31st International Conference on Massive Storage Systems and Technology (MSST 2015)

DDN STORAGE

Advanced IO Architectures

Leveraging Flash in Integrated, Scalable Systems

June, 2015

MIKE VILDIBILL

Vice President

Product Management & Emerging Technology Development

About Us

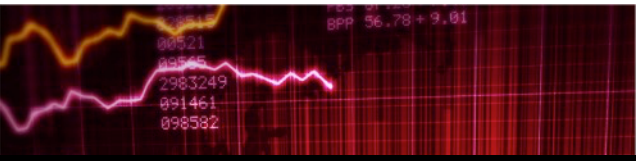
We Solve Big Data Lifecycle Management Challenges at Large Scale

Customers: 1,100+ in 50 Countries	**Employees:** 550+ in 20 Countries
Go-To-Market: Direct, Partner Assist, Specialized VARs	**History:** Founded in 1998, Growing & Profitable
Headquarters: Santa Clara, CA	**Manufacturing:** Chatsworth, CA

World-Renowned & Award Winning

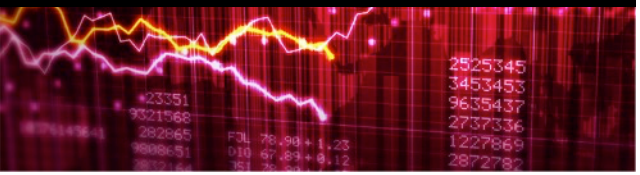
IDC Analyze the Future · Inc. · Gartner · the 451 group · HPCwire · STORAGE · Federal Computer Week

DDN HPC Technology Throughout the Enterprise



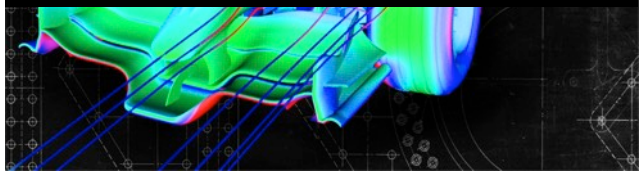
Financial Services

Powering 40% of Leading Global Investment Banks



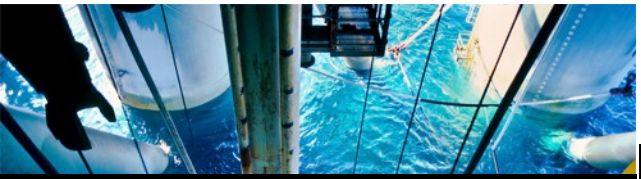
Manufacturing

Powering 30% of the World's Top Aero & Auto Manufacturers



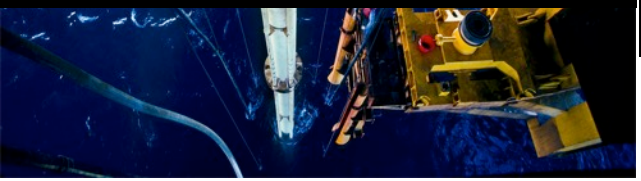
Media

Powering over 600 HD Workflow Customers



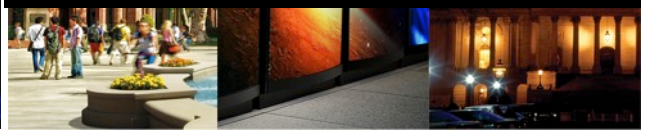
Oil & Gas

Powering Over ½ of the Largest Oil & Gas Sites



Supercomputing

Powering Over 2/3 of the Top100®



Life Sciences

Powering Over 1/3 of the Top Sequencing Centers

DDN Technology

Best of Class standalone and integrated solutions

Application Acceleration, Burst Buffer — **Infinite Memory Engine*™**

High- Performance Converged Media Storage

MEDIAScaler™

Parallel Workload Client

Integrated Cloud: S3, & Swift, Active Archive and Tape Tiering

Petascale Lustre® Storage

EXAScaler™

Up to 100,000 Clients

Open Source

Leading metadata & small file performance

Enterprise Scale-Out GPFS File Storage

GRIDScaler™
GS7K™

Up to ~16,000 Clients

Integrated Object Storage, Tape

NFS & CIFS

Snapshots, Replication

Unified Management

DirectMon™

Core Storage Platforms

SFA™12KX

48GB/s, 1.7M IOPS

1,680 Drives in 2 Racks

Embedded Computing

SFA™7700X

12.5GB/s, 450K IOPS

60 Drives in 4U

396 Drives in 20U

Flexible Drive Configuration

All Flash, Hybrid or HDD-only

SAS **SSD** **SATA**

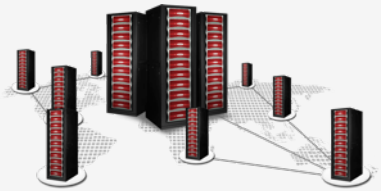
Automated Flash Caching

SFX™ Adaptive Transparent Flash Cache
SFX API Gives Users Control
[pre-staging, alignment, by-pass]

Content Distribution, Cloud Storage, Active Archive

WOS®
Web Object Scaler

- 32 Trillion Unique Objects
- Data Protection & Distribution via Replication or Erasure Coding
- Lowest latency data access and rebuild

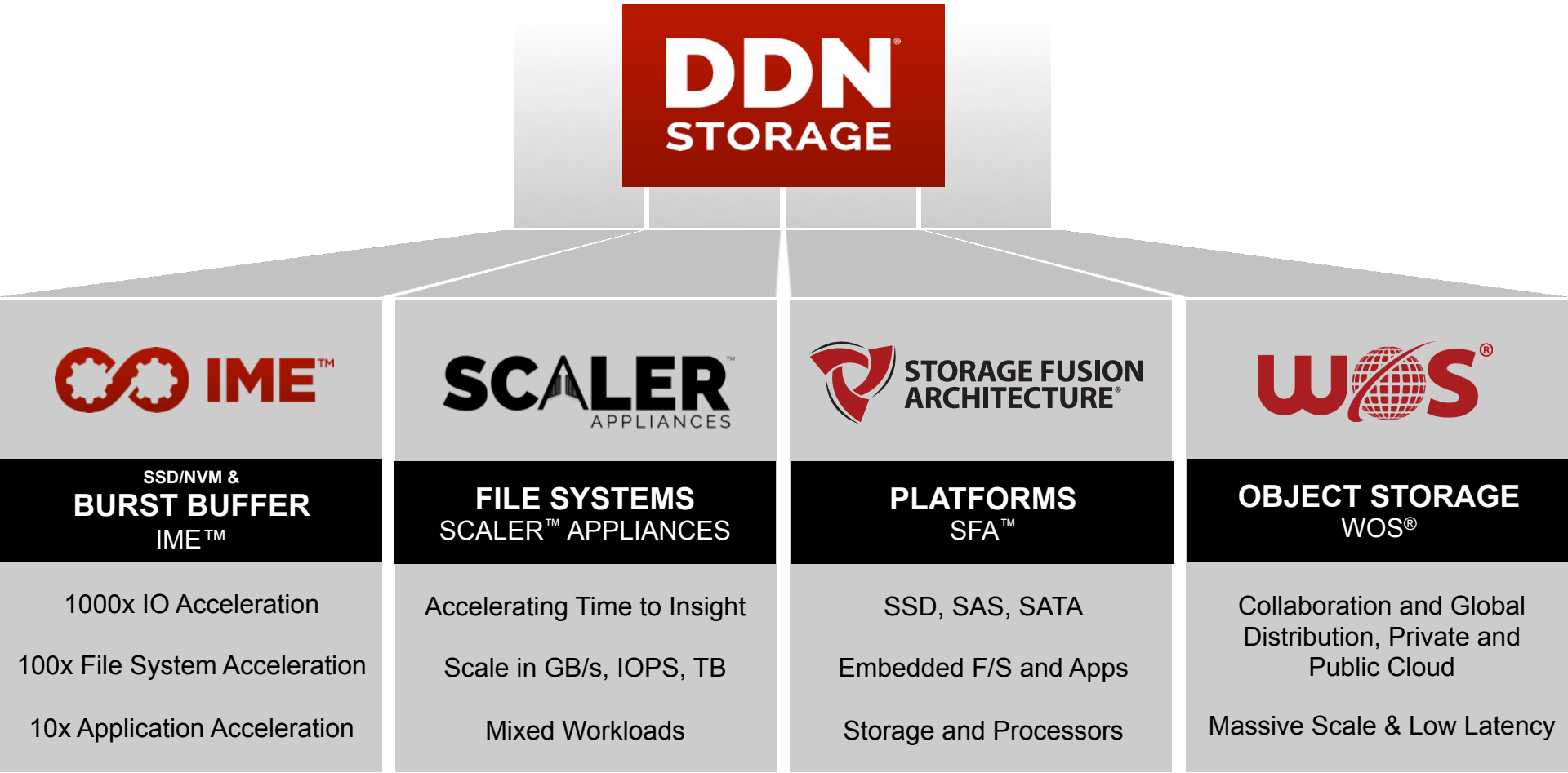


Options

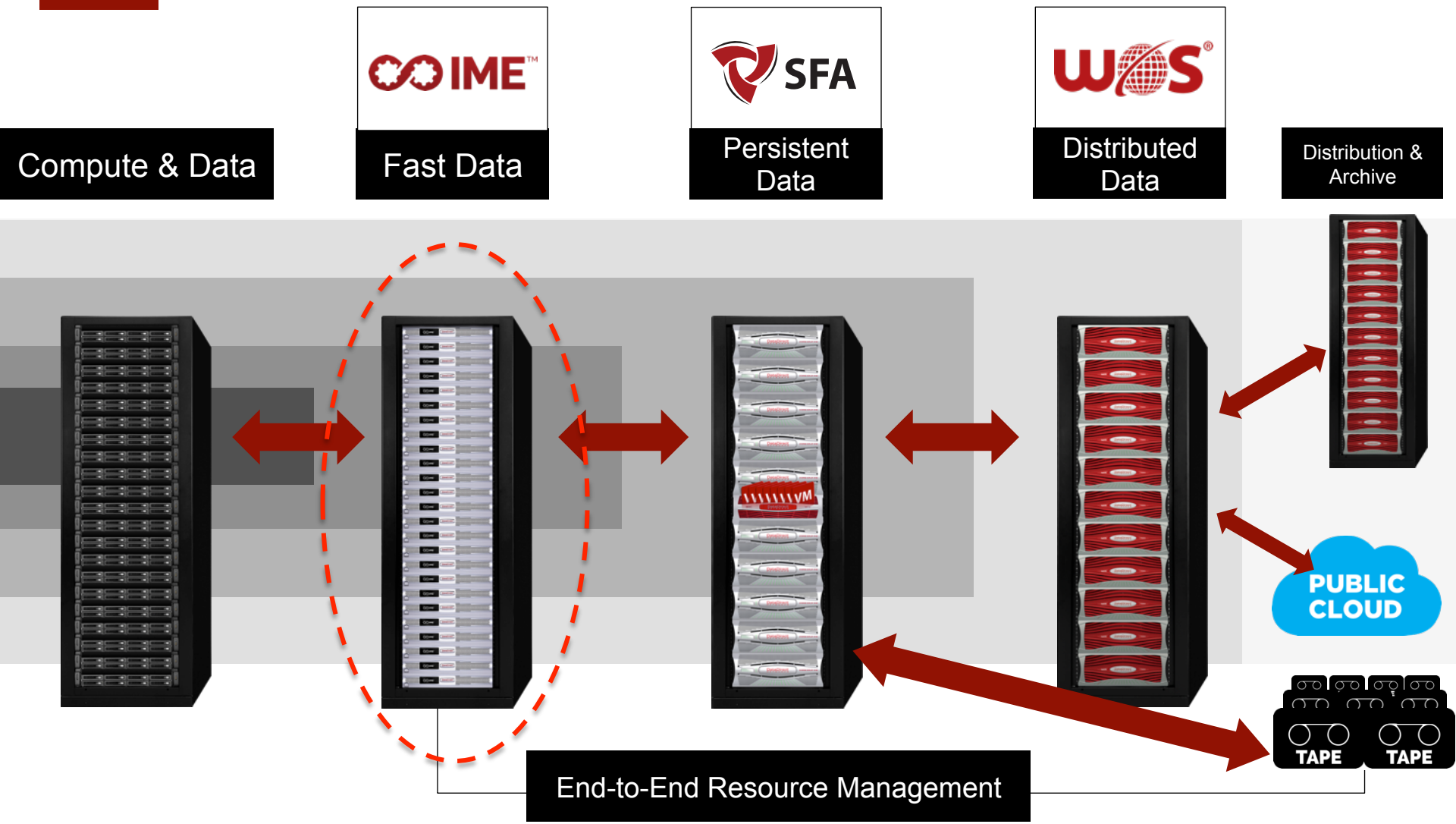
- Single 4U appliance up to 64 sites
- OCP-compliant Hardware
- Software-only

DDN High-performance Portfolio

Four Strong Pillars



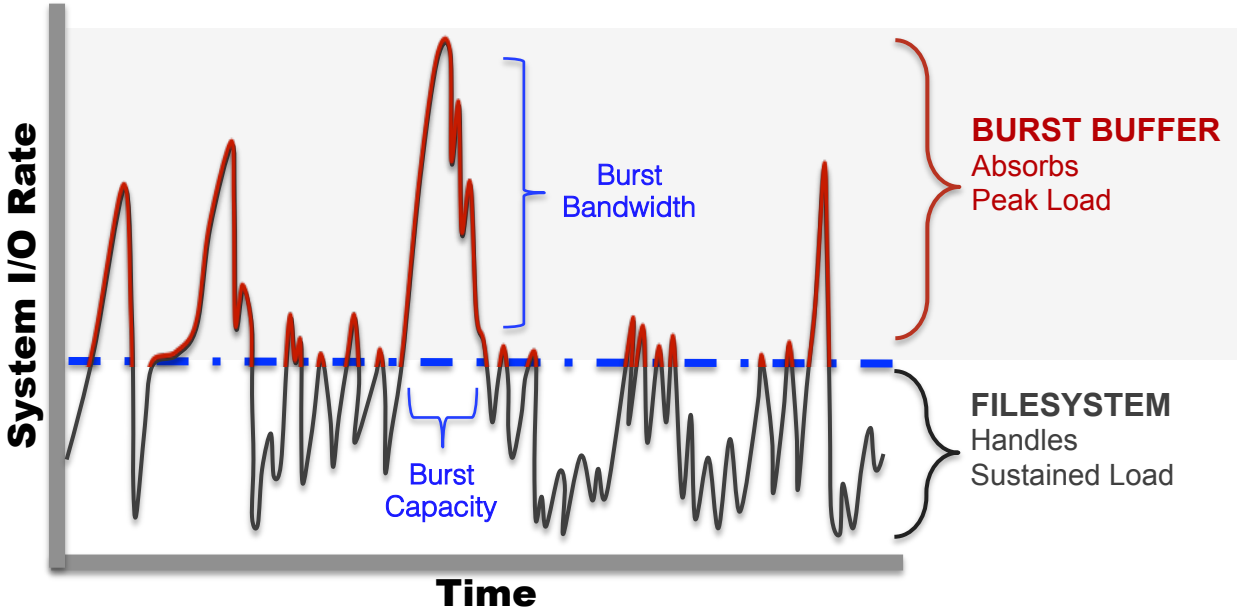
DDN Integrated Supporting End-to-End Workloads



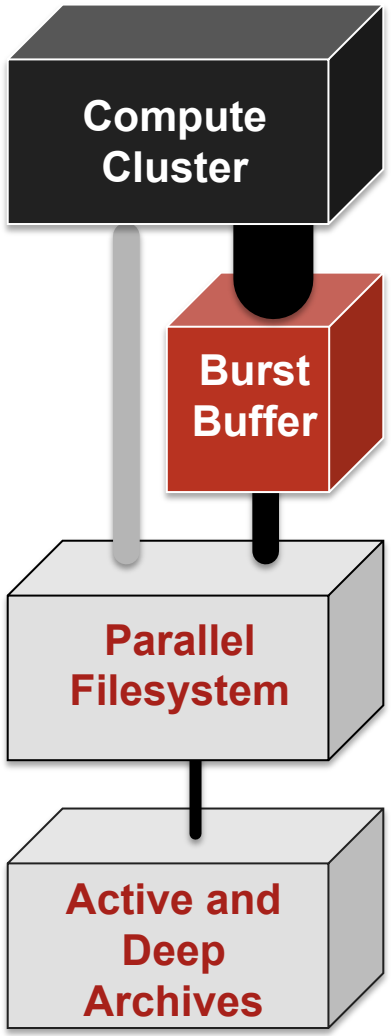
A New Paradigm in I/O Provisioning

Analysis of a major HPC production storage system

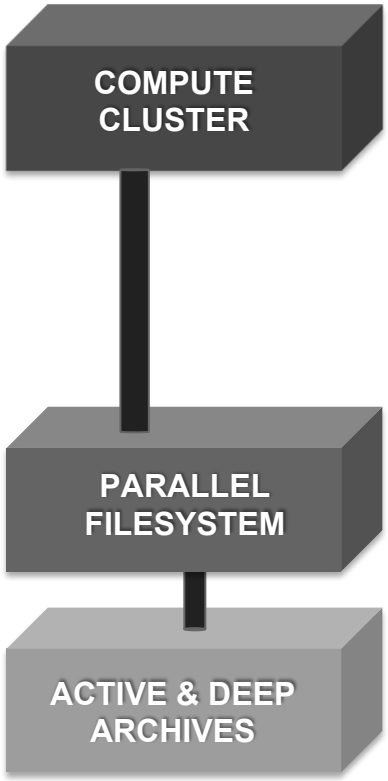
- 99% of the time, storage BW utilization < 33% of max
- 70% of the time, storage BW utilization < 5% of max



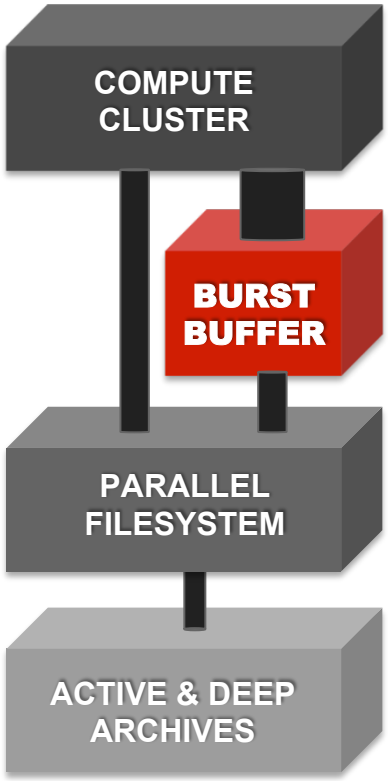
DDN's Infinite Memory Engine™ (IME) is a unified, distributed, non-volatile storage pool that is transparently accessible to parallel applications and is tightly integrated with proven high performance parallel filesystems. Configured as a Burst Buffer it changes how we provision I/O performance



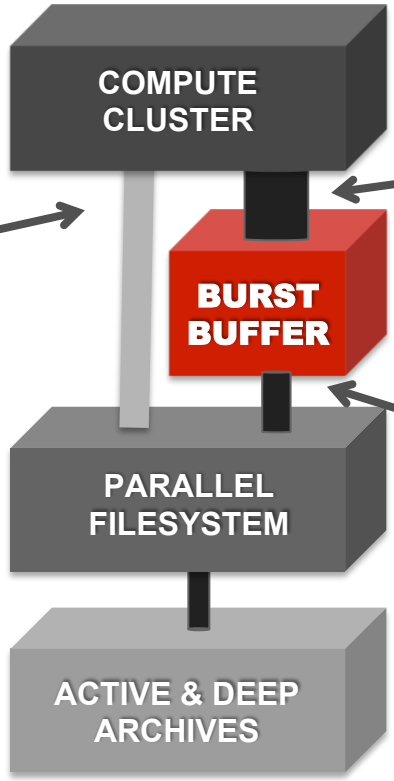
Infinite Memory Engine™



Infinite Memory Engine™



Infinite Memory Engine™



IME and PFS clients deployed here

APIs: POSIX, MPI-IO, HDF5, etc.

Network capacity for burst

Compute nodes can bypass IME

IME server software here

Network capacity for replay (IME-to-PFS) or pre-stage (PFS-to-IME)

Persistent storage and shared name space; Lustre, GPFS, etc.

IME Accelerates I/O in Several Ways
"Problem Application" Case Study: S3D (write)

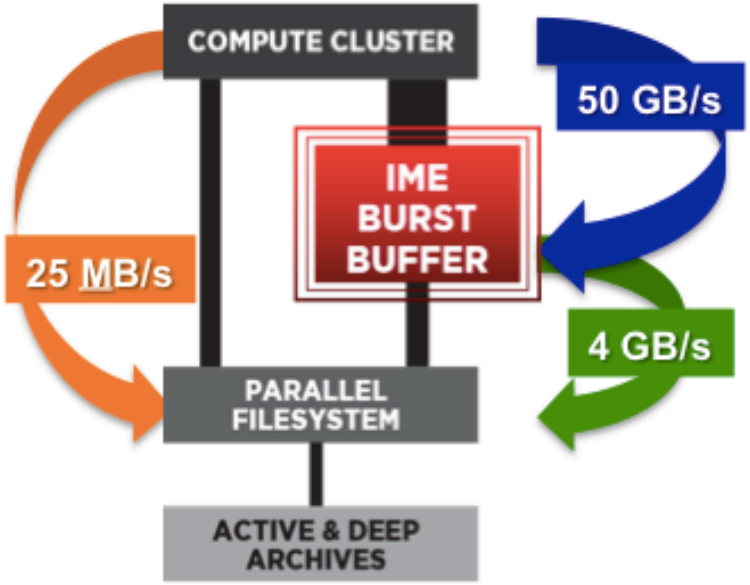
1) MITIGATES POOR PFS PERFORMANCE caused by PFS locking, small I/O, and mal-aligned, fragmented I/O patterns.

IME "**makes bad apps run well**" and also prevents a poor-behaving app from impacting the entire supercomputer.

This is especially valuable to **diverse workload** environments and **ISV** applications.

At SC14, we **demonstrated 1000x speed-up** on mal-formed I/O when using non-POSIX low-level communications.

1000x



Application: S3D
Turbulent Flow Model

10x

2) PROVIDES HIGHER PERFORMANCE I/O (bandwidth and latency) to the application.

Providing additional bandwidth here is relatively inexpensive. Configuring **10x more bandwidth** compared to PFS is typical.

3) IME DRIVES I/O MORE EFFICIENTLY TO THE PFS by re-aligning and coalescing data within the non-volatile storage.

At SC14, we demonstrated **100x speed-up** due to this efficiency. IOR benchmarks show a **3x – 20x speedup on I/Os <32KB**.

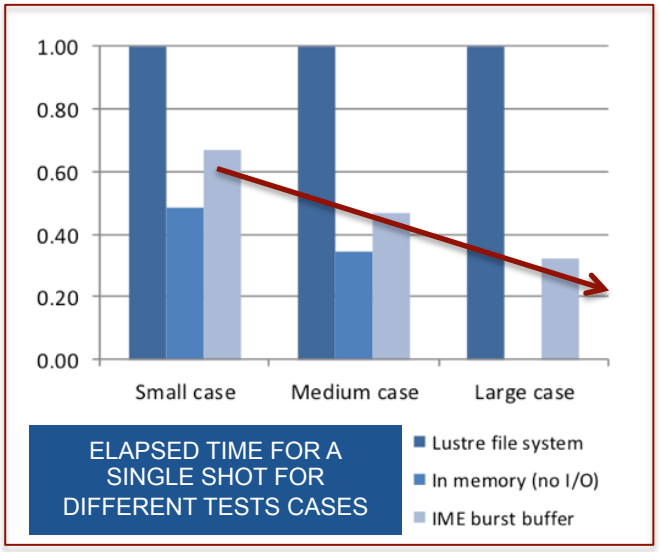
100x

DDN IME Testbed Program Q3'14 – Present

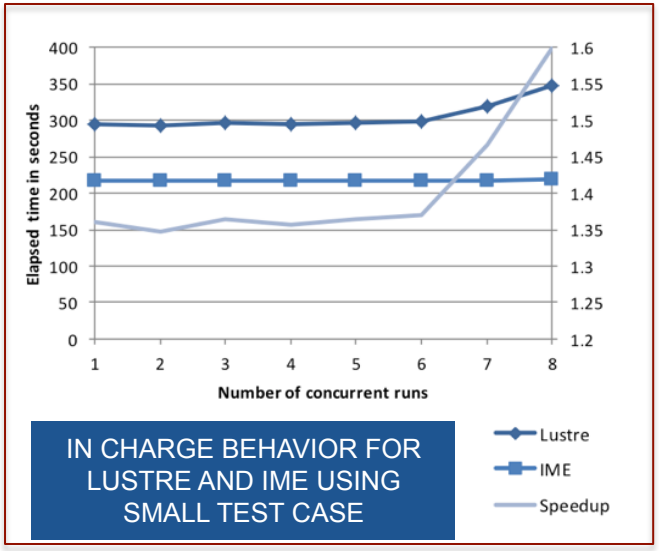
10+ IME Testbed deployments, demonstrating real-world app acceleration

Example: Tullow Oil's RTM code:

- **A modest 100MB seismic image, if reconstructed with 5000 snapshotting time steps will lead to writing and then re-reading 500GB of data, for a global data movement of 1 Terabyte**
- **A typical seismic survey will induce the generation of thousands of such images, usually computed in parallel with several independent jobs launched on the cluster, resulting in the movement of many petabytes of data during the execution of the parallel application**
- ***IME transformed this I/O-bound production app into a Compute-bound one***
- **3x full-app speed-up over Lustre alone**

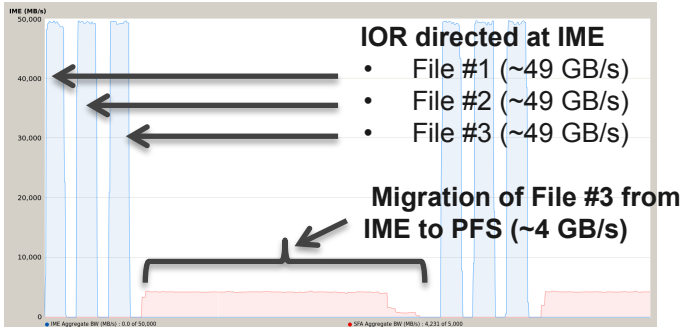


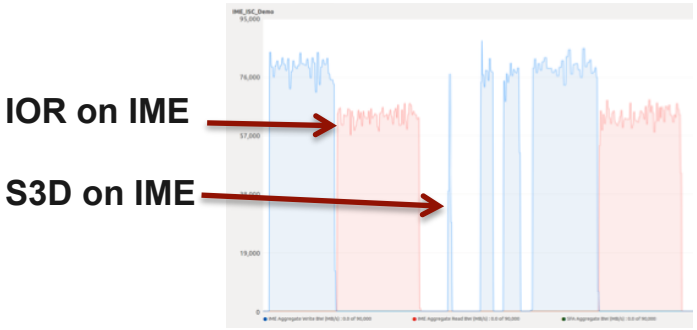
ELAPSED TIME FOR A SINGLE SHOT FOR DIFFERENT TESTS CASES



IN CHARGE BEHAVIOR FOR LUSTRE AND IME USING SMALL TEST CASE

DDN IME Public Demonstrations Q4'13 – Present





DDN SC'13 IME Prototype Demo

- First public demonstration
- IOR to shared-file, 4KB strided IO, ~100 MPI tasks
- ~49 GB/s write to IME
- ~4 GB/s "replay" to GPFS on SFA7700

DDN ISC'14 IME Testbed Platform Demo

- IME testbed hardware w/ ~140 MPI tasks
- IME MPI-IO / ROMIO interface
- IOR and S3D IO kernel
- Index fault tolerance (replication)
- ~80 GB/s write and ~67 GB/s read

DDN SC'14 IME Testbed Platform Demo

- Early preview of FUSE (POSIX) interface
- Ran IOR, S3D IO kernel, HACC IO kernel, and madBench2 through IME MPI-IO / ROMIO
- 300 – 600 MPI tasks
- ~80 GB/s write and ~80 GB/s read

How Does IME Help?

Supports a Much Broader Range of Applications & Data Patterns

IME's "Beyond Burst Buffer" capabilities extend beyond checkpoint/restart to support both large & small I/O, enabling the acceleration of your performance-hungry applications

Featured Use Cases

- **Burst Buffer I/O Acceleration**
  - Checkpoint-Restart
  - Write-back and Write-through Cache for File Alignment (direct effect) and Block Alignment (indirect effect)
  - Stage-in, Stage-out, Demand Loading
  - Isolation of ill-behaving applications
- **Out-of-Core I/O**
- **Data Analysis Support**
  - Post-processing
  - Visualization
- **Temporary Data Storage**
  - Sequential-job Data Sharing (many-task computing, …)
  - Concurrent-job Data Sharing (**coordinated sharing of data through several tasks**)
  - Intermediate Results

Thank You!

Keep in touch with us

sales@ddn.com

2929 Patrick Henry Drive
Santa Clara, CA 95054

@ddn_limitless

1.800.837.2298
1.818.700.4000

company/datadirect-networks

DDN STORAGE

ddn.com