Intelligent Controllers
Key to Solid State Storage Systems

Anil Vasudeva
President & Chief Analyst
IMEX Research.com
(408) 268-0800  (408) 489-0800 cell
imex@imexresearch.com
• Industry Trends
• SSD Adoption Challenges & Solutions
• Workload Characterization & Impact on Controller Designs
• Server and Storage Attach SSDs - Product Segments
• Key Take-aways
IT Industry Journey - Roadmap

Analytics – BI
Predictive Analytics - Unstructured Data
From Dashboards Visualization to Prediction Engines using Big Data.

Cloudization
On-Premises > Private Clouds > Public Clouds
DC to Cloud-Aware Infrast. & Apps. Cascade migration to SPs/Public Clouds.

Automation
Automatically Maintains Application SLAs
(Self-Configuration, Self-Healing®, Self-Acctg. Charges etc.)

Virtualization
Pools Resources. Provisions, Optimizes, Monitors
Shuffles Resources to optimize Delivery of various Business Services

Integration/Consolidation
Integrate Physical Infrast./Blades to meet CAPSIMS ®IMEX
Cost, Availability, Performance, Scalability, Inter-operability, Manageability & Security

Standardization
Standard IT Infrastructure- Volume Economics HW/Syst SW
(Servers, Storage, Networking Devices, System Software (OS, MW & Data Mgmt. SW))
Data Centers & Cloud Infrastructure

Source: SSD Industry Report 2012 © IMEXResearch2010-12, Copying prohibited. All rights reserved.
2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
Workloads: Key to Computers Design

* IOPS for a required response time (ms) *=(Channels\text{*}Latency-1)

© 2010-11 IMEX Research, Copying prohibited. All rights reserved.
2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
Workloads Characterization

- **Data Warehousing Workloads** are I/O intensive
  - Predominantly read based with low hit ratios on buffer pools
  - High concurrent sequential and random read levels
    - Sequential Reads requires high level of I/O Bandwidth (MB/sec)
    - Random Reads require high IOPS
  - Write rates driven by life cycle management and sort operations

- **OLTP Workloads** are strongly random I/O intensive
  - Random I/O is more dominant
    - Read/write ratios of 80/20 are most common but can be 50/50
    - Can be difficult to build out test systems with sufficient I/O characteristics

- **Batch Workloads** are more write intensive
  - Sequential Writes requires high level of I/O Bandwidth (MB/sec)

- **Backup & Recovery** times are critical for these workloads
  - Backup operations drive high level of sequential IO
  - Recovery operation drives high levels of random I/O

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
## Storage Targeted Metrics by Apps

<table>
<thead>
<tr>
<th>Target</th>
<th>Applications</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>App Acceleration</strong></td>
<td>Web 2.0</td>
<td>Latency, $/IOP</td>
</tr>
<tr>
<td></td>
<td>Volume HPC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VZ SAN/NAS</td>
<td></td>
</tr>
<tr>
<td><strong>Boot Drive</strong></td>
<td>All Server Apps</td>
<td>$/GB</td>
</tr>
<tr>
<td><strong>Performance Boost</strong></td>
<td>Web 2.0</td>
<td>$/IOP/GB</td>
</tr>
<tr>
<td></td>
<td>HPC, OLTP/OLAP</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CDN, VoD</td>
<td></td>
</tr>
<tr>
<td><strong>Capacity</strong></td>
<td>Data Warehousing, Backups, Archives</td>
<td>$/GB, Watt/GB</td>
</tr>
</tbody>
</table>
Enterprise Storage: Issues & Solutions

Storage Requirements by CAPSIMS

Reducing Storage Costs
- Traditional Data Growth
- Snapshots ~ 75%
- Virtual Clones ~ 80%
- Thin Provisioning ~ 30%
- DeDuplication ~ 25-95%
- Auto-Tiering 65-95%
- Thin Replication ~ 95%

Improving Availability
- Data Protection
  - RAID 1, 5, 6, 10
  - Backup & Disaster Recovery

Improving Performance
- Flash SSD in SAS/SATA ~ 75%
- PCIe based Server Flash Cache ~ 85%
Server to Storage - IO Issues

- For Each Disk Operation, Millions of CPU Ops. or Thousands of Memory Ops. can be accomplished
- PCIe used for HBAs to connect to (External Shared Storage) via Storage Switches/Fabric as SAN/NAS on HDDs front ended by DRAM Cache creating a gap of 100,000x latency gap

### I/O Gap

- A 7.2K/15k rpm HDD can do 100/140 IOPS

<table>
<thead>
<tr>
<th>Operation</th>
<th>Time (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 cache reference</td>
<td>0.5</td>
</tr>
<tr>
<td>Branch mispredict</td>
<td>5</td>
</tr>
<tr>
<td>L2 cache reference</td>
<td>7</td>
</tr>
<tr>
<td>Mutex lock/unlock</td>
<td>25</td>
</tr>
<tr>
<td>Main memory reference</td>
<td>100</td>
</tr>
<tr>
<td>Compress 1K bytes with Zippy</td>
<td>3,000</td>
</tr>
<tr>
<td>Send 2K bytes over 1 Gbps network</td>
<td>20,000</td>
</tr>
<tr>
<td>Read 1 MB sequentially from memory</td>
<td>250,000</td>
</tr>
<tr>
<td>Round trip within same datacenter</td>
<td>500,000</td>
</tr>
<tr>
<td>Disk seek</td>
<td>10,000,000</td>
</tr>
<tr>
<td>Read 1 MB sequentially from disk</td>
<td>20,000,000</td>
</tr>
<tr>
<td>Send packet CA-&gt;Netherlands-&gt;CA</td>
<td>150,000,000</td>
</tr>
</tbody>
</table>
SSDs Filling Price/Performance Gaps

SSD segmenting into
- PCIe SSD Cache
  - as backend to DRAM &
- SATA SSD
  - as front end to HDD

DRAM getting Faster (to feed faster CPUs) & Larger (to feed Multi-cores & Multi-VMs from Virtualization)

HDD becoming Cheaper, not faster

Best Opportunity to fill the gap is for storage to be close to Server CPU.
• **Improving Query Response Time**
  
  • Cost effective way to improve Query response time for a given number of users or servicing an increased number of users at a given response time is best served with use of SSDs or Hybrid (SSD + HDDs) approach, particularly for Database and Online Transaction Applications

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
Data Storage: Hierarchical Usage

I/O Access Frequency vs. Percent of Corporate Data

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

- SAS Arrays
  - Tables
  - Indices
  - Hot Data

- Cloud Storage
  - Back Up Data
  - Archived Data
  - Offsite DataVault

- SATA
  - Back Up Data
  - Archived Data
  - Offsite DataVault

65%
75%
85%
95%

1% 2% 10% 50% 100% % of Corporate Data % of I/O Accesses

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
• Importance of QoS in Enterprise
  - Usually **Multiple SSDs working in parallel**
  - Usually **large no. of threads all reading and Writing**
  - For even a single large latency spike (even from outliers) many threads will wait on that single SSD causing a stall until drive recovers
  - Unlike HDDs, **SSD latency** spikes are not aligned, thus **are additive**
  - For Example a SINGLE 90 ms delay per 10,000 writes has an **impact of >25%**

• **QoS Difficult to Manage/Deliver on SSDs**
  - NAND needs to be managed in the background
  - Background Operations can hold off host activity/operations
    - Garbage Collection
    - Disturb Algorithms,
    - Detention Algos,
    - Wear Leveling
  - Level of Difficulty Comes from Workloads in cases when:
    - **Write Workloads** higher than Read Workloads
    - **R/W Mixed Workloads** higher than Mostly Read Workloads
    - **Random Write Workloads** Higher than Sequential Write Workloads
    - **Small Block Writes** higher than Big Block Writes
    - **Larger queue depth** higher than Shorter Queue depth
Managing Factors Impacting Performance

**Hardware**
- CPU, Chipset, Interface,

**System SW**
- OS, App, Drivers, Cache Algorithms,

**Workload**
- Random/Sequential R/W Mix, Queues, Threads, Parallelism

**SSD Device**
- Flash Generation, Caching, Wear-Leveling, Garbage Collection, TRIM, Purge

**Write History**
- Pre-Conditioning Amount
- TB Written, spares…

**Performance**
- “Burst” First On Board (FOB), Steady State post xPE Cycles,

Using NewGen Controllers, performance of MLC SSDs matching SLC SSDs

Additional performance gains with interleaved memory banks, caching and other techniques

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
SSD Challenges & Solutions: Storage Optimization by Apps/Workloads

Goals

Establish Goals for SLAs
- Performance/Cost/Availability, BC/DR (RPO/RTO)

Increase Performance for DB, OLTP and OLAP Apps:
- Random I/O > 20x, Sequential I/O Bandwidth > 5x
- Remove Stale data from Production Resources to improve performance

Use Partitioning Software to Classify Data
- By Frequency of Access (Recent Usage) and
- Capacity (by percent of total Data) using general guidelines as:
  - Hyperactive (1%), Active (5%), Less Active (20%), Historical (74%)

Implementation

Optimize Tiering by Classifying Hot & Cold Data
- Improve Query Performance by reducing number of I/Os
- Reduce number of Disks Needed by 25-50% using compression SW

Match Data Classification vs. Tiered Devices accordingly
- Flash, High Perf Disk, Low Cost Capacity Disk, Online Lowest Cost Archival Disk/Tape

Balance Cost vs. Performance of Flash
- More Data in Flash > Higher Cache Hit Ratio > Improved Data Performance

Create and Auto-Manage Tiering
- Monitoring, Migrations, Placements without manual intervention

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
Meeting Enterprise Requirements

Enterprise Requirements

- **Always-On 24x7 Reliability** and performance supersede cost
- **Fast I/O Performance** required by business-critical applications and
- **5-Yr. Life Cycle Endurance** reqd. by mission-critical applications in Enterprise.
- **Use State-of-the-Art** new sophisticated controllers and firmware technologies to run mission critical applications in the enterprise, using
  - Robust ECC, Internal RAID, Wear Leveling (To reduce hot spots), Spare Capacity, Write Amplification, Avoidance, Garbage Collection Efficiency, Wear Out Prediction Management etc.

<table>
<thead>
<tr>
<th>SATA3 I/F</th>
<th>New Intelligent Controller (2nd Generation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHY</td>
<td>Optimized Write</td>
</tr>
<tr>
<td>Link</td>
<td>Block Mgmt/ Wear Leveling</td>
</tr>
<tr>
<td>Transport</td>
<td>Garbage Collection</td>
</tr>
<tr>
<td>Command</td>
<td>Read/Disturb Management</td>
</tr>
<tr>
<td>AES 256/128 Encryption</td>
<td>RAID w/o Std. Parity OH</td>
</tr>
<tr>
<td>TCG Compliance</td>
<td>55b/512B BCH ECC</td>
</tr>
</tbody>
</table>

New Gen Controllers allow SSDs to meet Enterprise Class Availability/Performance/Over 5-Year Life/ Scalability/ Auto-Configuration & Auto Data-Tiering
Managing I/O Performance w Storage-Tiering

LBA Monitoring and Tiered Placement
- Every workload has unique I/O access signature
- Historical performance data for a LUN can identify performance skews & hot data regions by LBAs
Managing NAND Media in SSDs

Leveraging Long History of managing HDD’s imperfect media & high error rates

- Characterizing the quality & capabilities of media
- Allocating data based on quality of media

**HDD Media**

- $10^{-4}$
- Adaptive Signal Processing for Media Rd/Wr/Erase
- Advanced Bit Detection & Error Correction Codes
- Defect Management
- $10^{-16}$

**Flash Media**

- $10^{-4}$
- Adaptive Signal Conditioning for Flash Media
- Auto Bit Detection & Error Correction Codes
- Defect Management
- $10^{-17}$

Leveraging Long History of managing HDD’s imperfect media & high error rates

- Endurance for Long Life Cycle
- Reliability through RAID of Flash Elements
- Adaptive Digital Signal Processing Technology
- Dynamically adjust Read/Write characteristics of each chip
- Tune adjustments over life of media
- ECCs - PRML Deploying Enhanced Error Correction Codes

Source: IMEX Research SSD Industry Report ©2011
Factors Affecting Endurance

What Affects SSD Endurance

Work Load – Writes - Sequential or Random

Amount of Spare Area - Additional Spare improves efficiency of background data move operations on Random Write workloads

NAND Type – SLC or MLC

Firmware Algorithms/SSD Architecture
- NAND disturb/Retention Algorithm Efficiency
- Wear Leveling Aggressiveness / Effectiveness
  - Distribution of writes evenly across all flash blocks in the SSD to maximize overall drive endurance
  - Ensuring the wear of blocks does not exceed a max. P/E cycles for Flash Memory
- Garbage Collection Algorithm efficiency
- SSD Error Avoidance Algorithms

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
Managing Endurance in NextGen SSDs

Over-provisioning
- Extra spare raw blocks are designed-in as headroom and included to replace those blocks that get overused or go bad. Additionally provide enough room for wear-leveling algorithms to enhance reliability of the device over its life-cycle.
- Typical SSD device’s specified GB device will actually contain 7% for Standard (and Client) and 20-50% extra raw capacity for High Endurance Data Centers

Using High Endurance Technologies
- Intel Labs has demonstrated value of HET vs standard MLC to be better than 38X in achieving lower RBER
Best Practices - SSD in Enterprise

Best-of-Breed Products, Tools & Best Practices for Implementing SSD in the Enterprise

Best Practices

- By leveraging Error Avoidance Algorithms, and Best Practices of Verification Testing, to keep total functional failure rate <=3% (with defects and wear-outs issues combined)
- In practice, endurance ratings are likely to be significantly higher than typical use, so data errors and failures will be even less.
- Capacity Over-provisioning will provide large increases in random performance and endurance.
- Select SSD based on confirmed EVT Ratings
- Use MLC within requirements of Endurance Limits

Use Best-of-Breed Controllers to achieve <=3% AFR and JEDEC Endurance Verification Testing for Enterprise SSDs

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
### Tiered Storage - Supplier Ecosystem

#### NAND Flash
- Intel
- Micron
- Samsung
- Hynix
- Toshiba

#### SSD Controllers
- Anobit/Apple
- Indilinx/OCZ
- Innodisk
- Memorigin
- SandForce/LSI
- Silicon Motion
- Marvell

#### Flash IC
- AMD
- Atmel
- Fujitsu FMI
- Integrated Silicon Macronix
- Micron
- Microsemi
- On Semiconductor
- Samsung
- Sharp
- Silicon Storage Tech
- Spansion
- STMicroelectronic

#### SSD Storage
- BitMicro
- Fusion i-o
- Intel
- LSI
- Nvelo
- OCZ
- SANDisk
- Seagate
- Smart Modular
- STEC
- Texas Memory
- Viking
- Virident

#### Storage Systems
- EMC / FAST
- EMC Isilon / Smart Pools
- NetApp / Virtual Storage Tier
- QLogic

#### Computer Systems
- Dell / Data Progression
- Fujitsu/
- HP / Adaptive Optimization
- HDS / Dynamic Tiering
- IBM / Easy Tier
- NEC/
- Oracle / Flash Cache
- Supermicro

---

*Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.*
Target Market – Servers Storage
SSD as backend storage to DRAM as the front end
36 PCIe Lanes Availability, 3/6 GB/s Performance (PCIe Gen2/3 x8),
Low Latency in micro sec, Low Cost (via eliminating HBA cost)

PCIe SSD’s attributes of high IOPS, high Bandwidth, Low Latency
and lower cost are a good match for Caching
PCI Express as an SSD Interface

- **High performance Features**
  - Full Duplex, multi-outstanding requests, OOO Processing
  - Scalable Port Width (x1 to x16)
  - Scalable Link Speed (2.5/5/8 GT/sec)
  - Low Latency (No HBA OH or Protocol Translation)

- **Lower Cost Features**
  - High Volume Commodity interconnect
  - Direct Attach to CPU eliminates HBA Cost

- **Power Management Features**
  - Link Power Management
  - Optimized Buffer Flush/Fill
  - Dynamic power Allocation
  - Slow power limit

- **Rapid Market Adoption**
  - Adoption by leading SSD manufacturers for both Server and Storage side as Caching for Application Acceleration
PCIe SSDs supporting Next Gen Universal Connectivity

Multi-FormFactors Capability, Multi-Interface (PCIe/SAS/SATA) Connector

2.5” PCIe SSD Form Factor

Multi-function SAS/PCIe bay
- Uses SFF-8639 Multi-function connector
- High performance (up to 25W per slot)
- Hot swap, serviceability (SAS)
- High availability (2 fault domains)
- Supports a range of devices
  - 12Gb/s SAS
  - 6Gb/s SATA
  - MultiLink SAS (4 SAS Ports)
  - PCIe SSDs (emerging)
  - NVMe, SOP-PQI, Proprietary
  - SATA Express

SFF-8639 Multifunction Connector
- SFF-8639 is an enterprise backplane connector for 2.5” storage connecting PCI Express®, SATA®, and SAS® devices
- SFF-8639 includes 6 lanes, only 4 lanes are used at one time
  - PCIe: 4 red lanes on CPU PCIe lanes
  - SATA & SAS: 2 blue lanes on HBA/RAID controller or chipset

Compatible with existing SATA and SAS devices

Source: Solid State Storage Industry Report 2012 © 2010-12 IMEX Research, Copying prohibited. All rights reserved.
# NextGen Storage Interfaces

<table>
<thead>
<tr>
<th>Drive Form Factors</th>
<th>SATA</th>
<th>SAS</th>
<th>PCIe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive Connectors</td>
<td>SFF-xxxx</td>
<td>SFF-8680</td>
<td>SFF-8639</td>
</tr>
<tr>
<td>Express Bay Compatible?</td>
<td>Yes, 2.5&quot;</td>
<td>Yes, 2.5&quot;</td>
<td>Yes, 2.5&quot;</td>
</tr>
<tr>
<td>Drive Power (Typical)</td>
<td>9W Typical</td>
<td>Upto 25W</td>
<td>Upto 25W</td>
</tr>
<tr>
<td>Max Bandwidth</td>
<td>0.6GB/s (x2)</td>
<td>9.6GB/s (x4)</td>
<td>8 GB/s (x4)</td>
</tr>
<tr>
<td>Host Driver Stack</td>
<td>AHCI</td>
<td>IHV</td>
<td>Common Driver</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(SOP/PQI)</td>
</tr>
</tbody>
</table>

## SATA
- **Drive Form Factors:** 1.8”, 2.5”, 3.5”
- **No of Ports/ Lanes:** 1, 2, 4
- **Command Set/Que Interface:** SATA / SATA-IO, SCSI / SAS
- **Transfer Rate:** 6Gb/s
- **Drive Connector:** SFF-xxxx
- **Express Bay Compatible?** Yes, 2.5"
- **Drive Power (Typical):** 9W Typical
- **Max Bandwidth:** 0.6GB/s (x2)
- **Host Driver Stack (Stg Cntlr/Direct Drives):** AHCI

## SAS
- **Drive Form Factors:** 2.5”, 3.5”
- **No of Ports/ Lanes:** 1, 2, 4
- **Command Set/Que Interface:** SCSI / SAS
- **Transfer Rate:** 12Gb/s
- **Drive Connector:** SFF-8680
- **Express Bay Compatible?** Yes, 2.5"
- **Drive Power (Typical):** Upto 25W
- **Max Bandwidth:** 4.8 GB/s (x2)
- **Host Driver Stack (Stg Cntlr/Direct Drives):** IHV

## PCIe
- **Drive Form Factors:** 2.5”
- **No of Ports/ Lanes:** 1, 2, 4, 8 (on card)
- **Command Set/Que Interface:** SOP/PQI
- **Transfer Rate:** 8 Gb/s
- **Drive Connector:** SFF-8639 (2.5”), CEM (Edge-Card)
- **Express Bay Compatible?** Yes, 2.5"
- **Drive Power (Typical):** Upto 25W
- **Max Bandwidth:** 9.6GB/s (x4)
- **Host Driver Stack (Stg Cntlr/Direct Drives):** IHV

## SOP/PQI
- **Drive Form Factors:** 2.5”
- **No of Ports/ Lanes:** 1, 2, 4
- **Command Set/Que Interface:** SOP/PQI
- **Transfer Rate:** 8 Gb/s
- **Drive Connector:** SFF-8639
- **Express Bay Compatible?** Yes, 2.5"
- **Drive Power (Typical):** Upto 25W
- **Max Bandwidth:** 8 GB/s (x4)
- **Host Driver Stack (Stg Cntlr/Direct Drives):** IHV

## NVMe Express
- **Drive Form Factors:** 2.5”
- **No of Ports/ Lanes:** 1, 2, 4 (on card)
- **Command Set/Que Interface:** NVMe Express
- **Transfer Rate:** 8 Gb/s
- **Drive Connector:** SFF-8639
- **Express Bay Compatible?** Yes, 2.5"
- **Drive Power (Typical):** Upto 25W
- **Max Bandwidth:** 8 GB/s (x4)
- **Host Driver Stack (Stg Cntlr/Direct Drives):** Common Driver (NVMe Express)
NVMe – NextGen Flash Memory IF

<table>
<thead>
<tr>
<th>Storage Stack</th>
<th>Time $^2$</th>
<th>Clock Cycles $^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCSI/SAS</td>
<td>6.0 µsec</td>
<td>19,500 cycles</td>
</tr>
<tr>
<td>NVM Express</td>
<td>2.8 µsec</td>
<td>9,100 cycles</td>
</tr>
</tbody>
</table>

Prototype Measured IOPS

Cores Used for 1M IOPs

Linux Storage Stack

User Apps

Kernel

VFS / File System

Block Layer

Req

SCSI

NVMe

SAS

2.8 µsecs

6.0 µsecs

2012 Storage Developer Conference. © Insert Your Company Name. All Rights Reserved.
Intelligent Controllers - Key Take-Aways

Intelligent Controller Designs key to successful SSD Systems

Caching location Server vs. Storage Based
PCle based Flash in Servers meeting market success

Data Staging
Sequentialize Writes to overcome Random Write shortcomings

NVMe rising as NextGen Interface for SSDs
Serves both Servers and Client market segment as Flash emulating DRAM for performance at lower cost of Flash, lower pwr, scalability (Bypasses SCSI/SAS stack)

Universal Connectivity SFF Connector
Connecting PCIe, SAS, SATA

Adding Storage System Level Features
Encryption, Compression, Deduplication, Snapshots