Storage Basics

Leah Schoeb, Member of SNIA Technical Council

SNIA Emerald™ Training

SNIA Emerald Power Efficiency Measurement Specification, for use in EPA ENERGY STAR®

July 14-17, 2014

SNIA Green Storage Initiative

SNIA Emerald™

Course Information

- **Who should attend?**
  - Information Technology professionals
  - Engineers
  - Consultants
- **Objectives – what you will learn**
  - Basics of enterprise storage technology
  - What are the initiatives for optimizing the data center
  - Current efficiency technologies used in storage
  - Understand Storage Performance basics
  - IO Generation tools are not all created equal

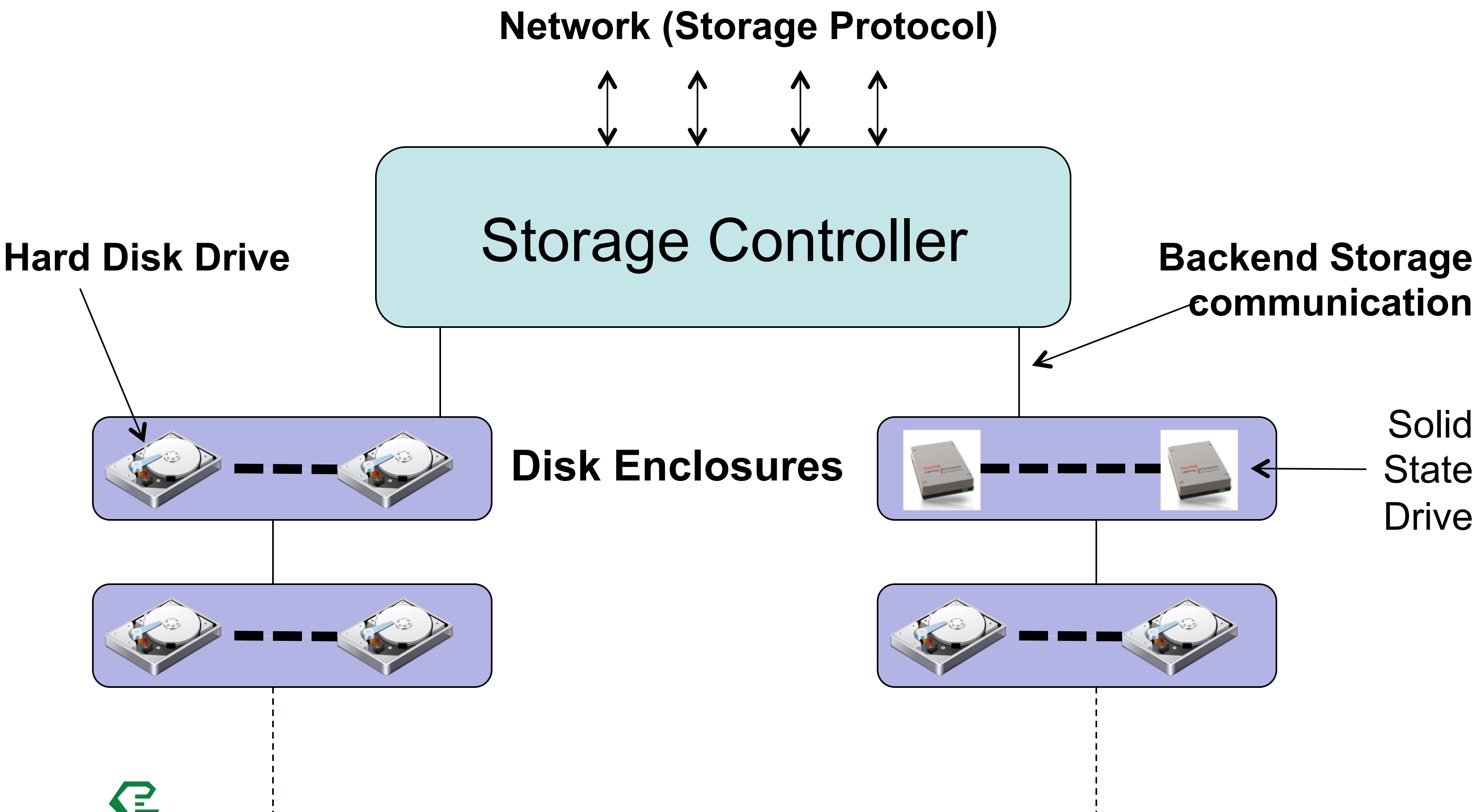
Agenda

- Storage 101
- Enterprise storage
- Enterprise Storage Performance and load generation
- Capacity Optimization
- Q&A
- Note: I will give time at the end for each section for review and update notes.

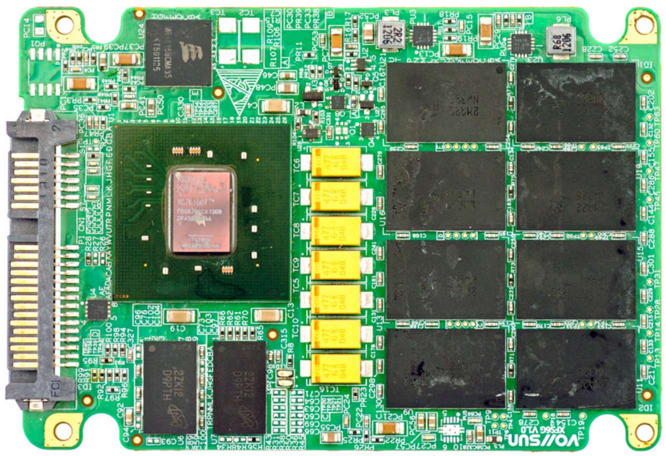
Storage 101

SNIA™
Green Storage Initiative

Typical Storage System Architecture

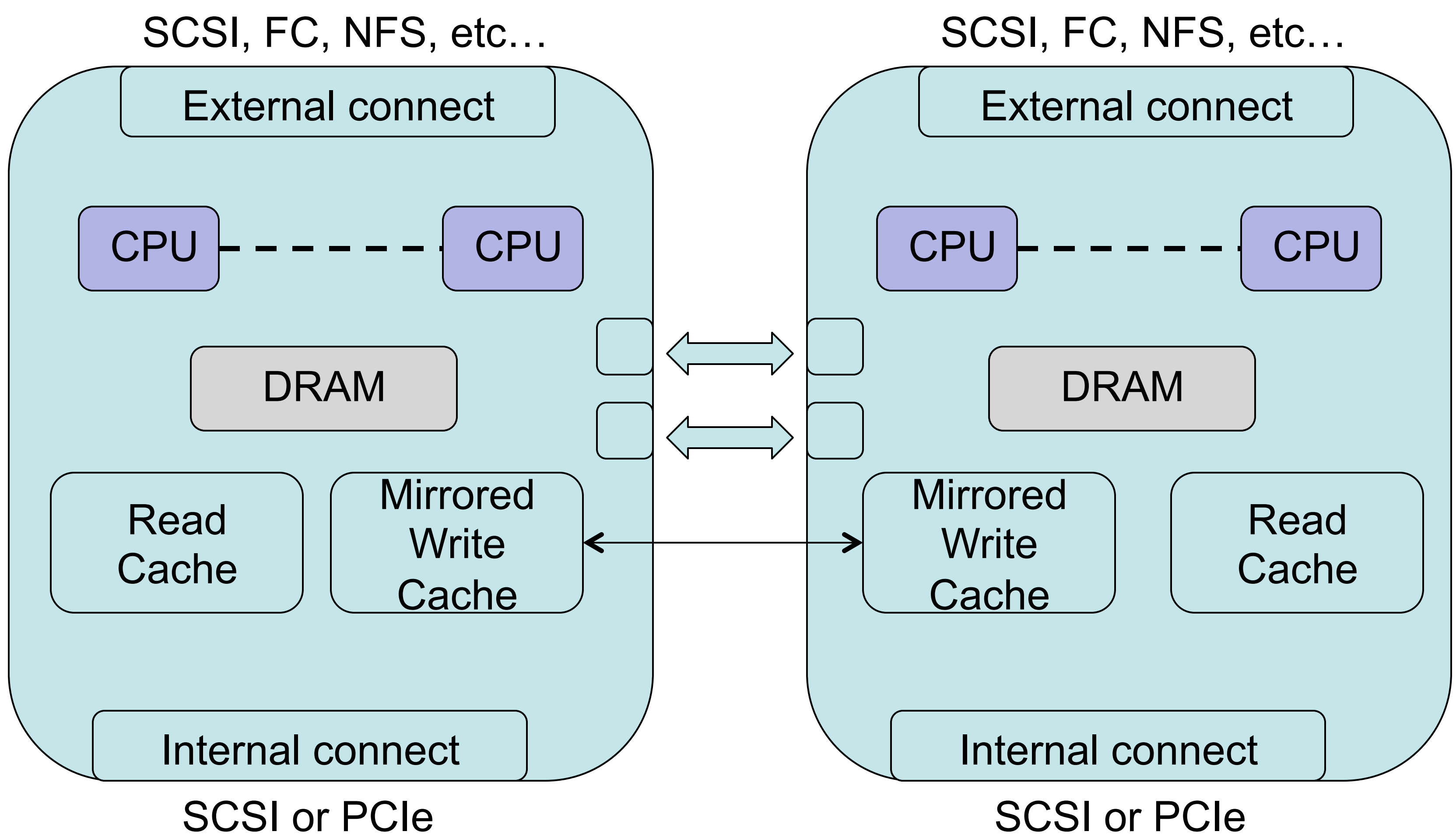


SNIA Green Storage Initiative

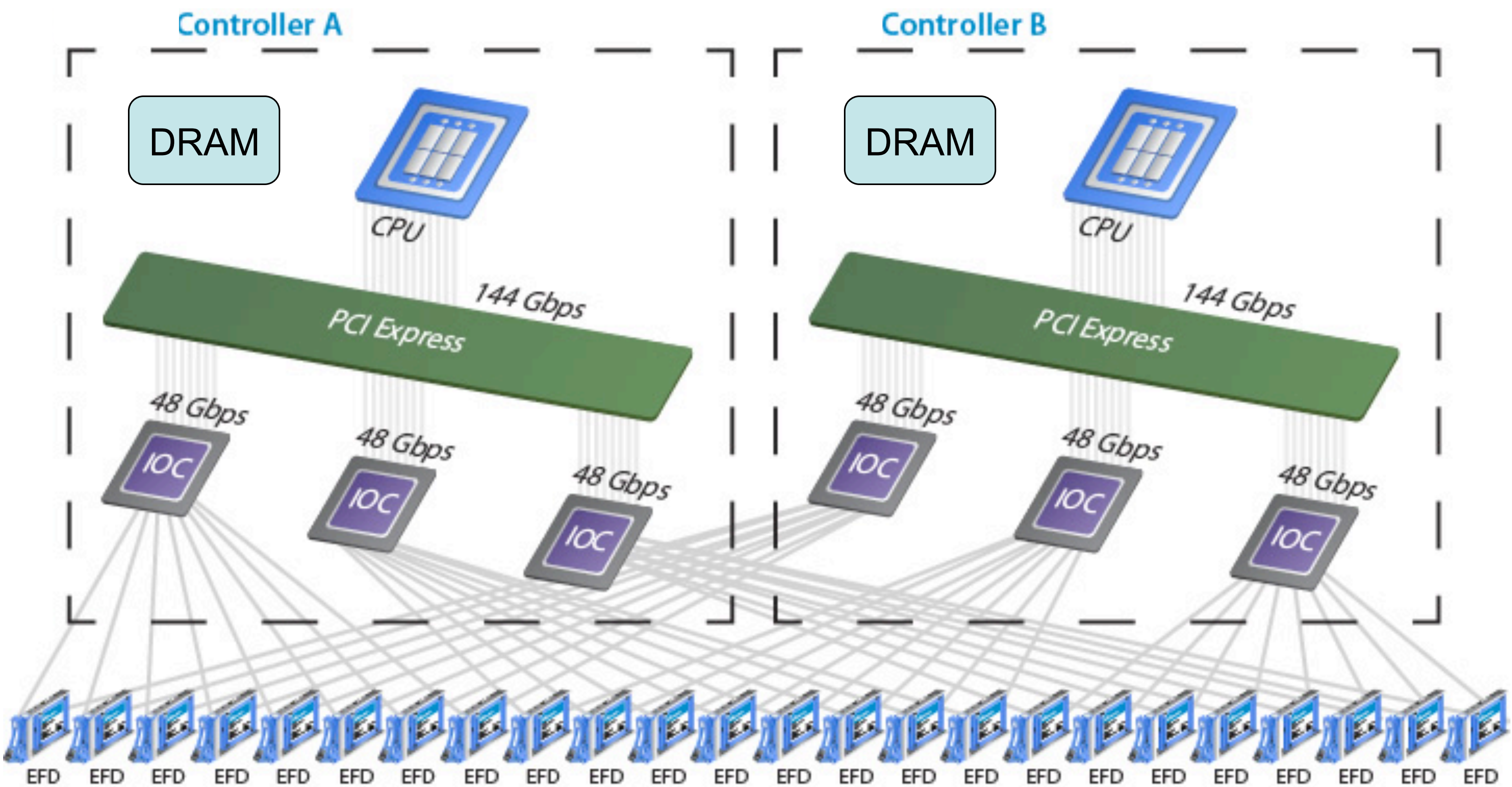


STORAGE CONTROLLER

Traditional Dual Storage Controller



Dual Flash Controller



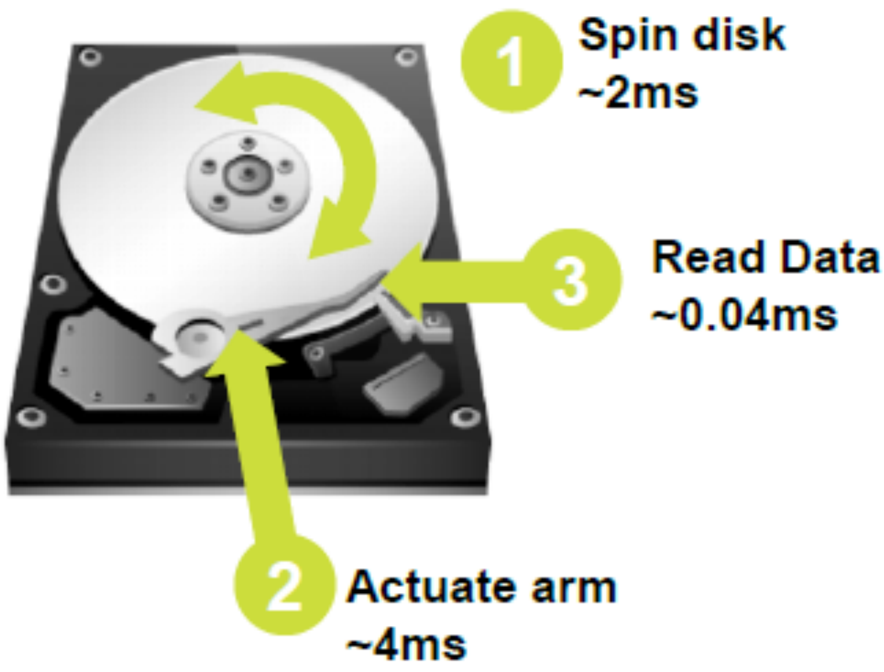
SNIA™
Green Storage Initiative

HARD DISK DRIVES

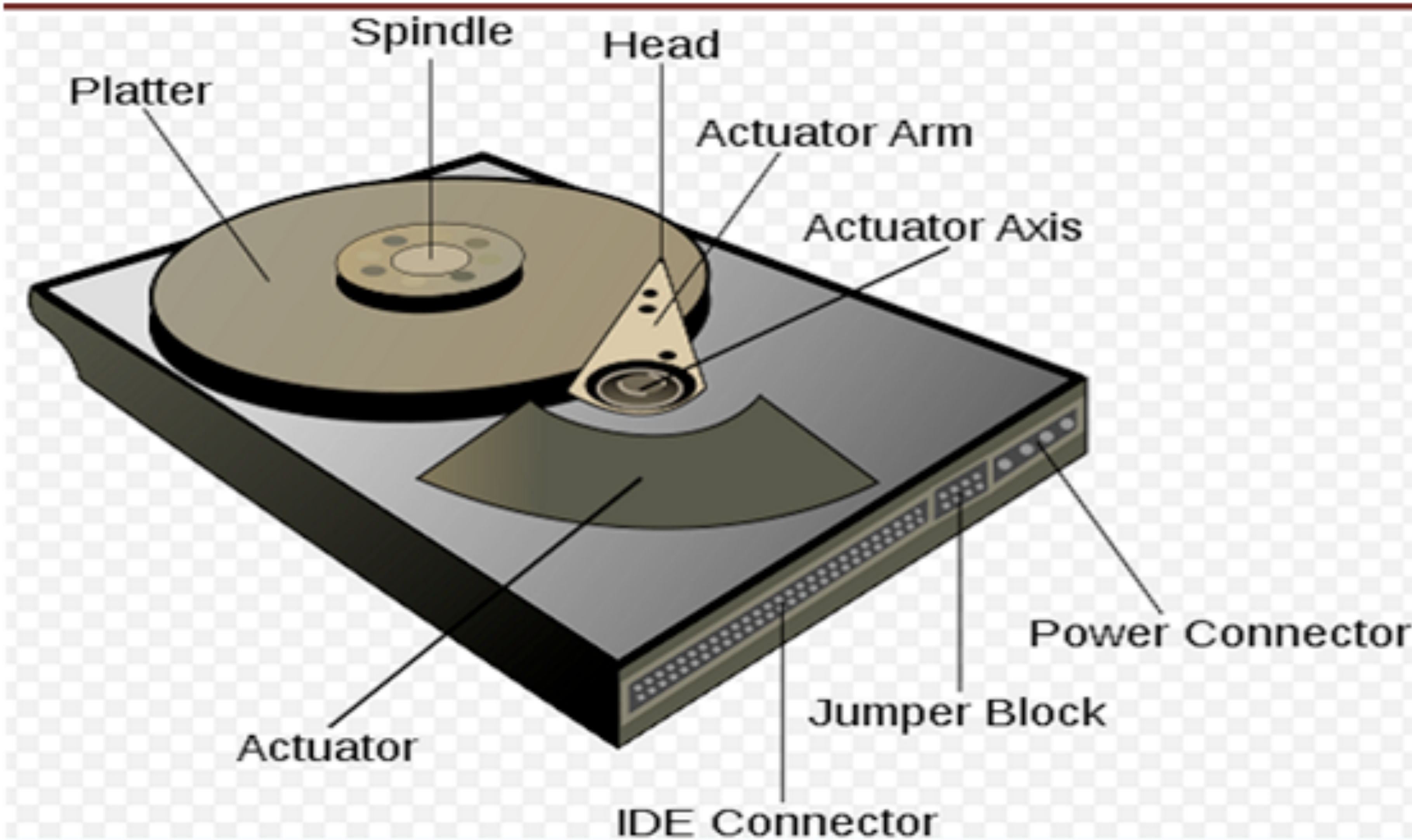
Hard Disk Drives (HDD)

- Electro-mechanics
- Disk storage uses spin motors and actuators
- Electro mechanical devices are limited by the mechanics
- Mechanisms wear, generate heat, consume power

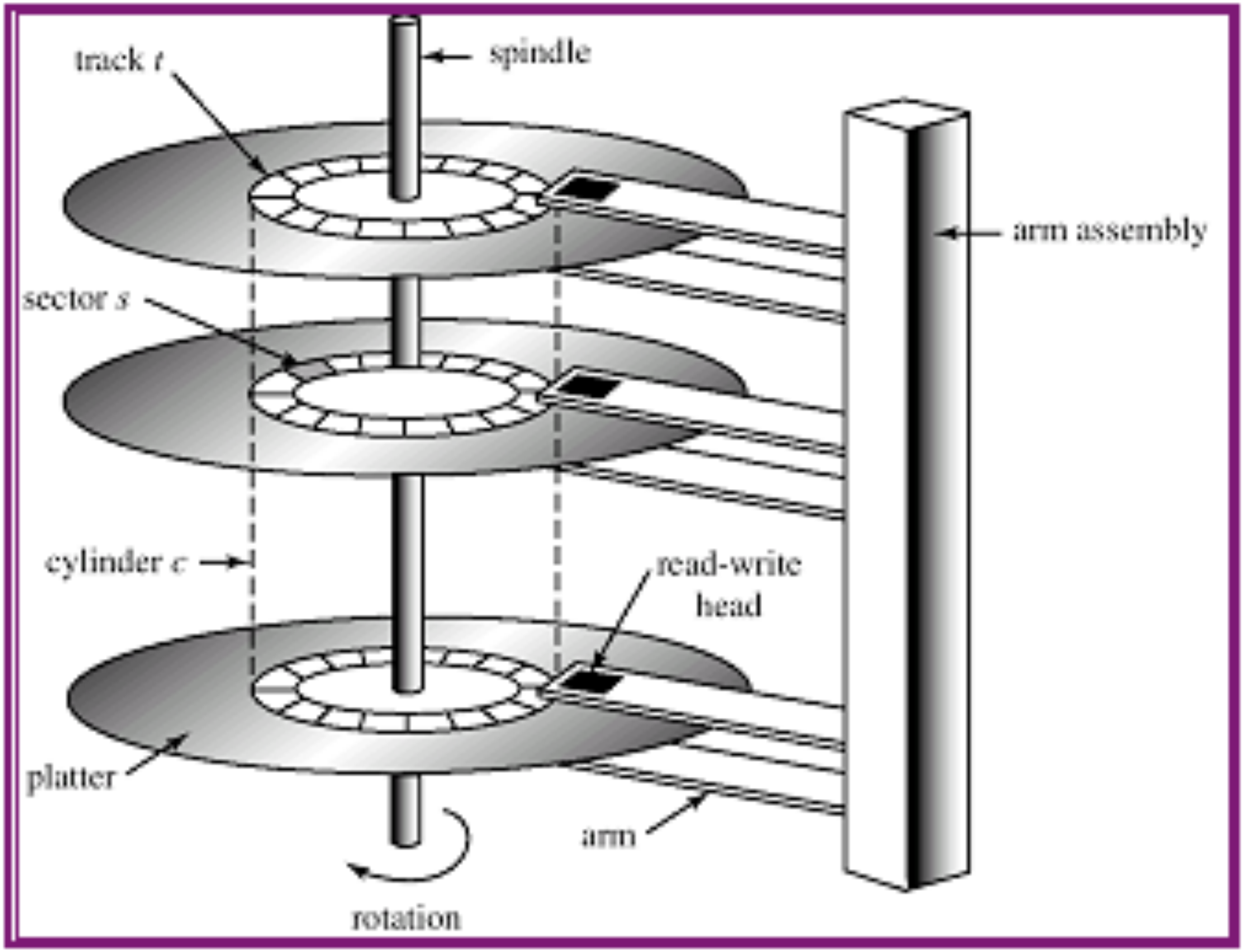
Read Operation



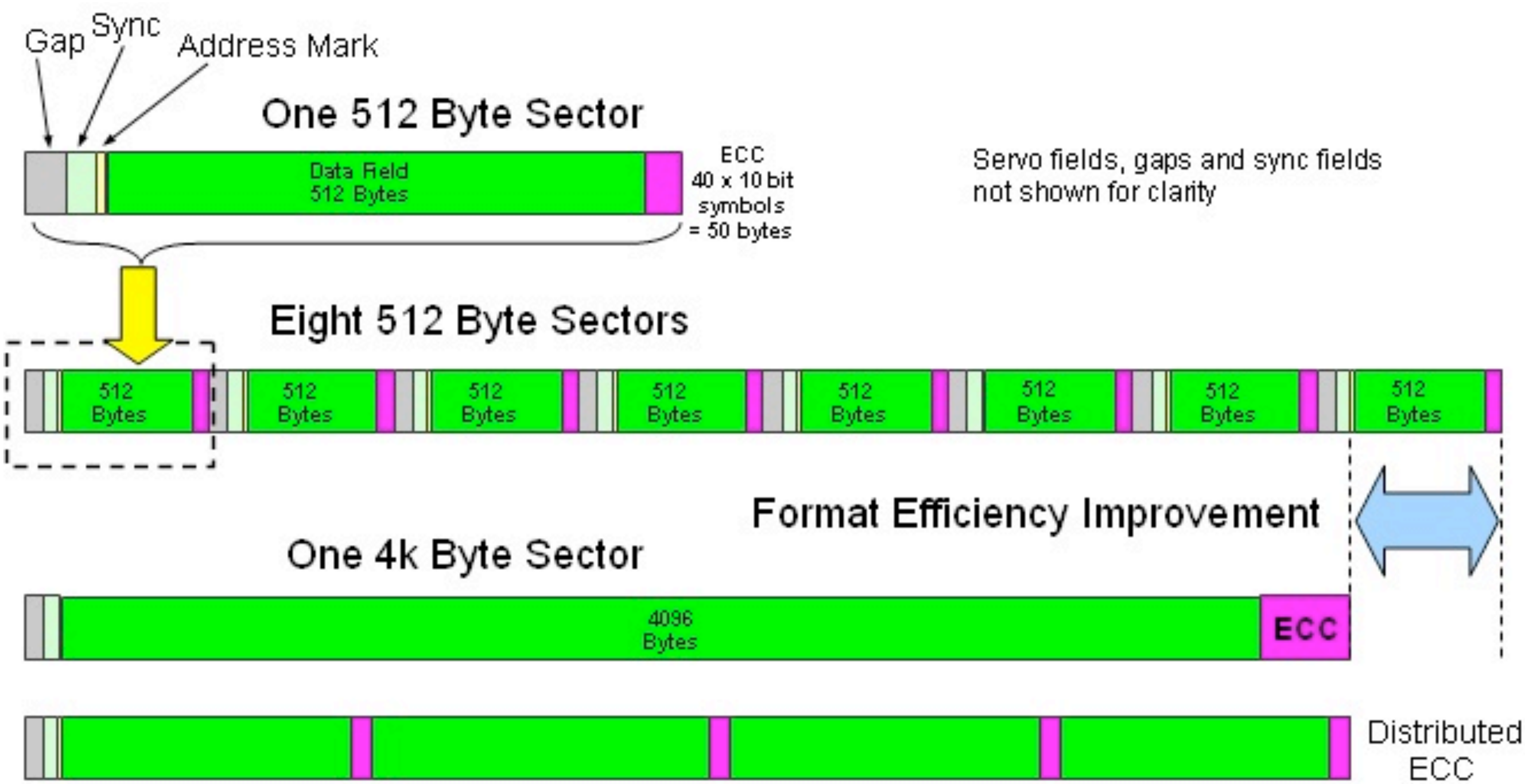
HDD Components



Multiple Platters, Cylinders, Tracks, and Sectors



Disk Sectors



Internal vs. External HDD

- Internal
  - fit inside a desktop computer, laptop, disk array enclosure, or server array
  - they are made to be enclosed in a computer or server
  - Internal hard drives can be sold separately and stored in an enclosure, which usually houses multiple hard drives
- External
  - work a little differently
  - designed for portability
  - Portable external drives can easily fit into cases, bags, and backpacks

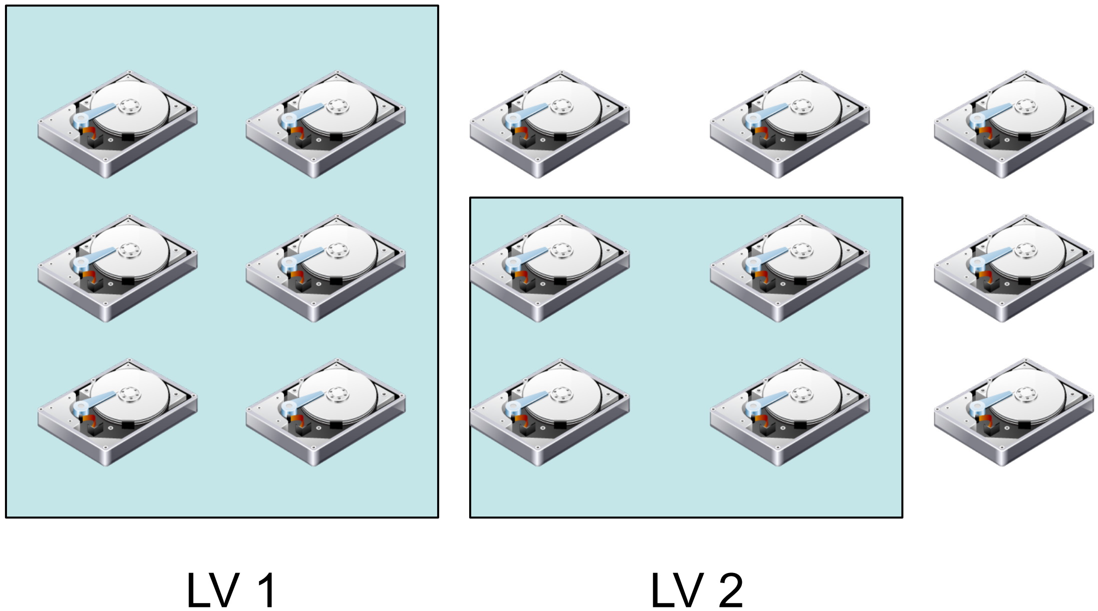
Rotating Media Selection

SNIA
Green Storage Initiative

Drive Type	Speed RPM	MB/sec	IOPS	Latency	LC Manage
FC 4Gb	15k	150	200	5.5ms	High Perf. Trans
FC 4Gb	10k	75	165	6.8ms	High Perf. Trans
SAS (6Gb, 12Gb)	10k	150	155	12.7ms	High Perf. Trans
SAS (6Gb, 12Gb)	15k	150	185	12.7ms	High Perf. Trans
SATA (6Gb,16Gb)	7200	140	38	12.7ms	Streaming/Nearline
SATA (6GB, 16Gb)	5400	68	38	12.7ms	Nearline

Disk Pools

- Logical volume
  - Pool disks together
  - Create one virtual disk
- Created either at server or storage controller levels
- Easier Management
- Creating RAID groups to protect from data loss

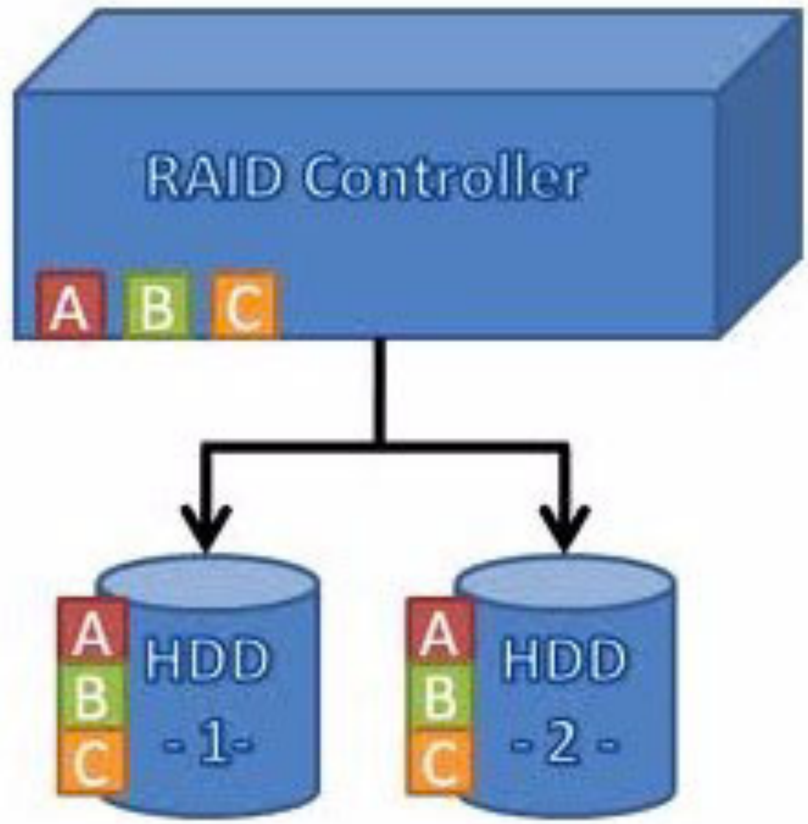


RAID Level Protection

- RAID – Redundant Array of Independent Disks
- RAID 0 - Striping
- RAID 1 - Mirroring
- RAID 3 – Striping + parity
- RAID 5 – Distributed Parity
- RAID 6 – Distributed Double Parity
- RAID 10 (0+1) – Combination of striping and mirroring

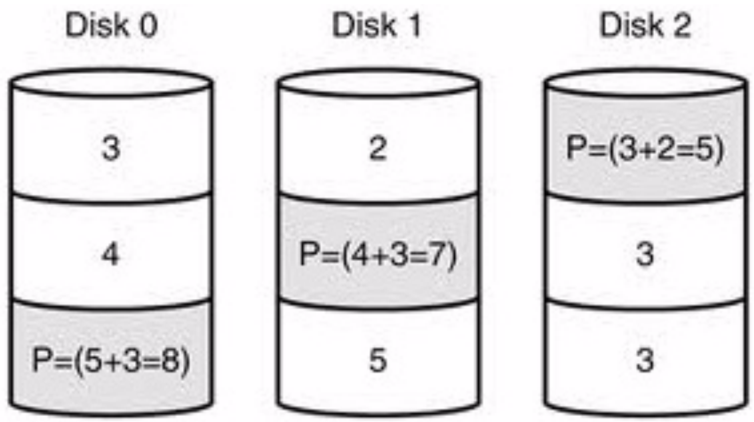
RAID 1- Mirroring
RAID 10 – Mirroring + striping

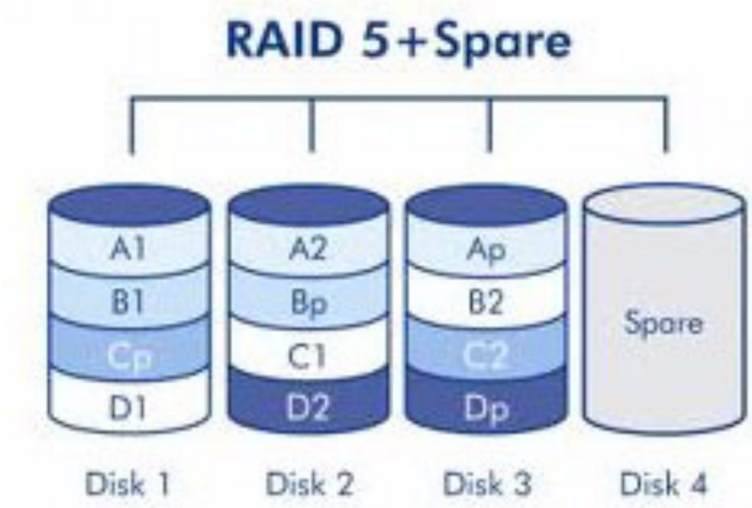
- 50% capacity utilization
- Mirrored copy in case of failure
- One read operation
- Two writes operations
- Expensive but faster
- Mirror + strip – 2 copies of the data distributed evenly across disks



RAID 5 – Distributed Parity
RAID 6 – Double Distributed Parity

- Uses Parity – Information is distributed to all disks in a RAID grouping
- Block Fails – parity information will recover the data
- Disk Fails – rebuild using parity data
- Minimum 3 disks to make a RAID 5 grouping
- RAID 5 + spare
- RAID 6 – double parity





SOLID STATE STORAGE

Technology Progression

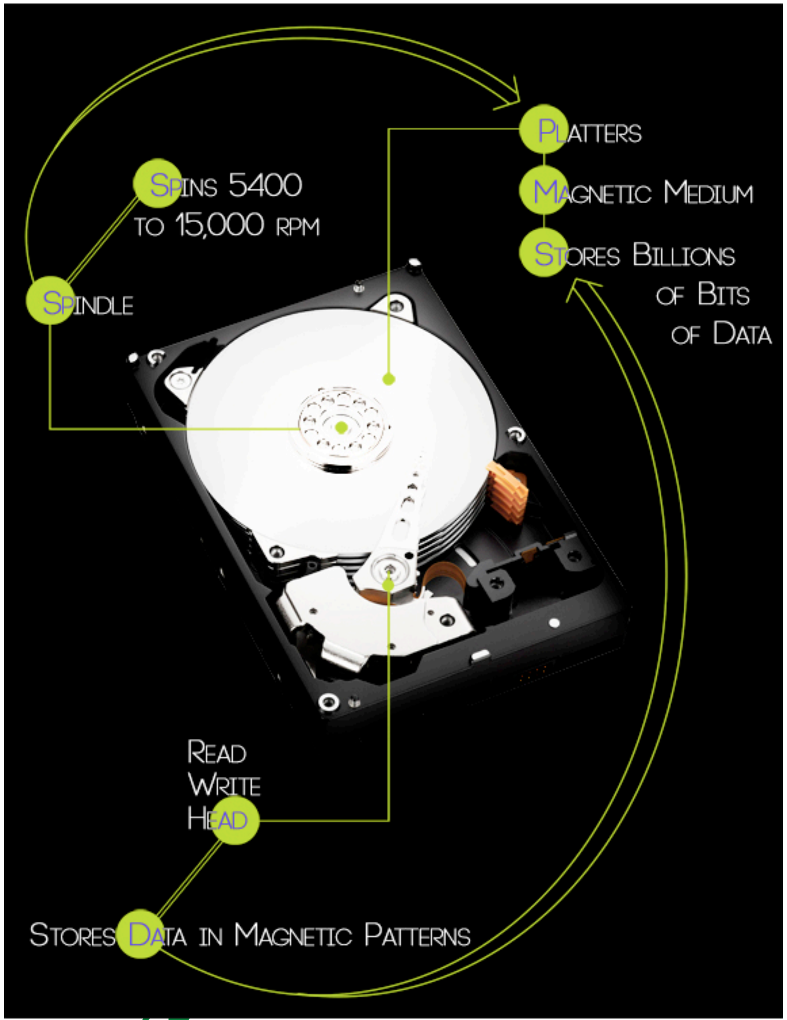
	1990	2010	Improvement
CPU	**0.05** MIPS/$	**147** MIPS/$	**2940x**
Memory	**0.02** MB/$	**25** MB/$	**1250x**
Addressable Memory	2^{16}	2^{64}	2^{48}x
Network Speed	**100** Mbps	**100** Gbps	**1000x**
Disk Data Transfer	**5** MB/sec	**130** MB/sec	**25x**

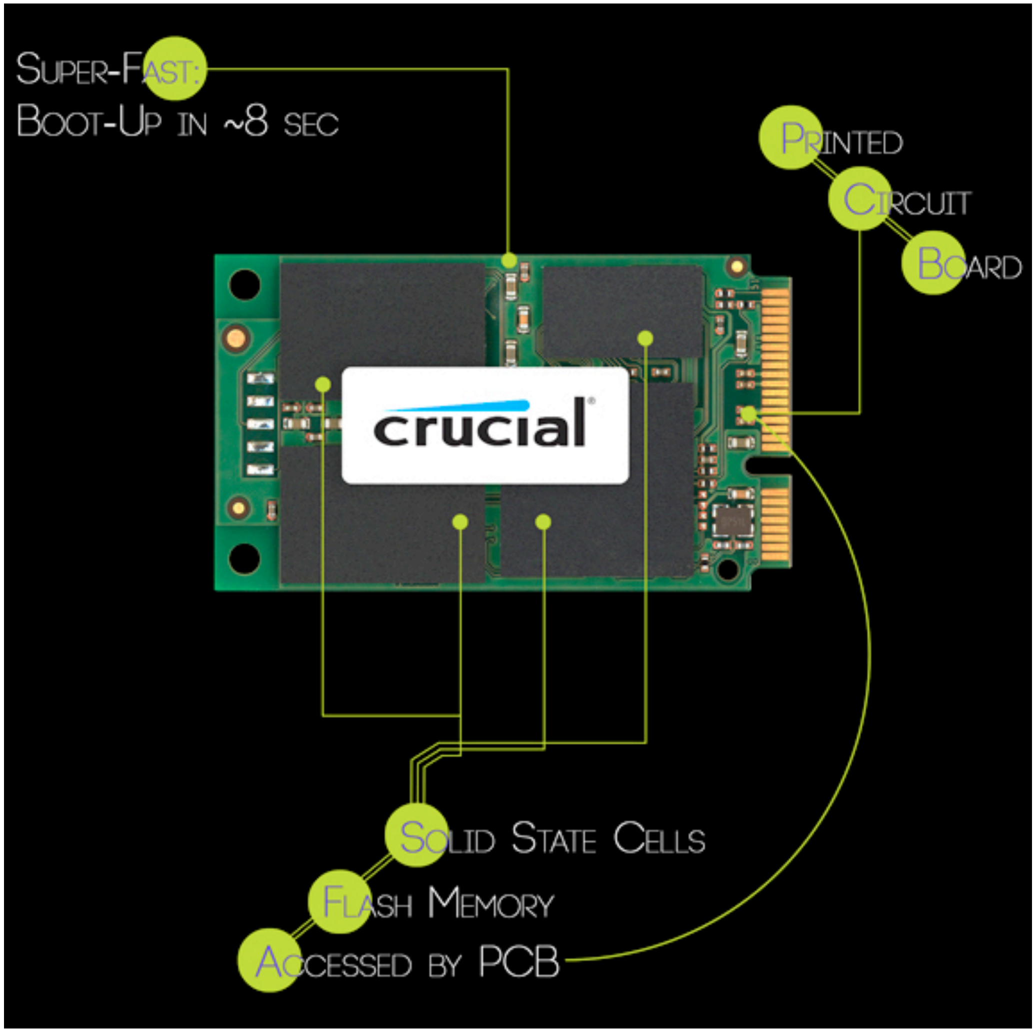
Why Solid State Technology?

SNIA
Green Storage Initiative

	Hard Disk Drive	Solid State (NAND Flash)	Increase
Performance (IOPS) Transaction Processing	200	6000	30x
Response Time	2-5 milliseconds	100-500 microseconds	1,000x

HDD vs SDD


Platters
Magnetic Medium
Stores Billions of Bits of Data
Spins 5400 to 15,000 RPM
Spindle
Read Write Head
Stores Data in Magnetic Patterns


Super-Fast: Boot-Up in ~8 sec
Printed Circuit Board
crucial
Solid State Cells
Flash Memory
Accessed by PCB

Hybrid Drives

- Both HDD and SSD
- NAND Flash Technology
- Adding speed w/ SSS
- Cost-effective capacity of HDD
- SSD acts as a cache for most frequently used data
- Data stored on HDD
- Can improve overall performance

Solid State Storage

- No all SSDs designed the same
  - NAND-based flash memory
  - DRAM-based (Random Access Memory)
  - Enterprise flash drives (EFDs)
  - Hybrid Drives
- Performance varies widely
  - Capacity
  - Compression
  - Wear leveling
  - Error Correction and bad block mapping
  - Metadata management
  - Garbage collection
  - Encryption

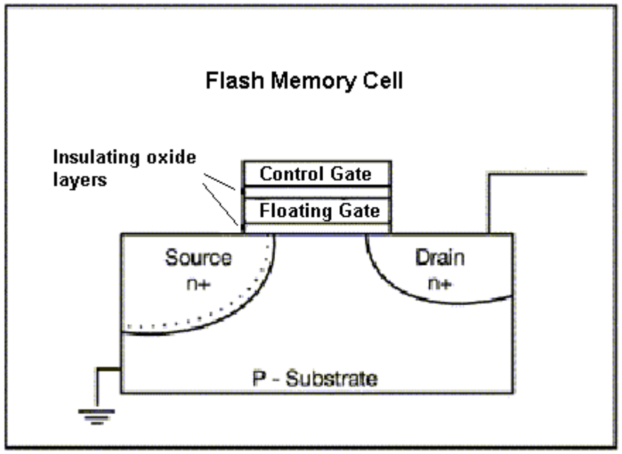
Solid State Technology – NAND Flash

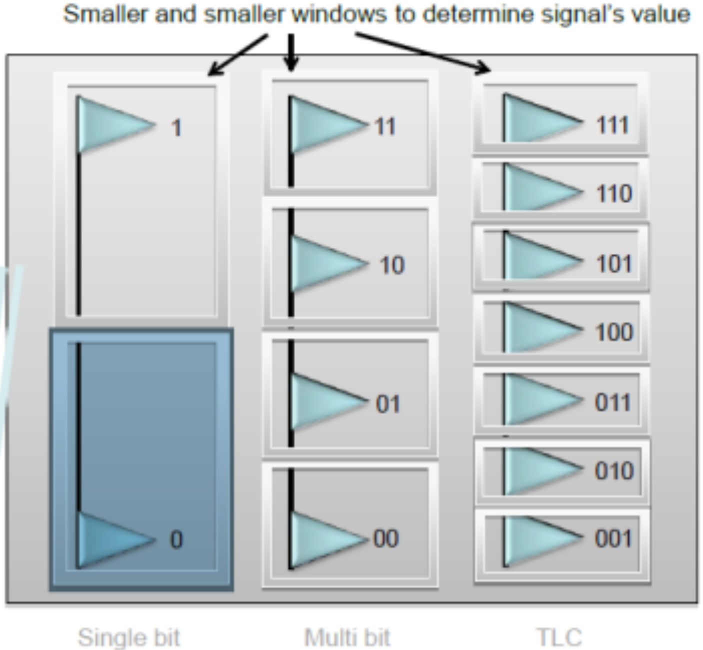
NAND Flash technology

- Continued capacity and endurance increases in technology
  - SLC – single level cell
  - MLC (and eMLC) – multi-level cell
  - TLC – triple level cell
  - 3D NAND Flash – stacked
    - greater endurance
    - 20%faster, 50% smaller
- 7 – 10 years in unpowered state

Trend – data reduction in solid state modules

- Compression & deduplication to ***multiply capacity and reduce number of writes required***





Solid State Storage

Metric	NAND Flash	
	SLC	MLC
Latency (microseconds)	100	200-300
Persistence	10x more persistent	Less reliable*
Cost	30% more expensive	More cost effective
Sequential read/writes	3x faster	Slower

*This can be overcome, even reversed by the internal design using higher over provisioning, interleaving, and changes to writing algorithms.

Solid State Storage Form Factors

SNIA
Green Storage Initiative

- **Server Side Solid State Storage (SSS)**
  - Solid State Devices (SSDs) and PCIe cards
  - Flash DIMMs
  - Caching Engines **(SW accelerators)**

PCIe SSD

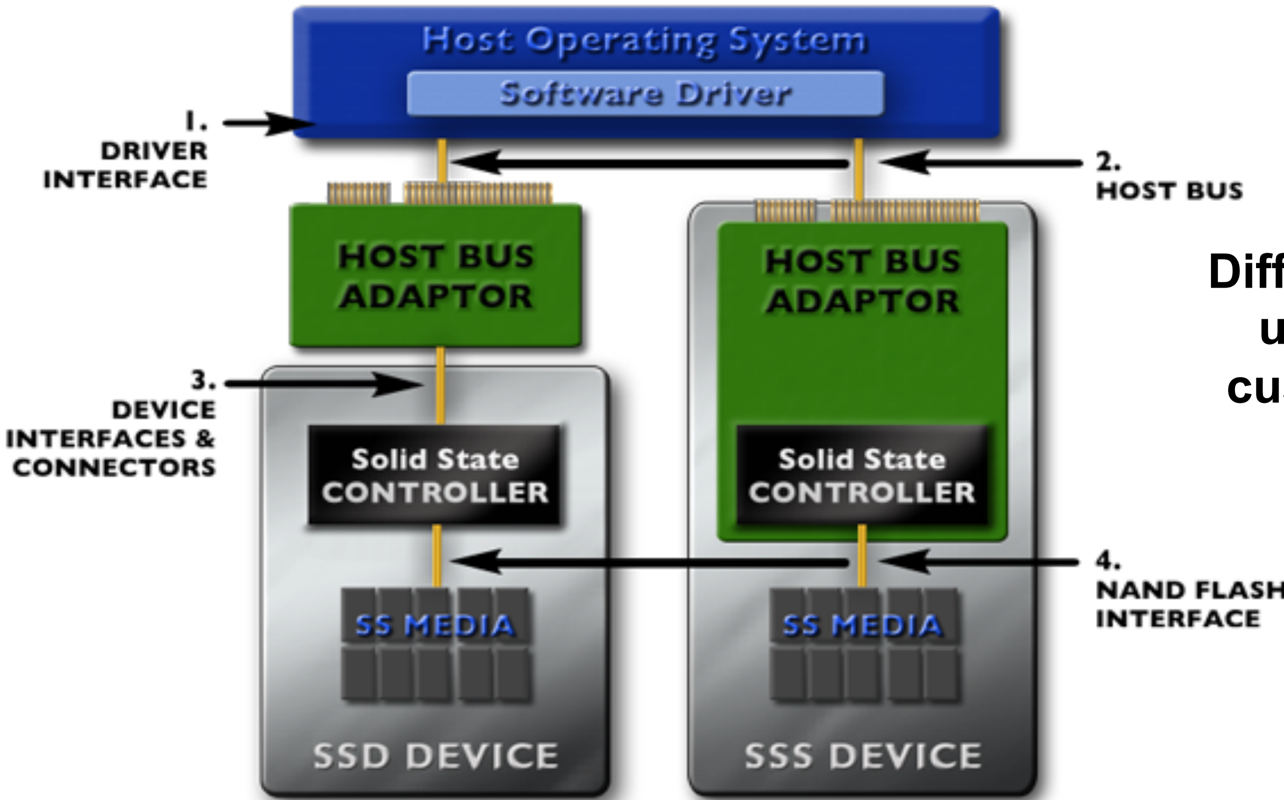
SAS, SATA. Etc… SSD

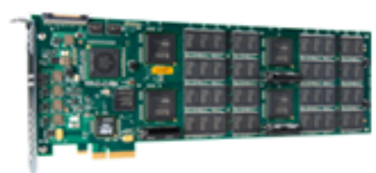
SSD DIMM

- **Solid Storage Arrays**
  - SAN SSS Arrays (FC or SAS)
  - Network SSS Arrays (NFS, SW iSCSI)

high-performance storage

Solid State Storage

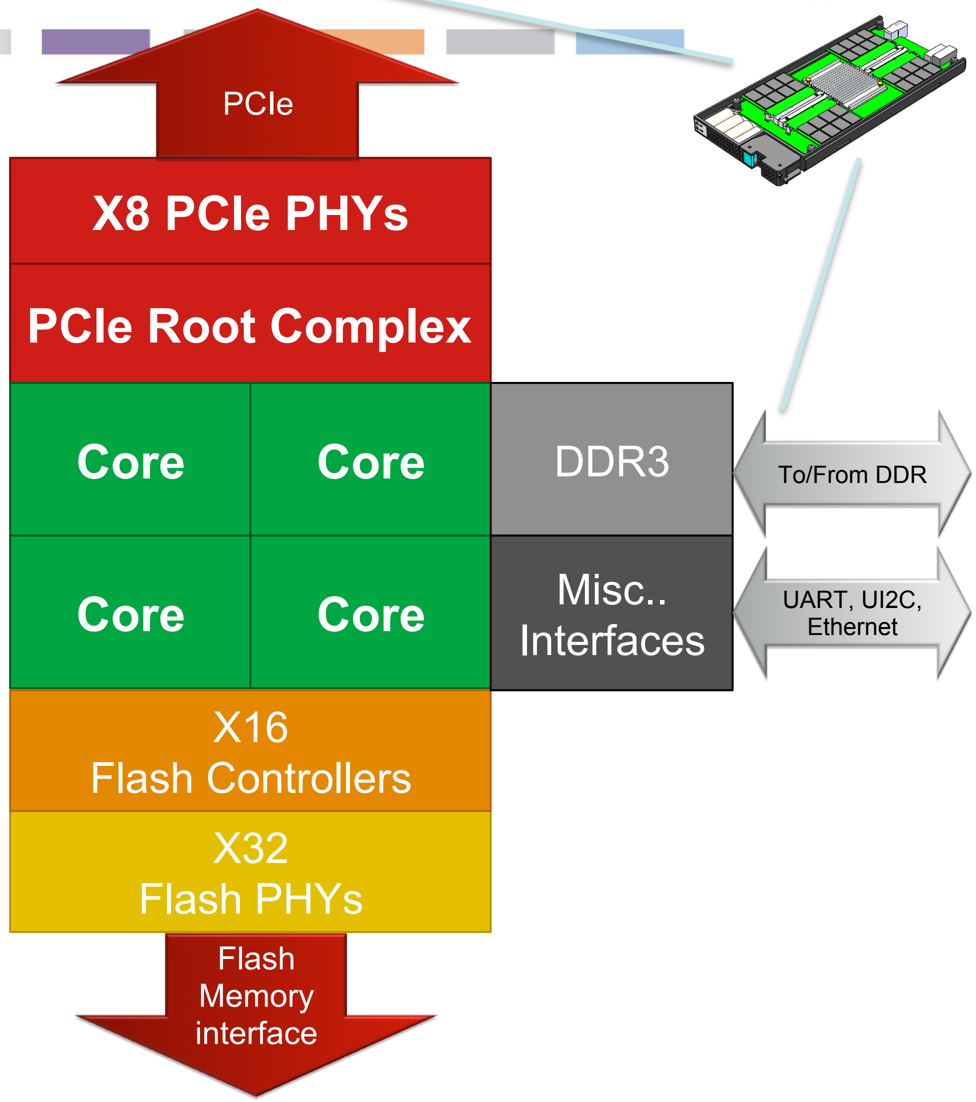




Difference between use of SSD's and custom solid state designs

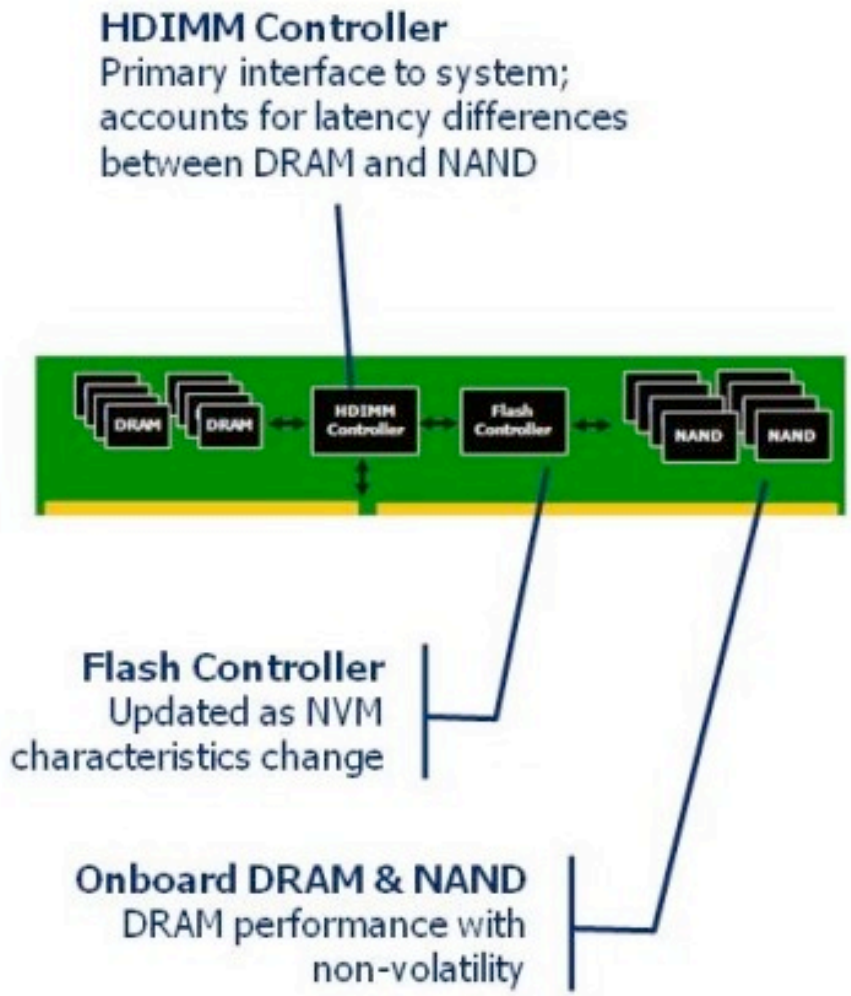
Flash Optimized – Using Flash Modules

- **Flash Module Devices**
  - Custom hardware design
  - Custom ASIC
- **Existing storage controllers can be 'flash optimized'**
  - Higher Performance
  - Unique Behaviors of solid state technology
- **Advanced storage features**
  - *Data reduction Technologies – reduces number of writes required*
    - *Deduplication*
    - *Compression*
  - *Thin privsioning*
  - *Copy Technologies (Snapshots, local and Remote replication)*
    - *Snapshot as a mapped technology*
  - *Self-healing techniques*



Flash DIMM

SNIA
Green Storage Initiative



Enables high capacity, SSD-like operation on the DDR bus

- Use Case #1: DIMM NAND-SSD with potentially large DRAM cache
- Use Case #2: NAND used as fast, local swap space for DRAM memory
- Use Case #3: Raw flash block storage with DRAM memory
- Requires significant software/ecosystem enablement to leverage full capabilities

Solid State Technology Future

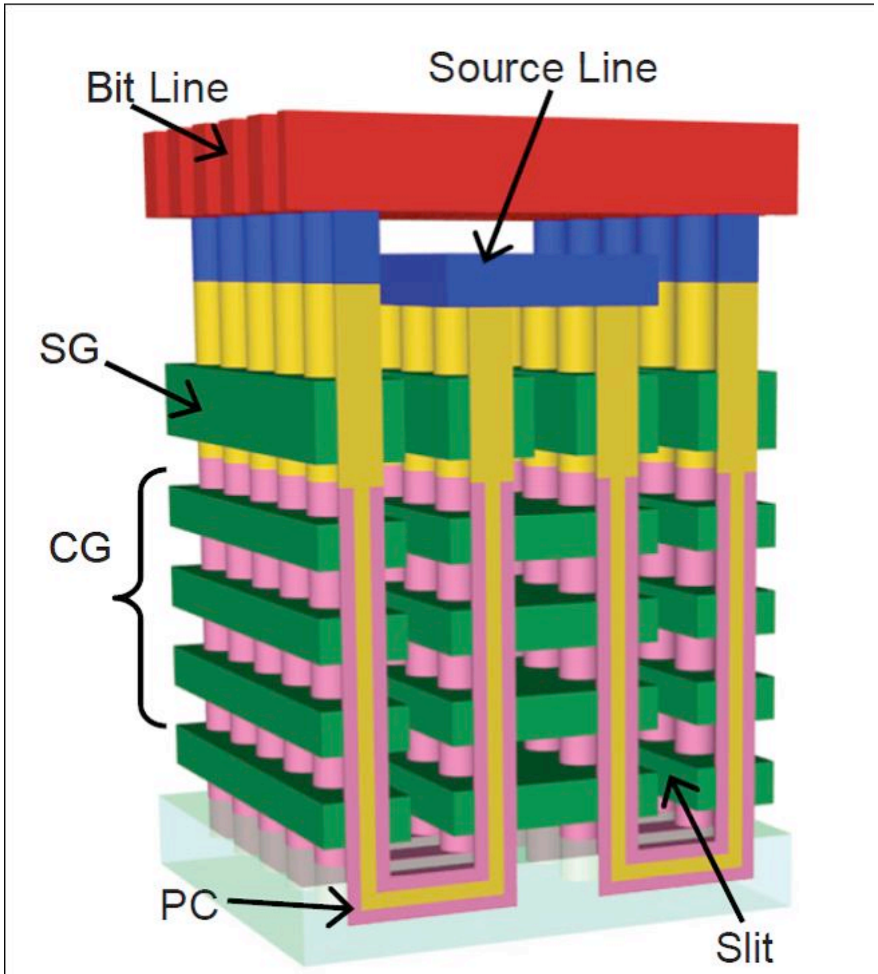
Current | New Technology

Characteristics	NAND Flash	3D NAND Flash	Resistive RAM**
Dimension	2D – SLC, MLC, TLC	3D (24 layers)	3D
Density	16 – 19 nm	19-20 nm	< 5 nm
Endurance	Wear Leveling	Wear Leveling	10x better*
Write Performance	7 MB/sec	14 MB/sec	140 MB/sec*
Program Energy	1360pJ/cell	1360pJ/cell	64pJ/cell
Retention	10 year	20 year	20 year

*Does not require an Erase prior to Programming or a wear leveling algorithm

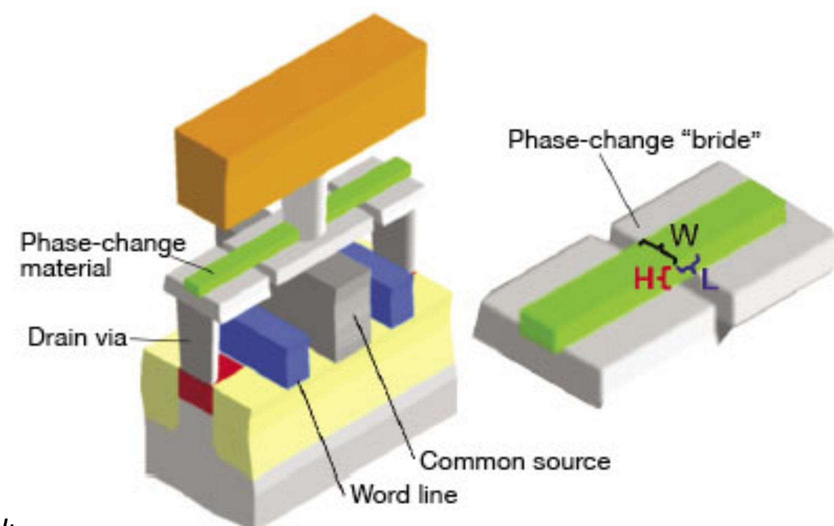
**Memristor technology

3D NAND Technology



Phase Change Memory

- PRM vs. Flash
  - PRM is Higher write Performance than Flash
  - PRM Cell degradation is much slower due to thermal than Flash
  - PRAM can’t be programmed before soldering due to high temps
- Micron has implemented this technology
- Draw Back
  - PRAM's temperature sensitivity
  - may require changes in the production process



Prototype Phase-Change Memory Switch Composed of germanium antimony, the new phase-change memory potentially can run 500 times faster than current Flash memory chips.

Block vs. File vs. Object

STORAGE TECHNOLOGY

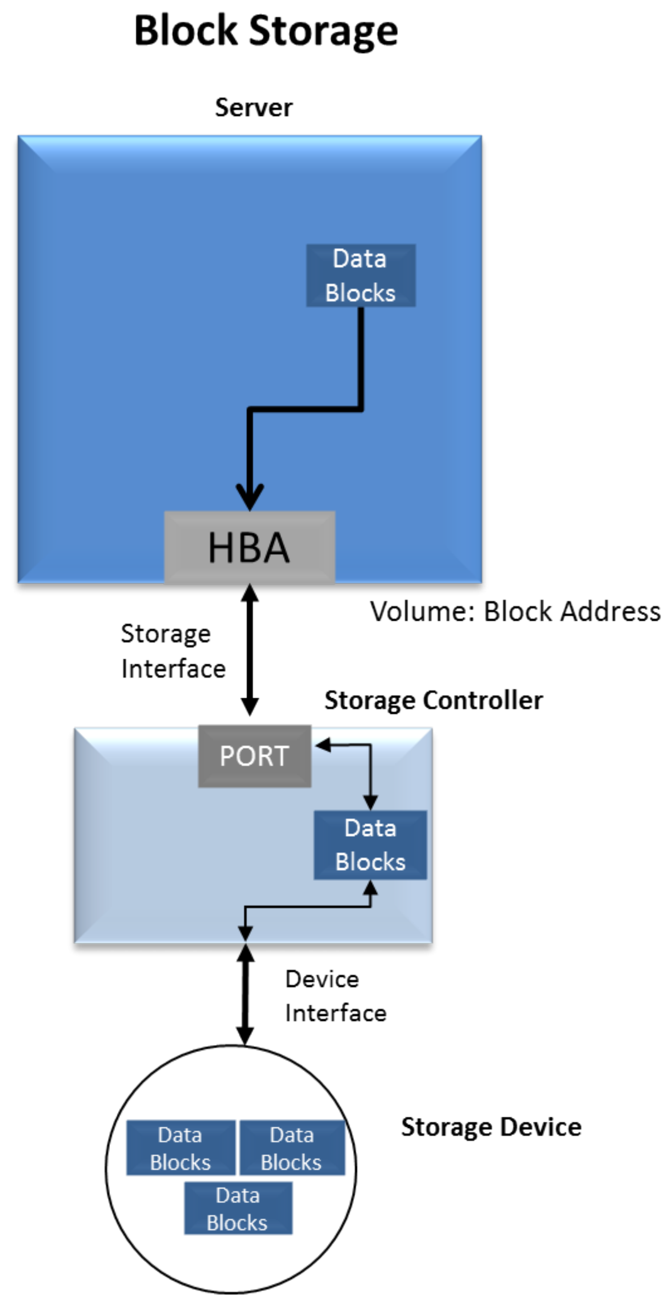
Block, File and Object

SNIA
Green Storage Initiative

- Block I/O vs. File I/O vs. Object I/O
  - Applications can do block I/O or file system I/O or object I/O
  - File systems turn file I/O into block I/O
  - Block I/O goes to specific device and reads or writes a block from/to that device
    - Linear address space of blocks
    - May do multiple blocks in single operation
    - Typically fixed length blocks
  - File I/O is represented by a file with file name and some offset into the file
    - Read or writes data in the file
    - Some number of bytes involved in the operation
  - Object I/O is storing data as objects with new control/metadata information

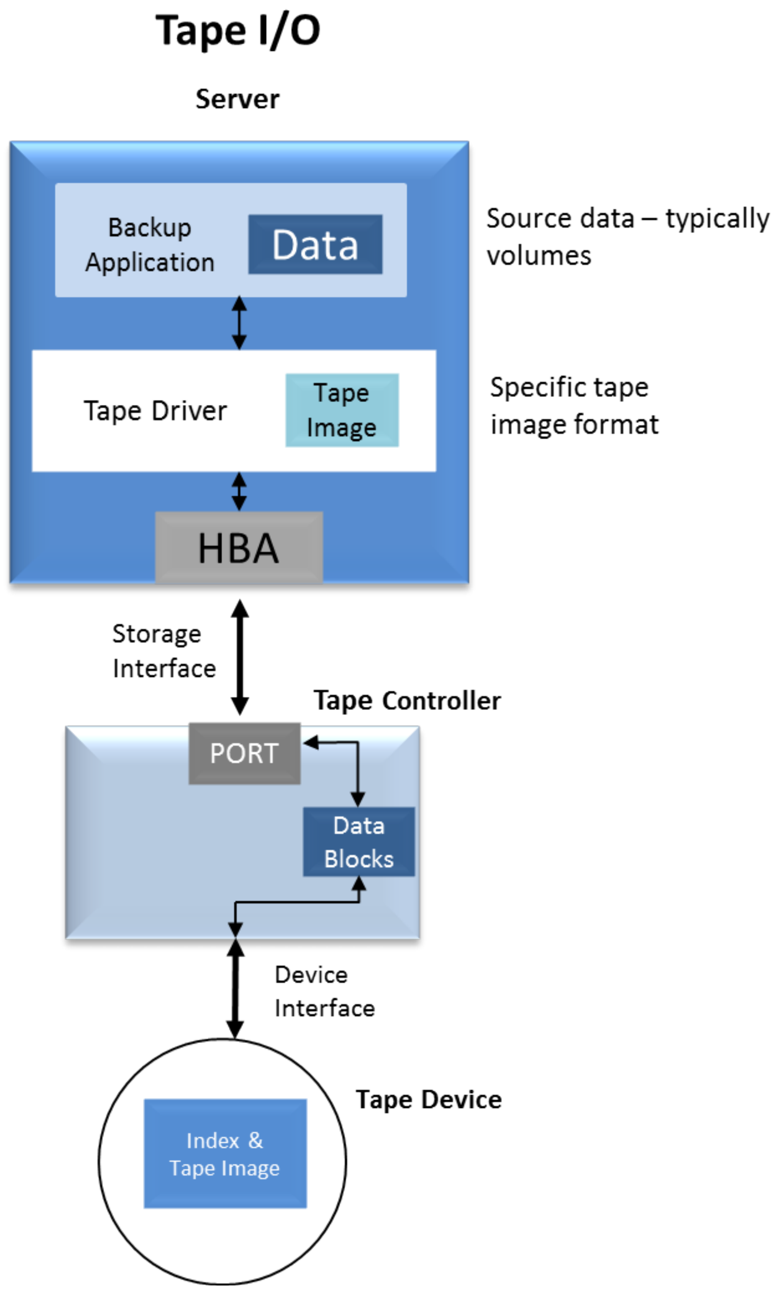
Block I/O

- Application writes data block
- Block goes to HBA and over storage interface
- Storage controller receives block
- Data written to device as data block



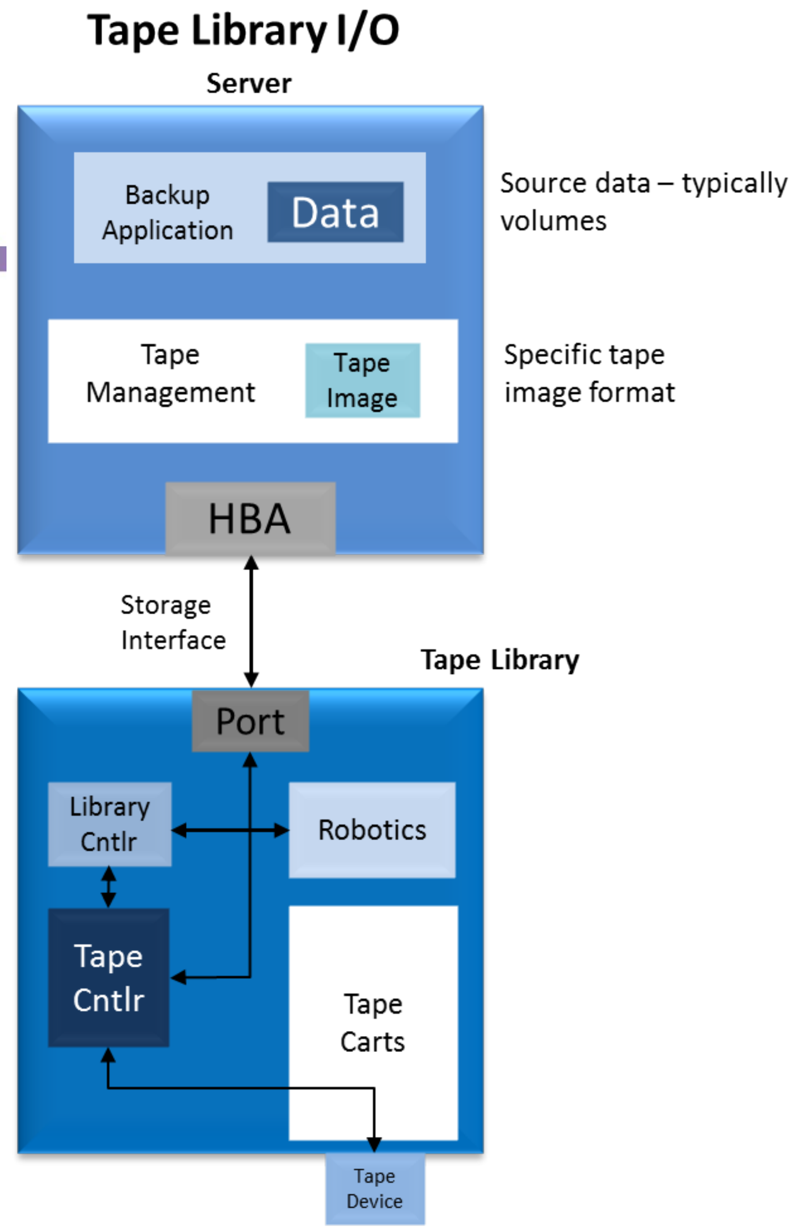
Tape I/O (Block I/O)

- Backup application writes data block to tape driver
- Block converted to tape image and goes to HBA and over storage interface
- Tape controller receives tape block
- Data written to tape as tape image block



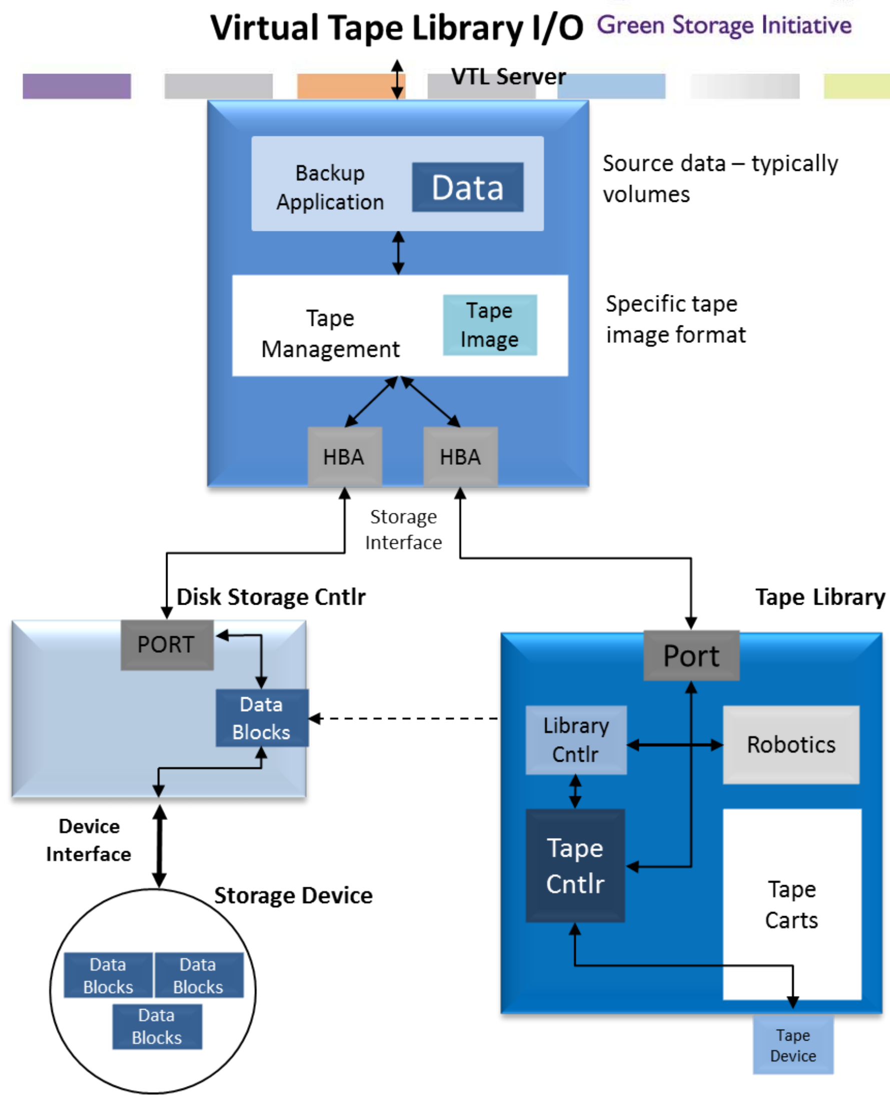
Tape Library I/O (Block I/O)

- Backup application writes data
- Block converted and tape volume identified
- Tape library receives tape block and volume information
- Data written to selected tape as tape image block



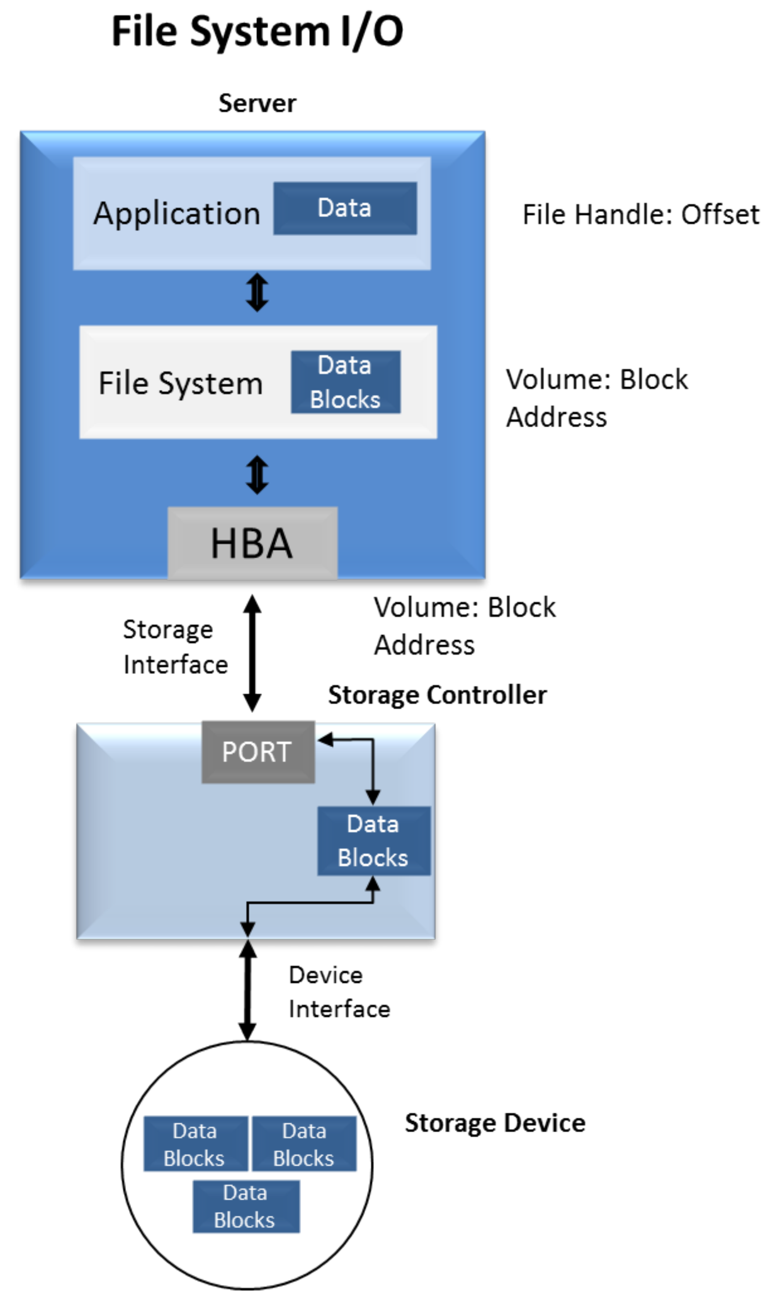
Virtual Tape I/O (Block I/O)

- Block converted to tape image
- Tape image written to disk controller
- Depending on controls, VTL reads tape image from disk and writes to tape library
- A few products can go direct to tape

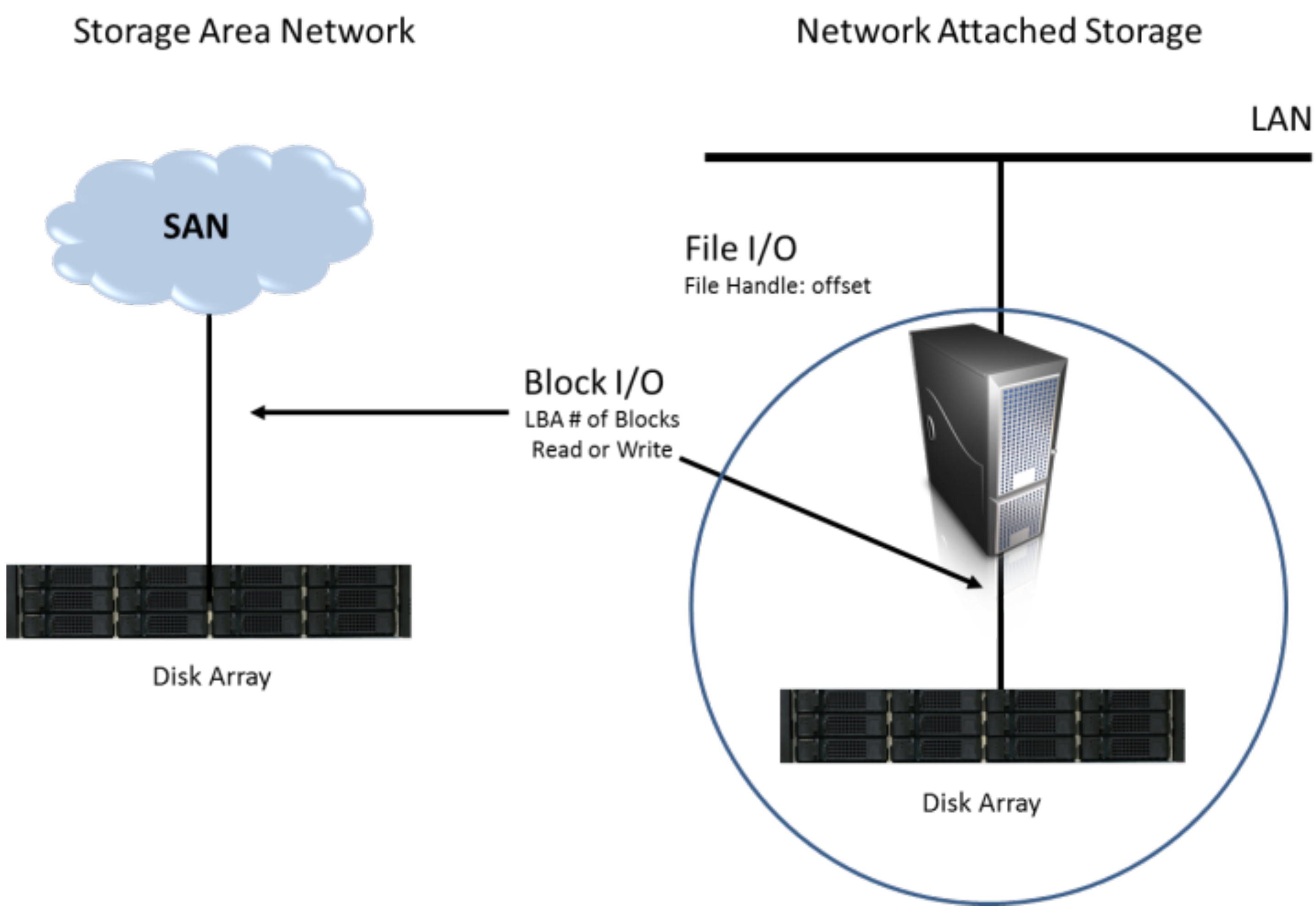


File I/O

- Application writes data block to a mounted file system
- Block goes to HBA and over storage interface
- Storage controller receives block
- Data written to device as data block
- Many Protocols

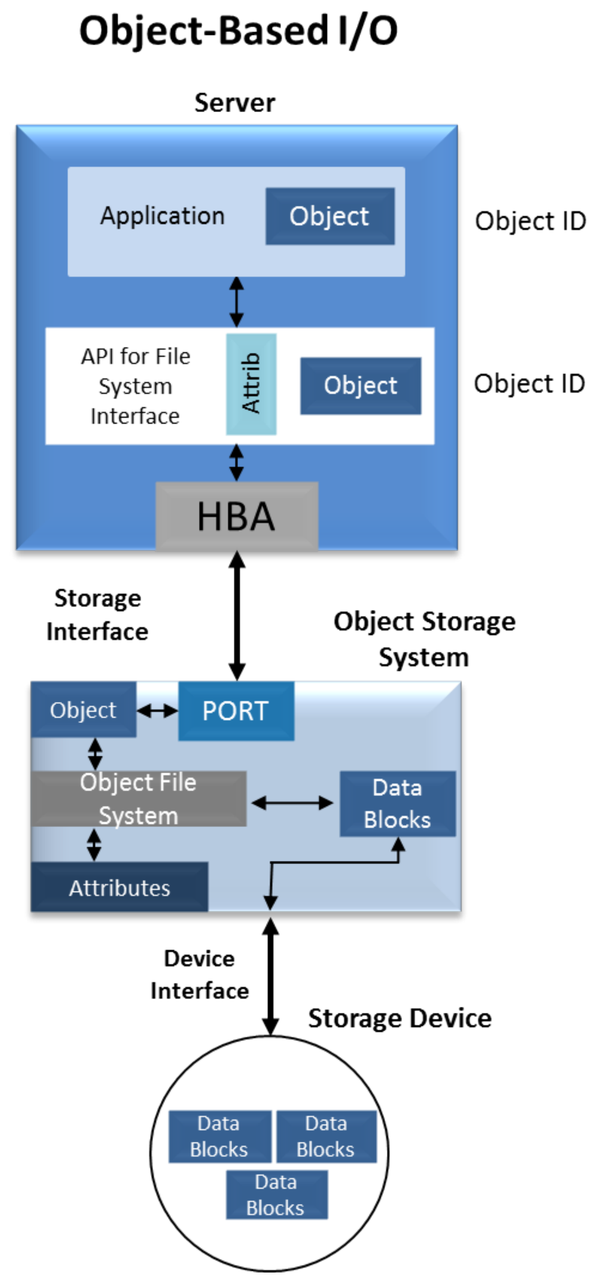


Block I/O vs. File I/O



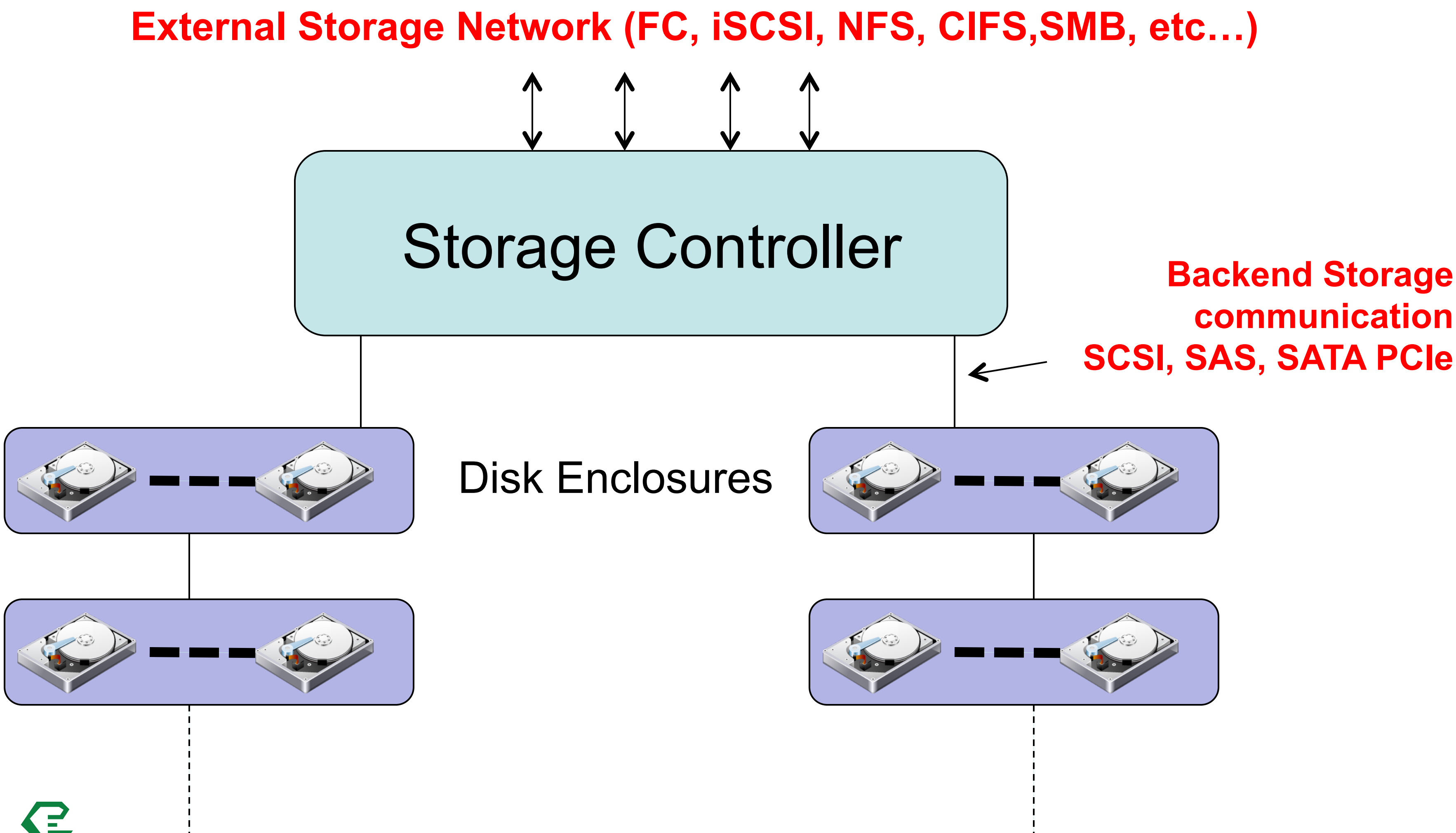
Object I/O

- Application writes object information
- Object file system creates attributes and sends object to HBA / NIC
- Storage controller receives object
- Data written to device as data block



STORAGE PROTOCOLS & DATA TRANSFER

Storage System Components



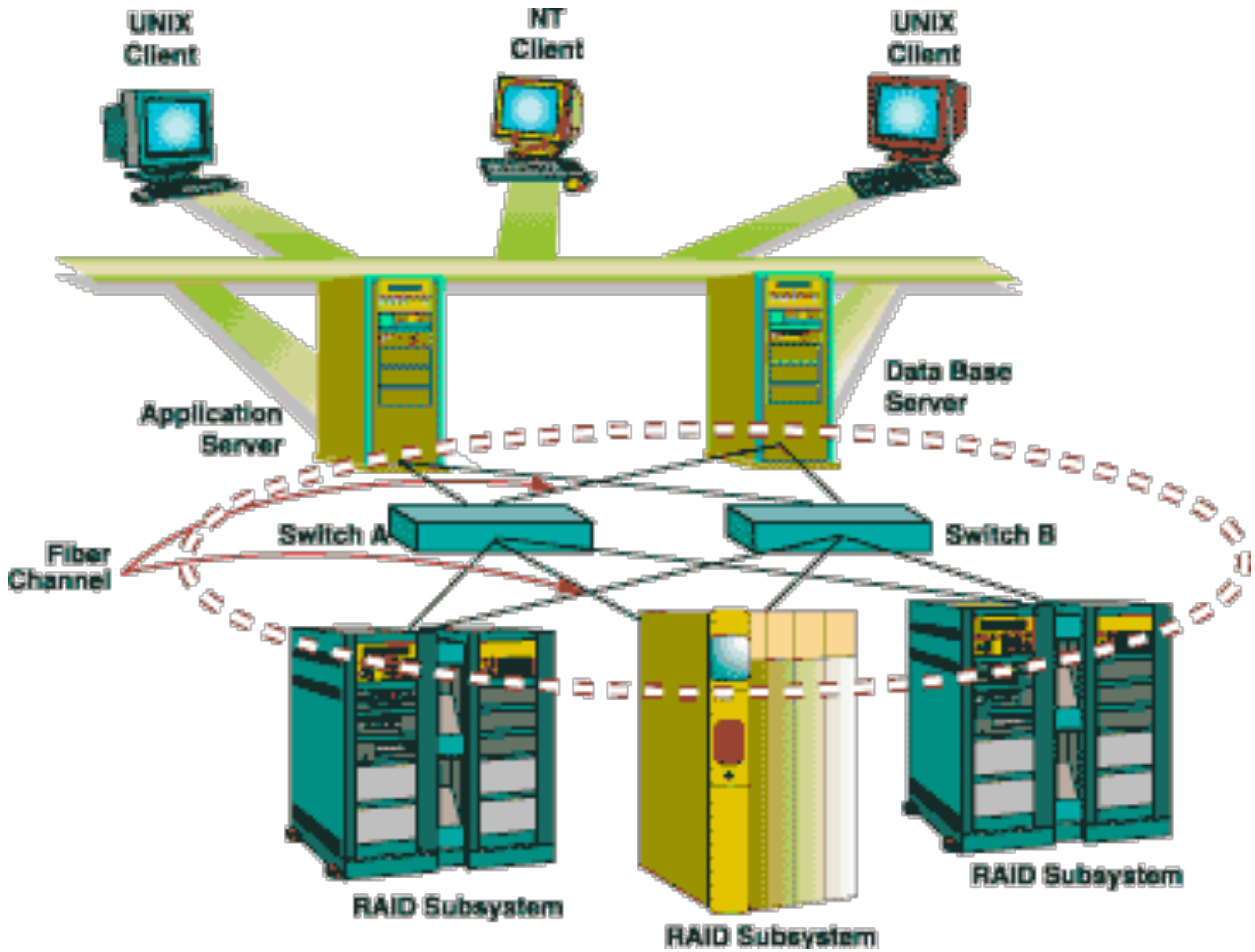
SAN and NAS

NETWORKING

Storage Area Network (SAN)

- Storage Area Network - A network whose primary purpose is the transfer of data between computer systems and block storage elements
- Physical Storage communication infrastructure for block storage (Fibre Channel (FC), iSCSI)
- Management layer – organize connections, storage elements, and computer systems.
- Uses switches and directors
- Shared storage design – many servers sharing a common storage utility

SAN

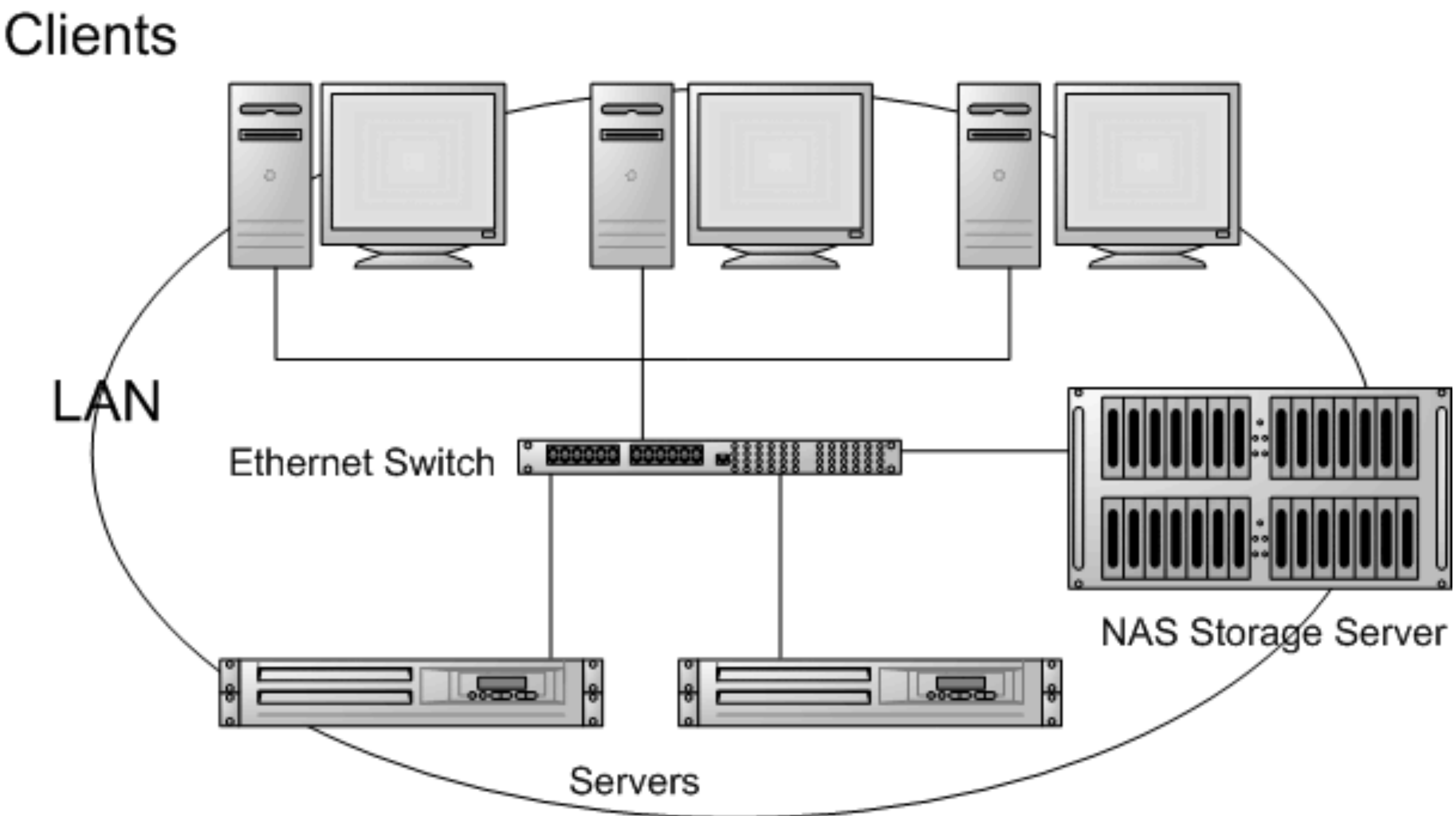

UNIX Client
NT Client
UNIX Client
Application Server
Data Base Server
Fiber Channel
Switch A
Switch B
RAID Subsystem
RAID Subsystem
RAID Subsystem

Network Attached Storage (NAS)

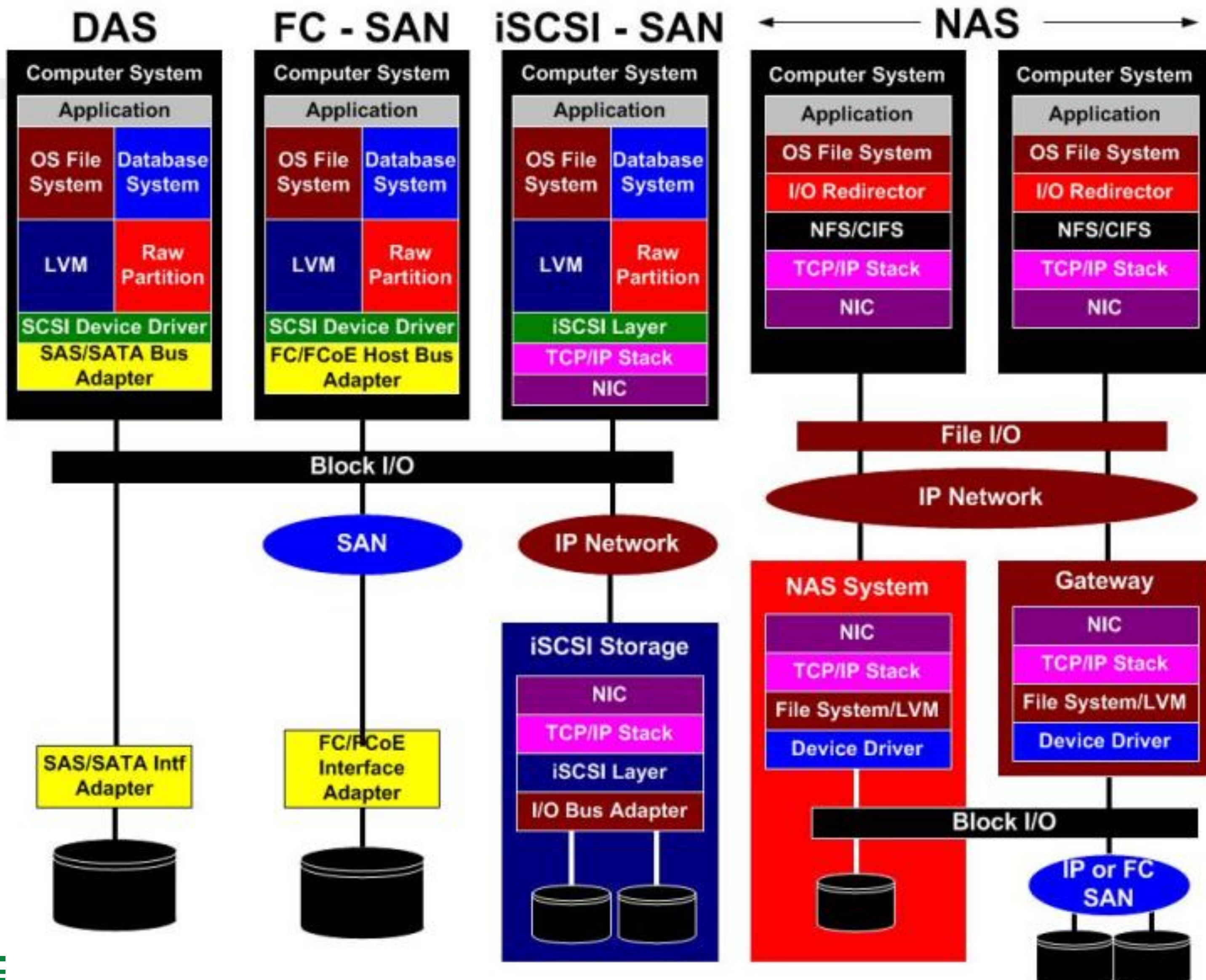
- Network Attached Storage - A network whose primary purpose is the transfer of data between computer systems and file storage elements
- Physical Storage communication infrastructure for file storage (NFS, CIFS, SMB, SW iSCSI, FCoE)
- Management layer – organize connections, storage elements, and computer systems.
- Uses LAN or WAN for communication.
- Shared storage design – many servers sharing a common storage utility

NAS

Network Attached Storage



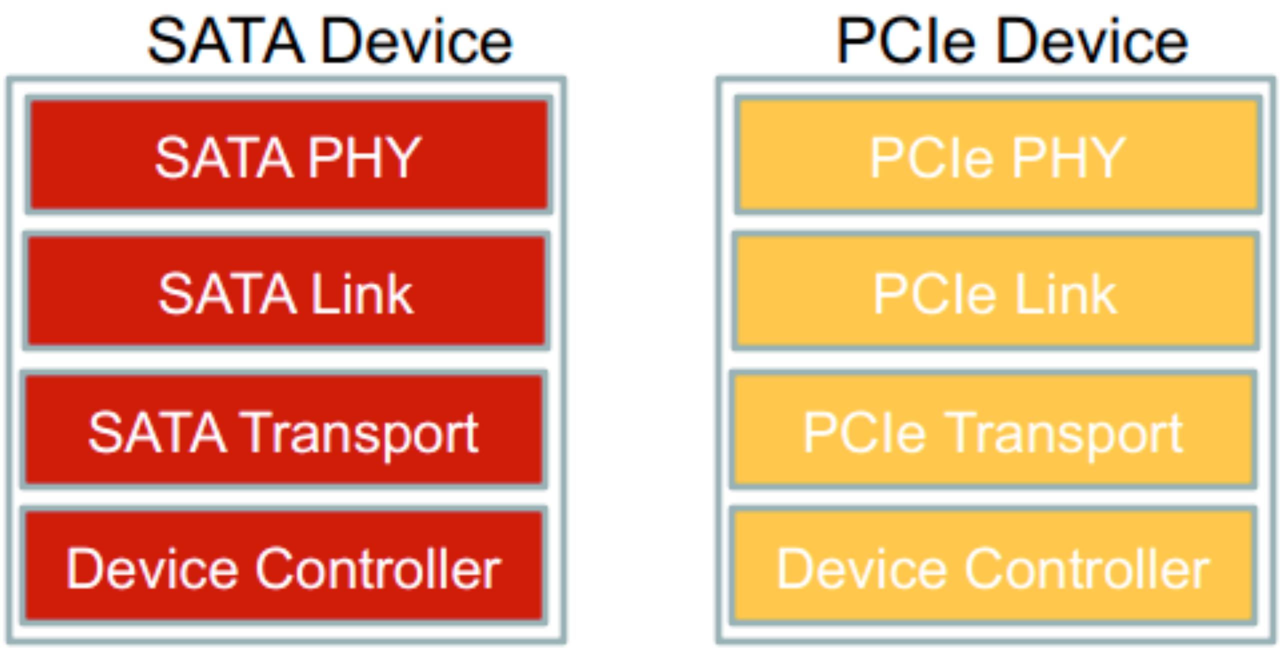
Storage Access



TIME FOR A BREAK

Internal System Protocols

- Serial ATA (SATA) Latest version 3.2 is 16 Gbits (1969 MB/sec)
- PCI express (PCIe)
- SAS



External System Protocols

- Block Storage (storage SAN switches)
  - Fibre Channel (FC)
  - iSCSI (HW)
- Networked Storage (Ethernet)
  - NFS
  - SMB
  - CIFS
  - iSCSI (SW)
  - Fibre Channel over Ethernet (FCoE)