Storage Research in an Open Cloud

Orran Krieger, Peter Desnoyers,
Storage at Scale lab & Collaborators
https://massopen.cloud/sas-group-team/

What I will talk about

- Approach to research and a few lessons learned
- Our efforts to create an open cloud, and its status
- The storage research directions driven by the MOC
- Some of the long term implications for systems and storage

Experience...

Platform	Hurricane OS
Compute & Networking Hardware	Hector Ethernet
Storage	Hard Drive NFS
Application Requirements	Performance Durability Recovery

Exploiting the advantages of mapped files for stream I/O, Krieger, et al, Usenix'92
HFS: A flexible file system for large-scale multiprocessors, Krieger, et al, DAGS'93
The Alloc Stream Facility: A redesign of application-level stream I/O, Krieger, et al, IEEE Computer'94
HFS: a performance-oriented flexible file system based on building-block compositions, Krieger, et al, IOPADS'96 & Trans on Computer Systems '97
Automatic Compiler-Inserted I/O Prefetching for Out-of-Core Applications, Mowry, et al, OSDI'96
Compiler-Based I/O Prefetching for Out-of-Core Applications, Demke Brown, et al, ACM Trans. on Computer Systems'01

Experience...

- HFS led to Tornado and then K42 - U of Toronto
- K42 led to Linux, Cell/PS3 and then virtualization (rHype, sHype, Xen, PHYP) - IBM
- Virtualization led to focus on unikernels (Libra, EbbRT, Seuss, UKL...) and the open cloud (vCloud Director, MOC) - IBM, VMware, MOC

None of these where “storage projects”, but storage was always a major element of a larger system.

Research philosophy/lessons

- Hypothesis of a radical change (e.g., 64 bit NUMA MP)
- Complete system; visibility into applications & technologies, and ability to work across layers
- Research based on real application demands; don't worry about innovation:
  - Start with something simple and evolve,
  - if problem is tough, research will happen to solve it
  - if your system is different, you will have novel insights
- Even if radical change takes time:
  - if hypothesis is eventually true, long term work will have an impact.
  - you will solve real problems, and a community will develop to enable radical change

Pulling together a small team with a shared vision leads to magic.

Hypothesis/vision: Open Cloud

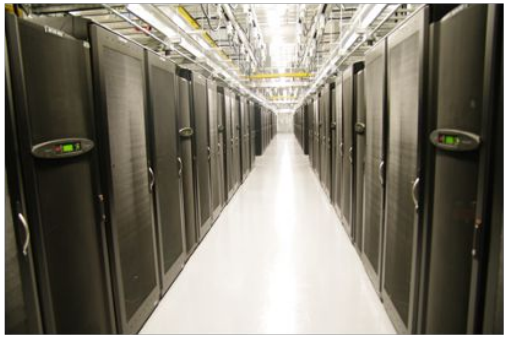
- All compute will move to cloud:
  - on demand access
  - economies of scale
  - massive number of services
- Today's clouds are black boxes, with a single company responsible for implementing and operating the cloud.
- An open model that enables competition/innovation by a broad community will eventually win.

Kittyhawk: Enabling cooperation and competition in a global, shared computational system, Appavoo et al. IBM Systems Journal '09
Enabling a marketplace of clouds: VMware's vCloud Director, Krieger, et al., OSR`10

In 2013 MGHPCC opened

Opened in 2013
More than 20,000 users
Hundreds of thousands of CPUs
150+ Petabytes of storage
2 Terabits/second network bandwidth
~1Terabit lit / ~ 10 providers

- Room for 1012 cabinets
- 12MW for compute; expandable to 24MW
- Land for expansion
- Carbon free energy
  - Local hydro and solar power



MIT Northeastern BOSTON UNIVERSITY UMASS HARVARD UNIVERSITY

The Mass Open Cloud (MOC)

We decided to build an open cloud

- servicing research users
- based on open source; enabling system research
- all the information available to open source and system research community

Describe, with a focus on storage:

- Real world problems we got hit with driven by real demands
- The simple first solutions and the research and insights that happened
- The communities/impact that has resulted

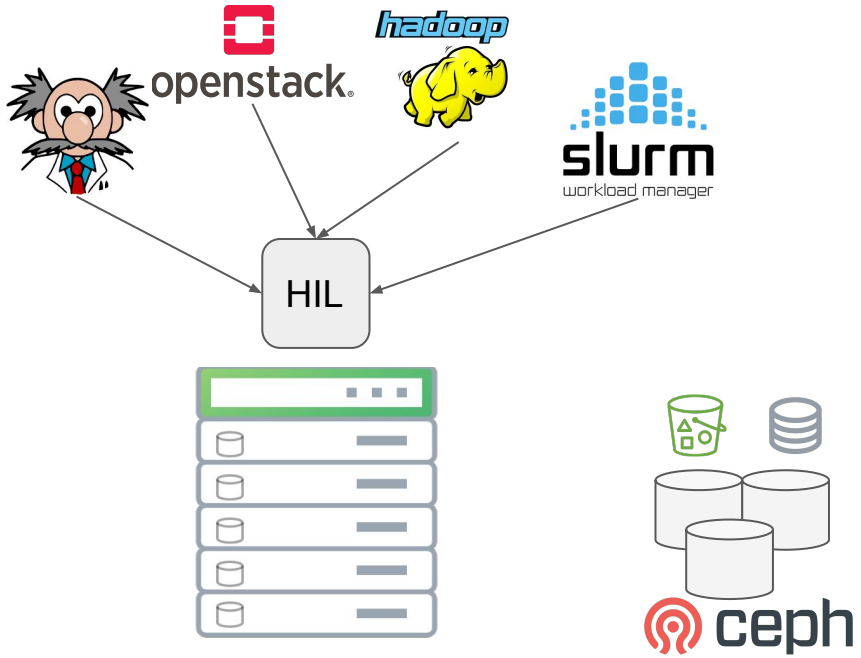
The Mass Open Cloud (MOC)

Toward an Open Cloud Marketplace: Vision and First Steps, Bestavros, et al., IEEE, IC'14
Using OpenStack for an Open Cloud eXchange (OCX), Desnoyers, et al., IC2E'15

MGHPCC

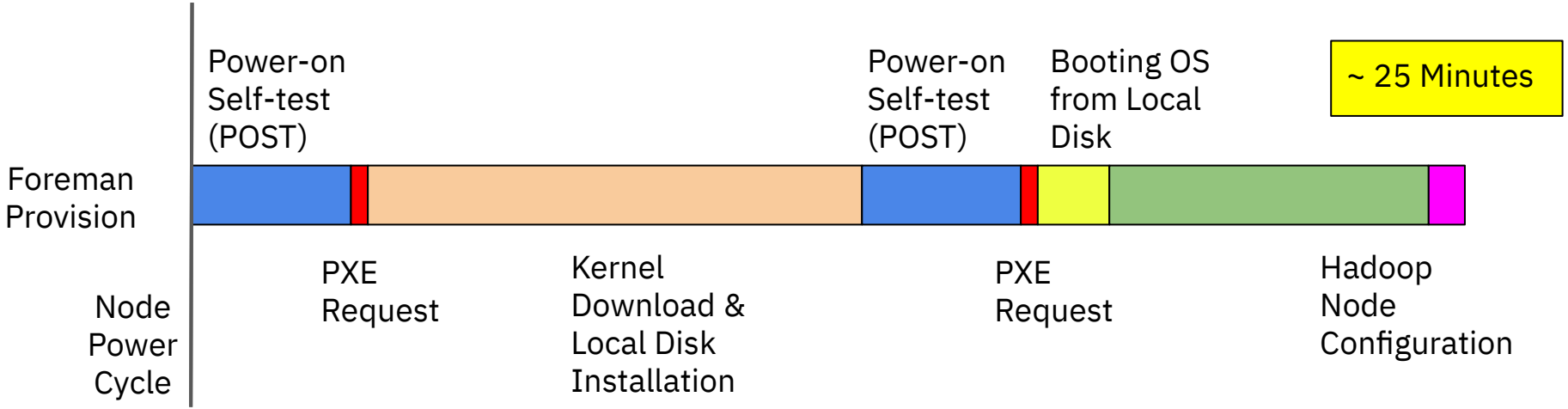
Problem want to support wide variety of platforms

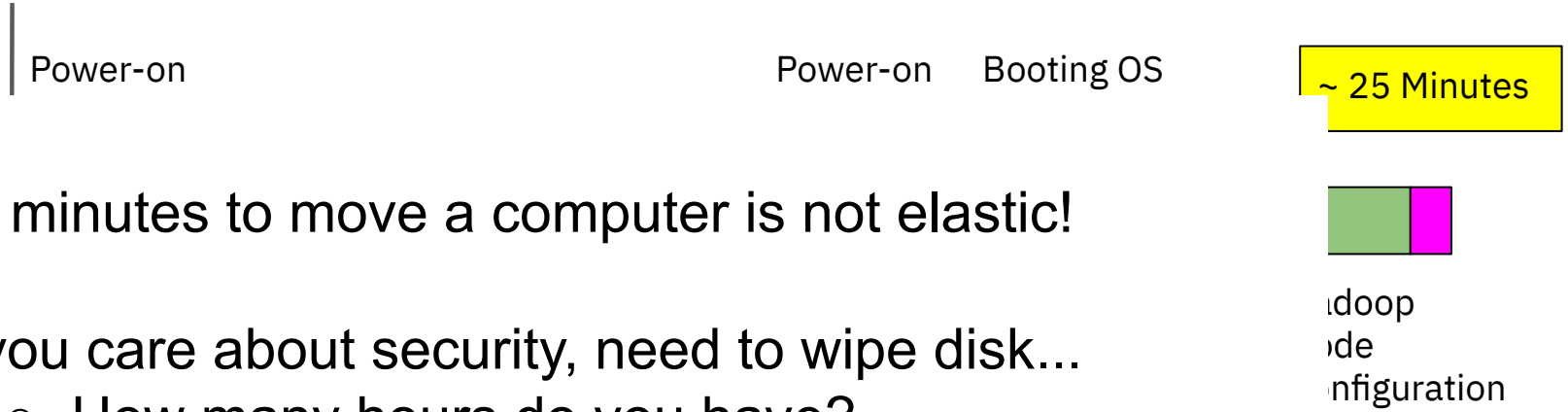
Hardware Isolation Layer (HIL): first simple system



HIL: Designing an Exokernel for the Data Center, Hennessey, et al., SoCC '16

Problem: moving computers is slow...

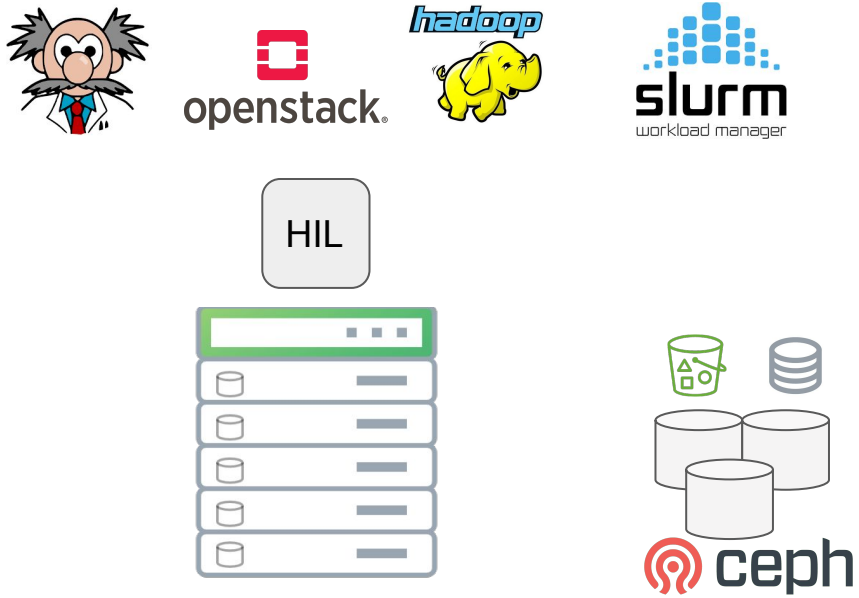



Power-on
Power-on
Booting OS
~ 25 Minutes
For
Pro
25 minutes to move a computer is not elastic!
If you care about security, need to wipe disk...
How many hours do you have?
How many passes?
Trust provider and everyone pays?
Trust tenant - covert channel?
doop
de
nfiguration

Insight: need to enable user control

Different systems/users have different:

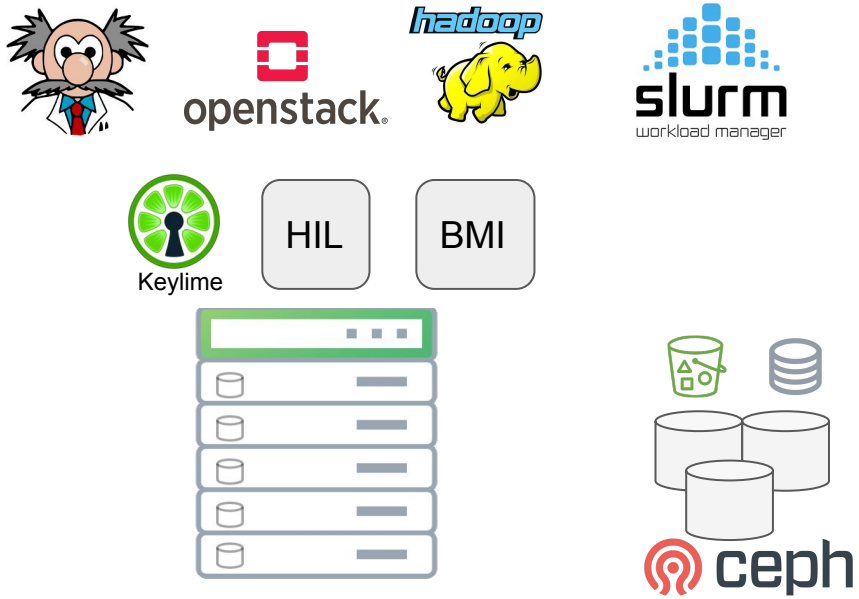
- provisioning systems
- security requirements
- compliance and operational concerns



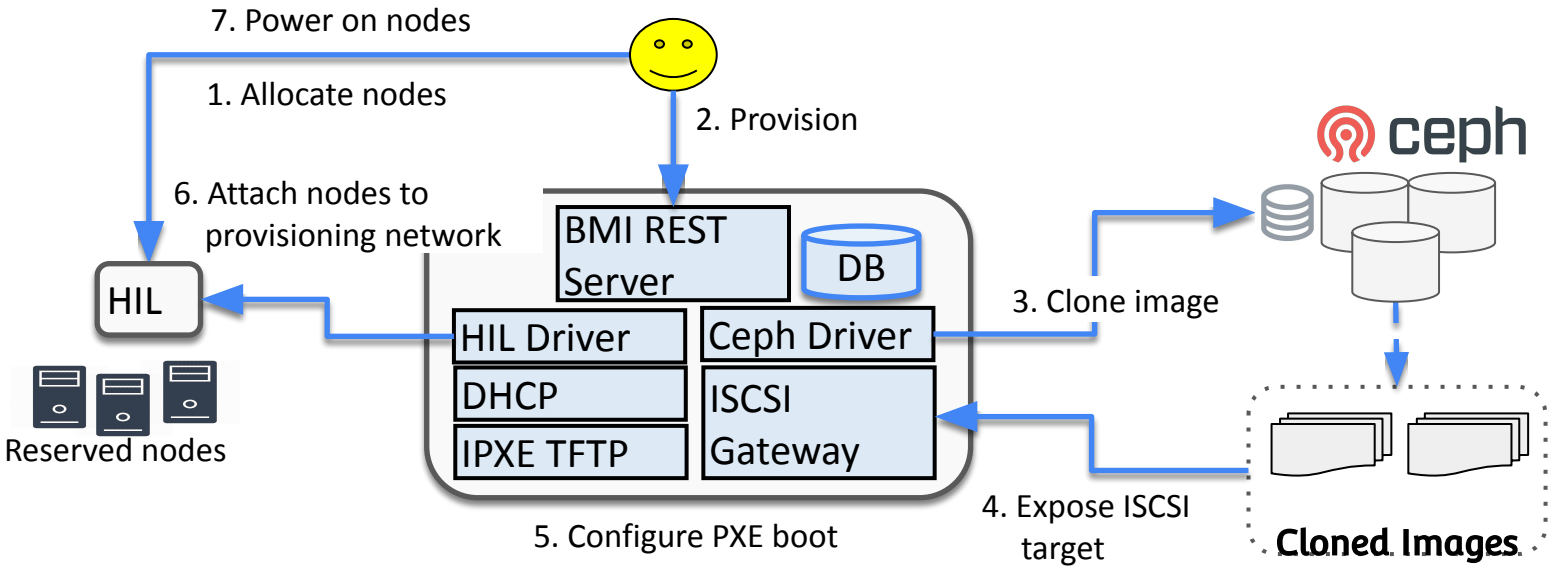
Tenant controlled security & provisioning

Integrate with HIL capability for tenants to:

- Encrypt volume storage
- Keylime: attest firmware not compromised before distributing keys
- Bare Metal Imaging: stateless provisioning

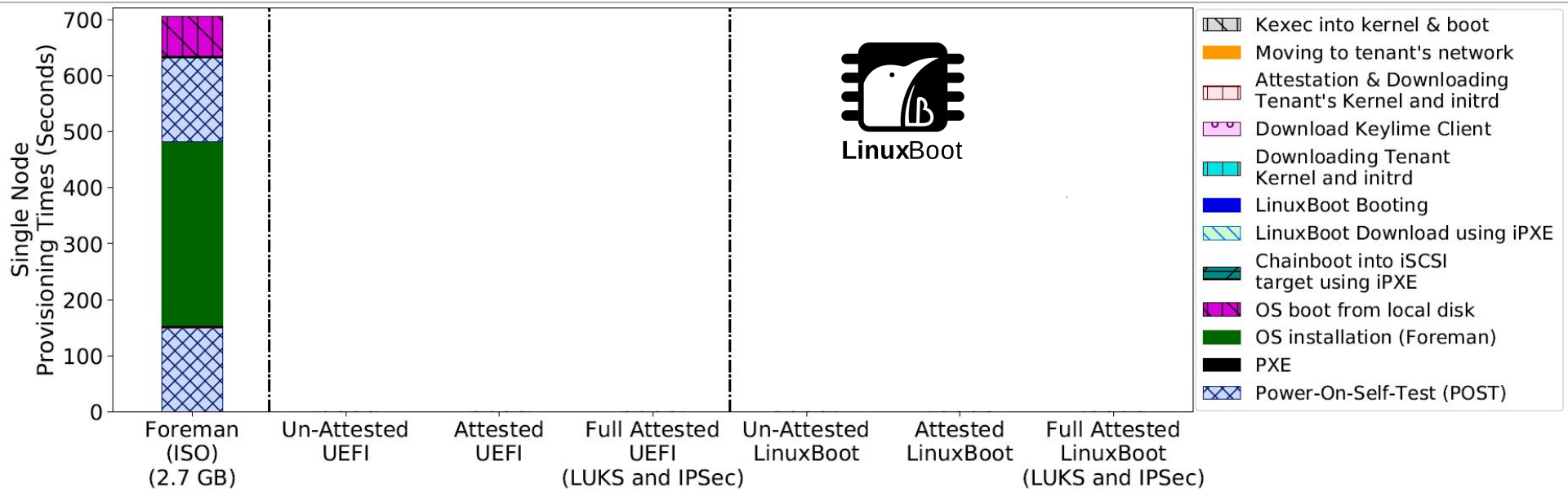


Bare Metal Imaging (BMI): first simple system

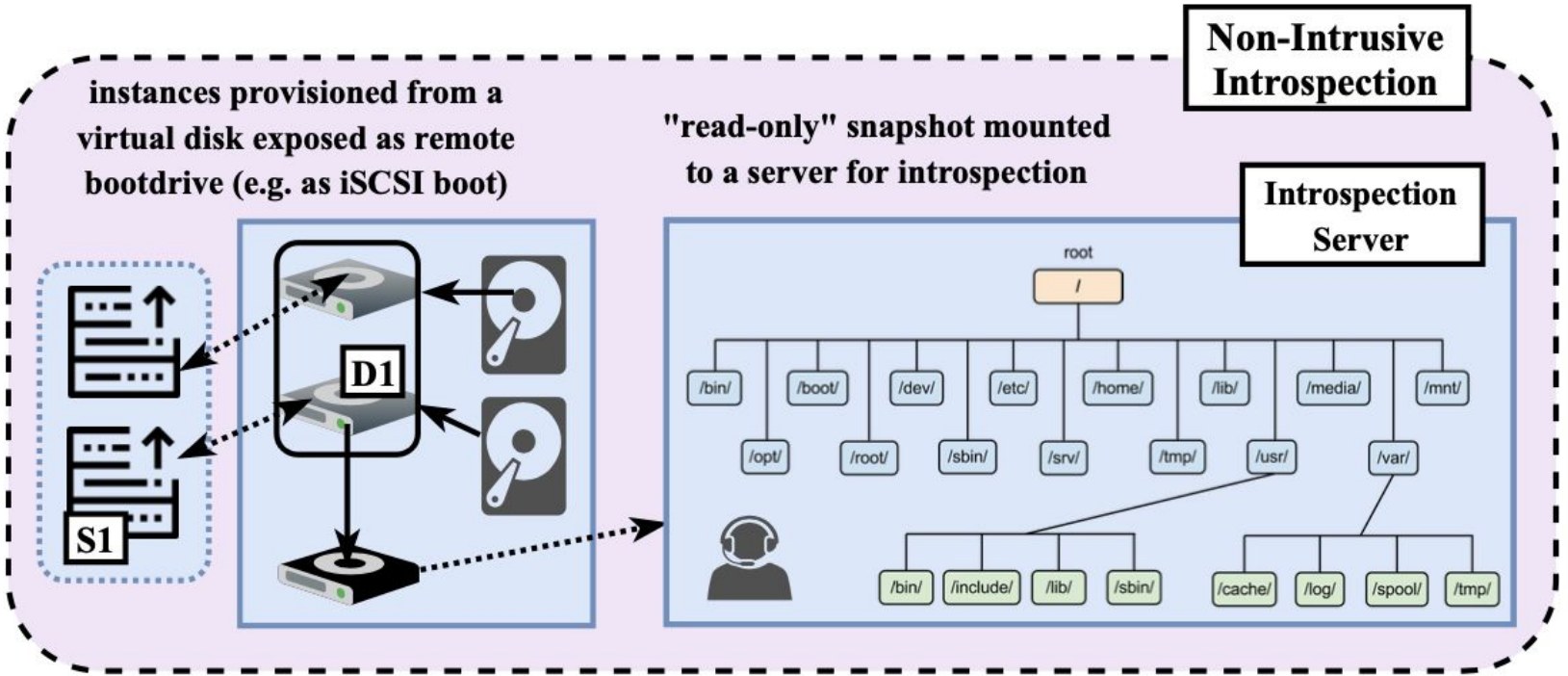


Boot Time

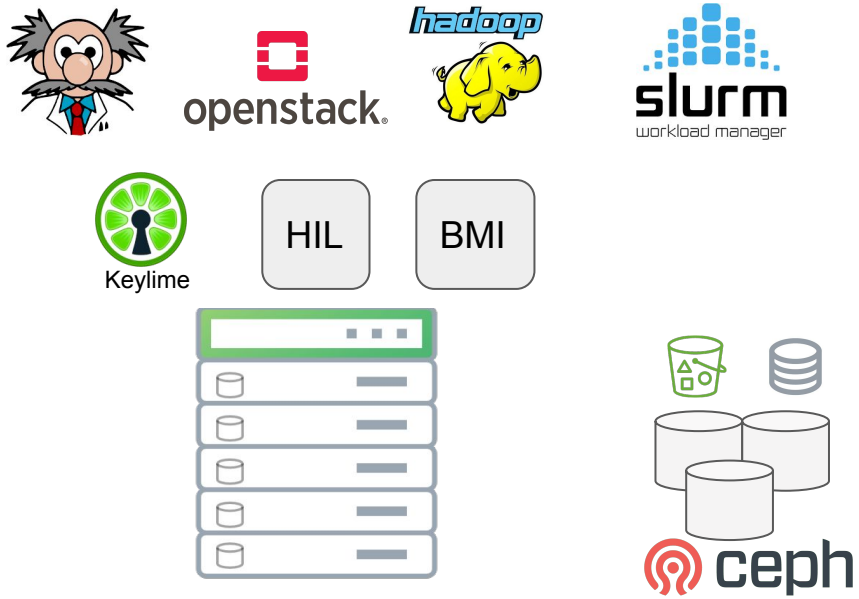
- Dell R630 server
  - 2 Xeon E5-2660 v3 2.6 GHz
  - 256 GB RAM

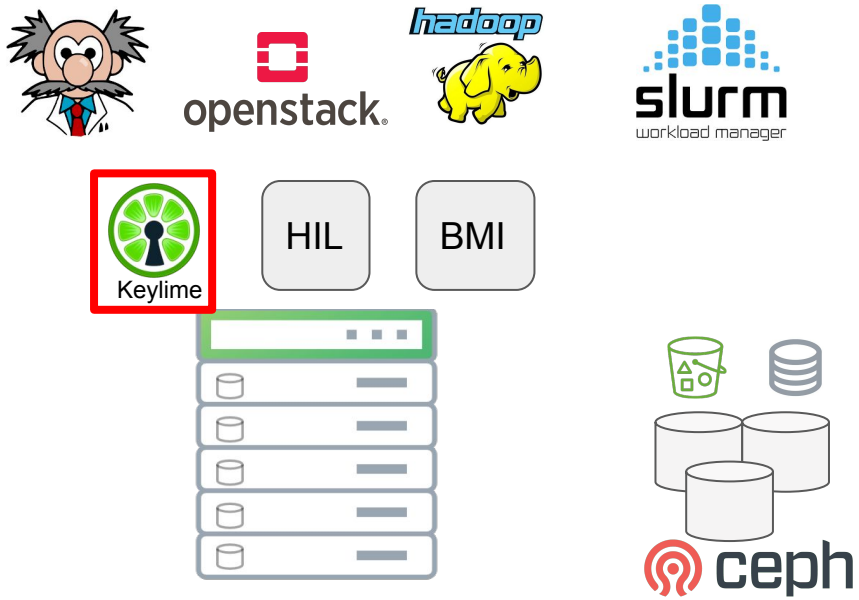


This also means



M2: Malleable Metal as a Service, Mohan, et al, IC2E'17
A Secure Cloud with Minimal Provider Trust, Mosayyebzadeh, et al, HotCloud'18
Supporting security sensitive tenants in a bare-metal cloud, Mosayyebzadeh, et al, ATC'19
Towards Non-Intrusive Software Introspection and Beyond, Mohan, et al, IC2E'20




openstack.
hadoop
slurm
workload manager
Keylime
HIL
BMI
ceph


openstack.

hadoop


slurm
workload manager


ESI


ceph

The Open Cloud Testbed (OCT): A platform for research into new cloud technologies, Zink, et al, CloudNet'21

The Dataverse® Project

OPENSHIFT

Red Hat OpenShift Data Science

APACHE Spark™

Open Cloud Testbed
Exploring Next-Generation Cloud Platforms

openstack.

hadoop

slurm
workload manager

ESI

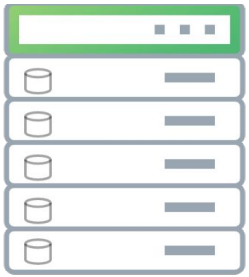
ceph

The Dataverse® Project

Red Hat OpenShift Data Science

APACHE Spark™

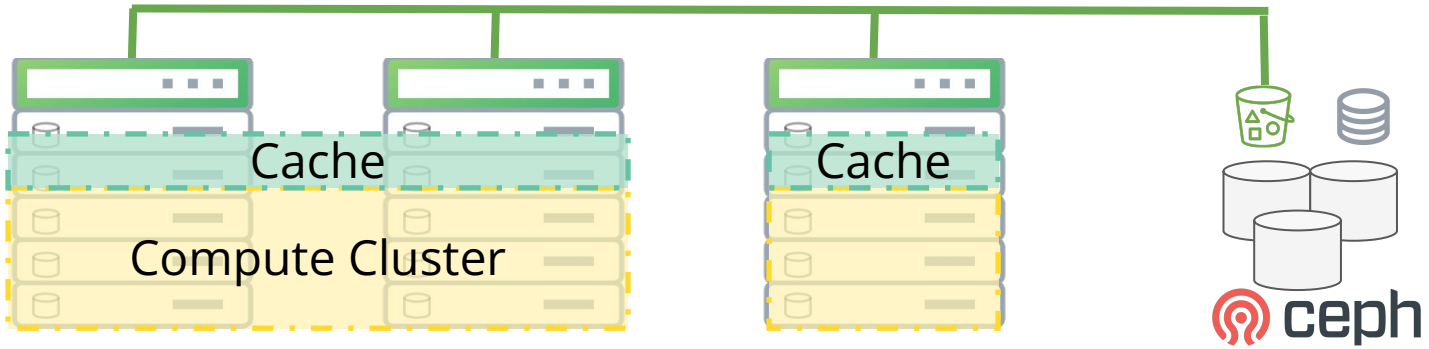
ESI



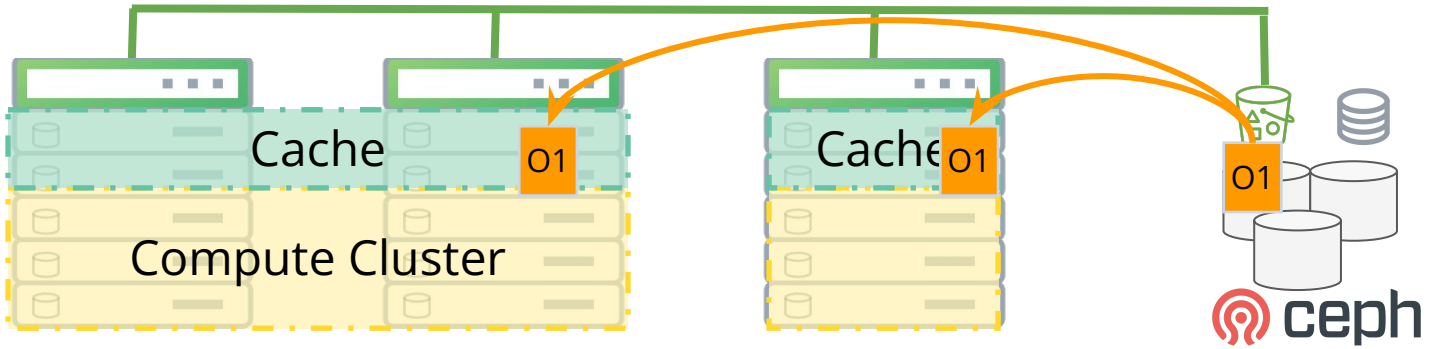
Problem: slow access to research data

- Limited bi-sectional bandwidth
- High capacity shared storage

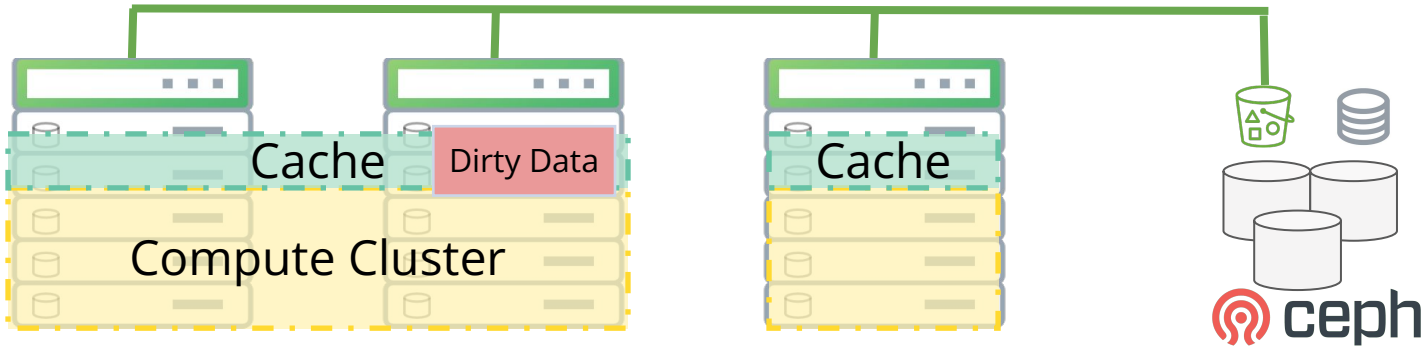
Normal solution using a Cluster/Framework Cache

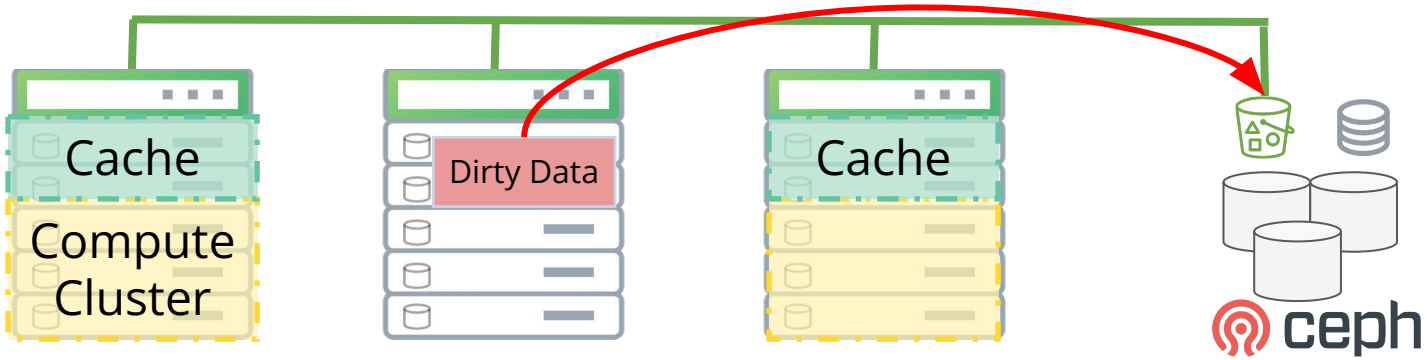


Problem: many of our users access the same data from different clusters and frameworks



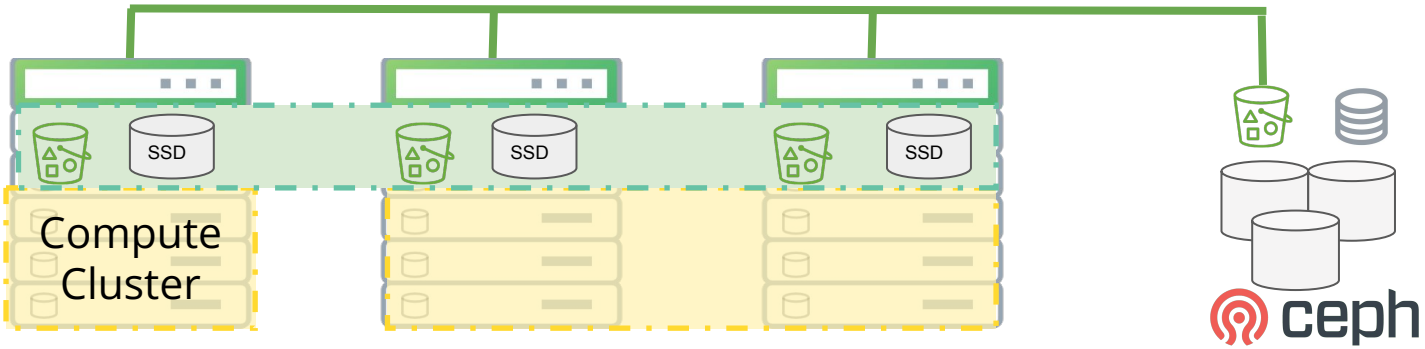
Problem: we want to use ESI to move machines around

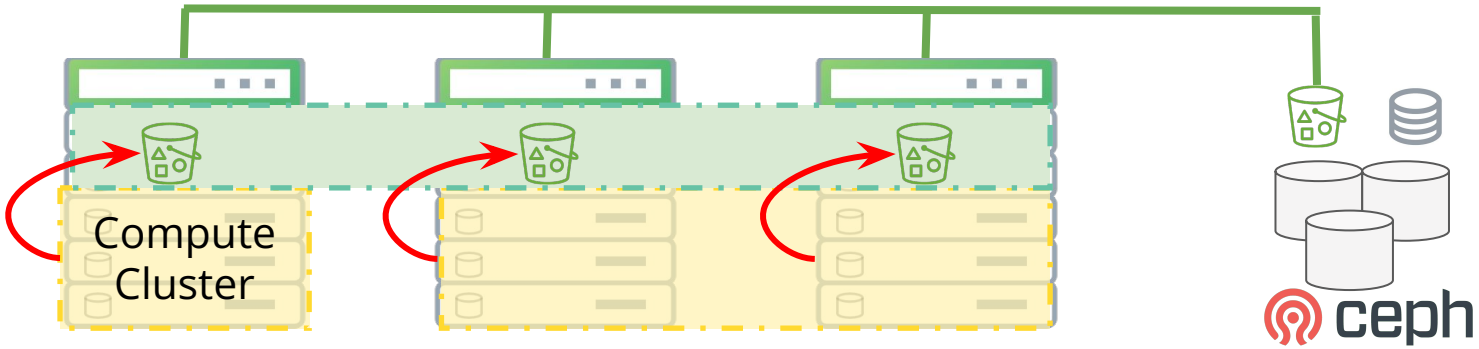


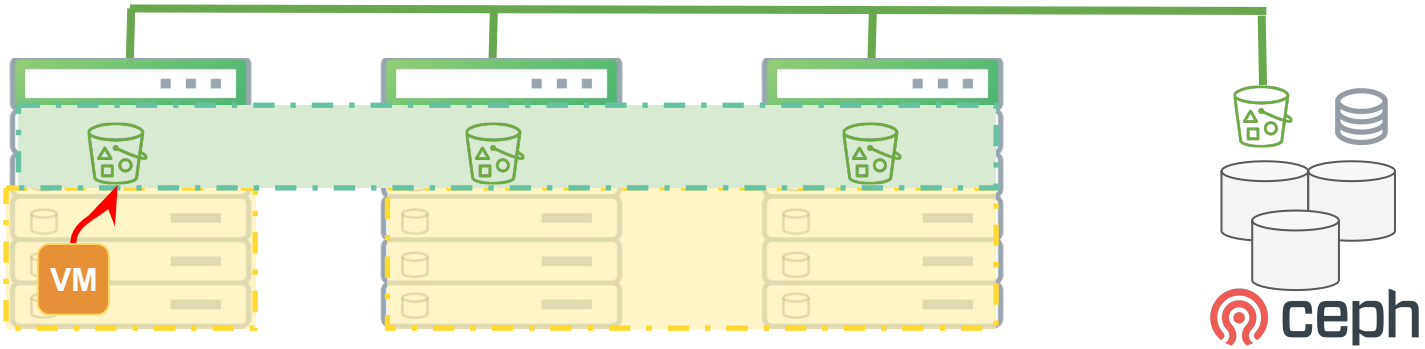

Cache
Compute
Cluster
Dirty Data
Cache
ceph

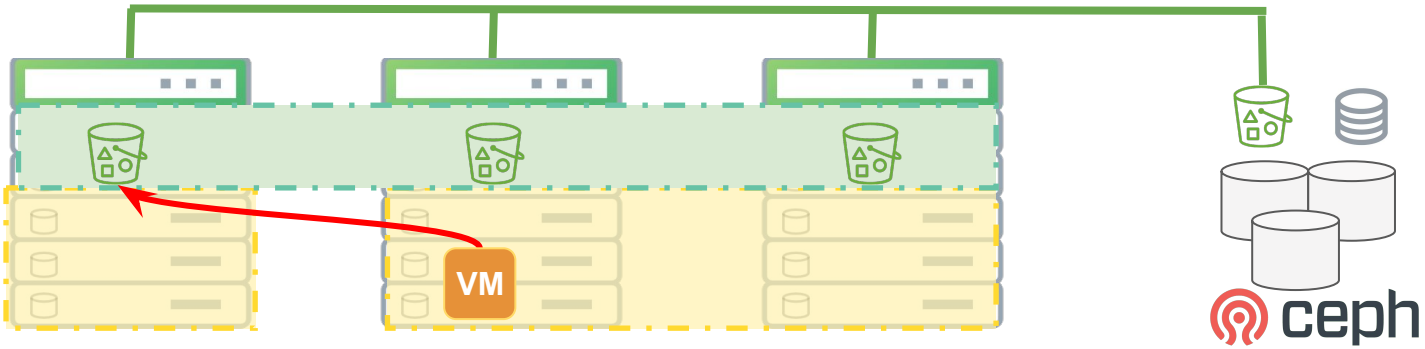
Cooperative cache extending the datalake:

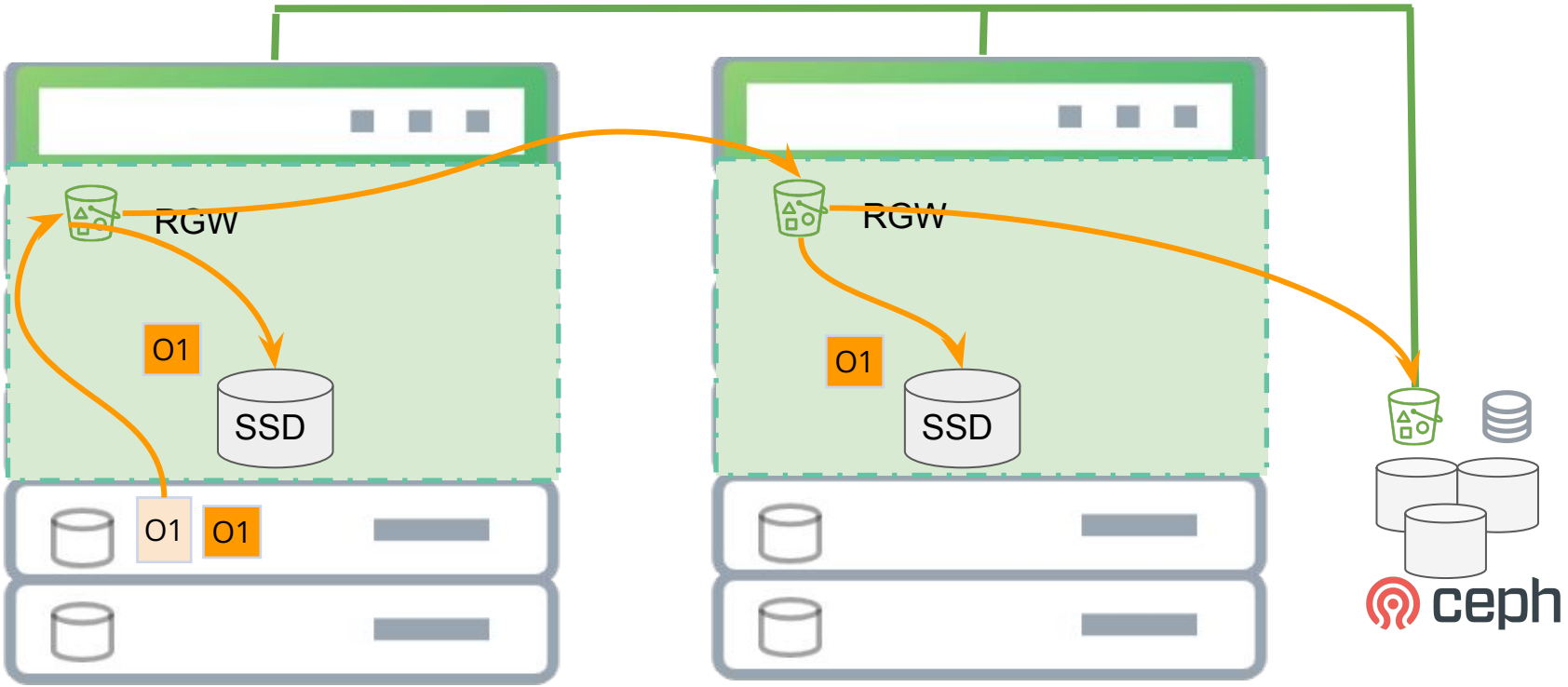
First simple D3N - DataCenter Data Delivery Network

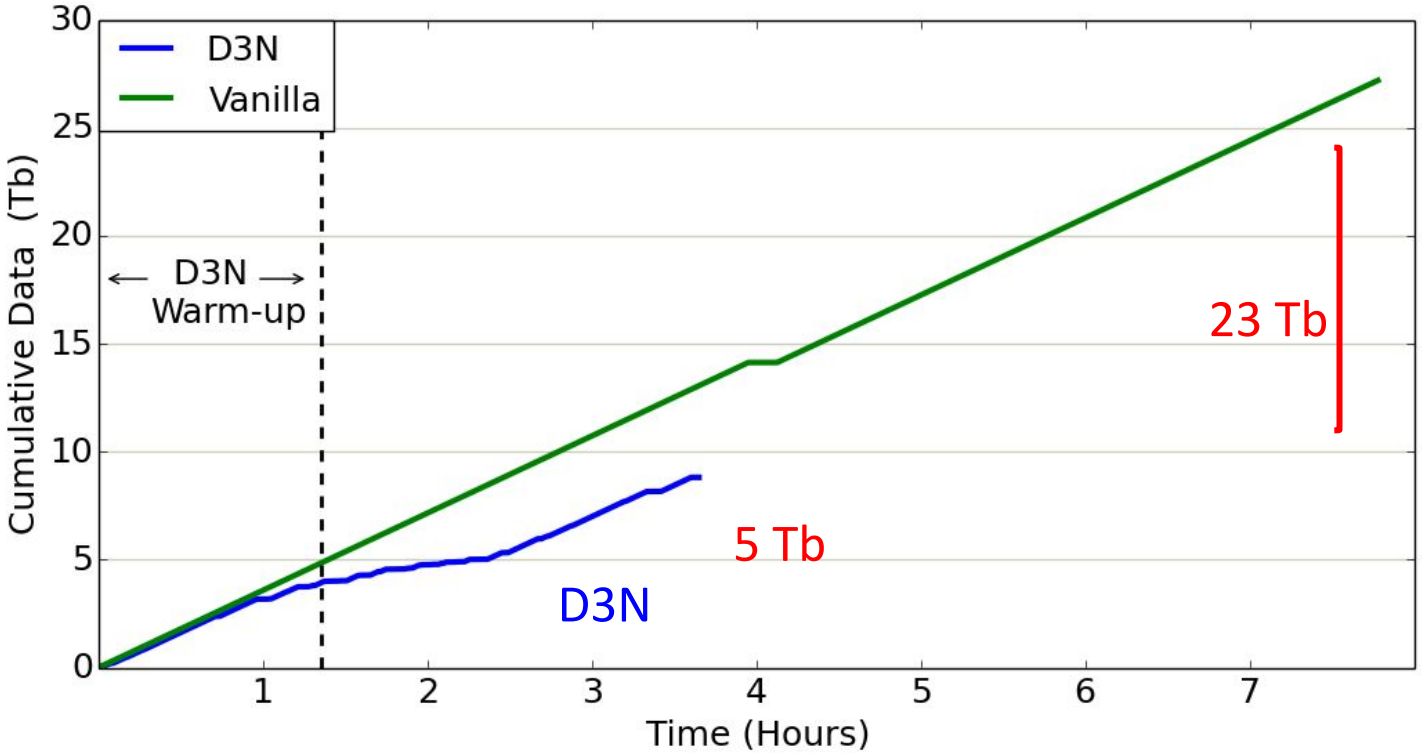


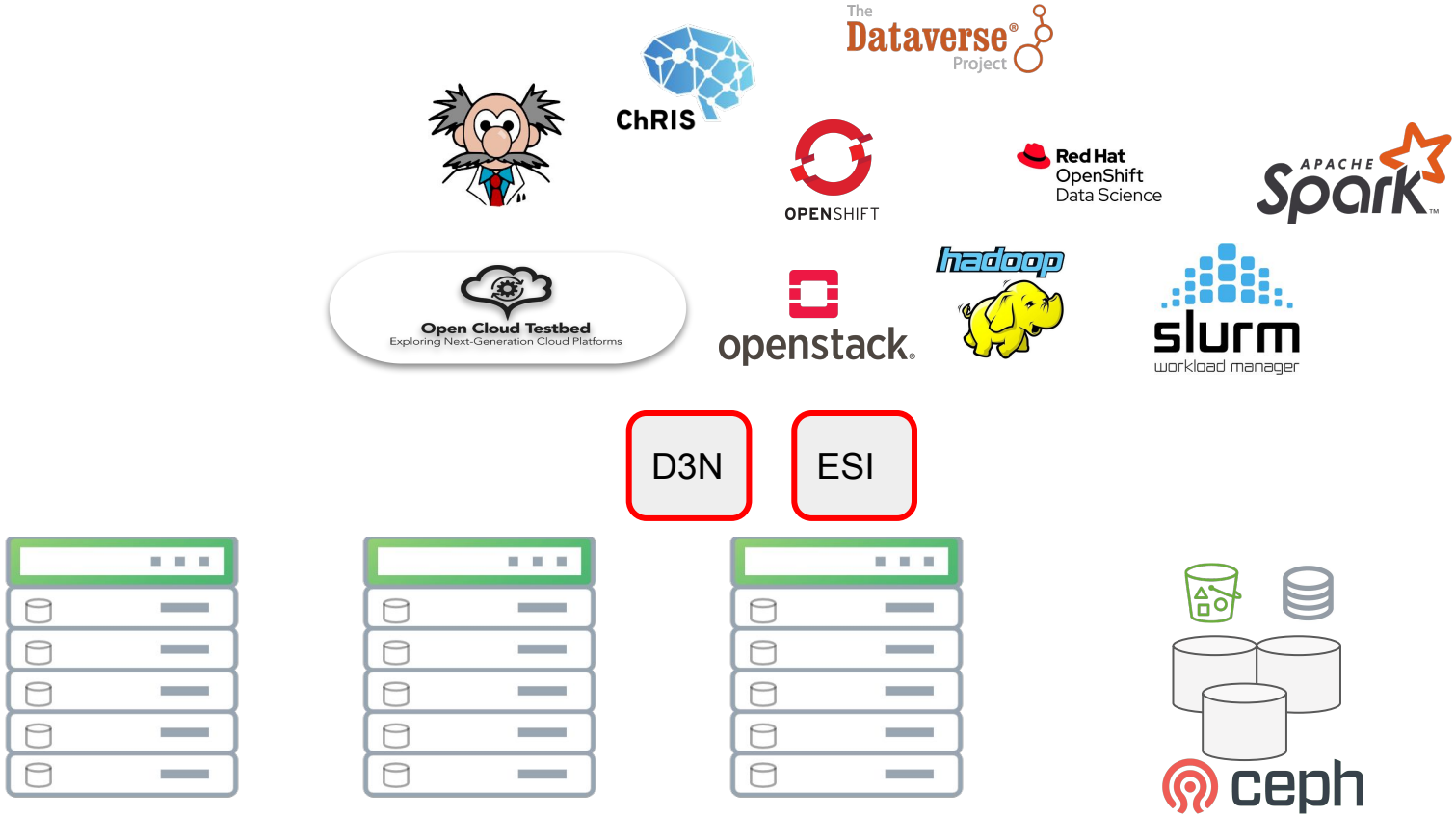

Compute
Cluster
ceph


VM
ceph


VM
ceph


RGW
O1
SSD
O1
O1
RGW
O1
SSD
ceph

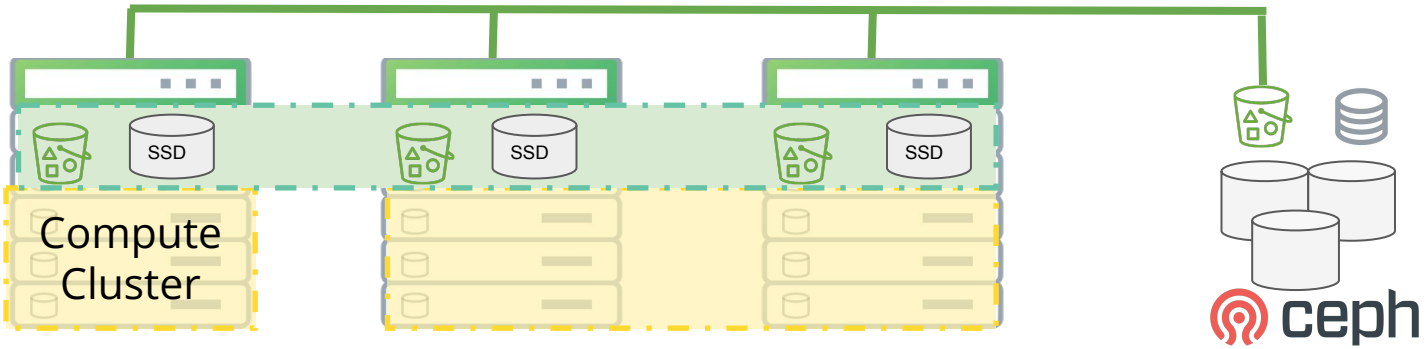

D3N
Vanilla
D3N
Warm-up
23 Tb
5 Tb
D3N
Cumulative Data (Tb)
Time (Hours)
0
5
10
15
20
25
30
1
2
3
4
5
6
7



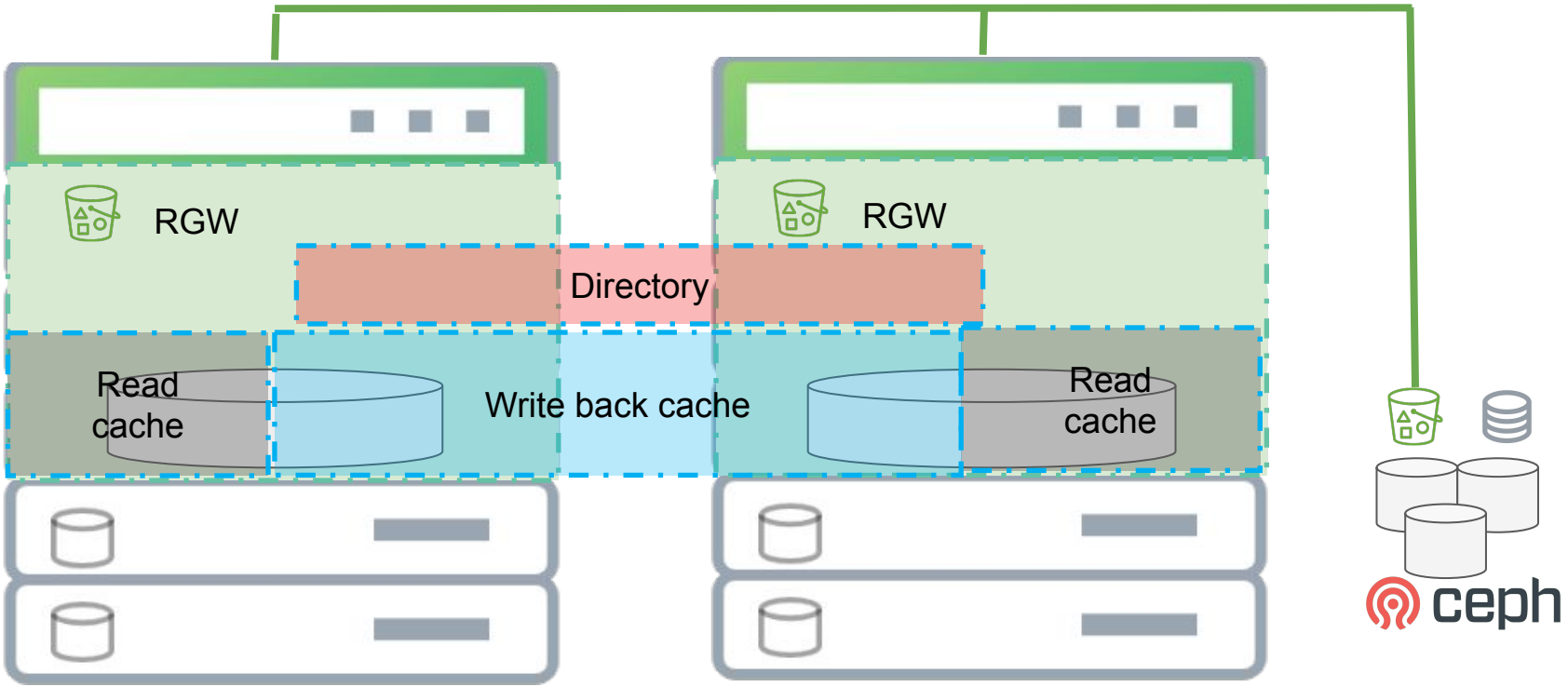
D3N: A multi-layer cache for the rest of us, Kaynar, et al, BigData'19

Problems

- Read cache not good enough
- Home node that is not using data adds overhead
- Wanted to explore more sophisticated cache management policies



D4N - Directory-based D3N

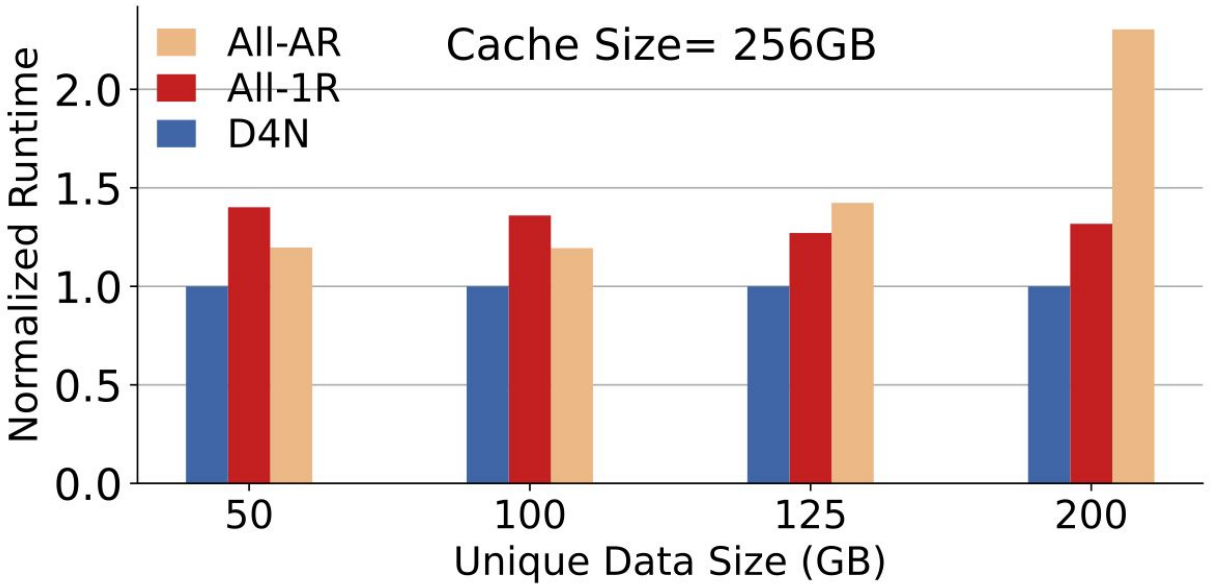


Workload Adaptability

Workload with an uniform distribution

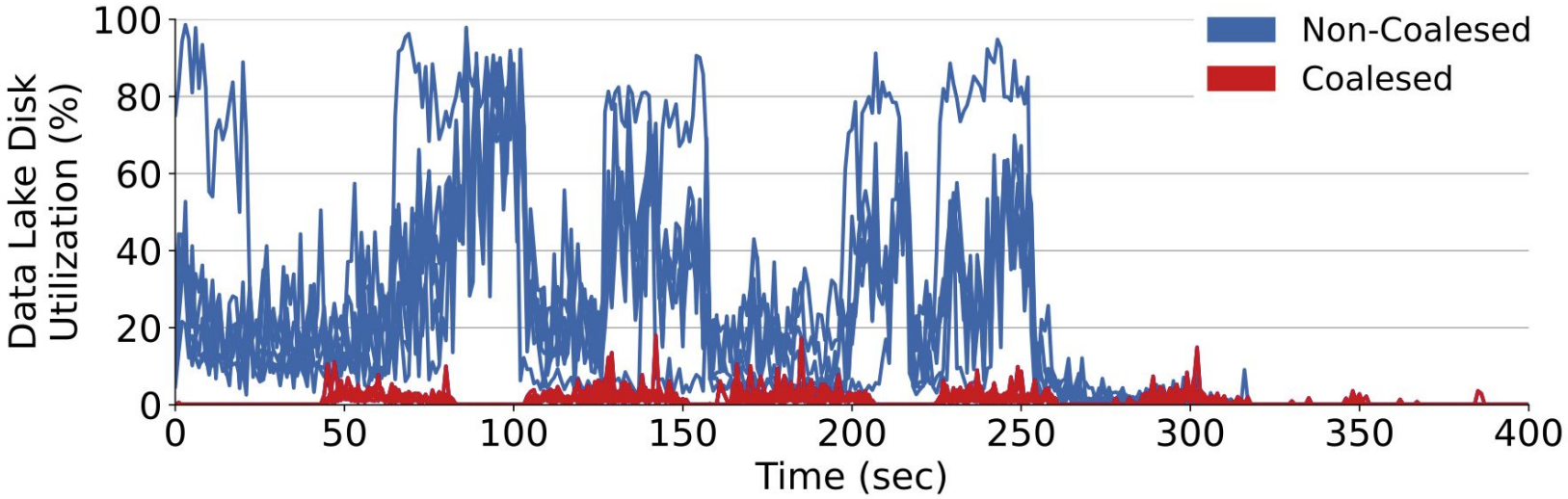
All-AR: Ideal for small working set size

All-1R: Ideal for large working set size



D4N automatically adapts replication to the working set of the demands

Insight: we can now transform data between high speed cache near compute & datalake



The Dataverse® Project

OPENSHIFT

Red Hat OpenShift Data Science

APACHE Spark™

Open Cloud Testbed
Exploring Next-Generation Cloud Platforms

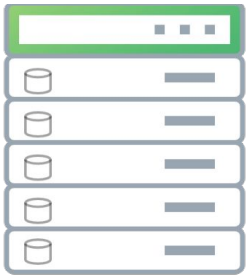
openstack.

hadoop

slurm
workload manager

D3N

ESI



The Dataverse® Project

OPENSHIFT

Red Hat OpenShift Data Science

APACHE Spark™

Open Cloud Testbed
Exploring Next-Generation Cloud Platforms

openstack.

hadoop

slurm
workload manager

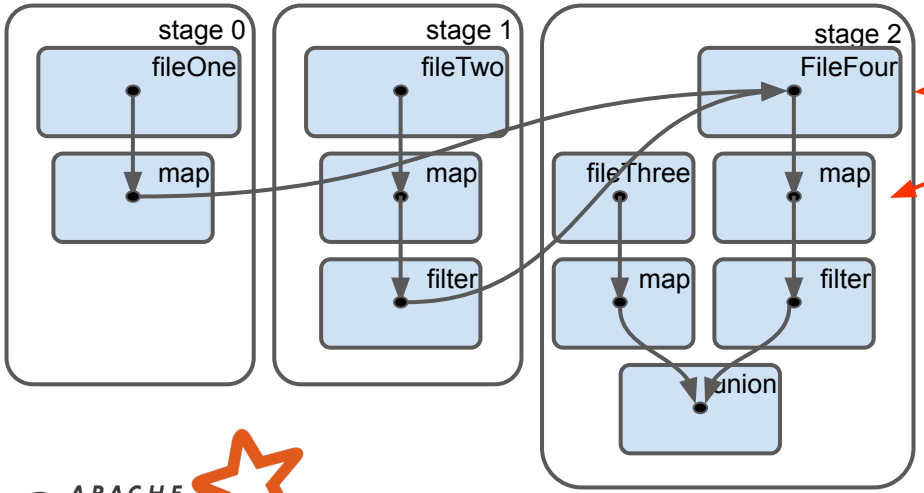
D4N

D3N

ESI

ceph

Insight: many platforms have knowledge



The Dataverse Project

Red Hat OpenShift

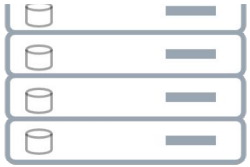
APACHE Spark

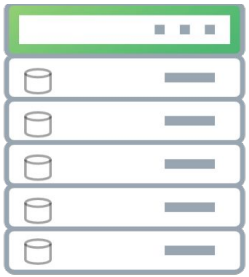
rm manager

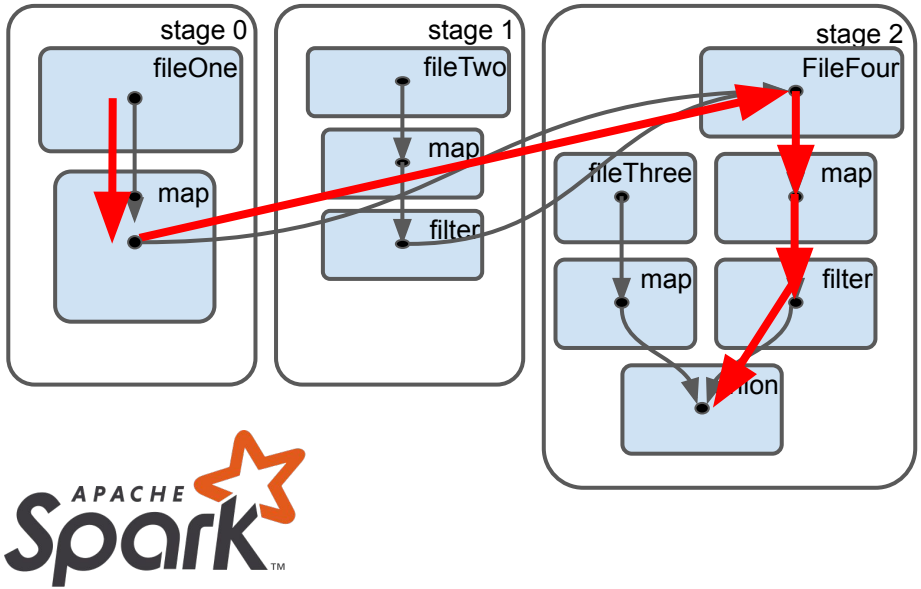
input size
+ operation
+ history
= predicted duration

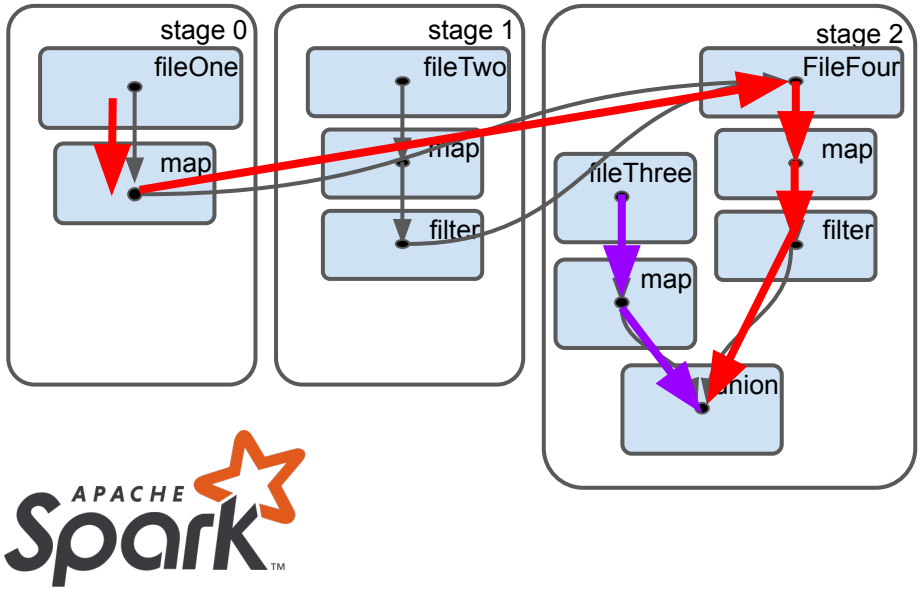
3N

ESI



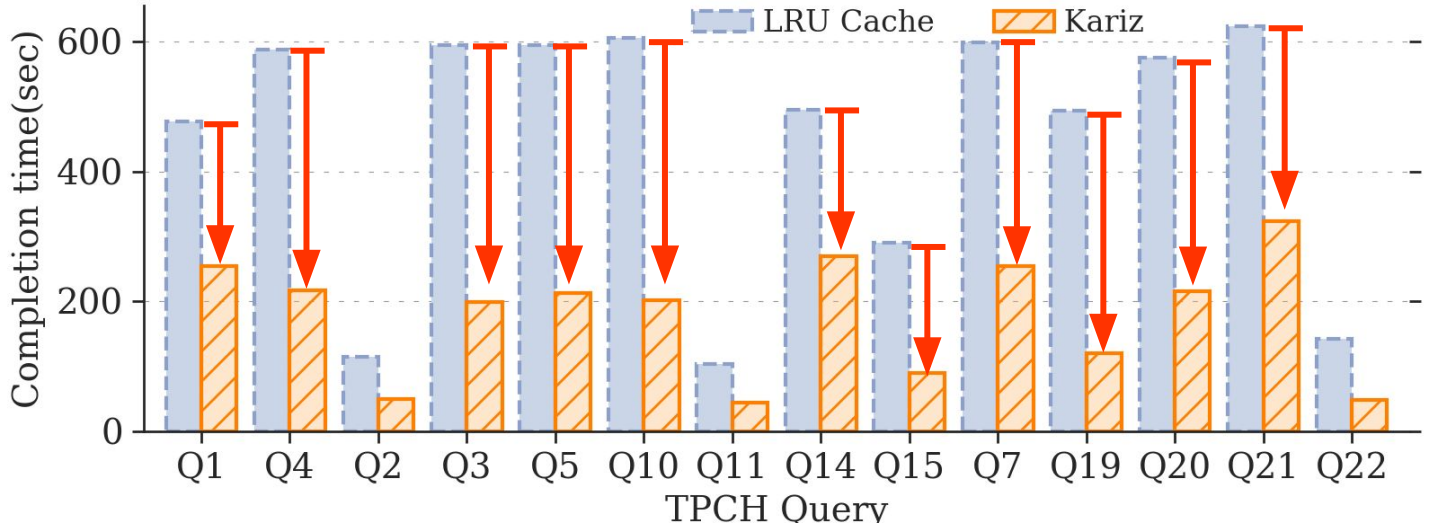


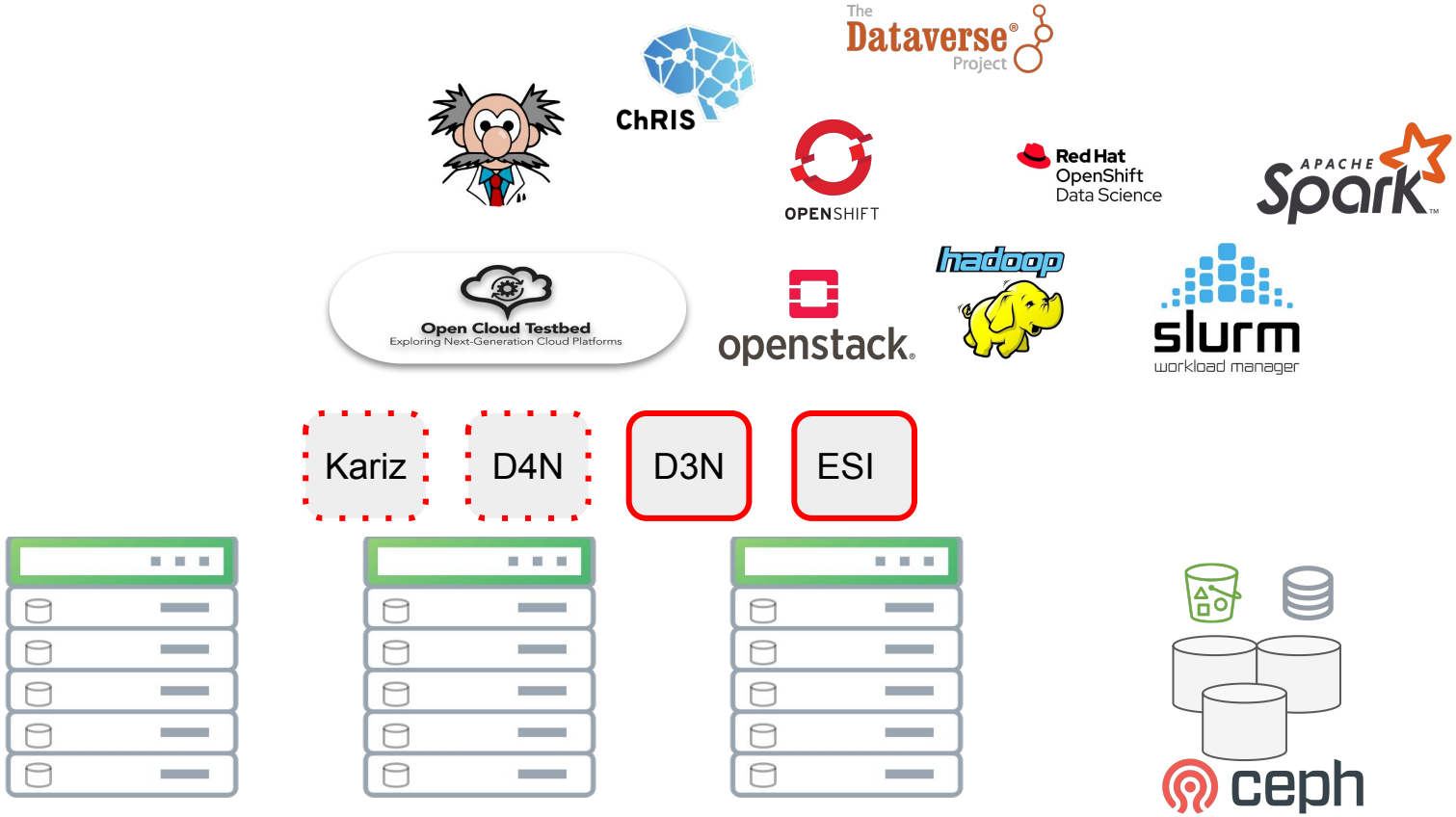

stage 0
fileOne
map
stage 1
fileTwo
map
filter
stage 2
FileFour
fileThree
map
map
filter
APACHE
Spark


stage 0
fileOne
map
stage 1
fileTwo
map
filter
stage 2
FileFour
map
filter
fileThree
map
union

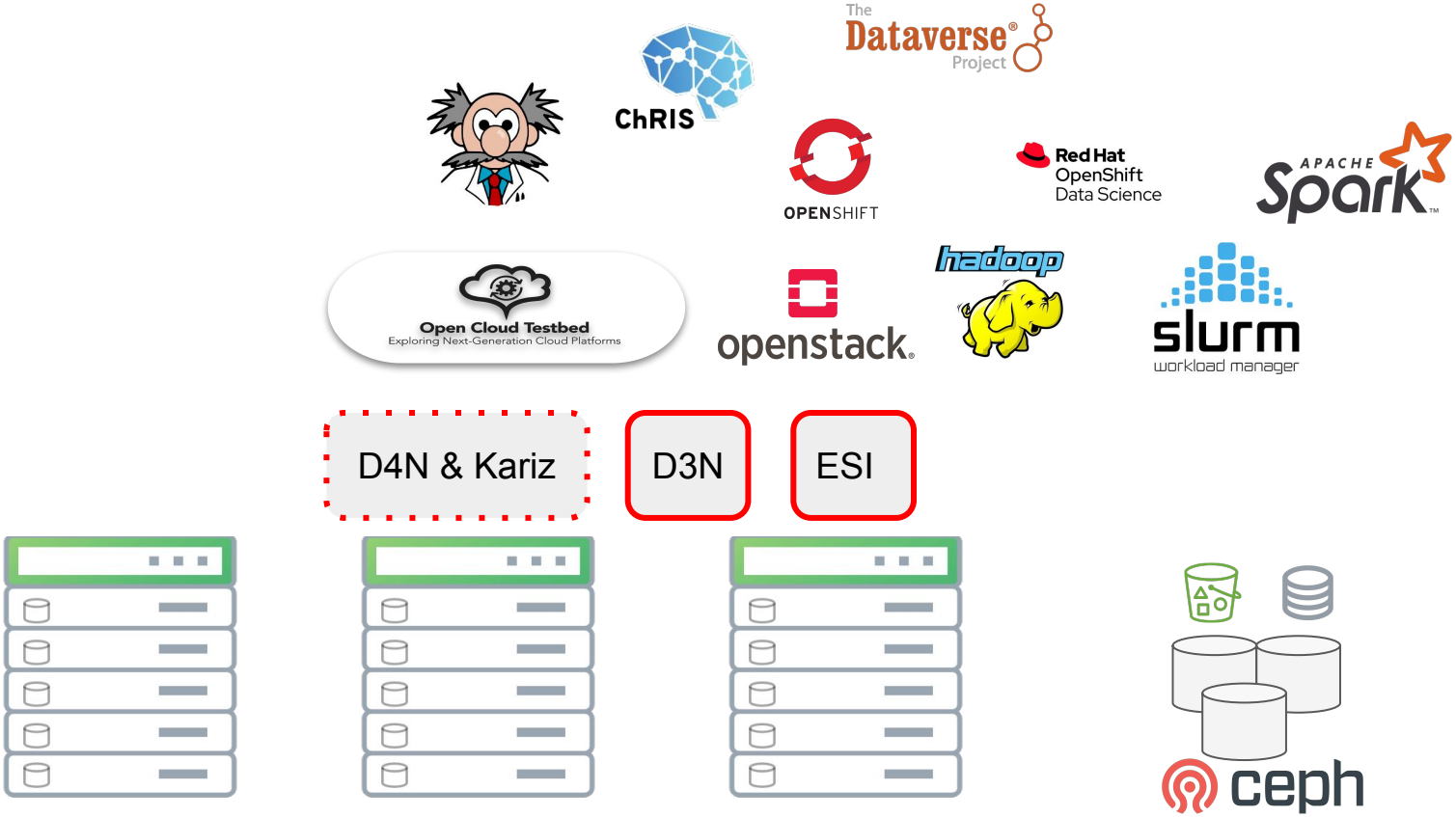
Turns out that you can improve caching when you know the future and contention

Up to 3.5x improvement on TPC-H queries





Caching in the Multiverse, Abdi et al, HotStorage'19
A Community Cache with Complete Information, Abdi et al, FAST'21


The Dataverse Project
ChRIS
OPENSHIFT
Red Hat OpenShift Data Science
APACHE Spark
Open Cloud Testbed
Exploring Next-Generation Cloud Platforms
openstack.
hadoop
slurm
workload manager
D4N & Kariz
D3N
ESI
ceph

Problem: slow access to research data

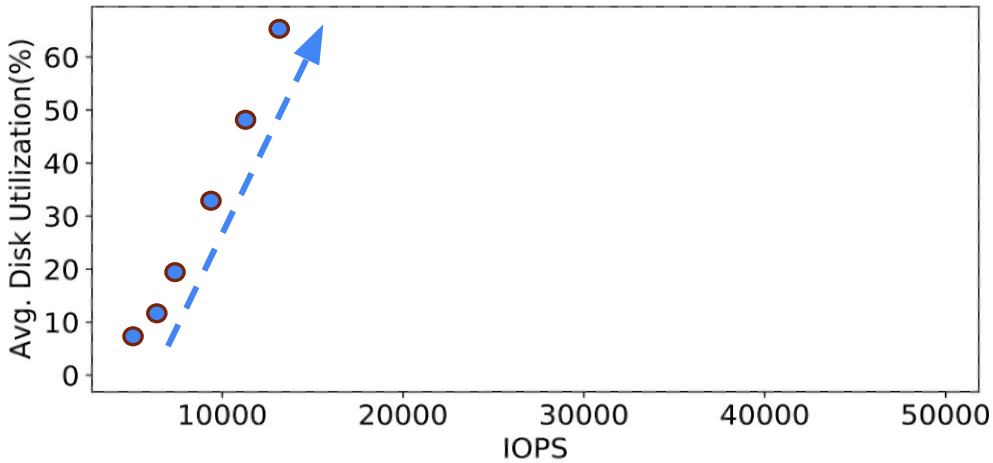
- Limited bi-sectional bandwidth
- High capacity shared storage

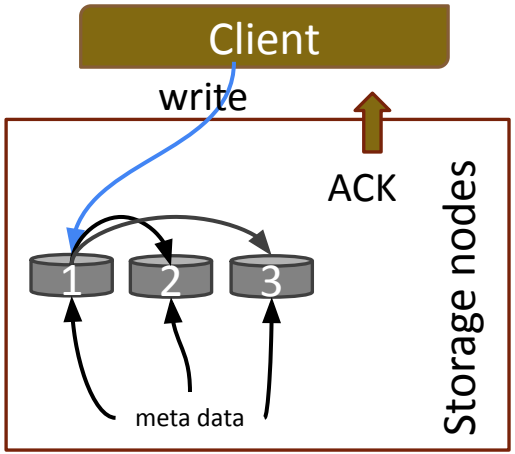
Problem: slow access to volume storage

- Limited bi-sectional bandwidth
- High capacity shared storage
- Write amplification

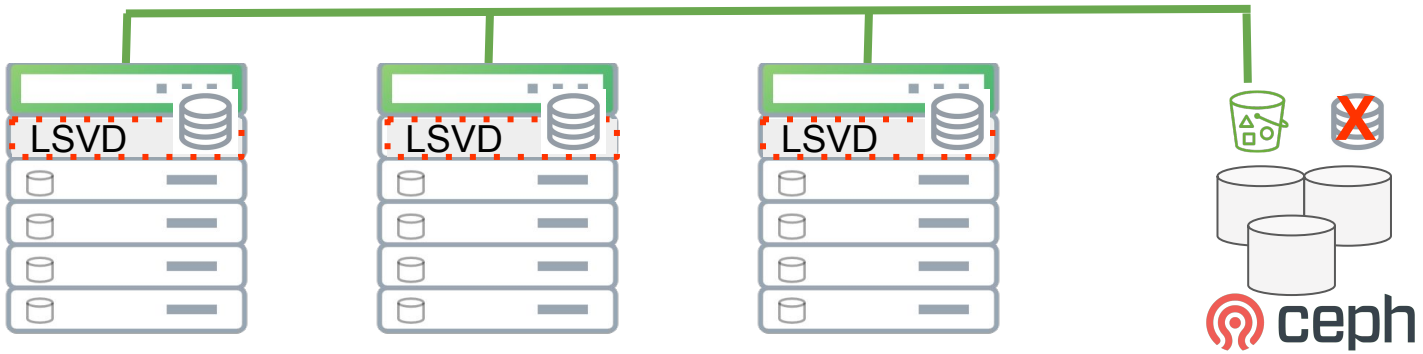
Write amplification

Increasing block devices from 1 to 32 in a backend with 9 servers and 56 spindles

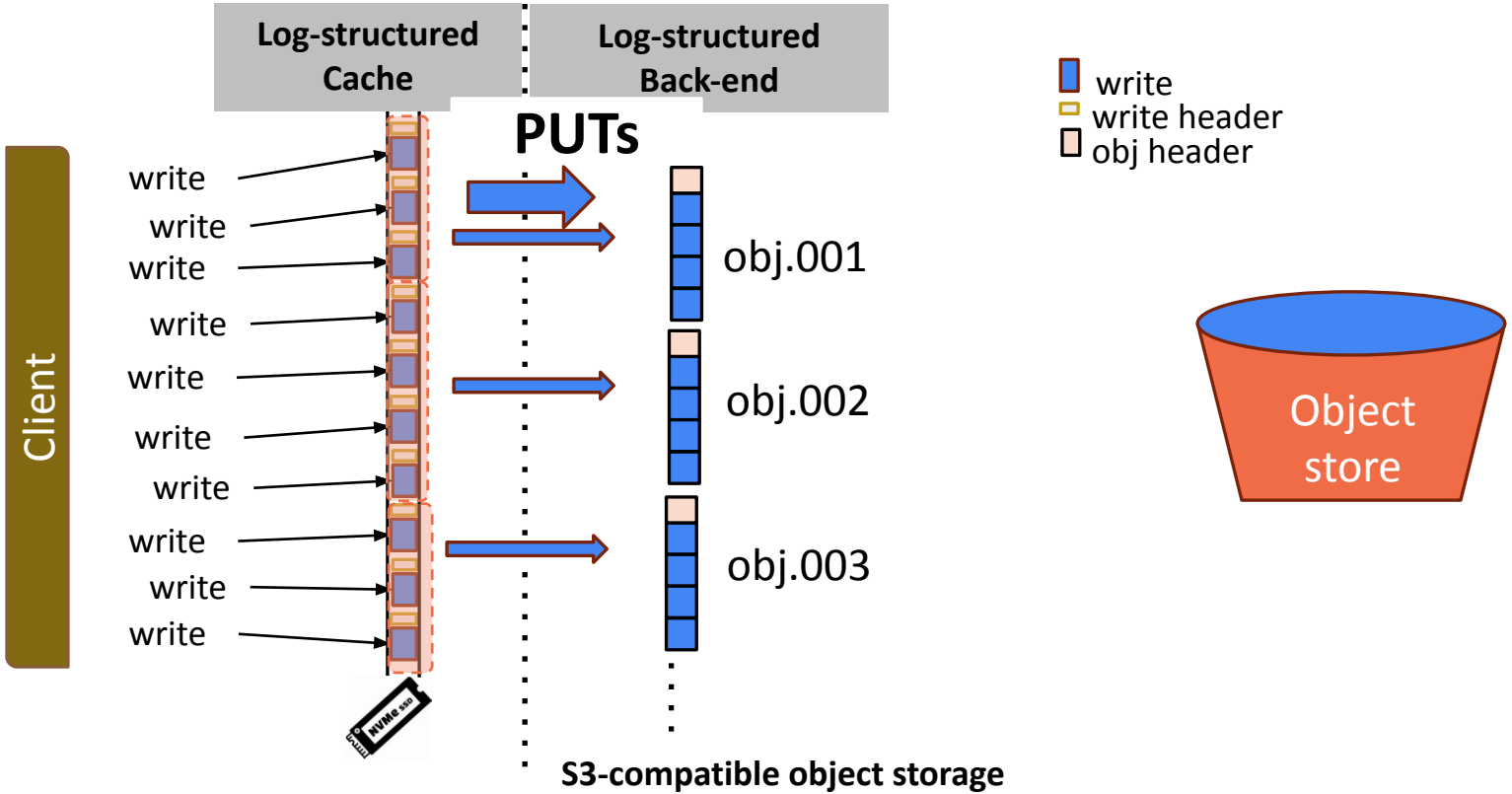




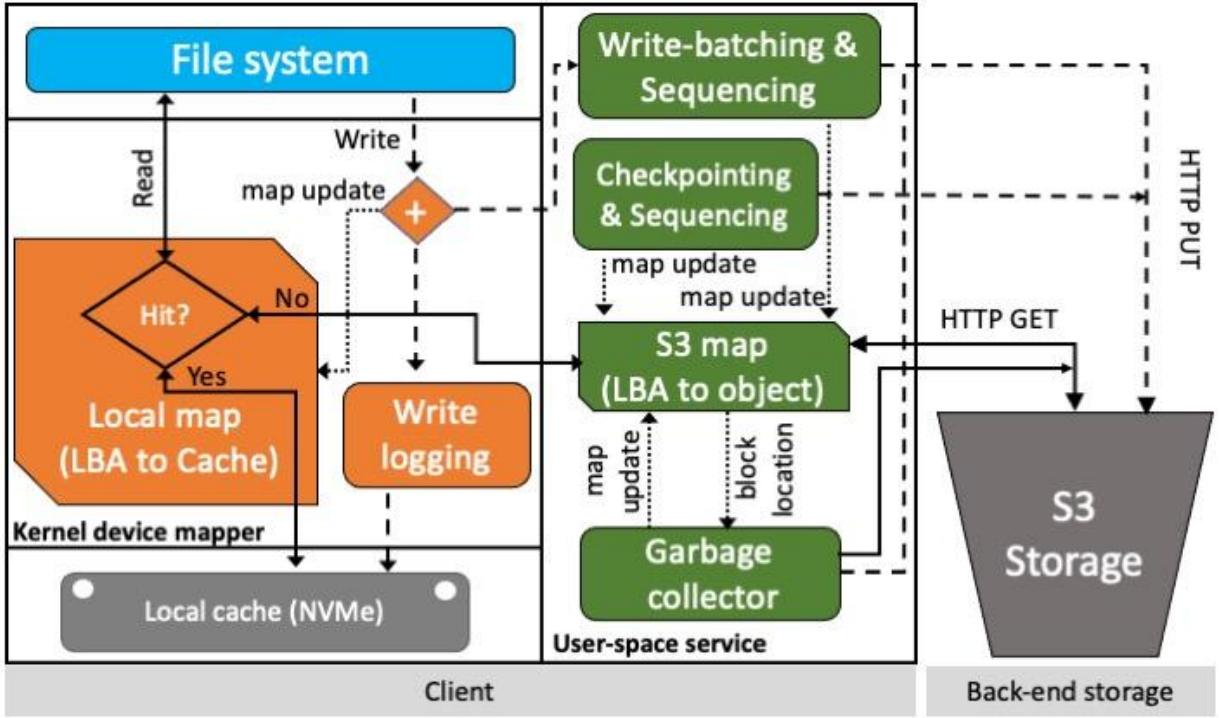
Insight: Mutate volume to object storage between high speed cache and slow datalake



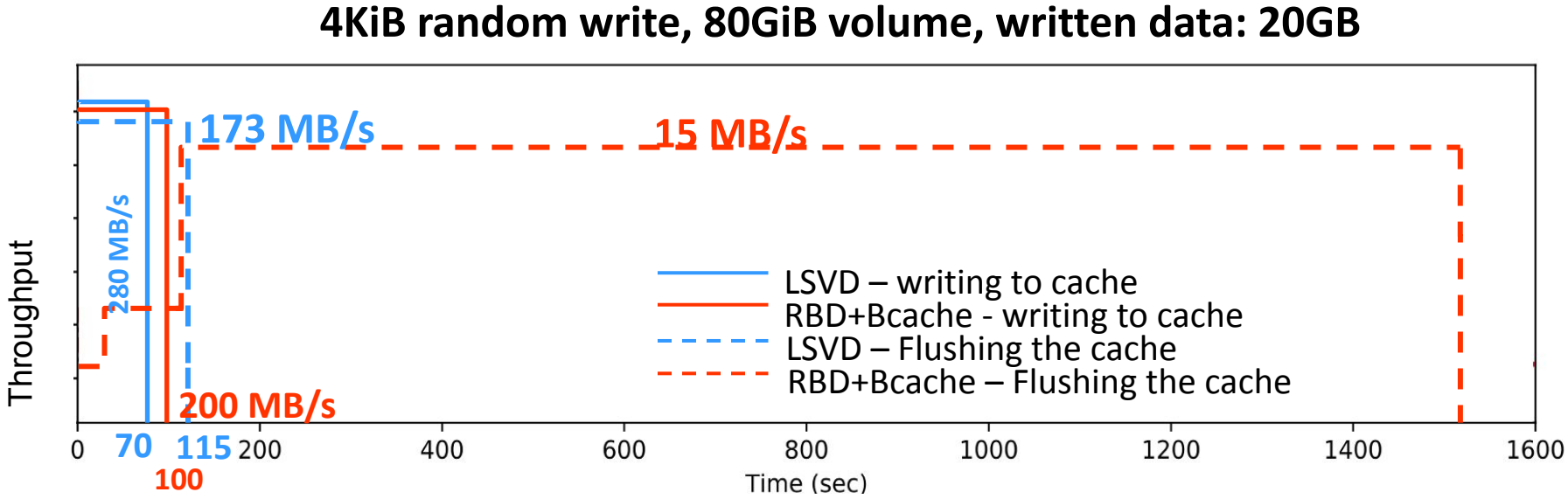
LSVD – Log structured Virtual disk

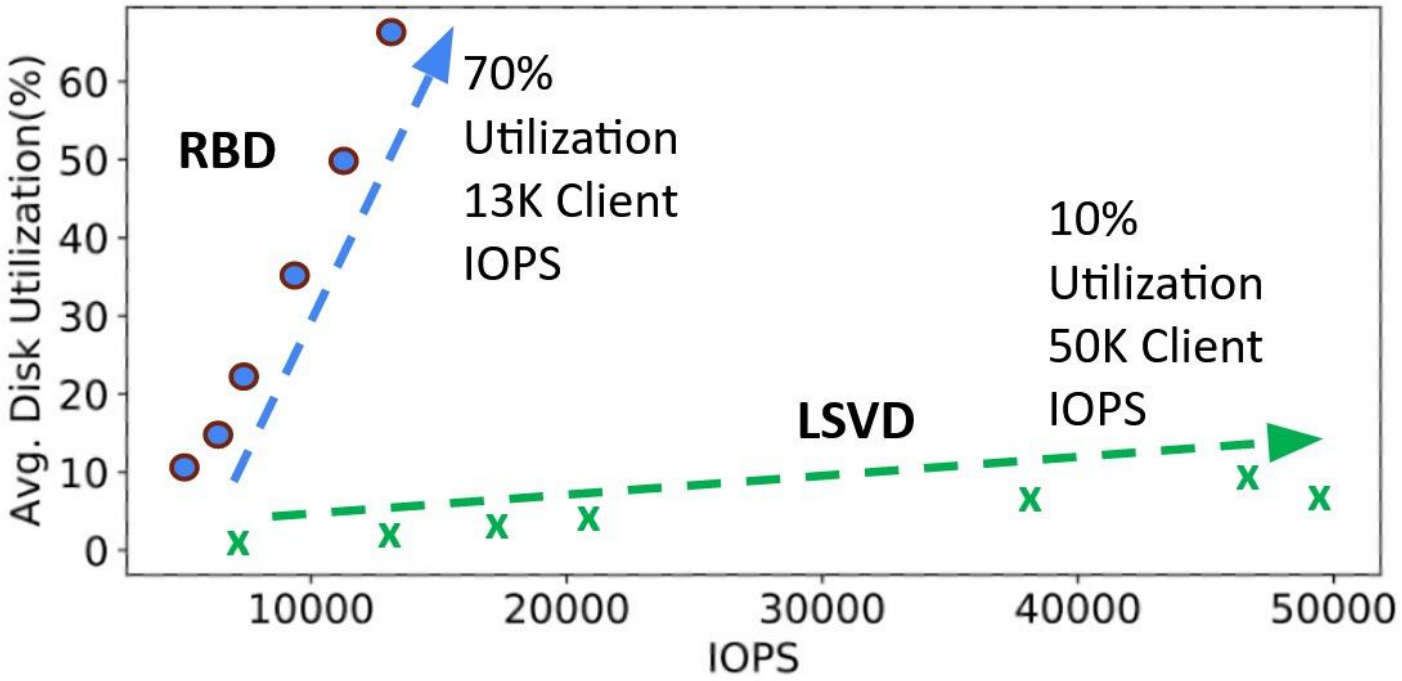


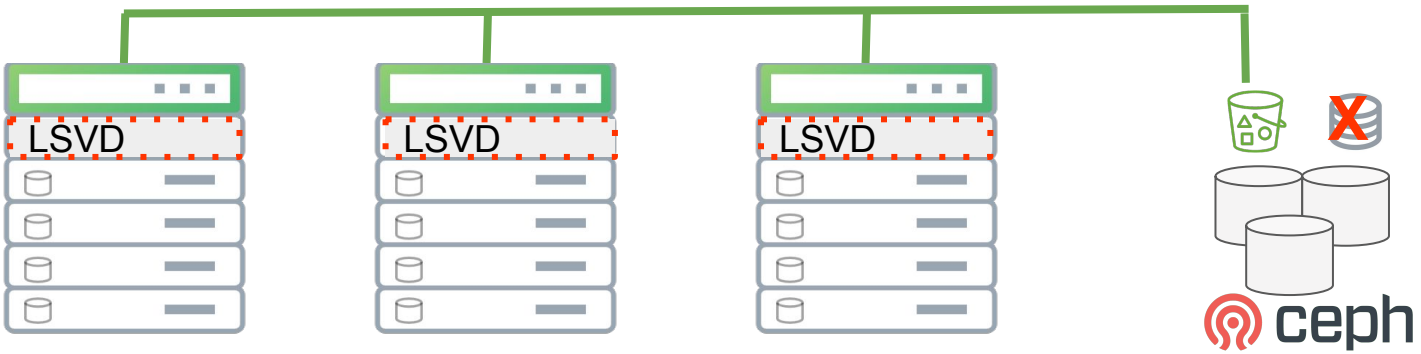
LSVD – Log structured Virtual disk



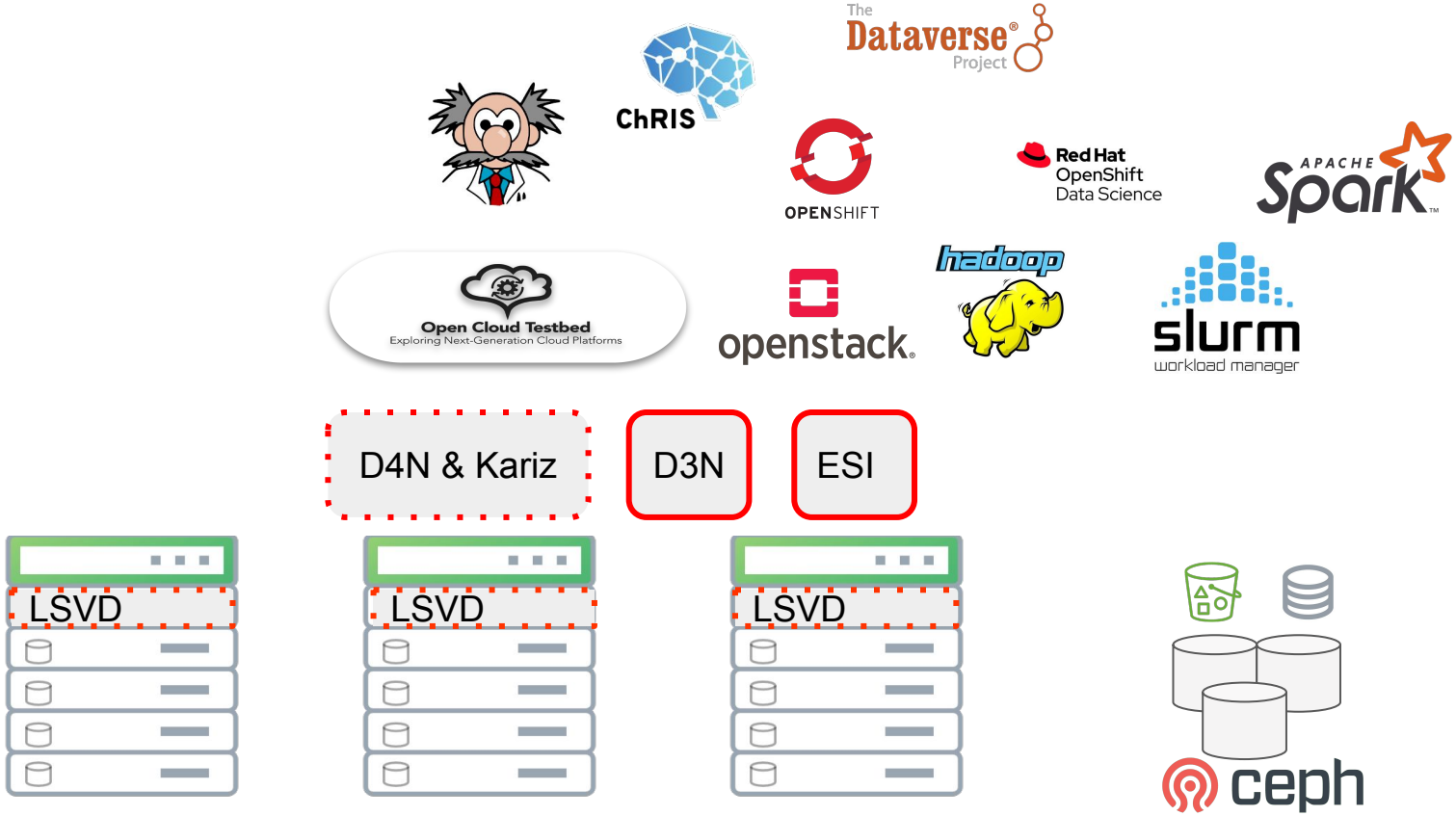
LSVD vs. RBD+Bcache – Burst Writes




Avg. Disk Utilization(%)
0
10
20
30
40
50
60
10000
20000
30000
40000
50000
IOPS
RBD
70%
Utilization
13K Client
IOPS
LSVD
10%
Utilization
50K Client
IOPS



Beating the I/O bottleneck: a case for log-structured virtual disks, Hajkazemi, et al, EuroSys'22


The Dataverse Project
ChRIS
OPENSHIFT
Red Hat OpenShift Data Science
APACHE Spark
Open Cloud Testbed
Exploring Next-Generation Cloud Platforms
openstack.
hadoop
slurm
workload manager
D4N & Kariz
D3N
ESI
LSVD
LSVD
LSVD
ceph

Larger messages

- Different services require control; no one size fits all.
- For compute fundamental building block is computers
- For storage long term storage is large immutable objects & services convert high-IOPS operations into large requests to shared storage
- Cross layer visibility & optimization critical
- Different hypothesis & whole system perspective leads to insights and new solutions to real problems

A community will develop to enable radical change?

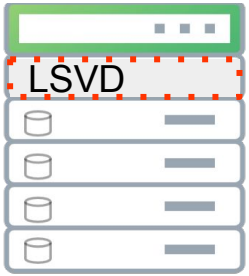
The community...

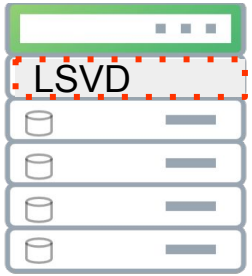
Coskun, Ayse (BU); Culbert, Jim (MGHPCC); Daitzman, Michael (BU);
Dempsey, Heidi (RH); Denhardt, Ian (BU); Desnoyers, Peter (NEU);
Dhangwattanotai, Jade (BU); Doddahonnaiah, Deeksha (NEU);
Edgar, Gwen (BU); Ferry, Sar… U); Finn, Daniel (BU);
Fonseca, Rodrigo (Brown); Fontecchio, Joseph (Umass);
Freudberg, Jeremy (BU); Freud… Ric (RH
Gilmore, Wayne (BU); Grosu, Pa… NEU);
Gudimetla, Ravi Santosh (NEU)
Hajkazemi, Mohammad …ssein (NEU); Hang, Lucas (BU);
Hayati, Arash Nemati (B… n Children's Hospital);
Hennessey, Jay (BU); Herbordt, Martin (BU); Hill, Chris (MIT);
Hudson, Trammell (2σ); Jia, Leo (BU); Jug, Sebastian (RH); Juma, Eric (BU);
…(NEU… ura (BU); Kanalavai, Nagasai Vinaykumar (NEU);
…Shas… ul, Robin (BU); Kaynar, Emine Ugur (BU);
…arthey (RH); Khare, Akshaya (NEU); Krieger, Orran (BU);
Kulkarni, Chaitanya (BU); Kumar, Gagan (NEU,RH); Kumar, Rajul (NEU);
Laskey, Richar… Huy (BU); Li, Hua (BU); Liberti, Kyl
Lu, Ming… (B… un (NE… krigiorgos, Dimitri (B
Maleki, Hoda … sta); Mandviwala, Huzefa (BU); Ma… (OCT);
Matsuura, Nicholas (BU); McGann, Laura (BU); Milanov, Rado (BU);
Mohan, Apoorve (NEU); Mosayyebzadeh, Armin (BU); Munakami, Milson (HU);
Munson, Charlie (MIT, LL); Murugesan, Sirushti (NEU);
Nadeau, Tom (Brocade,RH); Nagaraj, Lohith Kesaguli (NEU); Nikolla, Kristi (BU);

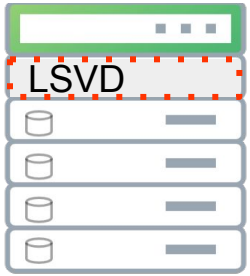
D4N & Kariz

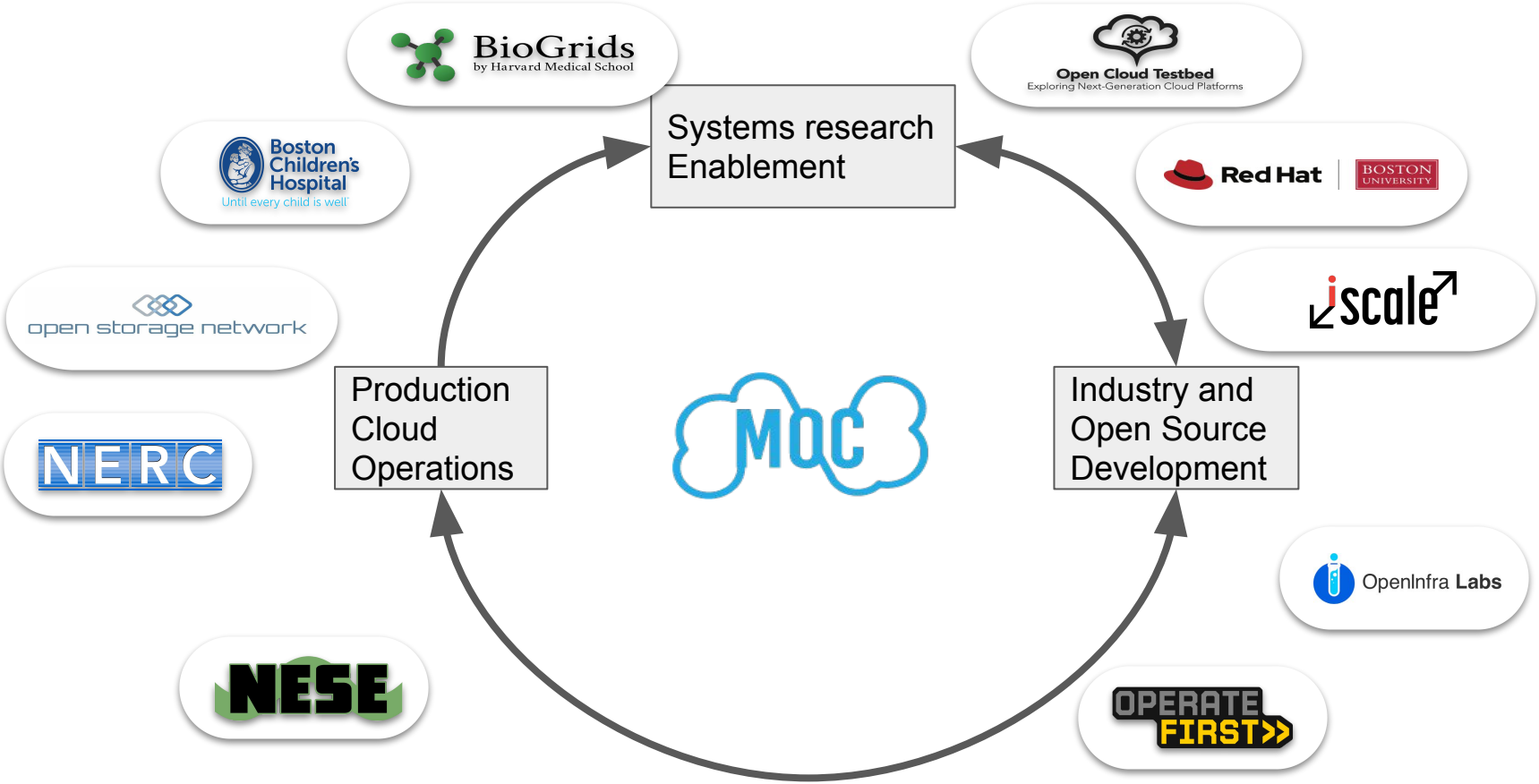
D3N

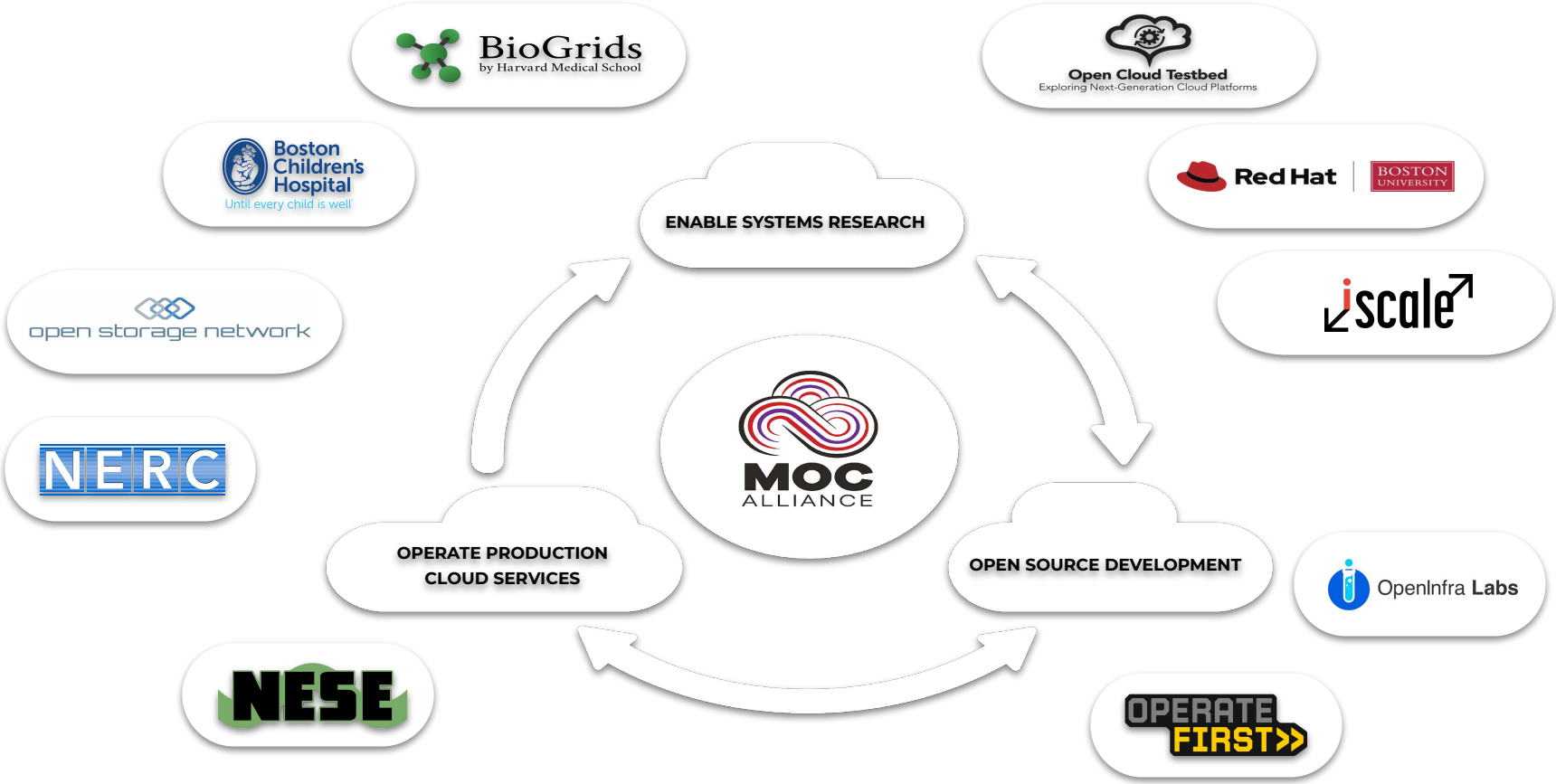
ESI








BioGrids
by Harvard Medical School
Open Cloud Testbed
Exploring Next-Generation Cloud Platforms
Boston Children's Hospital
Until every child is well
Red Hat
BOSTON UNIVERSITY
open storage network
jscale
Systems research
Enablement
Production
Cloud
Operations
MOC
Industry and
Open Source
Development
NERC
OpenInfra Labs
NESE
OPERATE FIRST


BioGrids
by Harvard Medical School
Open Cloud Testbed
Exploring Next-Generation Cloud Platforms
Boston Children's Hospital
Until every child is well
Red Hat
BOSTON UNIVERSITY
ENABLE SYSTEMS RESEARCH
open storage network
jscale
NERC
MOC
ALLIANCE
OPERATE PRODUCTION CLOUD SERVICES
OPEN SOURCE DEVELOPMENT
OpenInfra Labs
NESE
OPERATE FIRST

Production services available: containers, VMs, object and volume storage

- Available research, education and limited startup use
  - Individual PIs can sign up through regular procurement
  - Some institutions are acting as intermediary
    - understand requirements & facilitation
    - offer free tier for under resourced
- Charges (starting Aug 1) cover costs equipment, energy, leased rack space, operations staff, software licensing..
  - cpu & GPU < 1/2 comparable comparable public cloud on-demand
  - storage ~1/3 comparable offerings
  - no egress fees
- We expect rates to drop substantially as scale grows
- Institutions can become operationally involved, or replicate & federate MOC

Where are we going? What are the new demands...?

- Can we inform platform scheduling by where data is cached/accessed?
- OSN and NESE tape driving us to work on geographical distribution and automatic tiering
- Dataverse engagement driving research on data discovery and integrating compute with storage for privacy concerns and self sustainability
- Increasing desire from broad range of data users for integrating providence: taint tracking, optimization,
- Everything on immutable objects.

Remember this?

Platform	Hurricane OS
Compute & Networking Hardware	Hector Ethernet
Storage	Hard Drive NFS
Application Requirements	Performance Durability Recovery

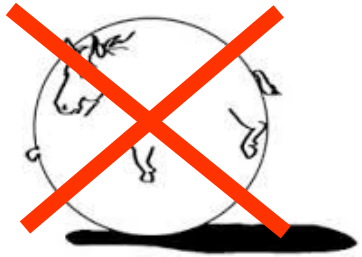
Story is more complicated today...

Platform	Kubernetes, Spark, Docker, Airflow, Redis, Kafka, Log4J, Mesos, RabbitMQ, TensorFlow, …
Compute & Networking Hardware	CPU (x86, ARM, RiscV), DPU, GPU, FPGA... Ethernet, InfiniBand, SDN, P4,...
Storage	FS: (NFS, HDFS, other non-POSIX), RDBMS, KV store (Memcache, Redis, Cassandra, Riak), object storage(S3), Volume (local SSD, local virt disk, remote virt disk, NVMe-o-F), Disks: (NVMe, QLC, SMR, pmem)...
Application Requirements	Data discovery, Cleaning, Security, Performance, Durability, Providence, Retention, Regulatory, Geographical access, Scale, Memoization ...

Approach still applicable...

- Hypothesis of a radical change (e.g., 64 bit NUMA MP)
- Complete system; visibility into applications & technologies, and ability to work across layers
- Research based on real application demands; don’t worry about innovation:
  - Start with something simple and evolve,
  - if problem is tough, research will happen to solve it
  - if your system is different, you will have novel insights
- Even if radical change takes time:
  - if hypothesis is eventually true, long term work will have an impact.
  - you will solve real problems, and a community will develop to enable radical change

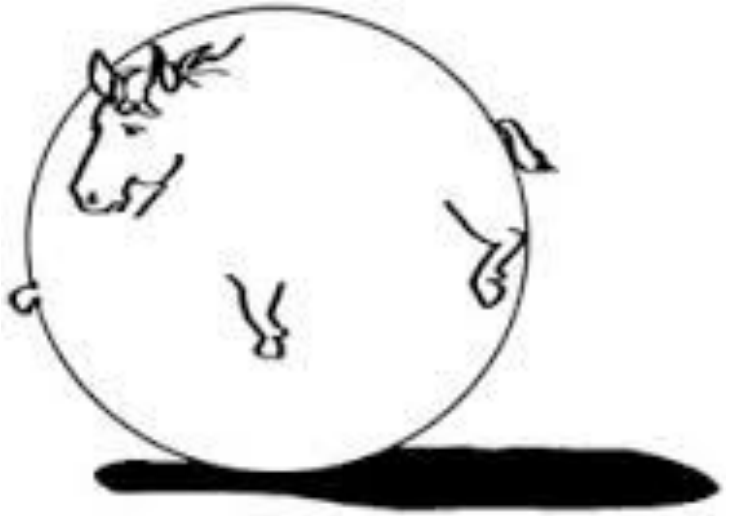
Concluding remarks



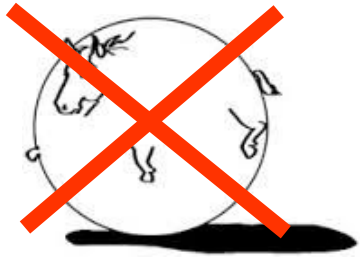
- The story isn’t simpler, but, research can be:
  - motivated by problems of a real cloud with access to real data, users, and scale
  - transition research ideas into capabilities that have an impact
  - engage with a community across diverse layers

A racehorse magnate offers a million dollars to anyone who could accurately identify race-winning horses....

Physicist answer... *assume a spherically symmetric horse travelling in a vacuum*.



Concluding remarks



- The story isn’t simpler, but, research can be:
  - motivated by problems of a real cloud with access to real data, users, and scale
  - transition research ideas into capabilities that have an impact
  - engage with a community across diverse layers
- In the past, I worked on compute problems, and I kept having to solve some storage problem to make progress.
- Today, most compute seems to be mutations on data, where you need to retain all the information about what you did...
- Is the future storage, with compute as a need to solve side effect...