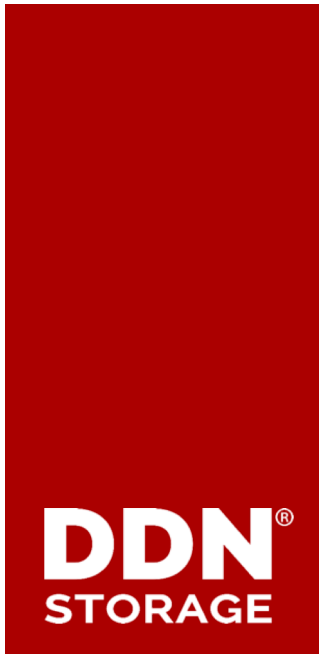


Optimizing storage stacks for AI

UK and World Wide Spectrum Scale User Group Meeting

May, 2019

Sven Oehme – Chief Research Officer DDN

DDN SFA | ALL-FLASH AND HYBRID BLOCK STORAGE PLATFORMS

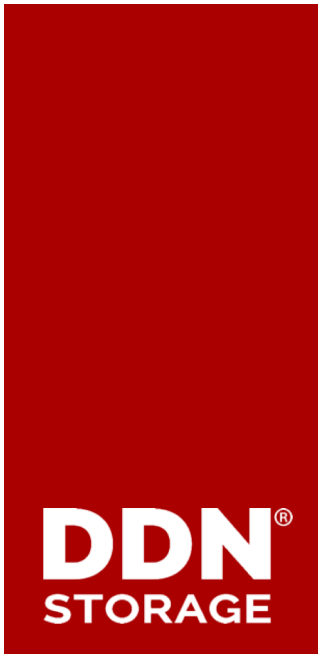
200NV	400NV	7990	14KX	18K
				
23GB/s 1M IOP/s	42GB/s 3M IOP/s	23GB/s 1M IOP/s	60GB/s 4M IOP/s	92GB/s 4M IOP/s
24 NVME Slots	24 NVME Slots	Up to 450 SSD/HDD	48 NVMe Slots Up to 1872 SSD/HDD	48 NVMe Slots Up to 1872 SSD/HDD
EDR IB (4), OPA (2) FC32 (8), FC (8)	EDR IB (8), OPA (4)	EDR IB (4), OPA (2) FC16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

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
- GRIDScaler V5 automatically bundles with RHEL 7

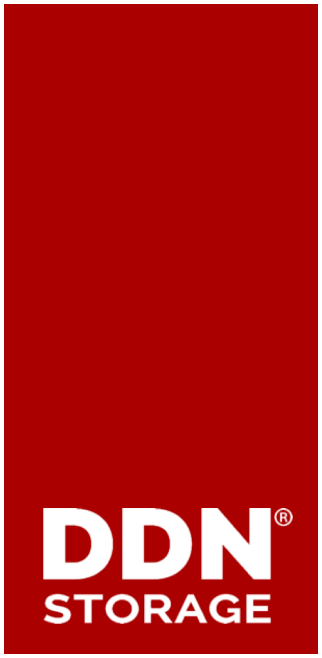


Optimizations for GRIDScaler

Optimizations for GRIDScaler

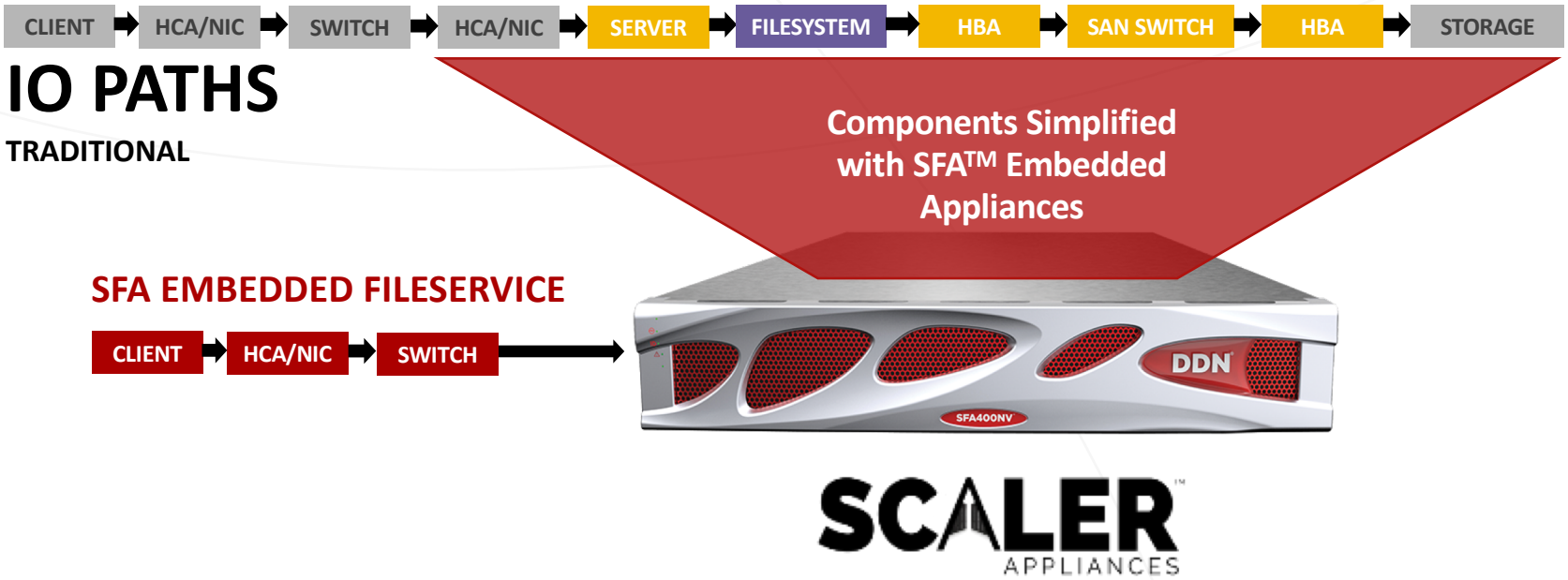
- Updated device drivers, OS and Scale tuning parameters and SFA multi-queue LUN support
- SFAOS Level enhancements for Large Block Support
- Embedded 14kx systems can now achieve > 1.5 Million random 4k read IOPS
- External SFA14KX NSD Server performance went from 1.25 Million to 2.96 Million*
- Some of 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

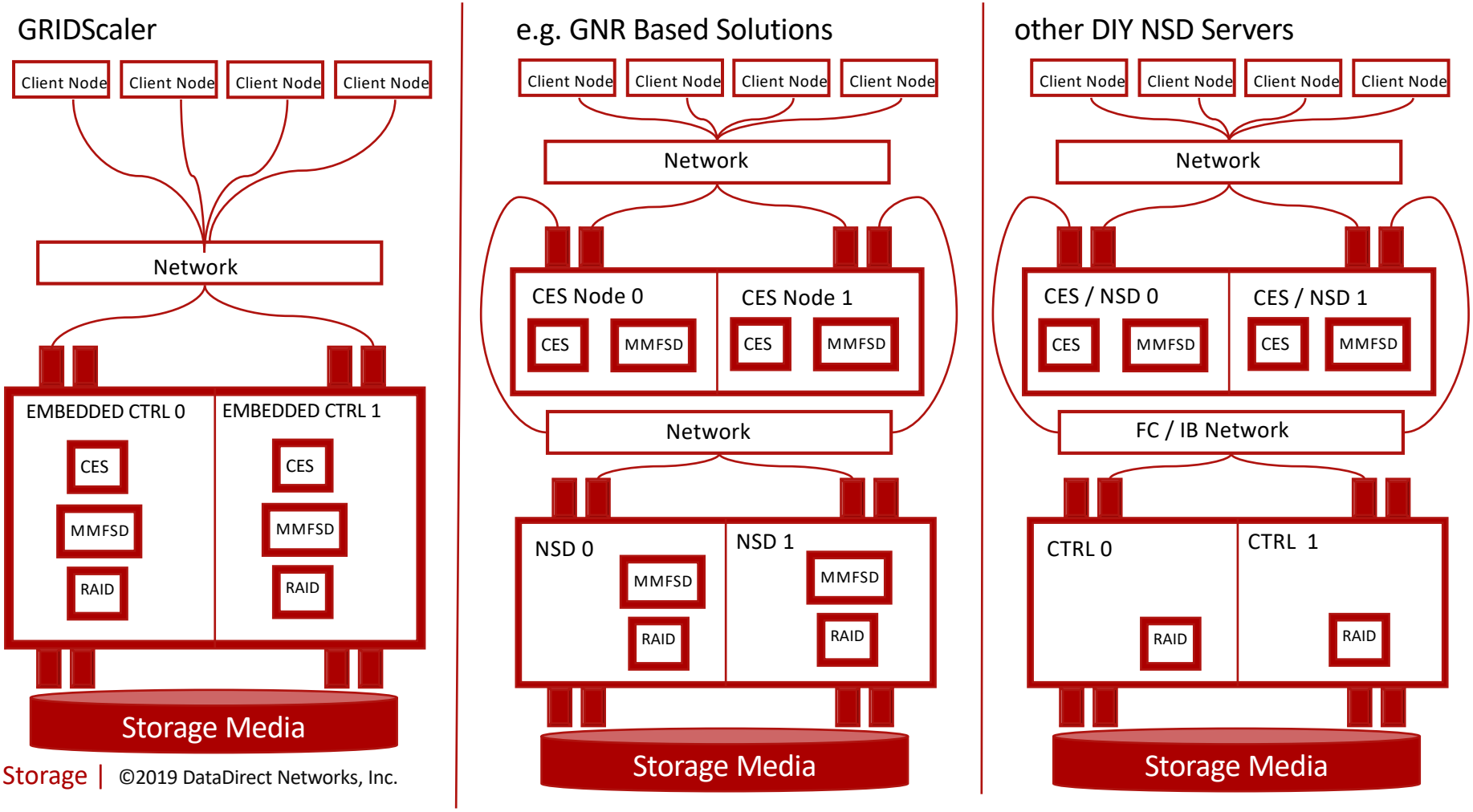


Platform Optimizations help significantly

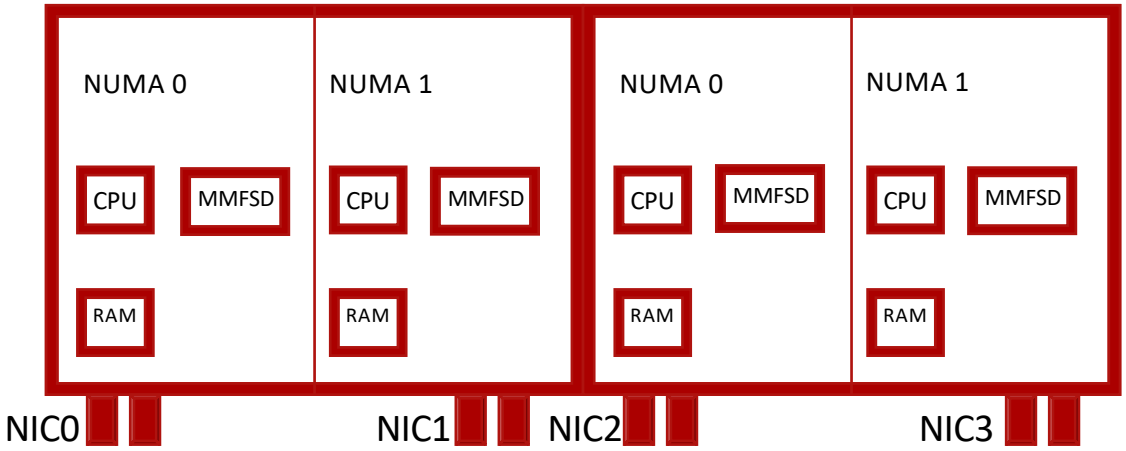
LOW LATENCY DESIGNED-IN



Collaps of layers improves simplicity and performance



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

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

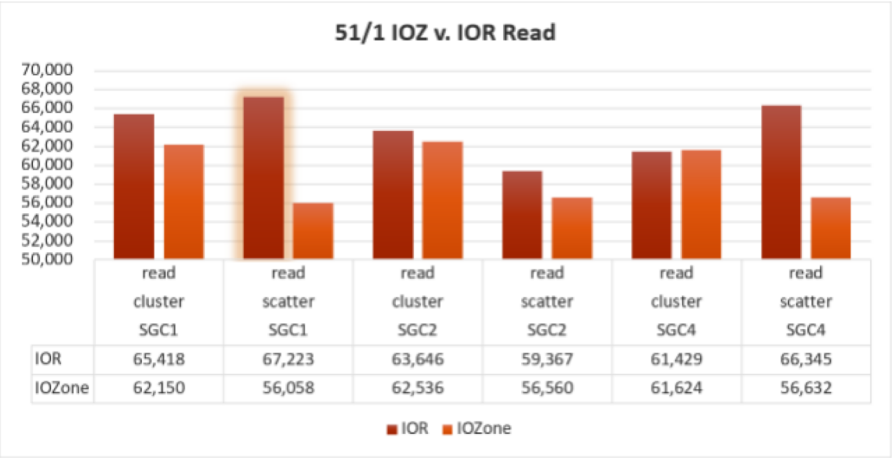
DIO Random 4k writes into a 100GB files

```
/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
```

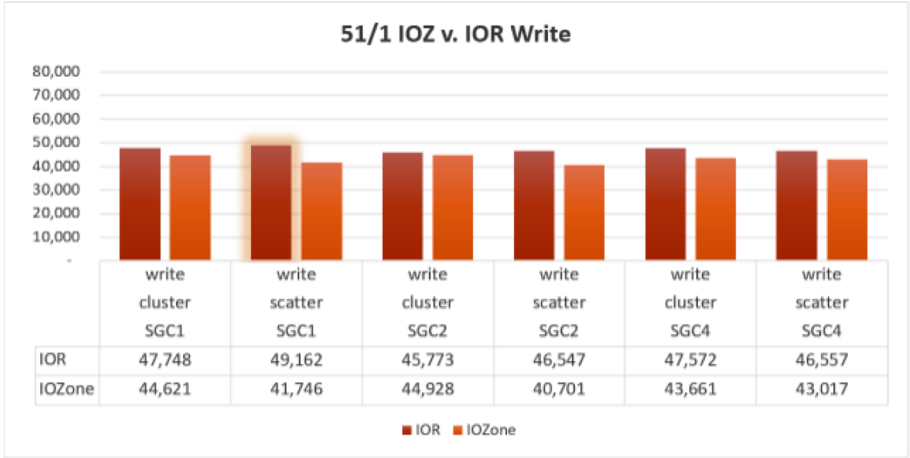
DIO Random 4k reads from a 1TB files (exceeds all cache by >4x)

```
/usr/lpp/mmfs/samples/perf/gpfsperf read rand /target/testfile-1t
  recSize 4K nBytes 1024G fileSize 1024G
  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 27982.49 Kbytes/sec, Op Rate was 6831.66 Ops/sec, Avg Latency
was 0.146 milliseconds, thread utilization 1.000, bytesTransferred 279830528
```

18k results with GRIDScaler V5 and SFAOS 11.5*



	read cluster SGC1	read scatter SGC1	read cluster SGC2	read scatter SGC2	read cluster SGC4	read scatter SGC4
IOR	65,418	67,223	63,646	59,367	61,429	66,345
IOZone	62,150	56,058	62,536	56,560	61,624	56,632



	write cluster SGC1	write scatter SGC1	write cluster SGC2	write scatter SGC2	write cluster SGC4	write scatter SGC4
IOR	47,748	49,162	45,773	46,547	47,572	46,557
IOZone	44,621	41,746	44,928	40,701	43,661	43,017

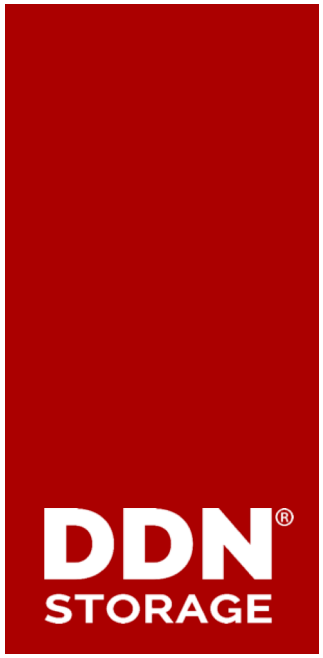
System Setup is a single SFA18K with 408 HDD's running SFAOS 11.5, with 8 Pools, each 51/1 RAID 6 with 2MB chunk size

We tested with 16 clients connected via single port EDR cable, data set size was >10x of all combined caches in each run

This translates in ~150 MB/sec per individual HDD for full Block random Read/writes in scatter mode

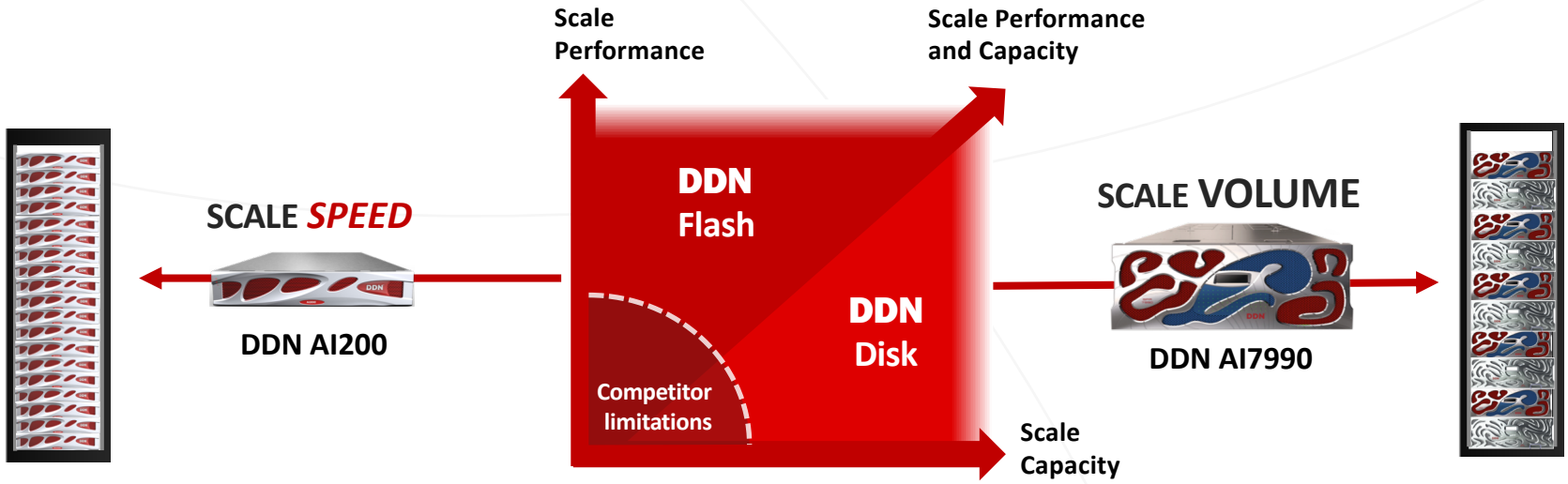
~150MB/sec/ HDD

*SFAOS 11.5 GA is June 2019



DDN A³I Solutions: Turnkey, integrated and optimized for NVIDIA DGX-1 and HP Apollo 6500

SCALE UP, SCALE OUT OR SCALE BOTH

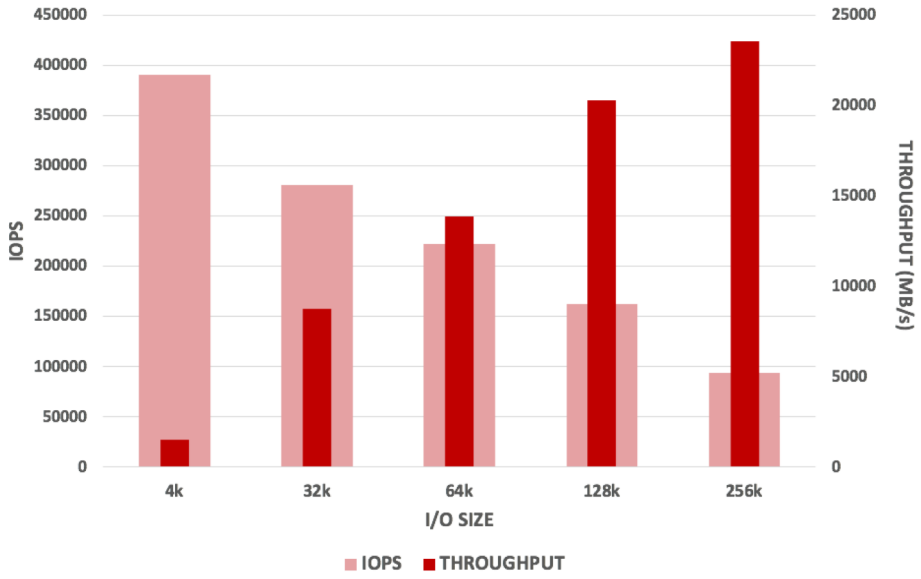


DDN A³I SOLUTIONS TO A SINGLE CONTAINER ON DGX-1

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

DDN A³I 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

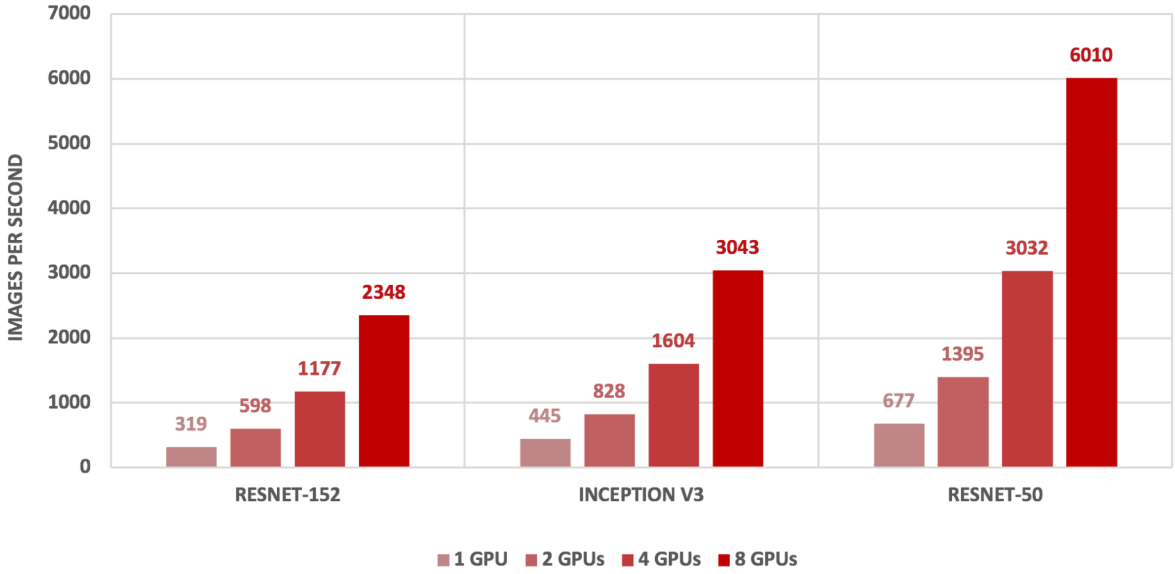
DDN A³I SOLUTIONS TENSORFLOW TRAINING PERFORMANCE

Fast, Consistent, Linear AI and DL Performance

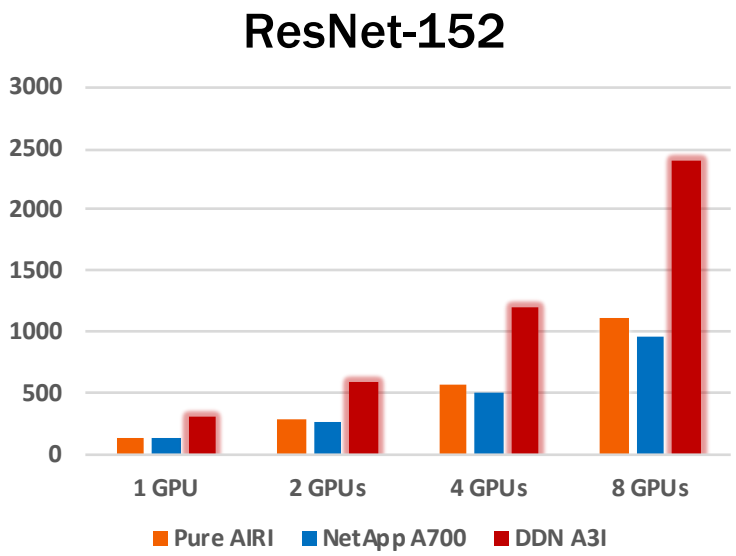
DL Training application performance scales linearly using multiple GPUs on DGX-1 with DDN parallel storage.

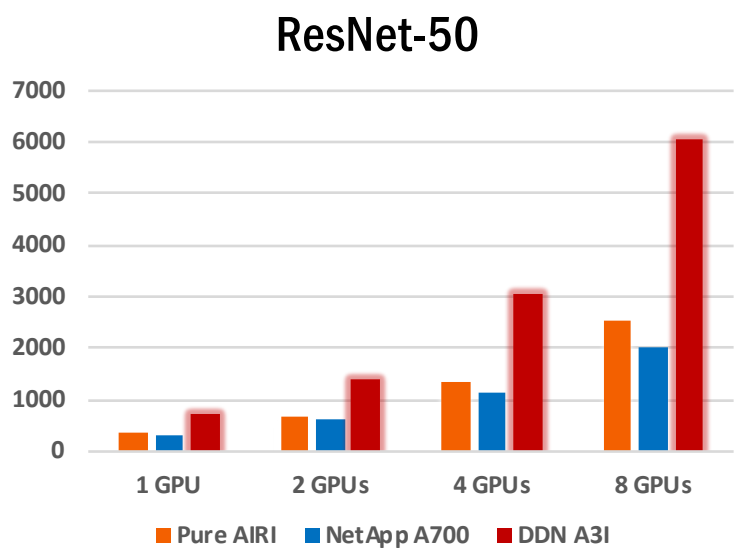
Parallel storage performance and shared architecture magnify end-to-end DL workflow acceleration.

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



DDN A³I SOLUTIONS LEADS PERFORMANCE FOR AI AND DL

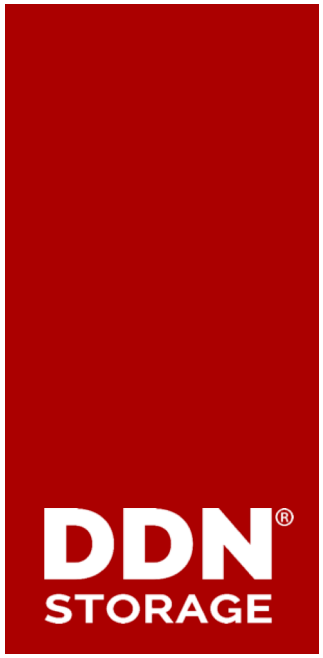




The Register®
Biting the hand that feeds IT

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

DDN.COM/A3I



World record SpecSFS 2014 with GRIDScaler*

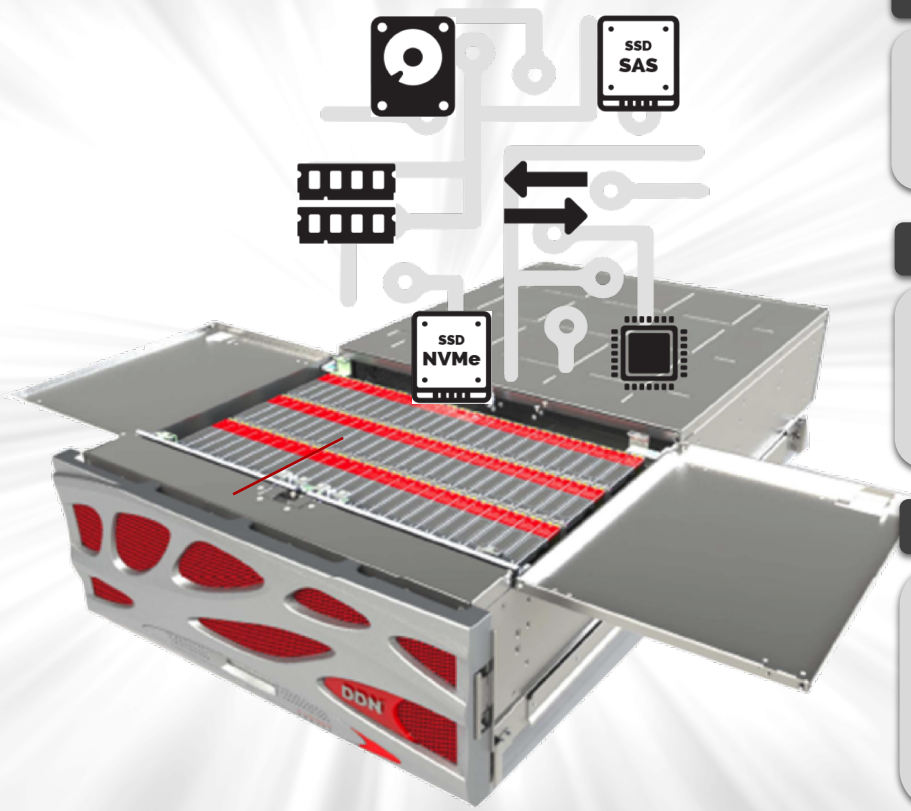
*world record has been broken with an 8 Storage System setup - we use ONE !

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

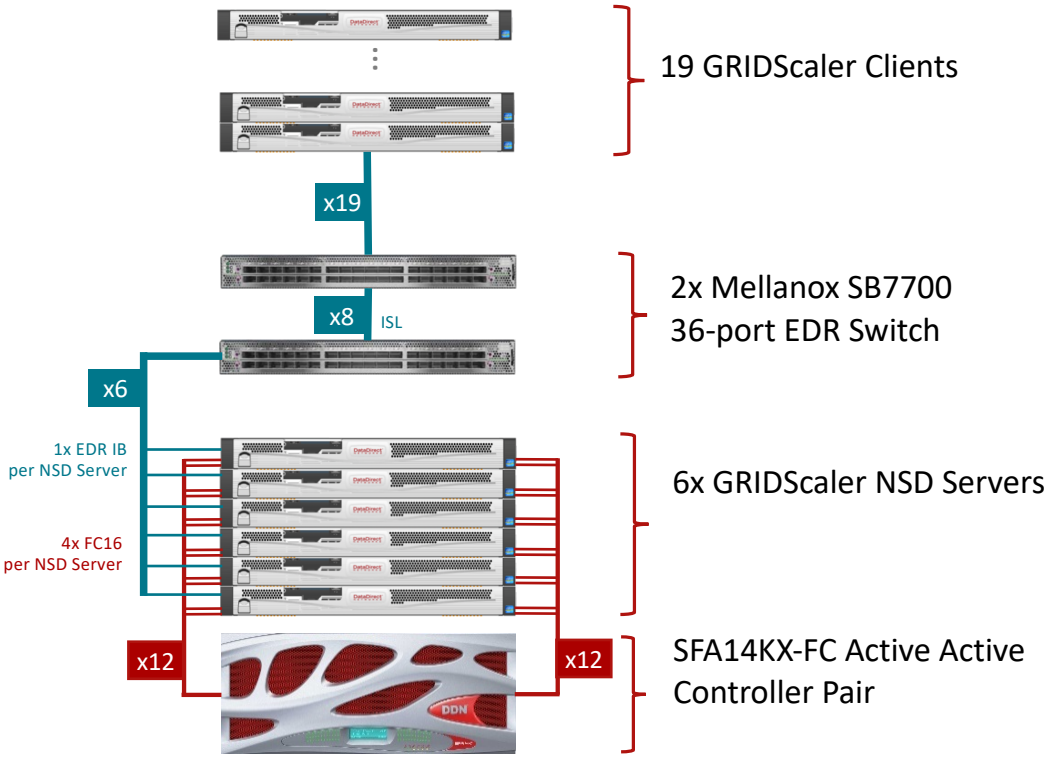
- ▶ 72 SAS SSD or 48 NVMe
- ▶ Up to 60 GB/sec throughput
- ▶ Up to 4 million IOPS

Best Data Protection

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

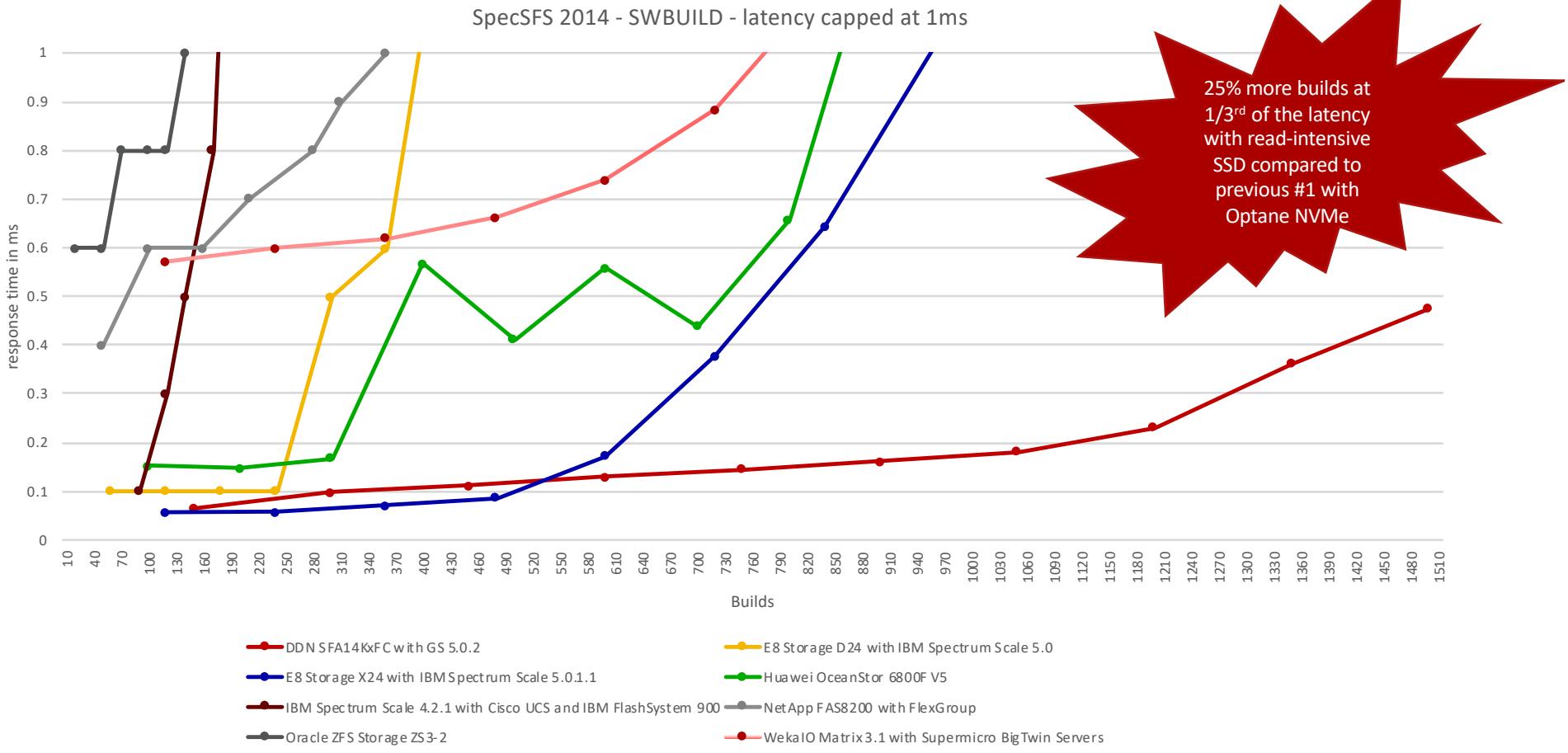
DDN SFA14KX with GRIDScaler

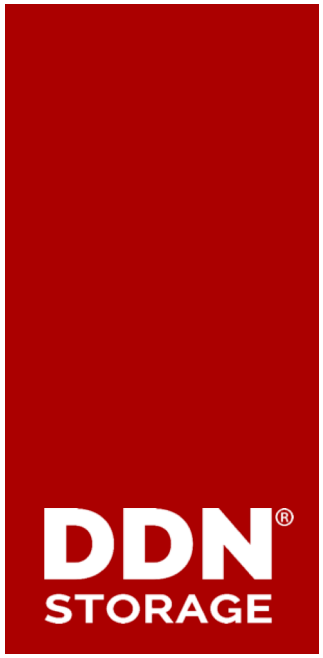
- With the SFA14KX and GRIDScaler parallel filesystem, DDN gains pole position for SPEC SFS
- DDN's SFA14KX running SFAOS with Declustered RAID and connecting to 6 GRIDScaler servers Sustains 25% more builds at 1/3rd of the Overall roundtrip latency with read-intensive SSD compared to previous #1 with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



System Benchmarked for SPEC SFS

SpecSFS 2014 – SWBUILD compare to other vendors

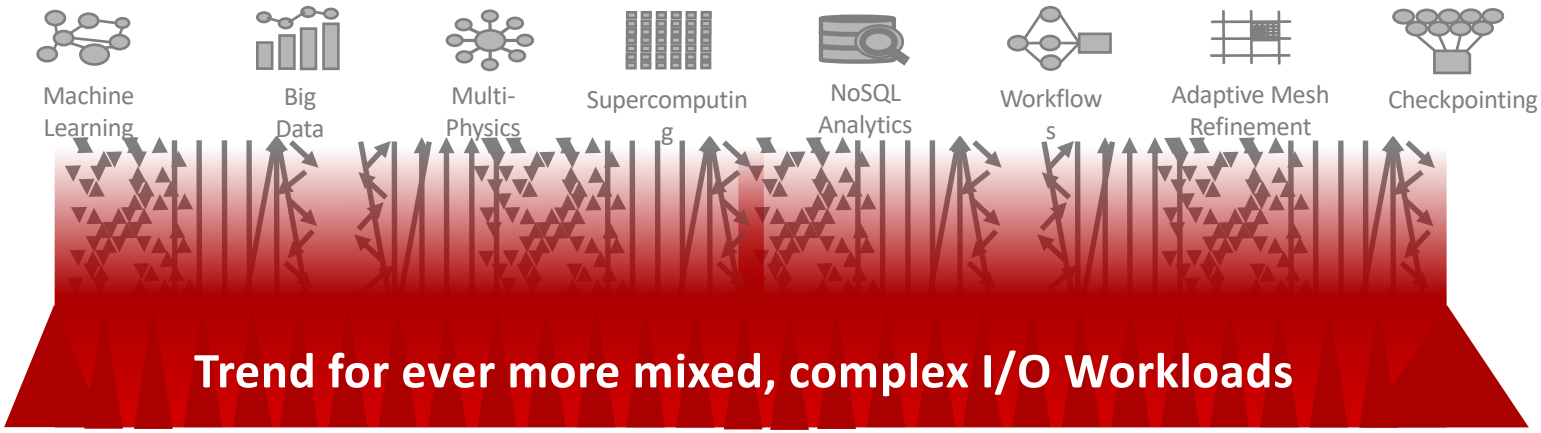




DDN IME - Workload Acceleration

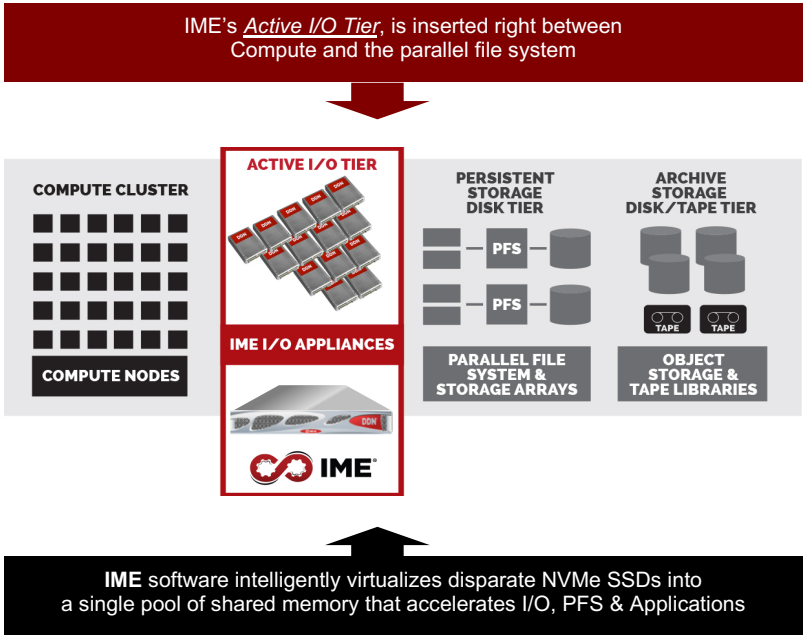
Challenges in I/O Performance and Behavior

- 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

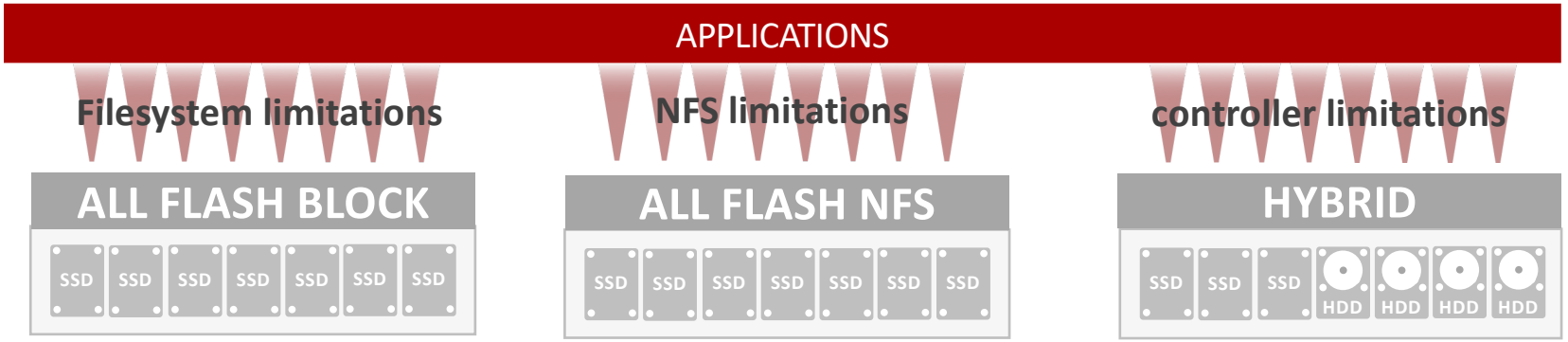


Expansion in Active Data Volumes requires a new economics for fast data at scale

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

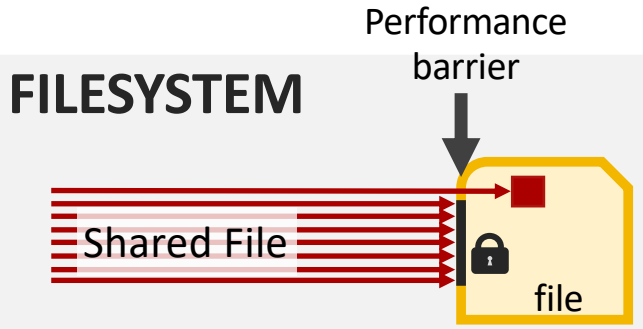
All-Flash NFS too slow and **too expensive** for real at-scale data problems

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



IME enables new levels of filesystem performance

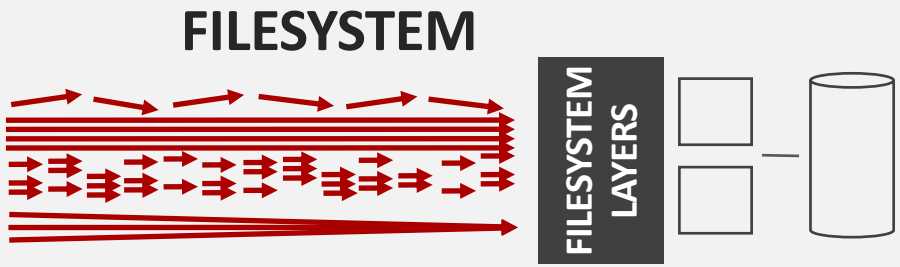
- Parallel File systems can exhibit extremely poor performance for shared file IO due to internal lock management as a result of managing files in large lock units



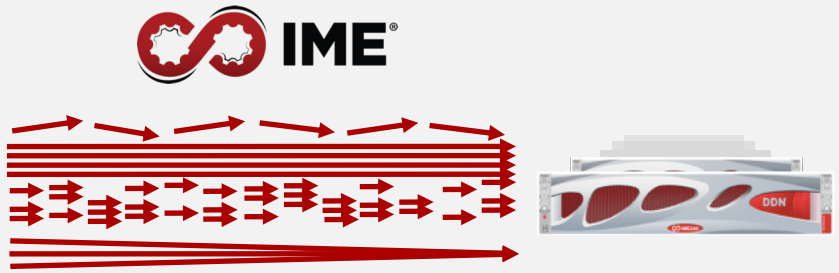
- IME eliminates contention by managing IO fragments directly, and coalescing IO's prior to flushing to the parallel file system

IME enables new levels of filesystem performance

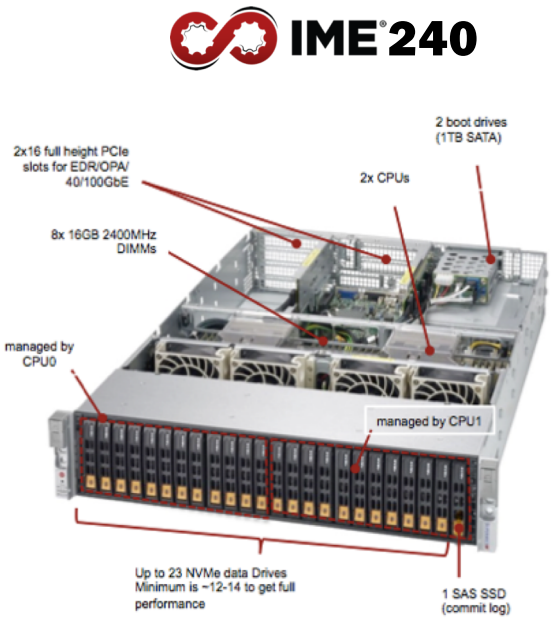
- Thick File system SW layers and traditional data layout severely restricts performance for tough workloads

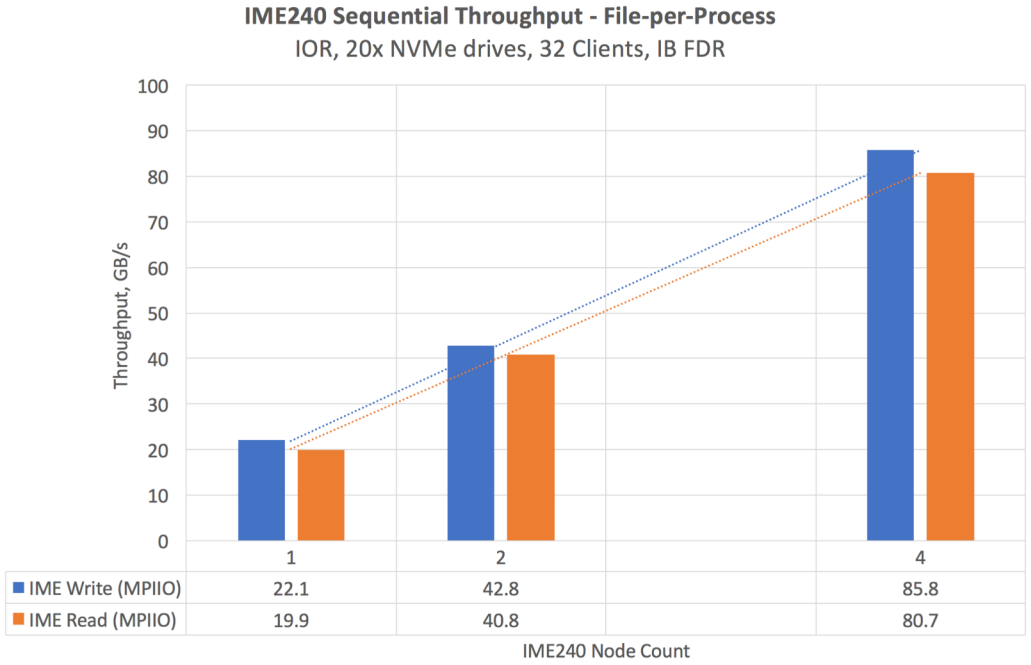


- IME's lean write anywhere, fully parallel IO completely removes the barriers that prevent your application seeing full performance



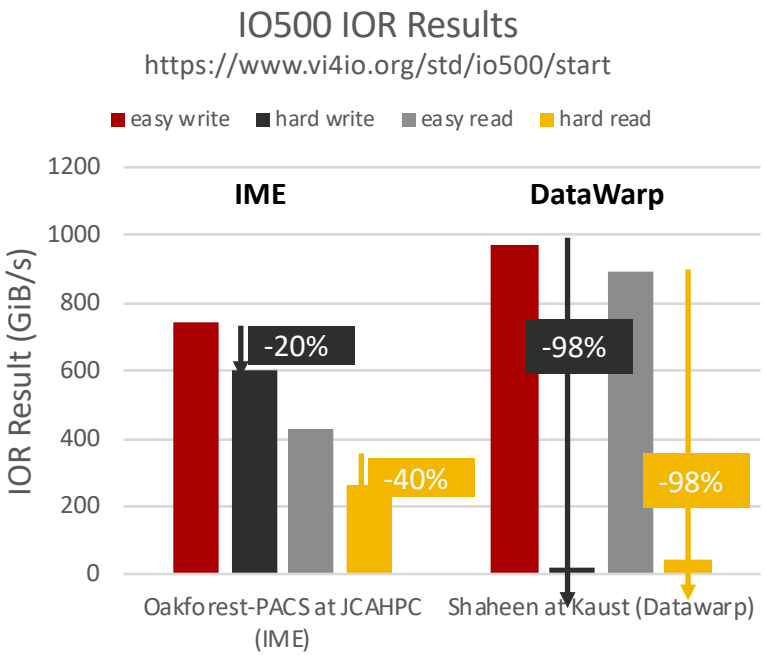
IME Performance Scalability & R/W Parity





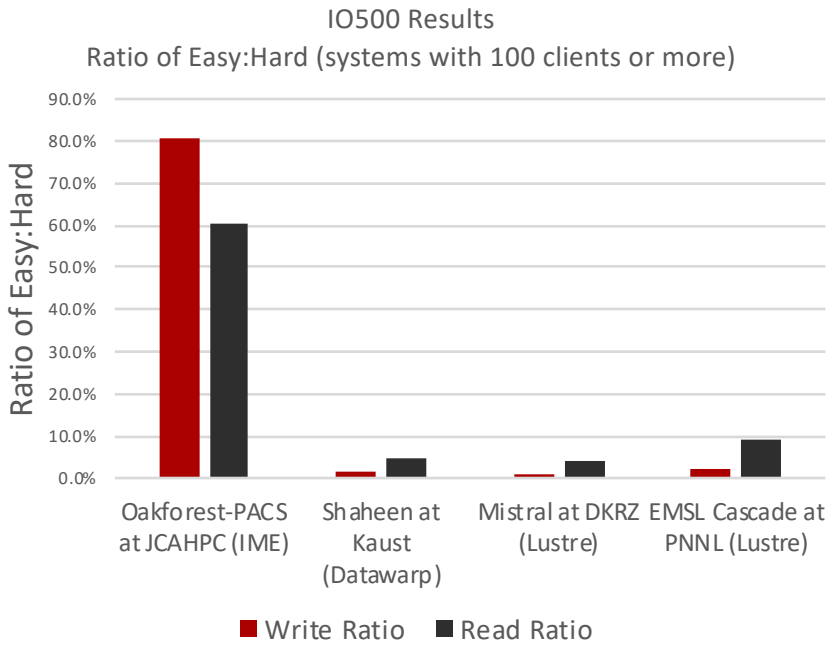
APPLICATION EFFICIENCY FOR THE REAL WORLD

- 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***



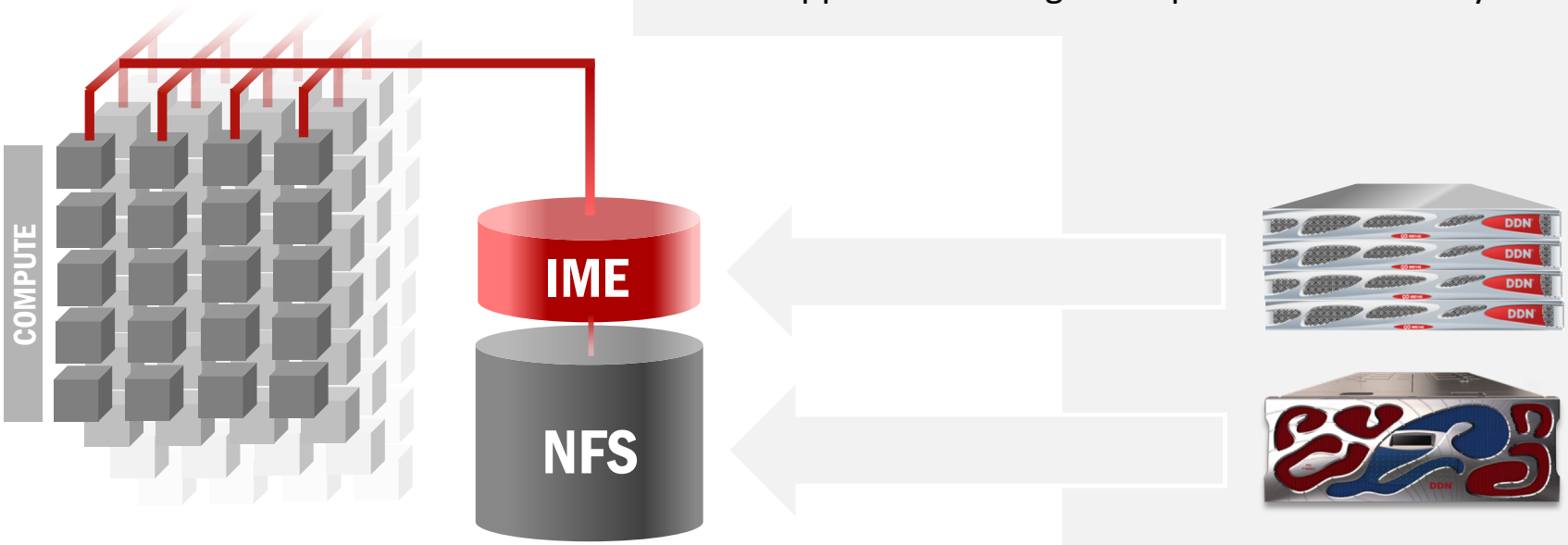
APPLICATION EFFICIENCY FOR THE REAL WORLD

- 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

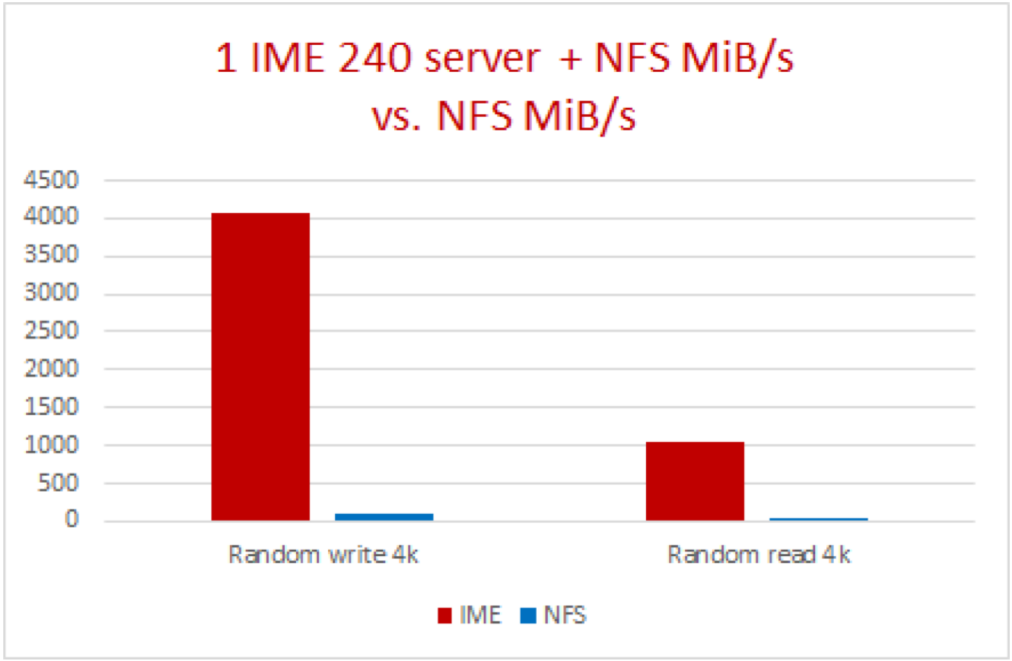
- Brings scale-out Flash native performance to NFS access
- Shield NFS server from ”tough" IO
- Increase IO throughput from NFS hardware
- Zero application changes - replace NFS mount by IME mount

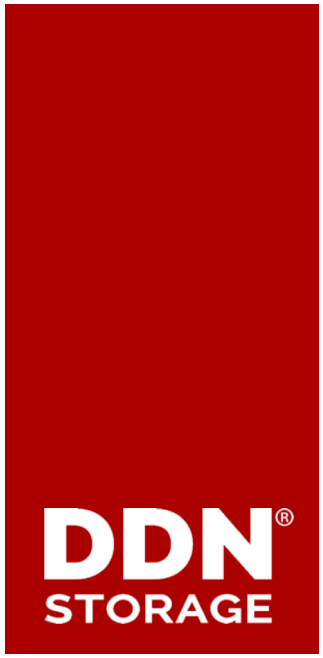


IME – Burst buffer for NFS

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





Dataflow – Hybrid Cloud data management at scale

DDN DATAFLOW MIGRATION MIGRATION OPTIONS

ACTUAL MIGRATION OPTIONS

Vendor proprietary
- Vendor exclusive
- Dedicated solution
- Mostly do not interact with other vendors

Robocopy/ Rsync
- Free tools
- Suitable if not specific constrains
- Mastering of tools and scripting

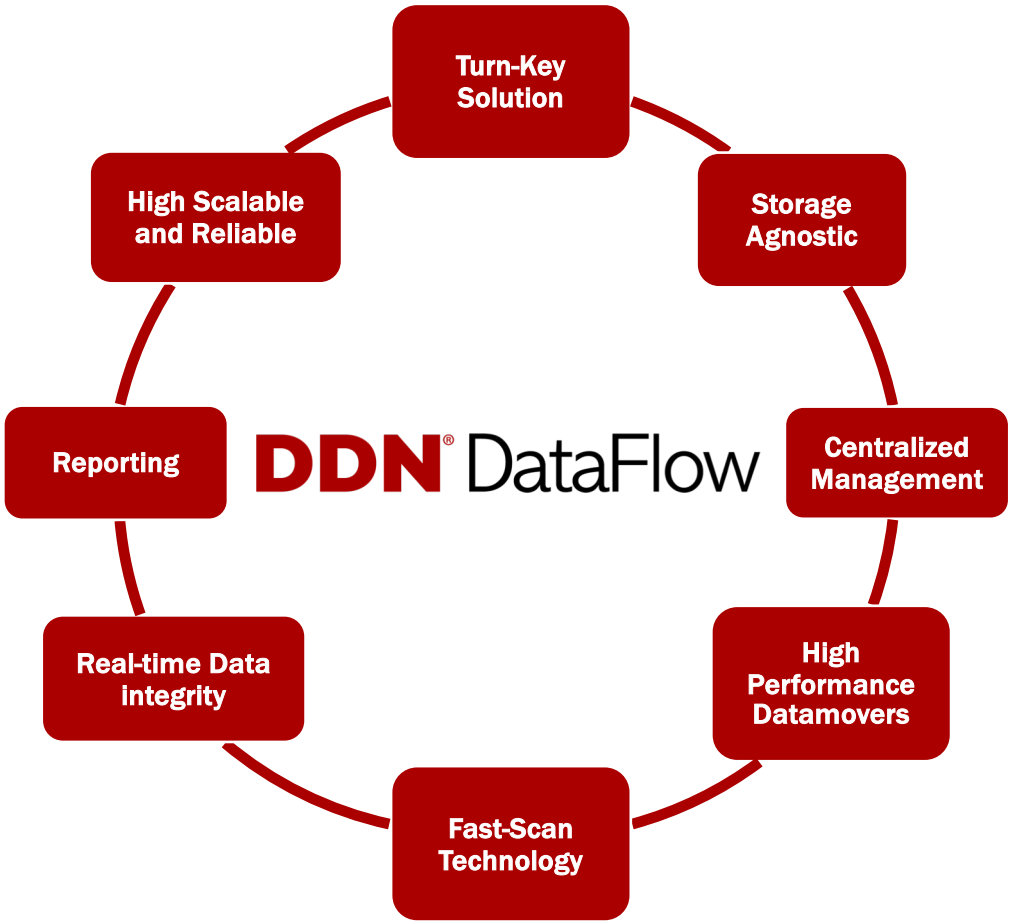
Professional Services
- Commonly based on Robocopy/Rsync tools

DDN Dataflow
- Commercial Solution
- Mature and proved solution

DDN DataFlow

- Cross technology solution, allows migration from any source/any vendor
- Simplified migration process with automatization of migration tasks through central management interface
- Mature and proved solution with professional support

DDN DATAFLOW MIGRATION PRODUCT FEATURES

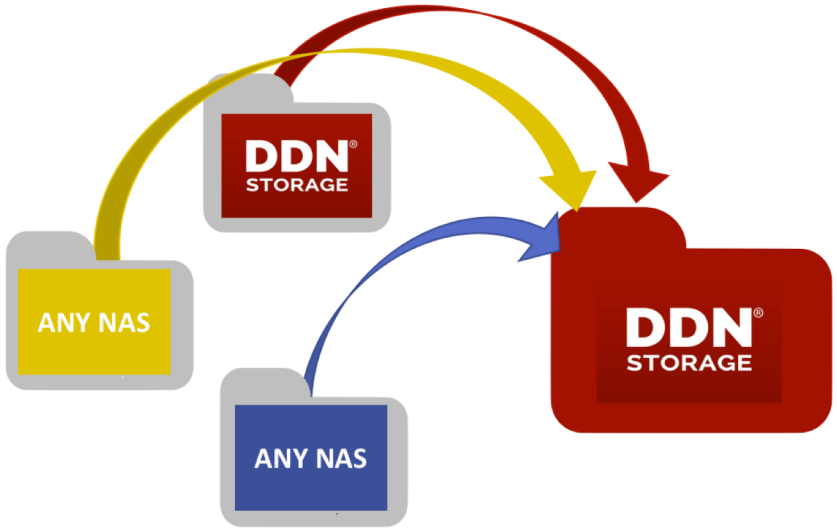


DDN DATAFLOW MIGRATION FROM ANY SOURCE

Storage agnostic solution allows from any source migration

Migrates data from any source including most common commercial NAS platforms: Isilon, NetApp, Panasas, parallel filesystems GPFS, Stornext and Lustre and Object storage, like DDN WOS.

It enables the consolidation of the data from multiple sources to a new DDN Scaler Solution.



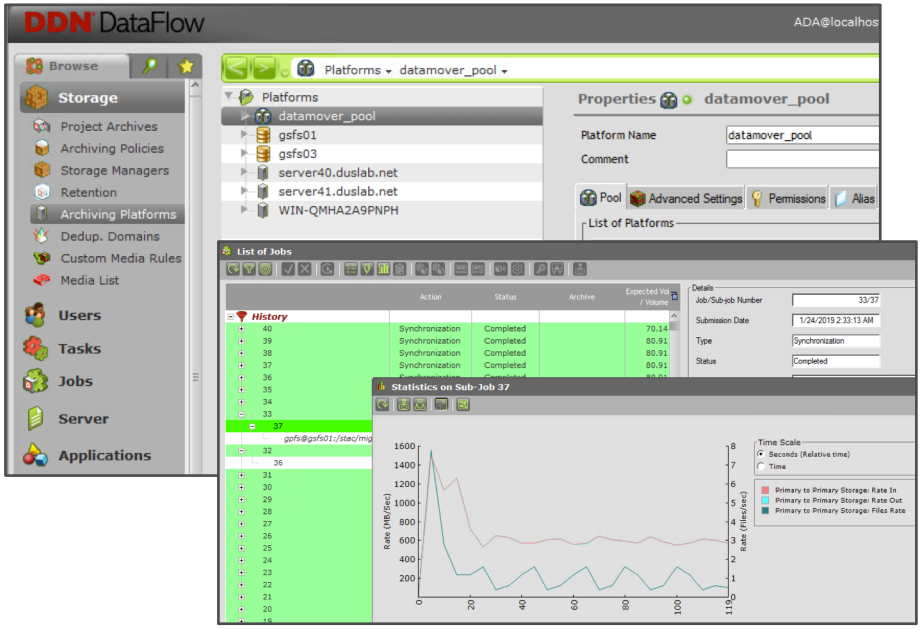
DDN DATAFLOW MIGRATION CENTRALIZED MANAGEMENT

Intuitive user interfaces for effortless productivity

The administrator console provides single pane of access for complete system configuration, workflow definition and process monitoring.

Historic and real time information of the migration tasks is available enabling customer to easily follow the migration process at all time.

Comprehensive CLI, web services and a C++ API are also available for automation and integration.



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