DDN® STORAGE

Software-Defined Memory for At-scale and High-Performance Computing

Kurt Kuckein
Director of Product Management

August, 2016

DDN | About Us

Solving Large Enterprise and Web Scale Challenges

History

- Founded in ’98
- World’s Largest Private Storage Company
- Growing, Profitable, Self Funded

Headquarters:

Santa Clara and Chatsworth, CA

World-Renowned & Award Winning

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

What Are the Challenges?

At a Macro Level

For the past two decades,
two trends have created massive I/O bottlenecks . . .

SPEED

Everything related to silicon evolution has gotten faster.

Systems, Storage, Processors, Servers, Interconnect

SCALE

The data to process, distribute, share is much larger.

Demanding Applications, Collaboration At-scale

DATA-INTENSIVE WORKFLOWS ACROSS ALL INDUSTRIES, ARE BRINGING IT INFRASTRUCTURE TO ITS KNEES

What Are the Challenges?

At an HPC Datacenter Level

As speed & scale grow, so does contention and inefficiencies in both your file system and infrastructure. . .

ISOLATING PROBLEM APPLICATIONS

That bring the PFS and entire cluster to its knees, killing SLAs

MINIMIZING IDLE RESOURCES

People & compute waiting for I/O from the PFS and spinning disk

IMPROVING BANDWIDTH PERFORMANCE

Needed to accelerate critical applications and workflows

INCREASING EFFICIENCY & DENSITY

To end the hardware sprawl caused by using HDDs for performance

GAME CHANGING TECHNOLOGY IS NEEDED TO BREAK THROUGH THE LIMITATIONS THAT HOLD BACK PROGRESS

5

What Are the Challenges?

At a Business Level

As speed & scale grow, so does the requirement for instantaneous access, insight and results

ACCELERATING TIME TO RESULTS

To gain competitive edge and improve top line

LOWERING COST

Creating new efficiencies to improve bottom line

INCREASING CAPABILITIES

Unlocking new possibilities, enabling more jobs in parallel

MINIMIZING CHANGE & LOCK-IN

Disrupt inefficiencies - not processes, while retaining flexibility

A NEW TECHNOLOGY IS REQUIRED THAT'S BUSINESS ENABLING AND DISRUPTS STATUS QUO

6

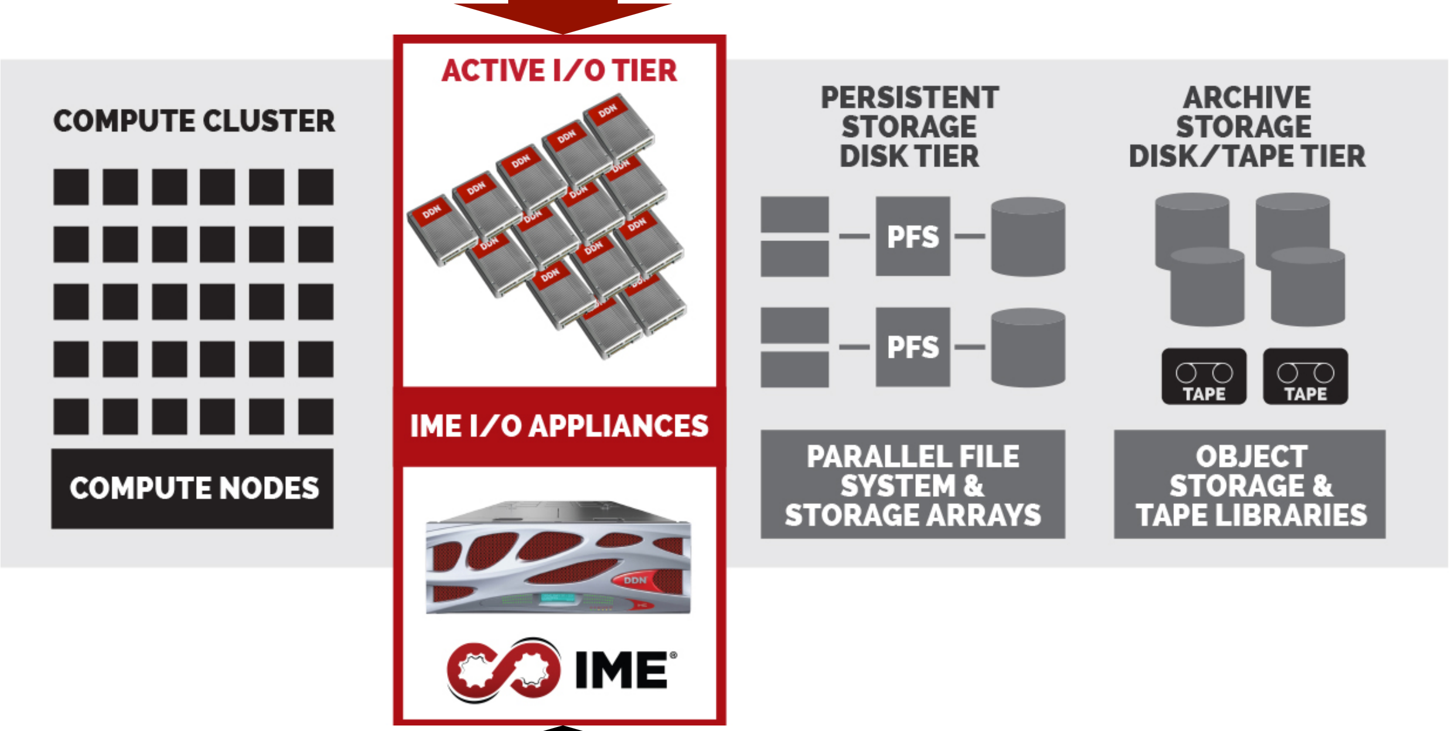
What is IME?

Speeds I/O, Applications and Workloads

7

IME® Is
The New I/O Acceleration Architecture

IME's *Active I/O Tier*, is inserted right between Compute and the parallel file system


COMPUTE CLUSTER
COMPUTE NODES
ACTIVE I/O TIER
DDN
IME I/O APPLIANCES
IME®
PERSISTENT STORAGE DISK TIER
PFS
PFS
PARALLEL FILE SYSTEM & STORAGE ARRAYS
ARCHIVE STORAGE DISK/TAPE TIER
TAPE
TAPE
OBJECT STORAGE & TAPE LIBRARIES

IME software intelligently virtualizes disparate NVMe SSDs into a single pool of shared memory that accelerates I/O, PFS & Applications

3 Key functions of IME

① **A Write Accelerating Burst Buffer**

Absorbing the bulk application data into the NVMe solid state cache significantly faster than the file system can absorb it

② **A File System Accelerator and Application Optimizer**

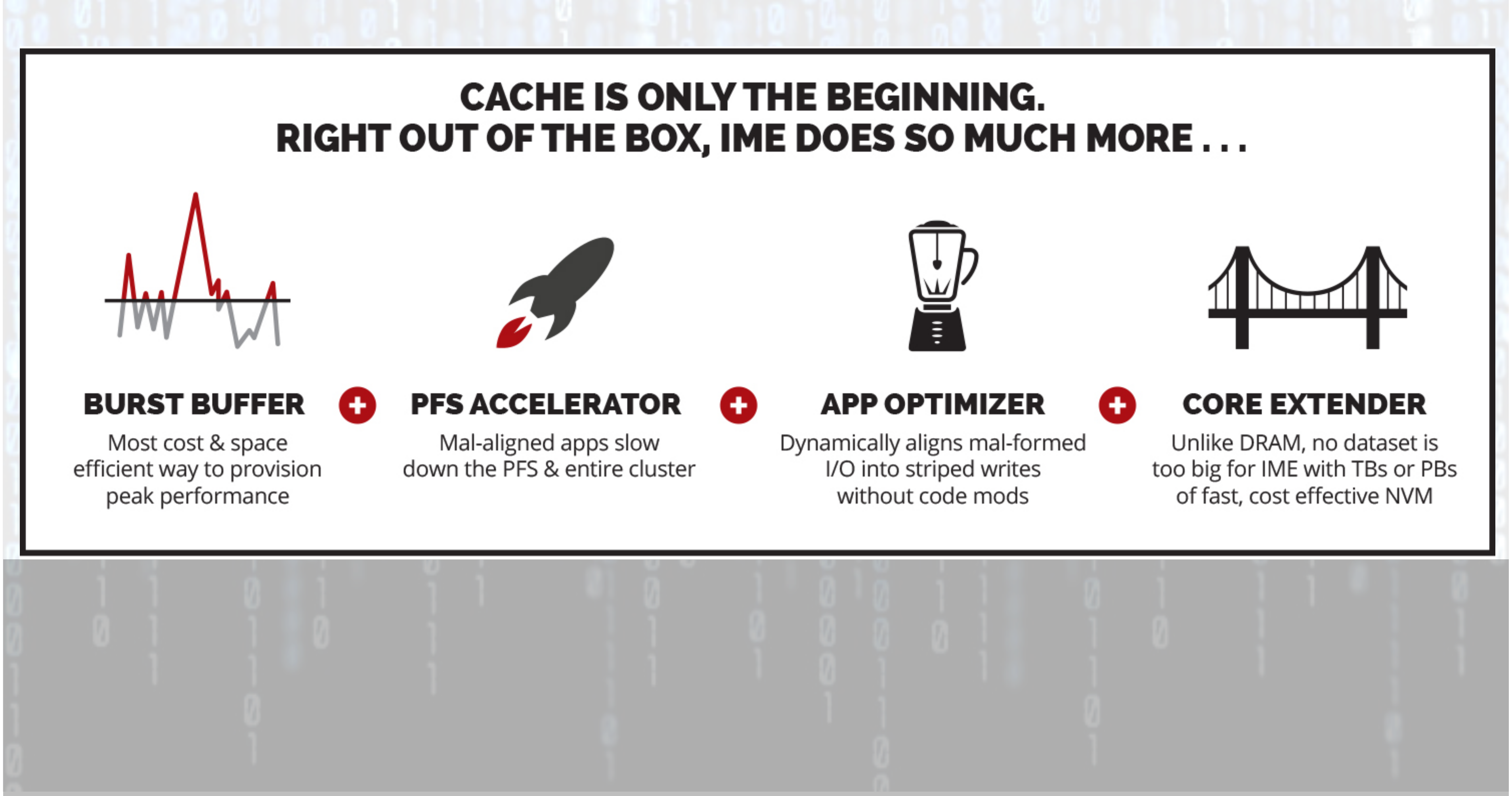
As IME reorders application I/O to optimize flushing the cache to long term storage (enabling purchasing as little expensive cache possible).

③ **A Read-optimized Application I/O Accelerator**

That enables out-of-band API configuration of the IME appliance to optimize both reads and writes, allowing more simultaneous job runs, shortening the job queue and enabling significantly faster application run time to the user. The API integrates IME with the job schedulers and pre-stages / warms the cache for new jobs, accelerating first read.

IME: A Burst Buffer & Way Beyond

Game Changing Technology, Enabling Your Next Leap Forward



CACHE IS ONLY THE BEGINNING.
RIGHT OUT OF THE BOX, IME DOES SO MUCH MORE . . .

BURST BUFFER +
Most cost & space efficient way to provision peak performance

PFS ACCELERATOR +
Mal-aligned apps slow down the PFS & entire cluster

APP OPTIMIZER +
Dynamically aligns mal-formed I/O into striped writes without code mods

CORE EXTENDER
Unlike DRAM, no dataset is too big for IME with TBs or PBs of fast, cost effective NVM

DDN STORAGE

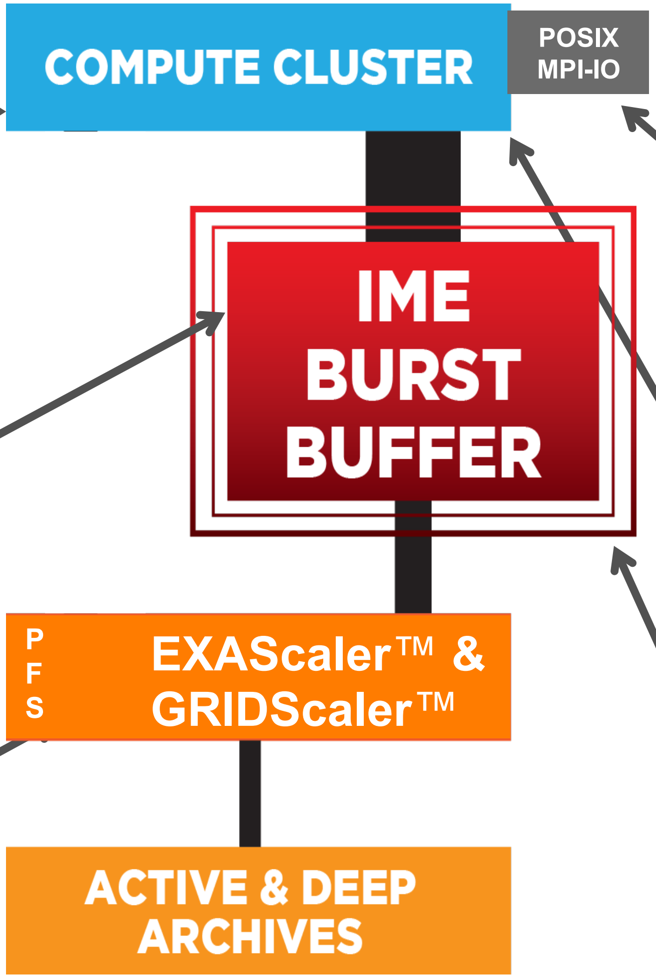
Introducing IME®

Technical Key Components and Operations

Transparent to Apps
A thin IME Client resides on cluster compute or I/O nodes.

Standards Embraced
Traditional Lustre® (or GPFS™) Client model with MDS (or NSD) interface

Application Transparency
The Parallel File System (PFS) is unaware that IME is an intermediary



No Source Modifications
Interfaces include: POSIX, MPI-IO, ROMIO, HDF5, NetCDF, etc.

Breakthrough Performance & Scalability
Low-level communications key value pair based protocol - not POSIX

Intelligent File Management
IME Server software does all the heavy lifting

11

Where IME Helps

Resolving challenges, environments, key industries, workflows

Where IME® Helps

SYMPTOMS . . .

- **Bursty I/O patterns**
- **I/O intensive, high bandwidth workloads**
- **Mal-aligned I/O HPC applications** (that run significantly below line rate into the PFS and/or slow down the compute cluster)
- **Regular checkpoint/restart operations** (on apps that scale beyond 20% of the size of the cluster)
- **Greenfield projects**
- **Limited power, space**
- **Exascale POC projects**

NAME ______________________________

ADDRESS ____________________ DATE __________

℞

PRESCRIPTION . . .

- **Maximize time available for computation**
- **Enable predictable performance for SLAs**
- **Accelerate application runtimes and/or time to discovery/results**
- **Reduce checkpoint times by 10x**
- **Run more jobs in parallel**
- **Run analytics/processing/ modeling in real-time**

Challenge of Bottlenecks At-scale

For Oil & Gas Companies

	ACQUISITION	PROCESSING	MODELING
I want . . .	Fastest Time to Oil	Fastest, High Fidelity Processing	Faster, More Accurate Models
Challenges	• Write-intensive operations • Growing datasets • Near-real time requirement	• Supporting Mixed I/O from ISV + homegrown • Sharpening data against historical models • Run more apps in parallel	• Mixed I/O (large & small) • Idle resources during data transfer • Growing complexity of code
Where am I blocked?	• PFS Performance • HDD Latency • POSIX locking	• Malformed I/O slowing PFS • Reading out of PFS, not cache • RTM's Out of core data runs slow on PFS	• Code may not be optimized for PFS • Slow reads from PFS • Ensemble behavior slowing down app
How will IME help?	• Accelerates apps • Ingests at line rate	• Aligns I/O into full stripes • Eliminates slow reads, now in cache • Runs large sets in cache	• Extends peak performance • Speeds apps • Reduces compute

Challenge of Bottlenecks At-scale

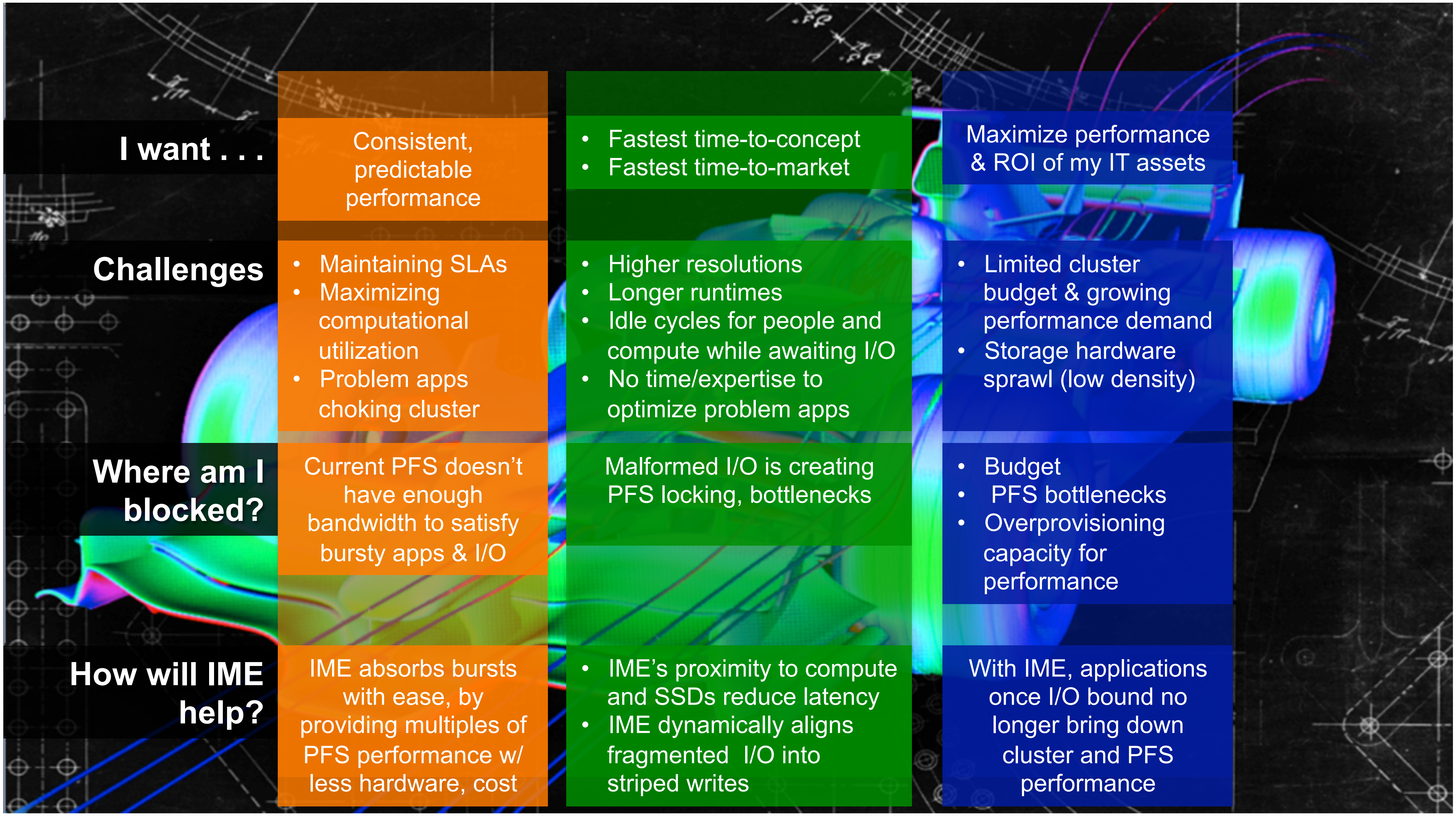
For National Labs & Universities

I want . . .	To eliminate “problem” apps	Provision performance with less hardware	Faster checkpoint/restart
Challenges	• Entire cluster slows down • Idle cluster time • Scientists (not programmers) writing apps	• Overprovisioning HDD capacity for performance • Overprovisioning compute to make-up for I/O wait times • Infrequent use of peak perf	• Expensive compute Idle during checkpoint • Writing checkpoints to scratch is time consuming & expensive • Restart from HDD is slow
Where am I blocked?	• POSIX locking • HDD latency • Malformed I/O • Ability/time to optimize apps	• IOPS & bandwidth come from HDDs • Inability to decouple capacity from performance • Buying HDDs for “Peak” requirements	• PFS & HDD latency • Fragmented I/O patterns choke PFS
How will IME help?	• Eliminates “problem” apps • Accelerates I/O • Aligns fragmented I/O into stripes	• Lower latency and increased I/O performance mean less compute	• Accelerate Checkpoint Restart • Checkpoints live in IME, no writes to scratch needed

DDN STORAGE

Challenge of Bottlenecks At-scale

For Manufacturing



I want . . .	Consistent, predictable performance	- Fastest time-to-concept - Fastest time-to-market	Maximize performance & ROI of my IT assets
Challenges	- Maintaining SLAs - Maximizing computational utilization - Problem apps choking cluster	- Higher resolutions - Longer runtimes - Idle cycles for people and compute while awaiting I/O - No time/expertise to optimize problem apps	- Limited cluster budget & growing performance demand - Storage hardware sprawl (low density)
Where am I blocked?	Current PFS doesn't have enough bandwidth to satisfy bursty apps & I/O	Malformed I/O is creating PFS locking, bottlenecks	- Budget - PFS bottlenecks - Overprovisioning capacity for performance
How will IME help?	IME absorbs bursts with ease, by providing multiples of PFS performance w/ less hardware, cost	- IME's proximity to compute and SSDs reduce latency - IME dynamically aligns fragmented I/O into striped writes	With IME, applications once I/O bound no longer bring down cluster and PFS performance

16

IME®
How it Helps

View Slide
in
Full Screen

17

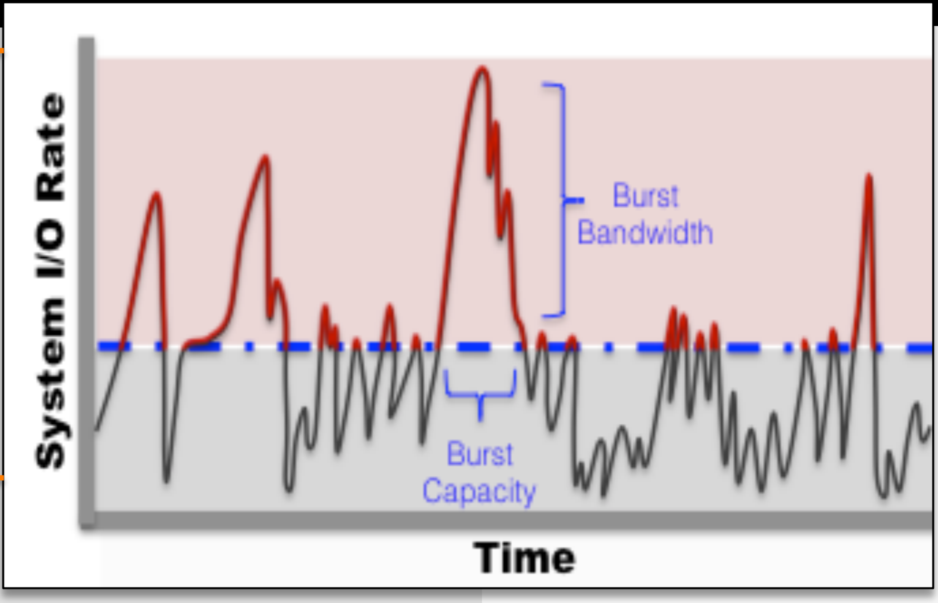
Provisioning Performance with HDDs is Highly Inefficient

BEFORE IME:

PFS Systems were designed to handle the entire performance load

This required lots of storage controllers, enclosures and drives to deliver full bandwidth –

which was rarely used **. . .**



STORAGE BANDWIDTH UTILIZATION OF A MAJOR HPC PRODUCTION STORAGE SYSTEM

- **99%** of the time **< 33% of max**
- **70%** of the time **< 5% of max**

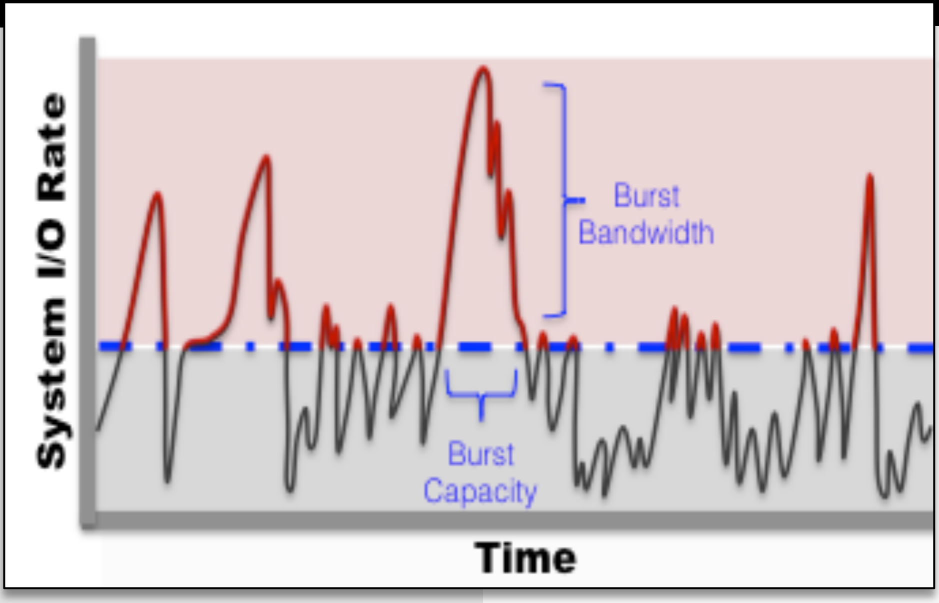
DDN STORAGE

Game Changing Bandwidth

IME Disrupts How Performance is Provisioned

View Slide in Full Screen

TODAY, WITH IME:



IME's BURST BUFFER Absorbs the Peak Load

PARALLEL FILE SYSTEM Handles the Sustained Load

IME enables peak performance to be provisioned with much less hardware, power, space

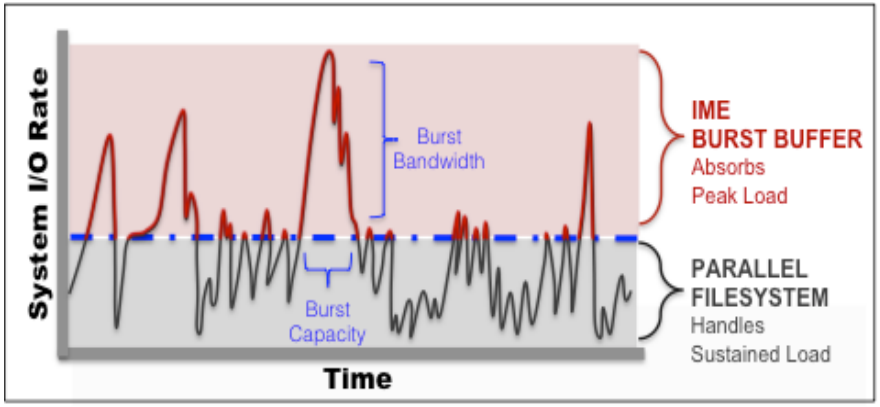
How Does IME® Help?

A more efficient way to provision IOPS and bandwidth than HDDs

STORAGE BANDWIDTH UTILIZATION OF A MAJOR HPC PRODUCTION STORAGE SYSTEM

- **99%** of the time **< 33% of max**

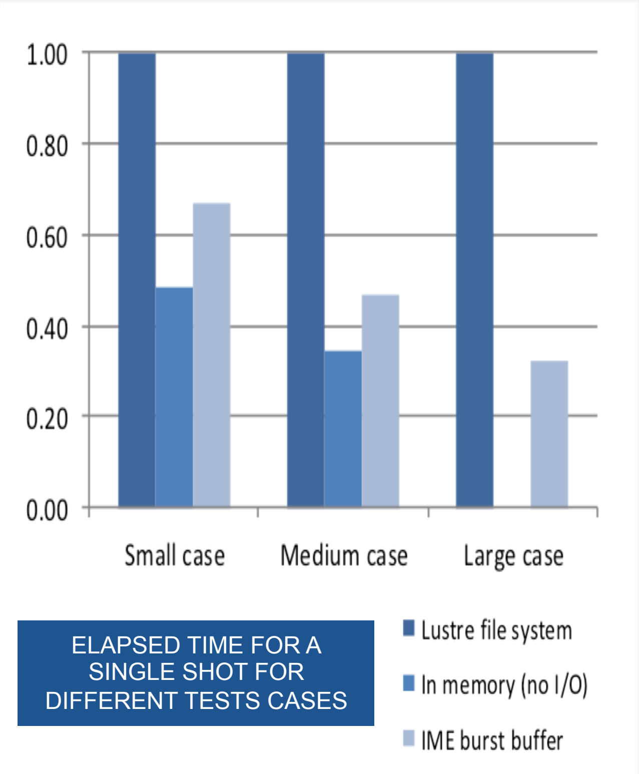
IME separates the provisioning of peak & sustained performance requirements with greater operational efficiency and cost savings



✓ **IME Reduces Storage Hardware**
Fewer systems to buy, power manage, maintain

Real World Results

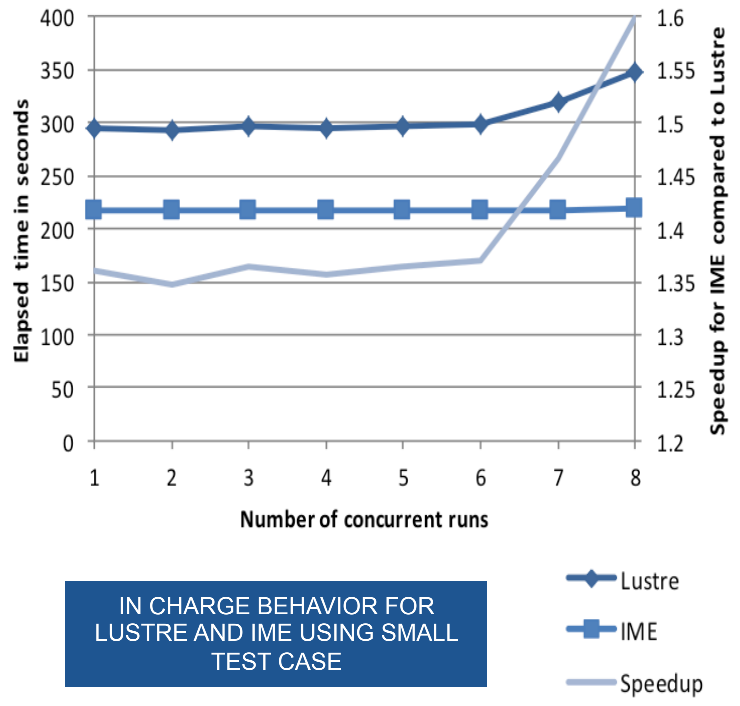
ICHEC researchers benchmarked their I/O bound, 3D seismic imaging application against IME and Lustre. Processing a _100 MB seismic image requires 1 TB of data movement_. Typical surveys consist of 1000s of images. Therefore, _1 PB of data movement is expected by this application for a typical survey_.



ELAPSED TIME FOR A SINGLE SHOT FOR DIFFERENT TESTS CASES

Left: Single "shot" profiles that include compute, communication, and IO time (normalized against Lustre baseline).

Right: Wall clock time for execution of multiple concurrent "shots."



IN CHARGE BEHAVIOR FOR LUSTRE AND IME USING SMALL TEST CASE

How Does IME® Help?

The only open file system acceleration product that supports any system environment

IME is a non-vendor-captive software approach, providing much greater flexibility in how users can architect their environments and a wider selection of specialty and commodity hardware platforms & components

- ✓ Compute Vendor-Agnostic
- ✓ Storage Vendor-Agnostic
- ✓ Interconnect-Agnostic
- ✓ Flash Form Factor-Agnostic
- ✓ Application-Agnostic
- ✓ Deployment-Agnostic

How Does IME® Help?

Multiplies compute performance

With every minute IME saves on job runs, your site can now run more jobs on the same compute resources

IME Eliminates:

- ✓ Parallel file system locking, limitations & bottlenecks
- ✓ Storage hardware, consumed floorspace
- ✓ Latency driving a loss of compute resources
- ✓ A considerable portion of Checkpoint/restart downtime

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