WHEN THE RULES CHANGE

Next Generation Oracle Database Architectures using Super-fast Storage

James Morle, EMC DSSD

DSSD

INTRO

THAR BE DRAGONS

perotsystems®

scale abilities

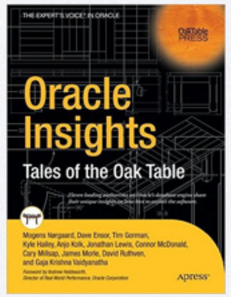
DSSD

1993

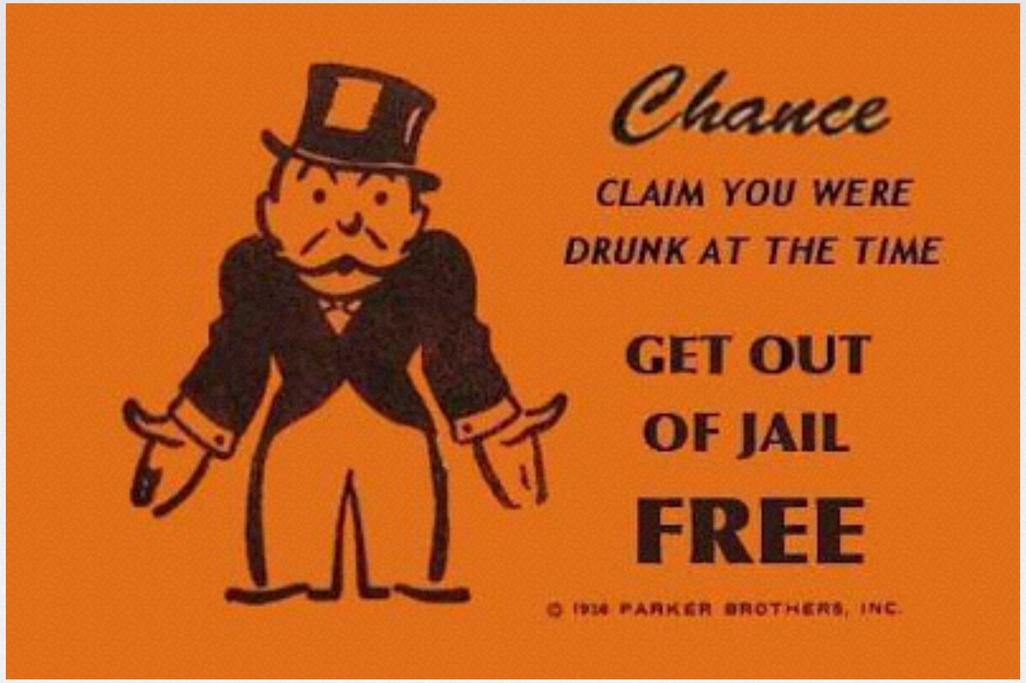
2001

2010

2015



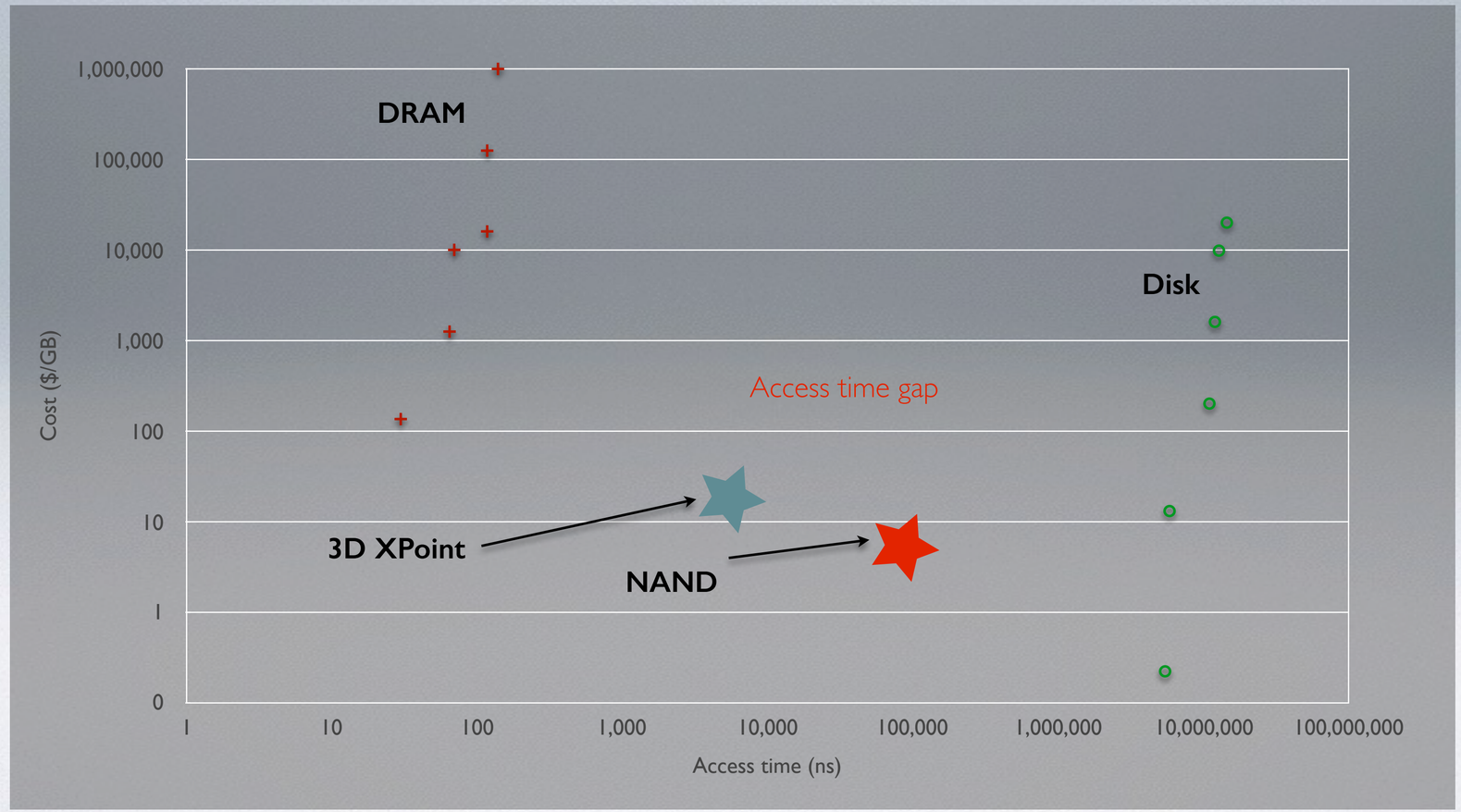
founded in



Disclaimer: I work for EMC these days, and use some of the corporate content, but all opinions here are my own - this is not an official company presentation.

"I/O certainly has been lagging in the last decade"
- Seymour Cray 1976

THE ACCESS TIME GAP



"Bandwidth is the work of man,
latency is the realm of <insert deity here>"

Jeff Bonwick, CTO and Founder, DSSD

DSSD

BIG PIPES ARE EASY


PERFORMANCE
FEAR IS THE MIND
LATENCY
KILLER

WHAT MATTERS WITH ORACLE WORKLOADS?

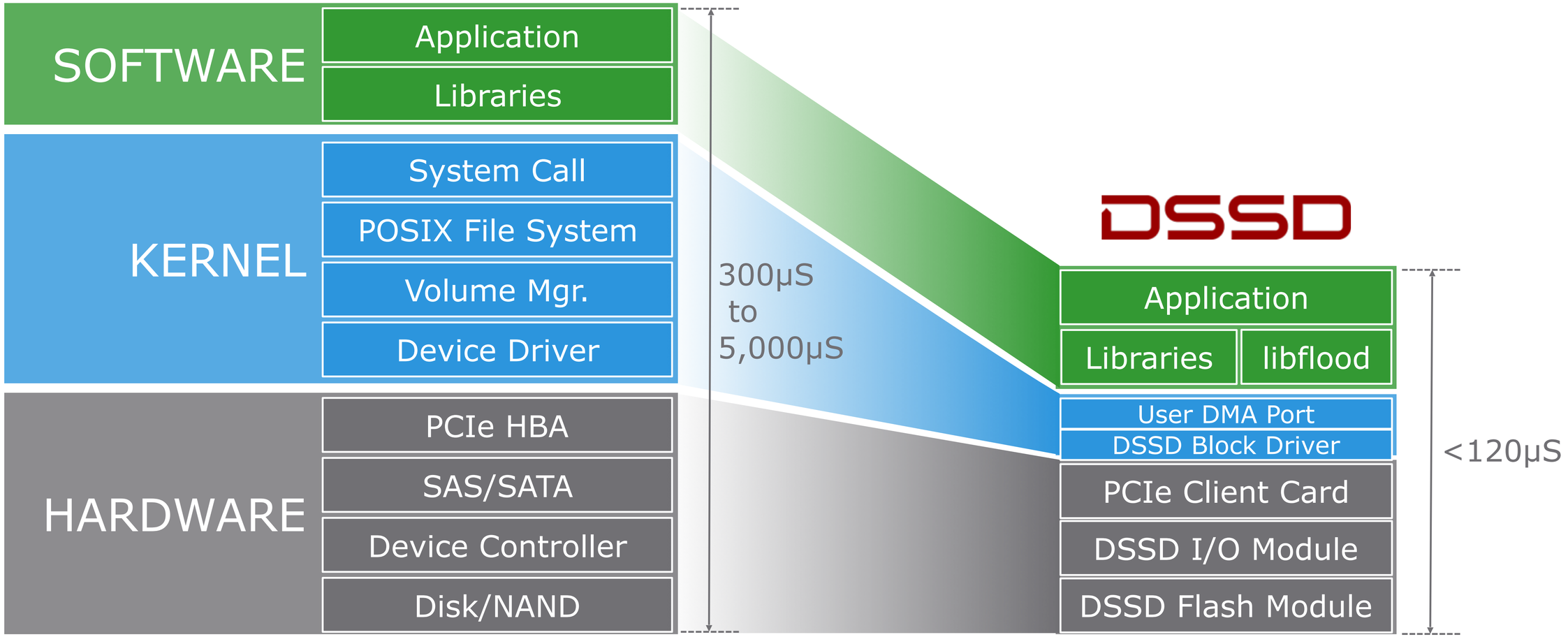
- **DW/BI Workloads:**
  - Multiblock read *bandwidth*
  - Sequential write *bandwidth* ***and*** *latency*
- **OLTP Workloads:**
  - Single block read *latency*
  - Sequential write *latency*

SO WHAT'S THE PROBLEM?

- Delivery of low latency I/O requires low latency **transport** in addition to low latency **media**
- We have the media, currently NAND flash, but…
- Fibre Channel often adds up to 200 microseconds of latency
- This needs something new, and fit for purpose… let's start with the software

DSSD Block Device Access to DSSD

A bit more latency due to kernel overhead



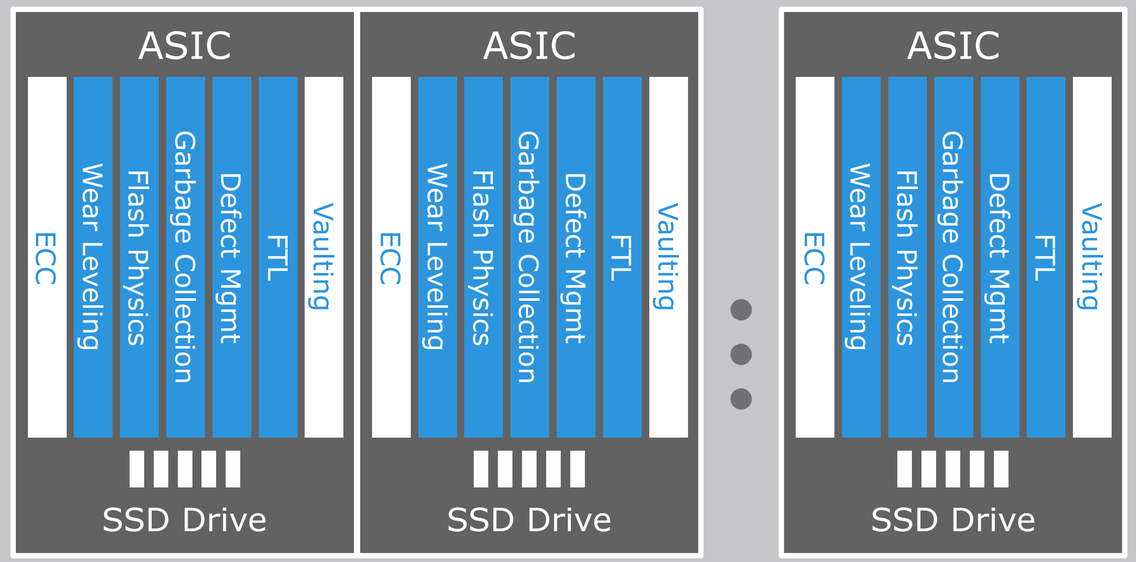
DSSD

EMC²

DSSD FM VS OTHER FLASH STORAGE

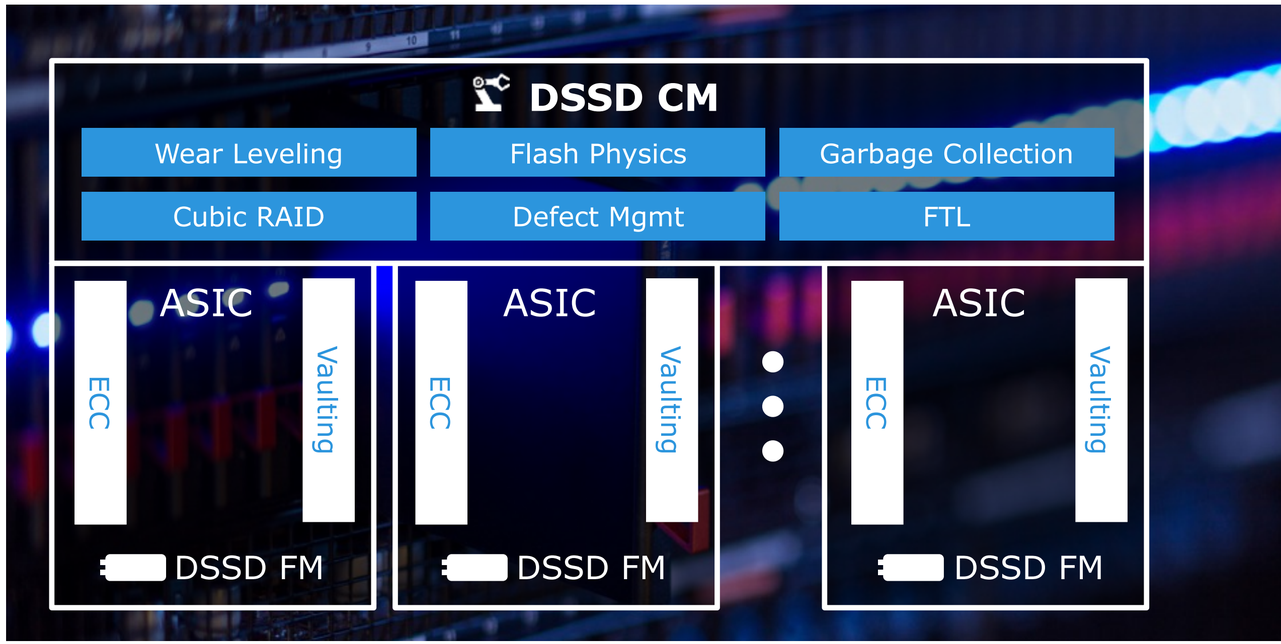
Simpler and Faster Flash Modules

Standard Flash Devices



- Complex firmware, limited power
- Independently managed media

DSSD D5

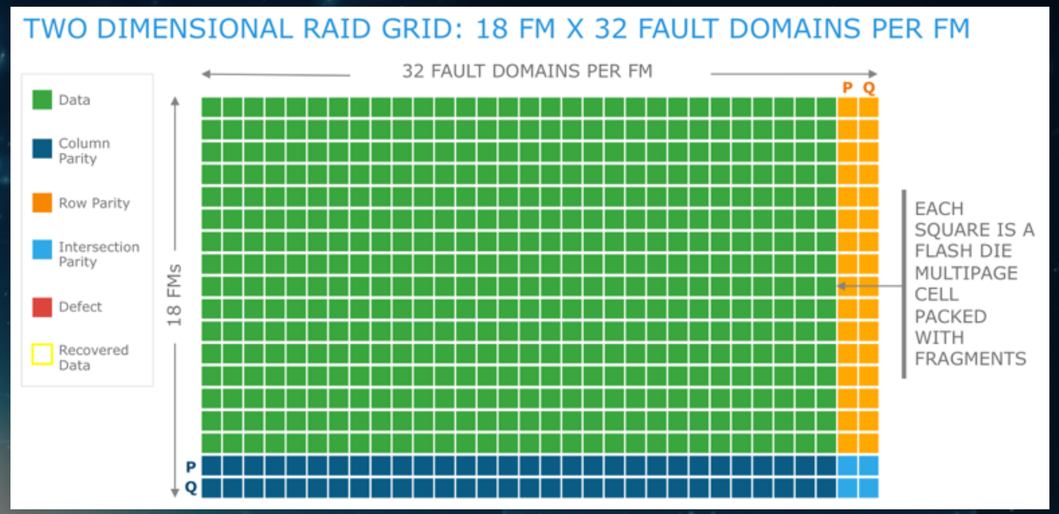


- DSSD has simple, fast Flash Modules
- Control Module with rich resources implements advanced global

HARDWARE + SOFTWARE RESILIENCE

Always On Cubic Raid

- Cubic RAID has 2x greater reliability of other RAID but has the same overhead (17%)
- Cubic RAID Grid is an interlocked, multi-dimensional array of multi-page “cells” of NAND die
- High performance – always on



System Wide Data Protection

Dense and Shared Flash

DSSD **D5** - 5U RACK SCALE FLASH PLATFORM

FLASH AND CMs

36 Flash Modules (FMs)
18 Flash Modules when Half Populated

2TB/4TB Flash Modules today
Larger FMs on the roadmap

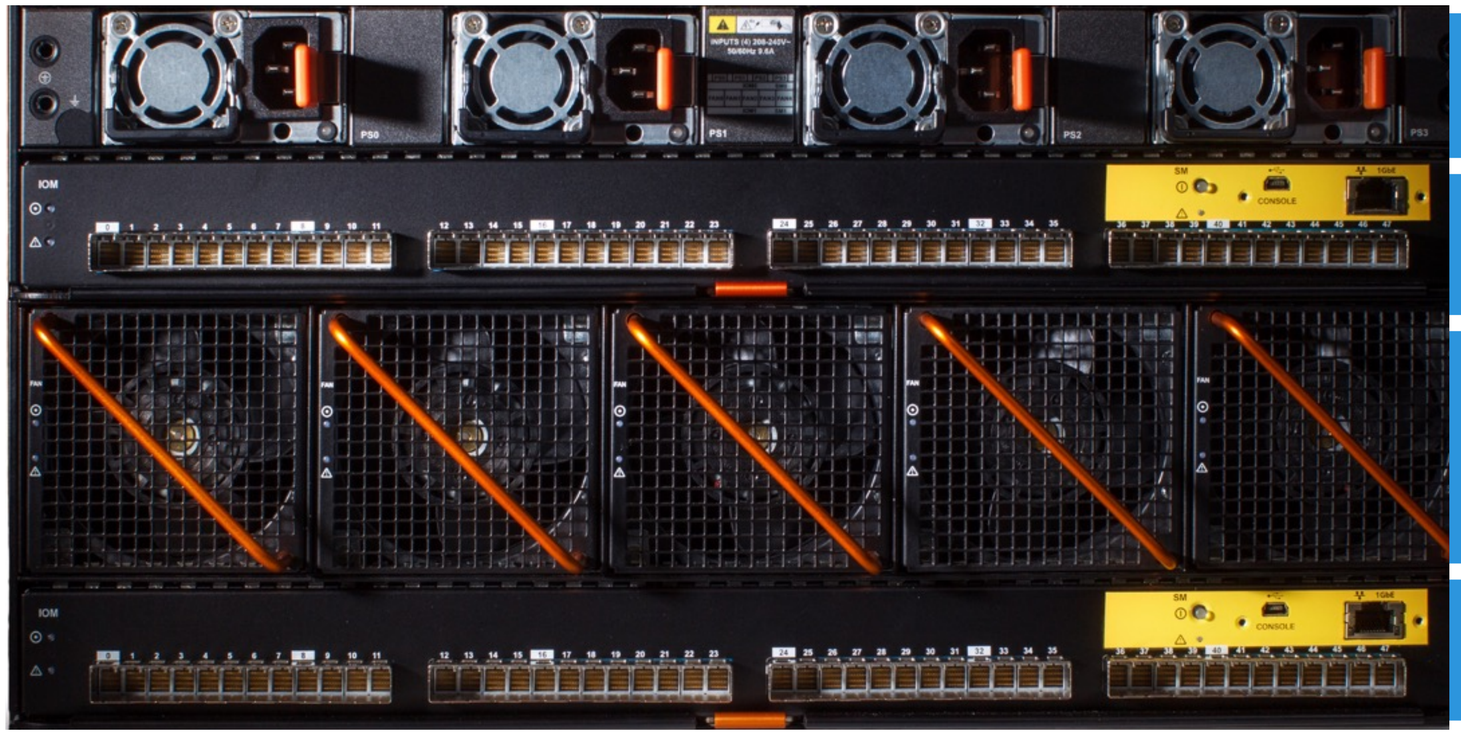
Dual Ported PCIe Gen 3 x4 per FM

Dual-Redundant Control Modules (CMs)
PCIe Gen 3 Connected

DSSD EMC²

Dense and Shared Flash

DSSD **D5** - 5U RACK SCALE FLASH PLATFORM



IOMs, Fans, Power Supplies

Redundant Power Supplies x4

Dual-Redundant IO Modules (IOMs)
PCIe Gen 3 Connected

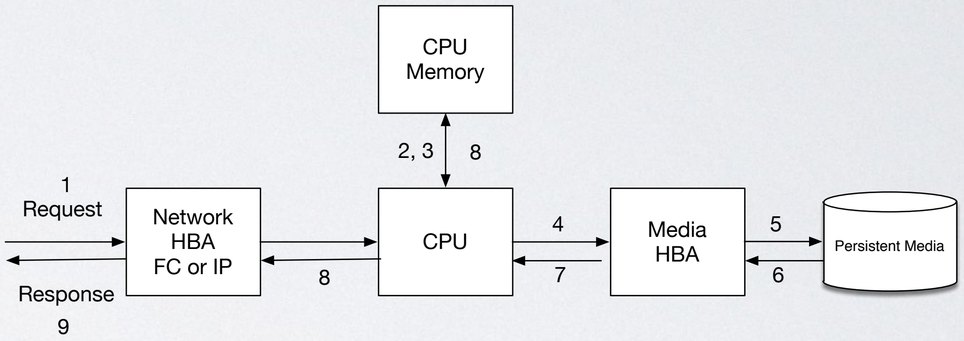
48 PCIe Gen 3 x4 Client Ports Per IOM
Total of 96 PCIe Gen 3 x4 Client Port Connections per D5

Redundant Fan Modules x5

DSSD EMC²

NOISY NEIGHBORS

- In all other (non-D5) storage solutions, **data is served by CPUs**
  - CPUs execute the code to service HBA requests, check caches, request data from media, and so on
  - CPU is a relatively scarce resource, and prone to abuse by certain sessions/systems/users – the noisy neighbors
  - When CPU resource is unavailable, response times degrade rapidly and exponentially

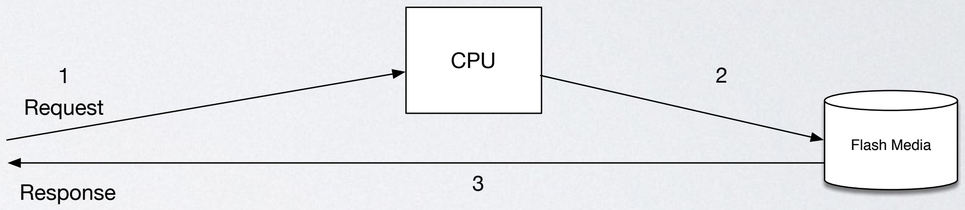


1. Request arrives,
2. CPU accepts interrupt, checks CPU memory for cached copy
3. If found, skip to 8. If not, continue
4. CPU forwards request to Media HBA
5. HBA makes request from persistent media
6. Media locates data and responds
7. HBA forwards data to CPU
8. CPU forwards data to Network HBA
9. Return data to host

NOISY NEIGHBOURS

- In DSSD D5, **data is self-service**
  - Hosts have full access to 18,432 flash chips, a much less scarce resource
  - Data is spread thinly across those chips, minimizing contention
  - All data transfers, re… …ect DMA between ho…
  - The D5 … much performance capacity, compared to other platforms, that the likelihood of a single errant system affecting others is greatly reduced

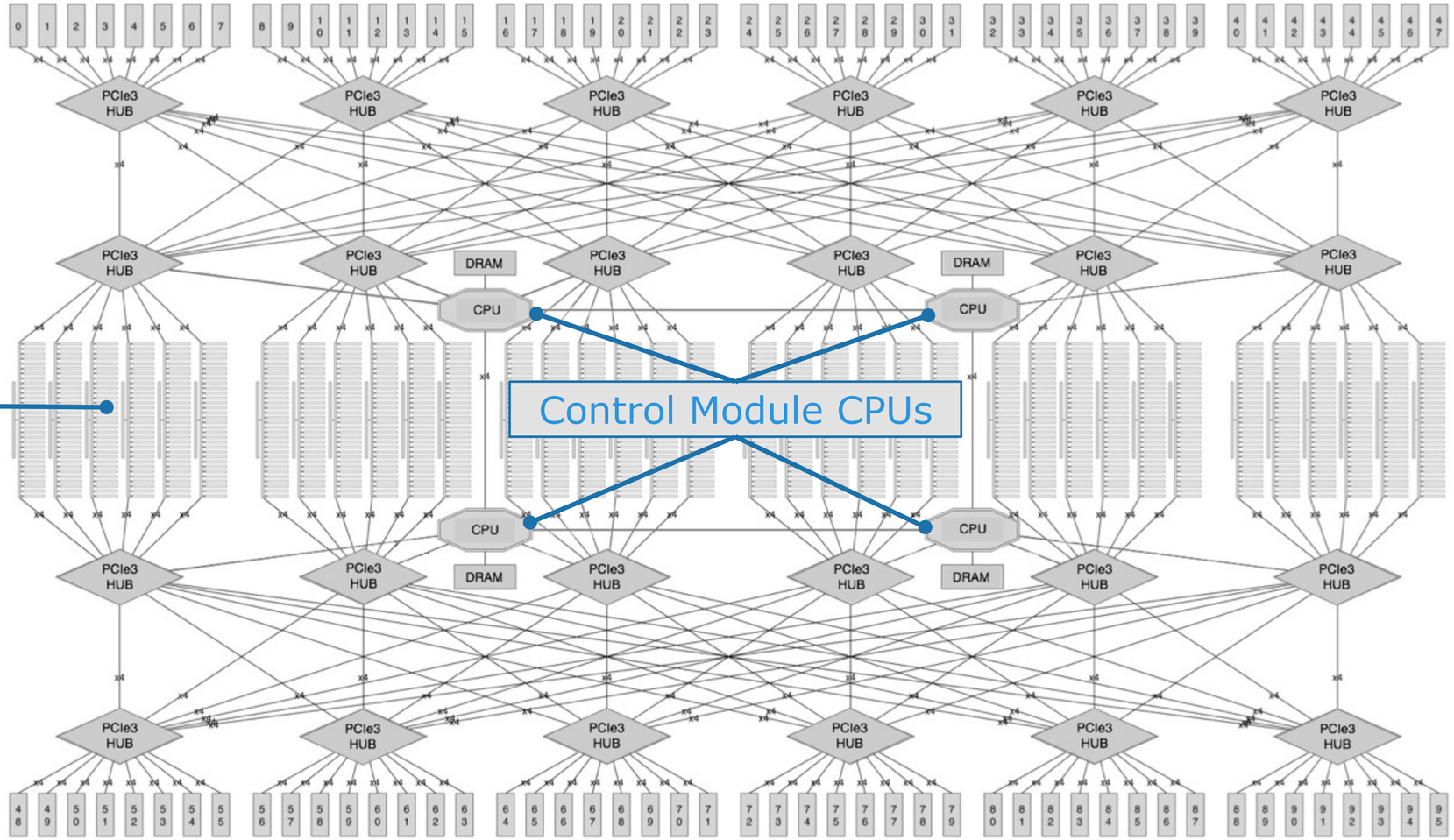
Much less prone to Noisy Neighbor Syndrome!



1. Request arrives (as DMA write of requested LBA)
2. CPU writes DMA directly to appropriate Flash Module
3. Flash Module returns data via DMA write to host

Performance Oriented Architecture

I/O Module PCIe ports



Flash Modules

I/O Module PCIe ports

WHAT DOES ALL THIS GIVE US?

- Marketing 'hero' numbers (real, but using artificial tools):
  - 100TB Usable
  - 100GB/s bandwidth
  - 100µs latency
  - 10 million IOPs (4KB)
  - 5U rack space

- Proven Oracle numbers
  - 100TB Usable
  - 60GB/s bandwidth into Oracle
  - 140µs latency
  - 5.3 million IOPs (8KB, SLOB)
  - 5U rack space

AND THERE'S MORE…

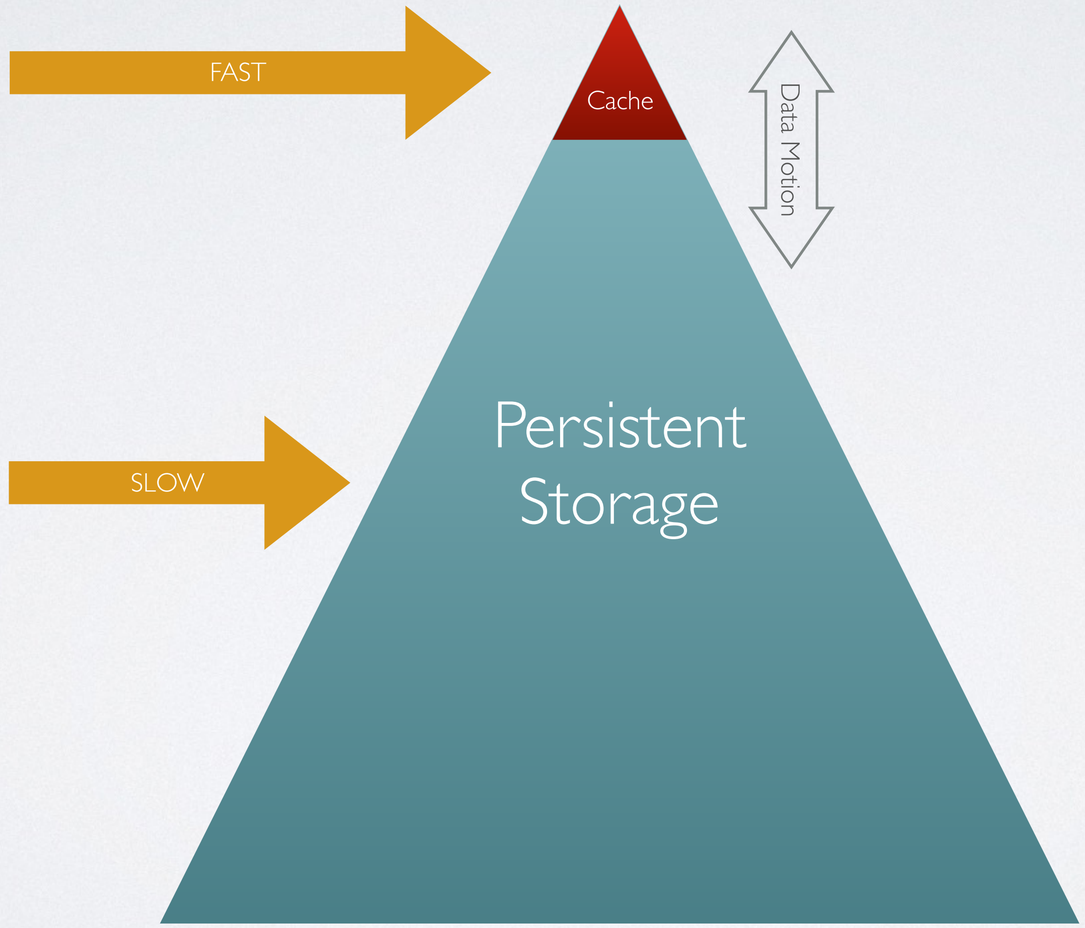
- Up to **two** D5s are currently supported on a single system

- Proven Oracle numbers
  - 200TB Usable
  - 120GB/s bandwidth into Oracle
  - 140µs latency
  - 10.6 million IOPs (8KB, SLOB)
  - 10U rack space

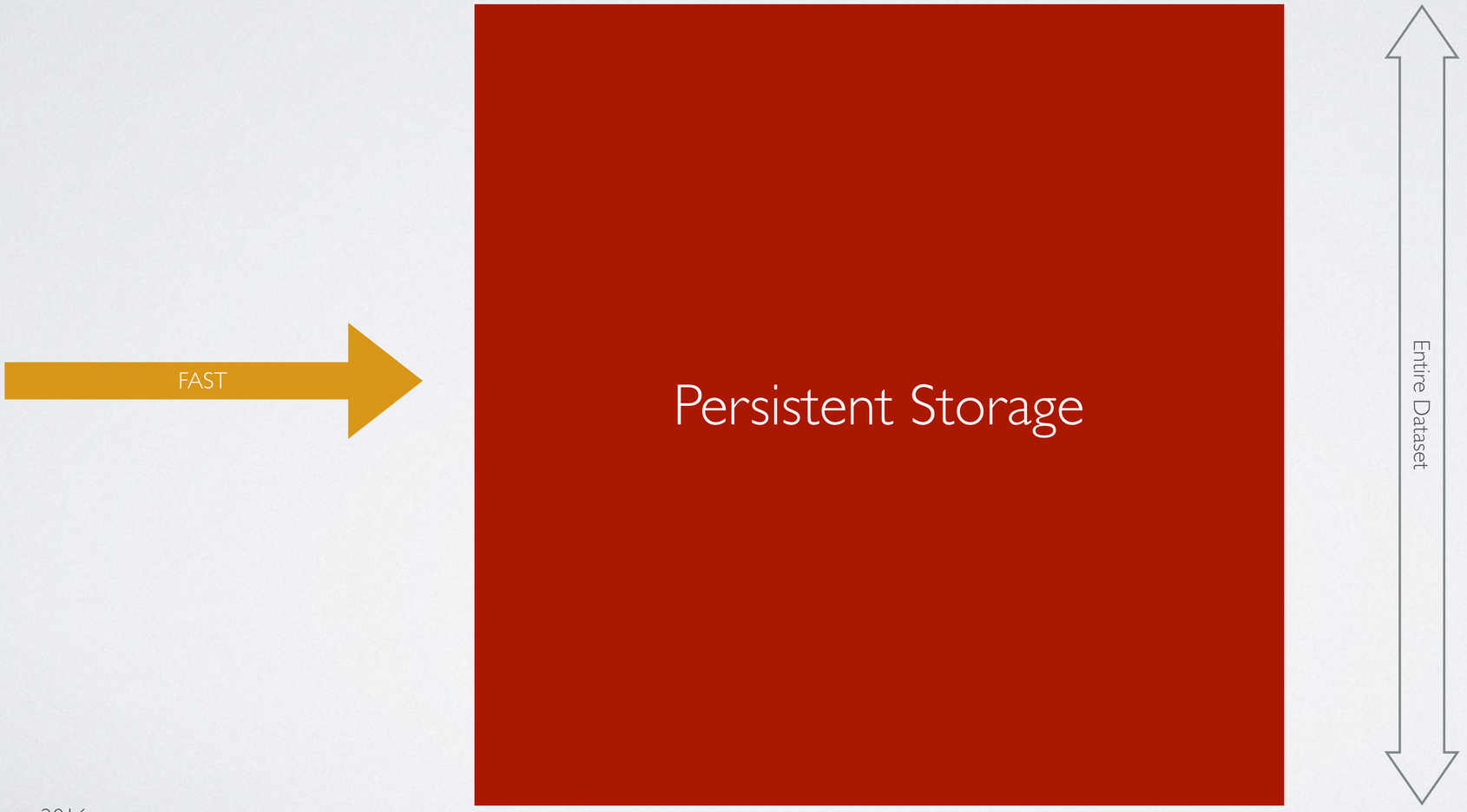
NEW RULES

- D5 has NO cache - Everything is fast
  - You just have a full **100TB usable** 'working set'

TRADITIONAL STORAGE



D5 STORAGE



WHAT DOES IT LOOK LIKE TO A DBA?

- Familiar block-driver interface:
  - i.e.: /dev/dssdXXXX devices
- Fully shared disk
- Multipathing is fully automatic and invisible
- No child devices exposed, no tunables
- Udev rules recommended to create friendly names

- Reference documentation is the "*Oracle Databases on DSSD D5 – Best Known Methods*" paper

WHAT DOES IT LOOK LIKE TO A DBA?

```
# ls -l /dev/asmdisks
total 0
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraOCR000441_00 -> ../dssd0030
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraOCR000441_01 -> ../dssd0031
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraOCR000444_00 -> ../dssd0028
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraOCR000444_01 -> ../dssd0029
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraRedo000441_00 -> ../dssd0000
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraRedo000441_01 -> ../dssd0001
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraRedo000444_00 -> ../dssd0026
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraRedo000444_01 -> ../dssd0027
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraVol000441_00 -> ../dssd0032
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraVol000441_01 -> ../dssd0033
lrwxrwxrwx 1 root root 11 Feb 11 20:19 OraVol000441_02 -> ../dssd0034
```

WHAT DOES IT LOOK LIKE TO A DBA?

```
SQL> l
  1* select group_number,path,name,failgroup,mount_status from v$asm_disk order by 1,4,3
SQL> /

GROUP_NUMBER PATH                                     NAME                 FAILGROUP            MOUNT_S
------------ ---------------------------------------- -------------------- -------------------- -------
           0 /dev/asmdisks/OraFRA000441_03                                                      CLOSED
           0 /dev/asmdisks/OraVol000441_11                                                      CLOSED
           0 /dev/asmdisks/OraVol000444_06                                                      CLOSED
           0 /dev/asmdisks/OraVol000444_03                                                      CLOSED
           0 /dev/asmdisks/OraRedo000444_00                                                     CLOSED
           0 /dev/asmdisks/OraVol000444_09                                                      CLOSED
           0 /dev/asmdisks/OraVol000444_01                                                      CLOSED


… etc
```

ELIMINATION OF COMPLEXITY

- No dm-multipath or Powerpath
  - Purpose built, high performance multipathing integral in client drivers
  - Only a single device name is exposed, all detail is handled by the driver
- No manipulation of I/O elevators
  - NOOP is forced
  - Everything is 4KB anyway (blkdev)
- DMA access and separate submission and completion queues
  - No queue tuning - DMA enqueues so fast that it is largely unnecessary - but we make an exception for redo

WHICH DIMENSION MATTERS?

- Bandwidth?
- Latency?
- IOPs?
- Nobody actually needs 5.3M IOPs, but they are a side effect of the bandwidth and low latency - which people DO need!

ANALYSIS OF DB TIME

- Low latency storage **dramatically** alters the split of time for a process
- Using SLOB:
  - Traditional storage: ~200µs CPU, 6000µs single block read. 30:1 ratio
  - D5: ~200µs CPU, ~200µs (at high load) single block read. **1:1 ratio**

LATENCY: SYNCHRONOUS I/O

- Oracle workloads are most frequently dependent on **synchronous I/O**
  - Index traversal and Nested Loop joins (serial I/O pathology)
  - Log writer (redo bandwidth is proportional to write latency)
- Latency is now *so low* that the returns are diminishing after this:
  - Reducing disk latency from 6ms->3ms was almost 2x speedup
  - But now the compute time is similar to the I/O time - halving I/O latency is **25%** speedup
    - OMG - if we **eliminate** I/O altogether, we can only go 2x faster. Where did orders of magnitude go?!

BANDWIDTH: BIG QUERIES

- It is rare that 'adhoc query' exists in reality:
  - Sure, submit the query
  - But it might not come back until next Tuesday
  - Oh, and everyone else will suffer while it runs

THE REALITY: THE DBA'S PLIGHT!

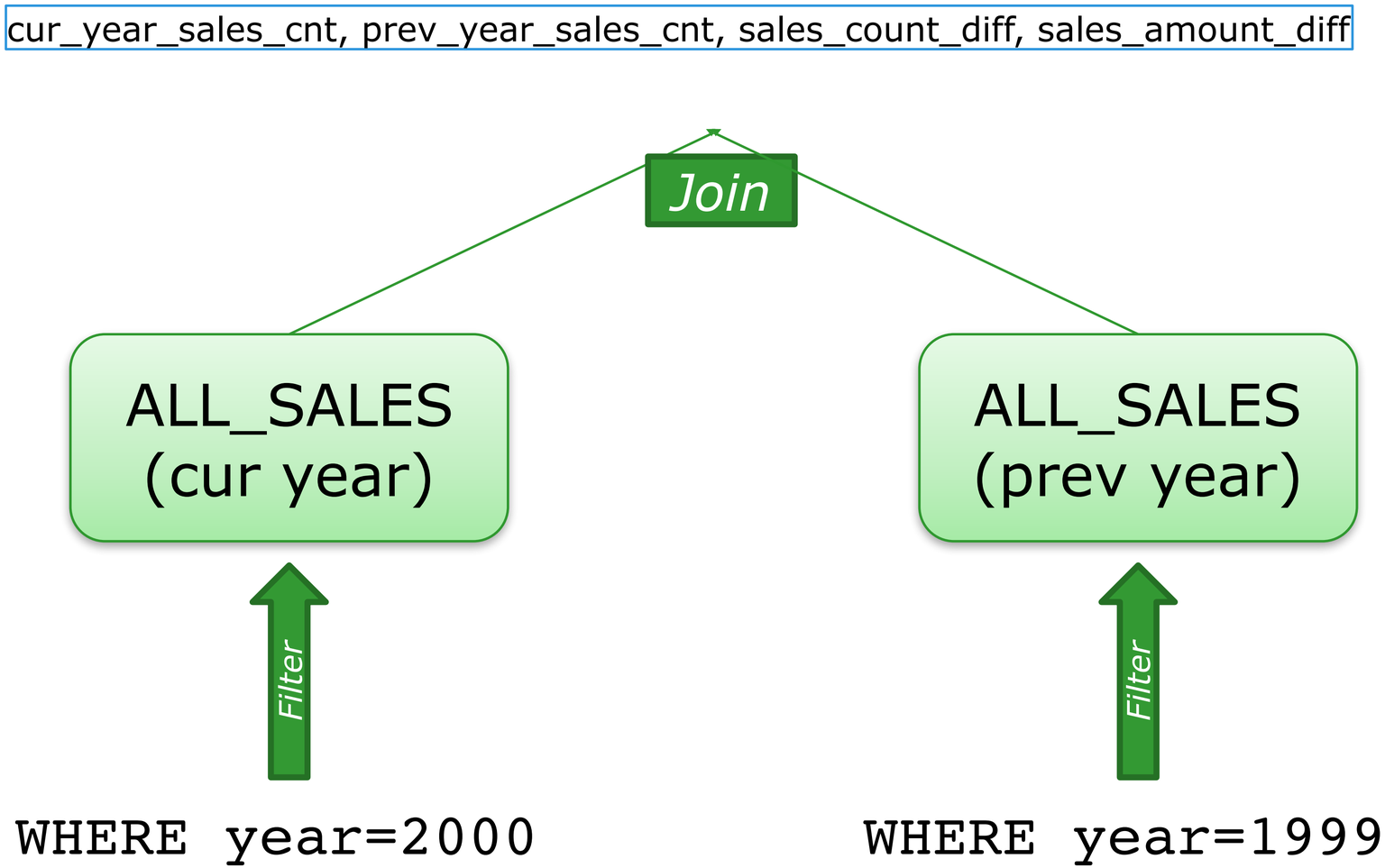
- Physical schema mitigations are adopted to minimize the data scan volume:
  - Materialized Views
  - Secondary Indexes
  - Fine grain subpartitioning
  - Even Smart Scan - a **non-deterministic workaround**

AN EXPERIMENT

- **Exam Question:** How much do Materialized Views actually help with runtimes when you have next-generation I/O horsepower?

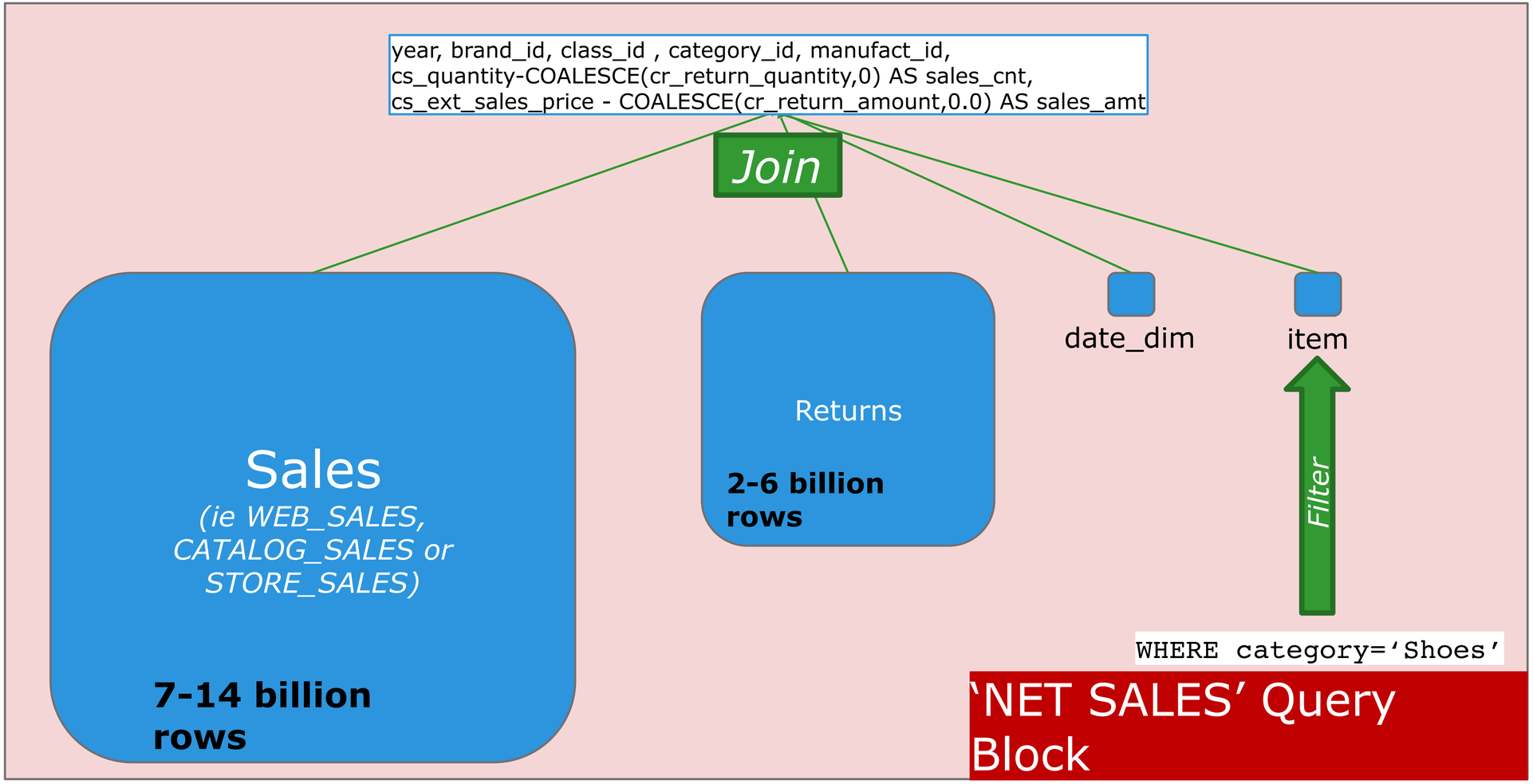
DISSECTING THE QUERY

TOP LEVEL



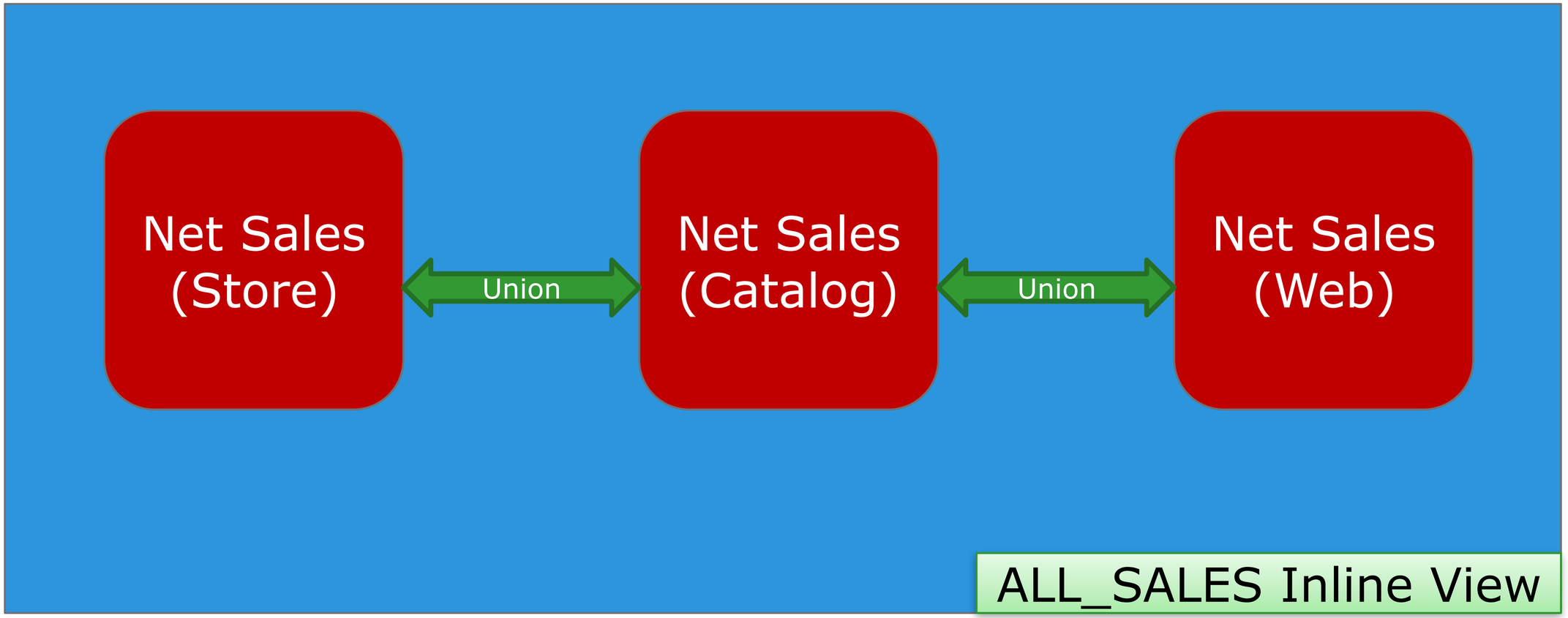
DISSECTING THE QUERY

MAIN BLOCK



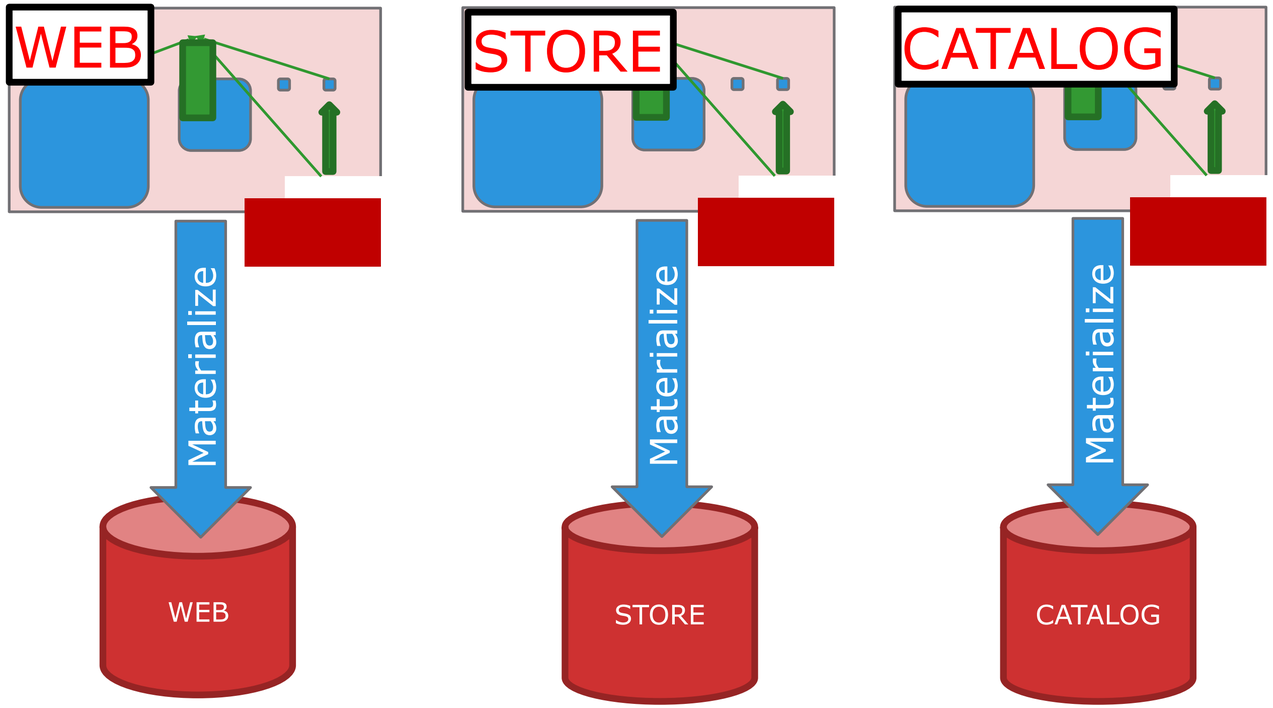
DISSECTING THE QUERY

UNION

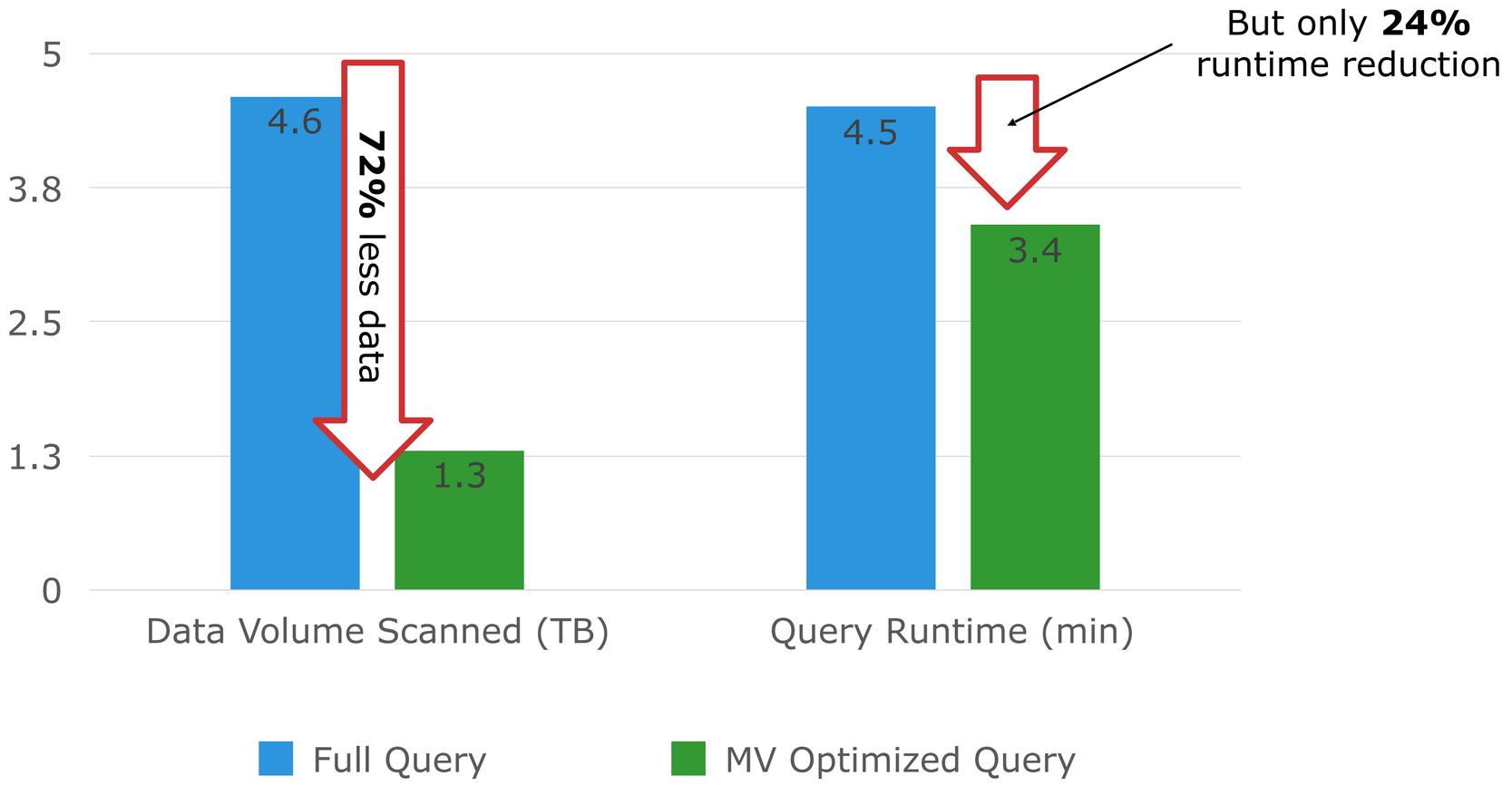


THE TEST

- Materialize the main query block of the three sales channels



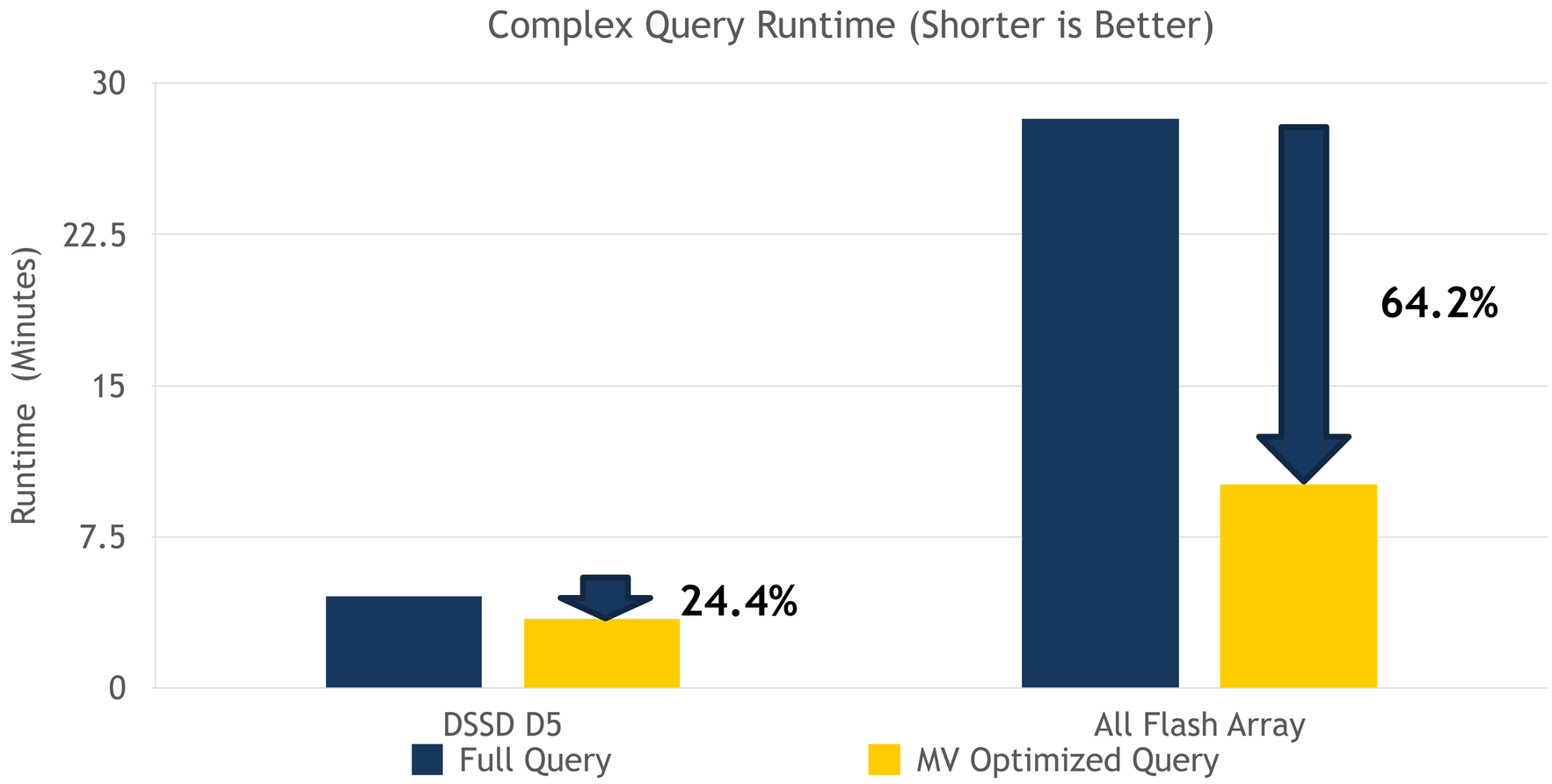
RESULTS



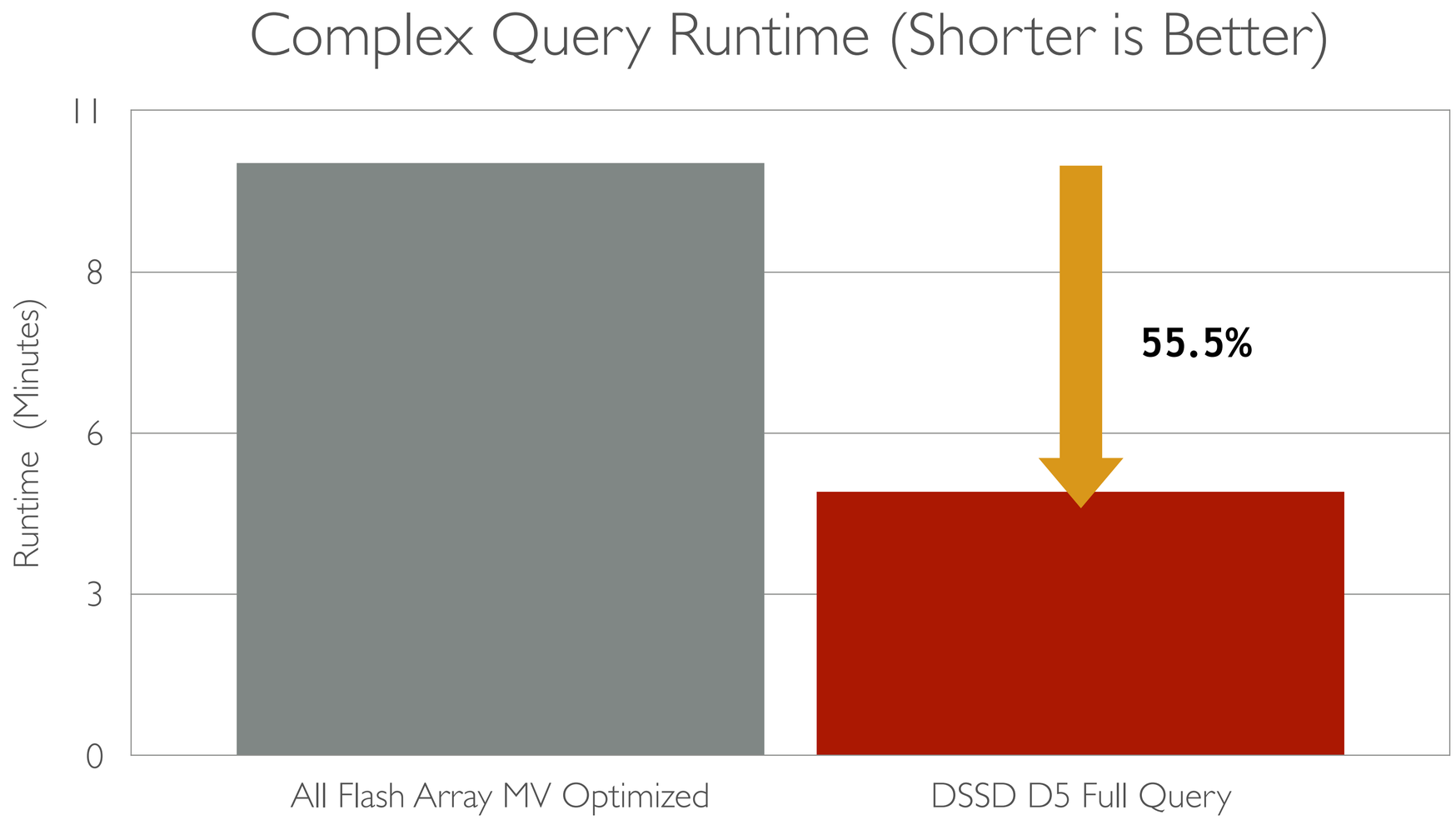
WHY ONLY A SMALL SPEEDUP?

- DSSD D5 makes the I/O portion of the query much **less significant** in the total runtime
- Remaining work, such as CPU compute, serialization, and inter-node communication remain constant

D5 Versus a Typical All-flash Array

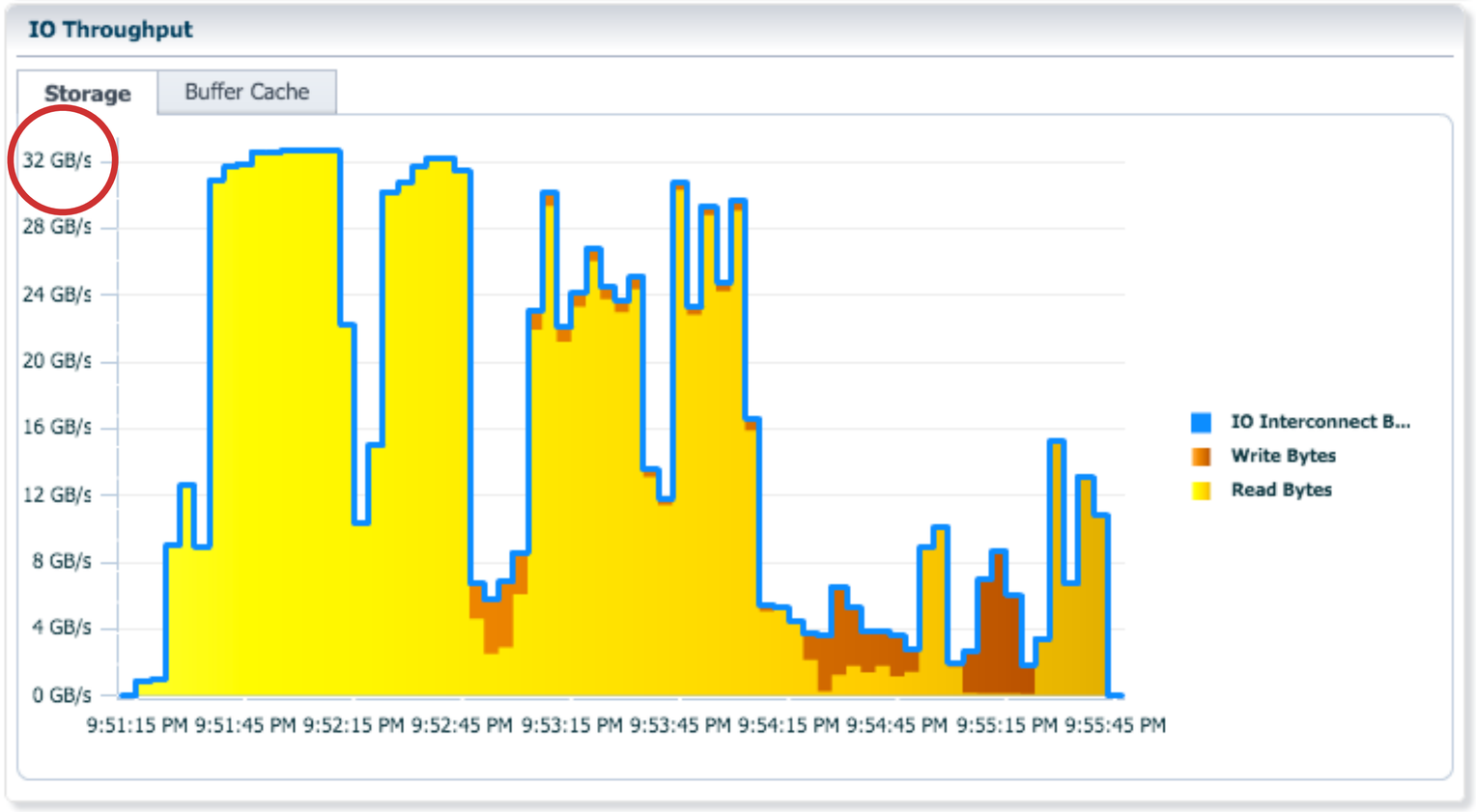


D5 Full Query vs. AFA Materialized View



BANDWIDTH MATTERS

- Full query running on DSSD D5 (with gas left in the tank):



SO WHAT?

- **“Extreme Performance” is not just for “Extreme Workloads”**
  - As a DBA, you have only been able to deliver that which the hardware allows
  - “Extreme Performance” is an *enabler* to business transformation

SOFTWARE: ALGORITHMS

- Until now, a cache miss meant certain death…
  - At least 50x slower, including code path
- Net result: algorithms carefully maximize cache hit, and optimizer aggressively favors cached access paths

SOFTWARE: ALGORITHMS

- Next-Gen Storage:
  - Cost of cache miss is much, much less
  - But algorithms remain largely the same
  - Algorithms could be significantly more speculative in approach

SQL OPTIMIZER

- Should push more I/O out as large physical I/O requests
  - Large index joins will become less relevant - synchronous/serial pathology and inefficient join algorithm at scale
  - Large PIO is async and parallel, and hash joins are highly effective (if you can spill to disk at a decent rate)

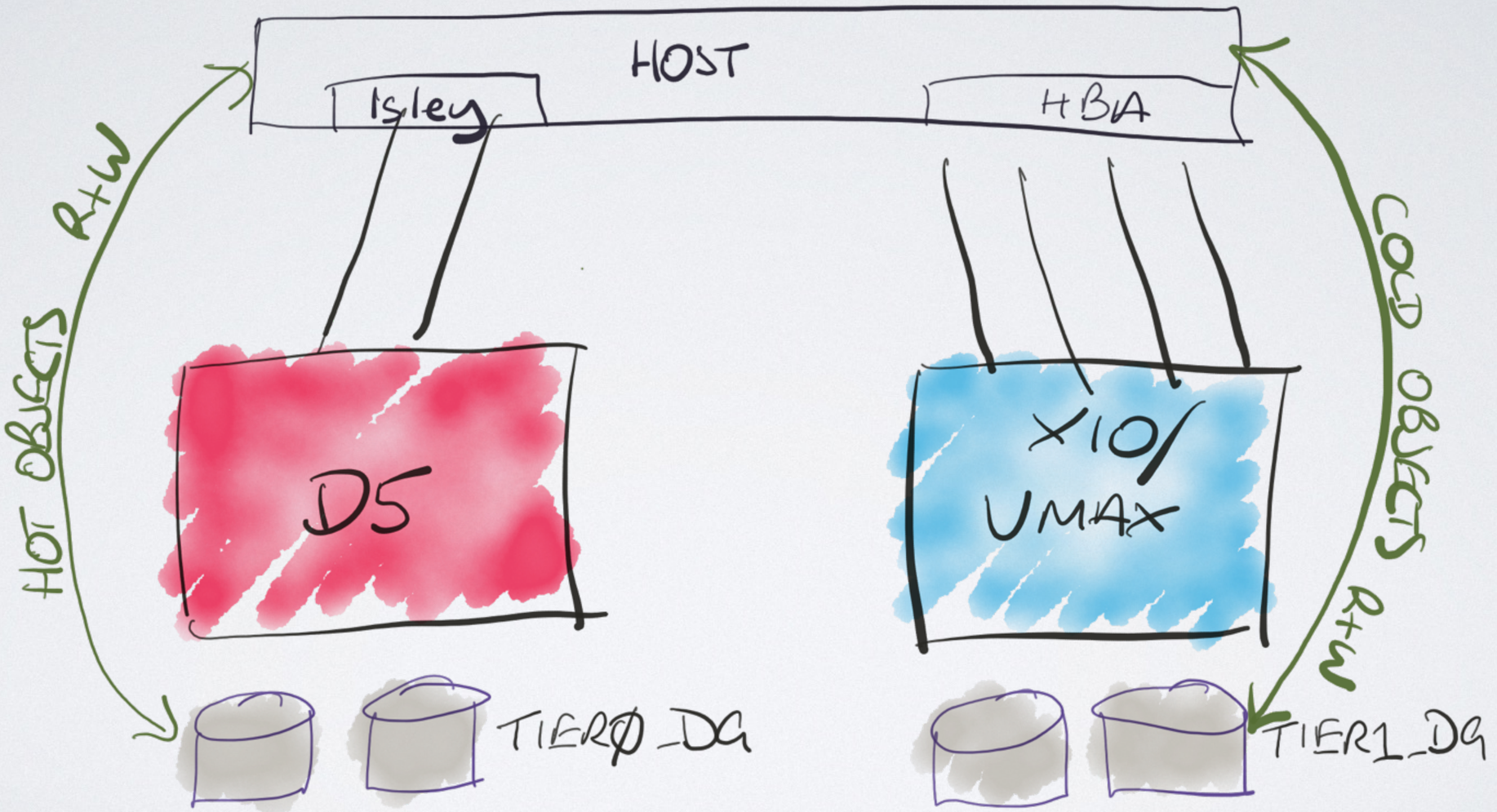
WHAT'S MISSING?

- Things that will probably never come:
  - Data Services
    - Compression
    - Dedupe
- Things that are coming:
  - Data Services (probably)
    - At-rest Encryption
    - Snapshots
    - Replication
  - **Full Non-disruptive Operations support (definitely, and soon)**

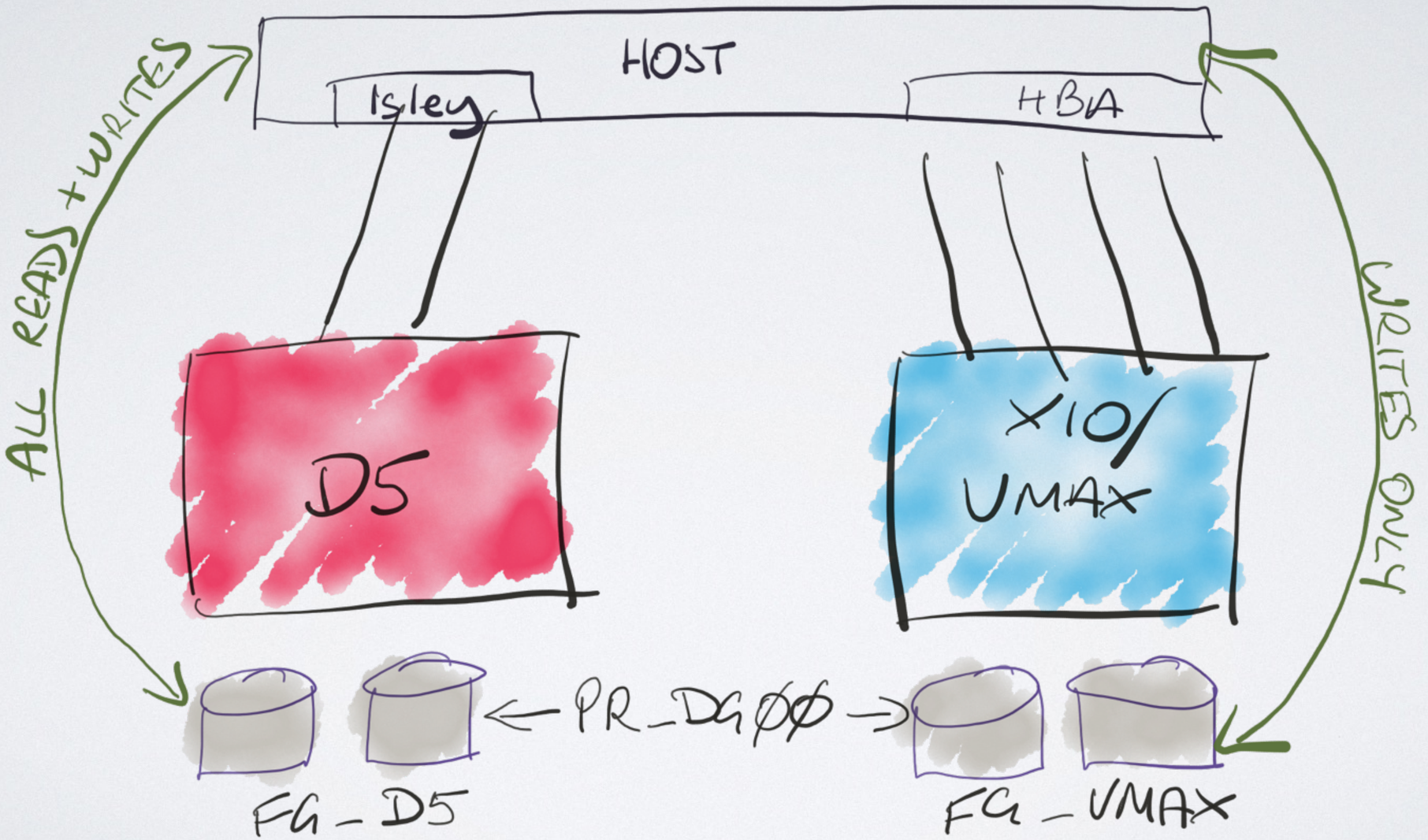
ARCHITECTURES

- Tiering with D5
- Preferred Read Failure Group

STORAGE TIERING



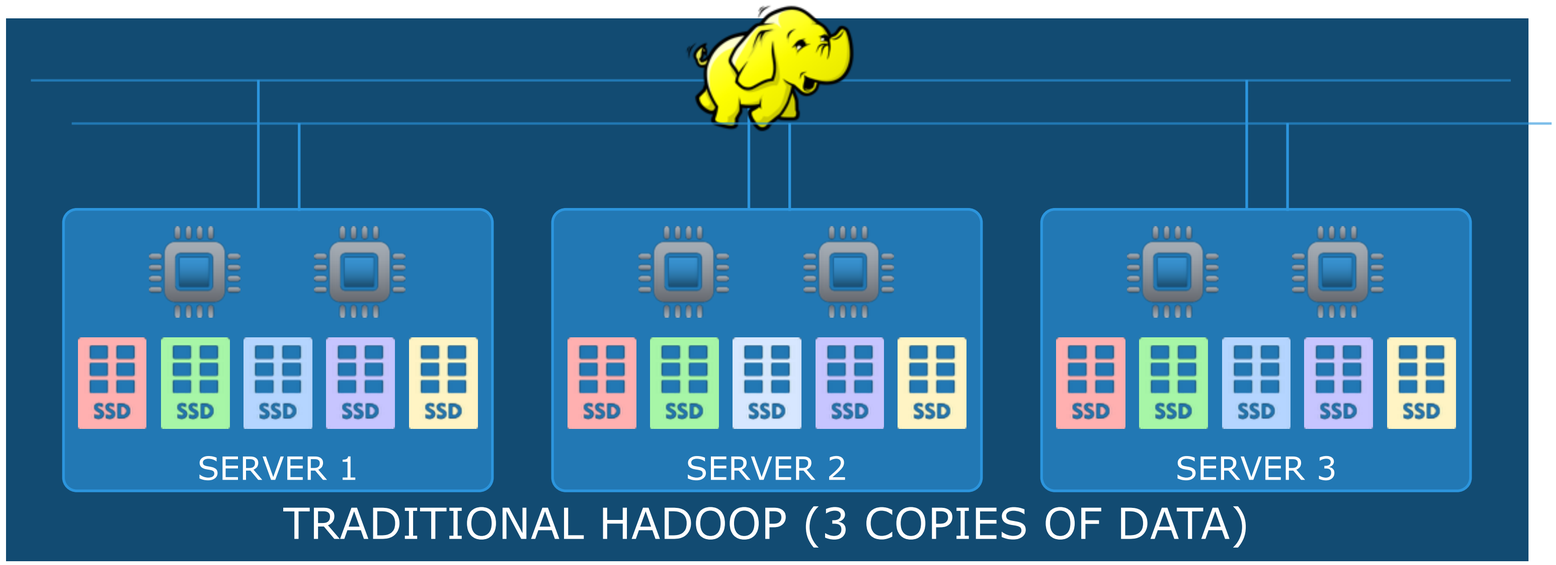
Preferred Read Failure Group



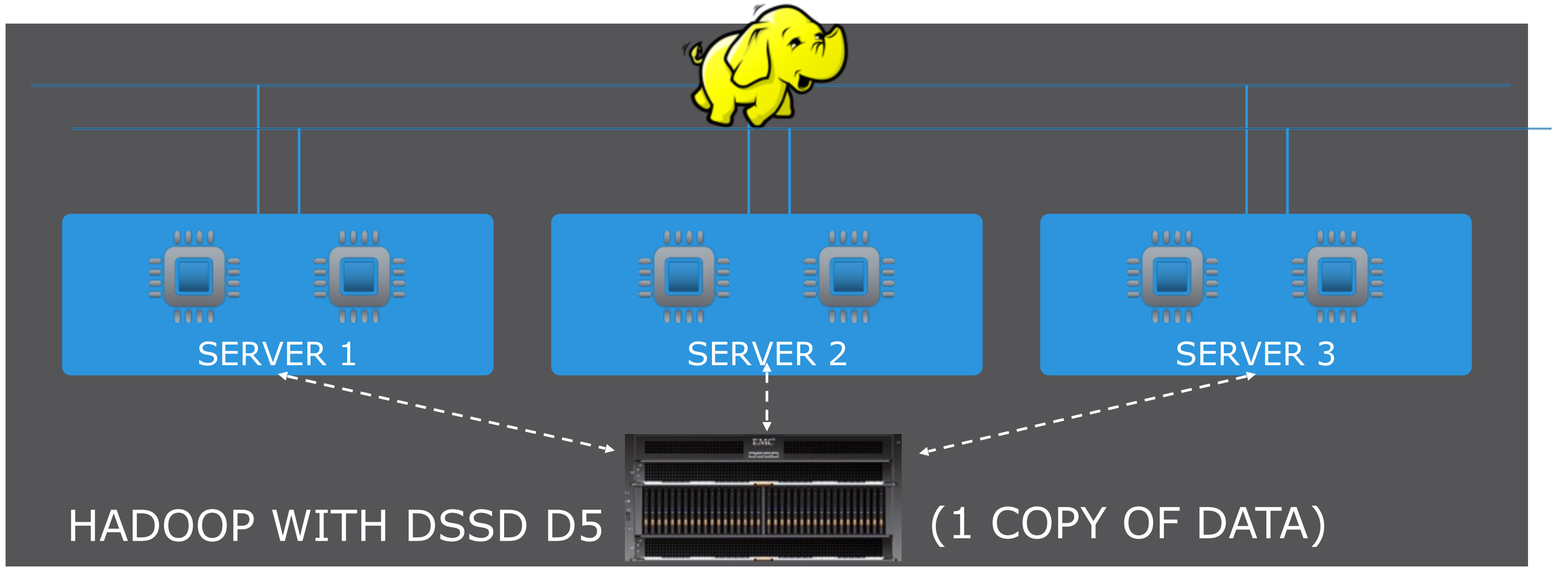
HADOOP/HDFS SUPPORT

- There is also an HDFS Datanode Plugin

FLASH OPTIMIZED HDFS



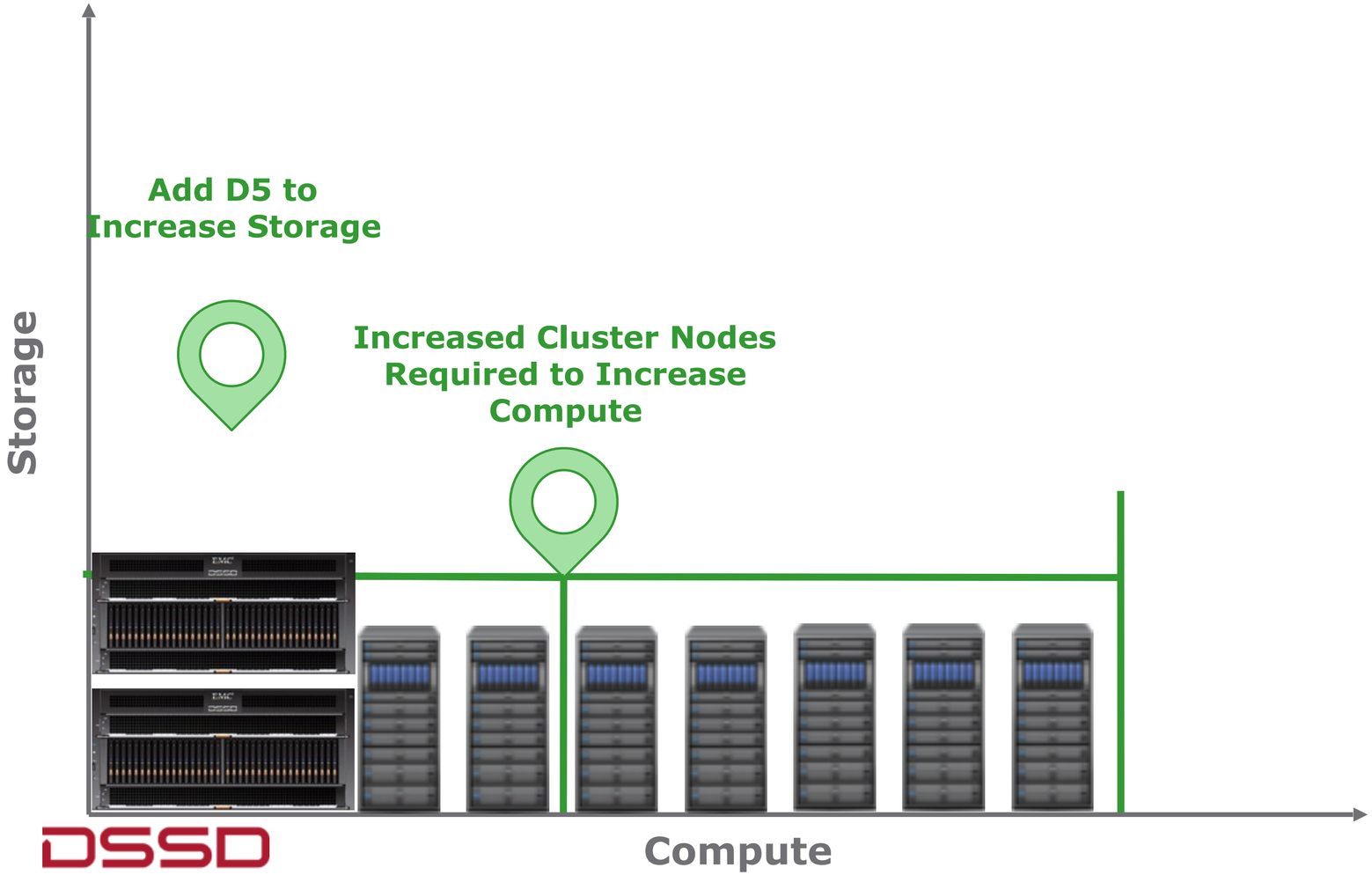
- HDFS uses a replication factor of at least 3 for availability
- Results in 3x+ data on persistent media
- Not economical for flash



- Stores just 1 copy of data regardless of replication factor
- Use entire flash capacity for data
- Increase data locality without using more capacity

SIMPLIFIED ARCHITECTURE

INDEPENDENT SCALING



HDFS on DSSD

- Scale compute *independent* of storage
- Achieve optimal asymmetric high performance balance
- Add additional performance as hardware evolves

HADOOP/HDFS SUPPORT

- Elimination of Replication
  - Storage savings make the D5 price competitive with local SSDs
  - Local data access is possible for every attached host *without* storage multiplication
    - **Eliminates any Key Hashing hotspots**
- Run all of this, Oracle, Hadoop, Filesystems, on the same storage platform

NEXT STEPS

- Moore's Law++ <—12mo doubling in storage density
- Controller CPU and memory is also subject to Moore's Law - balanced growth
- Optane/3DXpoint - another order of magnitude

THANK YOU!

- Any Questions?

- morlej@dssd.com

- @jamesmorle