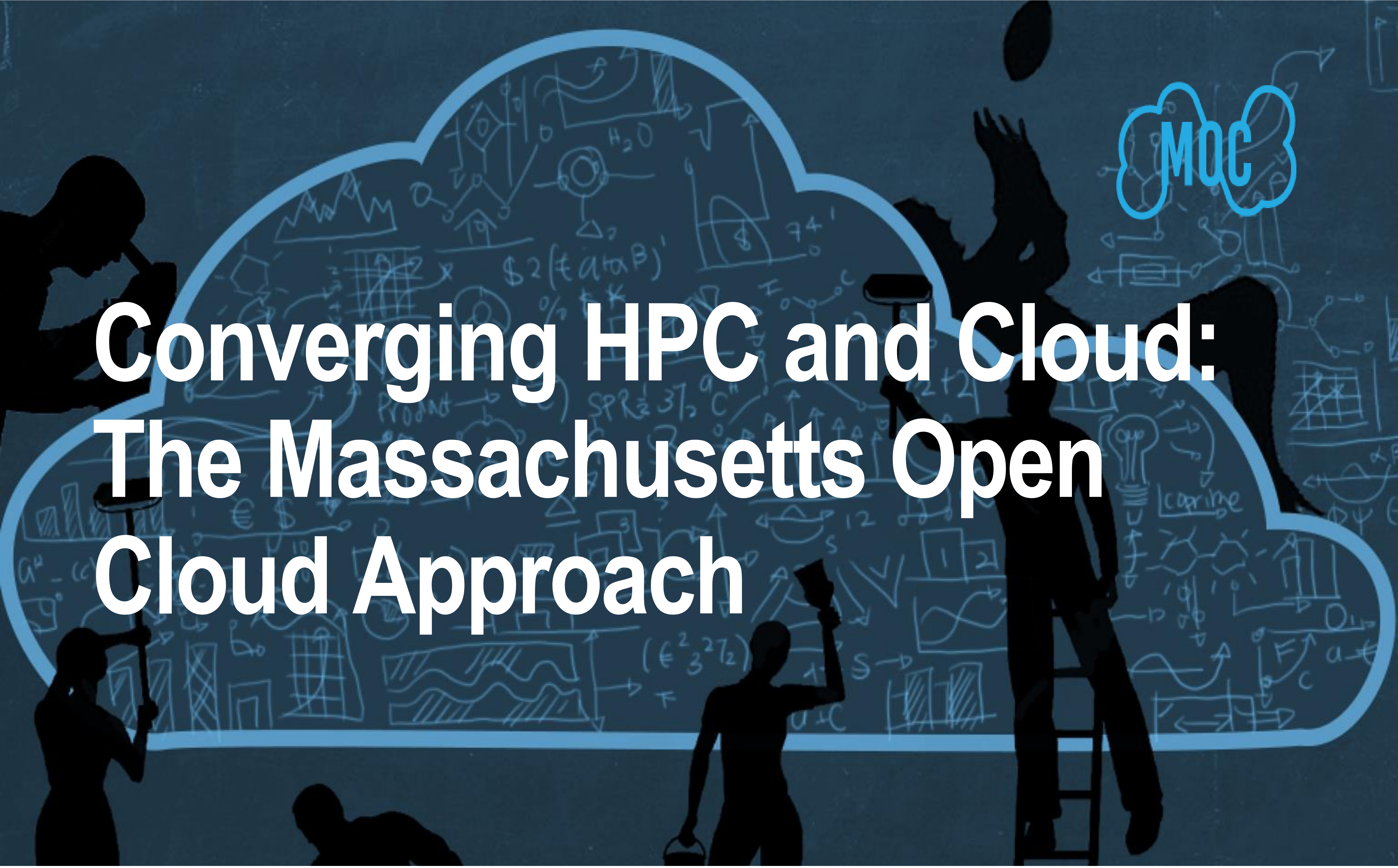


MOC

Converging HPC and Cloud: The Massachusetts Open Cloud Approach

MASSACHUSETTS TECHNOLOGY COLLABORATIVE

BOSTON UNIVERSITY

VE RI TAS

MGHPCC MASSACHUSETTS GREEN HIGH PERFORMANCE COMPUTING CENTER

MIT

Northeastern

UMASS

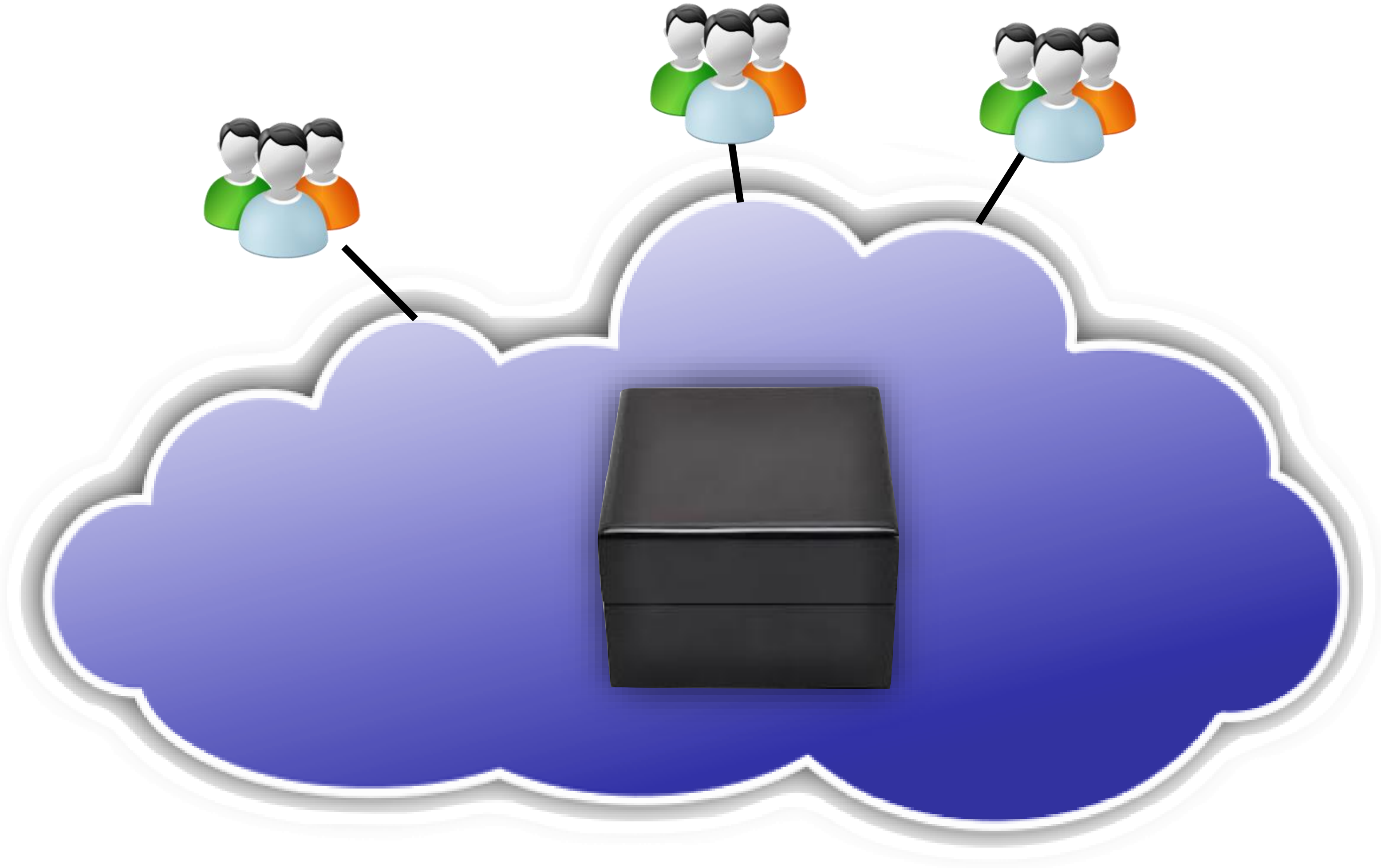
Cloud/HPC Convergence

- Google 3-year trailing CapEx, as of March 2017: $29.4 Billion (John Wilkes – RH Colloquium series 2017)
  - Cloud is the new commodity
- Increasing use of Accelerators:
  - FPGAS – MSFT (e.g., Catapult)
  - GPUs, TPUs – Google
- Full-bisectional bandwidth networks
- Enormous opportunity to share infrastructure
- HPC is a much larger marketplace than today's cloud – if we leave the ivory tower

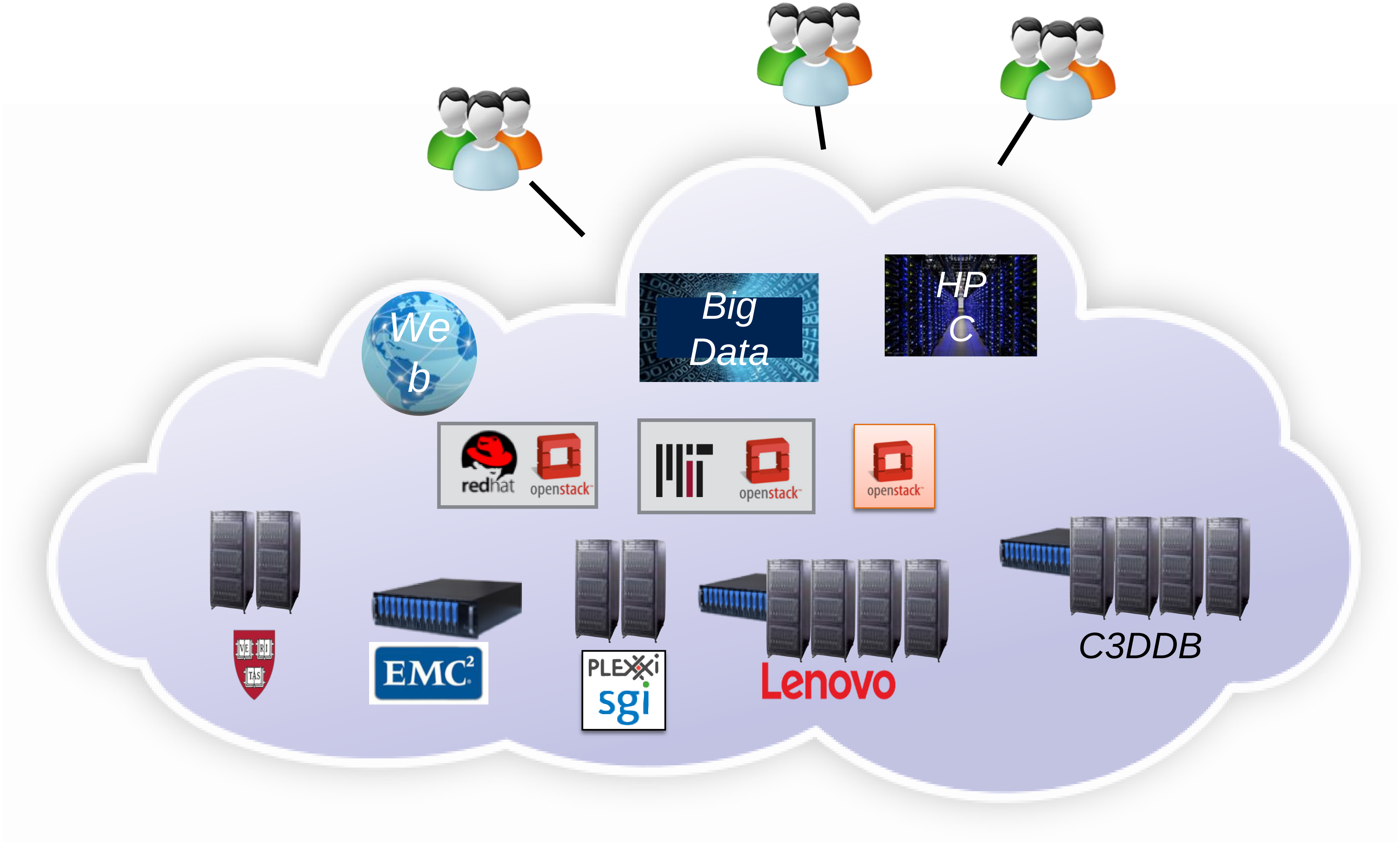
- We have an incredible opportunity to lead the convergence in this region

Today’s IaaS clouds

- One company responsible for implementing and operating the cloud
- Typically highly secretive about operational practices
- Exposes limited information to enable optimizations



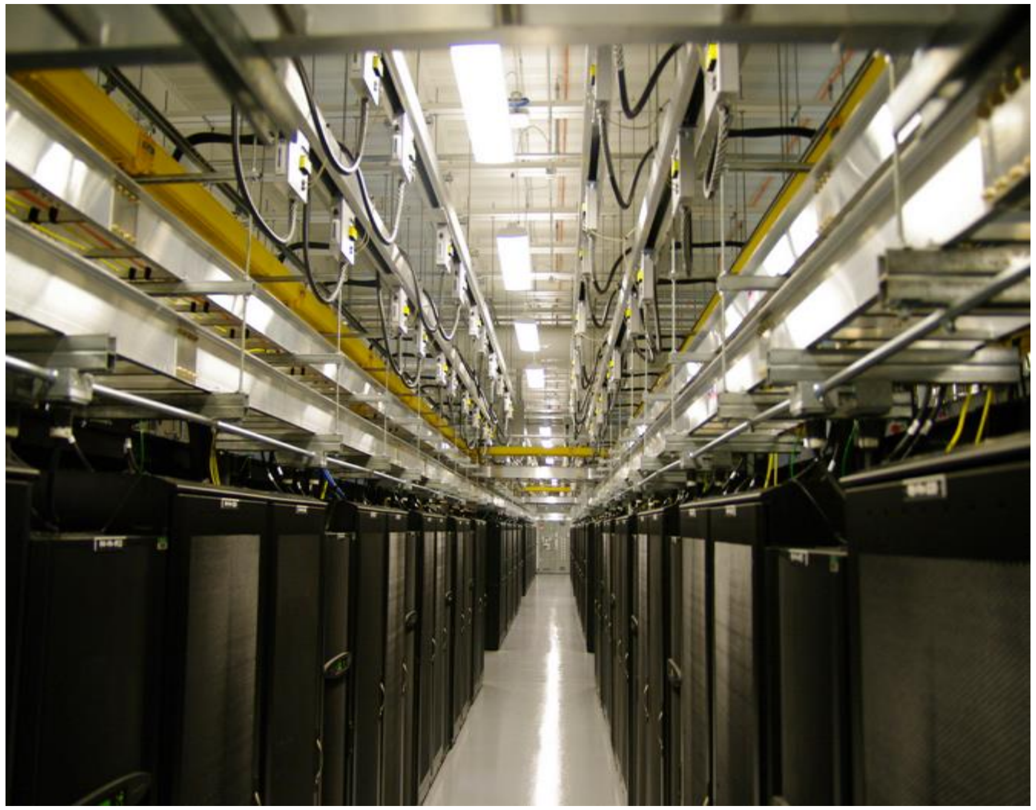
We are exploring a different model An "Open Cloud eXchange (OCX)"



This isn’t crazy… really

- Current clouds are incredibly expensive…
- Much of industry locked out of current clouds
- lots of great open source software
- lots of great niche markets; markets important to us…
- Price is terrible for computers run 24x7x365
- this doesn’t need to be AWS scale to be worth it
  - “Past a certain scale; little advantage to economy of scale” — John Goodhue

MGHPCC



15 MW, 90,000 square feet + can grow

The Massachusetts Open Cloud

ADVERTISEMENT

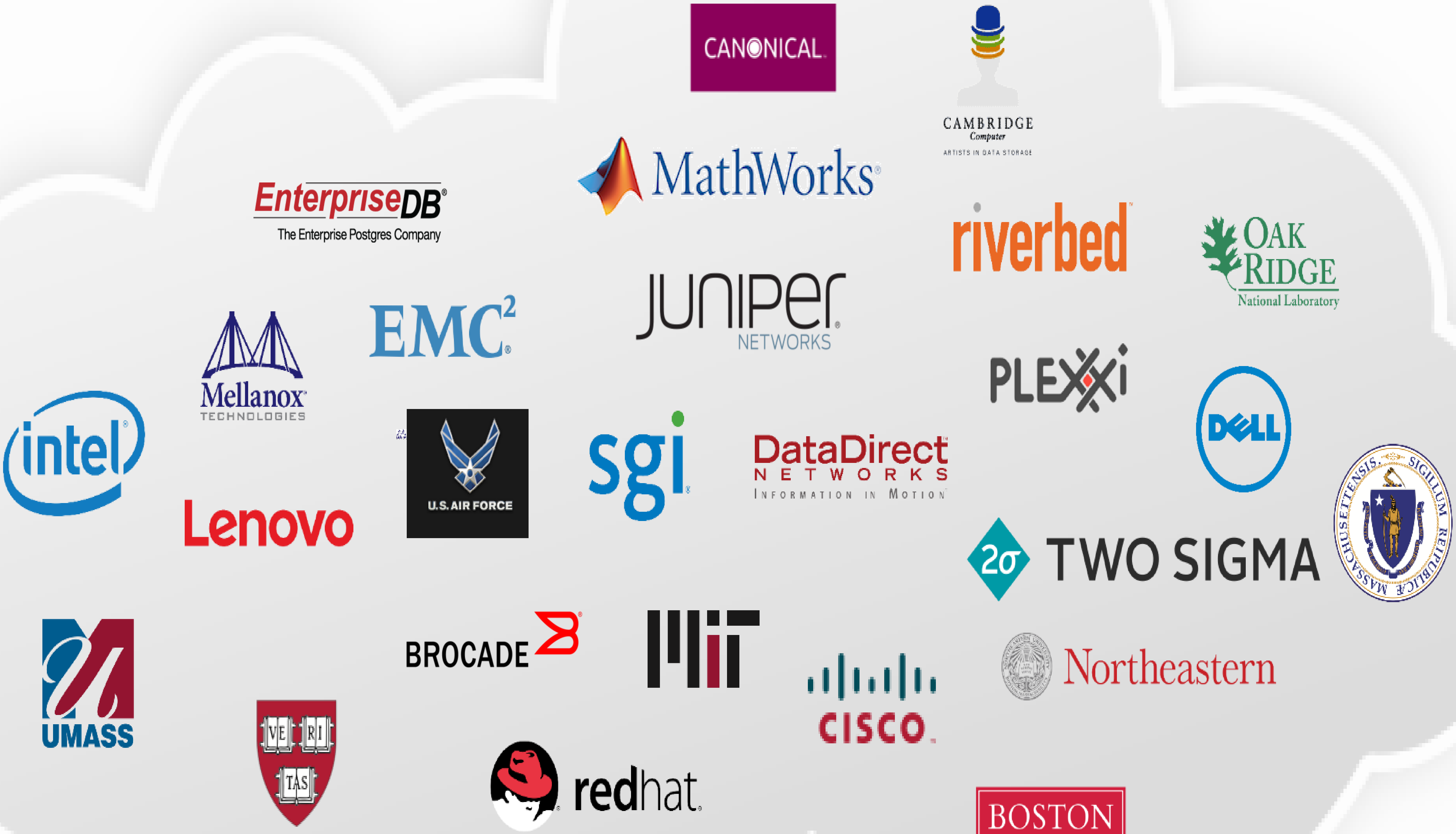
Governor Patrick Announces Funding to Launch Massachusetts Open Cloud Project

Mon, 04/28/2014 - 12:07pm

by Mass Open Cloud Project

Get the latest news in High Per
more - Sign up now!

THE MASSACHUSETTS COLLABORATORS



It’s real…

- Available now: Production OpenStack services…
  - Small scale, but growing (couple of hundred servers, 550 TB storage), 200+ users
  - VMs, on-demand Big Data (Hadoop, SPARK...),

- What’s coming:
  - OpenShift – Red Hat
  - Simple GUI for scientific end users
  - Federation across universities
  - Rapid/secure Hardware as a Service
  - Cloud Dataverse
  - Integration 20+ PB DataLake - NESE

A few relevant projects

HPC/HTC and Cloud
Making them work together efficiently

Rajul Kumar, NEU NEUkumar.raju@husky.neu.edu
Christopher Hill, MIT, cnh@mit.edu
Evan Weinberg, BU, weinbe2@bu.edu
…

HPC and Cloud convergence

High Performance Computing (HPC)

- HPC clusters tend to have infinite workloads
- Requires loads of dedicated resources to get the jobs done

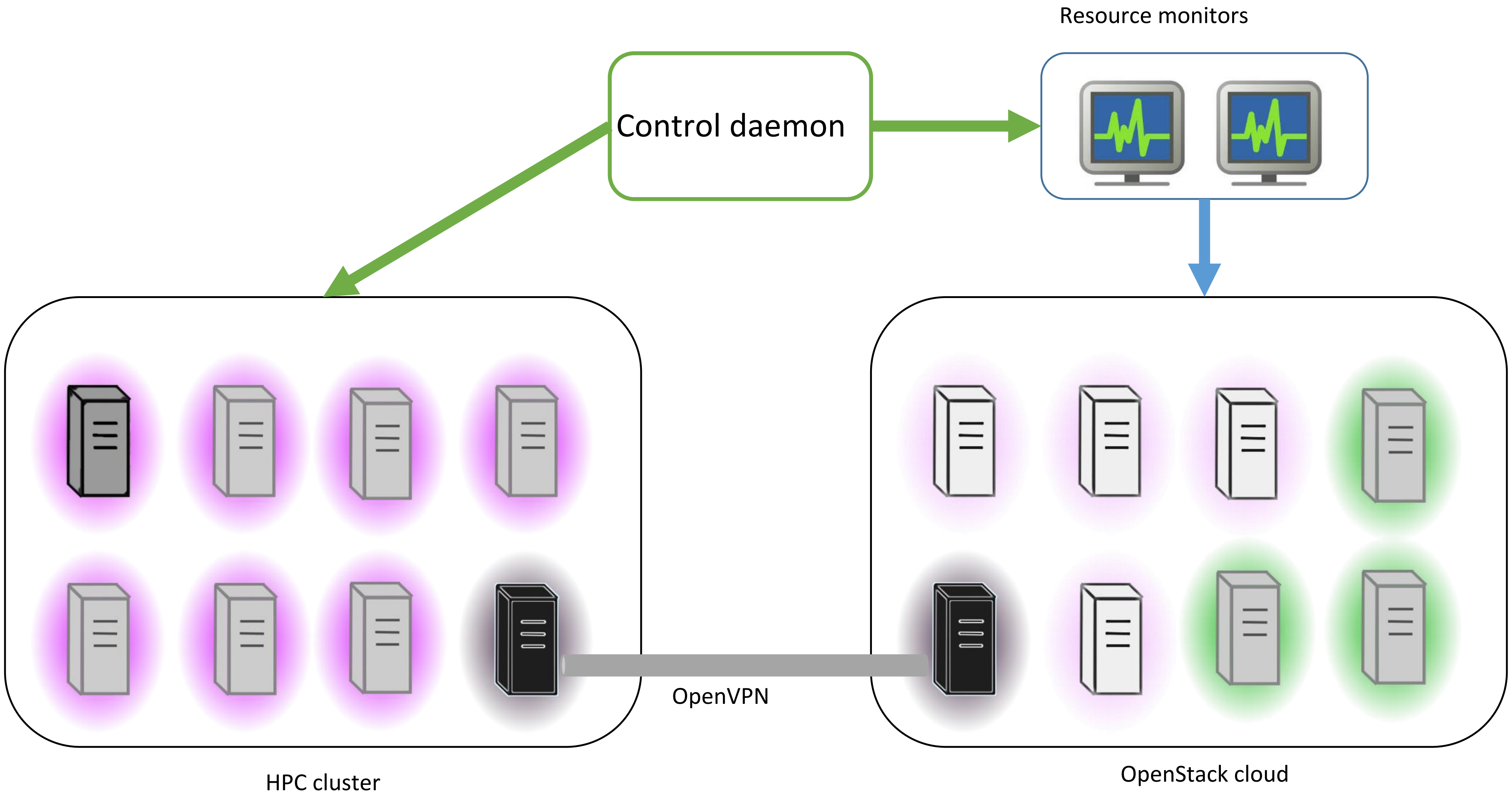
Cloud

- Overprovisioned to meet the peak workloads
- Underutilized most of the time with lots of idle resources

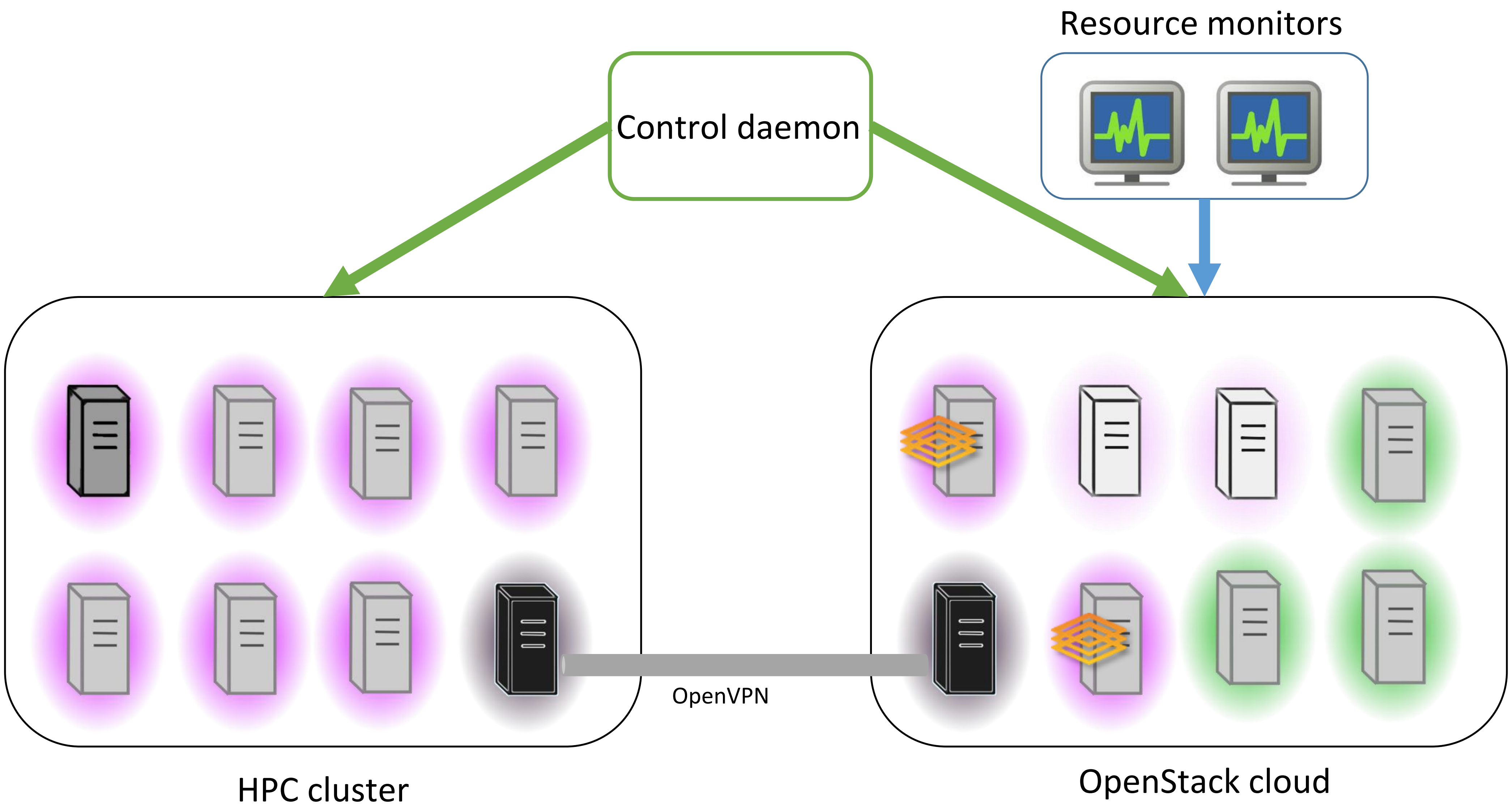
Focus first on HTC:

- Open Science Grid(OSG) High Throughput jobs: HPC cluster has OSG's HTC jobs backfilled to consume idle cycles
  - Kills HTC jobs for resources and requeues them
  - Leads to resource wastage and unproductive utilization

Architecture



Architecture


Resource monitors

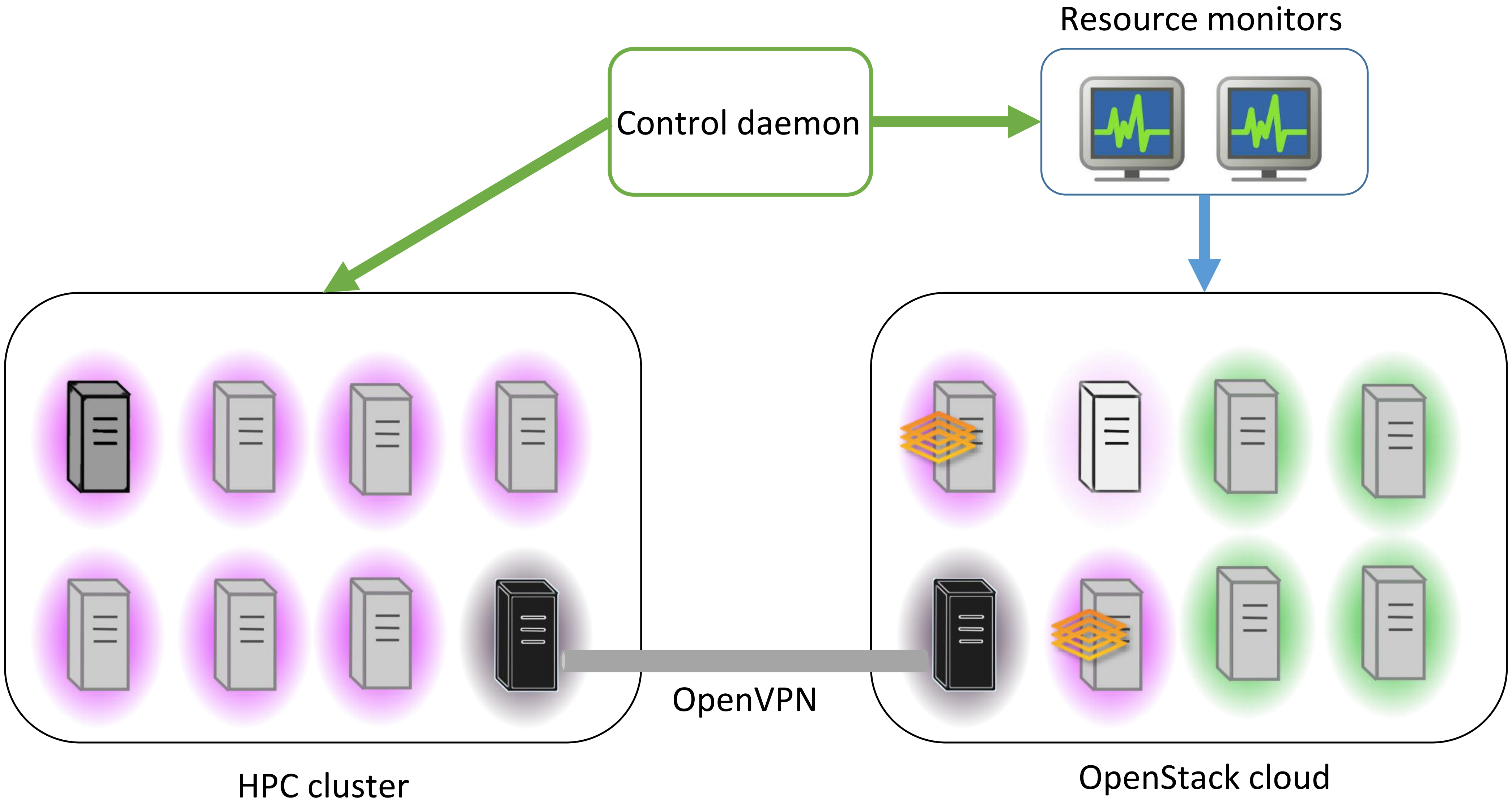
Control daemon

OpenVPN

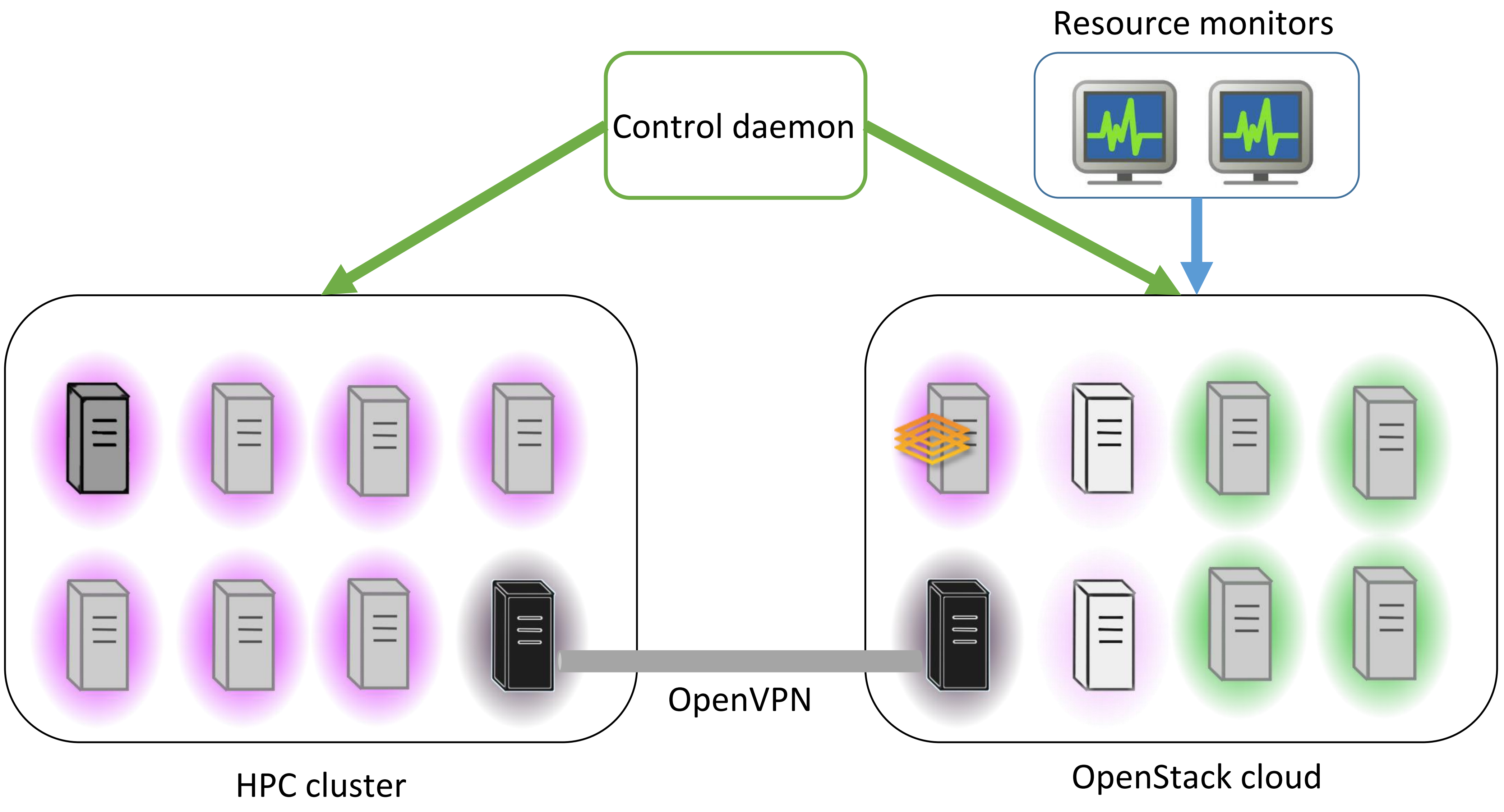
HPC cluster

OpenStack cloud

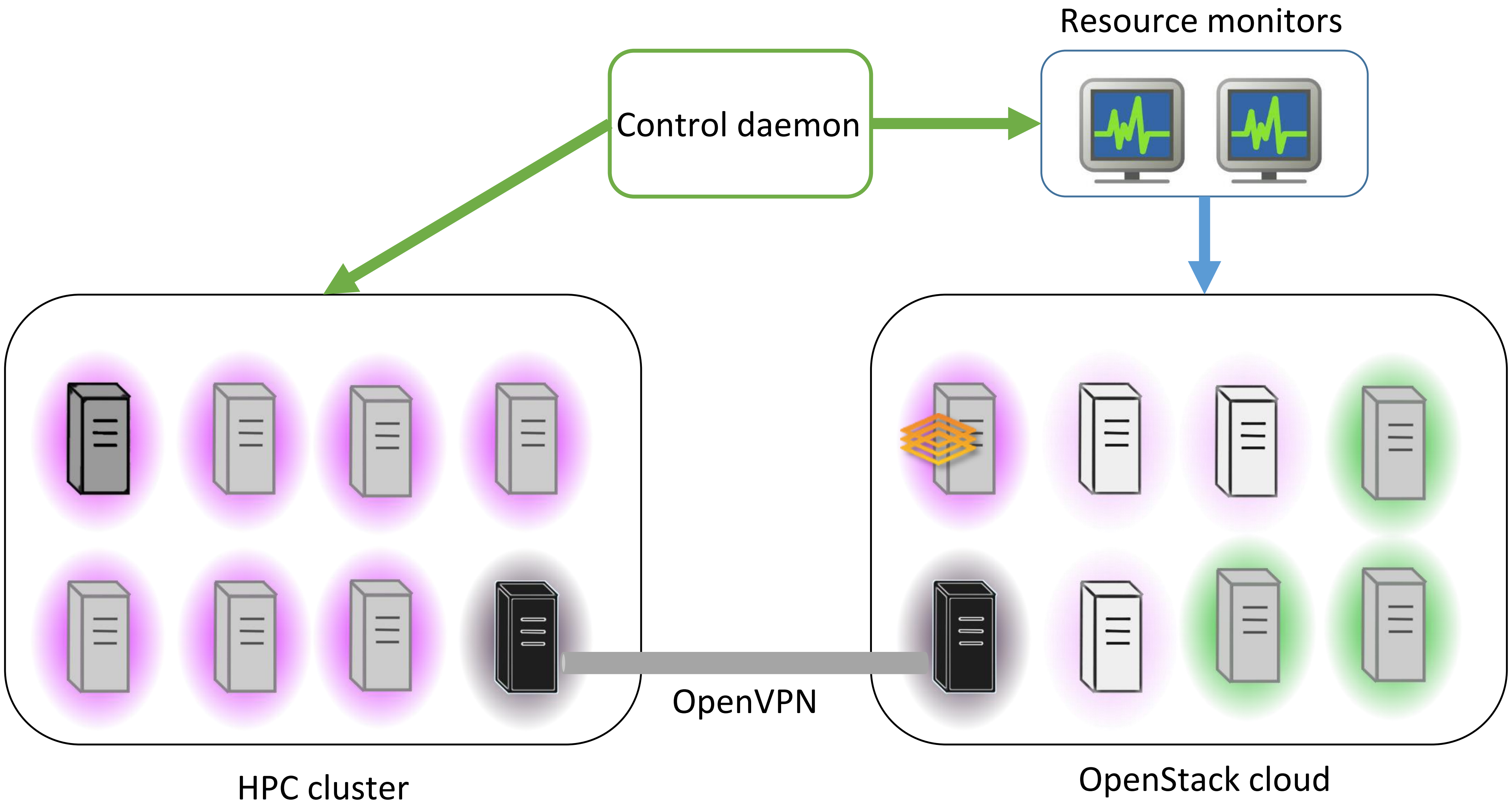
Architecture



Architecture

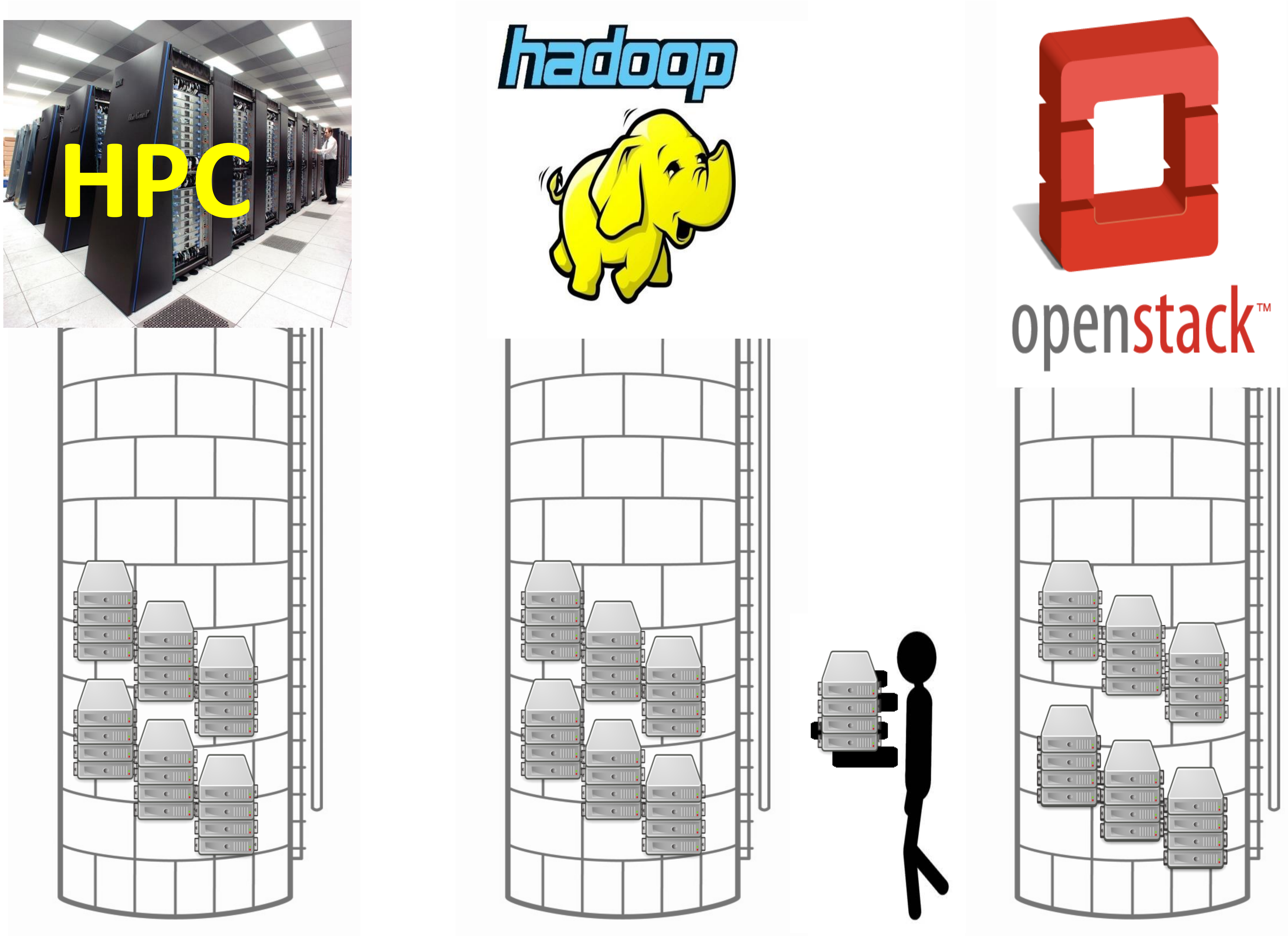


Architecture

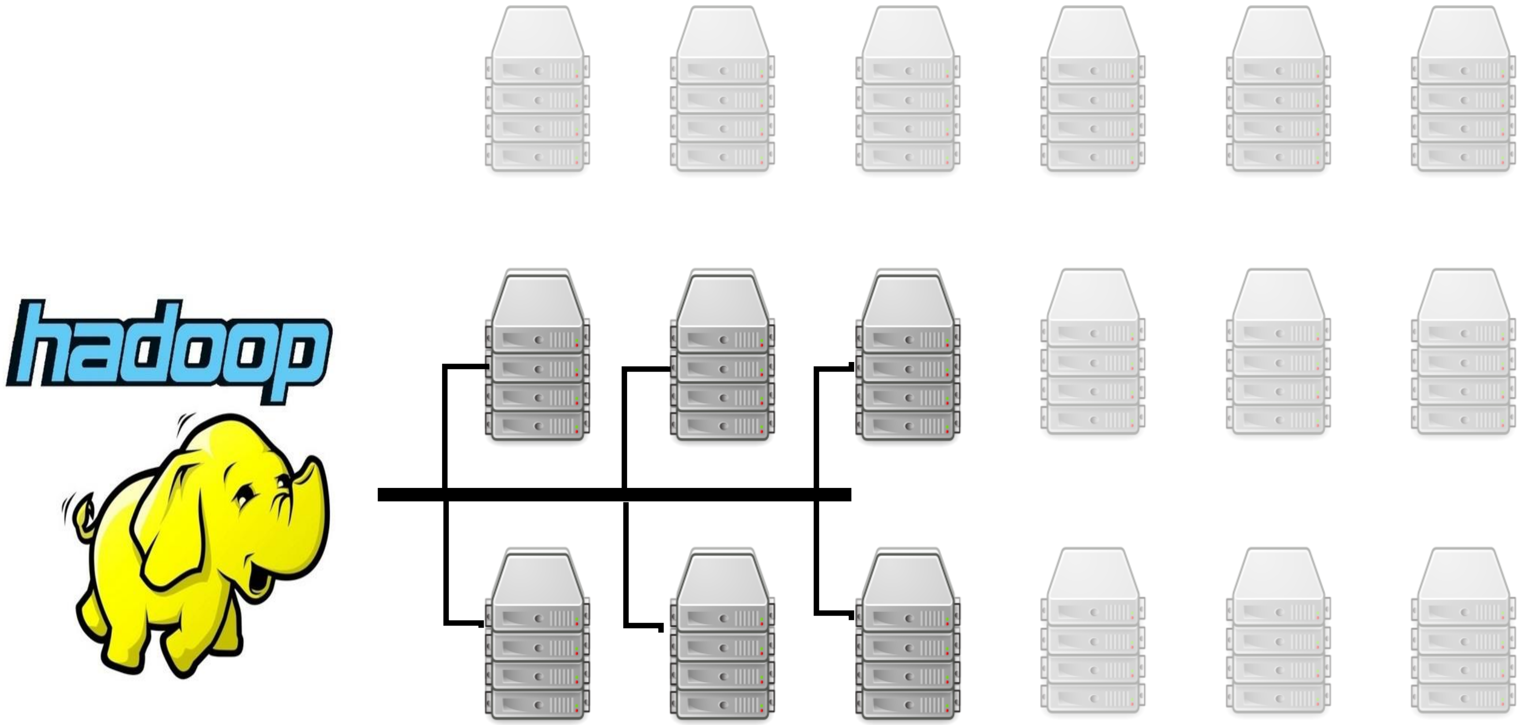


HIL: Hardware isolation layer

Datacenter has isolated silos



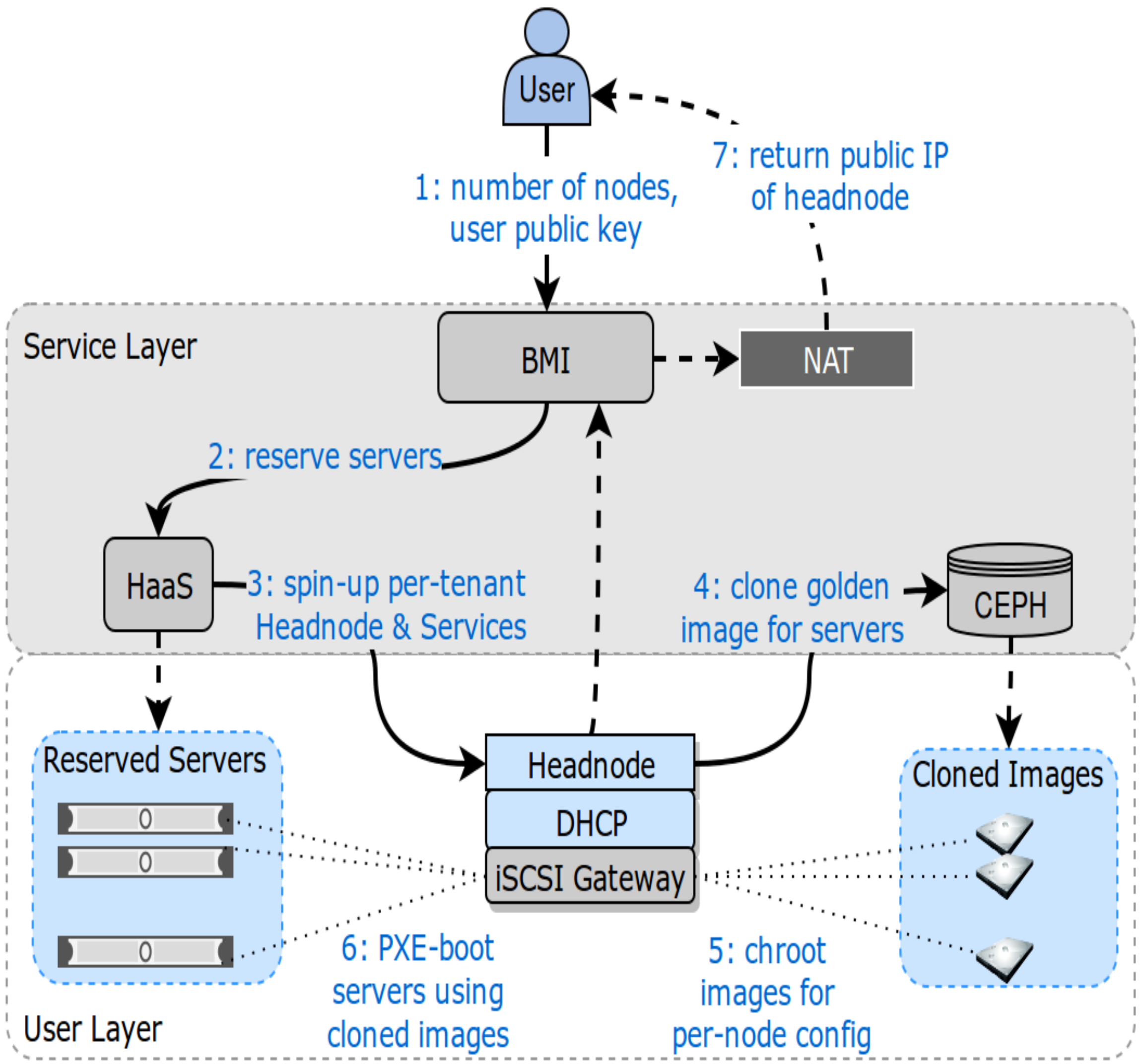
Hardware isolation layer



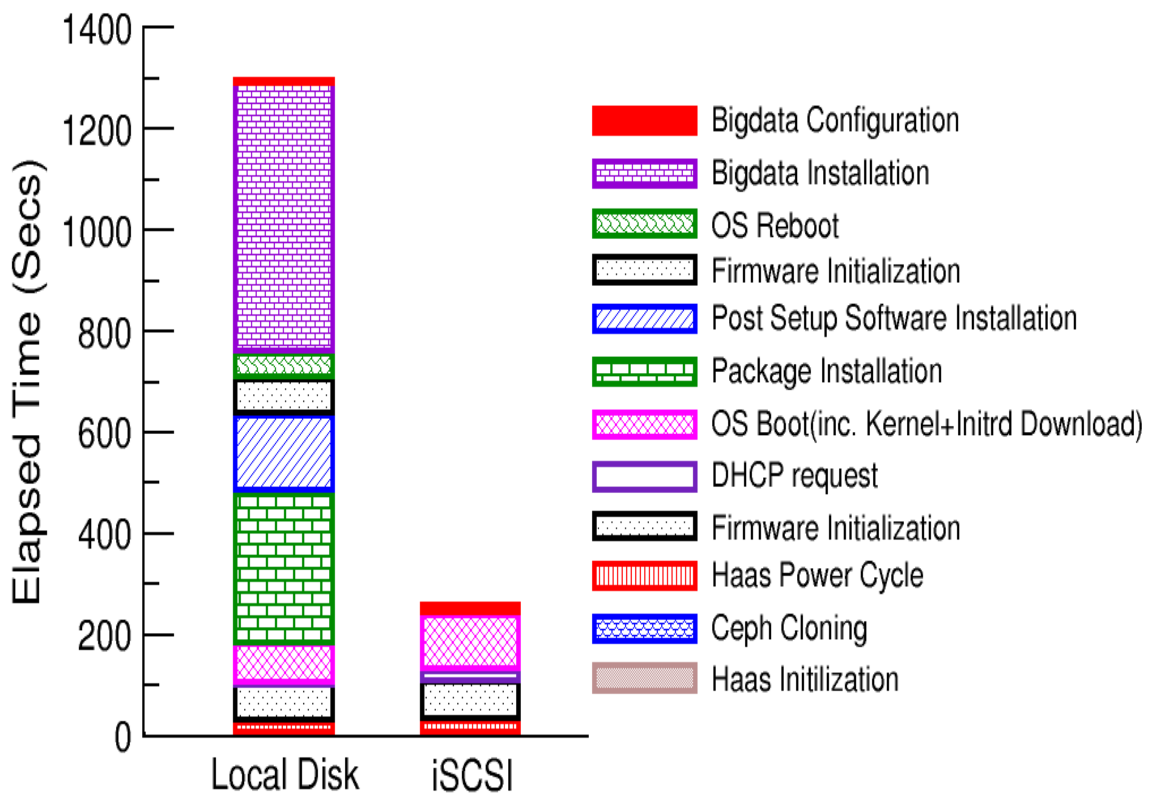
Connect nodes and networks

Bare Metal Imager:
VM-like disk image management

iSCSI-based

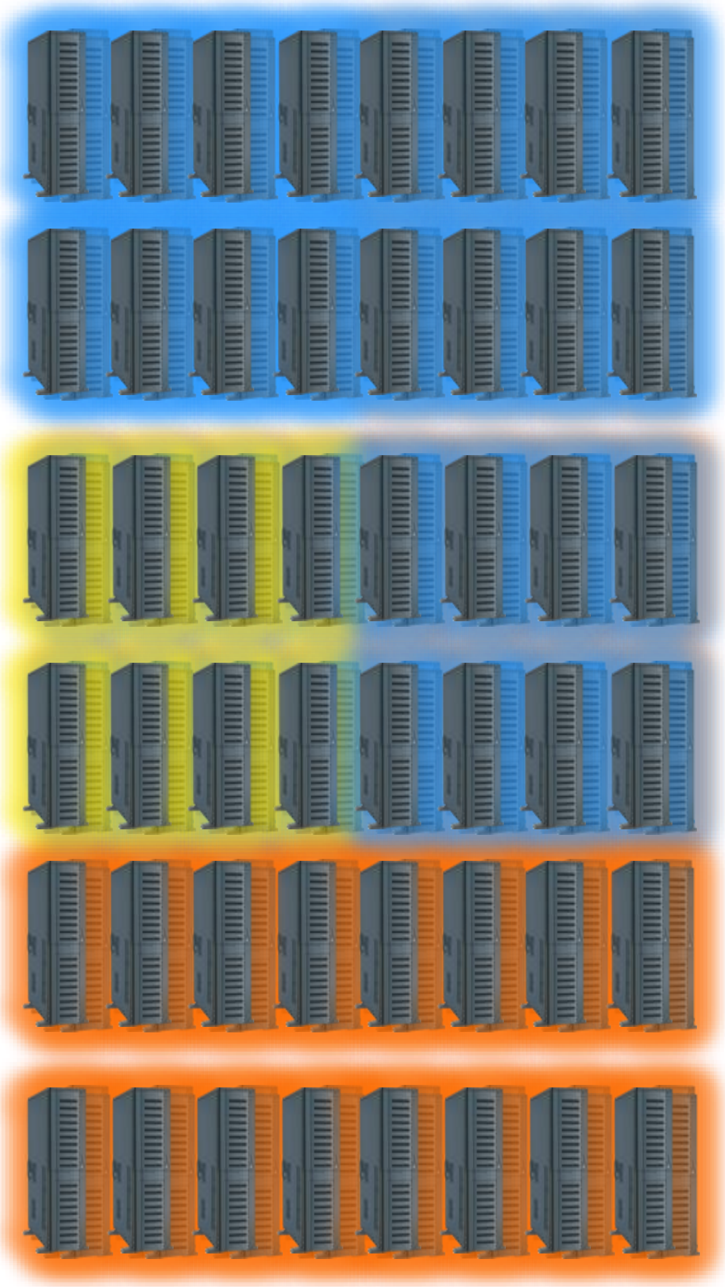

User
1: number of nodes, user public key
7: return public IP of headnode
Service Layer
BMI
NAT
2: reserve servers
HaaS
3: spin-up per-tenant Headnode & Services
4: clone golden image for servers
CEPH
Reserved Servers
Headnode
DHCP
iSCSI Gateway
Cloned Images
6: PXE-boot servers using cloned images
5: chroot images for per-node config
User Layer

Able to provision + boot in < 5 min



Turk, A., Gudimetla, R. S., Kaynar, E. U., Hennessey, J., Tikale, S., Desnoyers, P., & Krieger, O. (2016). An Experiment on Bare-Metal BigData Provisioning. In 8th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 16).

SLURM, PBS



Custom OS
(NeuroDebian?)

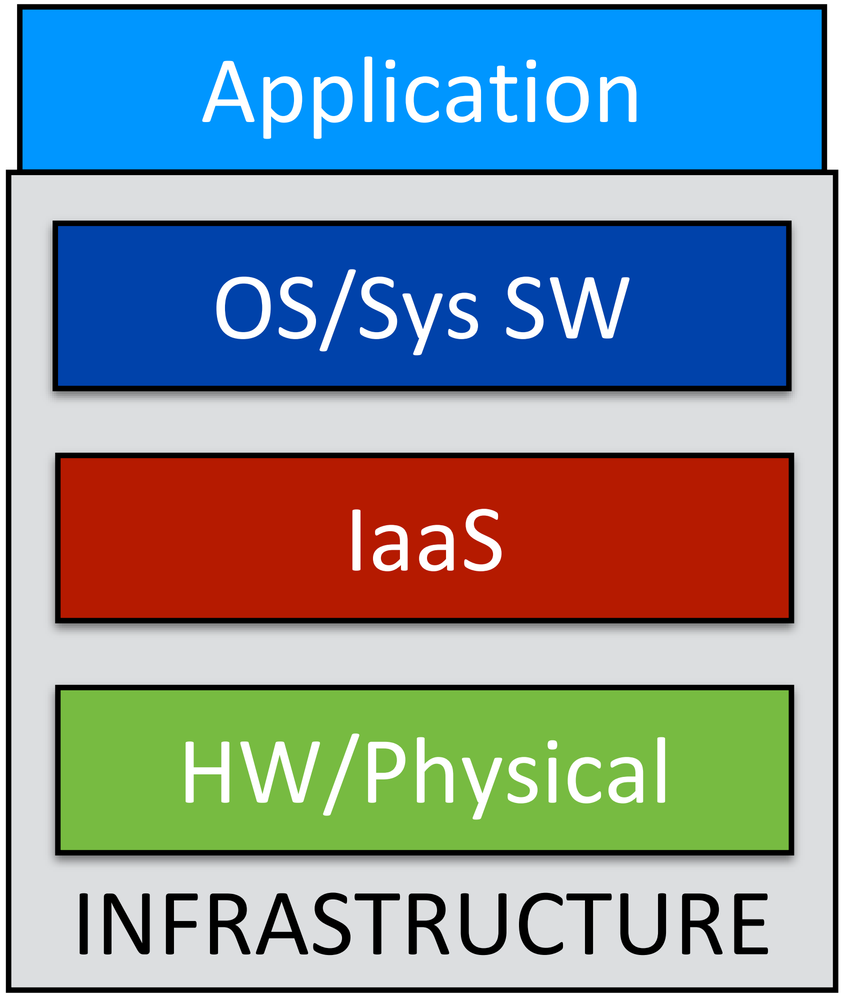
OpenStack

SESA and the MOC

Jonathan Appavoo, Han Dong, Jim Cadden, Dan Schatzberg

Imagine a cloud platform where a task can consume 1,000 CPUs for a few seconds…

- **S.E.S.A:** ***S**calable **E**lastic **S**ystem **S**oftware*
- System-level research
  - Previous work in multicore, datacenter-scale
- Highly-elastic software stacks
  - automatic and reactive resource allocation
  - usage sized proportionally to demand
- Opportunities arise by culmination of trends
  - Dense interconnection datacenter hardware
  - Fine-grain decomposition of cloud applications



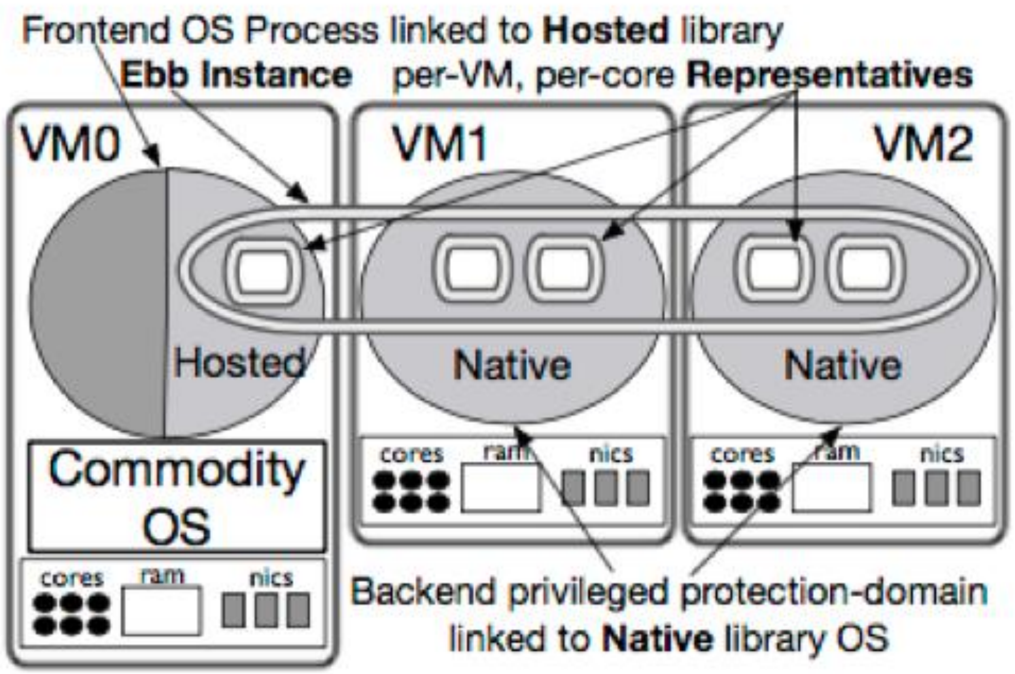
BOSTON UNIVERSITY

Scalable
Elastic
System
Architecture

EbbRT: Library OS for Cloud Computing

Elastic Building Block Runtime (EbbRT)

- a *framework* for building per-application library operating systems.



Frontend OS Process linked to **Hosted** library
Ebb Instance
per-VM, per-core **Representatives**
VM0
VM1
VM2
Hosted
Native
Native
Commodity OS
cores
ram
nics
cores
ram
nics
cores
ram
nics
Backend privileged protection-domain linked to **Native** library OS

Figure 1: High Level EbbRT architecture

- Elasticity as a first-order abstraction
- APIs for allocation, introspection and reactivity
- Fast-boot kernel, minimal coordination overheads
- Application events primitive unit of execution

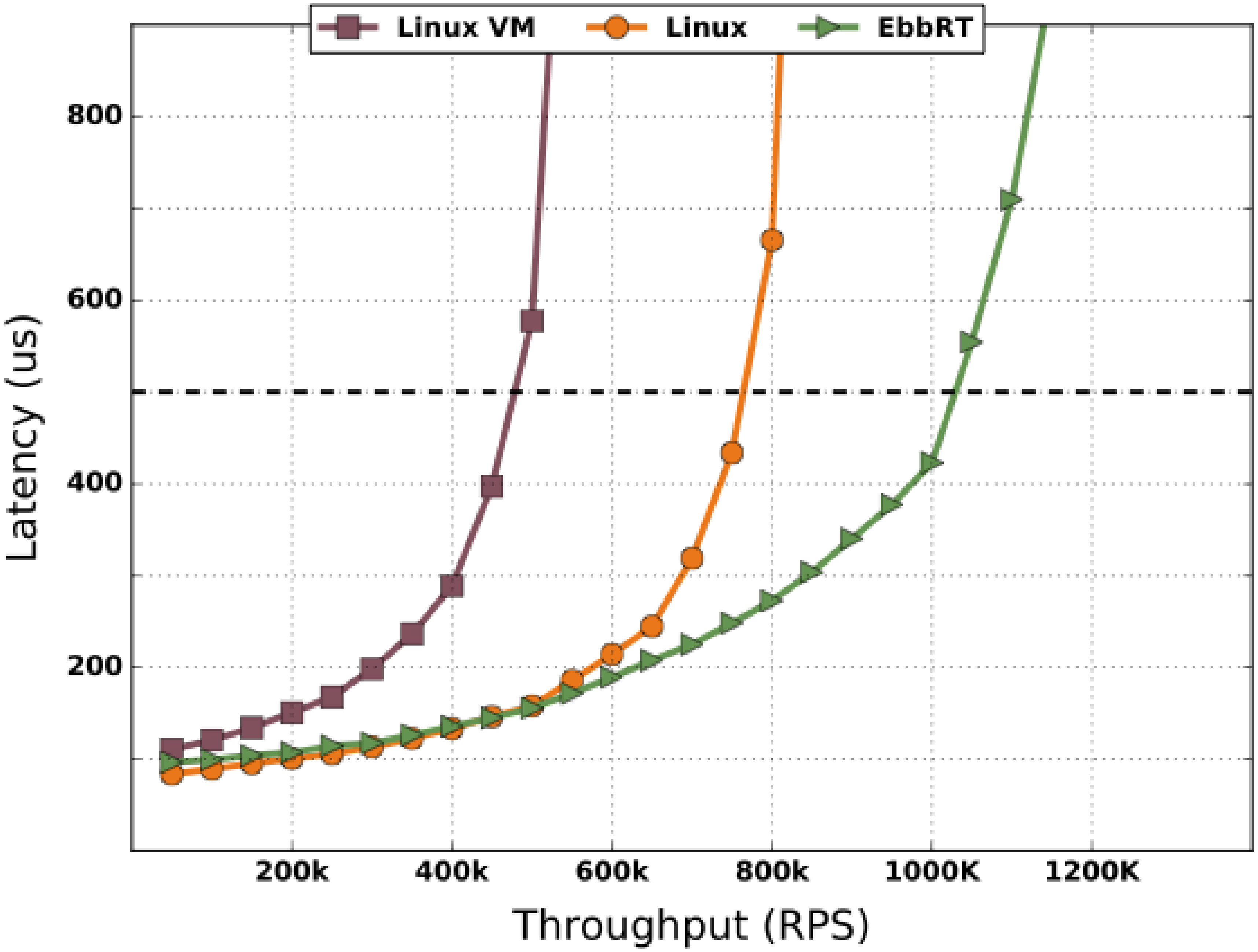
EbbRT: A Framework for Building Per-Application Library Operating Systems [OSDI 16]

`https://www.github.com/SESA/EbbRT`

BOSTON UNIVERSITY

Scalable
Elastic
System
Architecture

EbbRT: Construct customized environments for individual applications with reusable components.



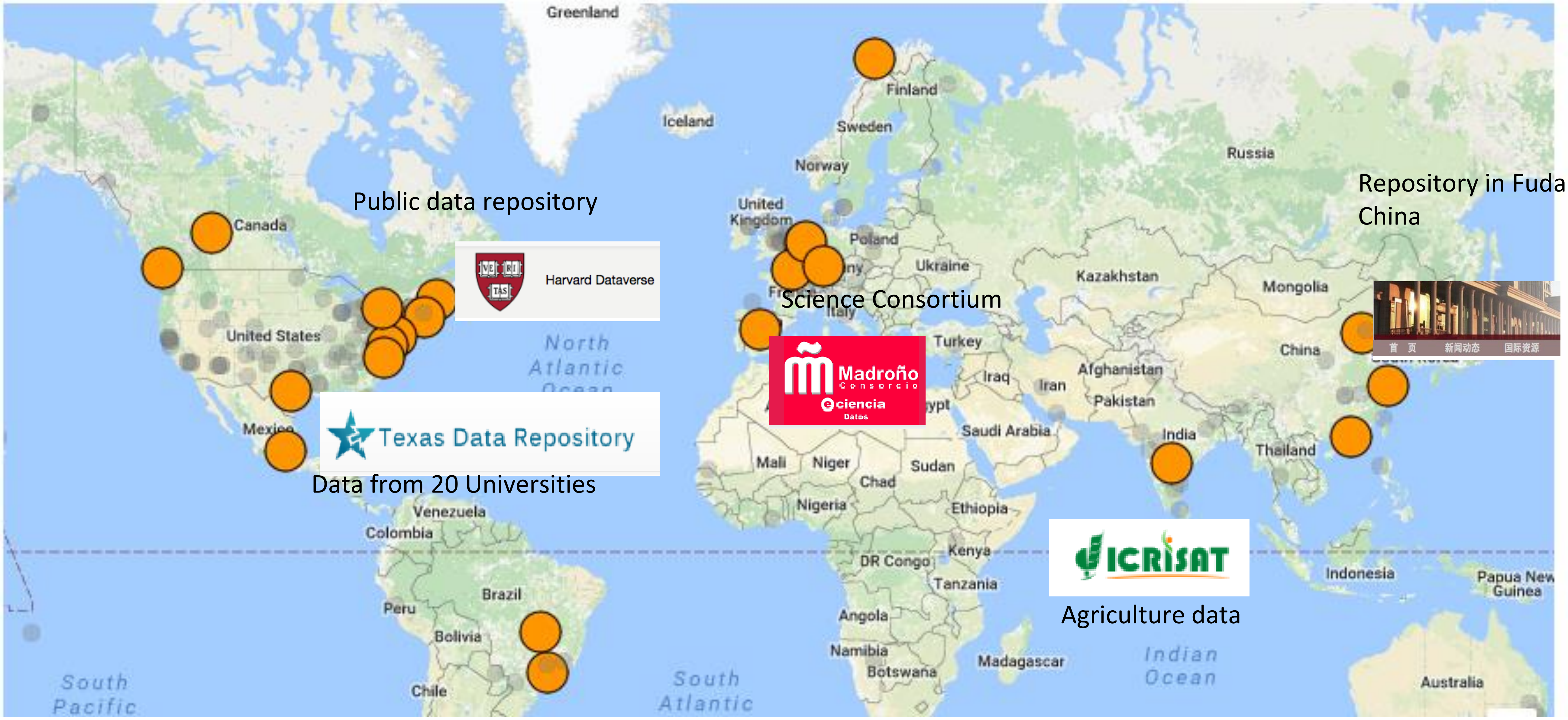
CLOUD DATAVERSE

The Dataverse Project

MOC

The Dataverse open-source platform

22 Installations 2,133 Dataverses 48,690 Datasets 2,400,322 Downloads



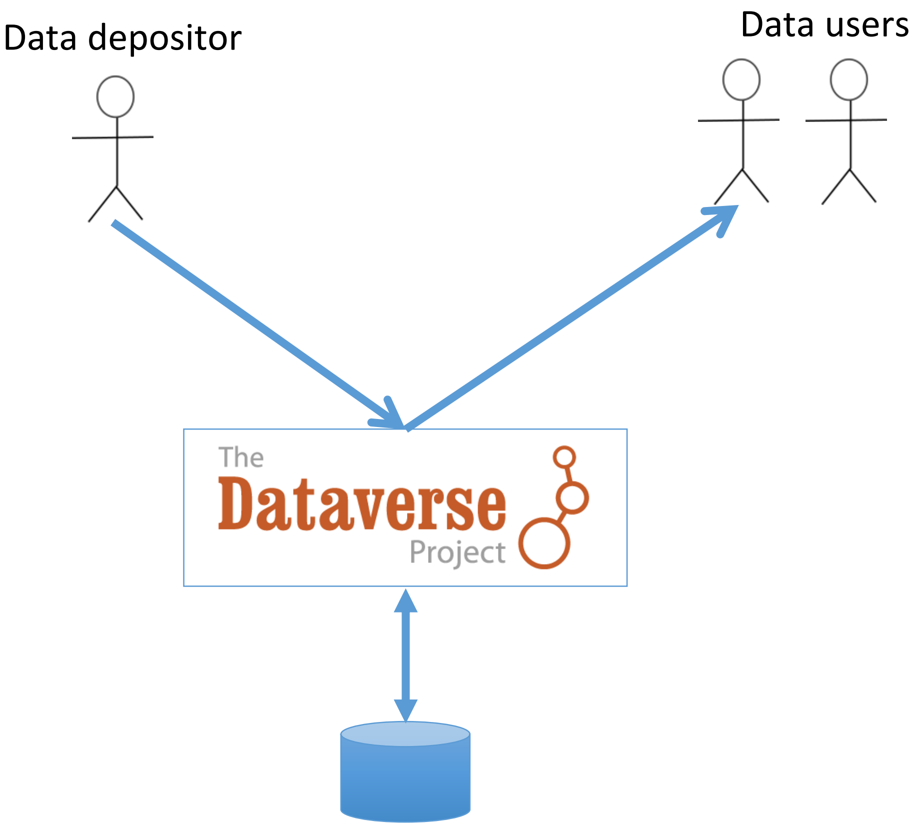
Public data repository

Data from 20 Universities

Science Consortium

Repository in Fudan, China

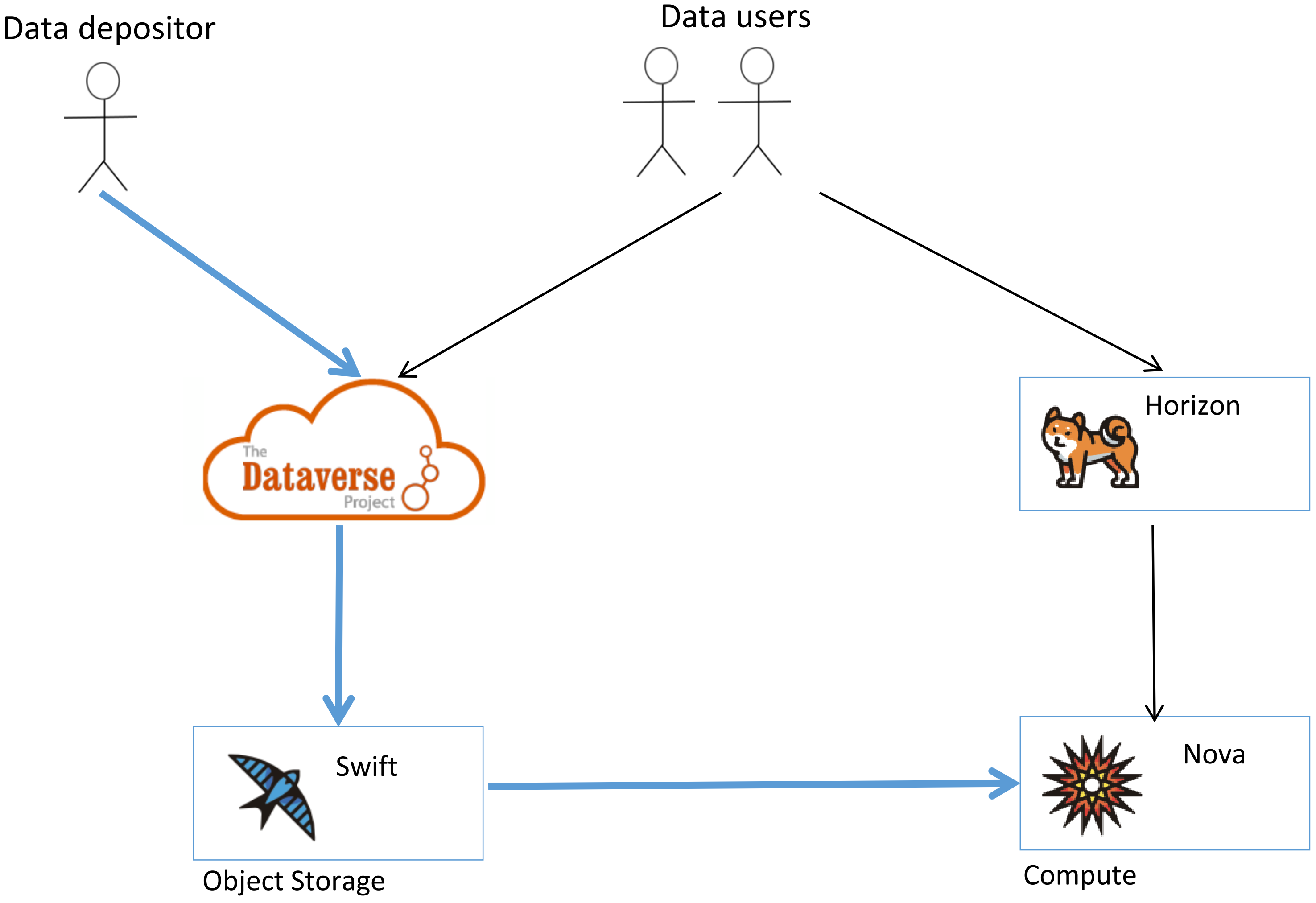
Agriculture data

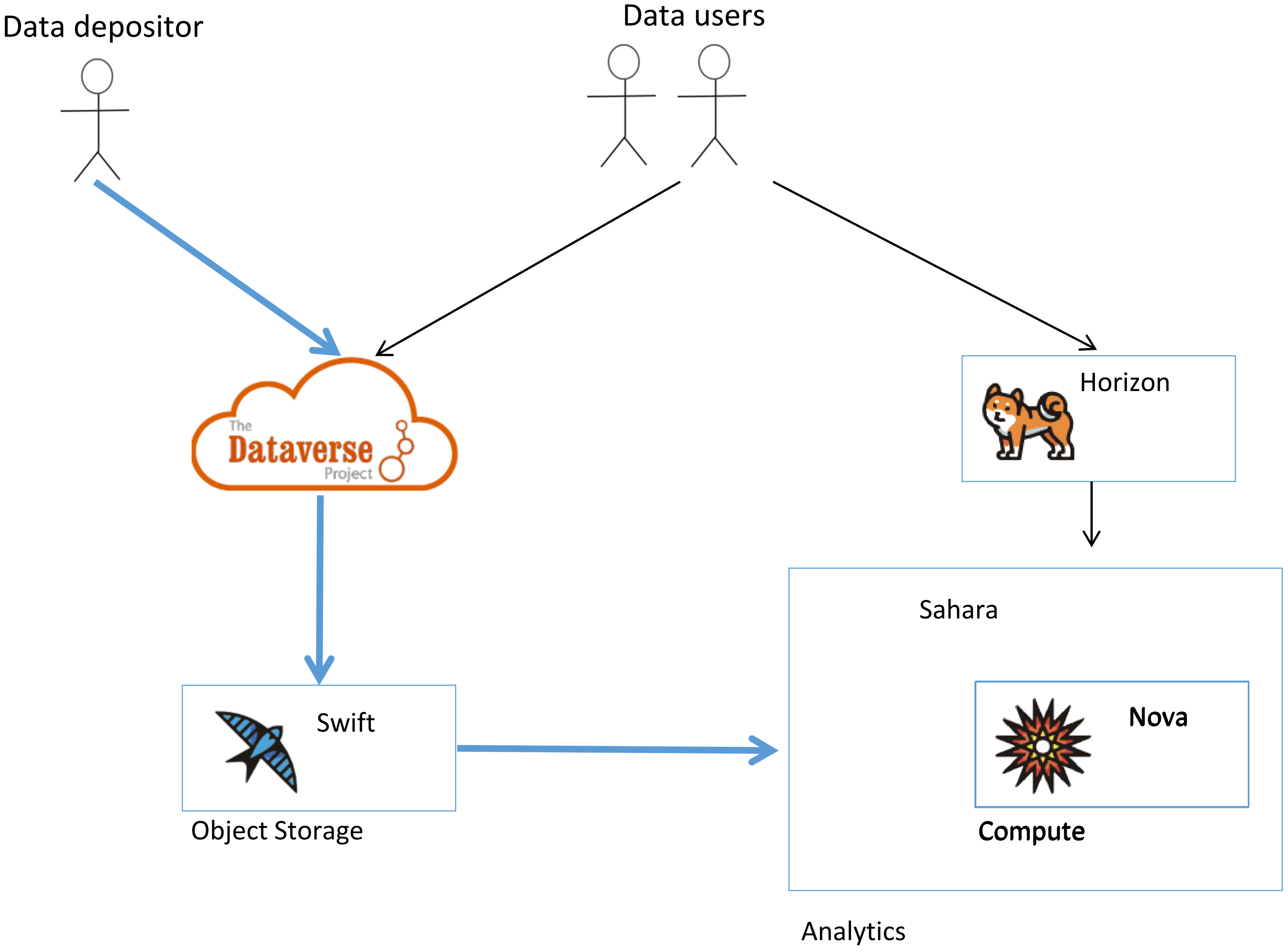


Data depositor

Data users

The Dataverse Project


Data depositor
Data users
The Dataverse Project
Horizon
Swift
Object Storage
Nova
Compute



Data depositor

Data users

The Dataverse Project

Horizon

Swift

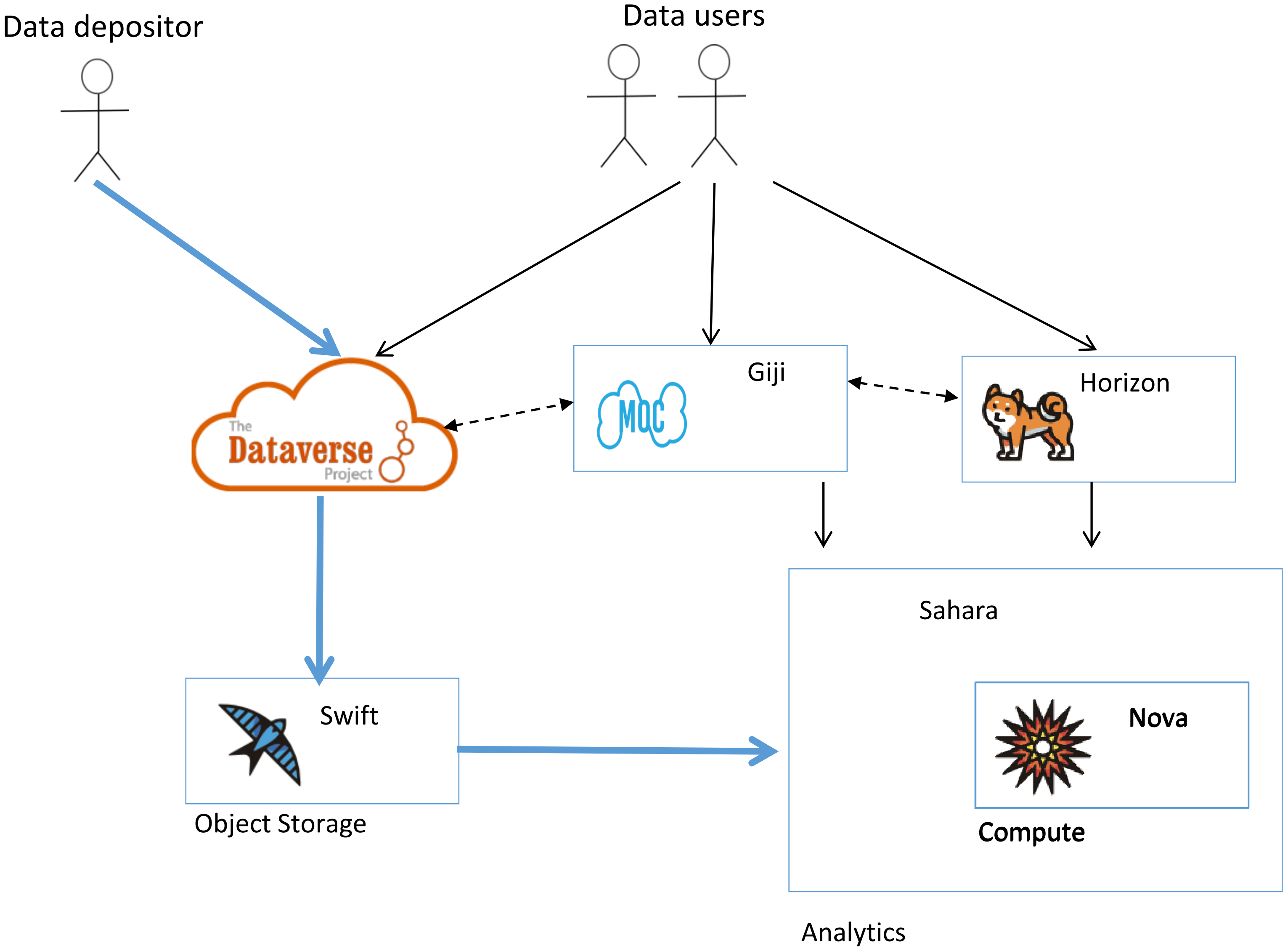
Object Storage

Sahara

Nova

Compute

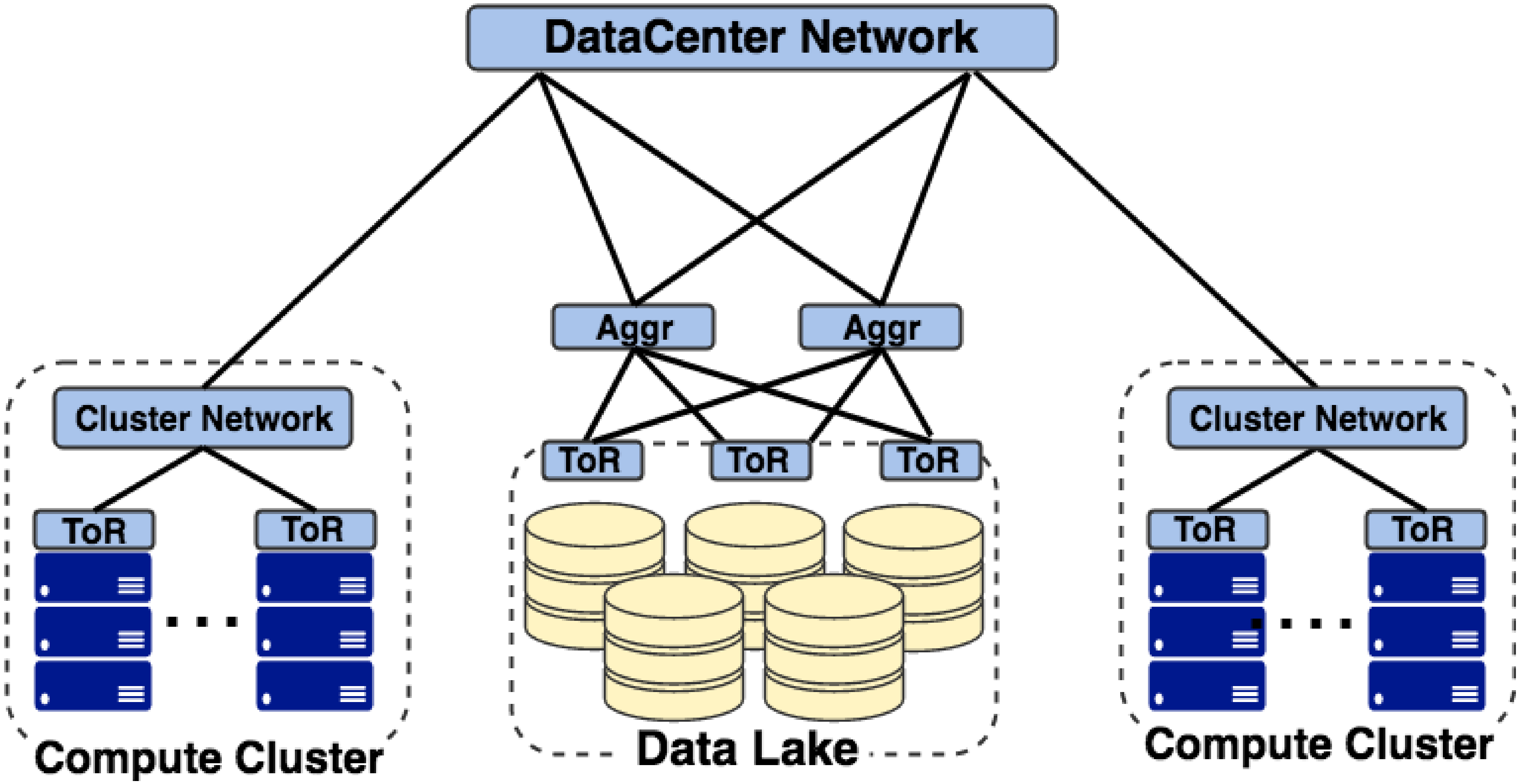
Analytics

Data depositor
Data users
Giji
Horizon
The Dataverse Project
Sahara
Swift
Nova
Object Storage
Compute
Analytics

Datacenter-scale Data Delivery Network (D3N)

MOC, Red Hat, Intel, Brocade, Lenovo, 2Sigma

Data Lake in a typical DC

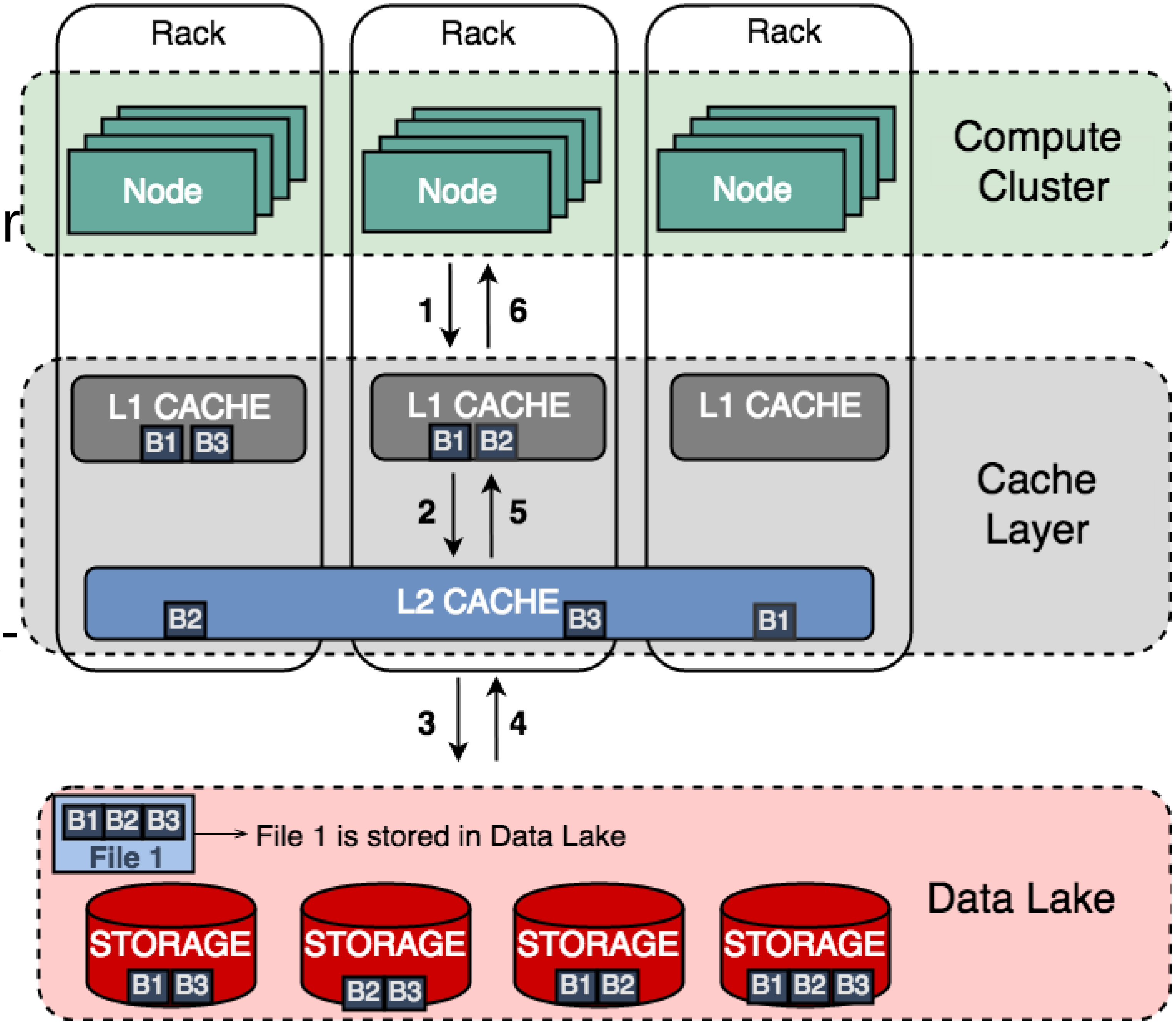


North Eastern Storage Exchange (NESE):
20+PB Harvard, NEU, MIT, BU, UMass

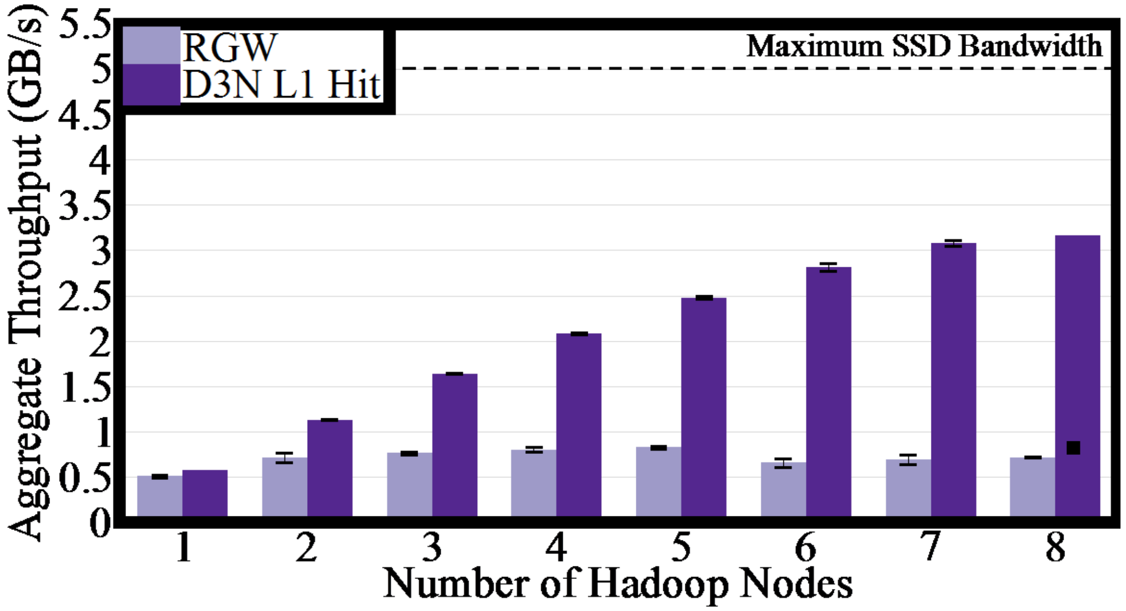
Datacenter scale Data Delivery Network (D3N)

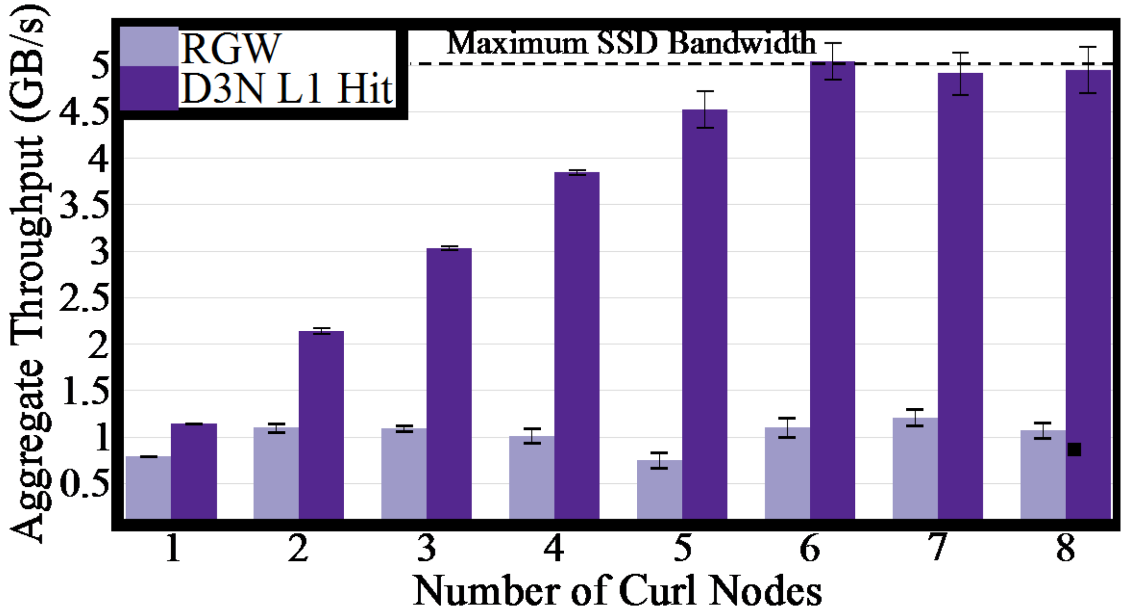
Simple deployment:

- Dedicated cache servers per rack
- **L1 :** Rack Local
  - reduce inter rack traffic
- **L2 :** Cluster Local
  - reduce clusters and back-end storage traffic
- Implemented by modifying **CEPH Rados Gateway**



D3N Results





- Exceeds maximum bandwidth Hadoop
- Demonstrates makes sense to share expensive SSDs – faster than local disk
- With extreme benchmark can saturate SSD & 40 Gb NIC
- Will be of enormous value with NESE data lake

Red Hat Collaboratory

- Mix & Match
- HIL & BMI (and QUADS integration)
- Big Data Analytics and Cloud Dataverse
- Datacenter-scale Data Delivery Network (D3N)
- Monitoring, Tracing, Analytics …
- OpenShift on the MOC
- Accelerator Testbed

Red Hat Collaboratory

- Mix & Match
- HIL & BMI (and QUADS integration)
- Big Data Analytics and Cloud Dataverse
- Datacenter-scale Data Delivery Network (D3N)
- Monitoring, Tracing, Analytics …
- OpenShift on the MOC
- Accelerator Testbed

End-to-end POC: Radiology in the cloud targeting OpenShift with accelerators

Final remarks

- HPC, Cloud, Big Data convergence
- Research has to go on in the context production cloud:
  - Scale is fundamental, need to tie innovation to a marketplace with real users, with real data sets, applications, with companies making money
- We have the opportunity to lead the convergence:
  - MGHPCC data center & established collaboration IT teams
  - MOC, Engage1, Dataverse, NESE
  - Rich Industry partnerships
  - Huge range of researchers in machine learning, FPGAs, HPC, Security, operating systems…
- Cloud means elasticity, economics, scale, broad applicability, industry, startups