DDN Update for 2017

PCCC

DataDirect Networks Japan, Inc

橋爪信明

Dec, 2016

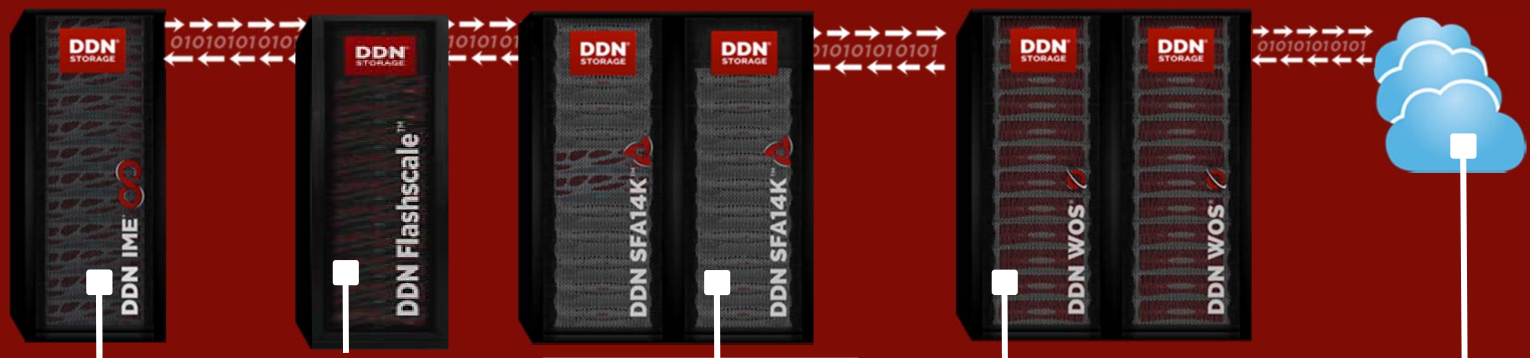
DDN® STORAGE

#1 IN HPC LEADERSHIP & INNOVATION
NVMe. SSD. FILE SYSTEM. ARCHIVE. CLOUD.

DDN END-TO-END DATA LIFECYCLE MANAGEMENT

IME® | FLASHScale™ | SFA14K™ | WOS®

BURST & COMPUTE | **Parallel File System HDD/SSD** | **LIVE ARCHIVE, SHARING & CLOUD**



- 1000X APP & FILE SYSTEM SPEED UP
- PREDICTIVE BURST BUFFER

- FASTEST, LOW LATENCY & BEST COST SSD
- 600GB/S & 60M IOPS, 5+PB/RACK

- MOST VERSATILE EMBEDDED FILE & BLOCK STORAGE
- EDR IB, 100 GbE, Omnipath, PCI-E

- FASTEST, SCALABLE OBJECT STORAGE
- GLOBAL MULTI-SITE SECURE TIERING

- OPEN SOURCE & HYBRID CLOUD
- BRIDGE S3, SWIFT, GPFS™ & LUSTRE®

Why IME?

cater better for real-world IO patterns

4

Why IME?

cater better for real-world IO patterns

All the time after that

4

DDN のFlash(IME)に対する取り組み

従来のHPC向け並列ファイルシステムでは解決出来ないIOワークロード（Big Data, AIなど）に向けてFlashを活用した新しいIOシステムを提案

- 2014 : Infinite Memory Engine(IME)発表
- 2015 : IMEを大規模スパコンに提案
- 2016 : IMEを国内3サイトに提供
- 2017 : 小さいIMEを展開する予定

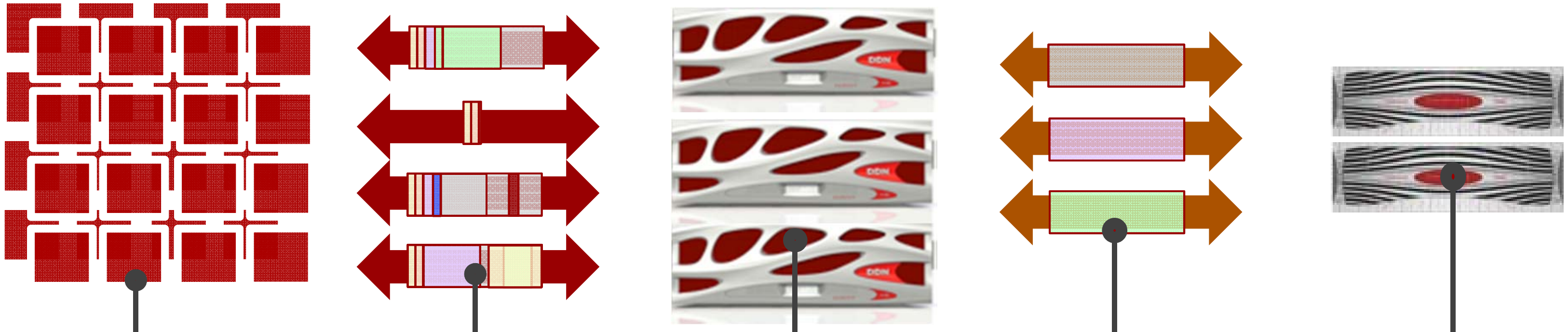
IME®

6

DDN | IME

Application I/O Workflow

Compute	IME	SFA
Diverse, high concurrency applications	Fast Data NVM & SSD	Persistent Data (Disk)



- Lightweight IME client intercepts application I/O. Places fragments into buffers + parity
- IME client sends fragments to IME servers
- IME servers write buffers to NVM and manage internal metadata
- IME servers write aligned sequential I/O to SFA backend
- Parallel File system operates at maximum efficiency

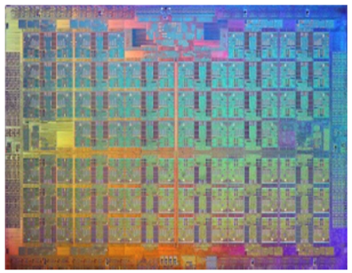
DDN | EXAScaler & Lustre Case Studies
JCAHPC System

- **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.

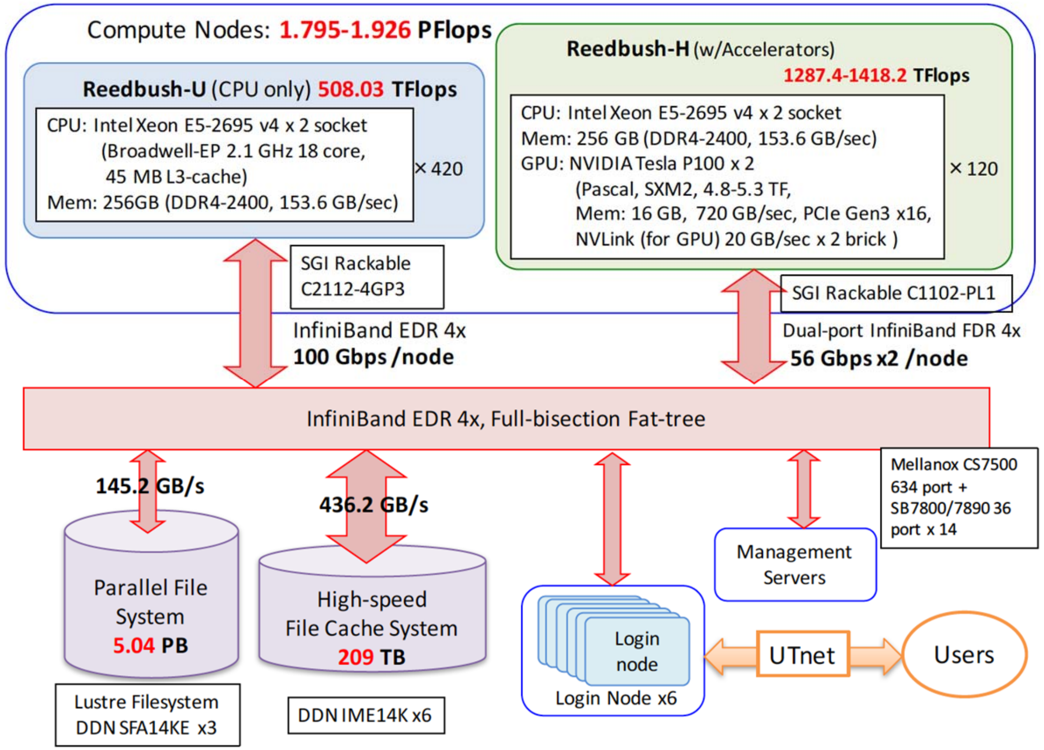
When fired up, the Oakforest-PACS will be the fastest supercomputer system in Japan for the moment. Twenty-five

DDN | EXAScaler & Lustre Case Studies
Reedbush Supercomputer System

DDN® STORAGE

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

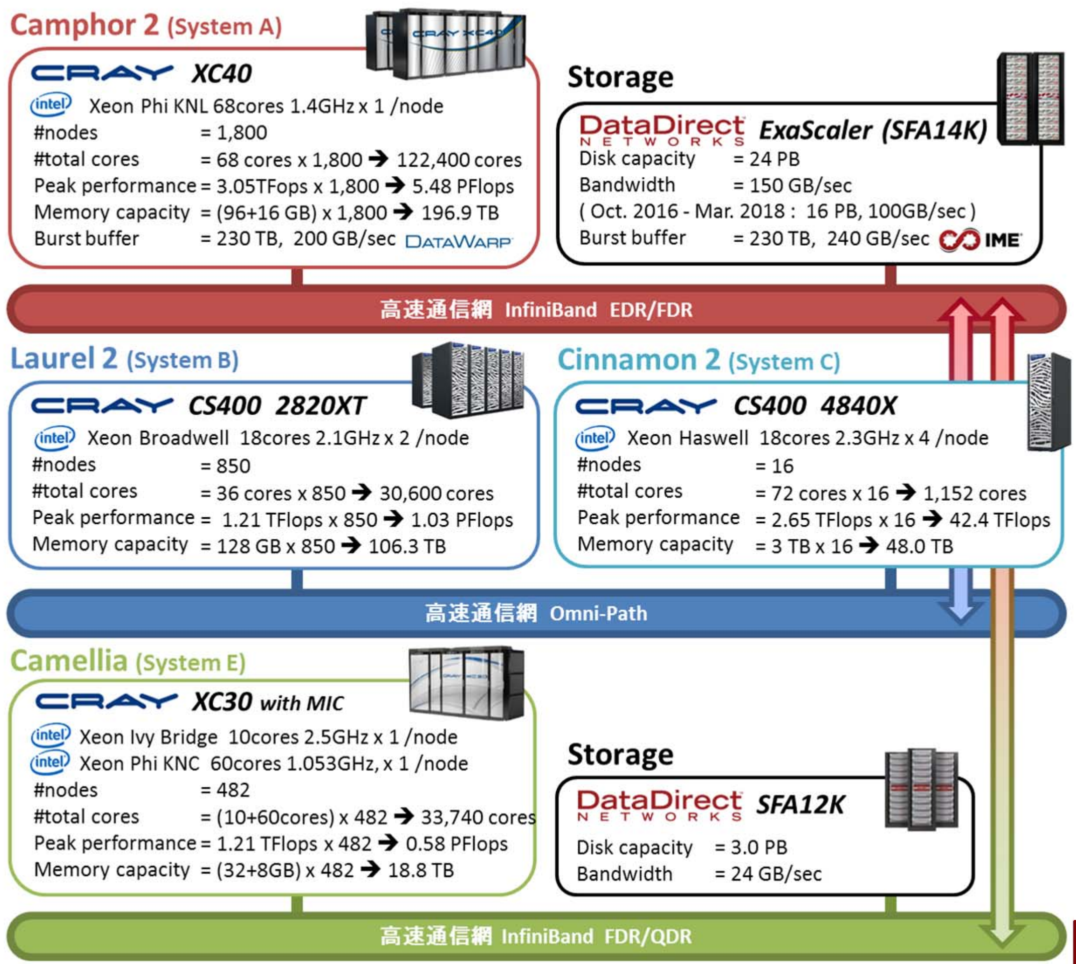
- **University of Tokyo**
- **I/O System by DDN**
  - 5 PB ExaScaler/Lustre @ 145.2 GB/sec
  - 200TB of IME Burst Buffer with NVMe @ 436.2 GB/sec



DDN | EXAScaler & Lustre Case Studies
Kyoto University Supercomputer System

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

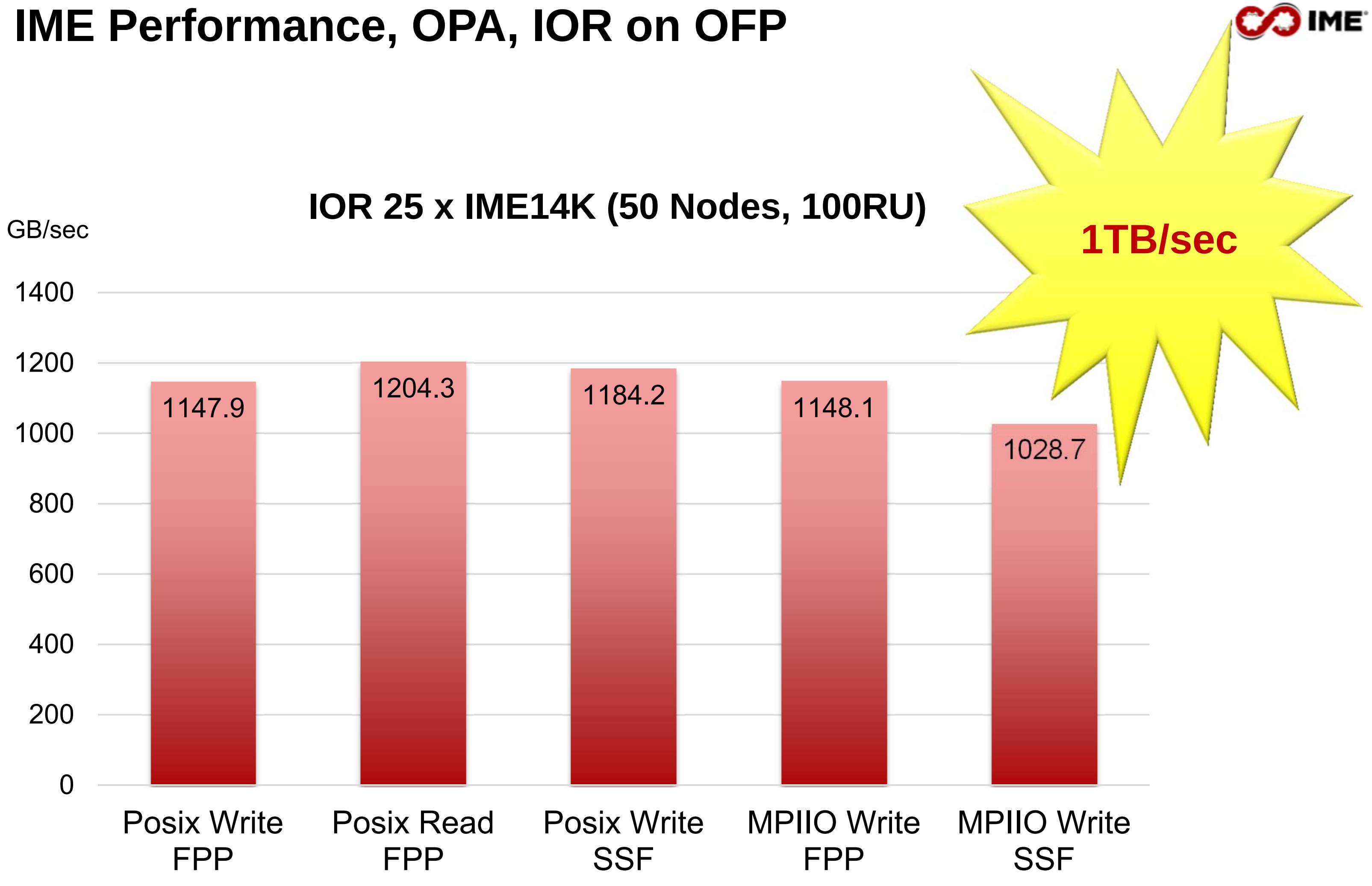
- **Kyoto University**
- **I/O System by DDN**
  - 24PB ExaScaler/Lustre @ 150 GB/sec
  - 230TB of IME Burst Buffer with NVMe @ 240 GB/sec



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)**

IME Performance, OPA, IOR on OFP

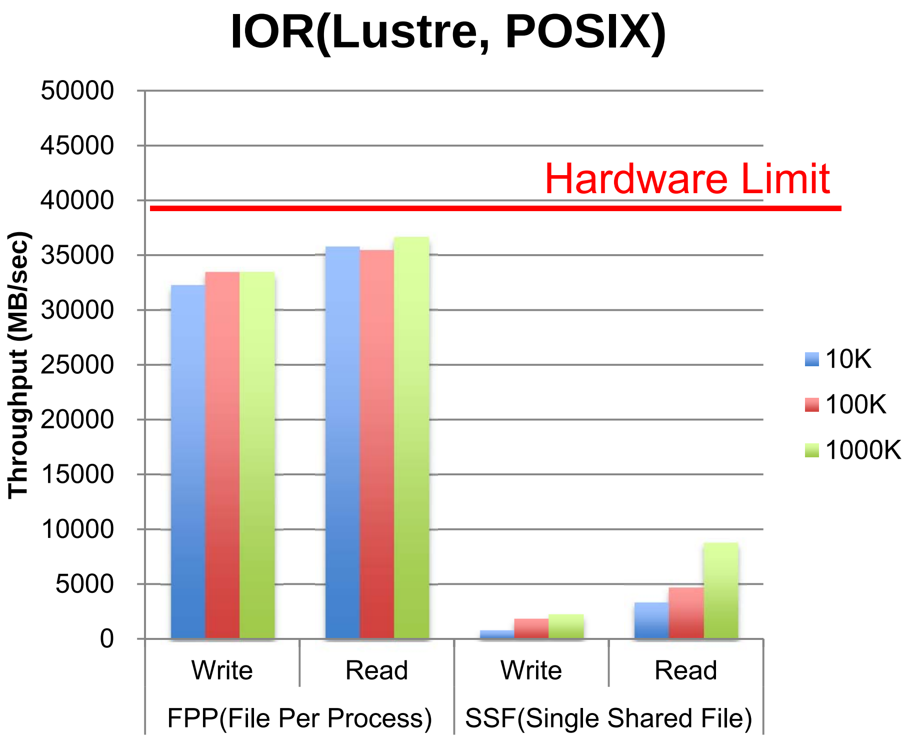


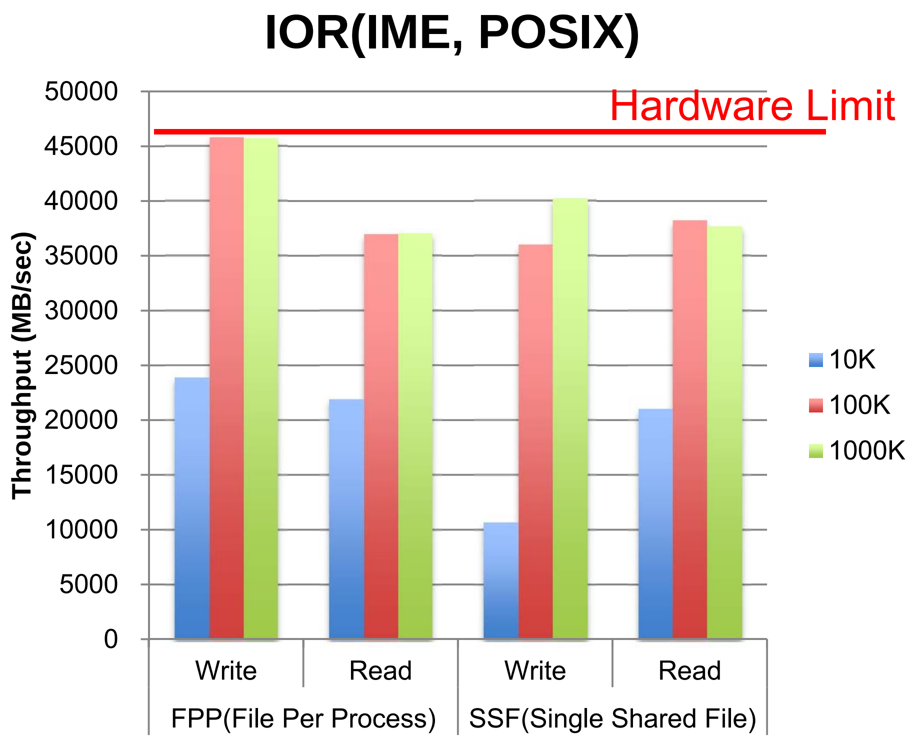
FPP : File Per Process
SSF : Single Shared File

FPP vs SSF

- **FPP(File Per Process)**
  - プロセス毎に独立したファイルを使用
  - メリット：並列ファイルシステムが得意、スケールしやすい
  - デメリット：超並列（何万・何十万プロセス）実行時など膨大なファイルによってメタデータ性能ネックになる
- **SSF(Single Shared File)**
  - 全てのプロセスで単一ファイルを共有
  - HDF5, NetCDFなどを利用
  - メリット：メタデータ性能ネックにならない
  - デメリット：FPPより場合によっては複雑

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

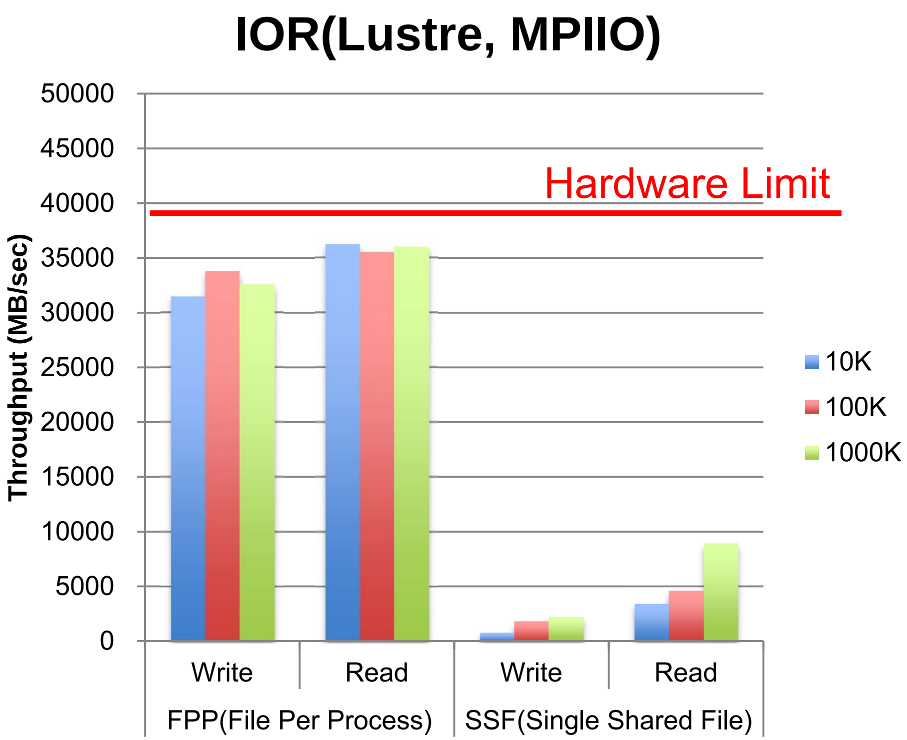


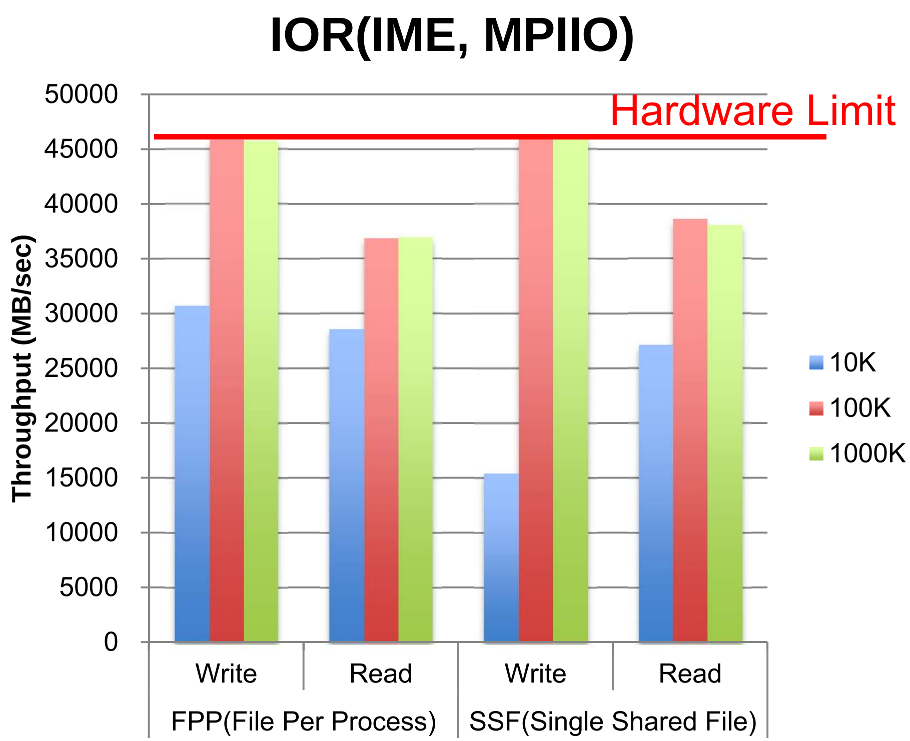


FPP I/O Efficiency ~84%(Write) ~90%(Read)
SSP I/O Efficiency ~5%(Write) ~22%(Read)

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

Lustre vs IME (IOR MPI-IO)
32 clients, 512 process, 3.3TB File Size



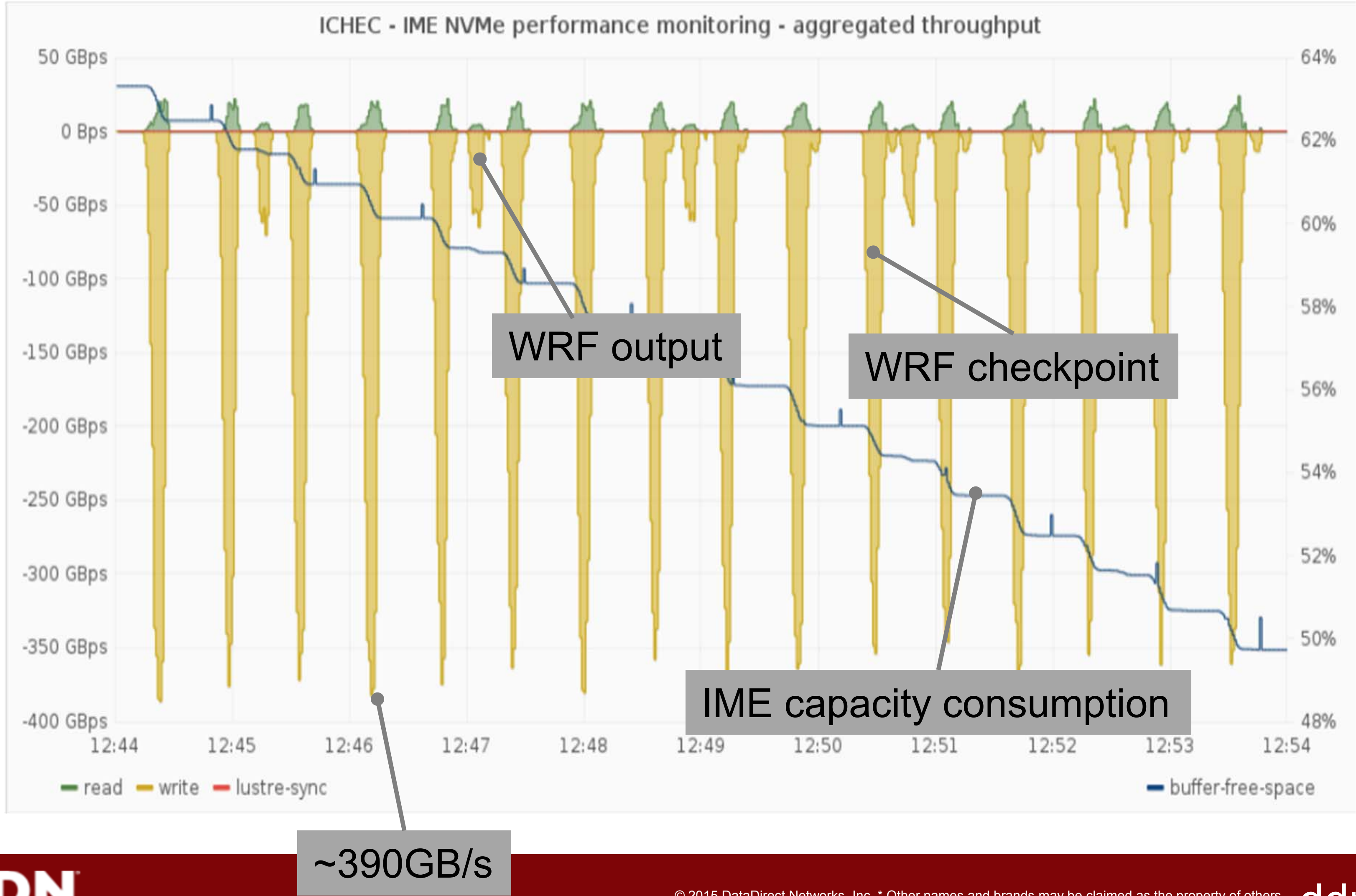


FPP I/O Efficiency ~84%(Write) ~90%(Read)
SSP I/O Efficiency ~5%(Write) ~22%(Read)

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

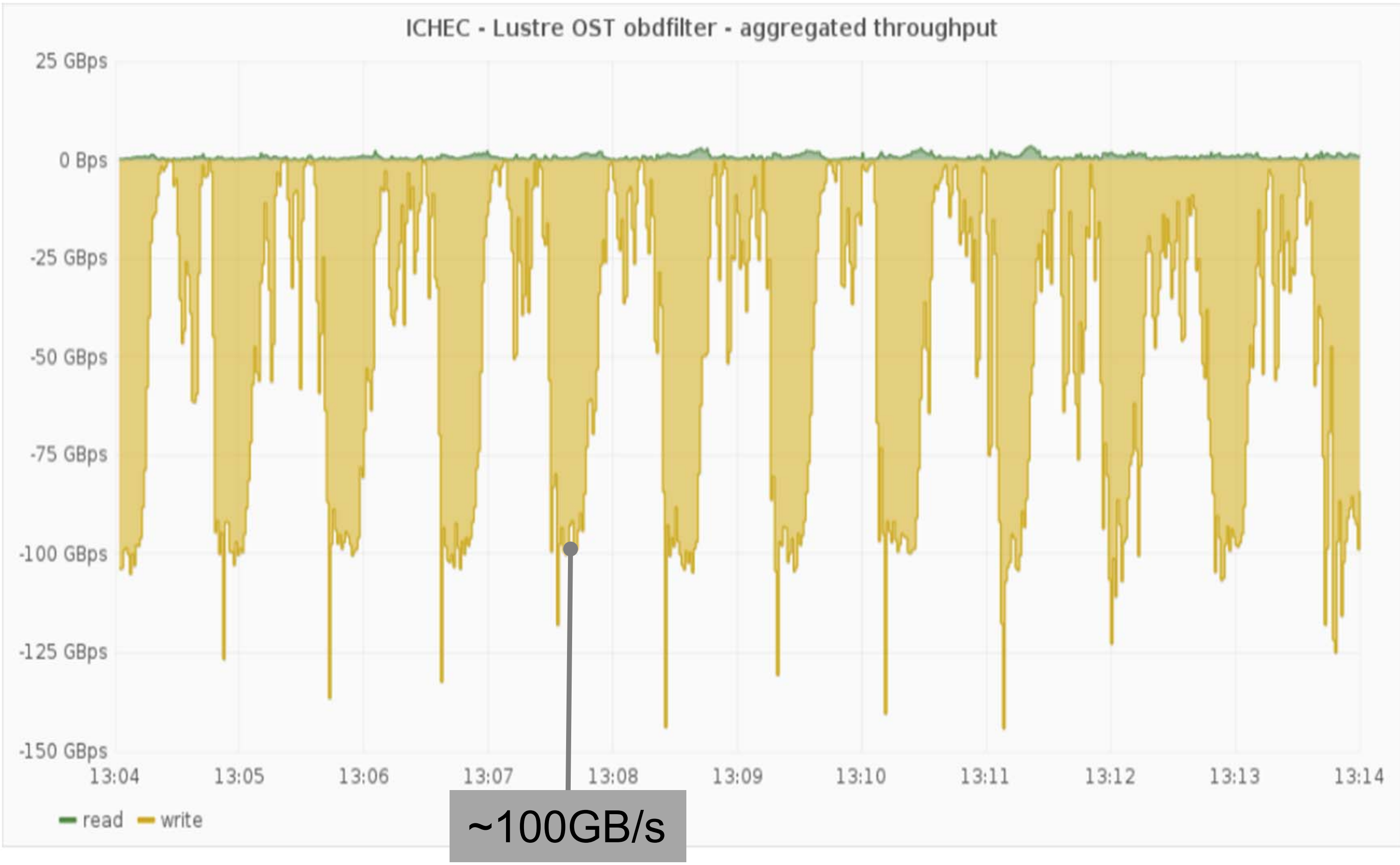
WRF on IME

48 job, total 240 compute node – throughput – focus, 5 minute window



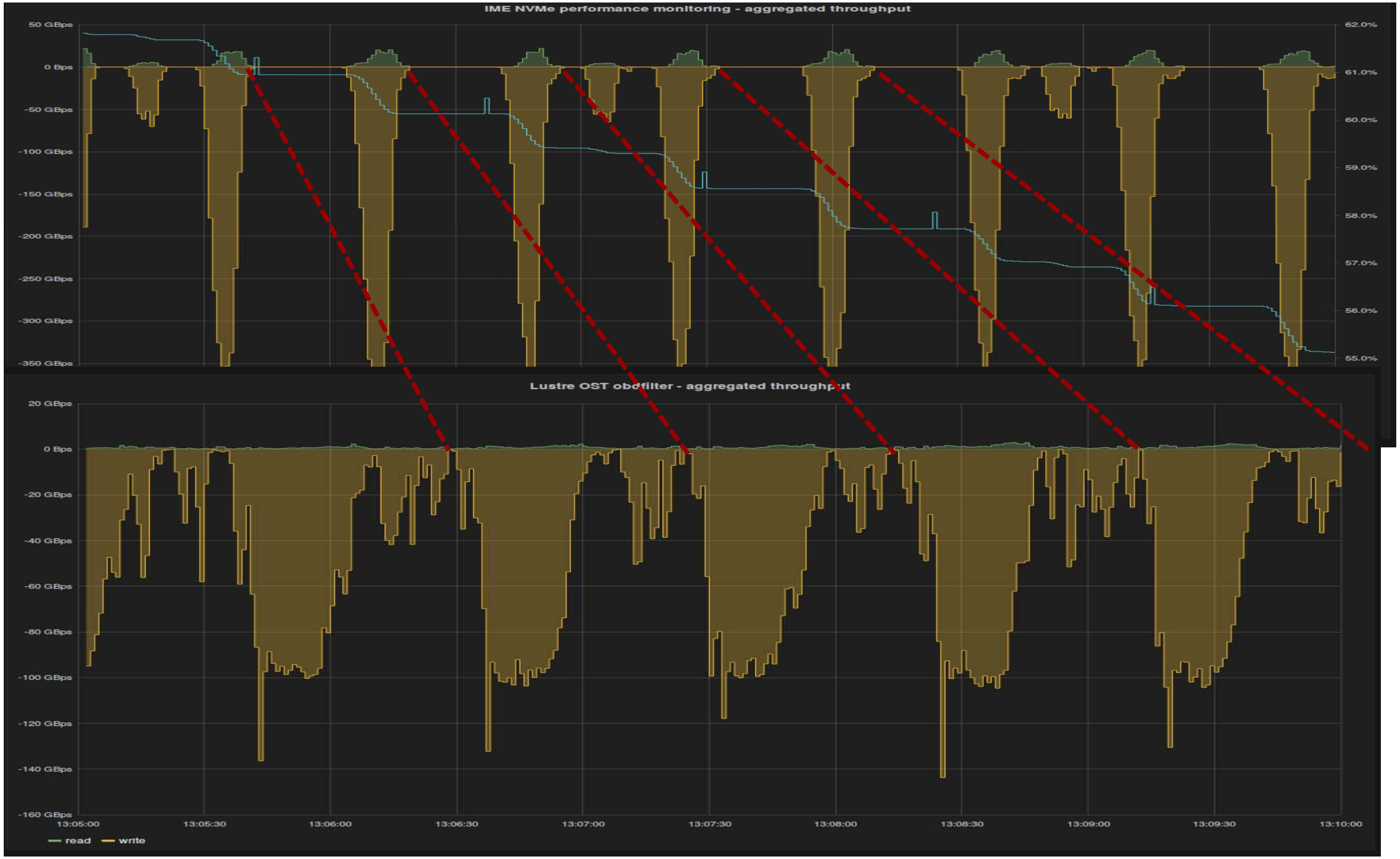
WRF on Lustre

48 job, total 240 compute node – throughput – focus, 5 minute window



WRF on IME at scale

Application runtime speedup over 1.5x



DDN | DDN IME Platforms

- **IME120 (SuperMicro Platform)**
  - Runs IME server software
  - 1P Broadwell, 1x EDR/OPA 1RU
  - 6 2.5” NVMe SSDs
  - 64-128 GB DRAM @2400
  - 4.6-12TB IME Capacity
  - Max 10GB/s throughput per server

- **IME240 (SuperMicro Platform)**
  - Runs IME server software
  - 2P Broadwell, 2x EDR/OPA, 2 RU
  - 20 2.5” NVMe SSDs
  - 128 – 256 GB DRAM @2400
  - 9.6-40TB IME Capacity
  - Max 20 GB/s throughput per server

- **IME14K Appliance**
  - Two controllers in 4 RU
  - 2P Haswell / Broadwell, EDR/OPA
  - 48 NVMe SSDs (V1.0.0) and 72 SAS SSDs (1H’16)
  - 38.4-86.4TB IME Capacity
  - 50 GB/s raw throughput per appliance

***Note:** Erasure coding topology will impact achievable peak performance!*

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