Using MLC NAND in Datacenters (a.k.a. Using Client SSD Technology in Datacenters)

Tony Roug, Intel
Principal Engineer

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- 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

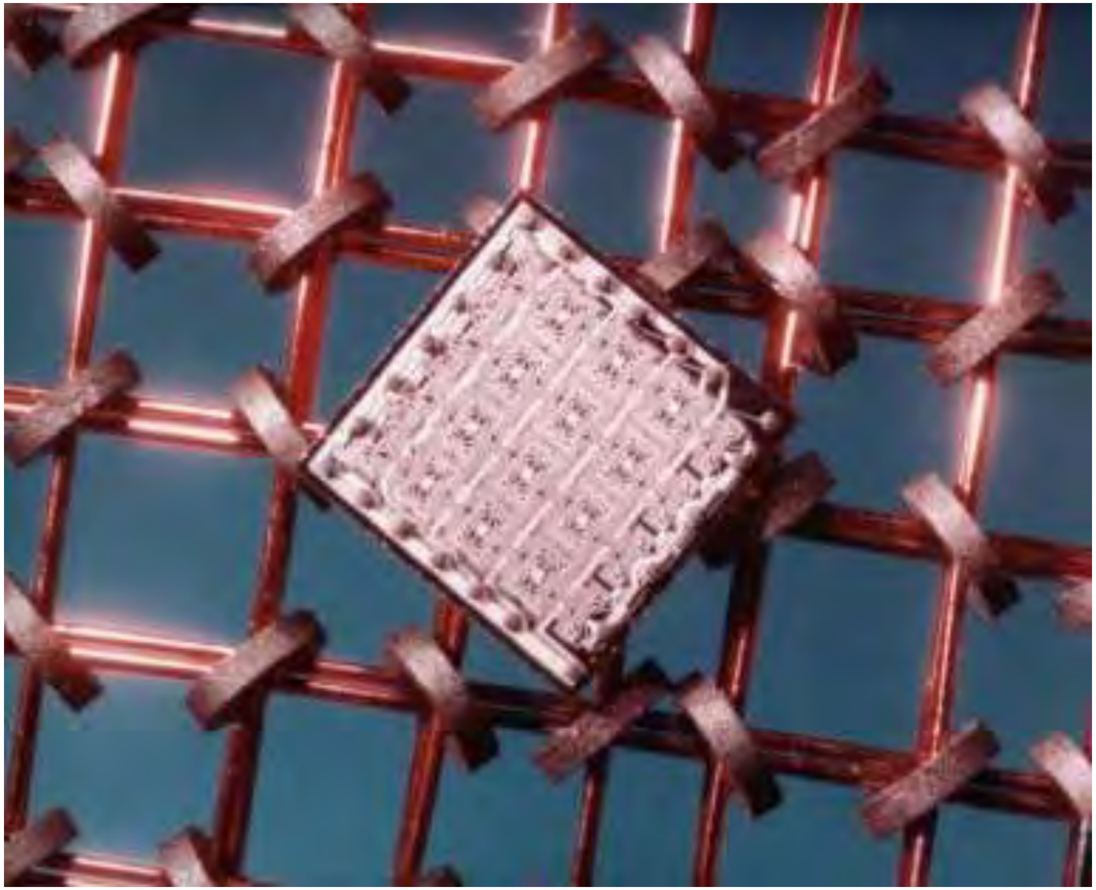
- SSDs are typically constructed using SLC as the datacenter NAND technology. Primarily because of the endurance and write performance of SLC NAND. For many usages, MLC NAND is more cost effective and can support the endurance and write performance required by the end user. This course outlines the different NAND usages in datacenter and highlights how MLC is a cost effective solution for datacenter applications.

- Learning Objectives
  - Understand tradeoffs for SLC NAND versus MLC NAND in datacenter
  - Understand how for specific applications MLC NAND is more cost effective
  - Understand how to tune MLC NAND to your application needs

Agenda

- Basics of NAND technology
- Basics of datacenter workloads
- MLC datacenter SSD
- Workload Examples

Question: What is the picture?



Integrated circuit foreground, core memory background

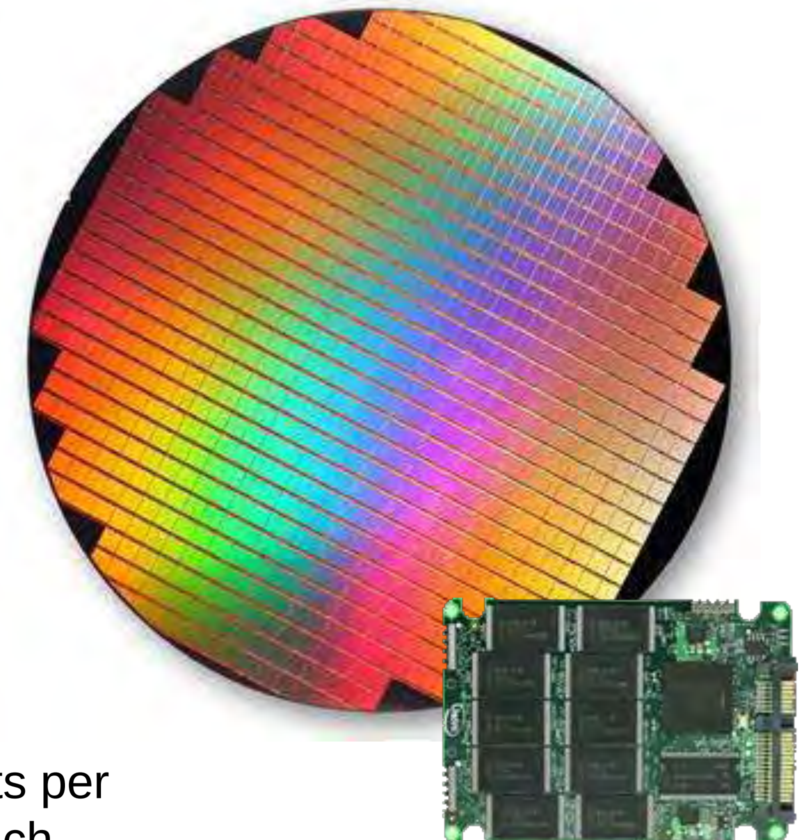
NAND…

<u>Hard Disk Drive Platter</u>

record data by directionally magnetizing ferromagnetic material

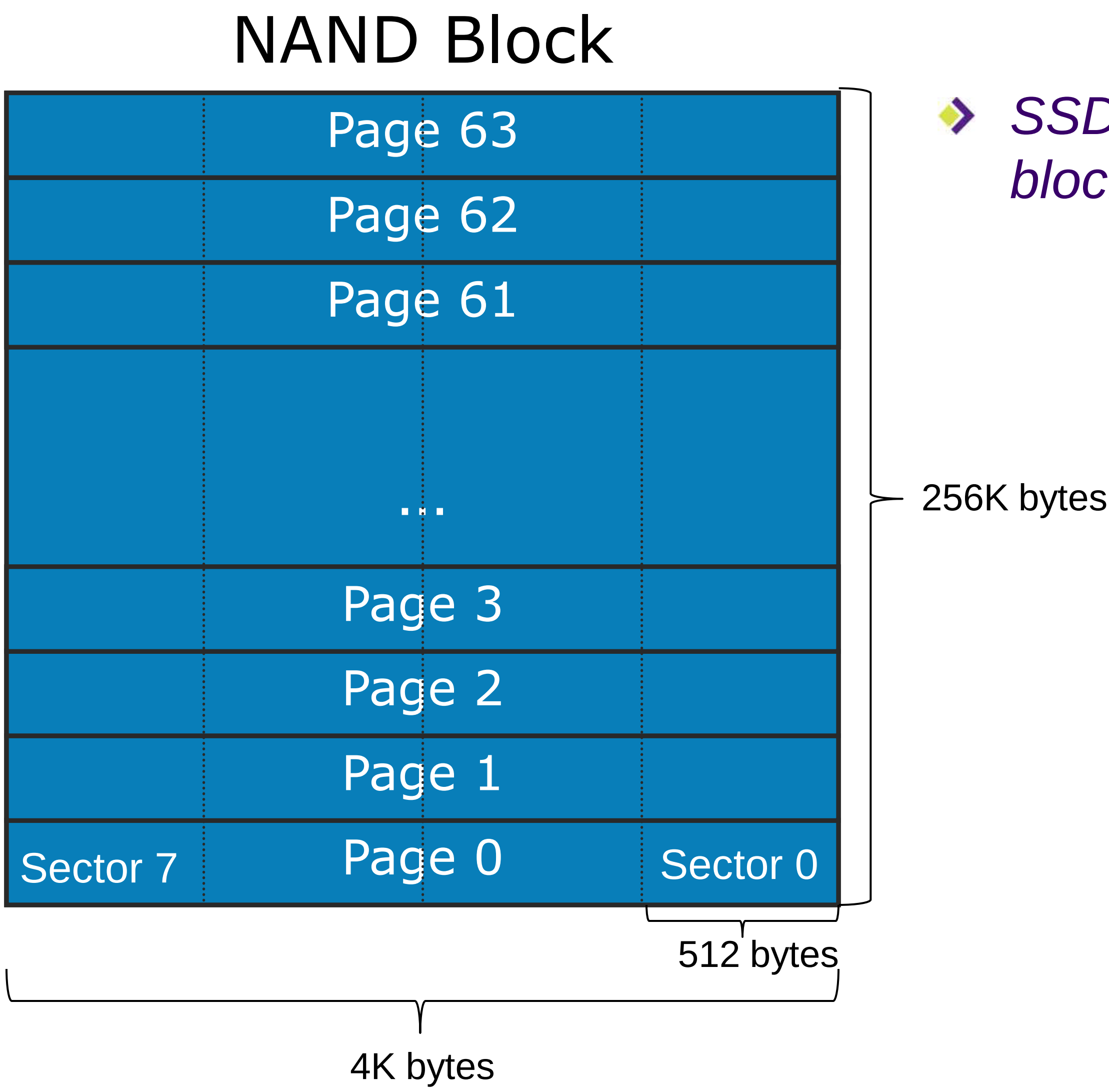
<u>NAND Silicon</u>

records data by “flashing” electron charges in an array of floating-gate transistors



>1 billion bits per square inch

NAND on solid-state drive



- *SSD NAND is arrange as blocks, pages, and sectors*

NAND on solid-state drive

Education SNIA

NAND Block

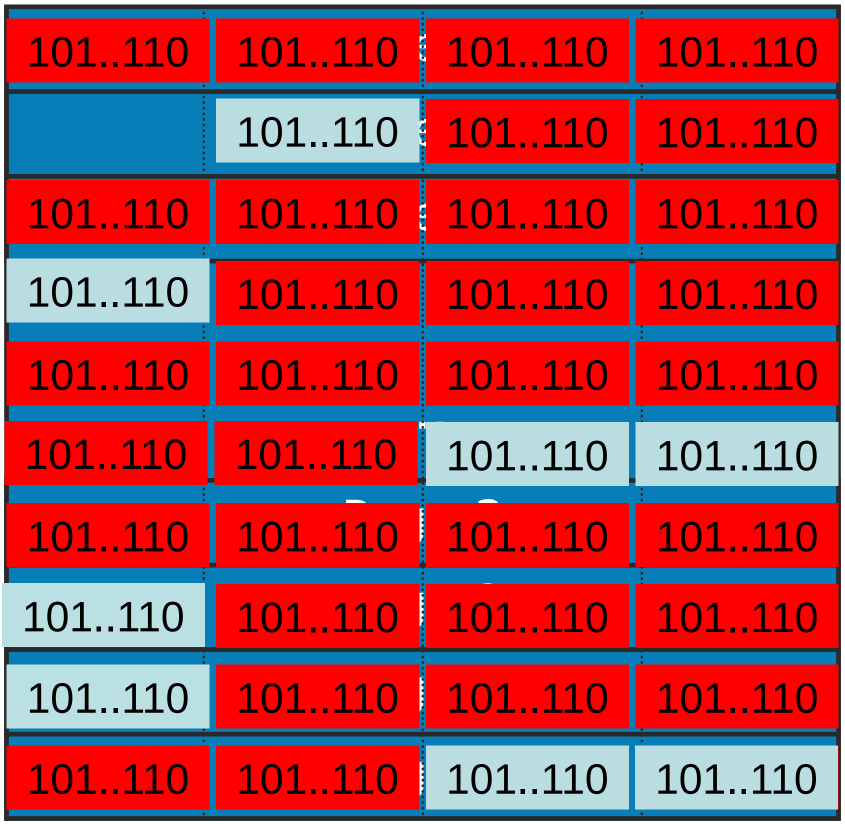
101..110	101..110	101..110	101..110
	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110
101..110	101..110	101..110	101..110

Valid Data
Invalid Data

- *NAND sector/page is **write once***
- *NAND sector/page is **read many***

NAND on solid-state drive

NAND Block

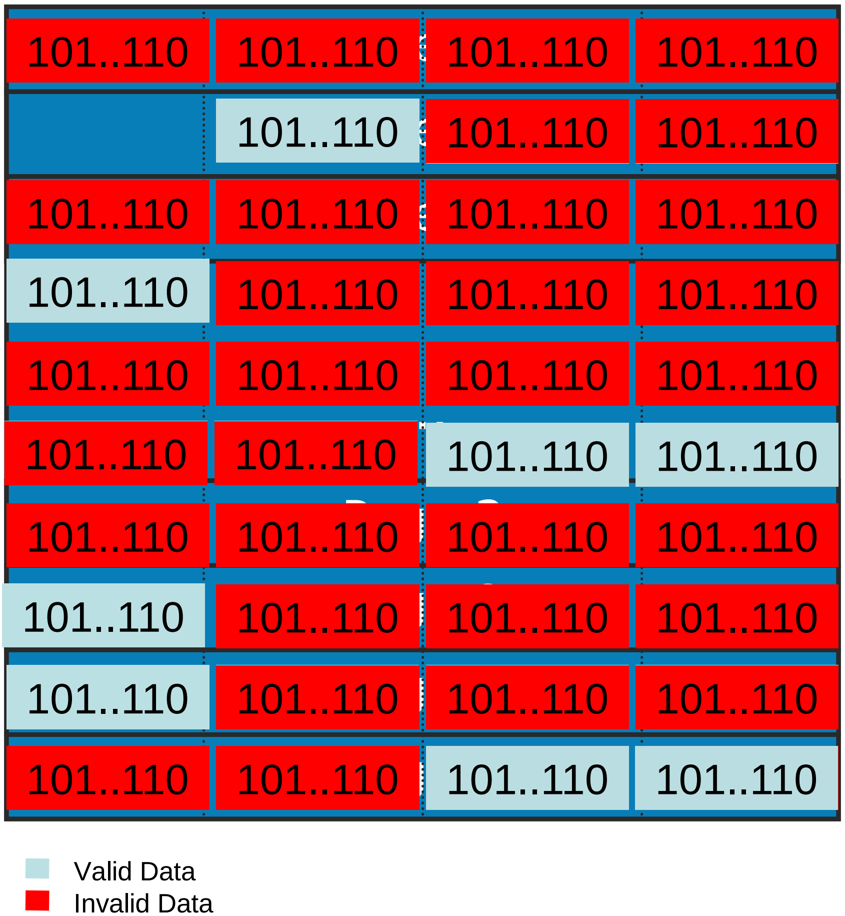


Valid Data
Invalid Data

- *NAND sector/page is **write once***
- *NAND sector/page is **read many***
- *When a NAND page is "full" and "aged", the page is first cleared, unused and cleared NAND creates in **write-amp (WA)***
- *and then **erased***

NAND on solid-state drive

NAND Block

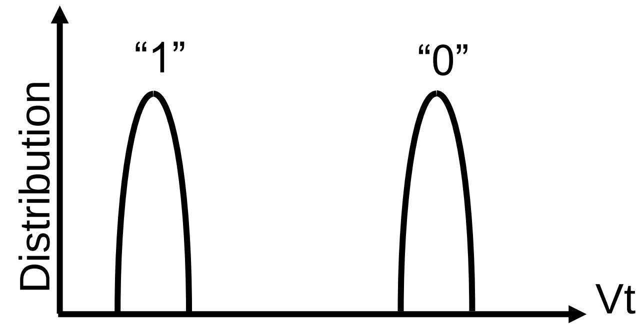


- *NAND sector/page is **write once***
- *NAND sector/page is **read many***
- *When a NAND page is “full” and “aged”, the page is first cleared, unused and cleared NAND creates in **write-amp (WA)***
- *and then **erased***
- *Each block erase is a **cycle***

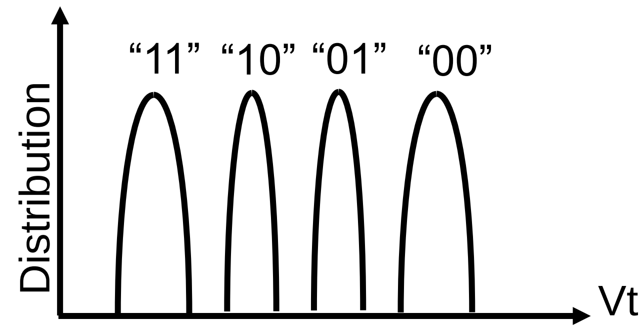
NAND SLC, MLC, etc

- Single Level Cell (SLC): 2 voltage levels
  - Level 0 = Erased = 0
  - Level 1 = Programmed = 1
- Multi-Level Cell (MLC): 4 voltage levels
  - Level 0 = Erased = 0
  - Level 1 = Programmed to L1 = 01
  - Level 2 = Programmed to L2 = 10
  - Level 3 = Programmed to L3 = 11
- Others
  - 2.5 bits per cell: 6 voltage levels
  - 3 bits per cell: 8 voltage levels
  - 4 bits per cell: 16 voltage levels

SLC: 2 Levels → 1 bit/cell



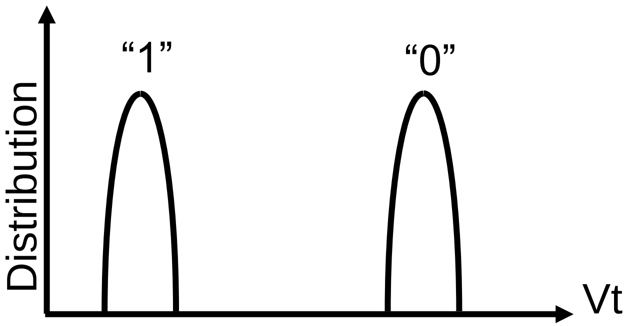
MLC: 4 Levels → 2 bit/cell



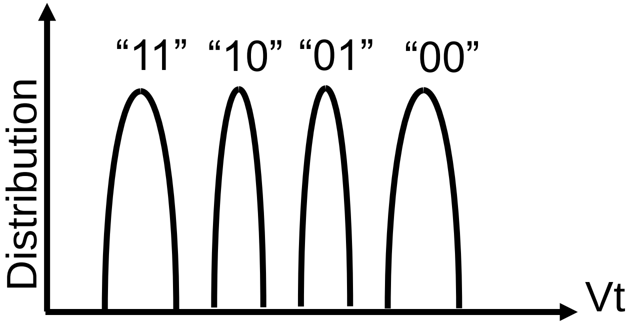
Basics of Solid State Drive NAND

Typical Specification	SLC	MLC
Bits per Cell	1	2
Page Size (K)	4	4
Pages/Block	64	128
Page Program (us)	250	900
Random Read (us)	25	50
Block Erase (ms)	2	2
Typical Program/ Erase Cycles	100,000	10,000

SLC: 2 Levels → 1 bit/cell



MLC: 4 Levels → 2 bit/cell

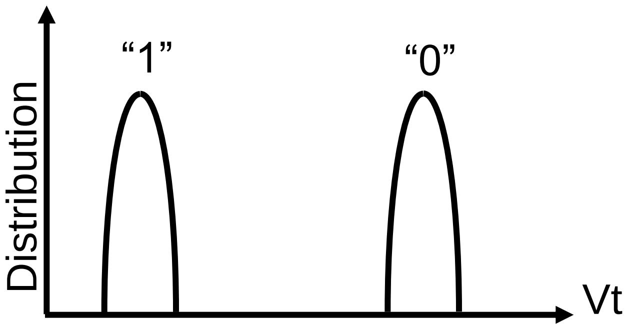


Highlighted specs affect ROI for SSD use in datacenter.

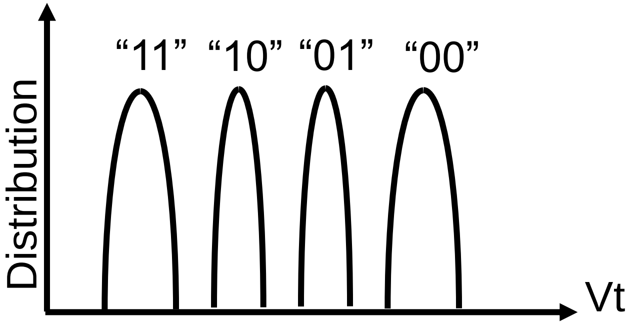
SLC & MLC Endurance/Lifetime

- JDEC shelf life 1 year
  - SLC - 100,000 cycles
  - MLC - 10,000 cycles
- Program/Erase Cycle
  - NAND block clear
  - Write amplification
- Bottom Line – It depends
  - Controller design
  - Firmware design
  - Usage Case

SLC: 2 Levels → 1 bit/cell



MLC: 4 Levels → 2 bit/cell



What Impacts Endurance?

Education SNIA

NAND Technology
erase cycles (SLC vs MLC)

Write Workload
Random vs Sequential

Spare Area
Capacity reserve / work space

Managed by:

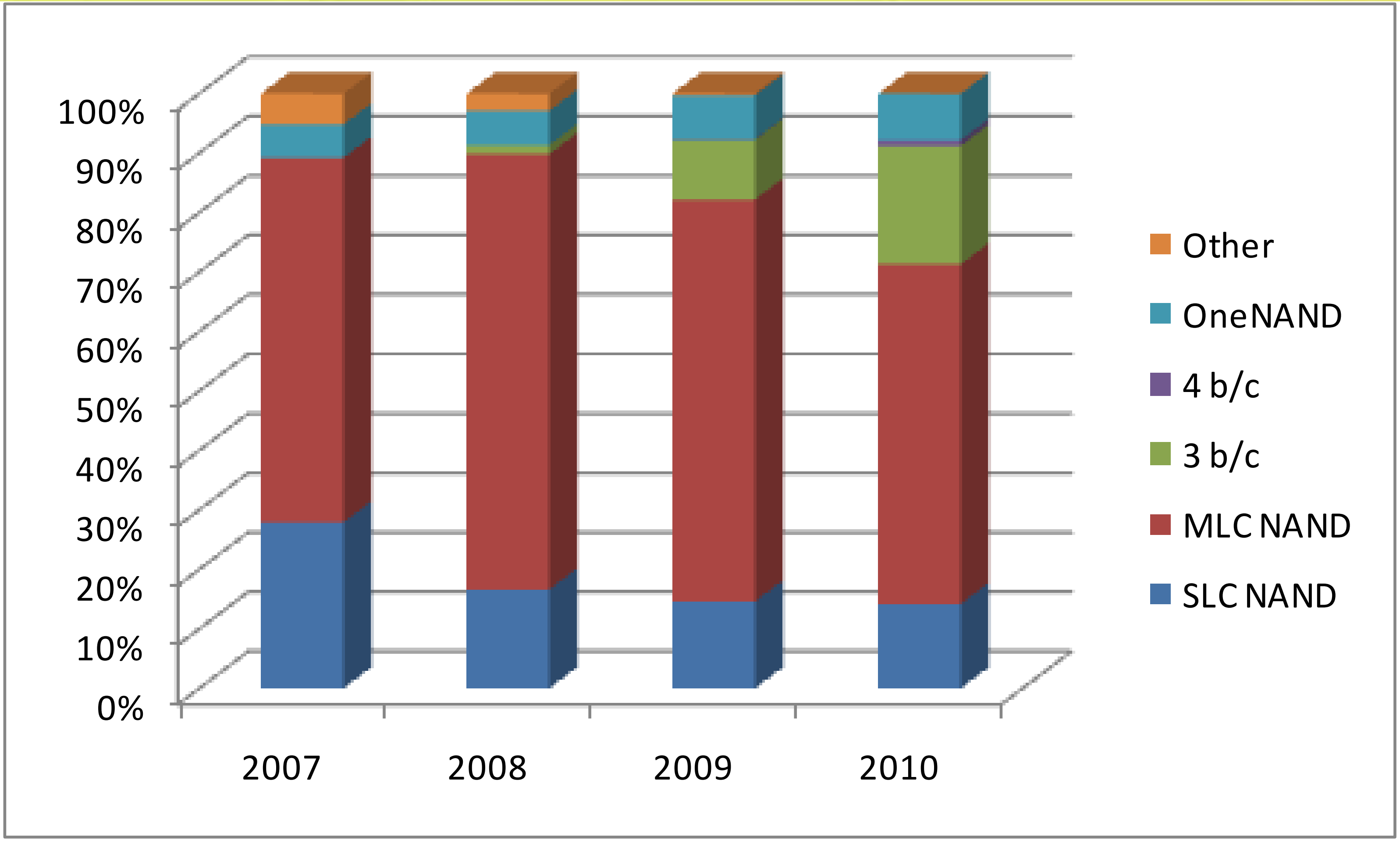
Firmware Algorithms
Efficiency of NAND writes (Write amplification) and wearleveling

Delivers:

Drive Endurance
Drive design and arch matters!

Lower write amplification→ Fewer NAND cycles→ Faster write perf
High Random Writes = Endurance Efficiency

SLC / MLC

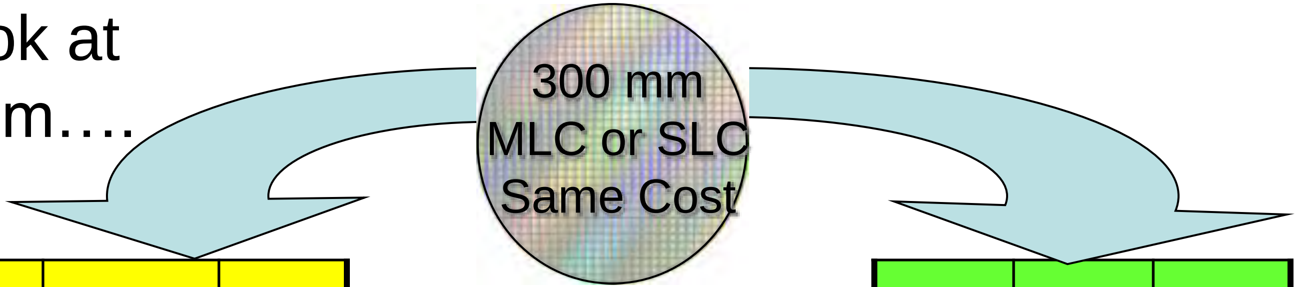


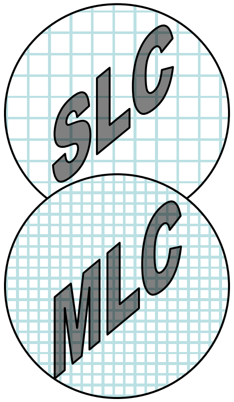
- NAND moving to greater charges per cell. Greater capacity at the expense of endurance and write speed

Choice of MLC vs SLC at Given Density

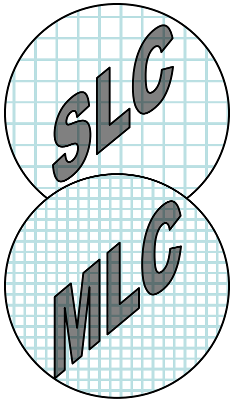
Education SNIA

Two Ways to look at the SLC Premium….





	Density /die	Die /Wafer	GB /Wafer
SLC	X	X	X
MLC	X	2X	2X



	Density /die	Die /Wafer	GB /Wafer
SLC	X	X	X
MLC	2X	X	2X

SLC requires an incremental 70-90% $/GB or suppliers produce other products

***Showing worst case for SLC - Actual MLC / SLC Delta ~ 80-90% assuming optimizations, timing issues, etc.**

Agenda

- Basics of NAND technology
- Basics of datacenter workloads
- MLC datacenter SSD
- Workload Examples

Legacy Storage in the Datacenter

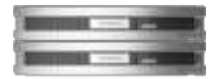
Storage Usage	Central Storage	Caching/ Proximity Tier	Server Attached
Cache 	**Local DRAM cache:** (e.g. OS block, hot application data)	**DRAM cache:** (Memcached, Storage virtualization appliance)	**Local DRAM cache:** (HDD Block Cache, Database Cache, NFS v4 cache)
Boot 	**Local boot data:** (SAN/NAS image)	**Local boot:** (Appliance image)	**Local boot data when not Pixie:** (Operating System, Hypervisor, SWAP, VM, Application Image)
Performance 	**Hot Application Data** (Database, Email, etc)		**Hot Application Data** (Web, Database, Email, Search, Videos, etc)
Capacity 	**Cold/Luke-warm Application Data** (Data warehouse, Documents, Backups, Archive, etc)		**Luke-warm Application Data** (Web, Email, Videos, etc)

Fabric (FC and GE/10GE)

Fabric (FC and GE/10GE)

Emerging NAND in the Datacenter

Education SNIA



Storage Usage	Central Storage	Caching/ Proximity Tier	Server/Workstation Attached
Cache ($/IOPS, Latency) 	**LBA Cache:** SAS/PCIe SSDs (**SLC**)	**LBA and Data Cache:** SAS/SATA/PCIe SSD (**SLC**)	**LBA Cache:** SATA SSD (**SLC**)
Boot ($/GB) 	**Local Boot Data:** SAS/SATA SSD (**MLC/3BC**)	**Local Boot Data:** SSD (**MLC/3BC**)	**Local Boot Data when not PXE:** SATA/SATA SSD (**MLC/3BC**)
Performance ($/IOP/GB) 	**Hot Application Data** SAS SSDs (**MLC**)		**Hot Application Data** SATA SSDs (**MLC**)
Capacity ($/TB, Watt/TB) 	**Cold/Luke-warm Application Data** SATA HDDs (future SSD 3BC/4BC?)		**Luke-warm Application Data** SATA HDDs (future SSD 3BC/4BC?)

Fabric (FC and GE/10GE)

Fabric (FC and GE/10GE)

The ROI Basics – SNIA TCO Calculator

Education SNIA

- For 3.5TB business intelligence database
  - 146G 15K SAS drives (SNIA default data)

1	What is the I/O transfer size in KB?	8K	example:
2	Which of the following most closely characterizes your application's I/O?	65/35	%reads/%writes

4K - 8K	16K - 32K	64K - 128K
OS, Transactions	Large File Transfer	Video Streaming

For questions 3 - 11, input information about your current Hard Drive configuration.

3	Select your current HDD size	2.5 Inch	
4	Select your HDD storage interface	SAS	
5	Select your current HDD RPM	15K rpm	
6	Select your current HDD per unit raw capacity	146 GB	
7	How many ***Hard Drives*** do you currently have in the application?	24	
8	What percentage of your Hard Drive capacity is ***consumed***?	100%	*(Include headroom if needed)*
9	What RAID configuration do you ***currently*** use?	No Raid	
10	Do you ***currently*** purchase maintenance plans for your hard drives?	no	
11	How many ***instances*** *(or systems)* of the above configuration do you have?	1	

  - Intel® X25E 64G SLC drives (www.newegg.com 3/21/10 pricing)

TCO Impact	I/O Performance Improvement	IOPS Gain	Reduction in power	Previous HDD Total	Total SSS	Total HDD Consumed	Total Usable SSS Capacity
($40,948)	2554%	285,150	58.3%	24	55	3504 GB	3520 GB

  - Intel® X25M 160G MLC drives (www.newegg.com 3/21/10 pricing)

TCO Impact	I/O Performance Improvement	IOPS Gain	Reduction in power	Previous HDD Total	Total SSS	Total HDD Consumed	Total Usable SSS Capacity
$57	663%	73,984	81.8%	24	24	3504 GB	3523 GB

- SNIA CO Calculator: www.snia.org/forums/sssi/programs/TCOcalc/

Agenda

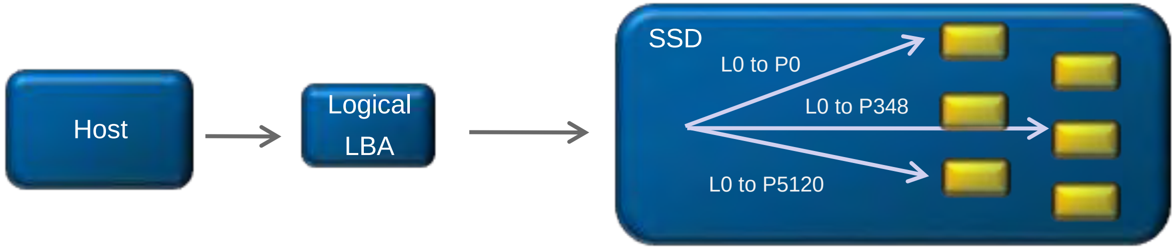
- Basics of NAND technology
- Basics of datacenter workloads
- MLC datacenter SSD
- Workload Examples

Nuances of SSD Performance/Endurance

1. Type of NAND
   - Single Level Cell (SLC)
   - Multi Level Cell (MLC)
   - Others (2.5BC, 3BC, etc)
2. Indirection system
   - Erasing and Writing Blocks
3. Host traffic pattern
   - Workload and Fullness of SSD
4. Spare area
   - SSD Workspace
5. Power off shelf life

Nuance of the "Indirection System"

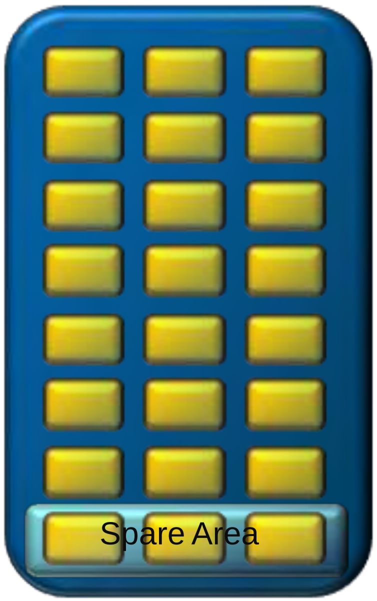
- Logical to physical LBA mapping removes need for atomic operations like read modify write (RMW)
  - The placement of new LBA information can be packed into pages that are at new physical locations
- Data placement in previously erased blocks makes foreground work (Host IO operations) faster
- Indirection "clean up" needs to reclaim invalid physical locations in background



SSD converts a Physical Page to Logical LBA. Logical LBA will not reside in the same physical location each time it is written

Host Traffic Pattern: Empty vs Full

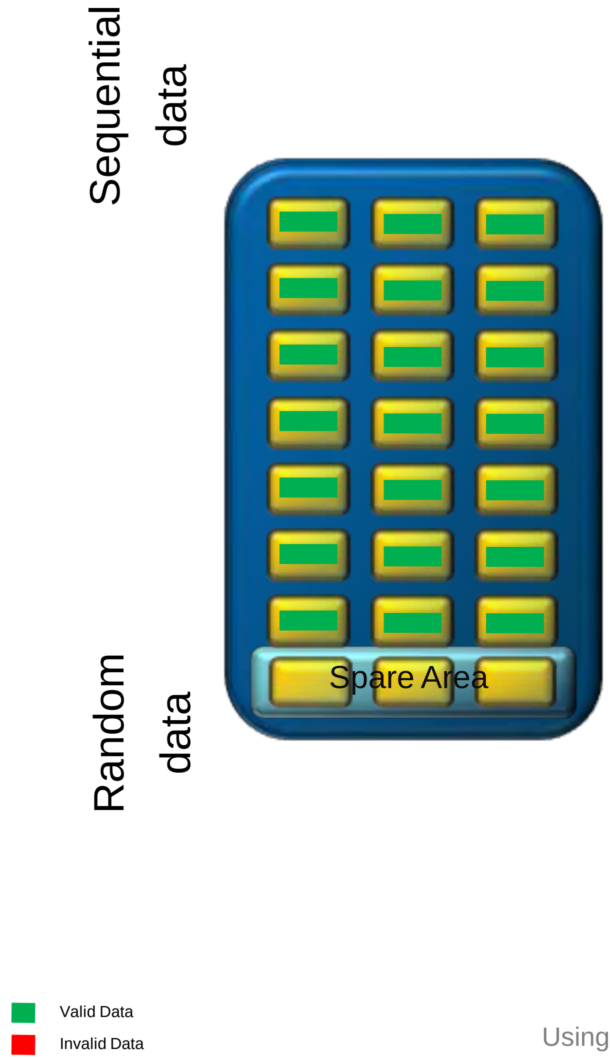
Sequential data



- An empty SSDs achieves its maximum write performance under all workloads
- Once initially filled performance will decrease
- Study State write performance is achieved when the SSD has settled into a consistent write latencies pattern
- A Steady State can be observed when
  - User capacity is full
  - Consistent work load is provided

SSDs steady state performance will have dependencies on the amount of spare area

Host Traffic Pattern: Sequential vs Random



- Steady state performance of an SSD full of sequential data is better than the steady state of an SSD full of random data
  - Sequential sectors will be invalidated in larger linear clusters than random.
  - Invalidation of sectors within a block is spotty in random writes.
- Changing the workload of an SSD from sequential to random will cause the performance to fall whereas changing from random to sequential will increase performance over time.

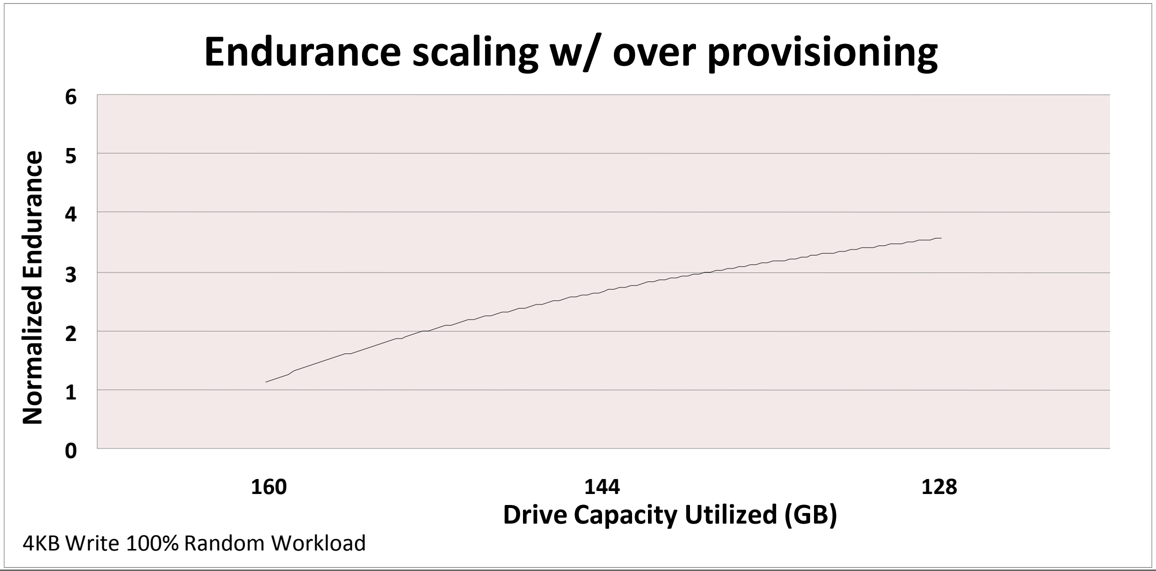
Spare Area: The Transitory Working Space

- Increasing the Spare Area helps performance by increasing the available “ready to be written” resource pool
- Larger work space allows
  - for less data movement to reclaim blocks.
  - Less erase cycles on the blocks as we do less background data movement
- SLC already maximizes spare area
- Increase MLC Performance by
  - Factory option - Set Max LBA to decrease user capacity and increase Spare Area
  - User option - Define a partition less than the max available capacity
  - Increasing spare capacity can boost performance by 10% or more . The main benefit it allows for more consistent performance.

Less background data movement increases performance

Spare Area Affects Endurance

- Increasing spare area increases endurance
  - Spare area beyond 27% of native capacity has diminishing returns
- Adjust SSD spare area by limiting drive capacity
  - ATA8-ACS Host Protected Area feature set is used (SET MAX ADDRESS)
  - Use ATA8-ACS SECURITY ERASE UNIT prior to limiting capacity
  - Setting partition to smaller size after erase is an option (less robust)



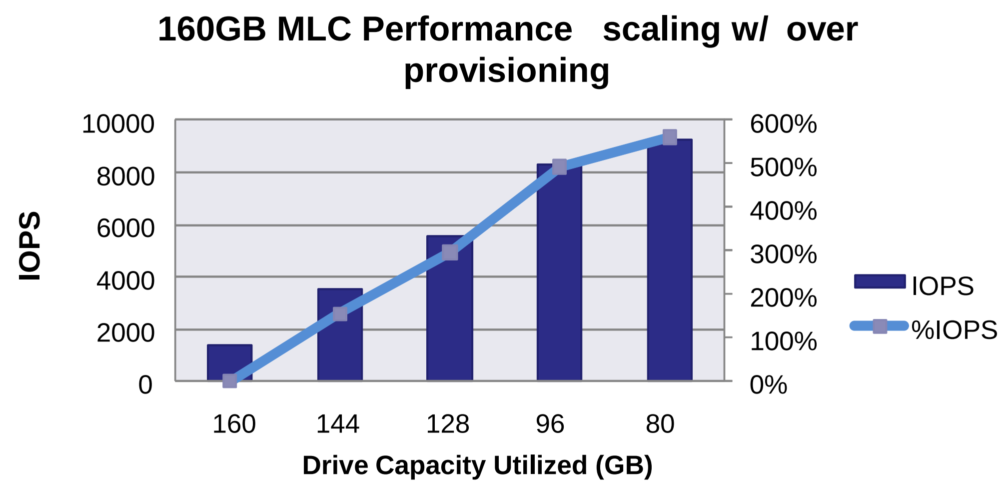
Data using Intel ®X25-M Mainstream SATA Solid-State Drive

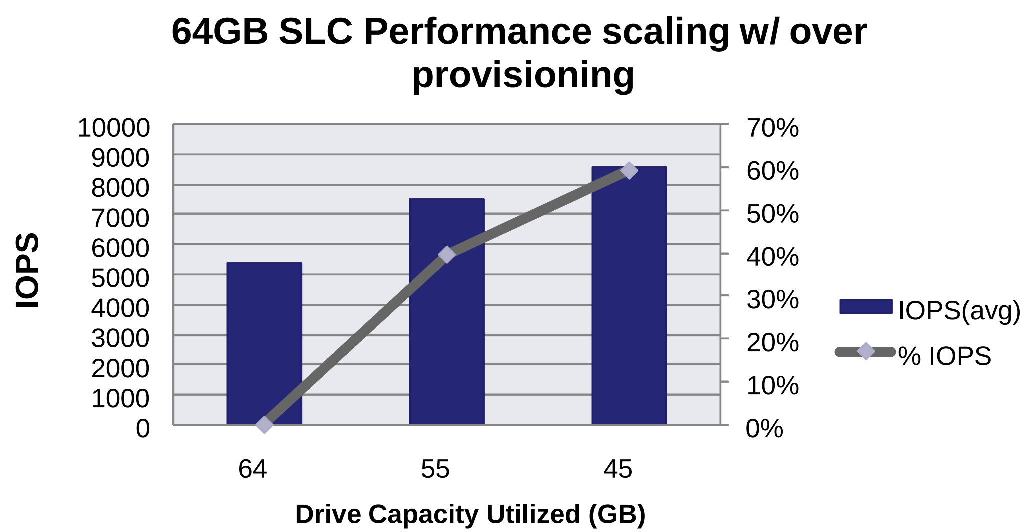
SSD Technology allows adjusting capacity to provide up to 3.5X improvement in endurance

Drive Performance vs Spare Area

Education SNIA

- As spare area increases so does performance
- MLC has a greater % performance increase due to the relative smaller spare area to start with





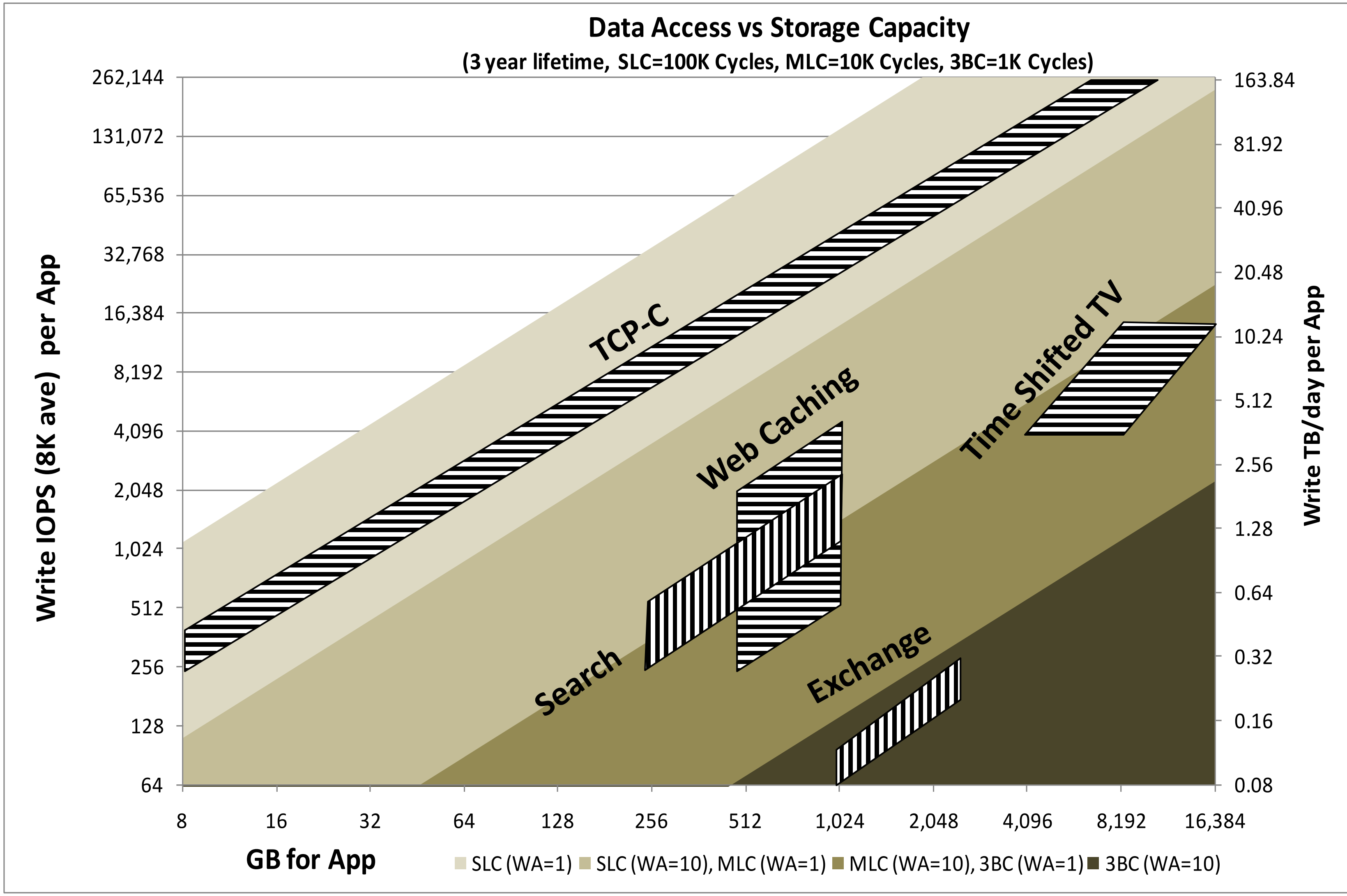
Data using Intel ®X25-M Mainstream SATA and Intel® X25-E Extreme SATA Solid-State Drives

IOPS scales with increase in spare area

Agenda

- Basics of NAND technology
- Basics of datacenter workloads
- MLC datacenter SSD
- Workload Examples

Mapping workloads to NAND

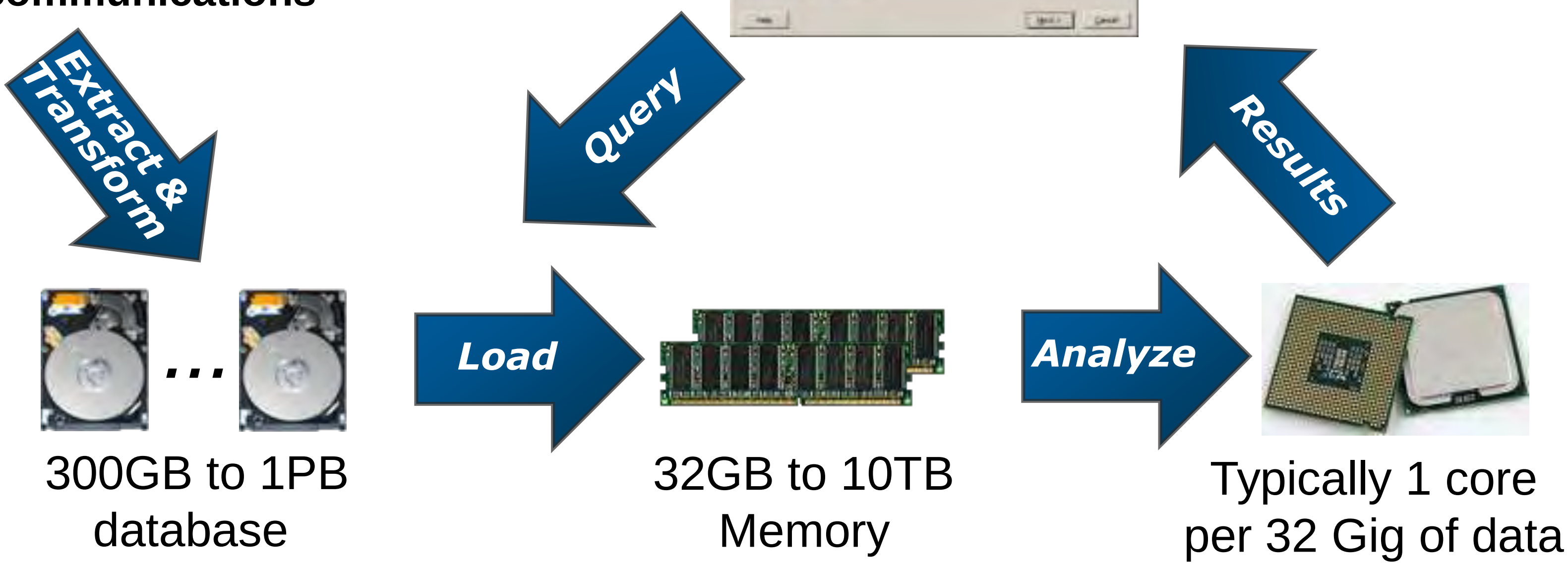


Business Intelligence: Example

Business data from industry

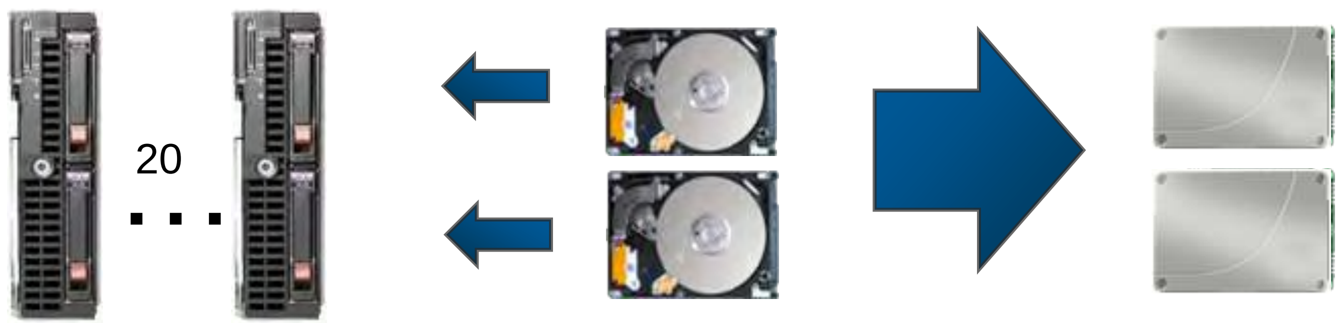
- **Digital Media**
- **Energy and Utilities**
- **Financial Services**
- **Health and Life Sciences**
- **Retail and CPG**
- **Telecommunications**

Speed of result is differentiation



Business Intelligence:

Financial services example



Quad-Core processor
48G DRAM

HDD Config:
2 - 146G
15K SAS HDDs

SDD Config:
2 – 160G
SATA SSDs
(config as 146G)

The Bottom Line Benefits

Time to complete
~6x
Decrease

SLC Solution Cost
~1.5x
more

MLC Solution Cost
~1.2x
more

Summary

- Best ROI achieve by focusing on solution
- MLC today for most application solutions meets
  - Endurance/Lifetime needs
  - Read/Write performance needs
- MLC cost today (and likely in future)
  - Significantly better than 2x less in $/G

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: tracksolidstate@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

- SNIA Education Committee

Tony Roug