Addressing Emerging Issues of Data at Scale

Keith Miller, VP Technical Services and Support

DDN | Who We Are

We Design, Deploy and Optimize Storage Systems Which Solve HPC, Big Data and Cloud Business Challenges at Scale

Main Office: Sunnyvale, California, USA
Installed Base: 1,000+ Customers in 50 Countries
Go To Market: Partner & Reseller Assisted, Direct
DDN: World's Largest Private Storage Company

World-Renowned & Award-Winning

All Time Winner

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

An Elite Collection Of HPC's Finest...

Our 1000+ Customers Include over 2/3 of the Top100

Big Data & Cloud Infrastructure

DDN Announced Product Portfolio

Cloud Tiering

Big Data Platform Management

DirectMon

Analytics Reference Architectures

SAP
sas
hadoop

Petascale Lustre* Storage

EXAScaler™
10Ks of Clients
1TB/s+, HSM
Linux HPC Clients
NFS & CIFS

Enterprise Scale-Out File Storage

GRIDScaler™
~10K Clients
1TB/s+, HSM
Linux/Windows HPC Clients
NFS & CIFS

Cloud Foundation

WOS® 3.0
32 Trillion Unique Objects
GeoReplicated Cloud Storage
256 Million Objects/Second
Self-Healing Cloud
Parallel Boolean Search

WOS7000
60 Drives in 4U
Self-Contained Servers

Storage Fusion Architecture™ Core Storage Platforms

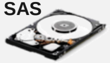
SFA12Kx
48GB/s/1.4M IOPS
1,680 Drives:
16U-2 Racks
Embedded Computing

SFA7700
13GB/s, 400K IOPS
60 Drives in 4U;
396 Drives in 20U

Flexible Drive Configuration

SATA

SAS



SSD

SFX Automated Flash Caching

- Read
- Write*
- Context Commit
- Instant Commit

* Future Release

Evolution of Cache-Centric Storage at DDN

ReACT™

- SFA Feature - Real-time intelligent cache management

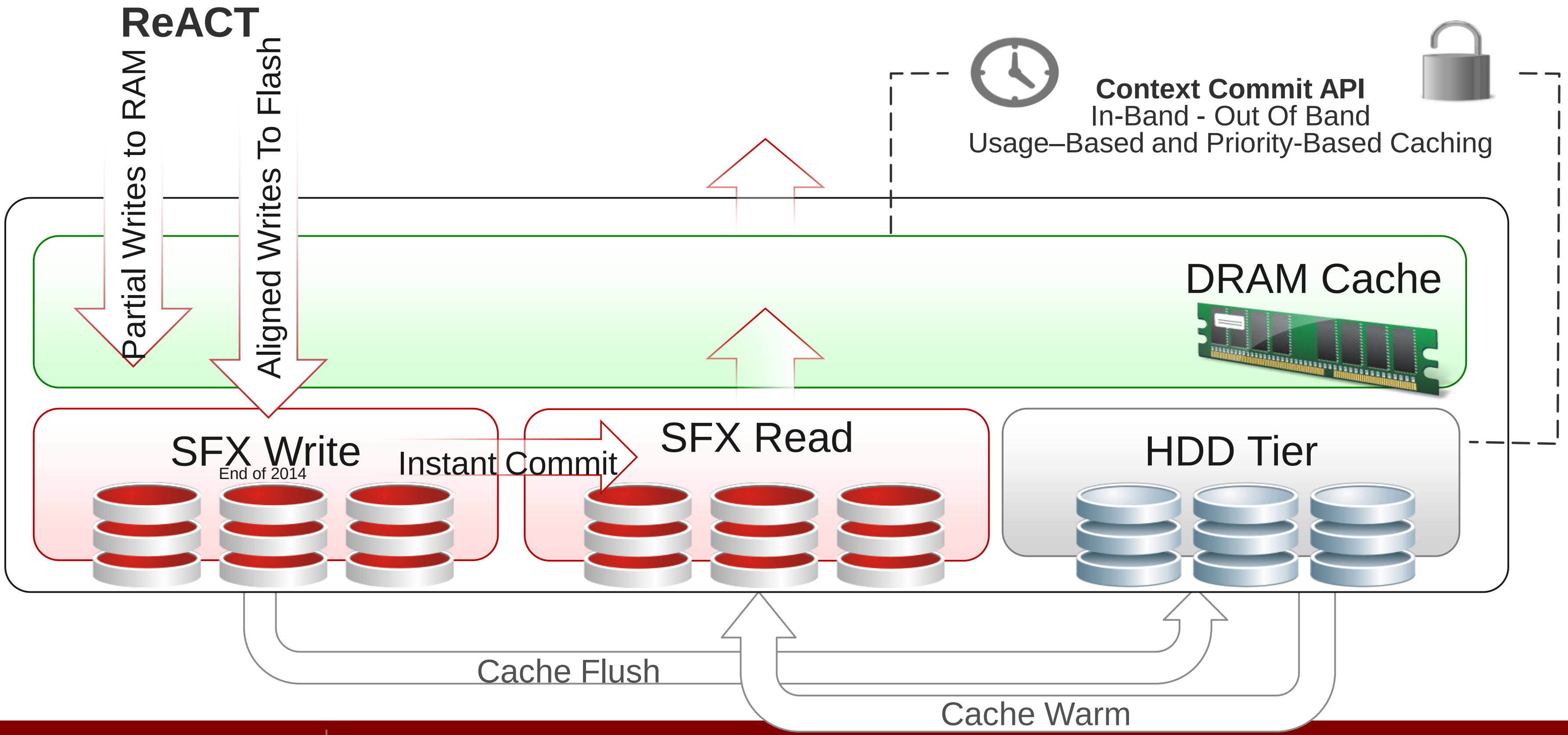
SFX

- Application-Aware Flash Caching

Infinite Memory Engine (IME)

- Exascale capable Burst Buffer

The Many Dimensions of SFA Acceleration



ReACT

- Caches unaligned IO
- Allows full stripe writes pass directly to disk
- Faster, safer data

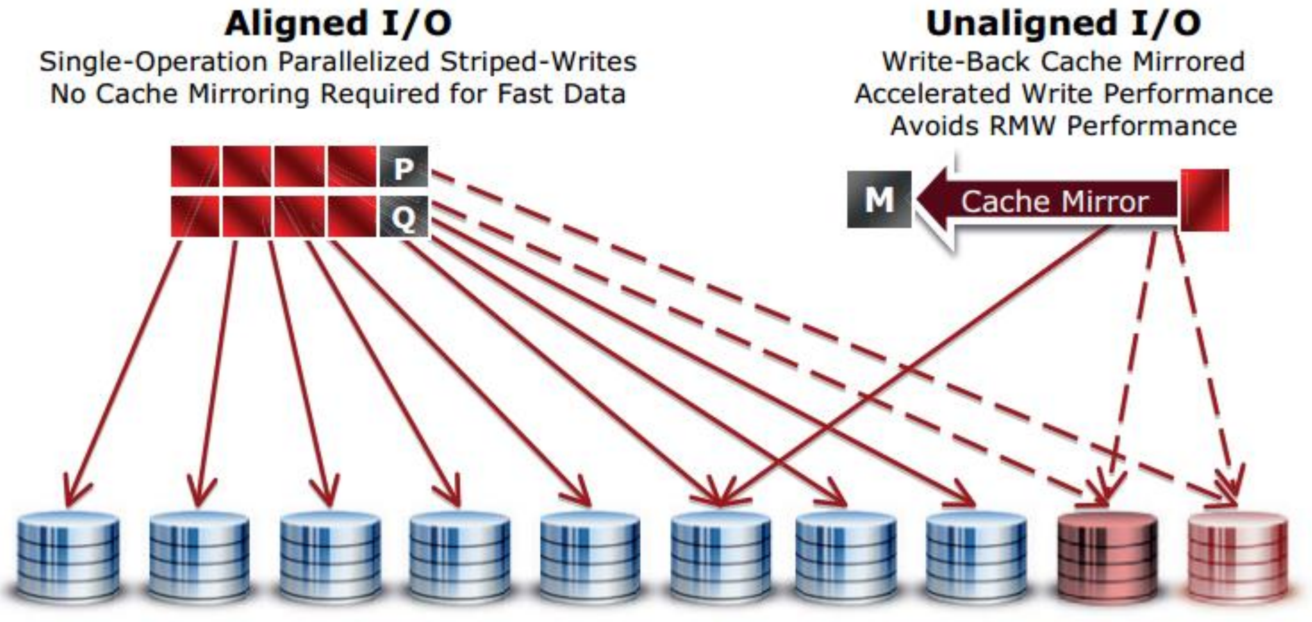


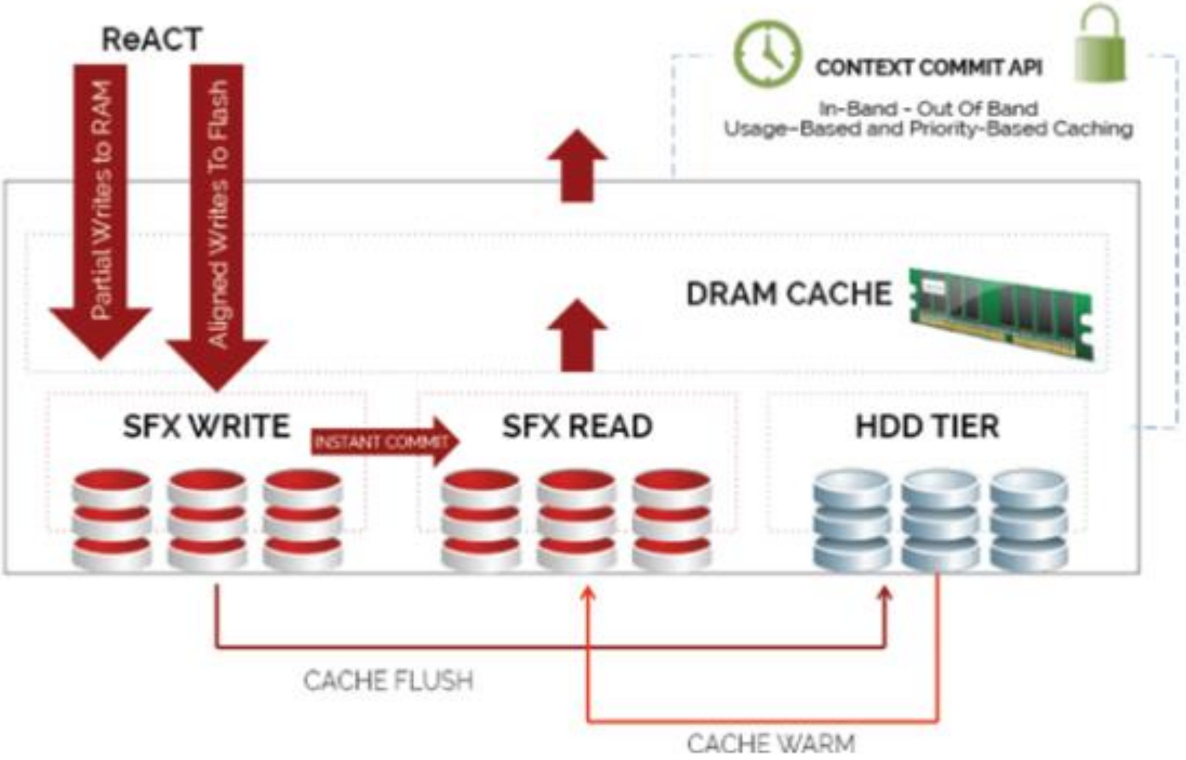
Figure 5 – Optimizing Cache Utilization with ReACT

SFX

ReACT works with SFX to bypass DRAM for aligned writes

In-Band Hints provided through API

Helps accelerate Big Data workloads with a combination of streaming as well as transactional IO

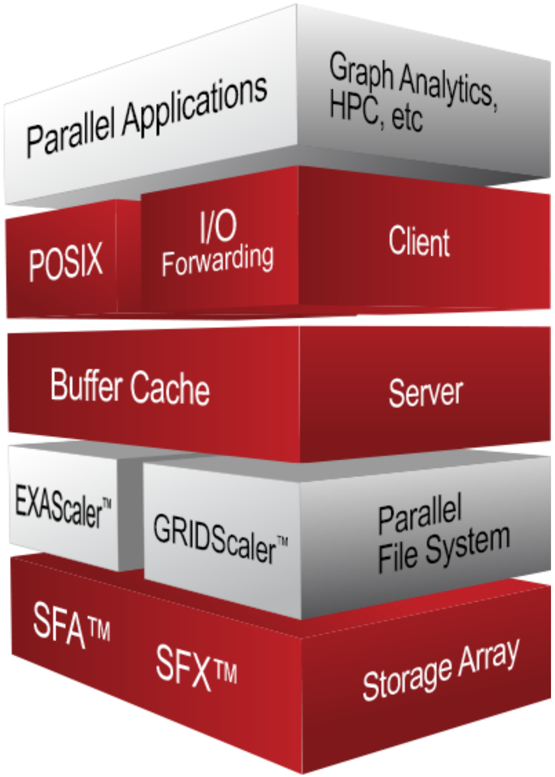


Infinite Memory Engine (IME)

What is Infinite Memory Engine (IME)?

A DDN-developed and patent-protected Distributed Hash Table (DHT) algorithm that manages distributed, non-volatile memory devices:

- High bandwidth, Low latency I/O
  - reads & writes
  - large and small
  - aligned or random
- Data integrity & protection
  - Cached application data
  - DHT metadata
- Massive scalability



What does IME Do?

Goal: Provide a scalable, high-bandwidth, system-level storage service / resource for accelerating I/O

Shrinks IO Phases for reduced time to solution for Petascale & Exascale computing systems:

- PFS I/O acceleration and file caching (checkpoint-restart)
- Accelerating HPC data analysis activities

Application I/O Acceleration

- Checkpoint-Restart
- File-based data cache
- Stage-in, Stage-out, Demand Loading

Out-of-Core I/O

Data Analysis Support

- Post-processing
- Visualization

Temporary Data Storage

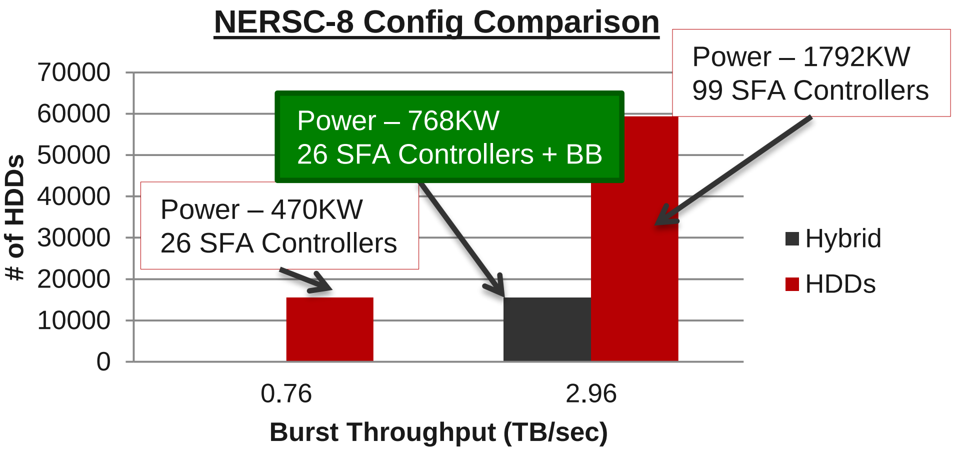
- Sequential-job Data Sharing (many-task computing, …)
- Concurrent-job Data Sharing (coordinated sharing of data through several tasks)
- Intermediate Results

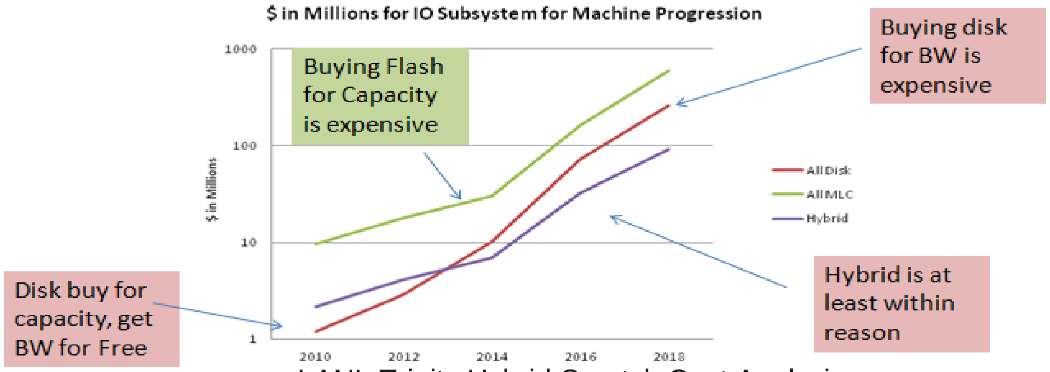
I/O Scaling Challenges

The first challenge is cost….

- TB/s I/O with HDDs is unwieldy and too expensive
  - Requires thousands of HDDs
  - Need spindles for performance but also get much more excess capacity
  - Power requirements become prohibitive
- From a system perspective there are too many moving parts to even build it
  - ~500,000 HDDs required for 100TB/s

Hybrid approach is necessary to meet bandwidth & capacity requirements





LANL Trinity Hybrid Scratch Cost Analysis

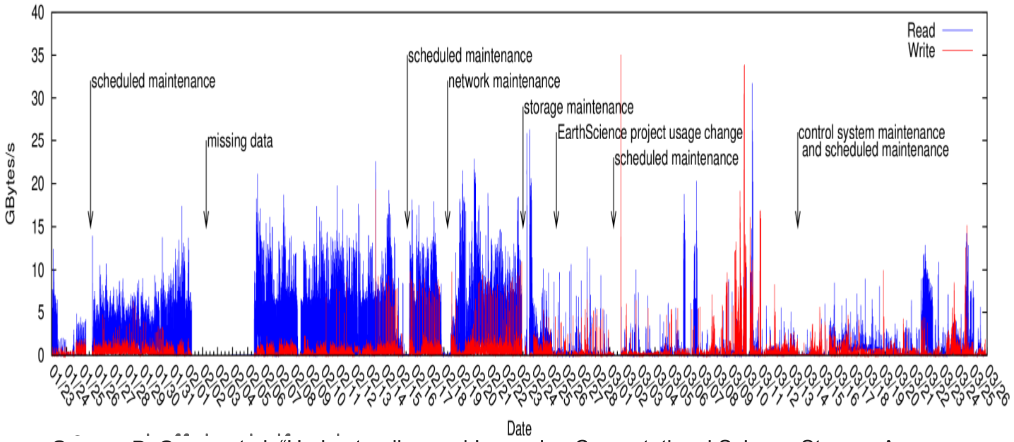
Source: Trinity FSIO NRE Presentation, LA-UR-11-11964

I/O Scaling Challenges

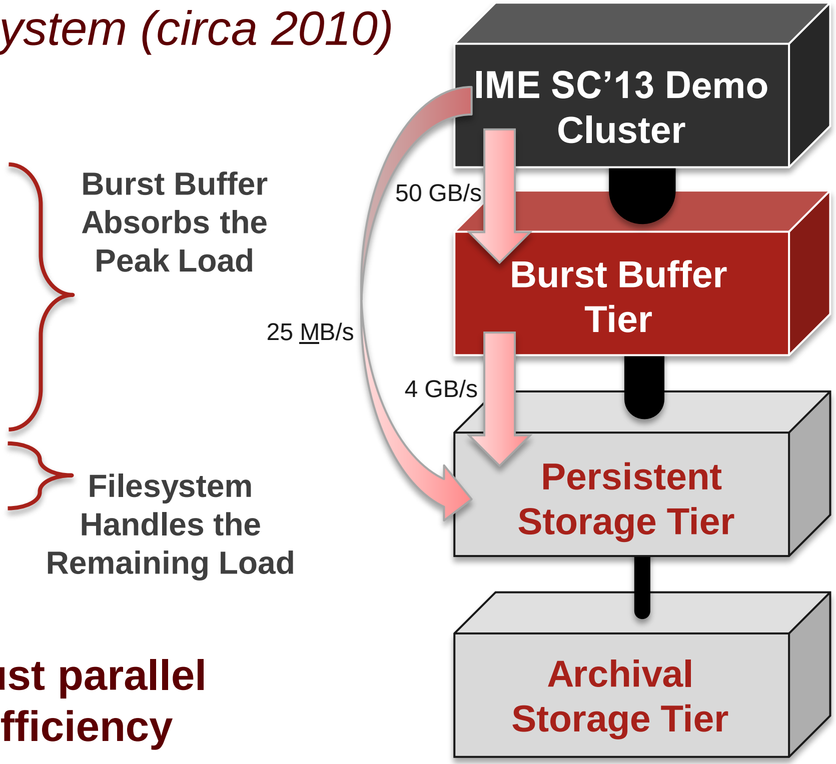
Another challenge is efficiency…

Analysis: Argonne's LCF production storage system (circa 2010)

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



Source: P. Carns, et al, "Understanding and Improving Computational Science Storage Access Through Continuous Characterization", Proceedings of MSST 2011, 2011



Trend: Burst Buffers will demand smaller, robust parallel filesystems that require very high bandwidth efficiency

Why is today's I/O efficiency so poor?

Serialization

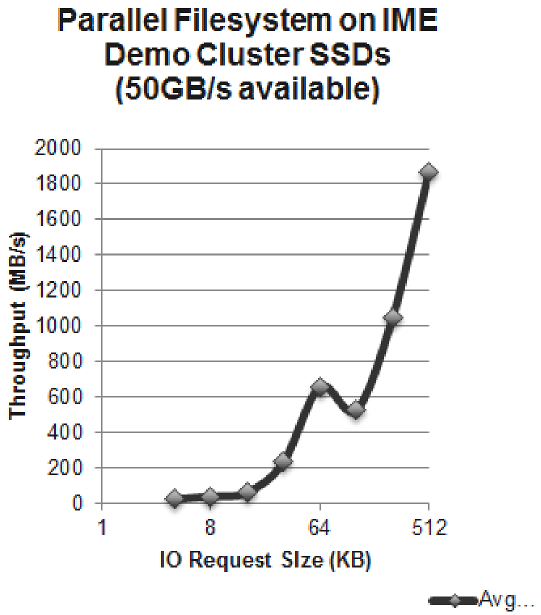
- Stripe and block alignment (PFS and RAID)
- Lock contention
- Exacerbated by poor I/O structuring in applications
- As compute resources get larger, lock contention worsens

Poor Horizontal Scaling

- PFS are only as fast as the slowest I/O component
- Oversubscribed or crippled I/O components affect the entire system performance
- As I/O requirements get larger and # of components increases the problem worsens (congestion)
- This weakest link can be all the way down to disks (RAID rebuilds/slow drives)

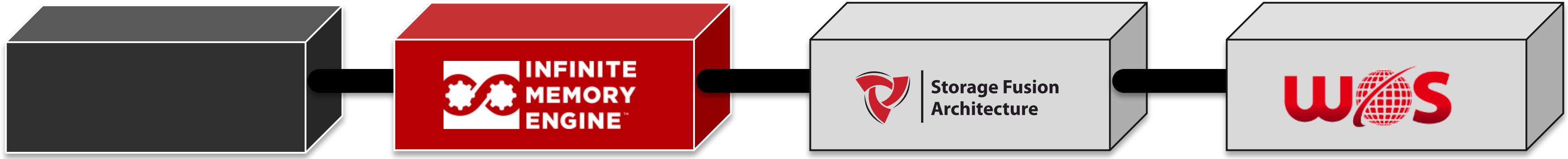
Scaling

- Faster media (SSDs) may not address the underlying PFS performance limitations



HPC Storage Hierarchy with IME

Exascale Compute Platform	**DDN (IME) Burst Buffer Tier**	**DDN SFA Persistent Storage Tier**	**DDN WOS Archival Storage Tier**



Massively Parallel Processors + Network	High-Speed File Caching Appliances	High-Speed, High-Capacity, Block & File	Power Efficient, High-Capacity Object Cloud

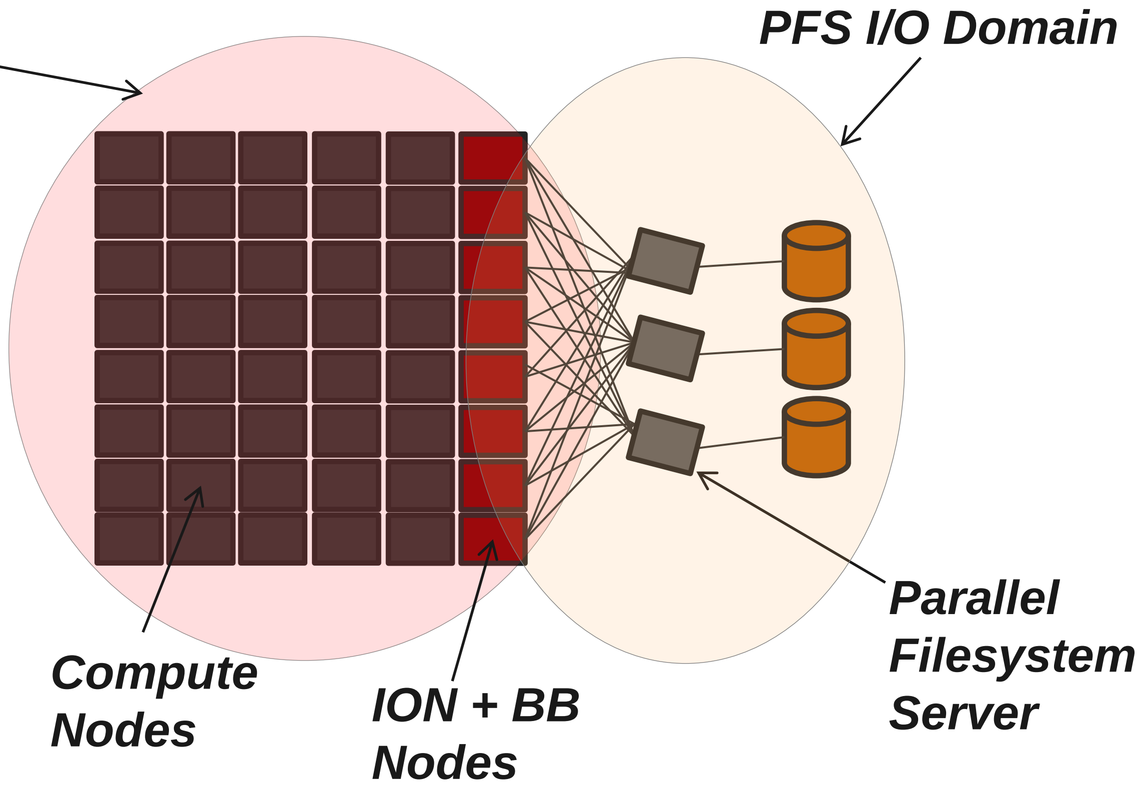
HPC Cluster Example (2018)		
Machine Size/Checkpoint	100M Threads, 2PB Checkpoint in 5m	
	Disk Only	Node Local IME + SFA
Peak Performance	7 TB/s	8 TB/s
Spinning Disks	33,400	3,240
Racks	100	12
Storage Cost	**$33M**	**$7M**

- Dramatically Reduces HPC Computing Cost
- Enhances Performance, Efficiency, & Power Usage by Orders of Magnitude
- Reduces Network & Server Amount & Complexity by Orders of Magnitude
- Patented Algorithms Deliver Unlimited Scale
- Liberates Applications from File System Serialization and Scalability Bottlenecks

IME Bulk Data Caching

Caching I/O Domain

1. Parallel Log-Structured Writes
   - Transforms random I/O into sequential I/O
   - Fast, Lockless
   - Highly entropic metadata
2. Dynamic load balancing of cached data
   - Eliminates bottlenecks on I/O path
3. O(1) lookup cost for cached data
   - Fast lookup of any I/O fragment
   - Normalize cached fragments during idle I/O periods



Managing IME's Bulk Data Cache

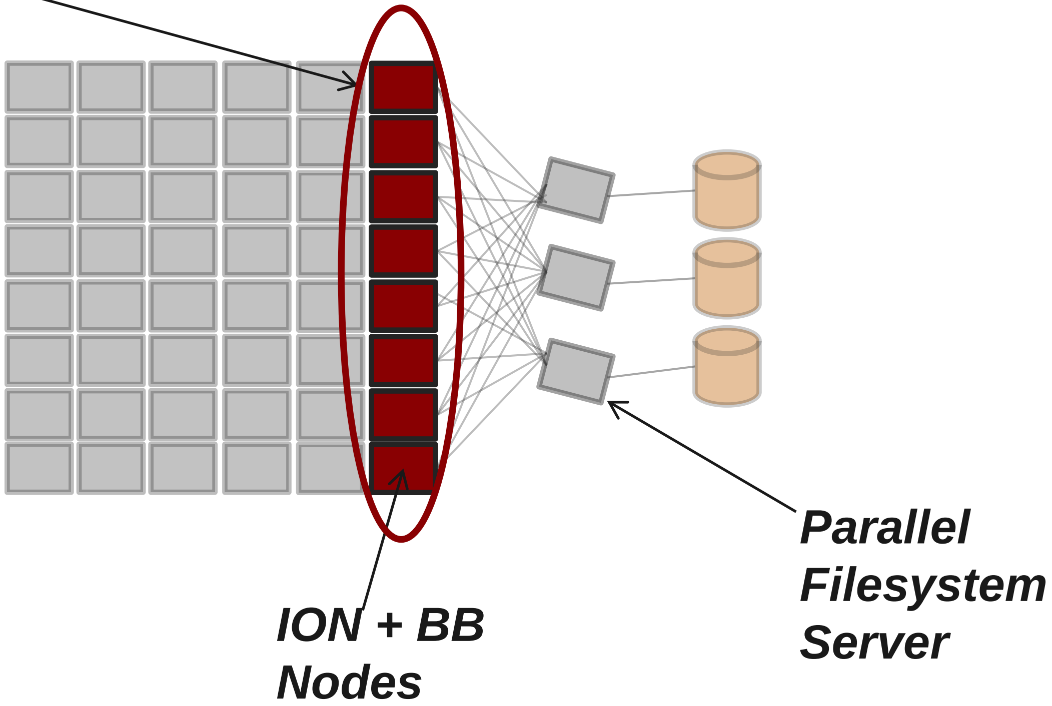
Cache Metadata Management Domain

1. Cache metadata describing bulk data is not passed to the PFS domain

2. Structures are managed in parallel and evenly distributed – highly scalable

1. Writing or pre-staging into IME automatically distributes log-structure metadata

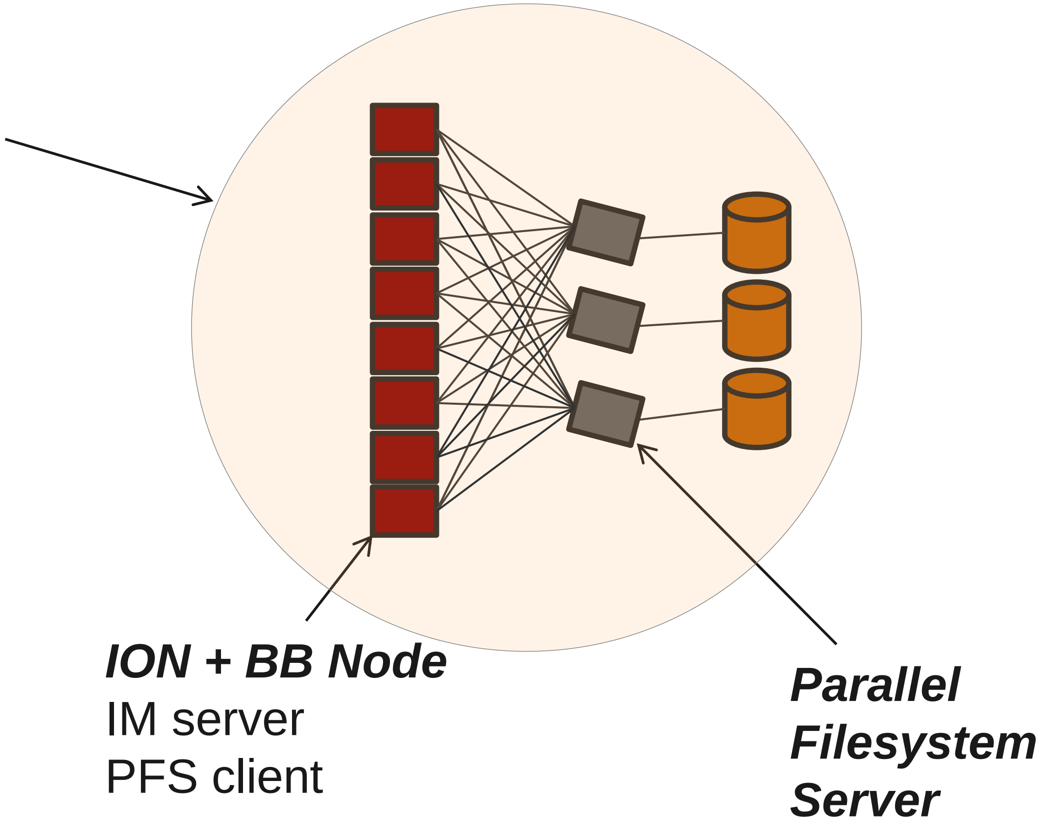
2. Metadata discovery is fast!



IME Interactions with PFS

PFS I/O Domain

1. Employs current parallel filesystem technology
2. Low risk – technologies have been in place for years
3. Performance problems arise when I/O patterns do not match PFS
4. Suitable for high GB/s to low TB/s throughput
5. High-capacity storage – Performance as an mid / end-tier archival system is sufficient



Questions?