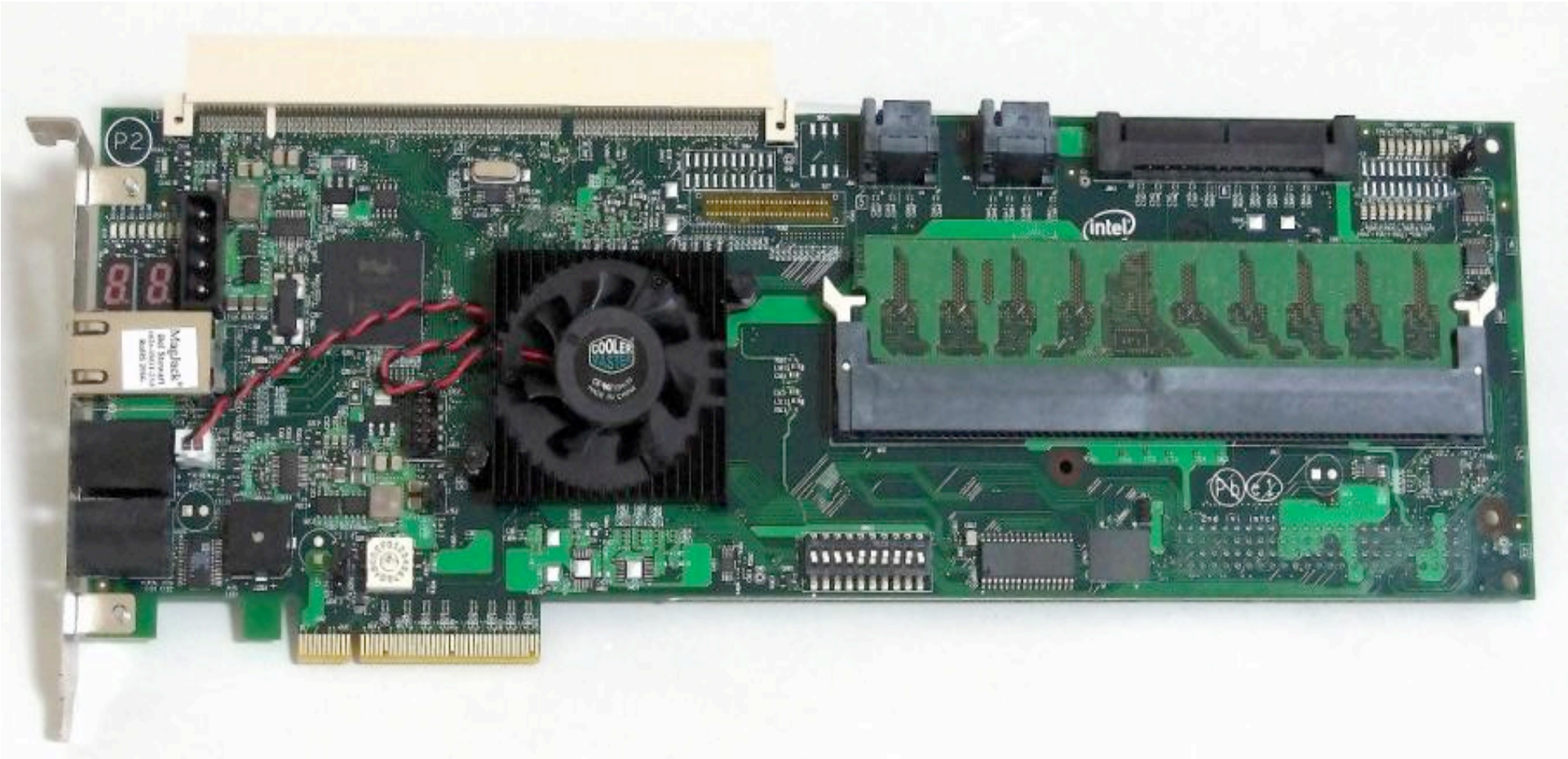
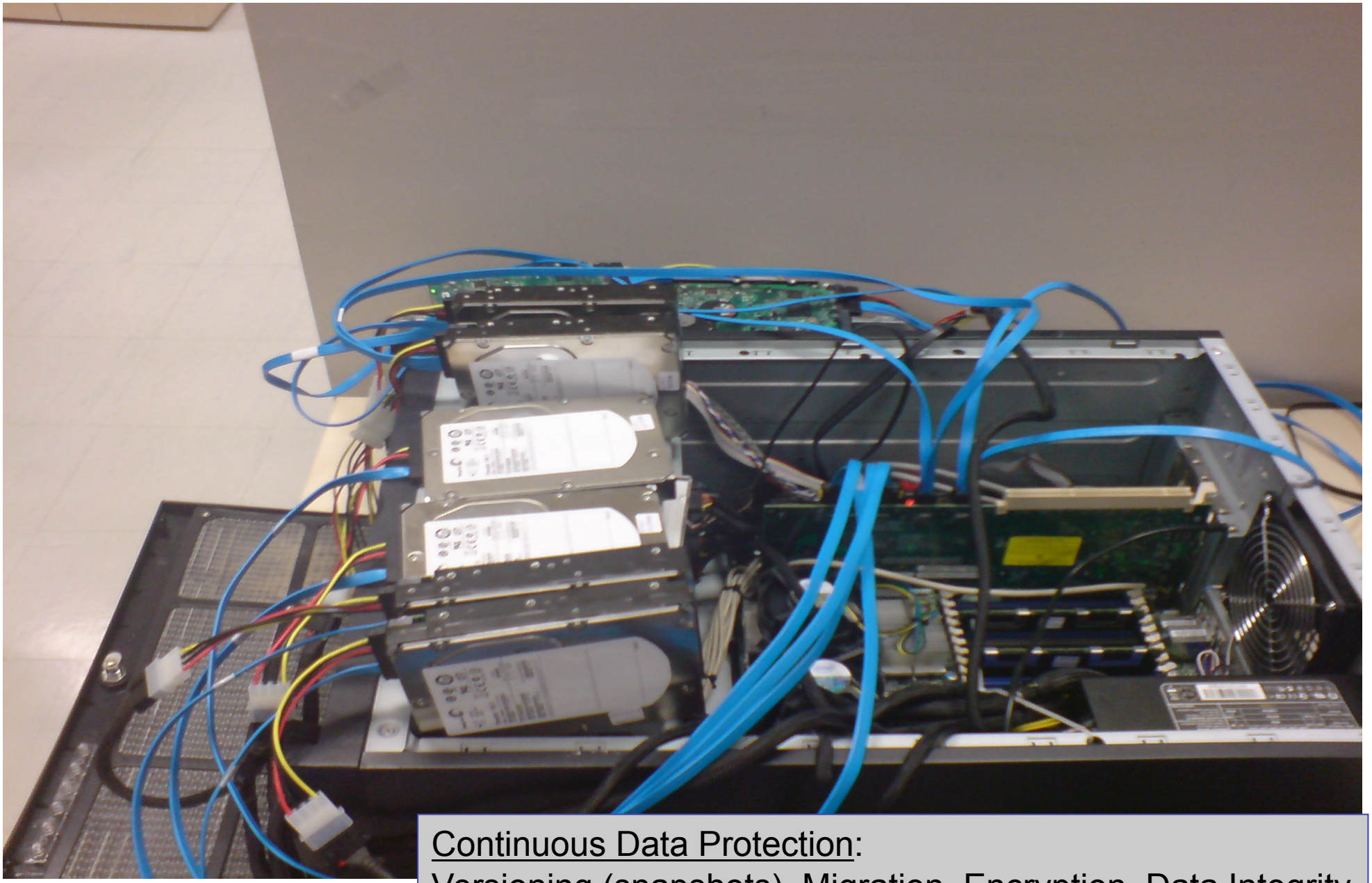
Linux Device Drivers: Case Study of a Storage Controller

Manolis Marazakis
FORTH-ICS (CARV)

IOP348-based I/O Controller



Programmable I/O Controller



Continuous Data Protection:
Versioning (snapshots), Migration, Encryption, Data Integrity

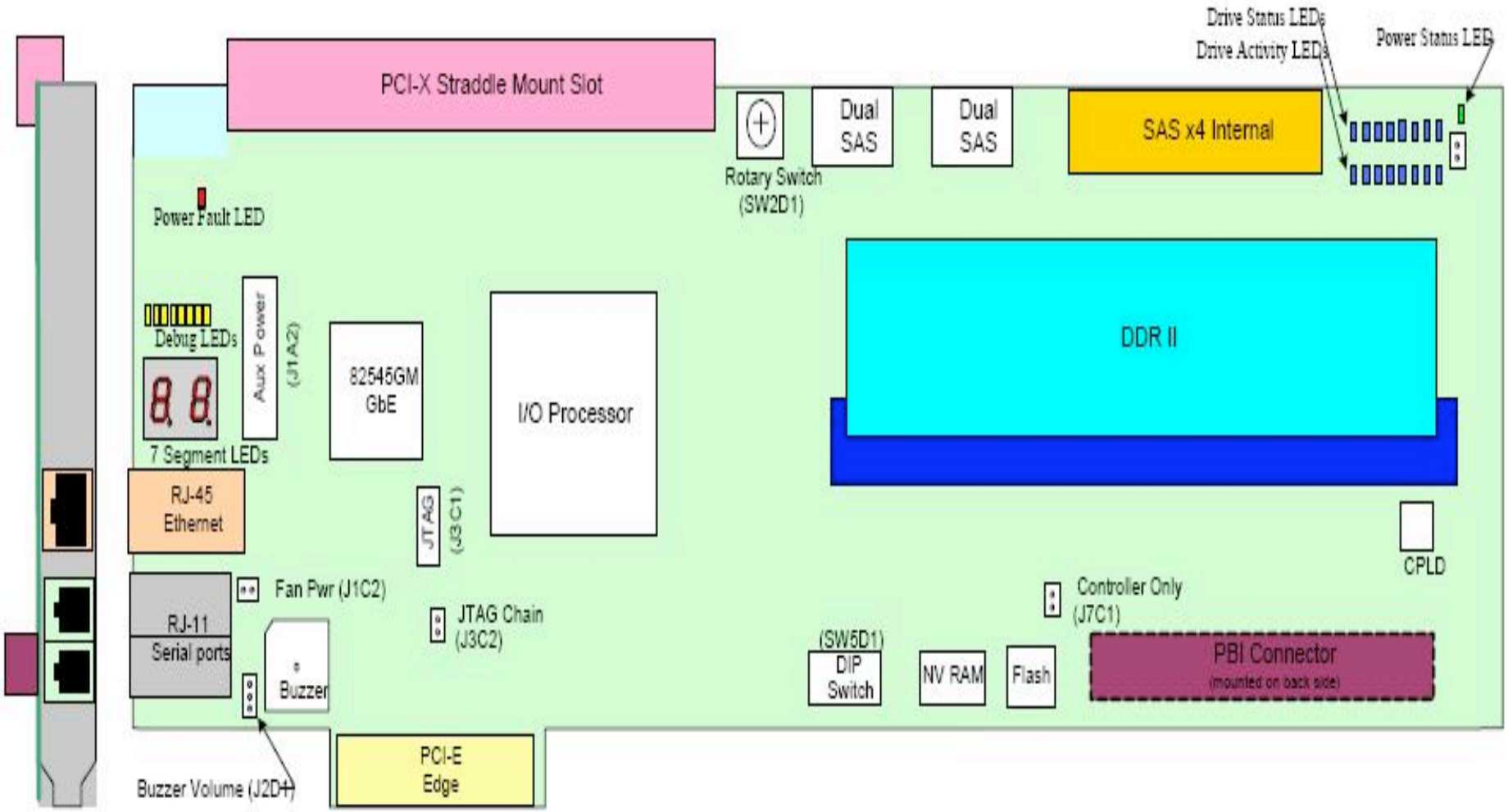
Misc. Storage Devices

SSD: 16+ GBytes
READ: 110–270 MB/sec
WRITE: 70-160 MB/sec
6000 – 10000 random IOPS
< 2 Watts

HDD: 500+ GBytes
READ/WRITE: 80-125 MB/sec
120-230 random IOPS
8.5-13.5 Watts

- Acceleration of commercial I/O-intensive workloads
- Reduction of power consumption

IOP348 - Board layout

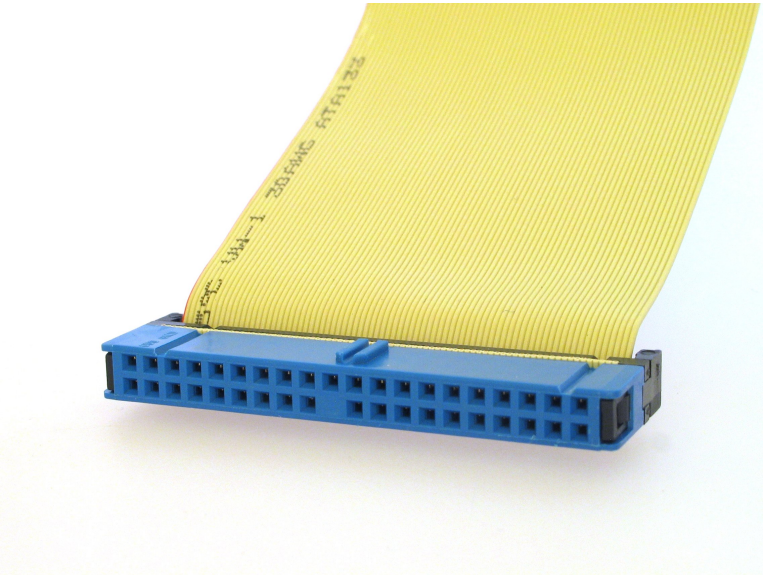


IOP348 – Boot Loader

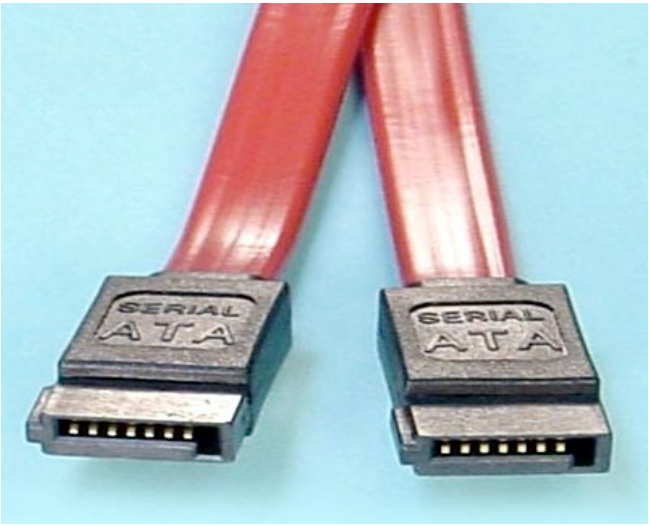
```
RedBoot> fis list
Name              FLASH addr  Mem addr    Length      Entry point
(reserved)        0xF0000000  0xF0000000  0x00200000  0x00000000
RedBoot           0xF0200000  0xF0200000  0x00040000  0x00000000
LinuxKernel       0xF0240000  0x01008000  0x00260000  0x01008000
FIS directory     0xF07E0000  0xF07E0000  0x0001F000  0x00000000
RedBoot config    0xF07FF000  0xF07FF000  0x00001000  0x00000000
RedBoot> fis load LinuxKernel
RedBoot> exec -c "console=ttyS0,115200 root=/dev/sda1" 0x1008000
Using base address 0x01008000 and length 0x00241f10
```

```
Intel IOP Redboot release
         version 2.3-IOP-RedBoot
         built 18:37:41, Dec 12 2006

Platform: IQ8134x SC (XScale) Core1, DDR2-533
IF_PCIX: 1, PCIe EP, PCI-X CR: PCIX-133
C1-Step, IB Speed: 400MHz, Core Speed: 1200MHz
Copyright (C) 2000, 2001, 2002, 2003, 2004 Red Hat, Inc.
Copyright (C) 2003, 2004, 2005, 2006 eCosCentric Limited

RAM: 0x00000000-0x10000000, [0x0001b7e8-0x0ffd1000] available
FLASH: 0xf0000000 - 0xf0800000, 64 blocks of 0x00020000 bytes each.
RedBoot> _
```

Disk Interface Technologies (I)



ATA



SATA

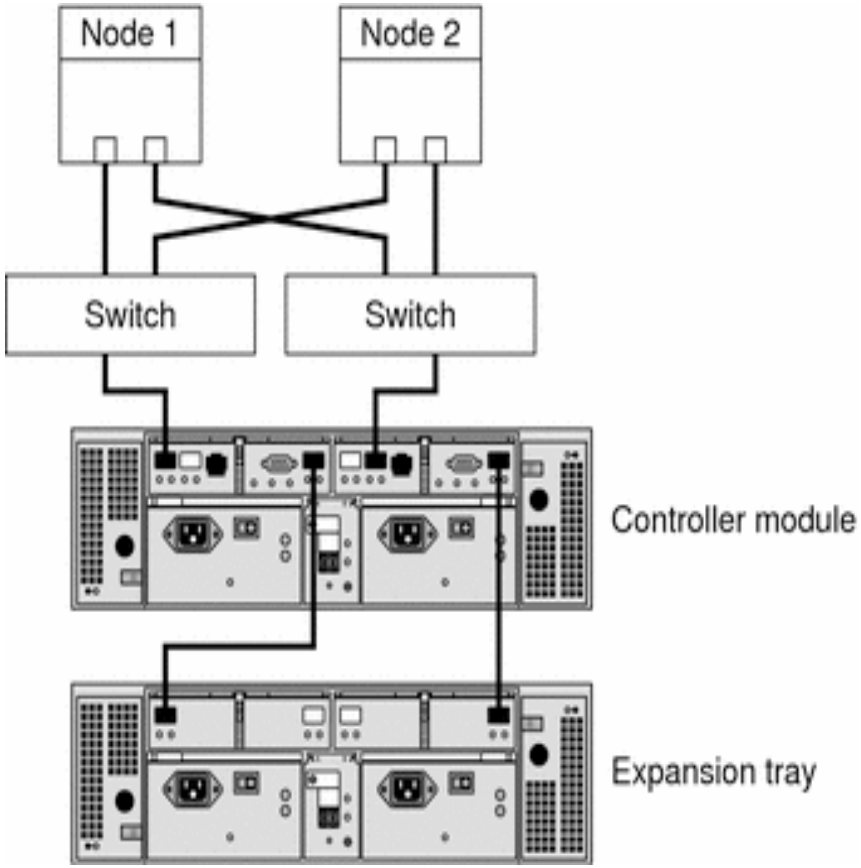
ATA: Advanced Technology Attachment
SATA: Serial ATA
SCSI: Small Computer System Interface
FC: Fibre Channel

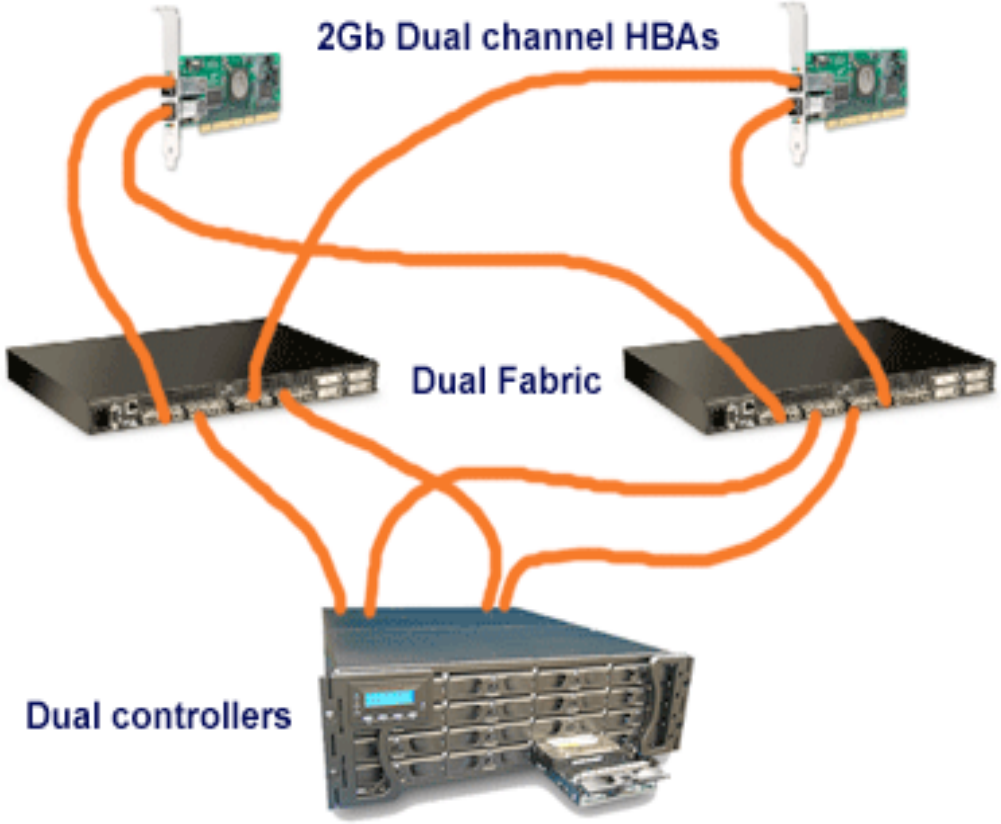
FC

Disk Interface Technologies (II)

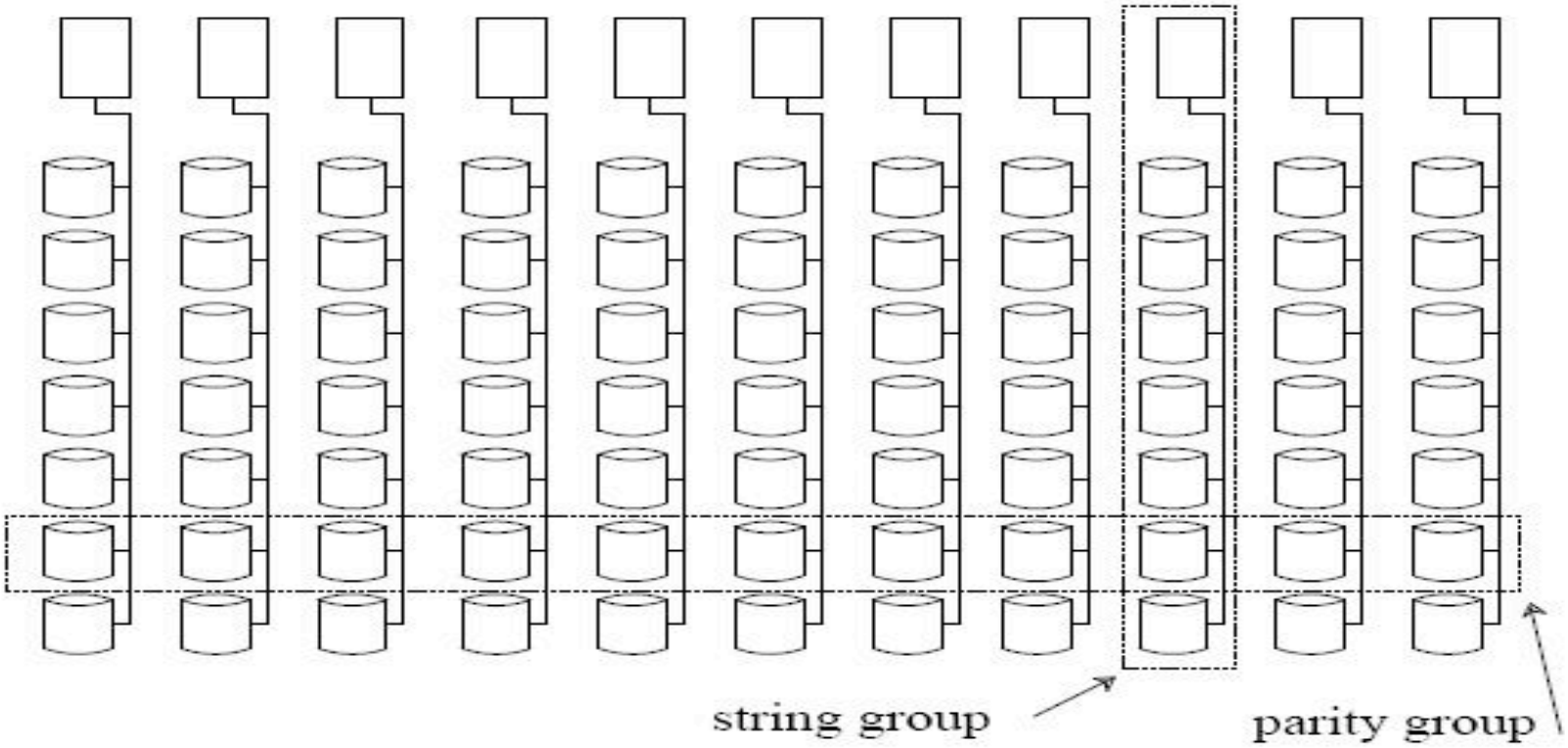
	Parallel ATA	Parallel SCSI	Fibre Channel	SATA	Serial Attached SCSI[1]
Performance					
Technology Introduction[2]	2000	2002	2001	2002	2004
Maximum Speed[3]	100 MB/s	320 MB/s	4.2 Gb/s (400 MB/s)	3.0 Gb/s (300 MB/s)	3.0 Gb/s (300 MB/s)
Topology	Shared bus master/slave	Shared bus	Arbitrated loop/ switched fabric	Point-to-point	Point-to-point
Number of Devices	2	15	1,000s	up to 15	100s

Example Storage Area Network Configuration



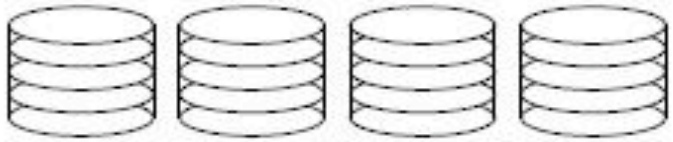


RAID: Redundant Array of Independent *(Inexpensive)* Disks

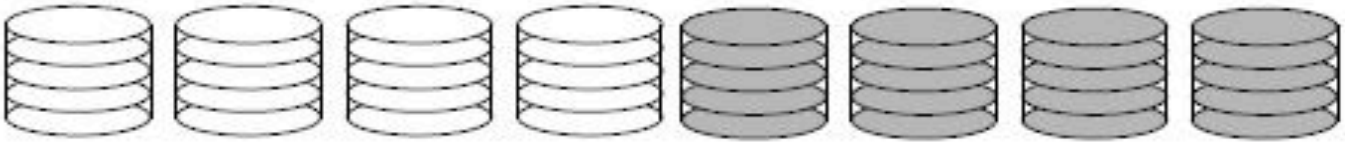


- Striping for parallel data transfer & load balancing
- Redundancy for failure protection
- Error-correcting codes

RAID Levels



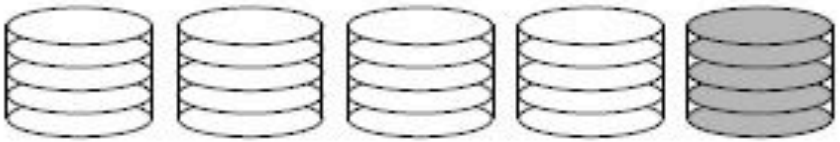
RAID Level 0: Non-redundant



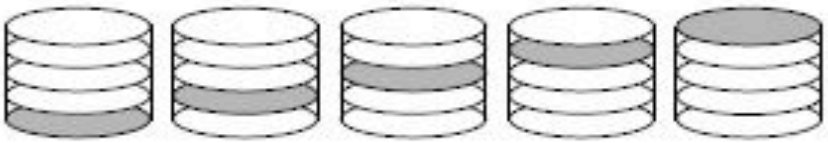
RAID Level 1: Mirroring

RAID Level 2: Byte-Interleaved, ECC

RAID Level 3: Byte-Interleaved, Parity

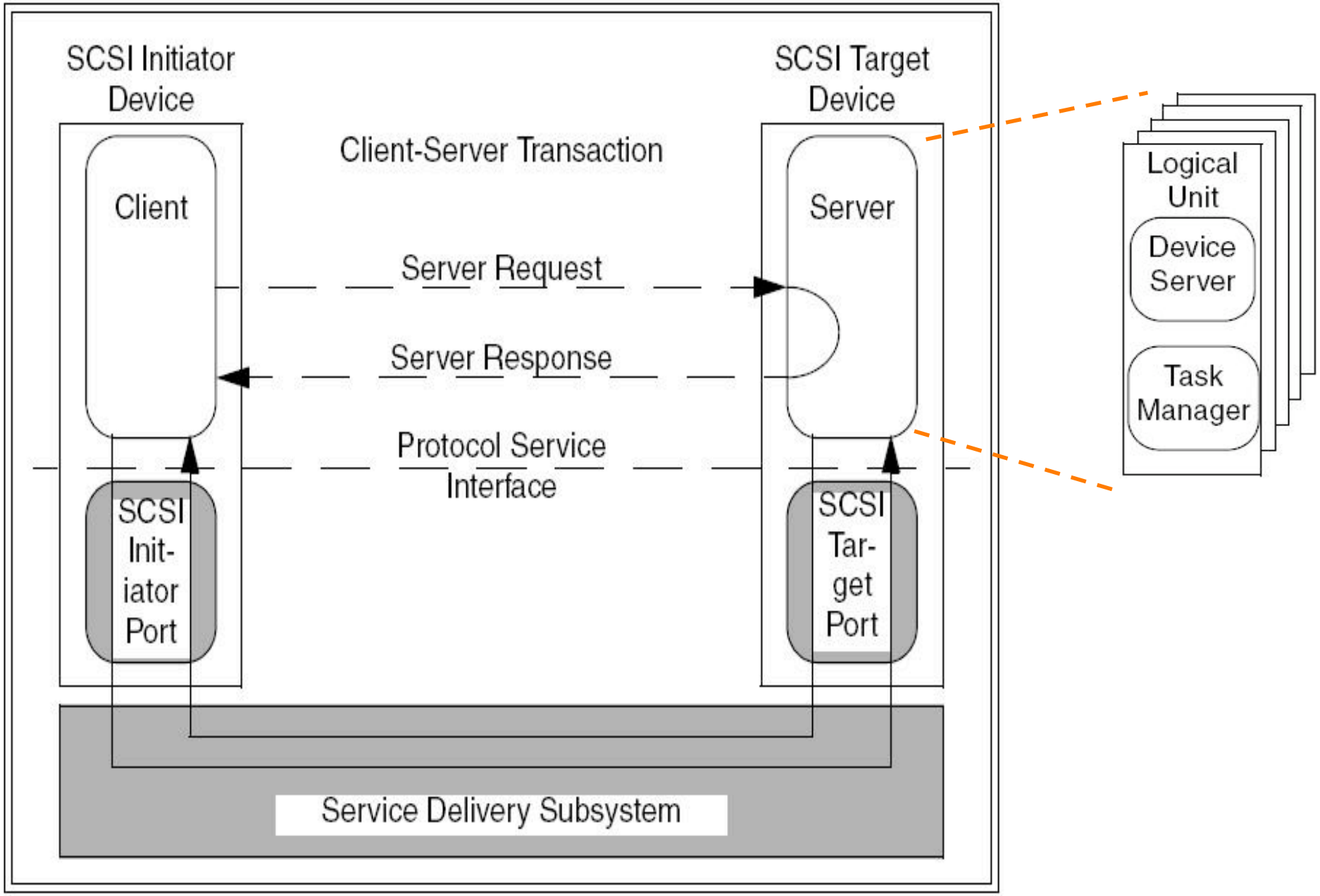


RAID Level 4: Block-Interleaved, Parity

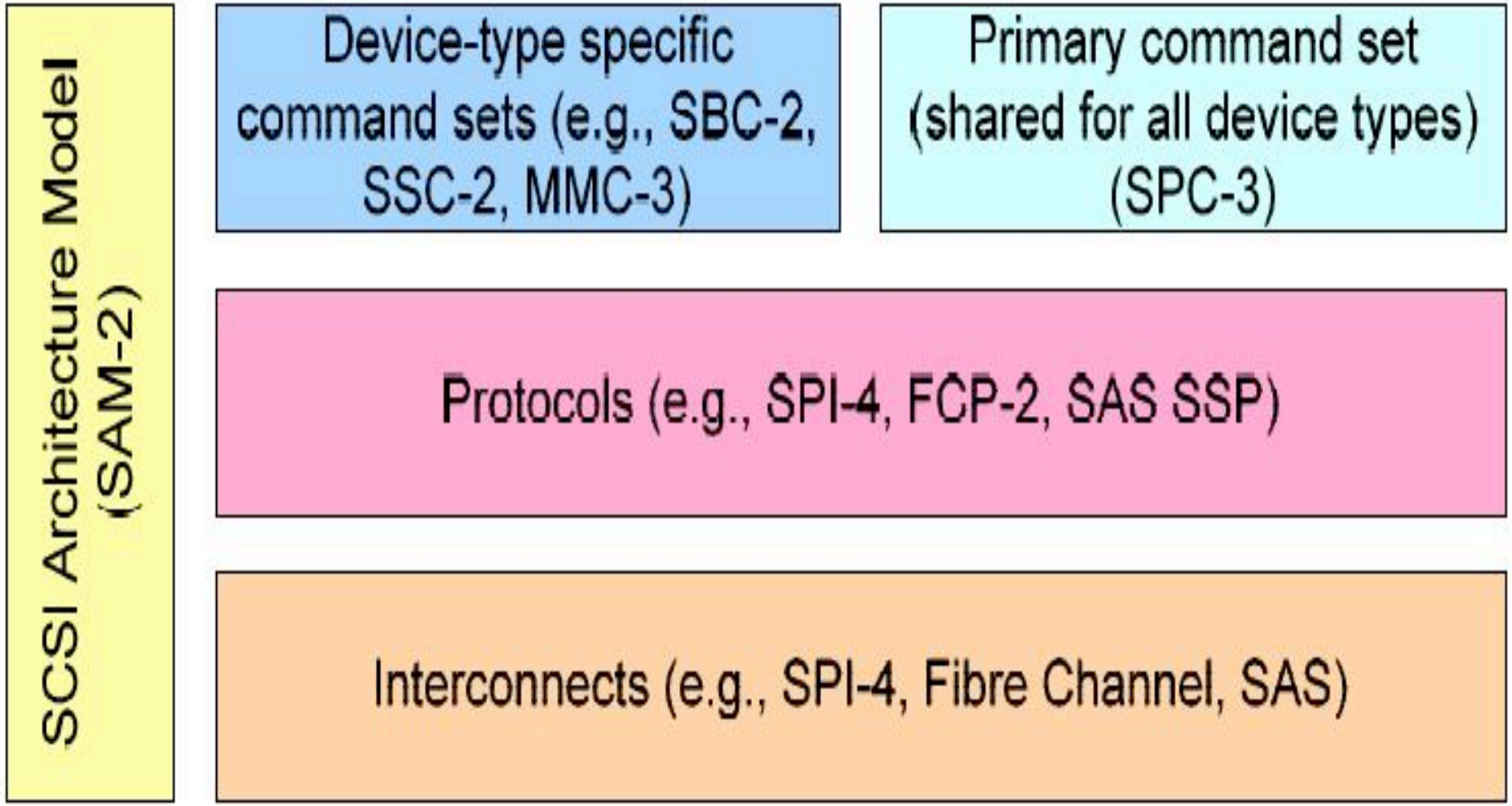


RAID Level 5: Block-Interleaved, Distributed Parity

SCSI (Small Computer System Interface)



SCSI Standards

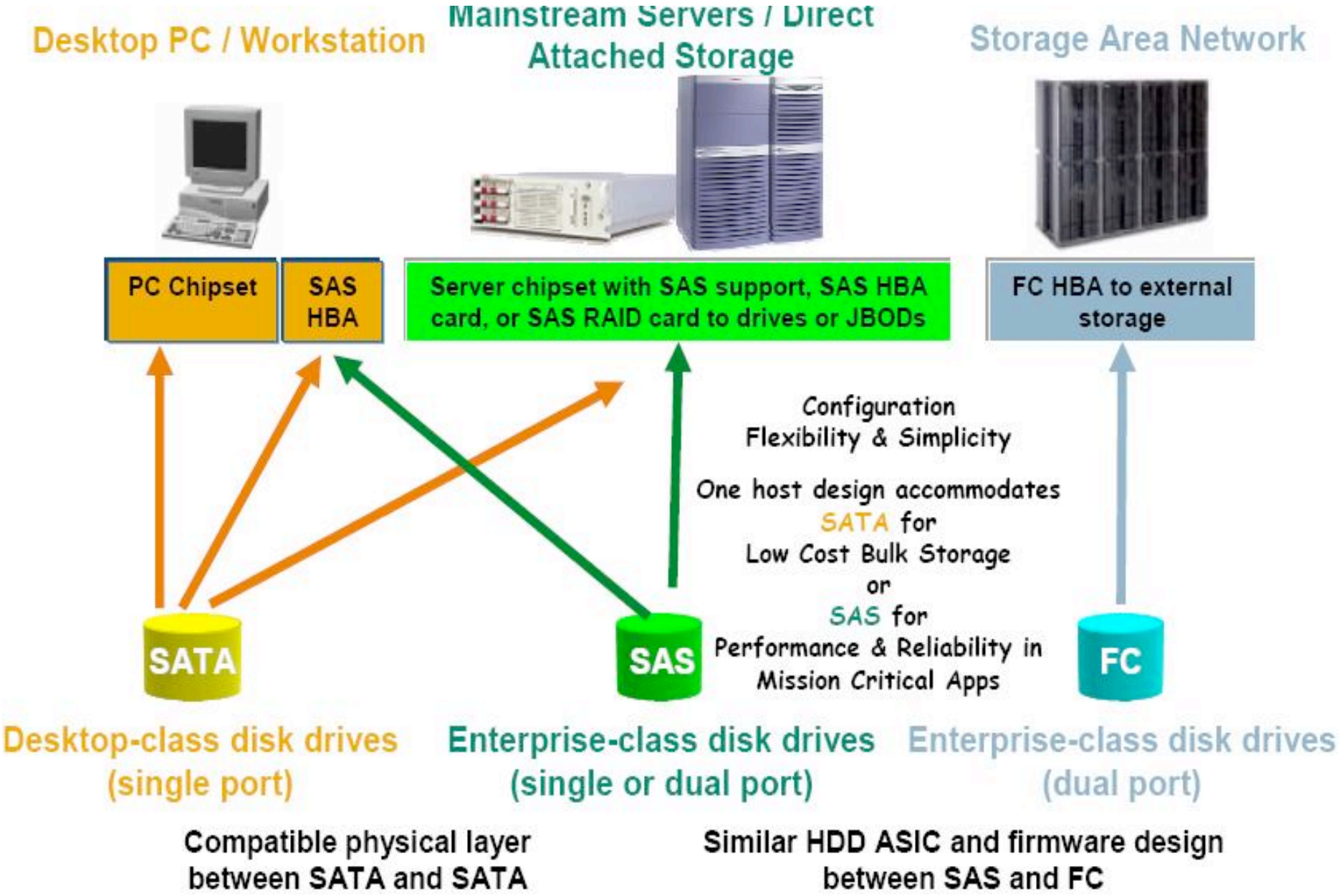


SCSI: Command Sets

Command name	OpCode	Command Support		Reference
		Fixed	Removable	
FORMAT UNIT	04h	O	O	RBC
INQUIRY	12h	M	M	SPC-2[1]
MODE SELECT(6)	15h	M	M	SPC-2[1]
MODE SENSE(6)	1Ah	M	M	SPC-2[1]
PERSISTENT RESERVE IN	5Eh	O	O	SPC-2[1]
PERSISTENT RESERVE OUT	5Fh	O	O	SPC-2[1]
PREVENT/ALLOW MEDIUM REMOVAL	1Eh	N/A	M	SPC-2[1]
READ (10)	28h	M	M	RBC
READ CAPACITY	25h	M	M	RBC
RELEASE(6)	17h	O	O	SPC-2[1]
REQUEST SENSE	03h	O	O	SPC-2[1]
RESERVE(6)	16h	O	O	SPC-2[1]
START STOP UNIT	1Bh	M	M	RBC
SYNCHRONIZE CACHE	35h	O	O	RBC
TEST UNIT READY	00h	M	M	SPC-2[1]
VERIFY (10)	2Fh	M	M	RBC
WRITE (10)	2Ah	M	M	RBC
WRITE BUFFER	3Bh	M	O	SPC-2[1]

[RBC - reduced block command set]

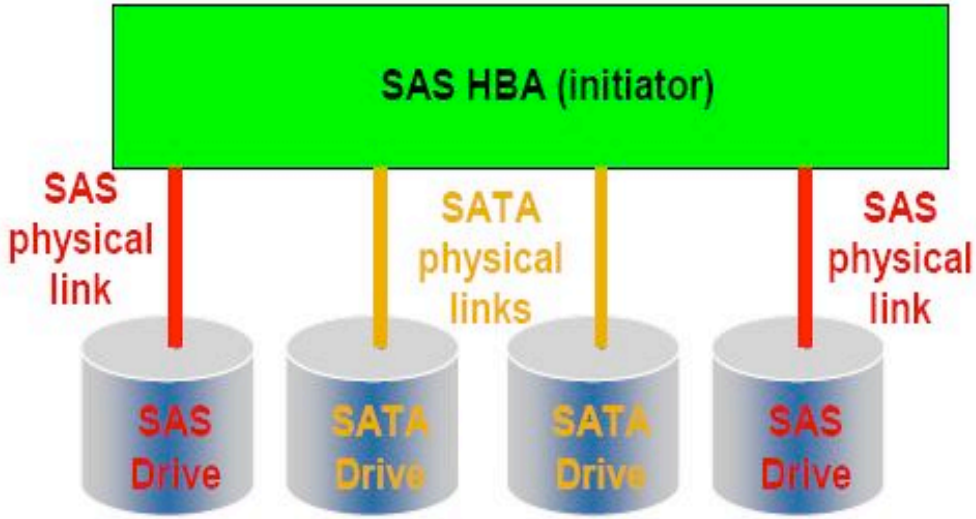
SAS: Serial Attached SCSI



SAS Host-Based Adapters

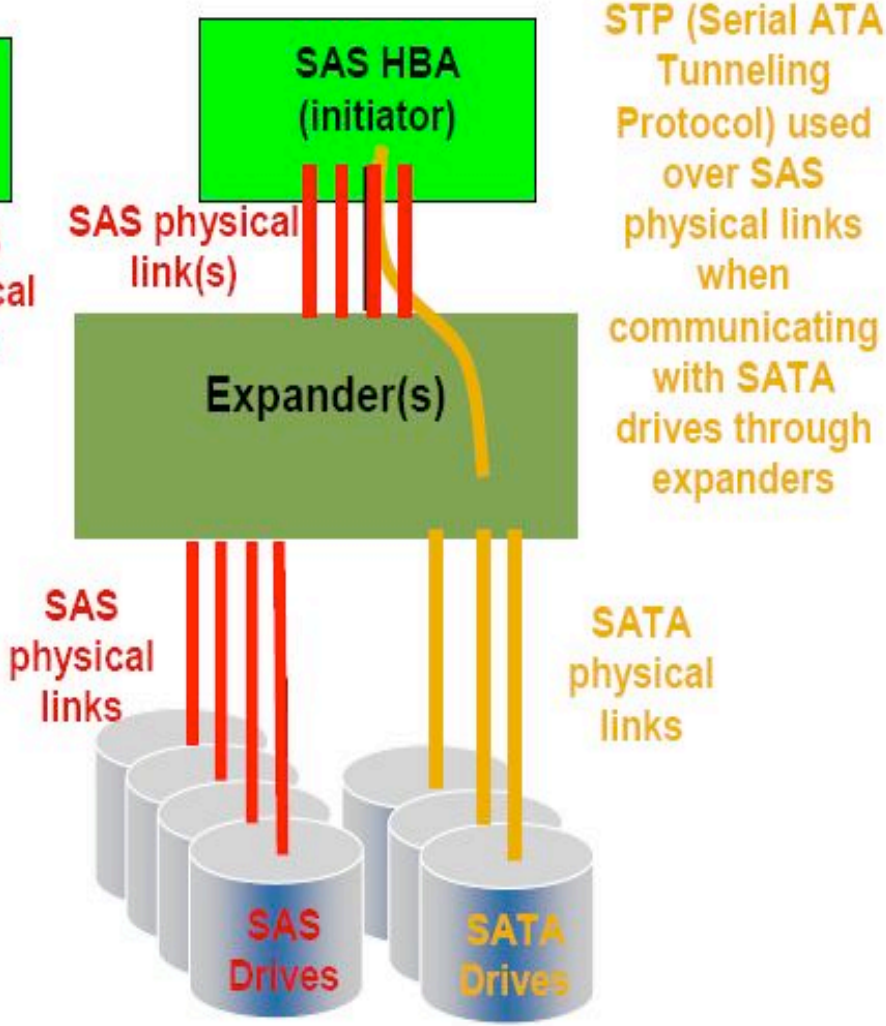
Direct attach =
Number of drives limited to number of ports in the HBA

Expander attach =
More drives than HBA ports



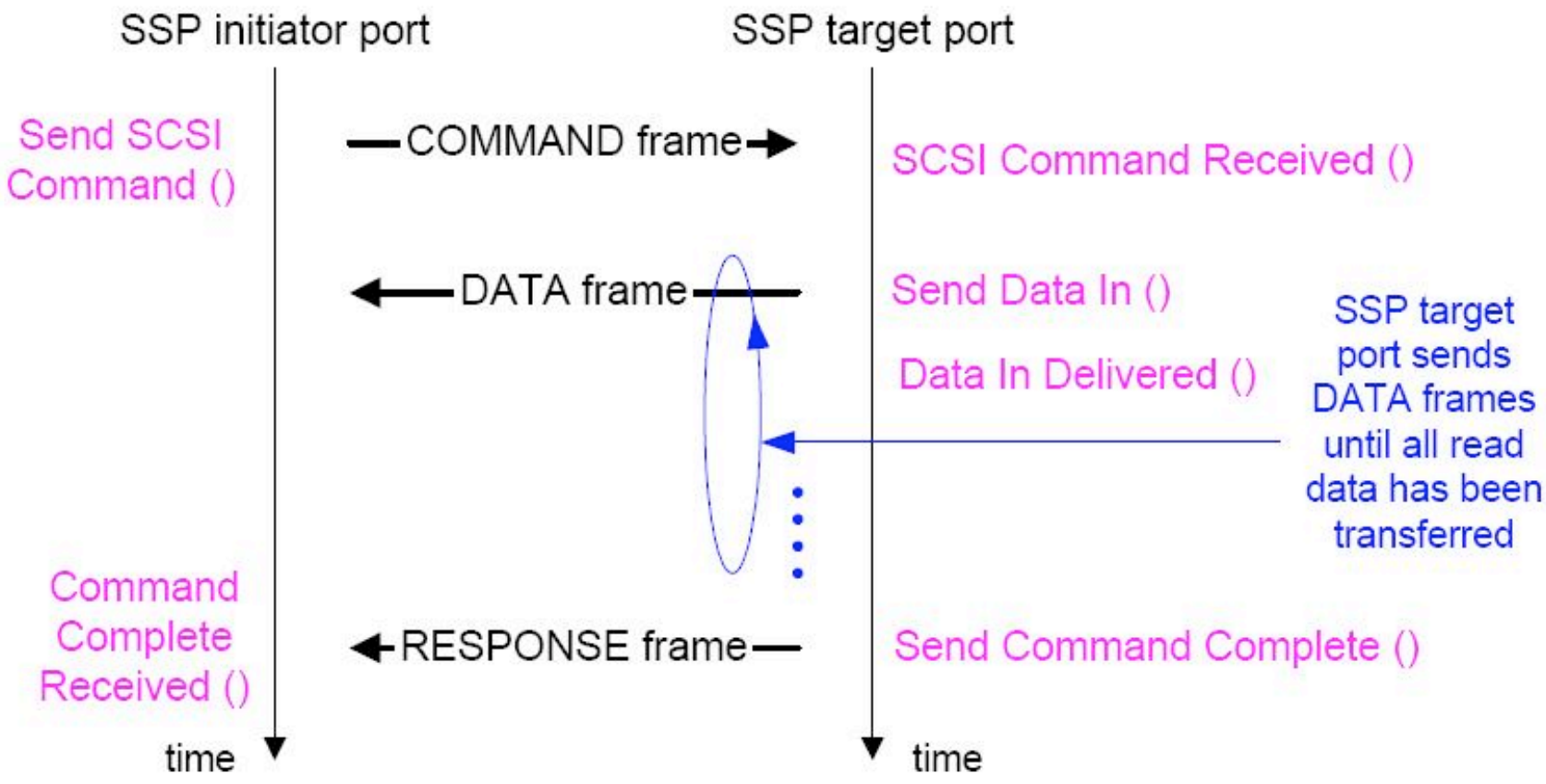
SSP (Serial SCSI Protocol) used to communicate with SAS drives

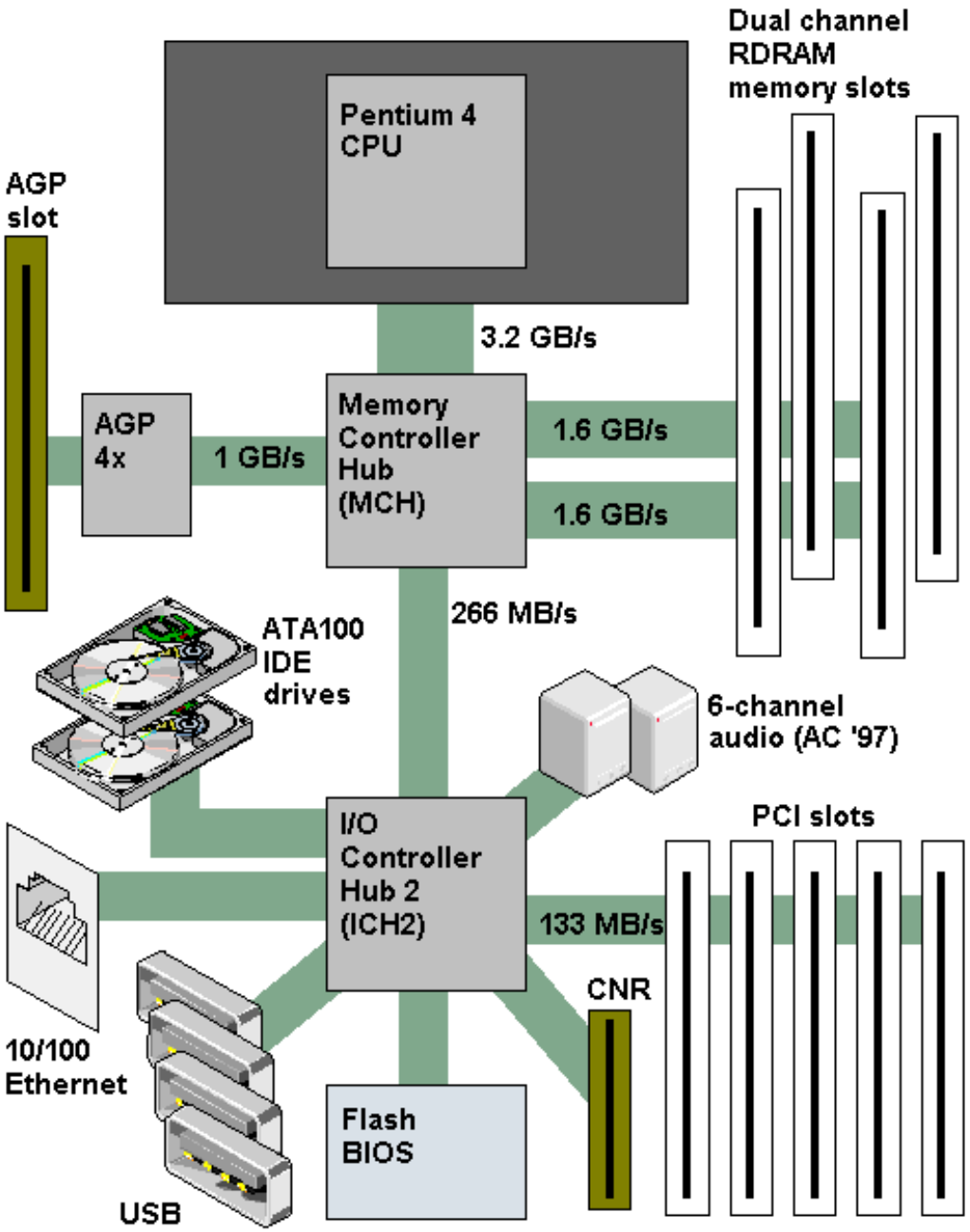
SATA (Serial ATA) used to communicate with SATA drives over SATA physical links



STP (Serial ATA Tunneling Protocol) used over SAS physical links when communicating with SATA drives through expanders

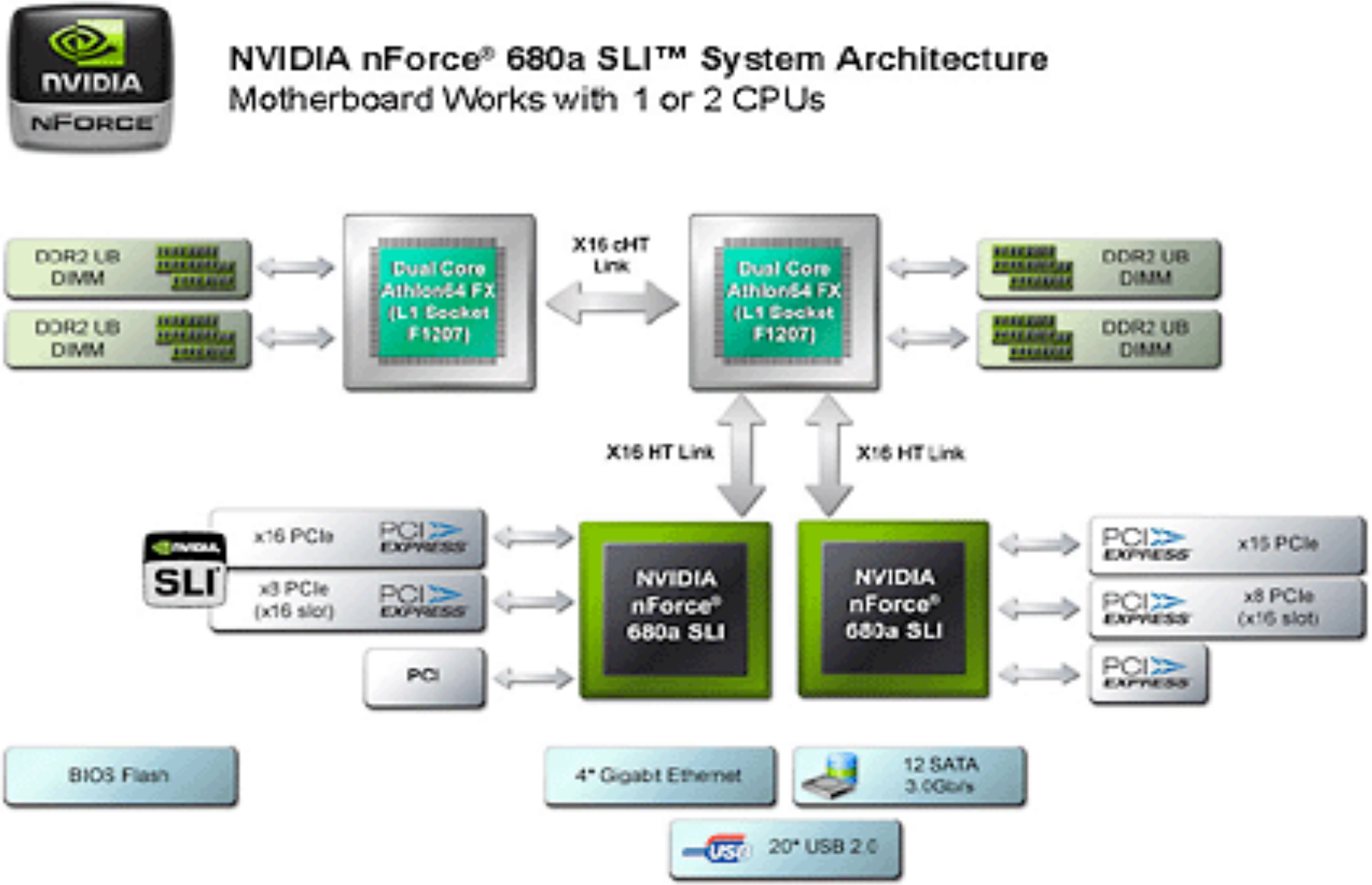
SAS: Read Command Sequence



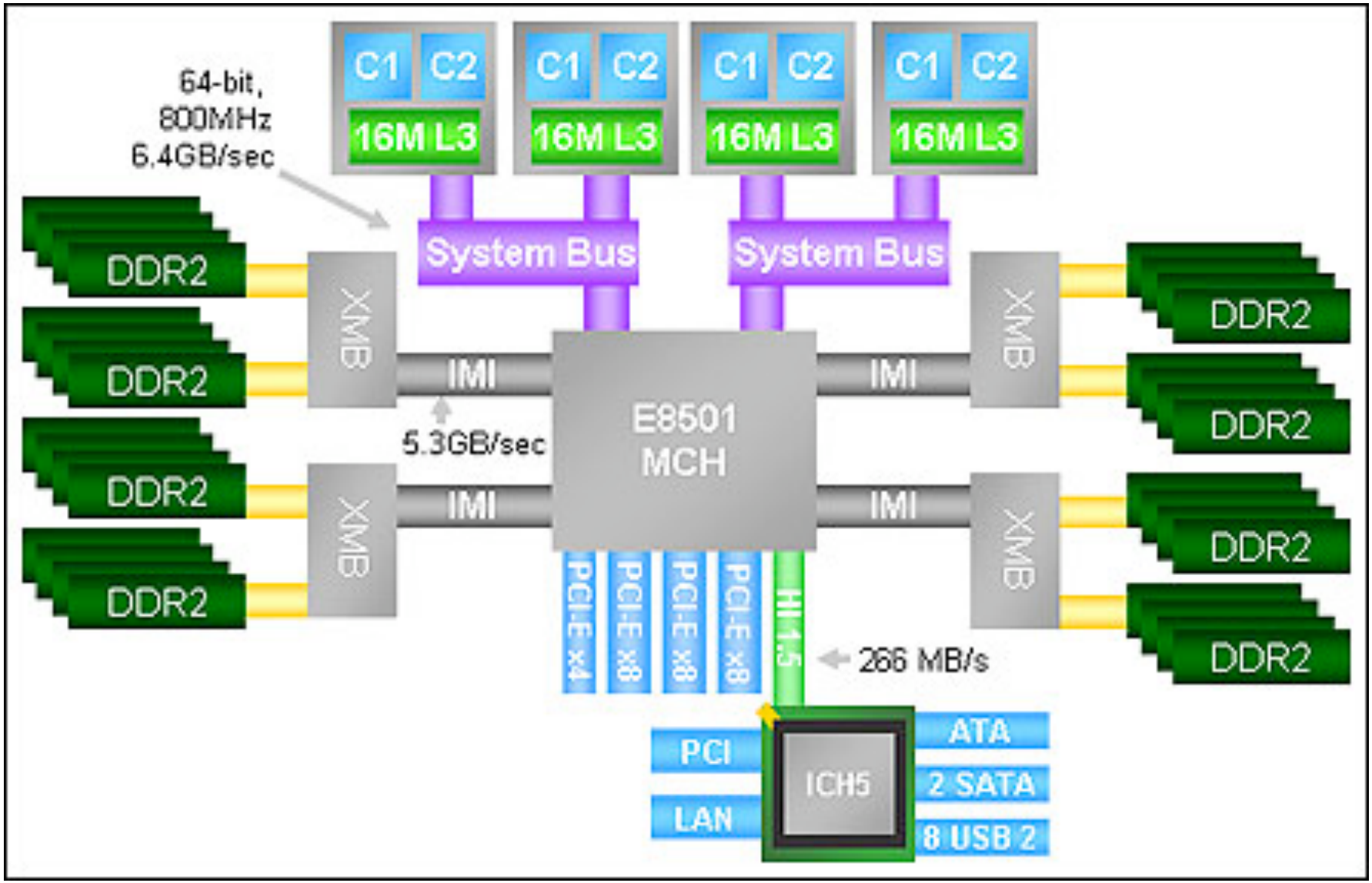


Chipset (Intel/ICH)

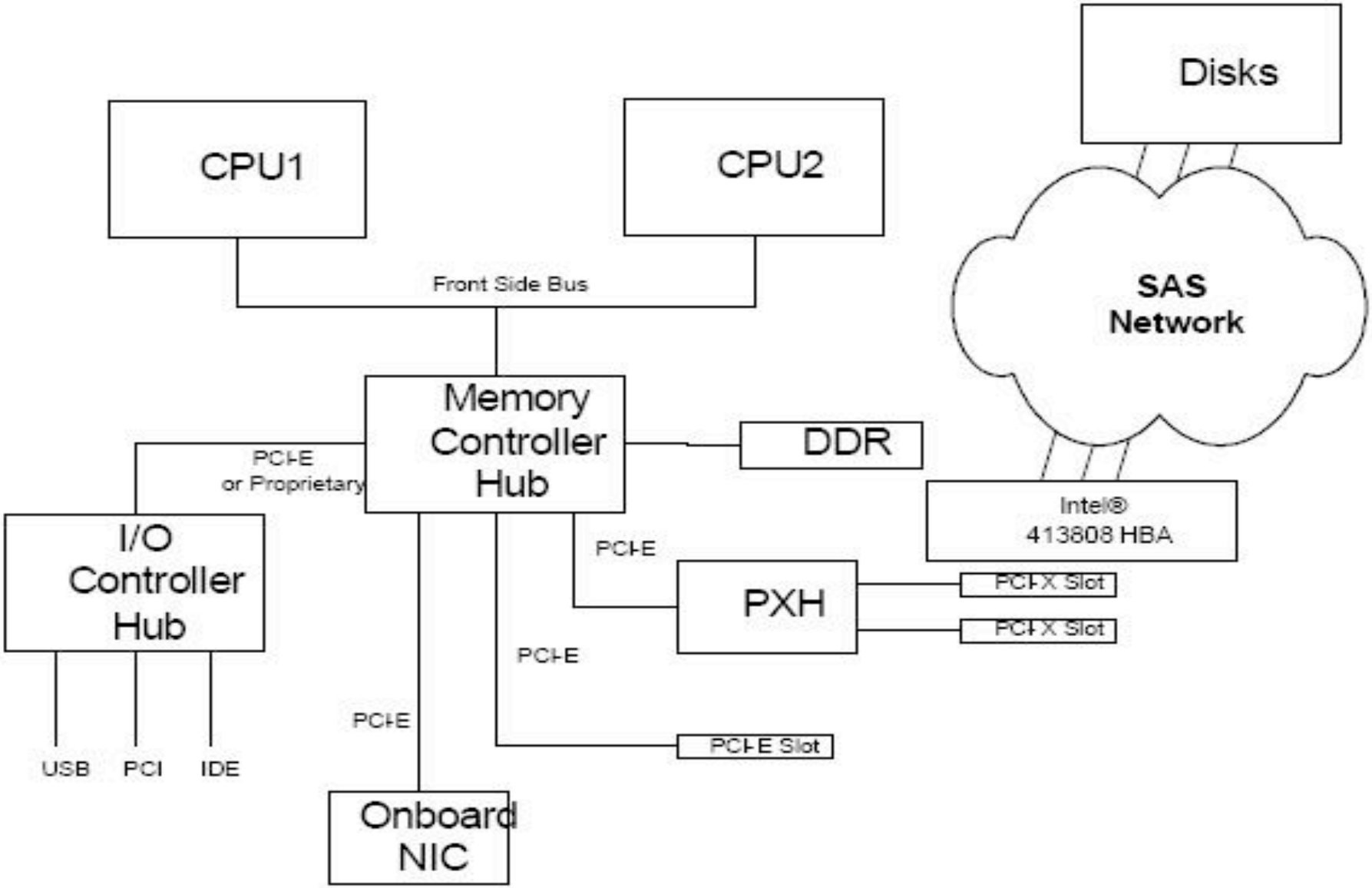
Chipset (AMD/NVIDIA)



Chipset (Intel/MCH)



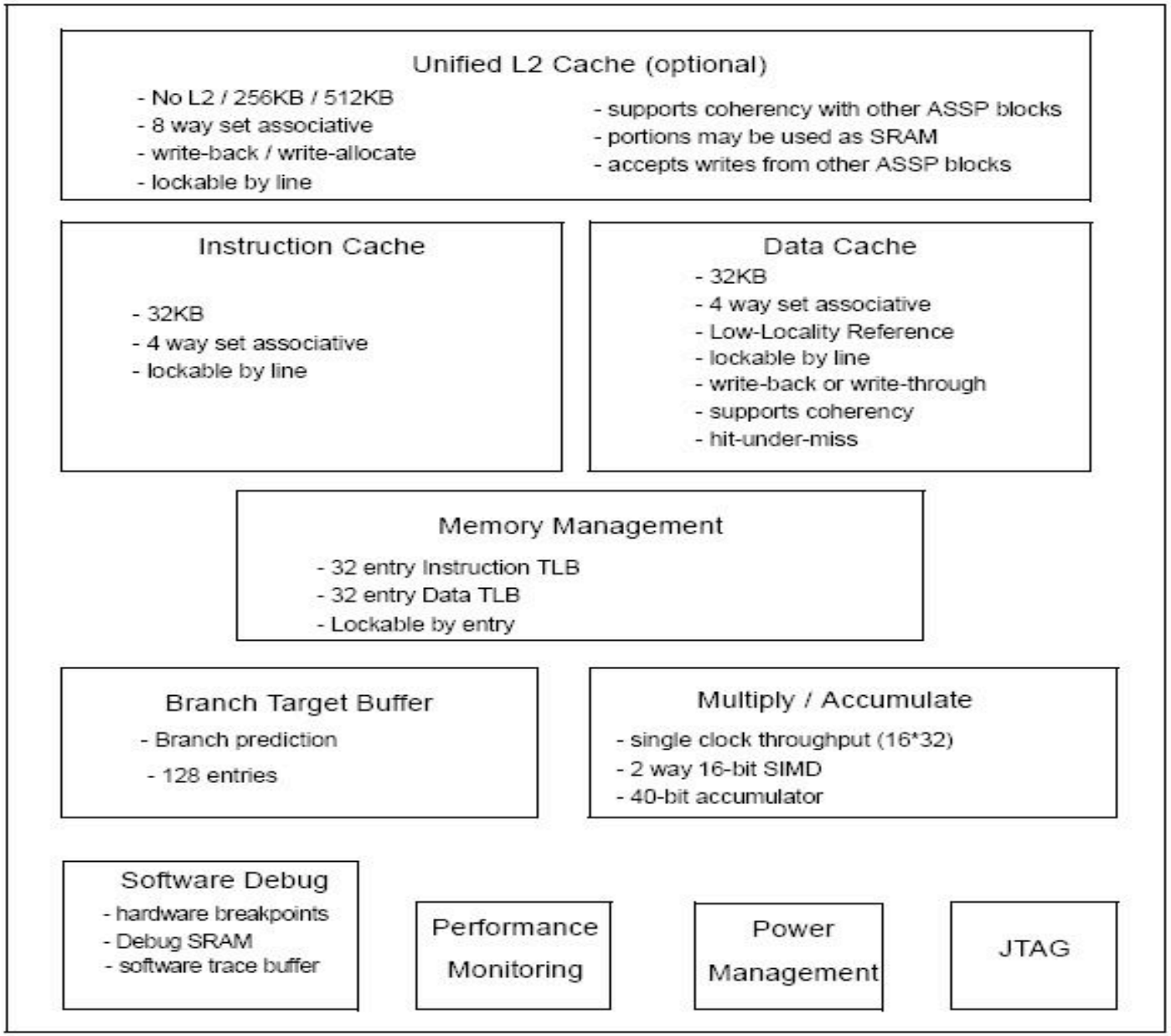
Host-Based Adapter



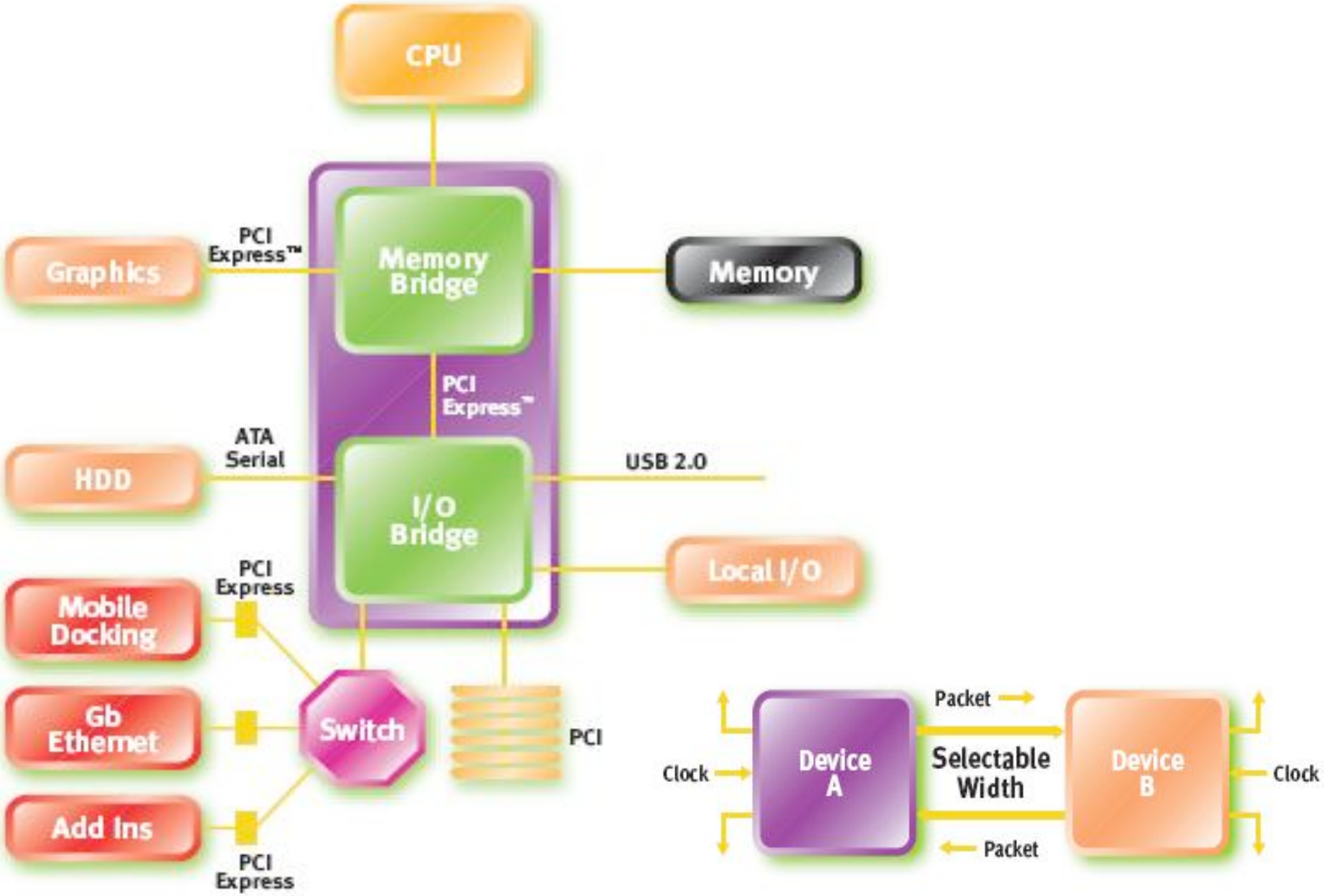
Intel 81348 Features (I)

- Host Interface:
  - PCI-X _or_ PCI-Express
  - 2-function PCI device: (ATU + MU, TPMI)
- Intel XScale Processor (ARM v5tel)
  - 2 cores, running at 1.2GHz
  - I-cache, D-cache per core (32KB each, 4-way)
  - Unified L2 cache (512KB, 8-way)
  - Inter-Processor Messaging Unit
- Internal Busses (North, South):
  - 128-bits wide, running at 400 MHz
  - Internal Bus System Controller: internal address bus arbitration, internal data bus arbitration, framing Address bus cycles, framing Data bus cycles, shared address & data paths
- DDR memory controller
- Timers:
  - 2 programmable timers per processor, 1 watchdog timer per processor
- I2C Bus Interface, 2 UART's, 16 GPIO, Peripheral Bus Interface (PBI), Performance Monitoring Unit (PMI)

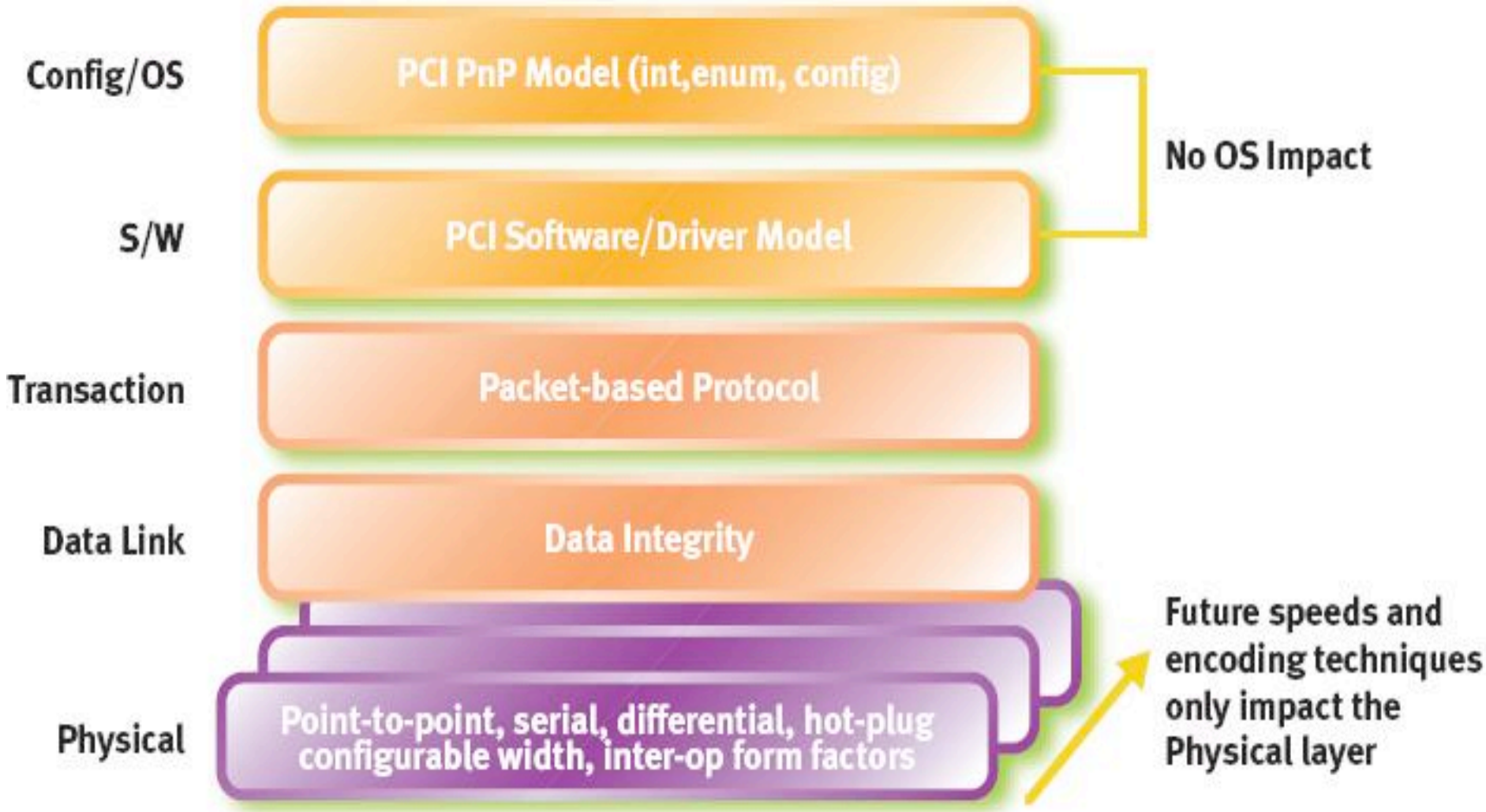
XScale Microarchitecture



PCI Express: Switch + Links



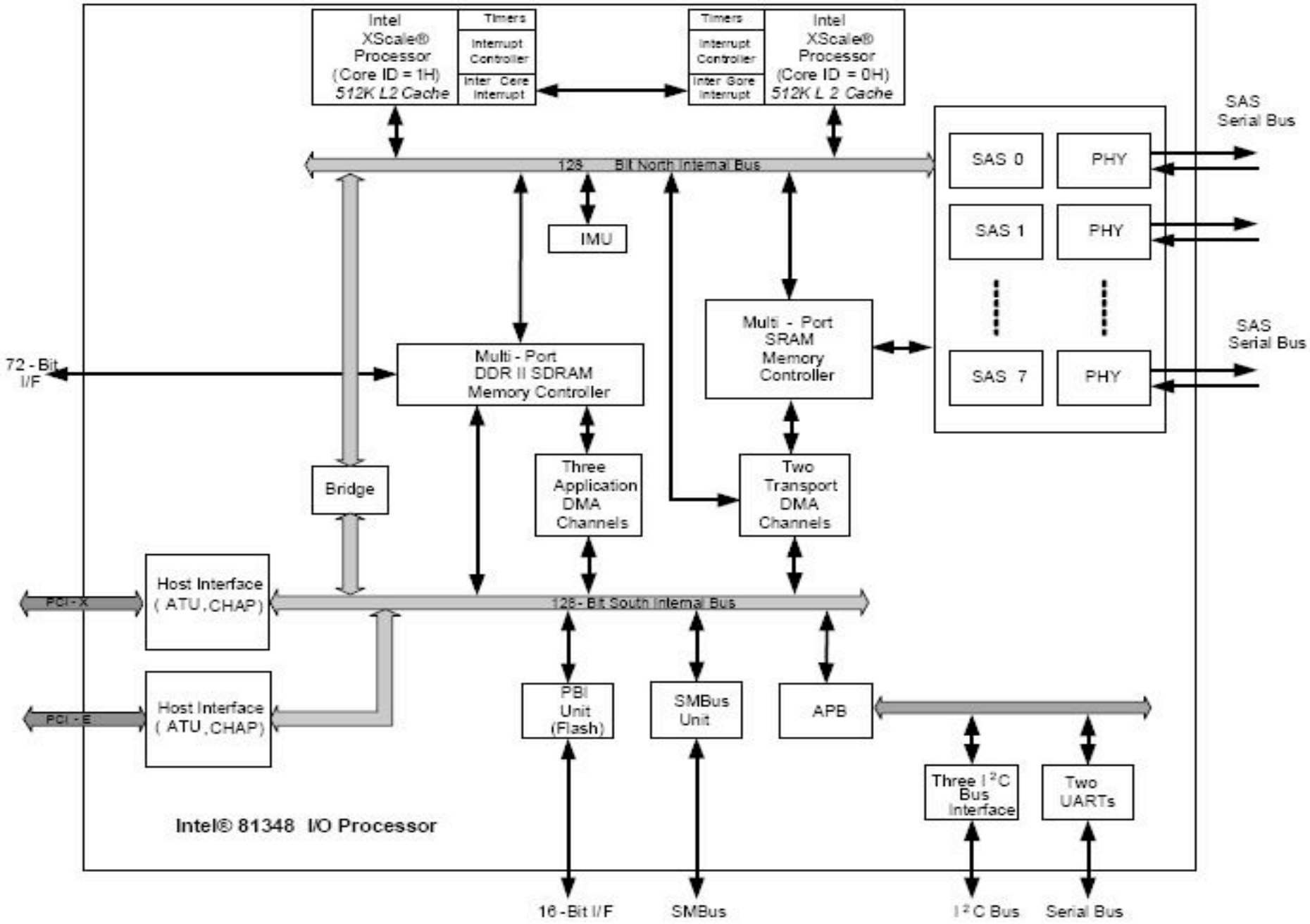
PCI Express: Layers



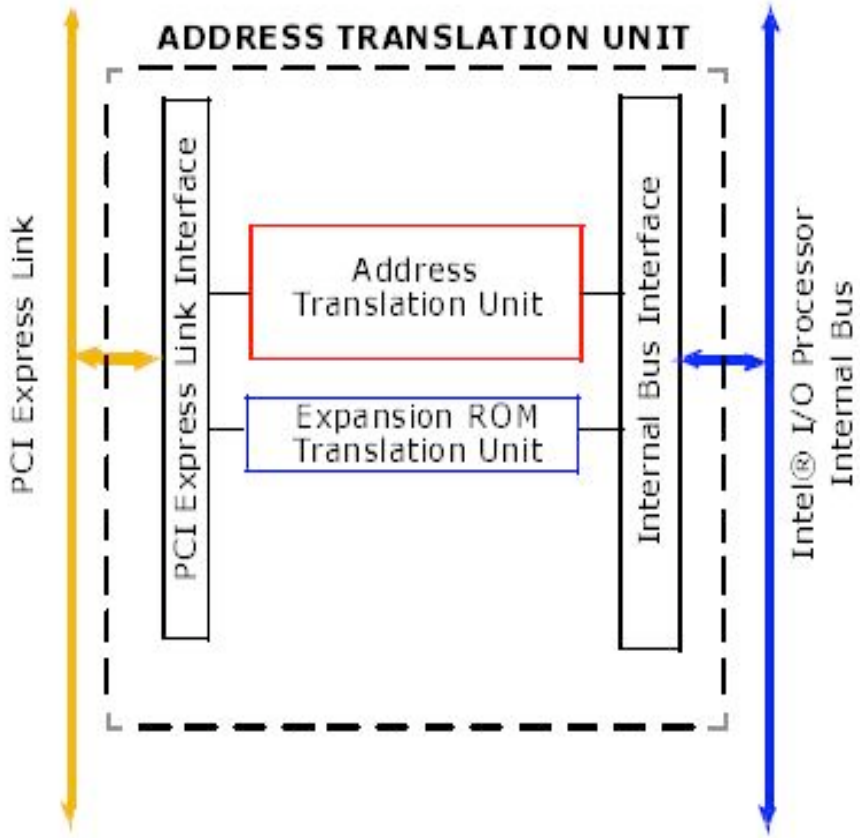
Intel 81348 Features (II)

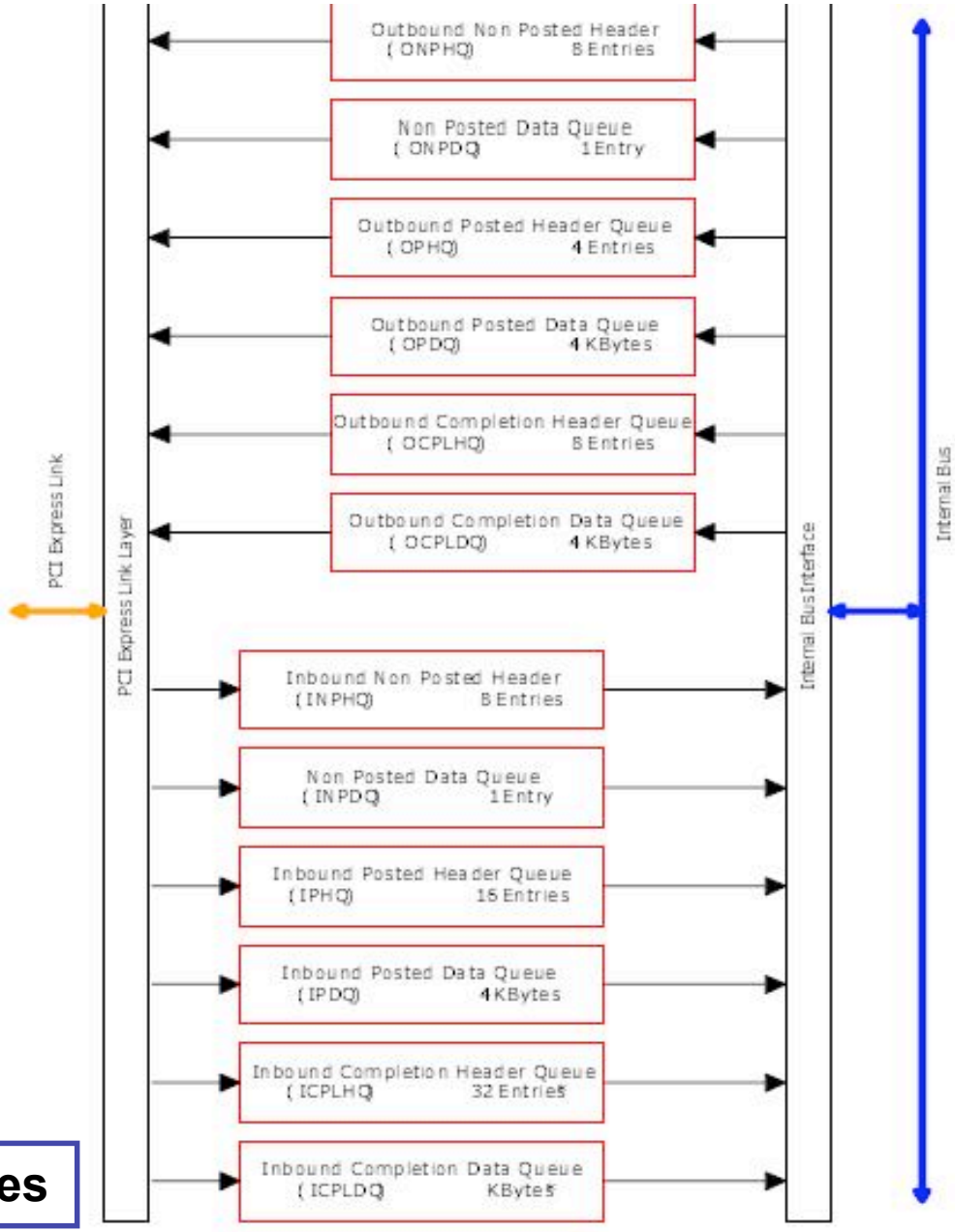
- 3 Application DMA Channels (ADMA)
  - Dual-ported: South Bus – SDRAM
  - Support L2 cache coherence
- Address Translation Unit (ATU)
  - Allow PCI Tx's direct access to local DDR SDRAM
  - Programmable registers to control address translation
- Messaging Unit (MU)
  - Data transfers between PCI system & 81348
    - Message passing, interrupt generation
  - Interrupts to notify each system when new data arrives
- FSENG block: 8 SATA/SAS engines

IOP348: Functional Blocks



Address Translation Unit (ATU)



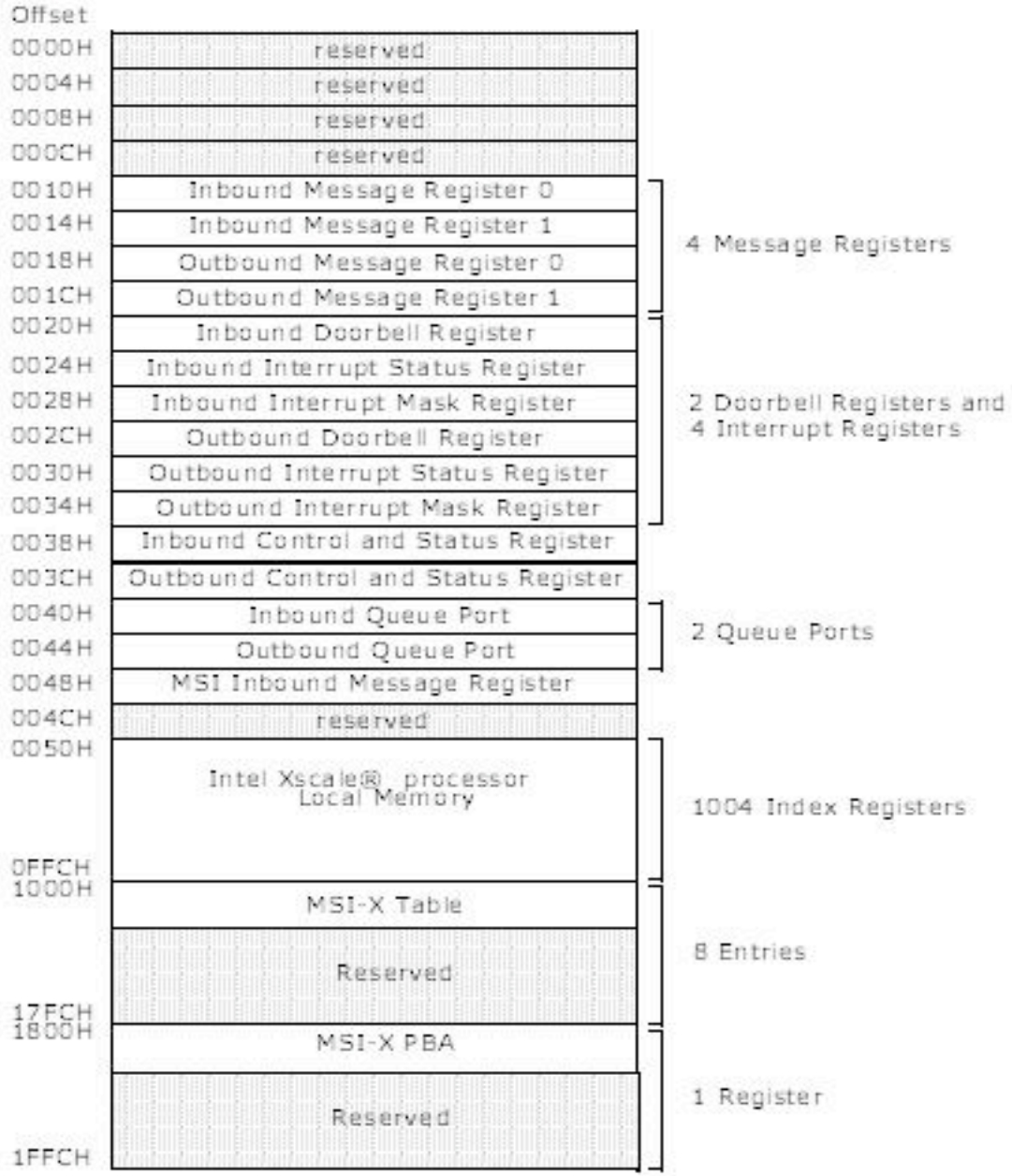


ATU Queues

Messaging Unit (MU)

The MU is accessed by an external PCI agent via ATU.

- Message registers
- Doorbell registers
- Circular queues
- Index registers



MU: inbound & outbound queues

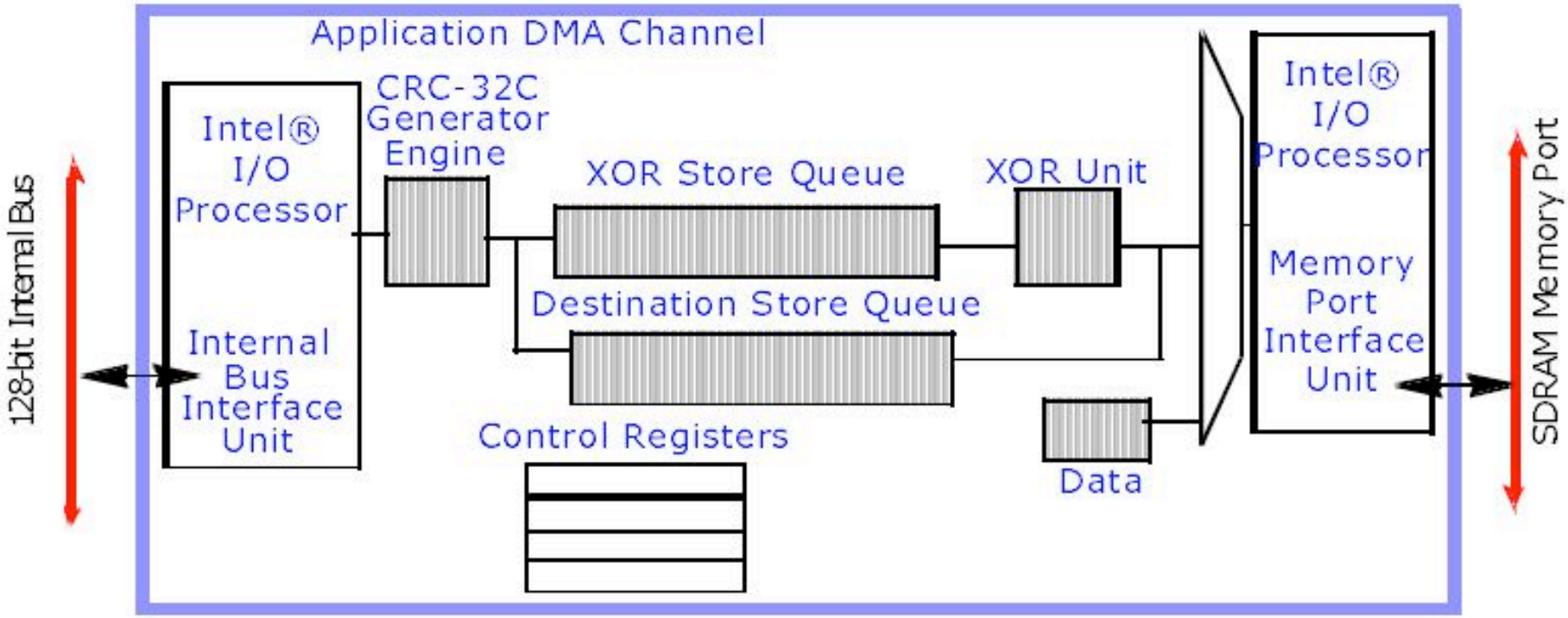
Inbound:

Queue Mnemonic	Queue Name	Queue Size (Bytes)
IWQ	Inbound Write Data Queue	4 KBytes (4*1KB)
IWADQ	Inbound Write Address Queue	4 Transaction Addresses
IRQ	Inbound Read Data Queue	4 KBytes (4*1KB)
IDWQ	Inbound Delayed Write address/data Queue	1 Transaction
ITQ	Inbound Transaction Queue	8 Addresses/Commands

Outbound:

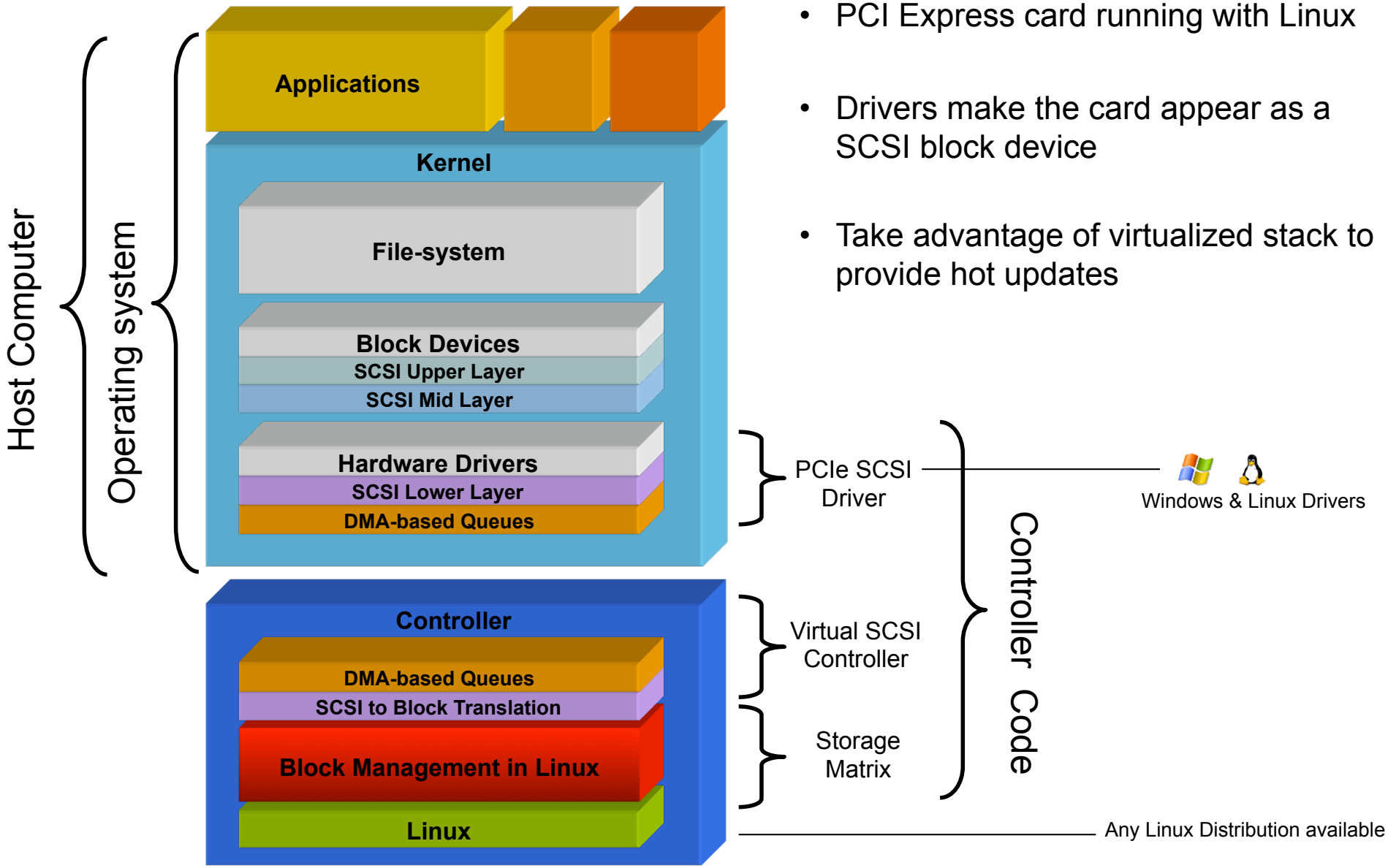
Queue Mnemonic	Queue Name	Queue Size (Bytes)
OWQ	Outbound Write Data Queue	4 KBytes (4*1024B)
OWADQ	Outbound Write Address Queue	4 Transaction Addresses
ORQ	Outbound Read Data Queue	2 or 4 KBytes (4* 512B or 4*1024B)[2]
OTQ	Outbound Transaction Queue	8 Addresses/Commands

Application DMA Channel (ADMA)



API: chain of Descriptors:
(SRC, DST, byte-count, control-bits) + link to next descriptor

IOP348 - Storage System Layers



- PCI Express card running with Linux
- Drivers make the card appear as a SCSI block device
- Take advantage of virtualized stack to provide hot updates

IOP348 - Controller Glue