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
Typical Storage System Architecture

Network (Storage Protocol)

Storage Controller

Hard Disk Drive

Disk Enclosures

Backend Storage communication

Solid State Drive

SNIA Emerald™ Training ~ July 14-17, 2014

www.sniaemerald.com
Traditional Dual Storage Controller

SCSI, FC, NFS, etc…

External connect

CPU

DRAM

Internal connect

SCSI or PCIe

Read Cache

Mirrored Write Cache

SCSI, FC, NFS, etc…

External connect

CPU

DRAM

Internal connect

SCSI or PCIe

Mirrored Write Cache

Read Cache

SCSI, FC, NFS, etc…

External connect

CPU

DRAM

Internal connect

SCSI or PCIe

Read Cache

Mirrored Write Cache

SCSI, FC, NFS, etc…

External connect

CPU

DRAM

Internal connect

SCSI or PCIe

Read Cache

Mirrored Write Cache
Dual Flash Controller

Controller A

Controller B

DRAM

CPU

PCI Express

144 Gbps

IOC

48 Gbps
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

1. Spin disk ~2ms
2. Actuate arm ~4ms
3. Read Data ~0.04ms
HDD Components
Multiple Platters, Cylinders, Tracks, and Sectors
Disk Sectors

One 512 Byte Sector

Data Field
512 Bytes

ECC
40 x 10 bit symbols
= 50 bytes

Servo fields, gaps and sync fields not shown for clarity

Eight 512 Byte Sectors

One 4k Byte Sector

Format Efficiency Improvement

Distributed ECC

SNIA Emerald™ Training ~ July 14-17, 2014  www.sniaemerald.com
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

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

- 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
## Technology Progression

<table>
<thead>
<tr>
<th></th>
<th>1990</th>
<th>2010</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU</strong></td>
<td>0.05 MIPS/$</td>
<td>147 MIPS/$</td>
<td>2940x</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>0.02 MB/$</td>
<td>25 MB/$</td>
<td>1250x</td>
</tr>
<tr>
<td><strong>Addressable Memory</strong></td>
<td>$2^{16}$</td>
<td>$2^{64}$</td>
<td>$2^{48}$x</td>
</tr>
<tr>
<td><strong>Network Speed</strong></td>
<td>100 Mbps</td>
<td>100 Gbps</td>
<td>1000x</td>
</tr>
<tr>
<td><strong>Disk Data Transfer</strong></td>
<td>5 MB/sec</td>
<td>130 MB/sec</td>
<td>25x</td>
</tr>
</tbody>
</table>
## Why Solid State Technology?

<table>
<thead>
<tr>
<th>Performance (IOPS) Transaction Processing</th>
<th>Hard Disk Drive</th>
<th>Solid State (NAND Flash)</th>
<th>Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Time</td>
<td>200</td>
<td>6000</td>
<td>30x</td>
</tr>
</tbody>
</table>

|            | 2-5 milliseconds | 100-500 microseconds | 1,000x   |
HDD vs SDD

Super-Fast Boot-Up in ~8 sec

- Solid State Cells
- Flash Memory
- Accessed by PCB

- Printed Circuit Board
- Crucial

- Spin 5400 to 15,000 RPM
- Helium Medium
- Stores Billions of Bits of Data

- Read Write Head
- Stores Data in Magnetic Patterns

SNIA Emerald™ Training ~ July 14-17, 2014

www.sniaemerald.com
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

*SNIA Emerald™ Training ~ July 14-17, 2014*
Solid State Storage

<table>
<thead>
<tr>
<th>Metric</th>
<th>SLC</th>
<th>MLC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (microseconds)</td>
<td>100</td>
<td>200-300</td>
</tr>
<tr>
<td>Persistence</td>
<td>10x more persistent</td>
<td>Less reliable*</td>
</tr>
<tr>
<td>Cost</td>
<td>30% more expensive</td>
<td>More cost effective</td>
</tr>
<tr>
<td>Sequential read/writes</td>
<td>3x faster</td>
<td>Slower</td>
</tr>
</tbody>
</table>

*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

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

- **Solid Storage Arrays**
  - SAN SSS Arrays (FC or SAS)
  - Network SSS Arrays (NFS, SW iSCSI)
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 provisioning
  - Copy Technologies (Snapshots, local and Remote replication)
    - Snapshot as a mapped technology

**Self-healing techniques**
Flash DIMM

**HDIMM Controller**
Primary interface to system; accounts for latency differences between DRAM and NAND

**Flash Controller**
Updated as NVM characteristics change

**Onboard DRAM & NAND**
DRAM performance with non-volatility

---

**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

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>NAND Flash</th>
<th>3D NAND Flash</th>
<th>Resistive RAM**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension</td>
<td>2D – SLC, MLC, TLC</td>
<td>3D (24 layers)</td>
<td>3D</td>
</tr>
<tr>
<td>Density</td>
<td>16 – 19 nm</td>
<td>19-20 nm</td>
<td>&lt; 5 nm</td>
</tr>
<tr>
<td>Endurance</td>
<td>Wear Leveling</td>
<td>Wear Leveling</td>
<td>10x better*</td>
</tr>
<tr>
<td>Write Performance</td>
<td>7 MB/sec</td>
<td>14 MB/sec</td>
<td>140 MB/sec*</td>
</tr>
<tr>
<td>Retention</td>
<td>10 year</td>
<td>20 year</td>
<td>20 year</td>
</tr>
</tbody>
</table>

*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

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

Storage Area Network

- SAN
- Disk Array
- Block I/O
  - LBA # of Blocks
  - Read or Write

Network Attached Storage

- LAN
- File I/O
  - File Handle: offset
- Disk Array
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

External Storage Network (FC, iSCSI, NFS, CIFS, SMB, etc…)

Storage Controller

Disk Enclosures

Backend Storage communication
SCSI, SAS, SATA PCIe
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
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

Clients

LAN

Ethernet Switch

Servers

NAS Storage Server
Storage Access

SNIA Emerald™ Training ~ July 14-17, 2014

www.sniaemerald.com
TIME FOR A BREAK
Internal System Protocols

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

![Diagram of SATA and PCIe devices and protocols]
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)