Beyond SAS-2

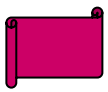
Presented to the SCSI Trade Association
(Posted as 08-077r0 on T10)

by Rob Elliott, HP ISS Server Storage Advanced Technology
11 January 2008

Features of interest

Feature	When needed
Save power	2008
Define denser connectors	2009-2010
Support active cables	2009-2010
Double the speed to 12 Gbps	2010-2011
Enhance the protocol	2010-2011
Split the standard	?

For a 15 minute presentation, skip after each slide labeled (summary) to the next section. For a 2 hour presentation, go through each slide.



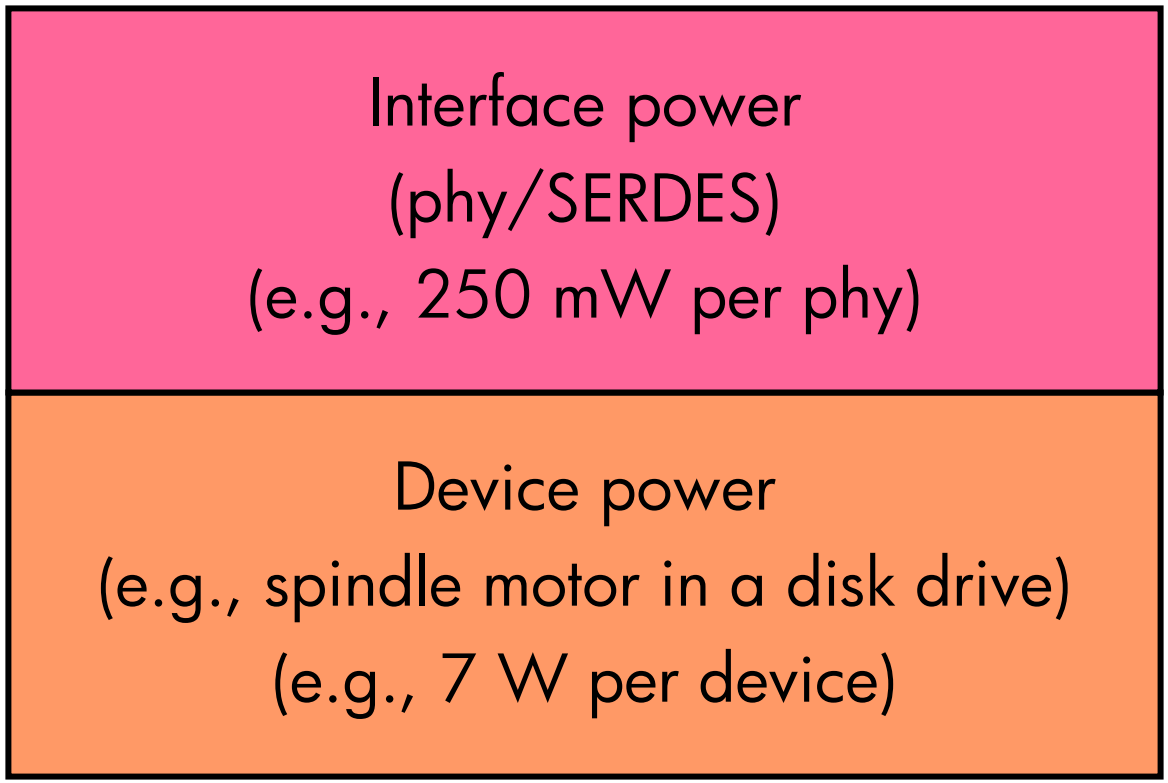
signifies a reference is available (see Reference slides at the end)

Baseline

- Best disk drive interconnect
  - For both SAS & SATA drives
- Good small fabric interconnect
  - Blade servers, adjacent racks
- High bandwidth
  - Direct-attached storage (DAS)
- Low cost
- High reliability
  - Bit Error Ratio of 10^{-15} or better

Save power

Two places to save power



Save power (summary)

- Saving interface power
  - Turn off phy when not needed
  - SATA interface power management
    - SAS initiators and expanders need to support this for attached SATA devices
  - SAS interface power management
    - Borrow from SATA, PCI Express, and Ethernet
- Saving device power
  - Adopt ATA device features

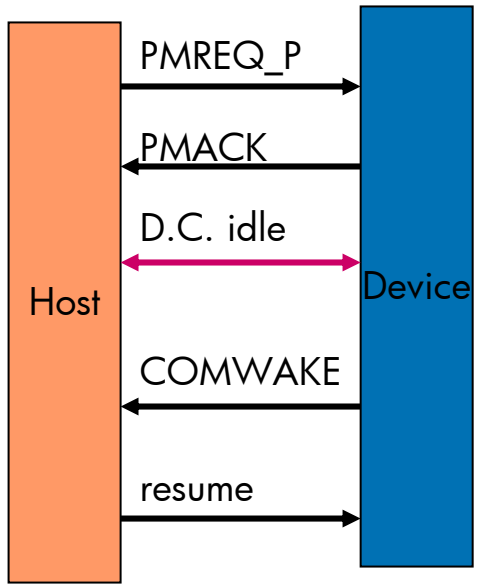
Saving interface power

- If phy is just transmitting idle dwords, turn it off
  - Turn it on when there is something interesting to send
- Other interfaces offer interface power management:

Interface	Power management feature
Serial ATA	Partial/slumber states
PCI Express	L0s state
Ethernet	Energy Efficient Ethernet IEEE 802.3az task force (underway)

SATA interface power management

- SAS initiators/expanders need to support SATA interface power management
  - SATA summary
    - Either phy requests entry with PMREQ_P or PMREQ_S
      - Partial (PMREQ_P): 10 μs wakeup latency
      - Slumber (PMREQ_S): 10 ms wakeup latency
    - Recipient allows/denies with PMACK and PMNAK
    - Either phy sends COMWAKE to wake up
  - Define SMP functions to control an expander
    - Accept or deny power management requests?
    - How aggressively to originate requests



SAS interface power management

- Borrow from SATA
  - Are two states overkill?
  - Could use COMINIT instead of COMWAKE
    - SAS doesn't consider COMINIT to be a "hard reset" like SATA
- Borrow from Energy Efficient Ethernet (and PCI Express)
  - Each direction is independent
  - Phy stops transmitting when it has no frames to send
  - Phy restarts transmitting when it has frames to send
  - Receiver PLL keeps running while interface is off
    - Faster recovery time
    - Periodically send a training pattern so the receiver PLL can keep running
    - Spread-spectrum clocking (SSC) increases recovery time

Saving device power

Feature	Description
Implement IDLE and STANDBY states	Already defined in SCSI and ATA
Add ATA SLEEP state	Wakeup on COMINIT START STOP UNIT mode 5h (like in T10 RBC) Interface falls asleep in this state too
Add ATA Advanced Power Management (APM)	1-byte value instructs device how aggressively to save power
Add low-rpm idle state	Drive spins slowly (not stopped) and retracts heads
Define multi-rpm operational modes	Perform media accesses at different rates (e.g. 15K while busy, 7.5K while not so busy)
More SAS spinup controls	Rather than specify “spin-up now”, specify how much power the device is allowed to use
Coordinate Multiple LUs	Enhance RAID controller power management

Define denser connectors

8-wide Mini SAS connectors

- Define 8-wide Mini SAS connectors
  - Already in SFF-8086, 8087, and 8088
  - PCI Express already defines an 8-wide version of its external cable

Connector	Description
Mini SAS 8i (internal)	Two SGPIO busses (16 total sidebands) •Hybrid with two (Mini) SAS 4i connectors on one end Stacked vs. wide?
Mini SAS 8x (external)	Add sidebands •Report cable length for passive cables •Support active cables •Provide adequate GROUNDs

Support active cables

Interest in active cables

- Some interest in SAS cables longer than 10 m
  - Match old 25 m parallel SCSI HVD cables
  - Optical cables have some pros and cons
- Don't define an "optical SAS" standard
  - Provide power to the connector interface
  - Let the cable assembly figure out how to get the signals to the other side
    - Optical or electrical
  - Avoid special optical SKUs
  - Options: QSFP, Active Mini SAS, or something new

Support active cables (summary)

- Quad Small Form Factor Pluggable (QSFP)
- InfiniBand active cables precedent
- Mini SAS active cables
  - Modify existing connector: preferred
  - Define derivative connector: not preferred
- Protocol caution for long cables

QSFP – Quad Small Form-factor Pluggable

- Similar to the Mini SAS 4x connector
  - 38 pins for 4 channels rather than 26 pins
  - Problem: Not plug compatible
- Problem: Uses InfiniBand-style pin assignments
  - Rx on one end, Tx on the other end
  - Differs from SATA connector, SAS Drive connector, and Mini SAS connectors
  - Differs from typical ASIC pinouts (where each SERDES colocates Tx/Rx)
  - SAS doesn't necessarily need to use the same pinout as others
- Being used by Fibre Channel, Ethernet, SONET/SDH, InfiniBand
- Developed by the QSFA Multisource Agreement consortium
  - http://www.qsfpmsa.net
- Published by SFF Committee as INF-8438
  - http://www.sffcommittee.org
  - As an INF document, no patent disclosures or RAND terms offerred
  - New SFF project underway that will follow normal rules

InfiniBand active cables

- InfiniBand added 12V/3V power to SFF-8470 pinout
  - Take over some GROUND pins
  - Compatible with original cables
- Many vendors of 5 Gbps active cables
  - Intel Connects (optical) up to 100 m
  - Zarlink ZLynx (optical) up to 100 m
  - Gore Extended Reach Cable Assemblies (copper) up to 25 m
  - EMCORE Media Converter (optical) up to 150 m
  - Quellan
- Jim Nadolny and Stefaan Sercu (FCI), Michael Kravets and Atul Gupta (Gennum). "Active Cable Assemblies for 10 Gigabit Ethernet." DesignCon 2003
  - http://portal.fciconnect.com/res/en/pdffiles/tlib/DC03_PaperActiveEyemax_FINAL.pdf

Intel® Connects Cables

Mini SAS active cables

- SAS could do the same for the Mini SAS connector
- Preferred: Add power to the Mini SAS 4x connector
  - Reuse some ground pins
    - Approach taken by InfiniBand
    - Circuitry must only provide power if active-to-active attachment is detected (no power to ground shorts!)
    - Reduces signal return quality/increases crosstalk
    - Maintains interoperability with existing cables
- Not preferred: define a new Mini SAS 4x connector with extra pins
    - Maintains current signal return quality (no loss of ground pins)
    - Loses interoperability with existing connectors/cables
    - Might as well use QSFP

Protocol caution for long cables

- SAS and SATA protocols assume short cables
  - Interlocked frame transmission
    - COMMAND, TASK, RESPONSE frames wait for an ACK before proceeding
  - 1024 byte frame size
    - Receivers typically advertise a low number of R_RDY credits
  - SATA flow control
    - Must respond to HOLD with HOLDA within 20 dwords
    - Receiver can add buffers to accommodate longer lengths, but won't know what an active cable needs
  - 25-100 m probably OK; longer is questionable

Double the speed to 12 Gbps

Double the speed (summary)

- 1200 Megabytes/second bandwidth
- Support same interconnects as SAS-2
- Follow PCI Express 3.0 if possible
- Support SATA 4.0
- 10 Gbps (e.g., with 64b66b) vs. 12 Gbps (with 8b10b)
  - Prefer 12 Gbps
- Receiver improvements
- Transmitter improvements
- General improvements
  - Forward Error Correction: avoid

1200 Megabytes/second bandwidth

- Exact multiple of 150, 300, and 600 Megabytes/sec
- Simplifies rate matching and multiplexing
  - Rate matching would need to insert different ALIGN rates than ½, ¾, etc.
  - Multiplexing would have to be account for extra ALIGNs
    - ALIGNs are considered inside the logical links today

Support same interconnects as SAS-2

- SAS-3 phys must work over every SAS-2 compliant interconnect
  - External cables: up to 10 m
  - Same channel models as used for SAS-2
    - including: HP01-14; HP24-28; 0.5, 1, 3, 6, 10 m cable
- SAS-2 phys must work over every SAS-3 compliant interconnect
  - At 6 Gbps
- Remember that SAS-1.1 phys might not work at 3 Gbps over all SAS-2 interconnects
  - Goal raised from 6m to 10 m

Avoid reinvention

- Should follow PCI Express 3.0 if possible
  - Want to continue to share SERDES technology as much as possible
  - PCIe is abandoning 8b10b, though
    - Allows slower rate: 8 Gbps rather than 10 Gbps
    - We may not want to do this
- Must be compatible with SATA 4.0
  - Not defined yet
  - Assume SATA 4.0 devices will continue to focus on 1 m internal cables
    - Low transmitter amplitudes (< 700 mV)
    - Limited transmitter deemphasis
    - Simpler receivers (probably no DFE)

10 vs. 12 Gbps

Rate Coding	UI Overhead	Notes
12 Gbps 8b10b	83.33 ps 25%	Tighter timing budget Higher SERDES power consumption Rate precedent: OIF-CEI-11G-LR (9.95-11.1 Gbps)
10 Gbps 64b66b or other	100 ps 3.125%	Redesign dword synchronization (no comma patterns) Reencode primitives (K28.5/K28.3 don't exist) Redesign SATA tunneling D.C. balance issues Detects fewer errors (disparity, invalid characters) Each real bit error creates 3 coded bit errors Rate precedent: 10GBASE-KR (10 Gigabit Ethernet over Backplane IEEE 802.3ap) (10.3125 Gbps), 10GFC, PCI Express 3.0 (8 Gbps)

- Prefer 12 Gbps with 8b10b if at all possible

Receiver improvements

Feature	Description
More DFE taps	Doubling speed means doubling the number of taps (to counter the same impulse response effects) • 3-tap at 6 Gbps, 6+ at 12 Gbps More complex adaptation algorithm than LMS
Adjust CDR to tolerate precursor ISI	Current DFE only accounts for post-cursor ISI
Edge equalization	Focus on edges, not center of eye
Maximum Likelihood Sequence Estimation (e.g., Viterbi Algorithm)	Similar to DFE, but separate boost for each unique previous n-bit pattern

Transmitter improvements

Feature	Description
More post-cursor deemphasis taps	6 Gbps just uses a single tap
Add precursor de-emphasis	May interfere with DFE/CDR in receiver, though
Provide feedback to transmitter to tune better to the channel	Difficult to add to protocol Hard to create an algorithm that doesn't oscillate with both sides tuning themselves Recommendation: avoid if possible
Let transmitter tune itself	Perform TDR (time domain reflectometry) test during startup (e.g. SATA calibration phase) Adjust parameters based on information from its own receiver
Have cable assembly report length	T10 proposal 05-100r0

General improvements 1

Feature	Description
Multi-level signaling (e.g., PAM-4)	Transfer 2 bits at a time (4 signal levels, 3 eyes) 9.5 dB SNR penalty (PAM-4) vs. less channel loss and NEXT noise (NRZ) Recommendation: stick with NRZ; reserve this for SAS-4 at 24 Gbps
Full duplex	Transmit and receive on the same wire Receiver cancels out the signal it knows its transmitter is transmitting used by 1/10 Gigabit Ethernet Recommendation: avoid
Lower the differential impedance	PCI Express 2.0 switched to 85 ohm for 5 Gbps Test equipment is all 100 ohm Recommendation: stay with 100 ohm

General improvements 2

Feature	Description
Near-end crosstalk (NEXT) cancellation	Used by 1/10 Gigabit Ethernet over Cat-5/6 cables May be used by Gigabit Ethernet over backplanes Recommendation: consider
Fractionally-spaced equalizers	Transmitter deemphasis or FFE receiver equalization at fractions of a UI Better counteracts the frequency response of the interconnect DFE only amplifies the y-axis, not the x-axis of an eye diagram; this stretches it from the sides Recommendation: consider

Forward Error Correction

- Burst Error Correcting Code (2112, 2080) as used in 10GBASE-KR
- Recommendation: avoid

Property	Impact
Overhead	2080 bits of payload (62 dwords) 32 bits of overhead (1 dword) Same overhead as 64b66 coding
Error handling	Corrects errors up to 11 bits Counteracts 64b66b degradation of CRC32 burst error detection "2 to 2.5 dB of coding gain"
Gates/memory	"15K gates + 33x(64+3+64) RAM"
Encoder latency	32 bits (1 dword)
Decoder latency	2211 to 4323 bits (69 to 135 dwords) 200 ns at 10 Gbps 2211 = 33 x (64 + 3) (with a 33 bit wide datapath) Painful for storage, where expander latency < 20 dwords

Enhance the protocol

Enhance the protocol (summary)

- Security
  - Authentication: maybe
  - Encryption: maybe
  - Secure zone management: yes
- Store-and-forward: avoid
- Other concepts
  - Continue multiplexing
  - Mandate transport layer retries
  - Solid state drive enhancements
  - Data compression

Security

Feature	Description
Authentication	Initiator-to-target (end-to-end) •Some sort of login Initiator-to-expander, target-to-expander •Cryptographically authenticated IDENTIFY address frames Borrow from FC-SP, IPsec, and IKEv2-SCSI Recommendation: maybe
Encryption	Must be end-to-end Encrypt frames or connections? Cover SSP, SMP, and STP Borrow from FC-SP, IPsec, and SCSI-ESP Recommendation: maybe
Secure zone management	Authenticated and encrypted zone manager password Recommendation: yes

Store-and-forward in expanders

- Multiplexing is simple but costly
  - Lots of gates in endpoints and expander crossbar switch
  - Fixed, not dynamic, scheme makes fabrics with multiple speed targets inefficient
- Could design a new a store-and-forward protocol for SSP
  - Don't try to store-and-forward SAS-2 traffic; just new traffic
  - New SSP frame header with full source and destination SAS addresses
  - Buffer-to-buffer and end-to-end credits like in FC
  - Could allow multicasting (instant RAID-0)
- Recommendation: stick with multiplexing

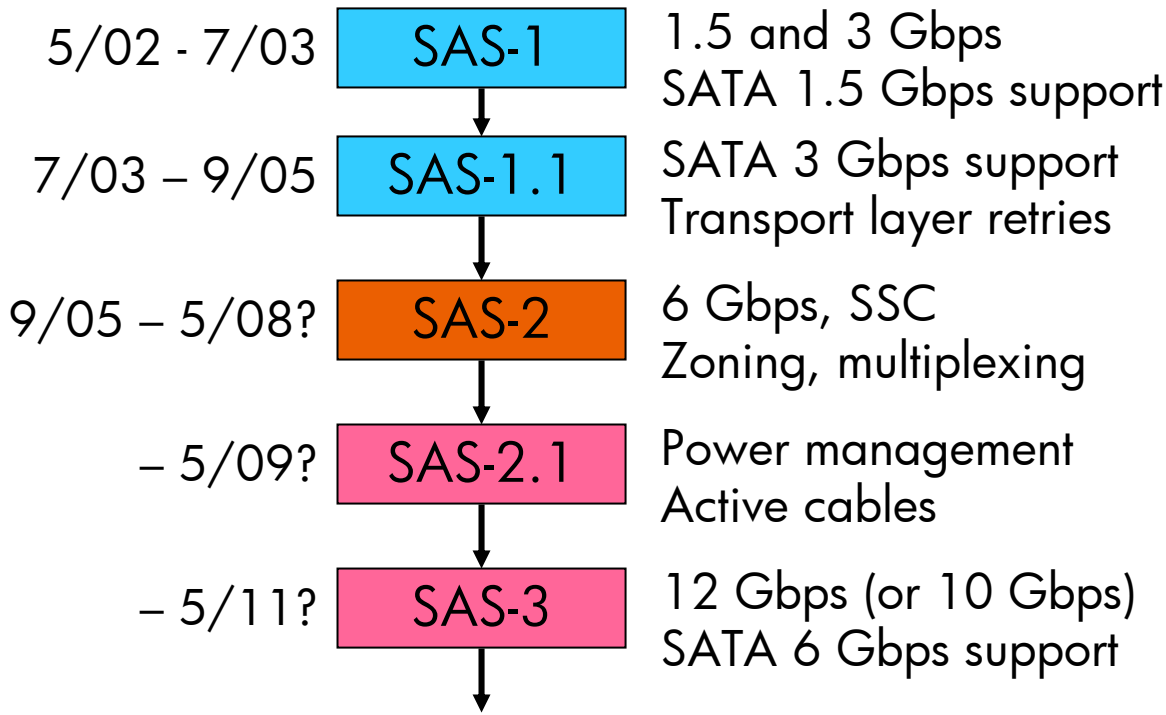
Other protocol concepts

Feature	Description
Continue 2:1 multiplexing	Assume disk drives will stay 6 Gbps until the next generation As SAS-3 rolls out, disk drives move to 6 Gbps As SAS-4 rolls out, disk drives move to 12 Gbps
Mandate Transport Layer Retries for all devices	Not just for tape drives Software error handling for block devices is not as robust as touted
Solid state drive enhancements	Optimize arbitration/connection management Application level identification/management of SSD
Data compression	Only beneficial if the data being transferred is not already compressed (or encrypted)

Split the standard?

Should we split physical and protocol documents?

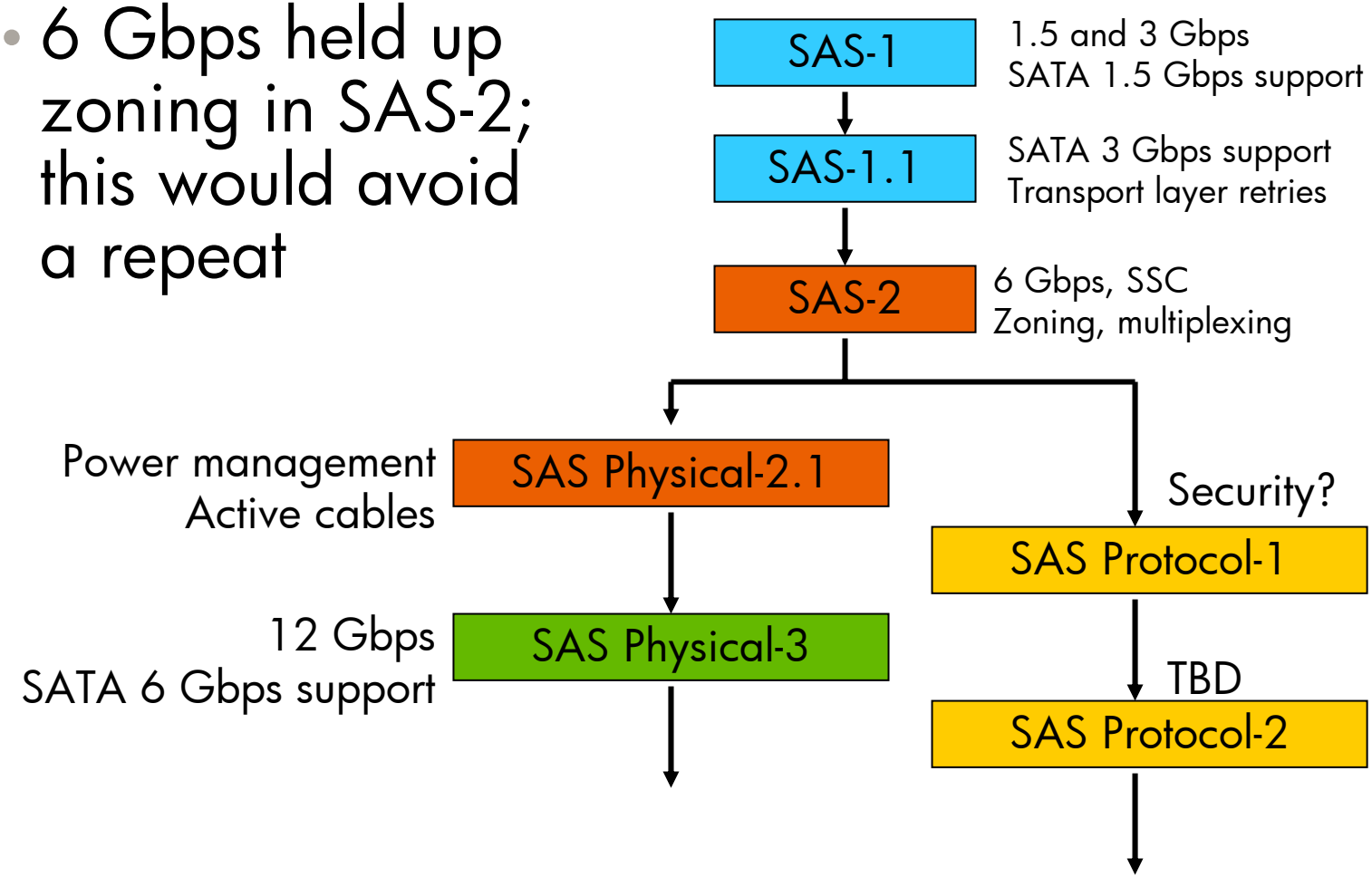
- Bug fixes have been included in each version
- Bug fixes for 6 Gbps and zoning will probably be needed before 12 Gbps



- Annoying for protocol to wait on physical

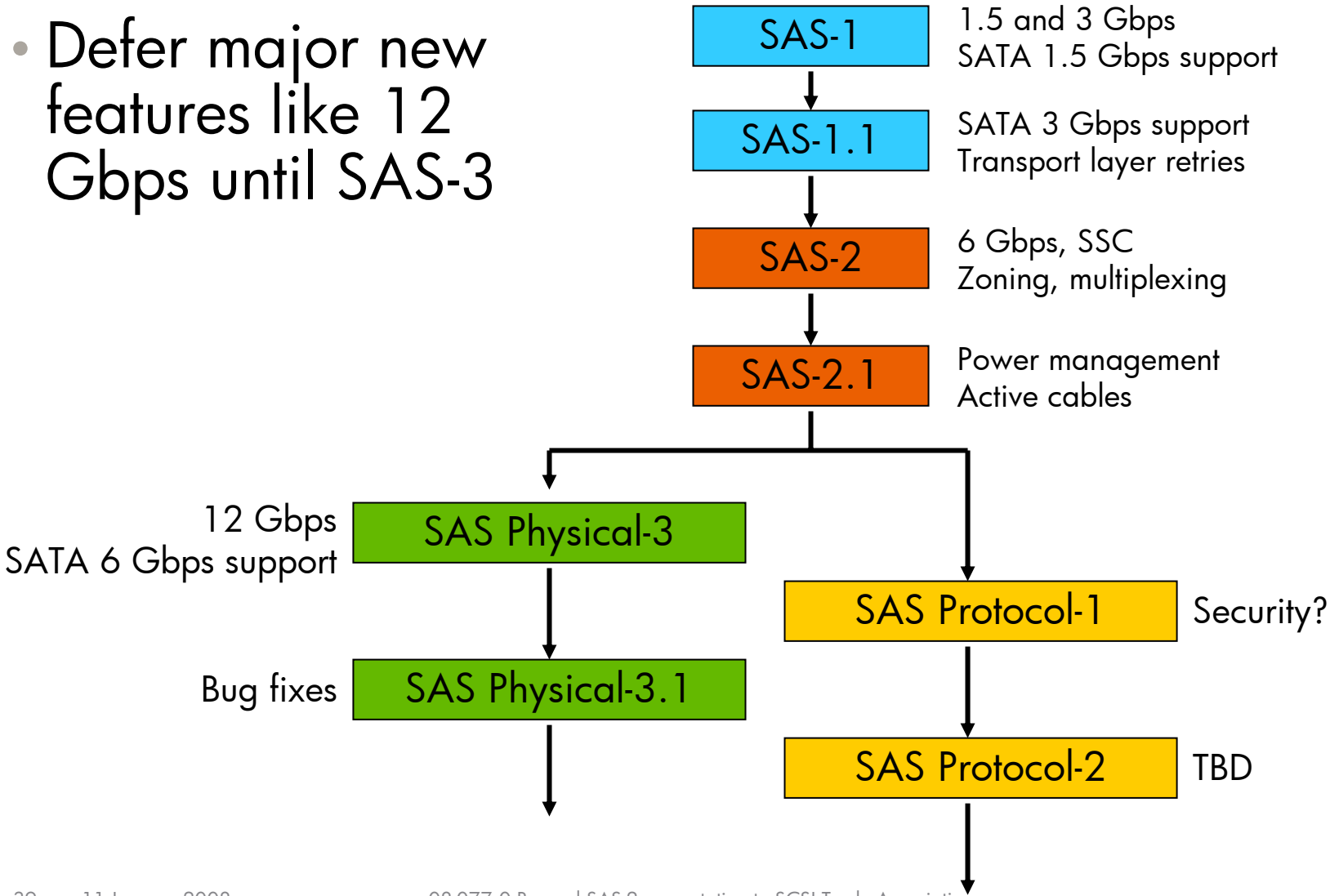
Could split after SAS-2

- 6 Gbps held up zoning in SAS-2; this would avoid a repeat



Could split after SAS-2.1

- Defer major new features like 12 Gbps until SAS-3



Opinion on splitting

- Don't split yet
- One document is easier
  - SATA ballooned into several specifications, yet reconverged
  - FrameMaker has no problem with the document size
- Target SAS-2.1 for mid-2009
  - Allow power management, active cables, bug fixes
  - Defer major new features like 12 Gbps and security into SAS-3 in mid-2011
- Reconsider splitting after SAS-2.1


富士山頂 〔吉田口登山道〕
Top of Mt.Fuji
(山梨県) (Yamanashi Pref.)
吉田口・富士浅間神社
Yoshidaguchi·Fuji-sengen Shrine
(山梨県) (Yamanashi Pref.)
河口湖口・富士スバルライン
Kawaguchikoguchi·Fuji-subaru Line
(Yamanashi Pref.)

References

Power management references

Topic	Reference
Energy Efficient Ethernet	www.ieee802.org/3/az/index.html
Serial ATA	www.sata-io.org
PCI Express	www.pcisig.org
ATA	www.t13.org
Add low-rpm idle state	Jim Wong (Hitachi). "Technology Innovation for Eco-Friendly HDDs." www.hitachigst.com
Multi-rpm drives	Sudhanva Gurumurthi. Dynamic RPM research. www.cs.virginia.edu/~gurumurthi

Coding scheme references

Topic	Reference
Coding schemes	Mel Belhadj (Cortina Systems). "Coding techniques for high-speed serial interconnect." Lightwave Magazine. http://lw.pennnet.com/articles/article_display.cfm?article_id=281518
DC balance	Howard Johnson (Signal Consulting Inc.). "Killer Packet." http://www.sigcon.com/Pubs/news/5_7.htm "DC Blocking Capacitor Value." http://www.sigcon.com/Pubs/news/7_09.htm
64b67b	Cisco, Cortina, SLE. "Interlaken Technology: New-Generation Packet Interconnect Protocol white paper." http://www.siliconlogic.com/pdfs/Interlaken_White_Paper-March_2007.pdf
8b10b and 64b66b	Wikipedia articles on "8b/10b" and "64b/66b." http://www.wikipedia.org

Transceiver improvement references 1

Topic	Reference
Adjust CDR to tolerate precursor ISI	J. Ren, H. Lee, Q. Lin, B. Leibowitz, et al. "Precursor ISI Reduction in High-Speed I/O." IEEE Symposium on VLSI Circuits Digest of Technical Paper, June 2007.
Edge equalization	Anthony Chan Carusone. "Edge Equalizer Adaption Algorithm to Reduce Jitter in Binary Receivers." http://www.eecg.utoronto.ca/~tcc/tcc06.pdf Brian Brunn, Xilinx. "Edge-Optimized Equalization Extends Performance in Multi-Gigabit Serial Signaling." DesignCon 2006
Maximum Likelihood Sequence Estimation (e.g., Viterbi Algorithm)	Gu, Y., Le-Ngoc, T., and Cheng, S. "Adaptive decision feedback equalization with MLSE based on predicted signals." IEEE International Conference on Communications 1993. Werner Rosenkranz and Chunmin Xia. "Electrical equalization for advanced optical communication systems." AEU - International Journal of Electronics and Communications Volume 61, Issue 3, 1 March 2007, Pages 153-157. MathWorks paper on linear equalizer, DFE, and MLSE. http://www.mathworks.com/products/communications/demos.html?file=/products/demos/shipping/comm/eqberdemo.html

Transceiver improvement references 2

Topic	Reference
Pre-cursor deemphasis	J. Ren, H. Lee, Q. Lin, B. Leibowitz, E-H. Chen, D. Oh, F. Lambrecht, V. Stojanovic, C.-K.K. Yang, J. Zerbe. "Precursor ISI Reduction in High-Speed I/O." IEEE Symposium on VLSI Circuits Digest of Technical Paper, June 2007
Lower differential impedance	Jan DeGeest, Dana Bergey, Stefaan Sercu (FCI) and John Lynch/Dennis Miller (Intel). "Improving System Performance by Reducing System Impedance to 85 ohms." DesignCon 2007 http://portal.fciconnect.com/res/en/pdffiles/7-TA4-2_DesignCon_2007_final.pdf
Near-end Crosstalk cancelation	Yin, et al. "Equalization and NEXT noise cancellation for 20 Gbps PAM-4 backplane serial IO interconnections." IEEE Transactions no Microwave Theory and Techniques Vol 53 No 1 January 2005.
Forward Error Correction	Andre Szczepanek (TI), Ilango Ganga (Intel), Cathy Liu (LSI), Magesh Valliappan (Broadcom). "10GBASE-KR FEC Tutorial." http://www.ieee802.org/802_tutorials/july06/10GBASE-KR_FEC_Tutorial_1407.pdf


hp
invent