Contribution to a Dialog Between HPC and Cloud Based File Systems

DDN STORAGE

Jean-Thomas ACQUAVIVA
jacquaviva@ddn.com

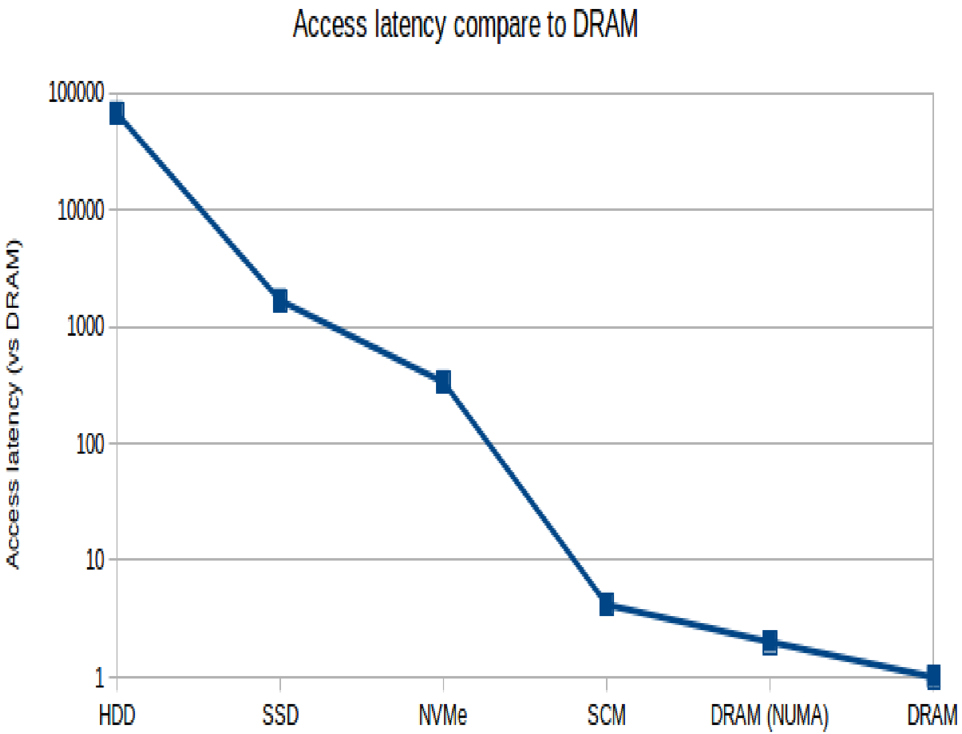
CS3

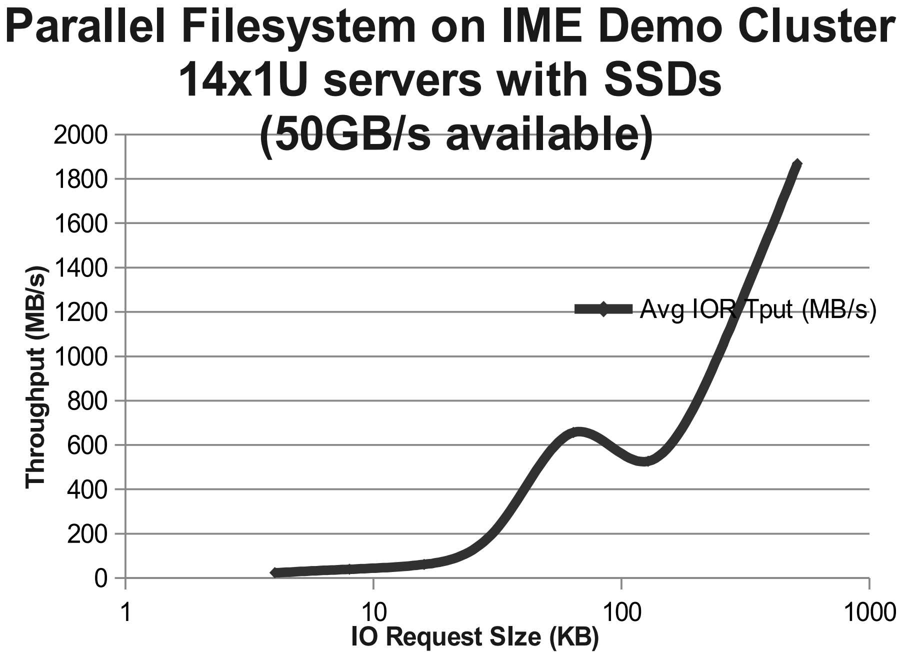
2018, JANUARY 31th

DDN overview

- 650 persons WW
- R&D in US, EU, ASIA
- 70% Top500

3 FLASH disruption: Software to be redesigned





INFINITE MEMORY ENGINE™ Flash Native IO Accelerator

Compute

Diverse, high concurrency applications

IME™

Fast Data NVM & SSD

SFA

Persistent Data (Disk)

Application issues IO to IME client. Erasure Coding applied

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

INFINITE MEMORY ENGINE™ Fast Forward Flash native

Improve the sys admin interface, thanks!

GA IME 1.0

Korea Institute of Science and Technology Information

Ohio Supercomputer Center

京都大学 KYOTO UNIVERSITY

A*STAR

University of Tokyo

NSCC SINGAPORE

JCAHPC

CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre

Oak Ridge National Laboratory

KISTI www.kisti.re.kr

SDSC San Diego Supercomputer Center

ICHEC Irish Centre for High-End Computing

NCAR

JÜLICH FORSCHUNGSZENTRUM

TACC Texas Advanced Computing Center

NERSC

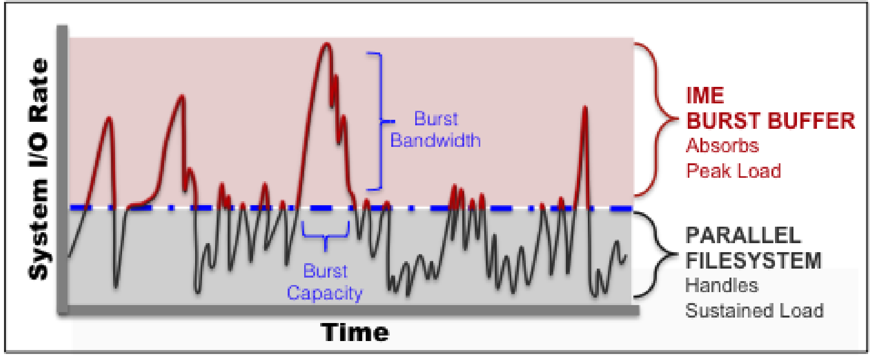
Tokyo Tech

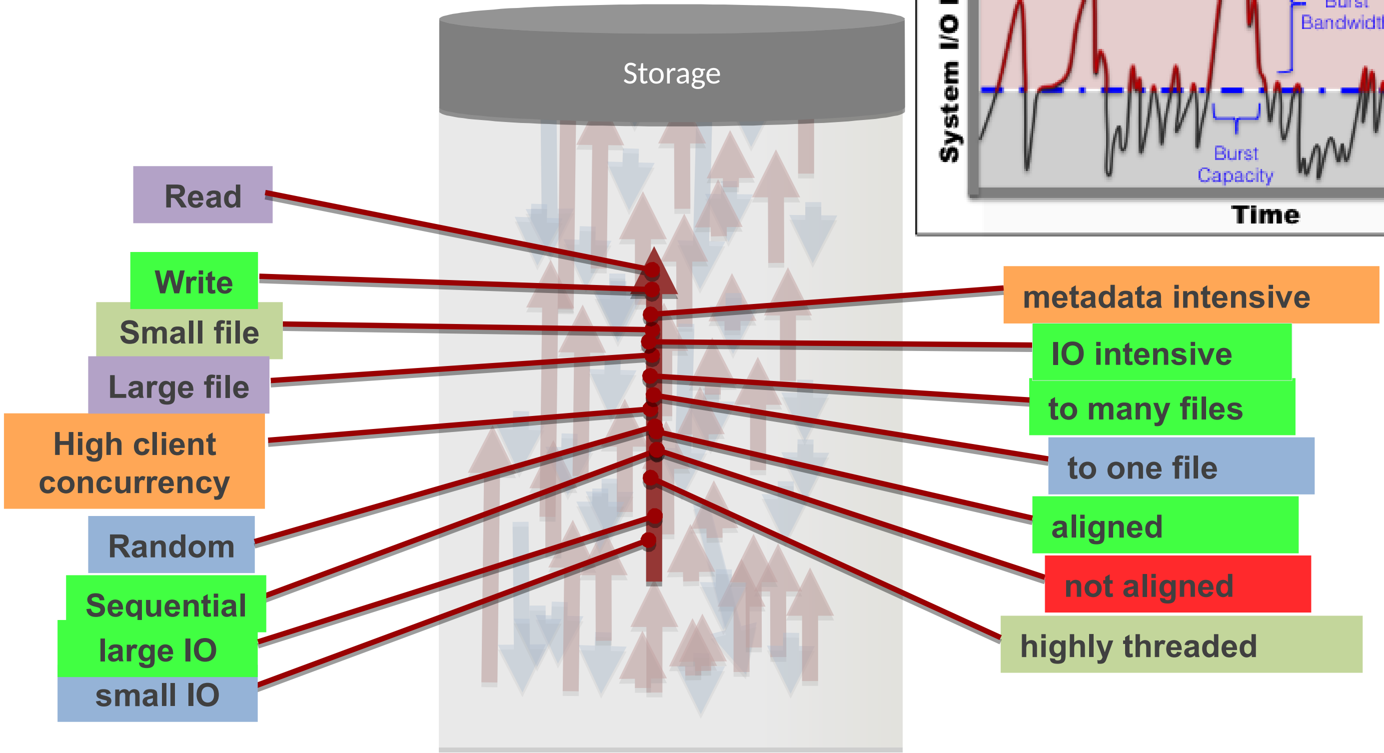
Argonne National Laboratory

2013–2014	2015–2016	2016–2017
IME Development	IME Testbed Program	IME Major Deployments

2013 2014 2015 2016 2017

Problem Statement – Diversity of loads





Diversity of Load: IO500

IO500 (November 2017)*

#	information				io500		
	system	institution	filesystem	client nodes	score	bw GiB/s	md kIOP/s
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	21.85
2	Shaheen	Kaust	DataWarp	300	70.90	151.53	33.17
3	Shaheen	Kaust	Lustre	1000	41.00	54.17	31.03
4	JURON	JSC	BeeGFS	8	35.77	14.24	89.83
5	Mistral	DKRZ	Lustre	100	32.15	22.77	45.39
6	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.38
7	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.58
8	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.57
9	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98

* io500.org

Oakforest-PACS ranked 9 TOP500 (November 2017)

- **Compute nodes**
  - 2048x Intel Xeon Phi 7250 (KNL) client nodes with Intel Omni-path network
- **IME 1.1 on 25x IME14K (=50 IME servers)**
  - 1200 NVMe SSDs (940 TB)
  - Theoretical peak B/W: 1,560 GB/s
  - Erasure coding 9+1 (per pool)
- **Backing File System: Lustre on SFA14KE nodes**
  - 500 GB/sec
  - 26.2 PB

http://jcahpc.jp/files/OFP-basic.pdf

IO500 Detailed Results

Rank	System	Institution	Filesystem	Client Nodes	Score	BW	MD	Detailed write: Easy Write	Detailed write: Hard Write	Detailed write: Hard vs. Easy	Detailed read: Easy Read	Detailed read: Hard Read	Detailed read: Hard vs. Easy
						GiB/s	kIOP/s	GiB/s	GiB/s		GiB/s	GiB/s	
1	Oakforest-PACS	JCAHPC	IME	2048	101.48	471.25	19.04	742.38	600.28	80.9%	427.41	258.93	60.6%
2	Shaheen	Kaust	DataWarp	300	70.9	151.53	33.17	969.45	15.55	1.6%	894.76	39.09	4.4%
3	Shaheen	Kaust	Lustre	1000	41	54.17	31.03	333.03	1.44	0.4%	220.62	81.38	36.9%
4	JURON	JSC	BeeGFS	8	35.77	14.24	89.81	30.42	1.46	4.8%	48.36	19.16	39.6%
5	Mistral	DKRZ	Lustre	100	32.15	22.77	46.64	158.19	1.53	1.0%	163.62	6.79	4.1%
6	Sonasad	IBM	Spectrum Scale	10	21.63	4.57	102.43	34.13	0.17	0.5%	32.25	2.33	7.2%
7	Seislab	Fraunhofer	BeeGFS	24	18.75	5.13	68.55	18.79	0.89	4.7%	22.34	1.86	8.3%
8	EMSL Cascade	PNNL	Lustre	126	11.17	4.88	25.59	17.81	0.39	2.2%	30.19	2.72	9.0%
9	Serrano	SNL	Spectrum Scale	16	4.25	0.65	27.98	1.08	0.22	20.4%	1.03	0.71	68.9%

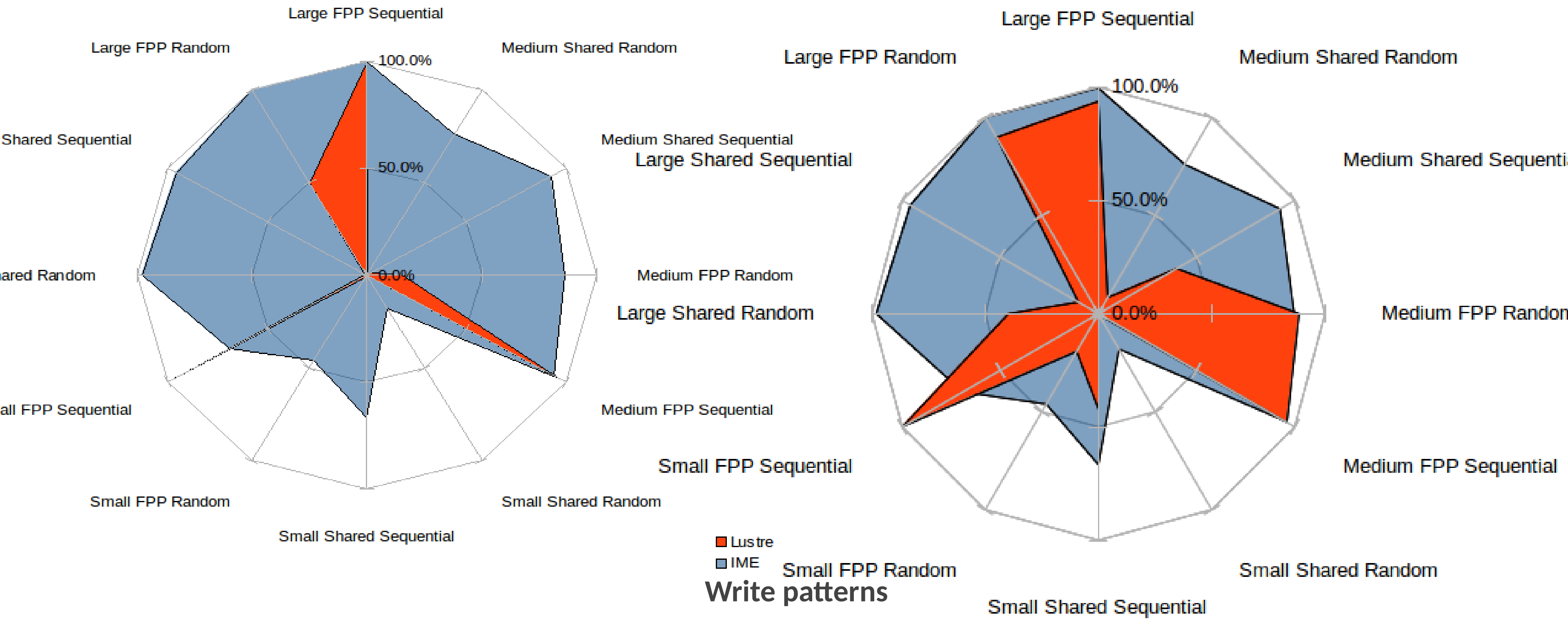
MDtest Benchmark

IOR Benchmark

Acknowledging multi-criteria performance metrics

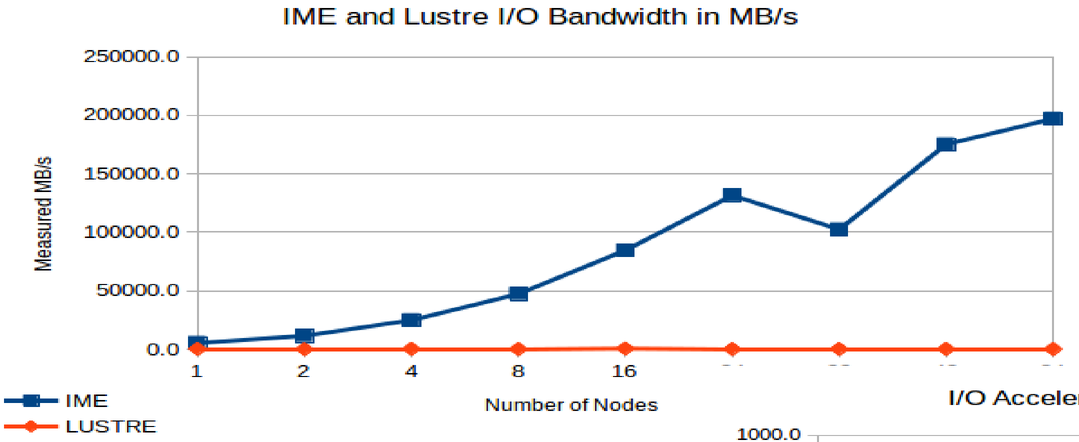
I/O Granularity	I/O control plane Pattern	I/O Data plane Pattern	
Large (>= 1MB)	File Per Process (= share nothing)	Sequential	IO500 Easy !
Large	File Per Process	Random	
Large	Single Shared File	Sequential	
Large	Single Shared File	Random	
Small	File Per Process	Sequential	
Small	File Per Process	Random	
Small (47008 Bytes)	Single Shared File	Sequential	IO500 Hard !
Small	Single Shared File	Random	

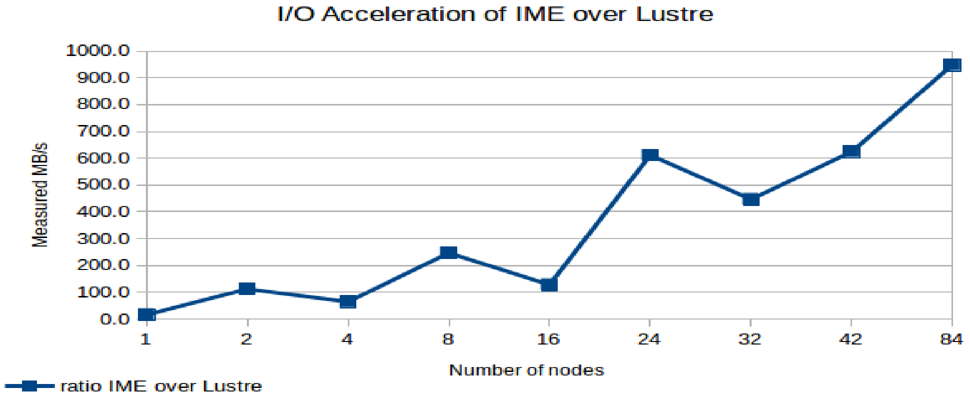
10 IO500 to a comprehensive picture: SDD vs HDD



Write patterns

Neurosciences with IME @ A*Star



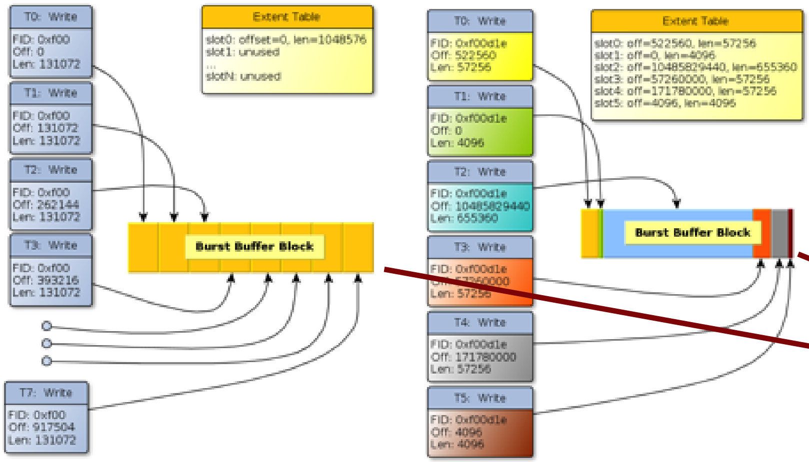


IME delivers x1000 more bandwidth than Lustre

Byte addressable device allows log structured

Sequential

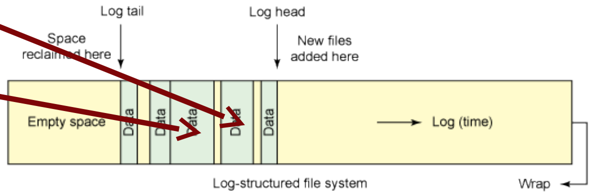
Non-sequential



'*Burst buffer blocks*' (BBB) are really just buffers generated at the client

Note the contents of the BBB can be aligned or not.

The same storage method is used for both blocks (despite the qualitative difference of their contents)!

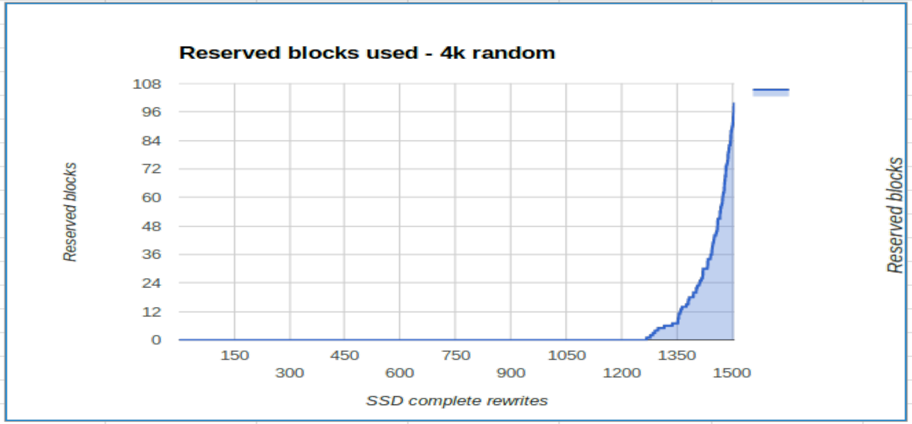


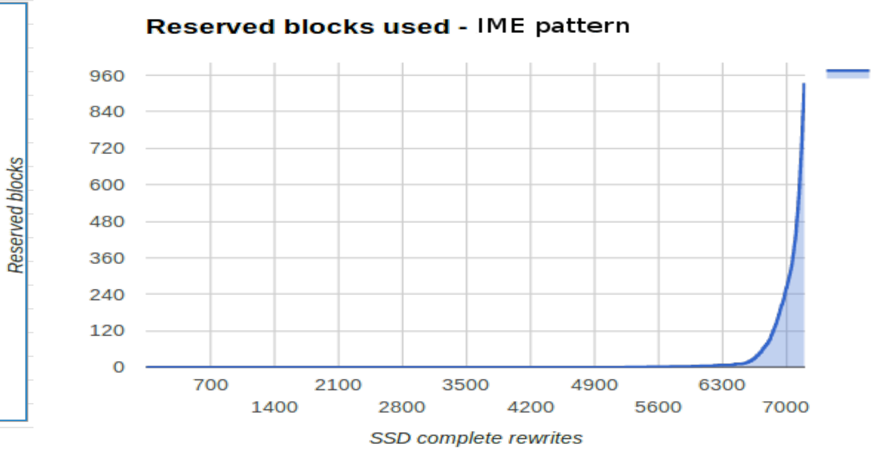
SSD brings its own complexity: Write Amplification

DWPD as a key endurance metric

→ Based on SNIA workload 4KB random

→ Fine grain control of device access pattern: life expectancy **x 4**

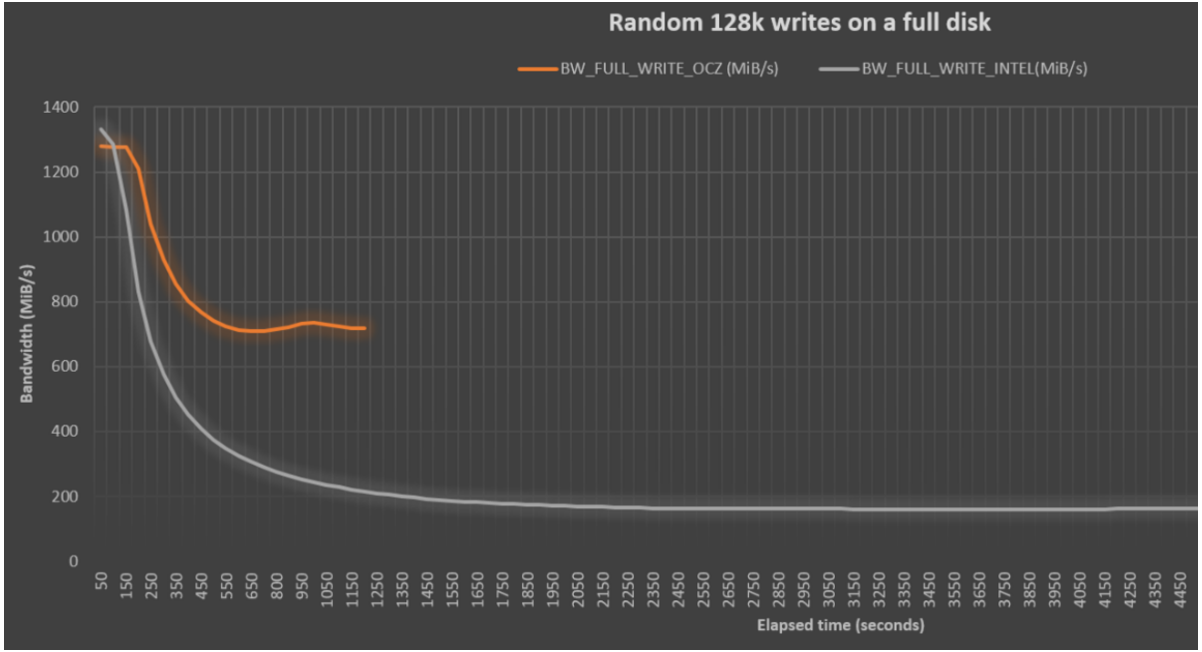




14

SSD brings its own complexity: Garbage Collector

Impact on performance on devices where ***unmap*** operation are not used*



Higher is better

Disks are first completely written. No unmap operations are used. 128K random write operations are then performed.

Intel NVMe P3520 1.2 TB

Start: **1350** MB/s

End: **180** MB/s

OCZ NVMe 6000 800GB

Start: **1300** MB/s

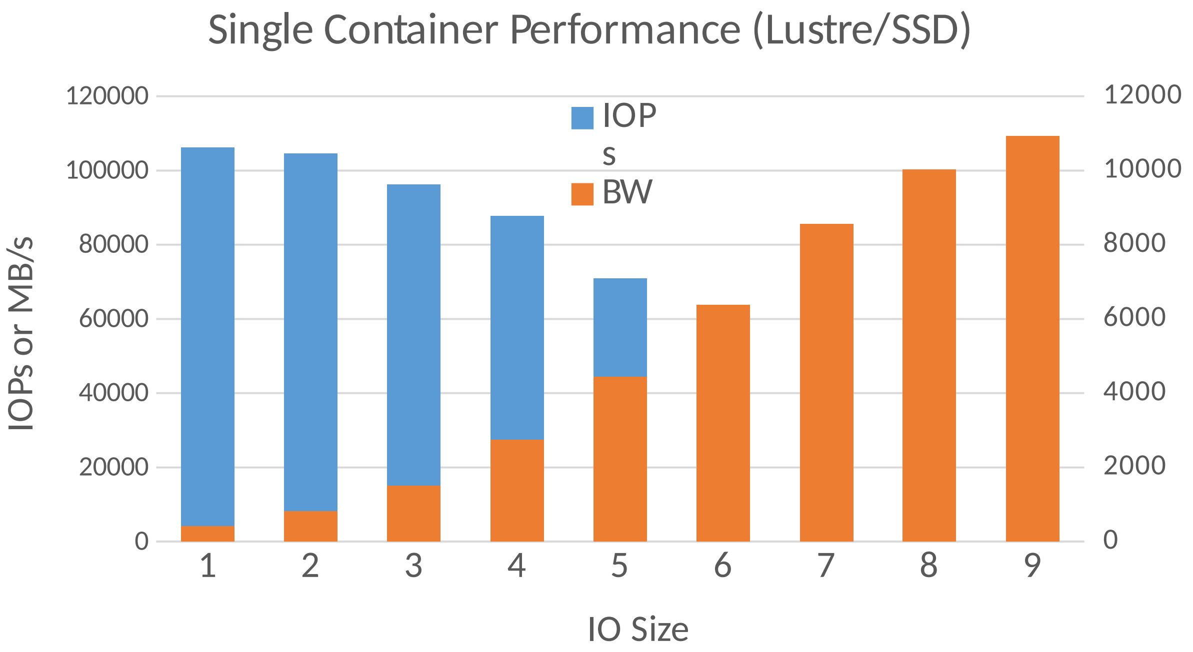
End: **750** MB/s

* *Joint work with UBO University of Western Brittany Brest*

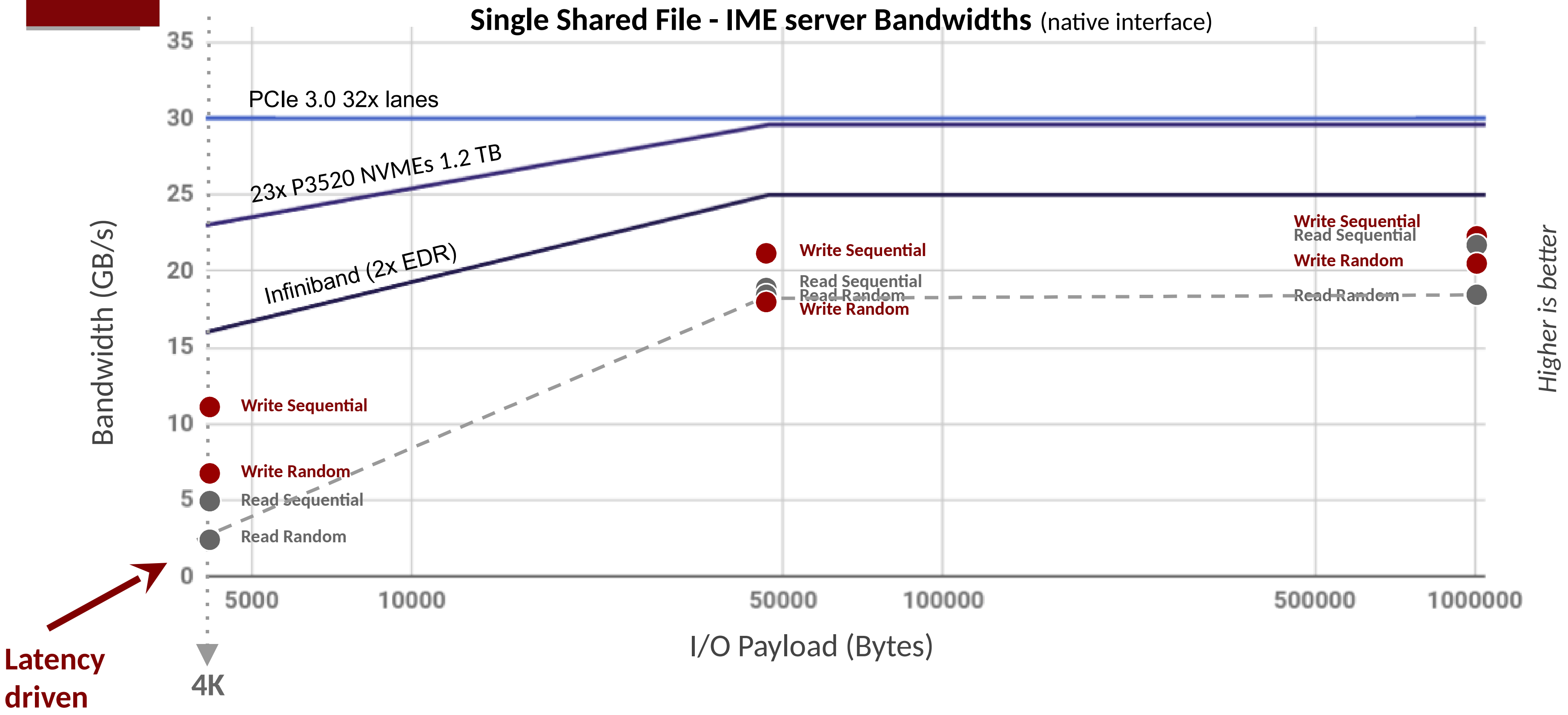
Note: IME sends unmap commands to the disks. Thus, this kind of performance drop does not happen.

DGX DDN EXAScaler Volume Performance

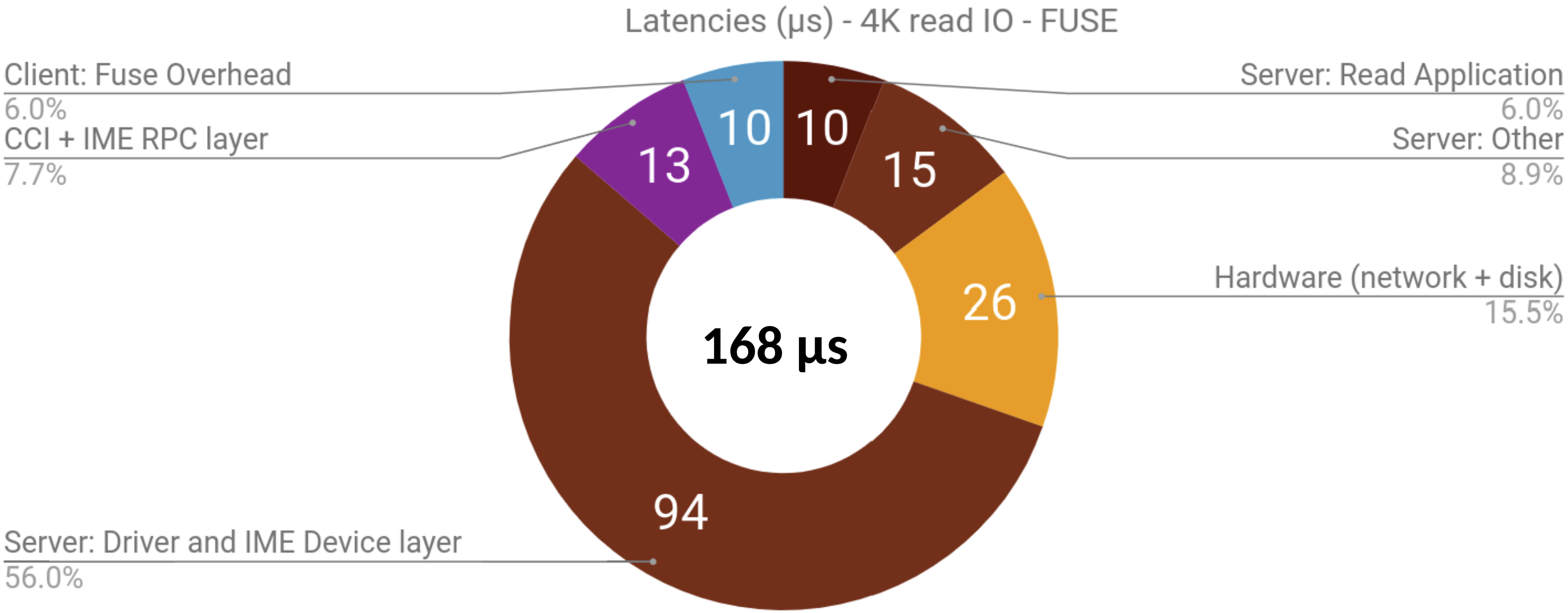
- Over 100K IOPs and 11GB/s to a single container



Toward an I/O roofline model

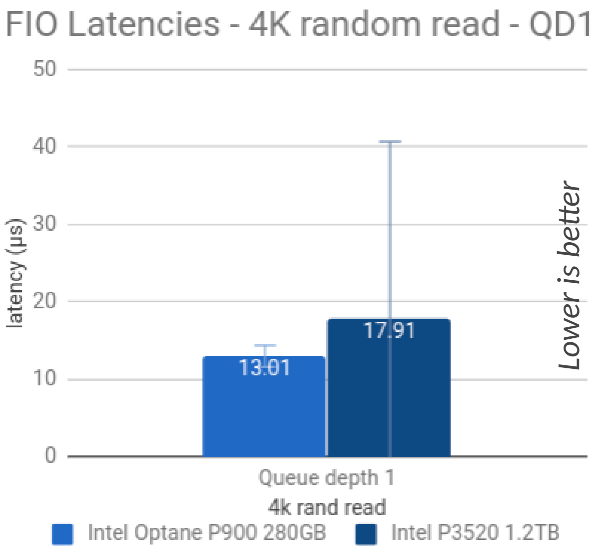


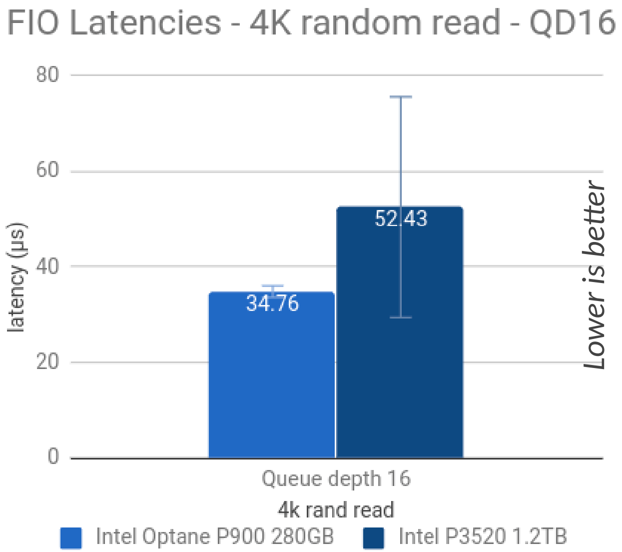
I/O critical path break down

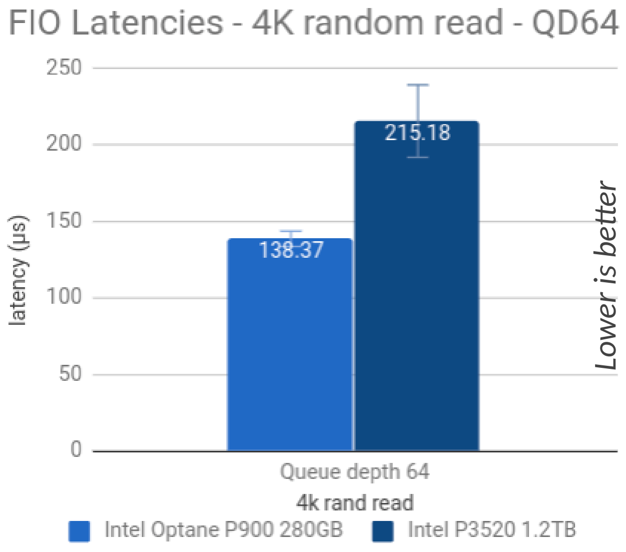


Software overhead dominates over HW latency

3D XPoint (Intel Optane) vs NAND

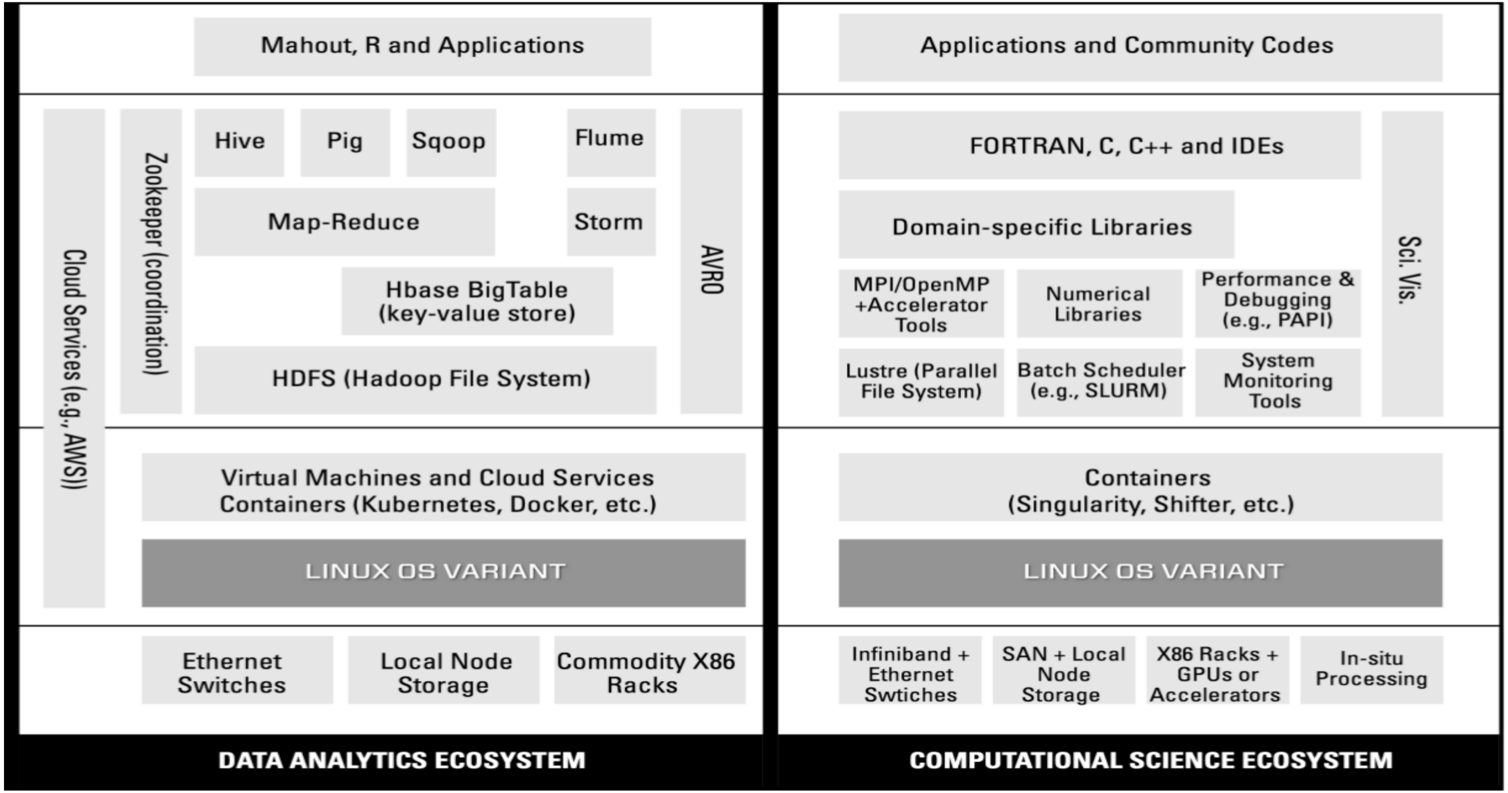






	Pros	Cons
Intel Optane	- Steady latencies *(fixed queue depth)* - No unmap required	- Price *(3.5x higher per GB)* - Latency close to high-end NVMes

Problem Statement – Diversity of stacks



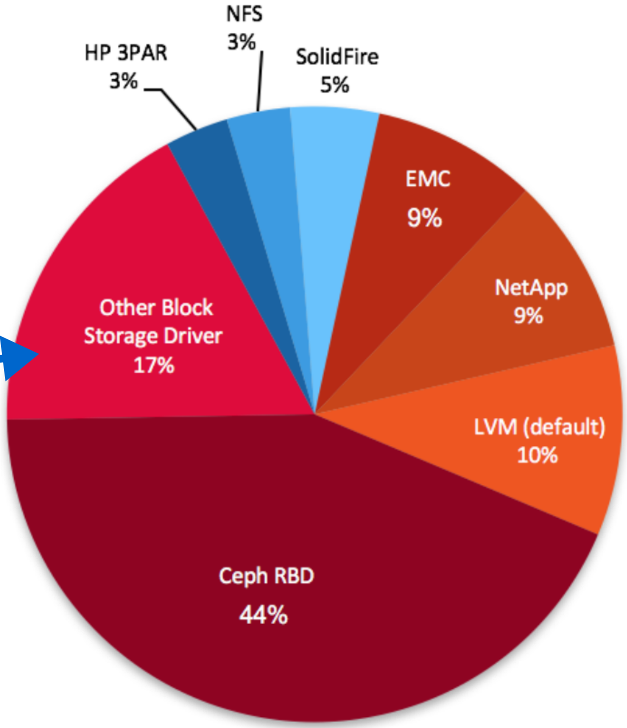
Source F. Bodin, Séminaire Maison de la Simulation, April 2017

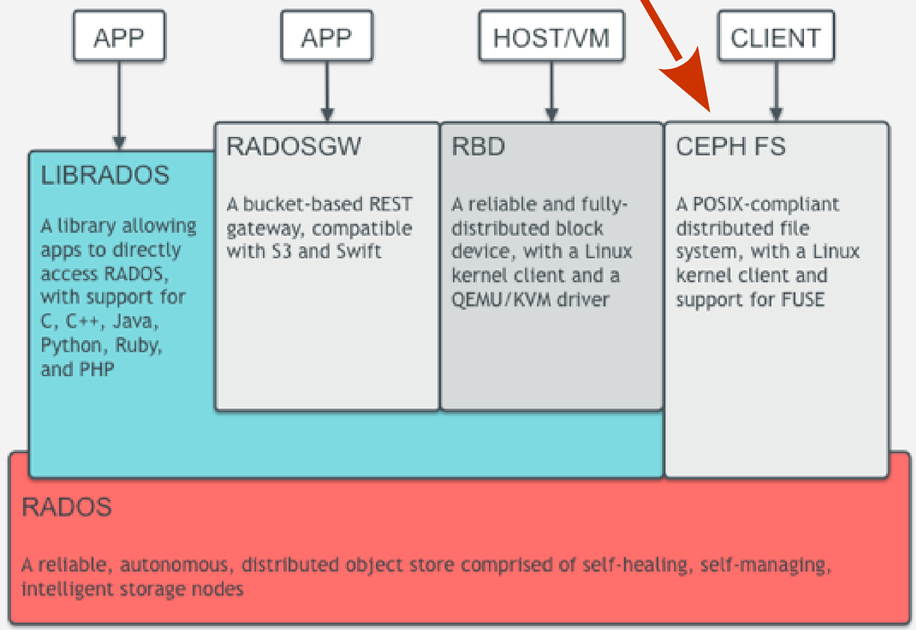
Problem Statement – Diversity of needs

HDFS < NFS < CephFS < {XFS, ext4}
CEPH file creation: few 1000s per second

e.g. OpenStack deployment

HPC FS here
GPFS 1%





Diversity of needs: Cloud Service SLA vs HPC

- Put/Get 10KB < 150 ms
- Put/Get 1.3 MB < 250 ms
- Put 10MB < 2 sec
- Get 10MB < 2.5 sec
- Availability 99.9%
- Durability 99.999999%

Extreme HPC workload:
→ Write ½ 2PB memory in a large file

Healthcare AI workload
→ 100s of thousands of 21KB I/O reads to train the network

Courtesy of Orange Object Storage Group
PER3S, Rennes Jan. 2018

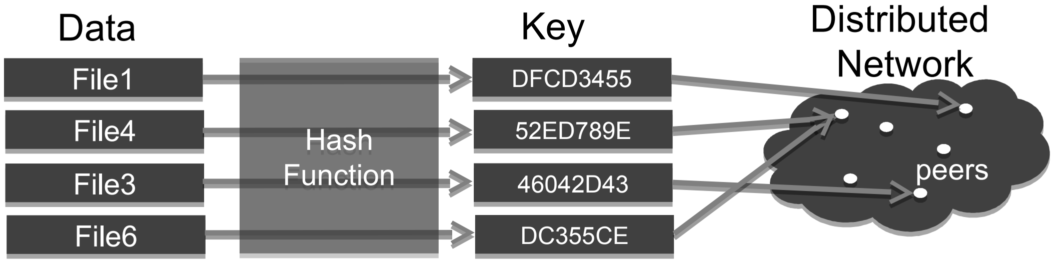
Metadata management DHT: ETCD

3 machines of 8 vCPUs + 16GB Memory + 50GB **SSD**
1 machine(client) of 16 vCPUs + 30GB Memory + 50GB **SSD**
etcd 3.2.0, go 1.8.3
With this configuration, etcd can approximately write:

NUMBER OF KEYS	KEY SIZE IN BYTES	VALUE SIZE IN BYTES	NUMBER OF CONNECTIONS	NUMBER OF CLIENTS	TARGET ETCD SERVER	AVERAGE WRITE QPS	AVERAGE LATENCY PER REQUEST	AVERAGE SERVER RSS
10,000	8	256	1	1	leader only	583	1.6ms	48 MB
100,000	8	256	100	1000	leader only	44,341	22ms	124MB
100,000	8	256	100	1000	all members	50,104	20ms	126MB

IME Metadata management: DHT

FABRIC-AWARE



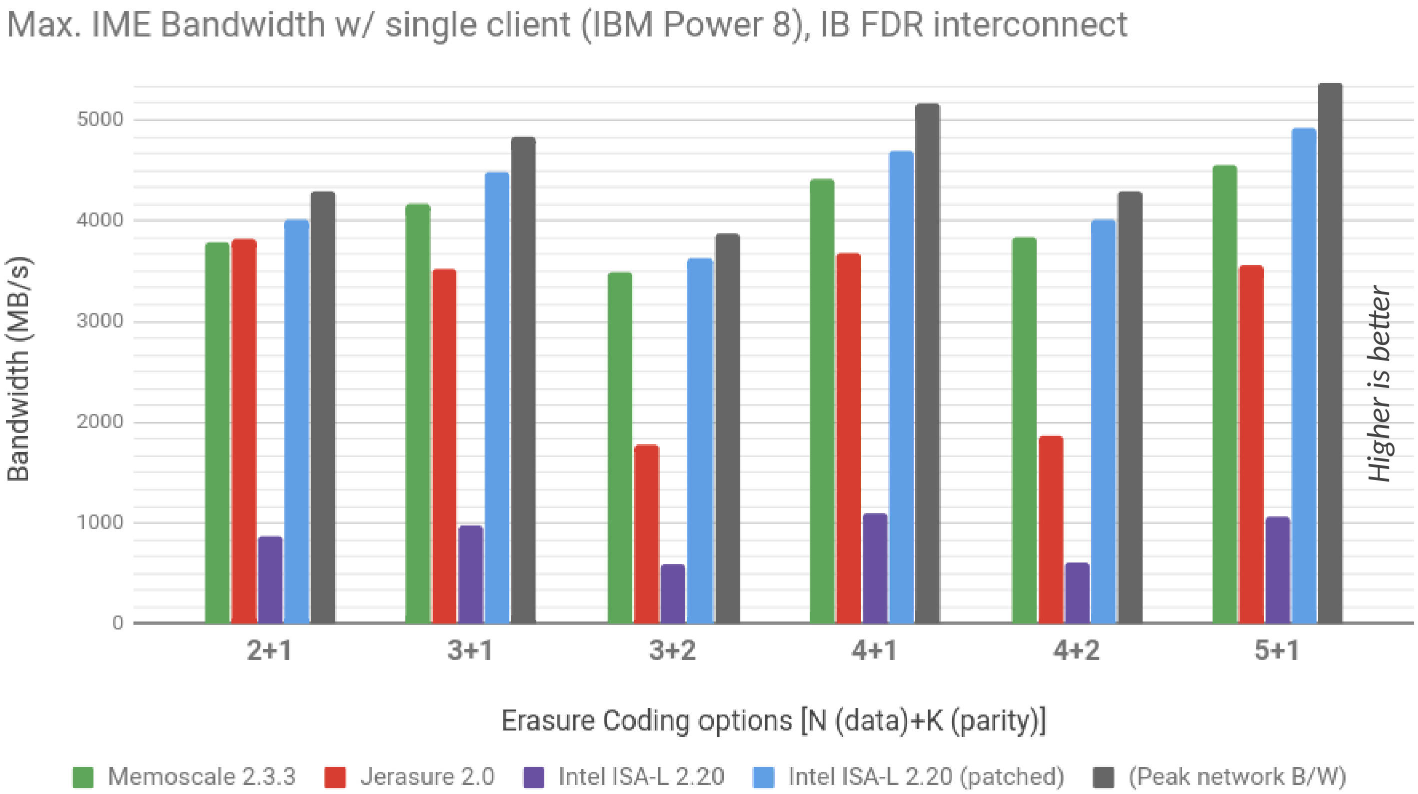
DHT provides foundation for
- Network parallelism
- Node-level fault tolerance
- Distributed metadata
- Self-Optimising for Noisy Fabrics

IME DHT relies on CRUSH* (as well as CEPH)

- ~2000,000 insertion / sec / server when protected by journal + RAFT

*Weil SA, Brandt SA, Miller EL, Maltzahn C. CRUSH: Controlled, scalable, decentralized placement of replicated data. In Proceedings of the 2006 ACM/IEEE conference on Supercomputing 2006 Nov 11 (p. 122). ACM.)

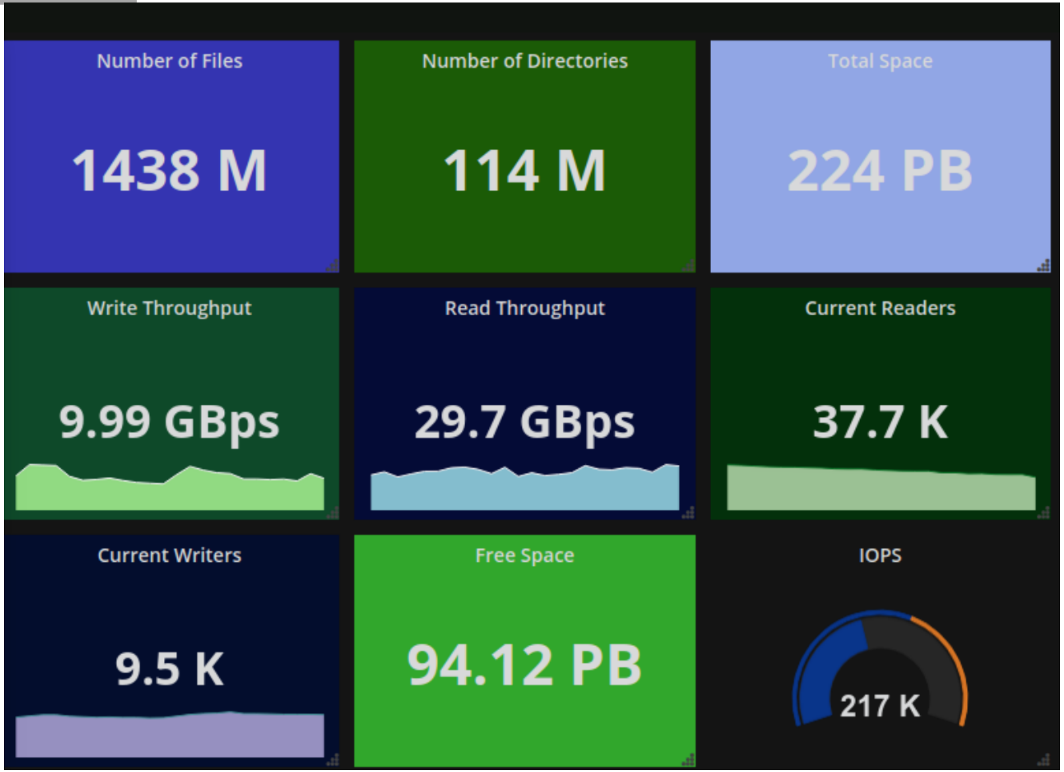
Erasure coding is cheap: redundancy is expensive



Using ***vpermxor*** vectorial instruction (Applies a permute and exclusive-OR operation on two byte vectors) in ISA-L.

Problem Statement – Scale?

EOS courtesy of Georgios Bitzes*, CERN



DDN AI installation

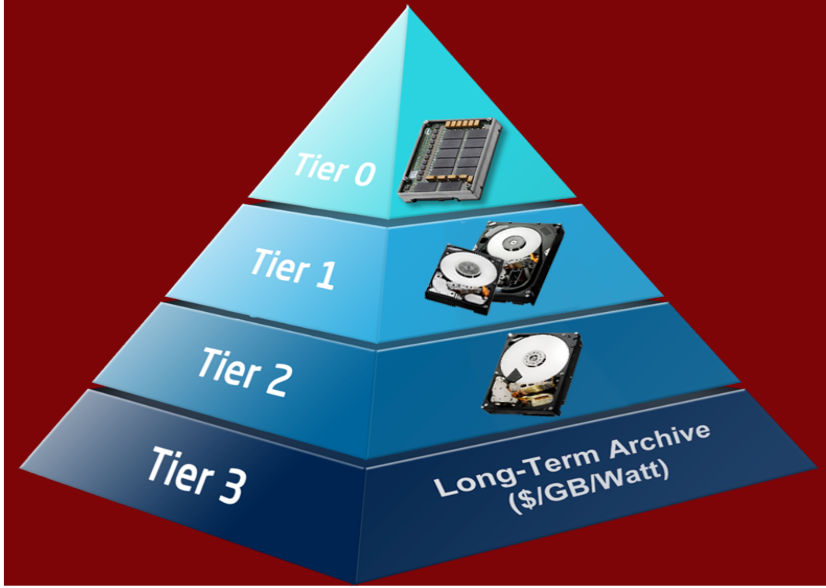
- Lustre Performance at Scale
- 175 Ethernet ES7KX
- Over 1.6TB/s Aggregate Performance
- Economical approach to very large capacities > 170PB
- Flat namespace performance for each Region

* Workshop on Performance and Scalability of Storage Systems, ISC-HPC, Frankfurt, 2017

26 Conclusion and openings

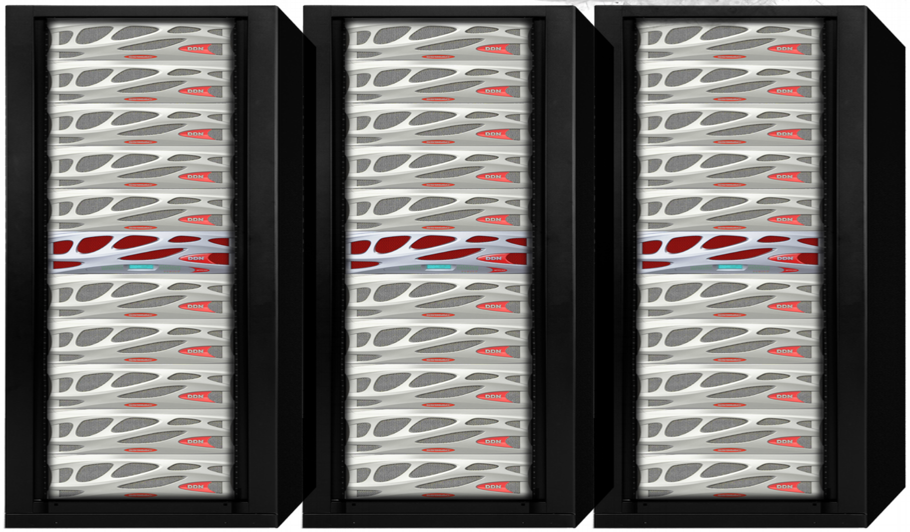
- ► **Convergence of Scale**
  - → no significant difference in size
- ► **Algorithmic Convergence**
  - → DHT for metadata
  - → RAFT for fault tolerance
  - → Erasure coding

Hardware tiering



Software is the new frontier

- **Software Defined Storage**
  - HPDA is reshuffling the deck
  - API & Service
  - Collaboration opportunity
  - At it end if run on silicon...



Thank You!

Keep in touch with us

sales@ddn.com

@ddn_limitless

company/datadirect-networks

9351 Deering Avenue
Chatsworth, CA 91311

1.800.837.2298
1.818.700.4000