DDN STORAGE

Optimizing storage stacks for AI

Spectrum Scale CIUK UG

December, 2018

Sven Oehme – Chief Research Officer DDN

DDN SFA | ALL-FLASH AND HYBRID BLOCK STORAGE PLATFORMS

200NV	400NV	7990	14KX	18K
				
23GB/s 1M IOP/s	42GB/s 3M IOP/s	23GB/s 1M IOP/s	60GB/s 4M IOP/s	92GB/s 3.2M IOP/s
24 NVME Slots	24 NVME Slots		48 NVMe Slots	48 NVMe Slots
		Up to 450 SSD/HDD	Up to 1872 SSD/HDD	Up to 1872 SSD/HDD
EDR IB (4), OPA (2) FC32 (8), FC (8)	EDR IB (8), OPA (4)	EDR IB (4), OPA (2) FC16 (8)	EDR IB (12 | 8) OPA (4), FC16 (24)	EDR IB (16), OPA (8)
NEW 2018	COMING 2019	NEW 2018		COMING 2019

DDN | GRIDScaler

Massively Scalable NAS & Parallel File Storage Appliance

	GS200NV	GS400NV	GS7990	GS14KX	GS18K
GRIDScaler v4	✓	✓*	✓	✓	✓*
v4 upgrade to v5	✓	✓*	✓	✓	✓*
GRIDScaler v5	✓	✓*	✓	✓	✓*

- Easy to deploy, All-in-One Appliance for All Flash Array with HDD, archive and cloud tiering options
- Scale-out building blocks architecture
  - Configurations scale from <100 TB to PBs of storage and 10s of TBs/sec of performance
- Flash Centric Architecture - custom embedded fabric delivers optimal SSD performance
- Feature-Rich, Enterprise Grade Quality and High Availability with no single point of failure
- Simple, Intuitive but Powerful DDN Insight monitoring solution

Optimizations for GRIDScaler V5

Optimizations for GRIDScaler V5

- Updated device drivers, OS and Scale tuning parameters and SFA multi-queue LUN support
- Embedded systems can now achieve up to 1.2 Million random 4k read IOPS
- External SFA14KX NSD Server performance went from 1.25 Million to 2.96 Million*
- This enhancements were used to produce the SpecSFS 2014 record publications

* test was using external NSD Server. all numbers are measured from network attached clients with GPFSPERF using one 100 GB file per client during random 4k reads using O_DIRECT

Platform Optimizations help significantly

ES200NV | LOW LATENCY DESIGNED-IN

CLIENT → HCA/NIC → SWITCH → HCA/NIC → SERVER → FILESYSTEM → HBA → SAN SWITCH → HBA → STORAGE

IO PATHS

TRADITIONAL

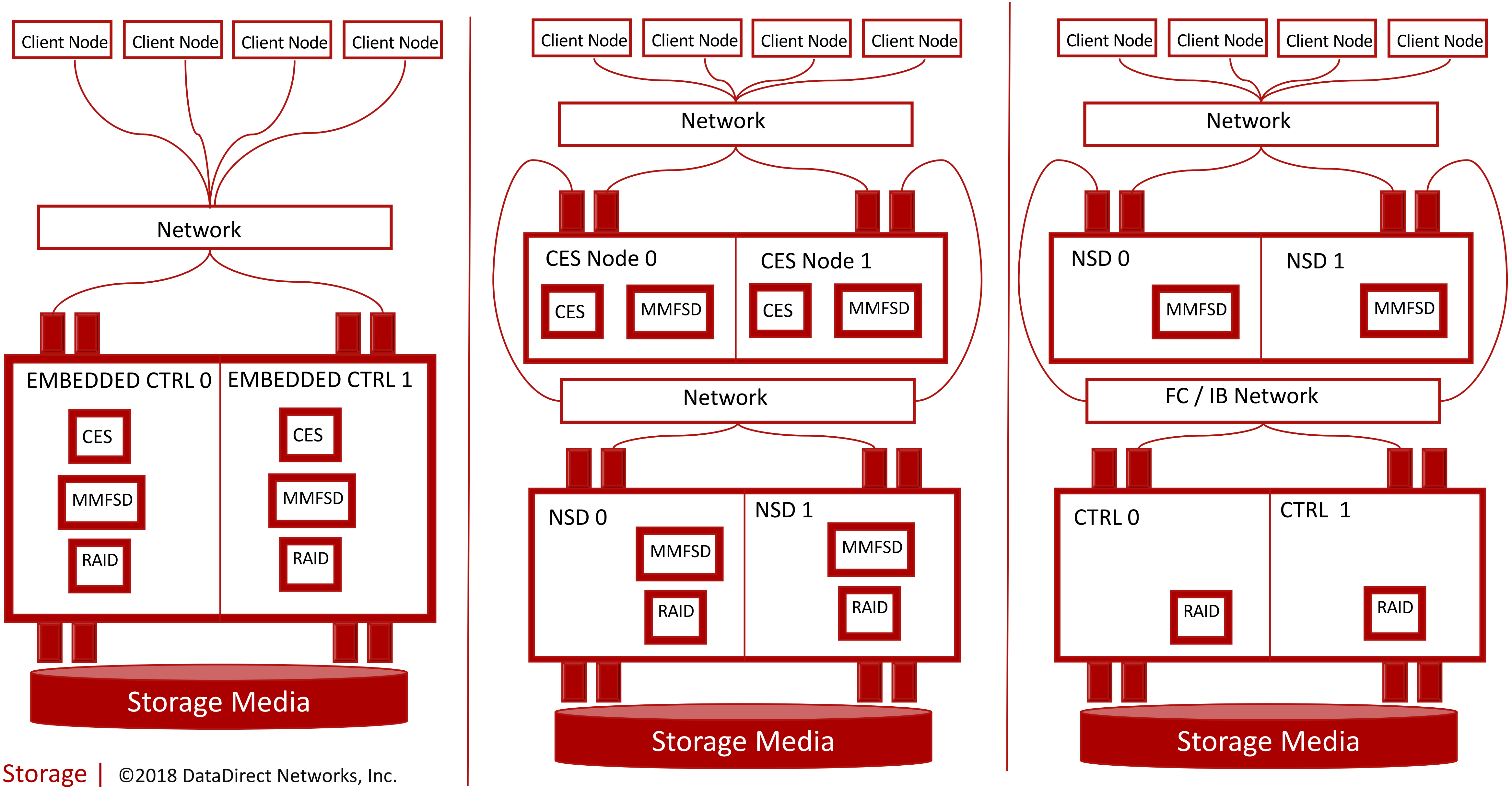
Components Simplified with SFA™ Embedded Appliances

SFA EMBEDDED FILESERVICE

CLIENT → HCA/NIC → SWITCH →

SCALER™ APPLIANCES

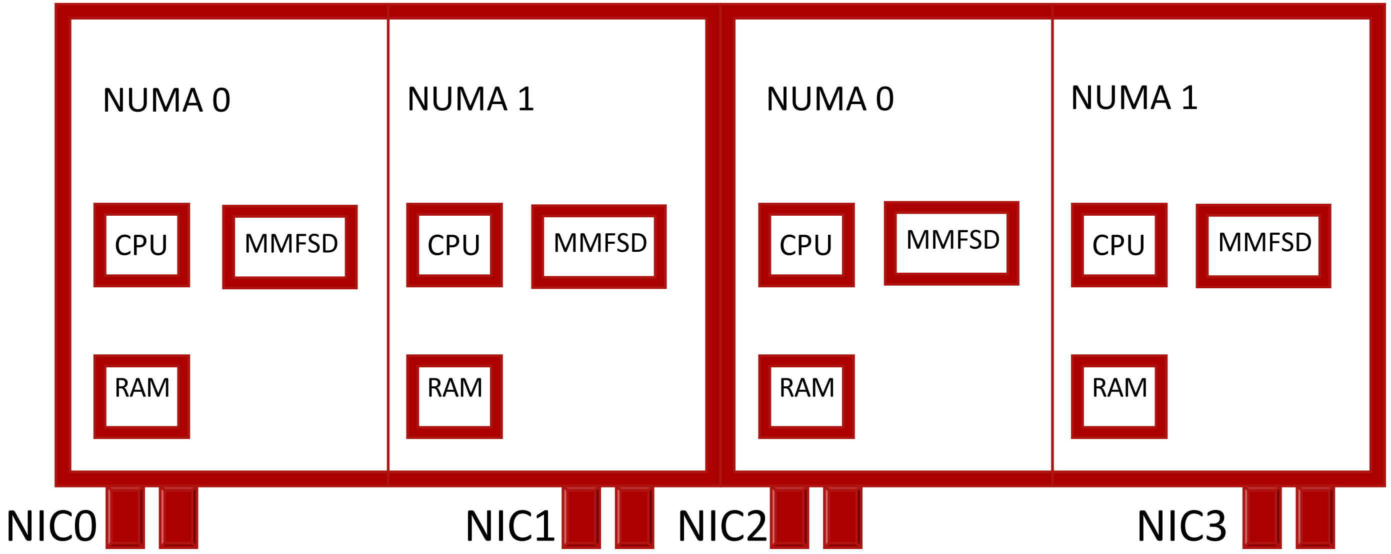
The fastest network hop is the one you can avoid



Why all this work, what’s to gain ?

- Remote NUMA region HW access in SW is one of the biggest issue to achieve HW capable performance targets
- even just a 2 NUMA Zone system (e.g. modern Intel 2 socket system) has significant overhead as without optimization on the SW, 50% of the access is remote, as larger the number of NUMA nodes as more overhead , each IBM power or modern AMD CPU has 2 NUMA nodes. So a 2 socket Power 8 system has 4 NUMA zones and a 75% chance your data is on the wrong side.
- Databases developers have spend years to optimize their SW stack to be NUMA aware, storage stacks are trying to catch up. On databases tests have show between 2-4x improvements with proper memory placement, for Storage the benefit can be even greater as it typically interacts with HW beyond memory that is NUMA dependent (e.g. HBA’s or HCA’s)
- Remote HW access significant increases latency and causes very unpredictable performance
- Linear scaling with increased core counts gets eliminated by contention on interconnects or lock overhead requiring synchronization between NUMA regions

SFA NUMA awareness



The system is perfectly balanced across numa nodes, which allows affinitizing of mmfsd threads to memory, core and network for lowest latency and consistent scaling

DIO Random 4k writes into a 100GB files

```
/usr/lpp/mmfs/samples/perf/gpfsperf write rand /target/sven-100g
  recSize 4K nBytes 100G fileSize 100G
  nProcesses 1 nThreadsPerProcess 1
  file cache flushed before test
  using direct I/O
  offsets accessed will cycle through the same file segment
  not using shared memory buffer
  not releasing byte-range token after open
  no fsync at end of test
    Data rate was 34659.88 Kbytes/sec, Op Rate was 8461.89 Ops/sec, Avg Latency
was 0.118 milliseconds, thread utilization 1.000, bytesTransferred 1039802368
```


DIO Random 4k reads from a 100GB files (exceeds all cache by 4x)

```
/usr/lpp/mmfs/samples/perf/gpfsperf read rand /target/sven-100g
  recSize 4K nBytes 100G fileSize 100G
  nProcesses 1 nThreadsPerProcess 1
  file cache flushed before test
  using direct I/O
  offsets accessed will cycle through the same file segment
  not using shared memory buffer
  not releasing byte-range token after open
    Data rate was 21763.50 Kbytes/sec, Op Rate was 5313.36 Ops/sec, Avg Latency
was 0.188 milliseconds, thread utilization 1.000, bytesTransferred 652910592
```


DIO Random 4k reads from a 100GB files (exceeds all cache by 4x) ETH

```
/work/oehmes/bin/gpfsperf read rand -r 4k -n 100g -th 1 -dio -millis 5000
/ai200g/test.sven
/work/oehmes/bin/gpfsperf read rand /ai200g/test.sven
  recSize 4K nBytes 100G fileSize 100G
  nProcesses 1 nThreadsPerProcess 1
  file cache flushed before test
  using direct I/O
  offsets accessed will cycle through the same file segment
  not using shared memory buffer
  not releasing byte-range token after open
    Data rate was 14888.68 Kbytes/sec, Op Rate was 3634.93 Ops/sec, Avg Latency
was 0.275 milliseconds, thread utilization 1.000, bytesTransferred 74448896
```


World record SpecSFS 2014 with GRIDScaler*

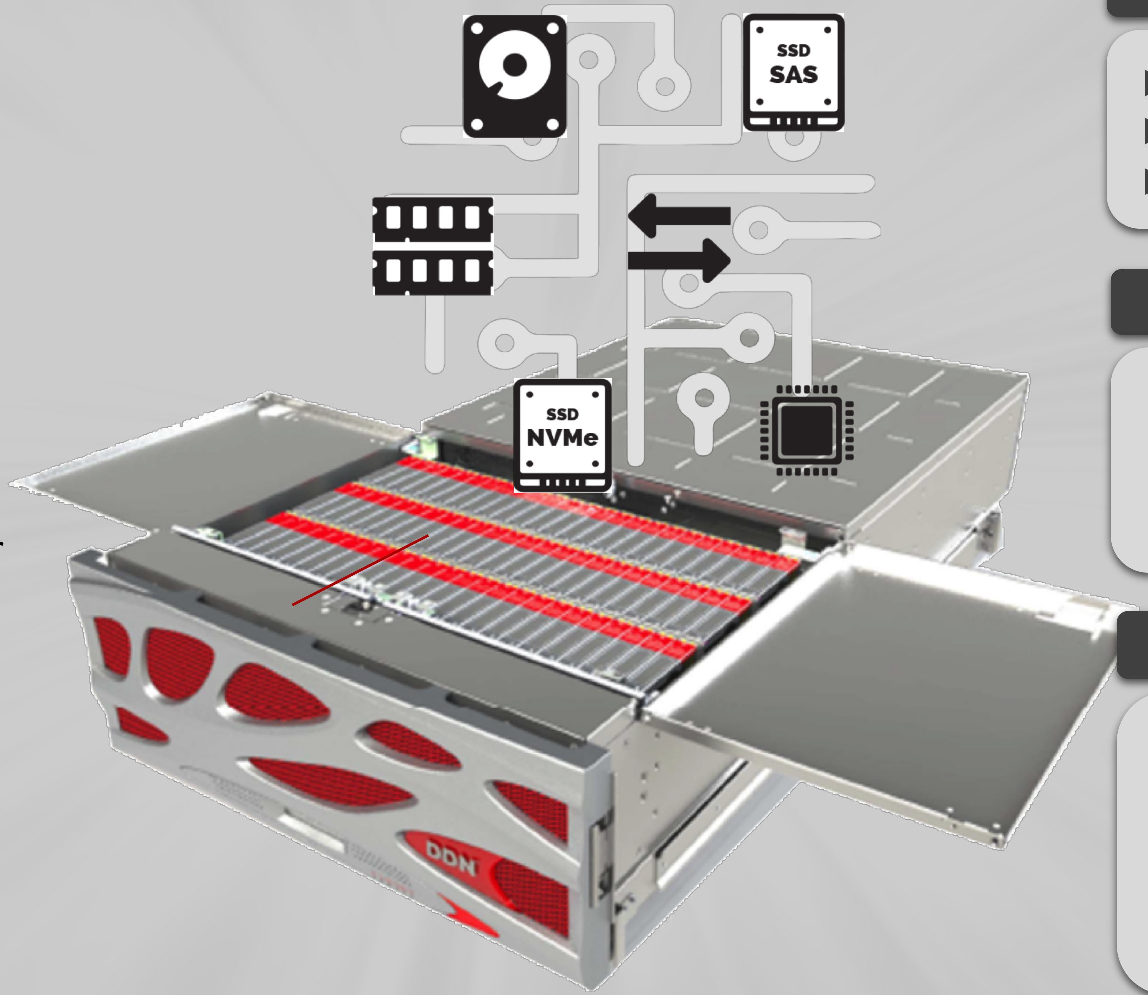
*world record has been broken with an 8 Storage System setup - we use ONE !

DDN SFA14KX

Fastest, Densest and Simplest at Scale

Low Latency, Highly Efficient Architecture

- All in one integrated design with expansion capability
- Dual Redundant Controllers
- 72 Drive High-Density 2.5" Enclosure with NVMe support for 48 2.5" dual ported NVMe
- Optimized Building Block for BW or IOPs
- Support for up to 20 SS9012 12Gb/s 90 drive Enclosures



Flexible Connectivity

- ▶ 10/40/100GbE
- ▶ IB and OmniPath
- ▶ 16/32Gb FC

Industry Leading Performance

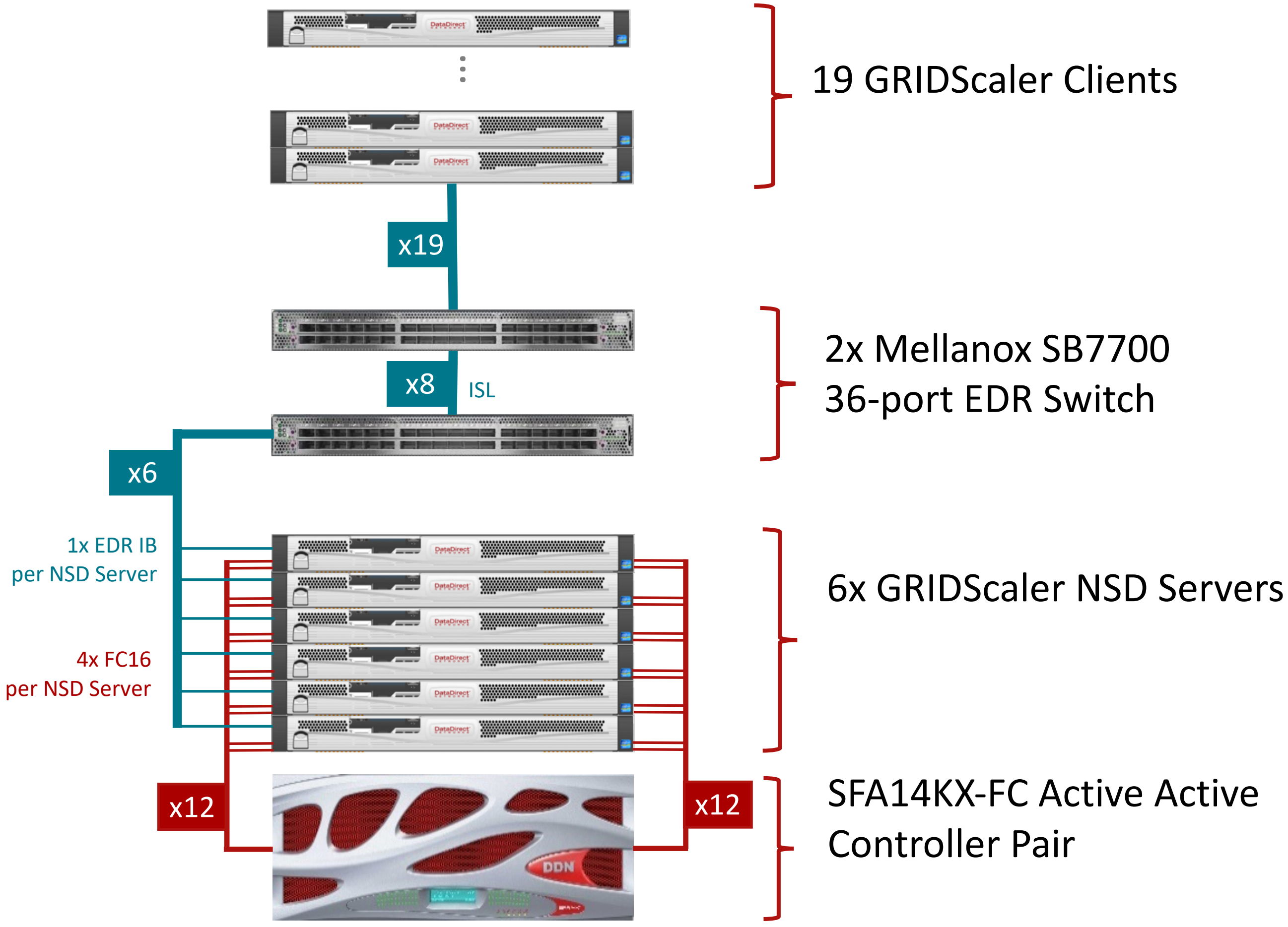
- ▶ 72 SAS SSD or 48 NVMe
- ▶ Up to 60 GB/sec throughput
- ▶ Up to 4 million IOPS

Best Data Protection

- ▶ Fully Declustered RAID
- ▶ Higher Data Availability
- ▶ Flexible Pool Management
- ▶ Optimized for both Random and Sequential IO

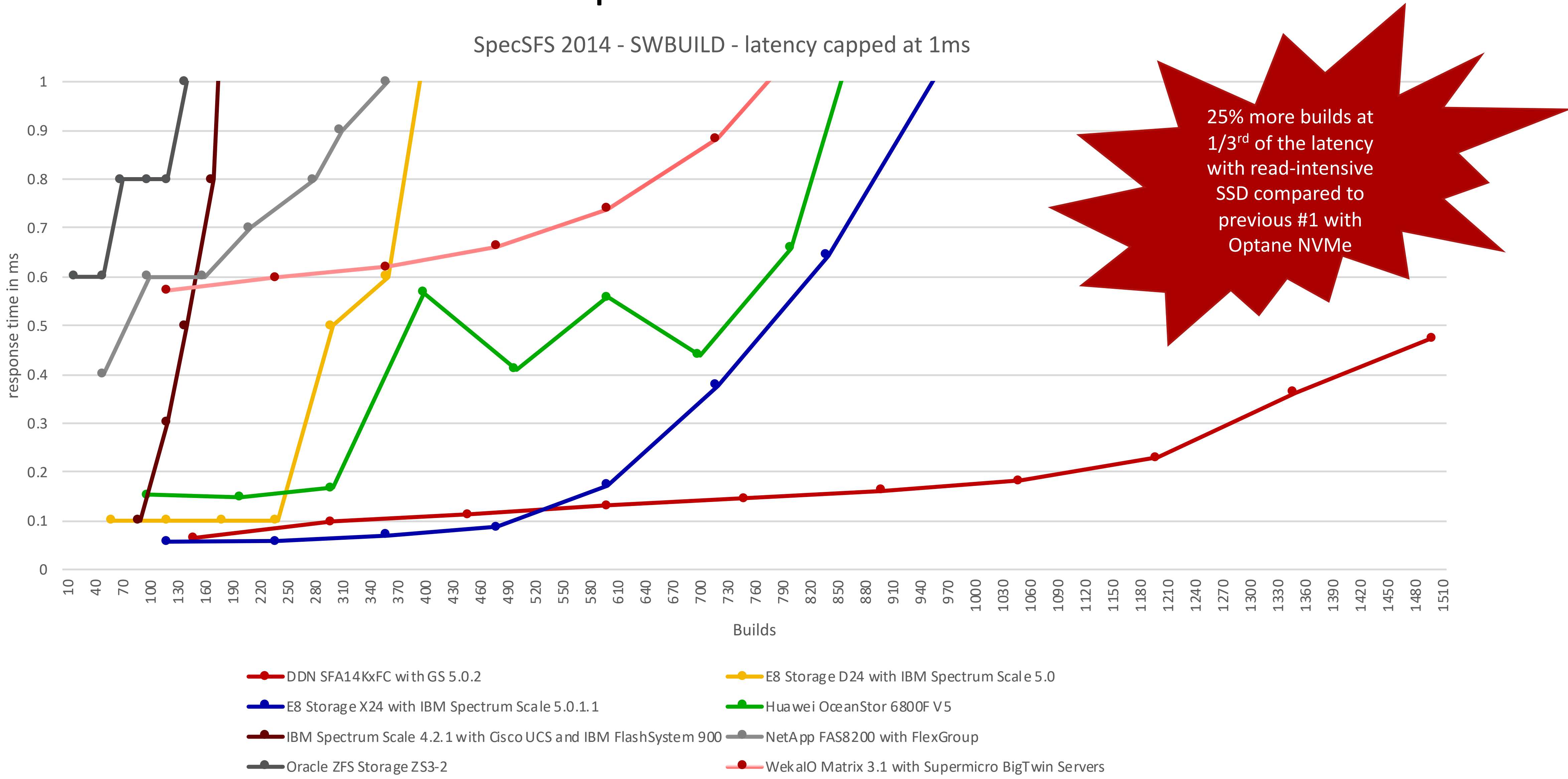
DDN SFA14KX with GRIDScaler

- With the SFA14KX and GRIDScaler parallel filesystem, DDN gains pole position for SPEC SFS
- DDN's SFA14KX running SFAOS with Declustered RAID and connecting to 6 GRIDScaler servers Sustains 25% more builds at 1/3rd of the Overall roundtrip latency with read-intensive SSD compared to previous #1 with Optane NVMeof - the next nearest competitor



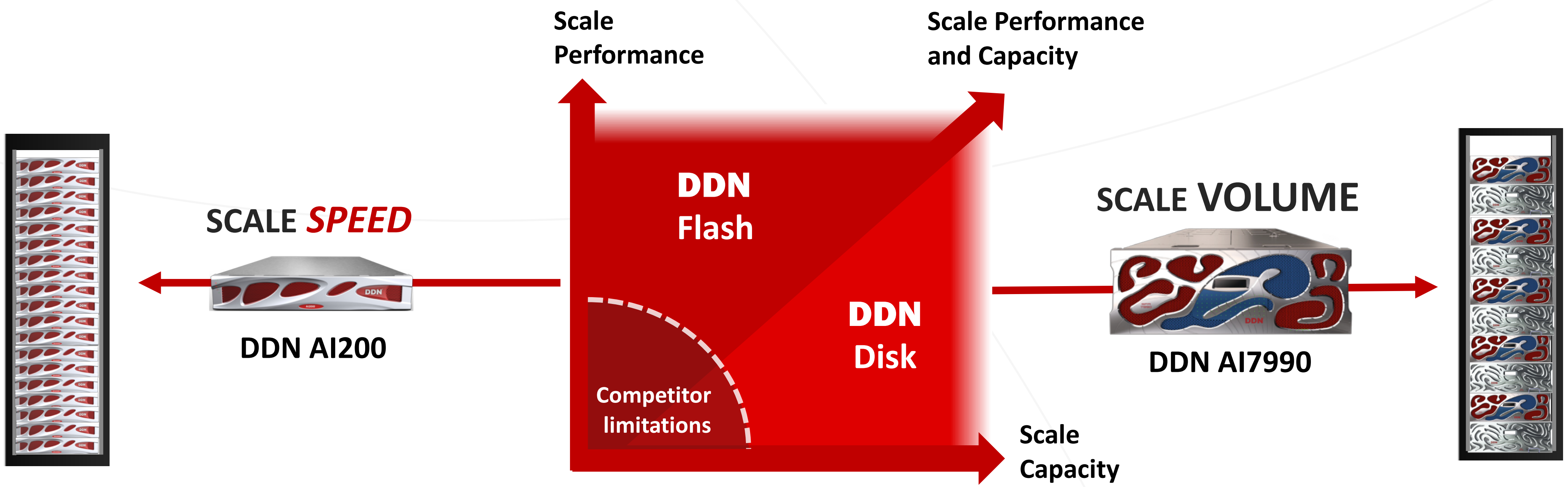
System Benchmarked for SPEC SFS

SpecSFS 2014 – SWBUILD compare to other vendors



DDN A^{3}I Solutions: Turnkey, integrated and optimized for NVIDIA DGX-1 and HP Apollo 6500

SCALE UP, SCALE OUT OR SCALE BOTH

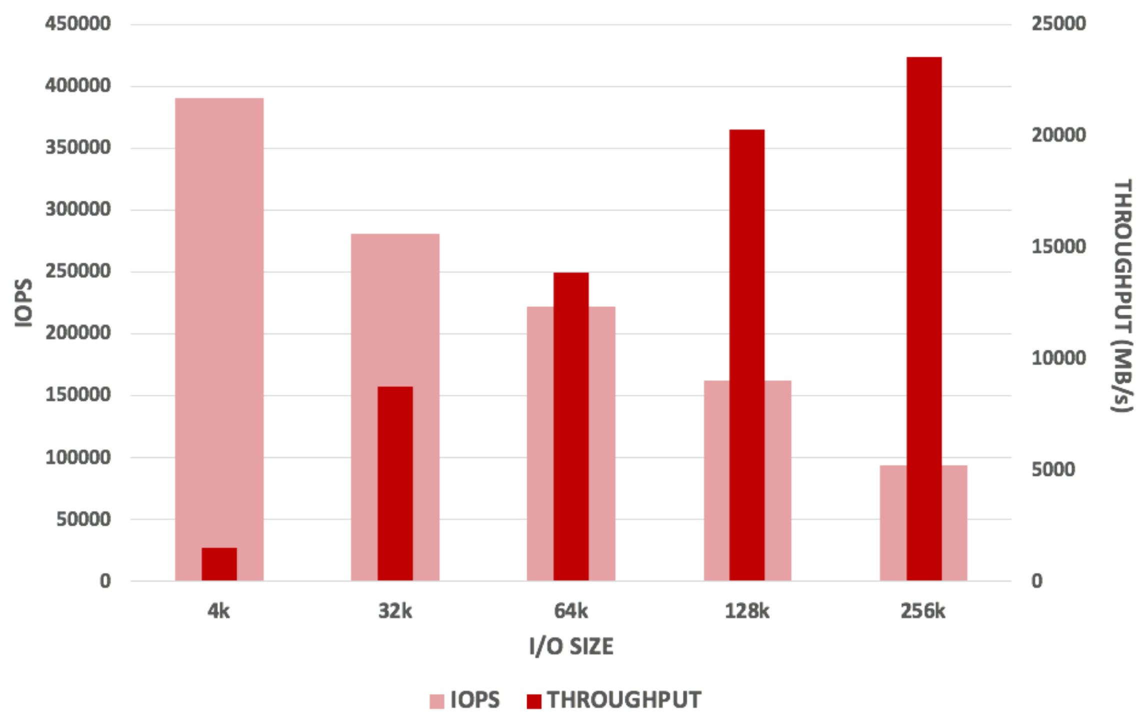


DDN A^{3}I SOLUTIONS TO A SINGLE CONTAINER ON DGX-1

23 GB/s and 395K IOPS to a single container*

DDN A^{3}I parallel storage client demonstrates over 23 GB per second and over 395K IOPS to a single container on DGX-1.

Typical deep learning codes perform IO using 128K size for which DDN delivers over 20 GB/s of sustained performance.



*numbers are with a single AI200 and was limited by client side performance of single DGX client

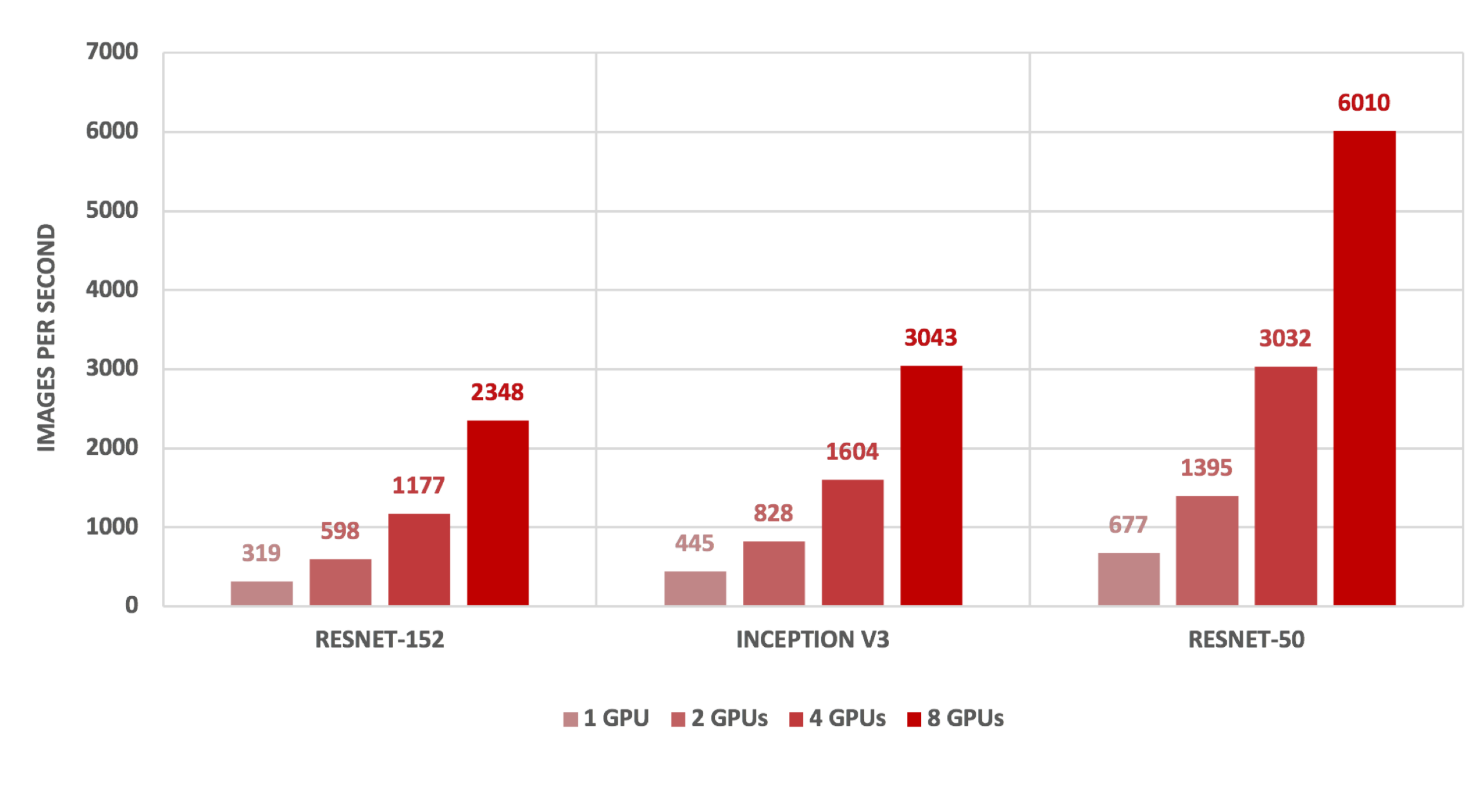
DDN A³I SOLUTIONS TENSORFLOW TRAINING PERFORMANCE

Fast, Consistent, Linear AI and DL Performance

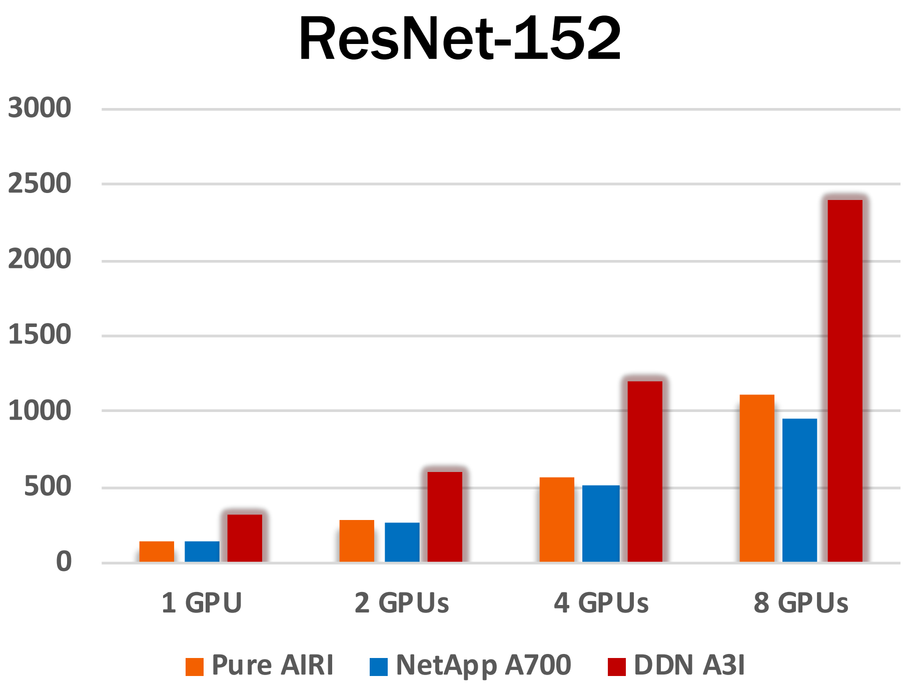
DL Training application performance scales linearly using multiple GPUs on DGX-1 with DDN parallel storage.

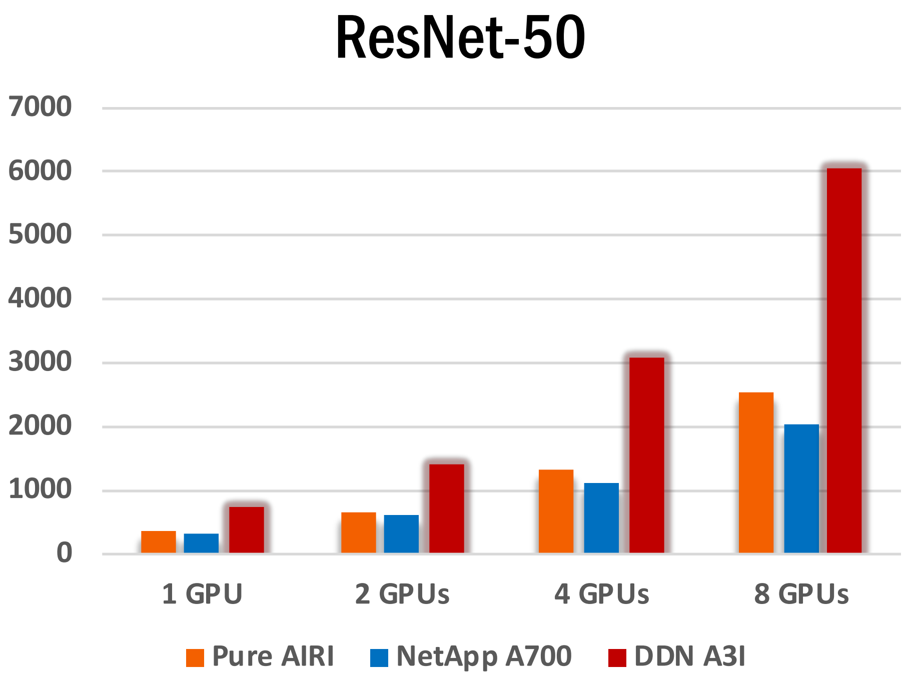
Parallel storage performance and shared architecture magnify end-to-end DL workflow acceleration.

Extensive application interoperability and performance testing has been engaged by DDN in close collaboration with NVIDIA and customers.



DDN A^{3}I SOLUTIONS LEADS PERFORMANCE FOR AI AND DL



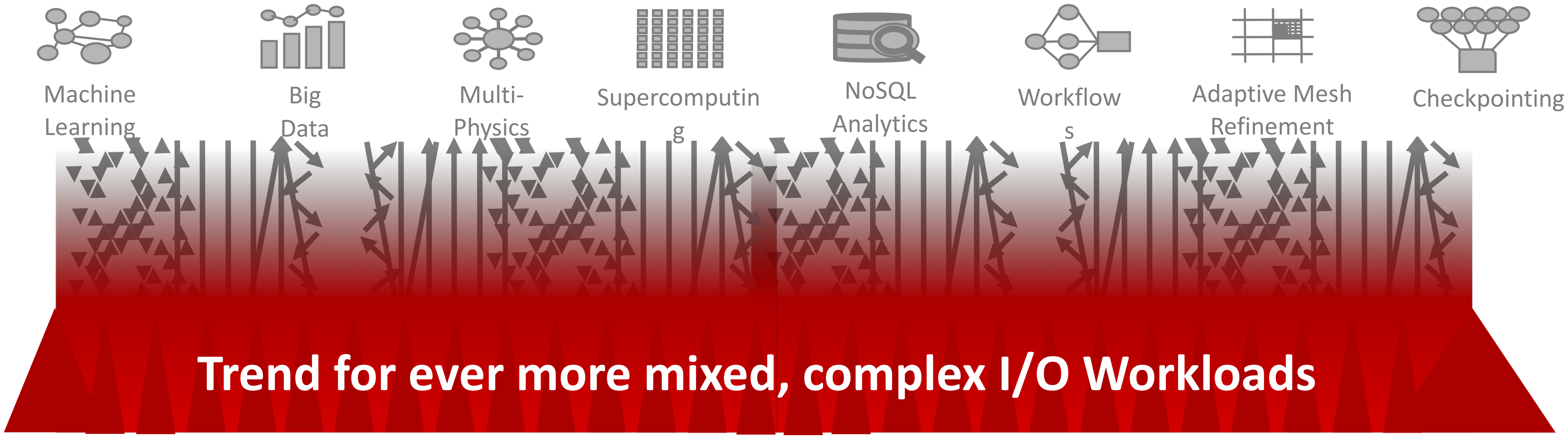


The Register® Biting the hand that feeds IT

"In the Resnet-152 and Resnet-50 tests, the AI200 tested faster than competing Pure, NetApp and Dell EMC systems."

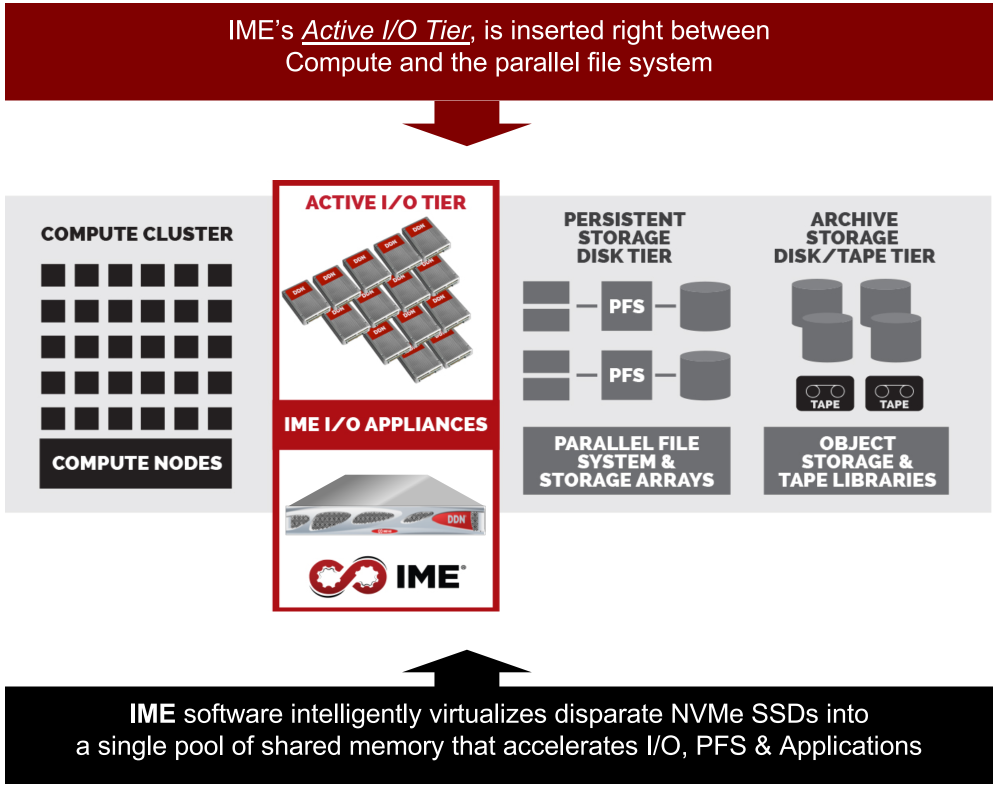
Challenges in I/O Performance and Behavior

- Newer applications need to operate on byte addressable data
- Significant shift from sequential to random I/O
- Multifold increase of metadata to data ratio
- Average data sizes are less homogeneous and are now fractions or multiples of previous workloads. gap between small and large data seems to wide (bytes on one end , GB's on the other end of the spectrum)
- Interactive, outcome and event driven analytics are driven by latency rather than bandwidth



WHAT IS IME?

- Scale-Out Flash Cache Layer using NVMe SSDs inserted between compute cluster and Parallel File System (PFS)
  - IME is configured as CLUSTER with multiple NVMe servers
  - All compute nodes can access cache data on IME
- Accelerates difficult IO patterns: small/random/shared file/high concurrency due to thin SW IO management layer
- configured as scale-out massive cache layer with huge IO bandwidth and IOPs

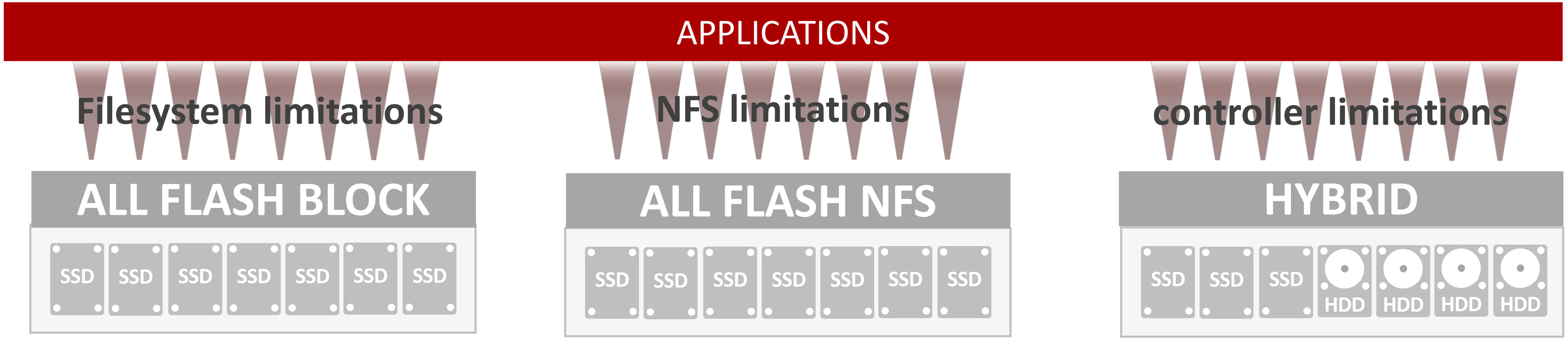


Expansion in Active Data Volumes requires a new economics for fast data at scale

All-Flash block doesn't solve the problem. **Block IOPs ≠ File IOPs**

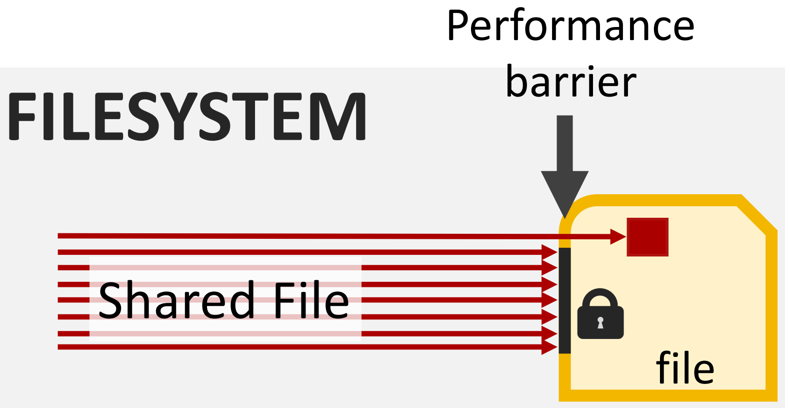
All-Flash NFS too slow and **too expensive** for real at-scale data problems

Traditional **Hybrid Approach doesn't enable flash at scale** – still limited by the storage controller



IME enables new levels of filesystem performance

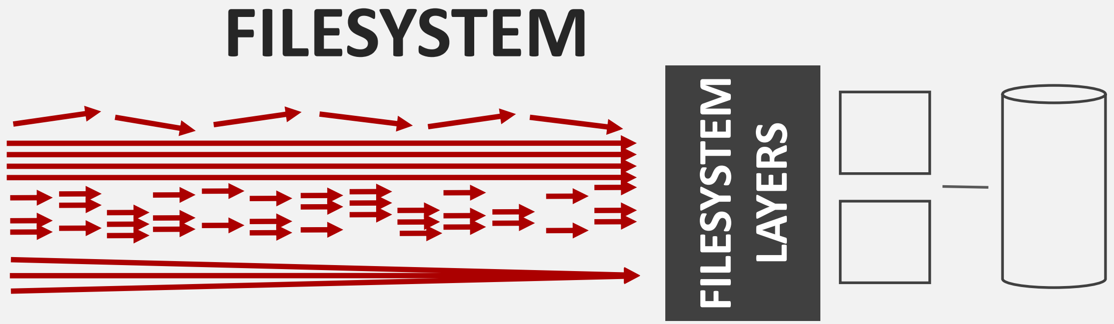
- Parallel File systems can exhibit extremely poor performance for shared file IO due to internal lock management as a result of managing files in large lock units



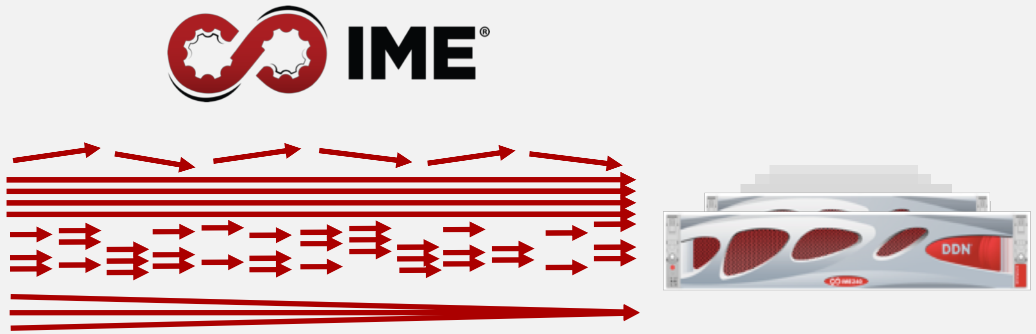
- IME eliminates contention by managing IO fragments directly, and coalescing IO's prior to flushing to the parallel file system

IME enables new levels of filesystem performance

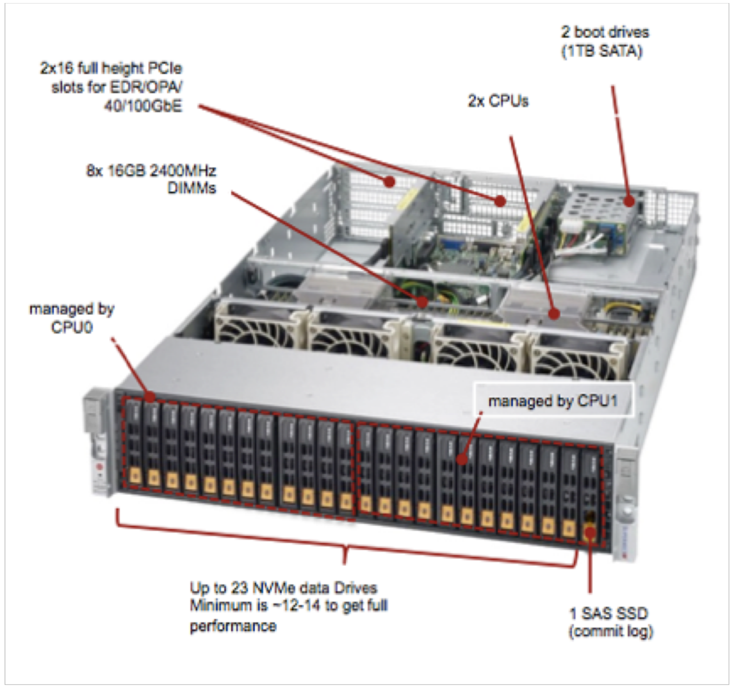
- Thick File system SW layers and traditional data layout severely restricts performance for tough workloads

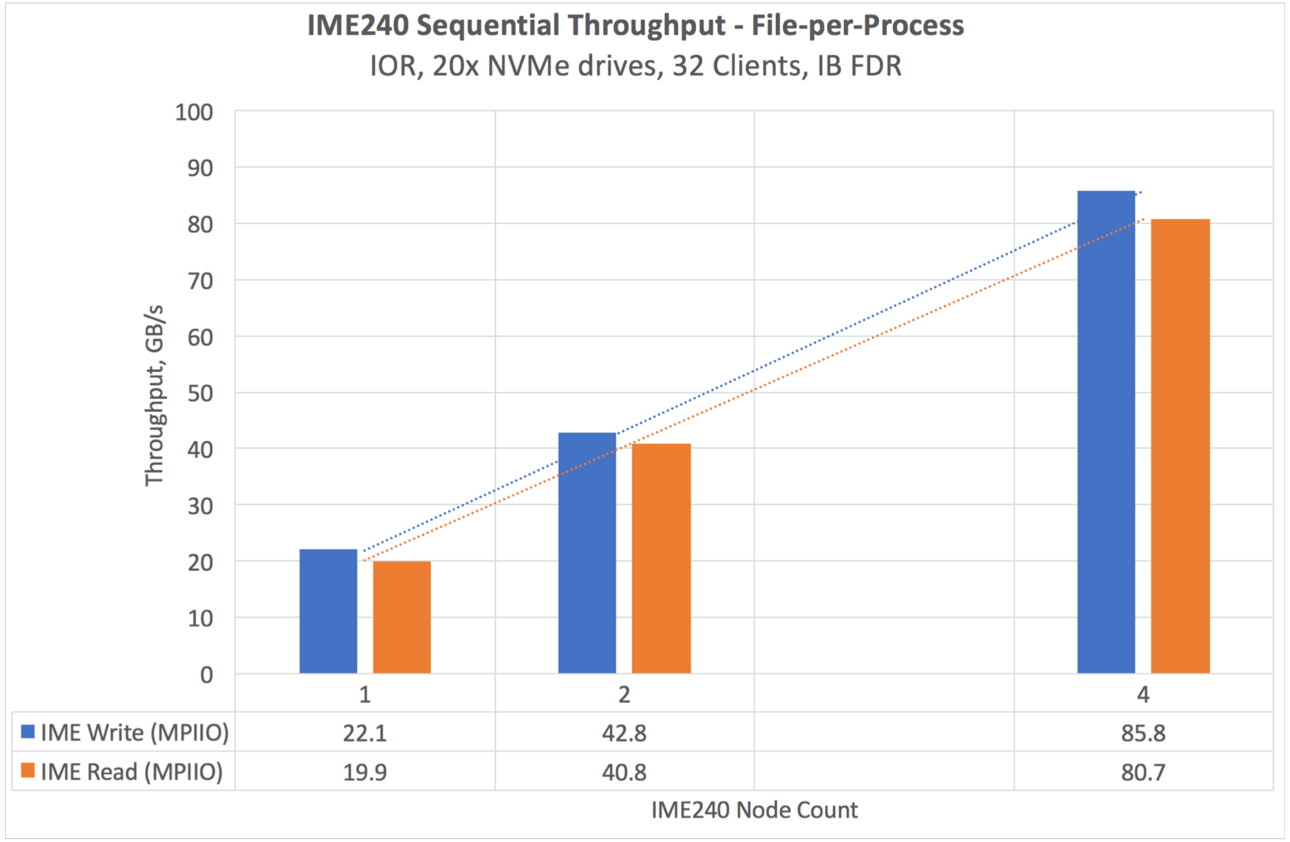


- IME's lean write anywhere, fully parallel IO completely removes the barriers that prevent your application seeing full performance



IME Performance Scalability & R/W Parity

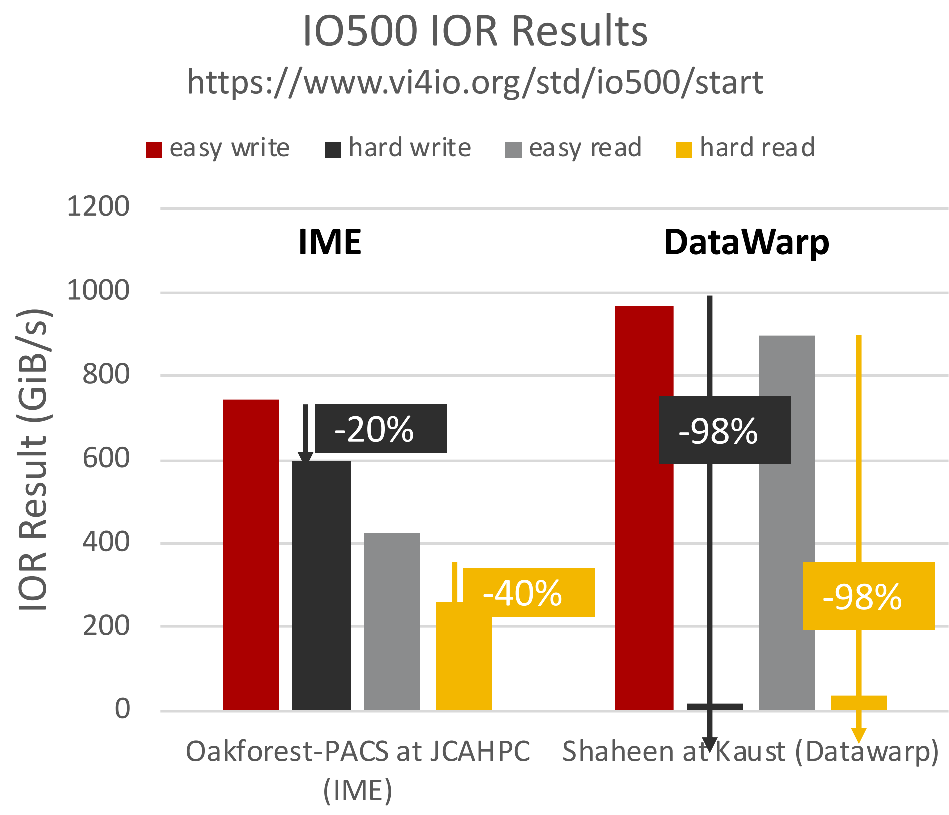




	1	2	4
IME Write (MPIIO)	22.1	42.8	85.8
IME Read (MPIIO)	19.9	40.8	80.7

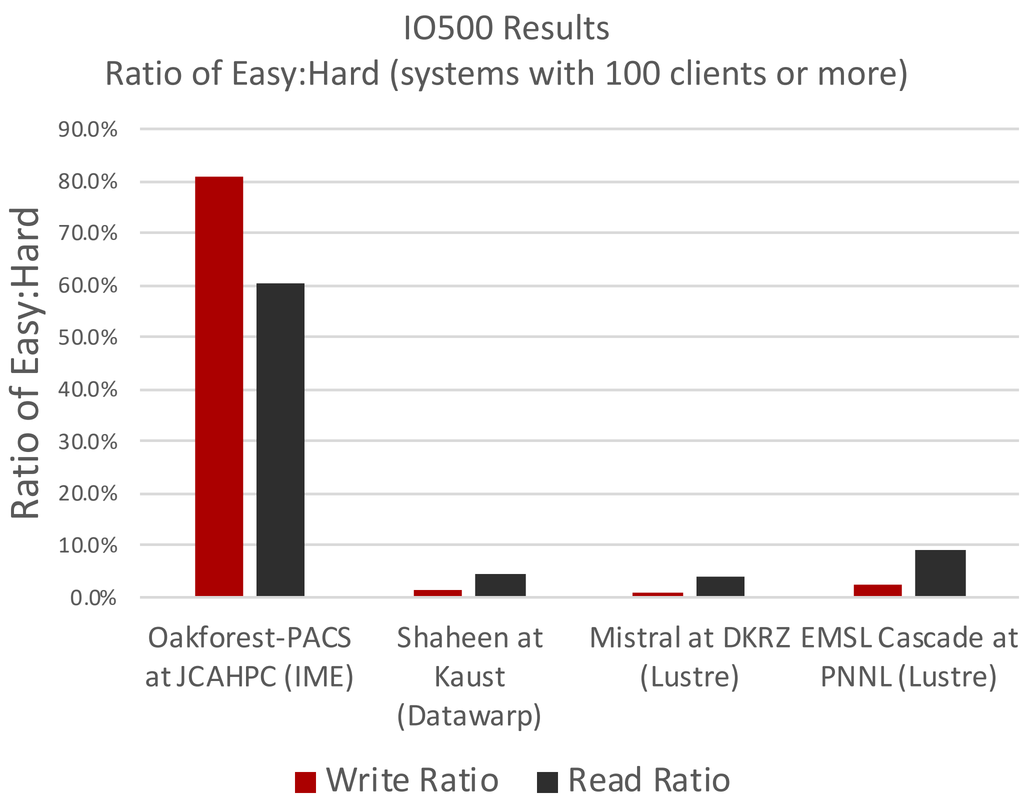
APPLICATION EFFICIENCY FOR THE REAL WORLD

- IME's datapath is designed to deliver the potential of flash to the application
- Other Burst Buffers use a conventional filesystem which severely limits the ability to deliver flash performance
- The IO500 uses "Easy" and "Hard" IOR benchmarks
  - IOR easy. You can set the parameters to be whatever you would like. You can use any of the modules such as HDF5 or MPI-IO. Typically people maximize performance by doing file-per-process and large aligned IO.
  - IOR hard. We enforce a particular set of parameters. Specifically, the IOs are 47008 bytes each interspersed in a single shared file. Your only control is to specify how many writes each thread does.
- ***Anyone can get good performance with enough equipment with the easy benchmark. Good Performance with the Hard Benchmark requires a new approach***



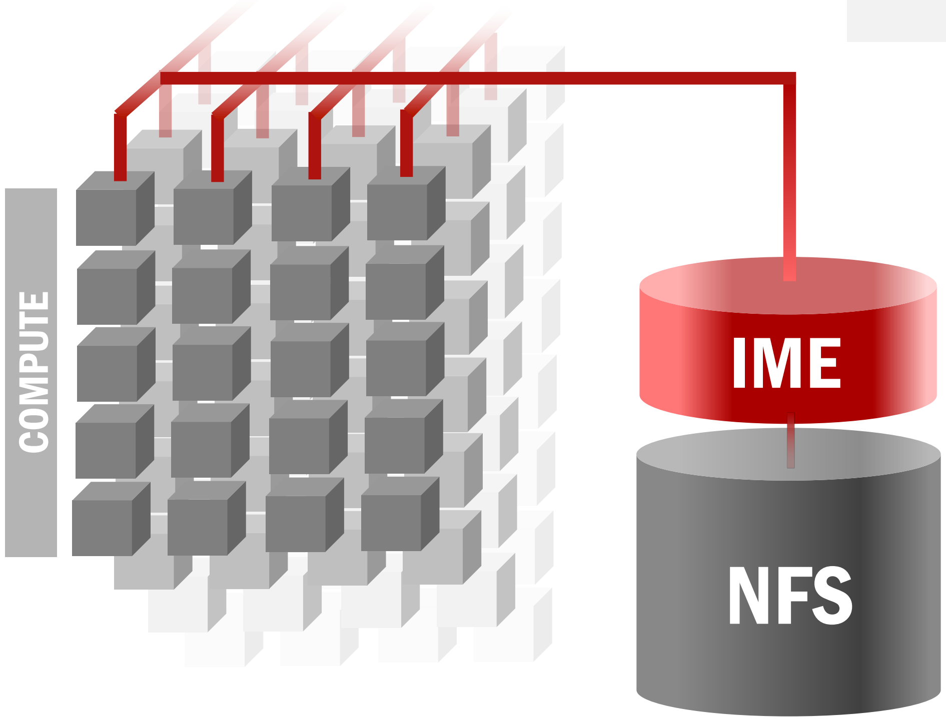
APPLICATION EFFICIENCY FOR THE REAL WORLD

- Extracting results from IO500 where the client count is 100 nodes or more
- Filesystem options show huge degradation when the IO patterns is tough.
- Only IME is able to present Flash to the applications efficiently



IME – Burst buffer for NFS

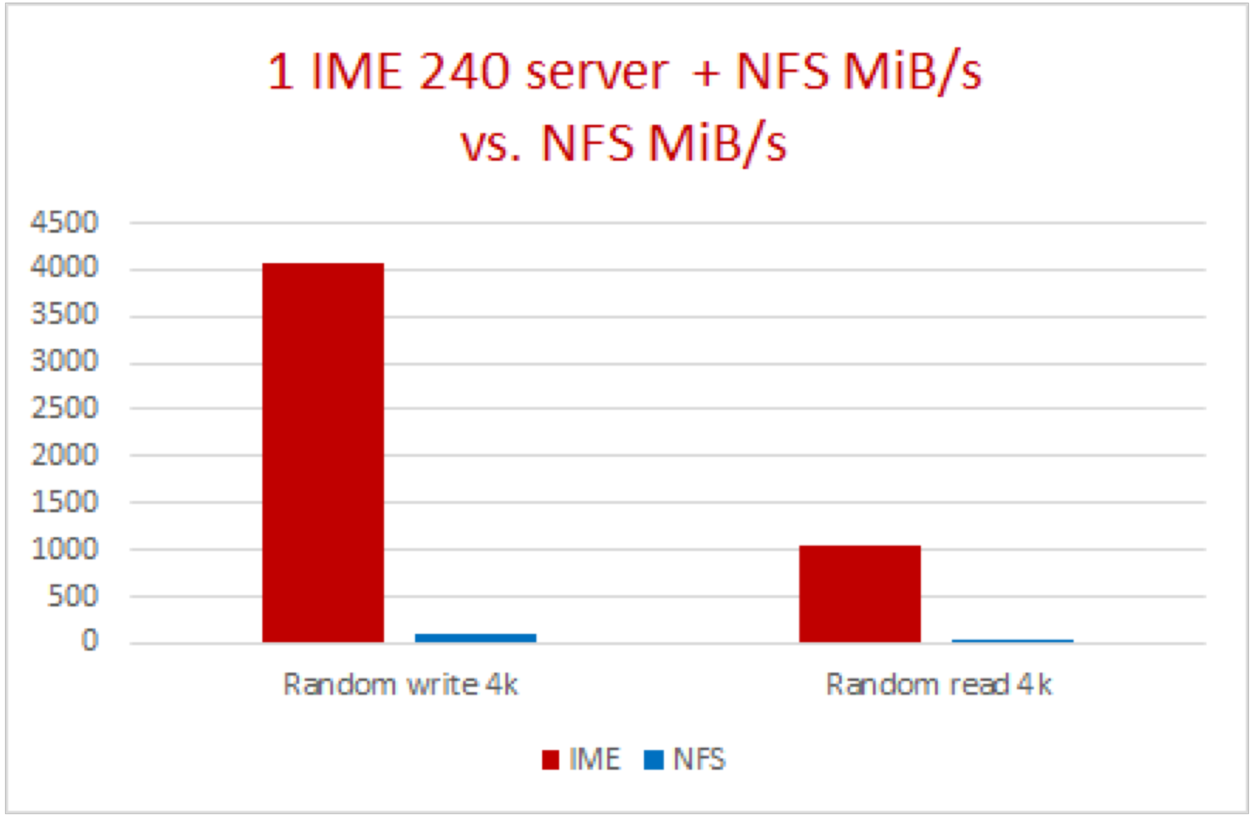
- Brings scale-out Flash native performance to NFS access
- Shield NFS server from ”tough" IO
- Increase IO throughput from NFS hardware
- Zero application changes - replace NFS mount by IME mount



IME – Burst buffer for NFS

IME with NFS

- Brings scale out Flash native performance to NFS Systems
- Removes complexity associated with Parallel Filesystems
- Shield NFS server for "bad" IO
- Increase IO throughput on top of NFS hardware
- No application changes - replace NFS mount by IME mount



DDN.COM/A3I

Thank You!

Keep in touch with us.

sales@ddn.com

@ddn_limitless

company/datadirect-networks

9351 Deering Avenue
Chatsworth, CA 91311

1.800.837.2298
1.818.700.4000