



Solving Latency Challenges with NVM Express SSDs at Scale

Chris Petersen
Facebook
Hardware Systems Technologist

Amber Huffman
Intel Fellow, Director Storage Interfaces
President, NVM Express, Inc.

Facebook @ Scale

facebook

Community Update

7.26.2017

Bringing the world closer together



2 Billion

on Facebook each month



100 Million

are members of meaningful groups



250 Million

use Stories each day



2 Billion

messages sent between
people and businesses
each month



250 Million

use Status each day



Connectivity

Aquila's second flight

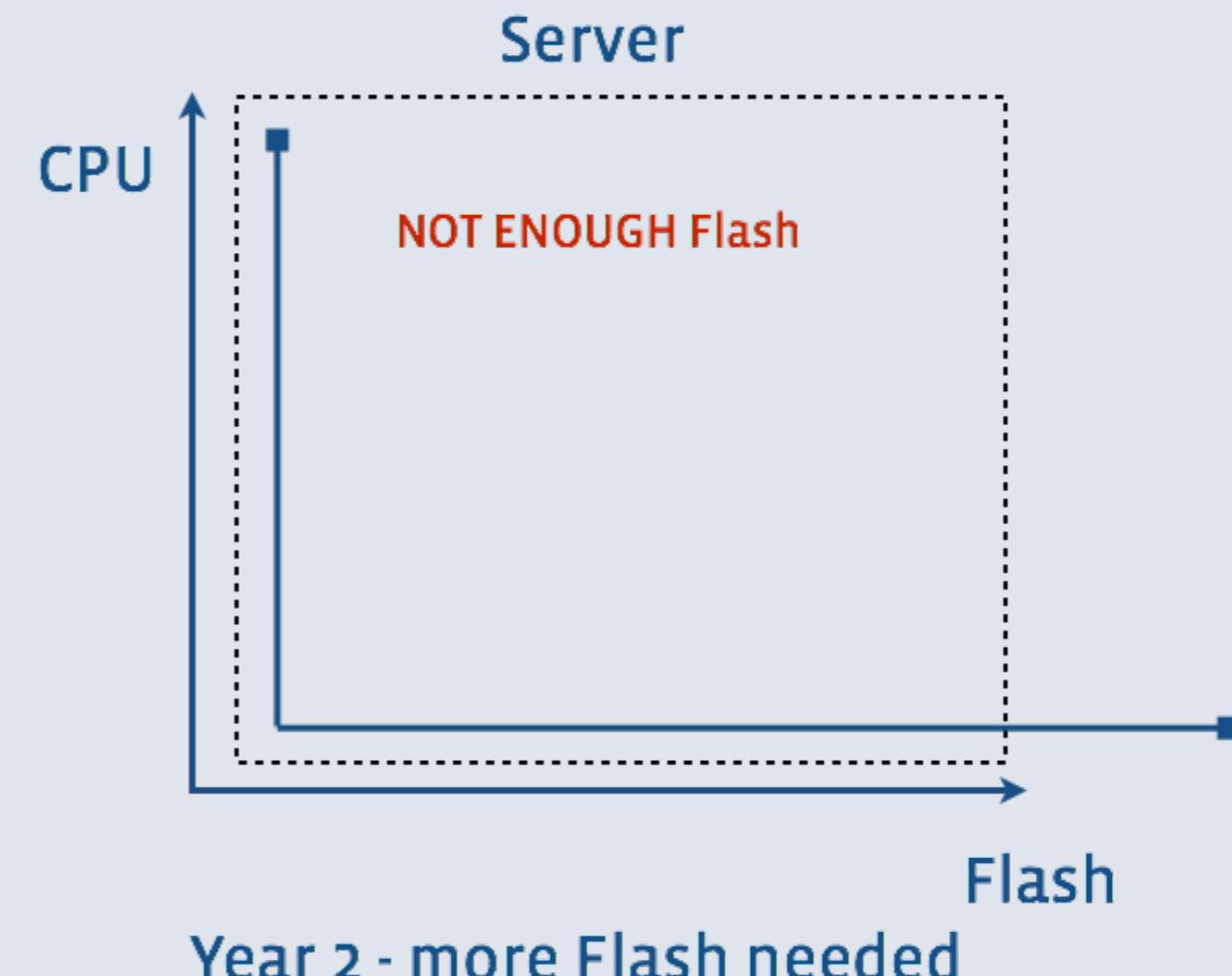
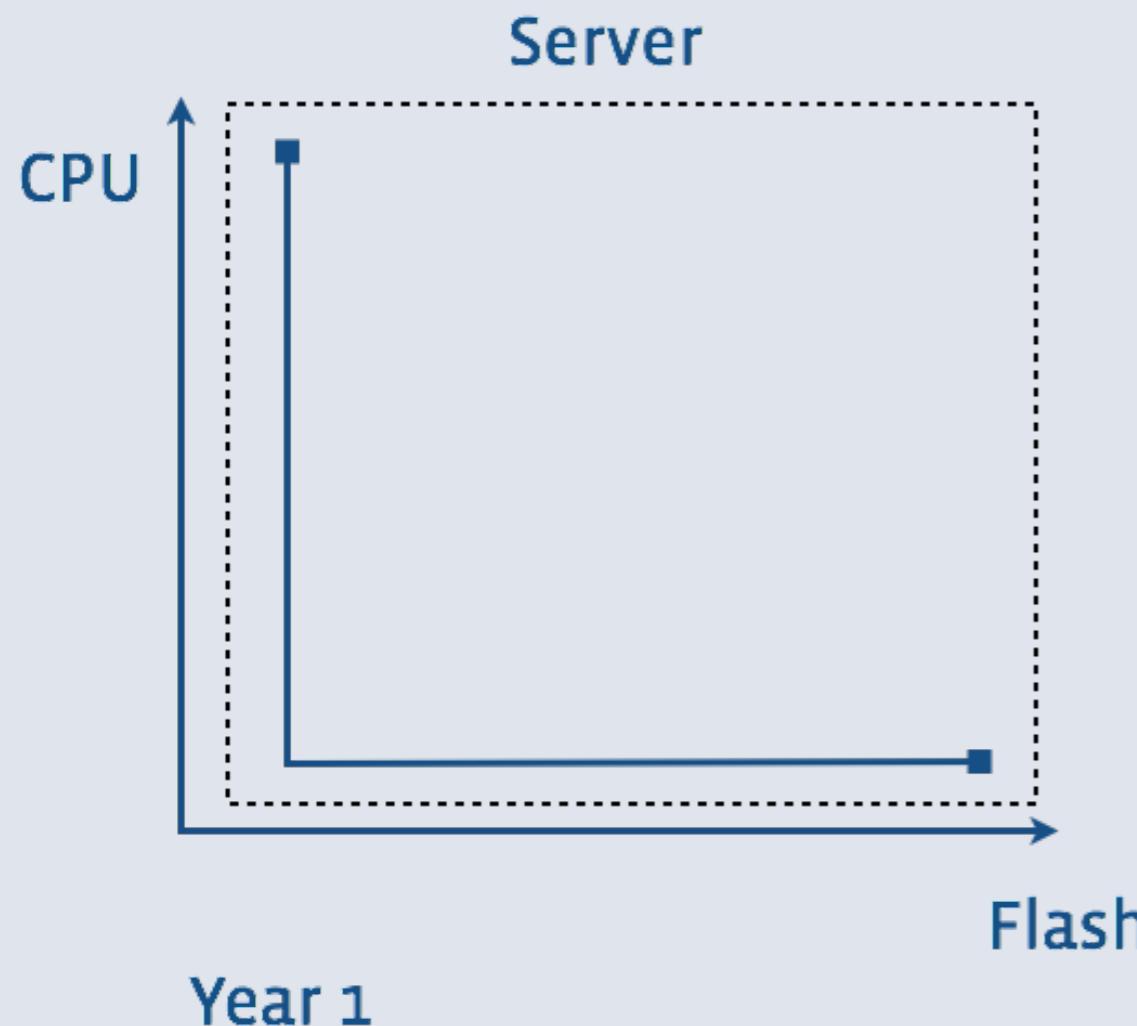


VR / AR

Launched Live from Spaces

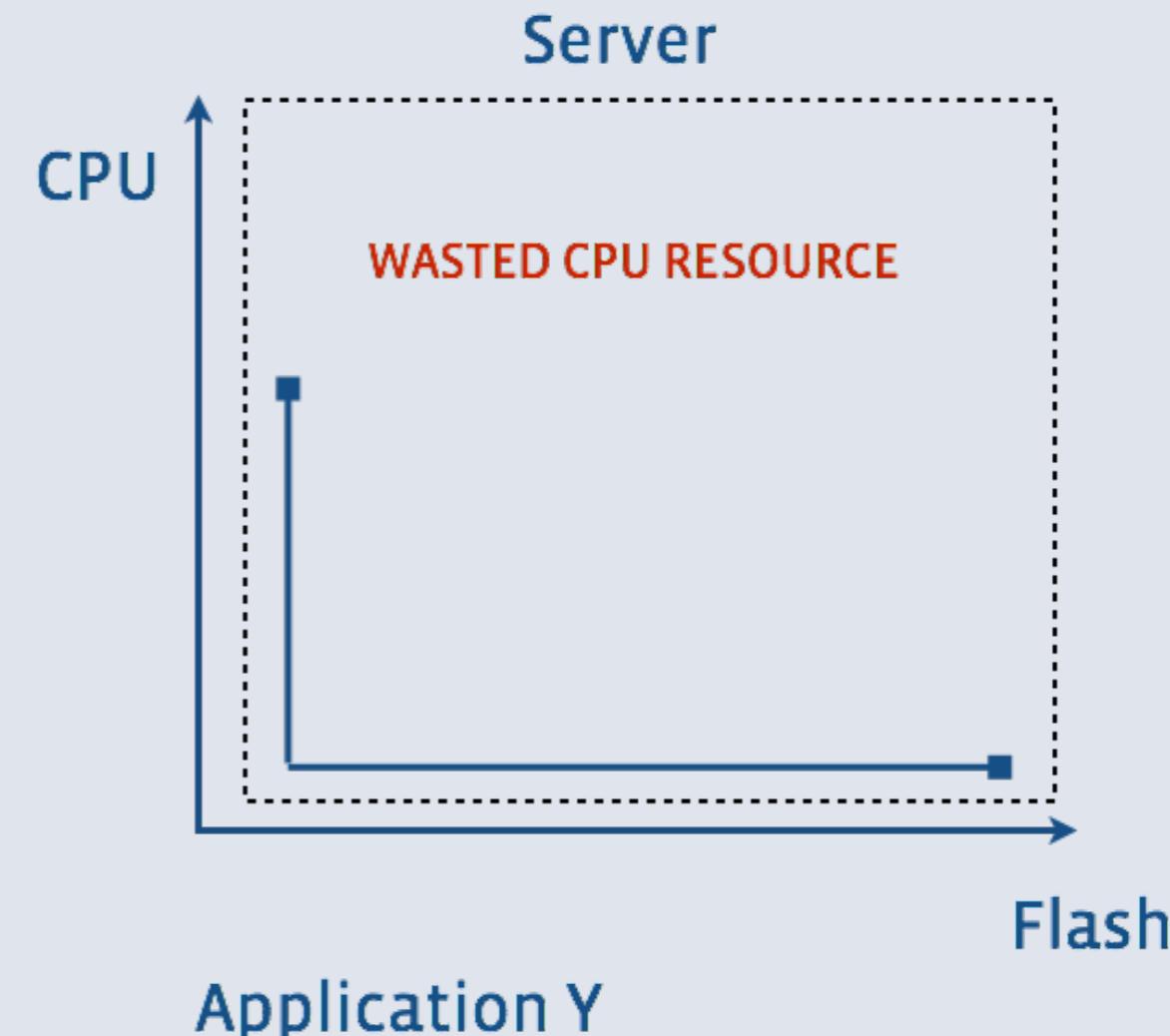
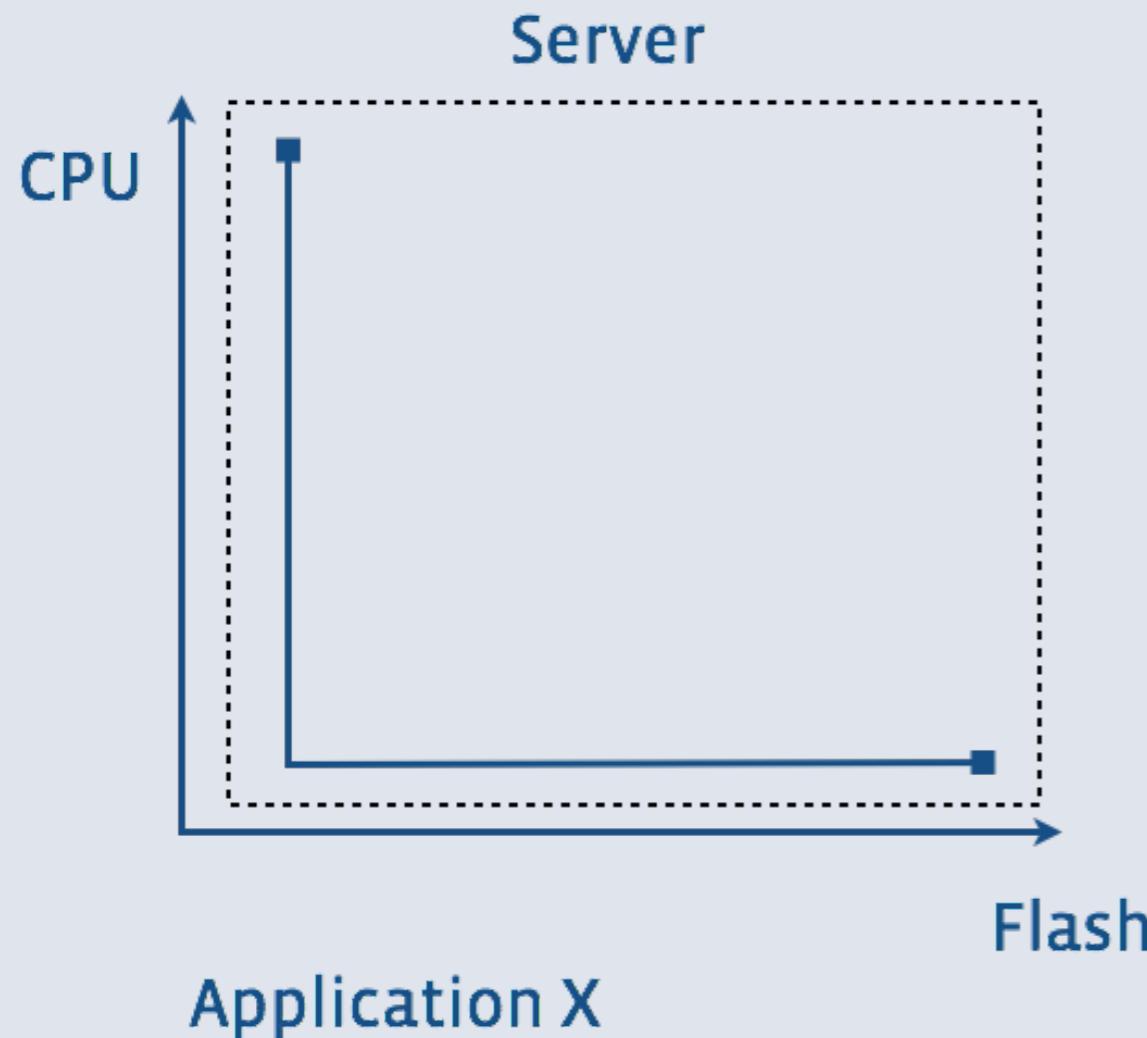
Disaggregated Flash

Applications change over time

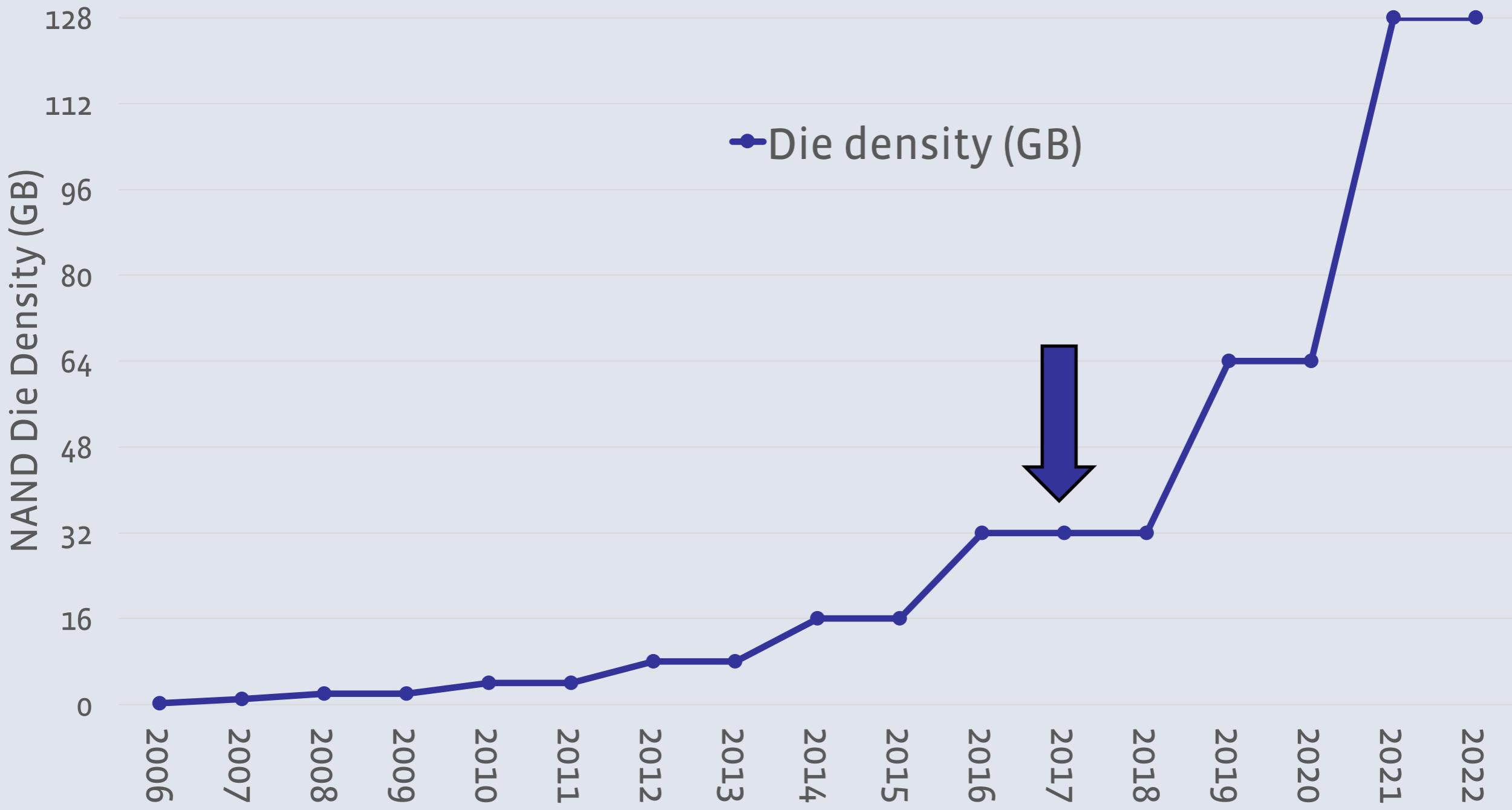


Disaggregated Flash

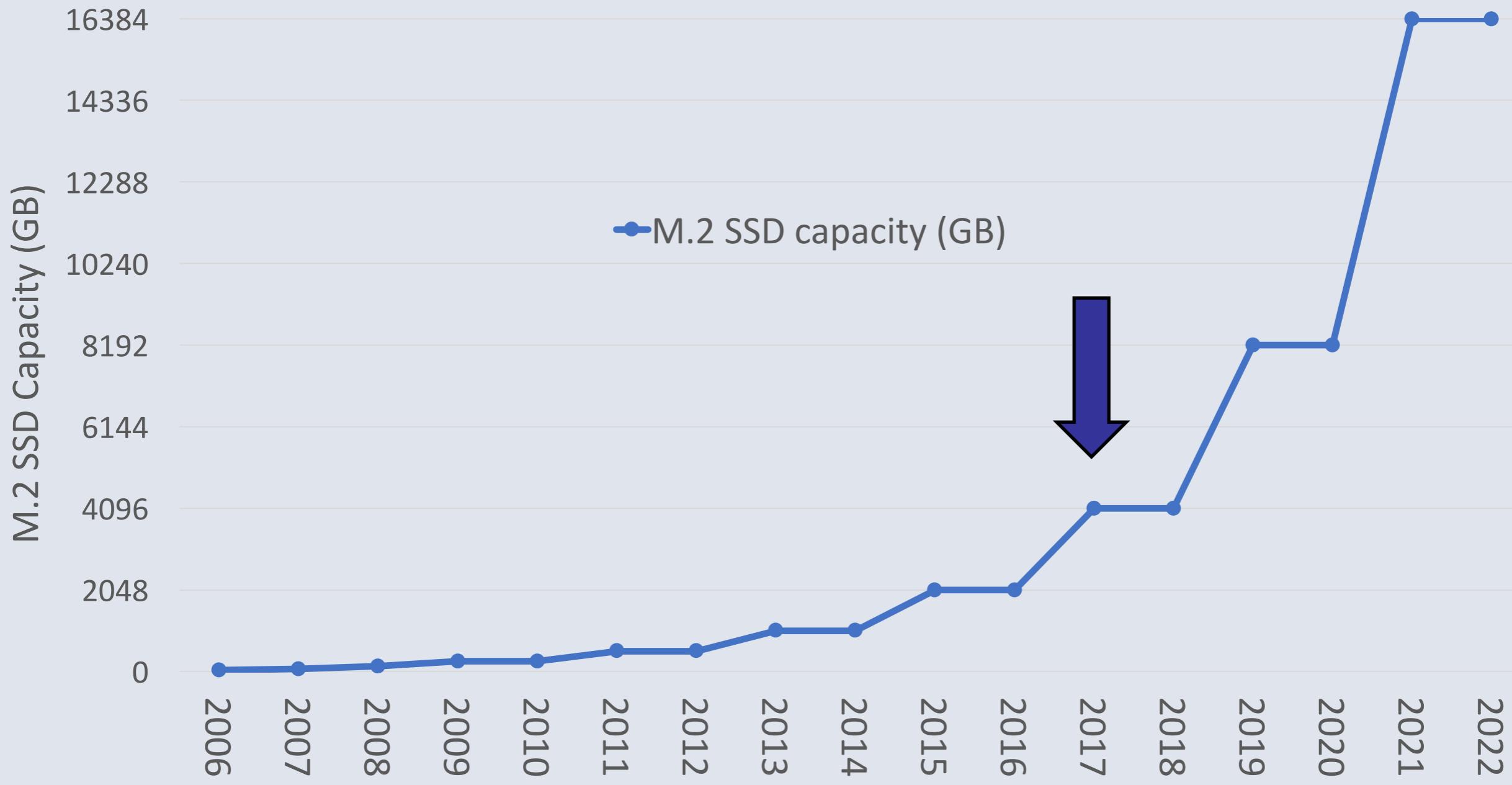
Applications have different needs



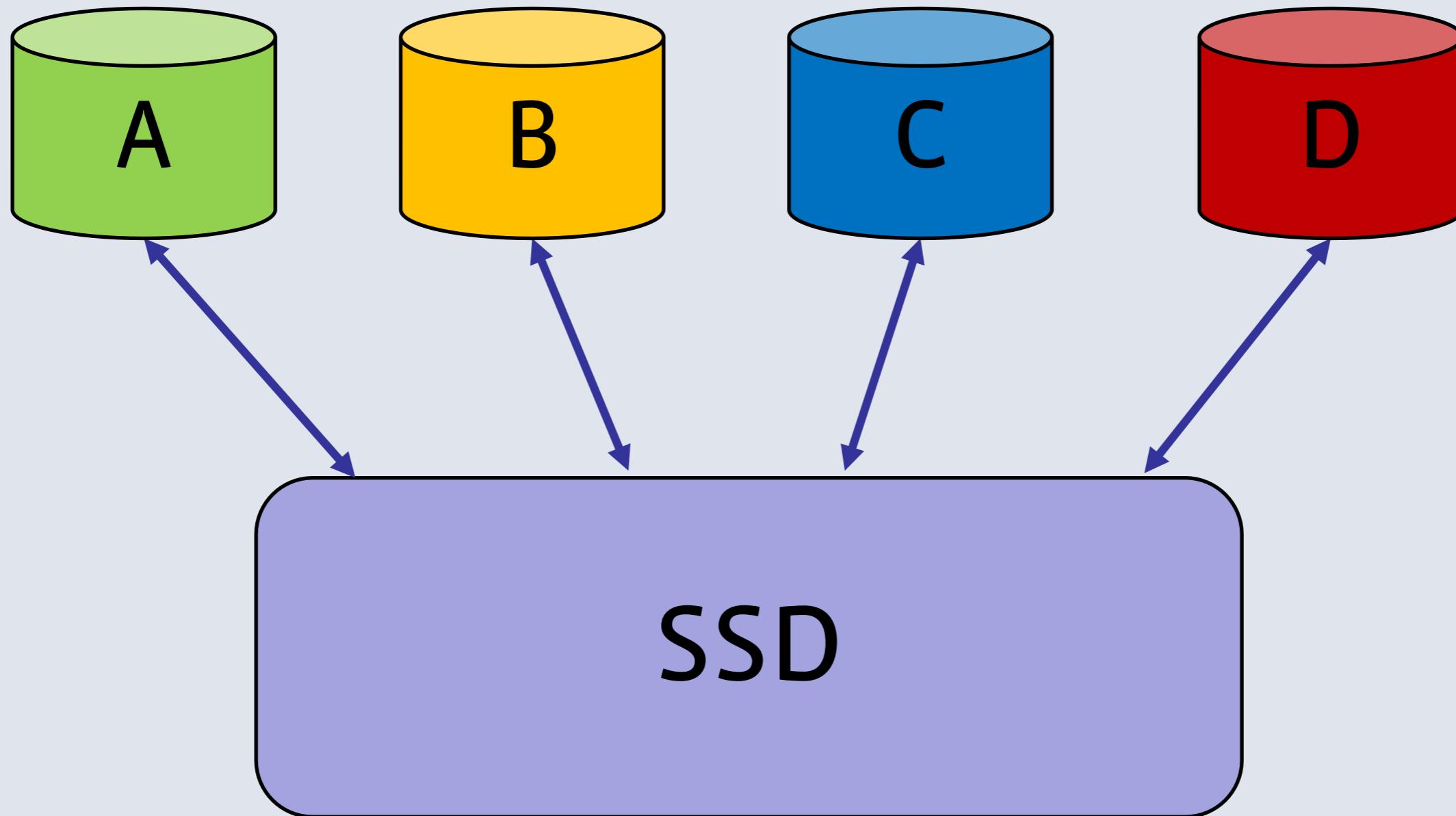
NAND Flash Trend



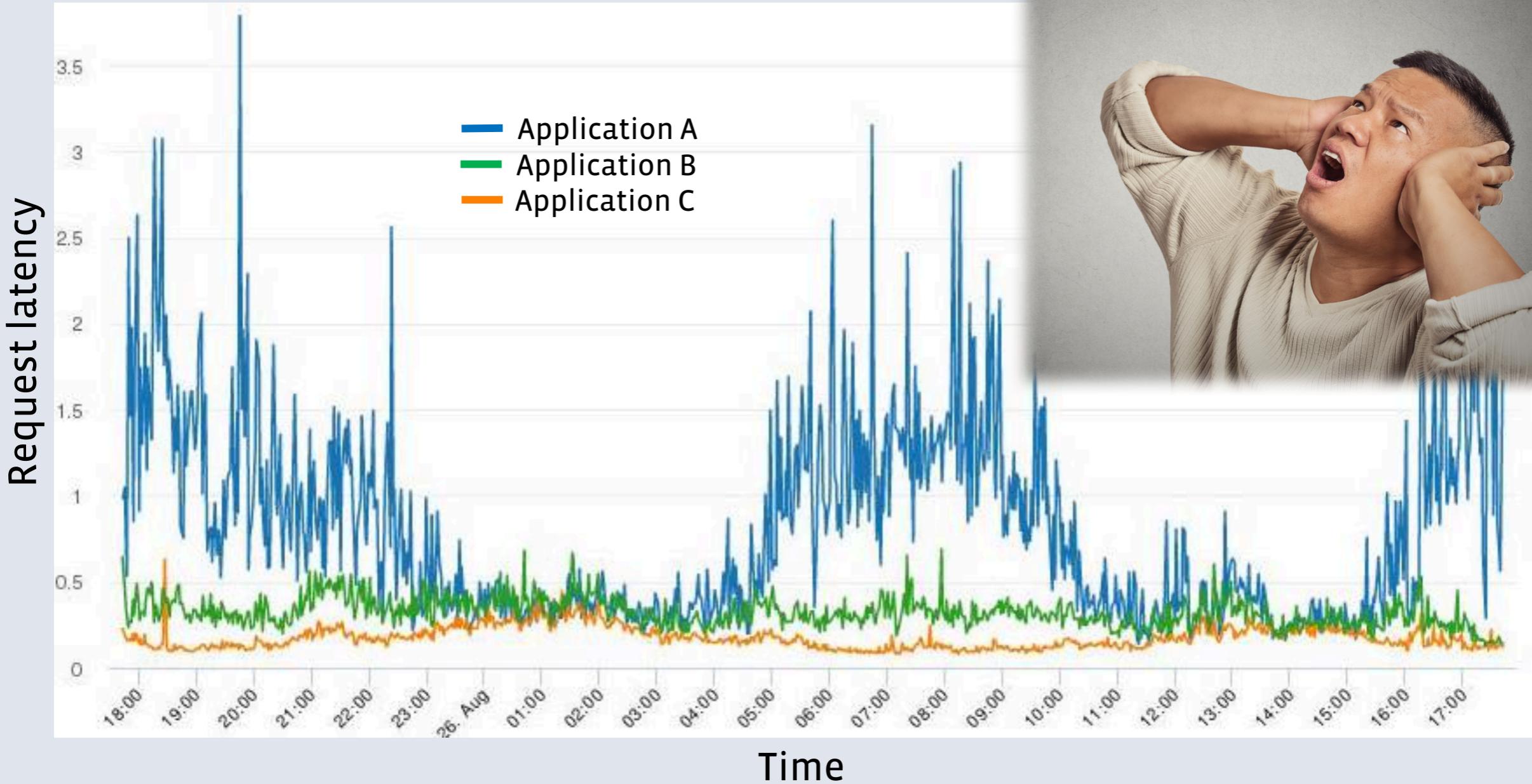
NAND Flash SSD Trend



SSD Capacity = Shared Resource



Noisy Neighbors

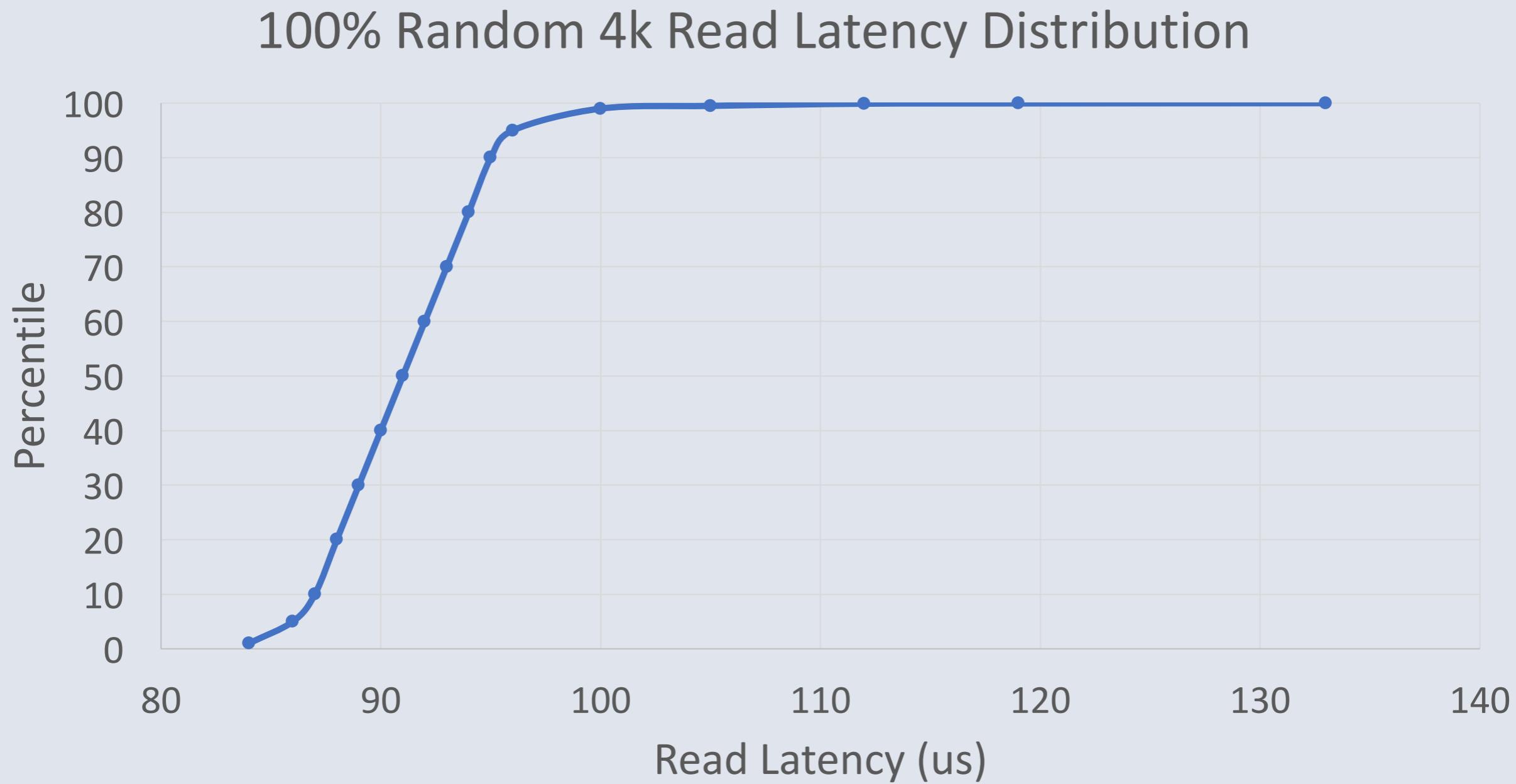


Latency vs. Bandwidth

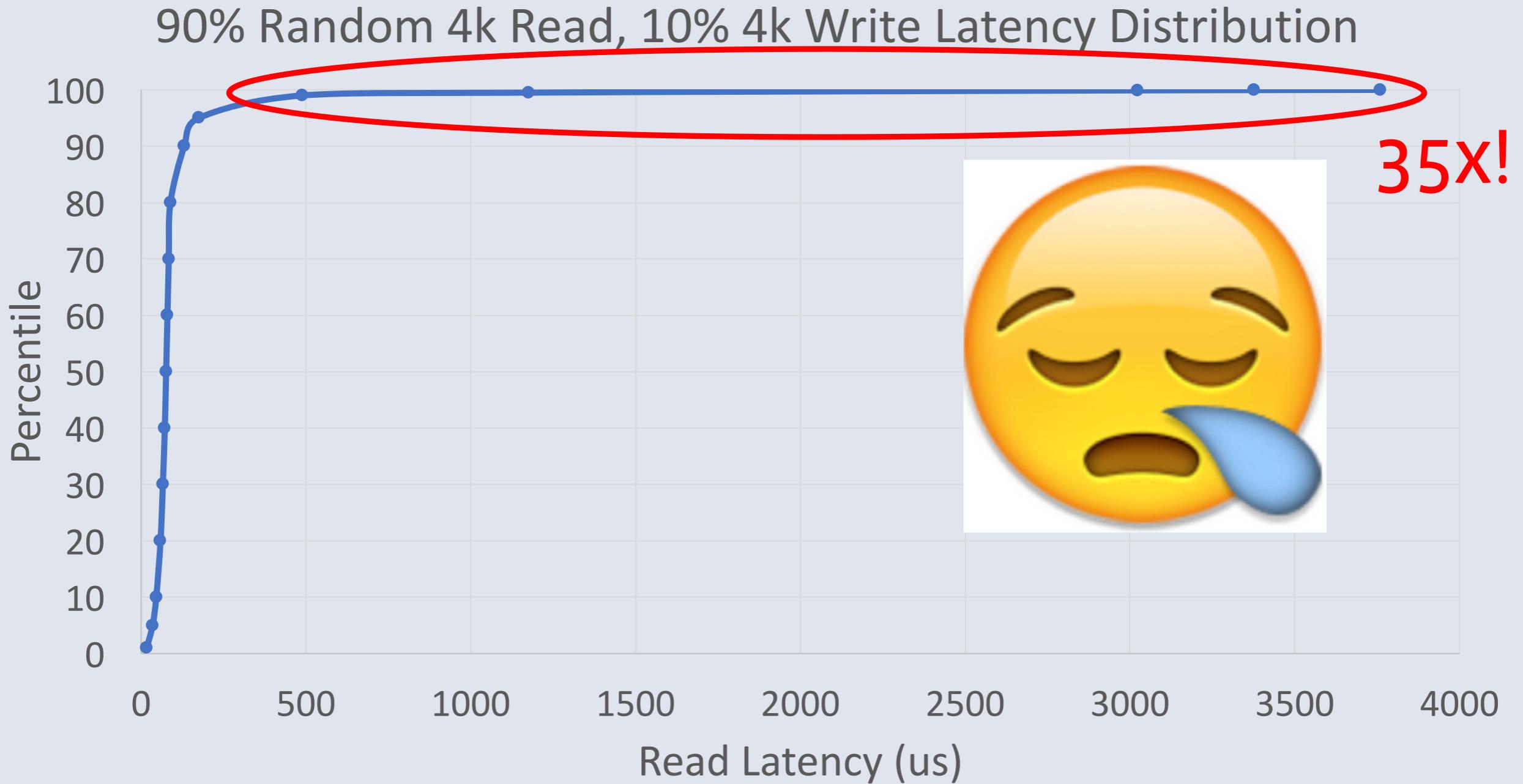


Image source: pixabay.com

Read Latency Challenge

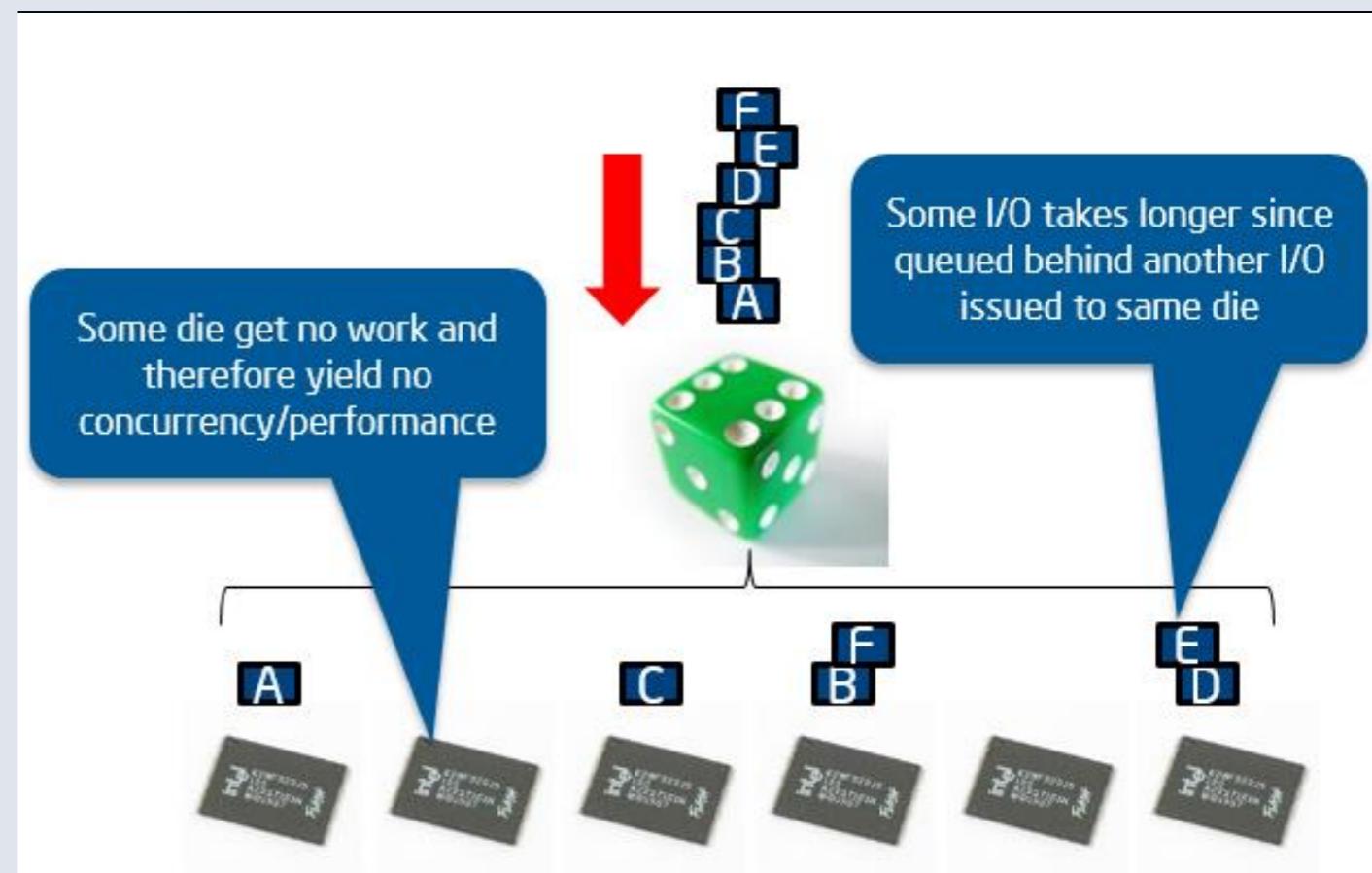


Read Latency Challenge



Read collisions

- Reads will collide with write or erase operations
- Writes are typically striped across many die to optimize BW
- For reads, no mechanism to target I/O to specific die
- Growing SSD capacities exacerbate this problem as IOPs/TB remains constant



“Yahtzee Effect: Statistical Clumping”
When rolling 6 dies with 6 faces, on average only 4 of the 6 values will come up

Hyper-scale SSD Requirements

- QoS-isolated, media-level partitions
- Simple SSD-to-host interface
- Media agnostic
- Guaranteed deterministic reads during some time periods

Re-negotiating the Data Center Storage Contract

1. What if the SSD exposed quality of service isolated regions?
2. What if the host only writes in certain time periods?
3. What if we trade the max error rate for the max read latency?

GENERAL SERVICE AGREEMENT

BETWEEN:

DATA CENTER

- AND -

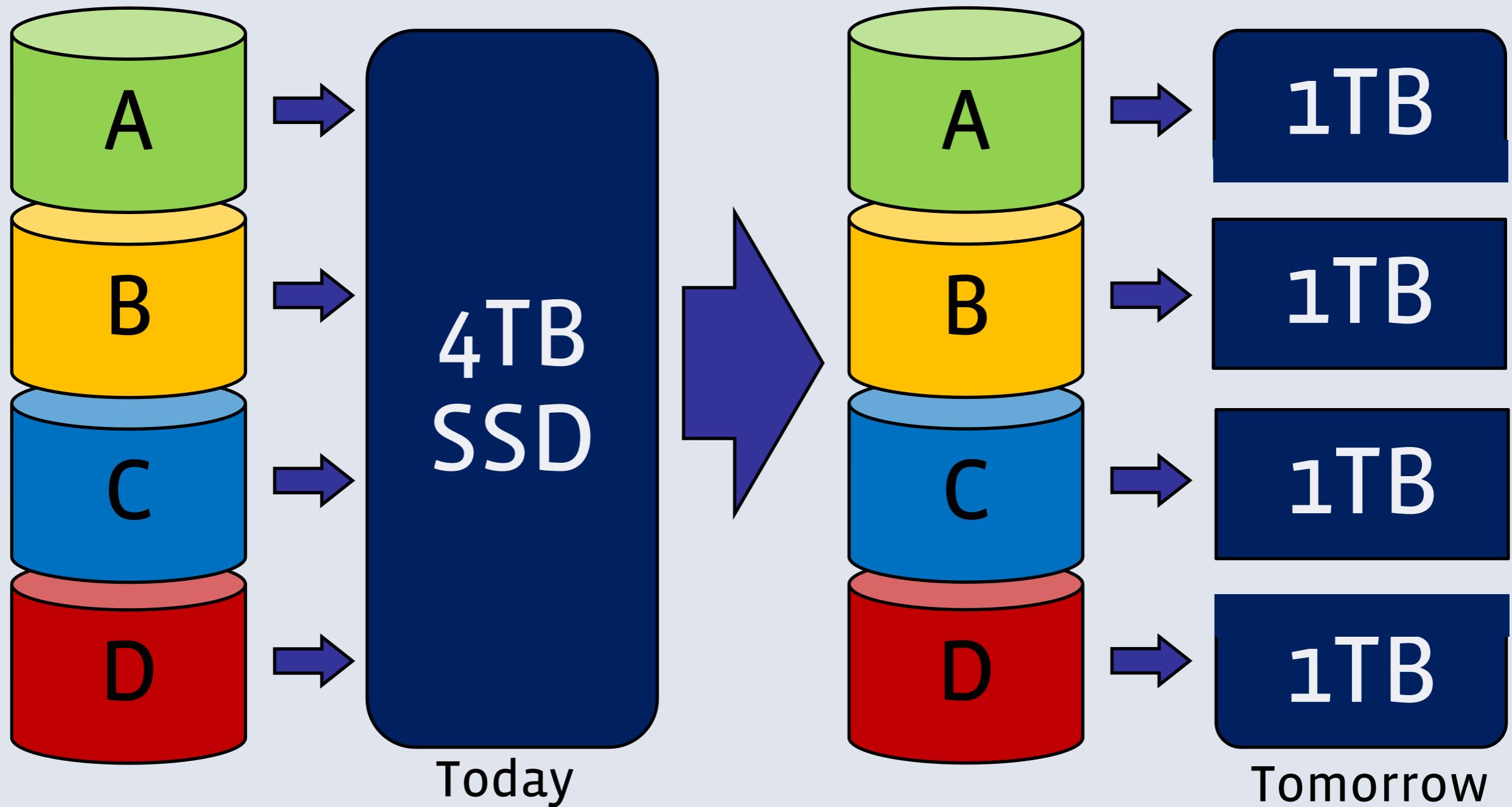
STORAGE SUPPLIER

IN CONSIDERATION of the matters described above and of mutual benefits and obligations set forth in this Agreement, the receipt and sufficiency of which consideration is hereby acknowledged, the Client and the Supplier (individually the "Party" and the collectively the "Parties" to this Agreement) agree as follows:

Term of Agreement
The term of the Agreement (the "Term") will begin on the date of this Agreement and will remain in force and effect until the completion of the Services, subject to earlier termination as provided in this Agreement.

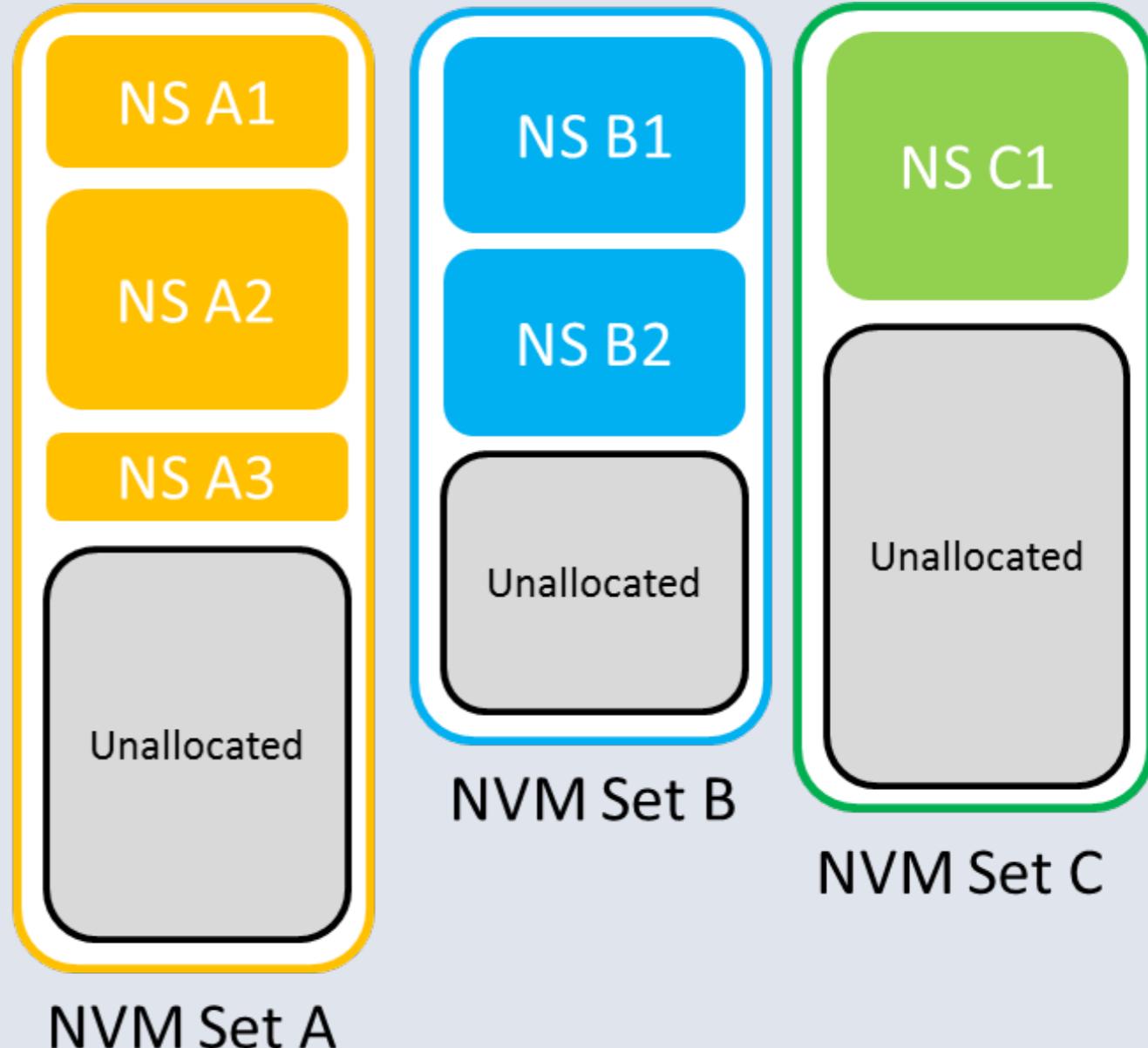
RENEGOVIALE

Quality of Service (QoS) Regions



NVM Sets

- New NVMe feature called:
I/O Determinism
- SSD is configured as multiple
NVM Sets (e.g. A, B, C)
- NVM Sets are QoS isolated regions
 - A write to Set A does not impact a read to Set B or C
- One or more namespaces are allocated from an NVM Set



Scheduling I/O

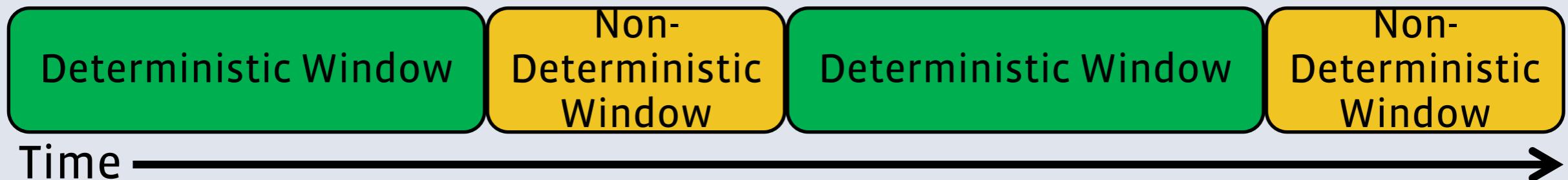
Predictable Latency mode = Two windows of time

1. Deterministic window

- Host issues only reads
- Drive does no background operations

2. Non-deterministic window

- Host can issue writes and TRIMs
- Host can issue reads but no latency guarantees
- Drive does background operations

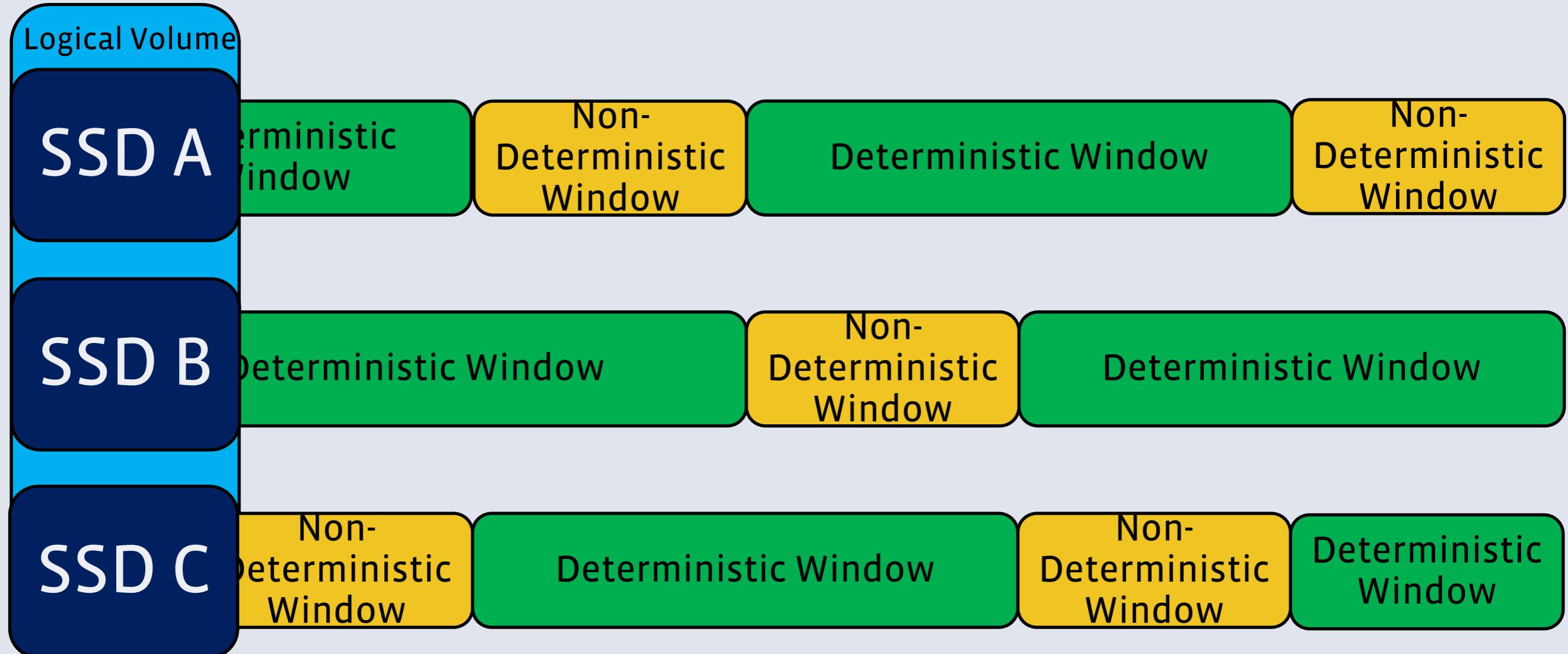


Reliable Estimates

- Users need **reliable estimates** of when background operations are required
- **Reliable** means a sufficiently accurate prediction (e.g. +/- 5%, but not 50%)
- Estimates are # of reads, # of writes, and time



Building a Solution



With predictable latency mode, reliable estimates, and a well-behaved host, it is possible to achieve excellent 99.99%+ read quality of service!!

Read Recovery Level

- Data Centers keep multiple copies of data (often erasure coded)
- Given the replication, there are tradeoff possibilities

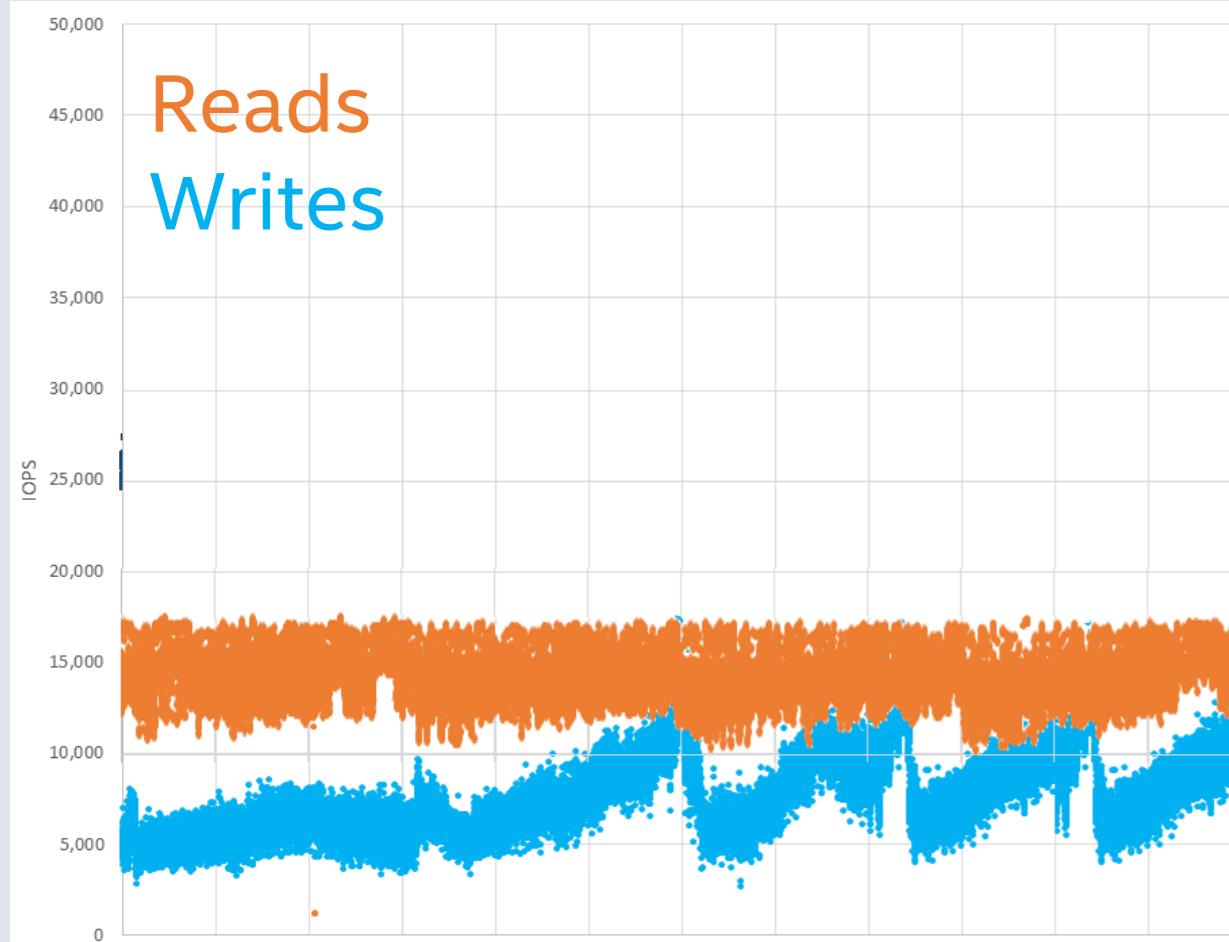
Read Recovery Level	99.99% Read Latency @ Queue Depth 1	Unrecoverable Bit Error Rate (UBER)
0 – “Fail Fast”	200 us	1e-14
1	400us	1e-16
2	1ms	1e-17

Contract Options: Quality of Service vs. Heroic Error Recovery

Challenges and solutions

- ✓ “Noisy Neighbor” => QoS isolated regions
- ✓ “Read collisions” => Predictable Latency Mode
- ✓ “Error handling outliers” => Read Recovery Level

Theoretically it should look like



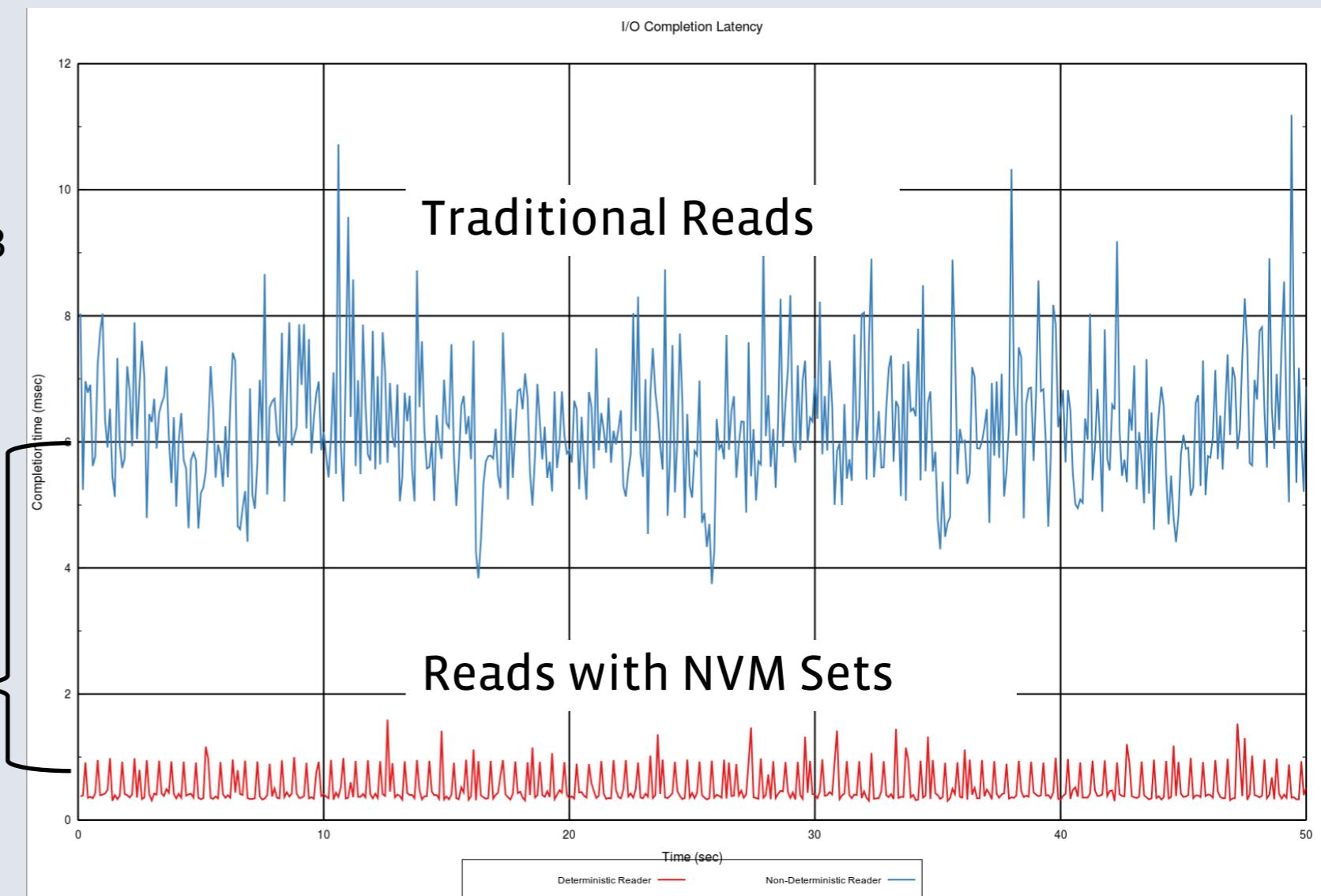
Theory: Increased read IOPs and tighter QoS distribution

Early Results meet Theory

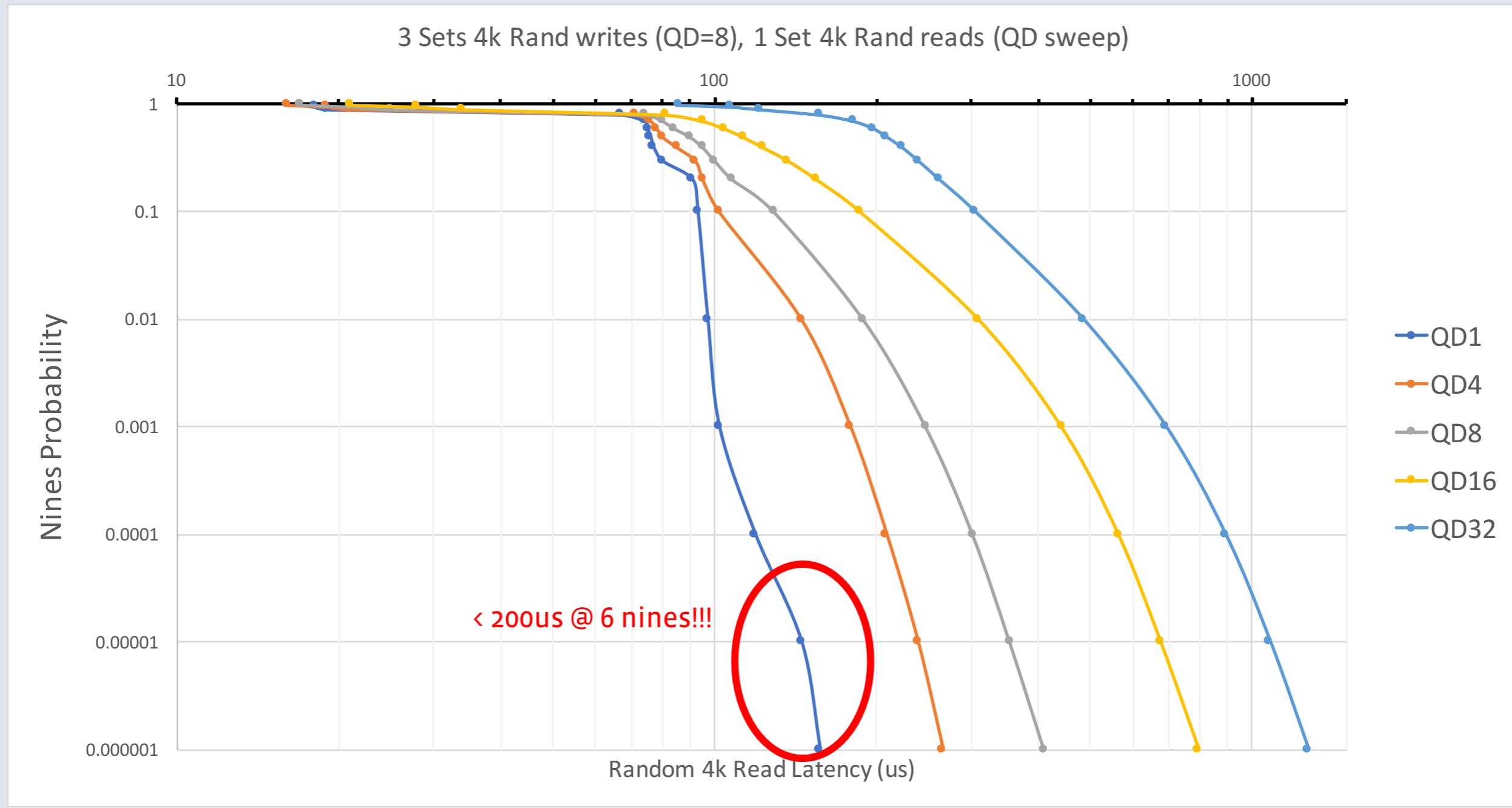
- Prototype with 2 NVM Sets
 - Tested with and without Sets
 - 4k Random Reads @ QD=8 to Set A
 - 4k Random Writes @ QD=8 to Set B

> 3X improvement in
max latency and MUCH
tighter variance

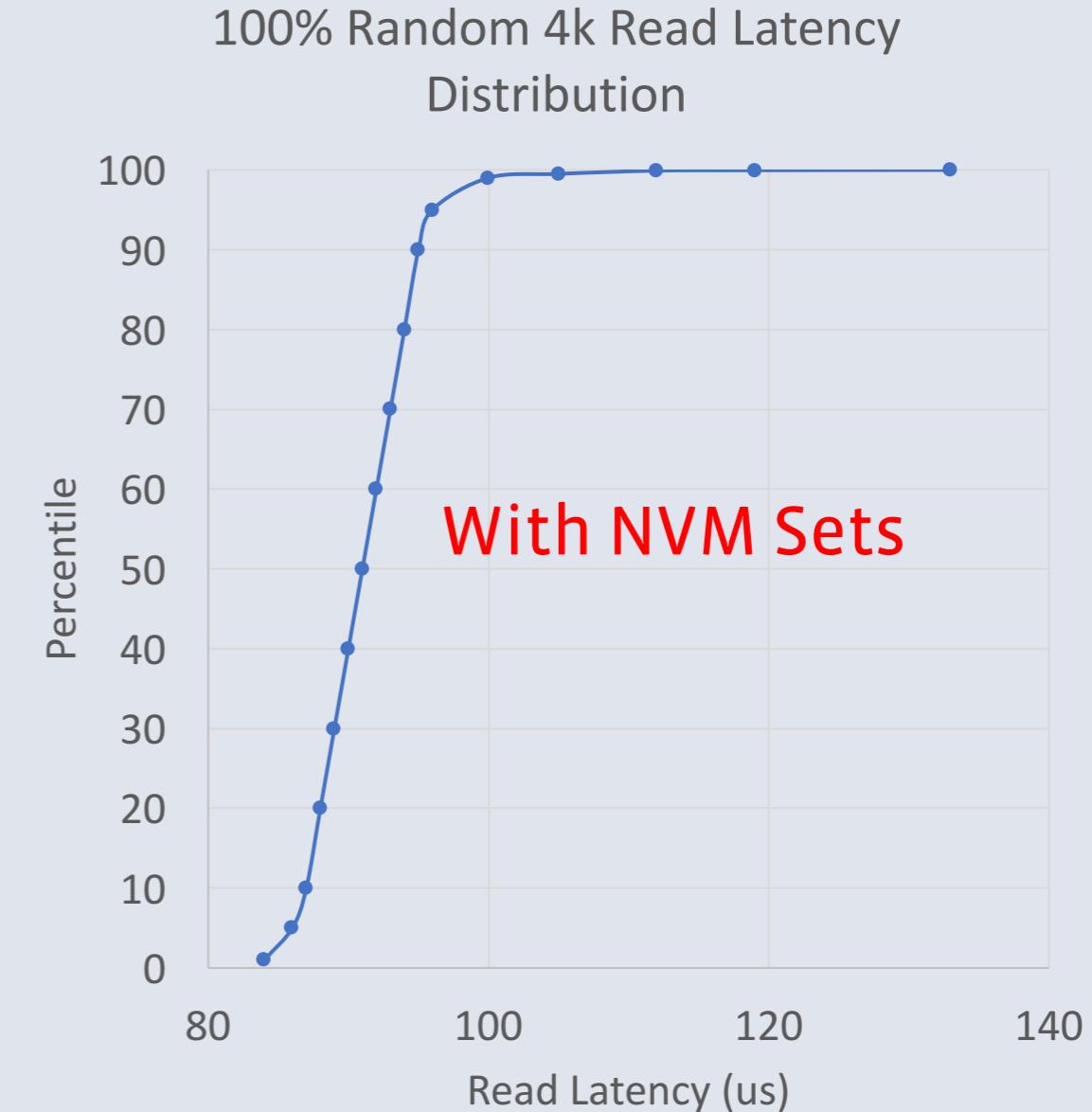
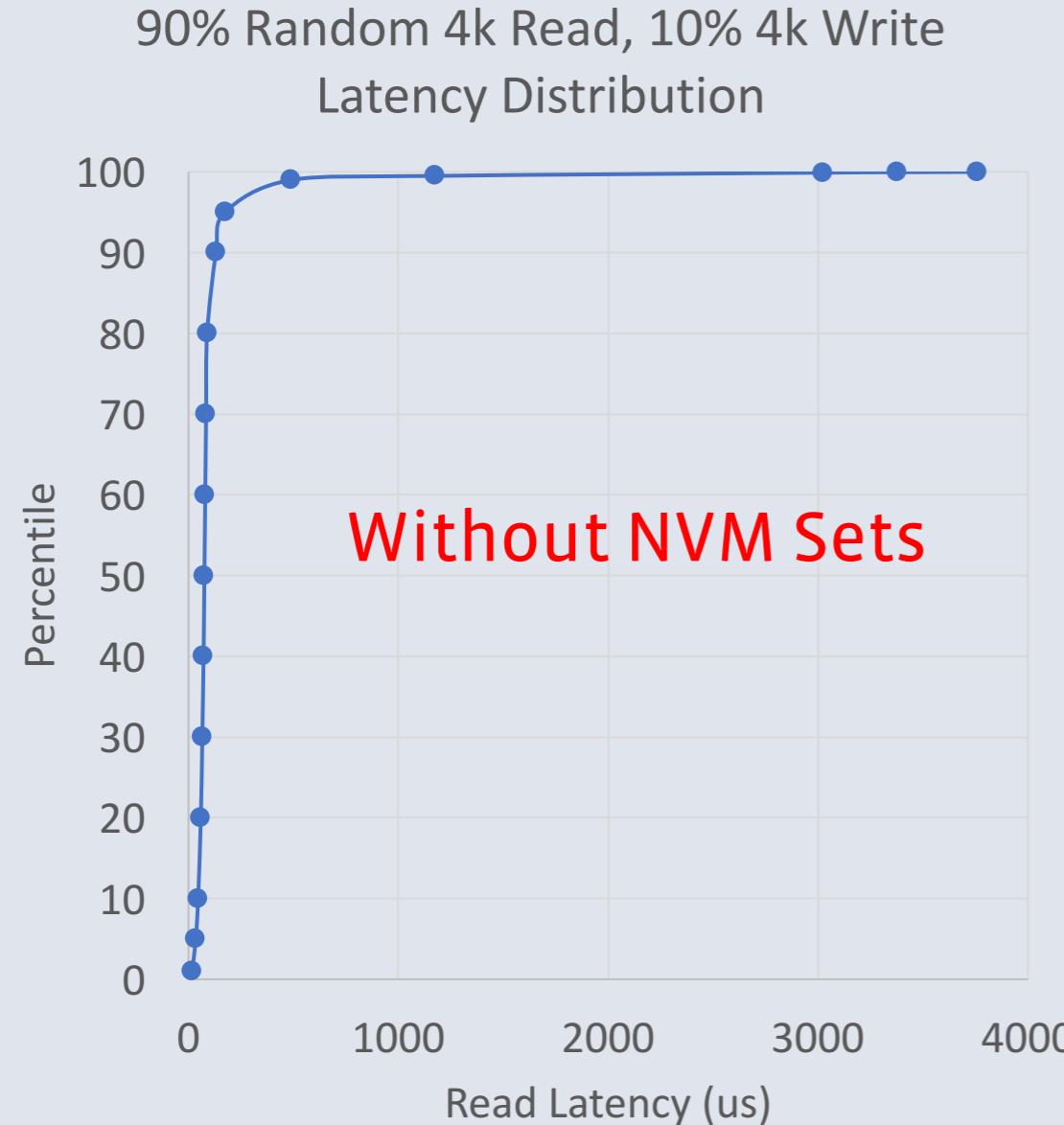
Max Completion Latency
(Measured in 100ms intervals)



Demonstrating QoS Isolation

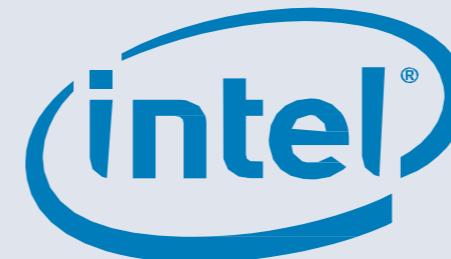


Revisit the read latency challenge



Conclusion

- Ratified NVMe specification will be available very soon!
- NVM Sets, Predictable Latency Mode, and Read Recovery Levels can each be implemented separately to fit **YOUR** use cases
- See the following booths for more information:



SAMSUNG

TOSHIBA

facebook

Thank You!