Solid-State Drive 2

Juncheng Yang

Harvard John A. Paulson School of Engineering and Applied Sciences

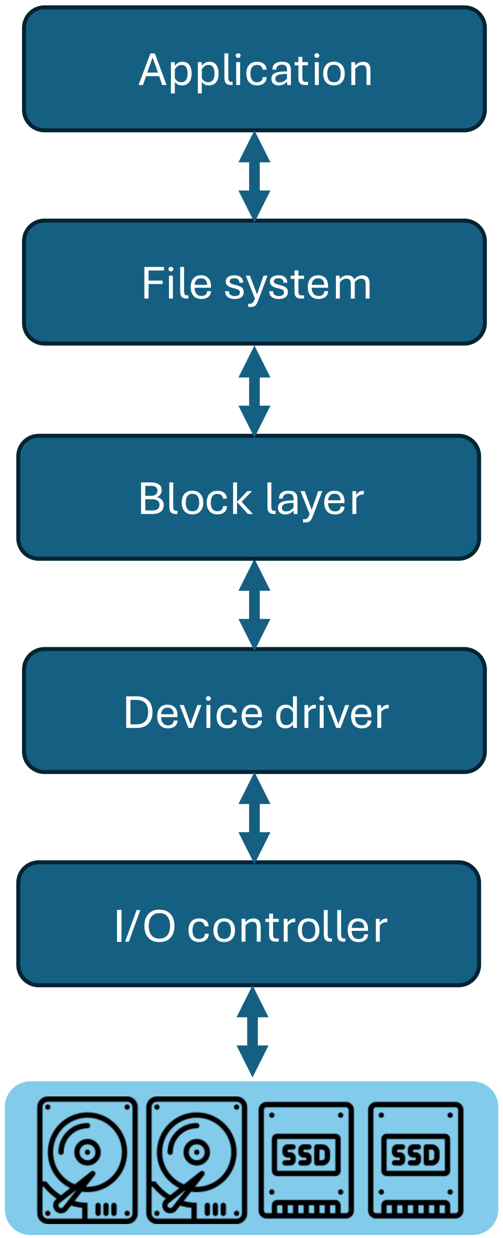
HARVARD UNIVERSITY MADSYS
Measurement and Design of Systems Lab

Recap: Solid State Drive

- Why SSD is better than HDDs
- SSD internals
  - how NAND cell works
  - why it wears out
  - different flash cells, e.g., SLC, MLC, TLC, QLC
  - flash translation layer (FTL) functions
    - mapping
    - garbage collection
    - wear leveling

Today

- SSD performance
- SSD reliability
- SSD density
- SSD cost and market
- Future trend



SSD performance

SSD internal parallelism

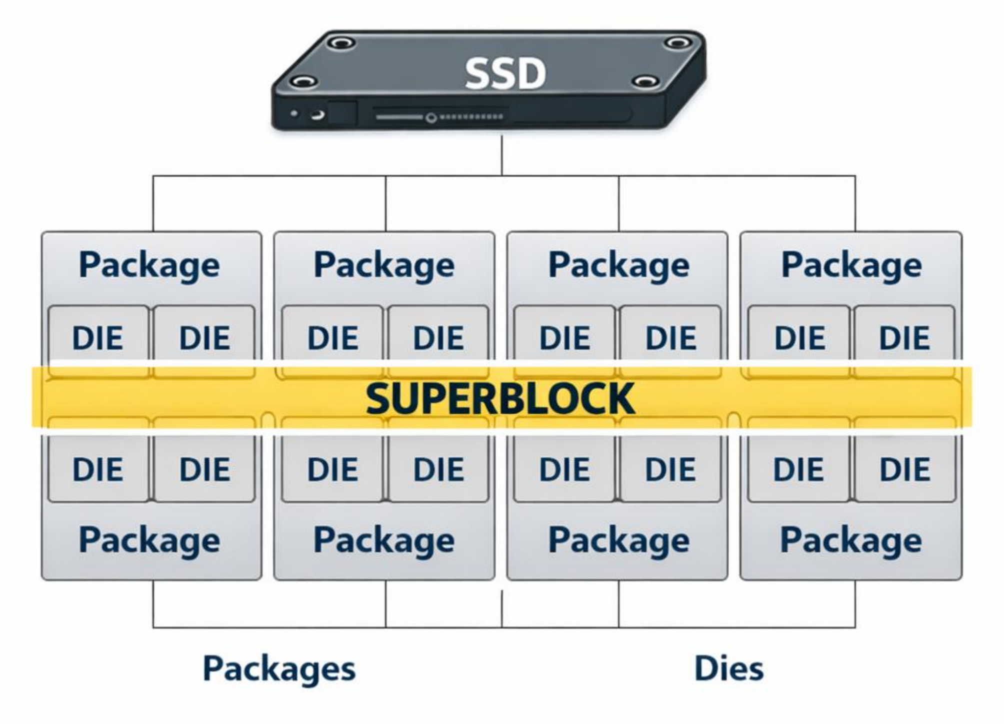
- Channel
  - a NAND interface on the SSD controller
  - electrically, a shared bus for command/address/data and control signals
  - way: the number of independently-selectable NAND targets per channel
    - e.g., an 8-ways channel connects to 8 packages
- A channel connects to multiple packages, and each package is connected to one channel
  - multiple channels avoid the channel becoming a bottleneck
  - read is serviced from the channel(s) where data reside
  - write can be placed across channels (striped)

SSD internal parallelism: hierarchical organization

- Channel and way parallelism
  - read / write to multiple packages / chips simultaneously
  - most important
- Die interleaving
  - pipeline operations within package
  - async die operations: channel bus is faster than die operations
- Plane parallelism: limited

SSD internal parallelism: superblock

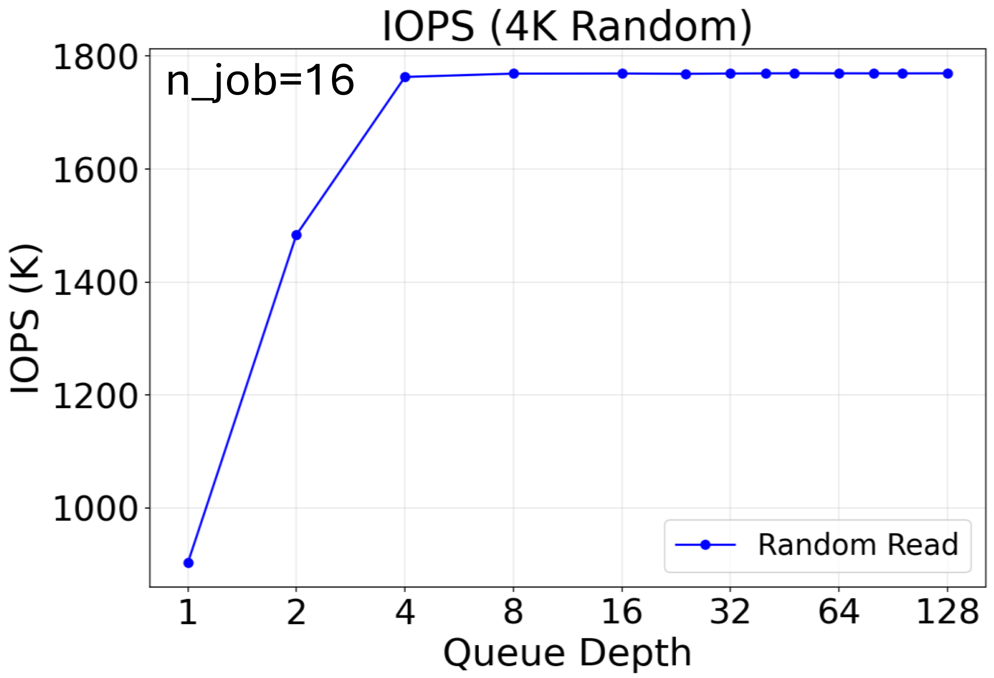
- Parallelism (striping)
  - logical grouping of physical blocks across packages and dies that allow them to be ***written and erased simultaneously***
- Reliability
  - use one block for parity (die-level RAID)
- Caveat
  - controller treat as one unit
  - GC penalty: update a 4 KB page requires more copies (write amplification)

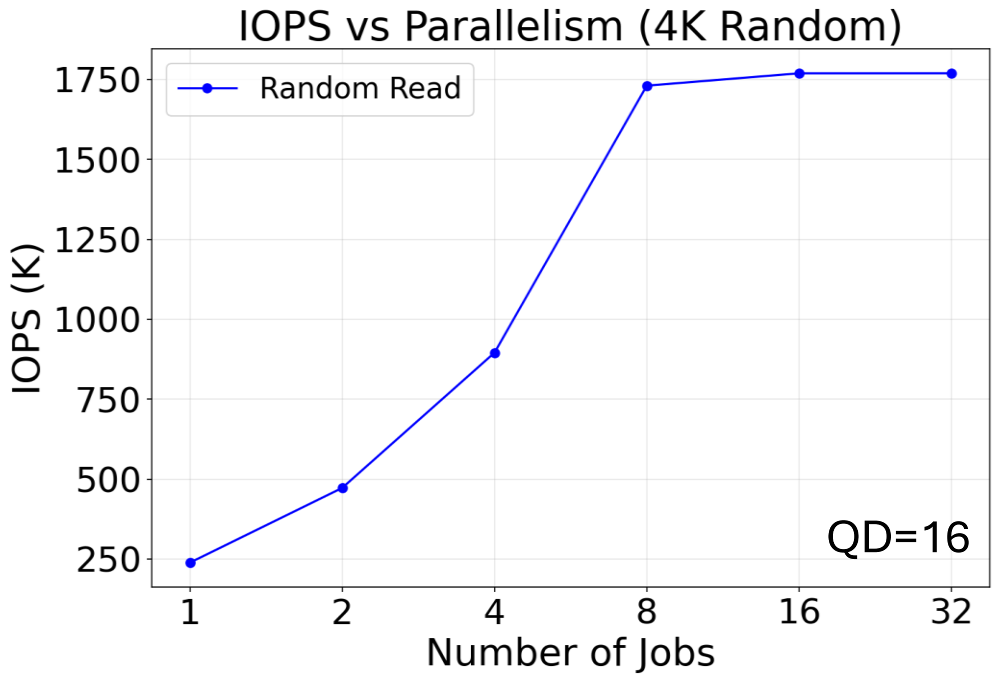


Note: this is different from file system superblock

SSD performance: bandwidth and IOPS

- Peak performance requires high concurrency
  - job: multiple proc to leverage queues and channels
  - queue depth: fill the pipeline and fully leverage parallelism





SSD performance: bandwidth and IOPS

- Write cliff: **Why?**
- Cause 1: running out of empty blocks
  - trigger blocking GC
- Cause 2: pSLC caching
  - treat a portion of TLC/QLC as SLC-mode cache for burst writes
  - when cache fills, writes fall back to native TLC/QLC speed (100s MB/s)
- Write performance therefore depends on workload and free space
- SSD expected sustained write bandwidth can be calculated using endurance
  - 1 TB TLC at 2000 PE and 5-year lifetime: 1 TB * 2000 / (5 * 365 * 24 *3600s) = 12 MB/s

SSD performance: latency

- Page read: 10s μs
  - lower than mechanical disks (10 ms), but higher than DRAM (10s ns)
- Page program: 100s μs
- Block erase: 1s ms
- Latency increases from SLC to QLC
  - harder to read, retry and error correction, *every extra bit doubles the latency*
  - harder to program, write latency increases more than read
- The latency above is for SSD medium
  - system level: read is 10-100s μs, write is shorter due to DRAM buffer
  - tail latency is more useful, but not always determined by device

Comparing NAND cell types

Cell Type	Bits / Cell	Read latency	Write latency	True read bandwidth (MB/s)	True write bandwidth (MB/s)	P/E cycles (Endurance)	Approx. cost (Rel.)
SLC	1	~25 µs	~250 µs	7,000 – 14,000	10,000+	50,000 – 100,000	Very High (5x–10x TLC)
MLC	2	~50 µs	~900 µs	5,000 – 10,000	~2,000	3,000 – 10,000	High (2x–3x TLC)
TLC	3	~75 µs	~1,500 µs	3,000 – 7,000	~800	1,000 – 3,000	Medium (Baseline)
QLC	4	~120 µs	~3,000 µs	1,500 – 4,000	~80-160	100 – 1,000	Low (0.6x–0.8x TLC)
PLC	5	~250+ µs	~5,000+ µs	500 – 1,000	~10-40	<100	Very Low (expected)

SSD reliability

SSD reliability overview

- More reliable than HDDs
  - no mechanical component
  - not susceptible to vibration
  - Uncorrectable Bit Error Rate (UBER): 1-2 order lower than HDD
  - AFR: half of HDD based on BackBlaze report*
- However, when SSDs dies
  - not recoverable

https://www.backblaze.com/blog/backblaze-drive-stats-for-q3-2025/

SSD failure modes

- NAND failure modes
  - wear out: P/E cycling
  - data retention loss: charge leakage over time (increase with wear and temperature)
  - read / program disturb: change the threshold distribution of adjacent cells
  - bad blocks: from manufacturing
- Firmware bug
- Controller, DRAM hardware failure: rare

Mitigation

- Over-provisioning
  - wear out and bad block
- Error correction code (ECC)
  - NAND stores parity in OOB/spare area
  - detect and correct errors during read
- Read-retry
- Refresh: prevent data retention loss

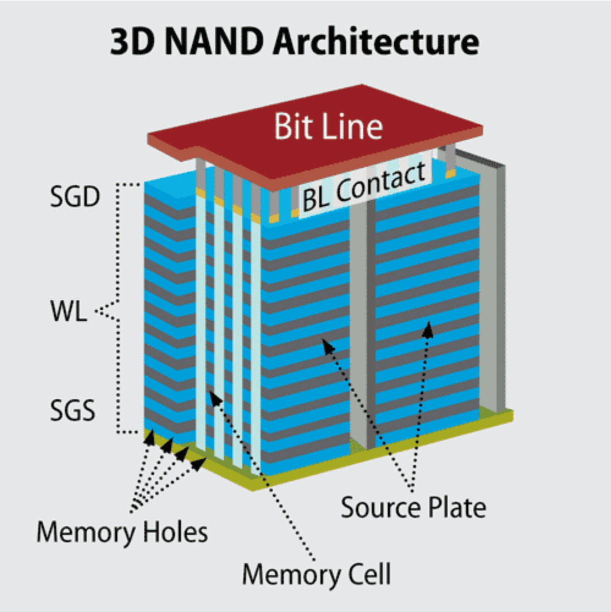
Endurance

- SSDs have limited endurance
  - 1 TB of TLC SSD only support 12 MB/s sustained write
- How do we measure endurance?
- TBW: TB written
- DWPD: disk write per day
  - consumer: 0.1–0.3
  - enterprise: 0.3 (read intensive)–3 (write intensive)
  - DWPD 0.3 allows users to write < 4MB/s per TB, lower than HDDs

SSD Density

SSD density

- Cell size: Moore’s law
- More bits per cell:
  - diminishing return
  - SLC->MLC: 100%
  - MLC->TLC: 50%
  - TLC->QLC: 33%
  - QLC->PLC: 20%



3D NAND: 2D -> 3D

- view layer as one floor
- read operation: controller selects one **Bit Line** (column) and one **Word Line** (floor), pinpointing a single cell in the 3D matrix
- modern drives: 300+ layers

Metric	**2D (Planar) NAND**	**3D NAND**
Scaling Direction	Horizontal (x, y)	**Vertical (z)**
Capacity Limit	~128 Gb per die	**2 Tb+ per die (and rising)**
Endurance	Low (tiny cells)	**Higher** (larger, more stable cells)
Cost per GB	High (stalled scaling)	**Low (scaling via layering)**

SSD Cost and Market

SSD cost and market

Component	Cost Share (Approx.)	Function
NAND Flash	**70% – 85%**	Storage media
DRAM Cache	**5% – 10%**	Mapping table & write buffer
Controller	**5% – 8%**	CPU, ECC Engine, PCIe PHY
PCB & others	**~5%**	Board, capacitors (PLP)
Assembly & Test	**~2%**	Packaging, Burn-in testing

To reduce cost: reduce DRAM and lower NAND cost

- Global Market Size (2025): ~$61.3 Billion
- Projected Growth: CAGR of ~16–18% through 2030
- Key Shift: In 2026, Enterprise SSDs are projected to overtake Client SSDs to become the largest segment of the NAND market by revenue.

SSD market segmentation

- **Enterprise / Data Center**: hyper-growth due to data gravity
- **Consumer**: stagnant and low growth due to softened PC demand

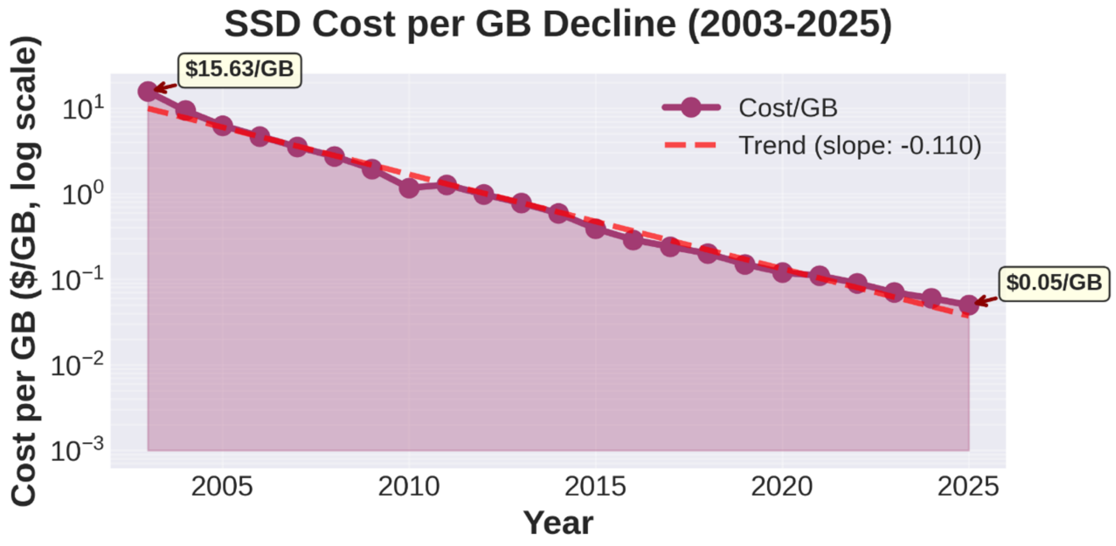
Manufacturer	Market Share (2025)	Strategy & Strength
Samsung	**~32%**	**Volume Leader.** Dominates both Client and Enterprise
SK Hynix / Solidigm	**~24%**	**The QLC King.** Solidigm (formerly Intel NAND) holds a unique lead in ultra-high capacity (60TB+) QLC drives
Kioxia (Toshiba)	**~15%**	**Mobile Focused.** Strong in smartphones (iPhone storage)
Micron	**~11%**	Often first to market with the highest layer counts

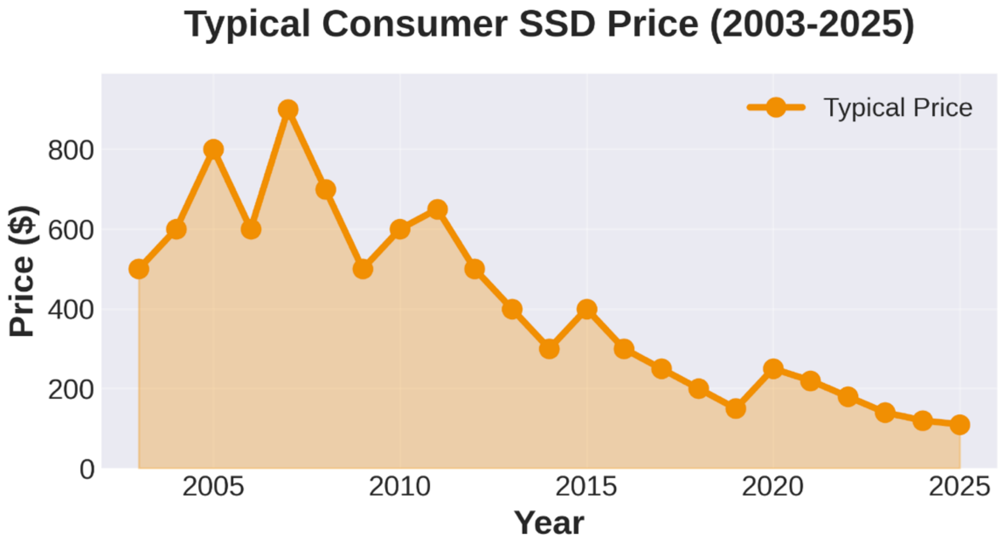
Note: not all manufactures produce both NAND and controller

Trend

Historical trend

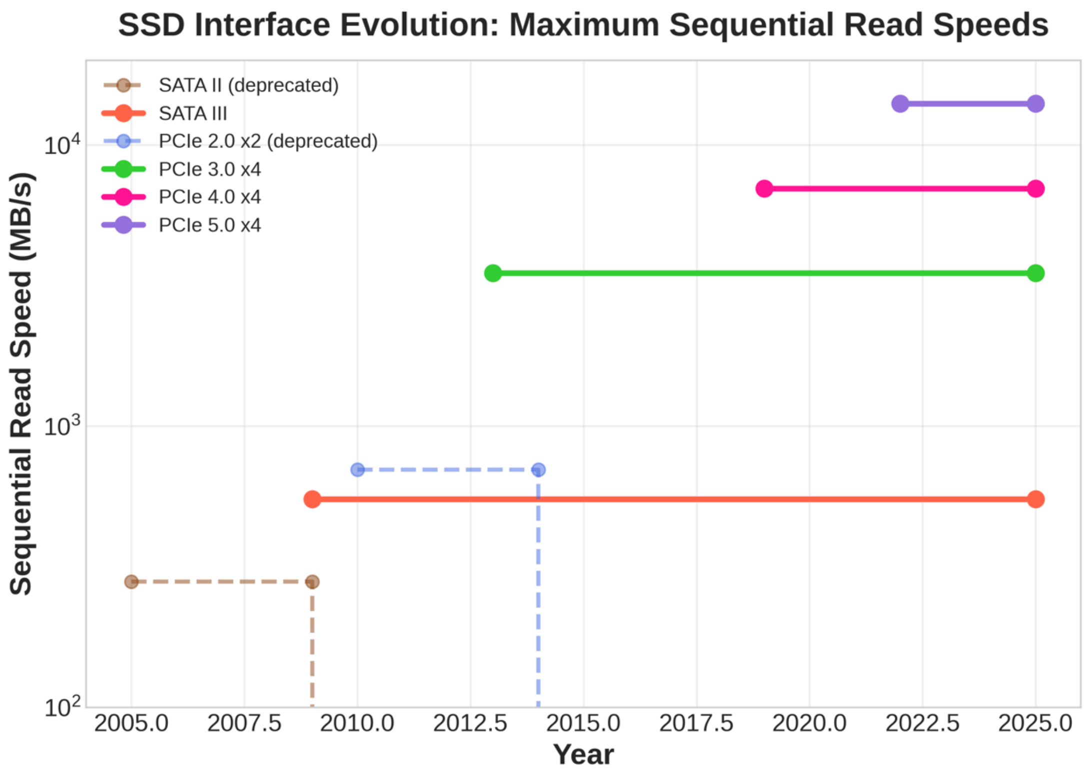
- Density improves close to exponentially

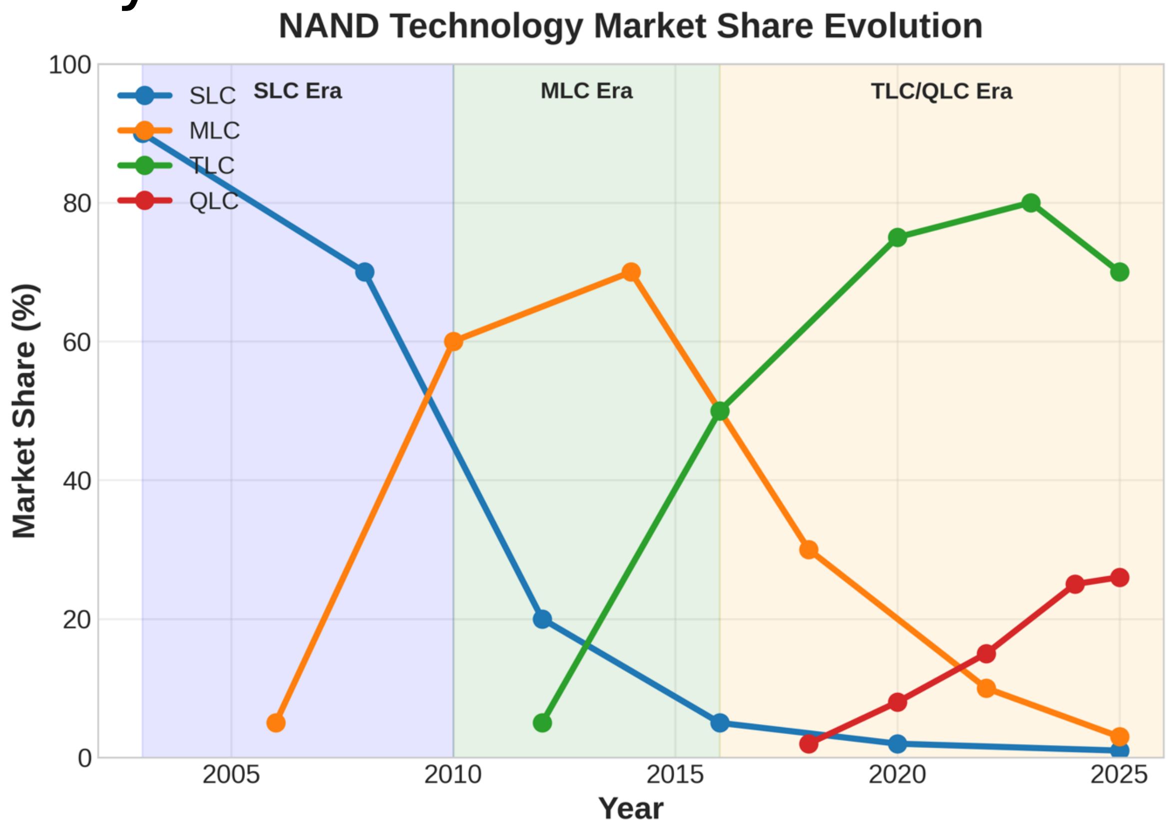




Historical trend

- Bandwidth is always close to interface speed
- Latency has been decreasing steadily





Future trend

- Denser flash
  - more bits per cell, will we have PLC or HLC? Doubtful
  - 3D NAND more layers **Yes!**
- Lower latency? Doubtful
  - we are already at ~10 μs, other components (software, PCIe, controller) will become bottleneck
- Higher bandwidth? **Yes!**
  - more parallelism, controller and interface improvement
- More IOPS? **Yes!** (same)
- Better power efficiency? **Yes!**

Future trend: implications

- Increasing enterprise adoption for
  - performance
  - physical footprint
  - sustainability (high-density QLC)
- Bandwidth gap between SSDs and DRAM narrows
  - many data can be offloaded to SSDs
  - but what should we do with limited endurance?
- Latency is increasingly lower
  - kernel I/O path has become the bottleneck

Summary

- SSD internal
  - NAND flash physics
  - SLC/MLC/TLC/QLC
  - internal organization
- SSD controller
  - FTL: mapping, garbage collection
- SSD performance
  - high bandwidth and IOPS through internal parallelism
- SSD reliability and SSD density

Three key questions

- What is Flash Translation Layer and what does it do?
- What is SSD's performance characteristics and Why?
- How should you extend SSD lifetime?

Next time

- Block layer
- Interface and protocols
- Device driver
- I/O controller