DDN Update Oct, 2016

JLUG2016

DataDirect Networks Japan, Inc

Nobu Hashizume

Oct, 2016

DDN Japan 2016

2016年はExaFlops Systemに向けてDDNが提唱するStorage SystemのFirst Versionを日本で提供できた年になりました

DDN Japan 2015

ちなみに去年のJLUGではIMEのコンセプトのお話しとSFA14Kのご紹介を話していました

DDN® STORAGE

#1 IN HPC LEADERSHIP & INNOVATION

NVMe. SSD. FILE SYSTEM. ARCHIVE. CLOUD.

DDN END-TO-END DATA LIFECYCLE MANAGEMENT

IME

FAST DATA & COMPUTE

- SPEED UP YOUR APPS 1000 TIMES!
- ACCELERATE YOUR LUSTRE & GPFS
- WORLD'S FASTEST BURST BUFFER

SFA14K™

PERSISTENT DATA & FILE SYSTEM

- HYBRID STORAGE
- EDR IB, 100GB ETHERNET
- MIX SSD AND DISK
- EMBED APPS AND F/S
- 60GB/S & 6M IOPS
- PCI-E FABRIC SPEED

WOS

LIVE ARCHIVE & COLLABORATION

- FASTEST OBJ. STORAGE
- HIGHEST SECURITY
- LOWEST COST ARCHIVE
- MULTI-SITE
- INTELLIGENT TIERING
- GLOBAL COLLABORATION

CLOUD

- HYBRID PUBLIC CLOUD
- GPFS™ AND LUSTRE® BRIDGE
- OPEN SOURCE ACCESS

DDN STORAGE

ddn.com

DDN | EXAScaler & Lustre Case Studies
JCAHPC System

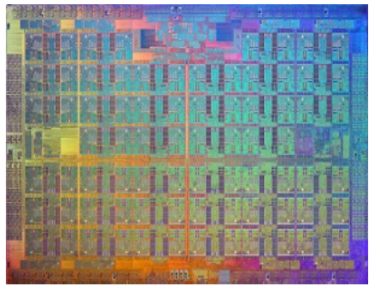
DDN STORAGE | 東京大学 THE UNIVERSITY OF TOKYO | JCAHPC | 筑波大学 University of Tsukuba | intel

- **University of Tokyo & University of Tsukuba**
- **25 PF System with 8208 KNL Nodes provided by Fujitsu**
- **I/O System by DDN**
  - **Intel Omnipath**
  - 26 PB ExaScaler/Lustre @ 400 GB/sec
  - 1 PB of IME Burst Buffer with NVMe @ 1400 GB/sec

May 10, 2016

Japan Unveils Details of 25 PFLOPS Machine to be Operational in December 2016

John Russell



Knights Landing Die Photo

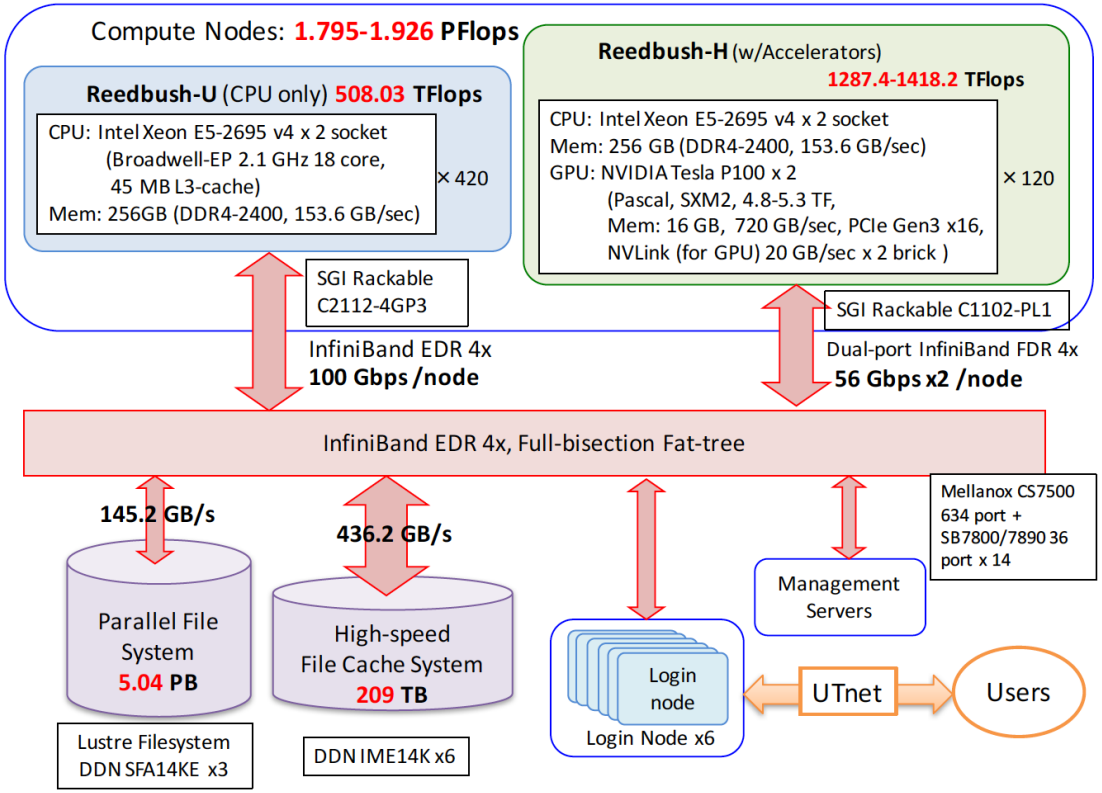
It's a good day to be Intel, Data Direct Networks (DDN), and Fujitsu. The Joint Center for Advanced High Performance Computing (JCAHPC) in Japan today released the details of its next generation supercomputer – Oakforest-PACs – which will deliver 25 PFLOPS, use Intel's Xeon Phi (Knights Landing) manycore processors and Omni-Path Architecture, be built by Fujitsu, and be operational in December 2016.

When fired up, the Oakforest-PACS will be the fastest supercomputer system in Japan for the moment. Twenty-five

DDN | EXAScaler & Lustre Case Studies
Reedbush Supercomputer System

東京大学情報基盤センタースーパーコンピューティング部門
Supercomputing Division, Information Technology Center
The University of Tokyo

- **University of Tokyo**
- **I/O System by DDN**
  - 5 PB ExaScaler/Lustre @ 145.2 GB/sec
  - 200TB of IME Burst Buffer with NVMe @ 436.2 GB/sec

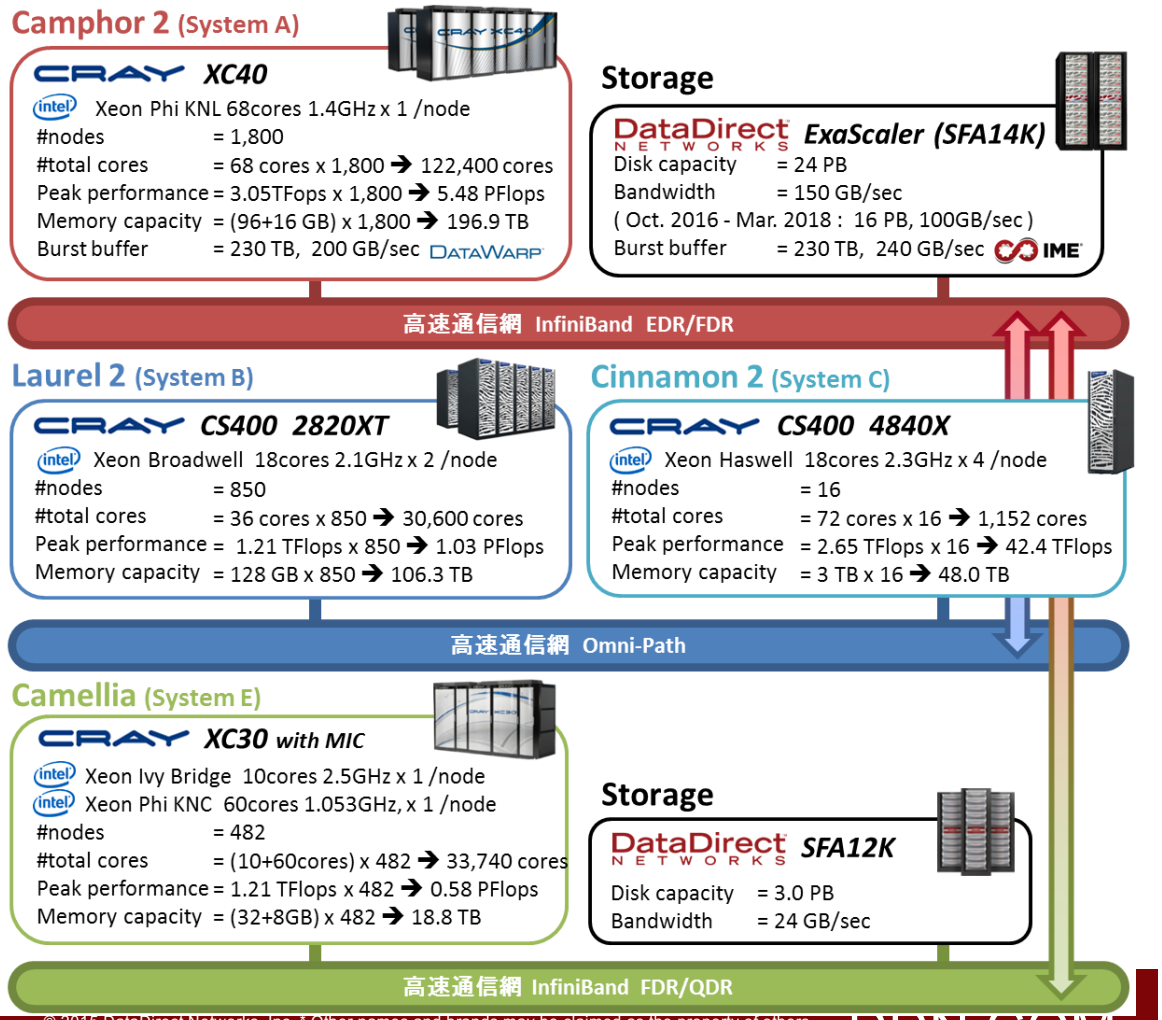


DDN | EXAScaler & Lustre Case Studies
Kyoto University Supercomputer System

DDN® STORAGE

京都大学学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University

- **Kyoto University**
- **I/O System by DDN**
  - 24PB ExaScaler/Lustre @ 150 GB/sec
  - 230TB of IME Burst Buffer with NVMe @ 240 GB/sec



IMEシステム@Japan

システム	IME	ExaScaler(Lustre)
JCAHPC Oakforest-PACS	物理960TB 25 x IME14K-OPA	物理32PB 10 x SFA14KXE-OPA
東大 Reedbush	物理230TB 6 x IME14K-EDR	物理6.25PB 3 x SFA14KE-EDR
京大 ACCMS2	物理230TB 6 x IME14K-OPA	物理24PB(P1: 16PB, P2: 8PB) 3 x SFA14K-EDR (P1) 2 x SFA14KXE-EDR (P2)
合計	**物理1.42PB**	**物理54.25PB (62.25PB)**

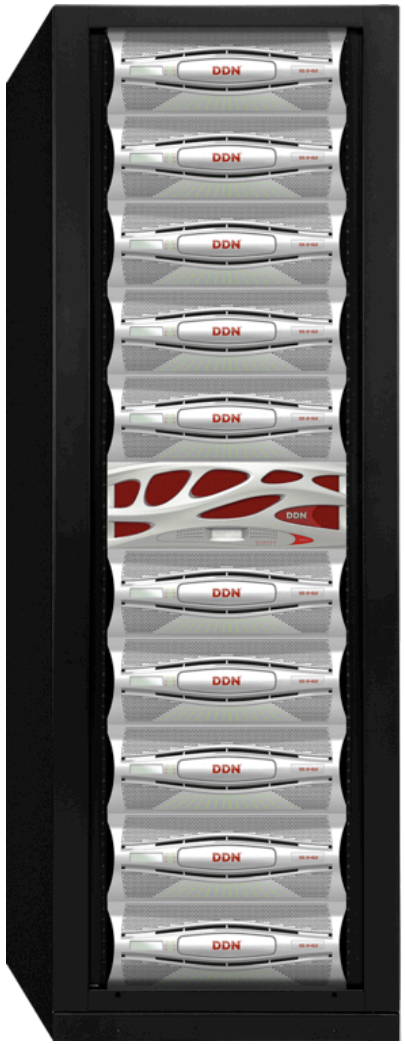
DDN SFA14K™

WORLD'S FASTEST HYBRID SSD & DISK, EMBEDDED & HYPER-CONVERGED

- Delivering 60GB/s and 6 Million IOPS in 4u
- NVMe and SSD Boosts I/O and Shrinks Latency
- Speeds Up SFA®, IME®, Lustre®, GPFS™
- Most Advanced Embedded PCI-e Fabric
- EDR IB, 100Gb/s Ethernet, 12Gb/s SAS, Omni-Path

THE FUTURE OF HPC STORAGE IS NOW SHIPPING!

DDN | SFA14K
Design Criteria



Overcome Intel 20GB/s Bottleneck

- No Effective Bandwidth Increases since Sandy Bridge
- PCIe-3 maximum speed has not increased

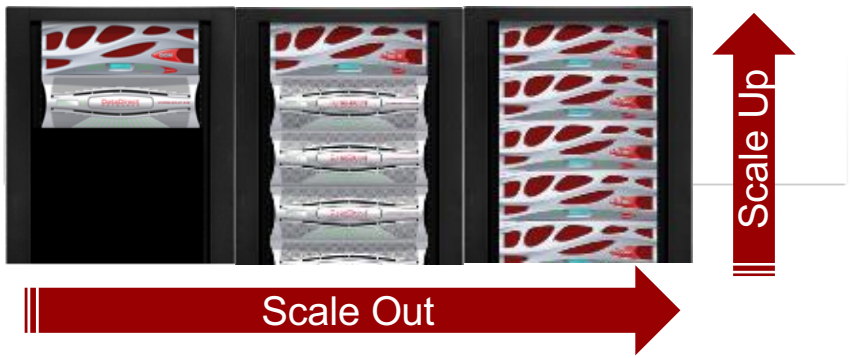
Accommodate New Interconnect Technologies

- Mellanox EDR, Intel Omni-Path, PCIe-3 x16, 100 Gb/s Ehternet, 32 Gb/s FC
- Adaptive Motherboard Design to Encompass Multiple Configurations

Integrate Cutting Edge Solid State Technologies

- 12Gb/S SAS Dual Port, 25W PCIe, NVMe Dual Port
- Most efficient IOPs, $/capacity, $/performance

Flexible Configurations for Multiple Use Cases



SFA14K

Most Versatile Next-Generation Storage Platform

SFA Powered By SSD & Disk

Maximize Bandwidth and IOPS Efficiency

- Data analytics acceleration
- Large and small file performance
- Up to 60GB/s bandwidth
- Support for controller clustering

Embedded Solutions

Parallel File System Appliance: [x]Scaler

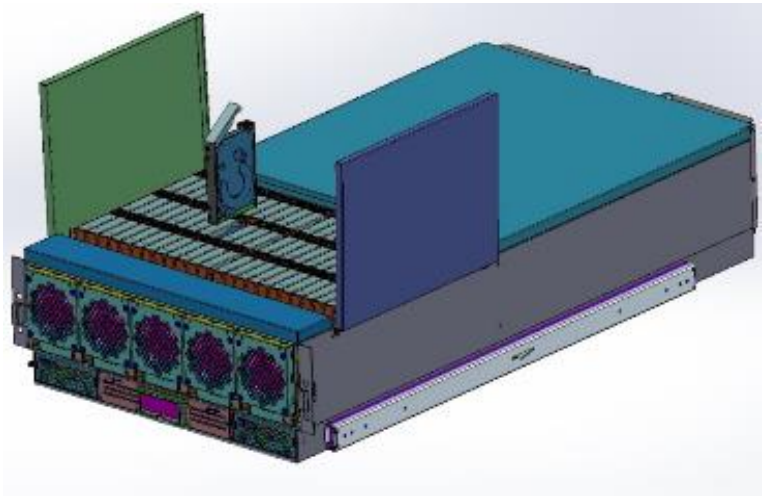
- Run in the controller head
- GPFS or Lustre
- Scale IOPS, throughput, capacity
- Up to 38GB/s bandwidth via file system

IME Application Acceleration

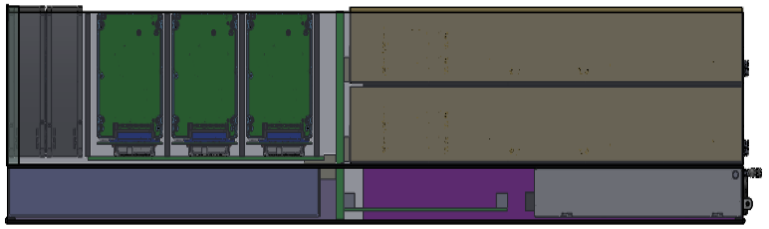
Embedded IME Technology

- Lower latency than flash arrays
- Application acceleration
- IO provisioning
- 50-60GB/s of buffer burst speed

- All in one integrated design with expansion capability
- Dual Redundant Controllers Intel Haswell / Broadwell Processor
- 72 Drive High Density 2.5” Enclosure
- SSDs become an Integral part of the platform
- 48 2.5” dual ported PCIe/SAS drives and 24 2.5” SAS 12 Gb/S dual ported drives
- Optimized Building Block for BW or IOPs
- Direct Connection for 10 – 8412 12Gb/S 84 drive Enclosures



Front/Side Isometric View



Right Side View

DDN | SFA14K

Next Generation Hybrid Storage Controller

Design Specs

Appliance	4RU Active/Active Dual-Controller - 48x PCIe NVMe/SAS 2.5” SSD Slots - 24x 12Gb/s SAS 2.5” SSD/HDD Slots - Power Fail-Safe
Internal Fabric External Expansion	Switched 600+ lanes PCIe Gen3 24 Quad 12Gb/s SAS
CPU – Memory Controller	Intel® Xeon® Processor E5 v3 (Haswell) Up to 2048GB DDR4 2133 Memory
Platform Performance	6M IOPS /4U 60M IOPS /Rack 60 GB/s /4U 600 GB/s /Rack
Host & Network Connectivity	32Gb/s FC EDR IB 100Gig E Intel OmniPath
Storage Scalability External Clustering	72 Internal 2.5” & up to 20 direct attached 84 slot 2.5”/3.5” expansion enclosures Controller Grid Fabric Interconnect ready
F/W and Software Options	SFAOS w SFX, Grid- & ExaScaler Embedded, IME, Open Platform

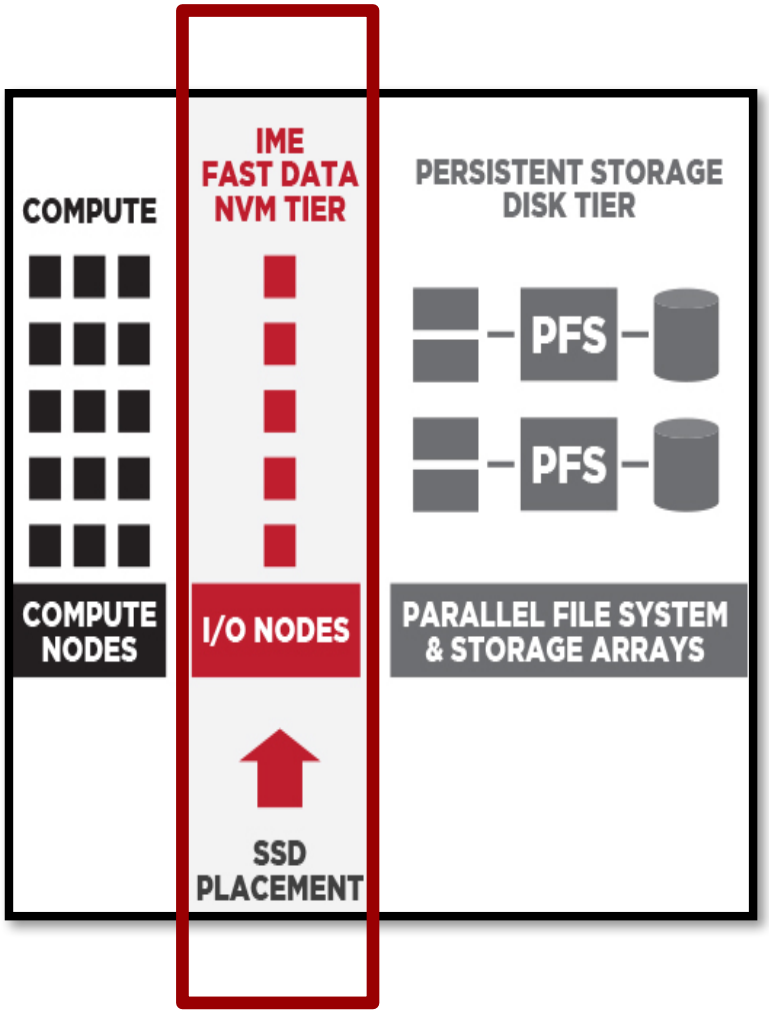
DDN IME®

WORLD'S MOST ADVANCED APPLICATION AWARE I/O ACCELERATION SOFTWARE

- Accelerate Your Applications Up to 1000 Times!
- Eliminate The Randomness of Your Workflows
- Boost Your Lustre® and GPFS™ File System Speed
- The Fastest and Most Reliable Burst Buffer

What is IME?

IME creates a new application-aware fast data tier that resides right between compute and the parallel file system to accelerate I/O, reduce latency and provide greater operational and economic efficiency



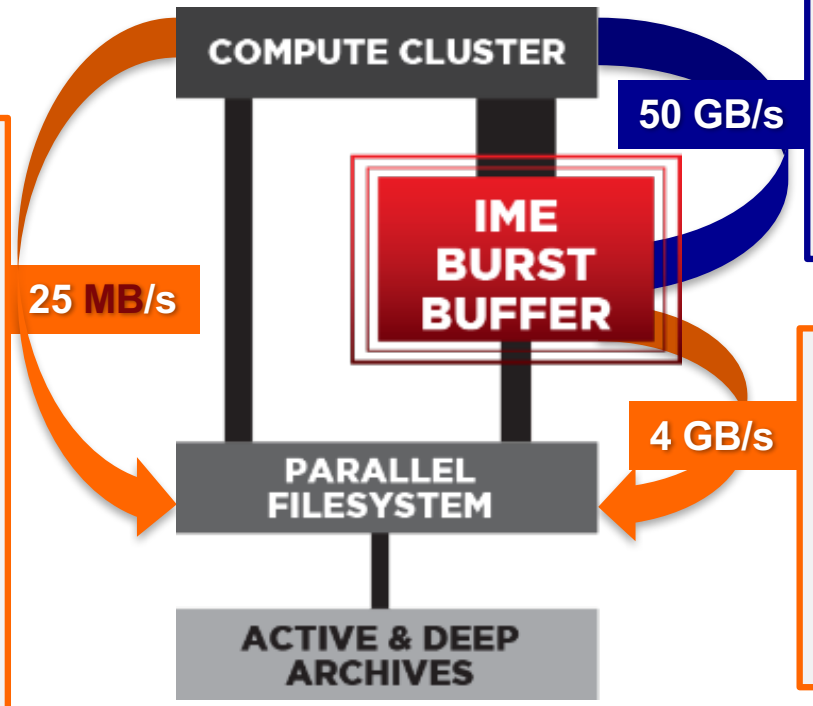
Where IME Provides Value

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.

IOR benchmarks indicate a **3x – 20x speedup on I/Os <32KB**.



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

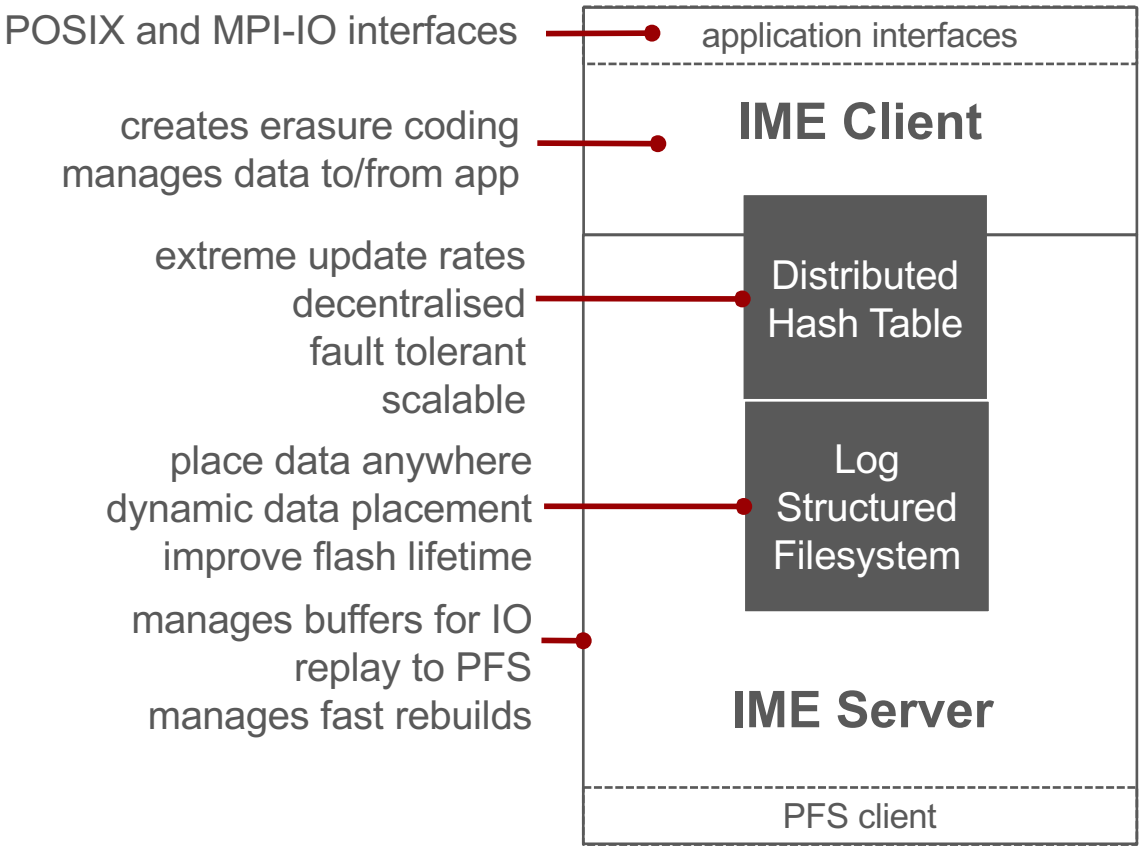
At ISC14, we **demonstrated three orders of magnitude speed-up** due to this high performance tier

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

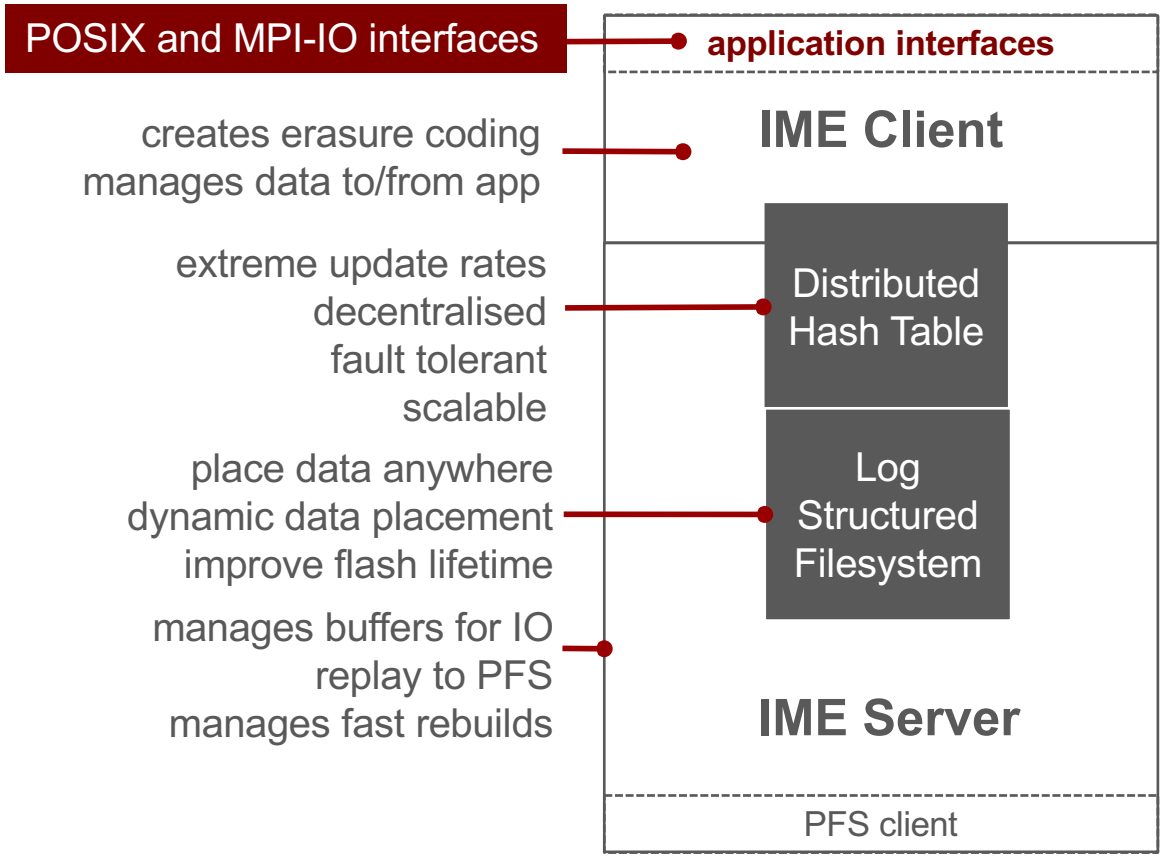
At ISC14, we demonstrated **two orders of magnitude speed-up** due to this efficiency

S3D Turbulent Flow Model

IME Architecture

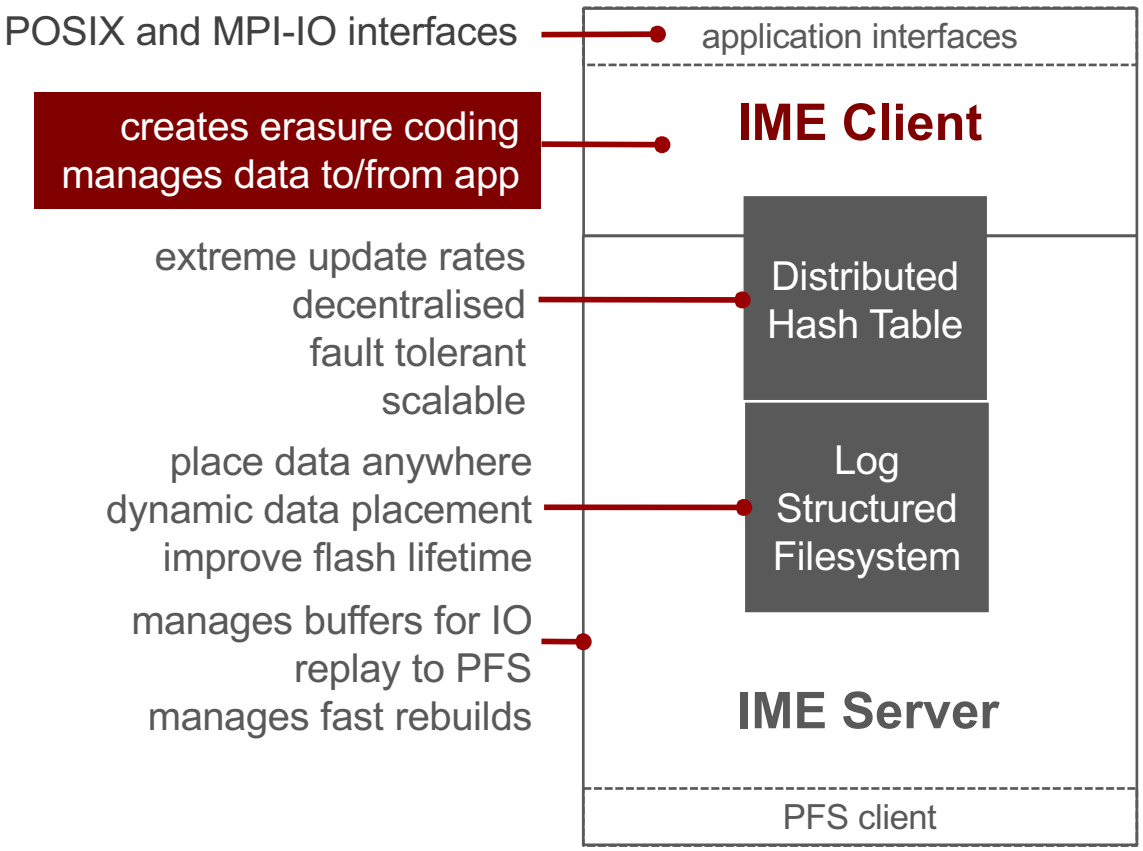


IME Architecture



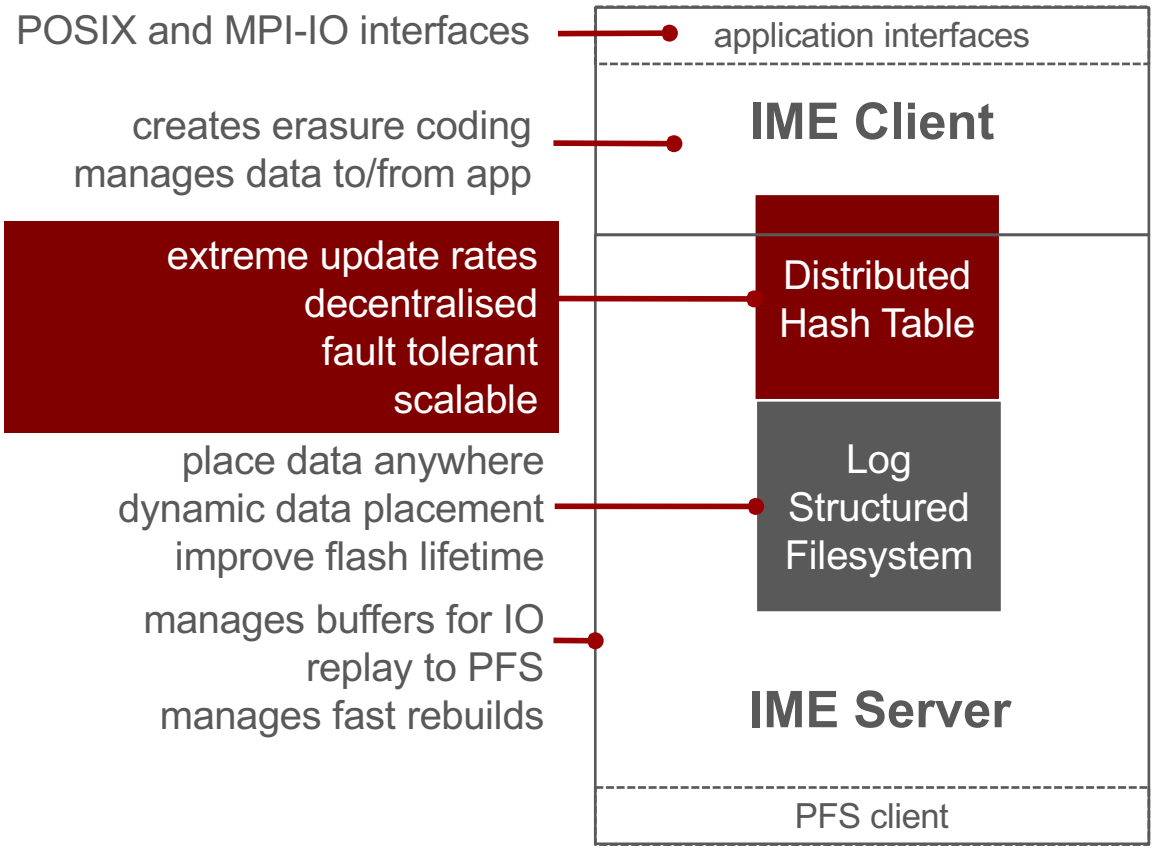
- **No application changes required**
- **Applications must be recomplied/relinked to utilise IME**

IME Architecture



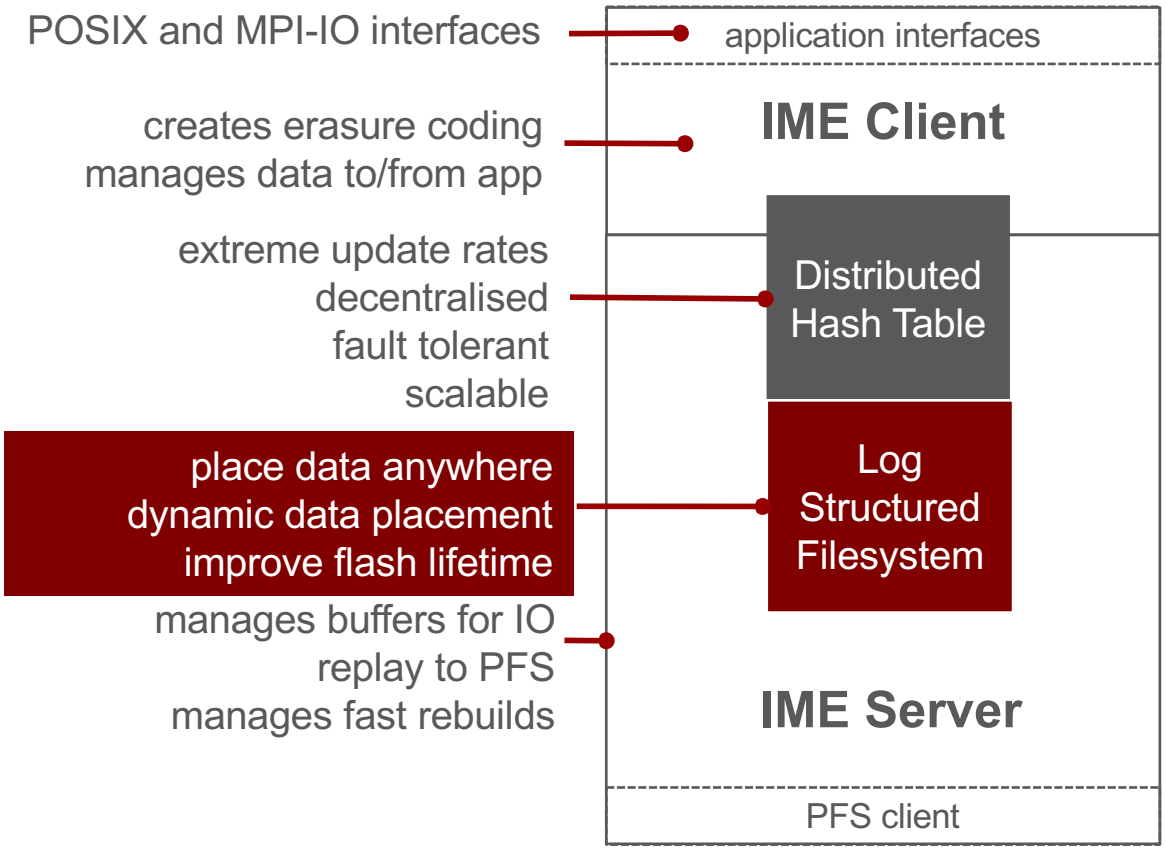
- **Client buffers IO fragments and sends to IME servers**
- **also (optionally) creates parity buffers according to RAID scheme chosen**
- **sends application and data to IME servers such that loss of a server or SSD does not result in data loss**
- **Heuristics in IME clients can intelligently pre-fetch IME-resident data to applications**

IME Architecture



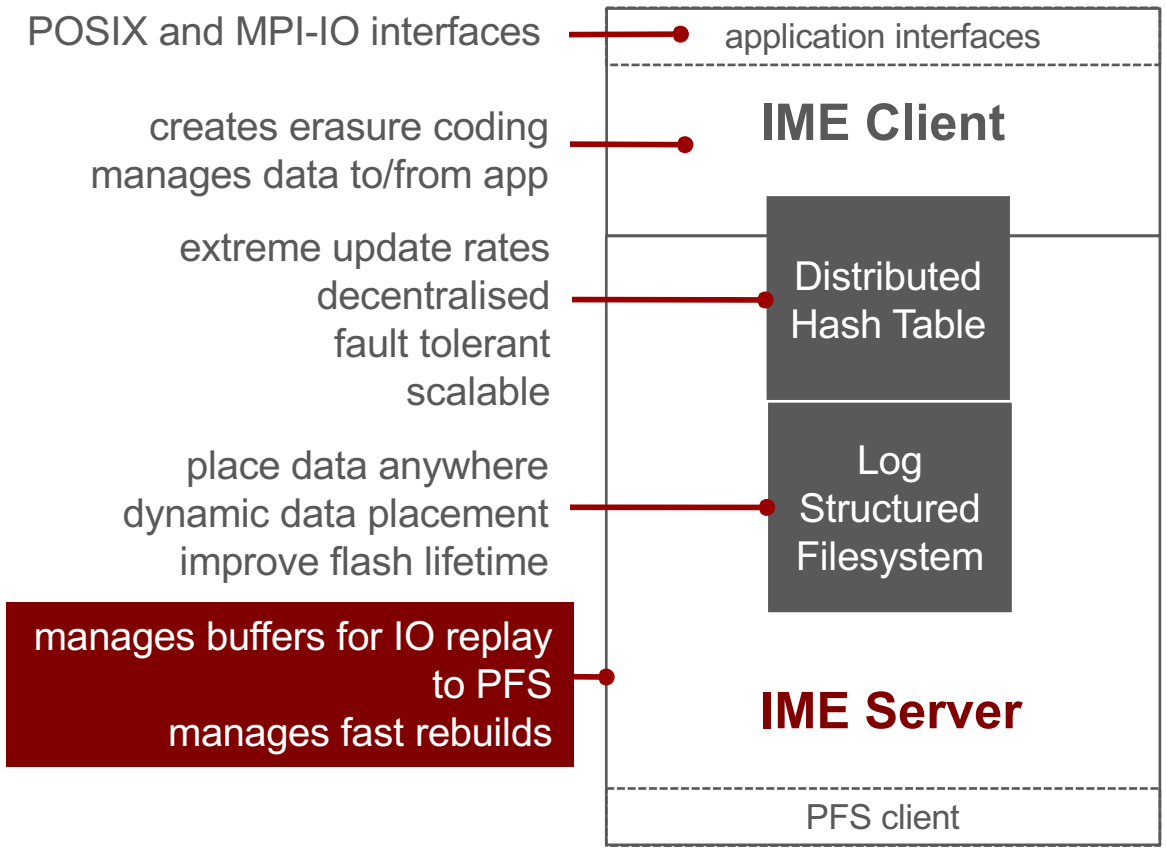
- **DHT is at the core of IME**
- **distributed index manages locations of files and objects**
- **Unique routing and data placement properties differentiate IME DHT from other DHTs**
- **Fast O(1) routing algorithm built on high-performance, non-cryptographic hash algorithms**
- **Load is uniformly distributed across DHT nodes**

IME Architecture



- **Data can be placed anywhere within IME – any server, any SSD**
- **IOPs to SSDs are minimised and optimised for NAND Flash**

IME Architecture



- **Actively reorganize I/Os to coalesce small block I/Os into large chunks and perform full stripe alignment before submission to the back end, Read acceleration is almost the reverse process**
- **Protect data by disseminating erasure coded chunks across the IME appliance cluster**
- **Pre-fetch data into IME from PFS using out-of-band commands and APIs**

IME Dataflow

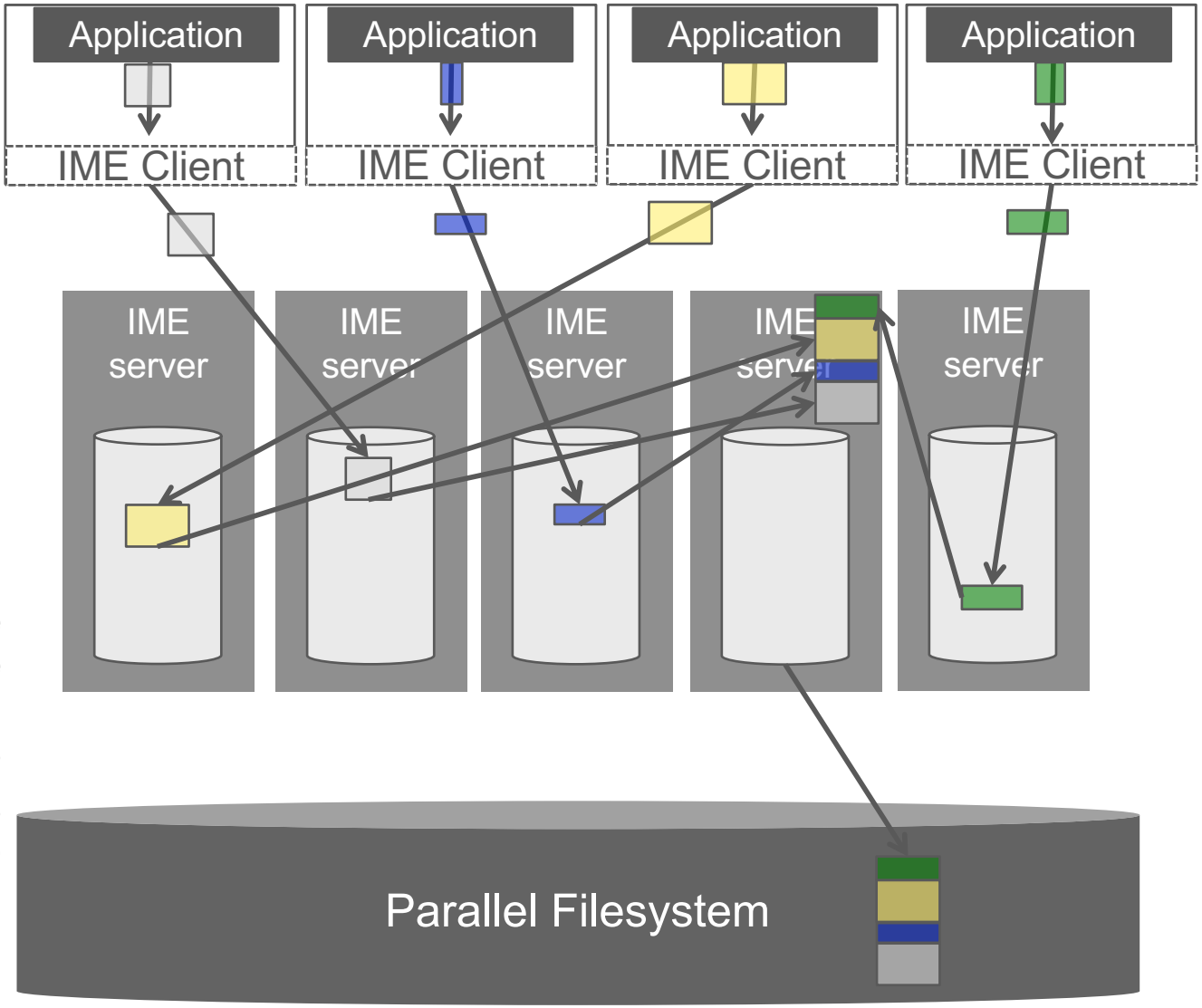
1. Application issues fragmented, misaligned IO

2. IME clients send **fragments** to IME servers

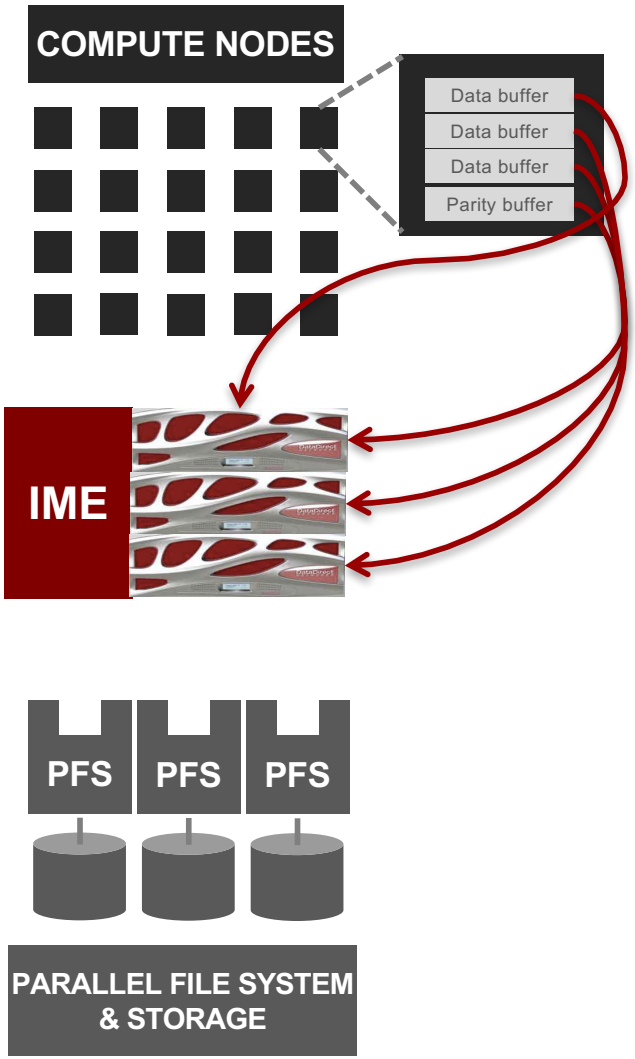
3. Fragments to IME servers accessible via DHT to all clients

4. Fragments to be flushed from IME are assembled into PFS stripes

5. PFS receives complete PFS stripe



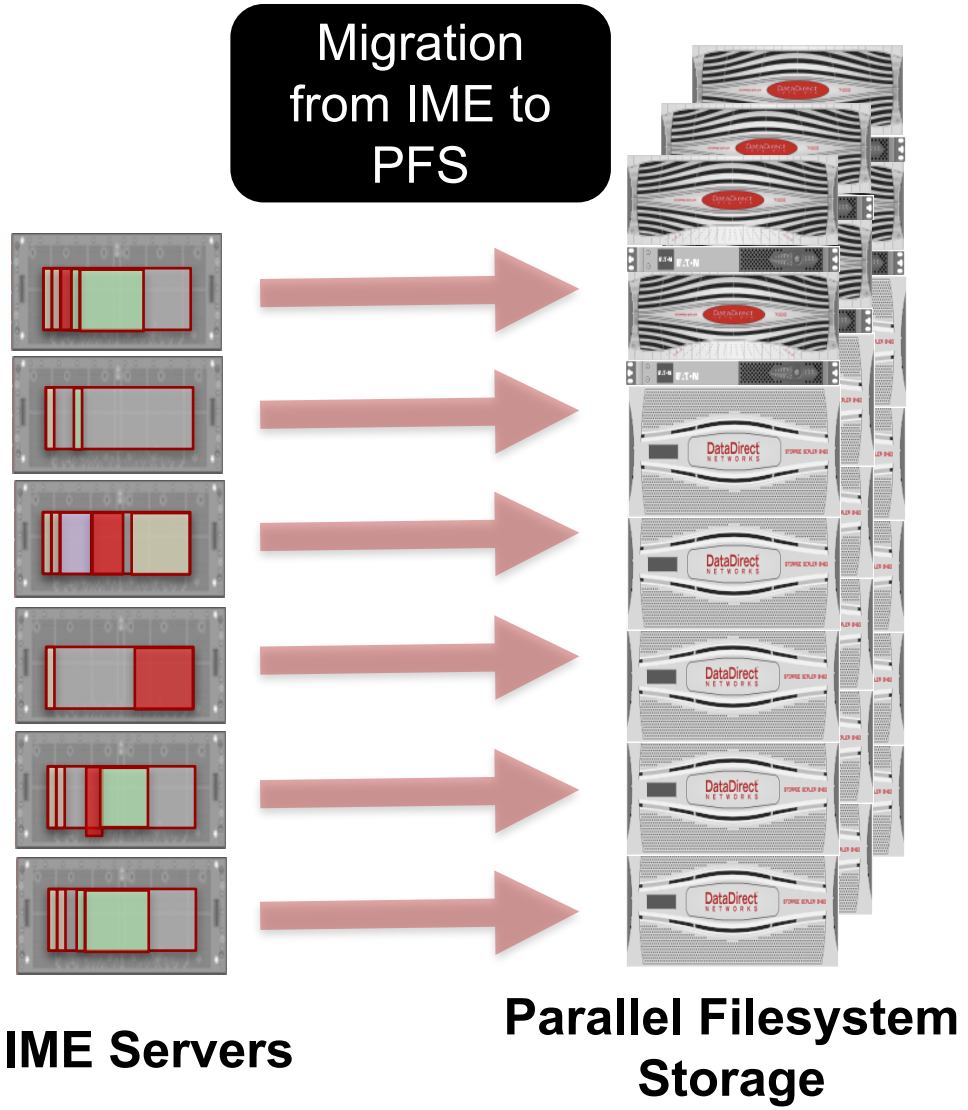
IME Erasure Coding



- **Data protection against IME server or SSD Failure is optional**
  - (the lost data is "just cache")
- **Erasure Coding calculated at the Client**
  - Great scaling with extremely high client count
  - Servers don't get clogged up
- **Erasure coding does reduce useable Client bandwidth and useable IME capacity:**
  - 3+1: 56Gb → 42Gb
  - 5+1: 56Gb → 47Gb
  - 7+1: 56Gb → 49Gb
  - 8+1: 56Gb → 50Gb

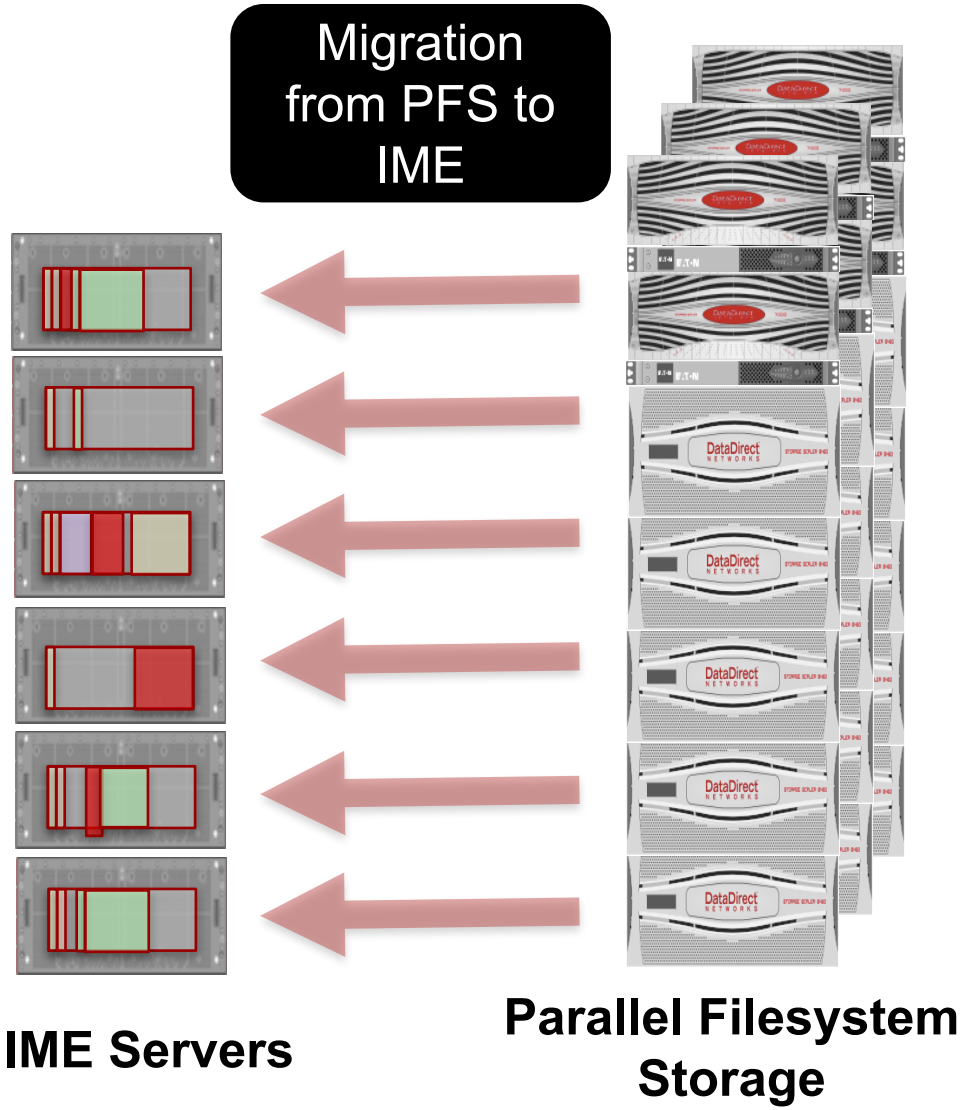
Data Migration Policy – Egress (to PFS)

- **Two modes:**
  - Fully dynamic
    - Data are flushed to PFS as they arrive in IME
    - .. Essentially an LRU
  - Explicit
    - Files are not migrated to the PFS by default
    - Command is required from user or scheduler
      - Shell command interface
    - These files *may* be flushed if IME has reached capacity



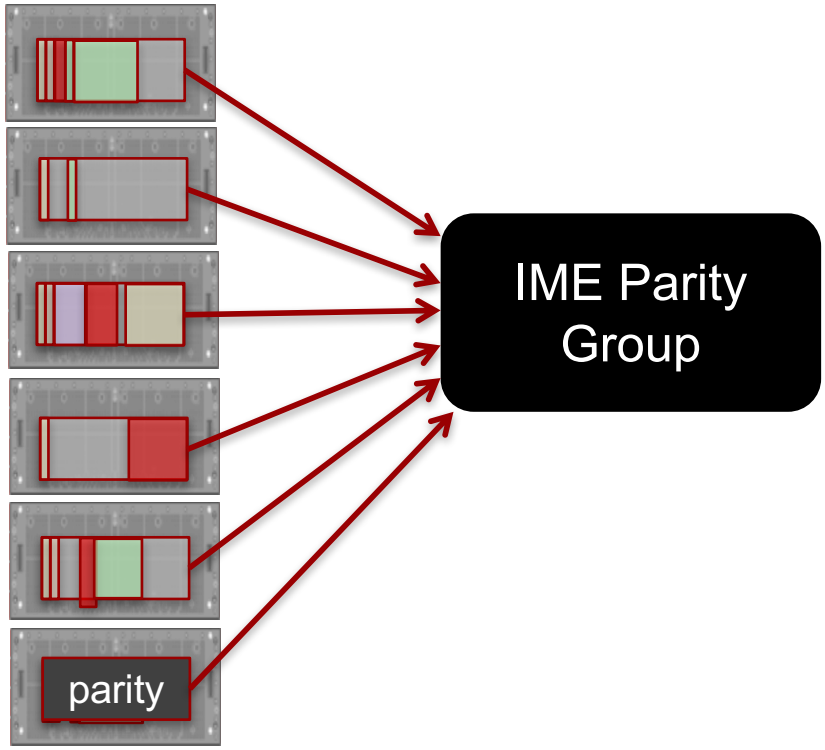
Data Migration Policy – Ingress (from PFS)

- **Fully parallel**
  - All IME servers participate in ingress procedures
- **Explicit only in IME 1.0**
  - Shell command interface
  - Issued by administrator or scheduler
- **Reads of non-cached data are transparently 'passed through to PFS**
  - Data is *not* stored on IME SSD



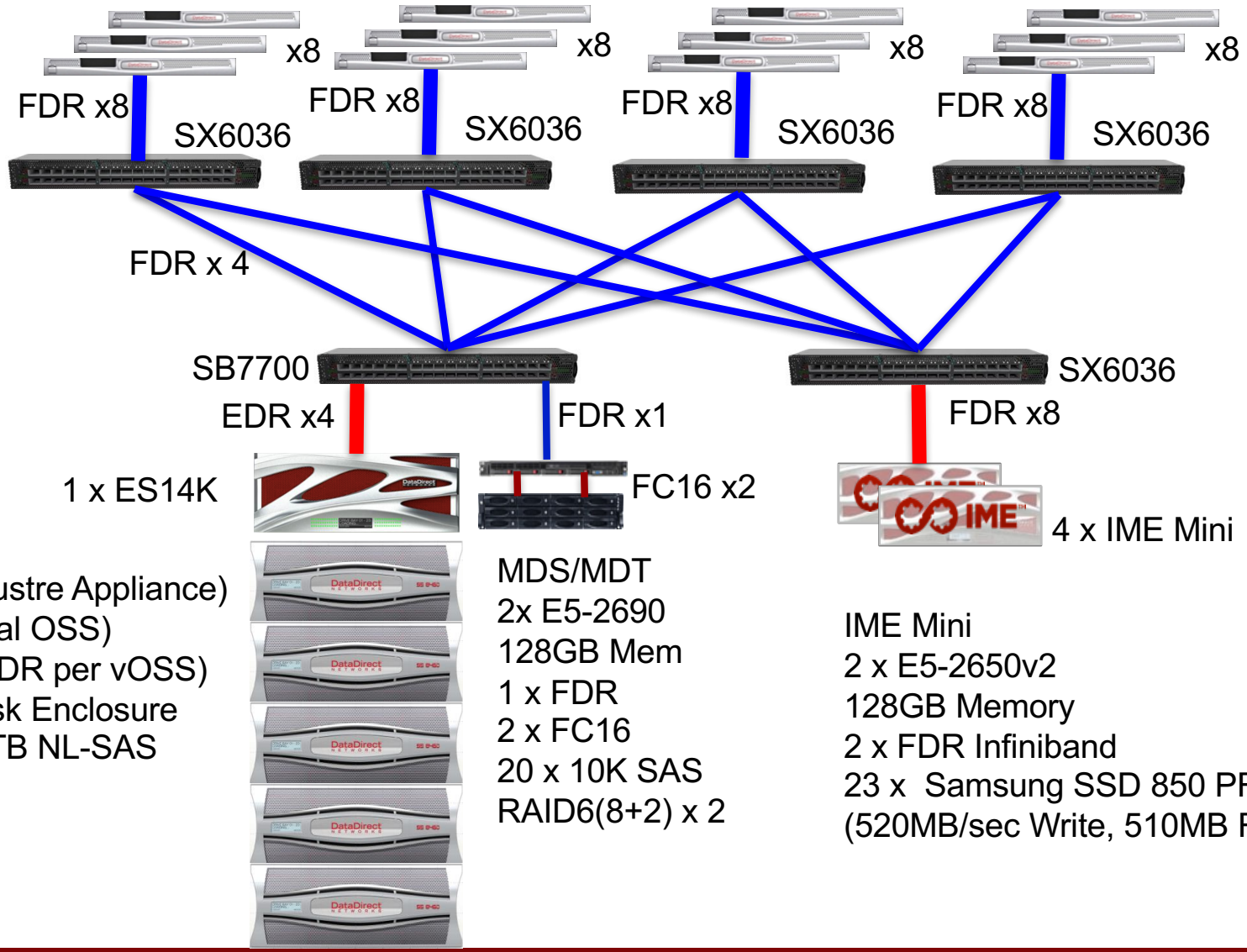
IME Fault Tolerance

- **Across IME servers**
  - *no* RAID within the IME server
- **Flexible based on configuration and workload**
  - N+M schemes
  - Replication
    - Small files
  - May be disabled
    - Per-file basis
- **Very High Performance**
  - Parity group may encapsulate non-contiguous fragments
  - fully non-blocking
  - No ‘read-modify-writes’ or network locking required



Benchmark Results

ベンチマーク環境



ES14K(DDN Lustre Appliance)
4 x vOSS(Virtual OSS)
4 x EDR (1 x EDR per vOSS)
5 x SS8462 Disk Enclosure
400 x HGST 6TB NL-SAS

MDS/MDT
2x E5-2690
128GB Mem
1 x FDR
2 x FC16
20 x 10K SAS
RAID6(8+2) x 2

IME Mini
2 x E5-2650v2
128GB Memory
2 x FDR Infiniband
23 x Samsung SSD 850 PRO
(520MB/sec Write, 510MB Read)

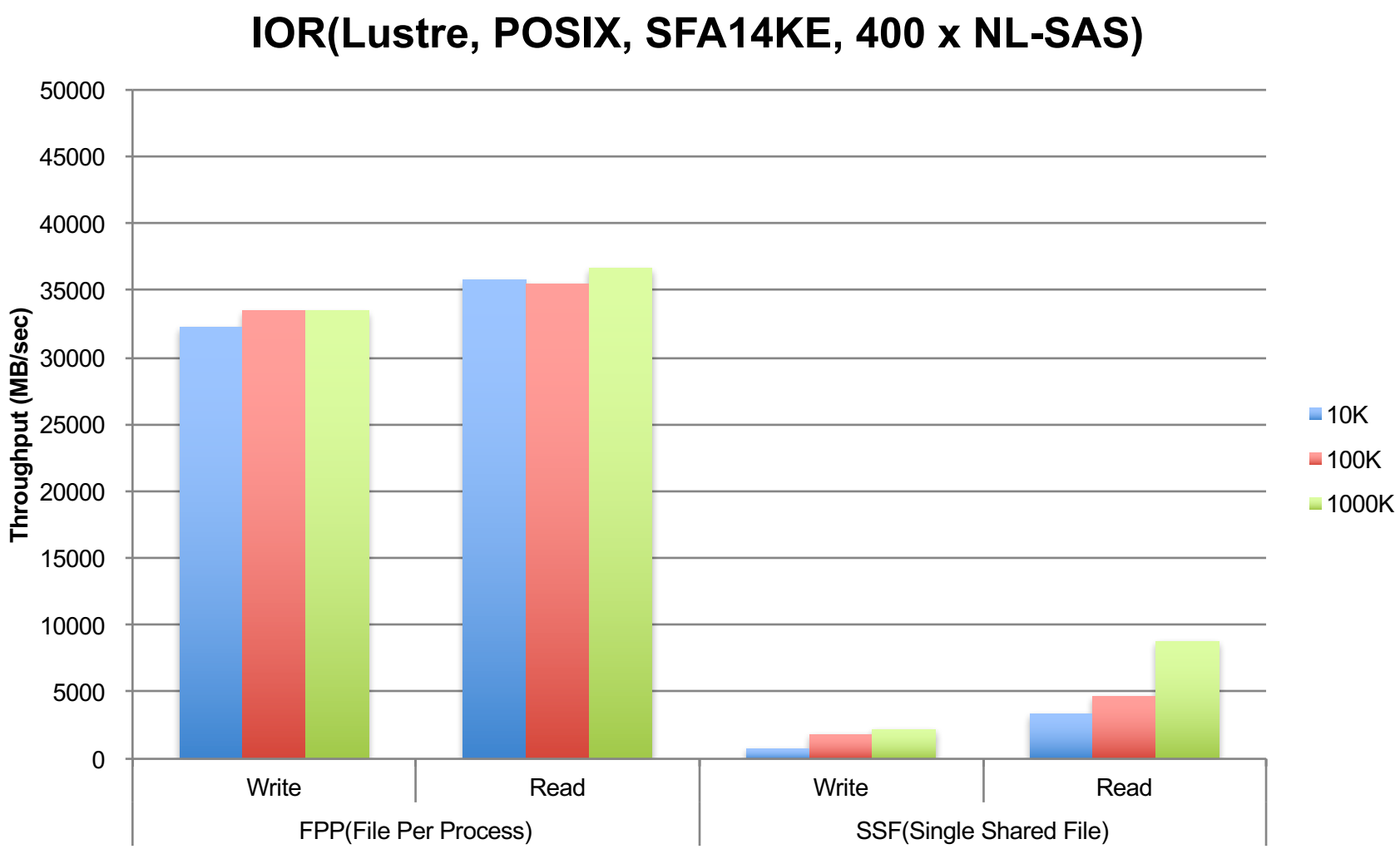
ハードウェア、ソフトウェア構成

- **Hardware**
  - Lustre OSS/OST
    - ES14K (SFA14KE/5 x SS8462, 400 x 6TB NL-SAS)
  - Mellanox EDR/FDR Switch (2:1 Blocking Configuration)
  - 1 x Lustre MDS/MDT
    - 1 x SuperMicro Server(2 x E5-2690, 128GB Memory, 1 x FDR)
    - 1 x EF4024(FC16, 10Krpm 600GB SAS x 20, RAID6(8+2) x 2)
  - 32 x Lustre Client
    - 16 x Dell R630 (2 x E5-2650 v2, 64GB Memory, 1x FDR)
    - 16 x Dell R620 (2 x E5-2640 v3, 64GB Memory, 1 x FDR)
- **Software Stack**
  - ES3.0 Beta(Lustre Server)
  - CentOS6.7 (Client)
  - Mellanox OFED-3.2-2.0.0
  - Lustre-2.5.41.ddn11

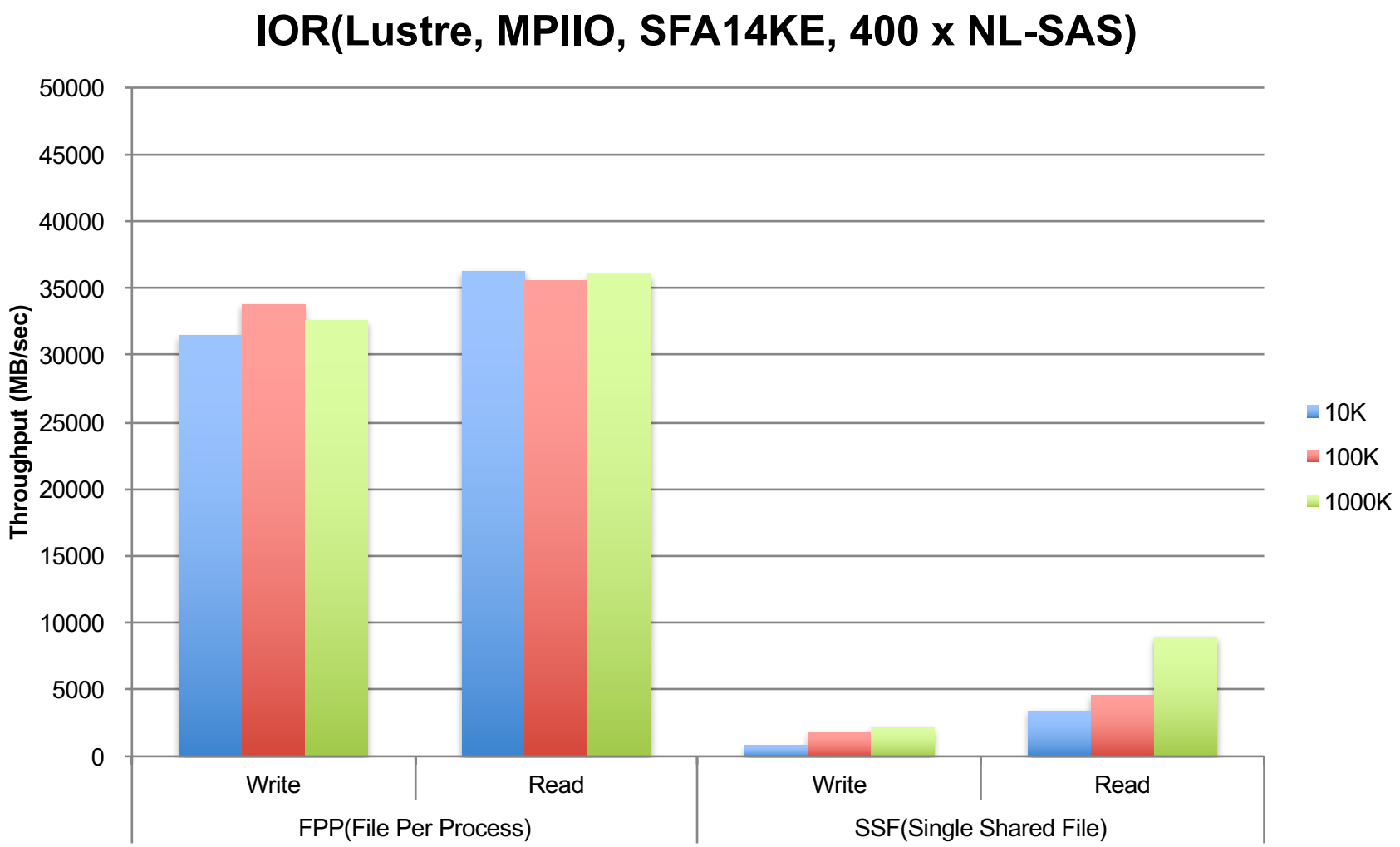
ベンチマーク概要

- **NeRSCの調達ベンチマークルールを使用**
  - http://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks
- **IOR**
  - POSIX I/O file per process(FPP)
  - POSIX I/O single shared file(SSF)
  - MPI/IO file per process(FPP)
  - MPI/IO single shared file(SSF)

IOR (POSIX, Lustre)
32 clients, 512 process, 3.3TB File Size

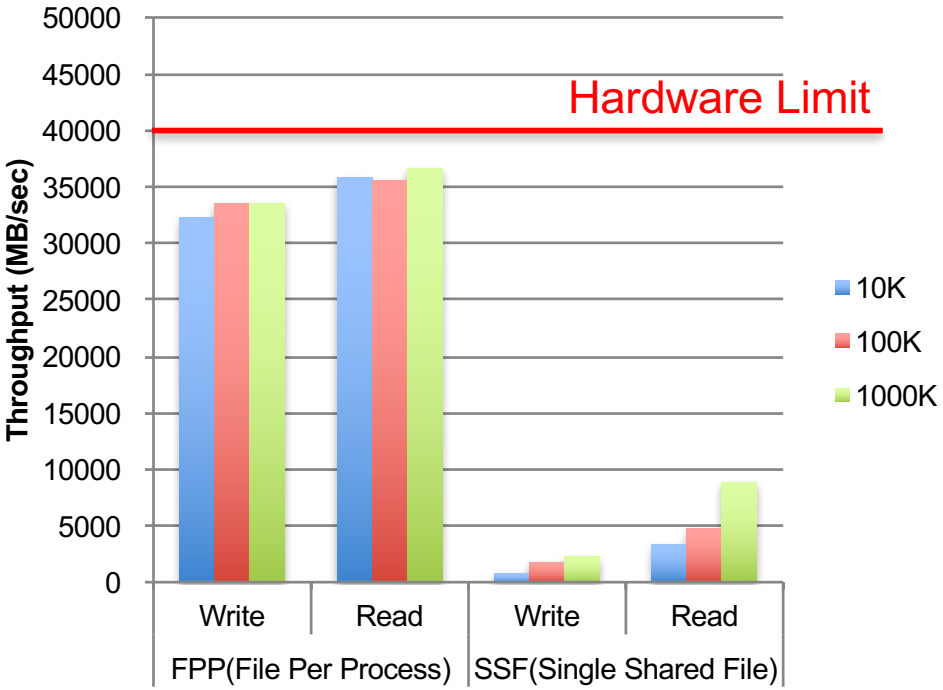


IOR (MPIIO, Lustre)
32 clients, 512 process, 3.3TB File Size



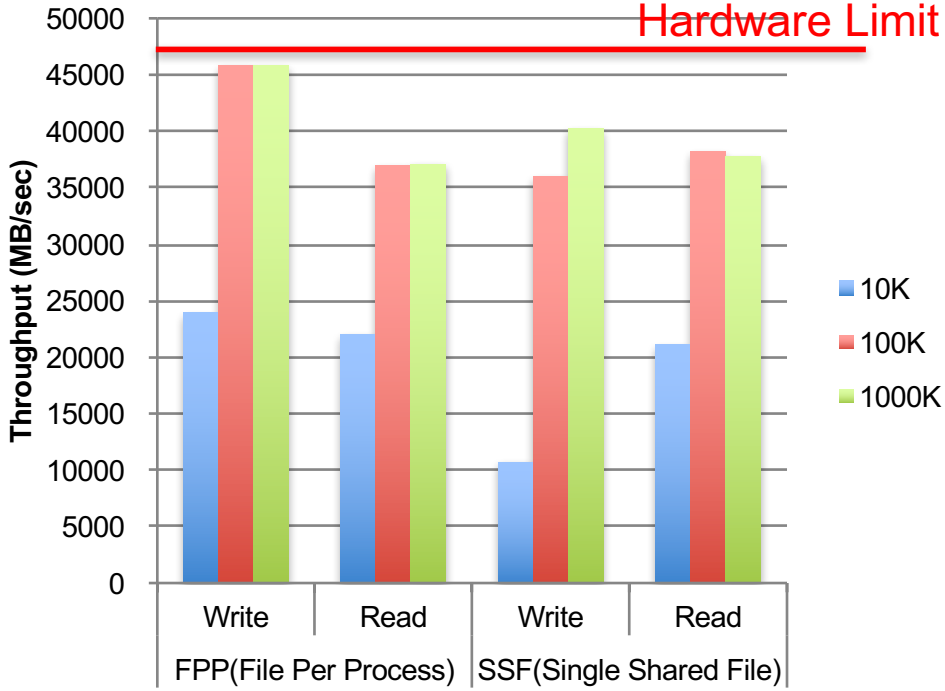
IOR (POSIX, Lustre and IME)
32 clients, 512 process, 3.3TB File Size

IOR(Lustre, POSIX)



FPP I/O Efficiency ~84%(Write) ~90%(Read)
SSP I/O Efficiency ~5%(Write) ~22%(Read)

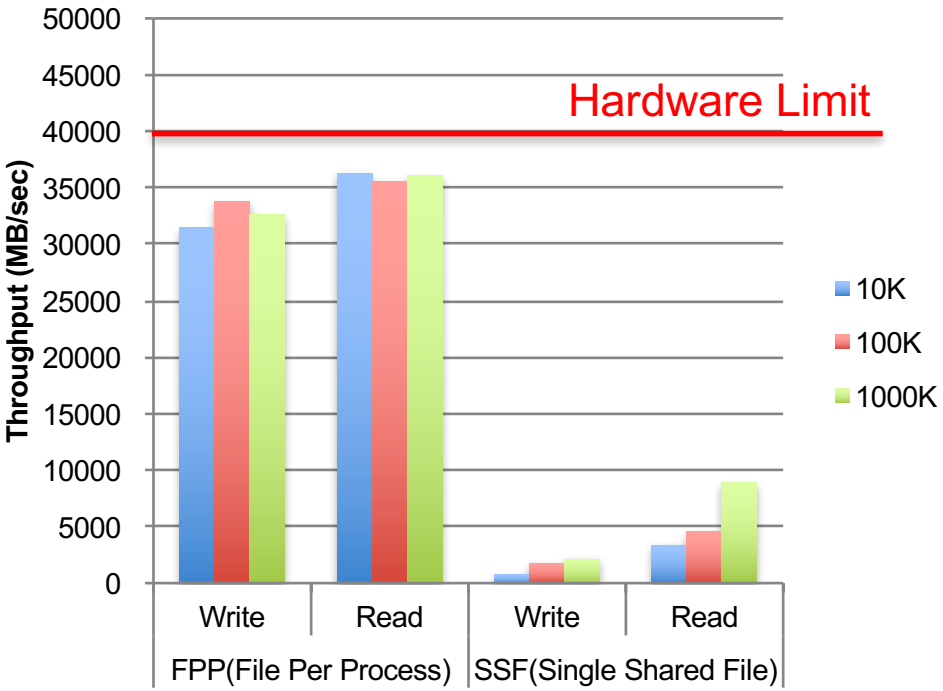
IOR(IME, POSIX)



FPP I/O Efficiency ~97%(Write) ~78%(Read)
SSP I/O Efficiency ~85%(Write) ~81%(Read)
Still under optimizations

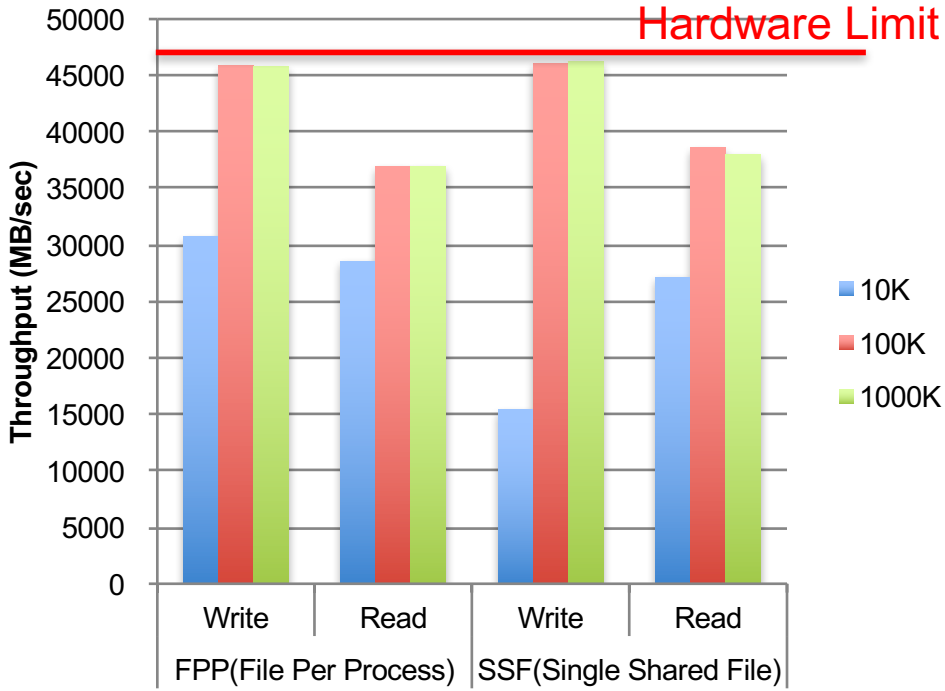
IOR (MPIIO, Lustre and IME)
32 clients, 512 process, 3.3TB File Size

IOR(Lustre, MPIIO)



FPP I/O Efficiency ~84%(Write) ~90%(Read)
SSP I/O Efficiency ~5%(Write) ~22%(Read)

IOR(IME, MPIIO)



FPP I/O Efficiency ~97%(Write) ~78%(Read)
SSP I/O Efficiency ~97%(Write) ~81%(Read)
Still under optimizations

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

DataDirect™ NETWORKS

ddn.com