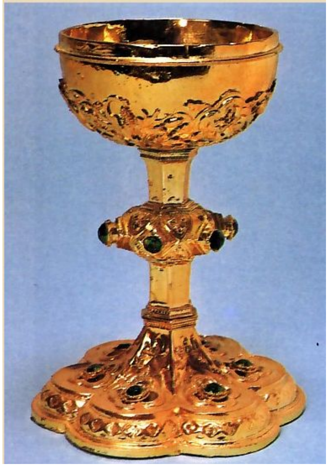
CSPs, Composable PFS, and the Role of Modern Virtual Block Device (IODC ‘24)

Paul Nowoczynski

niova

Have Cloud Service Providers (CSPs) Led us to the Holy Grail?

.. or the 'false grail'?



Introducing the Speaker

- .sys
  - Slash (PSC)
  - Zest (PSC)
  - Slash2 (PSC / NARA)
  - IME (DDN)
  - Niova-block (Niova)
  - PumiceDB (Niova)
- .org
  - Scale8 - Clustered CDN
  - PSC - HPC
    - Built several production archival solutions
    - HDD-base burst buffer
    - PLFS Paper Co-author (~300 citations)
  - DDN - HPC Storage
    - IME - first IO500 Winner
  - DigitalOcean - Cloud Storage
  - Niova - Distributed Block Storage
- .edu
  - B.S. Information Science
    - University of Pittsburgh

> 20 years exp implementing distributed storage software

> 1 million lines of C written

“before storage”

CSPs & IaaS

CSPs have focused a great deal of time and effort on ***Infrastructure as a Service***, as a result they have a set of on-the-fly provisions:

- Compute resources
- Highly reliability Blob Stores
- Fault Tolerant Block Devices
    - *EBS, Azure Managed Disk, GCP Persistent Disk*

Combining these enables the creation of Production-level PFS services!

EC2

Amazon S3

Azure Disk Storage

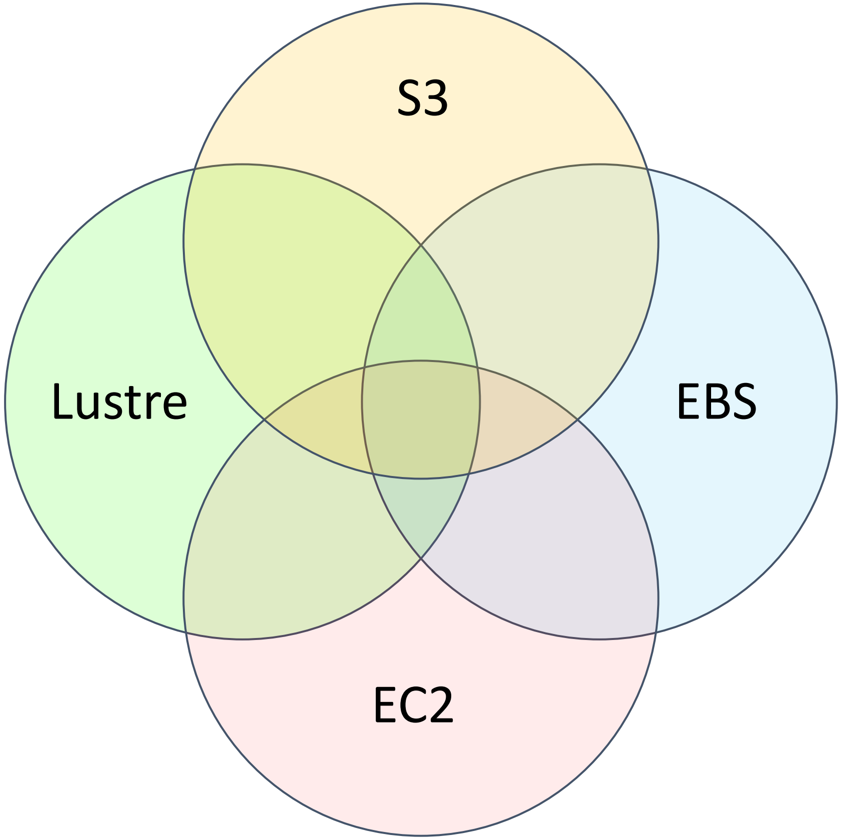
Microsoft Azure Blob Storage

Cloud Storage

Persistent Disk

Amazon EBS

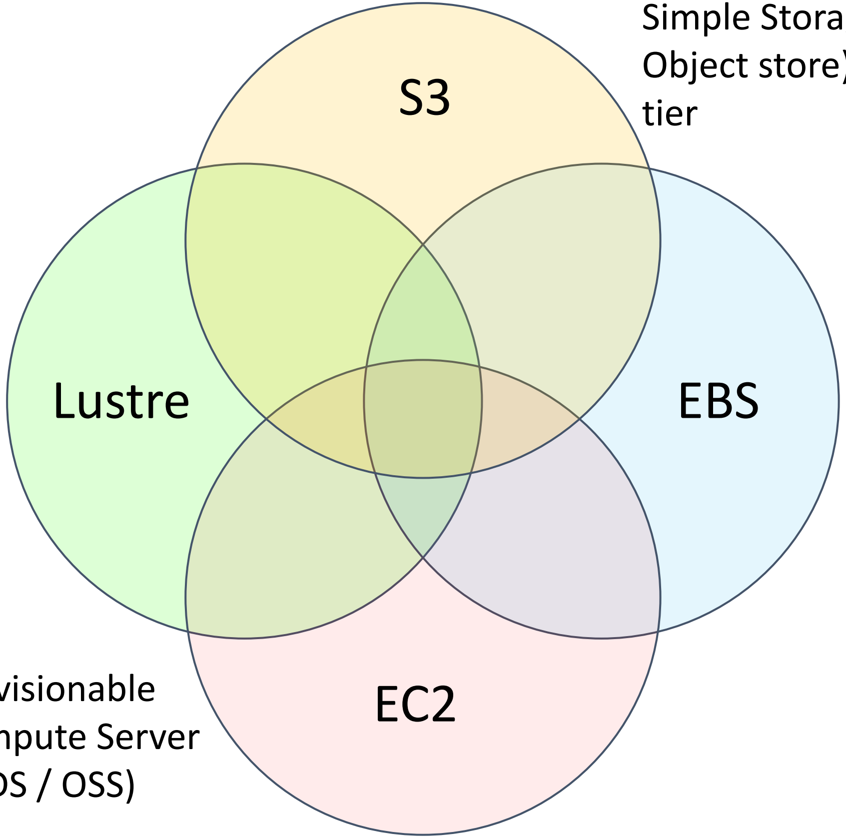
CSP IaaS + Existing PFS Solutions (Lustre + AWS Parlance)



CSP IaaS + Existing PFS Solutions (Lustre + AWS Parlance)

Simple Storage Service (aka Blob / Object store) – serves as archival tier

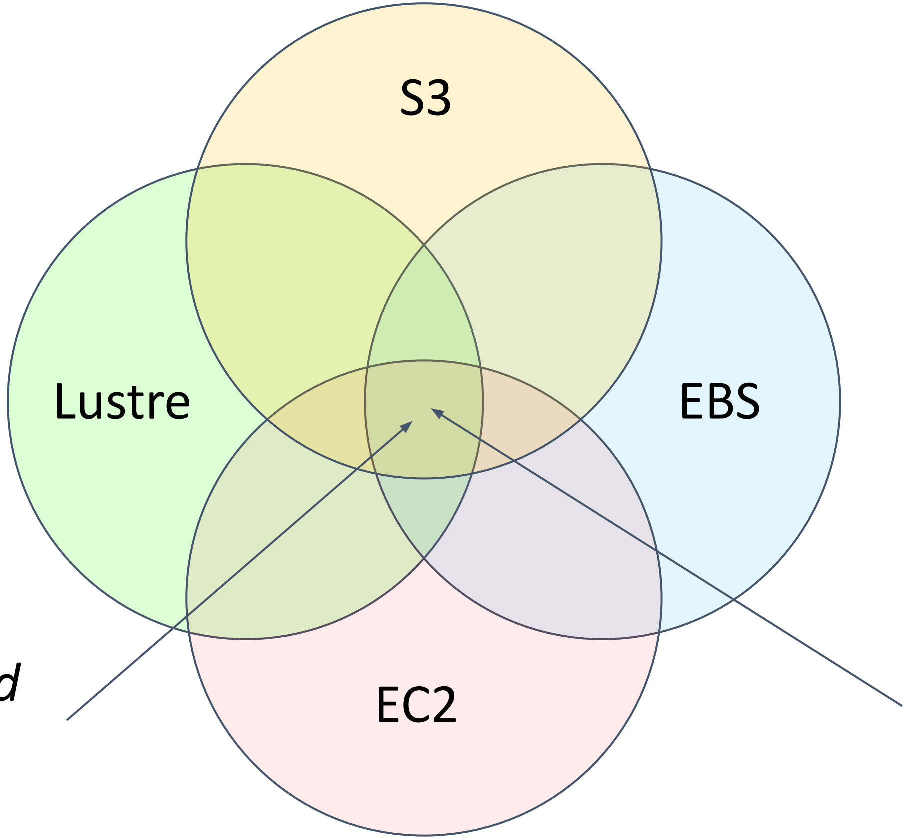
Open Source Parallel File System



Elastic Block Storage - provides reliable block device service for Lustre MDS & OSS

Provisionable Compute Server (MDS / OSS)

CSP IaaS + Existing PFS Solutions



Azure Managed Lustre (AMLFS)

AWS LustreFSx

What's Interesting about CSP Lustre Instances?

- Composable on-the-fly
- Integrated Archive / Lifecycle Mgmt
- Configurable Performance and Capacity
- H/A managed by the CSP

Azure Managed Lustre file system

CSP PFS: Reduces / Removes Inter-Job Interference

Poorly structured user workloads can degrade performance for all users

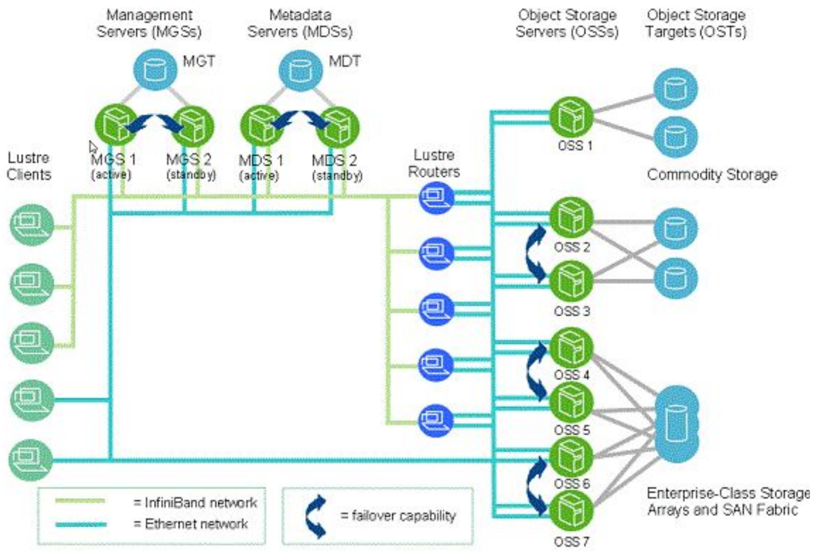
"***Users often setup a Slurm job script to ask for 2x the time they will need to run***" - *HPC R&D Staff Member at Top 5 HPC Site*

"***one of the most complex manifestations of performance variability on large scale parallel computers.***" - on parallel I/O contention

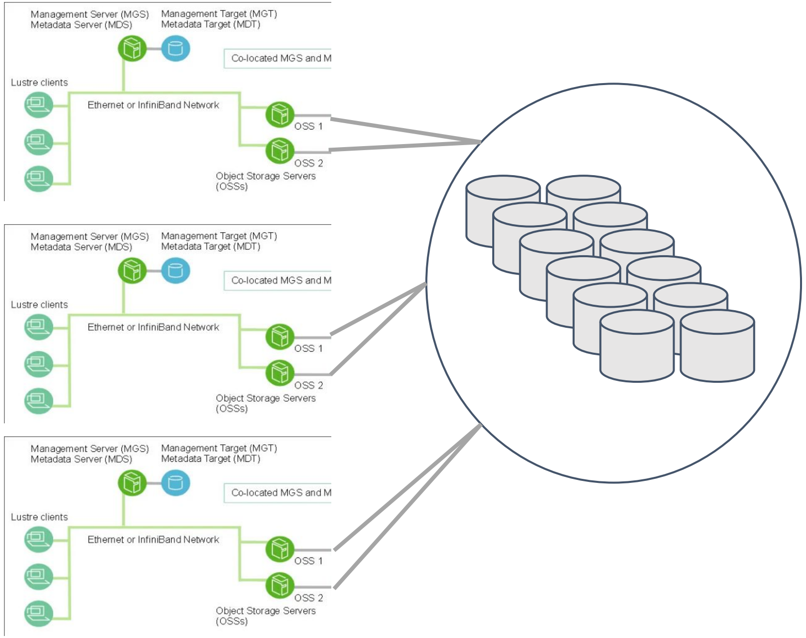
D. Skinner and W. Kramer, "Understanding the Causes of Performance Variability in HPC Workloads," in IEEE Workload Characterization Symposium, ***2005****, pp. 137–149.*

CSP PFS: Reduces / Removes Inter-Job Interference

How? Sharing is done at the block layer not the PFS



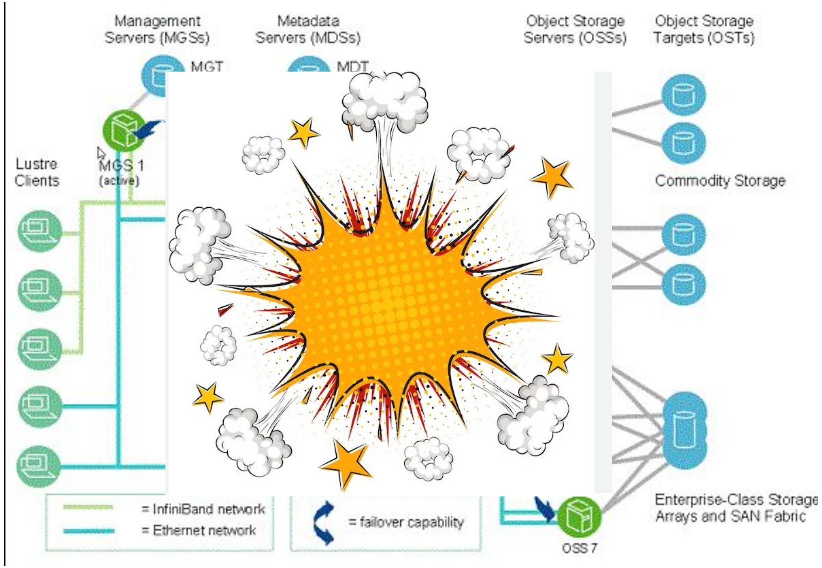
Typical HPC Config

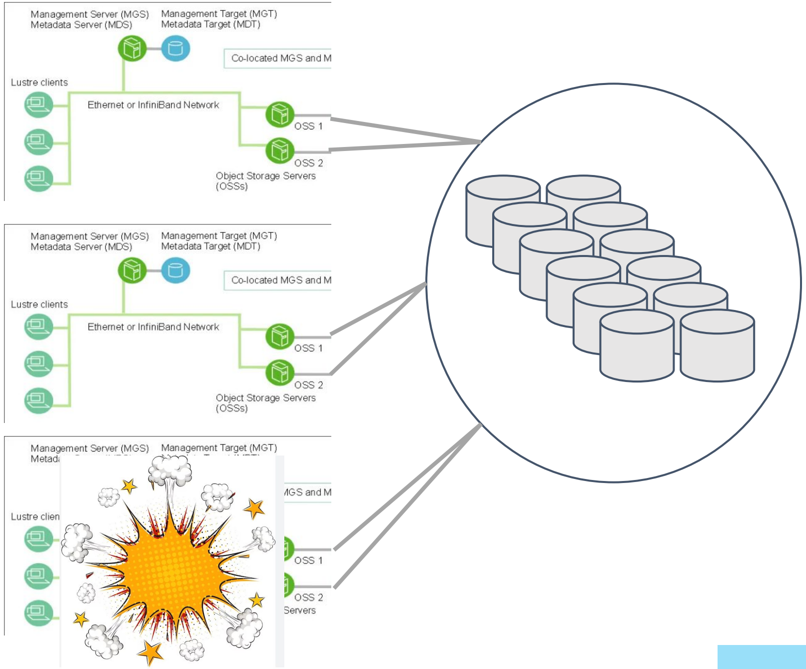


CSP Config

CSP PFS: Decreases Blast Radius

Caveat: Assumes unaffected Virtual Block Layer

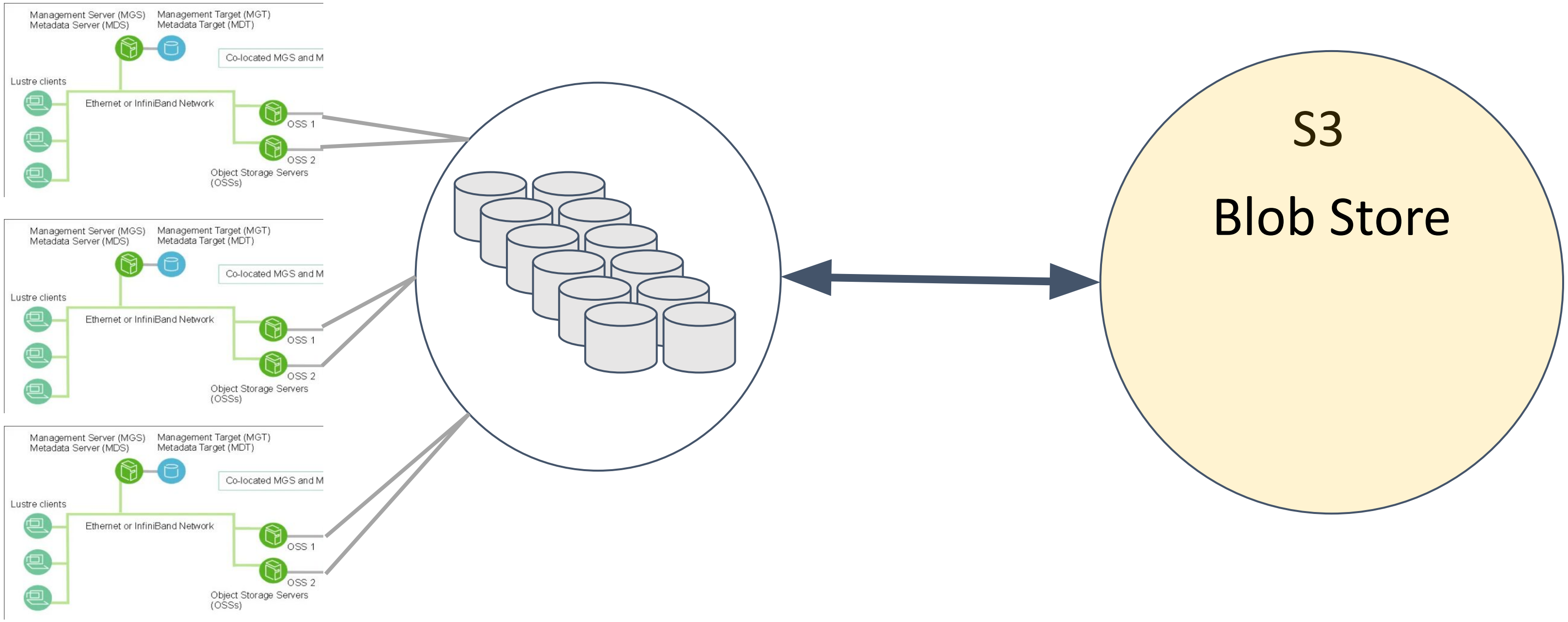




CSP PFS: Provisionable Performance and Capacity

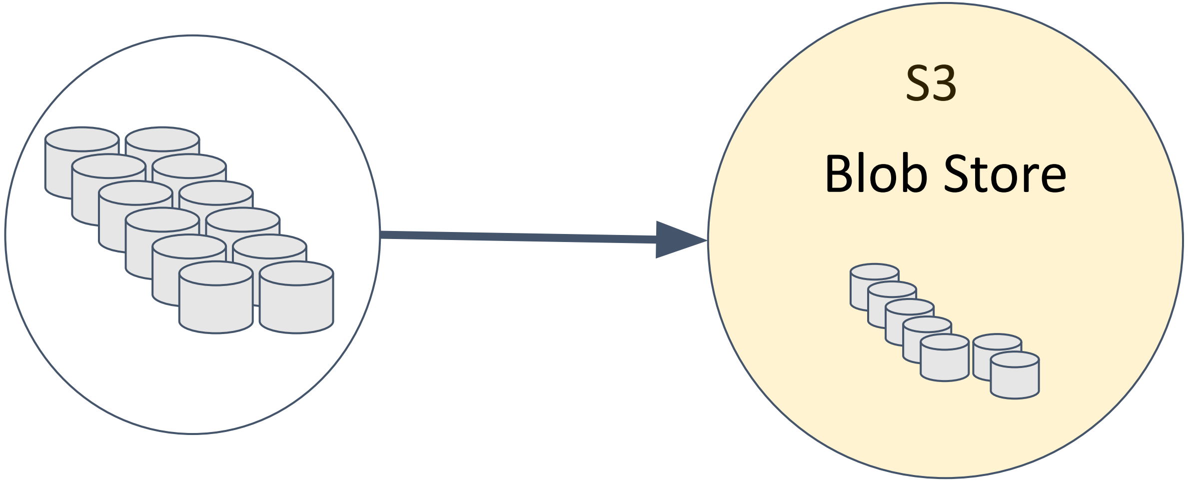
Even better.. IOPs and BW limits are enforceable at the Virtual Block Layer

CSP PFS: Transparent Archiving to Blob Store



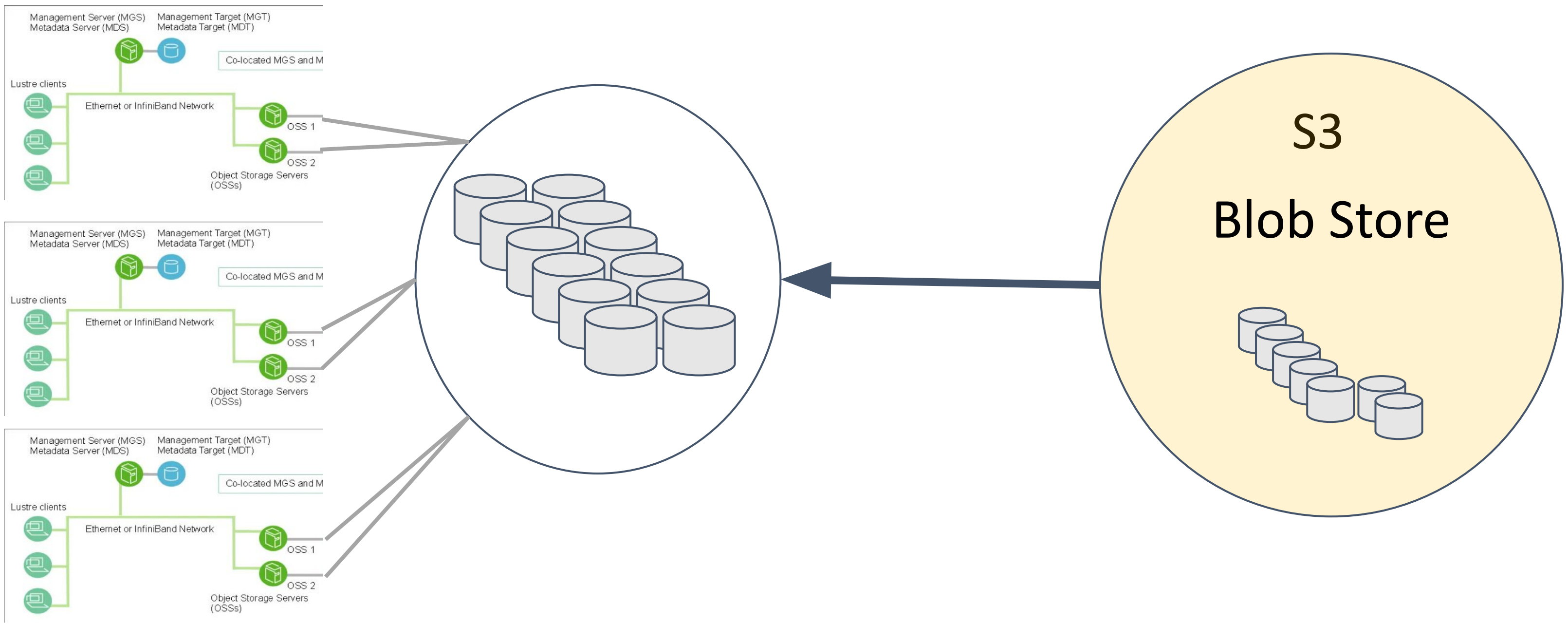
CSP PFS: Transparent Archiving to Blob Store

Instances can be torn down..



CSP PFS: Transparent Archiving to Blob Store

.. and rehydrated later

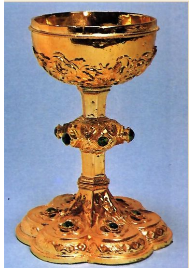


CSP PFS: How do they provide all these amazing things?

CSP Virtual Block Devices are Smart and Capable

- Snapshottable
  - Integration w/ Blob Store for low cost archiving
- Thin-Provisioned
  - *They don't charge that way, however*
- Network addressable
  - Follows the VM around the cluster
  - Reassignable via API
- Fault Tolerant
- Highly Available

So What's the Catch?

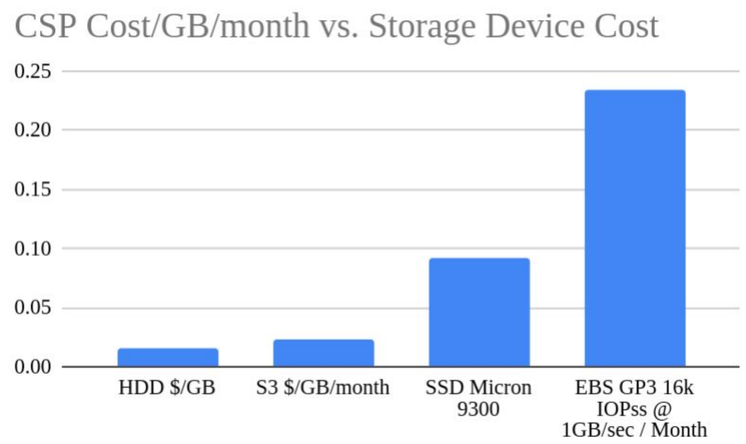


CSP managed disks and blob store are relatively expensive

- Especially viewed through HPC lens
- We have our own data centers!

HPC has employed RAID / Erasure Coding for decades

- CSP pricing implies replication



How can we get Erasure Coding: NVME over Fabric?

NVMEoF Lacks Important Capabilities

- ~~Snapshottable~~
  - ~~Integration w/ Blob Store for low cost archiving~~
- ~~Thin-Provisioned~~
- Network addressable
  - Follows the VM around the cluster
  - Reassignable via API
- Fault Tolerant
- Highly Available

Means static partitioning is required!

EC can be done via MD Raid

Ceph? 1TB/sec Study Reveals the Difficulty of Dist EC

Ceph offers thin provisioning but lacks performant EC

In practice triplication is used which increases system cost!

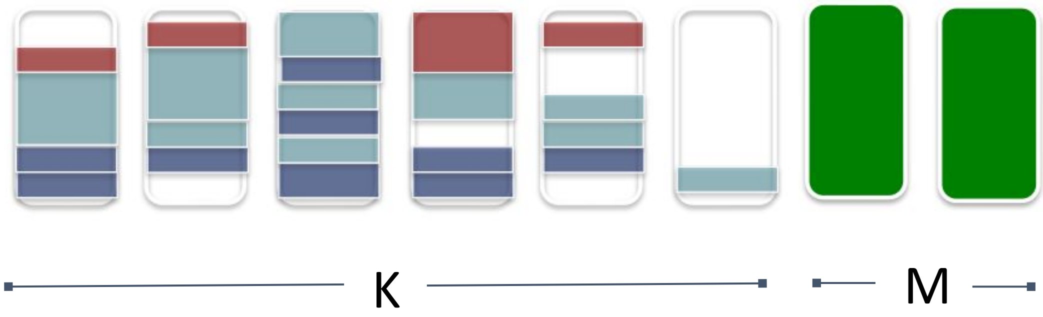
this is the config DigitalOcean operates..

https://ceph.io/en/news/blog/2024/ceph-a-journey-to-1tibps/

	630 OSDs (3x)	630 OSDs (EC62)
Co-Located Fio	Yes	Yes
4MB Read	1025 GiB/s	547 GiB/s
4MB Write	270 GiB/s	387 GiB/s
4KB Rand Read	25.5M IOPS	3.4M IOPS
4KB Rand Write	4.9M IOPS	936K IOPS

Existing Approach for “Efficient” Distributed EC

“Non deterministic” / unaffiliated EC sourcing has shown to be useful in removing read-modify-writes from the network EC storage path



Zest Checkpoint storage system for large supercomputers

December 2008

DOI: 10.1109/PDSW.2008.4811883

Source · IEEE Xplore

Conference: Petascale Data Storage Workshop, 2008. PDSW '08. 3rd

Paul Nowoczynski · Nathan Stone · Jared Yanovich · Jason Sommerfield

IME

silk

VAST

infinia

2018 IO500 IOR Hard

#	INFORMATION			IO500	IOR	
	SYSTEM	INSTITUTION	FILESYSTEM TYPE	SCORE	HARD WRITE	HARD READ
1	Data Accelerator	University of Cambridge	Lustre	158.71	7.44	46.78
2	Oakforest-PACS	JCAHPC	IME	137.78	692.74	287.09
3	ShaheenII	KAUST	DataWarp	77.37	139.59	392.93
4	Data Accelerator	University of Cambridge	BeeGFS	74.58	7.00	27.86
5	Oakforest-PACS	JCAHPC	Lustre	42.18	2.36	6.95
6	ShaheenII	KAUST	Lustre	41.00	1.44	81.38
7	JURON	JSC	BeeGFS	35.77	1.46	19.16

IME used EC in this configuration, DataWarp did not!

2018 IO500 IOR Hard

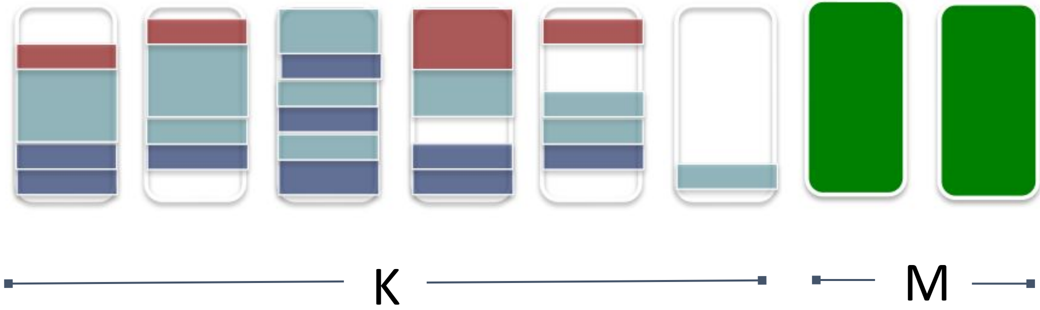
With Erasure Coding!

IOR	
HARD WRITE	HARD READ
7.44	46.78
692.74	287.09
139.59	392.93
7.00	27.86
2.36	6.95

Existing Approach for “Efficient” Distributed EC

“Non determinis… …ated EC sourcing has shown to be use… …ing read-modify-writes from the netwo… …e path

GC Method requires stateful tracking of individual extents which is expensive and difficult to implement. GC performance may be poor in cases..



non-deterministic EC sourcing method

.. the resulting garbage collection

Approach for Simplifying Distributed EC for Block

Method for Efficient Erasure Coded Group Management in shared Nothing Storage Clusters

Nowoczynski; Paul Joseph

uspto.report › / patents › / Nowoczynski; Paul Joseph › / Patent 17/105286 / Applicant

uspto.report › / patents › / Nowoczynski; Paul Joseph › / Patent 17/105286 / Inventors

Patent Application Summary

U.S. patent application number 17/105286 was filed with the patent office on 2021-05-27 for *method for efficient erasure coded group management in shared nothing storage clusters*. The applicant listed for this patent is Paul Joseph Nowoczynski. Invention is credited to Paul Joseph Nowoczynski.

Approach for Simplifying Distributed EC for Block

Abstract

A method that achieves high availability by employing distributed erasure coding instead of distributed replication and preserves and applies the positive attributes of distributed replication to that of distributed erasure coding. The results are improvements and simplifications to the otherwise difficult internal management processes found in distributed, shared-nothing, erasure coding systems. ***The key positive attributes of the distributed replication method are processing of a user's write request without requiring the presence of some set of adjacent blocks (ie a read-modify-write) and the ability of storage endpoints to perform garbage collection tasks with complete autonomy of one another.*** *The distributed block storage system simultaneously captures the capacity advantages of erasure coding and the positive attributes of fault tolerance management found in data replication.*

Approach for Simplifying Distributed EC for Block

Abstract

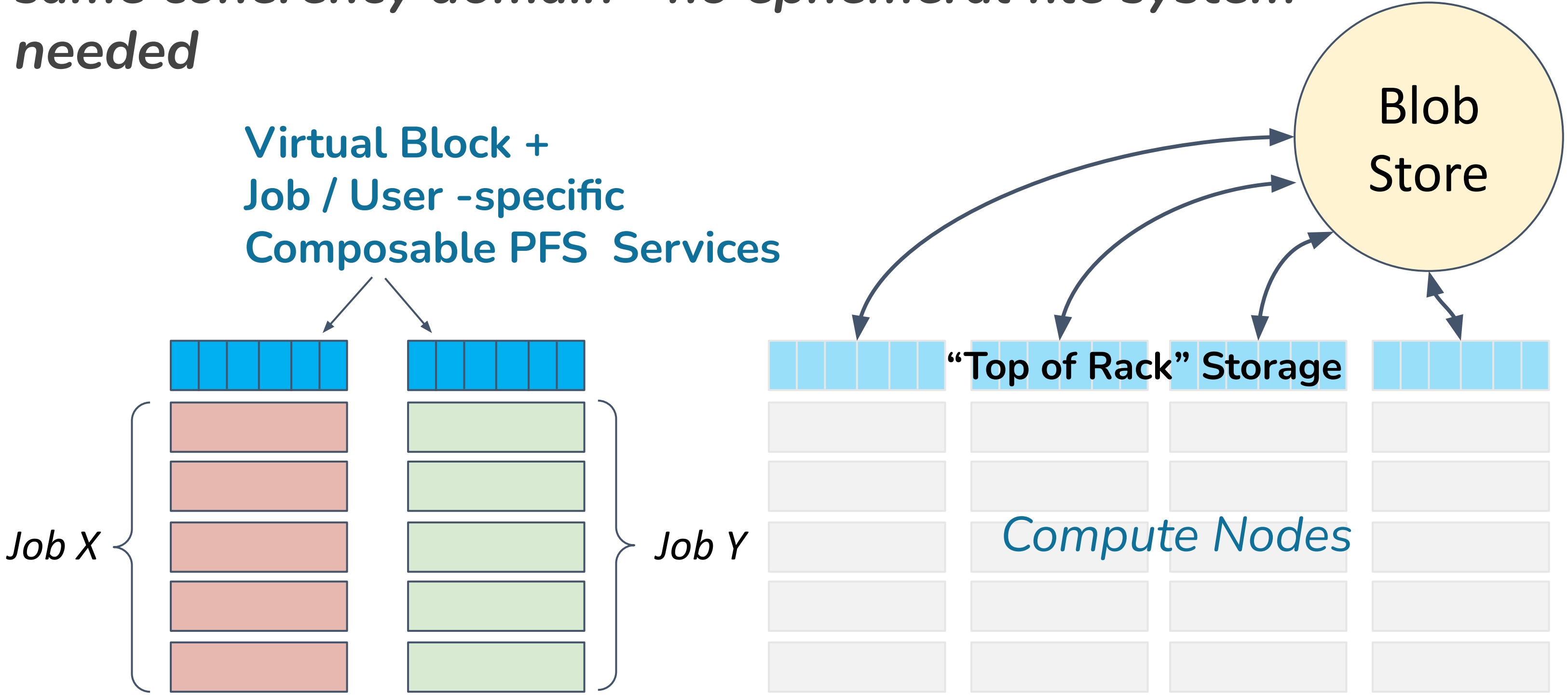
A method that achieves high availability by employing distributed erasure coding instead of distributed replication and preserves and applies the positive attributes of distributed replication to that of distributed erasure coding. The results are improvements and simplifications to the otherwise difficult internal management processes found in distributed, shared-nothing, erasure coding systems. The key positive attributes of the distributed replication method are processing of a user's write request without requiring the presence of some set of adjacent blocks (ie a read-modify-write) and the ability of storage endpoints to perform garbage collection tasks with complete autonomy of one another. ***The distributed block storage system simultaneously captures the capacity advantages of erasure coding and the positive attributes of fault tolerance management found in data replication.***

Moving Beyond the CSPs

If distributed block + efficient erasure coding are in reach what are the possibilities?

Moving Beyond the CSPs: *Transparent Locality Mgmt*

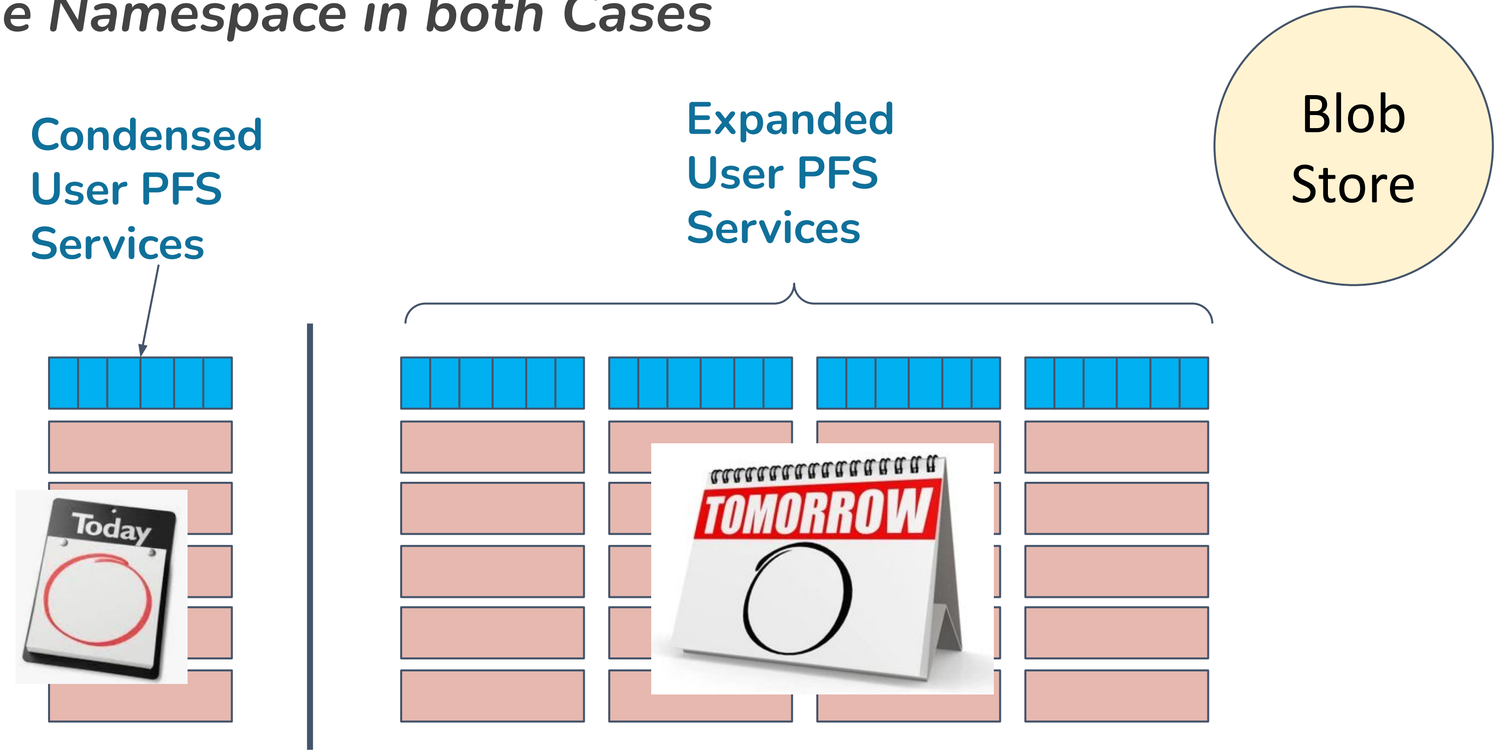
Data migration at the block level can be done within the same coherency domain - no ephemeral file system needed



Moving Beyond the CSPs: *Adaptable PFS Service Scaling*

Adjusts to Users' Job Size

Same Namespace in both Cases



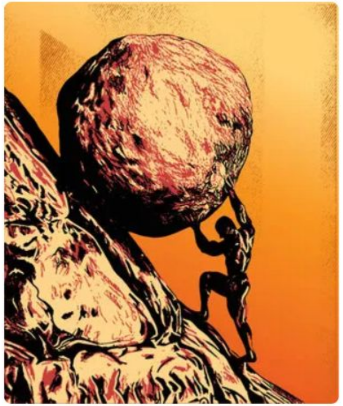
Moving Beyond the CSPs: *Enabling New PFS Tech*

IO500 Production List is a Full of the Known Players

These systems have taken millions of man hours to build..

Why? Recovery and Fault Tolerance are very difficult to implement

# ↑	BOF	INSTITUTION	SYSTEM	STORAGE VENDOR	FILE SYSTEM TYPE	CLIENT NODES	TOTAL CLIENT PROC.	SCORE ↑	BW (GIB/S)	MD (KIOP/S)	REPRO.
1	SC23	Argonne National Laboratory	Aurora	Intel	DAOS	300	62,400	32,165.90	10,066.09	102,785.41	✓
2	SC23	LRZ	SuperMUC-NG-Phase2-EC	Lenovo	DAOS	90	6,480	2,508.85	742.90	8,472.60	✓
3	SC23	King Abdullah University of Science and Technology	Shaheen III	HPE	Lustre	2,080	16,640	797.04	709.52	895.35	✓
4	ISC23	EuroHPC-CINECA	Leonardo	DDN	EXAScaler	2,000	16,000	648.96	807.12	521.79	✓
5	ISC24	Zuse Institute Berlin	Lise	Megware	DAOS	10	960	324.54	65.01	1,620.13	✓
6	SC23	Memorial Sloan Kettering Cancer Center	IRIS	WekaIO	WekaIO	36	4,248	308.94	104.79	910.80	✓
7	ISC22	China Telecom Research Institute	CTPAI	CTCLOUD	DAOS	10	200	187.84	25.29	1,395.01	-
8	ISC24	NHN Cloud Corporation	NHN CLOUD GWANGJU AI	DDN	EXAScaler	10	640	176.57	62.58	498.22	✓
9	ISC24	ACC Cyfronet AGH	Helios	HPE	Lustre	80	640	153.39	122.31	192.36	✓
10	ISC23	Imperial College London	Imperial - hx cluster	Lenovo	Spectrum scale	32	512	119.56	44.63	320.31	✓



Moving Beyond the CSPs: *Enabling New PFS Tech*

With Smart & Capable Virtual Block Devices current high performance ephemeral PFS tech could be brought closer to Production!

CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory

Osamu Tatebe, University of Tsukuba, Japan, tatebe@cs.tsukuba.ac.jp
Kazuki Obata, University of Tsukuba, Japan, obata@hpcs.cs.tsukuba.ac.jp
Kohei Hiraga, University of Tsukuba, Japan, hiraga@ccs.tsukuba.ac.jp
Hiroki Ohtsuji, Fujitsu Research, Fujitsu Limited, Japan, ohtsuji.hiroki@fujitsu.com

DOI: https://doi.org/10.1145/3492805.3492807

CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory

Osamu Tatebe, University of Tsukuba, Japan, tatebe@cs.tsukuba.ac.jp
Kazuki Obata, University of Tsukuba, Japan, obata@hpcs.cs.tsukuba.ac.jp
Kohei Hiraga, University of Tsukuba, Japan, hiraga@ccs.tsukuba.ac.jp
Hiroki Ohtsuji, Fujitsu Research, Fujitsu Limited, Japan, ohtsuji.hiroki@fujitsu.com

DOI: https://doi.org/10.1145/3492805.3492807

SPs: *Enabling New PFS Tech*

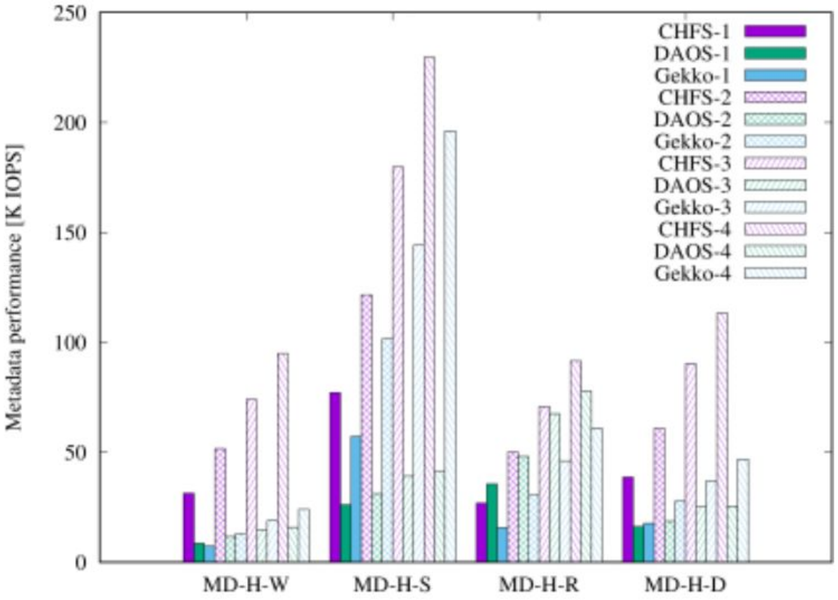
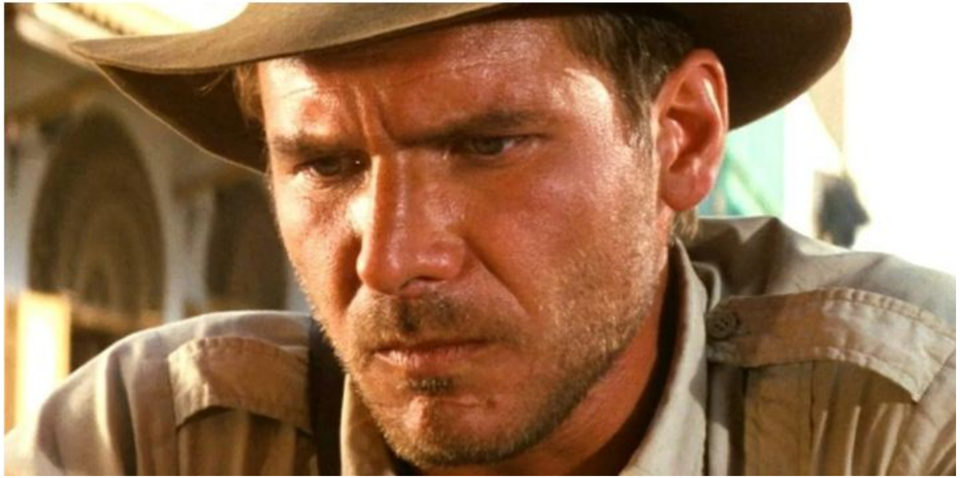


Figure 9: IO500 metadata performance of CHFS, DAOS, and GekkoFS in the hard case. MD-H-W, MD-H-S, MD-H-R and MD-H-D denote MDtest hard write, stat, read, and delete, respectively. CHFS displays the best and scalable

Have CSPs Led us to the Holy Grail?



.. unsure, TBH – it's complicated :)

Thank You!