Serial Attached SCSI
Phy layer

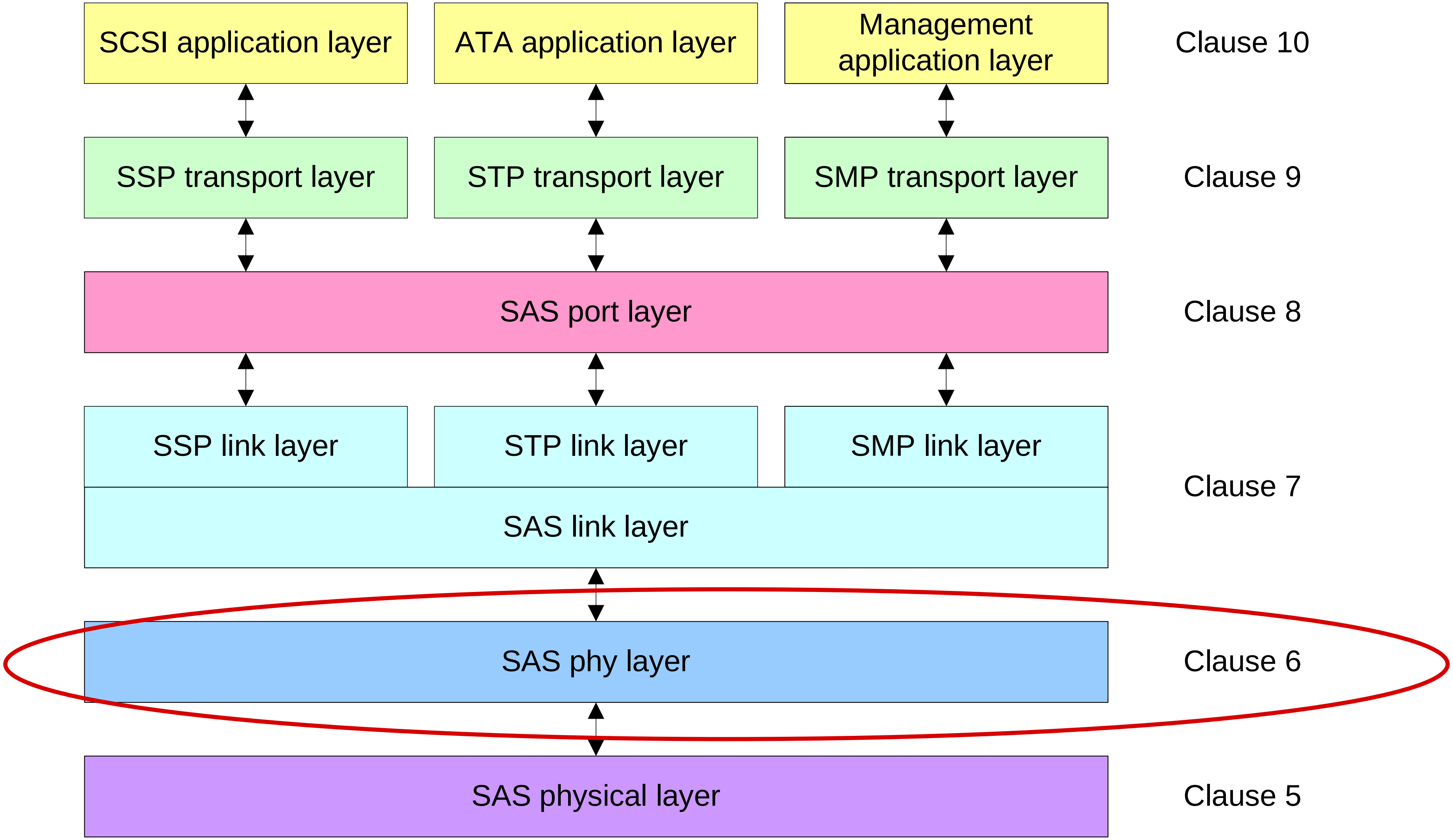
by Rob Elliott
HP Industry Standard Servers
Server Storage Advanced Technology
elliott@hp.com http://www.hp.com

30 September 2003

Notice

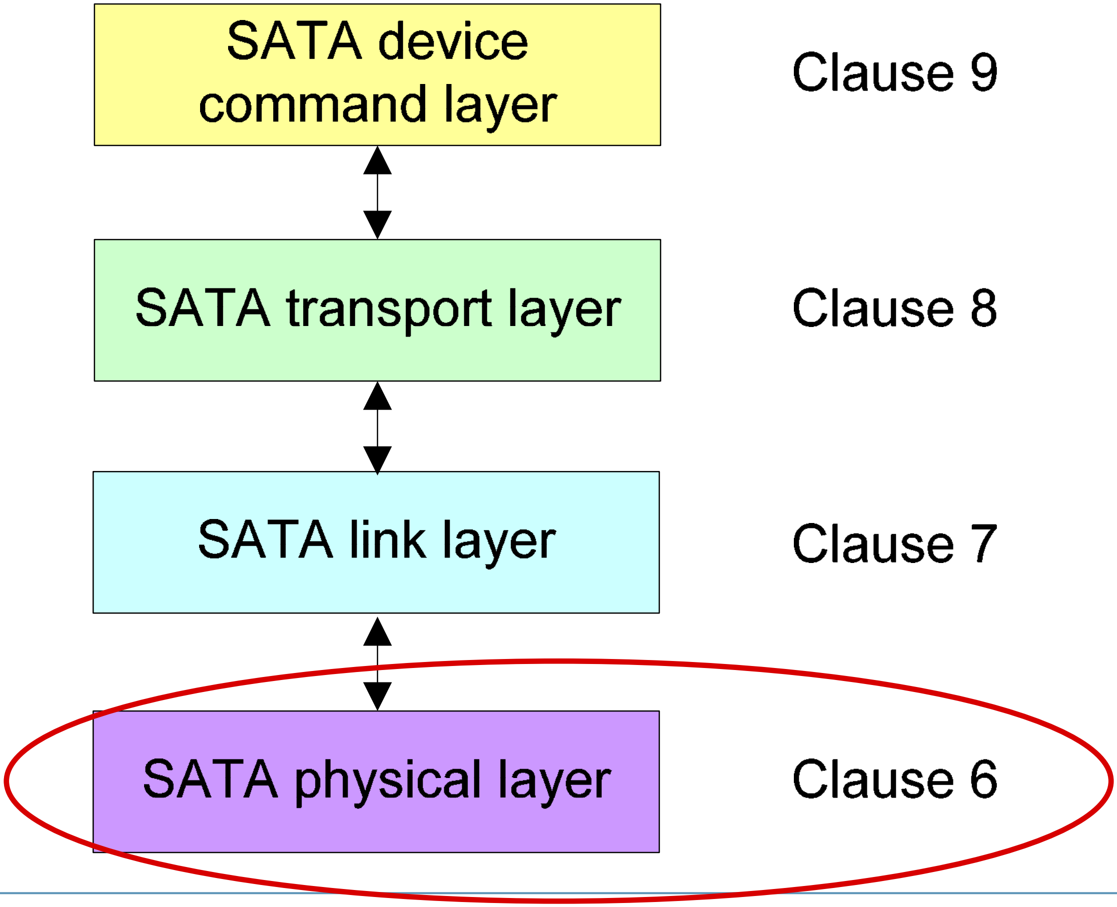
- These slides are freely distributed by HP through the SCSI Trade Association (http://www.scsita.org)
- STA members are welcome to borrow any number of the slides (in whole or in part) for other presentations, provided credit is given to the SCSI Trade Association and HP
- This compilation is © 2003 Hewlett-Packard Corporation

SAS standard layering



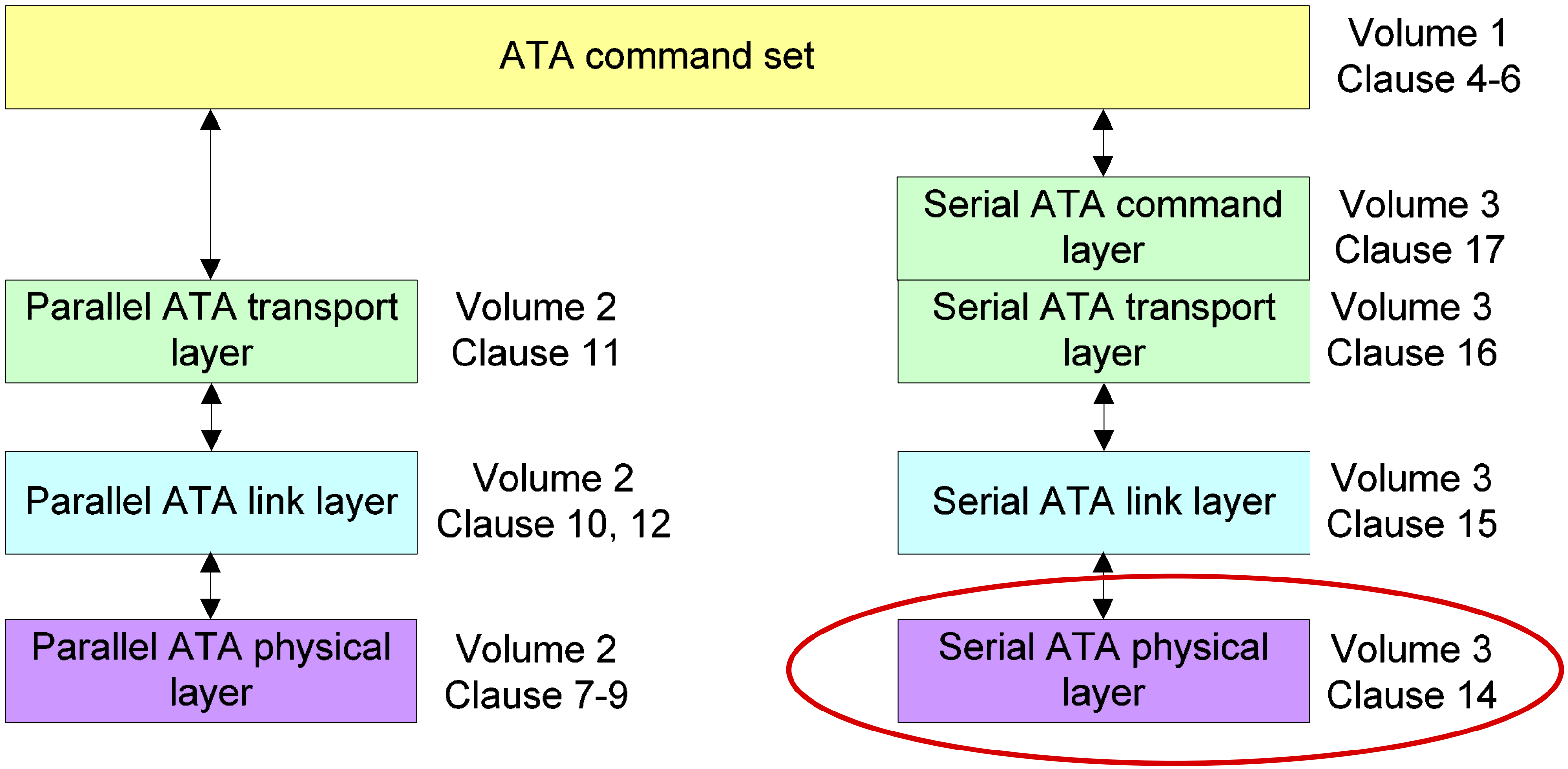
SATA 1.0a standard layering

- For SATA 1.0a from the private Serial ATA working group



ATA/ATAPI-7 standard layering

- For the public standard ATA/ATAPI-7
- Subject to change by T13 standards committee



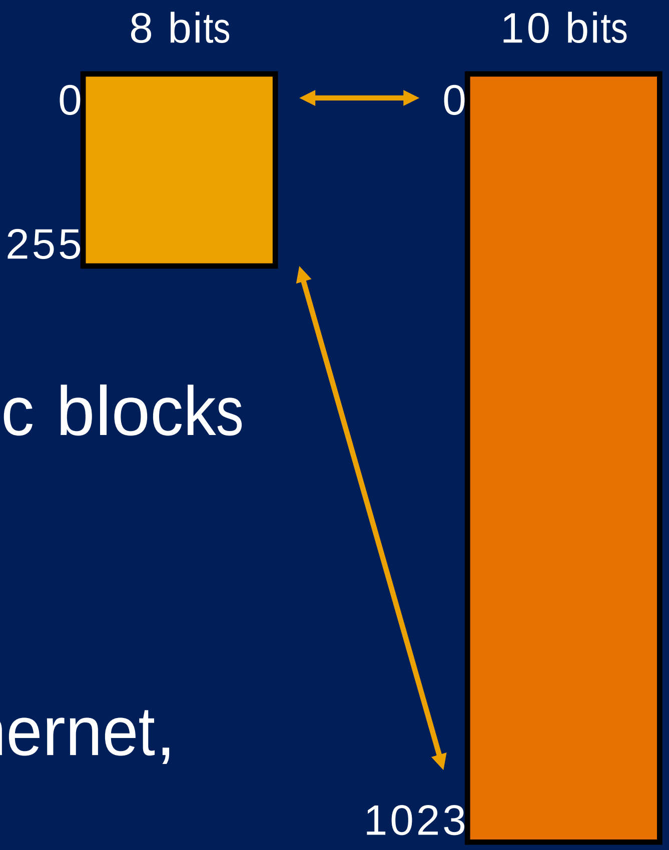
SAS clause 6 – Phy layer

- Encoding (8b10b)
- From bits to dwords
- OOB signals
- Phy reset sequence
  - OOB sequence
  - Speed negotiation sequence
- State machines
  - SAS SP, SP_DWS
  - SATA

Phy layer - Encoding

Encoding

- 8b10b coding converts 8-bit bytes into 10-bit data characters for transmission on the wire
- Reasons
  - Clock recovery
  - DC balance
  - Special characters
  - Error detection
- Mapping done with two simple logic blocks
  - 5b6b and 3b4b
  - Full table in SAS and SATA
- Invented by IBM in 1983
- Used by Fibre Channel, Gigabit Ethernet, 1394b, and many other standards



8b10b mapping

- Patterns without 6/4, 5/5, or 4/6 zeros/ones are considered invalid
- Longest stream of consecutive bits: 5 zeros or 5 ones
- Zxx.y nomenclature
  - Z is D for data characters and K for control characters
  - xx.y highlight the bits going into 5b6b and 3b4b encoding
    - "xx" represents bits 4:0 (0 through 31) for 5b6b
    - "y" represents bits 7:5 (0 through 7) for 3b4b
    - D00.0 means 00h
    - D01.0 means 01h
    - D00.1 means 20h

8b10b data characters

- Data characters (Dxx.y)
  - All 256 data bytes are mapped into data characters
  - Bytes that map into a 6/4 character also map into a 4/6 character
  - Some bytes map into one 5/5 character; others into two
  - Examples
    - D00.0 (00h) is100111_0100 (5/5) or 011000_1011 (5/5)
    - D01.0 (01h) is 011101_0100 (4/6) or 100010_1011 (6/4)
    - D00.1 (20h) is 100111_1001 (6/4) or 011000_1001 (4/6)
    - D00.4 (80h) is 100111_0010 (5/5) or 011000 _1101 (5/5)
    - D10.2 (4Ah) is 010101_0101 (5/5)
    - D21.5 (B5h) is 101010_1010 (5/5)
      - These two characters have the maximum number of transitions

8b10b control characters

- Control characters (Kxx.y)
  - 12 characters remain that are not mapped to data bytes
  - used for special purposes
  - K28.1, K28.5, K28.7 are the only characters containing a comma pattern (1100000 or 0011111)
  - Examples
    - K28.3 is 001111_0011 (4/6) or 110000_1100 (6/4)
    - K28.5 is 001111_1010 (4/6) or 110000_0101 (6/4)
  - In implementations, can use an 8-bit data byte with an extra Z bit set to indicate K or D

8b10b properties

- Running disparity
  - If a data character has two mappings, choose the mapping based on previous character sent that had two mappings
  - After sending a 6/4, next byte that is not 5/5 must be 4/6
  - E.g. 6/4, 5/5, 5/5, 5/5, 4/6, 6/4, 4/6, 5/5, 6/4, …
  - Positive = 6/4 was last
  - Negative = 4/6 was last
  - After power on, may start with either disparity
- Clock recovery
  - Each character has multiple 0 to 1 and 1 to 0 transitions
  - Phase Lock Loop (PLL) can construct a clock based on the data stream
  - No need for separate clock signal – clock is embedded in the data stream

Phy layer – From bits to dwords

Dwords

- Byte = 8 bits (xxh)
- Character = 10 bits (Dxx.y or Kxx.y)
- Dword = 4 characters (or 4 bytes, depending on context)
  - 40 bits on the wire
  - Usually represents 32 bits of data
  - A dword may flip disparity or leave it the same
  - A dword is either a data dword, a primitive, or an invalid dword

first	second	third	fourth
char	char	char	char

Data dwords

- 4 data characters

first	second	third	fourth
Dxx.y	Dxx.y	Dxx.y	Dxx.y

- For SAS (except for Serial ATA Tunneling - STP), big-endian byte ordering
- For SATA, little-endian byte ordering

first	second	third	fourth	
D00.0	D00.0	D00.0	D01.0	00000001h in SAS
D01.0	D00.0	D00.0	D00.0	00000001h in SATA
D00.4	D00.0	D00.0	D00.0	80000000h in SAS
D00.0	D00.0	D00.0	D00.4	80000000h in SATA

Primitives

- 1 control character and 3 data characters
  - First character is K28.5 (for SAS primitives), K28.3 (for SATA primitives), or K28.6 (special SATA error primitive)
    - K28.6 primitive serves as an invalid dword for SATA
  - Last three characters are data characters
  - Endianness does not matter
    - both SAS and SATA primitives always have the control character first on the wire

first	second	third	fourth
K28.5	Dxx.y	Dxx.y	Dxx.y
K28.3	Dxx.y	Dxx.y	Dxx.y
K28.6	Dxx.y	Dxx.y	Dxx.y

Dword examples

- Primitives are defined by the link layer
  - Data characters chosen for best Hamming distance from each other
  - 8 bit errors are needed to morph one valid primitive into another
- Phy layer uses D10.2 characters during SATA speed negotiation
- Phy layer uses ALIGN(0) primitive during OOB signaling and speed negotiation
- Phy layer uses ALIGN(1) primitive during SAS speed negotiation

	first	second	third	fourth
ALIGN(0)	K28.5	D10.2	D10.2	D27.3
ALIGN(1)	K28.5	D07.0	D07.0	D07.0

Invalid dwords

- Any primitive with a control character other than K28.5 or K28.3

first	second	third	fourth	
K28.n	Dxx.y	Dxx.y	Dxx.y	
K28.6	Dxx.y	Dxx.y	Dxx.y	Invalid per SATA

- Any primitive with a control character other than K28.5 or K28.3

		or	or
any	Kxx.y	Kxx.y	Kxx.y

- Any dword with an invalid character

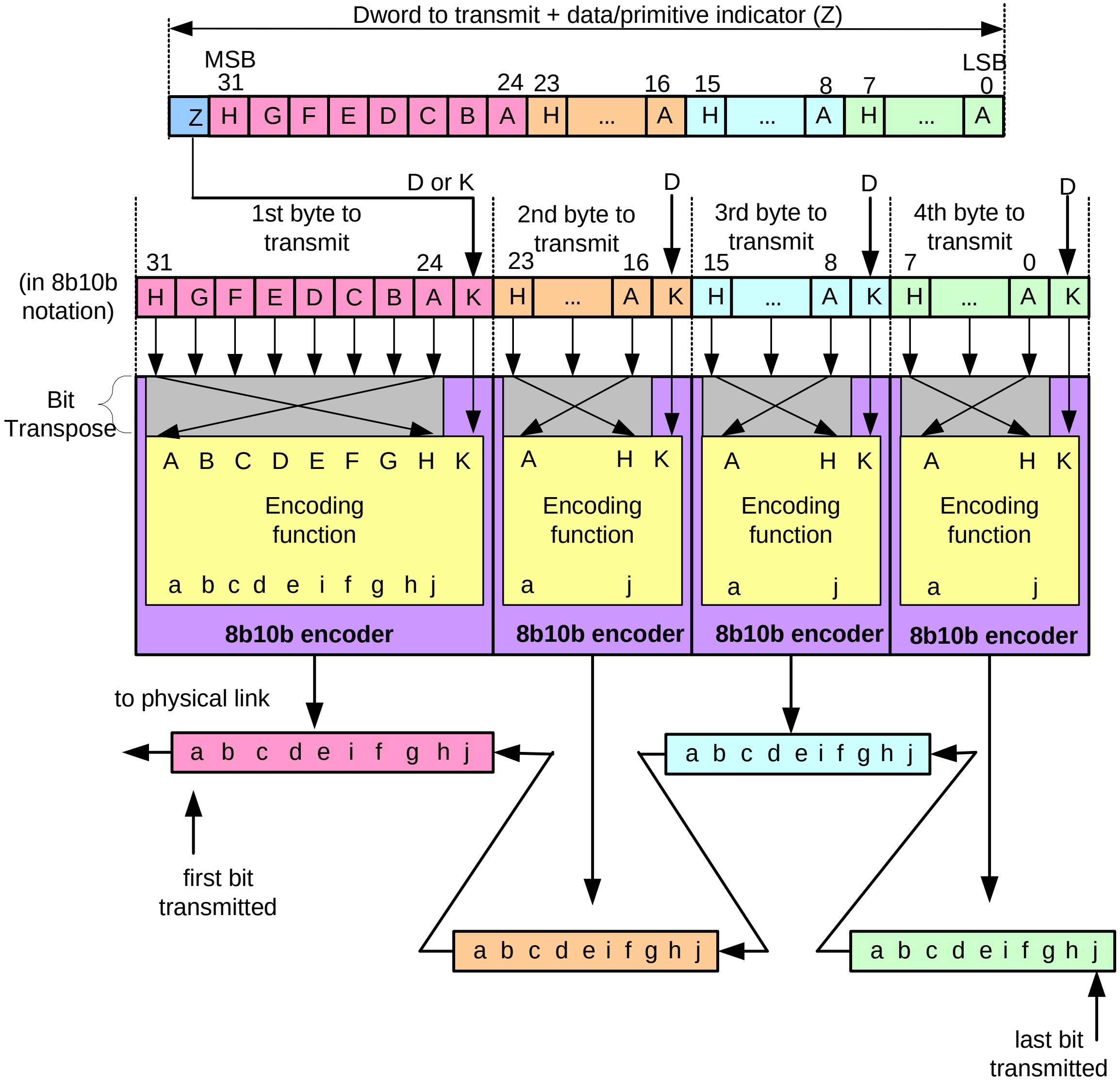
	or	or	or
bad	bad	bad	bad

- Any dword with a character with wrong disparity

	or	or	or
bad+-	bad+-	bad+-	bad+-

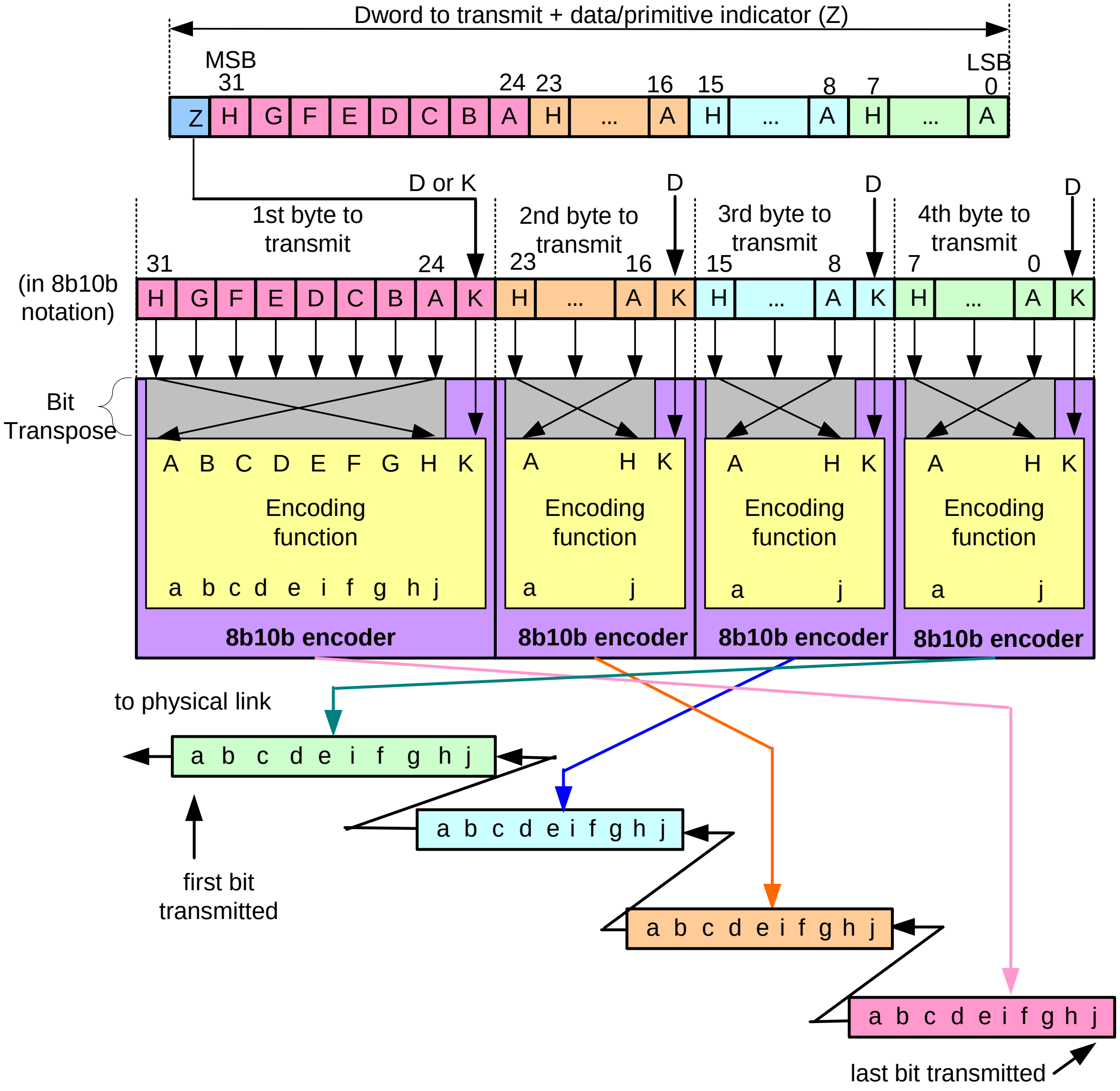
SAS transmit bit order

- A dword is a multibyte quantity, so endianness matters
- Most significant bit (MSB) is shown as bit 31 on the left
- Least significant bit (LSB) is shown as bit 0 on the right
- Selected 8b10b encoder input/output order effectively flips the bits
  - Requires CRC to also flip bits later to preserve coverage of burst errors crossing character boundaries
  - Following Fibre Channel mistake



SATA transmit bit order

- Most significant bit (MSB) is shown as bit 31 on the left
  - Same as SAS
- Least significant bit (LSB) is shown as bit 0 on the right
  - Same as SAS
- Byte containing bit 0 is transmitted first, not last
- For SAS STP, the SAS dword transmit bit order is used
  - conversion to little-endian done outside this logic (see link layer)

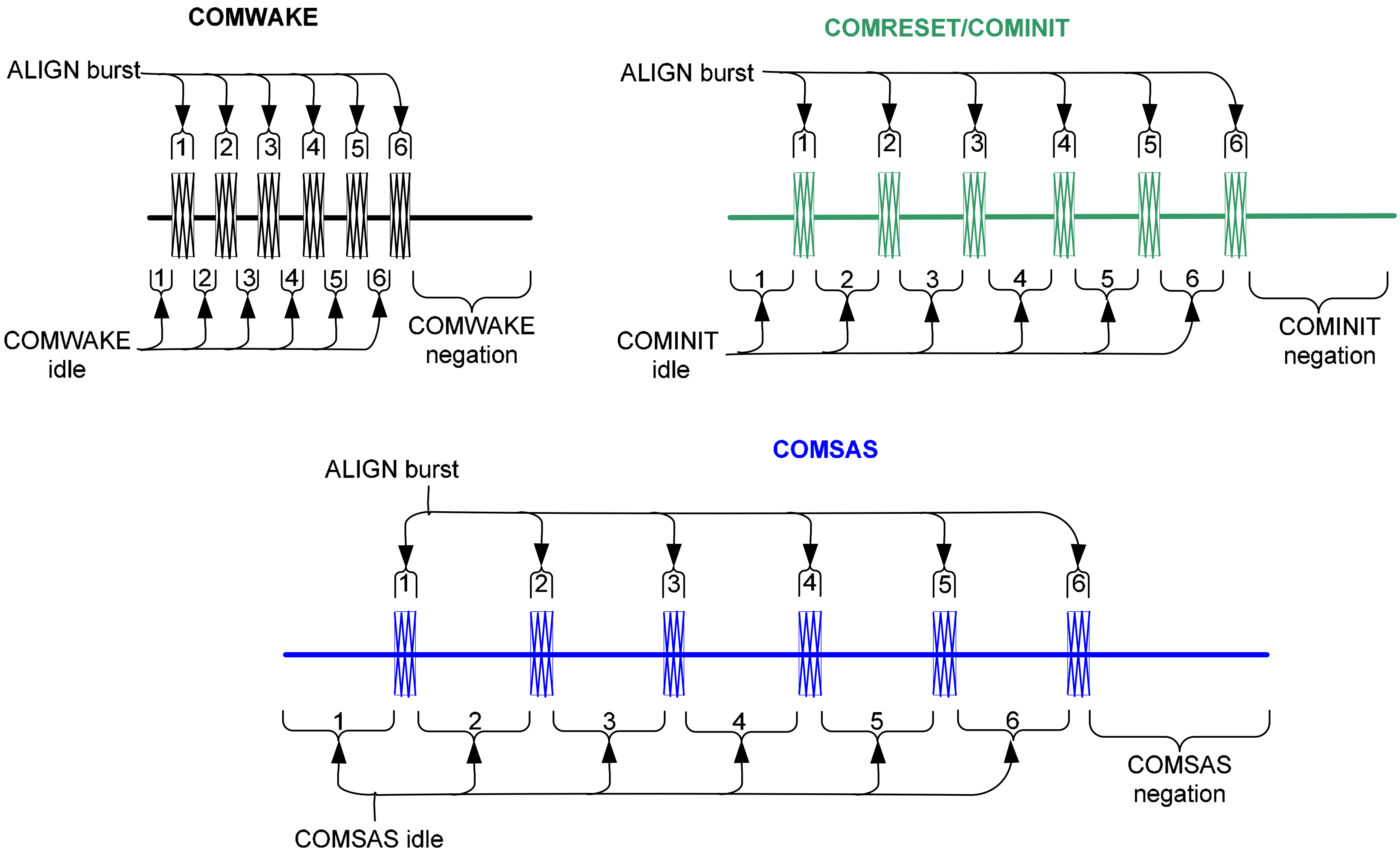


Phy layer – OOB signals

OOB signals

- OOB signal is a pattern of idle times and burst times
  - Idle time (and negation time)
    - Differential 0 V (Positive signal = negative signal)
    - No transitions (DC idle)
  - Burst time
    - Transmitted as a burst of ALIGN(0) primitives
    - Received as presence of edges (whether they are valid ALIGNs is irrelevant)
- Designed to be detectable by analog squelch detection logic
- Length of idle time distinguishes between OOB signals

Transmitting OOB signals

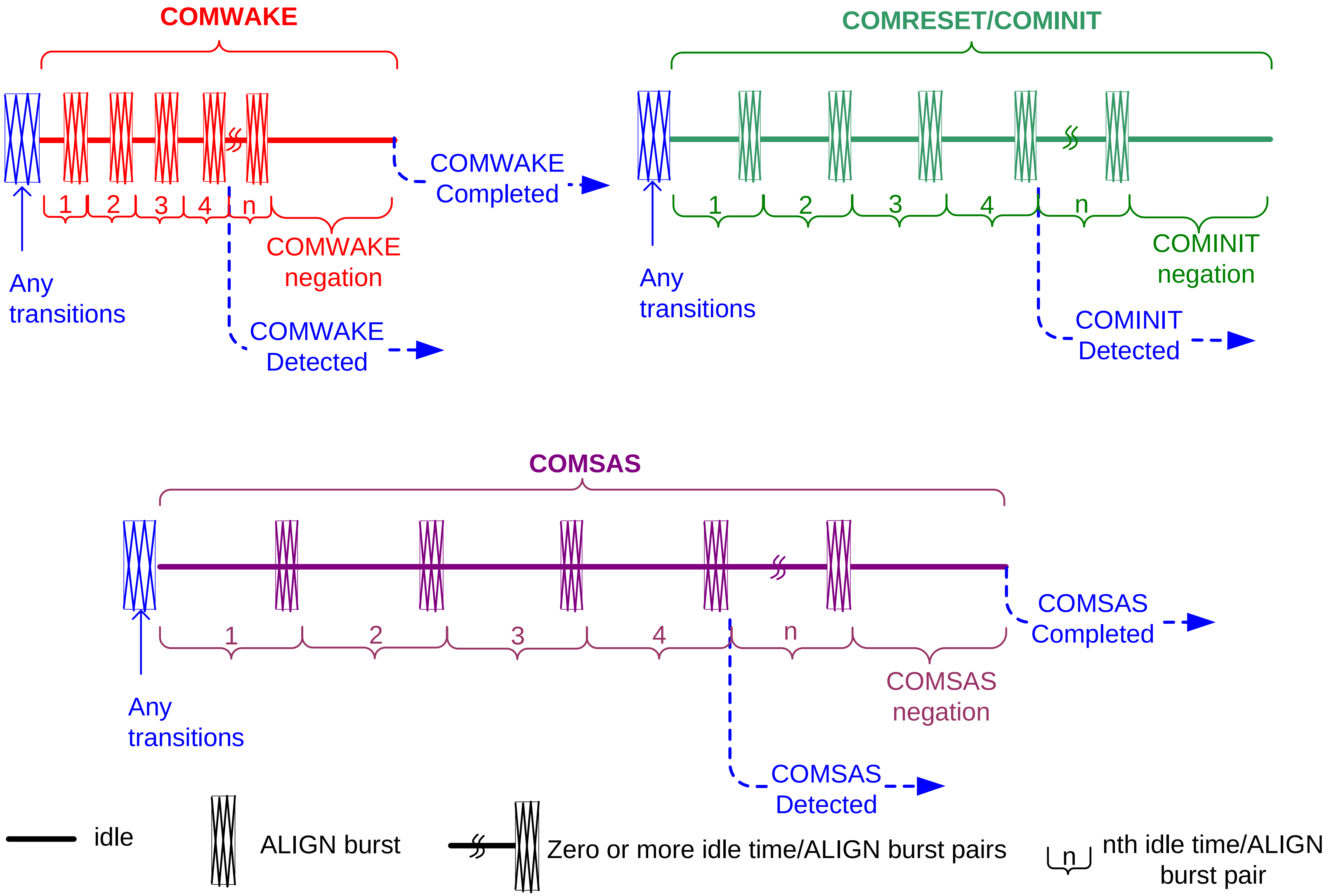


Transmit OOB signal timing

- Transmit six idle time/burst time pairs, then a negation time
- OOBI based times
- Unit Interval during OOB sequence
  - Looser clock tolerance than normal SAS unit interval
  - 666.600 to 666.734 ps

OOB Signal	Burst time	Idle time	Negation time
COMWAKE	160 OOBI (106.7 ns)	160 OOBI (106.7 ns)	280 OOBI (186.7 ns)
COMINIT/ COMRESET	160 OOBI (106.7 ns)	480 OOBI (320.0 ns)	800 OOBI (533.3 ns)
COMSAS	160 OOBI (106.7 ns)	1440 OOBI (960.0 ns)	2400 OOBI (1.6 µs)

Receiving OOB signals



COMWAKE

1 2 3 4 n

Any transitions

COMWAKE Completed

COMWAKE negation

COMWAKE Detected

COMRESET/COMINIT

1 2 3 4 n

Any transitions

COMINIT negation

COMINIT Detected

COMSAS

1 2 3 4 n

Any transitions

COMSAS Completed

COMSAS negation

COMSAS Detected

idle

ALIGN burst

Zero or more idle time/ALIGN burst pairs

n

nth idle time/ALIGN burst pair

Receive OOB signal timing

- Receive four idle time/burst time pairs
- Don't detect the same OOB signal until receiving a negation time
- Burst times irrelevant (100 ns minimum mentioned in the standard)
- Signals differentiated by idle times

Negation times

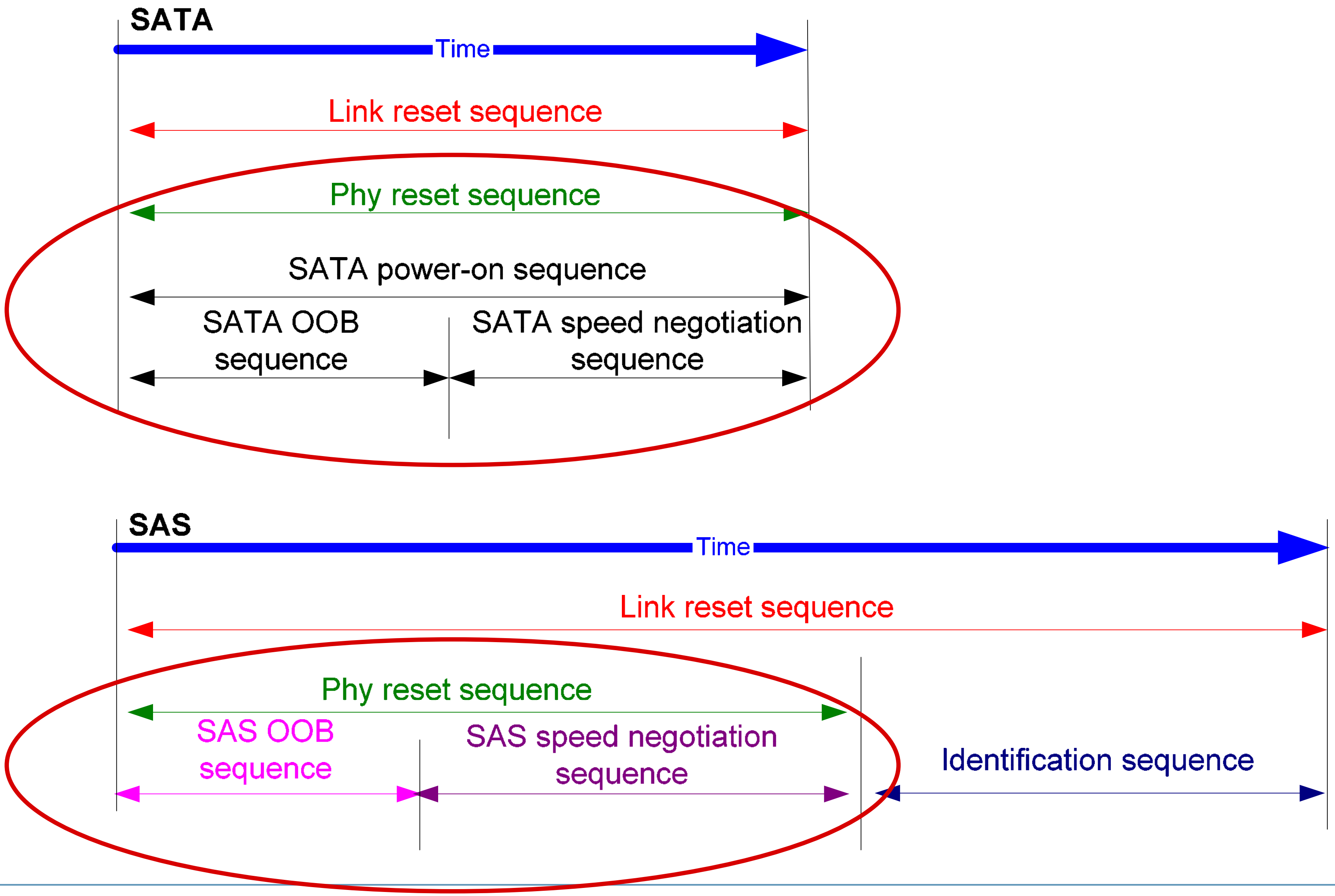
OOB Signal	Shall detect
COMWAKE	> 175 ns
COMINIT/ COMRESET	> 525 ns
COMSAS	> 1575 ns

Idle times

OOB Signal	May detect	Shall detect	Shall not detect
COMWAKE	55 to 175 ns	101.3 to 112 ns	< 55 or > 175 ns
COMINIT/ COMRESET	175 to 525 ns	304 to 336 ns	< 175 or > 525 ns
COMSAS	525 to 1575 ns	911.7 to 1008 ns	< 525 or > 1575 ns

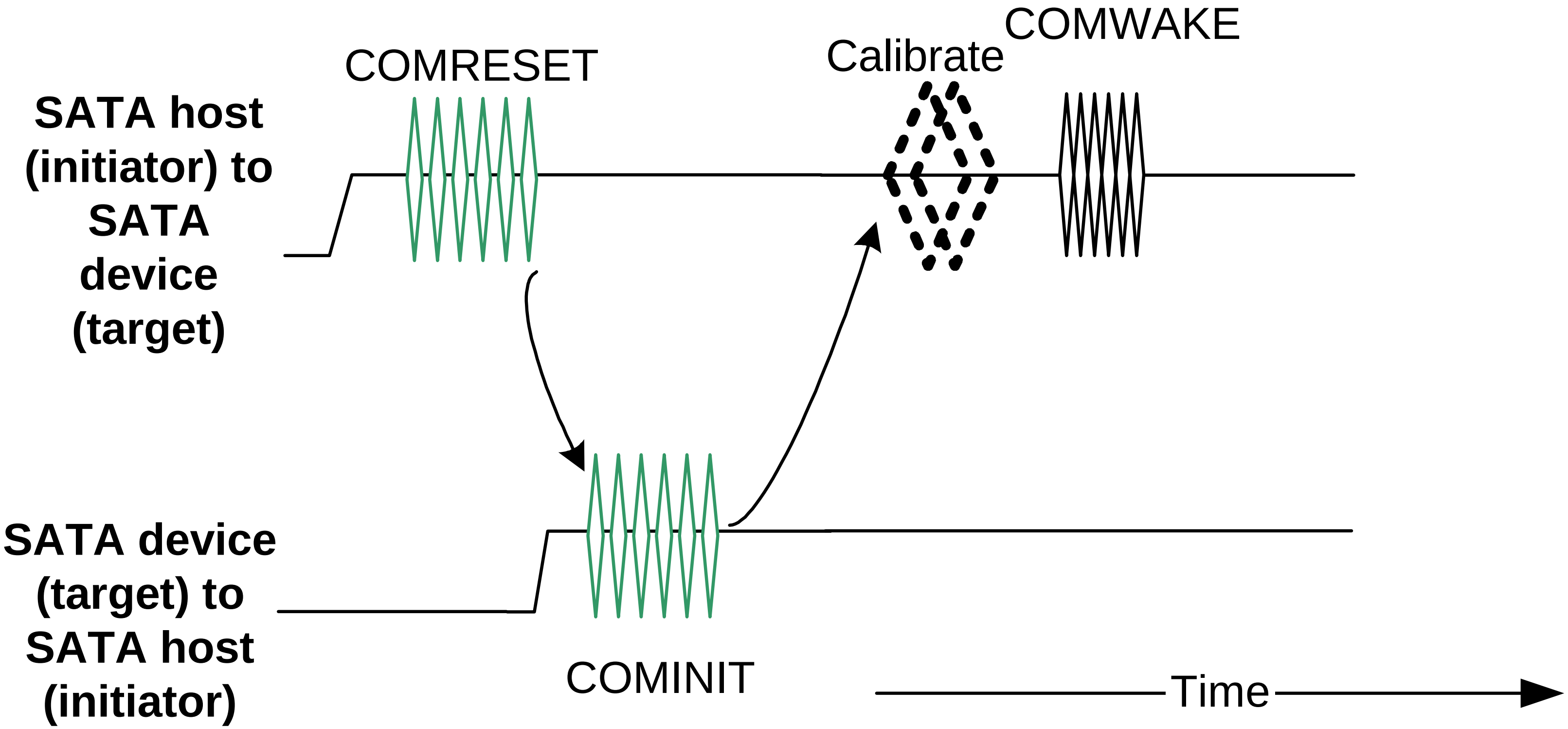
Phy layer – Phy reset sequences

Phy reset sequences



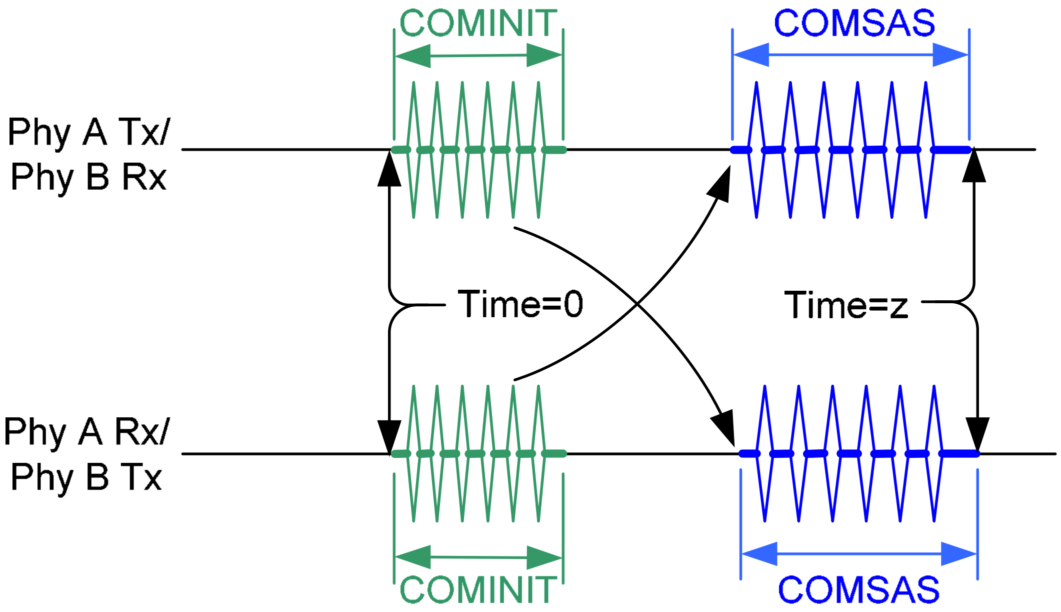
SATA OOB sequence

- SATA
  - Transmit and receive COMINIT
  - Host then sends COMWAKE


COMRESET
Calibrate
COMWAKE
SATA host (initiator) to SATA device (target)
SATA device (target) to SATA host (initiator)
COMINIT
Time

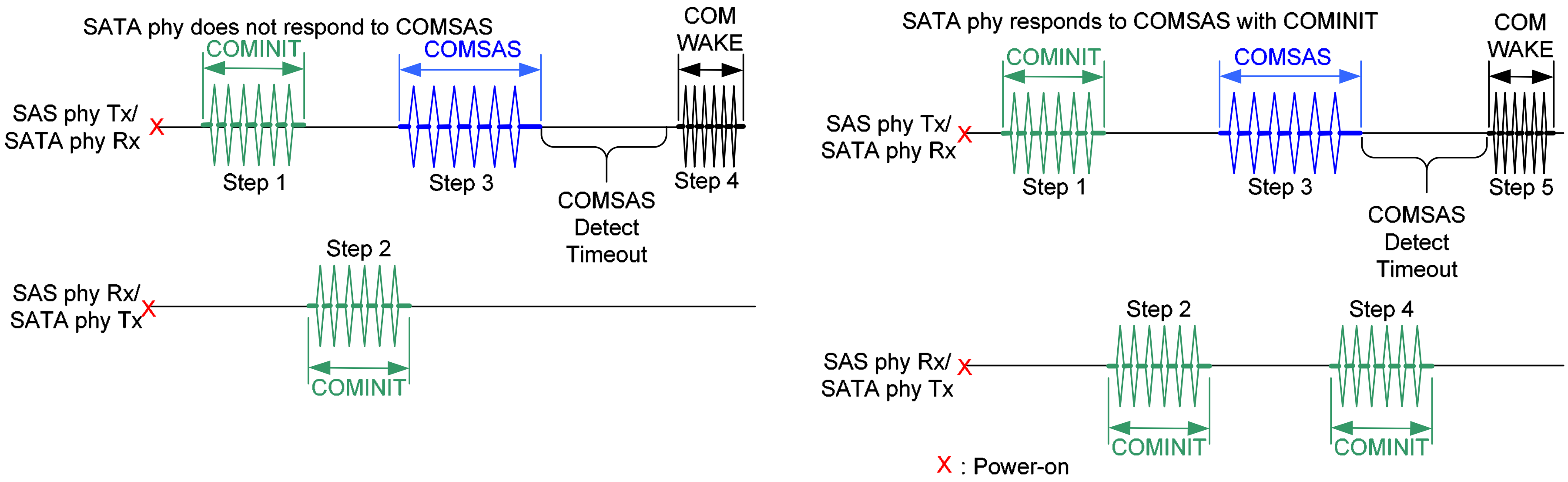
SAS OOB sequence

- Transmit and receive COMINIT
- Transmit COMSAS
  - If COMSAS received, physical link is SAS to SAS
  - If COMSAS is not received, physical link is SAS to SATA
- There are a few hot-plug situations where COMINIT leads directly to COMSAS; this is allowed



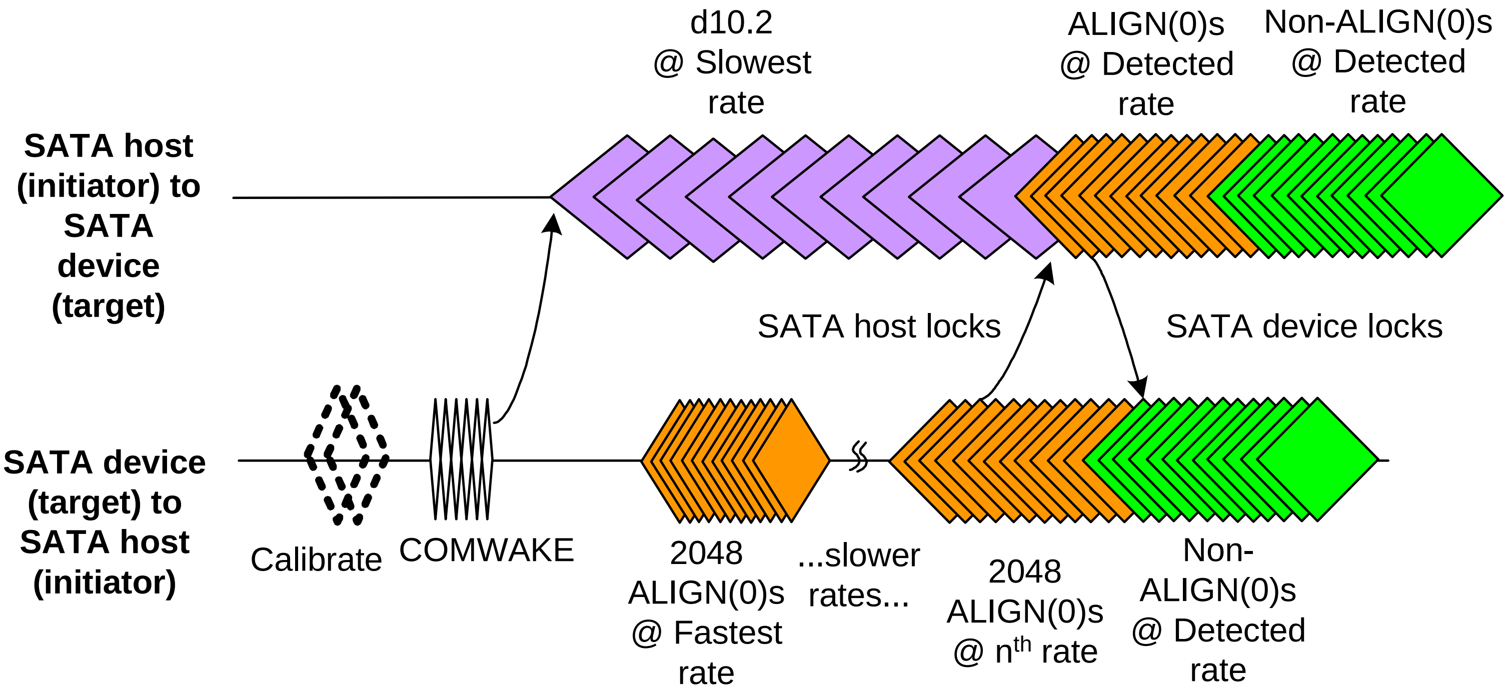
SAS to SATA OOB sequence

- If COMSAS is not received, physical link is SAS to SATA
- SAS OOB sequence morphs into the SATA OOB sequence
- Since the SAS device drives COMWAKE, it looks like a SATA host
- Only supports SATA device not SATA host



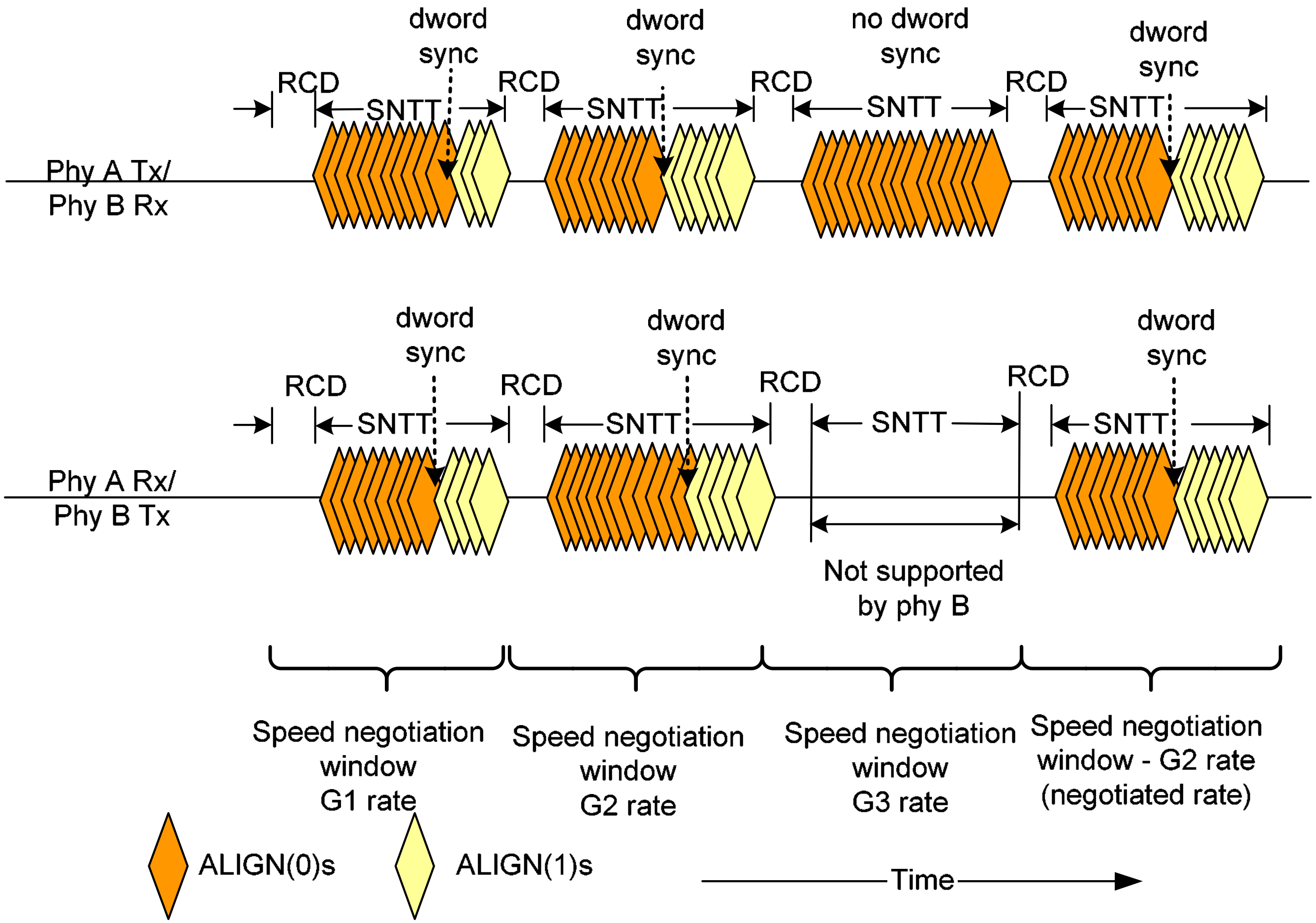
SATA speed negotiation sequence

- SATA device transmits COMWAKE
- Fast-to-slow algorithm
  - SATA device transmits ALIGN(0)s at fastest rate
  - looks for SATA host to reply with ALIGN(0)s at that same rate
  - After transmitting 2048, SATA device tries next slowest rate



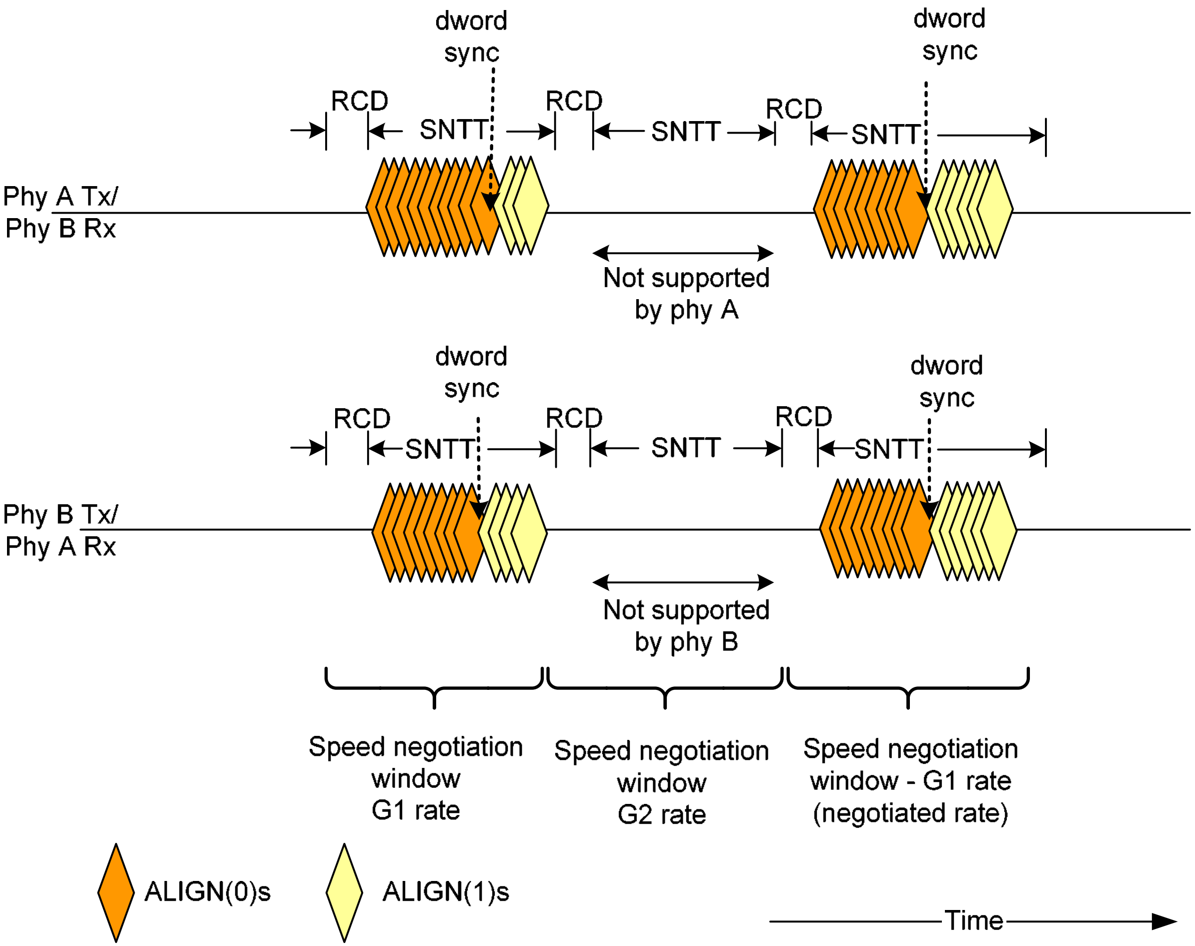
SAS speed negotiation sequence

- Slow-to-fast
- Both phys run same set of speed negotiation windows
- 1.5, then 3.0, then 6.0 (if needed), etc. until they find:
  - a supported rate; then
  - a (faster) non-supported rate
- Last window returns to the highest supported rate detected



SAS speed negotiation sequence 2

- Phy changes from ALIGN(0) to ALIGN(1) when it receives incoming ALIGN(0)s
- A window succeeds if ALIGN(1)s are both being transmitted and received, fails otherwise
- Phy transmits DC idle when it doesn't support a rate



SAS speed negotiation times

- SAS speed negotiation takes much longer than SATA
  - SATA exits as soon as lock is achieved
  - SAS waits through each window
  - SAS is symmetrical (no host/device roles)
  - 1.8 ms for 1.5 Gbps; 2.4 ms for 3.0 Gbps

OOB Signal	Time	Derivation
Rate change delay time (RCDT)	750,000 OOBI (500.0 µs)	Huge time requested by LSI Logic
Speed negotiation transmit Time (SNTT)	163,840 OOBI (109.2 µs)	4096 dwords
Speed negotiation lock time (SNLT)	153,600 OOBI (102.4 µs)	3840 dwords
Speed negotiation window time	913,840 OOBI (609.2 µs)	RCDT + SNTT

Hot-plug timeout

- If no reply to COMINIT
  - SAS targets – should try only once
    - One attempt announces presence
    - Since targets don’t originate traffic on their own, they just wait to be spoken to
  - Expanders – shall try again within 500 ms
    - Initiators depend on expanders to detect new devices
  - SAS initiators – should try again
    - Initiator can probe whenever it wants
  - No earlier than 10 ms between attempts

Phy layer – State machines

SAS SP state machine

- Runs OOB sequence
  - Always probes with COMSAS; responds as SATA host if necessary
- Runs speed negotiation sequence
  - Supports both SATA and SAS speed negotiation
- Three sets of states
  - SPnn:OOB_x – OOB sequence
  - SPnn:SAS_xxx – SAS speed negotiation
  - SPnn:SATA_xxx – SATA speed negotiation
- Goal is to reach SP15:SAS_PHY_Ready or SP22:SATA_PHY_Ready
  - Link layer transmits and receives dwords
- Works with SP_DWS state machine to obtain dword synchronization

SAS SP_DWS state machine

- Started by SP state machine during ALIGN(0) search portion of speed negotiation
- Searches for a comma pattern
  - appears in K28.5
  - K28.5 is the first character in an ALIGN (0)
- Looks for 3 primitives
  - If a comma is found out of position, starts over
  - If 3 valid primitives are found, declares dword synchronization achieved
- Afterwards, watches for loss of dword synchronization
- If several bad dwords arrive, can declare dword synchronization lost and start searching anew
  - SP may rerun the phy reset sequence if errors persist

SATA Phy Initialization state machines

- Separate state machines for hosts and devices
- Host Phy Initialization state machine
  - HP1:HR_Reset – transmits COMRESET
  - HP4:HR_COMWAKE – transmits COMWAKE
  - HP6:HR_AwaitAlign – transmits D10.2s
  - HP7:HR_SendAlign – transmits ALIGNs
  - HP8:HR_Ready – link layer transmits and receives
- Device Phy Initialization state machine
  - DP1:DR_Reset – idle
  - DP2:DR_COMINIT – transmits COMINIT
  - DP5:DR_COMWAKE – transmits COMWAKE
  - DP6:DR_SendAlign – transmits ALIGN
  - DP7:DR_Ready – link layer transmits and receives

Wrap up

Serial Attached SCSI tutorials

- General overview (~2 hours)
- Detailed multi-part tutorial (~3 days to present):
  - Architecture
  - Physical layer
  - Phy layer
  - Link layer
    - Part 1) Primitives, address frames, connections
    - Part 2) Arbitration fairness, deadlocks and livelocks, rate matching, SSP, STP, and SMP frame transmission
  - Upper layers
    - Part 1) SCSI application and SSP transport layers
    - Part 2) ATA application and STP/SATA transport layers
    - Part 3) Management application and SMP transport layers, plus port layer
  - SAS SSP comparison with Fibre Channel FCP

Key SCSI standards

- Working drafts of SCSI standards are available on http://www.t10.org
- Published through http://www.incits.org
  - Serial Attached SCSI
  - SCSI Architecture Model – 3 (SAM-3)
  - SCSI Primary Commands – 3 (SPC-3)
  - SCSI Block Commands – 2 (SBC-2)
  - SCSI Stream Commands – 2 (SSC-2)
  - SCSI Enclosure Services – 2 (SES-2)
- SAS connector specifications are available on http://www.sffcommittee.org
  - SFF 8482 (internal backplane/drive)
  - SFF 8470 (external 4-wide)
  - SFF 8223, 8224, 8225 (2.5”, 3.5”, 5.25” form factors)
  - SFF 8484 (internal 4-wide)

Key ATA standards

- Working drafts of ATA standards are available on http://www.t13.org
  - Serial ATA 1.0a (output of private WG)
  - ATA/ATAPI-7 Volume 1 (architecture and commands)
  - ATA/ATAPI-7 Volume 3 (Serial ATA standard)
- Serial ATA II specifications are available on http://www.t10.org and http://www.serialata.org
  - Serial ATA II: Extensions to Serial ATA 1.0
  - Serial ATA II: Port Multiplier
  - Serial ATA II: Port Selector
  - Serial ATA II: Cables and Connectors Volume 1

For more information

- International Committee for Information Technology Standards
  - http://www.incits.org
- T10 (SCSI standards)
  - http://www.t10.org
  - Latest SAS working draft
  - T10 reflector for developers
- T13 (ATA standards)
  - http://www.t13.org
  - T13 reflector for developers
- T11 (Fibre Channel standards)
  - http://www.t11.org
- SFF (connectors)
  - http://www.sffcommittee.org
- SCSI Trade Association
  - http://www.scsita.org
- Serial ATA Working Group
  - http://www.serialata.org
- SNIA (Storage Networking Industry Association)
  - http://www.snia.org
- Industry news
  - http://www.infostor.com
  - http://www.byteandswitch.com
  - http://www.wwpi.com
  - http://searchstorage.com
- Training
  - http://www.knowledgetek.com

hp®

invent