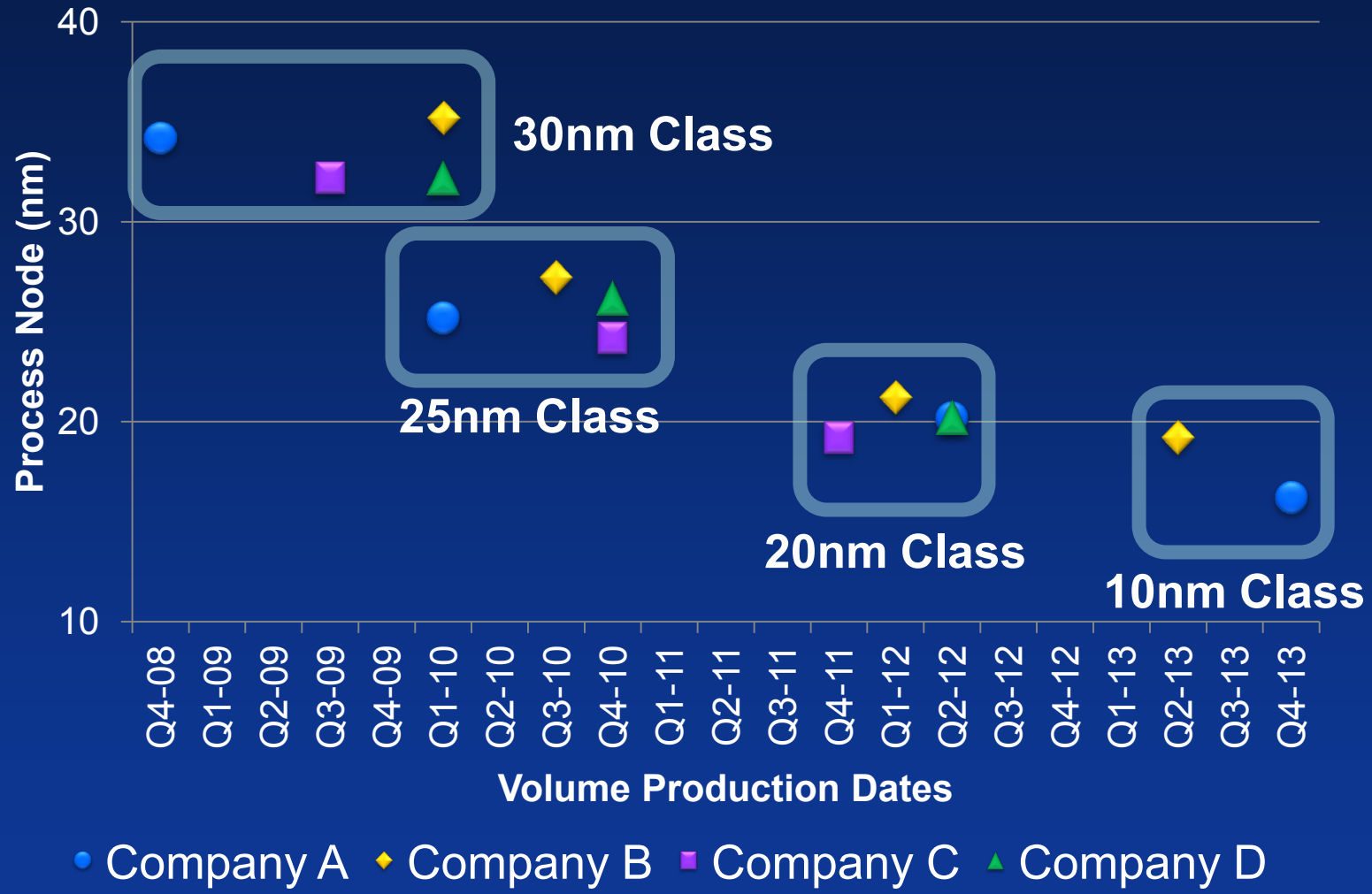
The Impact of NAND Lithography Trends on System Design

Michael Abraham (mabraham@micron.com)
NAND and Emerging Memory Architect
Micron Technology, Inc.

Abstract

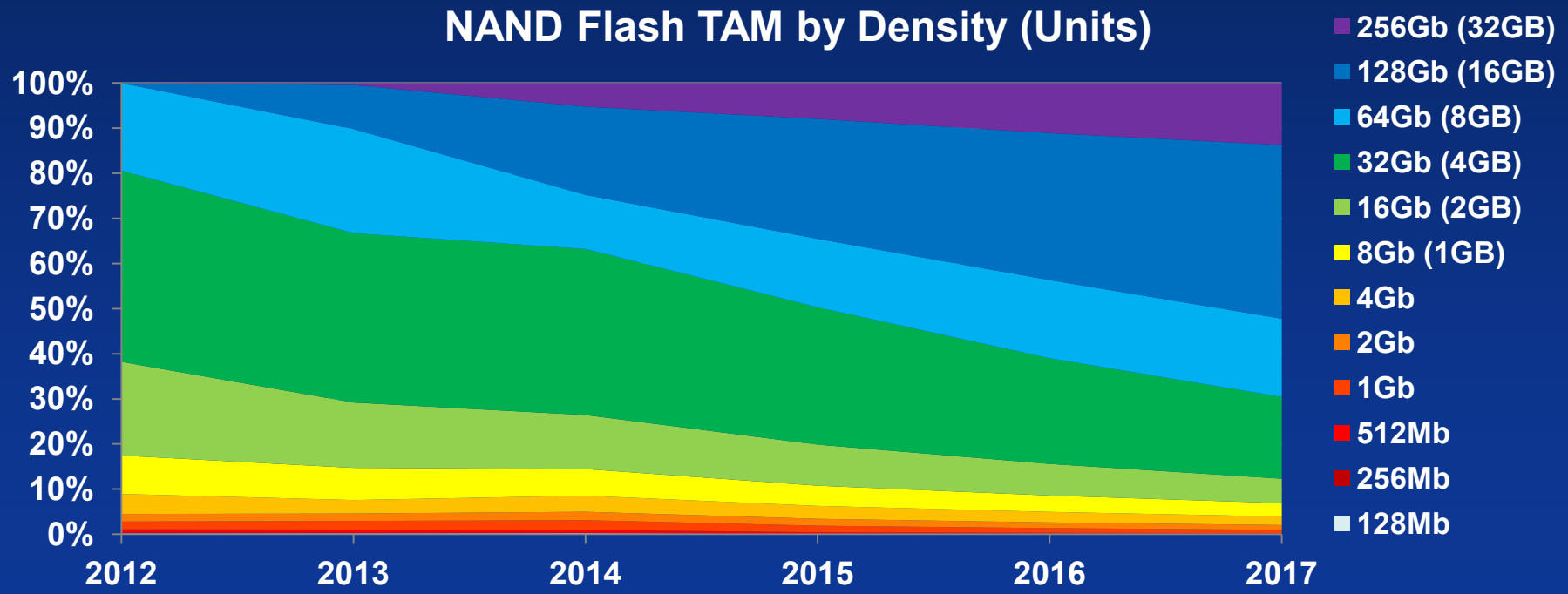
- NAND Flash is quickly moving to sub-20nm and 3D lithographies, making it the fastest scaling semiconductor technology ever
- This talk covers the impact that these shrinks have to NAND's architecture, performance, and reliability
- Prepare for these changes and to learn to counteract some of them through improved system design

NAND Process Migration: A Lithography Race to the Bottom



NAND Die Densities Continue to Grow

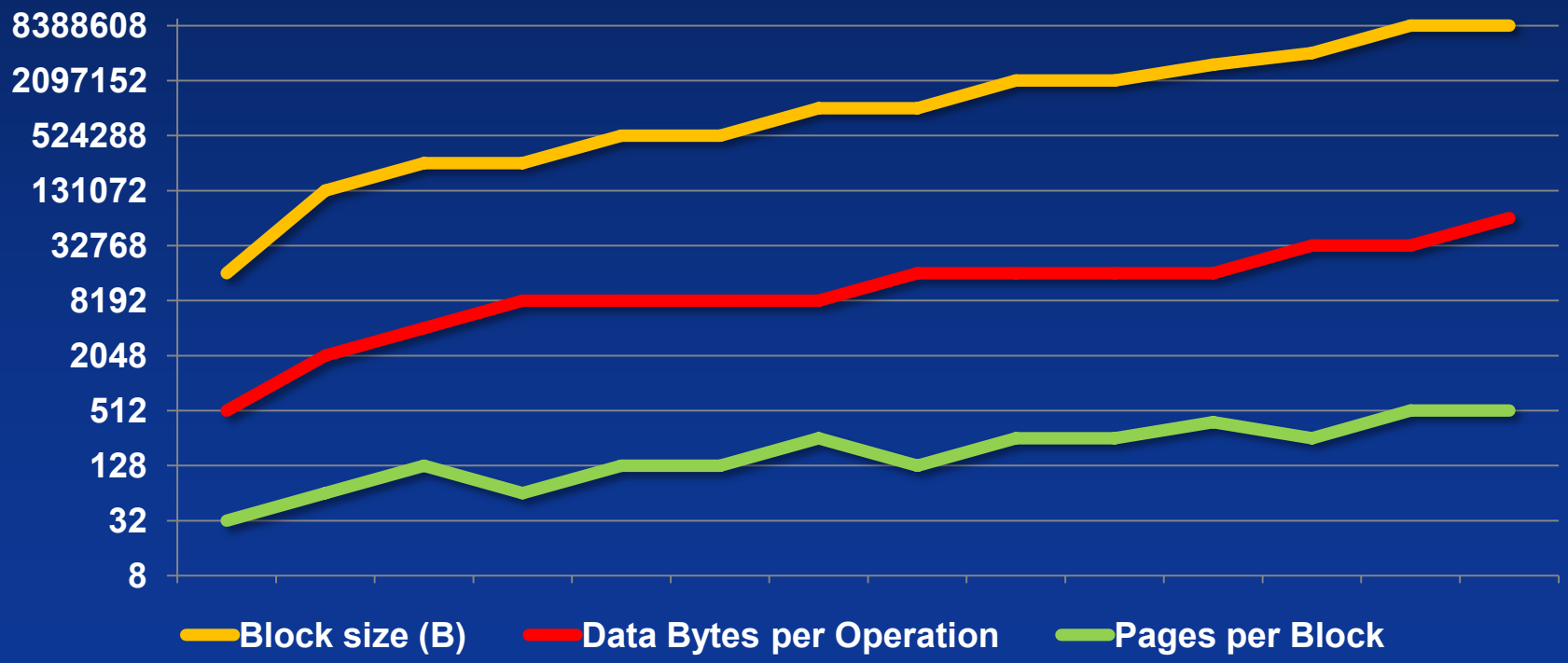
- As lithographies shrink, there is significant pressure to achieve lower costs per bit while maintaining or improving system performance
- For a given process node, the lowest cost per bit comes from increasing the total number of bits per die (e.g. the die density)



Source: iSuppli 1Q13

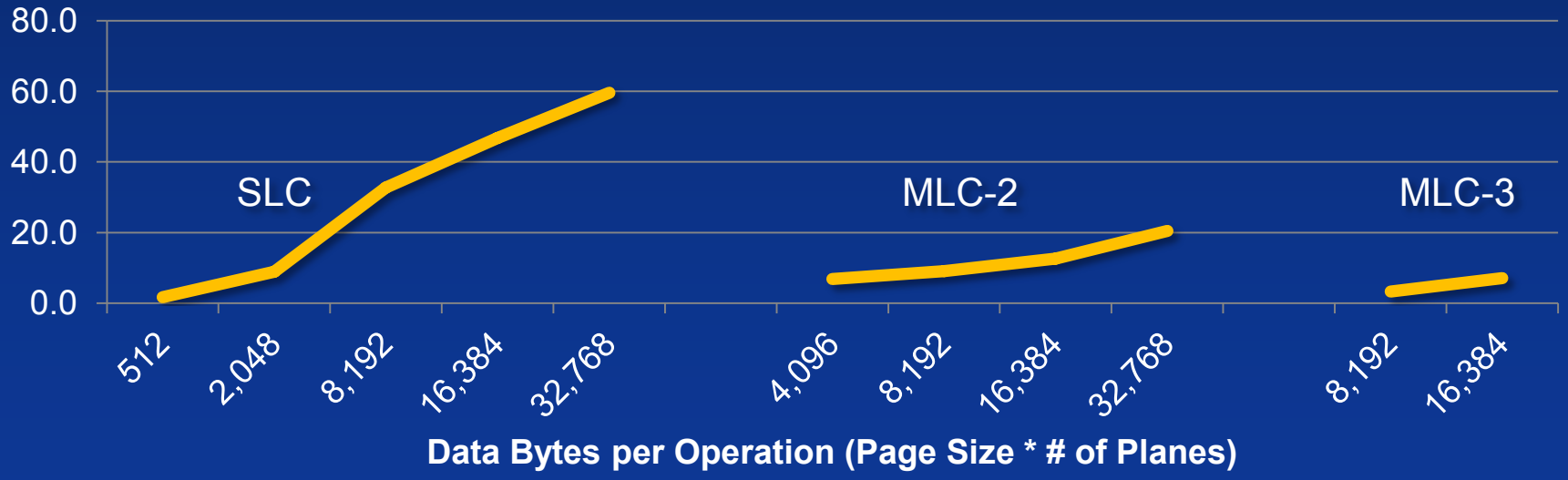
Memory Organization Trends

- NAND block sizes continue to increase
  - Though the number of erasable blocks per die remains stable at ~2,048 blocks per die, block sizes are increasing
  - Larger page sizes and more planes increase sequential throughput
  - For a given litho node, increasing the number of pages per block reduces the die size and therefore the cost per bit



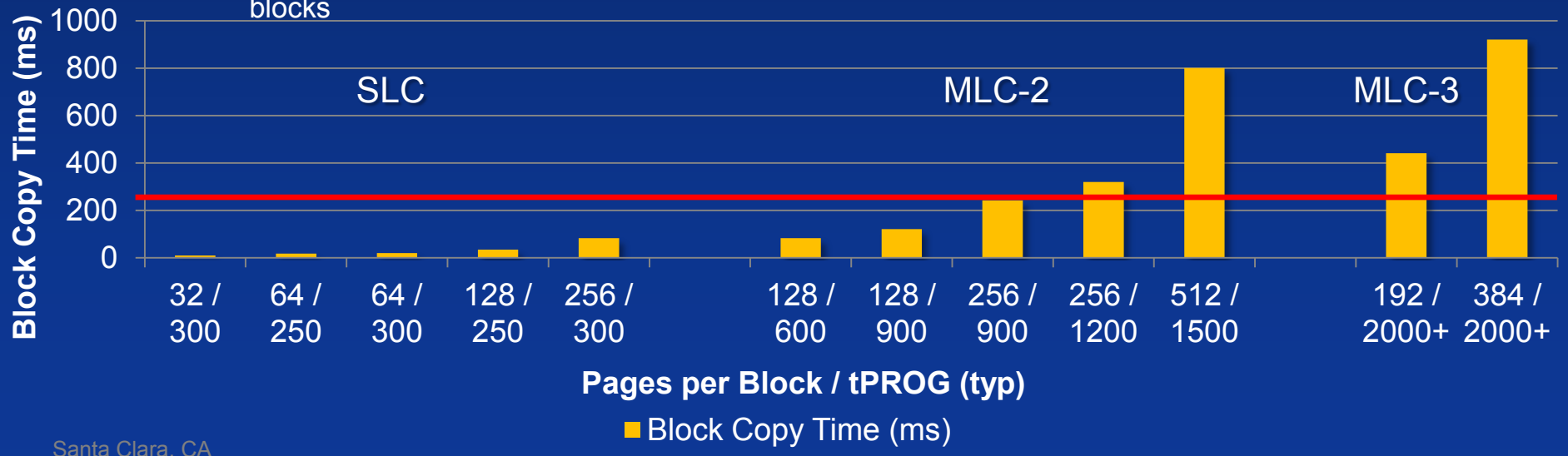
Larger Page Sizes Improve Sequential Write Performance

- For a fixed page size across process nodes, write throughput decreases as the NAND process shrinks
- NAND vendors increase the page size to compensate for slowing array performance
- Write throughput decreases with more bits per cell



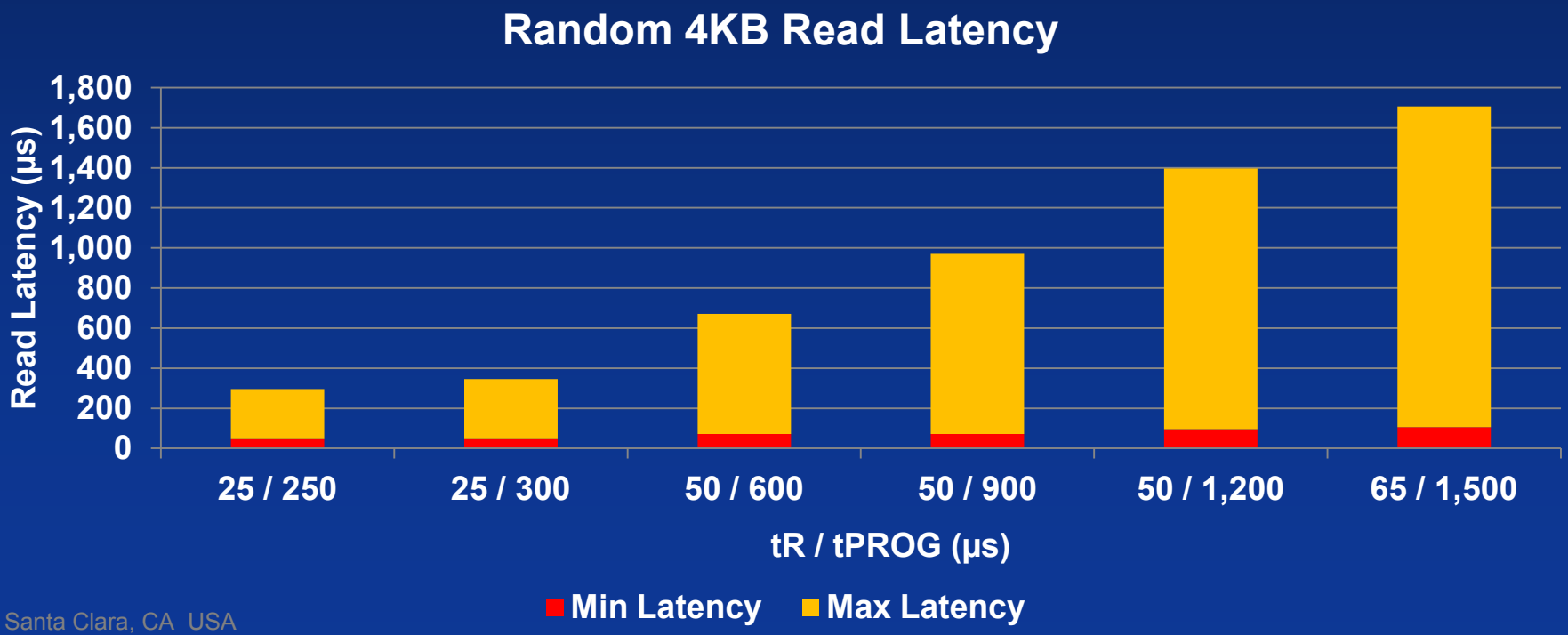
Larger Block Sizes Can Impact Random Write Performance

- Controllers translate host logical block addresses (LBAs) to NAND Flash's physical address space
  - Systems that use large DRAM buffers, like SSDs, are able to do this efficiently
  - Systems that use smaller SRAM buffers, like eMMC and SD cards, are not as efficient
- For systems with a small logical-to-physical (L2P) translation buffer, as the NAND block size increases, random write performance decreases
  1. Number of pages per block (most significant factor)
  2. Increase of tPROG
  3. Increase in I/O transfer time due to larger pages (effect not shown below)
- In order to reuse a block, it must be erased; all valid data must be copied out of it first
  - Some card interfaces have write timeout specs at 250ms
  - To improve random performance, block management algorithms manage pages or partial blocks

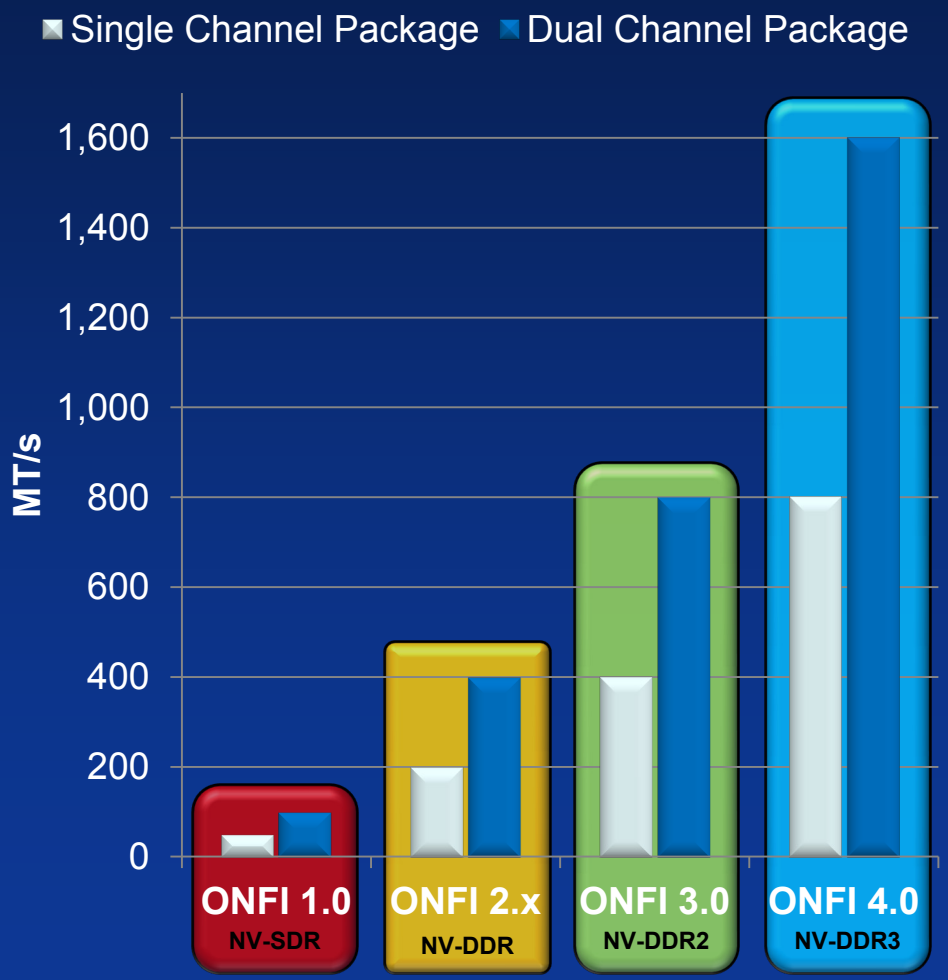


Slowing Array Operations Increase Random Read Latencies

- Most applications favor read operations over write operations
- Most read operations are 4KB data sectors
- As monolithic NAND density increases, less NAND die are being used for a fixed system density
- As tPROG increases, the latency of random 4KB sector reads becomes more variable in mixed-operation environments as the probability of needing to read from a die that is busy increases



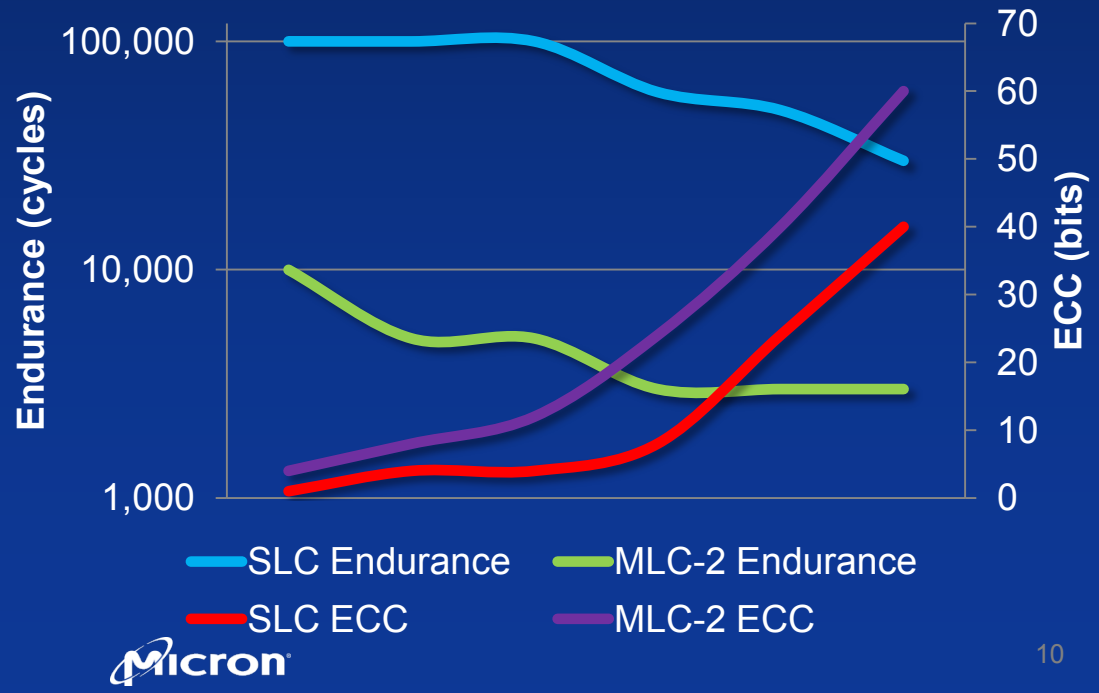
Faster Interface Speeds Reduce Latencies and Increase Read Throughput



- Read operations are still interface limited when two or more NAND die share the same I/O bus
- Almost all SSDs today use NAND interface speeds of 200MT/s or faster
- Some systems prefer multiple channels per package
  - 2-channel BGA
  - 4-channel BGA
- ONFI 4.0 is in definition
  - Up to 800MT/s throughput
  - Reduces energy per bit with 1.2V interface

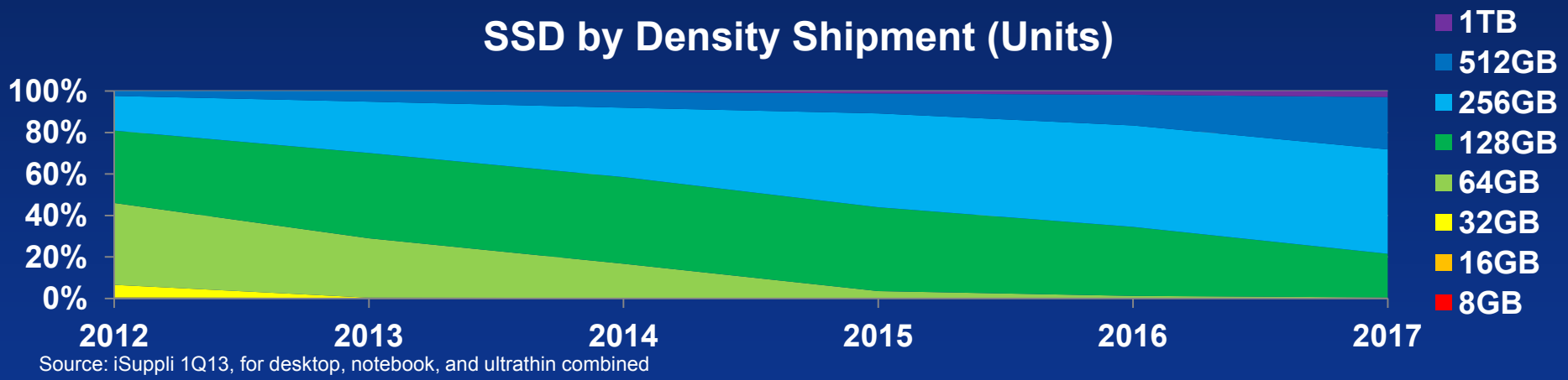
Fewer Electrons Per Cell Require More ECC to Maintain Data Retention and Endurance

- Process shrinks lead to less electrons per floating gate
- ECC improves data retention and endurance
- To adjust for increasing RBERs, ECC is increasing exponentially to achieve equivalent UBERs
- As ECC requirements increase, the spare area per NAND page continues to increase
- ECC algorithms are transitioning from BCH to LDPC and codeword sizes are increasing



NAND Flash Trends Can Be At Odds with SSD Requirements

- SSD densities aren't scaling as fast as NAND Flash
- The SSD interface is transitioning from SATA to PCIe providing higher interface bandwidth
- Market pressure to make each new generation of SSD faster than the one before
- Can be mitigated by using the second largest monolithic die density on a process node instead of the largest

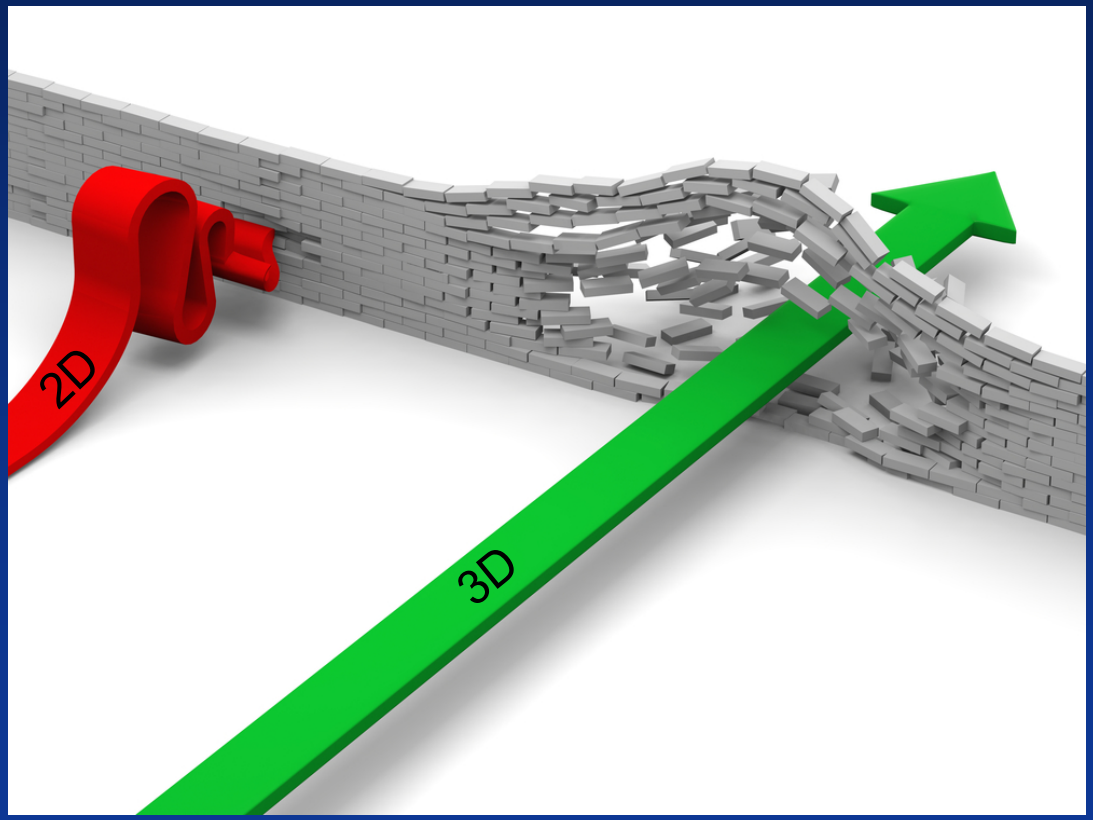


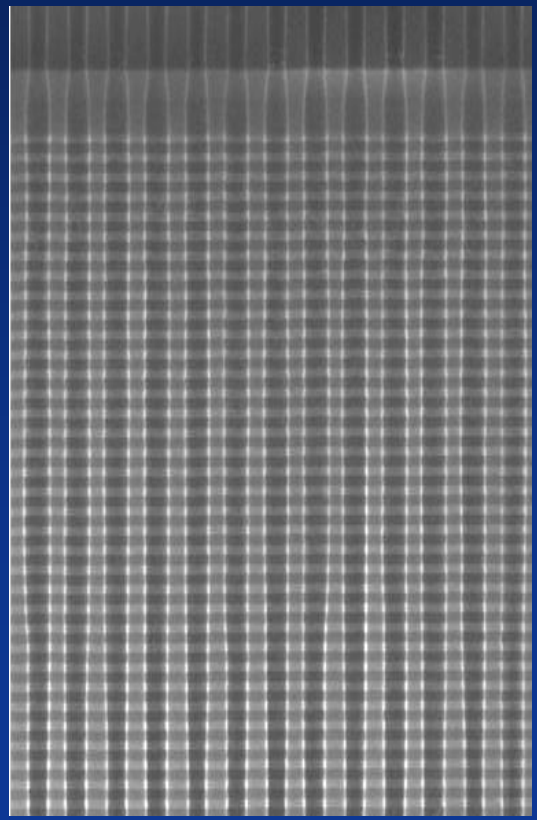
Source: iSuppli 1Q13, for desktop, notebook, and ultrathin combined

Node	NAND Density	Data size	# of Die	Seq Write (MB/s)
Gen 1	32Gb (4GB)	8192	32	230
Gen 2	64Gb (8GB)	16,384	16	184
Gen 3	128Gb (16GB)	32,768	8	133
Gen 3	***64Gb (8GB)***	***32,768***	***16***	***324***

Has NAND Scaling Hit the Wall?

- NAND will continue to scale using 3D structures, stacked like high-rise buildings
- Future architecture, performance, and reliability trends are similar to those of today





Questions?

About Michael Abraham

- Architect in the NAND Solutions Group at Micron
- Covers advanced NAND Flash and emerging memories
- IEEE Senior Member
- BS degree in Computer Engineering from Brigham Young University