DDN® STORAGE

Flash Memory Summit 18

Integrating Scale-out Flash into Production Workflows.

Optimizing flash to speed up random read, random write, shared file, high concurrency and streaming workloads

Kurt Kuckein
Sr. Director, Marketing

August 2018

THE RISE OF DATA:

Diversified workloads, more complexity, deeper workflows.

Machine Learning

Big Data

Multi-Physics

Supercomputin g

NoSQL Analytics

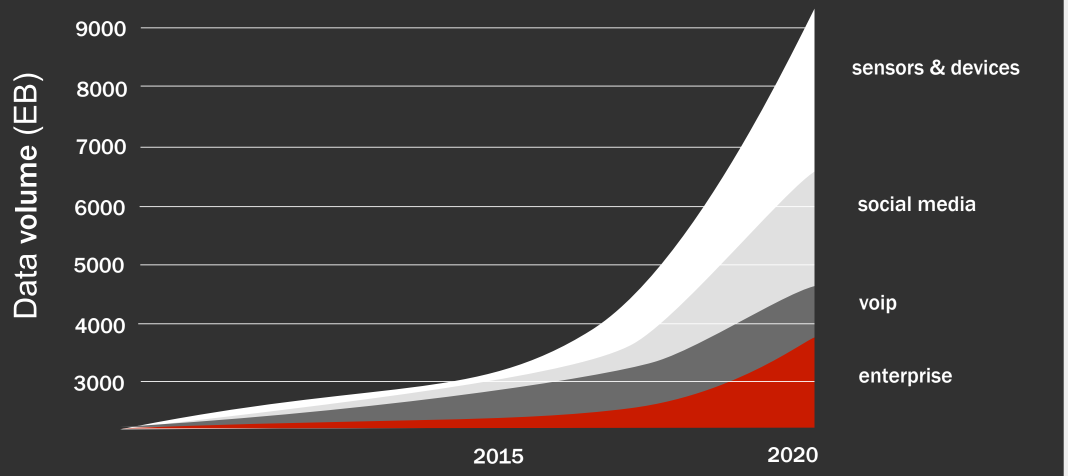
Workflows

Adaptive Mesh Refinement

Checkpointing

DDN STORAGE

Deep sophistication in data platforms reduces complexity for the business



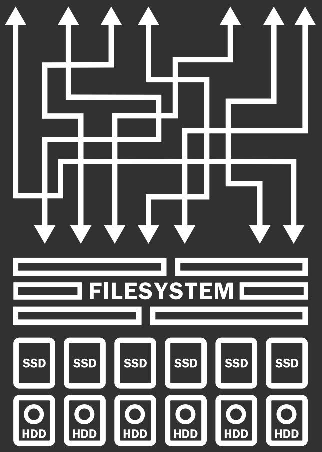
Active data Volumes & increasing application sophistication demanding new levels of **performance, scale** and **economy** from data platforms.

Despite the new emergence of all-flash, modern enterprise storage approaches are failing to address the challenges at scale.

Modern workloads, introduce tougher IO = Pain for filesystems, even parallel filesystems.

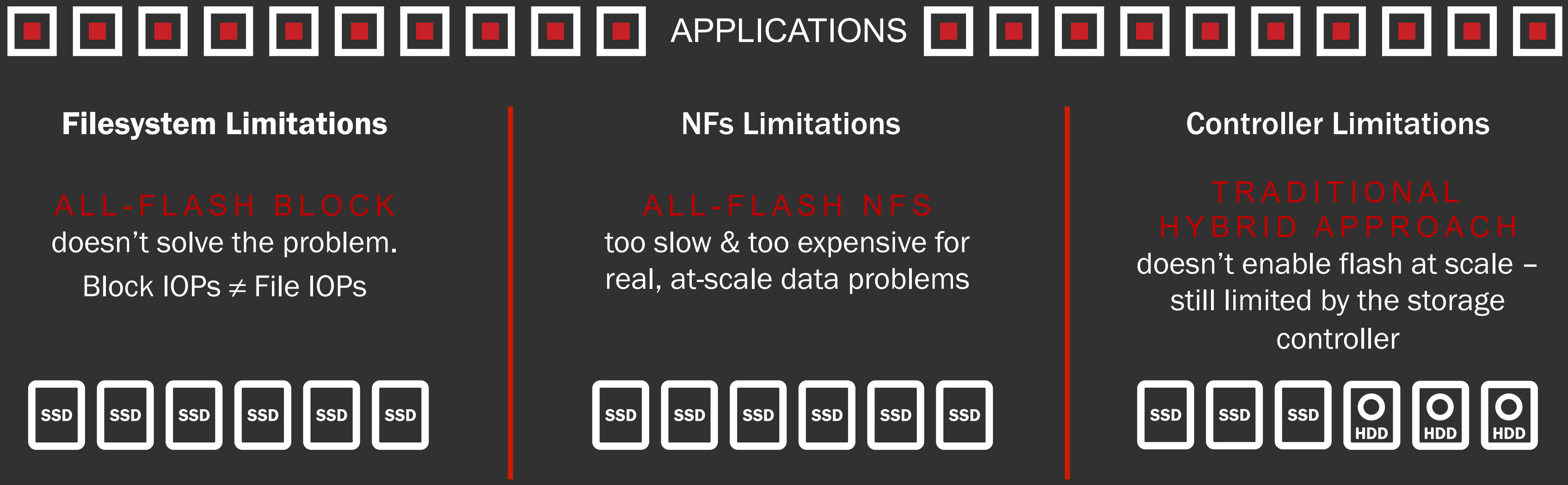
Machine Learning | Big Data | Multi-Physics | Supercomputing | NoSQL Analytics | Workflows | Adaptive Mesh Refinement | Checkpointing

Modern Workload IO patterns are increasingly mixed and tough: reads and writes, random and sequential, high thread counts, shared file access



Traditional Thick File system SW layers and fixed data layout severely restrict performance for tough workloads – even with SSDs

Expansion in active data volumes requires a new economics for fast data at scale.



APPLICATIONS

Filesystem Limitations

ALL-FLASH BLOCK
doesn't solve the problem.

Block IOPs ≠ File IOPs

NFs Limitations

ALL-FLASH NFS
too slow & too expensive for real, at-scale data problems

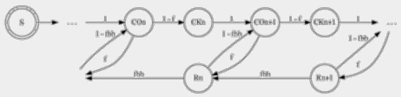
Controller Limitations

TRADITIONAL HYBRID APPROACH
doesn't enable flash at scale – still limited by the storage controller

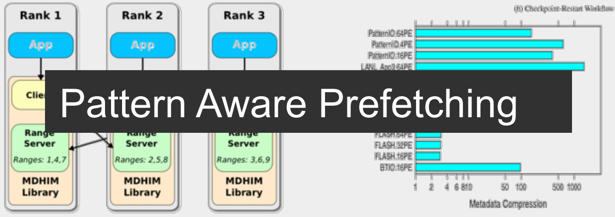
Flash Potential is Hard to Extract.
It Requires Productized Innovations

I/O Acceleration with Pattern Detection

Jun He, John Bent, Aaron Torres, Gary Grider, Garth Gibson, Carlos Maltzahn, Xian-He Sun



(b) Checkpoint-Restart Workflow with an Unreliable Burst Buffer



Pattern Aware Prefetching

MDHIM: A Parallel Key/Value Framework for HPC

Hugh N. Greenberg John Bent Gary Grider

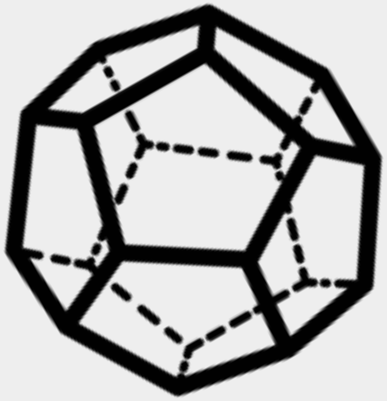
Write-Anywhere Layout

(b) Checkpointing with BAD-Check.

BAD-Check[§]: Bulk Asynchronous Distributed Checkpointing

John Bent[*‡] Brad Settlemyer[†‡] Haiyun Bao[*]
Sorin Faibish[*] Jeremy Sauer[†] Jingwang Zhang[*]

DDN STORAGE



On the Non-Suitability of Non-Volatility

John Bent[*†] Brad Settlemyer[†‡] Nathan DeBardeleben[†‡]
Sorin Faibish[*] Uday Gupta[*] Dennis Ting[*] Percy Tzelnic[*]

Software Defined Erasure

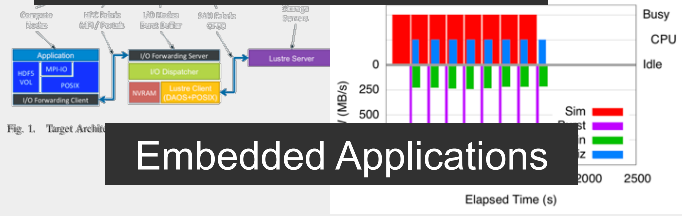


Fig. 1. Target Archit

Embedded Applications

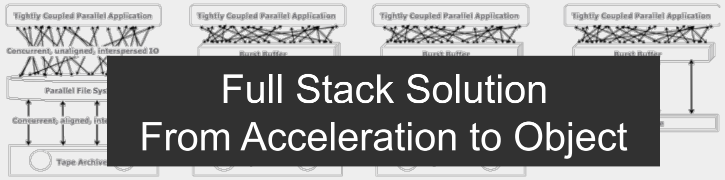
Jitter-Free Co-Processing on a Prototype Exascale Storage Stack

John Bent Sorin Faibish Jim Ahrens Gary Grider

Serving Data to the Lunatic Fringe

The Evolution of HPC Storage

JOHN BENT, BRAD SETTLEMYER, AND GARY GRIDER


Tightly Coupled Parallel Application
Concurrent, unaligned, interspersed IO
Parallel File Sys
Concurrent, aligned, inte
Tape Archive
Burst Buffer

Full Stack Solution
From Acceleration to Object

(a) 2002–2015 (b) 2015–2016 (c) 2016–2020 (d) 2020–

Figure 2: From 2 to 4 and back again. Static for over a decade, the HPC storage stack has now entered a period of rapid change.

DDN STORAGE | IME

INTRODUCING

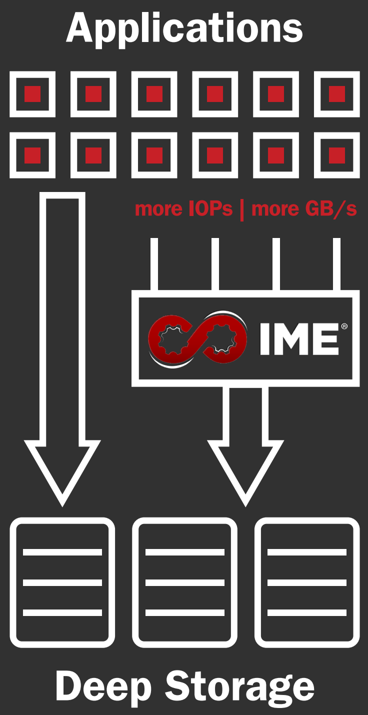
Scale-Out Flash

- Truly Software-Defined
- Commodity Hardware
- Highly Available
- Forever Scalable
- Low Power, High Density
- Removes Filesystem Bottlenecks
- 100% Flash Native

IME

Scale-Out NVMe Flash.

- IME forms a transparent, scalable cache which **delivers unprecedented performance** to applications
- **Zero Application modifications are needed** for IME to unleash the power of your next generation workloads
- IME **dramatically accelerates** random read, random write, shared file, high concurrency and streaming workloads

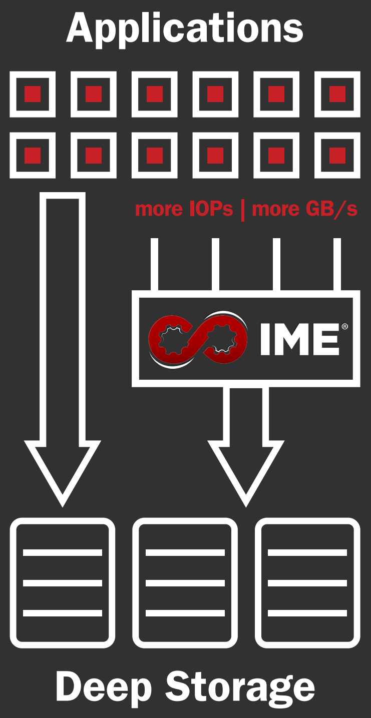


DDN Confidential | ©2017 DDN Storage

IME

Scale-Out NVMe Flash.

- Protects data against device and node failures and intelligently and transparently manages data movement
- Wirespeed-fast on RDMA and TCP networks for Reads and Writes
- Filesystem IOPS scales infinitely with zero penalty for file sharing



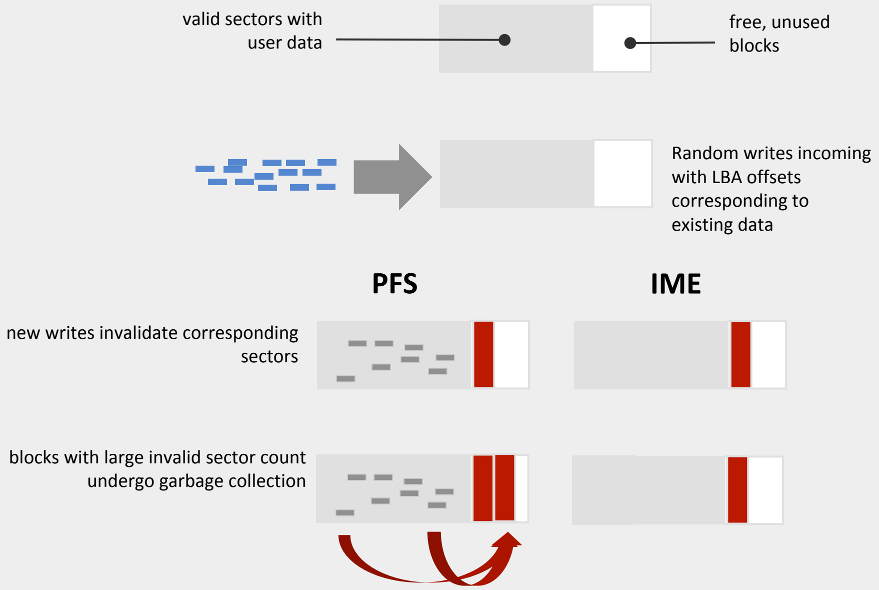
- extendable user/application workload
- Best $/IOPs & Throughput
- NVMe Flash Linear Scaling
- Best $/Capacity
- Deep Disk Capacity Scaling

Maximize Flash Endurance

Traditional file systems overwrite the same blocks.

Incur costly garbage collection.

Hurts performance and NVME lifetime



valid sectors with user data

free, unused blocks

Random writes incoming with LBA offsets corresponding to existing data

PFS

IME

new writes invalidate corresponding sectors

blocks with large invalid sector count undergo garbage collection

IME never overwrites physical blocks.

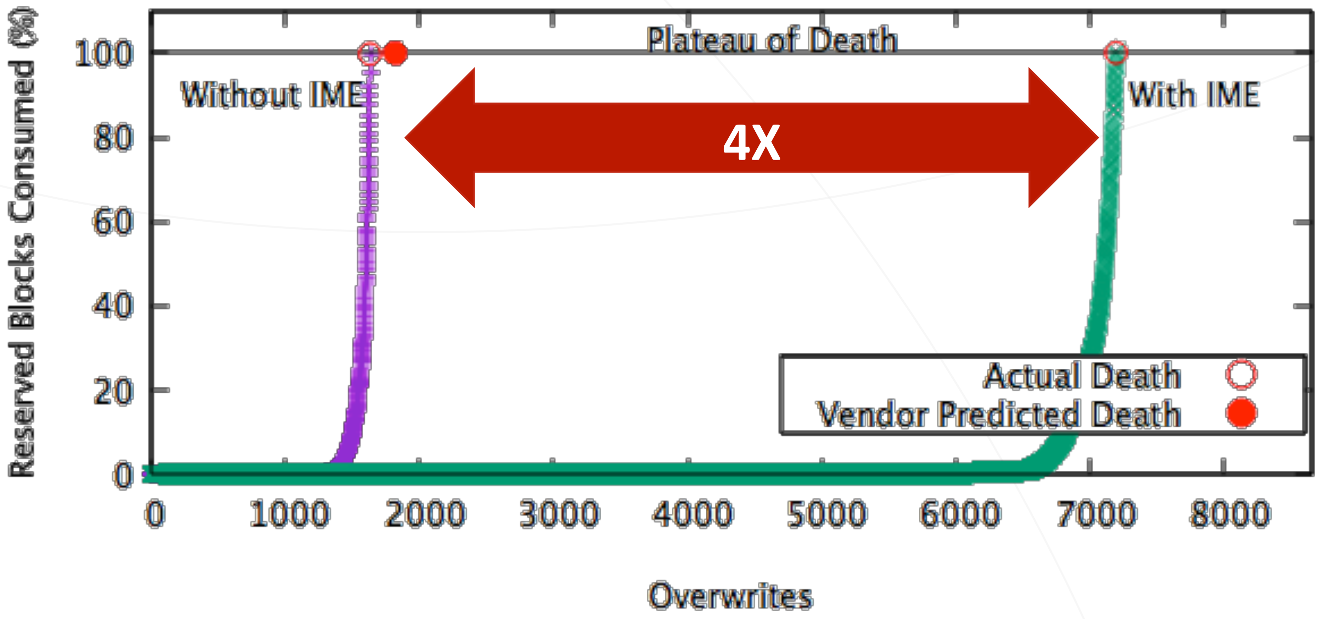
IME manages NVMe to to reduce garbage collection.

IME transforms small random into large sequential.

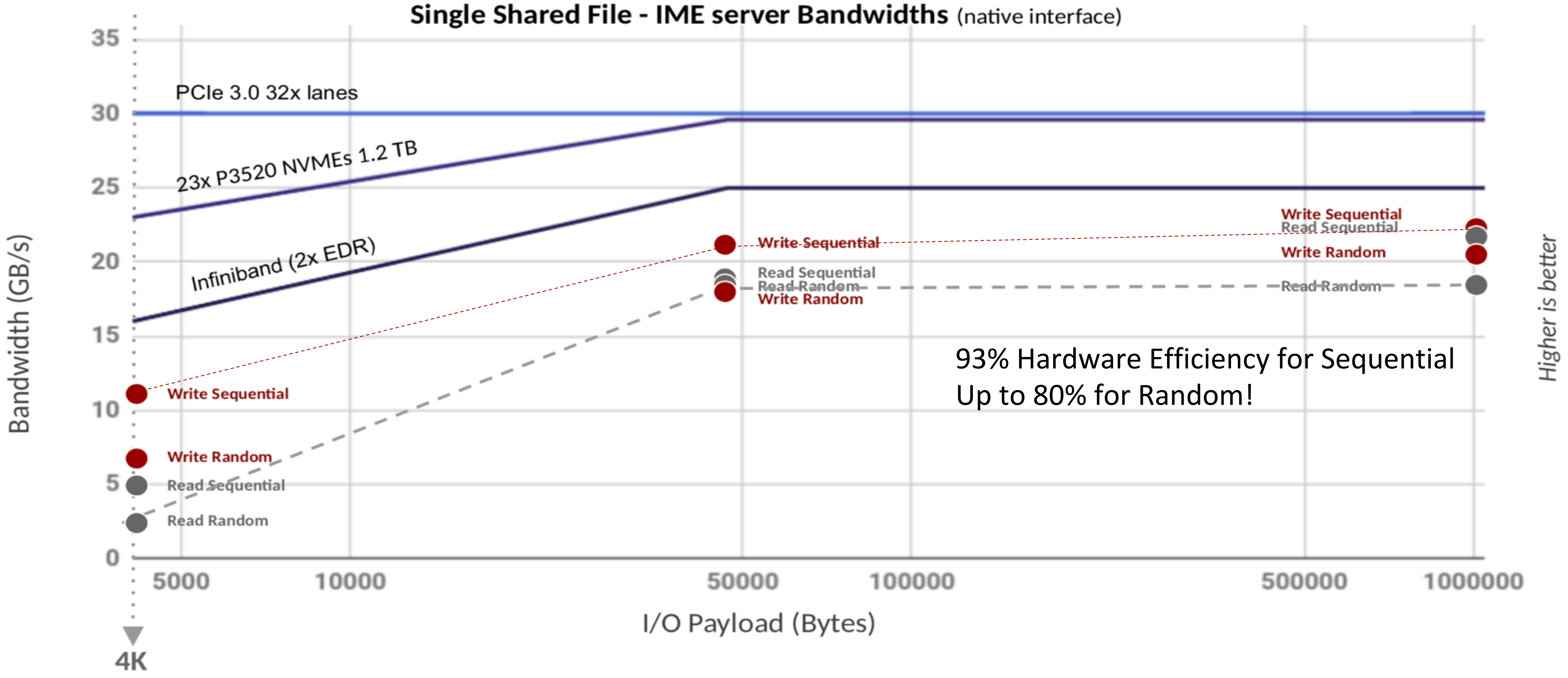
Improves performance and lifetime.

LONG-LASTING FULL FLASH POTENTIAL FOR YOUR WORKFLOWS

MAXIMIZE FLASH ENDURANCE

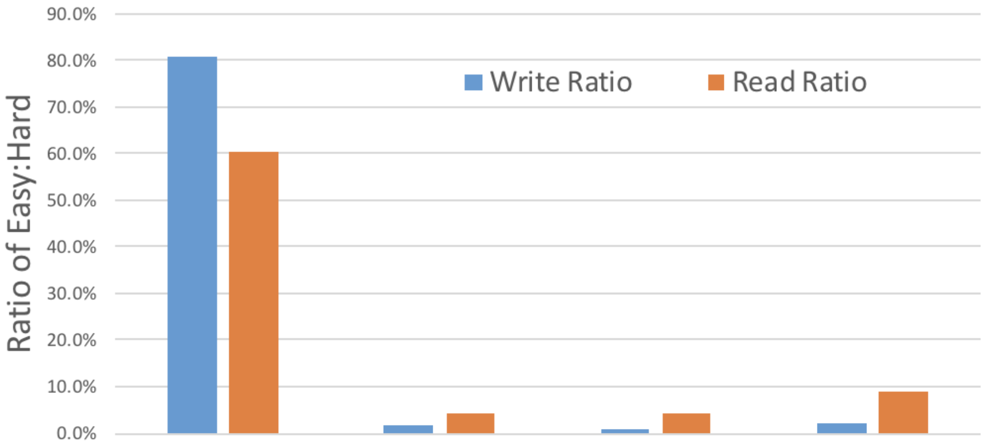


MAXIMIZE FLASH PERFORMANCE



93% Hardware Efficiency for Sequential
Up to 80% for Random!

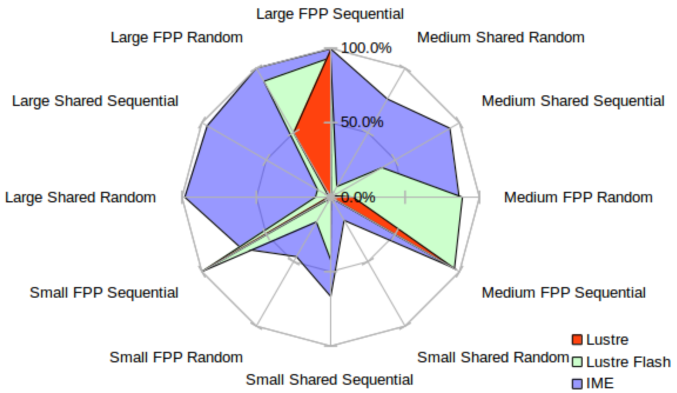
MAXIMIZE FLASH USABILITY



IO500 Results
Ratio of Easy:Hard (systems with 100 clients or more)

IME @ JCAHPC

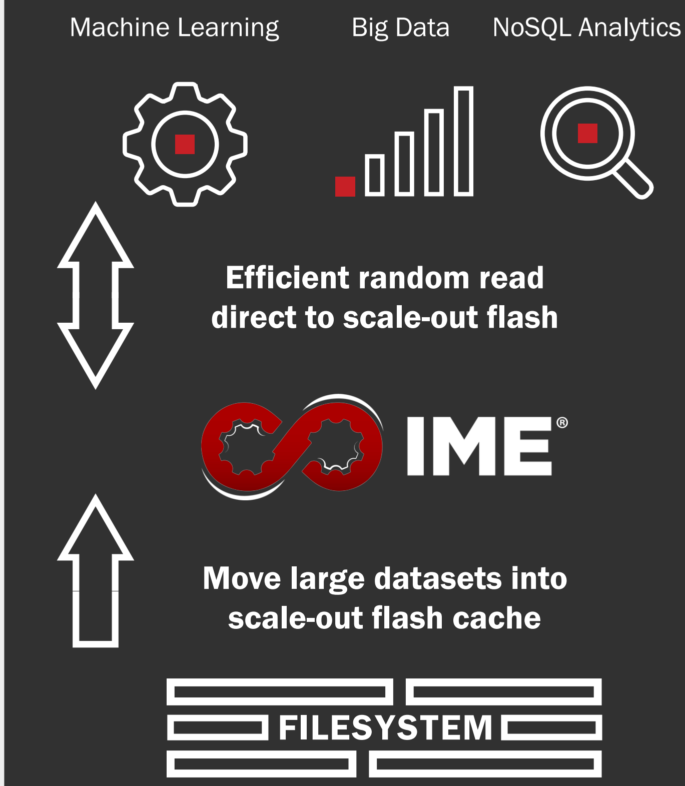
Other Flash Systems



HOW DOES IME HELP?

Analytics, Big Data & Machine Learning

- Analytical workloads are characterized by read-intensive, random IO over very large datasets
- Normal caching techniques are not scalable enough to cope with 10’s or 100’s of TBs of data
- IME Scale-out cache allows you to maintain even PB’s of hot data in flash cache.
- Random Reads served at 600K IOPs in 2U



Machine Learning

Big Data

NoSQL Analytics

Efficient random read direct to scale-out flash

Move large datasets into scale-out flash cache

IME

Machine Learning, at Scale.

Use Case
Large Scale, multiple DGX
(or Apollo6500 HP or Supermicro…)

Customer Pain Points
Coping with hot vs cool datasets managing economics at massive scale. Ensuring consistent scalable performance. Managing data movements

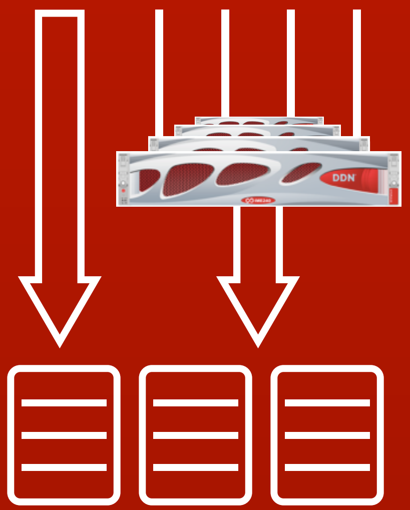
DDN Solution
IME with ES/GS14KX HDD

DDN differentiation
Ideal match of topology to requirement. Strong, Fast Native data management features across tiers. Super scalable performance flash tier and capacity tier

Actual Win
T.B.D. Life Sciences

Actions for Sales
Write up/share references, partner with your local NVidia Sales, know the high level pitch for ES DGX Solution, Talk to Nvidia Rental Partners (SCAN)

CPU/GPU Scale-Out Nodes



Large Capacity GS

Meet shared file demanding random IO workloads with scale out flash

Great APIs, parallel data movements, managed consistency

High Density ES14KX-based HDD provides a scalable cost optimized data lake

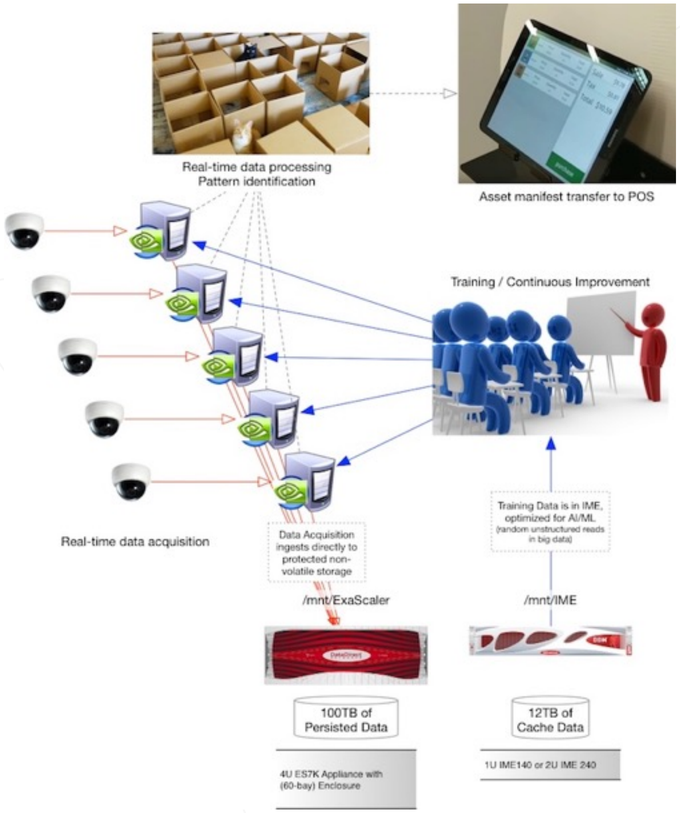
STANDARD COGNITION - REAL TIME CONSUMER DETECTION & BILLING

AUTOMATED CHECKOUT

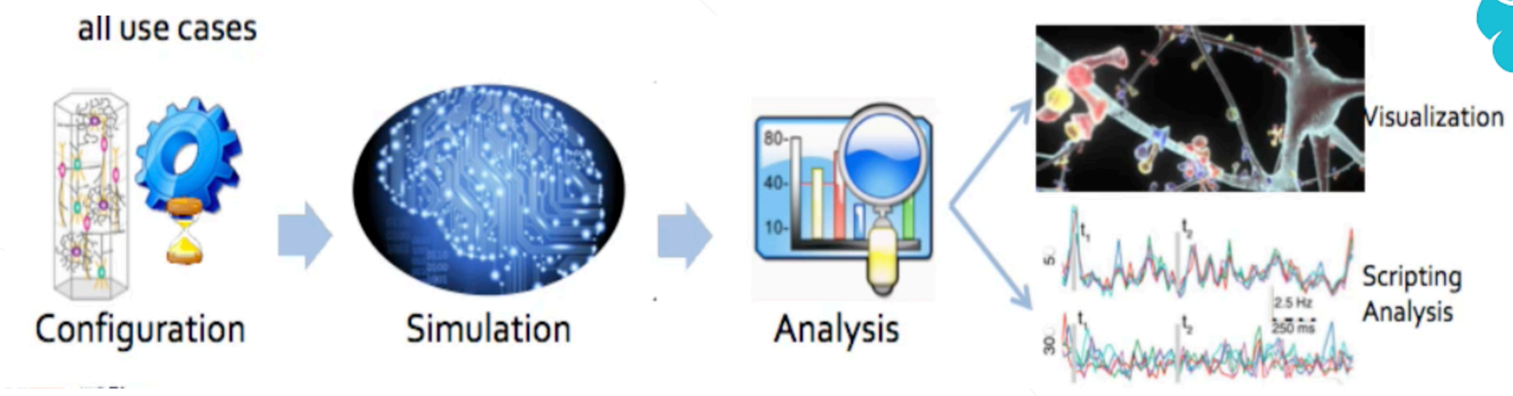
Cameras identify shoppers and bill them real time. Hosting on site.

Customer Needs

- Continuous Data Acquisition
- Ease of deployment/Integration
- Low-latency On-Prem Flash Tier


Real-time data processing
Pattern identification
Asset manifest transfer to POS
Training / Continuous Improvement
Real-time data acquisition
Data Acquisition ingests directly to protected non-volatile storage
Training Data is in IME, optimized for AI/ML (random unstructured reads in big data)
/mnt/ExaScaler
/mnt/IME
100TB of Persisted Data
12TB of Cache Data
4U ES7K Appliance with (60-bay) Enclosure
1U IME140 or 2U IME 240

EPFL HUMAN BRAIN PROJECT



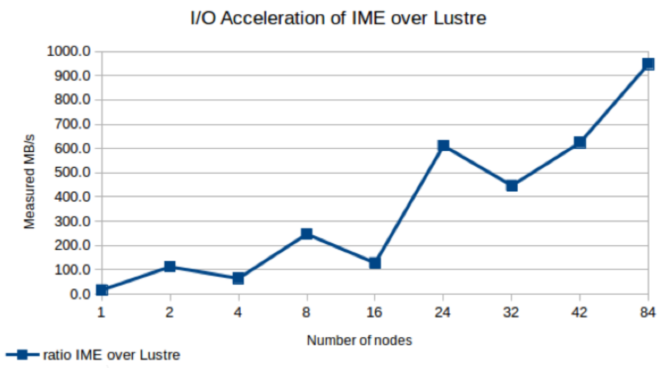
SIMULATION AND RECONSTRUCTION OF THE HUMAN BRAIN

Three Core Focuses

- Neuroscience, Medicine, Computing

Four IME nodes at EPFL

- Simpler, faster, more efficient science
  - Storage redendered transparent
- Scientists focus on science
- NeuroMap application 1000 X Speedup



JCAHPC

DATA AT SCALE FOR JAPAN'S LARGEST SUPERCOMPUTER

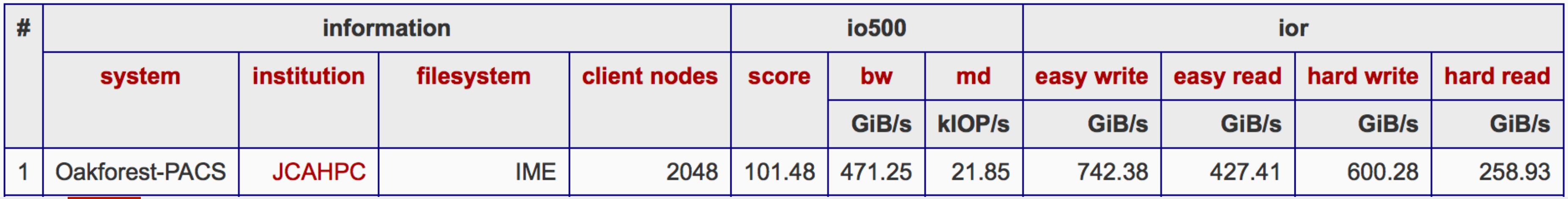
IME on Oakforest-PACS

#5 Supercomputer on Top 500
#1 Storage System on IO 500

Diverse University Applications

- Lattice Quantum Chromodynamics
- Ab-initio Real-Time Electron Dynamics Simulator (ARTED)
- Atmosphere and ocean coupling
- Earthquake simulations using GAMERA/GOJIRA
- First-order optical material science simulations

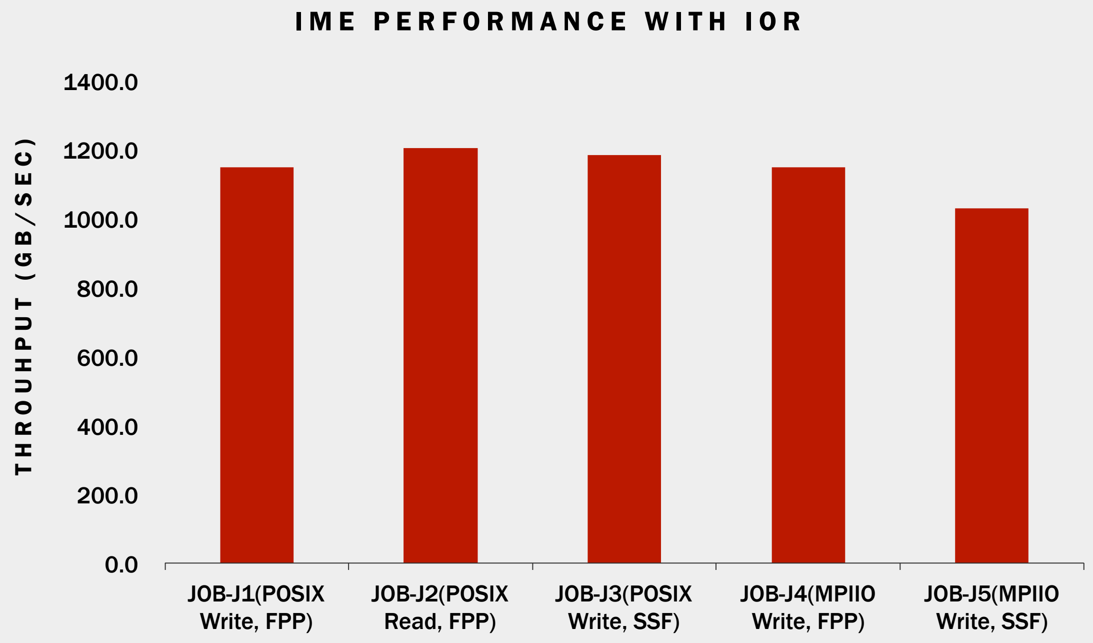
Diverse applications require storage that doesn't need tuning



#	information				io500			ior			
	system	institution	filesystem	client nodes	score	bw	md	easy write	easy read	hard write	hard read
						GiB/s	kIOP/s	GiB/s	GiB/s	GiB/s	GiB/s
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	21.85	742.38	427.41	600.28	258.93

IME at Scale

- Real world implementation of around 2 racks of IME
- ~1PB Flash
- Lustre Backing Filesystem
- Measured 1.2 TB/s
- Both File per Process AND
- Single Shared File



Thank You.

sales@ddn.com

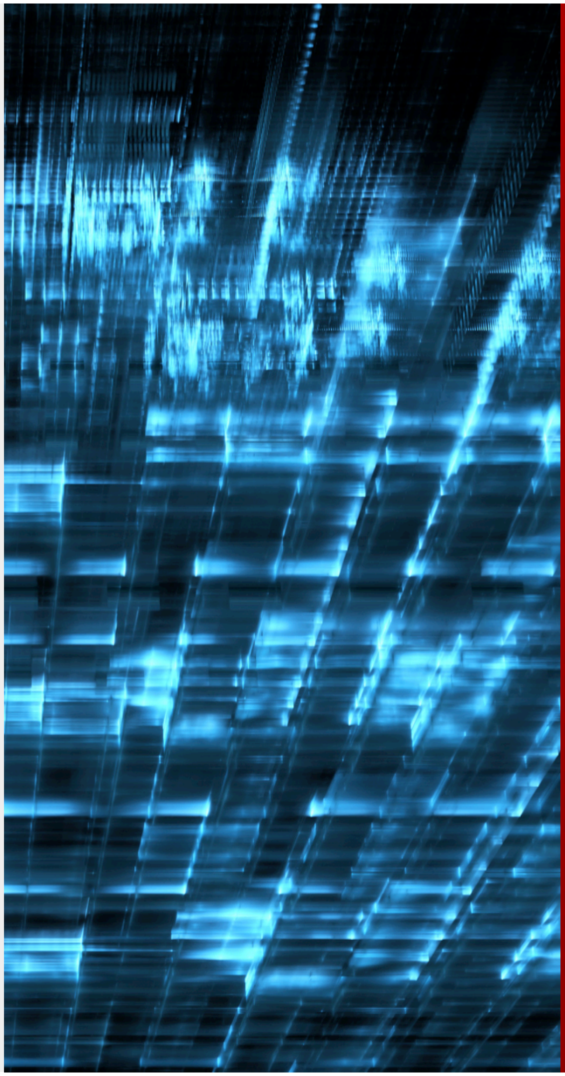
@ddn_limitless

company/datadirect-networks

9351 Deering Avenue
Chatsworth, CA 91311

☎ 1.800.837.2298 1.818.700.4000

DDN STORAGE



IME

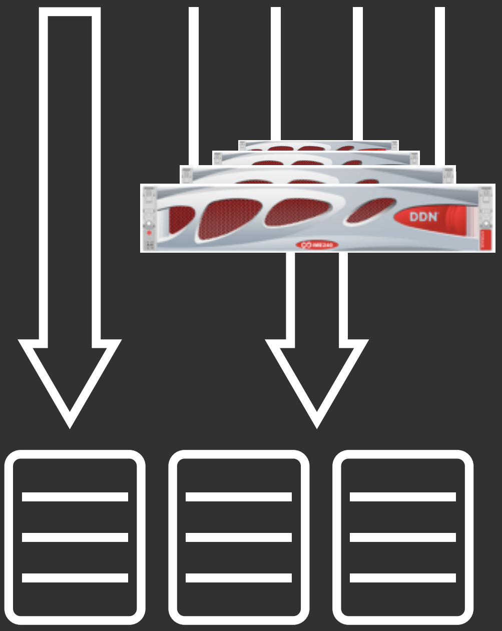
Reservoir Simulation Study

Real-world reservoir simulation study comparing large HDD-based GRIDScaler against a 4-node IME240 system

Varying job count (1,2), Model Size (50M, 100M, 200M), number of writers (50,100,200), etc across 50 compute nodes

4-5x IO time reductions with IME – speedup improves with concurrency

Reservoir Simulation



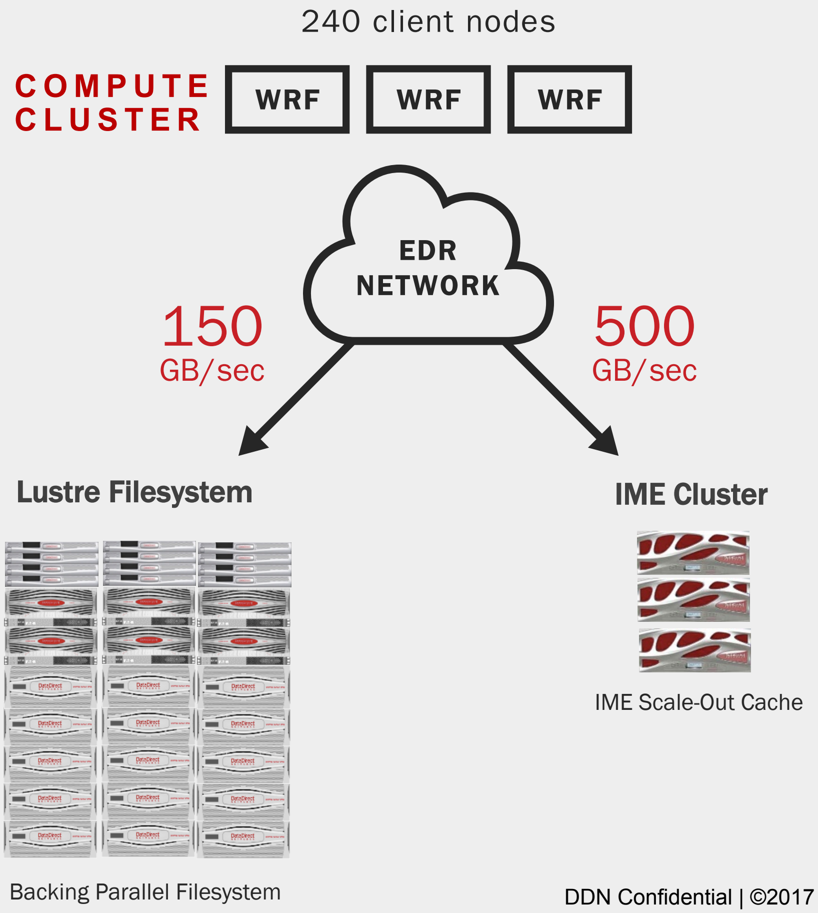
Large Capacity GS

WRF ON IME

Weather Code Performance

Simulating Ensemble runs of WRF – multiple jobs executing concurrently

3.8x I/O Speedup versus Lustre Filesystem in 1/20th RU and 1/10th Power Envelope



240 client nodes

COMPUTE CLUSTER

WRF

WRF

WRF

EDR NETWORK

150 GB/sec

500 GB/sec

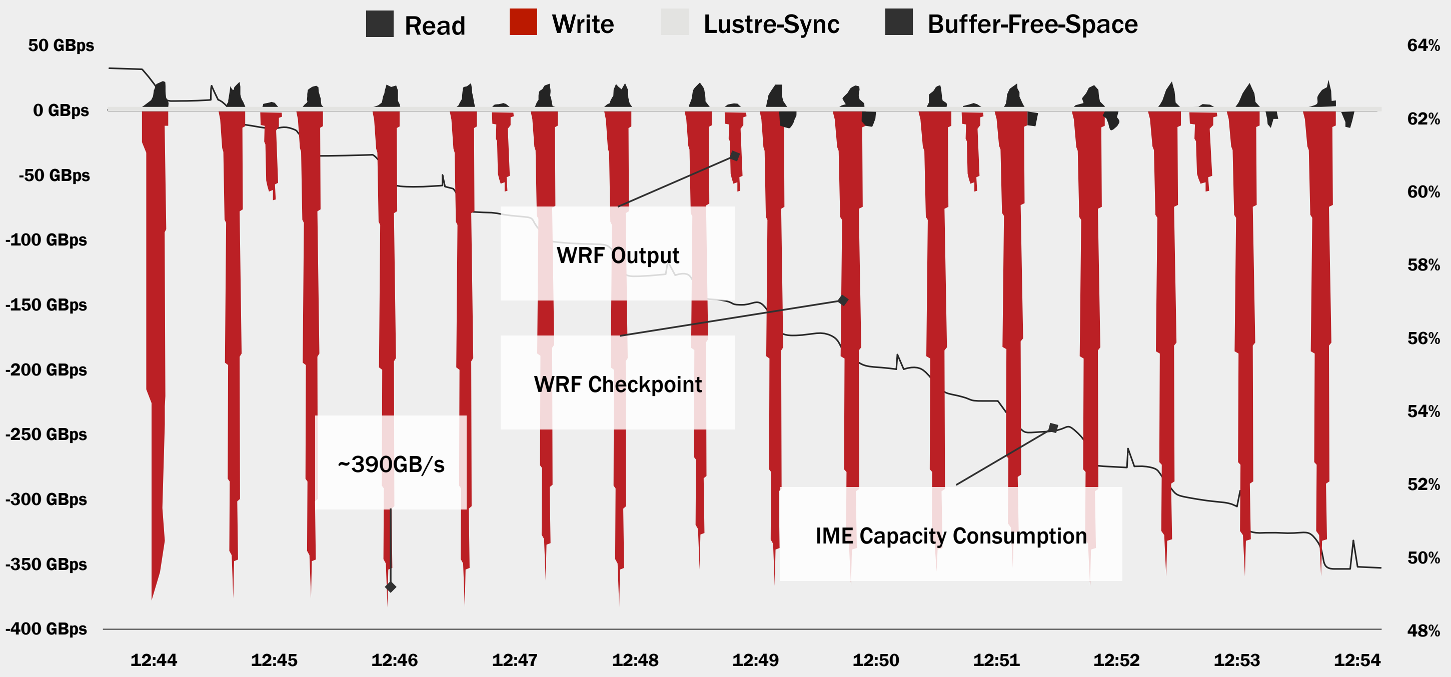
Lustre Filesystem

IME Cluster

IME Scale-Out Cache

Backing Parallel Filesystem

ICHEC – IME NVME PERFORMANCE MONITORING – AGGREGATED THROUGHPUT



WRF Output

WRF Checkpoint

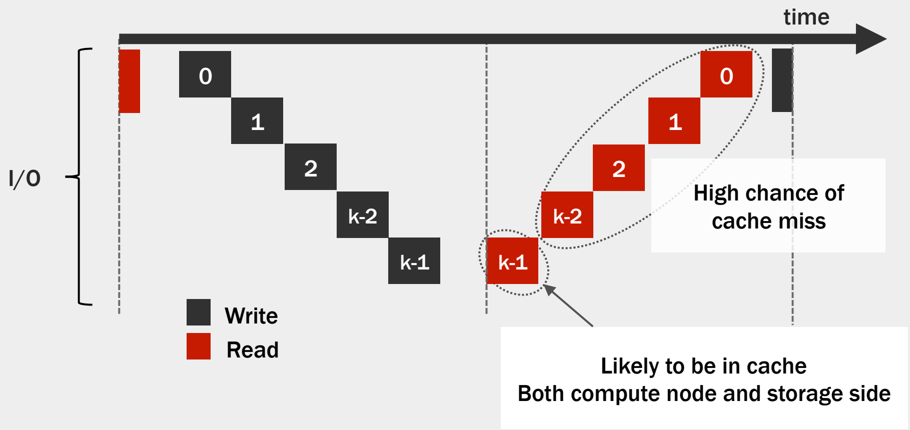
~390GB/s

IME Capacity Consumption

IME

Oil and Gas | Seismic

IME expands the cache volume from GBs to PBs and eliminates cache misses associated with Reverse Time Migration IO patterns



High chance of cache miss

Likely to be in cache
Both compute node and storage side

IME

TORTIA (Reverse Time Migration Code)

Seismic datasets are too large for RAM. IME delivers Reverse Time Migration speed ups 5-50x with IME over PFS

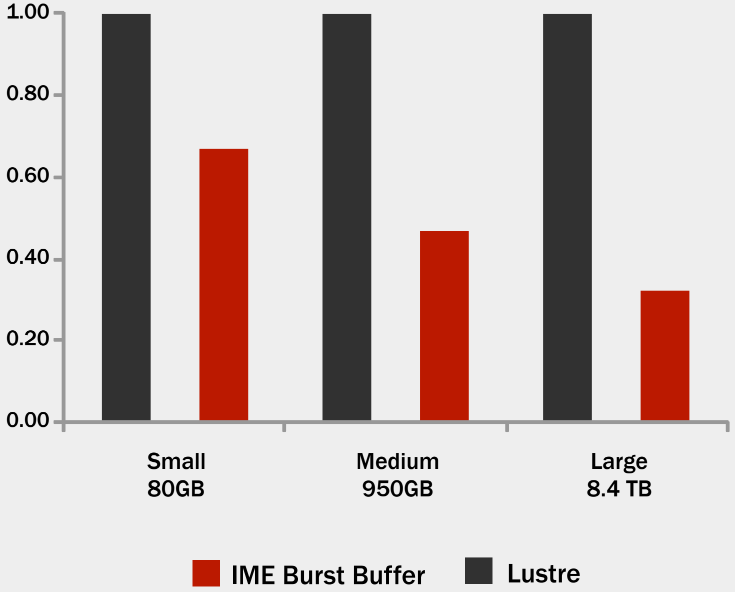
Other Novel solutions don't deliver the combination of performance, ease-of-use and capacity that and IME solution delivers

10x

Speedup in IO time

3x

Speedup in run time



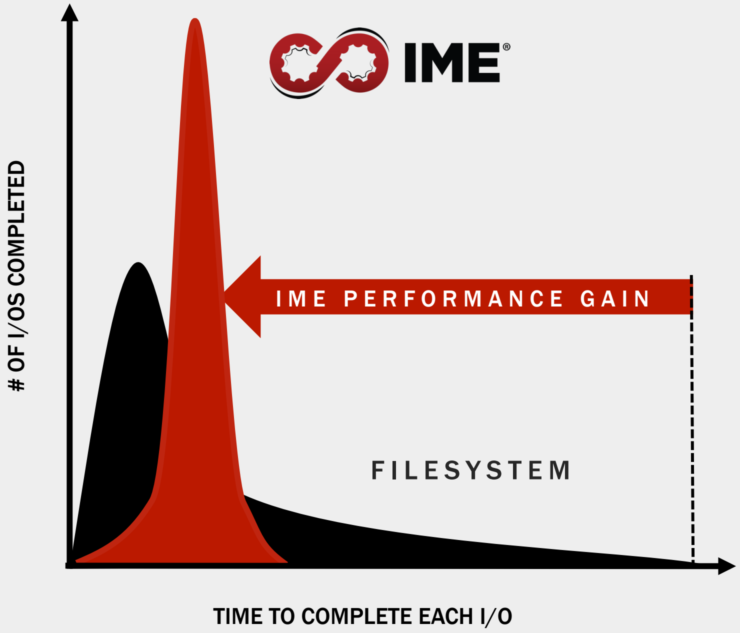
DDN STORAGE

IME

Performance Consistency

File systems, particularly with HDDs can exhibit extremely long tails for the completion of the last I/O. Typically an application only cares about completion of the last I/O

IME exhibits extremely consistent I/O times without the long tail effect that impacts applications.


IME
IME PERFORMANCE GAIN
FILESYSTEM
# OF I/OS COMPLETED
TIME TO COMPLETE EACH I/O