I attended the Flash Memory Summit last week and took the following notes. Since it is a multi-track conference, I chose the sessions I thought of greatest interest to enterprise applications. I have highlighted in red key thoughts or concepts. I had a schedule conflict and was unable to attend the final day.

Notes from the Flash Memory Summit 8/11/2009

Forum 1C 8:30a

Jim Handy Chair - SSDs in the enterprise

Sumeet Bansal

Database Acceleration Using Solid State Storage—Practical Examples

Wine.com case study. This is an On Line Transaction Processing System
10x ramp Nov-Dec in business volume due to seasonal fluctuation.
RAID 1 with synchronous mirroring.
AVG latency on Write down from 4 to 1 ms on ioDrive
AVG latency on Read down form 12ms to 1ms on ioDrive
SQL xaction from 345 to 88 msd
Full DB backup from 2 hours to 6 minutes
Full DB restore from 3 hours to 15 min
500 ms xaction in 1 hour window from 3011 to 163

Rob Peglar – VP Technology at Xiotech (partially owned by STX)

SSD Technology – *Where Does It Fit For Customer Applications*
Comparison to IBM PC model 5150
Access density issue gotten much worse IOPS/GB
Is it cache or is it disk? Is it memory or is it a peripheral?
Planned Downtime is an Oxymoron
Applications don't want disks they want space
Applications don't want IOPS they want time
Applications do IO because they have to but they don't really want to
Unstructured data is a
- Poor fit for SSD
- Exception small non growing tagged files
- OS images boot from flash page to DRAM

Structured data is a
- Excellent fit for SSD
- Exception large growing table spaces
- DB have key elements that are excellent fit for SSD's

SSD should be treated exactly like magnetic
SAN based SSD=Good
- Not captive to server, scales
- Add more SSD drives as demand grows, online

Clustered Storage Types
- Type 1 single access captive storage
- Type 2 dual acess captive storage

Type 3 multi-access non-captive storage [N controller nodes networked with N storage nodes] requires intelligent storage elements, sparing at storage node level – grid allocation, head-level IO & mapping, active recalibration

Sang-Won Lee – Sungkyunjwan University (works w Samsung)
A Case For Flash Memory SSD in OLYP
Joint work with Indilinx
Vision: "Flash is Disk, Disk is Tape, Tape is Dead" Jim Gray
Enterprise is easier sell than consumer
"Migrating Enterpise Storage to SSDs: Analysis of Tradeoffs" (European conf paper) no advantage except OLTP
IOPS crisis in OLTP due to Moore's law growth in demand for IOPS
Compared 8 hard disks to 1 Indilinx SSD.
Transactions Per Second change over time interesting graph

Doug Dimitru – Easyco
Optimizing Flash SSD Applications w Linearizing Block Remapping SW
[This software was also discussed by Doug at the Denali MEMCON. It is very interesting]
A few hundred GB of system data in DB
Trying to improve native structure by not doing random writes by dynamically moving blocks on media so that all writes are sequential.
Linearization SW
Writes delayed not reordered
Data written to disk with header and footer including address of data?
Fast Block Device FDB still appears in drive names
90+% of the drives available linear BW is used
Write amplification reduced to 3:1 or a little less
So 2x MLC devices outlast SLC
3/4X MLC practical for SSD applications
DRAM memory overhead 1MB/GB (smaller arrays) to 1.25MB/GB (larger arrays >2TB)
Dedicated freee space 30% for 24x7 server apps, smaller for workstation.
Allows use of lower quality, commodity flash SSD's for enterprise apps?
Real performance of M-tron 5 SSD RAID server 60/30k IOPS read/write
1.6:1 write amplification
Looking for licensees

Questions

Doug Dimitru [EasyCo] – Alignment with 4k file systems optimal. Minutes to come up when RAM not stored. Uses log file system. Mix of reads and writes scales differently due to latency (SSD bad with SAS port expander); bursts of latency with streamed writes interrupt reads typically ~100ms or less.

Rob Peglar – Storage in host system vs SAN; can be anywhere if don't have to take down to add or subtract SSD. Some customers can't have any downtime. Consider scalability and downtime.

Sumeet Bansal – Can open "can" due RAID 1 mirror and replace HW w/o downtime.

Doug [EasyCo] – As storage moves further from host latency increases. Usual speedup of application of 10-15 to 1 doesn't reflect raw SSD improvement of 100:1

Part II of Forum 1C 10:15am

Chair John Vrionis, Lightspeed Venture Partners
Larry Chiu, IBM, Almaden – Quicksilver IOPS project
Roadmap for Enterprise Systems SSD Adoption
Placing the right data on SSDs to maximize the performance/cost benefit- want to automate learning process to enable smart data placement; heat map of hot data regions.
Dynamically recognize usage and place in "right" tier of storage. Results in response time reduction of 60-70%.
Double IOPS with same latency by using SSD for hot data
300% improvement in throughput using DB2 and Smart Tiering
System implemented in various IBM HW/SW
Workload Exercising →Workload Learning Thru Smart Monitoring→Smart Data Placement→Autonomic Performance Improvement

Marco Sanvideo, TCG
Securing Flash and Solid State Drives
Security is not only about encryption
Core architecture, logical channel that allows access control using Storage Working Group [of Trusted Computing Group] commands

Steve Garceau, Viking Modular
Why Do SSDs Mimic HDD Form Factors
SSDs have inherently more flexibility in size
PCIe
PCIe mini card form factor with SATA IF
Slim Light SSD SATA SSD 70% smaller than 2.5"
CFast is CF form factor with 3G SATA IF Capacities to 32GB 2-4 channel
Cube or Stacked SSD focuses on increasing z-height to reduce footprint with 3GB SATA IF
Choose the right form factor and other metrics for the job

Munif Farhan, Dell, Client Storage Sr. Eng.
Insight into SSDs Impact on Client Notebooks
20 Business and Consumer Platform offerings

What should the next features be beyond what we have?
Great device level vs system level performance impact! And power impact!
Still fears about endurance need to be addressed

Forum 2A Solid State Drives 2:40-5:30 pm

Chair Tom Burniece

Phan Hoang, Vitruim Technology

Integrating Solid State Storage and DRAM Into Standard Memory Module Form Factors
Integrate to make smaller, ligher, higher performance, lower cost
Processor FSB Memory Hub DMI IO Hub now single chip with off chip SSD/DRAM =
Virtium SSDDR = complete storage subsystem
8Gb=50nm, 16Gb=42nm, 32Gb=3xnm
Industry standard SODIMM module boots about 15 vs 30 sec
Used today in 2 single board computers (AMC card?)

Tony Lavia, Flexstar

How to test SSDs compared to HDDs
[Slide with good list of SSD tests is replicated here]

	Test Process		***Problem***	***Applicable Test Process***
1	Multiple writes		Endurance	1,3,4,5,6,7,8,9,11,13,14
2	Performance verification		Performance-mfg. variability	2,10,12
3	Disturb testing / pattern writes		Bit failures / Data Retention	3,4,5,
4	***Power cycling***		Component Failures	3,4,5,6,78,9,
5	***Extended test at temperature***		Write splice	9,14
6	***Voltage margining***		Metadata corruption	6,14
7	***Four Corners***		Write performance	2,12
8	***Voltage margining & 4 corner***		Erase failures	11
9	Power cycling mid writes		Design Margin	2,3,4,6,7,8,9
10	Random I/O w/ power cycling		Wear leveling performance	1,3
11	Margined erase		Data Retention	5,13
12	Fragmentation tests		Reallocation errors	1,3,10
13	RW tests w/ power cycling			
14	Write splice, cold writes			

First tester for SanDisk several years ago
Initially did same tests as HDD – Functioned same as HDD
Can test lube on HDD? Write for fractional days at high temp on single track then move off track and try to read adjacent tracks.
Flexstar saw stuff [developing SSD technology] as it jelled because startup SSD companies sought them out.
Like arms dealers sell to SSD and HDD
Tests: Power management,

Controller/NAND
- Wear leveling -
- Error management, incl power off data loss and fusing:
  - write shutoff – no error on power loss.
- Power management

NAND
- cycling
- Disturb (proram or erase)
- Data Retention

Proposals; virtual RPM, Endurance, Write amplification
Endurance SanDisk proposal Long-term Data Endurance LDE using TBW, TeraBytes Written determines life expectancy of SSD based on workload scenarios
Write amplification proposed by Intel: measure resulting actual data written vs. host data
if writing 4kB of host data results in 32kB of written data then write amplification = 8
SSD also has "spin-up" power peak
SSDs have more problems than HDDs because most companies making them are too new.
MLC focus has been on write disturbs and data retention
Only about 3 different SSD controllers seen so far.
NAND itself is either flakey or good. Controller and FW are major issues.
Not much testing being done in competitive advantage.
SNIA proposal for performance over time
Revising motherboard for tester due to Fusion IO speed

Esther Spanjer, SMART Modular
What's Up with These Numbers?
The need for performance benchmarking standardization
No standard terminology – examples
IOPS, Block Size, R/W Mix surface- where are we [ex from Calypso Systems]
Performance over time: Pre-conditioning is a must. All management algorithms must be operating otherwise non-deterministic latency
Workload dependency 90% drop
- IOMeter x
- HDTach/H2benchw x
- Everest x
- HD Tune
- PCMark
- SysMark

X's are enterprise testing optimum
Test Sequence recommended
- Pre-condition drive [incl fill cache before beginning]
- Run IOMeter for 3D IOPS view
- Block size 512b-1Mb
- Entire R/W mix range
- Validate performace stability
- Validate workload independency
- Run sequential test, random test, sequential test,; rukn work load simulations

Standards Activitiy

- Technical Working Group SNIA standard for performance benchmarking 1st draft to public 4Q09
- JEDEC 64.8 spec for SSD endurance measurement
- SSDA testing of reliability (power cycling, data retention, endurance, etc) and OS compatibility (Windows 7)

Sang-Yun Lee, BeSang

3D IC Architecture for SSD-in-a-Chip

BeSang in Korean means rising high
SSD-in-a-Chip 1/10 cost of conventional SSD
Process temps below 400C, vertical transistor, 5 memory layers, 0.5 um thick layers (with metal and insulators ~2 um per layer, 0.1F^t
Two vertical transistors; no endurance- low leakage current, low Soft Error Rate, E/W=endurance window
Pictures of Si pillar down to 3nm
Deposit flash on top of DRAM
8 bits per layer, goal .5 b/mask
In lab yet; have flash process technology not control logic or DRAM
Practical die size limit for all technologies due to defects ~300mm^2 (therefore limit ~250mm^2)

Kent Smith, SandForce

Benchmarking SSDs—The devil is in the pre-conditioning details

Past Writes Affect Future Performance [Very Good talk on testing SSD's]
Conditioning Crossover
Sequential and Random performance very different
Pre-conditioning assures repeatability of test results
Issues

- Advanced Host Controller Interface and associated drives
- NCQ and queue depth
- Offset and alignment
- Operating system background operations
- Boot drive vs. Secondary drive

Recycling or Garbage Collection

- Only during initial out of box is there no garbage collection
- Secure erase can restore to out of box state (fast way to known state)
- Begins typically just before drive capacity is reached

Past writes affect future performance

- Sequential writes create a few large blocks (areas) of free space
- Random writes will generally leave many small blocks of free space that makes recycling slower

Conditioning Crossover

- Random writes change over time in transisiton form sequential steady state to random

Inverse also true random to sequential steady state transition slowlhy improve to steady state

Real world is in-between BUT sensitive to immediate prior history

Testing for short periods of time will not necessarily disclose steady state performance

Queue depth important will create new set of data

Precondition for testing only wears out drive.

May need to write entire drive 2 to 5x; overprovisioning affects

Advance garbage collection may preserve junk data and wear out drive faster

(Therefore don't recycle all of drive in idle times)

Secure erase doesn't necessarily mean that every block on the drive is erased (spares, block flagged as invalid or bad?)

Must initiate test immediately after pre-conditoning

Time between commands means different performance with different controllers

Day 2

Tutorial 1A Designing Products with Flash Memory

Chair Deepak Shankar

Jim Cooke, Micron

ONFI Update: Tastes Great Less Filling

ONFI 2.1 is current standard
Block Abstracted NAND is a managed solution
Working with JEDEC ~1yr to publish and
ISSUES:
As device geometry shrinks latency is increasing due to page size increases to 8KB and beyond
ONFI module is 2 channels
Traditional asynchronous IF would be 40 MB/s ONFI supports 200 MB/s (166 saturates) with 34nm Micron NAND
In ONFI 2.1 ECC bytes added
Downloadable at ONFI.org
Path to 400 MTransfers/s
Shorter channel, wider spacing between signals, on-die termination, complementary clock and DQS signals
On-die termination is magic that allows 400 MT/s
In ONFI 2.2 will be able to suspend erase when high priority read arrives
ONFI 2.1 is DDR IF; with 2 channels and ONFI 3.0 800MB/s possible
ONFI 3.0 to arrive middle of next year
Micron supports ONFI in all designs but needs DQS pin but no cost differential?

Lakshmi Mandyam, ARM

Storage SoC Controller Trends

Application requirements
Issues: power, performance, cost; enterprise primarily power but also power
Green IT, enterprise reliability and availability, security, application acceleration (SAP, Oracle, web cache driving performance)
Cost is not only Si but, dual sourcing, supporting component cost, pin count, development cost (tools, ease of programming, time to market), scalability, performance/$ (eg standard high performance math features) also debug and trace capability
Enterprise SSD architecture
Single core moving to dual core 400-1000 DMIPS, ECC support on all memories
Host IF – SATA 2 going to SATA 3 (6GB/s) IOPS at 10k going to >50k
Flash – Moving to ONFI 2.0 4-10 channels today moving to 16-20 channels plus enhanced ECC with separate small core
Cache sizes roughly 2x page size?

Deepak Shankar gives Takeshi Ohkawa's paper

Performance Impact of Flash Memory on Multi-Core Android based Smart Phone

Main movtivation power consumption reduction

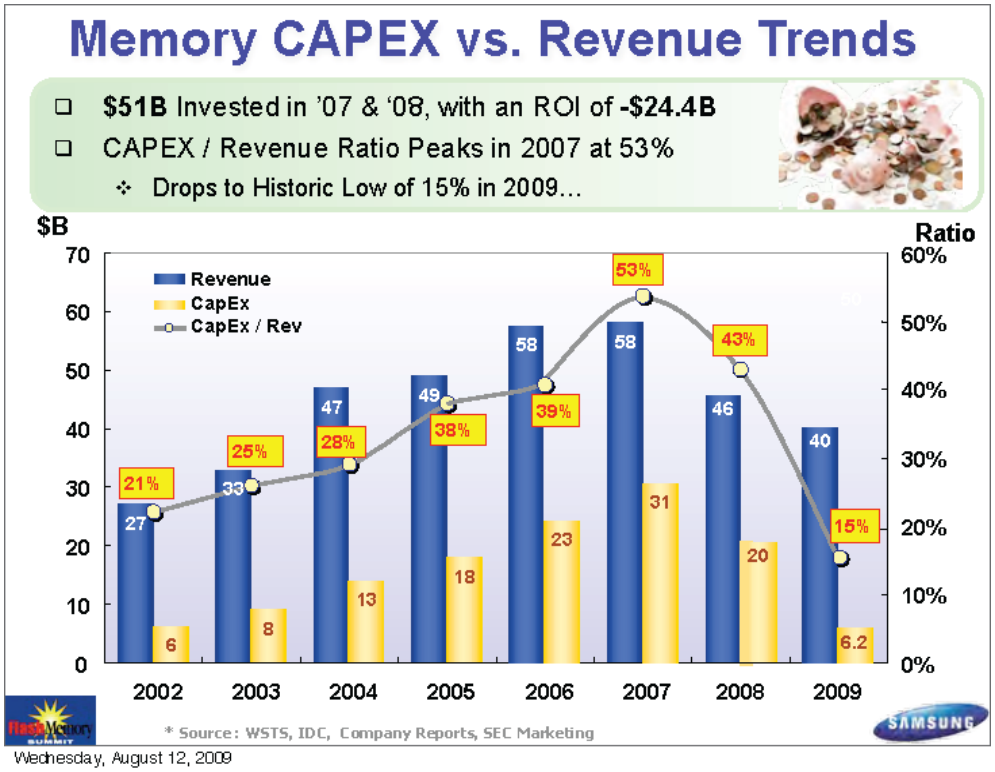
Virtual prototype using VisualSim platform runs on Windows and Linux under Software platform in QEMU w/ HW in VisualSim
Significant Market and Focus on Android
Used VisualSim to set up Performance and Power meters showing
Performance: flash, CPU, SDRAM, WiFi
Power: CPU, SDRAM, flash, WiFi, LCD, Touch Screen
Wanted to expand from phone to Netbook to Set Top Box, etc.
TOPS Systems makes custom multi-core processors for high performance computing
Using multicore architecture to solve power problem. More efficient than single core running at high speed
QEMU is like VMWare but open source
QEMU runs functions but VisualSim provides HW timing and power thru CORBA IF
Achieved 10-20 MIPS for a cycle-based and Approximately-Timed simulation running

Allesandro Fin, SMART Modular
PCIe Do We Need Anything Else
Defined in 2004 by IBM, HP, Intel, Dell
Serial, point-to-point, up to 32 lanes (full duplex Tx/Rx pair)
5.0 Gb/s x lane
Version 3.0 in 2010 8Gb/s/lane
Today at least 6 different IF's for Flash in enterprise, many limited in scale, BW, architecture
Using PCIe only limits IF to one, OS drivers to 1, form factors to 2, high performance scales, high capacity scales, single host controllers, high bandwith (multi-lanae)
This implies easier host design, shorter host driver debug cycle, easier mechanical design, easier migration to next generation performance/BW requirements
PCIe bus is there already with most CPU Chip sets therefore no host controller needed (SAS or SATA)
SATA SSD advantage is that it plugs into existing system; PCIe requires a proprietary custom PCIe driver from? OS supplier may not have core competency for PCIe driver and SSD vendor may have no OS core competence.
Intel NVMHCI advantage is standard
Mini PCIe is one lane only

Gilat Chitayat, QualiSystems, Israel
Automatic, Fast, and Thorough: Automatic Test of Flash Memory Cards
SW development co. SanDisk Israel QA lab is customer
Testing Includes:
Format/Partition, FW setting, HW measurement , RW errors and timing, etc.
Single cycle can take days; complete test cycle can reach a month per product
Biggest issue is manual result collection, data aggregation
TestShell Solution
SW drives and manages entire testing process incl HW IF
Central repository

TestShell allows writing test sequence w/o programming skills (GUI point and click), runs test and generates customizable reports (library provided) then store results in central repository
Interfaces with existing API DLLs plus all common test equip so very quick to bring up and customize
Set and measure current consumption
Set & measure signal and timing behavior through scope
Activate additional equip
Initiate SanDisk DLLs to format, partition, read, write and more
Test that originally took ~1 mo reduced to 2 days
Automatic means overnight and weekend operation possible as well (ie. Unmanned)
Uses main SQL DB

Jim Elliott, VP Memory Mktg Keynote
NAND Mkt update incl new killer apps
-13% Market reduction this year
$51B invested w ROI -$24B



Prolonged density life cycle → cost reduction decreasing
FCST price erosion ~-30% 2011, 12
Demand patterns shifting

Single use device vs. convergence eg. flip video, Amazon kindle
Mobile phones—holding own but Smart Pho9nes continue to grow: both numbers and memory content.
SSD—Key growth engine moving forward; no more than 64GB for enterprise laptop [still?!], gaming PC (due frames/s and load times),

Social Networking as a Killer-App
Facebook, YouTube Twitter, etc.
Twitter 16x CAGR, 44.5M unique visitors in June '09, Dell selling re-furbs w/ twitter
Just over 1M units in 2009 to 7M in 2013 in Enterprise ($2.2B)
Energy Star spec for servers now
How many people does it take to fail the internet? 1 if its Michael Jackson
Prevent the Twitter "Fail Whale" w/ SSD

Francois Piednoel, Intel (Sr. Performance Analyst)
8 core Intel desktop. pc
Memory → Nehalem: done
IO→ PCIe: done
Integration coming soon
Storage?

Intel SSD and Cache
Accessing data is issue
SSD in laptop is better for performance than adding a separate graphic card
New 80GB Intel SSD at $220
Sea of picture/video data example
Picture preview is an extra wasteful file
Calendar picture zoom-in [Intel custom app] can saturate 3.4GHz 8 thread processor (due to indexing of data and decode of .jpg?)
Larrabee will increase need for storage performance

Session 103

Chair Alan Niebel, Webfeet

Going beyond raw IOPS

Raj Parekh, Virident (founder and CEO) ex-CTO SUN and SGI

Flash in the Data Center

[Virident and Schooner both sell similar but different SSD appliances]

Focus on Cloud and Web 2.0-- cost effective scaling, agile provisioning (super-virtualization, and energy efficiency

Ex CTO's of Google and ? Other founders

30+% CAGR for IT spending

Big market with big barriers

Inefficiency multiplies due to server designs from pre-internet era

Design for failure to be contained in smallest envelope so it doesn't bring down entire data center

Must take account of user level,, application level, system SW level, device level and chip level

Custom ASIC w 3GB/s BW up to 2TBf flash 3-5M IOPS, 25MM Read cycles

Less than 100s (30s after server boot) to full warm cache after power fail or attack

When flash chips are designed for enterprise apps instead of PDA's and flash cards more performance gains possible.

John Busch, Schooner Information Technology (ex-SUN researcher, HP) sold by IBM

The DNA of Next Generation Data Centers

As commoditization continues then specialization and local optimization occurs around a set of standards which prevents taking full advantage of the underlying technology.

Scaling by adding more and more systems and GbE not able to take advantage of multi-core processors and flash memory

Replace with tightly coupled HW architecture and SW Administrator

New platforms optimized specifically for specific applications

Replace DRAM cache with flash gets order of magnitude improvement in performance, power. Also replace disk with Flash?

Morgan Littlewood, Violin Memory

Flash Appliances for the Data Center

Enterprise grade Si storage

Work w 10s to 100s of TB in most data centers

70/30 RW mix, 24/7 incl sustained writes Oracle, SMP, email → random writes important

Power—must reduce total data center power w/o taking out servers

Logical place for flash is on data center fabric to allow access from all resources

Treat as accelerator for all applications not just a few specific ones

Get ~ 100x IOPS per shelf vs. traditional

Product is purpose built memory appliance; no server, pure memory; unique RAID algorithm specific to flash (not RAID 5 or 6)

Power savings from reducing numbers of CPU's needed and reducing spindle count or higher power spindle count.
Move 80% of IOPS into flash
Cliff of death—when drive is full. Question, how steep is the cliff? Non-blocking erases [check this slide for more details]
Customer moves individual high activity LUNs from rotating to flash
Typical 20x improvement in latency and IOPS for Oracle

Adam Leventhal, Sun (ex-Cisco many years and sold startup to Cisco)
The Need for Higher-Level Software in Flash
Fishworks Flash Architect?
Flash combined into ZFS for traditional NAS box
Hybrid Storage Pool
Lithography Death March in Michael Cornwell KN
Need new IF for flash: PCIe or NVHMCI
Smarter SW enables all this: Allows use of dumber HW

Tony Roug, Intel
Flash Storage: Unlocking the Data Center IO Bottleneck
Principal Engineer in Digital Enterprise Group
Focuses on how servers are impacted by flash
Old TPC-C benchmark from HP data
tpmC? And $/tpmC
Storage costs dominate 42-74% of total
SW must change to take advantage of new HW, so for now data centers just replace HDD w SSD
IO to processor is bottleneck; mitigate with DRAM as cache and multiple HDD spindles w/ now substitution of NAND for DRAM cache and HDD.

Questions:
Raj Parekh
Concept of erase needs to be integrated into OS's. Not there because DRAM and HDD don't need it.
Let system do garbage collection and ECC. Allow system to see wearout of cells.

John Busch
It's really all about the software. Benefits realized today could be 5x with rewritten SW. Don't need more IOPS or cores. PCIe unnecessary. Need the application rewritten to take advantage of what's there. Rewrite takes cores from ~10% utilization to 100%. Raj agreed. Can run some applications flat out

Morgan Littlewood
Some applications better than others for SSD utilization as written.

Adam Leventhal—Need more direct raw access to flash to further optimize ZFS

Virtualization only at server level not cache level

All talking about balanced systems

VM Motion allows virtual machine to move from one system to another
Data motion is missing because it takes too long
Want to move VM w/o moving data

RAW NAND not delivered to industry because tuning of FW required for MLC

End of SSD session

Tutorial T2B: SNIA Tutorials — SSDs in Enterprise Storage (Part II)
Martin Czekalski, Seagate
SSD Enterprise: Ready or Not?
[This talk was the last in an session [SSDs in Enterprise Storage] that overlapped the previous one. Some of this presentation is missing. Some useful concepts were discussed.]
Support--- Forensic logging capabilities
Performance needs to be predictable and consistent
%Life remaining—T13 proposal; T10 SAS(TBD)
Cliff of death slide data from SNIA and Calypso Systems
STX proposed to JEDEC Endurance factors for application classes
Client 1,2 and Enterprise 1,2
Enterprise 2 example 2500GB/day writes 60/40 RW, 24/7, 55C 6mos data retention, 10^{-16} BER; no downtime
Robustness: Validation difficult. Need tests for IF compliance, exception and error handling, application compatibility
Infrastructure Maturity Issues
Optimize components in the stack:
- HBA and RAID controllers, drivers and storage protocols, file systems (trim [as proposed by T13 needs work has security hole??] and thin provisioning (SCSI, T10), applications

Management—Still in infancy
- How to install and migrate data onto new devices while minimizing disruption
- Added complexity of operations: tools to automate data placement and auto migration
- Effect on BU/Recovery, disaster recover, archive and ILM processes
- Inclusion vs. another island
- Tools for optimization of performance and cost

Sustainable Technology Roadmap
- Decreasing performance and endurance forecast is at odds with enterprise requirements!
- Will there be different architectures or approaches? Which to choose?
- Standards are mostly at an early state
- Flash is not like DRAM: many more considerations
- Second Sources required

What should users do?
- Homework understand applications and requirements
- Look to server/storage providers for integration and validation, application integration and validation and tools for ease of use.

END of NOTES