Future of PCI-X

Michael Krause

Fellow Engineer, Hewlett-Packard

Agenda

- What is PCI-X 2.0?
- “Need for Speed”
- Industry State of I/O
- PCI-X 2.0 Enablement Efforts

What is PCI-X?

- "PCI-X is high-performance backward compatible PCI for the future"
  - ✓ PCI-X uses the same PCI architecture
  - ✓ PCI-X leverages the same base protocols as PCI
  - ✓ PCI-X leverages the same firmware / BIOS as PCI
    - PCI-X supports the new PCI-SIG Firmware specification
  - ✓ PCI-X uses the same connector as PCI
    - "Any adapter, any slot, any time"
  - ✓ PCI-X and PCI products are interoperable
  - ✓ PCI-X uses same software driver models as PCI
- PCI-X is faster PCI
  - ✓ PCI-X 533 is up to 32x faster than the original version of PCI
  - ✓ PCI-X protocol is more efficient than conventional PCI

PCI-X 2.0 Naming

- "2.0" is the revision of the specification
  - ✓ Sometimes (erroneously) used to mean the new speeds
- PCI-X 1.0 Official Names
  - ✓ PCI-X 66
  - ✓ PCI-X 133
- PCI-X 2.0 Official Names
  - ✓ PCI-X 66
  - ✓ PCI-X 133
  - ✓ PCI-X 266
  - ✓ PCI-X 533

  } Mode 1

  } Mode 2

PCI-X 1.0 had 2 speed grades

PCI-X 2.0 has 4 speed grades

- Industry already executing PCI-X 2.0 across complete range of design points enabling solutions to scale to any level to meet customer needs
  - ✓ PCI-X 2.0 mode 1 compatible with PCI-X 1.0b
    - – Any PCI / PCI-X capable server is able to support PCI-X 2.0 without changing one line of software or requiring a hardware upgrade
  - ✓ PCI-SIG Compliance program has been testing PCI-X 2.0 mode 1 since specification completion over two years ago
- Industry beginning to execute PCI-X 2.0 mode 2 across select design points based on customer requirements
  - ✓ Recent PCI-SIG Compliance workshop included mode 2 capable platform
  - ✓ Multiple IHV will be bringing mode 2 devices out in the coming months

Agenda

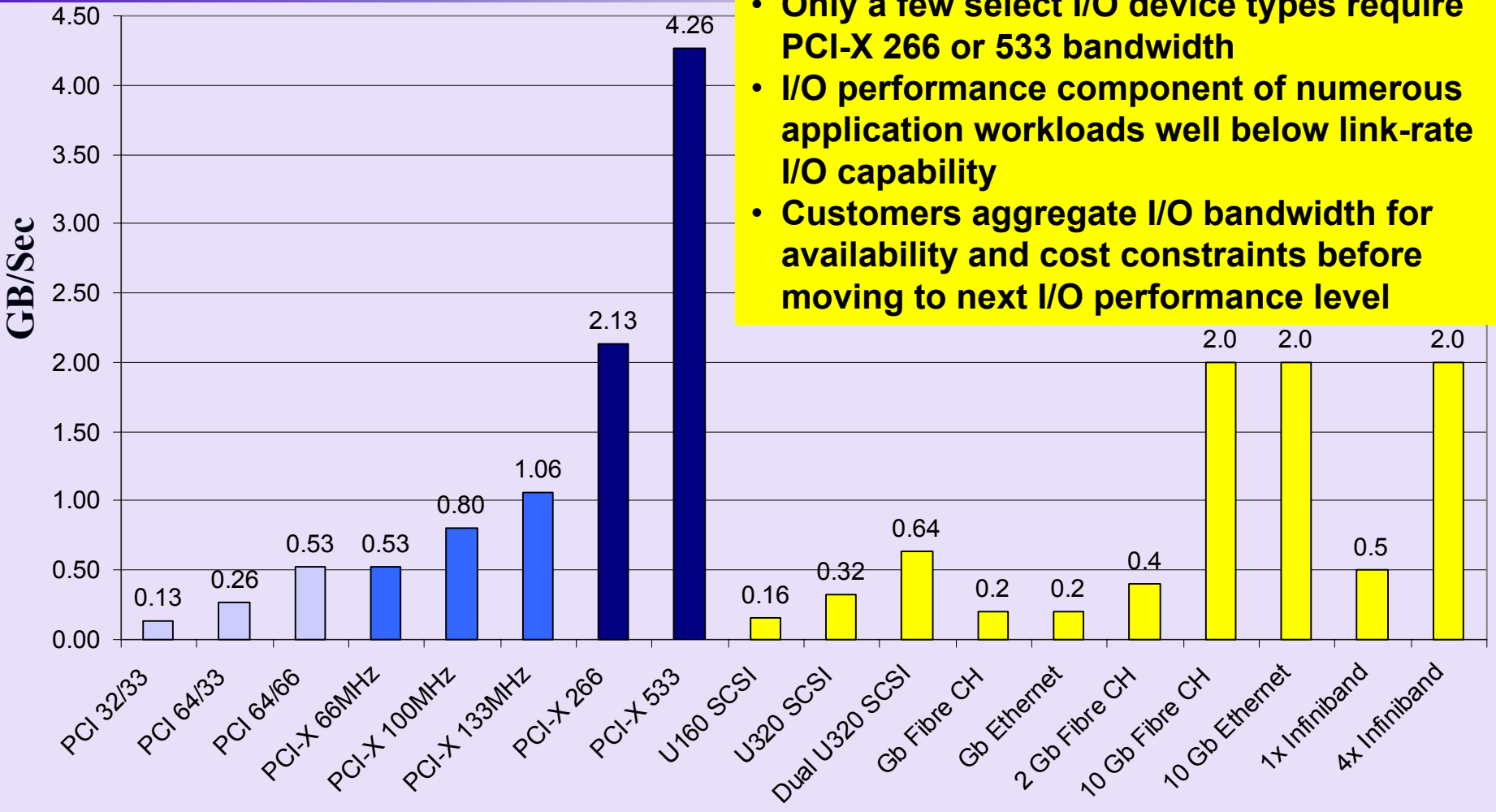
- What is PCI-X 2.0?
- “Need for Speed”
- Industry State of I/O
- PCI-X 2.0 Enablement Efforts

Is the Need for Speed Real?

- Real-world adoption of new high-speed I/O devices lags marketing “hype”
  - ✓ Customers focused on value not just performance
    - Each performance boost impacts more than just I/O device
    - Customers examining entire ecosystem and total cost of ownership
      - Cost of optics, cabling, switch infrastructures, management, storage controller performance, etc. all must be taken into account
  - ✓ Customers unwilling to pay any price forcing some technologies to take “baby steps” rather than “leaps”
    - Fibre Channel opting to migrate from 2 Gbps to 4 Gbps rather than 10 Gbps to maintain interoperability
      - Could see 8 Gbps if physical interoperability continues to be valued higher than raw I/O bandwidth
    - 2.5 Gbps Ethernet proposed as “free” bandwidth upgrade while maintaining existing cable and backplane infrastructure
    - Increased number of multi-port I/O devices on the market to provide transparent fail-over and aggregate I/O bandwidth at lower cost

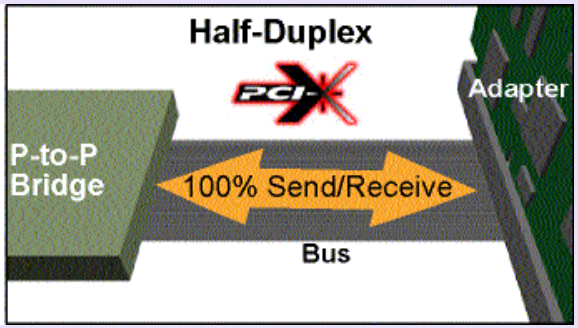
PCI-X 2.0 Bandwidth Span vs. Target Applications

- Only a few select I/O device types require PCI-X 266 or 533 bandwidth
- I/O performance component of numerous application workloads well below link-rate I/O capability
- Customers aggregate I/O bandwidth for availability and cost constraints before moving to next I/O performance level



PCI-X 266 / 533 have sufficient bandwidth to meet any I/O device requirements

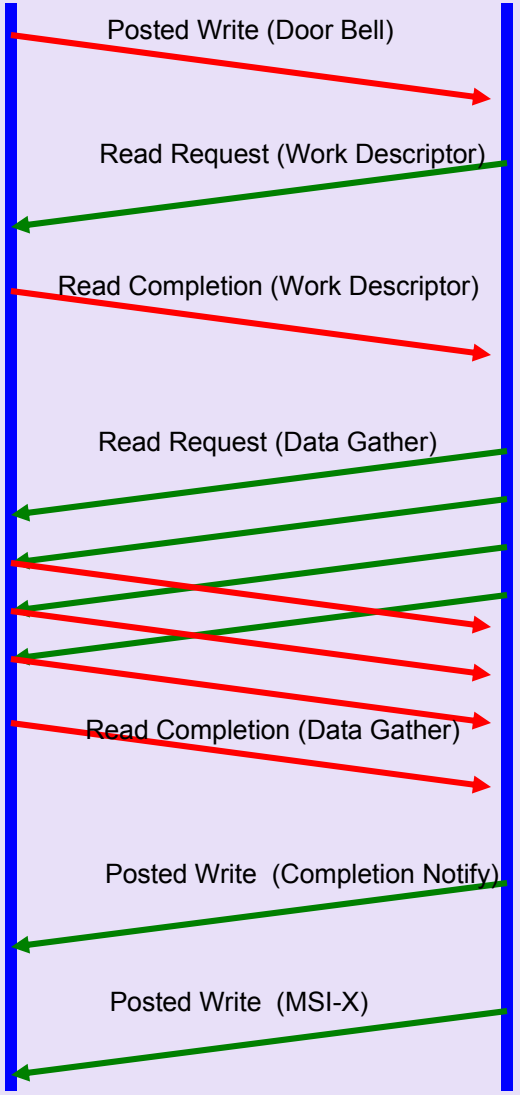
Importance of Push-Rate BW



- For servers and storage controllers, majority of workload is asymmetric in nature which can take advantage of the full bandwidth provided PCI-X
  - ✓ For example, network workload analysis shows the following:
    - 80 / 20 of inbound vs. outbound traffic
    - 80 / 20 inbound packets are small, i.e. under 256 Bytes
    - 80 / 20 outbound packets are large, i.e. approximately 1024 Bytes
  - ✓ For example, storage workloads analysis shows the following:
    - Major workloads such as OLTP, data warehouse, decision support, network or SAN back-up, etc. all move large blocks of data using a minimum of control messages
      - For efficiency, data movements range from 8KB to 1 MB in size
    - Control packets typically under 64 Bytes
    - Data typically moved using large scatter-gather lists for efficient placement while the device “chunks” storage block into full wire level units of transfer
      - For example, Fibre Channel segments / reassembles messages into 2 KB frames

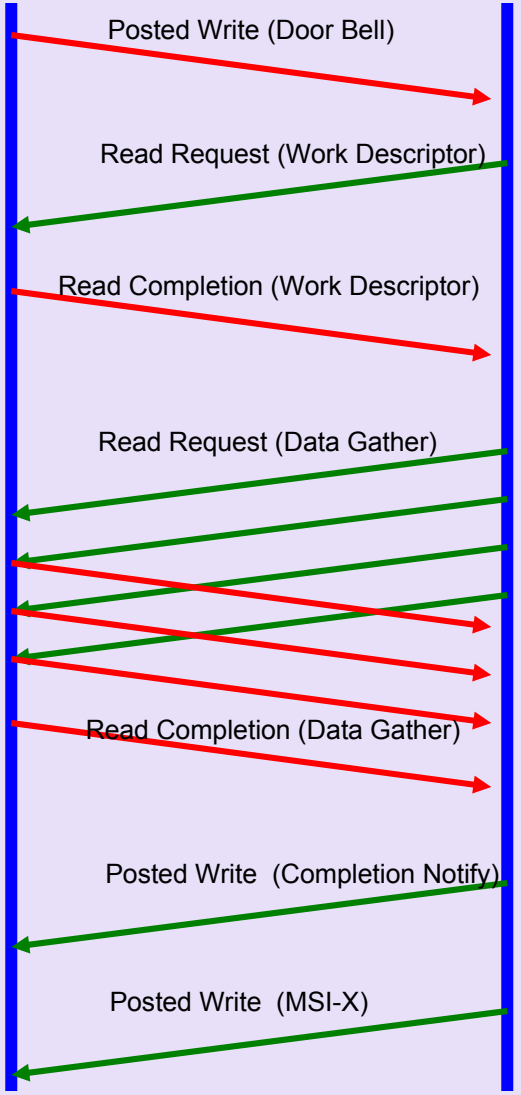
Performance Requirements (1)

- Each I/O operation = multiple I/O transactions
  - ✓ Each transaction consumes local I/O bandwidth
    - – Function of technology protocol efficiency as to how much bandwidth is available per I/O device operation
    - – PCI-X protocol efficiency varies between 40-96% as a function of transaction size and max transaction size
  - ✓ "Future proofing" is more marketing than reality
    - – System balance is the key to delivering a credible solution and must be designed in from the start at the system level – it is relatively independent of which PCI technology is used
  - ✓ Problem is in the design / implementation of the I/O subsystem and I/O device (including its driver paradigm) not just a matter of which local I/O is used
- I/O Latency to Memory Impact
  - ✓ Memory bandwidth is the gating factor in I/O performance
    - – Memory bandwidth increases perhaps 10% per year
  - ✓ If limited concurrent transactions, I/O latency to memory will have negative impact on delivered performance
    - – Potential 25-50% negative impact on device bandwidth



Performance Requirements (2)

- Concurrency
  - ✓ I/O taking same path as processors / memory
    - – Application transaction rate increasing faster than I/O operation rate
      - Time to process operation gets smaller with each new device bandwidth increase
    - – System chipsets and I/O devices must increase the number of concurrent I/O operations supported
      - Need to move beyond typical “single” operation to multiple concurrent operations
    - – Variety of methods to utilize in various components
      - Support multiple independently ordered / executed work descriptor queues
      - Support MSI-X to segregate completion traffic processing across multiple processors
      - Increase the number of outstanding split transactions to improve “pre-fetch” opportunities while avoiding “pipeline” execution stalls in the memory / I/O controllers and I/O device



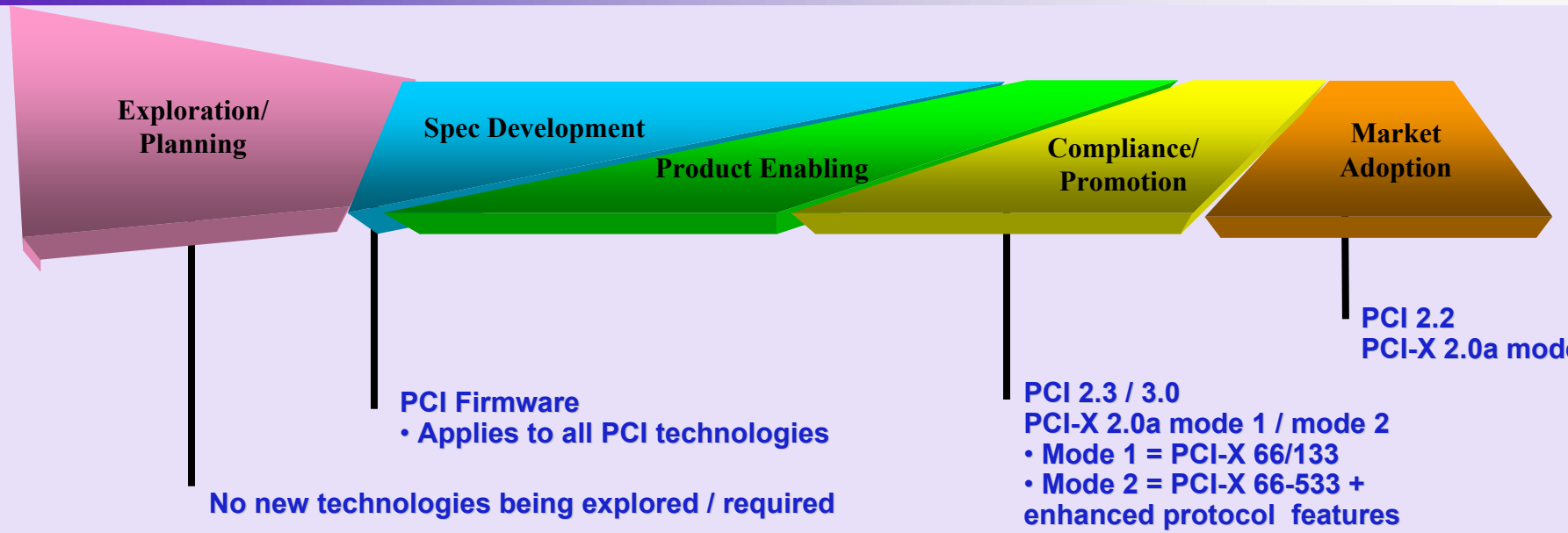
Agenda

- What is PCI-X 2.0?
- “Need for Speed”
- Industry State of I/O
- PCI-X 2.0 Enablement Efforts

I/O Technology Trends

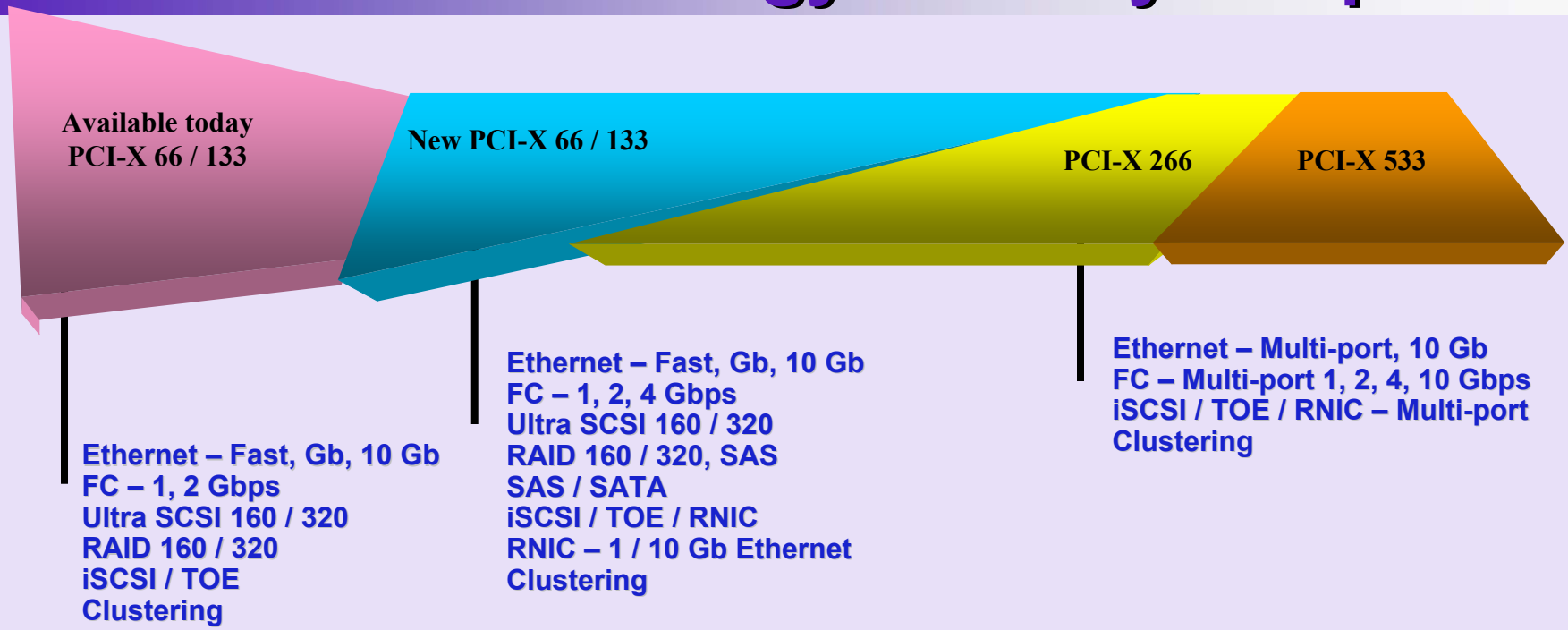
- Economics driving technology consolidation
  - ✓ Industry driving to:
    - Simplify solution delivery while reducing time-to-market
    - Constrain cost structures as technologies move to commodity
    - Improve “out-of-box” customer experience
    - Seek next level of innovation / value-add
    - Constrain cost structures with innovation / value-add
  - ✓ 99% of server and storage controller solutions constructed from just six basic high-volume, industry-standard I/O device types:
    1. Local storage: U160/320 SCSI, SAS, SATA
    2. RAID: U160/320 SCSI transitioning to SAS / SATA
    3. 1 / 2 / 4 / 10 Gbps Fibre Channel SAN
    4. 1 / 2.5 / 10 Gbps iSCSI Ethernet SAN
    5. 10 / 100 / 1000 / 10000 Ethernet Networking
    6. Cluster interconnects: E.g. InfiniBand, RNIC (RDMA / TOE)

State of PCI / PCI-X Technology



- Conventional PCI and PCI-X specifications are solid / stable
- Complete ecosystem – full product suite available for server solutions
  - ✓ IHVs have already executed 100's of designs through PCI-X 133
    - New I/O functionality continues to be delivered on PCI-X 66 / 133
  - ✓ IHVs executing PCI-X 266 where additional bandwidth is required

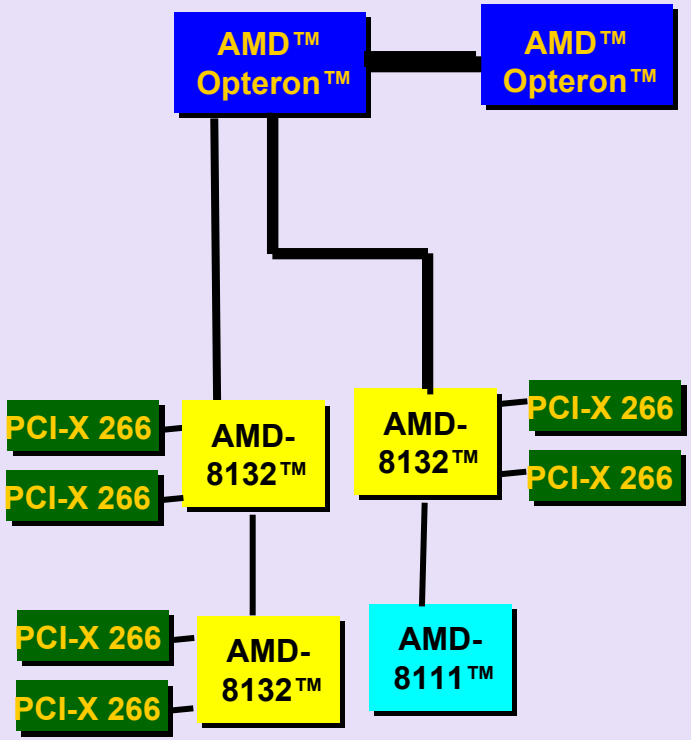
PCI-X Technology Industry Adoption



- Single-port and multi-port I/O devices exist over wide range of I/O device types providing maximum customer choice in how solutions are configured
  - Many devices support multiple functions, e.g. a multi-port Ethernet, Fibre Channel, SAS, etc. controllers. Multi-function devices reduce slot pressure and when combined with point-to-point attachment, deliver optimal performance
- With on-going I/O device type consolidation, the number of high-speed I/O devices needed to deliver a solid customer-driven solution is relatively small

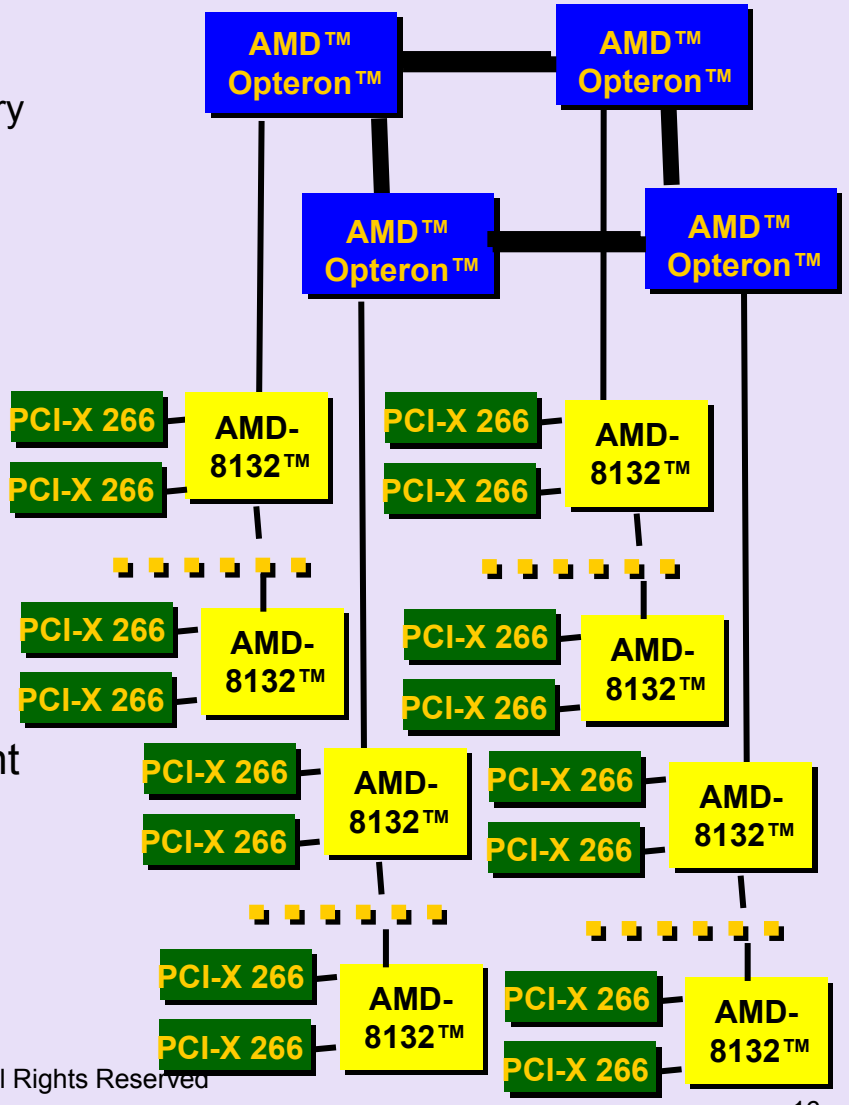
PCI-X 66-266 2P Server Example

- Highlights:
  - ✓ Integrated memory controller on each CPU
    - Aggregate I/O bandwidth can scale with memory bandwidth
    - High-speed access to memory via HyperTransport™ enables multiple I/O devices to operate at full-speed per link.
  - ✓ HyperTransport™ interconnect technology
    - High-speed access to memory via HyperTransport™ enables multiple I/O devices to operate at full-speed per link.
    - Up to 3 links per CPU at 6.4GBps gross bandwidth available for each IO chain
    - Devices can be chained
  - ✓ Point-to-point I/O attach
    - Complete error containment and fault isolation
    - Complete performance isolation
    - Hot-plug supported
  - ✓ Supports wide mix of PCI-X device attachment to deliver balanced server performance
    - 530 MBytes/s to 2.13 GBytes/s bandwidth per device
  - ✓ Allows system scalability based number of PCI-X interfaces desired



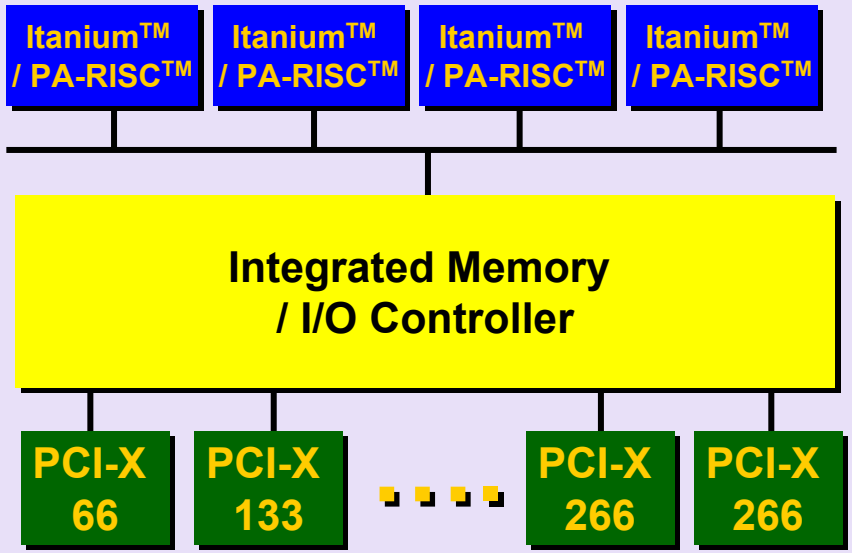
PCI-X 66-266 4P Server Example

- Highlights:
  - ✓ Integrated memory controller on each CPU
    - Aggregate I/O bandwidth can scale with memory bandwidth
    - High-speed access to memory via HyperTransport™ enables multiple I/O devices to operate at full-speed per link.
  - ✓ HyperTransport™ interconnect technology
    - High-speed access to memory via HyperTransport™ enables multiple I/O devices to operate at full-speed per link.
    - Up to 3 links per CPU at 6.4GBps gross bandwidth available for each IO chain
    - Devices can be chained
  - ✓ Point-to-point I/O attach
    - Complete error containment and fault isolation
    - Complete performance isolation
    - Hot-plug supported
  - ✓ Supports wide mix of PCI-X device attachment to deliver balanced server performance
    - 530 MBytes/s to 2.13 GBytes/s bandwidth per device
  - ✓ Allows system scalability based number of PCI-X interfaces desired



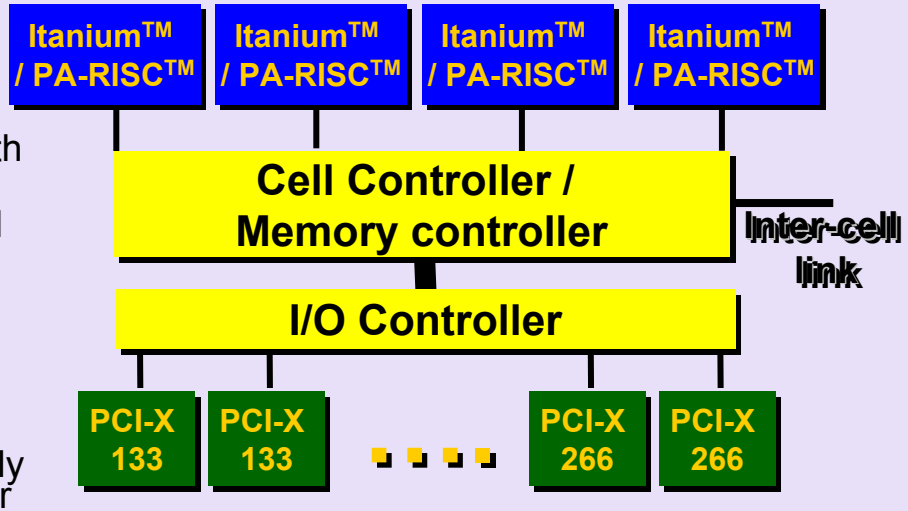
PCI-X 66-266 4P Server Example

- Highlights:
  - ✓ Integrated memory controller / I/O controller
    - Aggregate I/O bandwidth to scale with memory bandwidth eliminating
      - PCI / PCI-X architecture has never limited I/O bandwidth scaling – strictly an implementation issue
    - High-speed access to memory enables multiple I/O devices to operate at full-speed
  - ✓ Point-to-point I/O attach
    - Complete error containment and fault isolation
    - Complete performance isolation
    - Improved hot-plug operation
    - Simplified management and troubleshooting
    - Delivers any device / any slot / any time capability
  - ✓ Optional support for shared bus
    - Shared bus usage is an implementation decision not a PCI / PCI-X architectural requirement
    - A number of server designs have been shipping for 5+ years using per bus per slot architecture
  - ✓ Supports wide mix of PCI-X device attachment to deliver balanced server performance
    - 530 MBytes/s to 2.13 GBytes/s bandwidth per device



PCI-X 133 / 266 1-128P Server Example

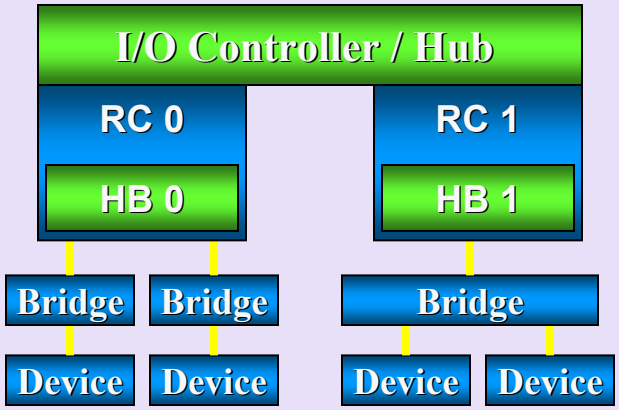
- Highlights:
  - ✓ High-speed I/O to memory interconnect
    - Enables I/O to scale with memory bandwidth
    - High-speed access to memory enables multiple I/O devices to operate at full-speed
  - ✓ System aggregate I/O scales with number of I/O controllers
    - I/O controllers can be accessed by multiple cells enabling applications to scale across large number of PCI-X devices
    - Various types of large-scale applications rely upon aggregate I/O as well as large number of devices to scale properly
      - E.g. OLTP, data mining, decision support, high-speed technical workloads, etc.
      - Today, customers deploy multiple PCI-X devices such as Fibre Channel to provide connectivity and performance
  - ✓ Point-to-point I/O attach
    - Complete error containment and fault isolation
    - Complete performance isolation
    - Improved hot-plug operation
    - Simplified management and troubleshooting
    - Delivers any device / any slot / any time capability



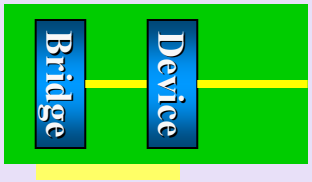
- ✓ Supports wide mix of PCI-X device attachment to deliver balanced server performance
  - 530 MBytes/s to 2.13 GBytes/s bandwidth per device

PCI Express to PCI-X Bridge

- Provides translation between PCI Express and PCI / PCI-X 2.0
- Three primary usage models:
  - ✓ PCI Express Root Port or Switch port attachment to PCI Express Bridge
    - Treats PCI Express as a system mezzanine bus to provide fan-out for PCI / PCI-X 2.0 device attachment (either system board or adapters)
  - ✓ Enable PCI / PCI-X 2.0 device to be embedded on a PCI Express form factor
  - ✓ Enable a PCI Express device to be attached to a PCI / PCI-X 2.0 chipset (system board or adapter) – this is referred to as a reverse bridge
- *Bridges are critical to the short and long-term success of PCI Express*
  - ✓ *Customer investment protection is critical to solution delivery / success as I/O transitions can take years to occur*



PCI Express as a Mezzanine Bus with PCI-X devices / slots



Adapter with embedded bridge – PCI Express to PCI-X 2.0 bridge or reverse bridge

Bridge Recommendations (1)

- Avoid taking ownership of transactions where possible
  - ✓ Support Max_Payload_Size to match I/O device unit of work
    - For example, Ethernet moves 1500B frames and Fibre Channel moves 2048B frames
    - Reduces buffer requirements for Requests (R / W) and Read Completions
      - Requests can be released once the link-level ACK is received
      - Majority of Read Completions are completed in a single transaction thus can release buffers faster and deliver better performance
- Support a 4096B Max_Read_Request_Size
  - ✓ Trivial for RC (thus common) while matching entire spectrum of endpoints
  - ✓ Simplified Read Completion processing as with a good Max_Payload_Size will complete worse case as two transactions thus mitigates any performance loss if transactions become interleaved in intervening switches.
  - ✓ Simplified control logic – should enable the bridge to support more concurrent requests which is critical to supporting high-speed I/O due to relatively large I/O-to-memory latency (memory speeds are not increasing as fast as processor and I/O speeds)
- Support Transaction End-to-end CRC
  - ✓ Must protect all data between a RC and a Bridge
  - ✓ Increases topology option flexibility and thus re-use of bridge across market segments

Bridge Recommendations (2)

- Supporting forwarding of ER_COR, ERR_NONFATAL, and ERR_FATAL errors from the secondary interface to the primary interface
  - ✓ SERR# should be set within the Command and Bridge Control register
- To deliver good performance:
  - ✓ Support only VC0 – simplify design / validation – increase bridge resource depths
    - Single VC will be the dominant design for server chipsets and I/O devices across the entire spectrum of server/ endpoint product offerings
  - ✓ Use 128B RCB (read completion boundary) in the RC
    - As PCI Express starts to ramp to volume in 2006 for servers, process technology will be primarily 90nm for chipsets and I/O making any cost arguments in favor of 64B rather moot
  - ✓ Operate the bridge in flood mode NOT store-n-forward
    - Store-n-forward increases end-to-end I/O operation latency thus harming performance
  - ✓ Coalesce read completions to improve PCI-X efficiency – see associated technical paper for PCI-X efficiency as a function of completion size
    - Majority of server I/O usage model is Ethernet, Fibre Channel, etc. which transfer larger I/O (typically <= 2048B)
  - ✓ PCI Express link should be ~2x the PCI-X bandwidth to overcome the inherent protocol overheads in PCI Express in order to deliver good PCI-X device performance

Agenda

- What is PCI-X 2.0?
- “Need for Speed”
- Industry State of I/O
- PCI-X 2.0 Enablement Efforts

PCI-SIG Compliance Program

- PCI-SIG Compliance Workshops fully support PCI-X 2.0 compliance and interoperability testing
  - ✓ Compliance checklists completed and available to all PCI-SIG members
  - ✓ Mode 1 and mode 2 test fixtures available
  - ✓ Multiple instrument vendors actively support PCI-X 2.0
  - ✓ New and existing mode 1 capable platforms and devices tested at each workshop
  - ✓ New mode 2 capable platforms and I/O devices started testing at most recent PCI-SIG compliance workshop

PCI-SIG "Yellow Pages"

- PCI-SIG website contains "yellow pages" directory to find PCI-X technology providers
  - ✓ Yellow pages cover the many areas such as:
    - PCI-X 2.0 devices, bridges, etc.
    - PCI-X 2.0 IP cores
    - PCI-X 2.0 test equipment, validation suites, test software, etc.
    - Etc.
  - ✓ See PCI-SIG website for additional details
- PCI-SIG website also contains a number of technical white papers and collateral materials to provide more in-depth understanding of PCI-X

Future of PCI-X Summary

- Natural follow-on to widely successful PCI-X 1.0
- Proven Infrastructure
  - ✓ Huge installed base of servers, storage controllers, etc.
  - ✓ 100's of I/O device and bridge designs to choose from
  - ✓ 10's of Millions of lines of software – complete range of OS, middleware, etc. support
  - ✓ OEM know how – relatively simple migration for systems
- PCI-X 2.0 is 100% Backward Compatible to PCI and PCI-X.
  - ✓ Spans all PCI, PCI-X, and PCI-X 2.0 devices
  - ✓ All existing PCI-X capable solution can use PCI-X 2.0 devices
  - ✓ All PCI-X 2.0 devices have ready homes in tens of millions of existing PCI-X 1.0 and PCI solutions.
- PCI-X 2.0 deliver the performance to drive all applications for the foreseeable future.
  - ✓ Capable of supporting 10 Gbps Ethernet, 10 Gbps Fibre Channel, cluster interconnects, etc.
- PCI-X slots will be needed to provide customer investment protection for many years
  - ✓ IT managers need to maintain backward compatibility
  - ✓ Over 90 million PCI-X slots deployed by 2006

Thank you for attending the
2004 PCI-SIG
Developers Conference.

For more information please go to
www.pcisig.com


PCI
SIG™