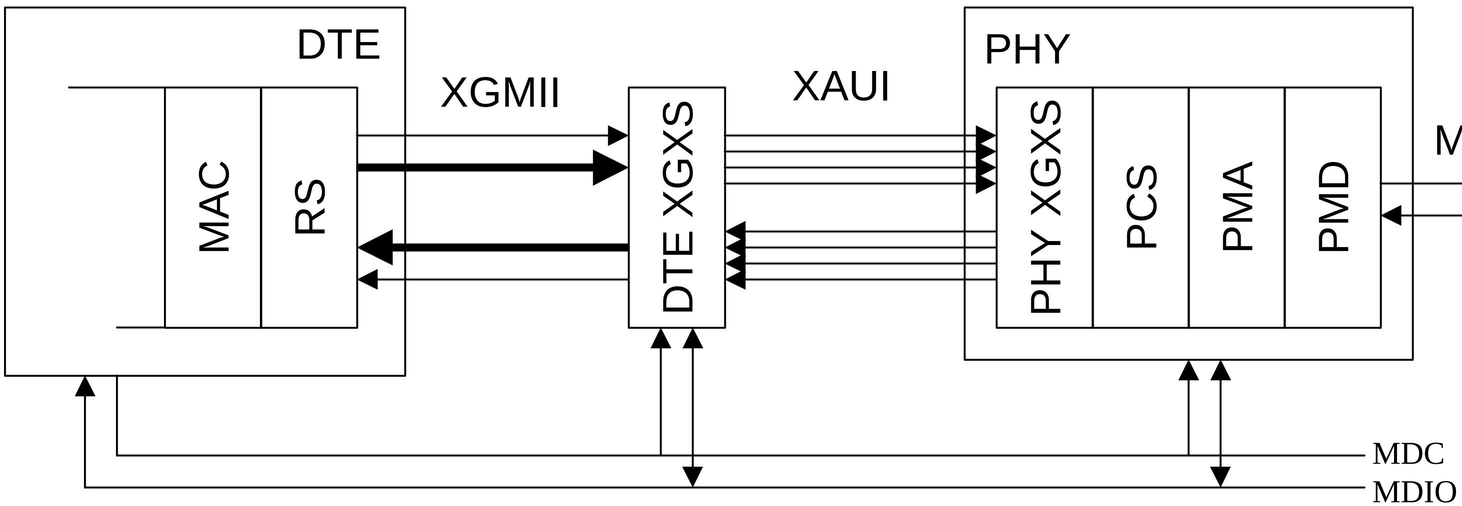


Diagram based on ‘XAUI/XGXS Proposal’, Rich Taborek et al, March 2000
URL: http://www.ieee802.org/3/ae/public/mar00/taborek_1_0300.pdf (Page 7)
also Brad Booth e-mail April 4th 2000 ‘XGMII a/k/r and XGXS - PCS Interface’
URL: http://www.ieee802.org/3/10G_study/email/msg02165.html

Issues

- Need to support expanded number of registers for ‘PMD’ use
- Other proposals may need register access
  - WIS
  - LSS
- Desire to provide larger register area for Vendor specified registers

Issues (Cont)

- Need to leave some space for the future
  - 100Gb/s ?
- Desire to support operation on same bus as existing PHY devices
- Bit and Register consumption means few Registers free in current address map

Use of existing registers

Register	15 … 0	Access
0	Control	R/W
1	Status	RO
2	PHY identifier	RO
3	PHY identifier	RO
4	AN advertisement	R/W
5	AN link partner ability	RO
6	AN expansion	RO
7	AN next page	R/W
8	AN link partner rcvd next page	RO
9	Master / Slave control	R/W
10	Master / Slave status	RO
11	Reserved	
12	Reserved	
13	Reserved	
14	Reserved	
15	Extended Status	RO

Unavailable   Available

Register	15 … 0
16	Vendor Specific
17	Vendor Specific
18	Vendor Specific
19	Vendor Specific
20	Vendor Specific
21	Vendor Specific
22	Vendor Specific
23	Vendor Specific
24	Vendor Specific
25	Vendor Specific
26	Vendor Specific
27	Vendor Specific
28	Vendor Specific
29	Vendor Specific
30	Vendor Specific
31	Vendor Specific

Unavailable

Current PMD Register Access Proposal

- New ST code proposal by Howard Frazier

  URL: http://www.ieee802.org/3/10G_study/public/sept99/frazier_3_0999.pdf

  – Proposed use of the ST sequence (00) for transactions with PMD
    - Used a new ST sequence to open up a fresh set of 32 registers and allowed PHY and PMD to be defined independently
  – Could be extended to provide another 64 registers by using all combinations of ST and OP codes
    - Appears not to be enough

New Proposal

- Use spare ST code (00) as proposed before
  - No more ST codes available
- Define new Indirect Address register access
  - Applicable to ST code 00 only
  - Access consists of a Address cycle followed by a Read or Write cycle
- Provides many more registers
  - 32 Ports as at present
  - 32 'Devices' per port
  - 65536 Registers per device

UNH Interoperability Study

– Investigation by UNH InterOperability Lab
  - Work undertaken by Alan Ames and Bob Noseworthy
– Tested existing PHY immunity to ST=00 frames
  - Tested single cycle reads and writes
  - Tested 2 concatenated frame accesses
– All 24 devices tested ignored frames with ST=00

Indirect Addressing Proposal

ST	OP	PHYADR	DTYPE	TA	ADDRESS/DATA

ST = 00

OP	Access Type
00	Address
01	Write
10	Read
11	Post Read Inc Address

Value	Device
00000	Reserved
00001	PMD
00010	XGXS PHY
00011	XGXS DTE
00100	WIS
-------	--------------

Access Type	Contents
Address	Address
Write	Write Data
Read	Read Data
Post Inc	Read Data

Indirect Addressing Example

ST	OP	PHYADR	DTYPE	TA	ADDRESS
00	00	0 0000	0 0001		0000 0100 1001 0001

OP = Address

Device Type = PMD

Address to be accessed

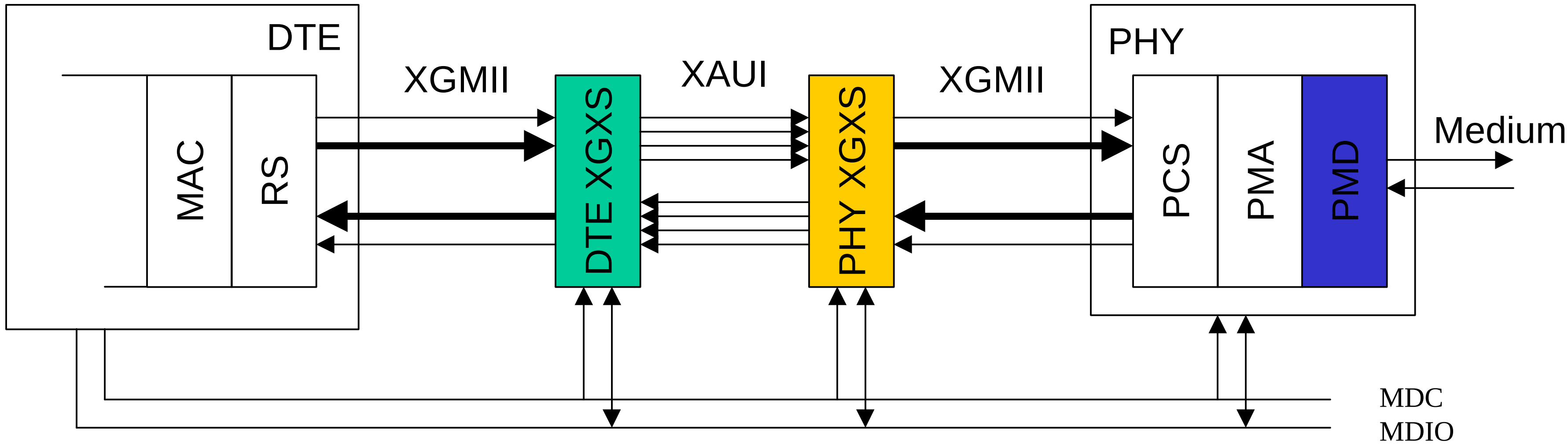
ST	OP	PHYADR	DTYPE	TA	DATA
00	10	0 0000	0 0001		0000 0100 1001 0001

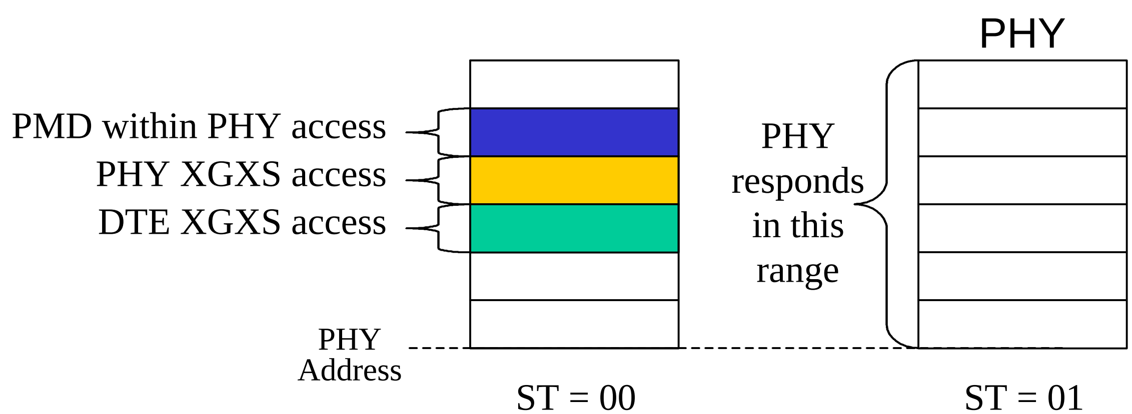
OP = Read

Device Type = PMD

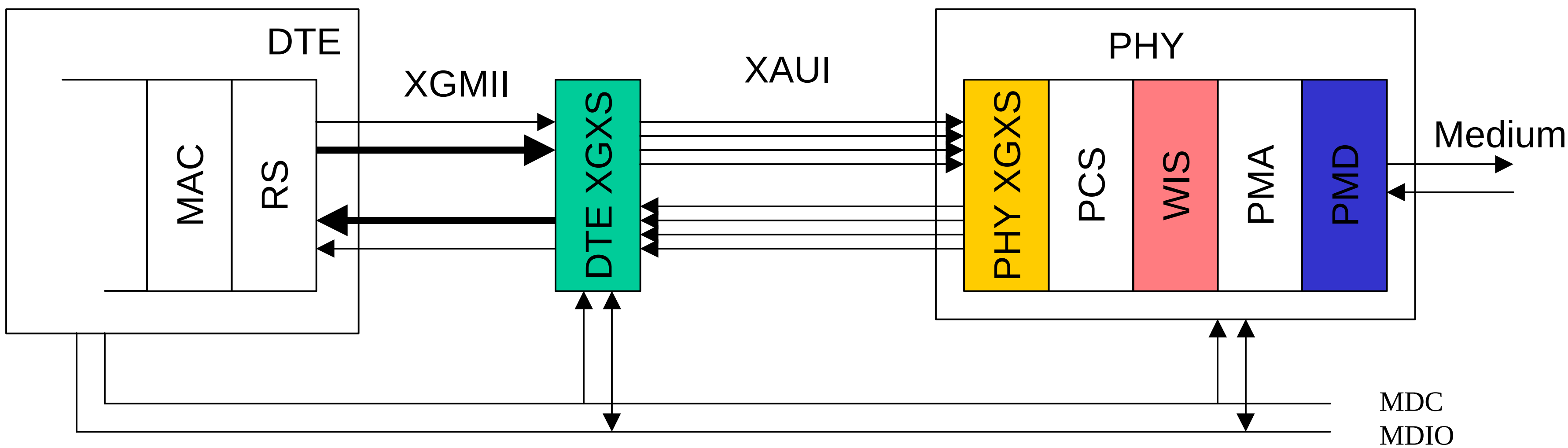
Data returned

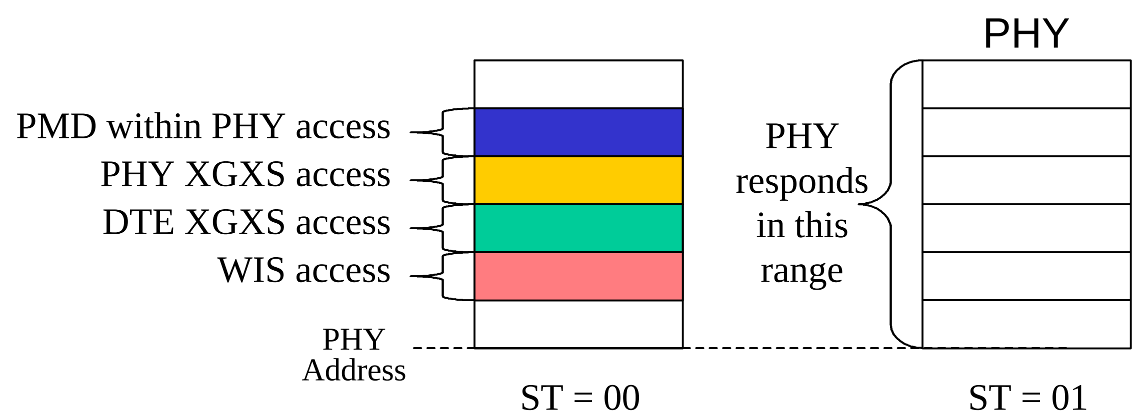
LAN PHY Example





WAN PHY Example





Summary

– Define new Indirect Address register access
  - Access consists of a Address cycle followed by a Read or Write cycle
  - 'PHY' registers already defined access as today
– Opens up many more registers
  - 32 Ports as at present
  - 32 'Devices' per port
  - 65536 Registers per device

IEEE P802.3ae 10Gb/s Ethernet Management MIB Baseline Proposal

David Law, Edward Turner - 3Com
Howard Frazier - Cisco Systems

Management Proposal

- Clause 30 - Protocol independent management definition
  - Add to, or modify, existing attributes, objects, capabilities and packages as required
- Annex 30A & 30B - GDMO MIB
  - Changes to match Clause 30 changes
- Annex 30C - SNMP MIB (Link Agg only)
  - No additions

Why no SNMP MIB ?

- Annex 30C only contains the SNMP MIB for the Link Aggregation managed object classes
  - To provide a full set of SNMP MIBs would require
    - Including current SNMP MIBs in Annex 30C
    - Investigating and fixing discrepancies between IEEE management and IETF Ethernet RFCs
  - This seems a lot of work and out of scope of the 802.3ae PAR
- SNMP MIBs produced by IETF as usual

Management Clause Changes

- Clause 30 Management
  - Changes to title and overview to add 10Gb/s
  - Changes to managed objects are summarised in following slides
    - Minor changes to behaviours to add 10Gb/s to the list of exclusion are not listed (invert the set)
- Annex 30A & 30B - GDMO MIB
  - Changes follow the Clause 30 changes
    - Counter sizes already fixed

Clause 30 Updates

- Layer management for DTEs (30.3)
  - oMAC Entity managed object class (30.3.1)
    - aMACCapabilities
      - Additional value used to indicate support for Rate Control
    - aRate
      - Attribute to control and report MAC Rate
      - May only be required for ‘Open-Loop’ Rate Control
  - oPHY Entity managed object class (30.3.2)
    - aPhyType & aPhyTypeList - additional values for new PHY types

Clause 30 Updates (Cont)

- oMACControlEntity managed object class (30.3.3)
  - No changes
- oPauseEntity managed object class (30.3.4)
  - No changes

- Layer management for repeaters (30.4)
  - oRepeater (30.4.1), oGroup (30.4.2) and oPort (30.4.3) managed object classes
    - Repeater not supported so no changes

Clause 30 Updates (Cont)

- Layer management for MAUs (30.5)
  - oMAU managed object class (30.5.1)
    - aMAUType - additional values for new PHY types
    - aMediaAvailable & aFalseCarriers - behaviour updates
    - This is where any new PHY specific additions would go. This will require more analysis once the 10Gb/s PHY selection becomes clearer
  - May be need for new object class such as oWIS for WIS related features

Clause 30 Updates (Cont)

- Management for link Auto-Negotiation (30.6)
  - oAutoNegotiation managed object (30.6.1)
    - Auto Negotiation not supported so no changes
- Link Aggregation Management (30.7)
  - oAggregationPort (30.7.2), oAggPortStats (30.7.3) and oAggPortDebugInformation (30.7.4) managed object classes
    - No changes required

IEEE P802.3ae 10Gb/s Ethernet Management Recommendation

- Changes to Clause 30 as listed
  - Changes to Annex 30A and 30B GDMO MIB to match as required
- No addition to Annex 30C SNMP MIB
  - SNMP MIB to be produced by IETF as normal

IEEE P802.3ae
10 Gigabit Ethernet Task Force

XGMII Update

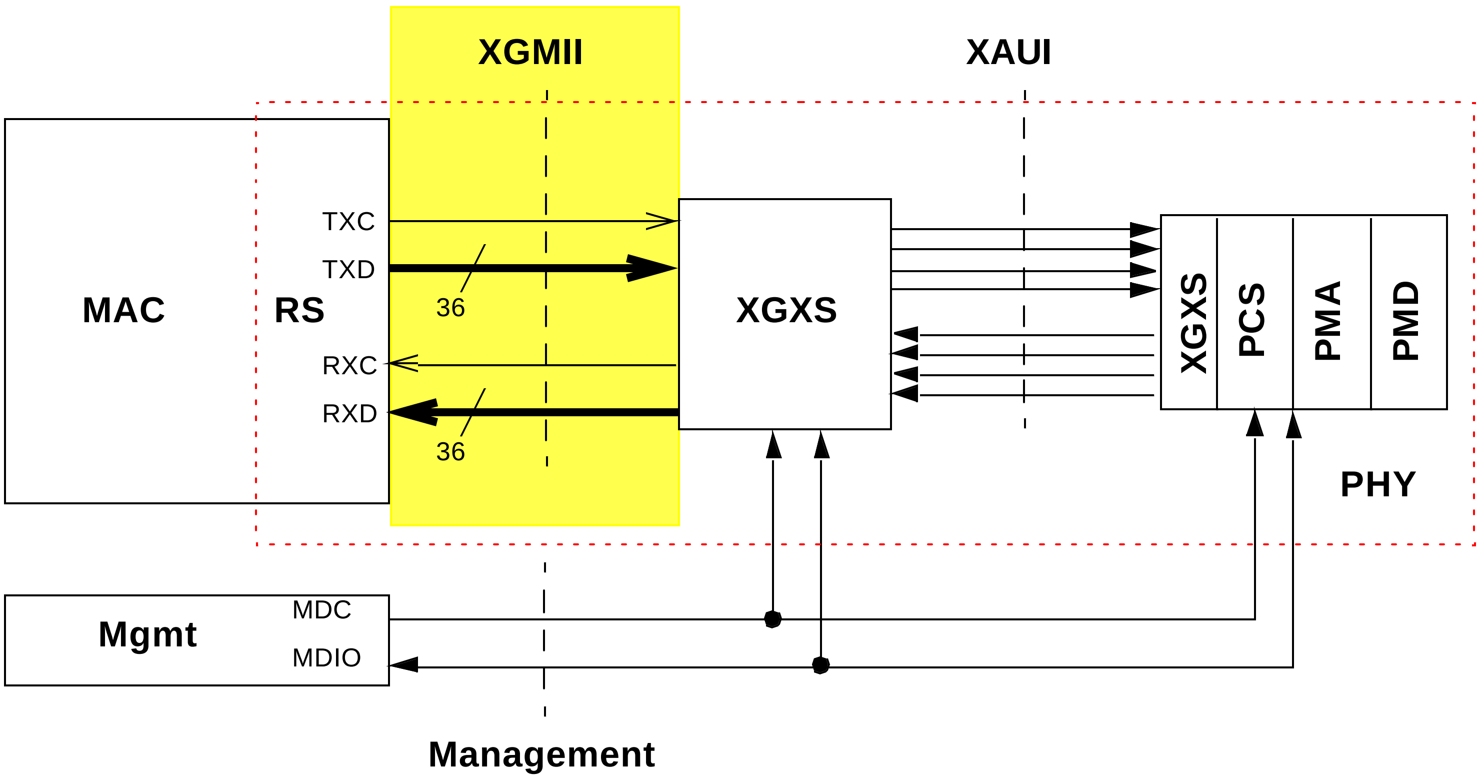
La Jolla, CA
11-July-2000

Howard Frazier - Cisco Systems

Goals and Assumptions

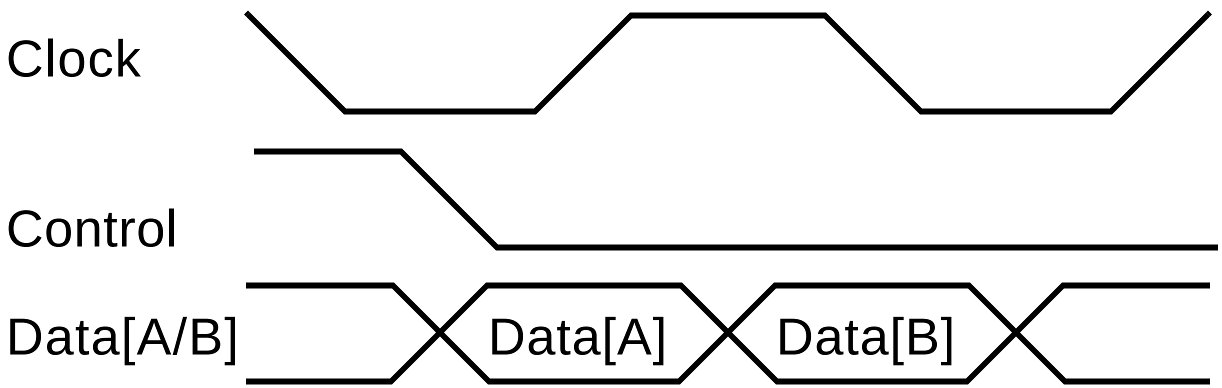
- Allow multiple PHY variations
- Provide a convenient partition for implementers
- Provide a standard interface between MAC and PHY
- Reference industry standard electrical specifications

Interface Locations



10 Gigabit Media Independent Interface

- 32 data bits, 4 control bits, one clock, for transmit
- 32 data bits, 4 control bits, one clock, for receive
- Dual Data Rate (DDR) signaling, with data and control driven and sampled on both rising edge and falling edge of clock



- 32 bit data paths are divided into four 8 bit “lanes”, with one control bit for each lane

10 Gigabit Media Independent Interface - Coding

- Use embedded delimiters rather than discrete signals
- Control bit (C) is “1” for delimiter and special characters
- Control bit (C) is “0” for normal data characters
- Delimiter and special character set includes:
  - Idle, Start, Terminate, Error
- Delimiters and special characters are distinguished by the value of the 8 bit data lane when the corresponding control bit is “1”
- Data (d) symbols are striped on lane 1, lane 2, lane 3, lane 0, etc.
  - Frames (packets) may be any number of symbols in length subject to minFrameSize and maxFrameSize

10 Gigabit Media Independent Interface - Coding

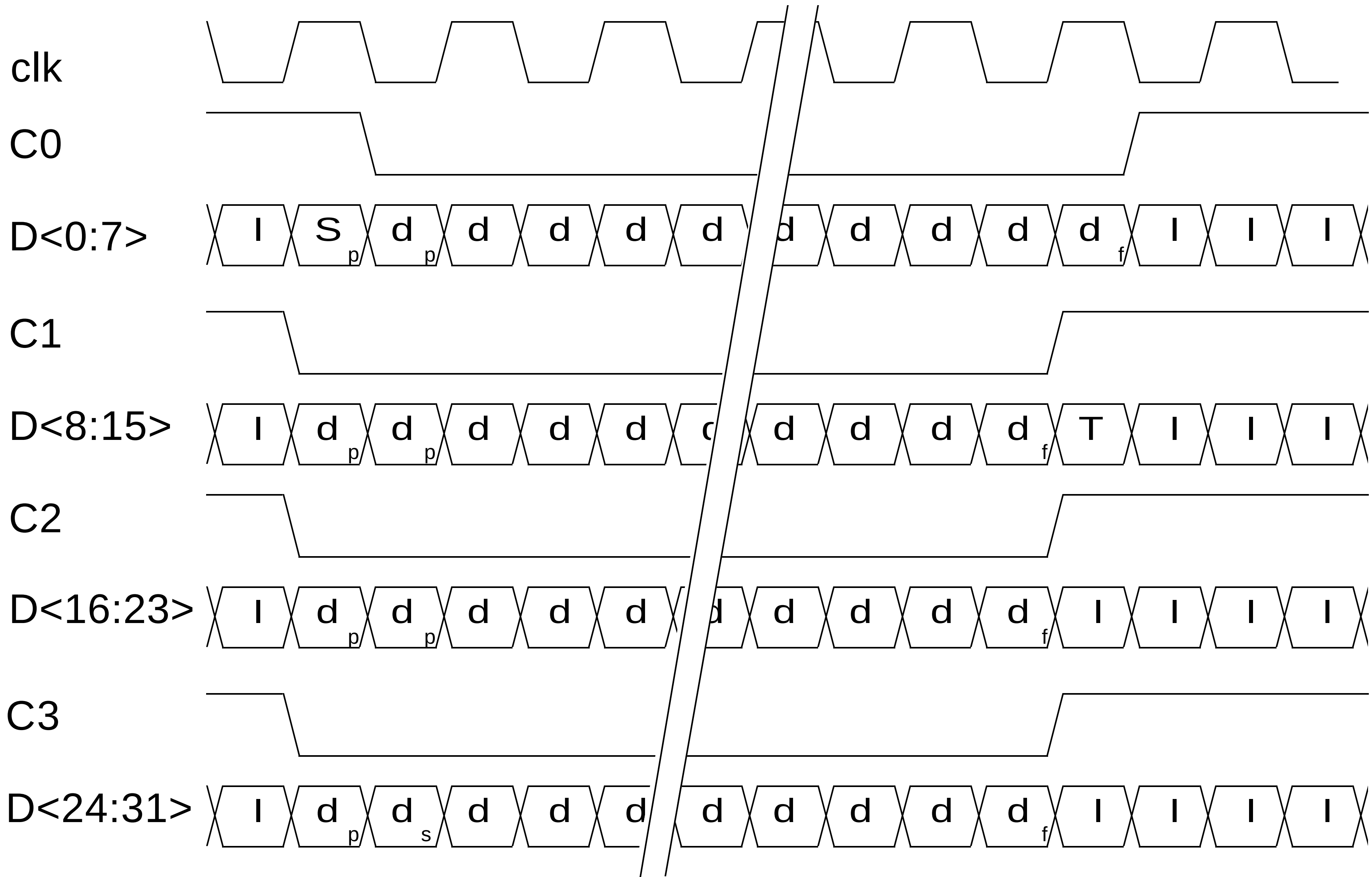
- Idle (I) is signaled
  - during the Inter-Packet Gap
  - when there is no data to send
- Start (S) is signaled
  - for one byte duration at the beginning of each packet
  - always on lane 0
- Terminate (T) is signaled
  - for one byte duration at the end of each packet
  - may appear on any lane
- Error (E) is signaled
  - when an error is detected in the received signal
  - when an error needs to be forced into the transmitted signal

10 Gigabit Media Independent Interface - Coding

NEW

Shorthand	Name	Code Point (Control)	Code Point (Data)
I	Idle	1	0x07
S	Start	1	0xFB
T	Terminate	1	0xFD
E	Error	1	0xFE
d	Data	0	0x00 - 0xFF

10 Gigabit Media Independent Interface - Example

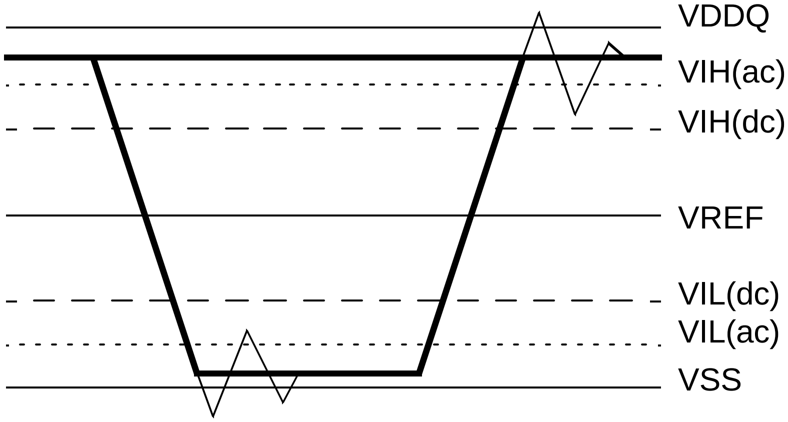


IEEE 802.3ae

10 Gigabit Ethernet

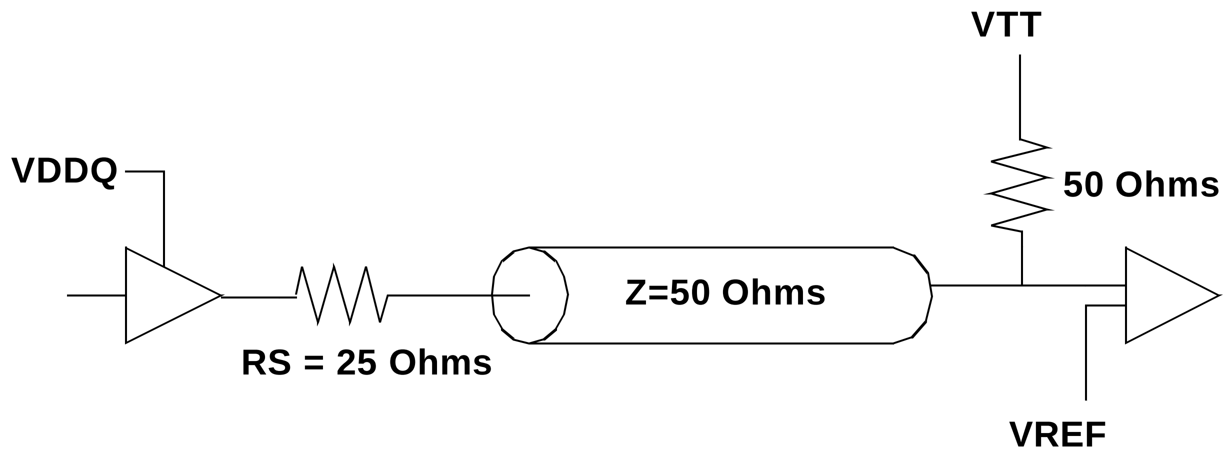
10 Gigabit Media Independent Interface - Electrical Characteristics

- Use Stub Series Terminated Logic for 2.5 Volts
  - SSTL_2
  - EIA/JEDEC Standard EIA/JESD8-9
  - Class I (8 ma) output buffers

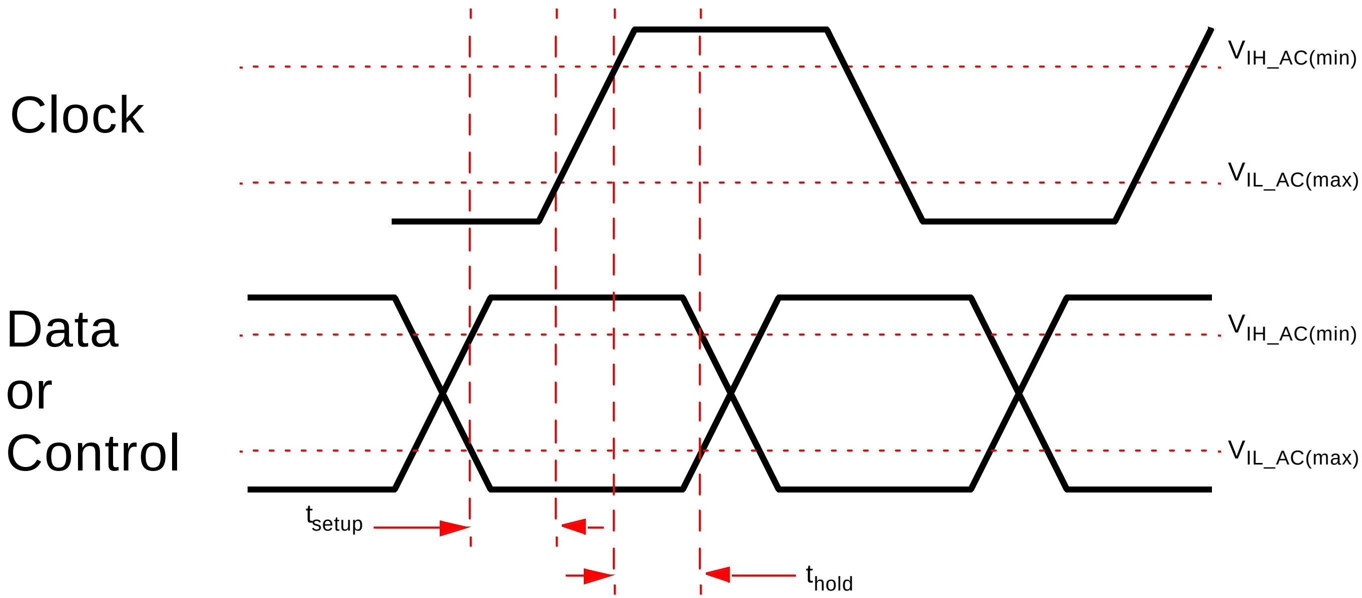


Symbol	Parameter	Min	Typ	Max
VDDQ	Supply Voltage	2.3	2.5	2.7
VREF	Reference Voltage	1.15	1.25	1.35
VTT	Termination Voltage	VREF-0.04	VREF	VREF+0.04
VIH(dc)	dc input logic high	VREF+0.18		VDDQ+0.3
VIL(dc)	dc input logic low	-0.3		VREF-0.18
VIH((ac)	ac input logic high	VREF+0.35		
VIL(ac)	ac input logic low			VREF-0.35

10 Gigabit Media Independent Interface - Circuit Topology Example



10 Gigabit Media Independent Interface - Timing



Symbol	Driver	Receiver	Units
t_{setup}	960	480	ps
t_{hold}	960	480	ps

Summary

- The XGMII coding proposal is stable
- The EIA/JEDEC SSTL_2 standard can be referenced for the XGMII electrical specification
- The timing proposal presented herein is a starting point for further discussion

XAUI/XGXS Proposal

By:

Don Alderrou, nSerial; Howard Baumer, Broadcom; Vipul Bhatt, Finisar; Brad Booth, Intel; Kirk Bovil, Blazel; Ed Chang, NetWorth Technologies; Ed Cornejo, Lucent; Robert Dahlgren, SV Photonics; Kevin Daines, World Wide Packets; John Dallesasse, Molex; Joel Dedrick, AANetcom; Thomas Dineen, Dineen Consulting; Schelto van Doorn, Infineon; Steve Dreyer, nSerial; Richard Dugan, Agilent; John Ewen, IBM; Howard Frazier, Cisco; Mark Feuerstraeter, Intel; Eric Grann, Blaze; Steve Haddock, Extreme Networks; Chuck Haymes, IBM; Ken Herrity, Blaze; Jay Hoge, JDS Uniphase; Osamu Ishida, NTT; Pat Kelly, Intel; Van Lewing, QED; David Lynch, Gennum; Jeff Lynch, IBM; Henning Lysdal, Giga; Kreg Martin, Brocade Communications; Ron Miller, Brocade Communications; Shimon Muller, Sun; Bob Musk, JDS Uniphase; Brian Peters, Blaze; Mark Ritter, IBM; Shawn Rogers, Texas Instruments; Koichiro Seto, Hitachi Cable; Dave Simmons, Gennum; Jeff Stai, Qlogic; Daniel Svensson, SwitchCore; Steve Swanson, Corning; Rich Taborek, nSerial; Bharat Tailor, Gennum; Jim Tavacoli, Accelerant Networks; Hemant Thakkar, Kinar Inc.; Tom Truman, Bell Labs/Lucent; Rick Walker, Agilent; Fred Weniger, Vitesse; Tony Whitlow, Molex; Bill Wiedemann, Blaze; Jim Yokouchi, Sumitomo Electric; Jason Yorks, Cielo; Nariman Yousefi, Broadcom;

Presentation Purpose

- Update of March '00 proposal
  - http://grouper.ieee.org/groups/802/3/ae/public/mar00/taborek_1_0300.pdf
- Inclusion of 8B/10B Idle EMI Reduction proposal
  - http://grouper.ieee.org/groups/802/3/ae/public/may00/taborek_1_0500.pdf
- Otherwise, no new material is introduced
- Proposal is ready for Prime Time!

Description

- XAUI = 10 Gigabit eXtended Attachment Unit Interface
- XGXS = XGMII eXtender Sublayer
- CDR-based, 4 lane serial, self-timed interface
- 3.125 Gbaud, 8B/10B encoded over 20” FR-4 PCB traces
- PHY and Protocol independent scalable architecture
- Convenient implementation partition
- May be implemented in CMOS, BiCMOS, SiGe
- Direct mapping of RS/XGMII data to/from PCS
  - XGMII proposed by Howard Frazier, Cisco, et. al.

    http://grouper.ieee.org/groups/802/3/10G_study/public/july99/frazier_1_0799.pdf

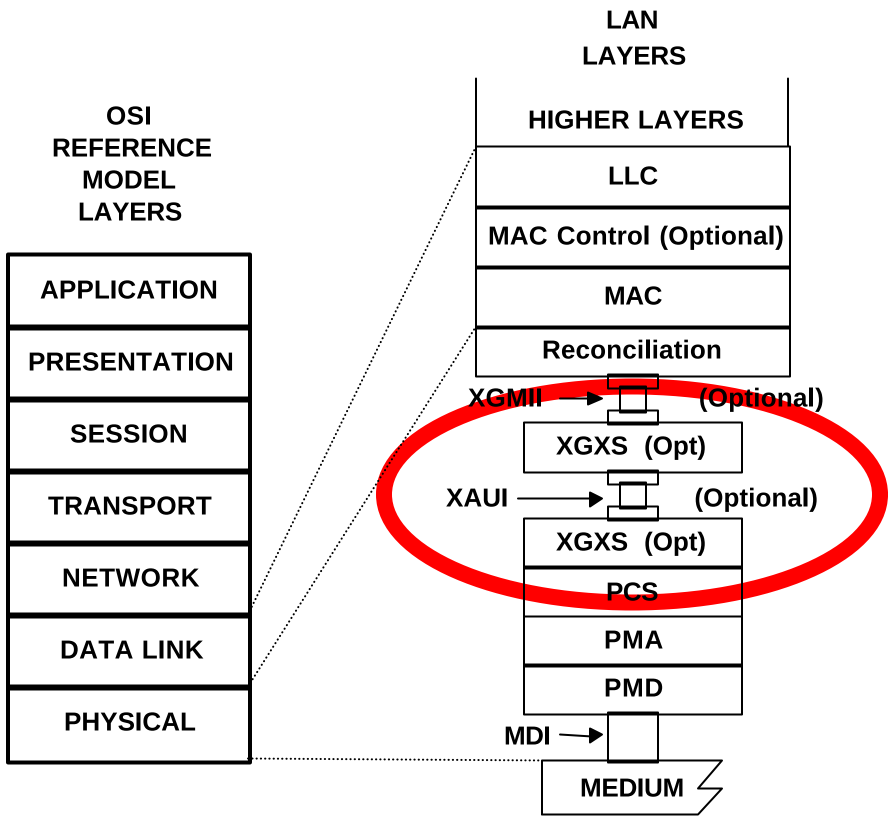
Applications

- Increased XGMII reach
- Low pin count interface = implementation flexibility
- Ease of link design with multiple jitter domains
- Lower power consumption re: XGMII
- Common transceiver module interface, enables SFF
- PCS/PMA agent for WWDM
  - Avoids excessive penalties for all other PHYs
- Self-timed interface eliminates high-speed interface clocks

Highlights

- Increased reach
  - XGMII is ~3” (~7 cm)
  - XAUI is ~20” (~50 cm)
- Lower connection count
  - XGMII is 74 wires (2 sets of 32 data, 4 control & 1 clock)
  - XAUI is 16 wires (2 sets of 4 differential pairs)
- Built-in jitter control
  - Chip-to-chip interconnect degrades XGMII source-synchronous clock
  - XAUI self-timed interface enables jitter attenuation at the receiver

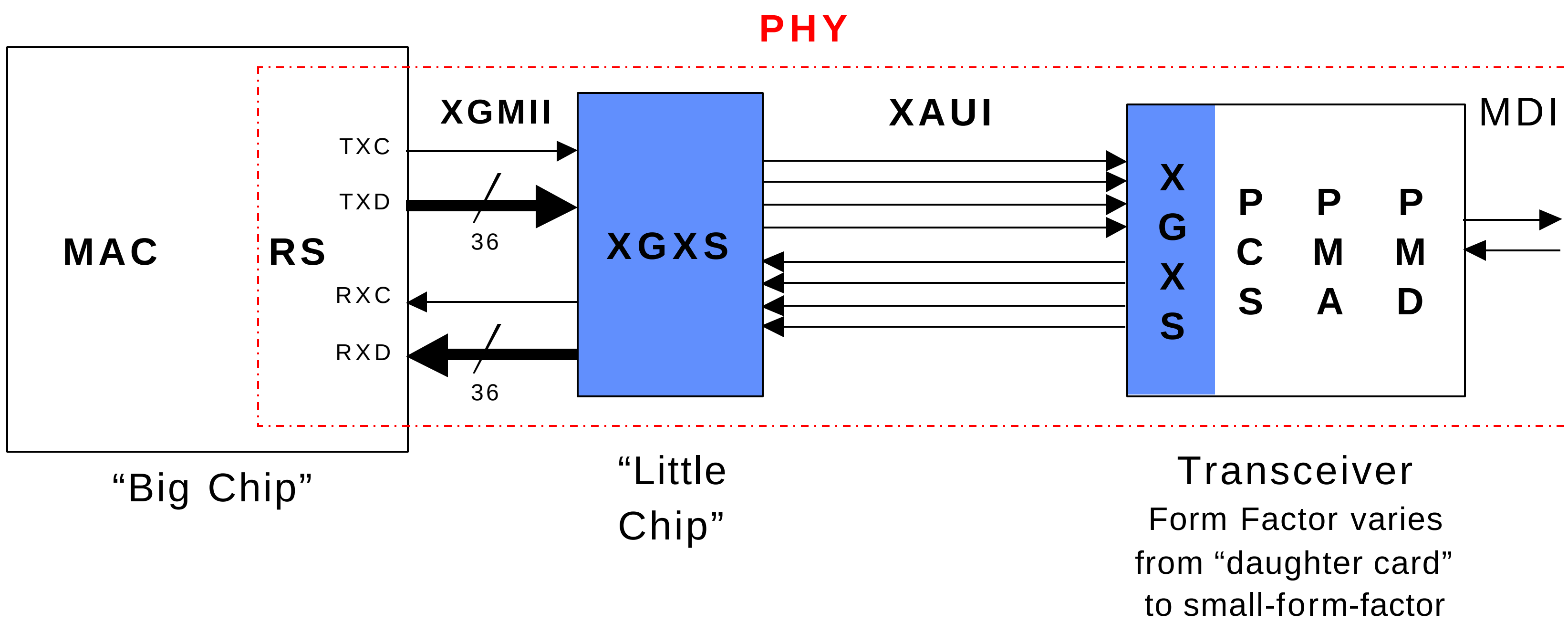
Location - Layer Model



MDI = Medium Dependent Interface
XGMII = 10 Gigabit Media Independent Interface
XAUI = 10 Gigabit Attachment Unit Interface
PCS = Physical Coding Sublayer

XGXS = XGMII Extender Sublayer
PMA = Physical Medium Attachment
PHY = Physical Layer Device
PMD = Physical Medium Dependent

Implementation Example



XGXS Functions

- Use 8B/10B transmission code
- Perform column striping across 4 independent serial lanes
  - Identified as lane 0, lane 1, lane 2, lane 3
- Perform XAUI lane and interface (link) synchronization
- Idle pattern adequate for link initialization
- Perform lane-to-lane deskew
- Perform clock tolerance compensation
- Provide robust packet delimiters
- Perform error control to prevent error propagation

Basic Code Groups

- Similar to GbE
  - No even/odd alignment, new Skip and Align

/A/ K28.3 (Align) - Lane deskew via code-group alignment

/K/ K28.5 (Sync) - Synchronization, EOP Padding

/R/ K28.0 (Skip) - Clock tolerance compensation

/S/ K27.7 (Start) - Start-of-Packet (SOP), Lane 0 ID

/T/ K29.7 (Terminate) - End-of-Packet (EOP)

/E/ K30.7 (Error) - Signaled upon detection of error

/d/ Dxx.y (data) - Packet data

“Extra” Code Groups

- The following are included in related proposals:

/Kb/ K28.1 (Busy Sync) - Synchronization/Rate control

/Rb/ K23.7 (Busy Skip) - Clock tolerance comp/Rate control

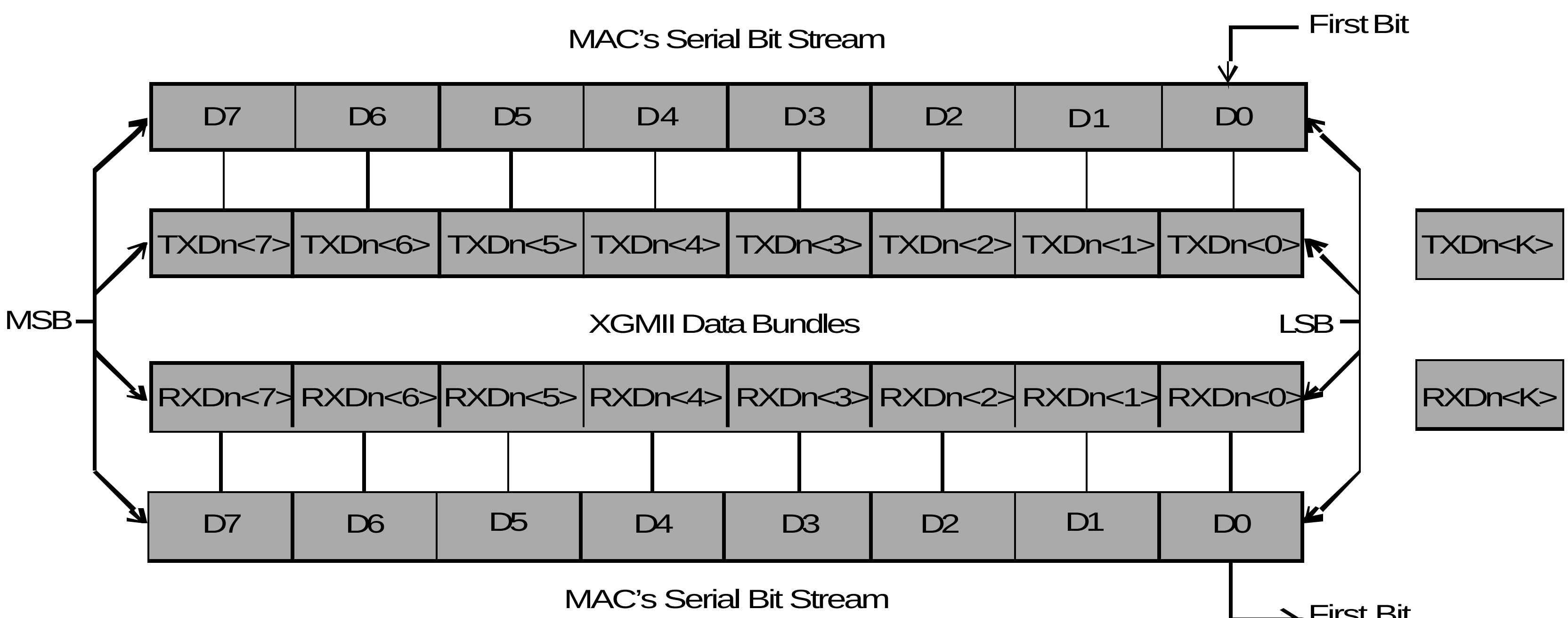
/LS/ K28.1 (Link Signaling) - LSS proposal

- The following remaining 8B/10B special code-groups are not used:

K28.2[1], K28.4, K28.6, K28.7

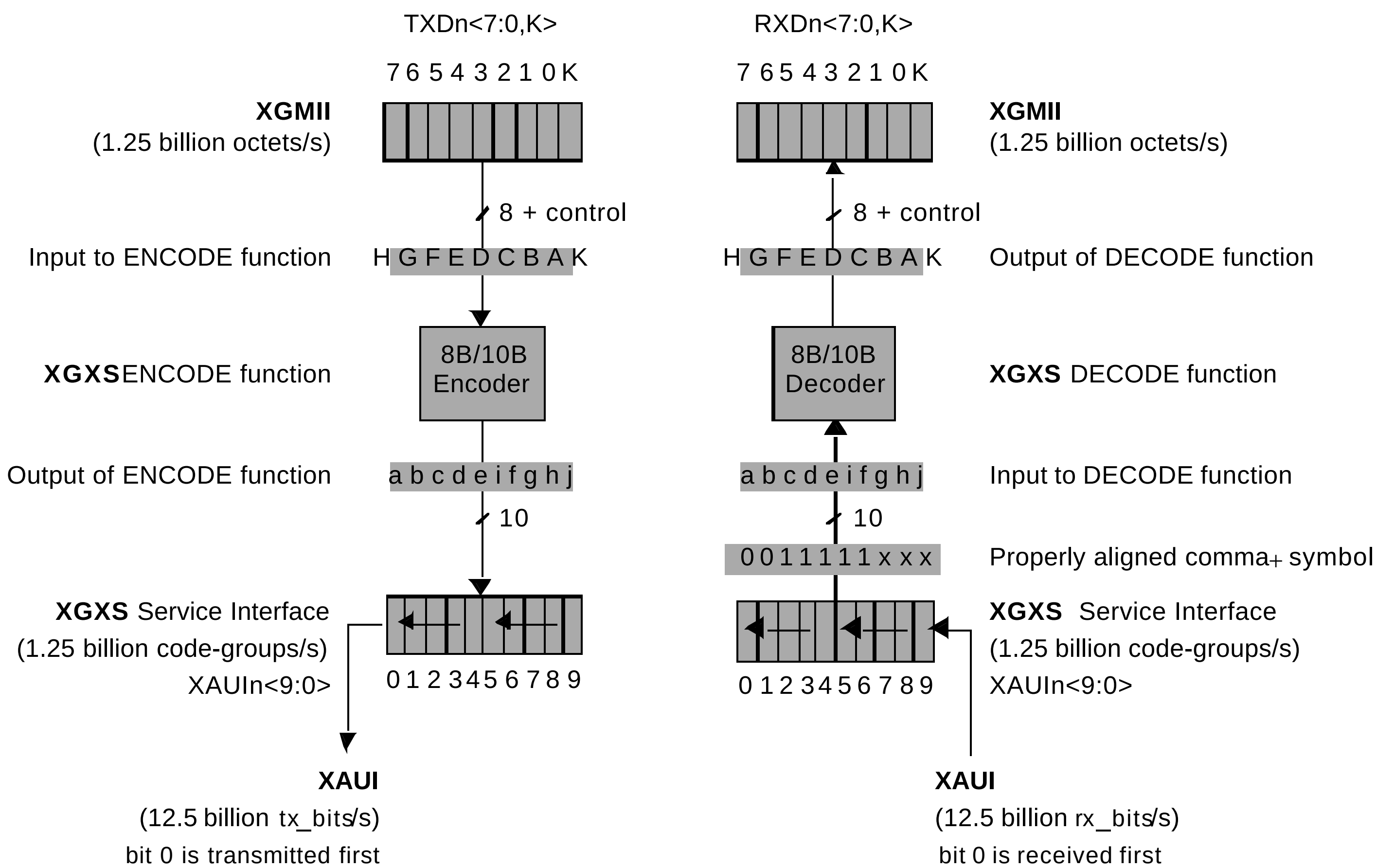
1 Reserved for Fibre Channel usage in NCITS T11 10 GFC project proposals

Data Mapping: MAC to XGMII



MAC is a bit serial interface
XGMII is a 32 bit data + 4 control bit interface
MAC octets represented by D7:0 map to 4 consecutive XGMII Data Bundles in rotating fashion of n=0:3

Data Mapping: XGMII to XAUI



Data Mapping Example

RS/XGMII Encoded Data

D<7:0,K0>	I	I	S	d_p	d	d	---	d	d	d	d_f	I	I	I	I
D<15:8,K1>	I	I	d_p	d_p	d	d	---	d	d	d_f	T	I	I	I	I
D<23:16,K2>	I	I	d_p	d_p	d	d	---	d	d	d_f	I	I	I	I	I
D<31:24,K3>	I	I	d_p	d_p	d	d	---	d	d	d_f	I	I	I	I	I

XGXS Encoded Data

Lane 0	K	R	S	d_p	d	d	---	d	d	d	d_f	A	K	R	K
Lane 1	K	R	d_p	d_p	d	d	---	d	d	d_f	T	A	K	R	K
Lane 2	K	R	d_p	d_p	d	d	---	d	d	d_f	K	A	K	R	K
Lane 3	K	R	d_p	d_p	d	d	---	d	d	d_f	K	A	K	R	K

Idle Encoding

- Idle (no data to send) is conveyed by the randomized pattern /A/ K/R/:

  +T-A+K-R-K+R+R+K-K+R+K-R-K+R+R+R+K (example pattern)

  -K+R+R+K-R-R-K+R+K-R-R-K+R+K-K+R+A ... on each XAUI lane

- ***/A/*** spacing is randomized: 16 min, 32 max (80-bit deskew capability)

- ***/K/R/***s between /A/s randomly selected (no discrete spectrum)

  See http://grouper.ieee.org/groups/802/3/ae/public/may00/taborek_1_0500.pdf for additional details

- /A/, /K/ and /R/ are all a hamming distance of 3 from each other

- Minimum IPG pattern is /A/K/R/ sequence, in order

Synchronization

- XAUI 4-lane link synchronization is a 2 step process
  1. Acquire sync on all 4 lanes individually;
  2. Align/deskew synchronized lanes.
- Loss-of-Sync on any lane results in XAUI link Loss-of-Sync
- Lane sync acquisition similar to 1000BASE-X PCS
  - Use hysteresis to preclude false sync and Loss-of-Sync due to bit errors
  - Re-synchronize only upon Loss-of-Sync (i.e. no "hot-sync")
  - Periodic Align (/A/-column) check a good link health check
- XAUI link sync is fast, straightforward and reliable
- See backup slides for an illustration

Deskew

- Skew is imparted by active and passive link elements
- XGXS deskew accounts for all skew present at the Rx
- Lane deskew performed by alignment to deskew pattern present in Idle/IPG stream: Align /A/ code-groups in all lanes

Skew Source	#	Skew	Total Skew
SerDes Tx	1	1 UI	1 UI
PCB	2	1 UI	2 UI
Medium	1	<16 UI	<16 UI
SerDes Rx	1	20 UI	20 UI
Total			<39 UI

- 40 UI deskew pattern needs to be 80 bits
- /A/ column Idle/IPG spacing is 16 columns (160 bits) minimum

XGXS Deskew

Skewed data at receiver input. Skew ~18 bits

Lane 0	K	K	R	A	K	R	R	K	K	R	K	R
Lane 1	K	K	R	A	K	R	R	K	K	R	K	R
Lane 2	K	K	R	A	K	R	R	K	K	R	K	R
Lane 3	K	K	R	A	K	R	R	K	K	R	K	R

Deskew lanes by lining up Align code-groups

Lane 0	K	K	R	A	K	R	R	K	K	R	K	R
Lane 1	K	K	R	A	K	R	R	K	K	R	K	R
Lane 2	K	K	R	A	K	R	R	K	K	R	K	R
Lane 3	K	K	R	A	K	R	R	K	K	R	K	R

Clock Tolerance Compensation

- The XGXS must restore the temporal fidelity of the signal by:

a. Repeating by amplifying and/or reshaping the signal w/<100% jitter transfer;

b. Retiming the data to a timing reference other than the received data.

- Idle pattern Skip (/R/) columns may be inserted/removed to adjust for clock tolerance differences due to retiming only
  - Skip columns may be inserted anywhere in Idle stream
  - Proper disparity Skip required in each lane
  - Any Skip column may be removed
- Clock tolerance for 1518 byte packet @ ±100 ppm is 0.76 UI/lane
  - A few bytes of elasticity buffering is sufficient to wait for many (~13) frames in case a Skip column is not available for removal.

Skip Column Insert Example

Lane 0	K	R	S	d_p	d	d	---	d	d	d	d_f	A	K	R	K
Lane 1	K	R	d_p	d_p	d	d	---	d	d	d_f	T	A	K	R	K
Lane 2	K	R	d_p	d_p	d	d	---	d	d	d_f	K	A	K	R	K
Lane 3	K	R	d_p	d_s	d	d	---	d	d	d_f	K	A	K	R	K

Skip column inserted here

Lane 0	K	R	S	d_p	d	d	---	d	d	d	d_f	A	R	K	R
Lane 1	K	R	d_p	d_p	d	d	---	d	d	d_f	T	A	R	K	R
Lane 2	K	R	d_p	d_p	d	d	---	d	d	d_f	K	A	R	K	R
Lane 3	K	R	d_p	d_s	d	d	---	d	d	d_f	K	A	R	K	R

Error Control

- Packets with detected errors must be aborted
  - 8B/10B code violation detection may be propagated forward
  - IPG special code groups are chosen to ensure that running disparity errors are detected
- Rule: Signal Error code upon detected error or in column containing EOP if the error is detected in the column following the EOP.
  - Error is signaled per lane since disparity is checked per lane
- XGXS checks received packets for proper formation

Electrical

- Electrical interface is based on low swing AC coupled differential interface
- AC coupling is required at receiver inputs
- Link compliance point is at the receiver
- Transmitter may use equalization as long as receiver specifications are not exceeded

XAUI Rx/Tx & Interconnect

Transmitter Parameter	Value
Vo Dif(max)	800 mV
Vo Dif(min)	500 mV
Voh	AC
Vol	AC
Iout nominal	6.5 mA
Differential Skew(max)	15 ps

Receiver Parameter	Value
Vin Dif(max)	1000 mV
Vin Dif(min)	175 mV
Loss 50W	9.1 dB
Differential Skew(max)	75 ps

Interconnect Parameter	Value
Tr/Tf Min, 20%-80%	60 ps[1]
Tr/Tf Max, 20%-80%	131 ps[1]
PCB Impedance	100 ±10W
Connector Impedance	100 ±30W
Source Impedance	100 ±20W
Load Termination	100 ±20W
Return Loss	10 dB[2]

1. Optional if transmitter meets the receiver jitter and eye mask with golden PCB
2. SerDes inputs must meet the return loss from 100 MHz to 2.5 GHz (0.8 x 3.125 Gbaud)

XAUI Loss Budget

Item	Loss
Connector Loss	1 dB
NEXT + FEXT Loss	0.75 dB
PCB Loss	7.35 dB
Loss Budget	9.1 dB

PCB Condition	Normal	Worst
MSTL Loss Max (dB/in)	0.32	0.43
Max Distance (in)	23”	17.1”

Normal PCB was assumed with loss tangent of 0.22. Worst case it was assumed high temperature and humidity 85/85.

Better FR4 grade may reduce loss by as much as 50%.

PCB Condition	Normal	Worst
STL Loss Max (dB/in)	0.41	0.55
Max Distance (in)	18”	13.4”

HP test measurement for 20" line showed 5.2 dB loss or 0.26dB/ in based on the eye loss, the loss assumed here is very conservative.

XAUI Jitter

Jitter Compliance Point	Tx[1]	Rx
Deterministic Jitter	0.17 UI	**0.41 UI**
Total Jitter	0.35 UI[2]	**0.65 UI**
1-sigma RJ @ max DJ for 10^{-12} BER[3]	4.11 ps	**5.49 ps**
1-sigma RJ @ max DJ for 10^{-13} BER[3]	3.92 ps	5.23 ps

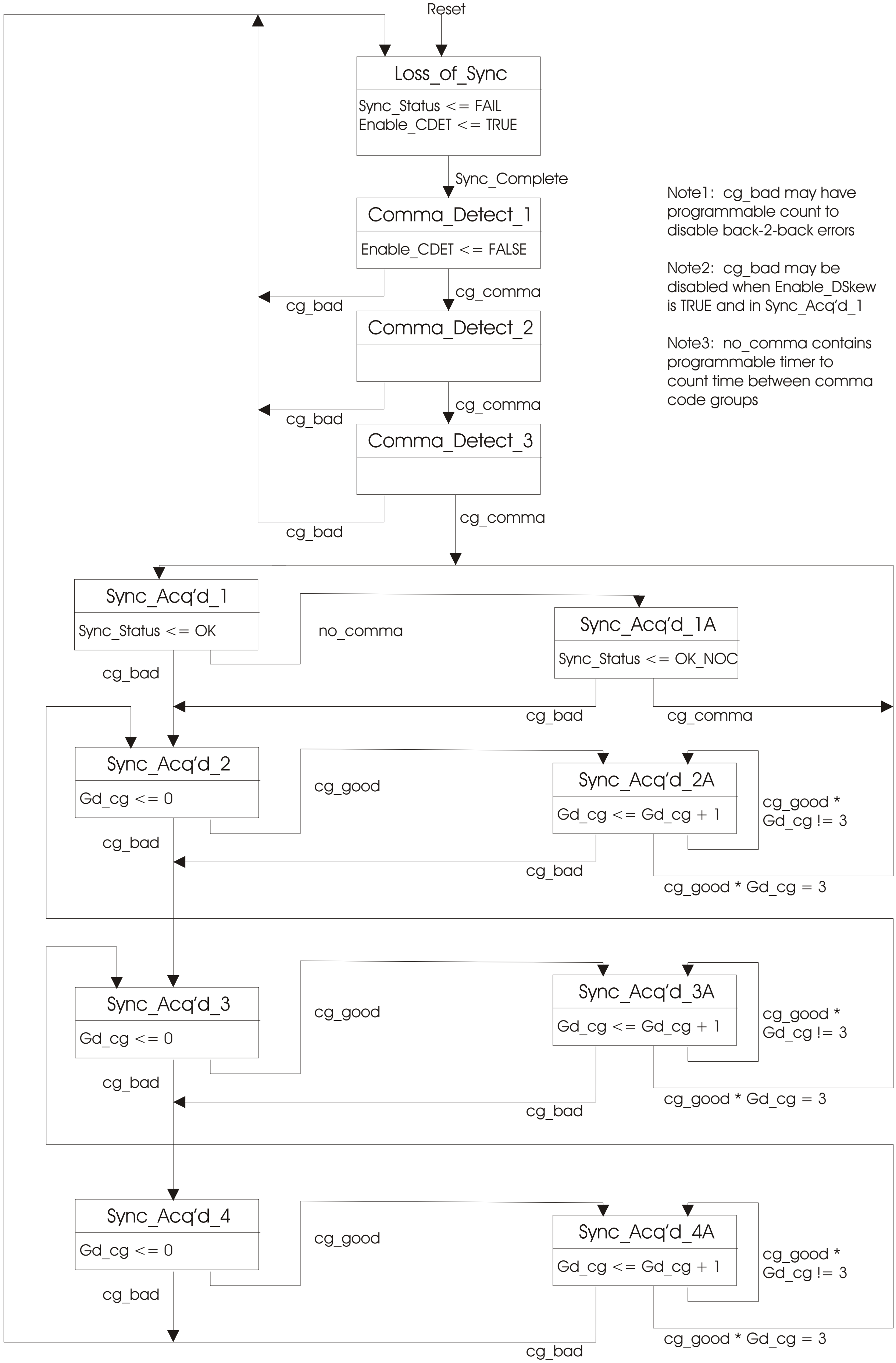
1. Tx point is for reference. Rx point is for compliance.
2. **The SerDes component should have better jitter performance than specified here to allow for system noise.**
3. 1-Sigma value listed here are at maximum DJ, if the DJ value is smaller then the 1-Sigma RJ may increase to the total jitter value.

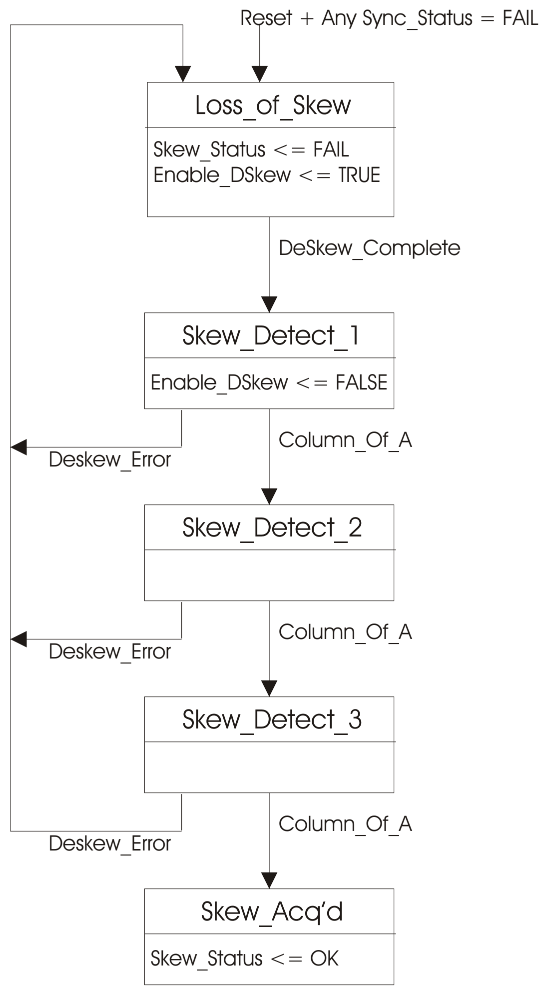
Summary

- Meets HSSG Objectives and PAR 5 Criteria
- Provides PHY, Protocol & Application independence
- Based on generic 10 Gbps chip-to-chip interconnect
- Resembles simple and familiar 1000BASE-X PHY
- Low Complexity, Low Latency, Quick Synchronizing
- May be integrated into MAC/RS ASIC, eliminating XGMII

Backup Slides

- XGXS Synchronization state diagrams


Reset
Loss_of_Sync
Sync_Status <= FAIL
Enable_CDET <= TRUE
Sync_Complete
Comma_Detect_1
Enable_CDET <= FALSE
cg_comma
cg_bad
Comma_Detect_2
cg_comma
cg_bad
Comma_Detect_3
cg_comma
cg_bad
Note1: cg_bad may have programmable count to disable back-2-back errors
Note2: cg_bad may be disabled when Enable_DSkew is TRUE and in Sync_Acq'd_1
Note3: no_comma contains programmable timer to count time between comma code groups
Sync_Acq'd_1
Sync_Status <= OK
no_comma
Sync_Acq'd_1A
Sync_Status <= OK_NOC
cg_bad
cg_bad
cg_comma
Sync_Acq'd_2
Gd_cg <= 0
cg_good
Sync_Acq'd_2A
Gd_cg <= Gd_cg + 1
cg_good * Gd_cg != 3
cg_bad
cg_bad
cg_good * Gd_cg = 3
Sync_Acq'd_3
Gd_cg <= 0
cg_good
Sync_Acq'd_3A
Gd_cg <= Gd_cg + 1
cg_good * Gd_cg != 3
cg_bad
cg_bad
cg_good * Gd_cg = 3
Sync_Acq'd_4
Gd_cg <= 0
cg_good
Sync_Acq'd_4A
Gd_cg <= Gd_cg + 1
cg_good * Gd_cg != 3
cg_bad
cg_bad
cg_good * Gd_cg = 3



Reset + Any Sync_Status = FAIL
Loss_of_Skew
Skew_Status <= FAIL
Enable_DSkew <= TRUE
DeSkew_Complete
Skew_Detect_1
Enable_DSkew <= FALSE
Deskew_Error
Column_Of_A
Skew_Detect_2
Deskew_Error
Column_Of_A
Skew_Detect_3
Deskew_Error
Column_Of_A
Skew_Acq'd
Skew_Status <= OK

8B/10B Idle Pattern
for
12-byte IPG

Rich Taborek, Don Alderrou nSerial

Presentation Purpose

- Modify 8B/10B Idle pattern to handle 12-byte IPG:
- Maintain all 8B/10B Idle pattern benefits

8B/10B Idle Pattern

- Current proposed 8B/10B Idle pattern
  - Fixed /A/K/R/ followed by randomized /A/ spacing and /K/R/ sequence
  - /K/ used to pad EOP column
- Problem: 12-byte IPG could compromise /R/ availability
  - Affects ability to perform clock tolerance compensation
  - Can't simply rearrange to place /R/ first:
    - Causes /A/ or /K/ starvation, and/or,
    - /R/ deletion may compromise EOP robustness
- Solution: Modify fixed /A/K/R/ to guarantee /R/
  - Start with random /A/K/ as first column following EOP
  - Second column is fixed /R/
  - Third and subsequent columns randomize /A/ spacing and /K/R/ sequence

Data Mapping Example

RS/XGMII Encoded Data

D<7:0,K0>	I	I	S	d_p	d	d	---	d	d	d_f	T	I	I	S	d_p
D<15:8,K1>	I	I	d_p	d_p	d	d	---	d	d	d_f	I	I	I	d_p	d_p
D<23:16,K2>	I	I	d_p	d_p	d	d	---	d	d	d_f	I	I	I	d_p	d_p
D<31:24,K3>	I	I	d_p	d_p	d	d	---	d	d	d_f	I	I	I	d_p	d_p

PCS Encoded Data

Lane 0	R	K	S	d_p	d	d	---	d	d	d	T	A	R	S	d_p
Lane 1	R	K	d_p	d_p	d	d	---	d	d	d_f	K	A	R	d_p	d_p
Lane 2	R	K	d_p	d_p	d	d	---	d	d	d_f	K	A	R	d_p	d_p
Lane 3	R	K	d_p	d_p	d	d	---	d	d	d_f	K	A	R	d_p	d_p

Summary

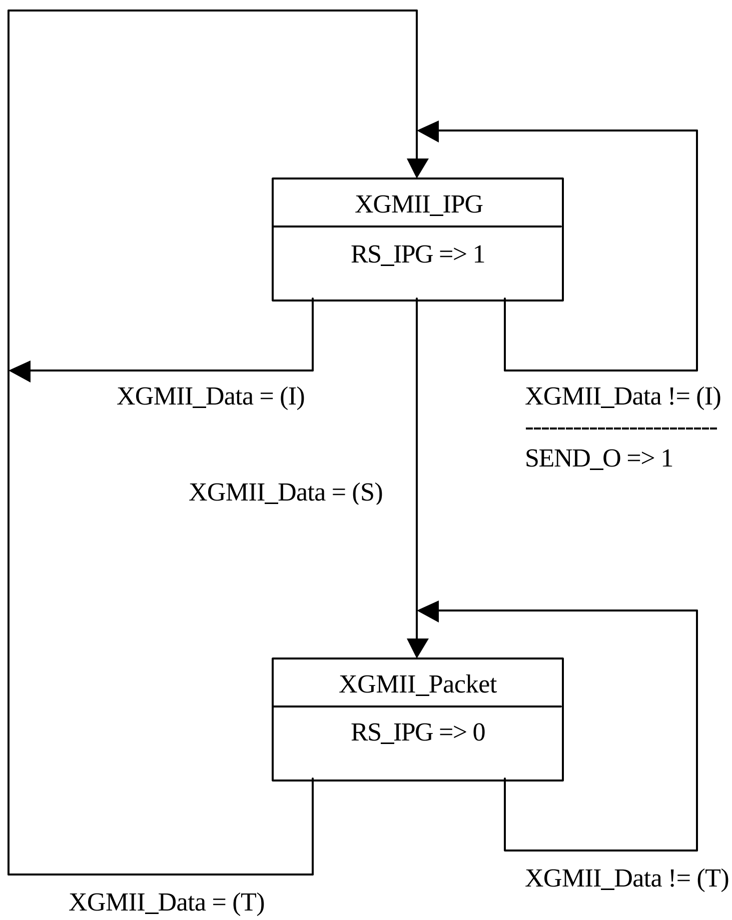
- Concerns of 8B/10B Idle pattern for 12-byte IPG addressed
- Solution is simple rearrangement of fixed Idle start pattern
- Retain all benefits of 8B/10B-based PCS and PMA
- Retain all benefits of XAUI/XGXS protocol
- No additional burden on receiver
- Retain all benefits of Idle EMI enhancements
- All benefits applicable to PCB traces & 4 Channel PMDs

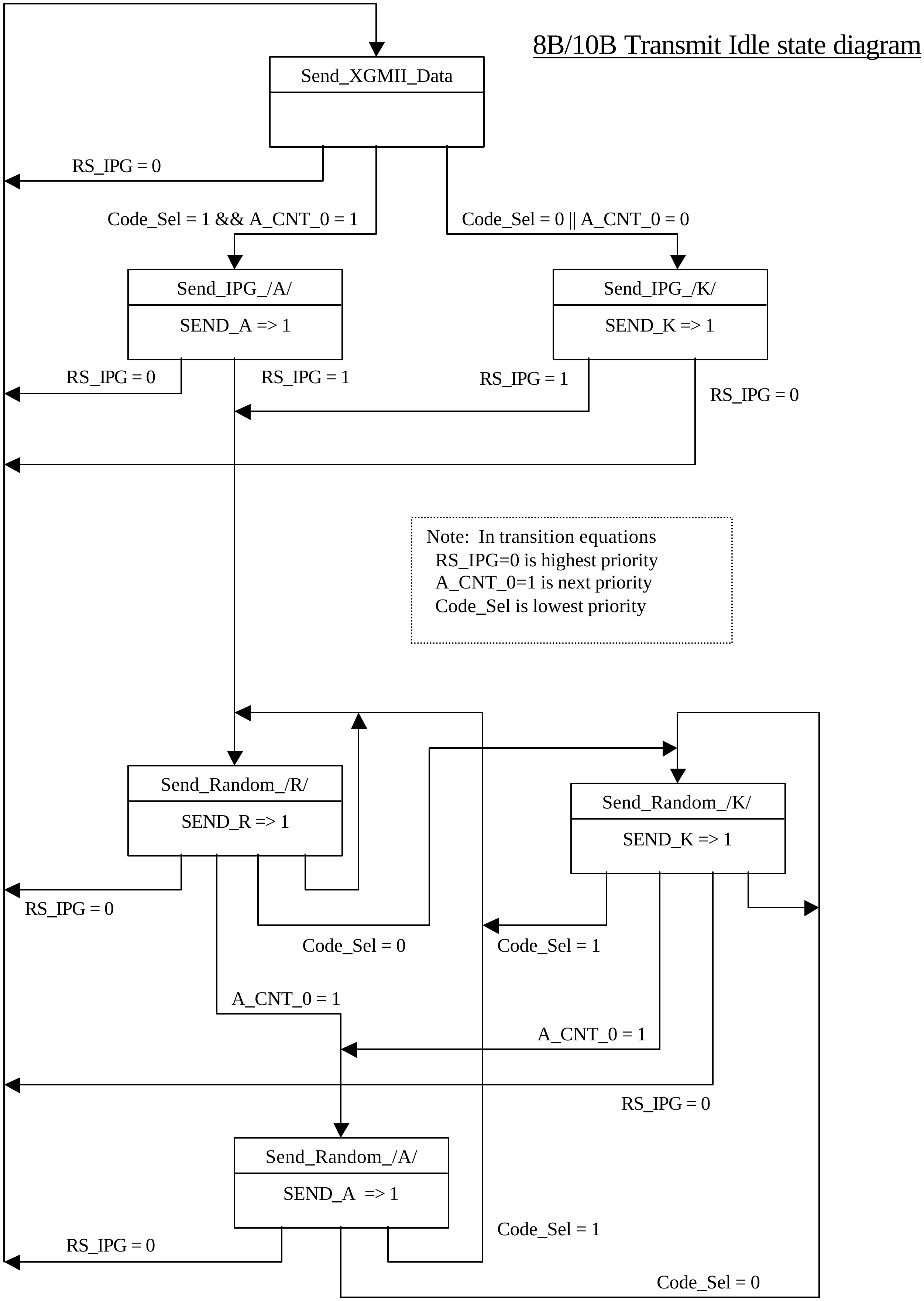
Supplementary Slides

Intended for those that REALLY want to know how this stuff works

- 8B/10B Transmit state diagram
  - Transmit IPG, SOP, EOP or Other (e.g. LSS)
- 8B/10B Transmit Idle state diagram
  - Generate IPG/Random AKR Idle
- 8B/10B Transmit Idle logic diagram
  - AKR Randomizer
- 8B/10B Transmit Data multiplexer diagram
  - Multiplexing of XGMII input and Random AKR Idle

8B/10B Transmit state diagram



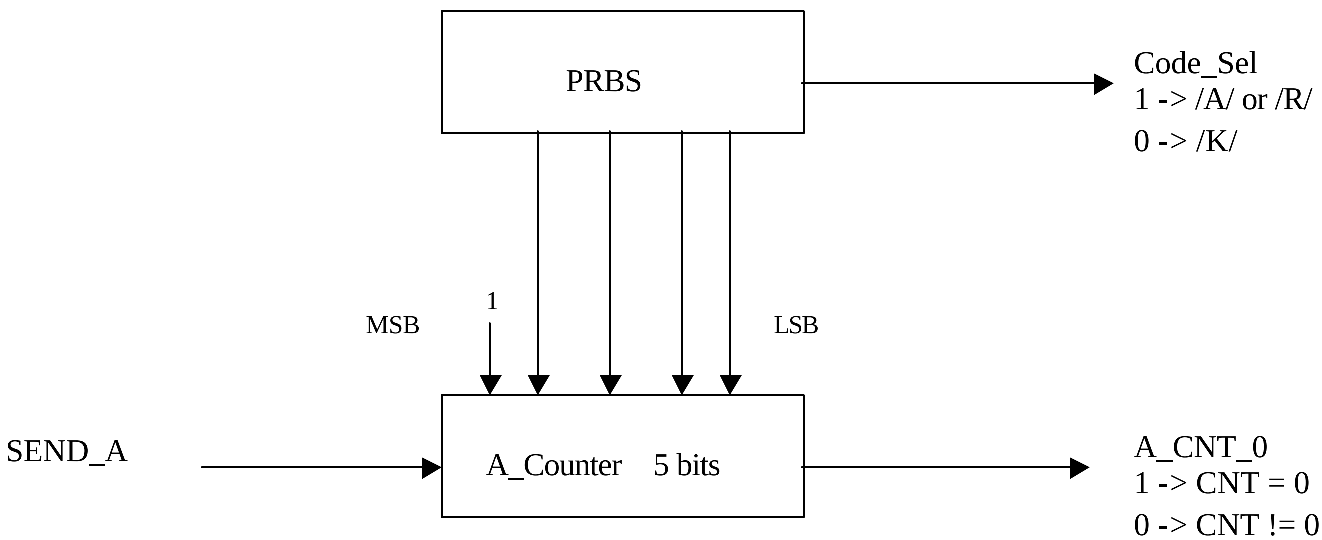


8B/10B Transmit Idle state diagram
Send_XGMII_Data
RS_IPG = 0
Code_Sel = 1 && A_CNT_0 = 1
Code_Sel = 0 || A_CNT_0 = 0
Send_IPG_/A/
SEND_A => 1
Send_IPG_/K/
SEND_K => 1
RS_IPG = 0
RS_IPG = 1
RS_IPG = 1
RS_IPG = 0
Note: In transition equations
RS_IPG=0 is highest priority
A_CNT_0=1 is next priority
Code_Sel is lowest priority
Send_Random_/R/
SEND_R => 1
Send_Random_/K/
SEND_K => 1
RS_IPG = 0
Code_Sel = 0
Code_Sel = 1
A_CNT_0 = 1
A_CNT_0 = 1
RS_IPG = 0
Send_Random_/A/
SEND_A => 1
Code_Sel = 1
RS_IPG = 0
Code_Sel = 0

8B/10B Transmit Idle logic diagram

The polynomial for the Pseudo-Random Bit Sequencer (PRBS) which has been simulated and tested in the lab is X^7 + X^3 + 1.

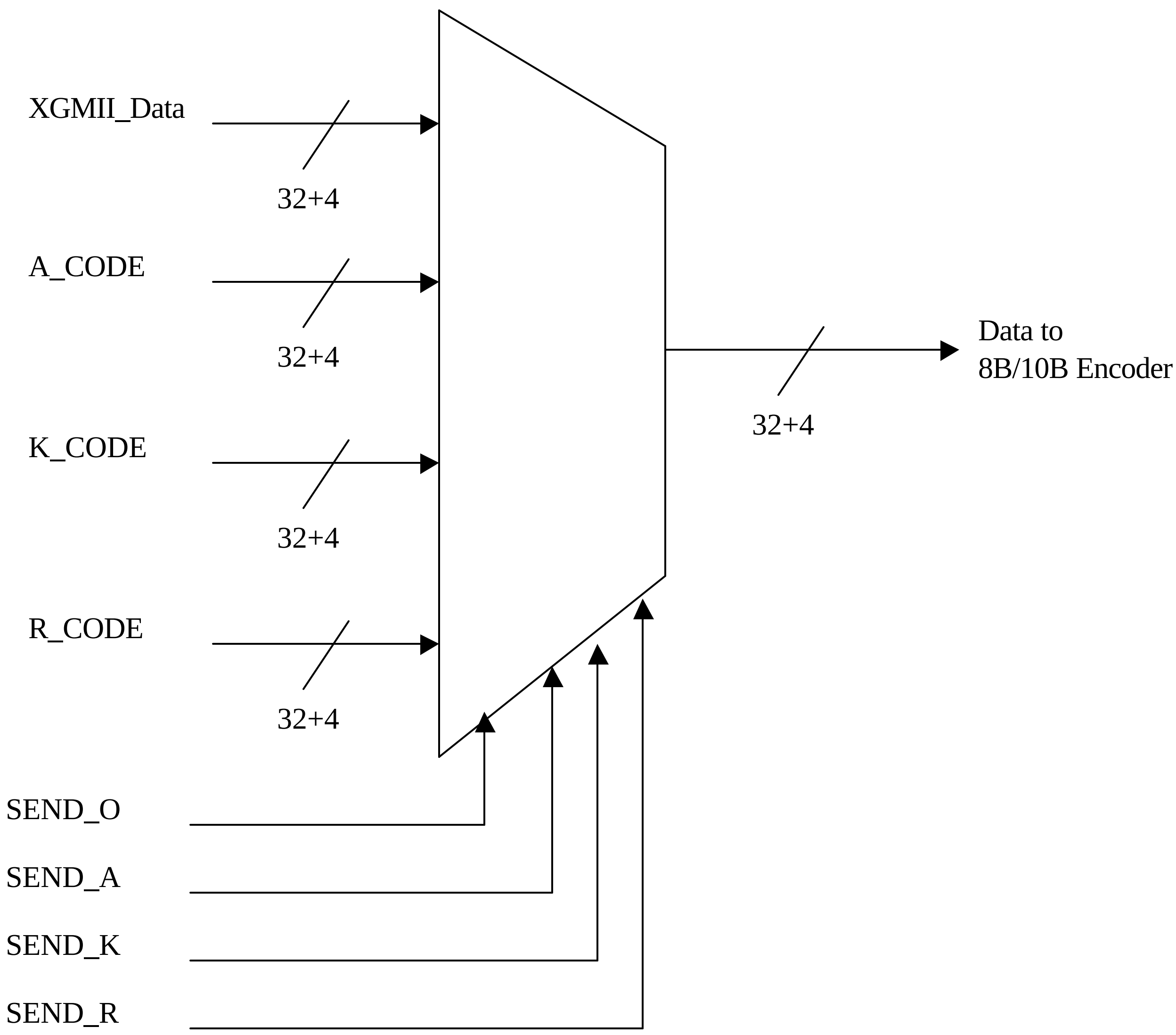
Note: it runs serially (one data bit shift per clock) at the byte clock rate.



The A_Counter counts down to zero and is parallel loaded with a random 4-bit pattern from four of the PRBS stages. The MSB is always loaded with a 1 to give the random count between /A/ codes of 16 to 31. It is loaded by the SM when an /A/ code is sent and signals a zero count back to the SM.

8B/10B Transmit Data multiplexer diagram

The data multiplexer selects either the XGMII 32-bit data & 4-bit control or one of the special codes. If none of the SEND_x signals are active, then the XGMII data & control is selected. The SEND_O signal has priority over the other SEND_x signals and will select the XGMII data & control.



64b/66b PCS

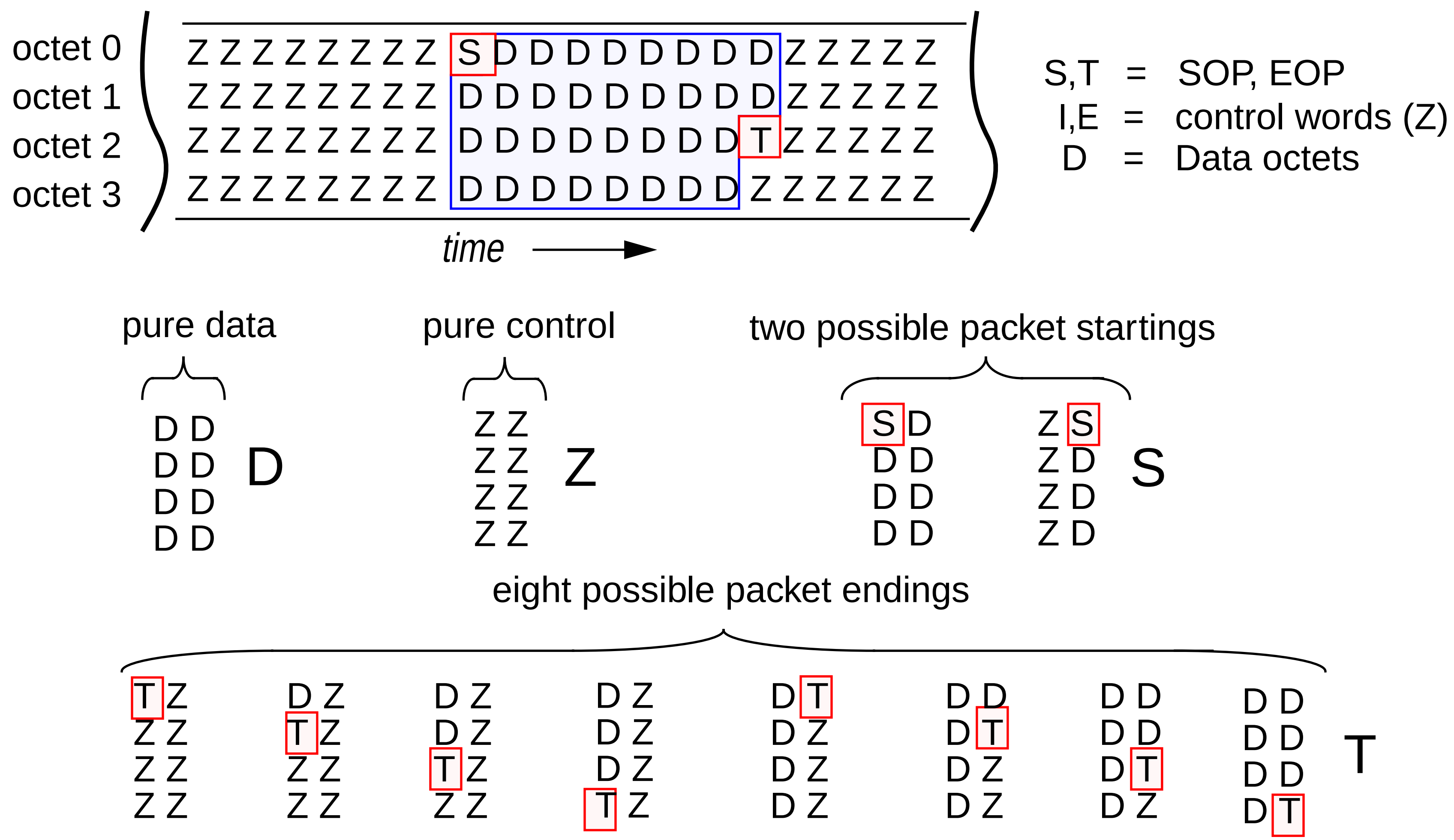
updated 6/30/2000

Rick Walker	Agilent	Howard Frazier	Cisco
Richard Dugan	Agilent	Paul Bottorff	Nortel
Birdy Amrutur	Agilent	Shimon Mueller	Sun
Rich Taborek	nSerial	Brad Booth	Intel
Don Alderrou	nSerial	Kevin Daines	World Wide Packets
John Ewen	IBM	Osamu Ishida	NTT
Mark Ritter	IBM	Jason Yorks	Cielo
Al Bezoni	Lucent	Henning Lysdal	Giga/Intel
Drew Plant	Agilent	Justin Chang	Quake

Topics

- Code review and update
- Test vectors
- Bit ordering sequence
- Frame sync algorithm and state machine
- TX,RX error detection state machines
- Optional code features
- Summary

Building frames from 10GbE RS symbols



Code Overview

Data Codewords have “01” sync preamble

Mixed Data/Control frames are identified with a “10” sync preamble. Both the coded 56-bit payload and TYPE field are scrambled

00,11 preambles are considered code errors and cause the packet to be invalidated by forcing an error (E) symbol on coder output

Code Summary

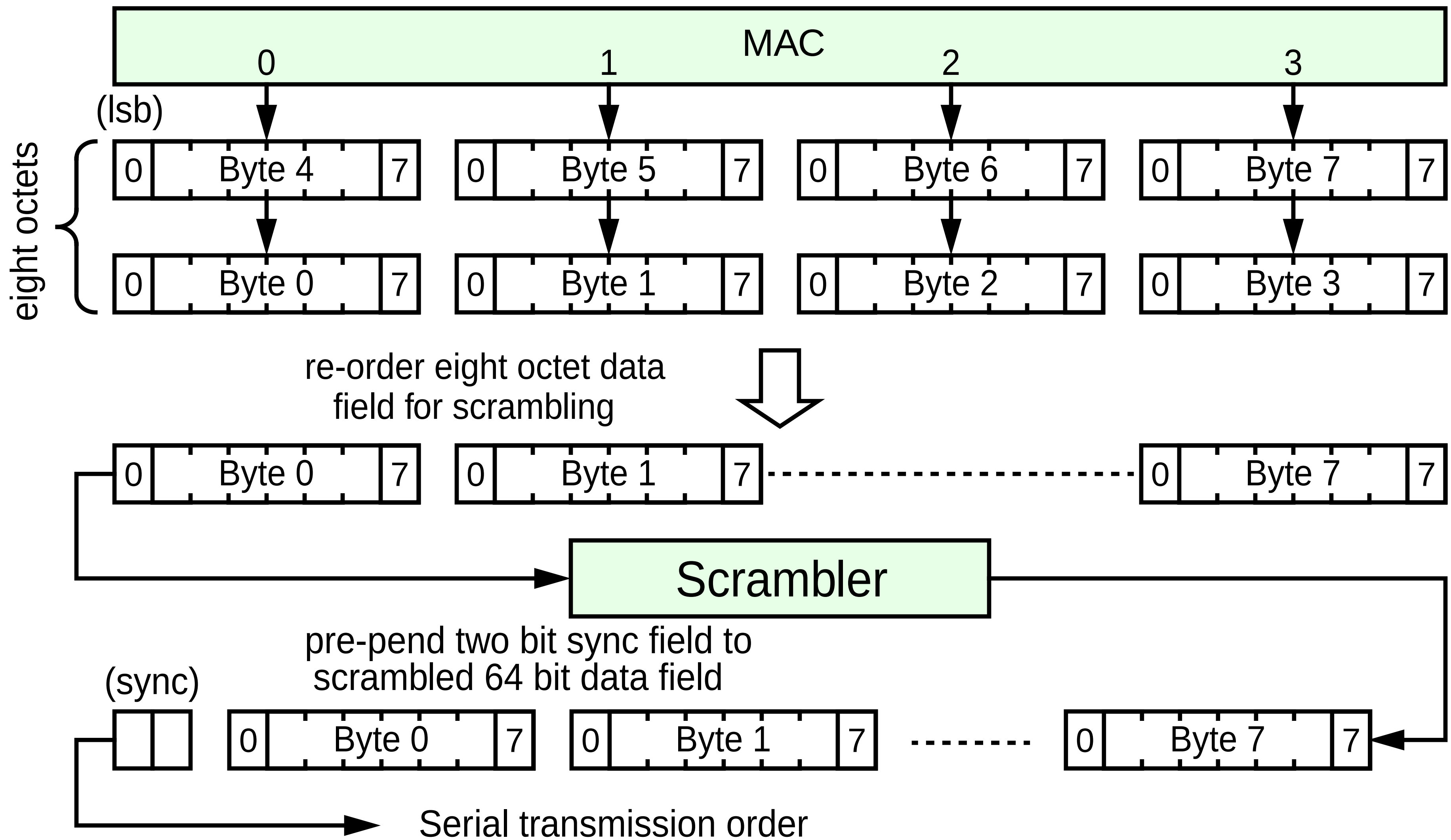
Input Data	Sync		Bit fields								
(first RS transfer / second RS transfer)	[0]	[1]	[2]								[65]
$D_0D_1D_2D_3/D_4D_5D_6D_7$	0	1	D_0 [0] [7]	D_1 [0] [7]	D_2 [0] [7]	D_3 [0] [7]	D_4 [0] [7]	D_5 [0] [7]	D_6 [0] [7]	D_7 [0] [7]	
$Z_0Z_1Z_2Z_3/Z_4Z_5Z_6Z_7$	1	0	0x1e “01111000”	C_0 [0] [6]	C_1 [0] [6]	C_2 [0] [6]	C_3 [0] [6]	C_4 [0] [6]	C_5 [0] [6]	C_6 [0] [6]	C_7 [0] [6]
$Z_0Z_1Z_2Z_3/S_4D_5D_6D_7$	1	0	0x33	C_0	C_1	C_2	C_3		D_5	D_6	D_7
$S_0D_1D_2D_3/D_4D_5D_6D_7$	1	0	0x78	D_1	D_2	D_3	D_4	D_5	D_6	D_7	
$T_0Z_1Z_2Z_3/Z_4Z_5Z_6Z_7$	1	0	0x87		C_1	C_2	C_3	C_4	C_5	C_6	C_7
$D_0T_1Z_2Z_3/Z_4Z_5Z_6Z_7$	1	0	0x99	D_0		C_2	C_3	C_4	C_5	C_6	C_7
$D_0D_1T_2Z_3/Z_4Z_5Z_6Z_7$	1	0	0xaa	D_0	D_1		C_3	C_4	C_5	C_6	C_7
$D_0D_1D_2T_3/Z_4Z_5Z_6Z_7$	1	0	0xb4	D_0	D_1	D_2		C_4	C_5	C_6	C_7
$D_0D_1D_2D_3/T_4Z_5Z_6Z_7$	1	0	0xcc	D_0	D_1	D_2	D_3		C_5	C_6	C_7
$D_0D_1D_2D_3/D_4T_5Z_6Z_7$	1	0	0xd2	D_0	D_1	D_2	D_3	D_4		C_6	C_7
$D_0D_1D_2D_3/D_4D_5T_6Z_7$	1	0	0xe1	D_0	D_1	D_2	D_3	D_4	D_5		C_7
$D_0D_1D_2D_3/D_4D_5D_6T_7$	1	0	0xff	D_0	D_1	D_2	D_3	D_4	D_5	D_6	

- all undefined bit fields (in yellow) are set to zero for 10GbE

RS “Z” code to 7 bit “C” field mapping

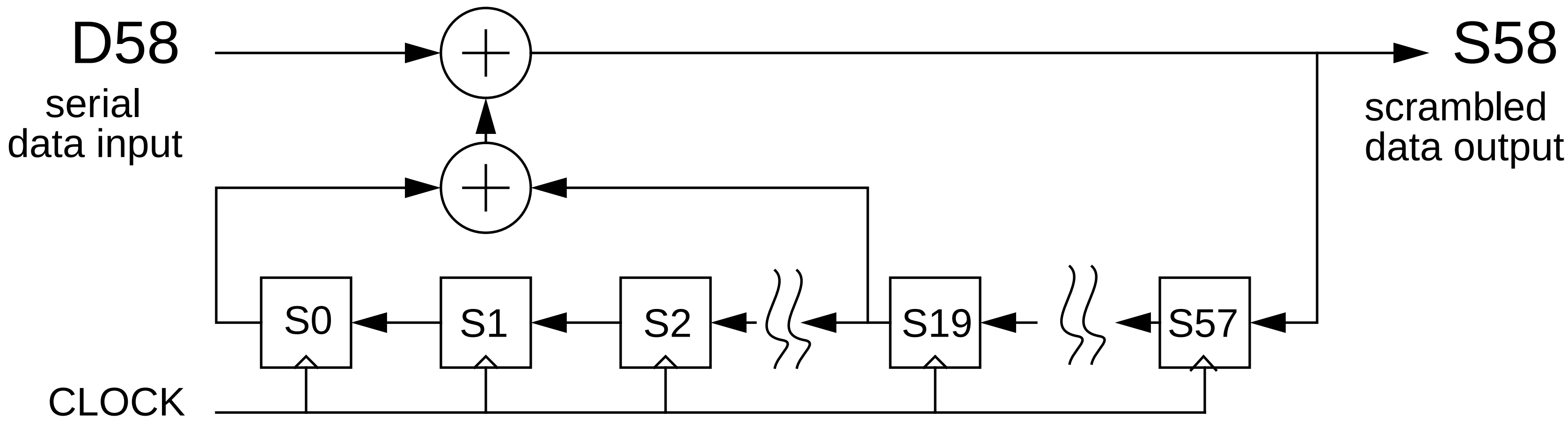
RS Z value	name	shorthand	7-bit C field line code
0x07,1	idle	[I]	0x00
0xfb,1	start	[S]	encoded by TYPE byte
0xfd,1	terminate	[T]	encoded by TYPE byte
0xfe,1	error	[E]	0x1e
0x1c,1	reserved0	-	0x2d
0x3c,1	reserved1	-	0x33
0x7c,1	reserved2	-	0x4b
0xbc,1	reserved3	-	0x55
0xdc,1	reserved4	-	0x66
0xf7,1	reserved5	-	0x78

Bit ordering sequence



Scrambler definition

Serial form of the Scrambler:



The serial form of the scrambler is shown here for bit ordering purposes. Parallel implementations could also be used. For details see:

http://grouper.ieee.org/groups/802/3/ae/public/mar00/walker_1_0300.pdf

Sample 64b/66b Test Vector

- Start with a minimum length (64 byte) Ethernet packet with preamble and CRC

```
55 55 55 55 55 55 d5 08 00 20 77 05 38 0e 8b 00 00 00 00 08 00 45 00 00 28 1c 66 00 00 1b 06 9e
d7 00 00 59 4d 00 00 68 d1 39 28 4a eb 00 00 30 77 00 00 7a 0c 50 12 1e d2 62 84 00 00 00 00 00
00 00 00 93 eb f7 79
```

- Add SOP, EOP, Idles and convert to RS indications

```
07,1 07,1 07,1 07,1 07,1 07,1 07,1 07,1  fb,1 55,0 55,0 55,0 55,0 55,0 55,0 d5,0
08,0 00,0 20,0 77,0 05,0 38,0 0e,0 8b,0  00,0 00,0 00,0 00,0 08,0 00,0 45,0 00,0
00,0 28,0 1c,0 66,0 00,0 00,0 1b,0 06,0  9e,0 d7,0 00,0 00,0 59,0 4d,0 00,0 00,0
68,0 d1,0 39,0 28,0 4a,0 eb,0 00,0 00,0  30,0 77,0 00,0 00,0 7a,0 0c,0 50,0 12,0
1e,0 d2,0 62,0 84,0 00,0 00,0 00,0 00,0  00,0 00,0 00,0 00,0 93,0 eb,0 f7,0 79,0
fd,1 07,1 07,1 07,1 07,1 07,1 07,1 07,1
```

- Arrange bytes into frames with type indicators and sync bits

```
"10" 1e 00 00 00 00 00 00 00 "10" 78 55 55 55 55 55 55 d5 "01" 08 00 20 77 05 38 0e 8b
"01" 00 00 00 00 08 00 45 00 "01" 00 28 1c 66 00 00 1b 06 "01" 9e d7 00 00 59 4d 00 00
"01" 68 d1 39 28 4a eb 00 00 "01" 30 77 00 00 7a 0c 50 12 "01" 1e d2 62 84 00 00 00 00
"01" 00 00 00 00 93 eb f7 79 "10" 87 00 00 00 00 00 00 00
```

- Scramble and transmit left-to-right, lsb first, (scrambler initial state is set to all ones)

```
"10" 1e 00 00 00 80 f0 ff 7b "10" 78 15 ad aa aa 16 30 62
"01" 08 e1 81 c5 6e 7c 76 6a "01" e6 30 28 80 cc aa f4 8d
"01" 83 ee 49 ae 6d 93 db 2c "01" f3 46 70 db 82 5a 90 74
"01" 1e 51 79 6b 1a 25 7a c5 "01" 41 1f bf d4 0c 44 ca 4a
"01" 09 28 12 d2 b5 2d 3f 2c "01" 49 92 de c8 b3 33 0e 32
"10" 2a a3 3a c8 d7 ad 99 b5
```

Frame alignment algorithm

Look for presence of “01” or “10” sync patterns every 66 bits

This can be done either in parallel, by looking at all possible locations, or in serial by looking at only one potential location.

In either case, a frame sync detector is used to statistically qualify a valid sync alignment.

In the parallel case, a barrel shifter can immediately make the phase shift adjustment. In the serial case, a sync error is used to cycle-slip the demultiplexor to hunt for a valid sync phase.

So what algorithm should be used for reliable and rapid frame sync detection?

Frame sync criteria

If misaligned, then sync error rate will be 50%. We must quickly assert loss of sync and "slip" our alignment to another candidate location

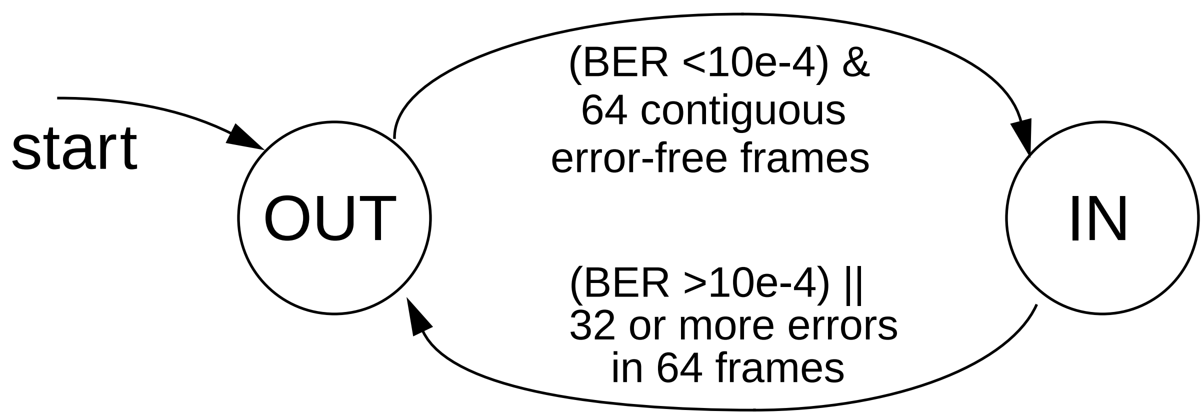
If already aligned with good BER (<10e-9), then we want to stay in sync with very high reliability

If BER is worse than10e-4 we should suppress sync, to avoid likelyhood of False Packet Acceptance due to CRC failures

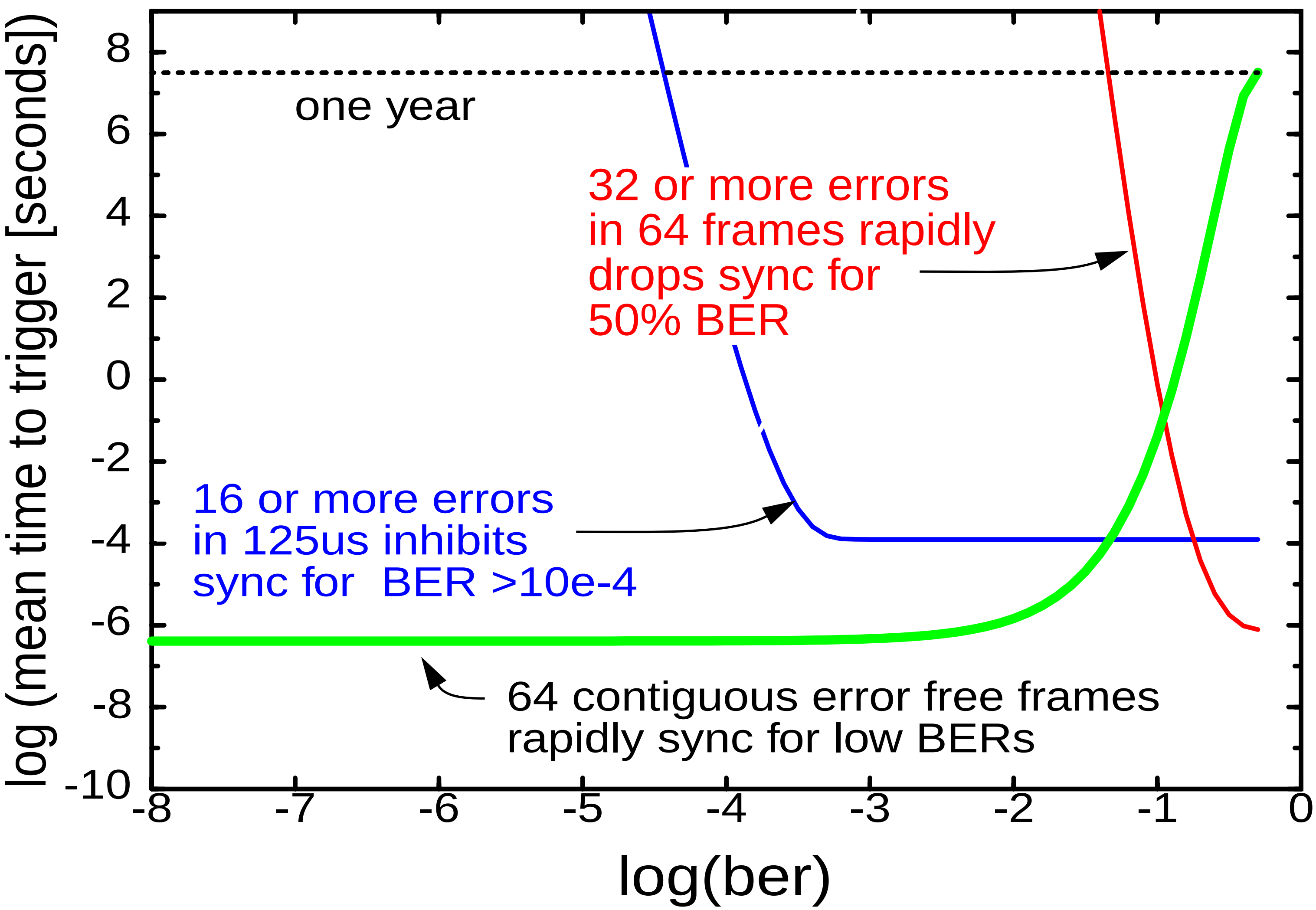
BER	current sync state	next sync state	notes
~50%	in	out	should be fast
>10e-4	in	out	prevents MTTFPA events, can be relatively slow to trigger
<10e-9	out	in	should be fast

Frame sync algorithm

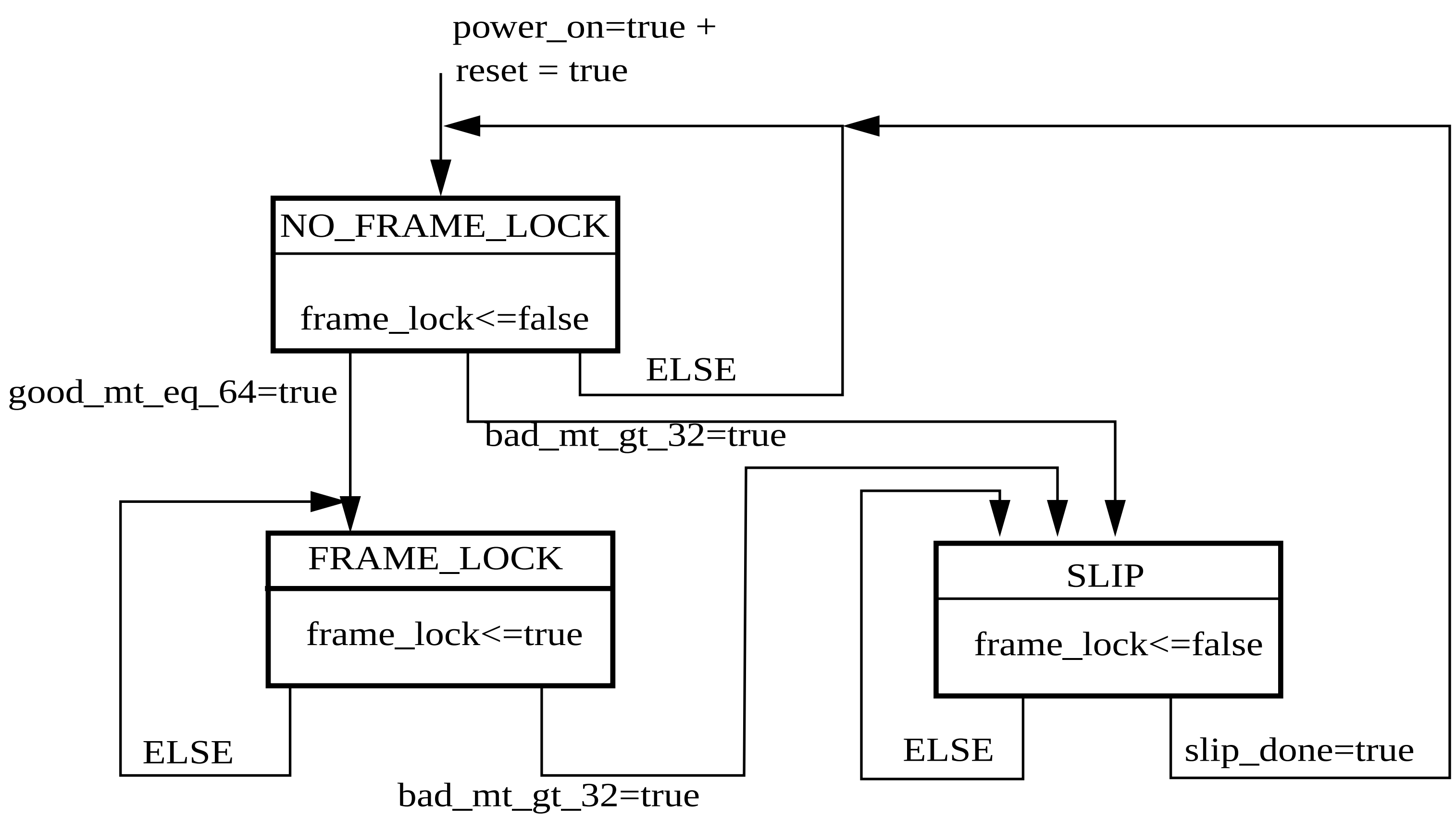
- frame sync is acquired after 64 contiguous frames have been received with valid “01” or “10” sync headers
- frame sync is declared lost after 32 “11” or “00” sync patterns have been declared in any block of 64 frames
- In addition, if there are 16 or more errors within any 125us time interval (~10e-4 BER), then frame sync is inhibited



64/66 frame sync performance

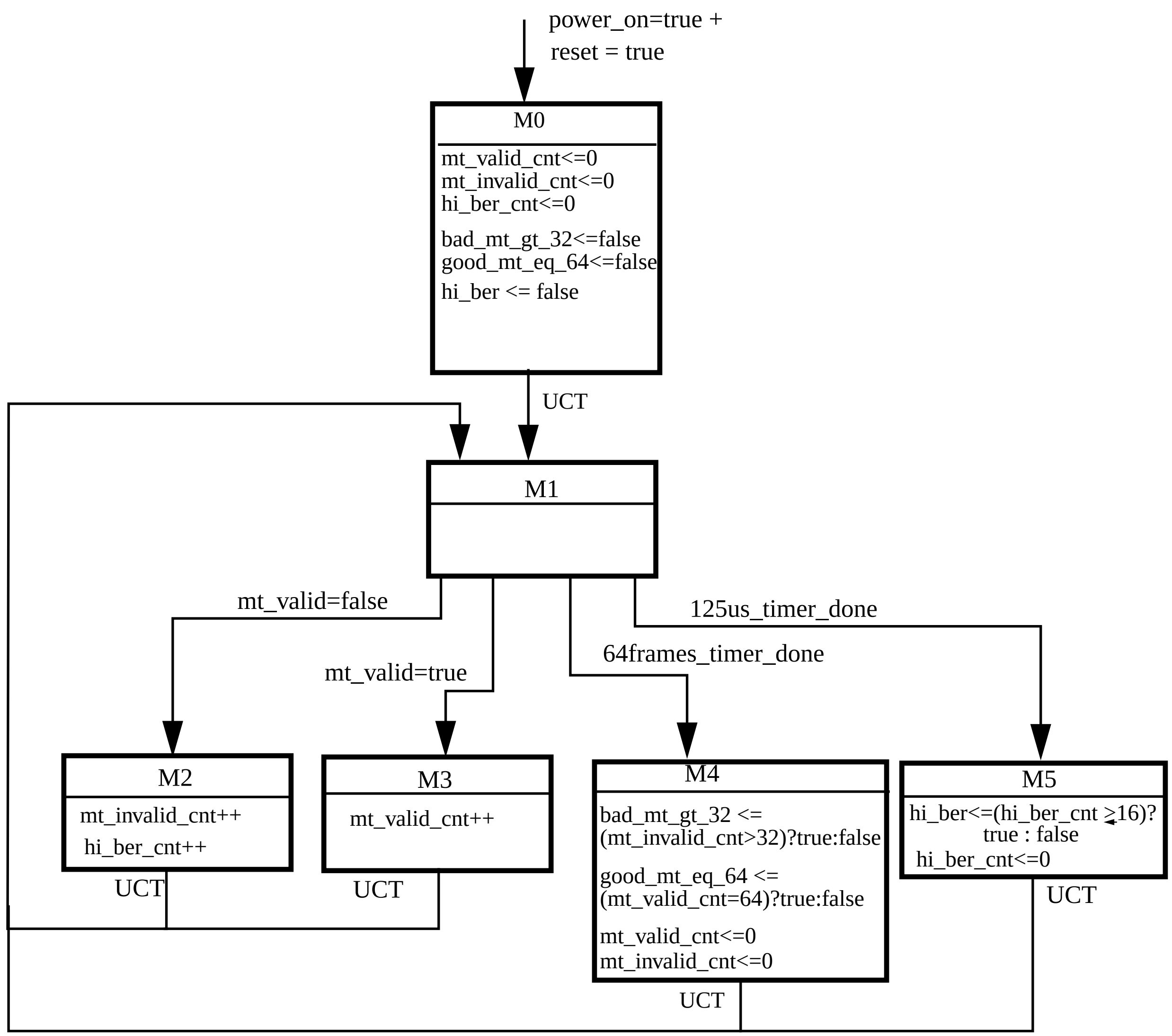


Frame lock process



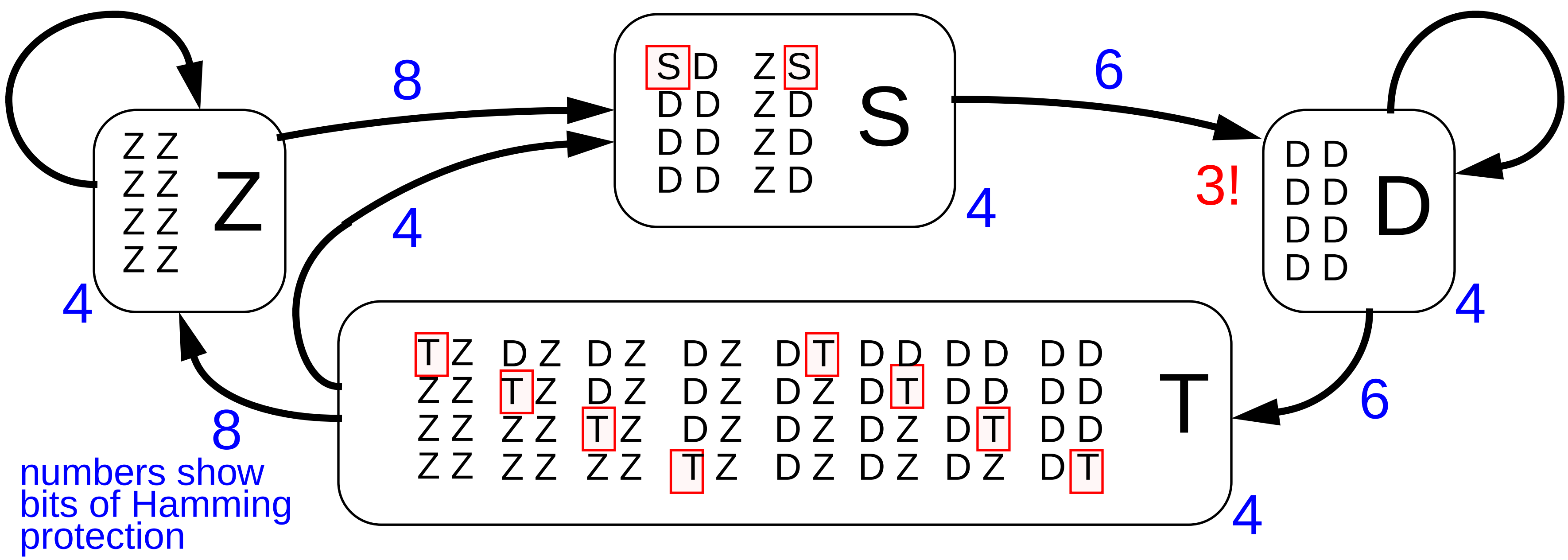
Receiver Synchronization condition
sync_done <= frame_lock=true * hi_ber=false

BER monitor process

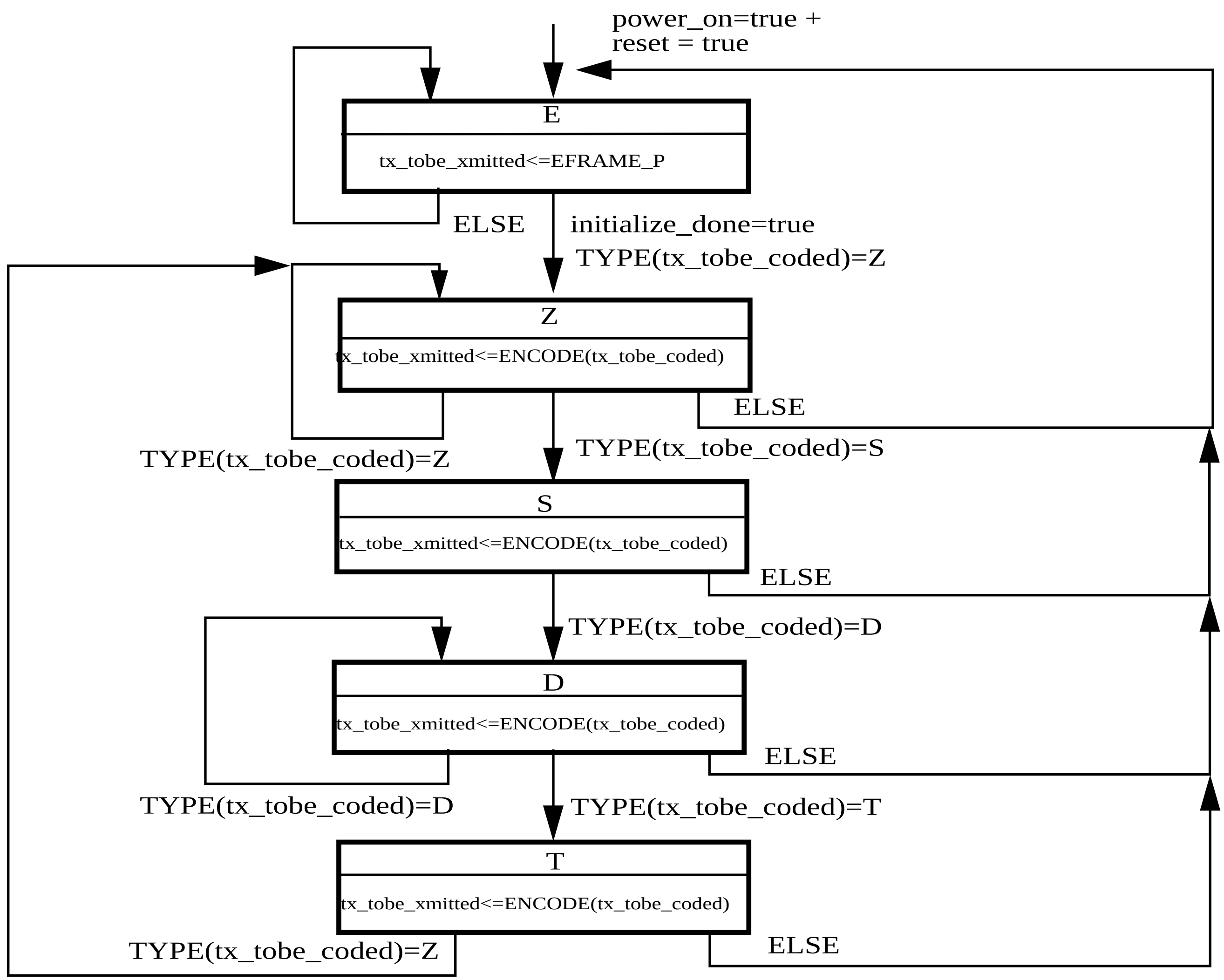


Packet boundary protection

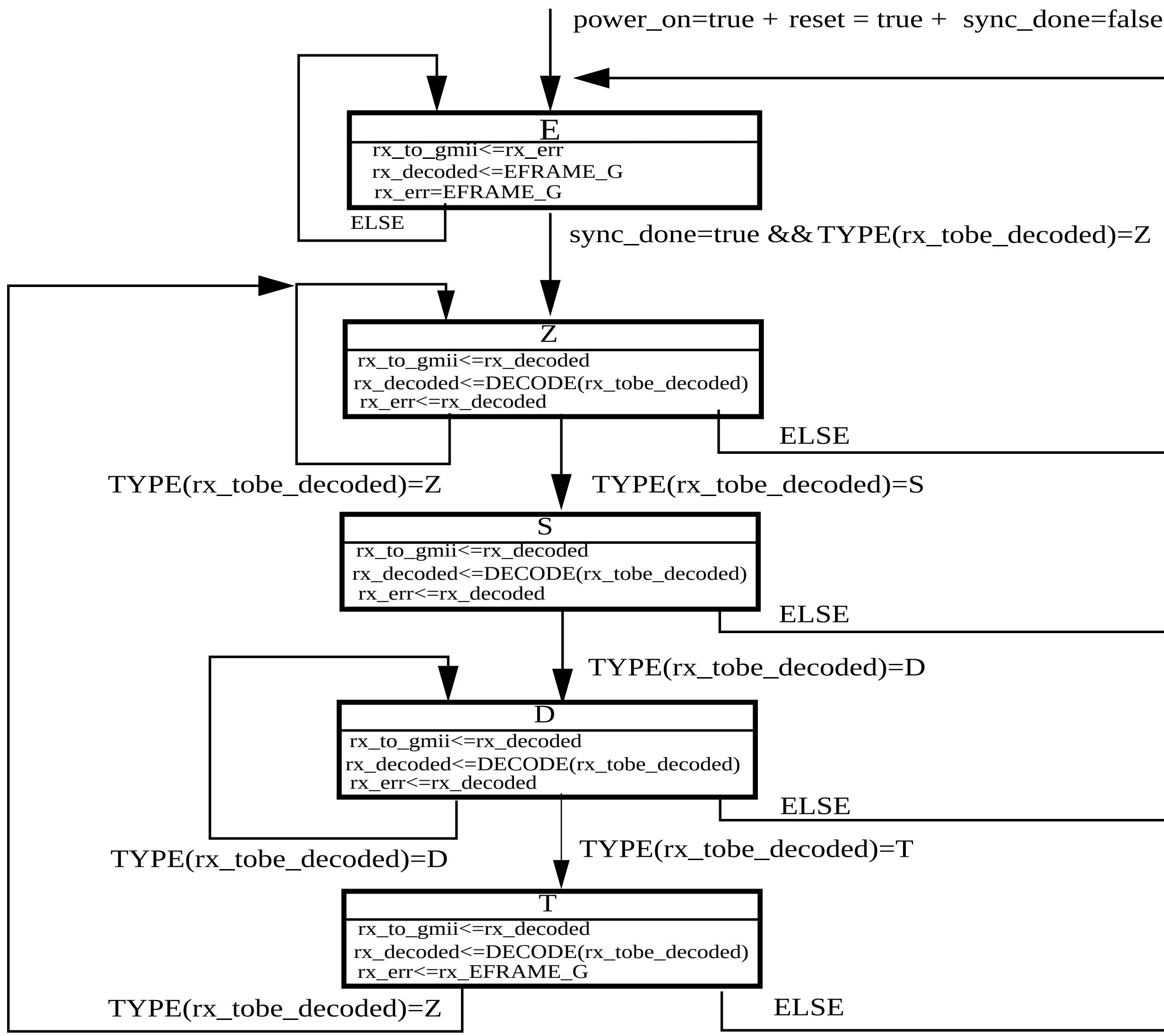
- A 2 bit error in the sync preamble can convert a packet boundary (S,T) into a Data frame (D) and vice-versa. However, all such errors violate frame sequencing rules unless another 4 errors recreate a false S,T packet (a total of six errors). Frame sequence errors invalidate the packet by forcing an (E) on the coder output.



TX process


power_on=true +
reset = true
E
tx_tobe_xmitted<=EFRAME_P
ELSE
initialize_done=true
TYPE(tx_tobe_coded)=Z
Z
tx_tobe_xmitted<=ENCODE(tx_tobe_coded)
ELSE
TYPE(tx_tobe_coded)=S
TYPE(tx_tobe_coded)=Z
S
tx_tobe_xmitted<=ENCODE(tx_tobe_coded)
ELSE
TYPE(tx_tobe_coded)=D
D
tx_tobe_xmitted<=ENCODE(tx_tobe_coded)
ELSE
TYPE(tx_tobe_coded)=D
TYPE(tx_tobe_coded)=T
T
tx_tobe_xmitted<=ENCODE(tx_tobe_coded)
TYPE(tx_tobe_coded)=Z
ELSE

RX process


power_on=true + reset = true + sync_done=false
E
rx_to_gmii<=rx_err
rx_decoded<=EFRAME_G
rx_err=EFRAME_G
ELSE
sync_done=true && TYPE(rx_tobe_decoded)=Z
Z
rx_to_gmii<=rx_decoded
rx_decoded<=DECODE(rx_tobe_decoded)
rx_err<=rx_decoded
ELSE
TYPE(rx_tobe_decoded)=Z
TYPE(rx_tobe_decoded)=S
S
rx_to_gmii<=rx_decoded
rx_decoded<=DECODE(rx_tobe_decoded)
rx_err<=rx_decoded
ELSE
TYPE(rx_tobe_decoded)=D
D
rx_to_gmii<=rx_decoded
rx_decoded<=DECODE(rx_tobe_decoded)
rx_err<=rx_decoded
ELSE
TYPE(rx_tobe_decoded)=D
TYPE(rx_tobe_decoded)=T
T
rx_to_gmii<=rx_decoded
rx_decoded<=DECODE(rx_tobe_decoded)
rx_err<=rx_EFRAME_G
TYPE(rx_tobe_decoded)=Z
ELSE

Optional Code Features

- Special frames are reserved to support ordered sets for both Fiber Channel and 10GbE Link Signalling Sublayer (LSS)
- x,y ordered-set IDs are “1111” for FC and “0000” for 10GbE LSS

XGMII Pattern	Sync		Bit fields 0-63								
ZZZZ/ODDD	1	0	0x2d	Z0	Z1	Z2	Z3	y	D5	D6	D7
ODDD/ZZZZ	1	0	0x4b	D1	D2	D3	x	Z4	Z5	Z6	Z7
ODDD/ODDD	1	0	0x55	D1	D2	D3	x	y	D5	D6	D7
ODDD/SDDD	1	0	0x66	D1	D2	D3	x	y	D5	D6	D7
SDDD/DDDD	1	0	0x78	D1	D2	D3	D4		D5	D6	D7
undefined	1	0	0x00	reserved for future expansion							

rs value	name	shorthand	7-bit line code
0x5c,1	FC ordered-set	[Of]	encoded by TYPE byte
0x9c,1	10 GbE Link Signalling	[LS]	encoded by TYPE byte

Summary

- We've shown a simple and reliable algorithm for 64b/66b frame sync detection
- Bit ordering has been clarified to be compatible with Ethernet CRC definition
- The TX and RX error control state machines have been presented
- A simple test vector has been produced to help to verify new implementations
- Optional 64b/66b extensions exist to support FC ordered sets and LS signalling

Supplementary slides

State machine notation conventions

Variables

TXD<35:0> ....TXD signal of GMII

RXD<35:0> ....RXD signal of GMII

tx_tobe_coded<71:0> .... 72 bit vector which is to be encoded by the PCS before transmission to the PMA.It is formed by concatenation of two consecutive TXD vectors. With the most recently received TXD word in the 35 : 0 bit locations.

tx_tobe_xmitted<65:0> ....A 66 bit vector which is the result of a PCS ENCODE operation and is to be transmitted to the PMA.

rx_tobe_decoded<65:0> ....A 66 bit vector containing the most recently received code word from the PMA.

rx_decoded<71:0>.........................72 bit vector which is the result of the PCS DECODE operation on the received bit vector, rx_tobe_decoded

rx_to_gmii<71:0>..........................72 bit vector which is a pipelined delayed copy of rx_decoded. This is sent to GMII in two steps of 36 bits each. Bits 71:36 are sent first to RXD, followed by bits 35:0.

rx_err<71:0>.................................This holds either a pipeline delayed copy of rx_decoded or the error frame EFRAME_G

state ....Holds the current state of the transmit or the receive process.

sync_done ....Boolean variable is set true when receiver is synchronized and set to false when receiver looses frame lock.

frame_lock ....boolean variable is set true when receiver acquires frame delineation

mt_valid ....boolean variable is set true if received frame rx_tobe_decoded has valid frame prefix bits. I.e, mt_valid = rx_tobe_decoded[65] ^ rx_tobe_decoded[64]

mt_valid_cnt ....Holds the number of frames within a window of 64 frames, with valid prefix bits

mt_invalid_cnt ....Holds the number of frames within a window of 64 frames with invalid prefix bits

good_mt_eq_64 ....Boolean variable is set true when there are 64 contiguous valid prefix bits

bad_mt_gt_32 ....Boolean variable is set true when there are at least 32 invalid prefix bits within a block of 64

hi_ber_cnt ....Holds the number of with invalid prefix bits, within a 125us period

hi_ber ....Boolean is asserted true when the hi_ber_cnt exceeds 16 indicating a bit error rate $>=10^{-4}$

slip_done.......................................Boolean variable is set true when the hi_ber_cnt exceeds 16 indicating a bit error rate $>=10^{-4}$

State machine notation conventions

Constants

const enum FRAME_TYPE = { Z, S, T, D} .... Each 72 bit vector, tx_tobe_coded and the 66 bit vector, rx_tobe_decoded, can be classified to belong to one of the four types depending on its contents. The frame types Z,S, T, D are defined in TBD.

EFRAME_G<71:0> ....72 bit vector to be sent to the GMII interface and represents a error octet in all the eight octet locations

EFRAME_P<65:0> ....66 bit vector to be sent to the PMA and represents a error octet in all the eight octet locations,

Functions

ENCODE(tx_tobe_coded<71:0>) .... Encodes the 72 bit vector into a 66 bit vector to be transmitted to the PMA

DECODE(rx_tobe_decoded<65:0>) .... Decodes the 66 bit vector into a 72 bit vector to be sent to the GMII

TYPE(tx_tobe_coded<71:0>)

TYPE(rx_tobe_decoded<65:0>) .... Decodes the FRAME_TYPE of the tx_tobe_coded<71:0> bit vector or the rx_tobe_decoded<65:0>

Timers

64frames_timer_done ....Timer which is triggered once every 64 of the 66-bit frames in the receive process

125us_timer_done ....Timer which is triggered once every 125us (is approximately 2^{14} 66-bit frames in the receive process).

WAN Interface Sublayer (WIS) Update

IEEE P802.3ae
La Jolla
July 2000

Norival Figueira, Paul Bottorff, David Martin, Tim Armstrong, Bijan Raahemi:…………	Nortel Networks
Richard Dugan:…………	Agilent
Tom Palkert:…………	AMCC
Juan Pineda, Bill Rivard:…………	Bravida Corporation
Howard Frazier:…………	Cisco Systems
Steve Haddock:…………	Extreme Networks
Nan Chen:…………	Force10 Networks
Michael McDonald:…………	Galileo Technology
Kevin On:…………	Infineon Technologies
Pankaj Kumar, Bradley Booth, Bob Grow:…………	Intel
Bjørn Liencres:…………	Juniper Networks
Nader Vijeh:…………	Lantern Communications
Enrique Hernandez (Bell Labs), Nevin Jones (Microelectronics):……….	Lucent
Iain Verigin, Stuart Robinson, Tom Alexander, Farzin Firoozmand:……	PMC Sierra
Lee Yong-Hee, Won Jonghwa:…………	Samsung Electronics
Shimon Muller:…………	Sun Microsystems
Frederick Weniger:…………	Vitesse

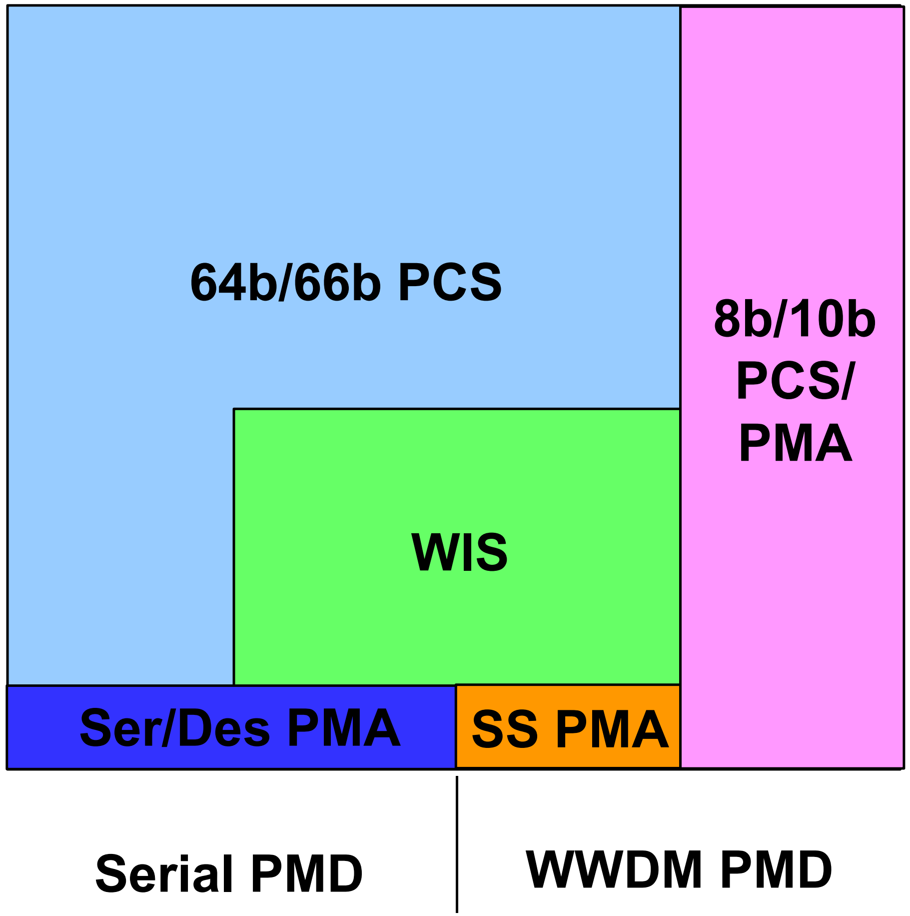
V1.1

Agenda

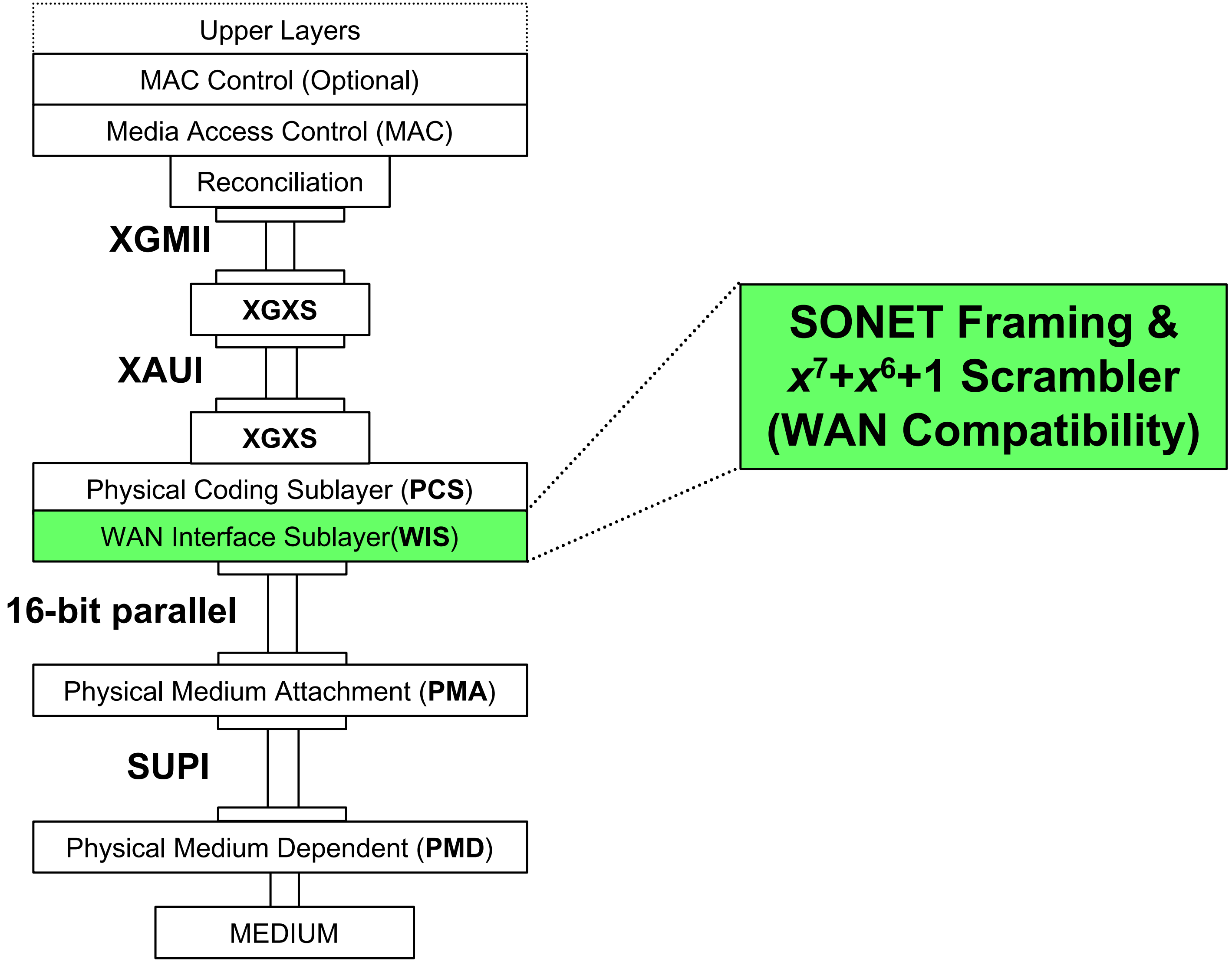
- **WIS**
  - $x^7 + x^6 + 1$ scrambler
  - SONET framing
  - overheads
  - frame synchronization

- **How to write the WIS Clause by cross-referencing ANSI T1.416-1999**
  - Defining required changes and additions
  - Keeping SDH compatibility

UniPHY Components

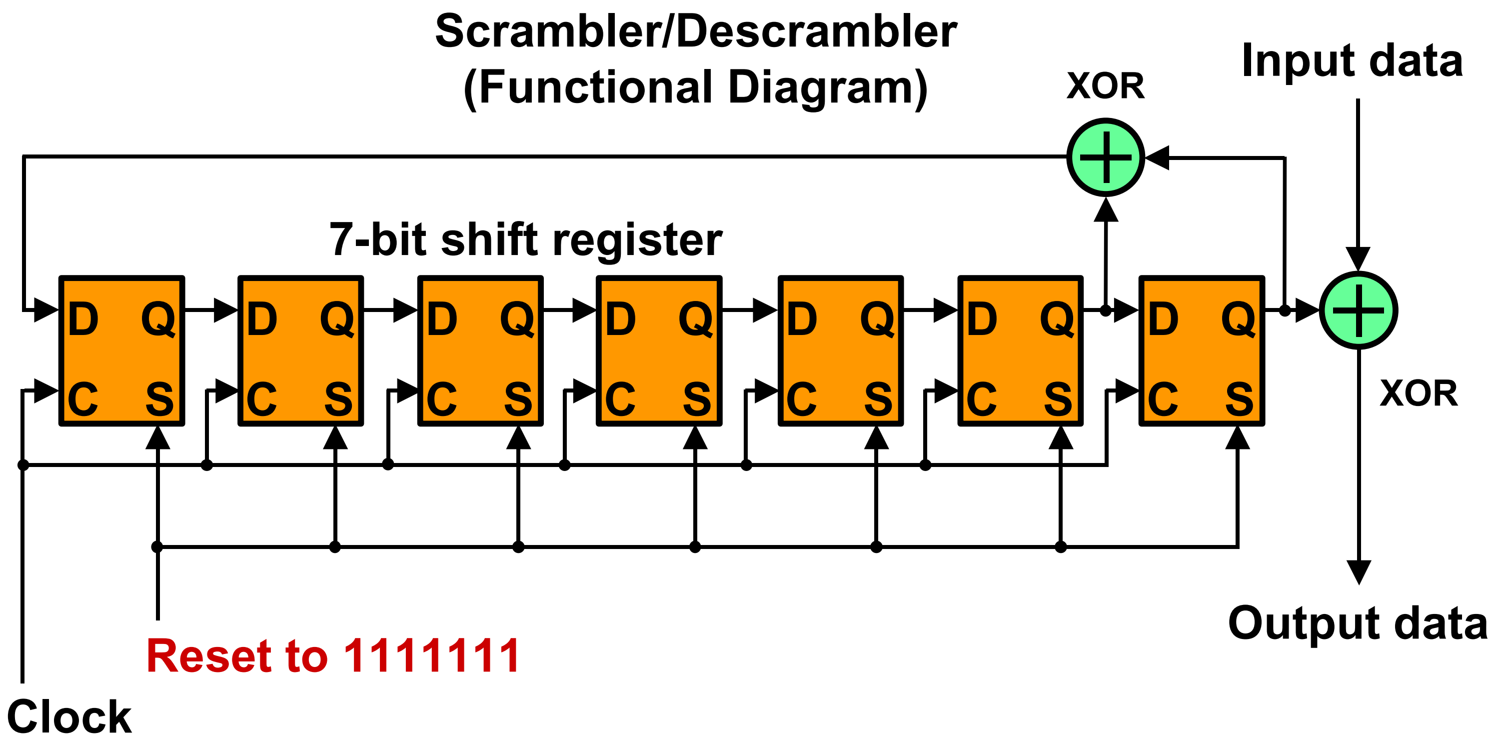


WAN-PHY and UniPHY Layer Model

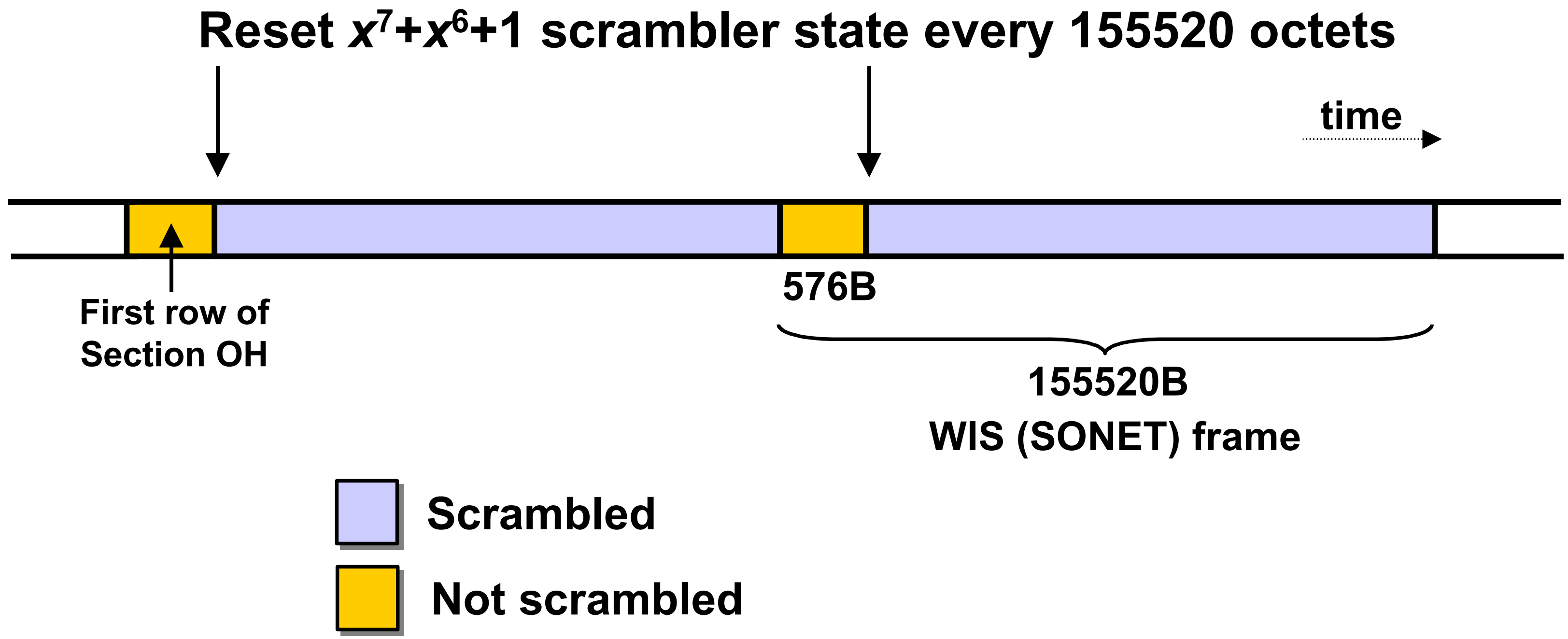


WIS x^7+x^6+1 Scrambler

- **Provides high randomization**
  - Assures adequate number of transitions for line rate clock recovery at the receiver

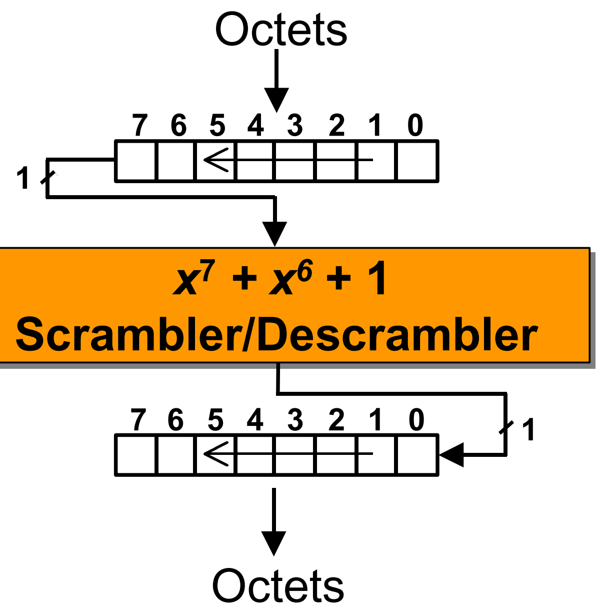


State is Periodically Resynchronized



Bit Order of Scrambling/Descrambling

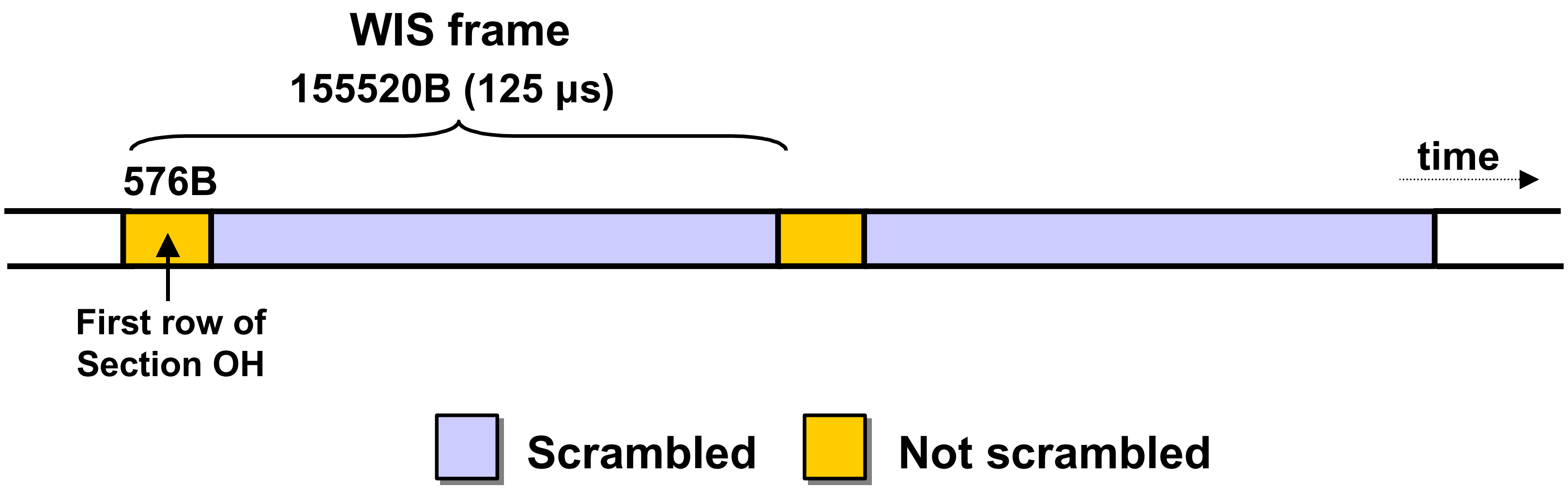
- **Most significant bit (MSB) first**



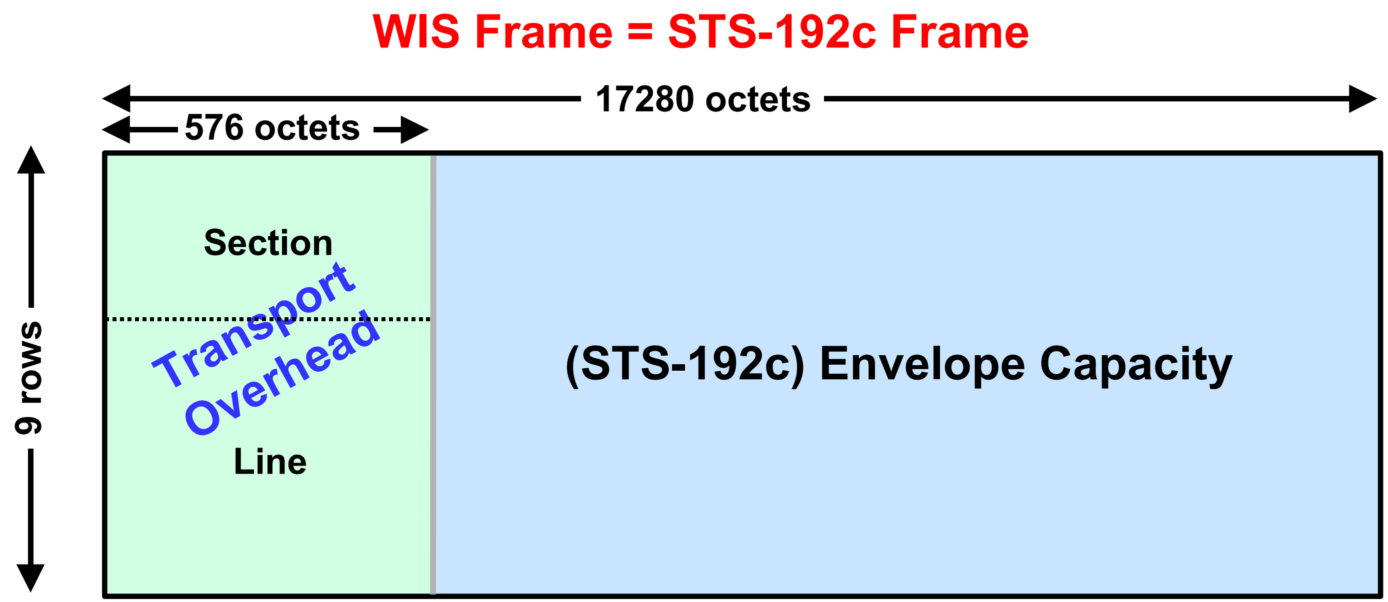
(Functional diagram)

WIS SONET Framing

- **SONET frame with minimum overhead support**
  - Overheads are out of band management used to control SONET networks
  - While the WIS frame is compatible with SONET, it does not provide full SONET management
- **Sequence of 155520 octets (125 µs)**



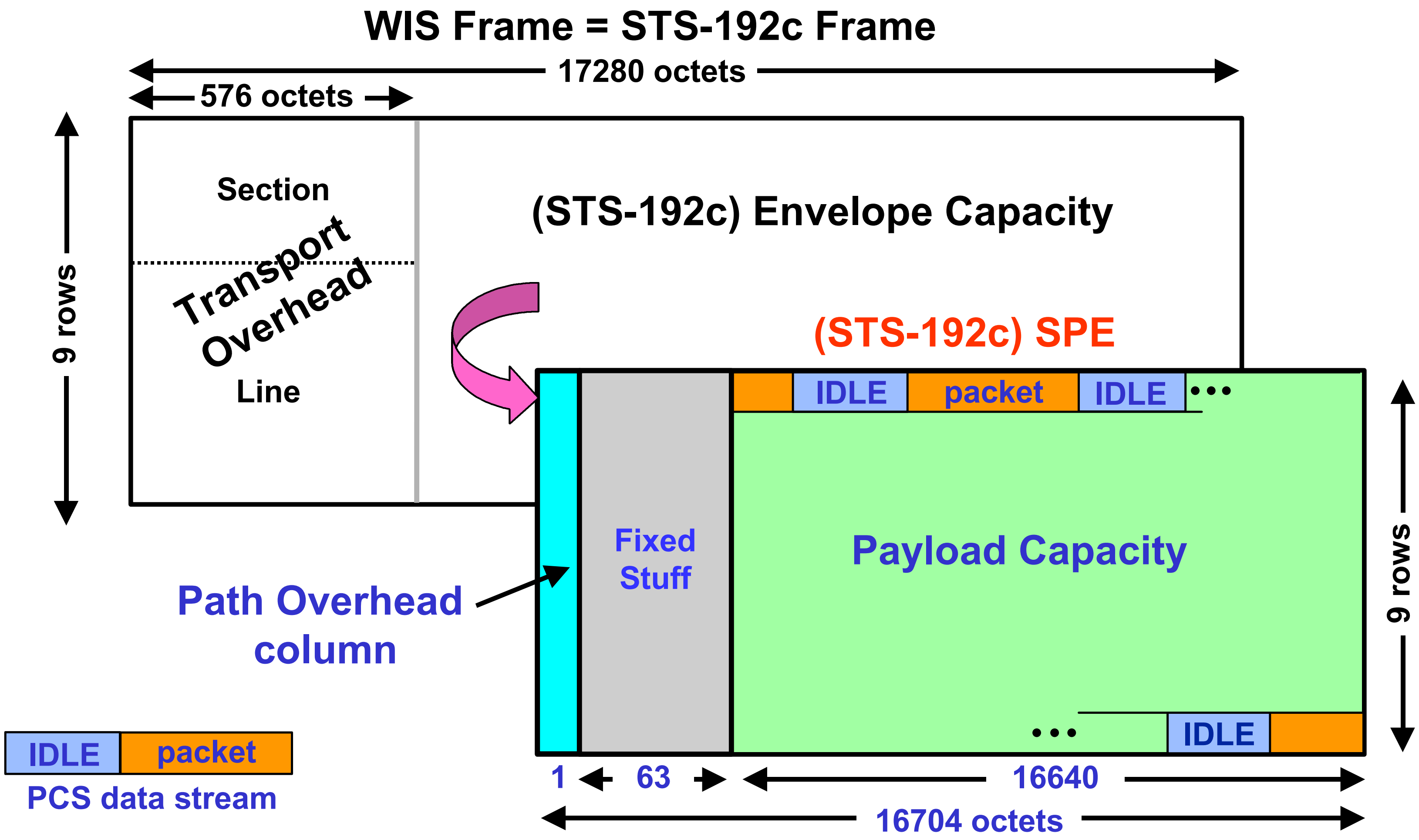
WIS Frame: Viewed as 9×17280 Octets



STS-192c = Synchronous Transport Signal – level 192, c = concatenated.

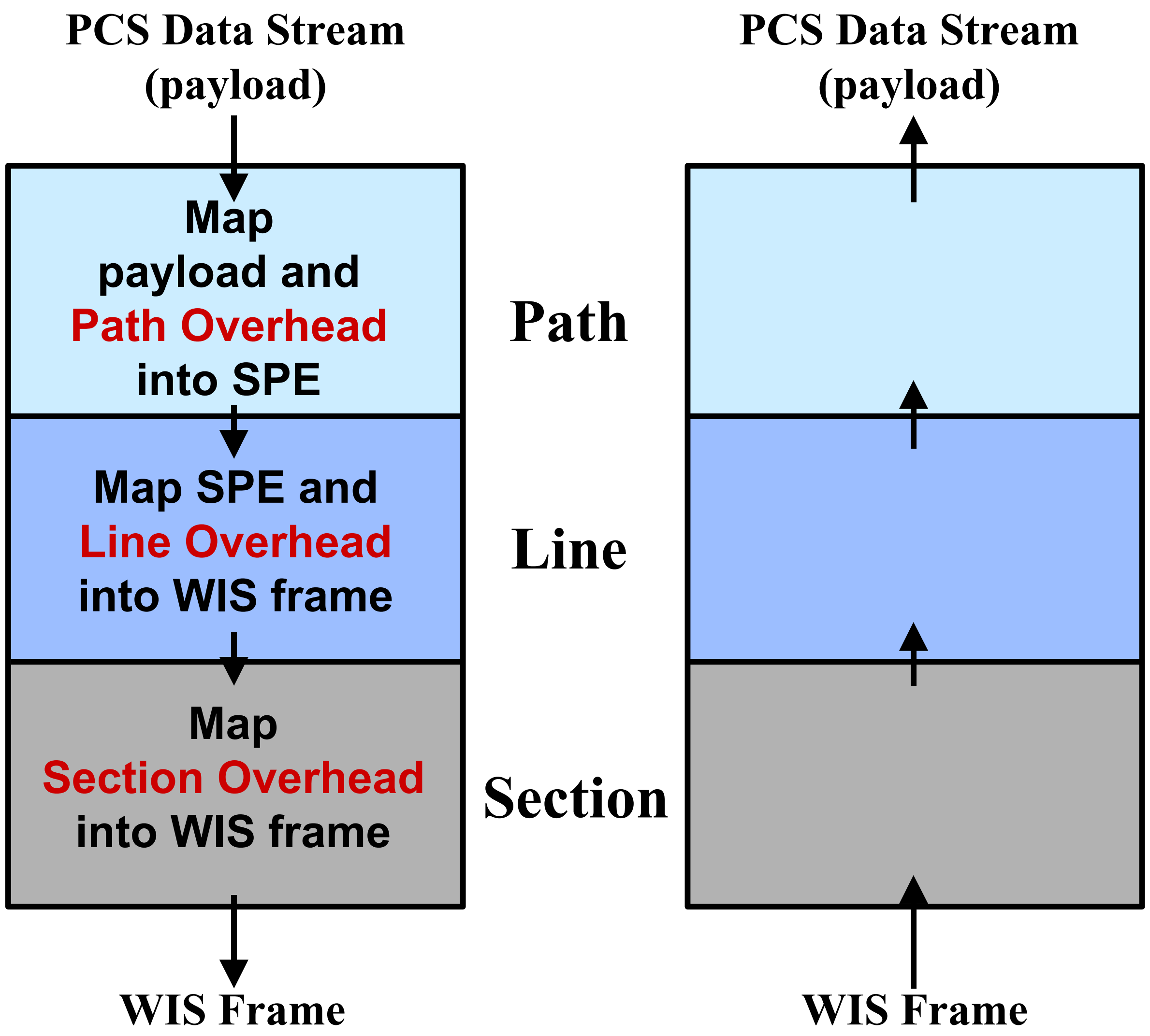
Transmission order: top to bottom, row-by-row, left to right.

Payload Capacity (9.58464 Gb/s)

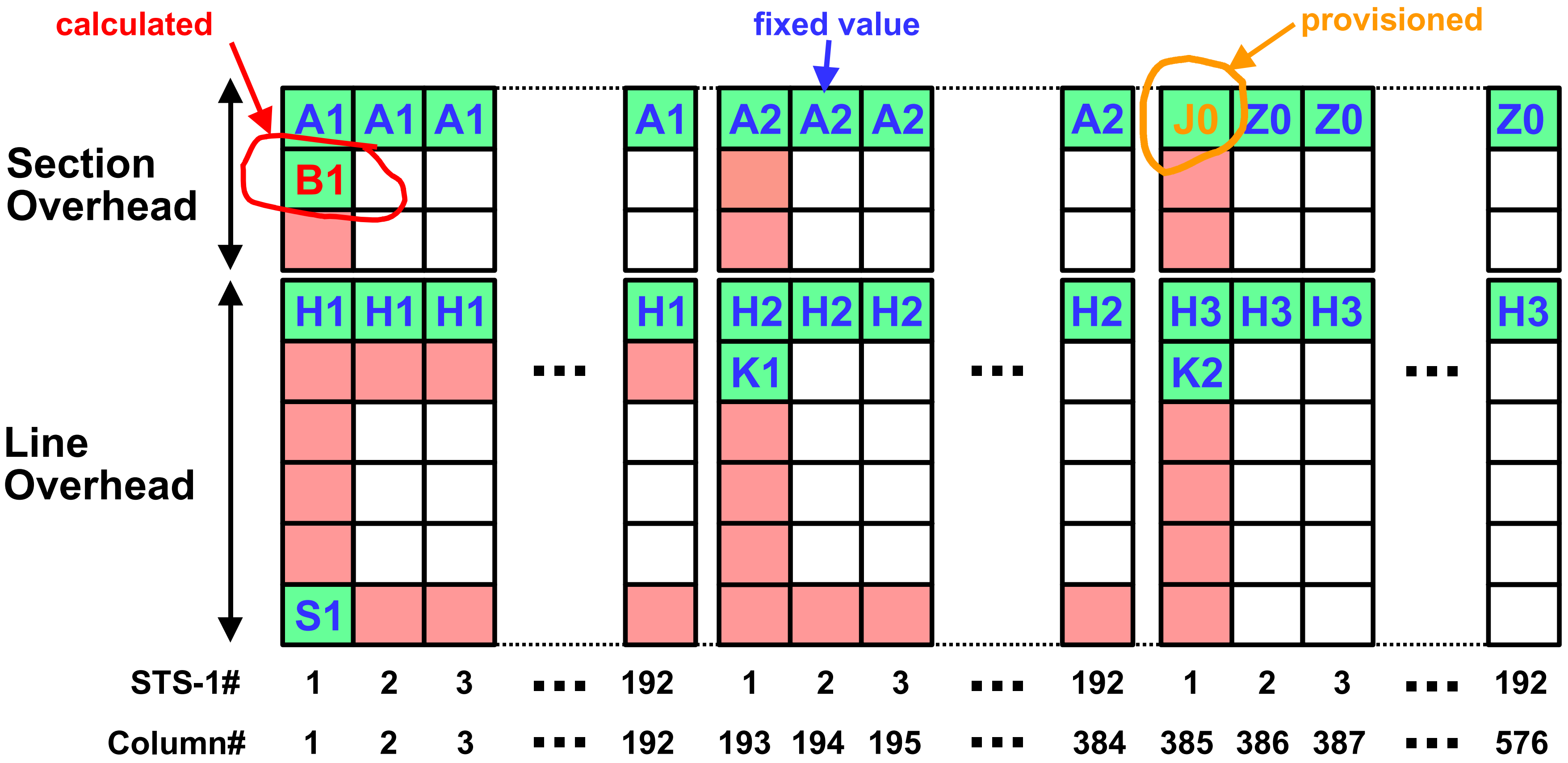


STS-192c = Synchronous Transport Signal – level 192, c = concatenated
SPE = Synchronous Payload Envelope

WIS Overhead Layers



Transport Overhead



= Undefined overhead octets (set to zero)

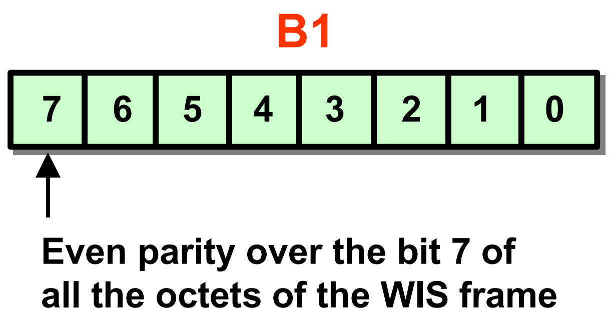
= Defined overhead octets (B2, E1-2, F1, D1-12, M1, Z1-2), unused by 10GE WAN PHY (set to zero)

Section Overheads

- **A1 and A2 (“Framing octets”)**
  - Fixed value: A1 = 11110110, A2 = 00101000
  - A1/A2 transition is used for WIS frame synchronization
- **J0 (“Section Trace”)**
  - Allows a receiver to verify its continued connection to the intended transmitter
  - Provisioned Value
    - when no value is provisioned, J0 shall be set to 00000001)
- **Z0 (‘Section Growth”)**
  - Fixed value: 11001100

Section Overheads (cont.)

- **B1 ("Section BIP-8")**
  - Used as a Section error monitoring function
  - Calculated value:
    - BIP-8 code (using even parity) over all the bits of the last transmitted WIS frame <u>after</u> scrambling

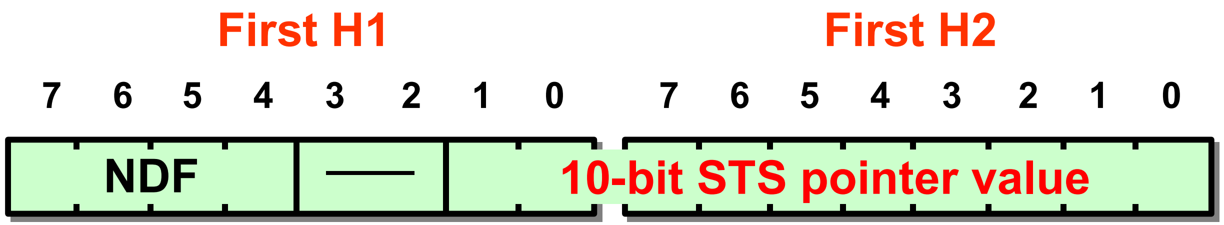


BIP-8 (Bit-Interleaved Parity-8) with even parity: The i^{th} bit of the code provides even parity over the i^{th} bit of all the covered octets.

BIP-8of the bit sequence 11110000 00001111 is 11111111.

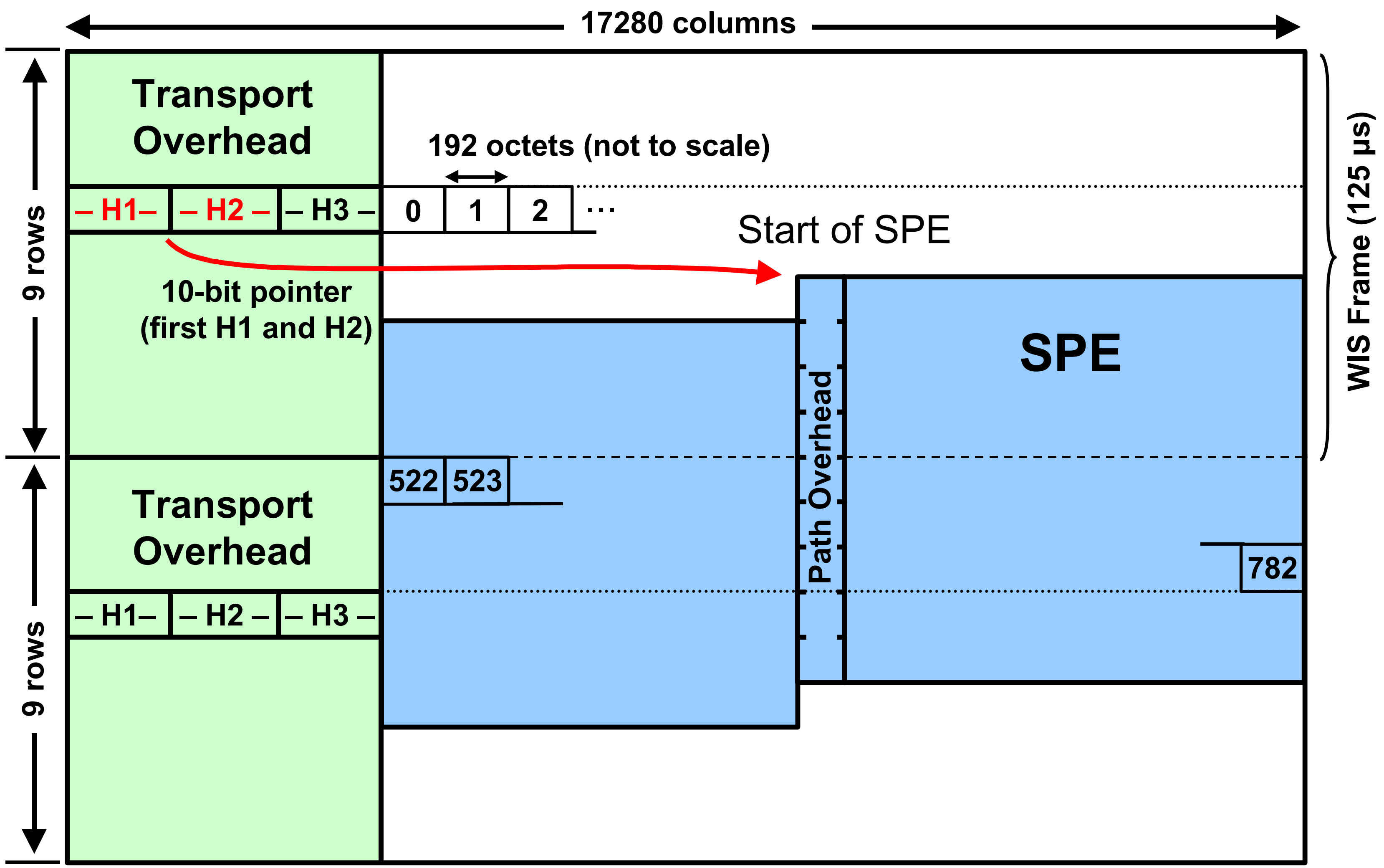
Line Overheads

- **First H1 and H2 ("Payload Pointer")**
  - 16-bit word containing 10-bit pointer in the range of 0 to 782
  - Transmits fixed values: H1 = 01100010 and H2 = 00001010 (i.e., pointer = 522)
  - Receiver 10GE WAN PHY shall be able to process arbitrary pointer values (which may be changed by a transport network)
- **Second to last H1 and H2**
  - Fixed Values: H1 = 10010011 and H2 = 11111111



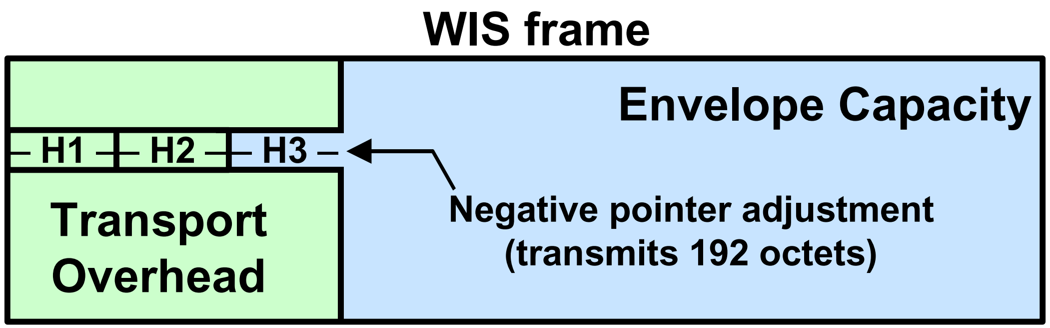
NDF (new data flag) field

H1/H2 Pointer and SPE Position



Line Overheads (cont.)

- **H3 ("Pointer Action Bytes")**
  - Allows an LTE to have slightly different clocks at the receiver and transmitter paths
  - Carries 192 extra SPE (payload) octets in the event of a "negative pointer adjustment," which may be required when the receiver clock is faster than the transmitter clock
  - Set to zero when not used

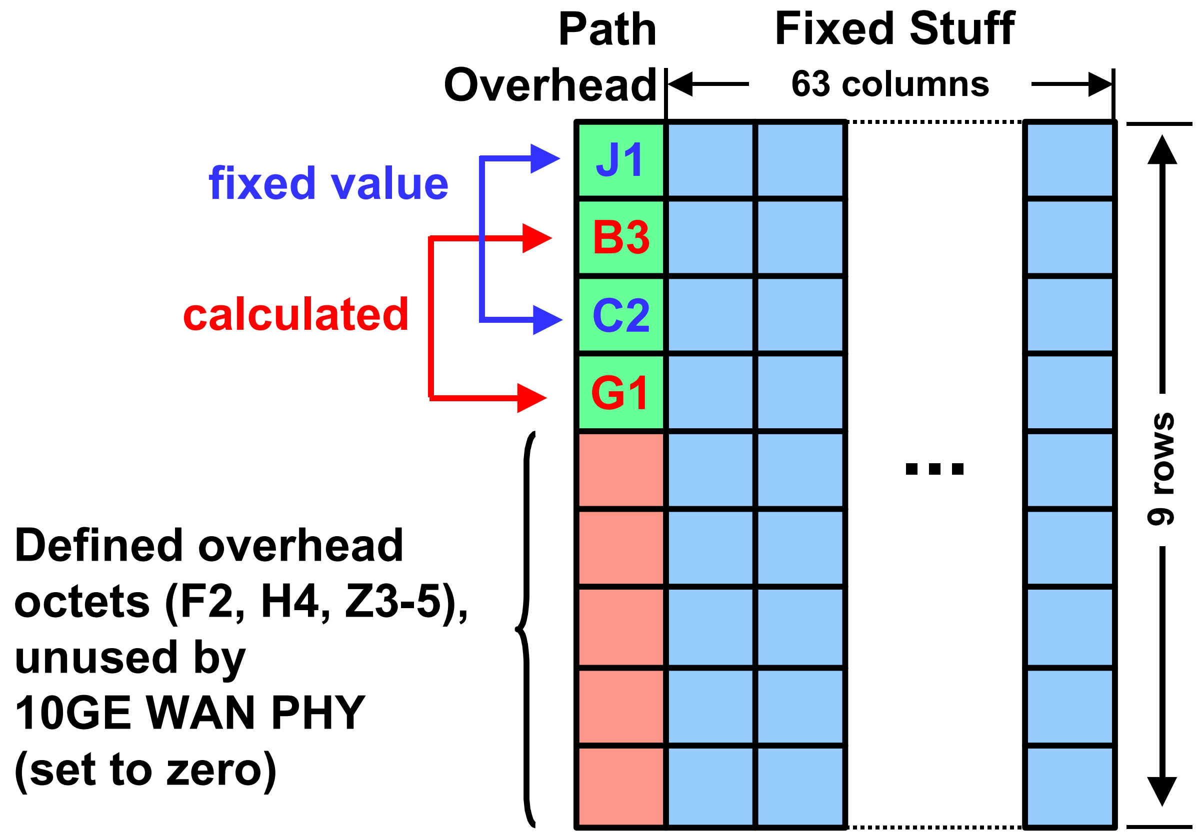


Line Overheads (cont.)

- **K1 and K2**
  - Fixed values: K1 = 00000001, K2 = 00010000
  - K1 and K2 are used on the protection line for automatic protection switching signaling. Above settings indicate a working channel rather than the protection channel.

- **S1**
  - Fixed value: 00001111
  - Indicates quality clock information to receiver. Above setting indicates “don’t use for synchronization”

Path Overhead and "Fixed Stuff"



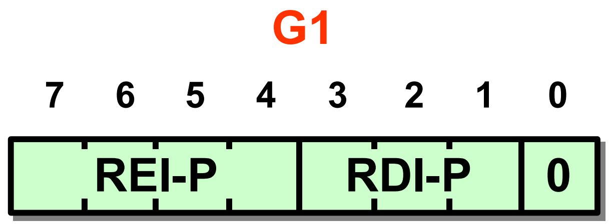
"Fixed Stuff" columns provide compatibility with SONET/SDH byte-interleaving and concatenation rules (set to zero)

Path Overheads

- **J1 (“Path Trace”)**
  - Fixed value: 00000000
- **B3 (“Path BIP-8”)**
  - Used as a Path error monitoring function
  - Calculated value: BIP-8 code (using even parity) over all the octets of the last transmitted SPE before (x^7+x^6+1) scrambling
- **C2 (“Path Signal Label”)**
  - Identifies the contents of the STS SPE (i.e., 10GE WAN PHY)
  - Fixed value: 00011010 (provisional value assigned to 10 GE)

Path Overheads (cont.)

- **G1 ("Path Status")**
  - Conveys the Path terminating status and performance back to the transmitter (i.e., a PTE)
  - Calculated value:
    - REI-P field = number of bit errors detected with the B3 octet of the last received SPE
    - RDI-P field = Detected defects on the received signal

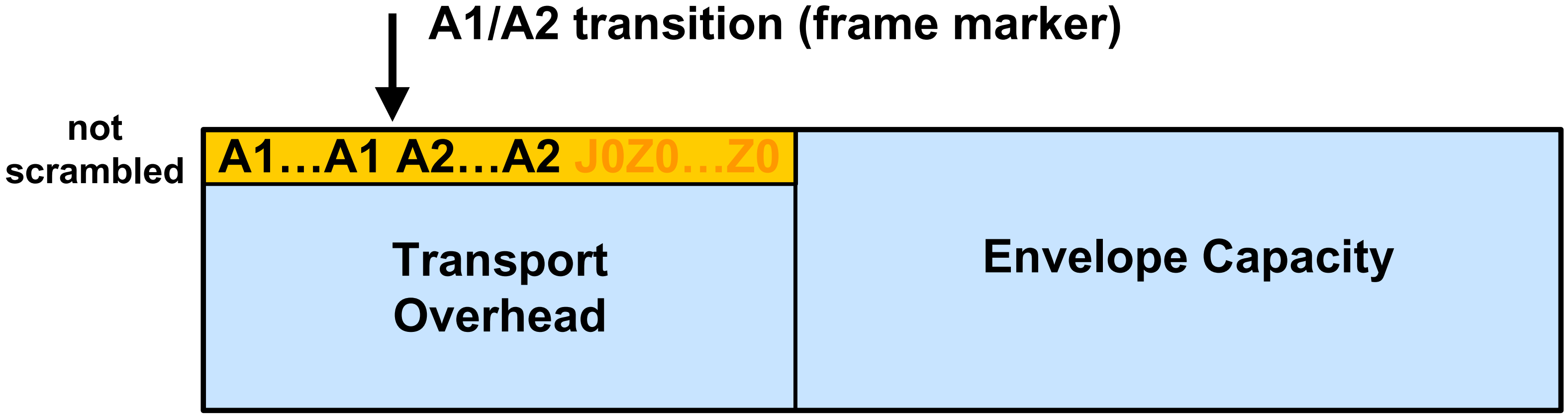


REI-P = Path Remote Error Indication
RDI-P = Path Remote Defect Indication

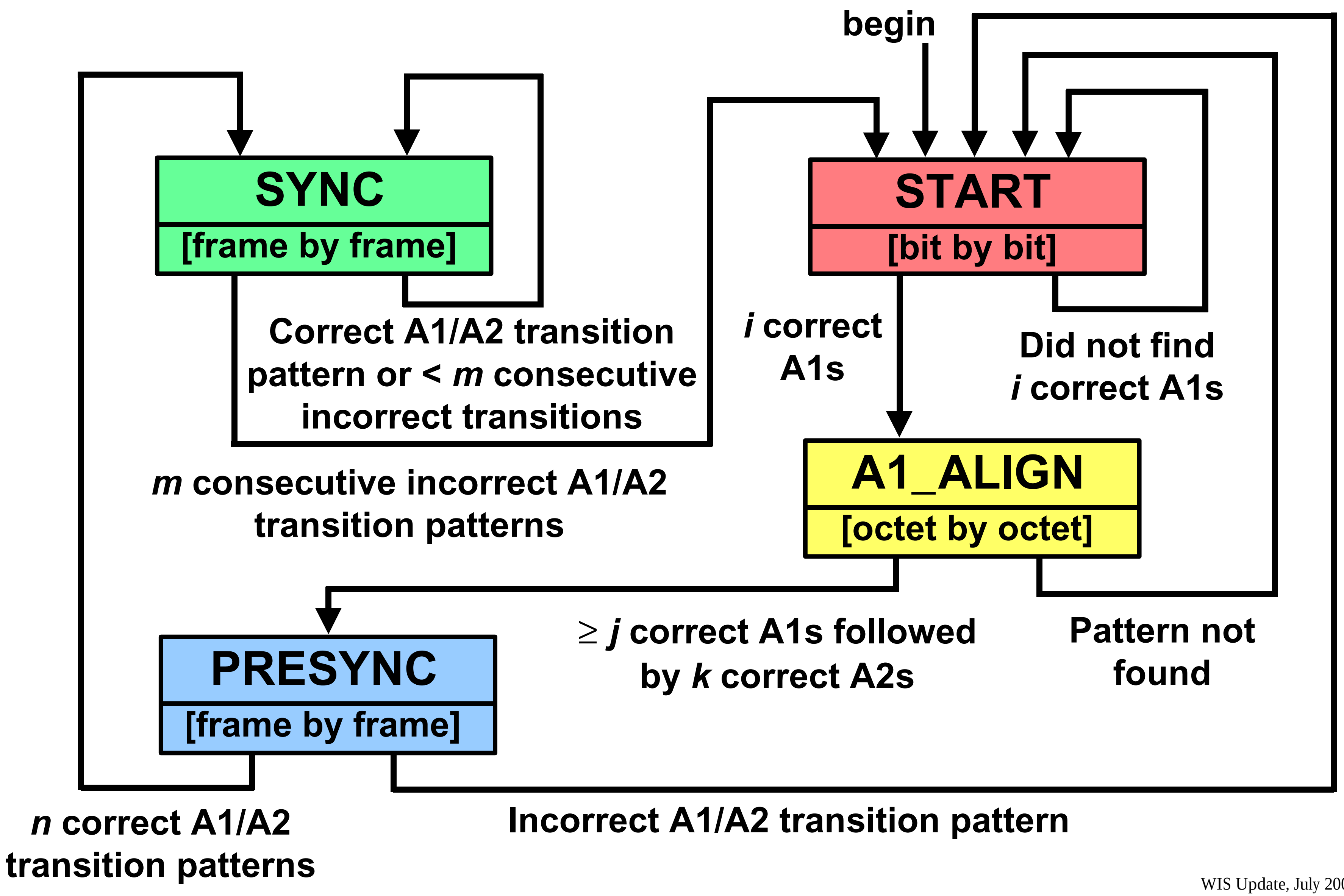
REI-P field
0000 to 1000 = 0 to 8 errors
when received, 1xx1 = 0 errors

WIS Frame Synchronization

- **Uses A1/A2 transition (i.e., frame marker) for frame and octet delineation**
- **Looks for the A1/A2 framing pattern consistently**
  - Expects it to appear once every 155520 octets (length of the frame)
  - When the framing pattern appears in the right place enough times, correct frame synchronization is assumed



Frame Sync Example: State Diagram

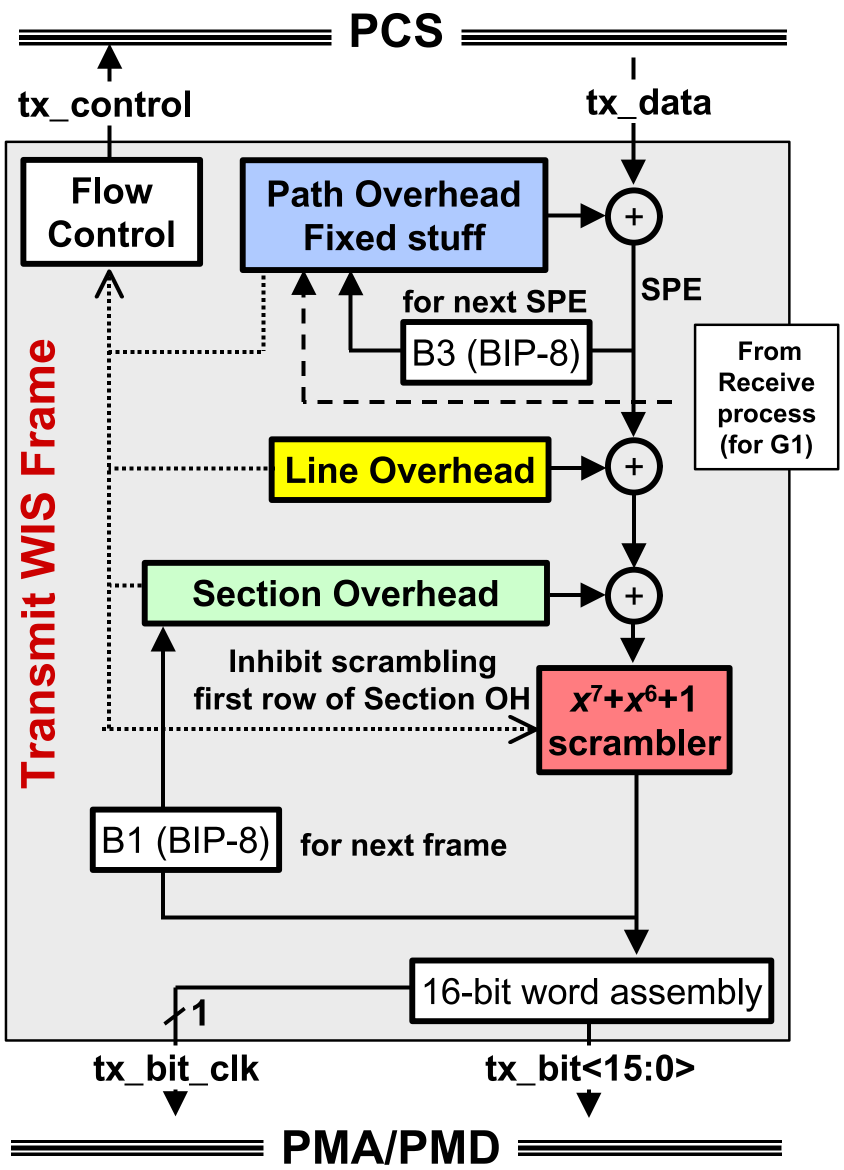


WIS Frame Sync. Performance

- **Example for *m* = 4, A1/A2 transition pattern = 2 A1/A2s**
  - Probability of frame loss ≈ $1.049 \times 10^6 \times BER^4$
    = $1.049 \times 10^{\mathbf{-42}}$ (@ BER = 10^{-12})
  - Average interval to frame loss
    - ≈ 3.7×10^{30} years (@ BER = 10^{-12})
      (> estimated age of observable universe, i.e., ~ 10^{10} years)

- **More robust implementations are possible, e.g., see**
  - **"10GE WAN PHY Delineation Performance"**
  - **http://grouper.ieee.org/groups/802/3/10G_study/public/email_attach/delineation_perf.doc**

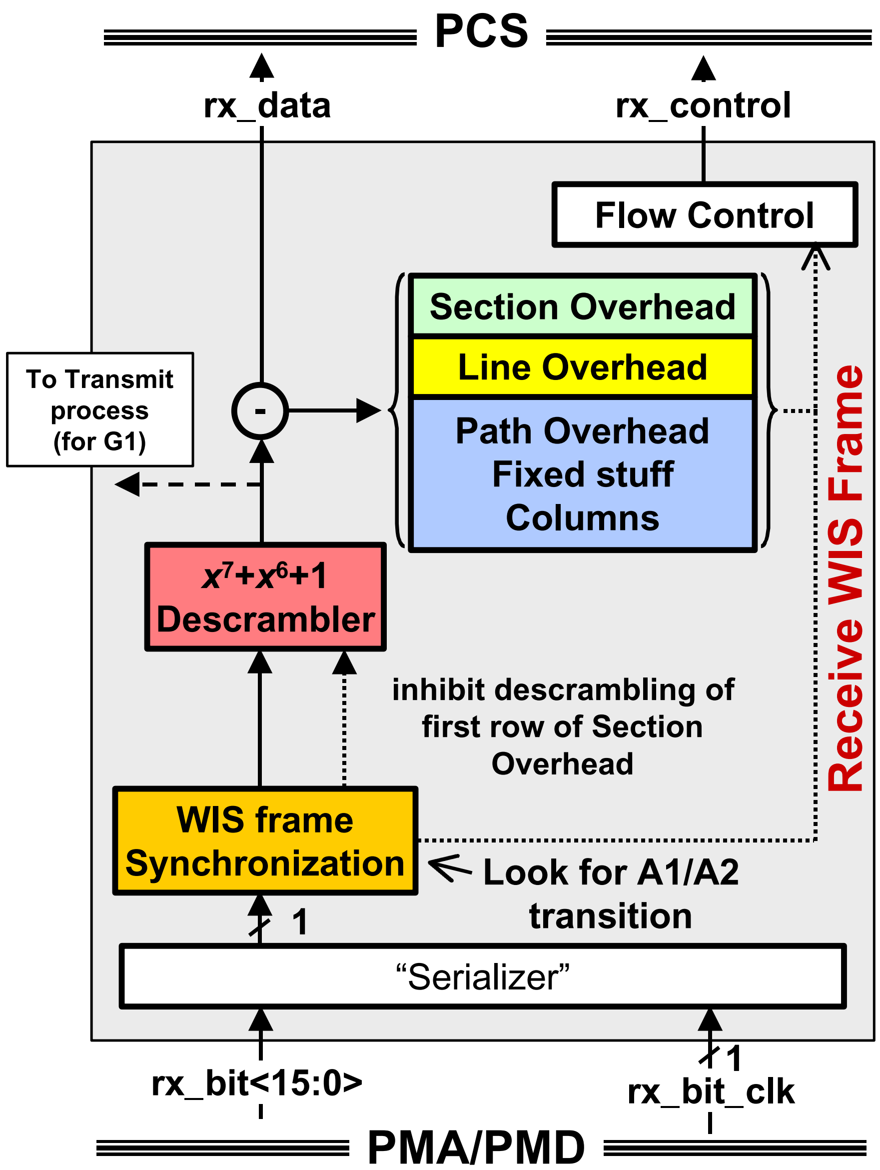
Reference Diagram: Transmit WIS Frame



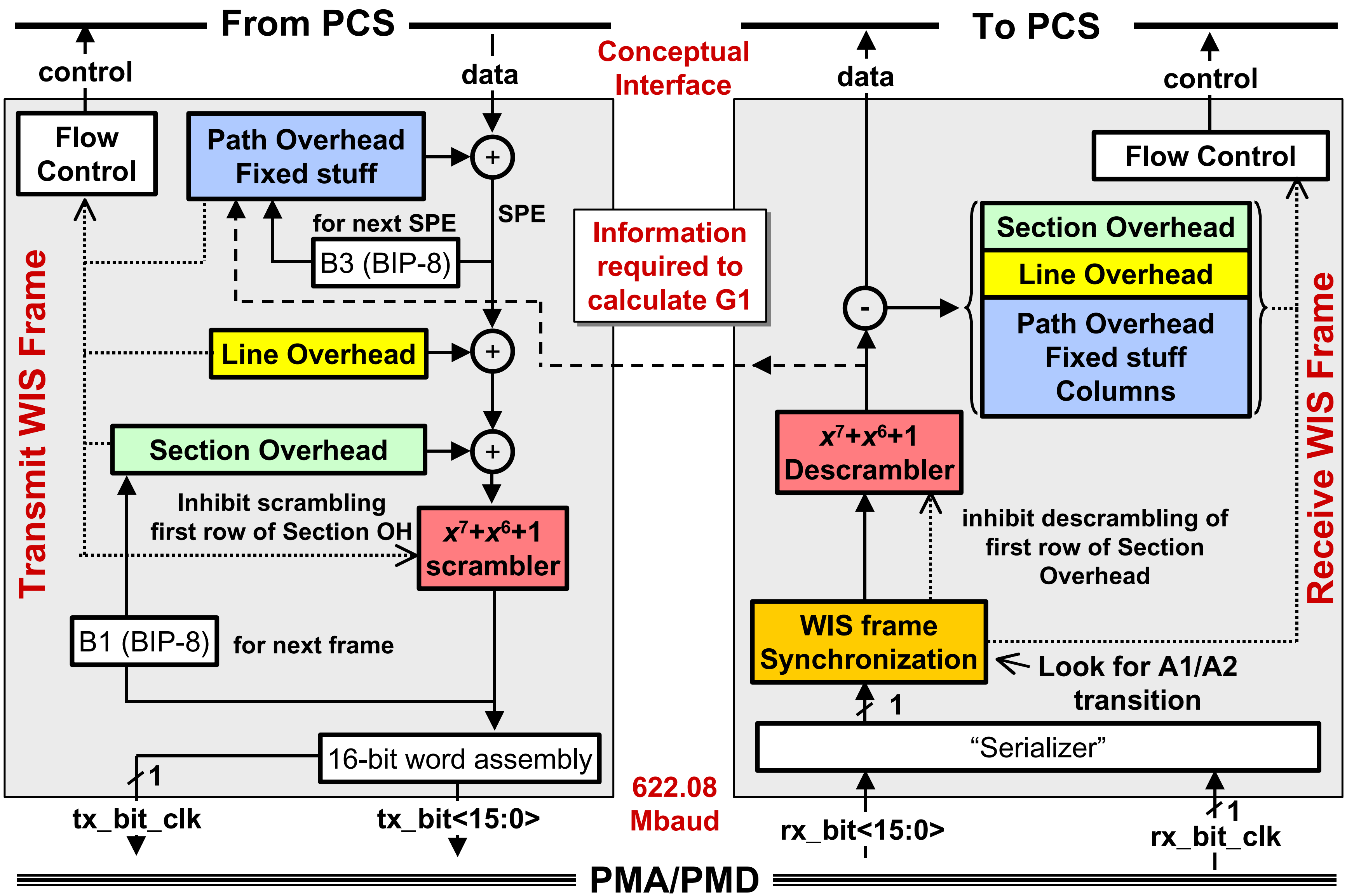
- **Functional View**
- **WIS frame formation (stages)**
  - (1) Path Overhead and fixed stuff columns
  - (2) Line Overhead
  - (3) Section Overhead
  - (4) Scramble with x^7+x^6+1 (first row of Section Overhead, i.e., A1/A2, J0, and Z0, is <u>not scrambled)</u>
  - (5) 16-bit words are transmitted to PMA/PMD (for 16-bit Parallel I/F)

Reference Diagram: Receive WIS Frame

- **Functional View**
- **WIS frame processing (stages)**
  - (1) “Serialize” received signal (figure shows 16-bit Parallel I/F)
  - (2) WIS frame synchronization and octet delineation
  - (3) Descramble with x^7+x^6+1 (first row of Section Overhead is <u>not descrambled)</u>
  - (4) Extract Section Overhead, Line Overhead, Path Overhead, Fixed Stuff columns
  - (5) Remaining octets = payload



WIS Reference Diagram



Writing the WIS Clause by Cross-Reference

- **How to write the WIS Clause by cross-referencing ANSI T1.416-1999**
  - WIS Clause proposed in "IEEE P802.3ae Document Structure Update" http://grouper.ieee.org/groups/802/3/ae/public/may00/booth_1_0500.pdf
  - ANSI T1.416-1999 can be obtained at the following URL: http://www.atis.org/atis/docstore/index.asp

- **WIS as described here**
  - With optional text to add support to B2/M1 and J1 (provisionable) and ±20 ppm reference clocks (if desired)

ANSI T1.416-1999

- **Title: “Network to Customer Installation Interfaces -- Synchronous Optical NETwork (SONET) Physical Layer Specification: Common Criteria”**
  - Contains definitions and references to other documents providing a complete specification of network and customer installation interfaces compatibility

- **Presentation provides definitions that allow for SDH compatibility**

Cross-References to ANSI T1.416-1999

- **Section 1 “Scope”**
  - Applicable as is
- **Section 2 “Normative References”**
  - Applicable as is
- **Section 3 “Definitions, Abbreviations, and Acronyms”**
  - Applicable as is

Cross-References (cont.)

- **Section 4 “Common Criteria”**
  - Applicable with changes to Table 1 (SONET Overheads at NIs), as indicated below
  - Following “optional” overheads are not supported
    - Section: D1-D3, E1, F1 (all set to 00000000)
    - Line: D4-D12, E2, Z1, Z2 (all set to 00000000)
    - Path: Z3-Z4, J1 (all set to 00000000)
      If J1-provisionable support is added, remove J1 from the above list and define a default value, say 00000000, or a default Path Trace message
  - Add that Z0 (Section Growth) is set to 11001100

Note: H1 “ss” bits do not compromise SDH compatibility, since the ITU now specifies that the receiver ignores them

Cross-References (cont.)

- **Section 4 "Common Criteria" (cont.)**
  - Following "required" overheads are not supported
    - Section: B2 (set to 00000000), M0-M1 (set to 00000000)
      If B2/M1 support is added, remove B2 and M1 from the above list
    - Line: S1 (set to 00001111, i.e., "don't use for synchronization")
  - Following "application specific function" overheads are not supported
    - Line: K1 (set to 00000001), K2 (set to 00010000) -- These settings indicate a working channel rather than the protection channel
    - Path: F2 (set to 00000000), H4 (set to 00000000), N1 (set to 00000000)
  - Add that C2 (STS Path Label) is set to 00011010
    (This is the provisional value assigned to 10GE)
  - VT Path Overheads are not applicable (not supported)

Cross-References (cont.)

- **Section 5 "Jitter"**
  - Not applicable. IEEE P802.3ae defines jitter specification
- **Section 6 "Synchronization"**
  - Not applicable
  - Add (not necessarily to Clause 48) that 10 Gigabit Ethernet signal is defined to be within ±100 ppm of the nominal rate (if required, replace ±100 ppm with ±20 ppm)
- **Section 7 "Maintenance"**
  - Sections that are not applicable
    - Section 7.2.2 "VT1.5 rate - Electrical Interface"
    - If B2/M1 support is added: Section 7.4.2 "VT1.5 rate"
otherwise: Section 7.4 "Line"
    - Section 7.6 "Performance and Failure Alarm Monitoring"
    - Section 7.7 "Performance Monitoring Functions"

Cross-References (cont.)

- **Section 7 “Maintenance” (cont.)**
  - Section 7.1, Table 2 “Near-end events and far-end reports”, <u>only</u> the following is supported
    - Defects: LOS (as defined in Section 7.2.1)
      SEF/LOF (as defined in Section 7.3)
      LOP-P (as defined in Section 7.5)
      AIS-P (as defined in Section 7.5)
      ERDI-P (as defined in Section 7.5)
      In addition, PLM-P (which is not listed in Table 2) is supported (as defined in Section 7.5)
    - Anomalies: BIP-N(S) (as defined in Section 7.3)
      If B2/M1 support is added:
        BIP-N(L) (as defined in Section 7.4.1)
        REI-L (as defined in Section 7.4.1)
      BIP-N(P) (as defined in Section 7.5)
      REI-P (as defined in Section 7.5)

Cross-References (cont.)

- **Section 7 “Maintenance” (cont.)**
  - Sections 7.2.1, 7.3, 7.4.1 (only if B2/M1 support is added), and 7.5 are applicable with the exclusion of defects and anomalies not listed in the previous slide
  - Section 7.2.1
    - Make T = T’ = 125 / 3 μs (i.e., three row periods)
      - Comment: Ambiguity in this value has long been an annoyance in SONET/SDH. Proposed value falls in the middle of the suggested range and gives vendors a single convenient value to implement. Removal of LOS would then take 125 μs.

Cross-References (cont.)

- **Annex A**
  “Normative -- SONET VT1.5 Line Interface Common Criteria”
  — Not applicable
- **Annex B**
  “Informative -- SONET maintenance signals for the NI”
  — Not applicable
- **Annex C**
  “Informative -- Receiver Jitter Tolerance and Transfer”
  — Not applicable
- **Annex D**
  “Informative -- Bibliography”
  — Applicable as is

Summary

- **WIS**
  - $x^7 + x^6 + 1$ scrambler
  - SONET framing, overheads, and frame synchronization

- **How to write the WIS Clause by cross-referencing ANSI T1.416-1999**
  - All required changes and additions are indicated
  - Provides SDH compatibility

XBI - Optional PMA Service Interface for Serial PMD's

IEEE P802.3ae La Jolla Meeting

July 10-14, 2000

Optional PMA Interface for Serial PMD's

IEEE P802.3ae La Jolla Meeting
July 10-14, 2000

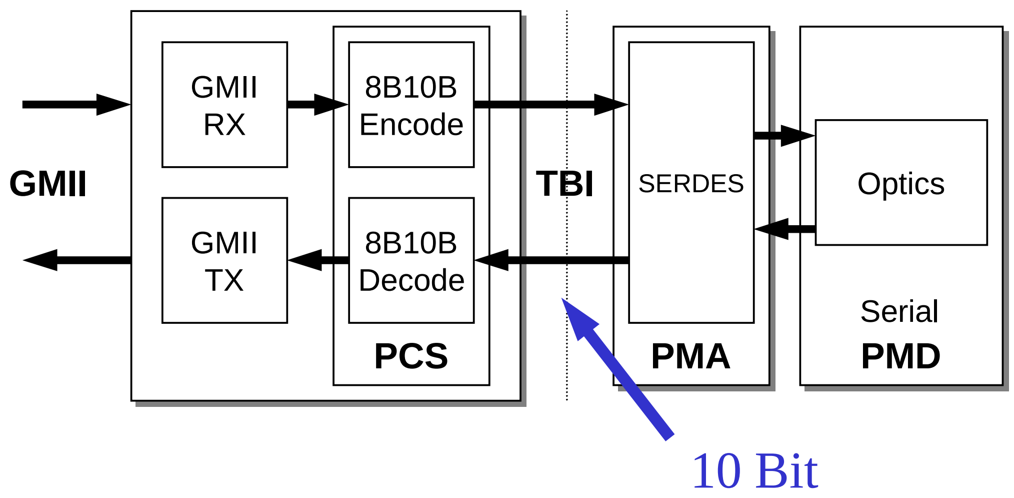
By Richard Dugan, Del Hanson, Agilent, Tom Palkert, AMCC,
Mike Lerer Avici, Mike Dudek, Jason York, Todd Hudson, Bob Mayer, Cielo,
Vipul Bhatt, Finisar, Joel Goergen, Som Sikdar, Force10 Networks,
John Ewen, Ladd Freitag, Jeff Lynch, IBM, Brad Booth, Intel,
Ramesh Padmanabhan, Juniper Networks, Ed Cornejo Lucent Technologies,
Scott Lowrey, Network Elements, Paul Bottorff, David Martin, Nortel Networks,
Don Alderrou, Steve Dreyer, Rich Taborek, nSerial, Osamu Ishida, NTT,
Van Lewing, QED, Tom Alexander, Gary Bourque, Joel Dedrick,
Stuart Robinson, PMC Sierra

Optional PMA Interface Spec needed

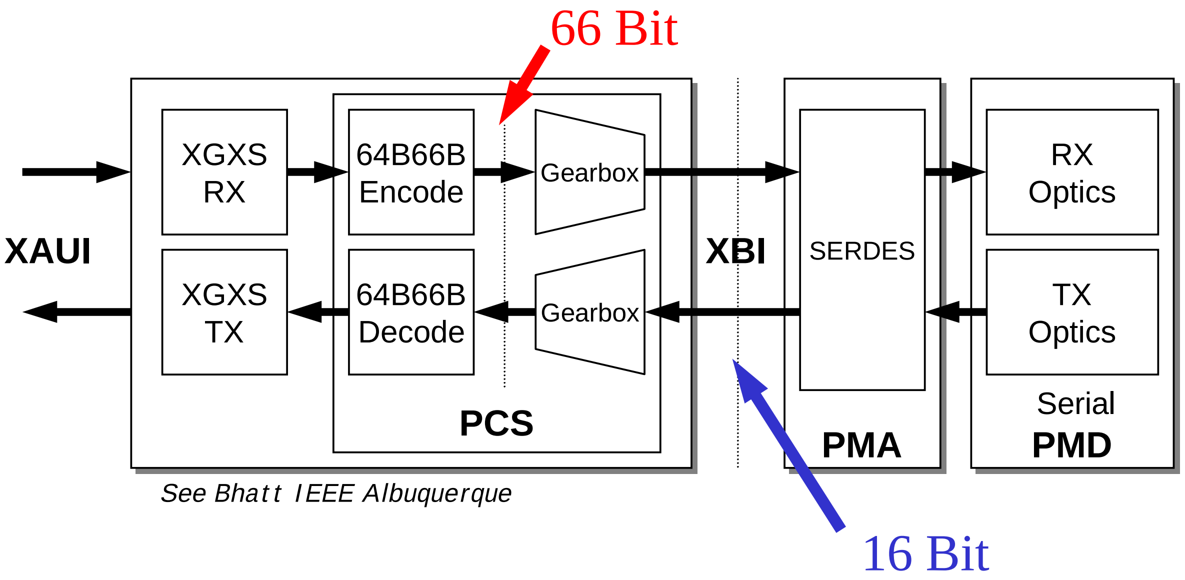
- An Optional PMA Interface (XBI) definition is needed
  - Ensure interoperability between Serial WAN/LAN PCS and SERDES chips (within optical module).
- PCS to PMA interface logical technology split
  - PCS likely in CMOS
  - PMA SERDES likely in SiGe, GaAs, Silicon Bipolar etc.
  - Potential to have these devices come from different vendors.
  - Interoperability definition required.

PMA Interface Precedent

- Gigabit Ethernet
- IEEE 802.3 1998 defines the Ten Bit Interface for serial transmission.
- Physical Instantiation of PMA (Clause 36.3.3 to 36.3.6).
- 8B/10B output is 10 bits wide.
- Narrow enough to use as the PMA Interface.



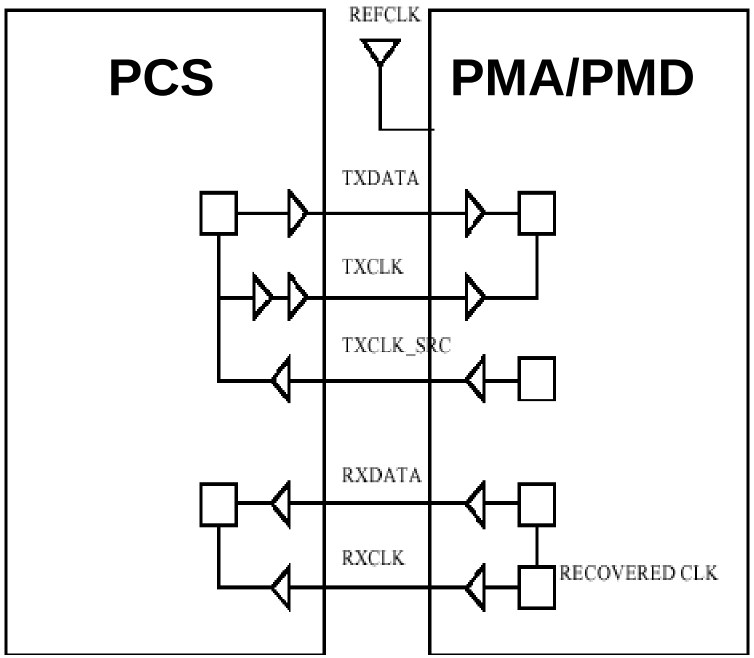
- 10 Gigabit Ethernet Serial LAN PHY
- 64B66B coder output is 66 bits wide.
- Gearbox solution to reduce pins to 16, a manageable number.



See Bhatt IEEE Albuquerque

PMA Service Interface XBI Proposal

– Aggregate rate of 9.953 - 10.3 Gbit/s.
– 16 differential pairs with 622-645 MHz operation, LVDS I/O
– 622-645 MHz Source synchronous clocking.
– REFCLK remains unspecified.



SFI-4 16 Bit SERDES Interface

- OIF SERDES interface for OC-192 (SFI-4)
  - Aggregate rate of 9.953 Gbit/s.
  - 16 differential pairs with 622 MHz operation.
  - LVDS I/O (IEEE Std 1596.3-1996).
  - 622MHz Source synchronous clocking.
  - SFI-4 Applicable to speeds up to 10.66 Gbit/s.
- Status:
  - Specification in final ballot now. (reference doc number OIF1999.102).
  - Interface has been demonstrated in working silicon.
- 10GE Serial LAN PHY Rate Accommodated by Existing Spec
  - “Other reference clock frequencies in addition to the 622.08 MHz are allowed”
  - We are within the bounds of SFI-4 as long as encoded bit rate is less than 10.6Gbit/s.
  - Use SFI-4 16x622 as base - set operating range 622 to 645 MHz for 10.3 Gbit/s.
  - Relaxation of SFI-4 may be necessary for Ethernet applications.

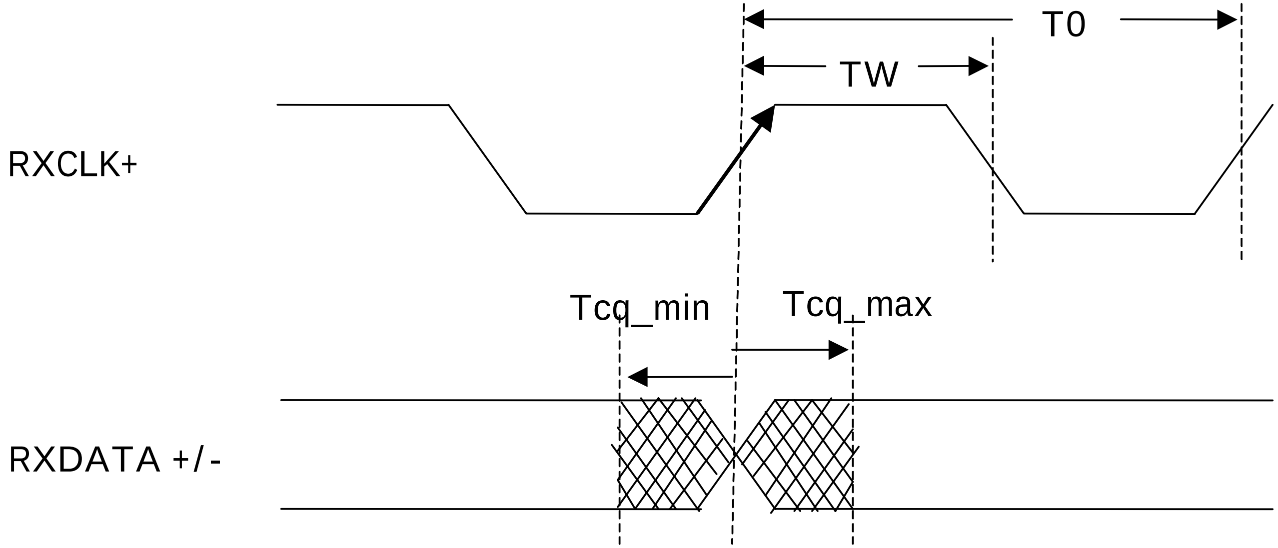
Why Add XBI to the IEEE 802.3ae standard

- OIF SFI-4 16x622 work has been done.
- Current SFI-4 spec allows higher freq, but does not specify them.
  - Thus, it is not guaranteed that vendors will build LAN PHY rate in an interoperable fashion
- OIF not a standards body (they create specifications for implementor's agreements) thus the IEEE P802.3ae cannot reference the SFI-4 specification.
- IEEE P802.3ae needs to control the PMA Interface definition so that it is not changed by the OIF.

XBI Interface Signals

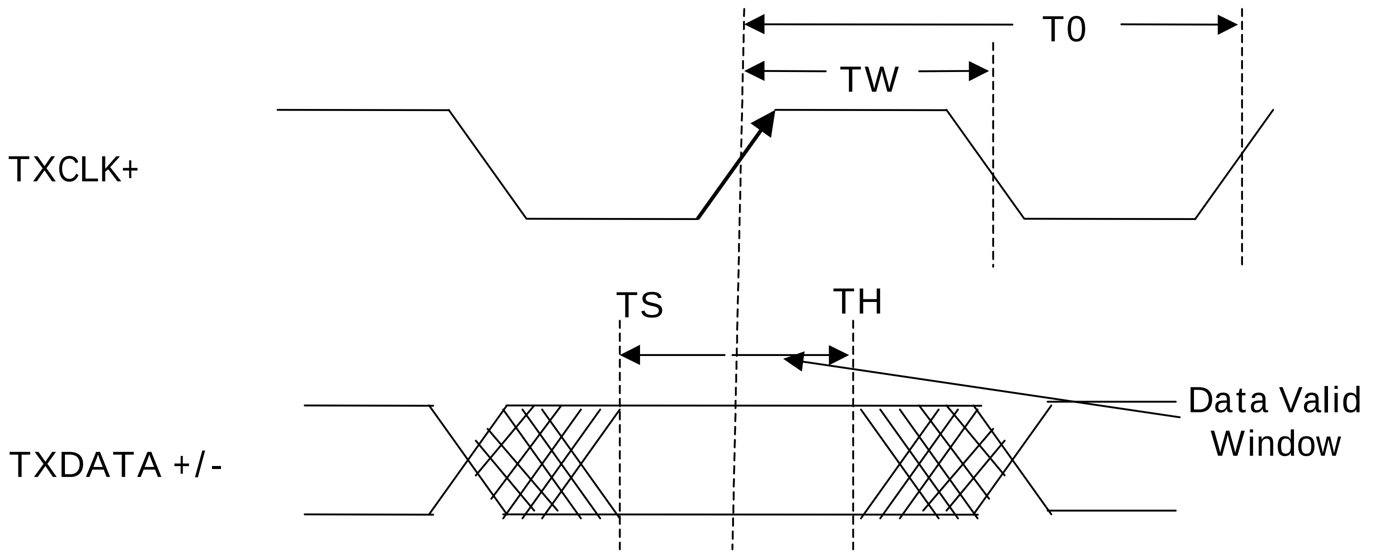
Symbol	Signal Name	Signal Type	Active Level	Description
PMA_TXDATA+<15:0> PMA_TXDATA-<15:0>	Transmit Data	I-LVDS	Diff	16 bit transmit data from the PCS to PMA.
PMA_TXCLK+ PMA_TXCLK-	Transmit Clock.	I-LVDS	Diff	Transmit clock to latch data into PMA. Ranges from 622 MHz to 645 MHz with +/- 100ppm tolerance.
PMA_TXCLK_SRC+ PMA_TXCLK_SRC-	Transmit Clock Source	I-LVDS	Diff	Transmit clock from the PMA to the PCS. May be used by PCS to generate the transmit clock.
PMA_RXDATA+<15:0> PMA_RXDATA-<15:0>	Receive Data	I-LVDS	Diff	16 bit received data presented to the PCS from the PMA.
PMA_RXCLK+ PMA_RXCLK-	Receive Clock	I-LVDS	Diff	Receive clock to latch data into PCS. Ranges from 622 MHz to 645 MHz with +/- 100ppm tolerance.

XBI PMA LVDS Output Waveforms



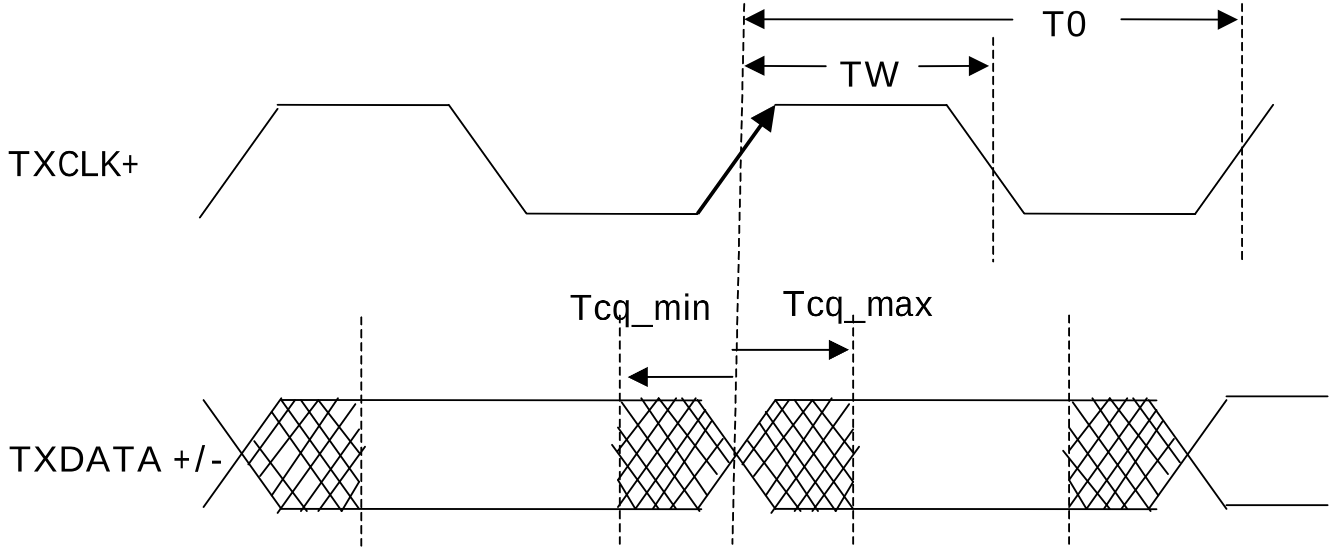
Parameter	Description	Value	Units
T0	Clock period	1.552 to 1.608	ns
TW/T0	duty cycl	0.45 < TW/T0 < 0.55	
TR, TF	20-80% rise, fall times	100-250	ps
Tcq_min, Tcq_max	Clock to out times	200, 200	ps

XBI PMA LVDS Input Waveforms



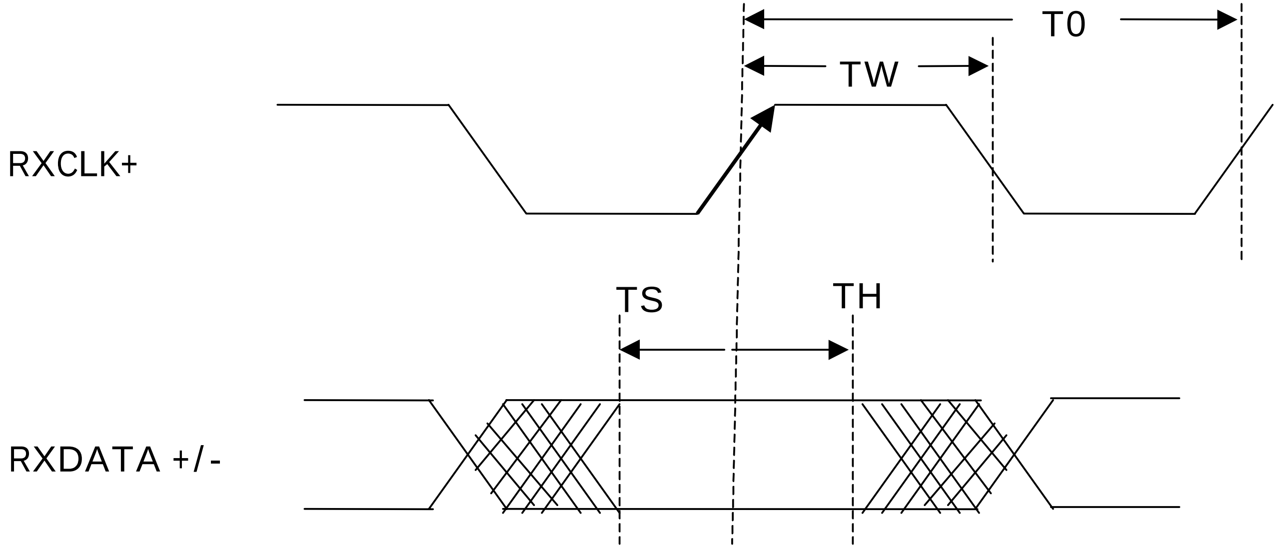
Parameter	Description	Value	Units
T0	Clock period	1.552 to 1.608	ns
TW/T0	duty cycl	0.4 < TW/T0 < 0.6	
TR, TF	20-80% rise, fall times	100-300	ps
Tcq_min, Tcq_max	Clock to out times	300, 300	ps

XBI PCS LVDS Output Waveforms



Parameter	Description	Value	Units
T0	Clock period	1.552 to 1.608	ns
TW/T0	duty cycle	0.4 < TW/T0 < 0.6	
TR, TF	20-80% rise, fall times	100-250	ps
TS, TH	Clock to out times	200, 200	ps

XBI PCS LVDS Input Waveforms



Parameter	Description	Value	Units
T0	Clock period	1.552 to 1.608	ns
TW/T0	duty cycle	0.45 < TW/T0 < 0.55	
TR, TF	20-80% rise, fall times	100-300	ps
TS, TH	Clock to out times	300, 300	ps

Issues to Resolve

- Determine appropriate jitter requirements.
  - To be addressed at a meeting at this plenary.

Summary

- An Optional Instantiation of the PMA Service Interface needs to be defined for the Serial PHYs in IEEE P802.3ae
  - Ensure interoperability between Serial WAN/LAN PCS and SERDES chips (in optical module).
  - Promotes multi-vendor chip interoperability.
- PCS-PMA Logical technology split.
- Builds on the precedent of Gigabit Ethernet TBI (Clause 36.3.3).
- Simply re-use OIF work to achieve Time to Market
  - SFI-4 16x622 specification is complete.
  - Cannot reference OIF SFI-4.
  - SFI-4 can accommodate both LAN & WAN PHY rates.
    - 622 – 645 MHz LVDS within current process capabilities
    - 622 – 645 MHz board implementation understood.
    - Relaxation may be necessary for Ethernet environments.

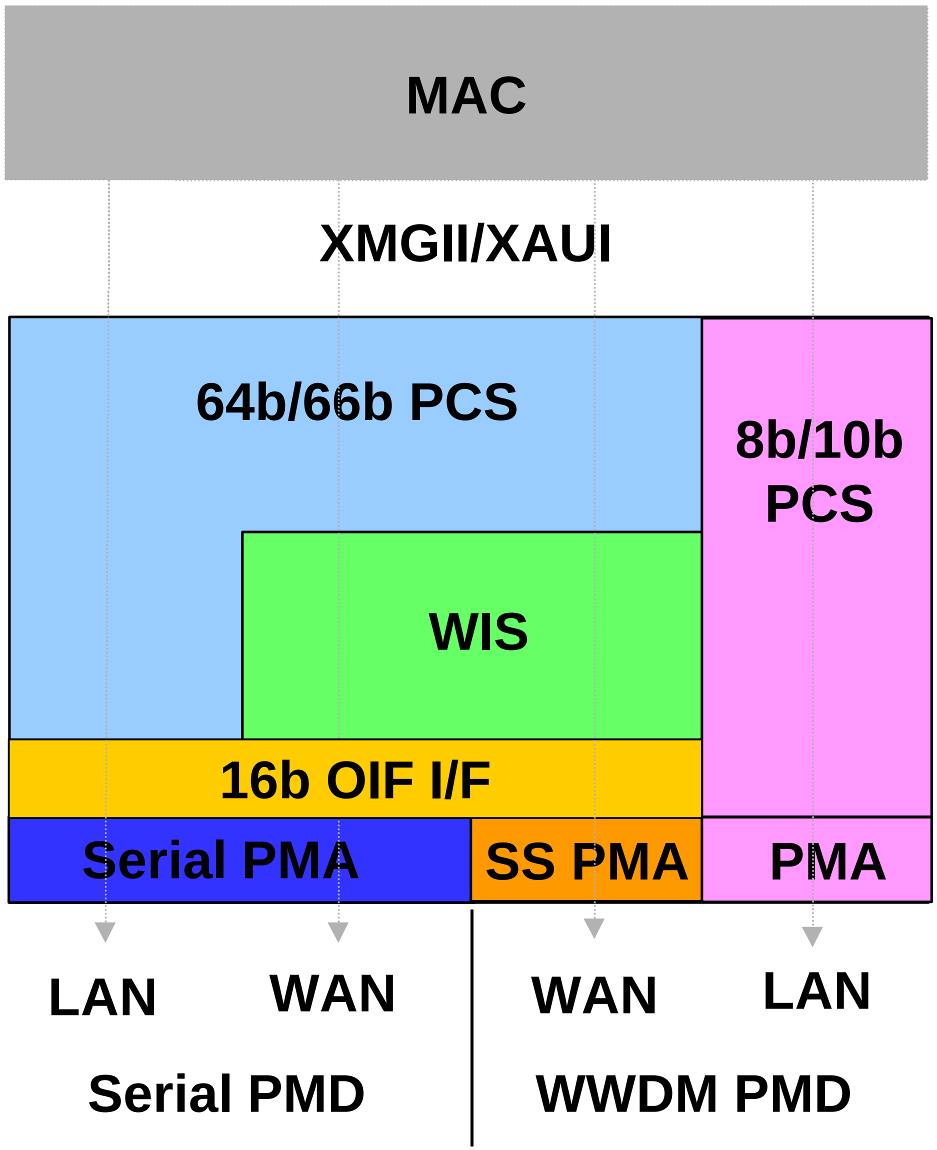
SUPI Update

IEEE P802.3ae
La Jolla
July 2000

Norival Figueira, Paul Bottorff, David Martin, Tim Armstrong, Bijan Raahemi: ……………………	Nortel Networks
Howard Frazier:………………………………………..	Cisco Systems
Enrique Hernandez (Bell Labs), Nevin Jones (Microelectronics):………………………………….	Lucent
Tom Palkert:………………………………………….	AMCC
Iain Verigin, Stuart Robinson, Tom Alexander, Farzin Firoozmand:………………………………….	PMC Sierra
Nader Vijeh:…………………………………………..	Lantern Communications
Frederick Weniger:…………………………………….	Vitesse
Shimon Muller:………………………………………..	Sun Microsystems
Kevin On:…………………………………………….	Infineon Technologies
Richard Dugan:………………………………………	Agilent
Nan Chen:……………………………………………..	Force10 Networks

V0.1

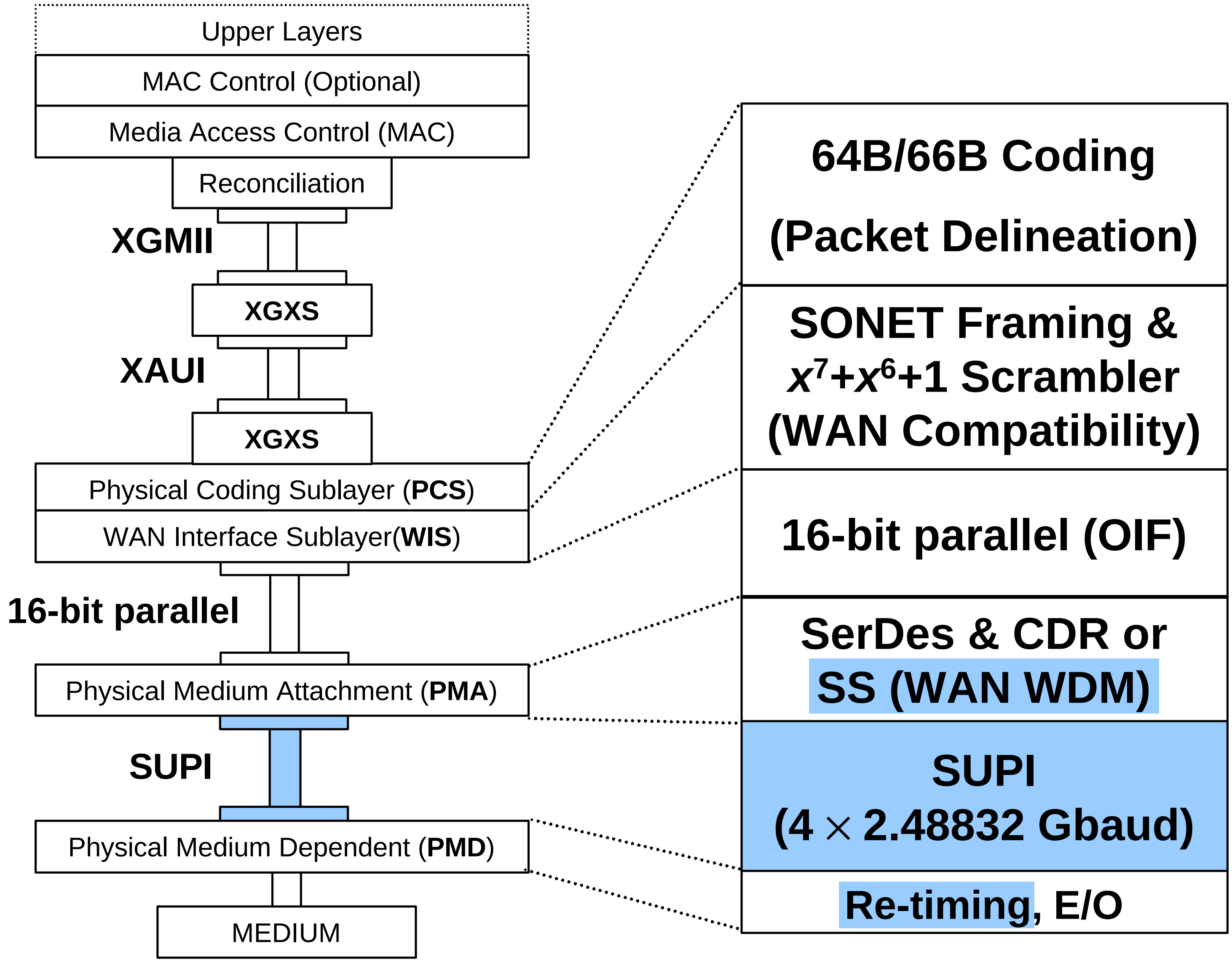
UniPHY Components



Attaching WWDM PMD to WAN PHY

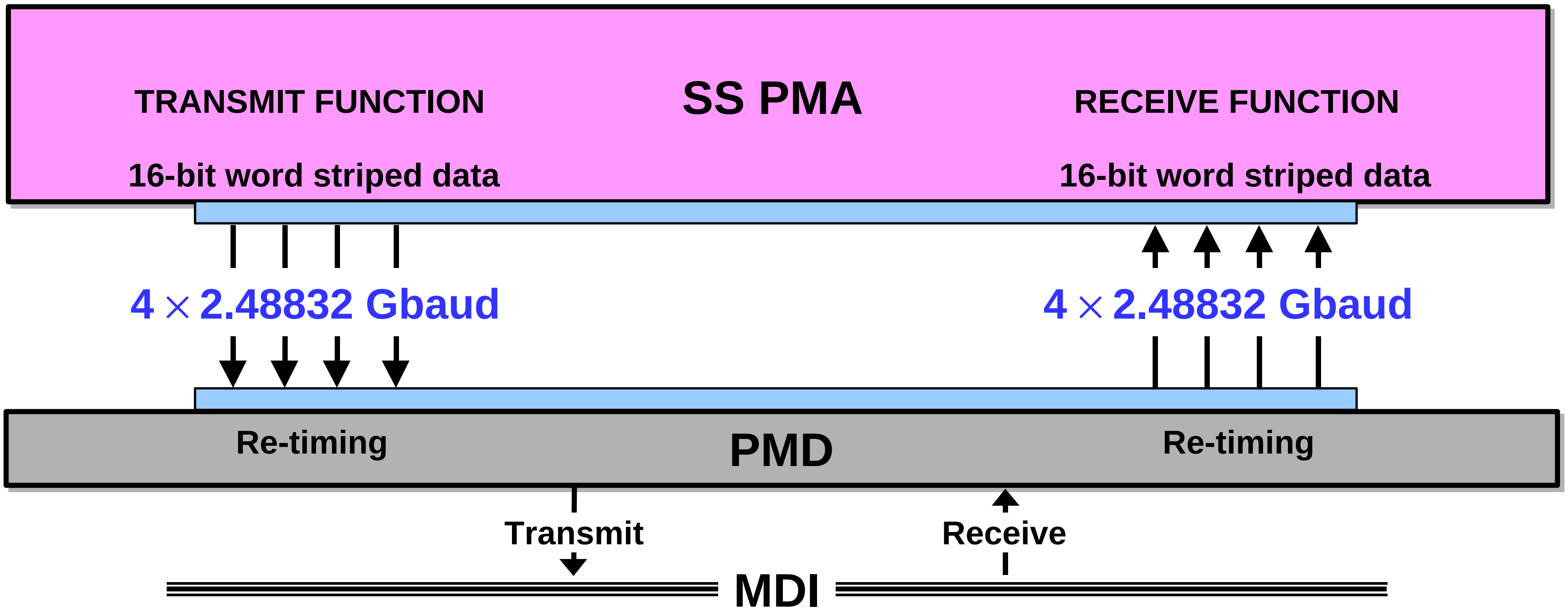
- **XAUI like attachment does not work because WAN PHY data area is pseudo random. WAN PHY data has no frame or gap codes.**
- **To operate on WWDM WAN-PHY must have a PMA function to generate the 4 lanes.**
  - Skew correction is needed between lanes
  - Techniques based on IFG codes can not be used due to the randomization of data

WAN-PHY and UniPHY Layer Model



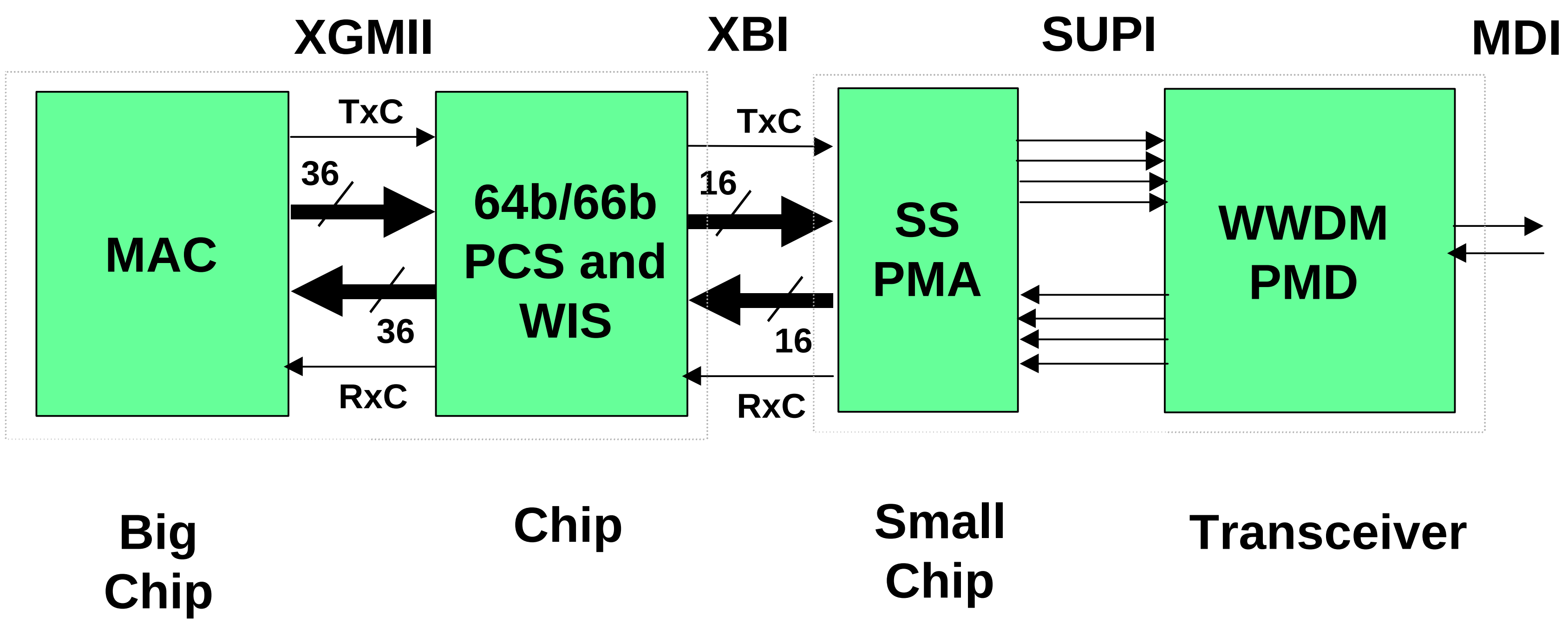
SS = SUPI Sublayer

SUPI (WDM PMD Service Interface)



SS = SUPI Sublayer

SS PMA Implementation Example

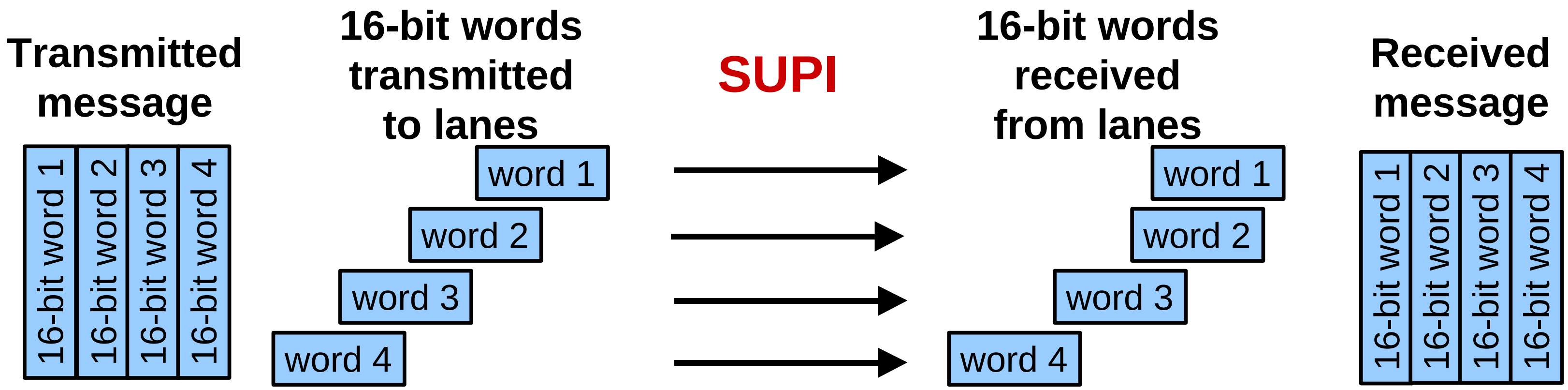


SUPI

- **Used for WWDM and 4× parallel PMDs**
- **Can use a recovered clock to reset jitter**
- **Can provide up to 62.5 usec skew correction**

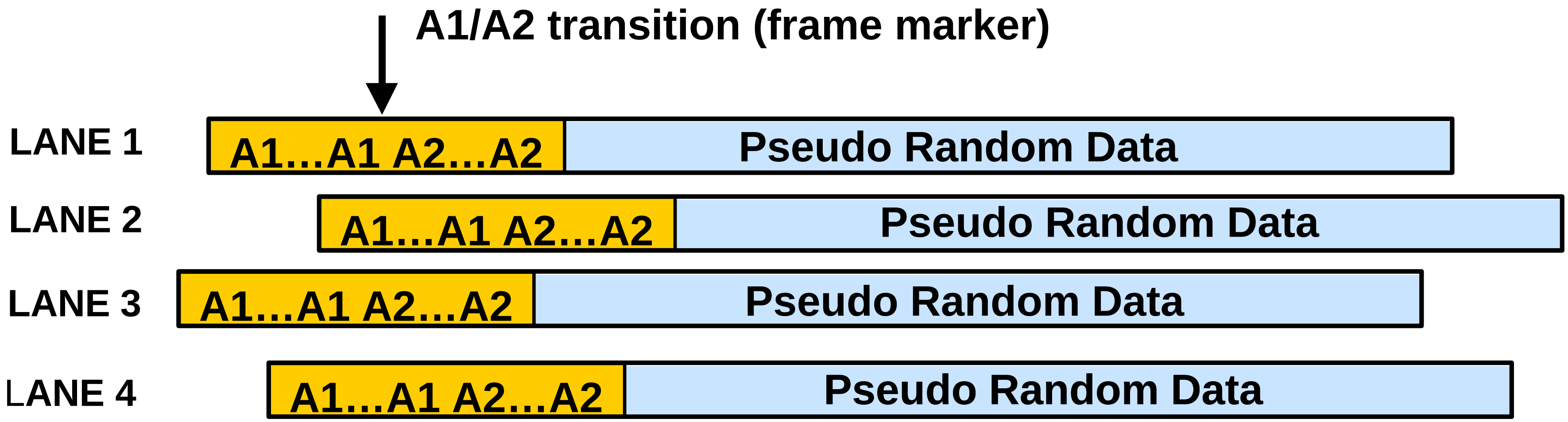
SUPI (cont.)

- **16-bit word striped data transmitted on each lane**
- **Each lane has 1/4 of the (SONET) A1/ A2 framing bytes for lane deskew and synchronization**
  - — Word synchronization from A1/A2 transition
  - — For fixed lane assignment, allows for large skew

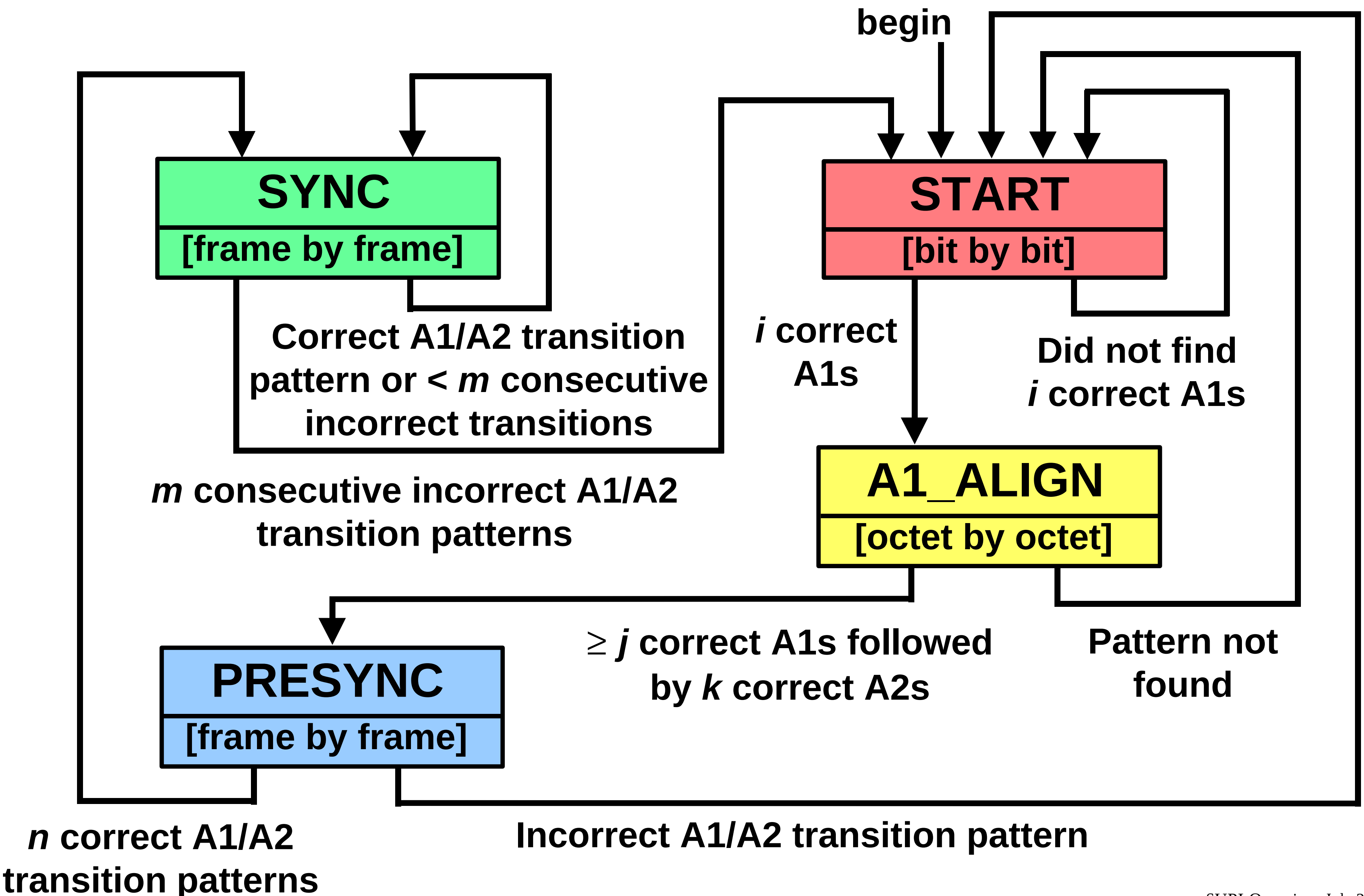


SUPI LANE Deskew

- **Uses A1/A2 transition (i.e., frame marker)**
- **Looks for the A1/A2 framing pattern consistently**
  - Expects it to appear on each lane once every 38880 octets
  - Each lane locks on the synchronization pattern



Lane Sync: State Diagram



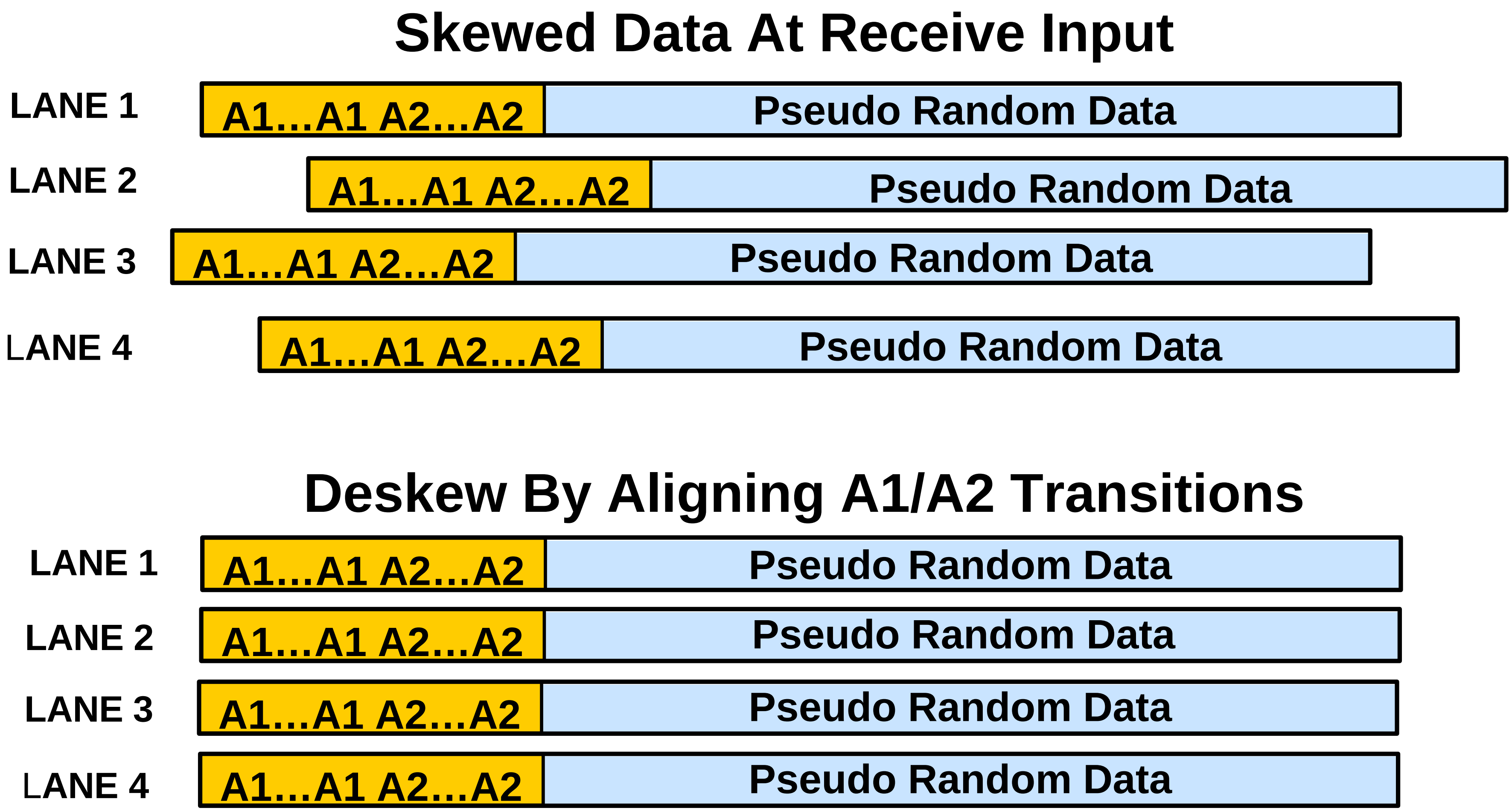
Deskew

- **Skew is imparted by active and passive link elements**
- **SS PMA deskew accounts for all skew present at the Rx**
- **Lane deskew performed by alignment to A1/A2 pattern present every 125 usec**

Skew Source	#	Skew	Total Skew
SerDes Tx	1	1 UI	1 UI
PCB	2	1 UI	2 UI
Medium	1	<16 UI	<16 UI
SerDes Rx	1	16 UI	16 UI
Total			< 35 UI

- **Required deskew is much less than possible 77,760 UI**

Deskew Example



- **Uses**

Summary

- **SUPI**
  - WAN WWDM PMD Service Interface
  - 4 $\times$ 2.48832 Gbaud
  - 16-bit word striped data transmitted on each lane
  - Each lane has 1/4 of the (SONET) A1/ A2 framing bytes for lane deskew
  - Word synchronization from A1/A2 transition

Proposed Set Of Three 10 Gigabit Ethernet PMDs & Related Specifications

Del Hanson & Piers Dawe, Agilent Technologies
Vipul Bhatt, Finisar
Mike Lerer, Avici Systems
Wenbin Jiang, E2O Communications
Brad Booth & Bob Grow, Intel
Ed Cornejo, Lucent
Stuart Robinson, Tom Alexander, & Gary Bourque, PMC-Sierra
Shimon Muller, Sun
Kevin Daines, World Wide Packets

10GbE Interim Meeting, Ottawa   May 23-35, 2000

Purpose

- To propose a set of three PMD implementations that meet all the distance objectives of P802.3ae.
- The set consists of :
  - WWDM at 1310 nm
  - Serial at 1310 nm
  - Serial at 1550 nm
- Target specifications for these three PMDs are described
- There will be separate presentations on other PMD cases

Figure 38-1 (equivalent for WWDM)

- This is the same for WWDM on a per lane basis

802.3z Figure 38-1 shows PMA, PMD, Fiber Optic Cabling (channel) and four test points

Table 38-6 (equivalent for WWDM)

Operating range for 10000BASE-LX WWDM over each optical fiber type

Fiber type	**Modal BW @ 1300 nm (min. overfilled launch) (MHz*km)**	Minimum range (meters)
62.5 um MMF	500	2-300
50 um MMF	400	2-240
50 um MMF	500	2-300
10 um SMF	N/A	2-10,000

Table 38-7 (equivalent for WWDM)

10000BASE-LX WWDM transmit characteristics

Description	62.5 um MMF, 50 um MMF, 10 um SMF	Unit
Transmitter type	Longwave Laser	
Signaling speed per lane (range)	3.125 +/- 100 ppm	GBd
Wavelength (range), four lanes	1270-1355	nm
Lane center wavelengths	1275.7, 1300.2, 1324.7, 1349.2 +/- 5.7	nm
Lane separation	24.5	nm
Trise/Tfall (max. 20-80% response time)	100	ps
Side-mode suppression ratio (SMSR), (min)	0.0	dB
RMS spectral width (max)	0.62	nm
Average launch power, four lanes (max)	3.5	dBm
Average launch power, per lane (max)	-2.5	dBm
Average launch power, per lane (min)	-7.5	dBm
Avg. launch power of OFF transmitter, per lane (max)	-30	dBm
Extinction ratio, (min)	7	dB
RIN (max)	-120	dB/Hz

Table 38-8 (equivalent for WWDM)

10000BASE-LX WWDM receive characteristics

Description	62.5 um MMF 50 um MMF	10 um SMF	Unit
Signaling speed per lane (range)	3.125 +/- 100 ppm		GBd
Wavelength (range), four lanes	1270 to 1355		nm
Lane center wavelengths	1275.7, 1300.2, 1324.7, 1349.2 +/- 5.7		nm
Lane separation	24.5		nm
Average receive power, four lanes (max)	3.5		dBm
Average receive power, per lane (max)	-2.5		dBm
Return loss (min)	12		dB
Receive electrical 3 dB upper cutoff frequency (max)	3750		MHz
Receive sensitivity	-15.5	-16.5	dBm
Stressed receive sensitivity	-10.3	-15.0	dBm
Vertical eye closure penalty	3.60	0.74	dB

Table 38-9 (equivalent for WWDM)

Worst case 10000BASE-LX WWDM link power budget and penalties

Parameter	62.5 um MMF	50 um MMF		10 um SMF	Unit
Modal bandwidth as measured at 1300 nm, (min, overfilled launch)	500	400	500	N/A	MHz*km
Link power budget	8.0	8.0	8.0	9.0	dB
Operating distance	300	240	300	10,000	m
Lane insertion loss	2.46	2.37	2.46	7.14	dB
Link power penalties	4.63	5.13	5.13	1.82	dB
Unallocated margin in link power budget	0.91	0.50	0.41	0.04	dB

Note 1: MMF parameters are calculated with link model having DCD_DJ = 25.0 ps
Note 2: SMF parameters are calculated with link model having DCD_DJ = 20.5 ps

Table 38-10 (equivalent for WWDM)

10000BASE-LX WWDM jitter budget

Compliance point	Total jitter		Deterministic jitter	
	UI	ps	UI	ps
TP1	0.240	76.8	0.100	32.0
TP1 to TP2	0.284	90.9	0.100	32.0
TP2	0.431	138.0	0.200	64.0
TP2 to TP3	0.170	54.4	0.050	16.0
TP3	0.510	163.4	0.250	80.0
TP3 to TP4	0.332	106.2	0.212	67.8
TP4	0.749	239.6	0.462	147.8

Figure 38-1 (equivalent for serial SMF links)

- Almost the same as in 802.3z
- The mode conditioning patch cord does not apply
- TP1 and TP4 are not likely to be physically accessible interfaces

(802.3z Figure 38-1 shows PMA, PMD, Fiber Optic Cabling (channel) and four test points)

Table 38-6 (equivalent for Serial SMF links)

Operating range for serial links using 10 um SMF links with two source types

Fiber type	Source Type	Target range (meters)
10 um SMF	1310 nm Laser	2-10,000
	1550 nm Modulator	2-40,000

Note 1. Operating ranges are targets because the attenuation of the outside plant is not guaranteed by standards.
Note 2. Shortest 1550 nm links may require an attenuator to avoid over-driving the receiver.

Table 38-7 (equivalent for Serial SMF links)

Transmit characteristics for serial 10 um SMF links

Description	Value		Unit
Transmitter type	Single longitudinal mode laser	Modulator	
Signaling speed (range)	10.3125 +/- 100 ppm		GBd
Wavelength (range),	1290-1330*	1530-1565	nm
Trise/Tfall (max. 20-80% response time)	40	33	ps
Side-mode suppression ratio (SMSR), (min)	30.0*	30.0*	dB
RMS spectral width (max)	0.40*	0.034*	nm
Average launch power, (max)	1.0	+2	dBm
Average launch power, (min)	-4.0	-2	dBm
Avg. launch power of OFF transmitter, (max)	-30		dBm
Extinction ratio, (min)	6*	8*	dB
RIN (max)	-130	-140	dB/Hz

- Notes on following slide

Table 38-7 (continued)

- Note 1: Change to Optical Modulated Amplitude (OMA) specification is proposed (actually OMA is a power).
- Note2: The 1310 nm link spectral characteristics are being review to possibly accommodate 1300 nm VCSELs.
- Note 3: The low spectral width of 1550 nm link is a temporary representation, a placeholder, for further work to be done regarding dispersion accommodation.
- Note 4: The 1310 nm case uses directly modulated laser where low extinction ratio helps the laser speed. The 1550 nm case uses a modulator which can deliver high extinction ratio. 8 dB is near the ITU/SONET specification.
- Note 5: SMSR reduction may improve cost effectiveness. This is currently under review.

Table 38-8 (equivalent for Serial SMF links)

Receive characteristics for serial 10 um SMF links

Description	Value		Unit
Signaling speed (range)	10.3125 +/- 100 ppm		GBd
Wavelength (range)	1290-1330	1530-1565	nm
Average receive power, (max)	1.0	-8.0	dBm
Receive sensitivity	-14.0	-20.0*	dBm
Return loss (min)	12		dB
Stressed receive sensitivity	-11.45	-15.41	dBm
Vertical eye closure penalty	1.71	2.72	dB

*Note: This is too optimistic. Further design work needed.

Table 38-9 (equivalent for Serial SMF links)

Worst case 10000BASE-LX Serial 10 um SMF link power budget and penalties

Parameter	1310 nm transmitter	1550 nm transmitter	Unit
Link power budget	10.0	18.0	dB
Operating distance	10	40	km
Link insertion loss	7.04	13.0	dB
Link power penalties	2.27	3.36	dB
Unallocated margin in link power budget	0.69	1.64	dB

Note 1: Table parameters are calculated with link model having DCD_DJ = 8.0 ps

Table 38-10 (equivalent for Serial SMF links)

10000BASE-LX serial SMF link jitter budget

Compliance point	Total jitter		Deterministic jitter	
	UI	ps	UI	ps
TP1	0.240	23.3	0.100	9.7
TP1 to TP2	0.284	27.5	0.100	9.7
TP2	0.431	41.8	0.200	19.4
TP2 to TP3	0.170	16.5	0.050	4.8
TP3	0.510	49.5	0.250	24.2
TP3 to TP4	0.332	32.2	0.212	20.6
TP4	0.749	72.6	0.462	44.8

Further Work

- WWDM specifications are stable.
- Serial link specification issues (indicated by *)
  - Operating ranges are targets due to unspecified fiber loss
  - Optical modulated amplitude (OMA) may replace extinction ratio (ER)
  - SMSR reduction will be reviewed to explore performance/cost trade-off
  - Serial jitter budgets will benefit from optimization work
- Additional 1550 nm 40 km link specifications issues
  - Spectral width and receive sensitivity will be reconsidered
  - Increasing the link length beyond 40 km will need OC-192 optical engineering. Shortest links require an attenuator.
  - It it likely that the 40 km specification can be achieved without using optical amplifiers or avalanche photodiodes.