



Flash Solid State Storage Reliability

Are we there yet?


Presenter: David Flynn



NAND Flash Reliability/Availability

- The GOOD:
 - No moving parts
 - Predictable wear out
- The BAD:
 - Bit error rate increases with wear
 - MLC wear rate (higher capacity) worse than SLC
 - Higher density NAND Flash increases bit error rate
 - Program and Read Disturbs
- The UGLY:
 - Partial Page Programming
 - Data retention is poor at high temperature
 - Early-life die failure is a factor (due to large number of parts...)

SNIA⁷





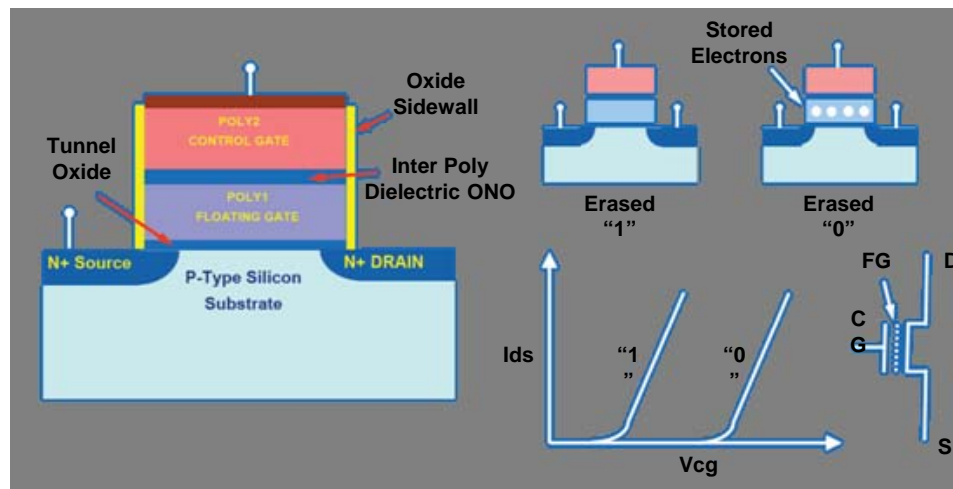
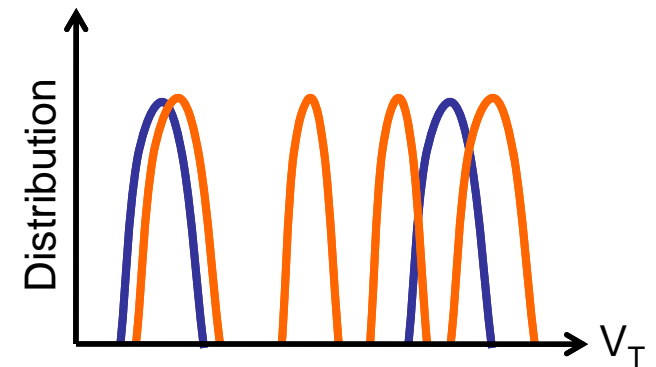
SNW
COMPUTERWORLD

April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida

Theory Behind NAND Flash

- What is persistent memory?
 - The ability to hold a state or store data without power
- What is NAND Flash?
 - Single cell uses a floating gate to modulate threshold voltage
 - Programming injects a charge (electrons) in to the floating gate
 - Sensing circuits detects the V_T to determine its state or value

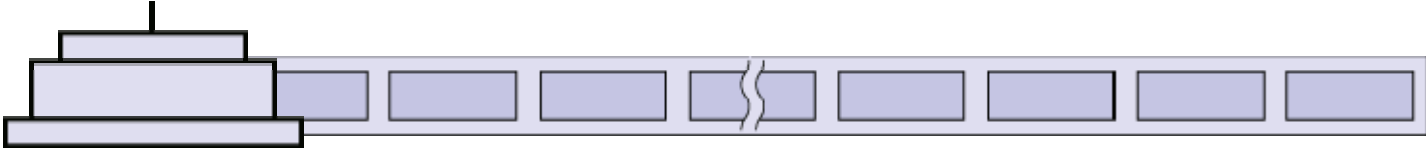
SLC = 1bit/cell = 2 levels 
 MLC = 2bit/cell = 4 levels 



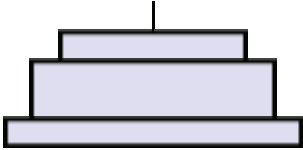
SNIA⁷
SNW
COMPUTERWORLD
April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida

The Hardware “Stack”

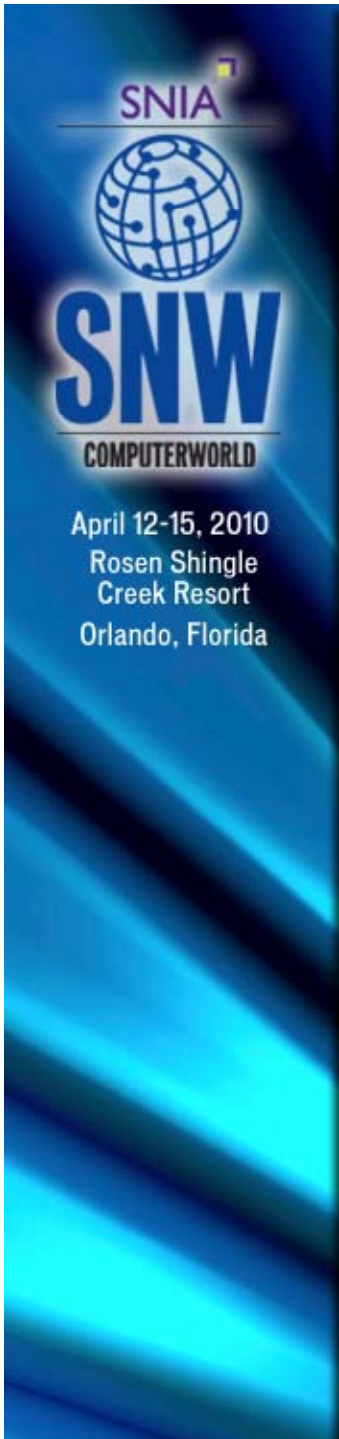
Cell to Plane



A Word Line or Page

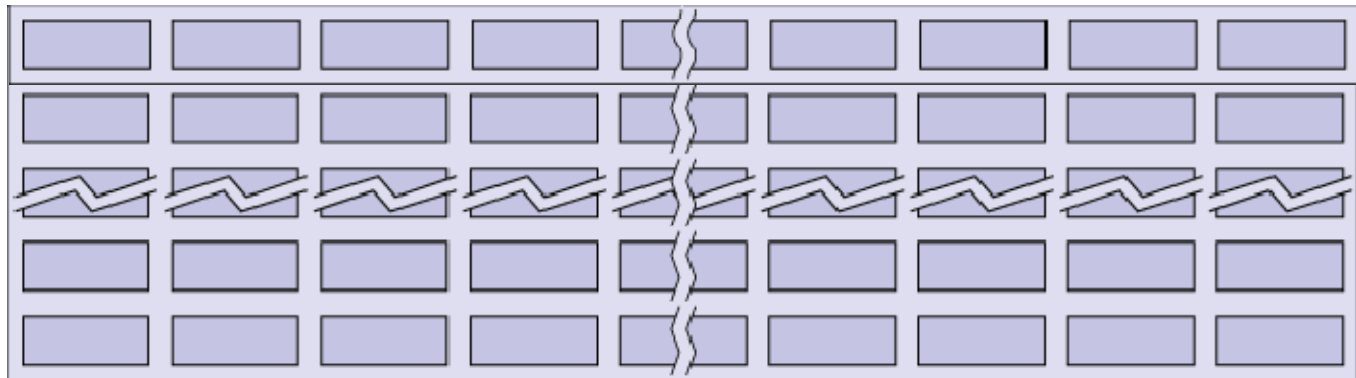


Cell




The Hardware “Stack”

Cell to Plane



An EB is a page wide (~4KB) x 2^N (~512) pages deep

SNIA⁷



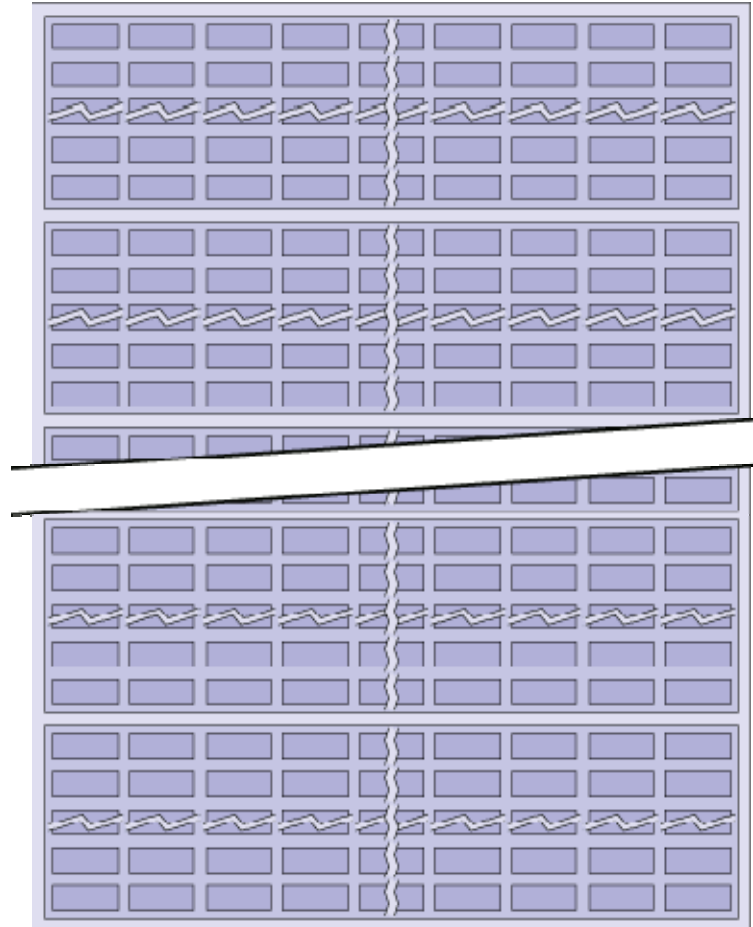
SNW

COMPUTERWORLD

April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida


The Hardware “Stack”

Cell to Plane



A Die Plane has ~4K EBs

SNIA⁷



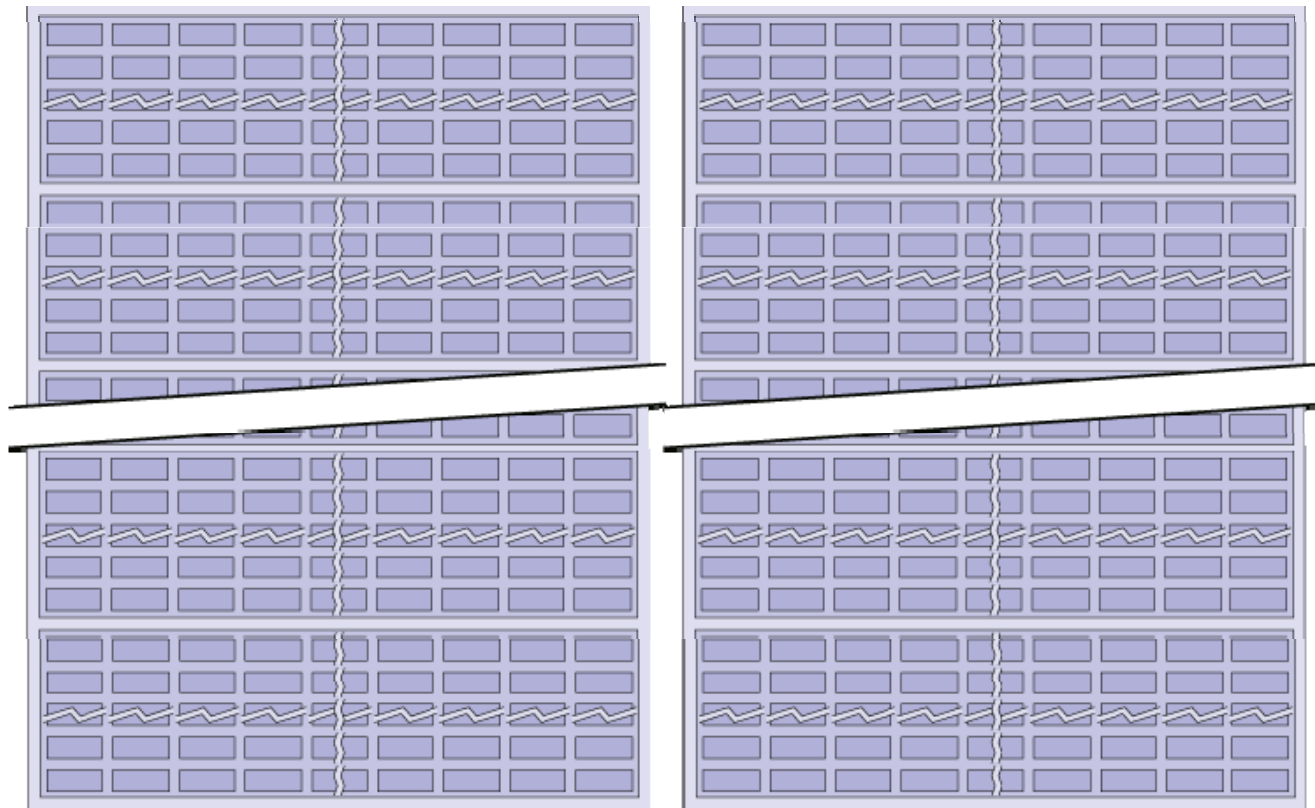
SNW

COMPUTERWORLD

April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida

The Hardware “Stack”

Die to Sub-System



Multiple Planes make up a Die

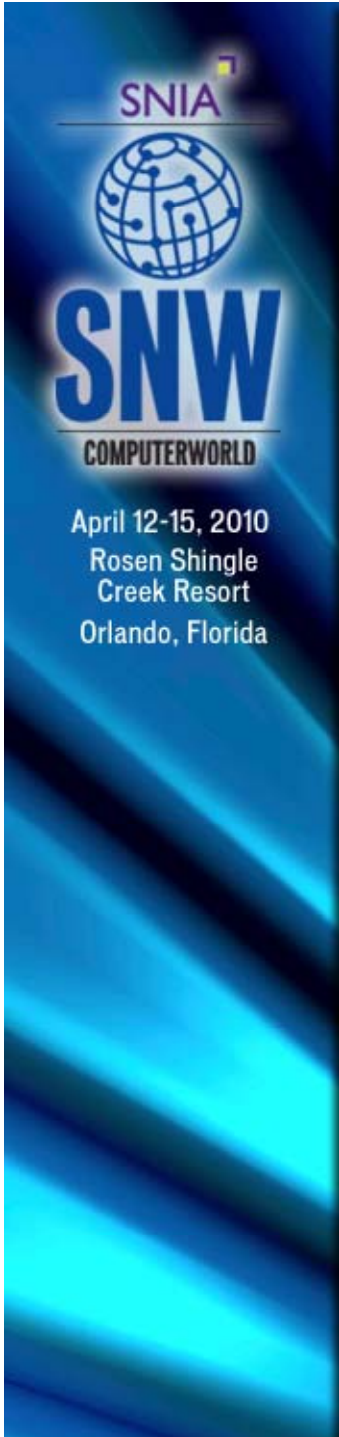


The Hardware “Stack”

Die to Sub-System

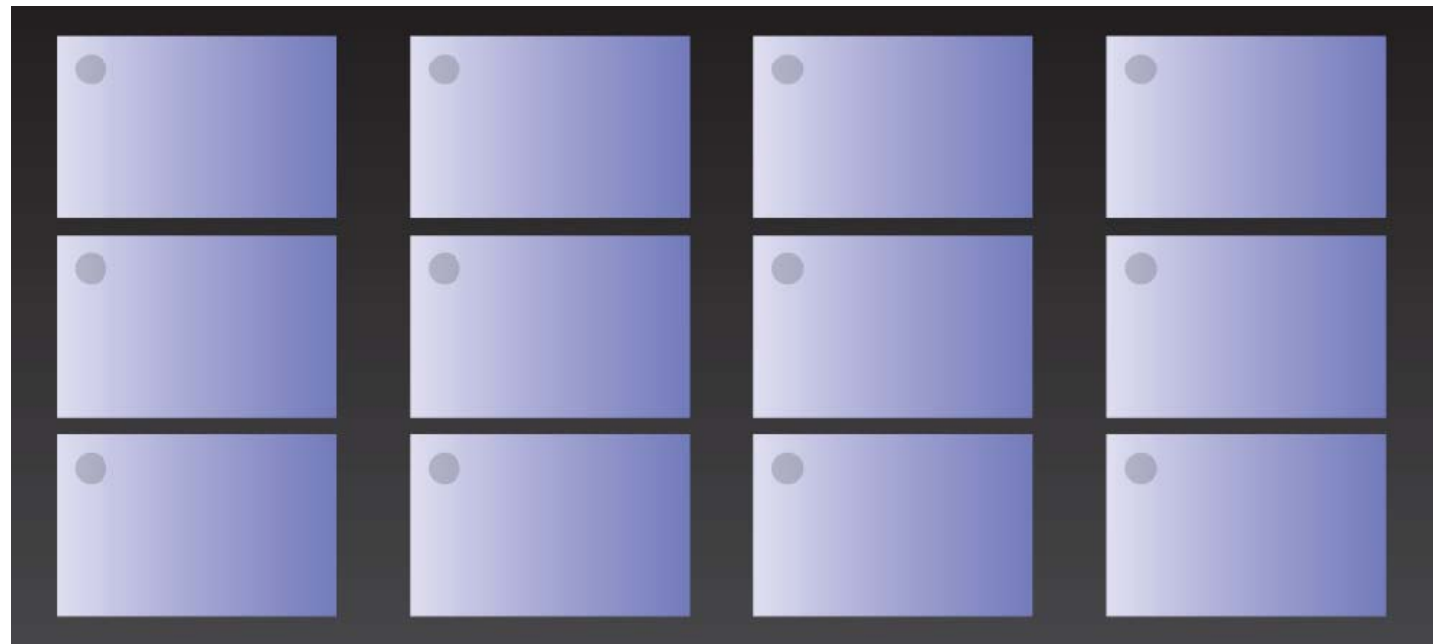


Multiple Die make up a Stack



The Hardware “Stack”

Die to Sub-System

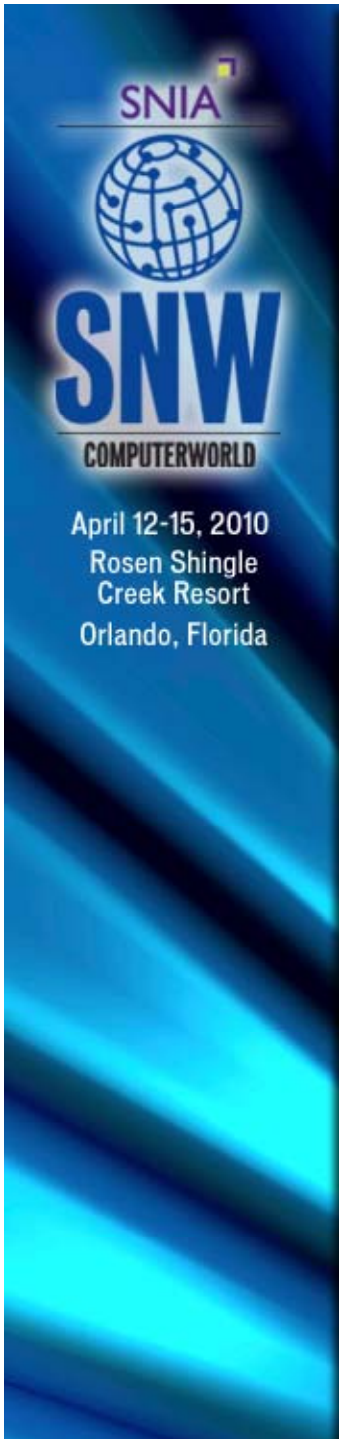


A Sub-system contain many Stacks



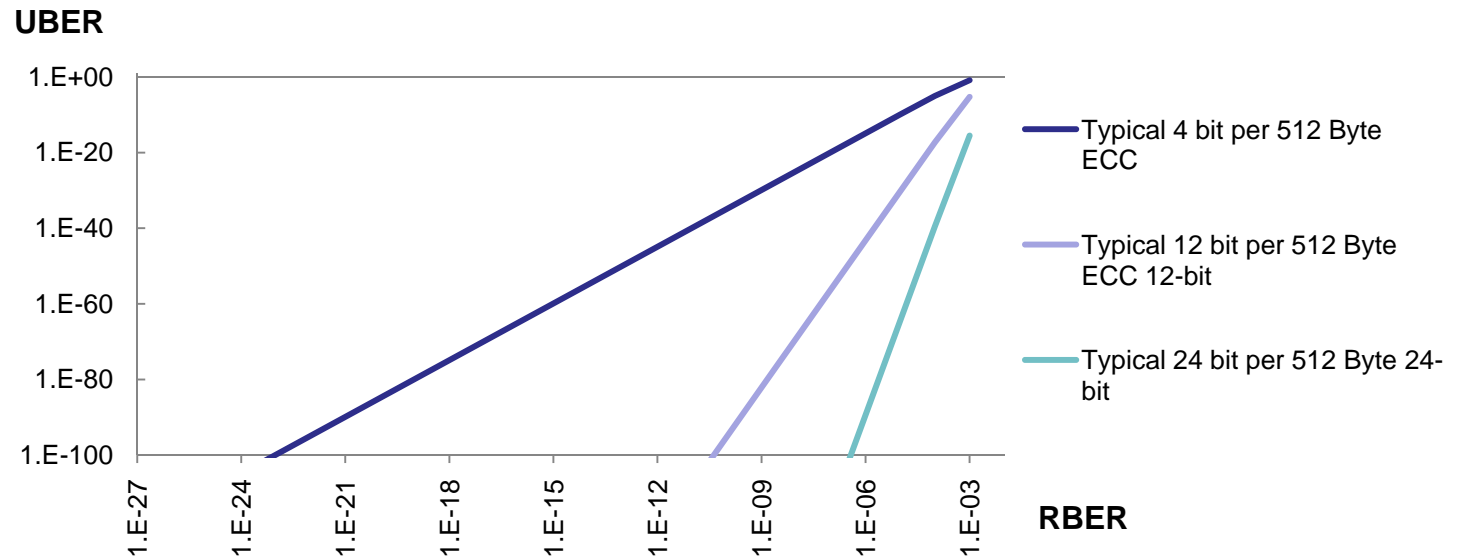
HDD Compared to SSD


	NAND Flash	Magnetic Media
MTBF	~ 2,000,000 Hrs	~1,000,000 Hrs
RBER	< 1e-10 (new)	~ 1e-7 (constant)
UBER	>= 1e-20	>= 1e-16
Bandwidth (in MB/s)	<= 1600 MB/s	<= 125 MB/s (SAS)
IOPS	<= 240,000	<= 300
Capacity	<= 640 GB	<= 1TB (SATA)
Target Life	5 years	5 years



The Need For Robust Error Correction Codes (ECC)

- NAND Flash is a lossy Media
- Lossiness increases with use
- RBER is increasing with feature size reduction







 COMPUTERWORLD

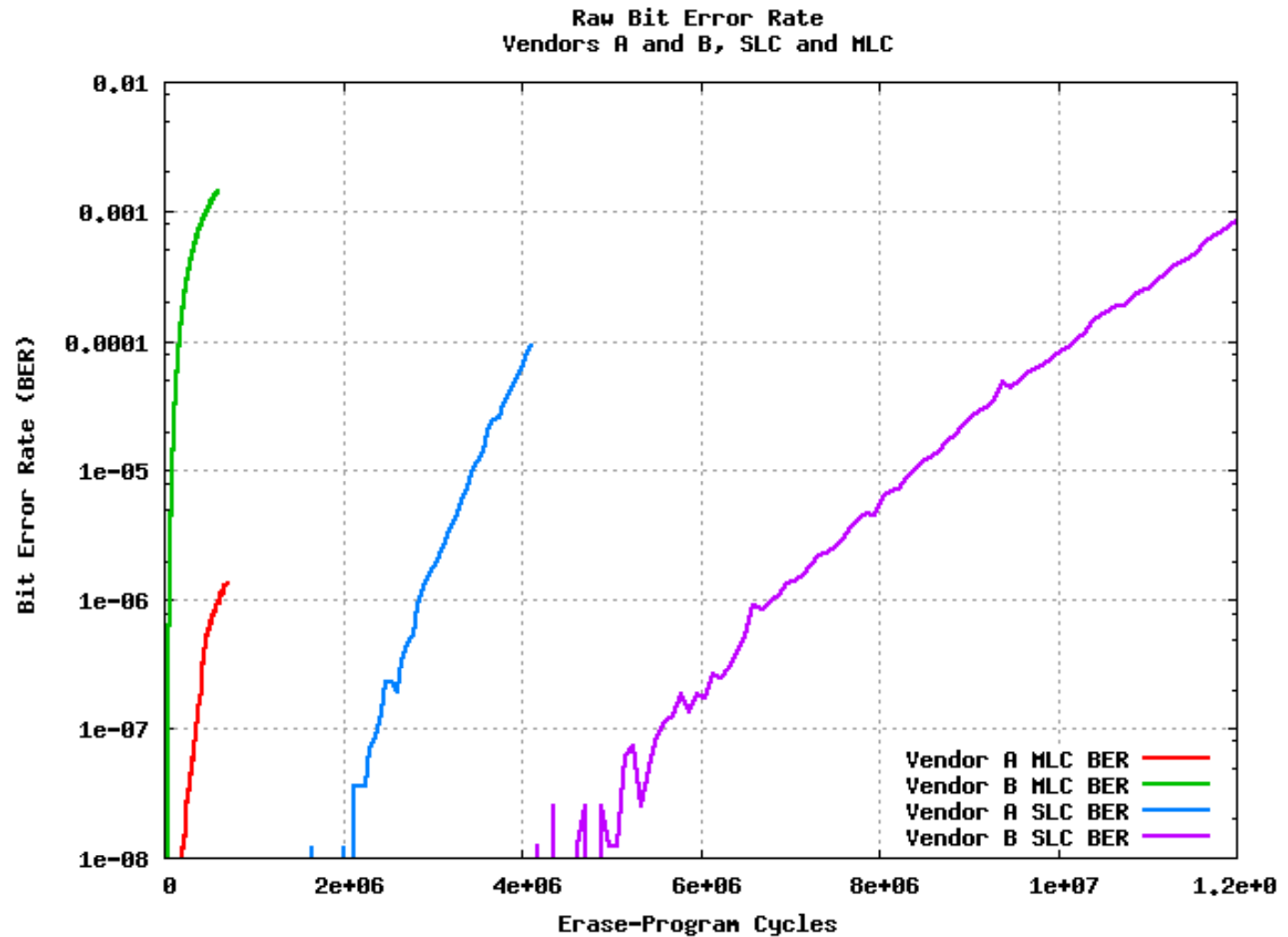
 April 12-15, 2010

 Rosen Shingle

 Creek Resort

 Orlando, Florida

Raw Bit Error Rate (SLC, MLC)



SNIA⁷

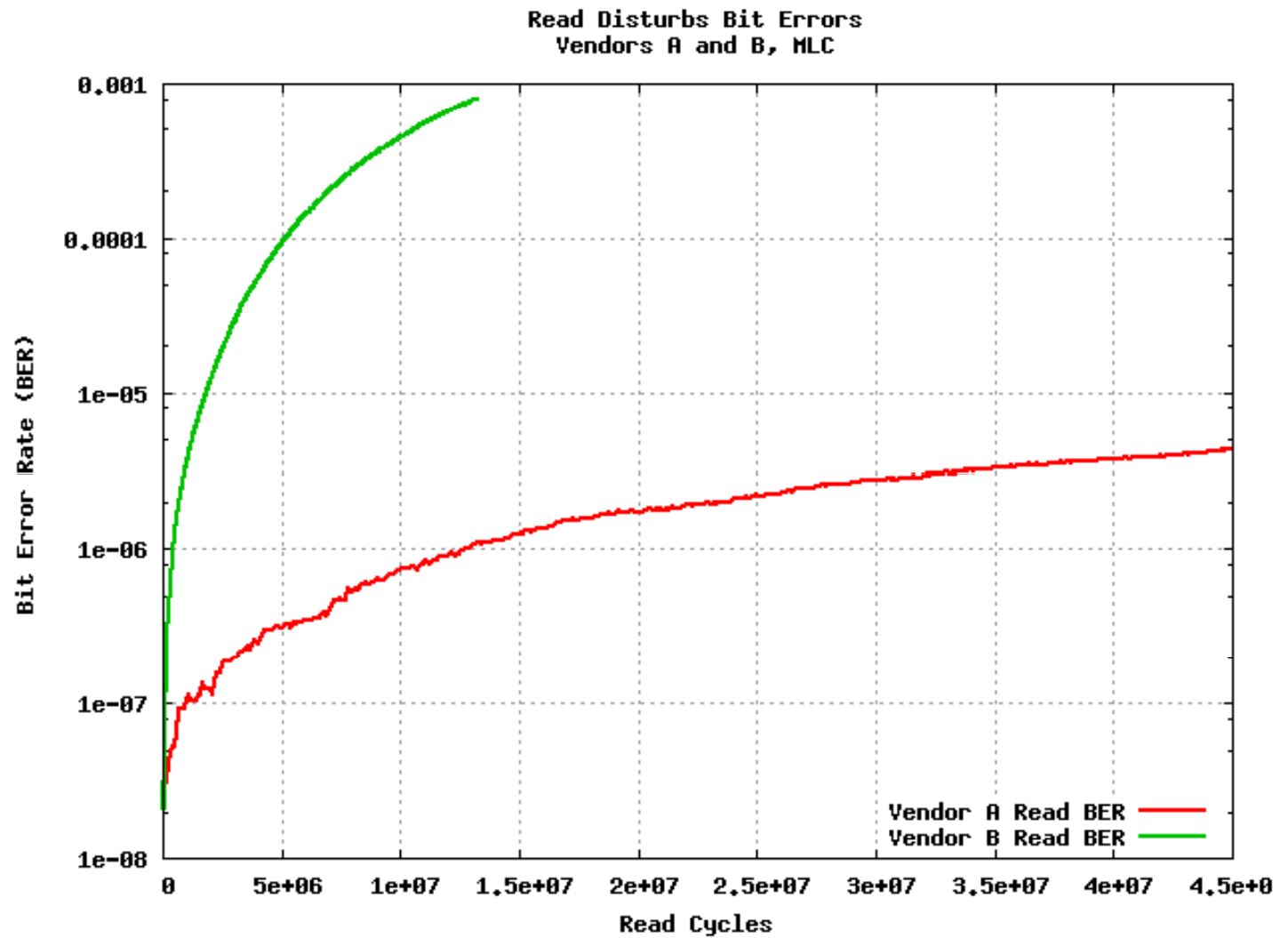


SNW


COMPUTERWORLD

April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida

Read Disturb Raw Error Rate (MLC)



SNIA⁷

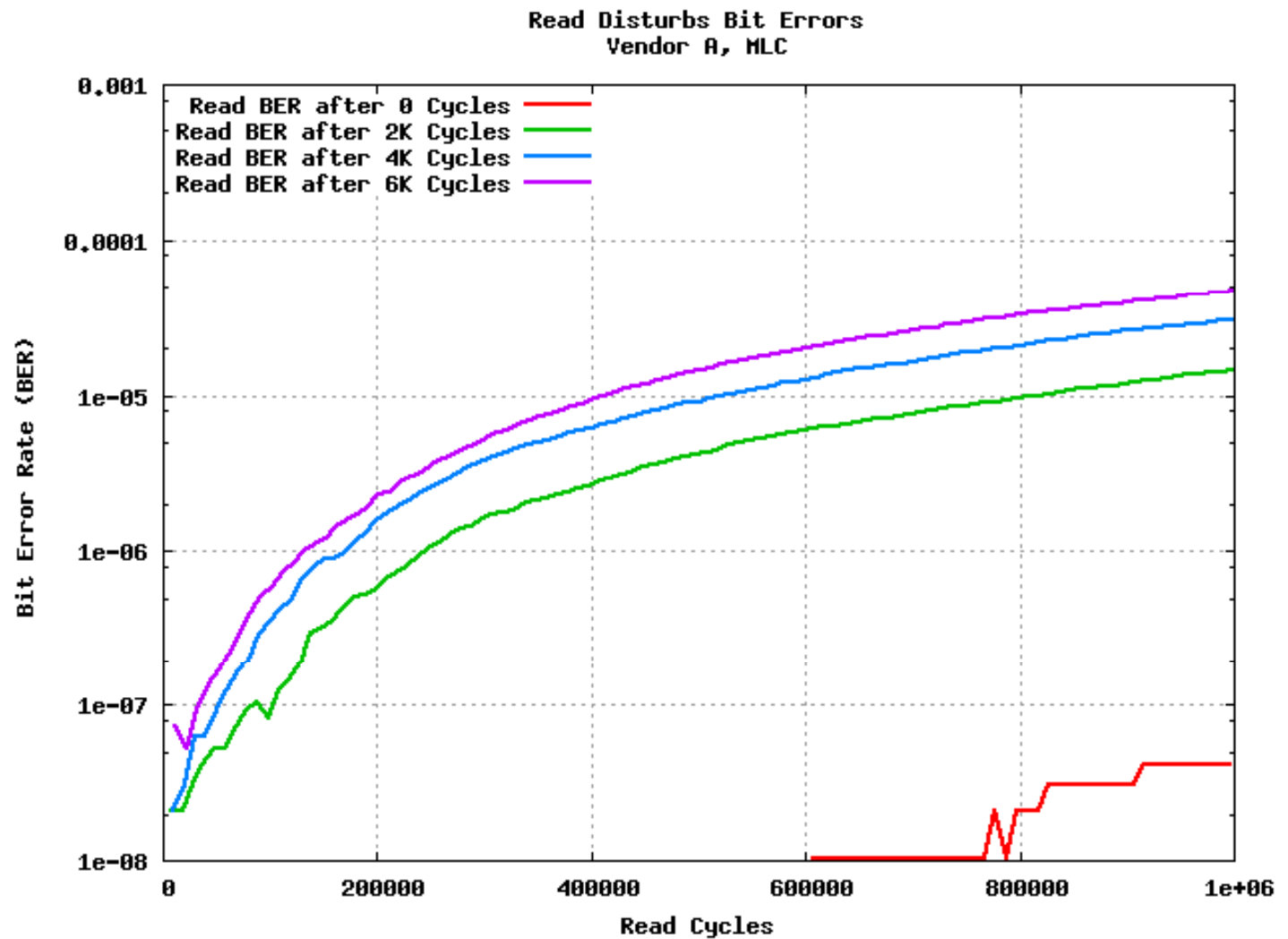


SNW


COMPUTERWORLD

April 12-15, 2010
 Rosen Shingle
 Creek Resort
 Orlando, Florida

Cycle Count Effect on Read Disturbs



SNIA⁷

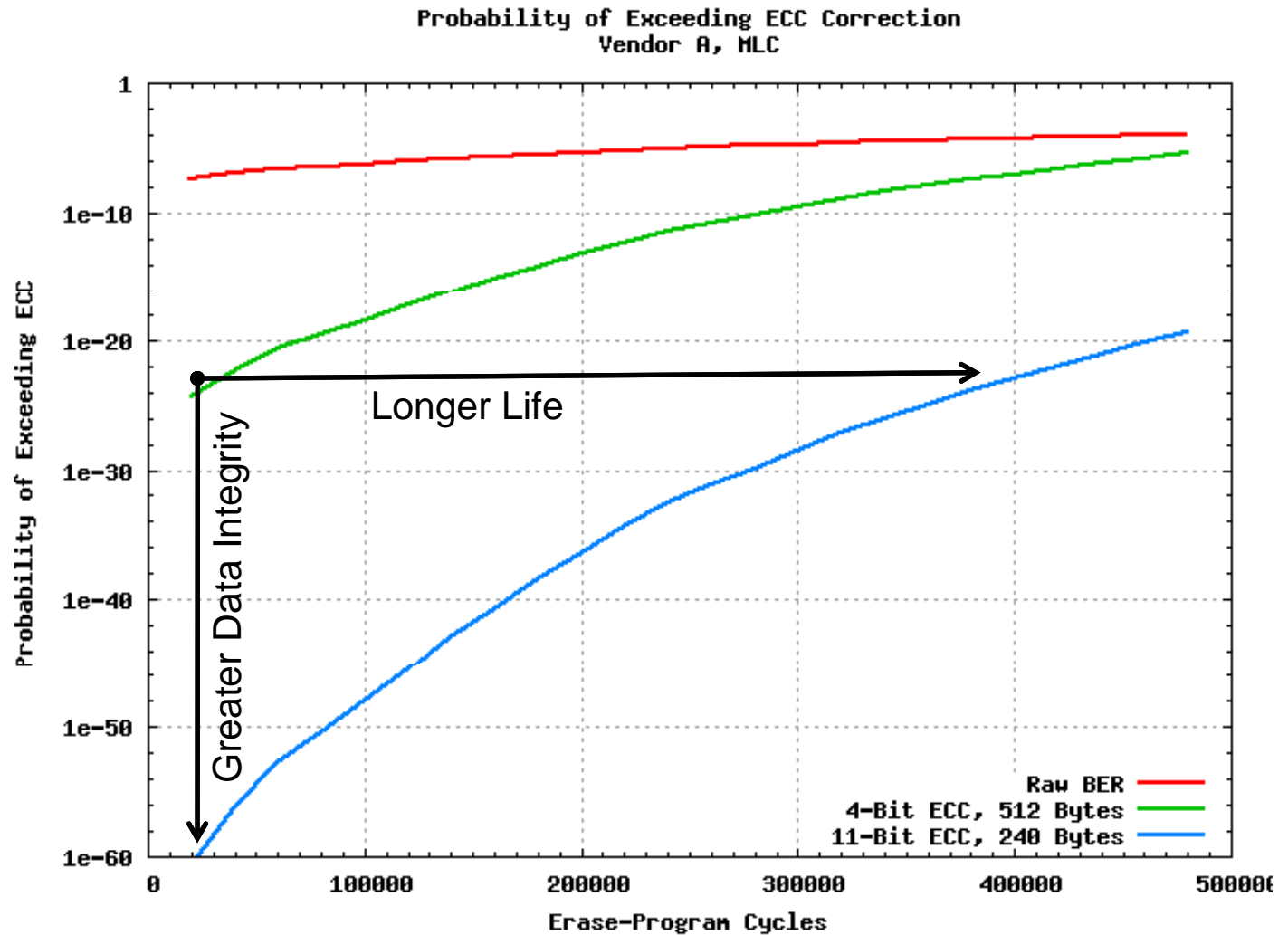


SNW

COMPUTERWORLD

April 12-15, 2010
 Rosen Shingle
 Creek Resort
 Orlando, Florida

Probability of Exceeding ECC (MLC)





SNIA⁷

SNW

COMPUTERWORLD

April 12-15, 2010
 Rosen Shingle
 Creek Resort
 Orlando, Florida

Expected Life Examples

	Life Expectancy	Average Rate	Used per Day	Consumed per Year	Life Expectancy
Consumer Vehicle	200K miles	45 MPH	1.5 Hrs	25K Miles	8.1 years
Commercial Vehicle	5,000K miles	55 MPH	11 Hrs	221K Miles	22.6 years
Consumer (MLC)	3.2 PBW	100 MB/s	7 Hrs	276 TB	11.6 years
Enterprise (SLC)	48 PBW	250 MB/s	24 Hrs	3,150 TB	15.2 years

Your Mileage May Vary



SNIA
COMPUTERWORLD

April 12-15, 2010
Rosen Shingle
Creek Resort
Orlando, Florida

SNW

Example of How to Calculate Expected Life

	"Enterprise"		"Consumer"
	300,000	Program-Erase Cycles	10,000
	160,000,000,000	Capacity (Bytes)	320,000,000,000
	48,000,000,000,000,000	Total Bytes Written	3,200,000,000,000,000
	24	Duty Cycle (Hrs / Day)	7
	40% Write	Read / Write Ratio	30% Write
	250,000,000	Write Bandwidth (B/s)	100,000,000
	8,640,000,000,000	Bytes Written per day	756,000,000,000
	3,150,000,000,000,000	Bytes Written per yr (B/yr)	276,000,000,000,000
	15.2	Expected Life in years	11.6



Enterprise Grade NAND Must Protect Data Integrity

- SECDED protection in host memory
- Robust ECC
- Minimize [Hardware | Software] components in the data path
- Write amplification avoidance
- Smart wear leveling
- Metadata and data protection
- Transport error protection
- Pro-active suspect device retirement



Enterprise Grade NAND Must Be Reliable and Available

- Support redundancy, fault-tolerance, over-provisioning, fail-in-place techniques
- Provide fast recovery and accurate diagnostics (improved serviceability)
- Minimize component count
 - Including Single-Points-of-Failure (SPoF)



Enterprise NAND Flash Must Meet Enterprise Life Expectations

- Use the right NAND Flash for application
- Robust ECC
- Apply redundancy & fault-tolerance optimized for NAND Flash devices
- Smart system management techniques
- Error monitoring & management
- Graceful retirement



Additional Data Protection Mechanisms

- Power loss protection
- Data scrubbing
- Read / program disturb management
- Data retention management
- Thermal management
- Partial page programming avoidance

Lossy Media + Great Controller → Great SSS Solution



To RAID or Not to RAID?

- Useful at the system level
 - Handling of redundant swappable devices
 - Software-layer integration akin to remote replication and other HA practices
- Not optimal at device level
 - RAID-0 = No redundancy
 - RAID-1 = $\frac{1}{2}$ usable capacity
 - RAID-5 = Fraction of performance
 - RAID failures require manual service, device must be physically serviced or lose capacity
 - Supercaps and batteries required for in-flight data flushes in the event of power loss – adds failure points that require maintenance
 - Cannot recognize corrupt data



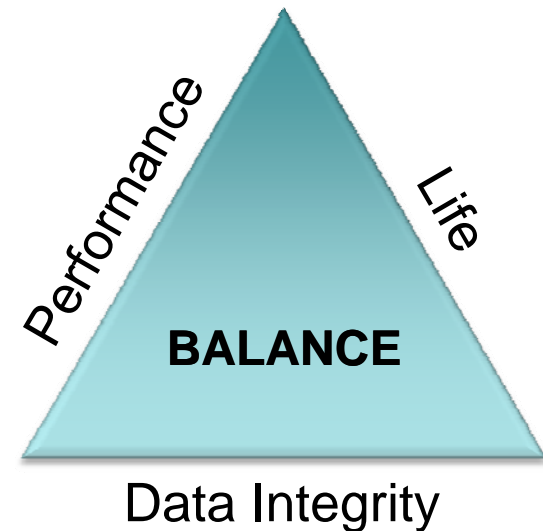
Alternatives to RAID

- Chip-level Redundancy
 - Parity-based (N+1)
 - Multi-tier (fine & coarse grained) substitution
 - No capacity loss
- Wire-speed performance
 - No Supercaps or Batteries Required
- Natively implemented in NAND Flash Controller
 - NAND Flash Specific Redundancy & Fault-handling
 - Self-healing – no user intervention required
 - Integrated Data Integrity Checking
- Optimized (fast & light-weight) for NAND Flash performance and failure semantics



Sub-Optimal Approaches

- Bandwidth throttling
 - May improve device life
 - May not translate to the system/data center or more global reliability gains.
- Partial page programming
 - May decrease write amplification
 - Severely reduces data integrity
- SSS tuned for performance at a block size you don't use
 - Write amplification at block size you do use
 - Increased wear
 - Decreased data integrity
 - Reduced life





The Net ...

- NAND Flash, like any other media, is not perfectly reliable
- Making imperfect media reliable at enterprise levels is the stuff of systems and reliability engineering.
- We can apply standard and proven methods to solve these challenges
- At the device level: these methods can make NAND Flash extremely reliable
- At the system level: NAND Flash performance enables re-architecture that greatly reduces system complexity



Questions?