facebook

Enhancing SSD Control of NVMe Devices for Hyperscale Applications

Luca Bert - Seagate

Chris Petersen - Facebook

Agenda

- Introduction & overview (Luca)
- Problem statement & proposed solution (Chris)
- SSD implication and proto work (Luca)
- Test results and analysis (All)
- Conclusion (All)

facebook SEAGATE

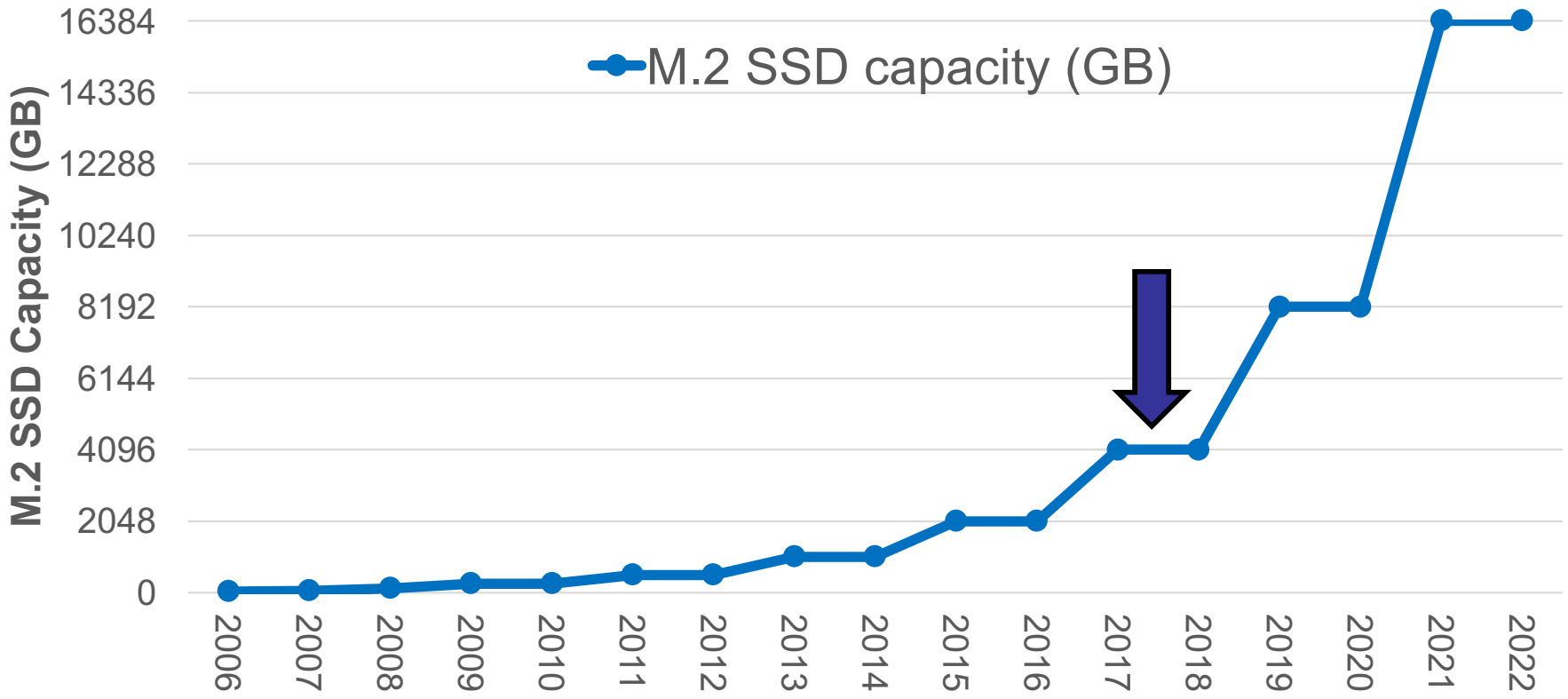
facebook

SEAGATE

Problem Statement & Proposed Solution

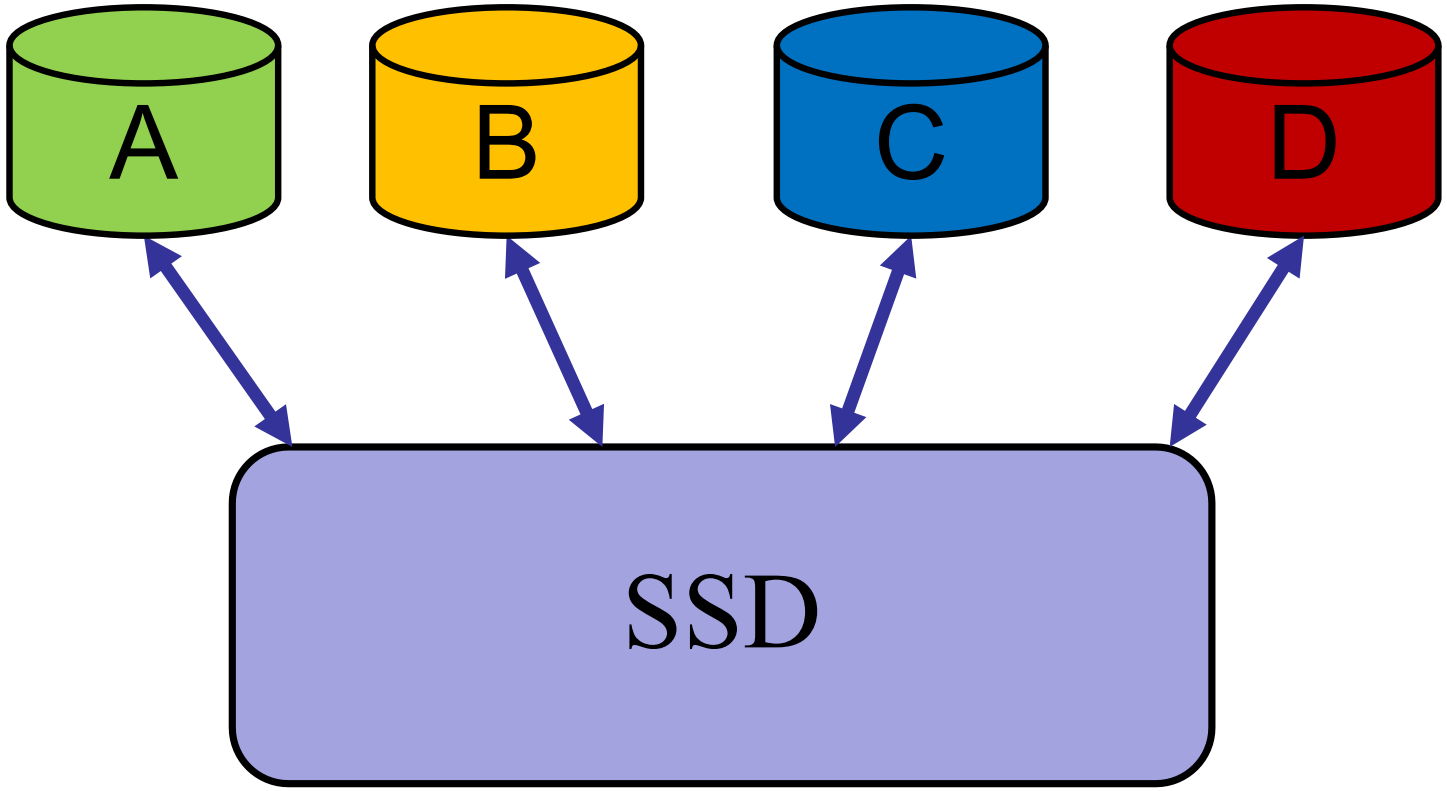
Flash Memory Summit

NAND Flash Trends



facebook SEAGATE

SSD Capacity = Shared Resource



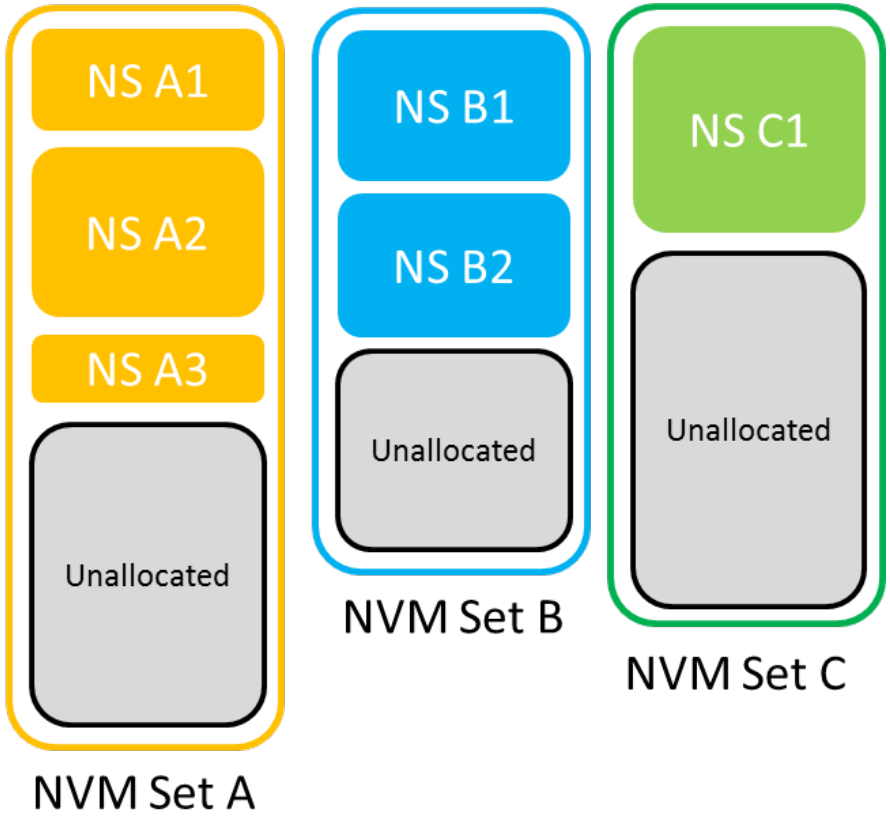
facebook SEAGATE

Noisy Neighbors

facebook SEAGATE

Introducing I/O Determinism

- 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



facebook SEAGATE

facebook

SSD Implications and Prototype Work

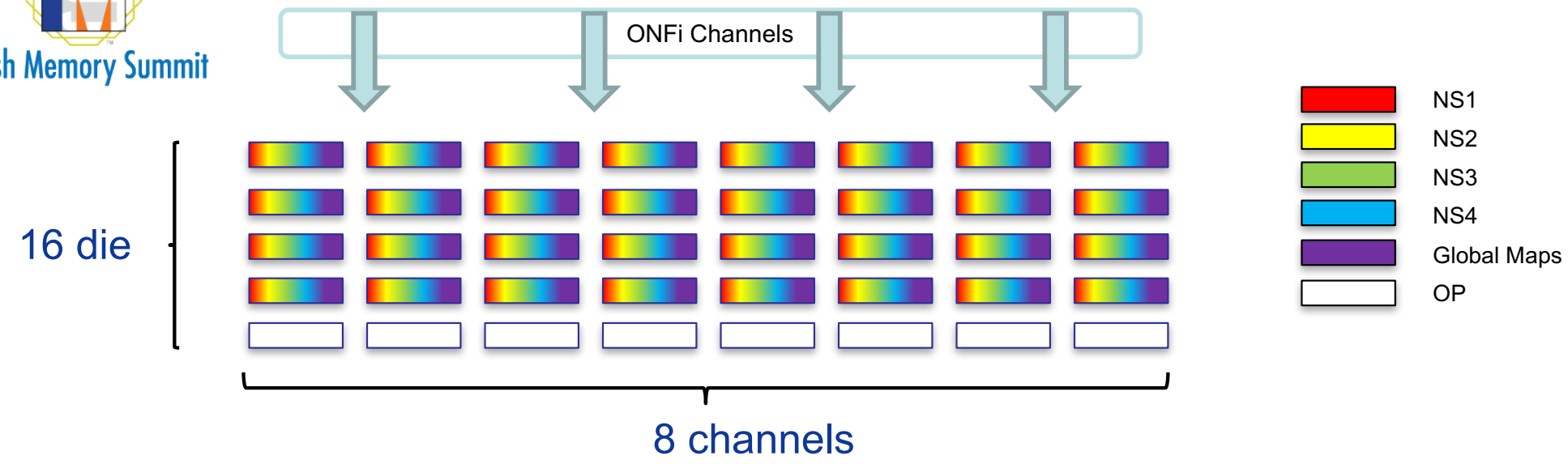
SSD Data Layout Impact

Key Hurdles to Address

- Physical resource segregation
  - Die have to be allocated to VM Sets
- "Localization of the Globals"
  - **Data**: Maps, Journals, Logs, Metadata… are global and need to be redesigned to be local to VM Sets
  - **Processes**: Same for GC and other background processes
  - **Resources**: OP, cache, memory... can be local or global

facebook SEAGATE

Baseline: Conventional SSD with 4 NS



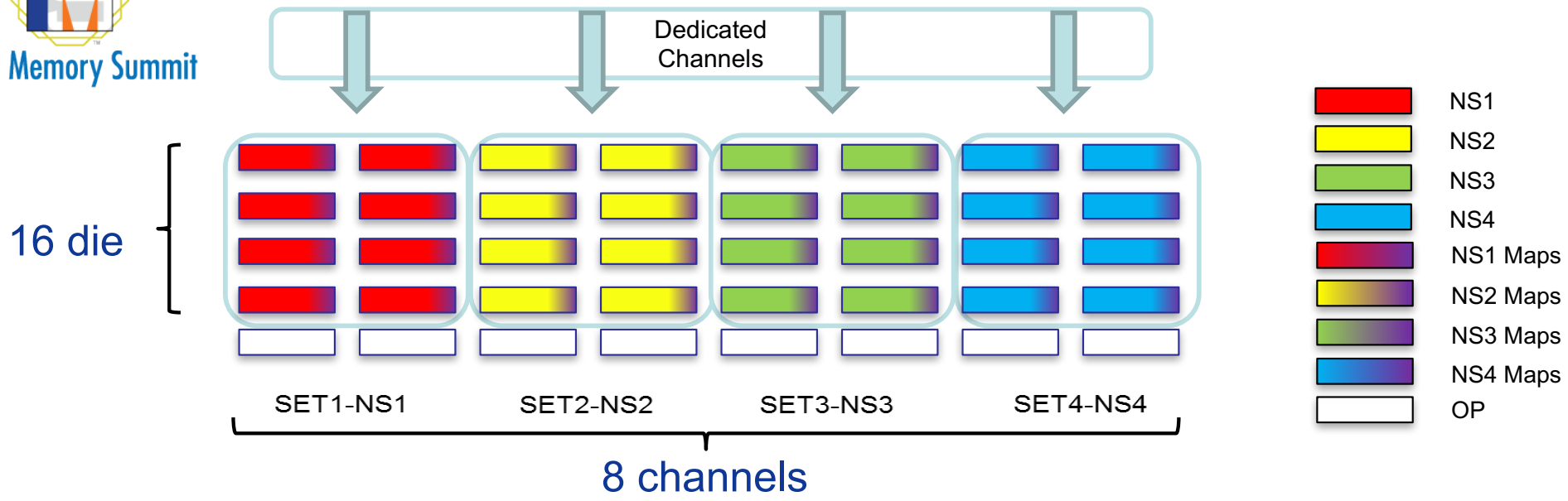
Pros

- Simplest configuration
- No logical – physical affiliation
- Most balanced model

Cons

- IO pattern conflicts

IOD: Vertical Sets (VS)



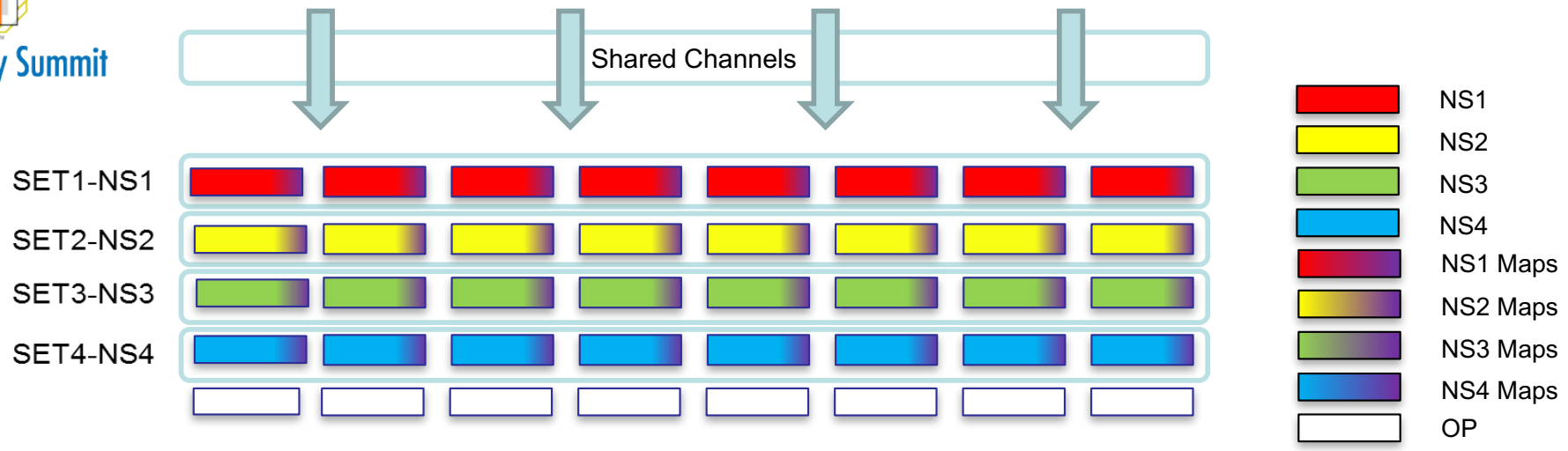
Pros

- Few dedicated channels for each Set
- No conflict in channel or Die
- Every Set runs in parallel resources

Cons

- 1/4 BW possible per Set (limited by number of dedicated channels)

IOD: Horizontal Sets (HS)



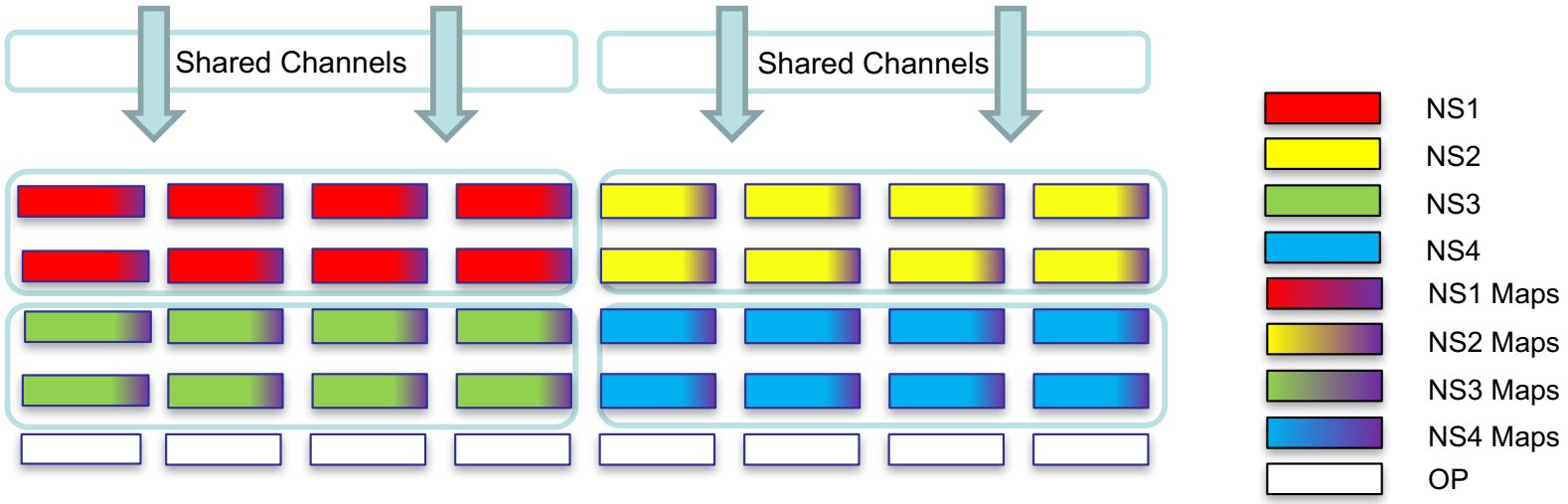
Pros

- All channels are utilized by all Sets
- Full BW possible with single active Set (or shared BW with all Sets)
- No Conflict in Die

Cons

- No Dedicated channels or queues for Set

IOD: Mixed Sets (MS)



Pros

- Few dedicated channels for group of Sets
- No conflict in Die
- Groups of Sets runs in parallel resources

Cons

- 1/2 of the BW possible per Set (limited by number of dedicated/shared channels)

Areas Not Covered in Proto Work

- R and W path are HW assisted
  - No complete separation between them
  - Conflicts do happen
- Cache memory is shared across VS
  - Conflicts happen but limited by test nature
- R and W have same priority
  - No priority to any specific IO
- No Erase Suspend

facebook SEAGATE

facebook

SEAGATE

Test Results and Analysis

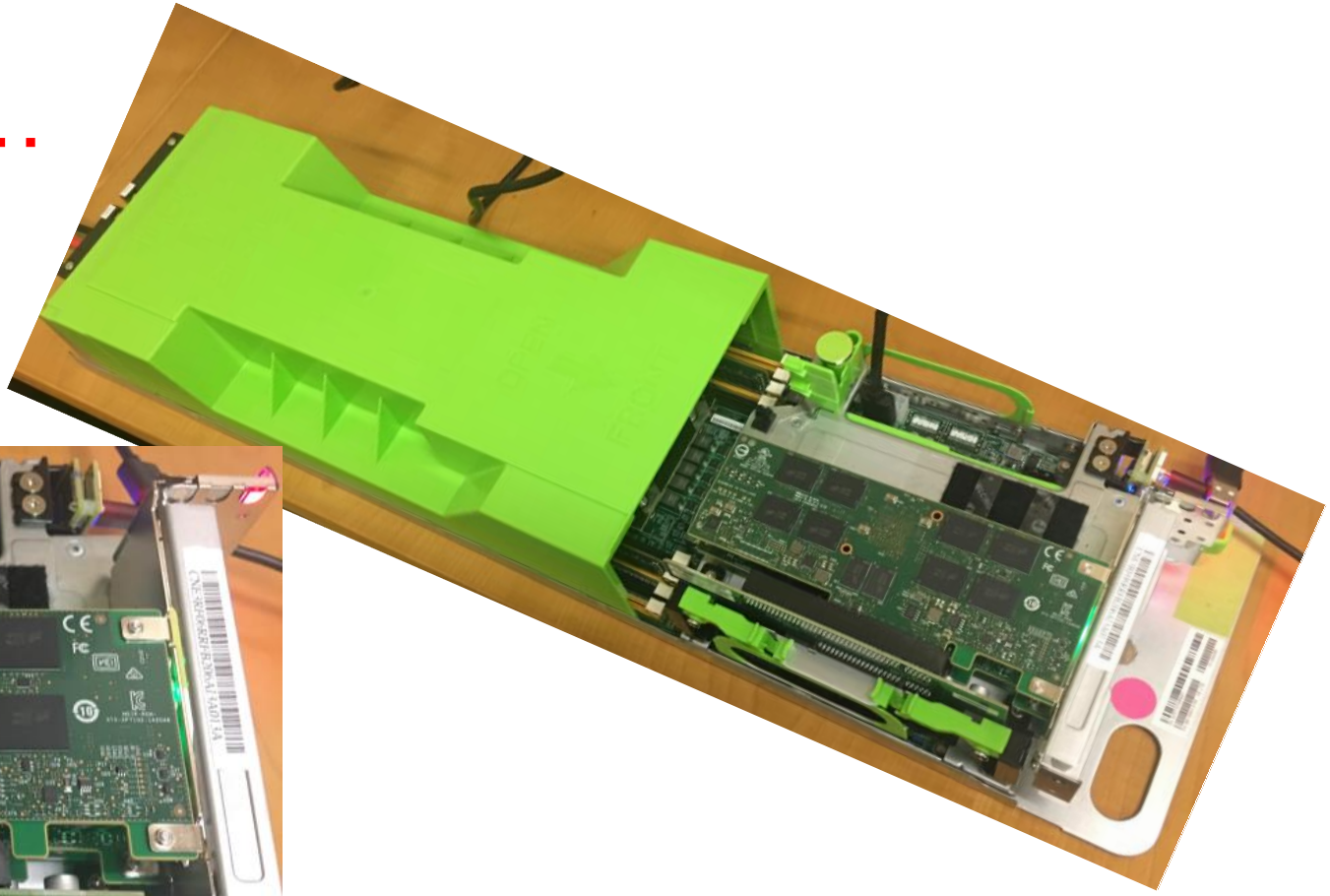
Configuration Under Test

- Facebook OCP Leopard System
  - 2-socket Intel® Xeon® v4 CPU
  - 32GB DRAM
- STX Nytro XP7102 (2TB eMLC),
  - 128 dies on 8 channels, ONFi @400 MT/s
- Tested with Baseline (No IOD), VSets and Hsets
- Writes = 32KB SW @QD4
- Reads = 4KB RR @QD8

facebook SEAGATE

Flash Memory Summit

PICS…



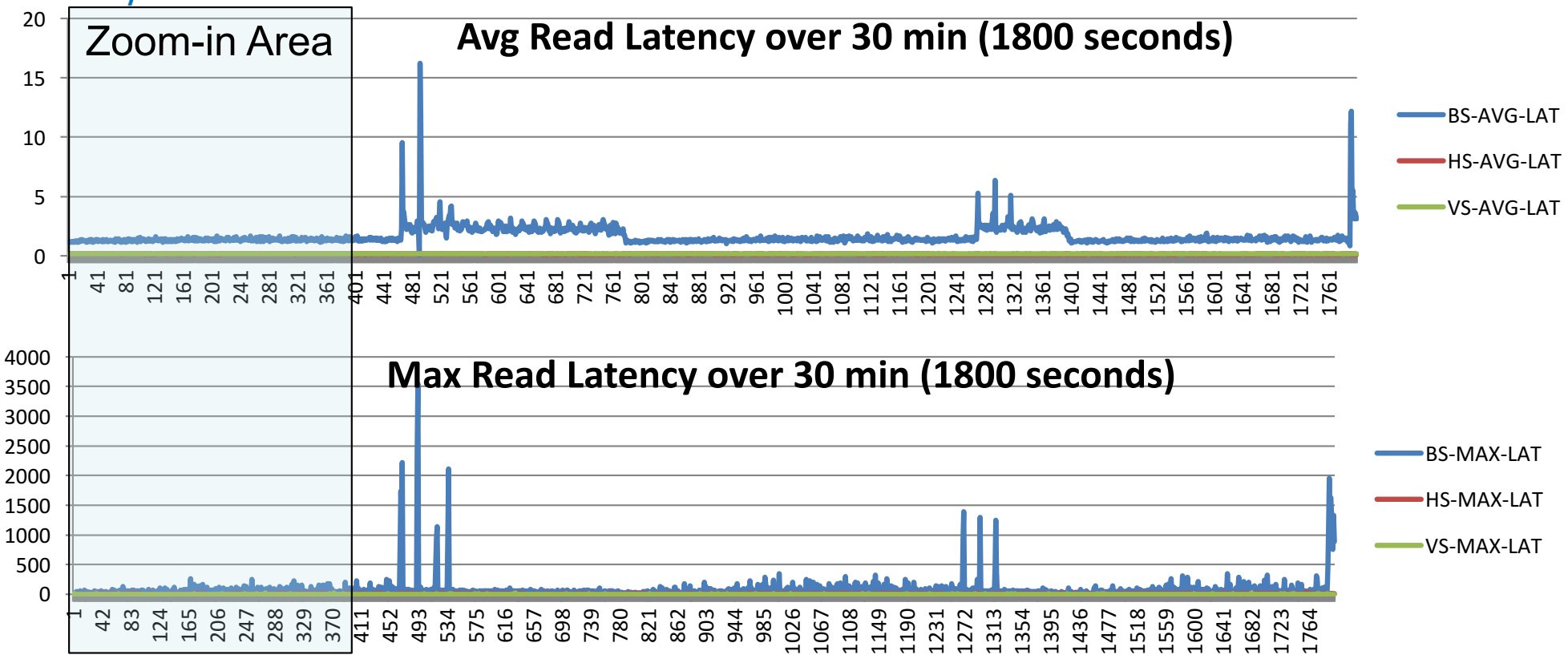
facebook SEAGATE

"The Noisy Neighbor"

- One Set under heavy writes, the other under heavy reads
  - Write is 32K Seq Write with QD=4
  - Read is 4K Random Read with QD=8
- Expected Outcome: minimal to no impact on Reads from the "noisy" Writes

facebook SEAGATE

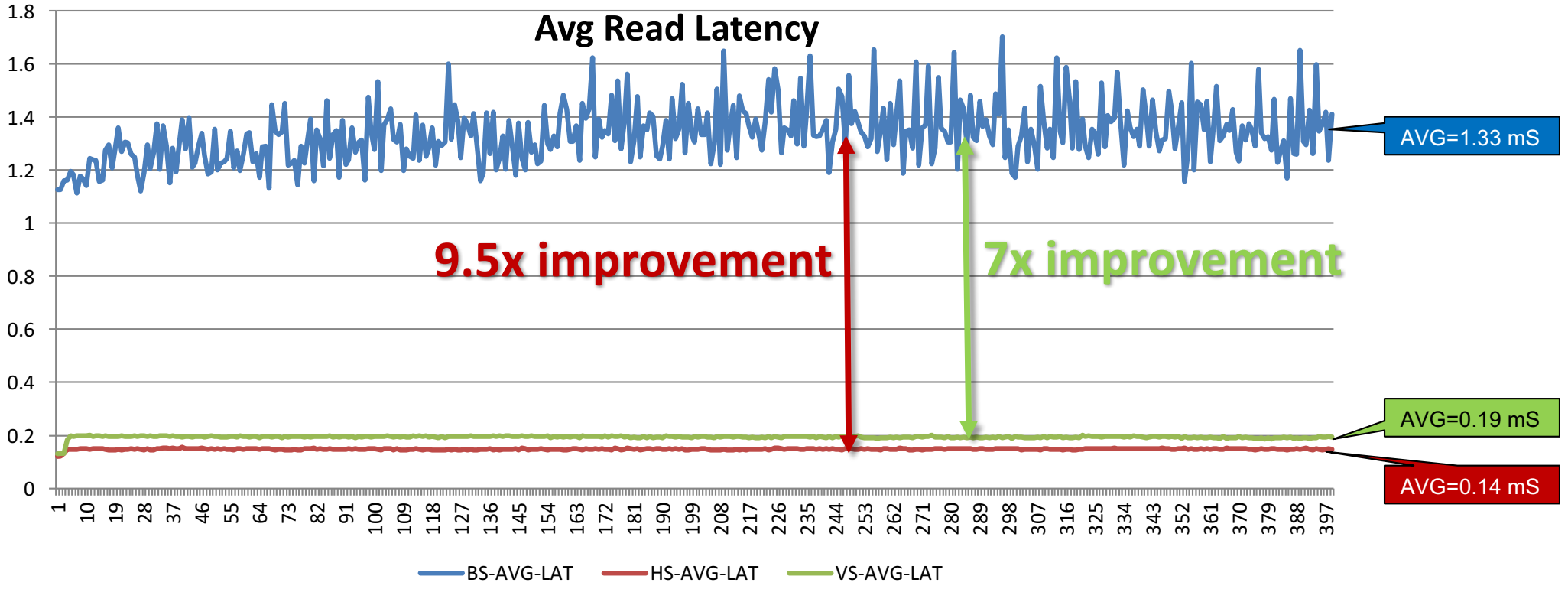
WR on NS1 + RD on NS2



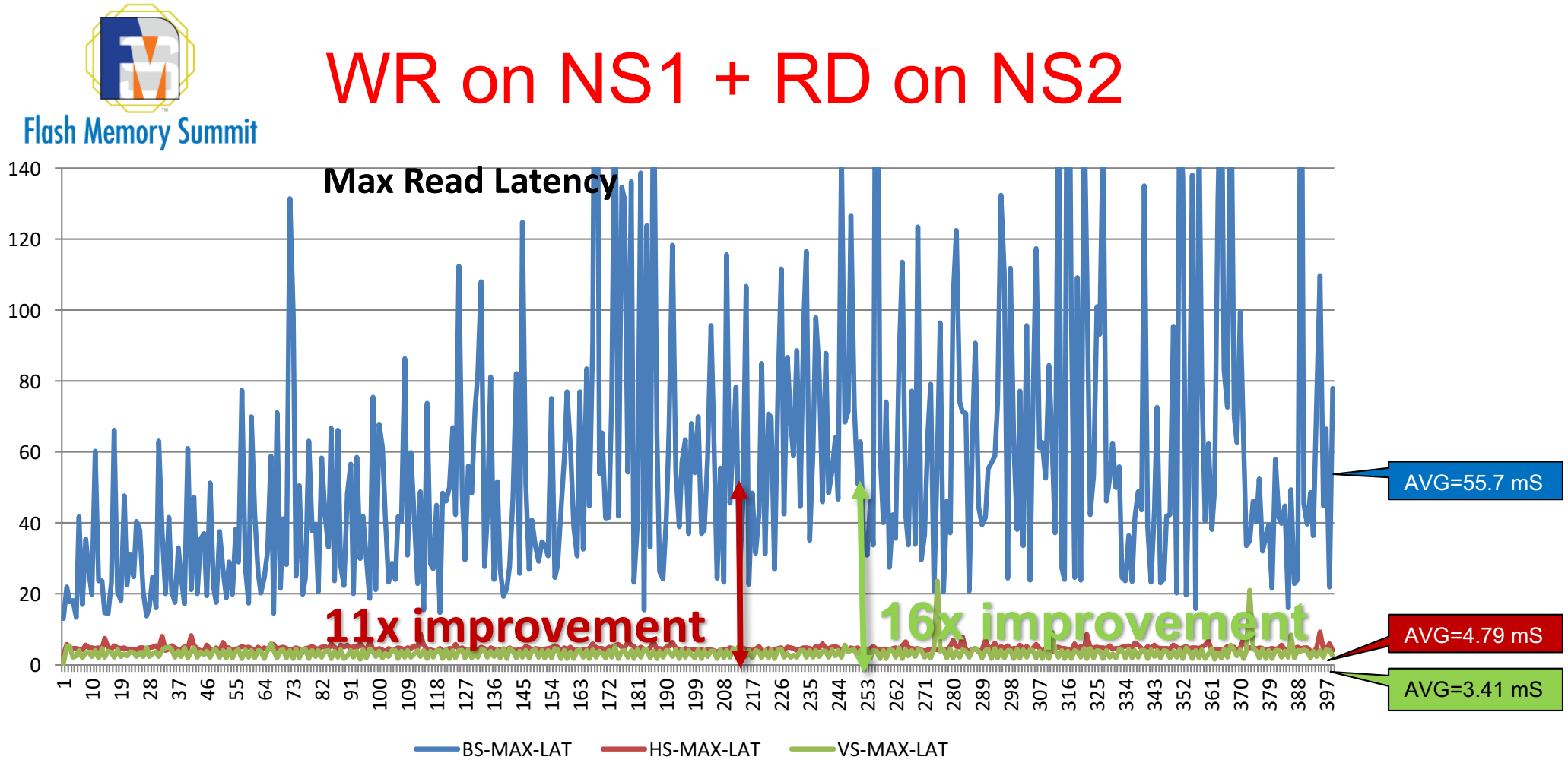
facebook SEAGATE

Flash Memory Summit

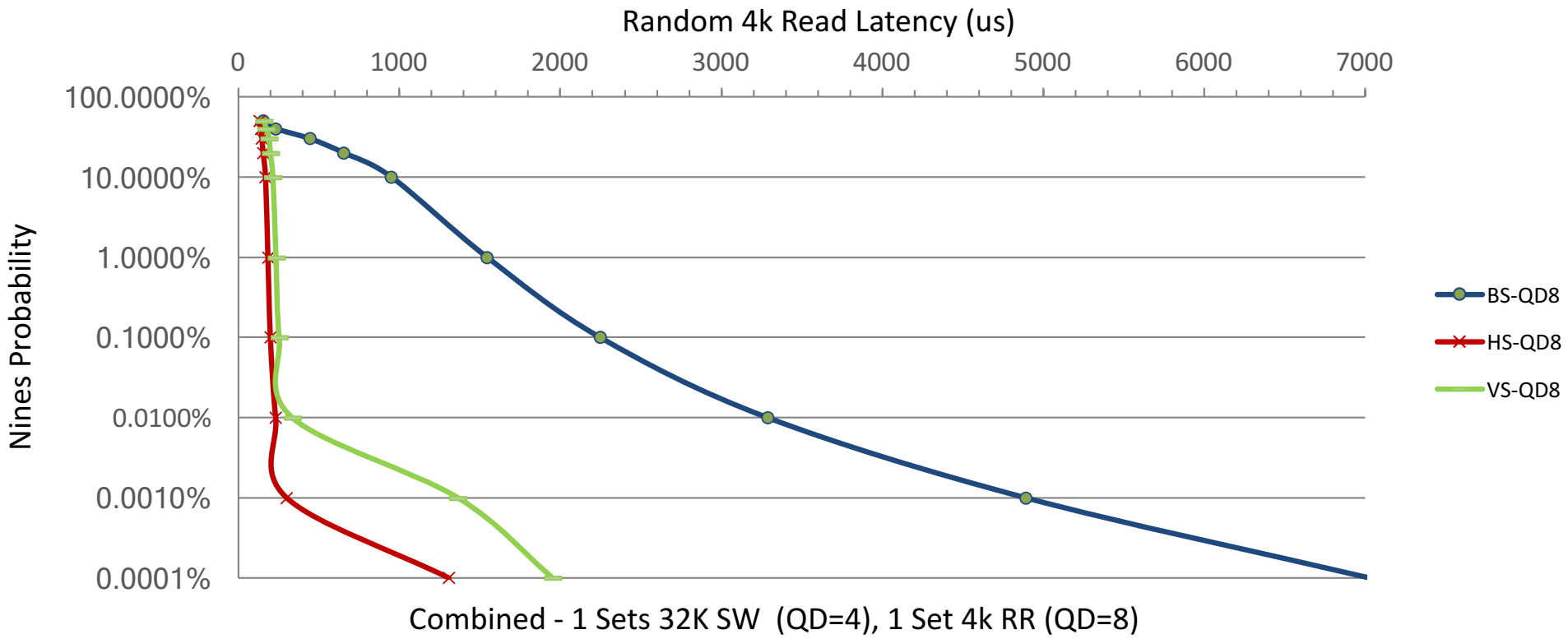
WR on NS1 + RD on NS2



facebook SEAGATE

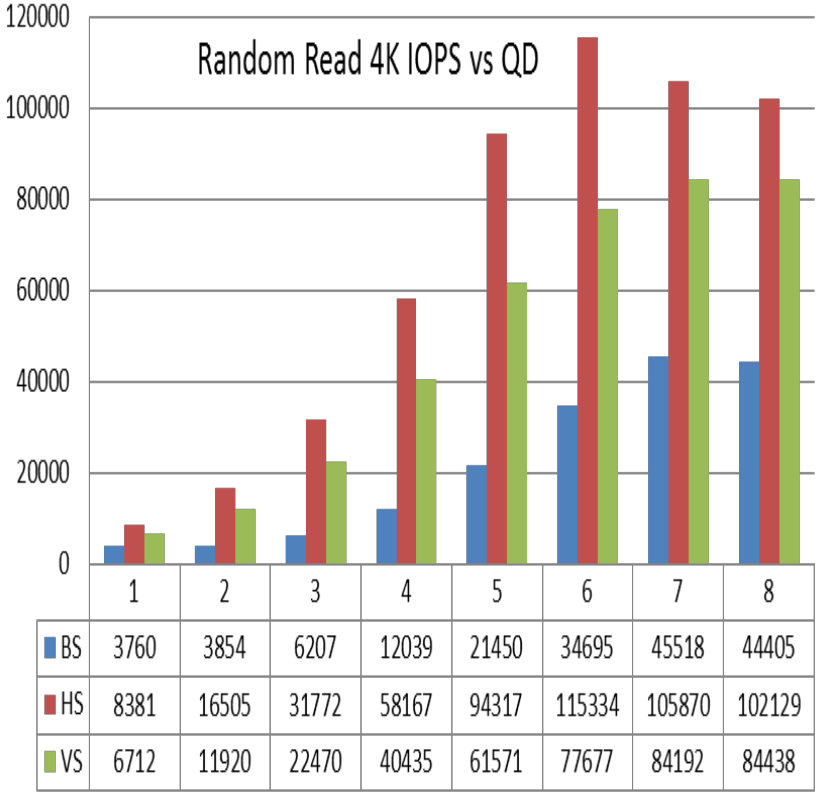

Flash Memory Summit
WR on NS1 + RD on NS2
Max Read Latency
140
120
100
80
60
40
20
0
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
226
235
244
253
262
271
280
289
298
307
316
325
334
343
352
361
370
379
388
397
11x improvement
16x improvement
AVG=55.7 mS
AVG=4.79 mS
AVG=3.41 mS
BS-MAX-LAT
HS-MAX-LAT
VS-MAX-LAT
facebook
SEAGATE

WR on NS1 + RD on NS2

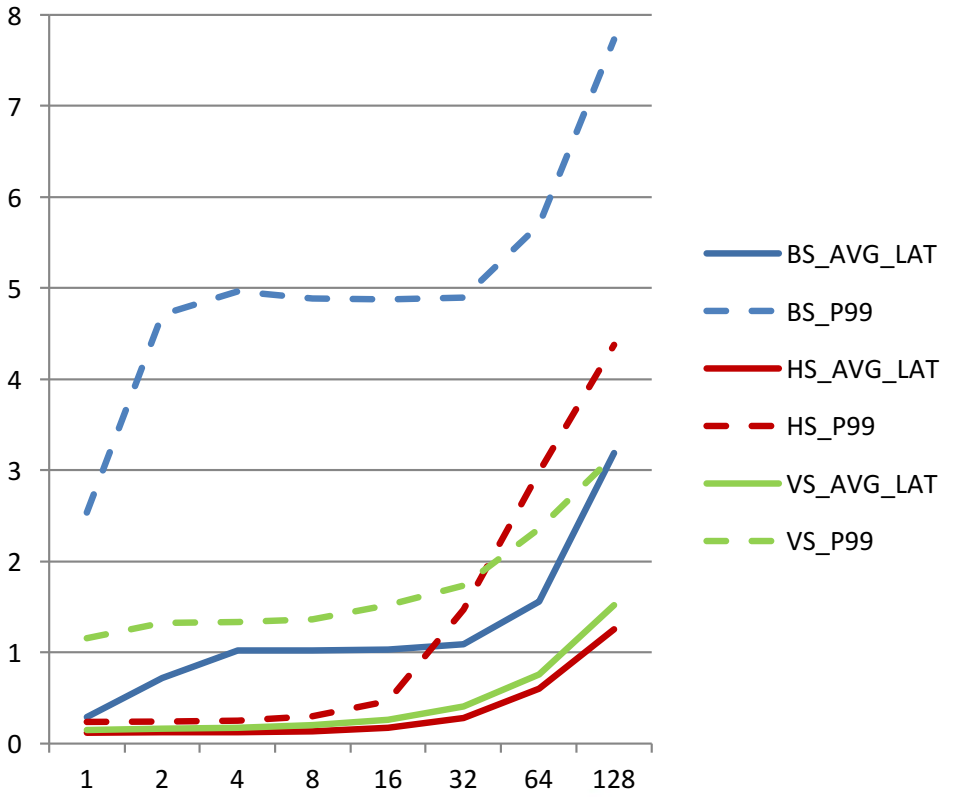


facebook SEAGATE

WR on NS1 + RD on NS2



	1	2	3	4	5	6	7	8
BS	3760	3854	6207	12039	21450	34695	45518	44405
HS	8381	16505	31772	58167	94317	115334	105870	102129
VS	6712	11920	22470	40435	61571	77677	84192	84438



facebook SEAGATE

facebook

Conclusions

Conclusions

- Average values: IOD is behaving largely as expected
- Max values: there is more under the hood to address but even proto show huge advantages
- IOD is all about die separation but Controller, DRAM, ONFi, and other resources play key roles under specific conditions
- Most of all… improvements are so high that there may be other usage models outside Hyperscalers

facebook SEAGATE

facebook

SEAGATE

Visit Seagate Booth #505

Learn about Seagate's portfolio of SSDs, flash solutions and system level products for every segment.

www.seagate.com/Nytro