Things You Should Know About Solid State Storage
Snippets from SNIA Tutorials and other Giblets

Marty Czekalski - President SCSI Trade Association,
Sr. Staff Program Manager - Seagate Technology
The material contained in this tutorial is copyrighted by the SNIA unless otherwise noted.

Member companies and individual members may use this material in presentations and literature under the following conditions:
- Any slide or slides used must be reproduced in their entirety without modification
- The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA Education Committee.

Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.

The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.
Abstract

What You Should Know About Solid State Storage

This session will appeal to Data Center Managers, Development Managers, and those that are seeking an overview of Solid State Storage. It’s comprised of excerpts from SNIA Solid State Tutorials and other sources. It touches on the highlights of the driving factors behind Solid State Storage, it’s performance, endurance, interfaces, and future directions.
Topics

- Why and Where for Solid State Storage
- Performance
- Endurance
- Interfaces
- Future Directions
Cost Structure of Memory/Storage Technologies

**Cost** determined by
- cost per wafer
- # of dies/wafer
- memory area per die [sq. µm]
- memory density [bits per 4F²]
- patterning density [sq. µm per 4F²]

Chart courtesy of Dr. Chung Lam, IBM Research updated version of plot from 2008 IBM Journal R&D article
New Storage Hierarchy in NGDC & Clouds

I/O Access Frequency vs. Percent of Corporate Data 2015

- **SSD**
  - Logs
  - Journals
  - Temp Tables
  - Hot Tables

- **FCoE/SAS Arrays**

- **Cloud Storage**
  - Primary Capacity Storage
  - Back Up Data
  - Archived Data
  - Offsite DataVault

<table>
<thead>
<tr>
<th>% of Corporate Data</th>
<th>% of I/O Accesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>95%</td>
<td></td>
</tr>
</tbody>
</table>

What You Should Know About Solid State Storage
© 2013 Storage Networking Industry Association. All Rights Reserved.
End-to-End Flash Categories

A  Pure Flash in Array
   - All flash

B  Flash in Storage Controller
   - Flash hardware and software
   - “Behind wire”
   - E.g. Flash Cache

C  Host-side Flash Software
   - Software only, may be tied to particular flash hardware

Flash as DAS / Cache
   - Flash hardware, stores persistent data
   - May be combined with software to form cache

Flash-based Virtual Storage Appliance
   - Software

Network-based Flash
   - Flash Hardware and software
   - “Bump in the wire”

Hybrid Flash / HDD Array
   - Mixed flash / HDD
   - E.g. Flash Pool

What You Should Know About Solid State Storage
© 2013 Storage Networking Industry Association. All Rights Reserved.
FC HDD plus Flash Cache Example

Benchmarked Configurations

<table>
<thead>
<tr>
<th>Configuration</th>
<th>FC Baseline</th>
<th>FC + Flash Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drives/Storage Capacity</td>
<td>224/64TB</td>
<td>56/16TB</td>
</tr>
</tbody>
</table>

- **75% Fewer Spindles**

SPECTsfs2008 Performance

- **Throughput (k-ops/sec)**
  - FC Baseline
  - FC + Flash Cache

- **Response Time (ms)**
  - WORSE
  - BETTER

- **Purchase price is 50% lower** for FC + Flash cache compared to Fibre Channel baseline

- **FC + Flash cache yields 67% power savings and 67% space savings**


SPEC® and SPECTsfs2008® are trademarks of the Standard Performance Evaluation Corp.
System Evolution

1980

Logic

Memory

Active Storage

Archival

CPU → RAM → DISK → TAPE

→ fast, synch

slow, asynch →

2010

CPU → RAM → FLASH SSD → DISK → TAPE

2013+

CPU → RAM → Storage Class Memory → DISK → TAPE

* e.g. Phase change memory

Memristor

Solid Electrolyte

Racetrack memory
Performance

- SPC-1C
- SNIA PTS (Enterprise and Client)
- Others
  - Reviews
  - OEMs
  - Application unique
SPC-1C Benchmark

SPC-1C is comprised of a set of I/O operations designed to demonstrate the performance of a small storage subsystem while performing the typical functions of a business critical application.

SPC-1C represents a segment of applications characterized by predominately random I/O operations and requiring both queries as well as update operations (for example: OLTP systems, Database systems, or Mail Server applications).

SPC-1C focuses on Small storage solutions 1 to 24 drives
All transfers 4K aligned

- Audited results
- http://www.storageperformance.org
Where Do We Report Performance?

RND/4K Write Saturation, Group By Classes

Log Scale!
Tests Contained In PTS-C 1.0 SPEC

**Client IOPS**
- **Random Access**
  - R/W:
    - 100/0, 95/5, 65/35, 50/50, 35/65, 5/95, 0/100
  - BS:
    - 1024KiB, 128KiB, 64KiB, 32KiB, 16KiB, 8KiB, 4KiB, 0.5KiB
- **Range Restriction:**
  - 100% & 75% LBA
  - 2048 Segments
- **Active Footprint Restriction:**
  - 8 & 16 GiB

**Client TP**
- **Sequential Access**
  - R/W:
    - 100/0, 0/100
  - BS:
    - 1024KiB
  - **Range Restriction:**
    - 100% & 75% LBA
  - 2048 Segments
- **Active Footprint Restriction:**
  - 8 & 16 GiB

**Client Latency**
- **Random Access**
  - R/W:
    - 100/0, 65/35, 0/100
  - BS:
    - 8KiB, 4KiB, 0.5KiB
  - **Range Restriction:**
    - 100% & 75% LBA
  - 2048 Segments
- **Active Footprint Restriction:**
  - 8 & 16 GiB
Tests Contained In PTS-E 1.0 SPEC

- Enterprise Performance Test Specification (PTS-E) V1.0 encompasses:
  - A suite of basic SSS performance tests
  - Preconditioning and Steady State requirements
  - Standard test procedures and reporting requirements

<table>
<thead>
<tr>
<th>Write Saturation</th>
<th>Enterprise IOPS</th>
<th>Enterprise TP</th>
<th>Enterprise Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random Access</strong></td>
<td><strong>Random Access</strong></td>
<td><strong>Sequential Access</strong></td>
<td><strong>Random Access</strong></td>
</tr>
<tr>
<td><strong>R/W:</strong></td>
<td><strong>R/W:</strong></td>
<td><strong>R/W:</strong></td>
<td><strong>R/W:</strong></td>
</tr>
<tr>
<td>• 100% Writes</td>
<td>• 100/0, 95/5, 65/35, 50/50, 35/65, 5/95, 0/100</td>
<td>• 100/0, 0/100</td>
<td>• 100/0, 65/35, 0/100</td>
</tr>
<tr>
<td><strong>BS:</strong></td>
<td><strong>BS:</strong></td>
<td><strong>BS:</strong></td>
<td><strong>BS:</strong></td>
</tr>
<tr>
<td>• 4KiB</td>
<td>• 1024KiB, 128KiB, 64KiB, 32KiB, 16KiB, 8KiB, 4KiB, 0.5KiB</td>
<td>• 1024KiB, 128KiB</td>
<td>• 8KiB, 4KiB, 0.5KiB</td>
</tr>
</tbody>
</table>
Tests Contained In PTS-E 1.1

PTS-E 1.1 adds:

- **Host Idle Recovery**
  - Examines effect of idle (no IO) on small block RND writes
  - RND/4KiB Writes

- **Cross Stimulus Response**
  - Examines switching between large block SEQ and small block RND writes
  - SEQ/1024KiB & RND/4KiB Writes

- **Demand Intensity – Response Time Histograms**
  - Performance and detailed response time statistics under various workload types
  - R/W=65/35 %, RND/8K
  - R/W=90/10 %, RND/128K
  - Response Time Histograms at various operating points

- **Enterprise Composite Workload**
  - Performance and detailed response time in a mixed IO Enterprise environment
  - R/W=60/40 %
  - BS from 0.5-64KiB
  - Three LBA Groups
Example: MLC/SAS
## The Enterprise Composite Workload

The ECW is a R/W=40/60%, random access pattern with a distribution of Block Sizes, each with a pre-defined Access Probability, plus restrictions on Access Range Probability Distribution.

<table>
<thead>
<tr>
<th>Block Size in Bytes (KiB)</th>
<th>Access Probability Within Each Measurement Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>512 bytes (0.5 KiB)</td>
<td>4%</td>
</tr>
<tr>
<td>1024 bytes (1 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>1536 bytes (1.5 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>2048 bytes (2 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>2560 bytes (2.5 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>3072 bytes (3 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>3584 bytes (3.5 KiB)</td>
<td>1%</td>
</tr>
<tr>
<td>4096 bytes (4 KiB)</td>
<td>67%</td>
</tr>
<tr>
<td>8192 bytes (8 KiB)</td>
<td>10%</td>
</tr>
<tr>
<td>16,384 bytes (16 KiB)</td>
<td>7%</td>
</tr>
<tr>
<td>32,768 bytes (32 KiB)</td>
<td>3%</td>
</tr>
<tr>
<td>65,536 bytes (64 KiB)</td>
<td>3%</td>
</tr>
</tbody>
</table>

Total: 100%

<table>
<thead>
<tr>
<th>% of Access within 1 Measurement Period</th>
<th>Active Range Restriction</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>First 5%</td>
<td>LBA Group A</td>
</tr>
<tr>
<td>30%</td>
<td>Next 15%</td>
<td>LBA Group B</td>
</tr>
<tr>
<td>20%</td>
<td>Remaining 80%</td>
<td>LBA Group C</td>
</tr>
</tbody>
</table>
ECW: Demand Intensity (PCIe, MLC)
Endurance

- What causes wear?
- Workloads and use cases
- JEDEC Endurance Specification
What is Flash Wear?

Electrons tunnel through oxide to charge the floating gate. Electron traps interfere with charge.

Floating Gate
What is SLC/MLC/TLC?

Smaller and smaller windows to determine signal’s value

Incoming Signals

Single bit Multi bit TLC

What You Should Know About Solid State Storage
© 2013 Storage Networking Industry Association. All Rights Reserved.
Write Amplification

“You need to move, I am going to erase this block.”

“You need to move, I am going to erase this block.”
Read Write Disturb
JEDEC SSD Standards

- **JESD218A**, *Solid State Drive (SSD) Requirements and Endurance Test Method*
  - SSD Definitions
  - SSD Capacity
  - Application Classes
  - Endurance Rating
  - Endurance Verification

- **JESD219**, *Solid State Drive (SSD) Endurance Workloads*
  - Client
  - Enterprise
Drive Writes Per Day versus TBW

- Sometimes endurance is expressed in the number of drive writes per day for an SSD
  - Can range from >25 to <1
  - What does drive writes per day that really mean?
  - It is a great expression in terms of SSD capacity, but make sure the SSD manufacturer can answer these questions:
    - What determines end of life?
    - Does it mean that the SSD had to wear level or was the workload such that it did not require the drive to do any wear leveling?
    - Was the data content random or compressible?

- TBW (TeraBytes Written) is a defined JEDEC endurance term that has specific guidelines for defining how it is verified and what it means
Application Classes

❖ Application classes:
   - Client
   - Enterprise

❖ Application class attributes:
   - Workload
   - Daily active use
   - Data retention
   - BER
JEDEC Endurance Rating - TBW

JESD218A SSD Classes and Requirements table:

<table>
<thead>
<tr>
<th>Application Class</th>
<th>Workload (see JESD219)</th>
<th>Active Use (power on)</th>
<th>Retention Use (power off)</th>
<th>Functional Failure Requirement (FFR)</th>
<th>UBER Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client</td>
<td>Client</td>
<td>40 °C 8 hrs/day</td>
<td>30 °C 1 year</td>
<td>≤3%</td>
<td>≤10⁻¹⁵</td>
</tr>
<tr>
<td>Enterprise</td>
<td>Enterprise</td>
<td>55 °C 24hrs/day</td>
<td>40 °C 3 months</td>
<td>≤3%</td>
<td>≤10⁻¹⁶</td>
</tr>
</tbody>
</table>

From JESD218A, Copyright JEDEC. Reproduced with permission by JEDEC

These requirements and classes define what the JEDEC TBW endurance rating really means.
Workload Impact On WAF

Workload factors that impact Write Amplification Factor (WAF):

- Sequential versus random
- Large transfers versus small transfers
- Boundary alignment
  - Transfer size vs program page size/alignment
  - Transfers crossing erase blocks
- Data content/patterns (especially for SSDs using data compression)
Endurance Rating Workload

- **Enterprise workload:**
  - Leveraged from SPC-1 but not totally on 4K boundaries
  - Writes 100% of LBA’s, including more writes to some areas than others

- **Client workload:**
  - Leveraged from a 9 month trace on client application
  - Includes trim commands

- Both workloads use a random data payload
Endurance Rating Temperature

Table C.1 — Expected retention (weeks) at different use temperatures

From JESD218A, Copyright JEDEC. Reproduced with permission by JEDEC

These numbers reflect data retention for NAND after 100% P/E cycles. Less than 100% has much longer data retention.
Interfaces and Form Factors

» Typical Storage Interfaces
  • SATA – 6Gb/s
  • SAS – 12Gb/s (2HCY2013)
  • Fibre Channel – 16Gb/s (4Gb/s for drives)
  • Infiniband – 14Gb/s

» Std. Disk form factors, rack mount, custom, etc

» PCIe
  • Form Factors
    › Express Bay (SFF-8639), SATA Express, CEM (Std. PCIe Cards), M.2, Mezzanine, etc.
  • Protocols
    › NVM Express (NVMe), SCSI Express (SCSle), Advance Host Controller Interface (AHCI), Proprietary
Why PCIe Storage Standards?

Areas to Address

- **Performance Trends**
  - Processor vs. Storage
  - Gap Increasing

- **Serviceability**
  - Internal Access
  - Cold-Plug

- **Interoperability**
  - Card Form Factor
  - Varying Card Sizes

- **Scalability**
  - Performance & Capacity

PCle SSD Benefits

- **Minimize Gap**
  - Improved Latency
  - Improved IOPs

- **Remove Constraints**
  - External Access
  - Hot-Pluggable

- **Common Form Factor**
  - Drive Form Factor
  - Multi-protocol

- **Increased Slots**
  - External Slots
  - “Live” Scaling

What You Should Know About Solid State Storage
© 2013 Storage Networking Industry Association. All Rights Reserved.
More on Latency …

**Today: Low OIO (Outstanding IO)**
- High latency NVM and legacy stack can diminish interface latency benefits

<table>
<thead>
<tr>
<th>Stack</th>
<th>PCIe</th>
<th>NVM</th>
<th>Stack</th>
<th>PCIe</th>
<th>NVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PCIe</th>
<th>SAS+PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Today: High OIO**

**Future: Upcoming advances**
- Parallelism reduces NVM and stack aggregate latency, seen today in database work loads
- Future NVM can achieve low latency even at low OIO
- Path latency of a PCIe solution can be much lower than SAS solution, providing significant performance improvement

<table>
<thead>
<tr>
<th>Stack</th>
<th>PCIe</th>
<th>SAS</th>
<th>NVM</th>
<th>PCIe</th>
<th>NVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PCIe</th>
<th>SAS+PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Express Bay Concept

Express Bay
- SFF-8639 multifunction connector
- Supports multiple protocols/interfaces
- PCI-SIG electrical specification
- Up to 25 Watts

Objectives
- Preserve the enterprise storage experience for PCI Express storage
- Meet SSD performance demands
- Serviceable, hot-pluggable Express Bay opens up new possibilities…
NVM Express Overview

- NVM Express is a scalable host controller interface designed for Enterprise and Client systems that use PCI Express* SSDs
  - Includes optimized register interface and command set
- NVMe was developed by industry consortium of 80+ members and is directed by a 10 company Promoter Group
- NVMe 1.1 published on October 11, 2012, available at nvmexpress.org
SCSI Express Components

- **SCSI**
  - Storage command set

- **PCI Express**
  - Leading server I/O interconnect

- **SCSI Over PCIe (SOP)**
  - Packages SCSI for a PQI queuing layer

- **SFF-8639 Connector**
  - Accommodates PCIe, SAS and SATA drives

- **PCIe Queuing Interface (PQI)**
  - Flexible, high performance queuing layer
# Efficient SSD Performance

## AHCI

<table>
<thead>
<tr>
<th>Feature</th>
<th>AHCI</th>
<th>nvm EXPRESS</th>
<th>SCSI EXPRESS</th>
</tr>
</thead>
</table>
| **Uncacheable Register Reads**        | 4 per command
8000 cycles, ~ 2.5 µs                                                |             | 0 per command |
| **MSI-X and Interrupt Steering**      | No                                                                   | Yes         |              |
| **Parallelism & Multiple Threads**    | Requires synchronization
lock to issue command                                                      | No locking, doorbell
register per Queue                                                        |              |
| **Maximum Queue Depth**               | 32                                                                   | 64K Queues
64K Commands                                                               |              |
| **Efficiency for 4KB Commands**       | Command parameters require two serialized
host DRAM fetches                                                           | Command parameters
in one 64 byte fetch                                                        |              |
A SATA Express (or M.2) will be able to accept either SATA or PCIe storage devices

- A signal on the device connector tells the host whether the device is SATA or PCIe
- Host then uses the appropriate driver depending protocol (SATA, AHCI, NVMe, or SCScle)
Future Directions

- NV Memories
  - Flash
  - Phase Change Memory
  - MRAM
  - Memristor (RRAM)
  - More …

- New entrants will challenge DRAM first
  - Much faster than Flash
  - Can be either block or byte accessible
  - More memory like than traditional storage

- The tradition IO Stacks and interfaces may no longer be sufficient
  - Programming model needs to change
  - BUT, don’t throw the baby out with the bathwater
    - Still need security, data protection, naming, etc.
  - Add new capabilities to existing interfaces
Building Towards APIs

Evolving NVM

Extensible Programming Model

API

Concepts…
• Sync and Async I/O
• Atomic Updates
• Memory-mapping
• Names
• Permissions
• Backing up Data
A Proposed Programming Model
Covering All Three Paths

1. **Standard Access**
   - Raw Access
   - File Access
   - File System
   - Block Layer
   - NVM Driver
   - NVM Hardware

2. **NVM regions exposed as files**
   - File or Raw Access
   - Management Applications (GUI, CLI, CIM)
   - Open NVM Kernel API
   - Naming Layer
   - NVM User-space API

3. **NVM API**
   - Middleware (e.g. JVM)
   - Optimized Applications
   - Control Path
   - Data Path
   - User
   - Kernel
   - Hardware

What You Should Know About Solid State Storage
© 2013 Storage Networking Industry Association. All Rights Reserved.
References

- www.snia.org/education/tutorials
- www.snia.org/nvmsummit
- www.ssdf ormfactor.org
- www.nvmexpress.org
- www.scsita.org
- www.t10.org
- www.sata-io.org
Questions?
Attribution & Feedback

The SNIA Education Committee would like to thank the following individuals for their contributions to this Tutorial.

Authorship History

Name/Date of Original Author here:
Marty Czekalski
Gary Kotzur
Ester Spanjer
Luanne Dauber
David Dale
Anil Vassudeva
Easen Ho
Alvin Cox
Andy Rudoff

Updates:

Additional Contributors

Too many to mention here:
See SNIA tutorials for complete list

Please send any questions or comments regarding this SNIA Tutorial to tracktutorials@snia.org