Avoiding Costly Read Latency Variations in SSDs Through I/O Determinism

Steven Wells

Toshiba America Electronic Components, Inc.

Santa Clara, CA
August 2017

Flash Memory Summit

Excellent Latency

- SSD designs have been about high average bandwidth for the last 30 years
- At best we spoke about bandwidth consistency
- A new paradigm focusing on latency is emerging from the needs of hyperscale data centers
- “IO Determinism”

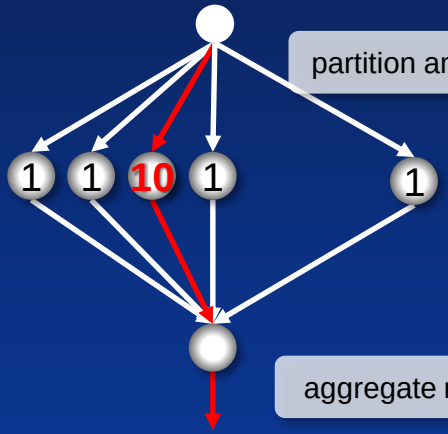
Latency Tail Impact

Legacy Mindset

1 worker
n iterations
Latency O(n)

Execution time ≈ ***Average*** lookup latency

Hyperscale Mindset



n workers
1 lookup each
Latency O(1)

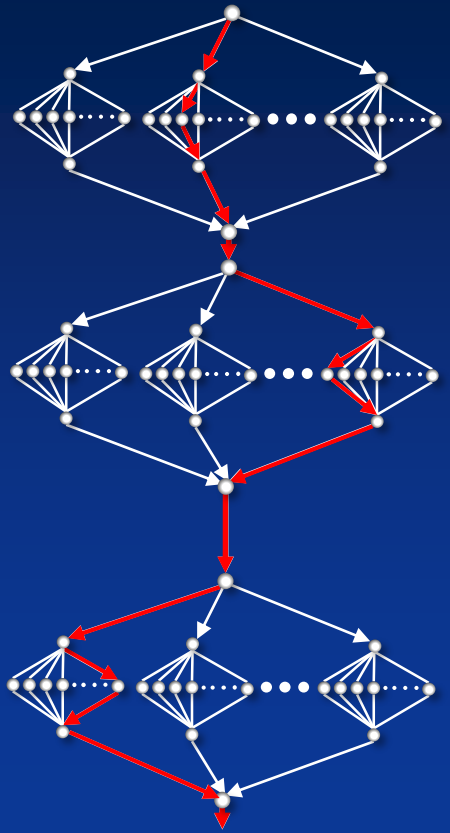
Execution time ≈ ***Longest*** lookup latency

Real Implementations are...

"In practice, a single user request may result in thousands of subqueries, with a **critical path that is dozens of subqueries long**."

"The fork/join structure of subqueries causes latency outliers to have a **disproportionate effect on total latency**, and the large number of subqueries would cause slowdowns or unavailability to quickly propagate…"

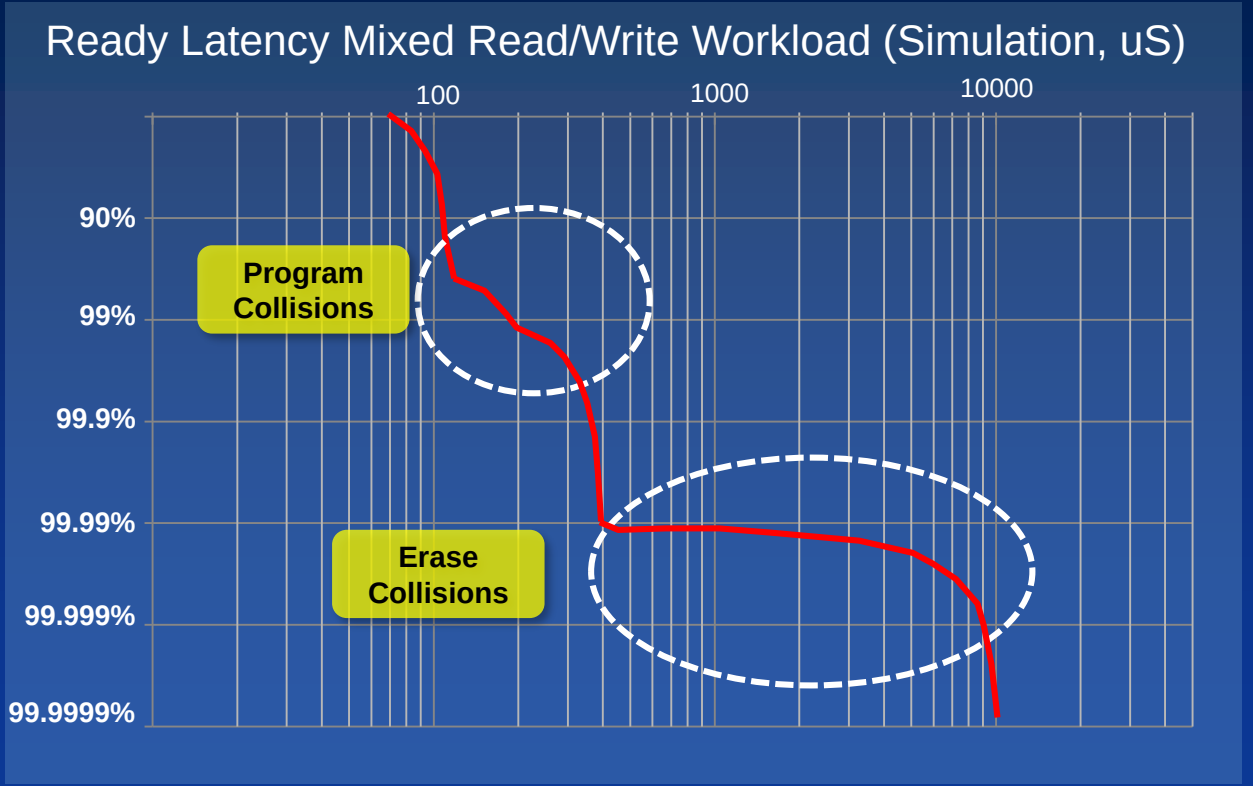
Challenges to Adopting Stronger Consistency at Scale
- Ajoux et. Al., (Facebook & USC)



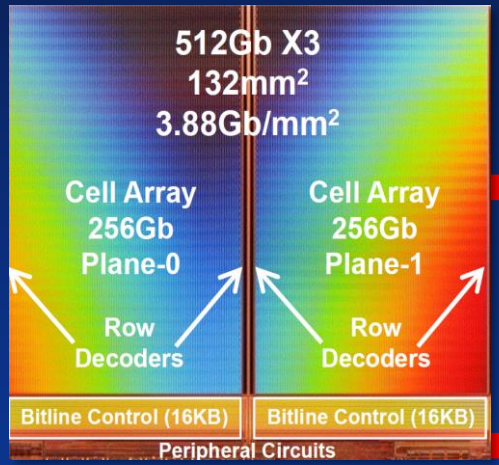
Flash Memory Summit

Latency is Affected by Maintenance

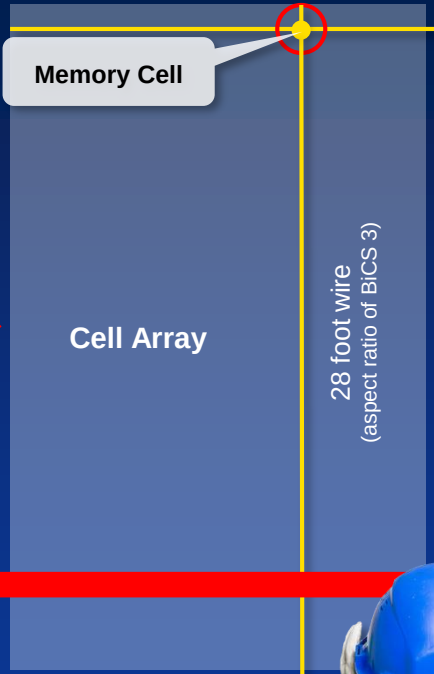
The Maintenance of an SSD



Suspends is Not a Solution for Deterministic Reads



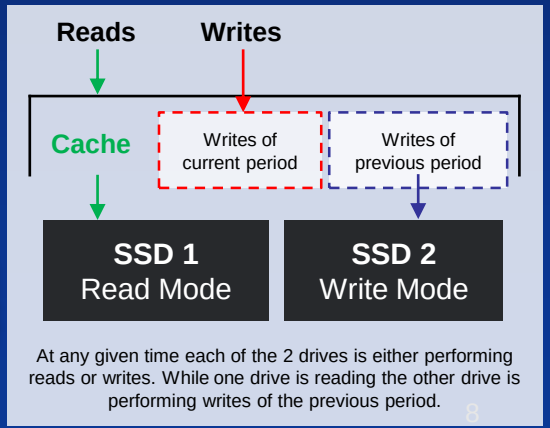
Toshiba BiCS FLASH™
64-Layer 3D Memory



8 foot wire
(aspect ratio of BiCS 3)

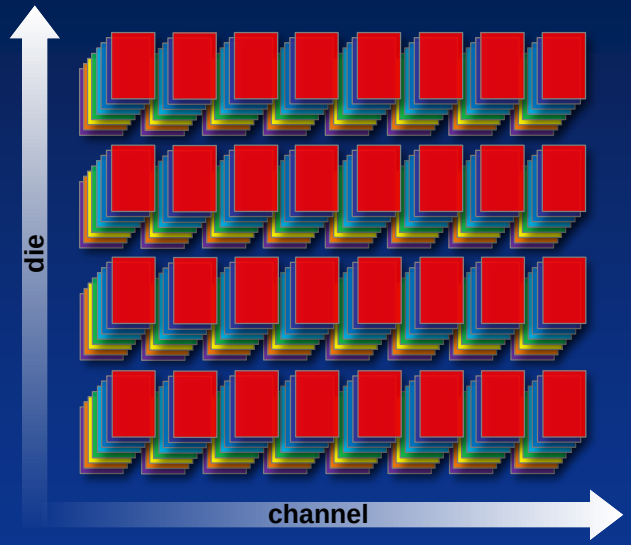
A New Working Model and Prior Work

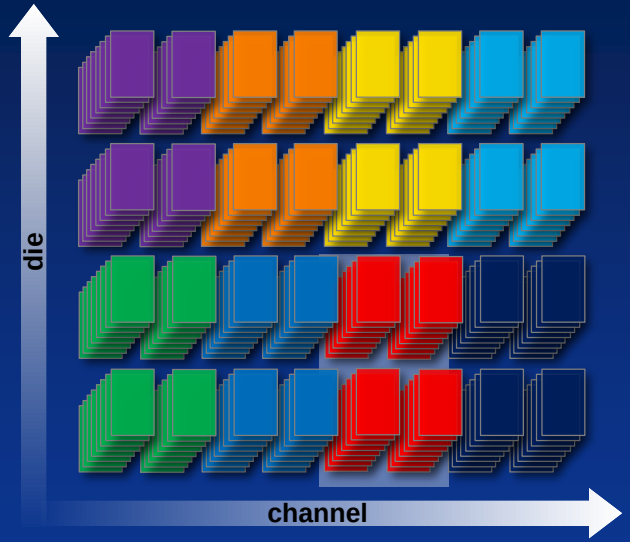
- Analogy: Fire department has multiple engines. Each is taken offline periodically for maintenance. But not all at the same time!
- Prior System Work: Flash on Rails - Consistent Flash Performance through Redundancy (USENIX 2014)
- With SSD size scaling into 16TB and beyond, SSD level isn't the right solution



Flash Memory Summit

NVM Sets Architecture

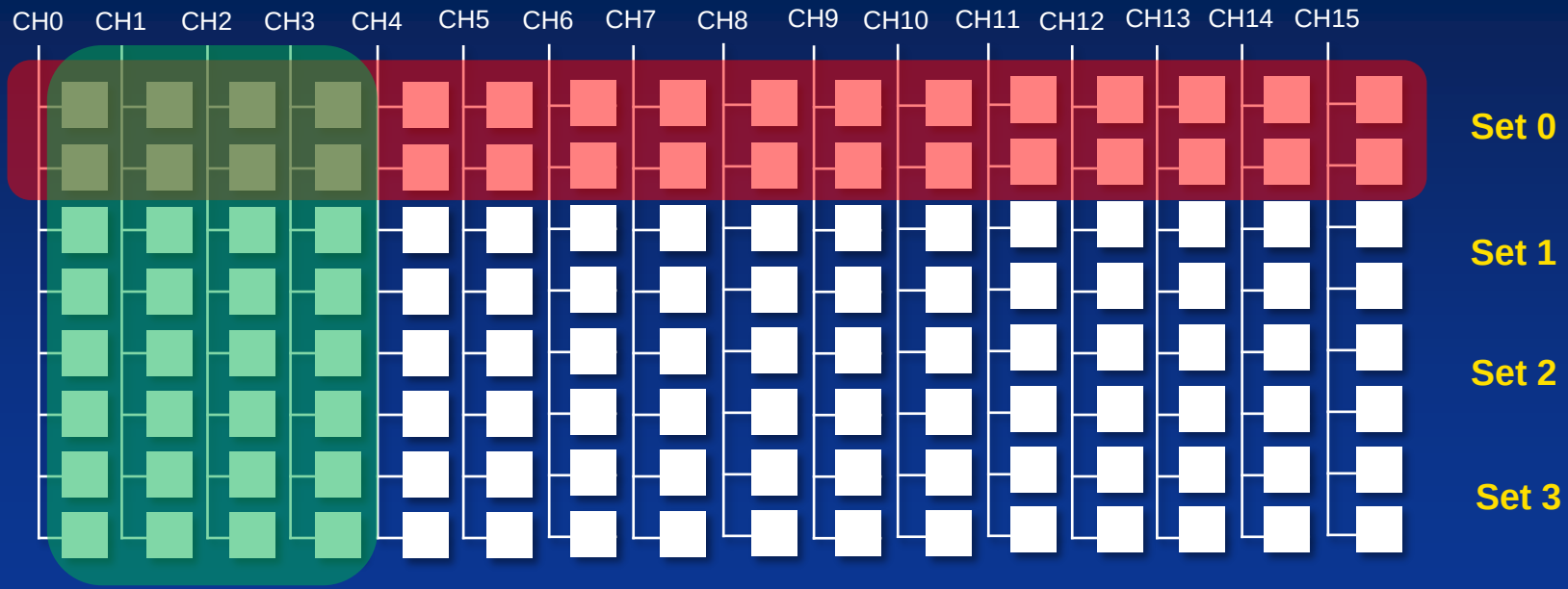




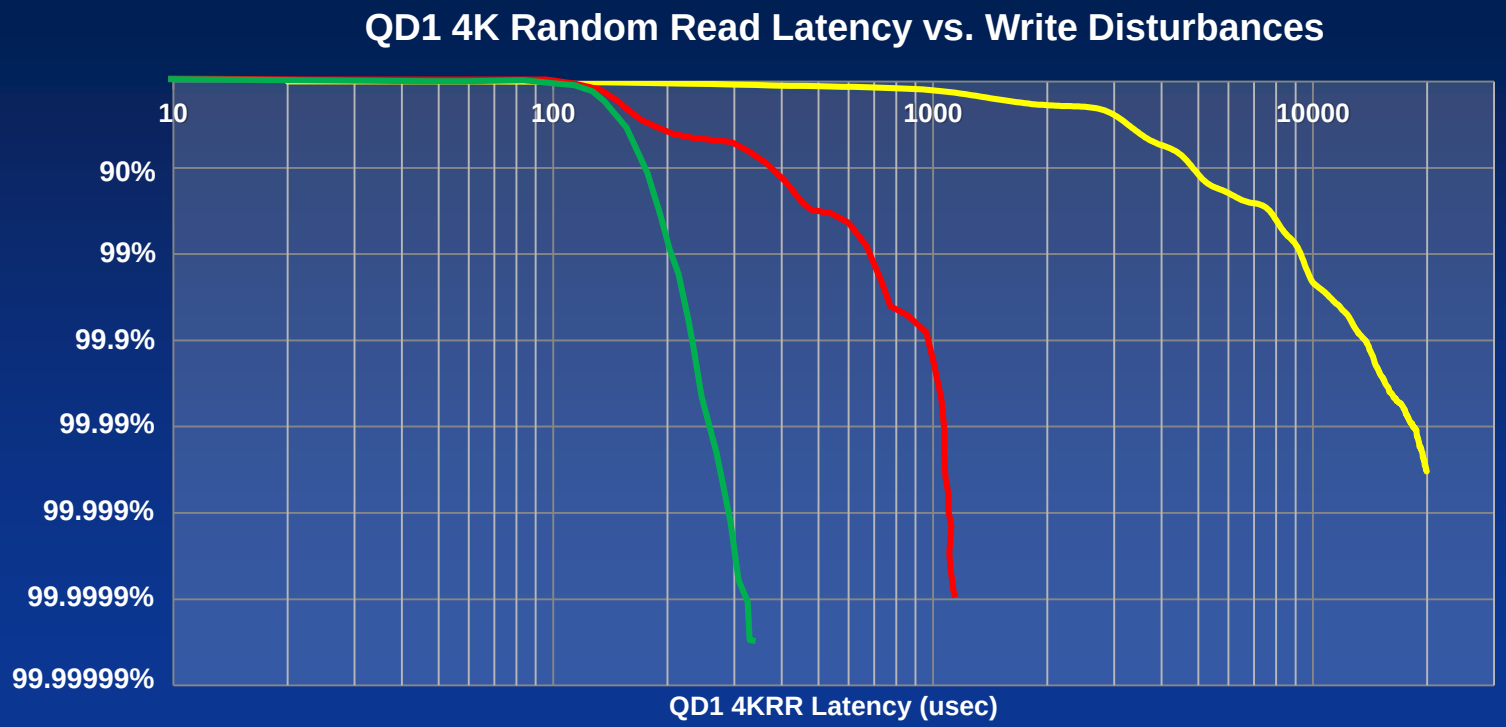
- Classic SSD architecture uses “bands” of devices on every channel to maximize bandwidth. Maintenance is also on every channel and every device
- New SSD array architecture creates independent NVM Sets ~1TB/set

Flash Memory Summit

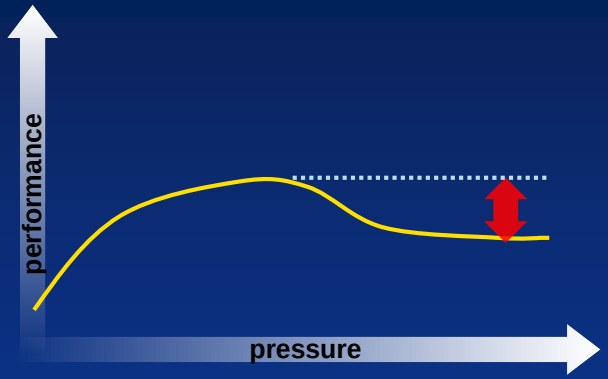
Proof of Concept



Proof of Concept Results



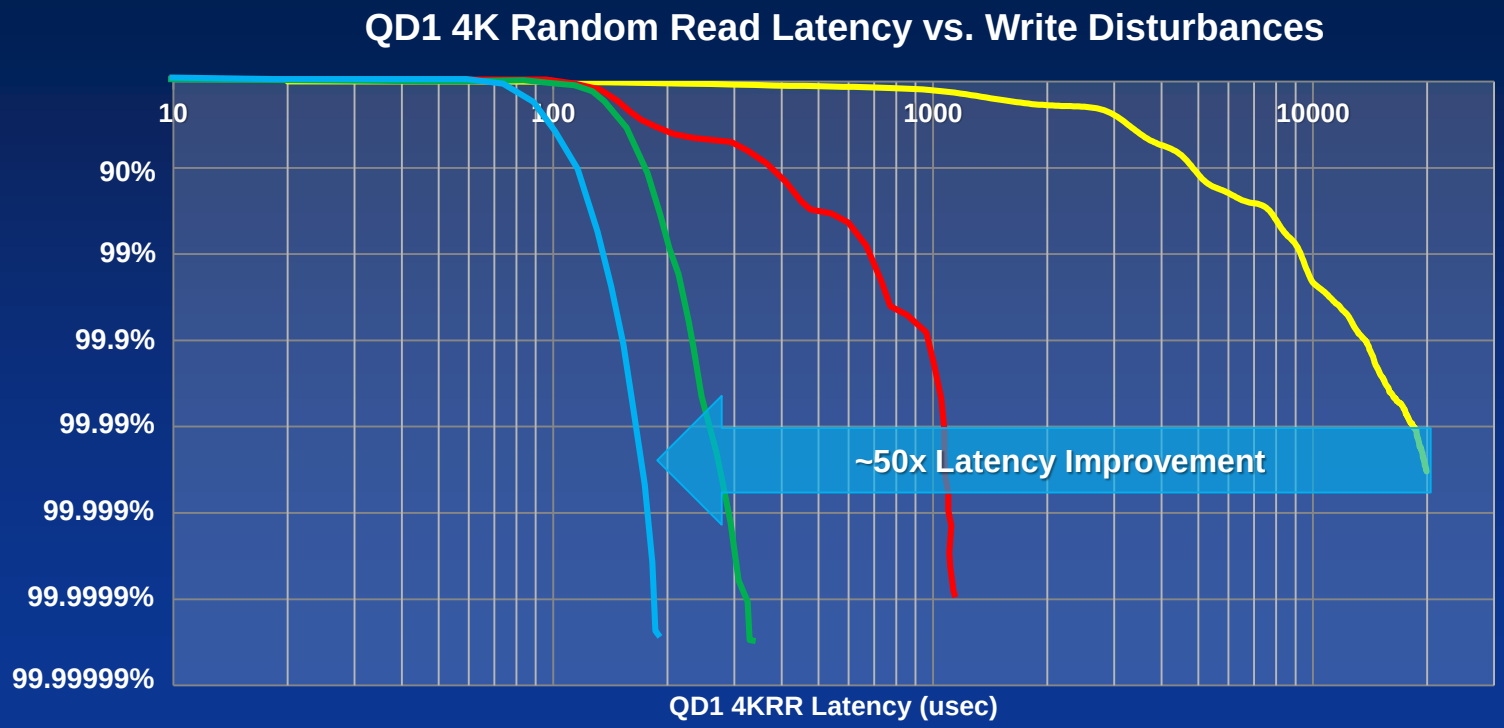
Managing Pressure Fatigue



- Producer pressure well beyond the capability of the consumer can degrade performance
- Pressure can be both host and internal
- Managing pressure allows reads to immediately take a priority position with a minimum of latency

Flash Memory Summit

Proof of Concept Final Result



Data from our floor demo

- For a fixed QD workload:
  - 50x latency improvement
  - 20x increase in read IOPS
  - 4x reduction in write IOPS

Closing Thoughts

- ~2 order of magnitude single SSD read latency improvement
- New SSD array architecture creating isolated “NVM Sets”
- A well behaved host is responsible for “ping-ponging” and other activities such as set to set wear leveling
- FW optimizations are required to support the new paradigm

Flash Memory Summit

Stop by the Toshiba Booth!

Booth #407