X64 Workshop

I/O Architecture

Michael Mullins

Systems-TSC

X64 Workshop

Introduction

- Welcome
  - > Introduction
- Scope
  - > What's covered?
  - > What's not covered?

Agenda

- Architectures
  - PCI (Peripheral Component Interconnect)
  - PCI-X
  - PCI-e
- Supported Options
- Current issues

X64 Workshop ---- PCI

- Bus Architecture
  - PCI and PCI-X busses are multi-drop parallel interconnect in which many devices share one bus
- PCI bus speeds 0 – 66Mhz (PCI)
- Devices are known as either Initiator or Targets (PCI)
- Typical PCI device either IC or PCI expansion card
- PCI bus loading
- Half duplex

X64 Workshop --- PCI-X

- PCI-X bus speeds 0 – 133Mhz (Higher speeds ?)
- Bus load specific
  - (1 Slot = 133Mhz, 2 Slot = 100Mhz, 4 Load = 66Mhz)
- Devices are known as either Requester / Completer (PCI-X)

X64 Workshop --- PCI / PCI-X

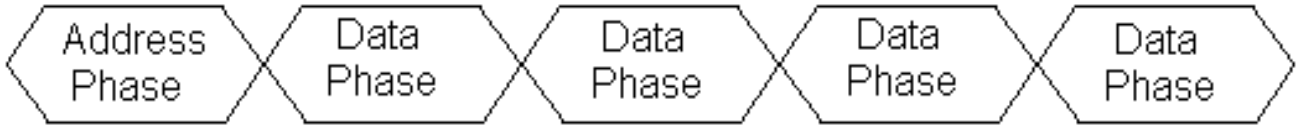
- Steps in a PCI Transaction:
  - Step 1: Bus Grant
    - The PCI arbiter gives a requesting initiator permission to use the bus
  - Step 2: Address Phase
    - The target for the transaction and the type of transaction to take place are identified.
  - Step 3: Data Phase(s)
    - Data is transferred between the agents. There may be multiple data phases if the current transaction is a burst transfer.

X64 Workshop --- PCI / PCI-X

- Transaction type are either Single date phases or Burst transfer

Single data phase transaction

Burst transfer



X64 Workshop --- PCI / PCI-X

- PCI Transaction Model
-
  - Programmed IO (PIO)
  - Direct Memory Access (DMA)
  - Peer to Peer
-
  - PCI Retry Protocol
  - PCI Disconnect Protocol

X64 Workshop --- PCI/PCI-X

- Programmed IO
  - A transaction initiated by the CPU and that targets a peripheral device, is referred to as a programmed IO transaction. Software commands the CPU to initiate a memory or IO read/write cycle on the host bus targeting an address mapped in a PCI devices address space. Once arbitration of the PCI bus is gained the address is driven onto the PCI bus and all targets decode the address and one device will then claim the transaction. Data is then transferred between master and target under control of the CPU

X64 Workshop --- PCI/PCI-X

- Direct Memory Access (DMA)
  - A PCI device becomes the master of the PCI bus under control of the software running on this device, the device may initiate a request to talk to memory. The PCI bus device arbitrates for the bus and when it becomes the bus master initiates a PCI memory bus cycle. The north bridge (NB) which decodes the address acts as the target for the transaction. During the data phase of the bus cycle data is transferred between PCI device and the NB. The NB in turn generates the memory cycle and communicates with system memory. Once the data transfer is completed the PCI device generates an interrupt to indicate transfer is complete.

X64 Workshop --- PCI/PCI-X

- Peer to Peer
  - A peer to peer transaction is the direct transfer of data between two PCI devices. A device that wishes to initiate a transaction arbitrates, wins ownership of the bus and starts a transaction. A target that recognizes the address claims the bus cycle.
  - For a write bus cycle, data is moved from initiator to target.
  - For a read bus cycle, data is moved from the target to the initiator.

X64 workshop --- PCI/PCI-X

- PCI retry protocol
  - When a PCI master initiates a transaction to access a target device and the target device is not ready, the target signals a transaction retry. (wait states may be inserted in the data phase if only a few are needed)
    A retry tells the master to end the bus cycle prematurely without transferring any data. The PCI device releases the PCI bus and must re-arbitrate for the PCI bus to attempt the same bus cycle again. Hopefully the target device will now be ready otherwise and the transaction can continue otherwise a retry transaction is again issued and the cycle continues.

X64 Workshop --- PCI/PCI-X

- PCI disconnect protocol
  - When a PCI master initiates a transaction to access a target device and if the target device is able to transfer part of the data but cannot complete the entire data transfer, it disconnects the bus cycle at the point at which it cannot continue the data transfer. (wait states may be inserted in the data phase if only a few are needed). The PCI device releases the PCI bus and must re-arbitrate for the PCI bus to attempt the same bus cycle again. Hopefully the target device will now be ready otherwise and the transaction can continue otherwise a retry transaction is again issued and the cycle continues.

X64 Workshop --- PCI/PCI-X

- PCI Bandwidth
-
- 33MHz 32bit 133MB/s
- 33MHz 64bit 266MB/s
- 66MHz 32bit 266MB/s
- 66Mhz 64bit 512MB/s

- PCI-X Bandwidth
-
- 100MHz 64bit 800MB/s
- 133MHz 64bit 1GB/s
- 266Mhz 64bit 2GB/s

X64 Workshop --- PCI-e

- Common PCI-e Terms
-
  - > Serdes (Serializer / Deserializer)
  - > 8B/10B encoding / decoding
    - – 8B character
    - – 10B symbol
    - – Among the areas in which 8B/10B encoding finds application are HyperTransport, PCI Express, IEEE 1394b, Serial ATA,SAS, Fibre Channel, SSA, Gigabit Ethernet, InfiniBand,XAUI, Serial RapidIO, DVI (Transition Minimized Differential Signaling), DVB Asynchronous Serial Interface (ASI)

X64 Workshop --- PCI-e

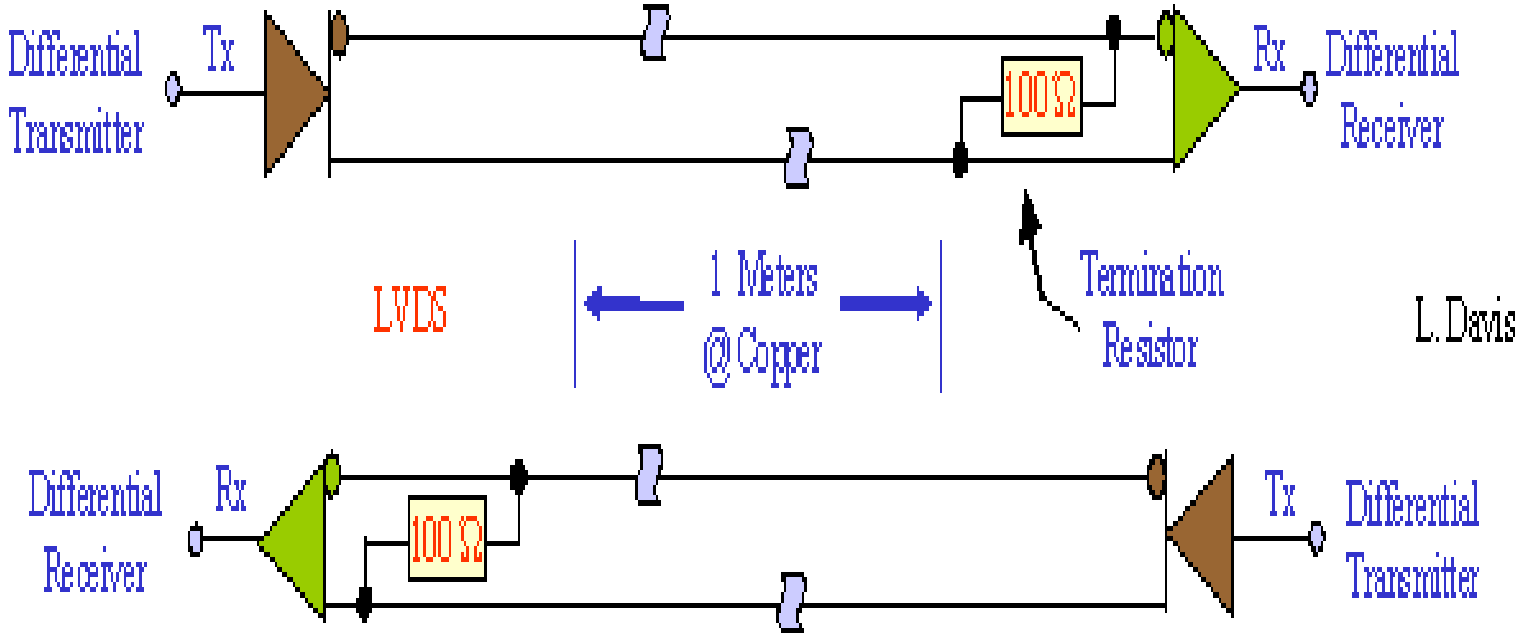
- 8b/10b encoding
  - As the scheme name suggests, 8 bits of data are transmitted as a 10-bit entity called a symbol. The low 5 bits of data are encoded into a 6-bit group and the top 3 bits are encoded into a 4-bit group. The data symbols are often referred to as Dxx.y where xx ranges from 0-31 and y from 0-7. Because 8B/10B encoding uses 10-bit symbols to encode 8-bit words, each of the 256 possible 8-bit words can be encoded in two different ways, one the bit-wise inverse of the other. Using these alternative encodings, the scheme is able to effect long-term DC-balance in the serial data stream.
  - The encoding is normally done entirely in hardware, based on lookup tables.

X64 Workshop --- PCI-e

- PCI-e implements a serial point to point type interconnect for communication between two devices
- A PCI-e interconnect that connects two devices together is referred to as a link
- A link consists of either x1, x2, x4, x8, x16 or x32 signal pairs in each direction
- These signals are referred to as lanes

X64 Workshop --- PCI-e

- Differential Transmitter / Receiver



X64 Workshop --- PCI-e

- Hardware protocol summary
  - Physical Layer.
    - At the electrical level, each lane utilizes two unidirectional low voltage differential signaling (LVDS) pairs at 2.5 Gbit/s. Transmit and receive are separate differential pairs, for a total of 4 data wires per lane.
  - Data Link Layer.
    - The Data Link Layer implements sequencing of Transaction Layer Packets (TLPs) that are generated by the Transaction Layer
  - Transaction Layer.
    - PCI Express implements split transactions (transactions with request and response separated by time), allowing the link to carry other traffic while the target device gathers data for the response.

X64 Workshop --- PCI-e

- Link initialization and training is a physical layer process that configures and initializes a devices physical layer, port and associated links so normal packet traffic can proceed on the link.
-
- There are two sub-blocks within the physical layer and these are split into transmit logic and receive logic, which allows for dual simplex communications

X64 Workshop --- PCI-e

- PCI-e Bandwidth
- Speed is per direction (transmit + receive)
- x1 250MB/s (500MB/s)
- x2 500MB/s (1GB/s)
- x4 1GB/s (2GB/s)
- x8 2GB/s (4GB/s)
- x16 4GB/s (8GB/s)
- x32 8GB/s (16GB/s)

X64 Workshop --- Supported Options

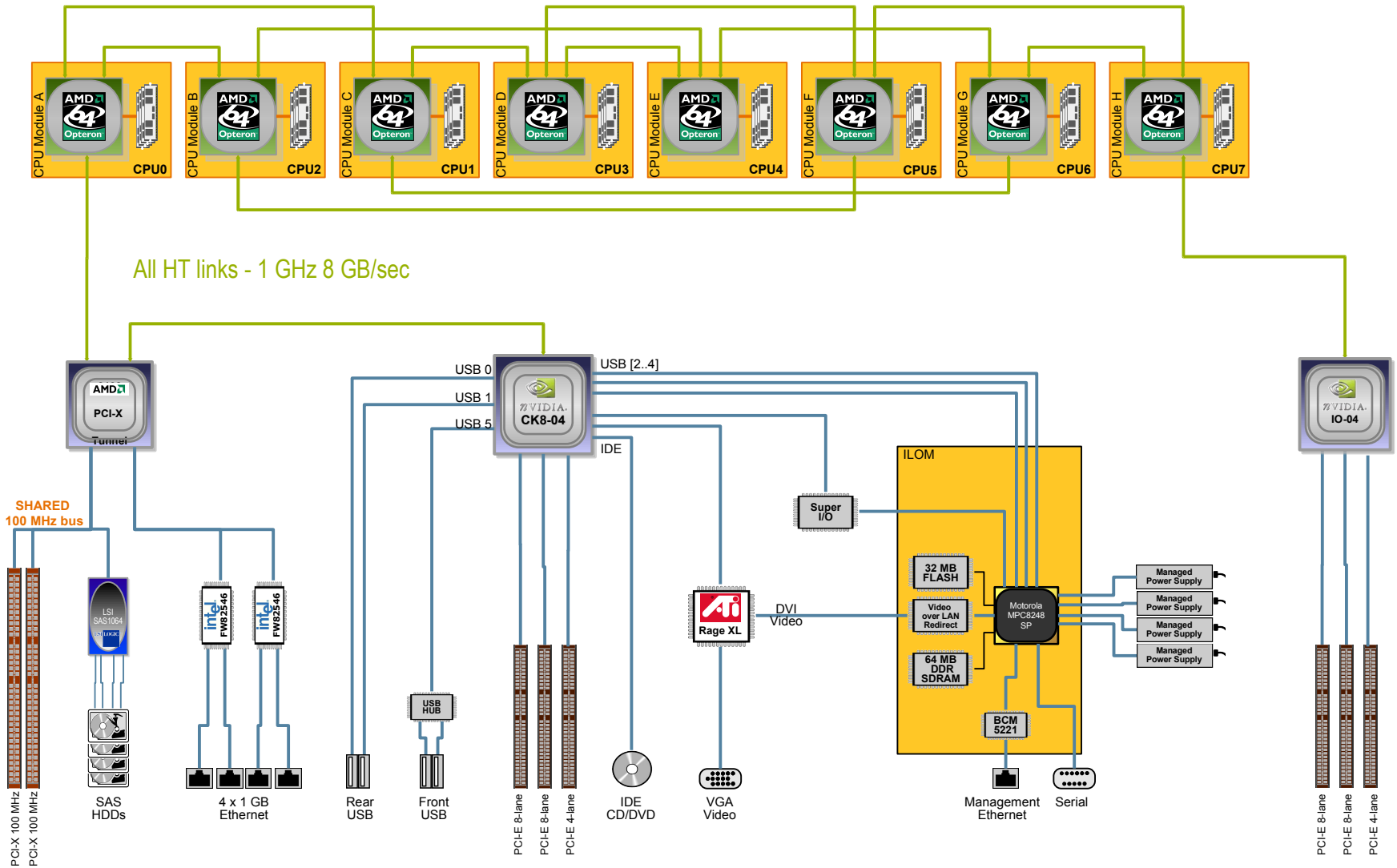
- PCI-X
  - Scsi
  - Fibre channel
  - Ethernet

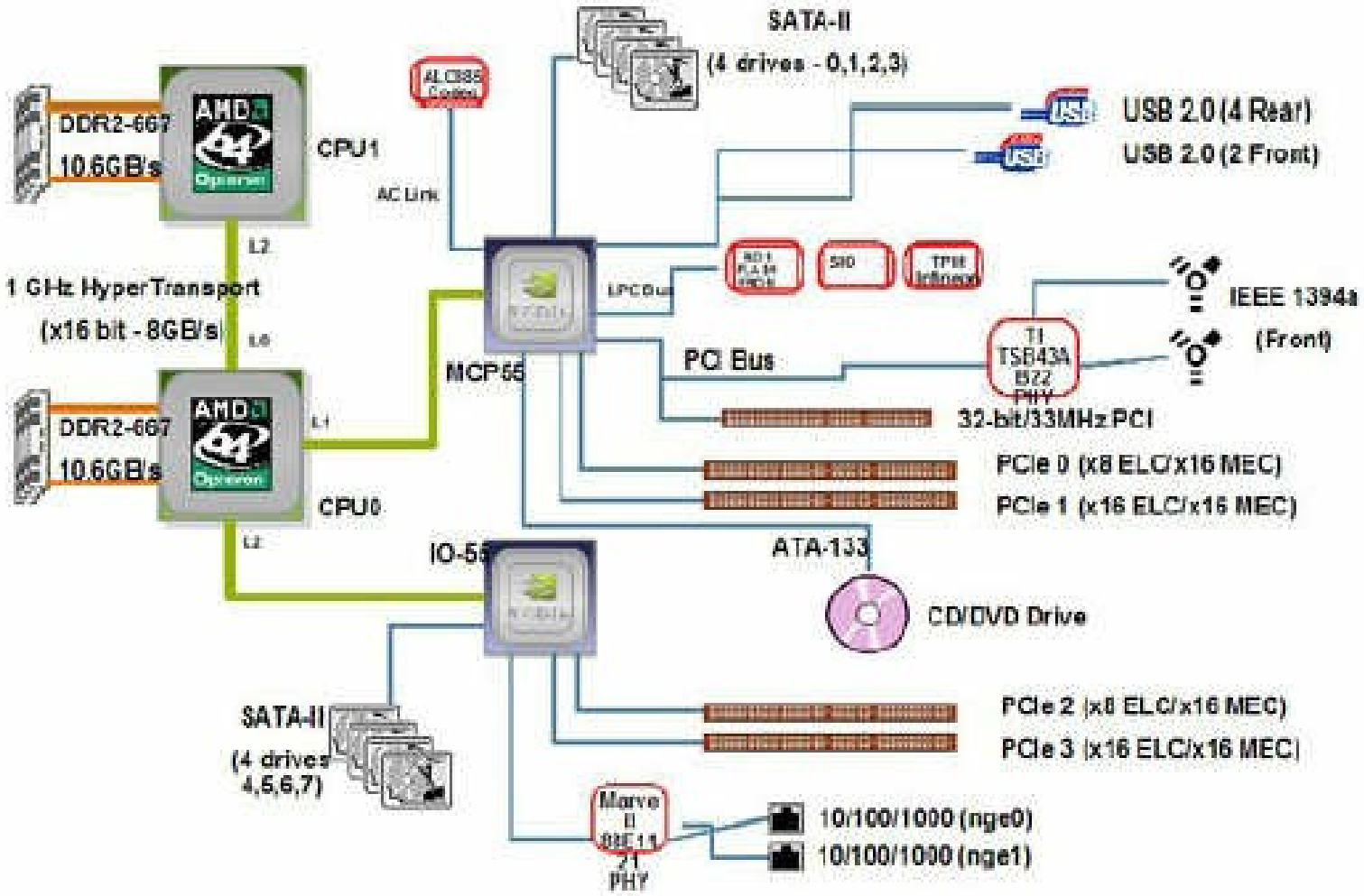
  Further options added almost daily

- PCI-e:
  - Infiniband
  - Fibre channel
  - Ethernet
  - Crypto Accellerator
  - Scsi

  Further options added almost daily

Block Diagram – 8 CPU Modules




SATA-II
(4 drives - 0,1,2,3)
USB 2.0 (4 Rear)
USB 2.0 (2 Front)
DDR2-667
10.6GB/s
CPU1
AC Link
L2
1 GHz HyperTransport
(x16 bit - 8GB/s)
L0
MCP55
LPC Bus
SIO
IEEE 1394a
(Front)
PCI Bus
32-bit/33MHz PCI
DDR2-667
10.6GB/s
L1
CPU0
L2
PCIe 0 (x8 ELC/x16 MEC)
PCIe 1 (x16 ELC/x16 MEC)
IO-55
ATA-133
CD/DVD Drive
SATA-II
(4 drives
4,5,6,7)
PCIe 2 (x8 ELC/x16 MEC)
PCIe 3 (x16 ELC/x16 MEC)
Marvell
88E1121
PHY
10/100/1000 (nge0)
10/100/1000 (nge1)

X64 Workshop --- PCI / PCI-X

- Current Issues.

Within the TSC platform pages there is an entry under each hardware platform a known issues page

Useful Links

http://systems-tsc/twiki/bin/view/Main/WebHome /* TSC home */

http://www.pcisig.com/specifications/ /* PCI specifications */
http://www.sun.com/io_technologies/ /* Third party support */

End

Presenter's Name
michael.mullins@sun.com