Virtualization: Control Your Stack

Rick Wagner
HPC Systems Manager
rpwagner@sdsc.edu
SDSC Summer Institute 2016

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Staffing

SDSC: Project management, system managements, systems software (Nucleus)

IU: User support, client software (Cloudmesh)

Virtual Clusters

Goal:

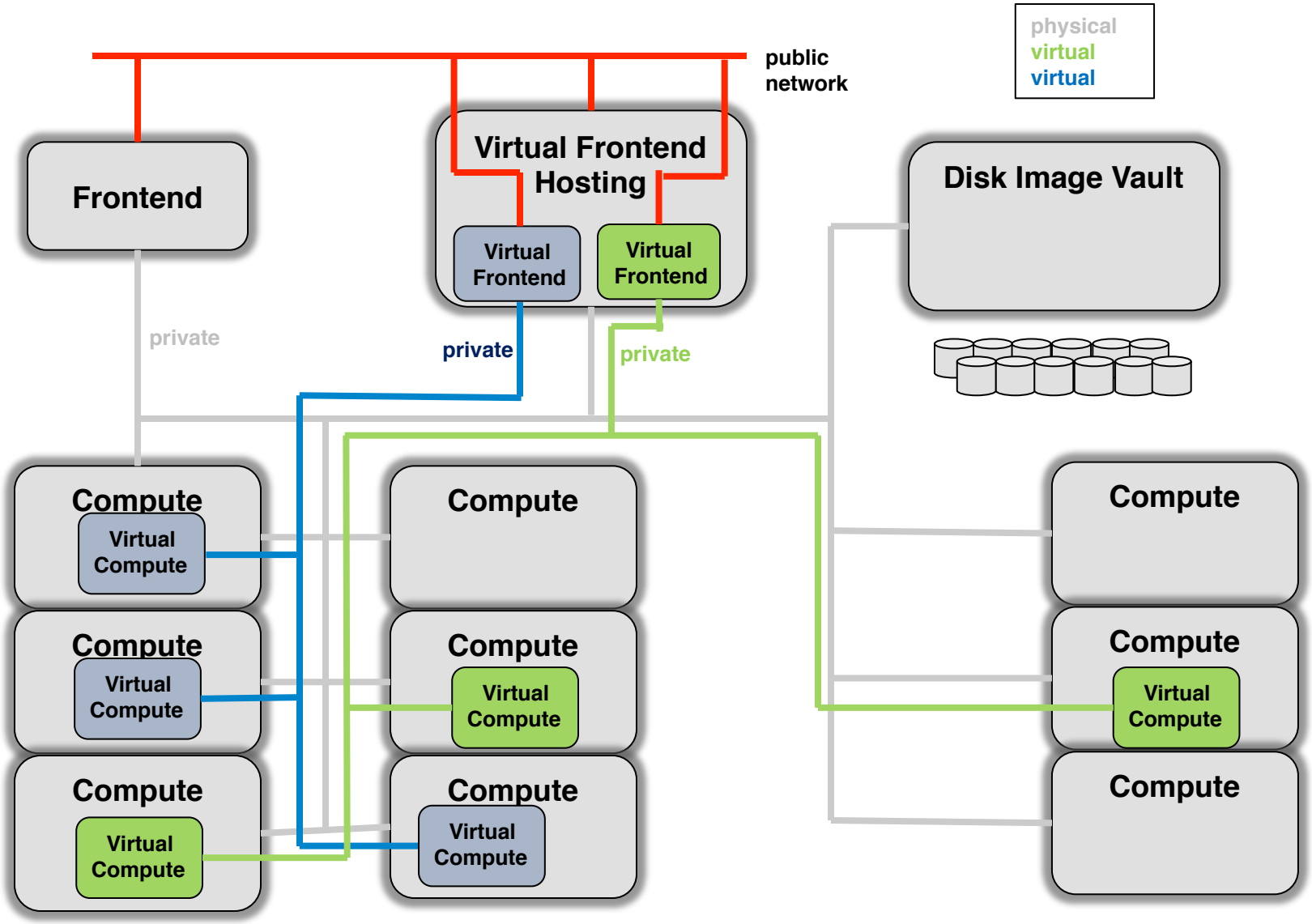
Provide a near bare metal HPC performance and management experience

Target Use

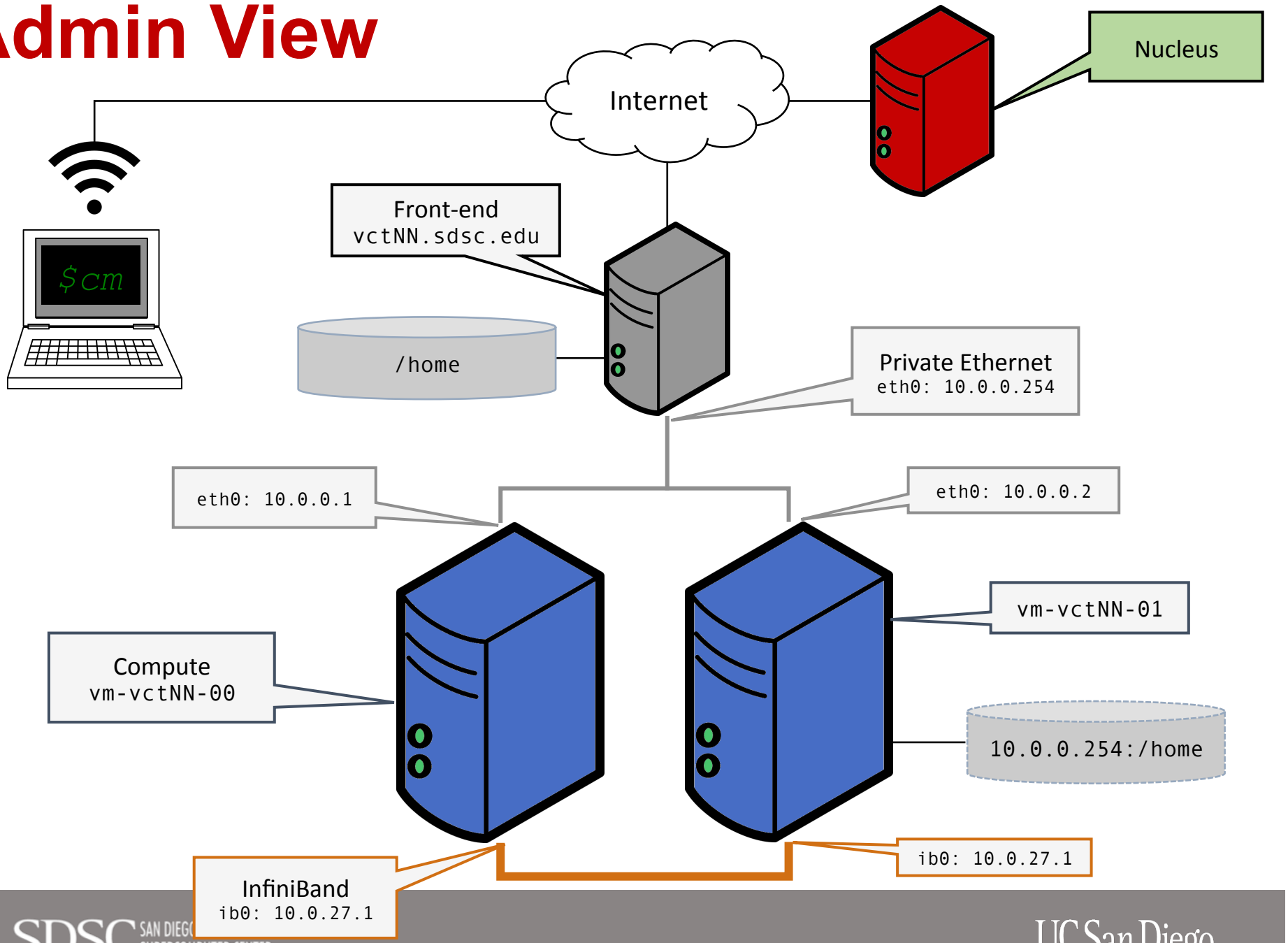
Projects that can manage their own cluster, and:

- can't fit our batch environment, and
- don't want to buy hardware or
- have bursty or intermittent need

User-Customized HPC

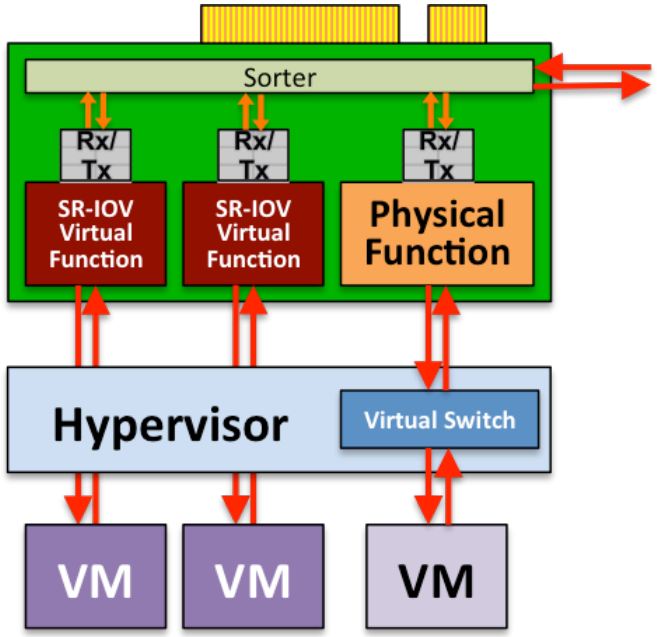


Admin View


Nucleus
Internet
Front-end
vctNN.sdsc.edu
$cm
/home
Private Ethernet
eth0: 10.0.0.254
eth0: 10.0.0.1
eth0: 10.0.0.2
vm-vctNN-01
Compute
vm-vctNN-00
10.0.0.254:/home
ib0: 10.0.27.1
InfiniBand
ib0: 10.0.27.1

Single Root I/O Virtualization in HPC

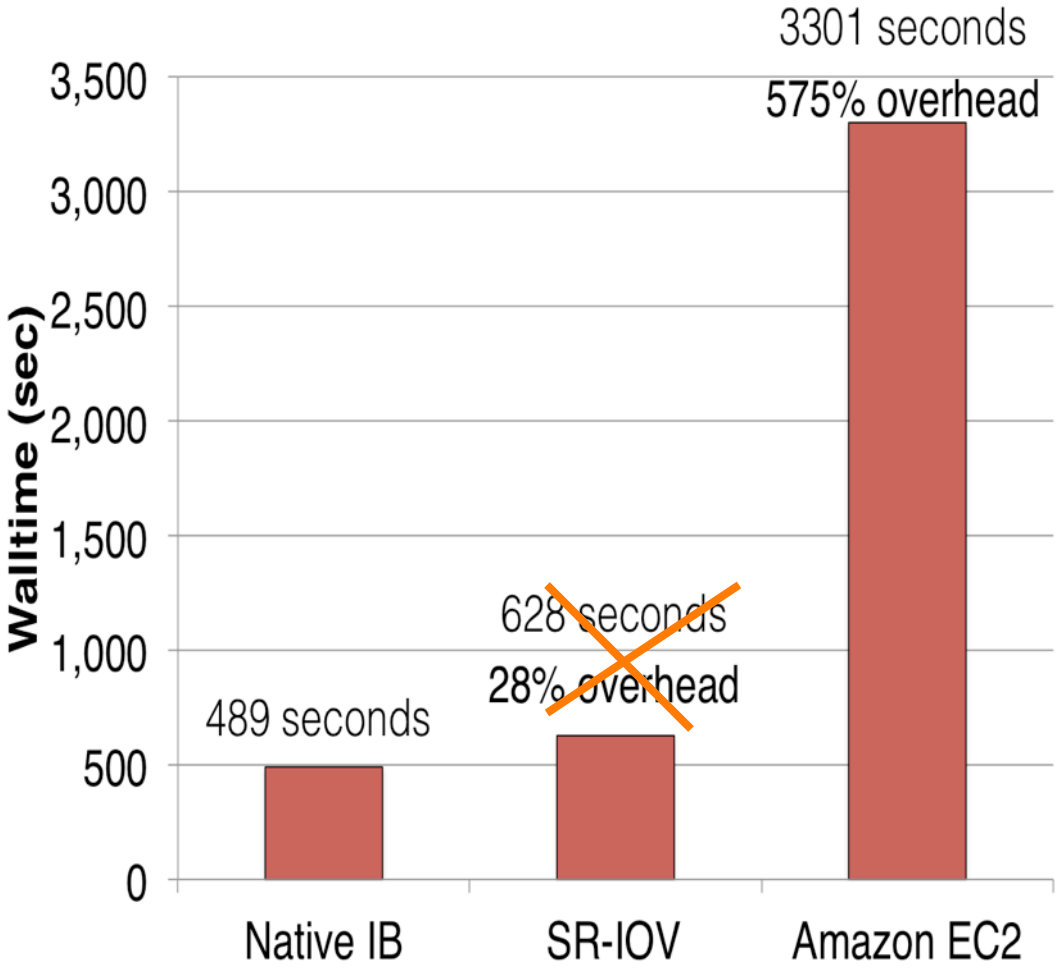
- Problem: Virtualization generally has resulted in significant I/O performance degradation (e.g., excessive DMA interrupts)
- Solution: SR-IOV and Mellanox ConnectX-3 InfiniBand host channel adapters
  - One physical function → multiple virtual functions, each light weight but with its own DMA streams, memory space, interrupts
  - Allows DMA to bypass hypervisor to VMs
- *SRIOV enables virtual HPC cluster w/ near-native InfiniBand latency/bandwidth and minimal overhead*



Mellanox TECHNOLOGIES

Quantum ESPRESSO: ~~28%~~ 8% (!) Overhead

- 48-core (3 node) calculation
- CG matrix inversion - irregular communication
- 3D FFT matrix transposes (all-to-all communication)
- ~~28% slower w/ SR-IOV vs native IB~~
- 8% slower w/ SR-IOV vs native IB!
- SR-IOV still > 500% faster than EC2 Despite 20% slower CPUs

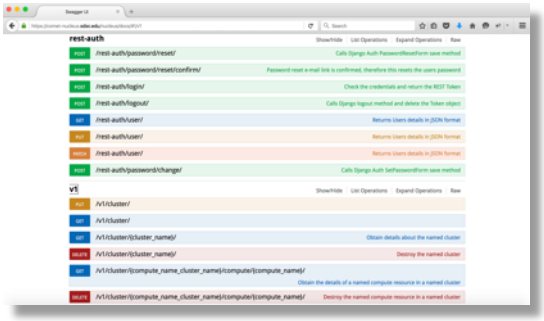


Quantum Espresso 5.0.2 – DEISA AUSURF112 benchmark

Selected Technologies: Enabling Major Impact

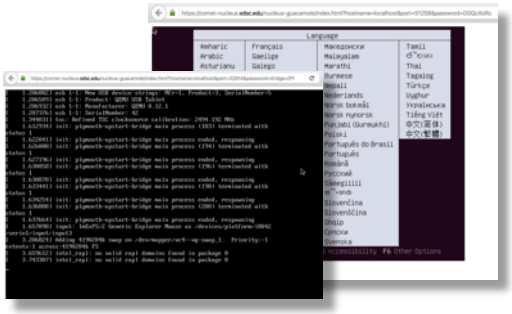
- KVM—Let's us run virtual machines (all processor features)
- SR-IOV—Makes MPI go fast on VMs
- Rocks—Systems management
- ZFS—Disk image management
- VLANs—Isolate virtual cluster management network
- pkeys—Isolate virtual cluster IB network
- **Nucleus—Coordination engine (scheduling, provisioning, status, etc.)**
- **Client – Cloudmesh**

Accessing Comets Virtual Cluster Capabilities



- REST API
- Command line interface
- Command shell for scripting
- Console Access
- (Portal)

```
Usage:
   comet ll [CLUSTERID] [--format=FORMAT]
   comet cluster [CLUSTERID]
                 [--format=FORMAT]
   comet computeset [COMPUTESETID]
   comet power on CLUSTERID [--count=NUMNODES] [NODESPARAM]
                [--allocation=ALLOCATION]
                [--walltime=WALLTIME]
   comet power (off|reboot|reset|shutdown) CLUSTERID [NODESPARAM]
   comet console CLUSTERID [COMPUTENODEID]
   comet image list
   comet image upload [--imagename=IMAGENAME] PATHIMAGEFILE
   comet image attach IMAGENAME CLUSTERID [COMPUTENODEID]
   comet image detach CLUSTERID [COMPUTENODEID]
   comet node rename CLUSTERID OLDNAME NEWNAME
```



```
(ENV)big:client big$ cm help

Documented commands (type help <command>):
========================================
EOF       comet    group      key       open     refresh      server   var
banner    context  h          launcher  pause    register     shell    verbose
check     debug    help       limits    portal   reservation  ssh      version
clear     default  history    list      py       reset        submit   vm
cloud     echo     hpc        man       q        rsync        sync
cluster   exec     image      network   quit     secgroup     timer
color     flavor   inventory  nova      quota    select       usage
```

User does NOT see: Rocks, Slurm, etc.

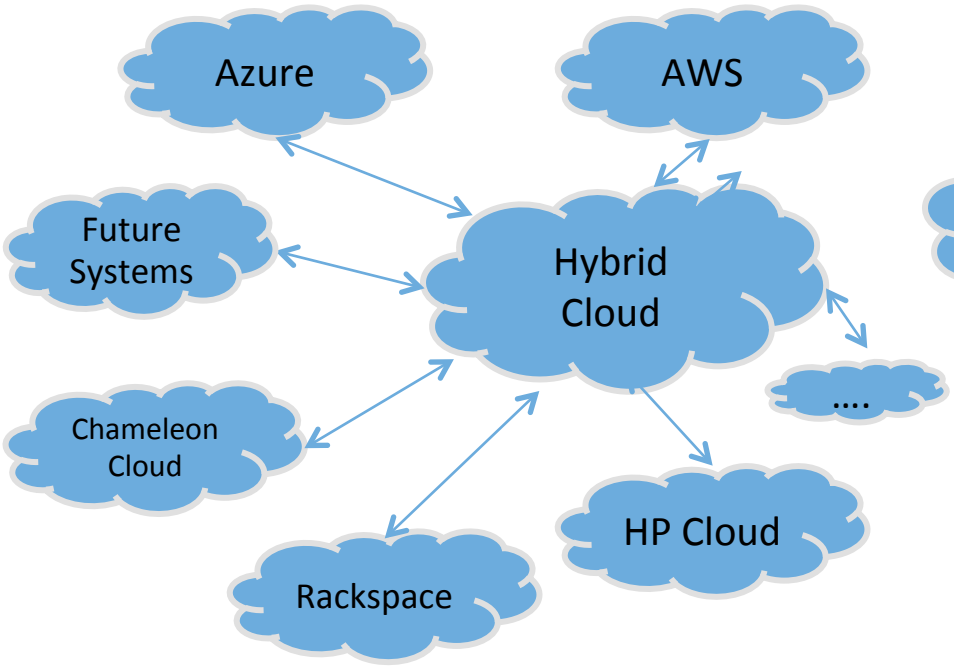
Cloudmesh

Hybrid Cloud Management

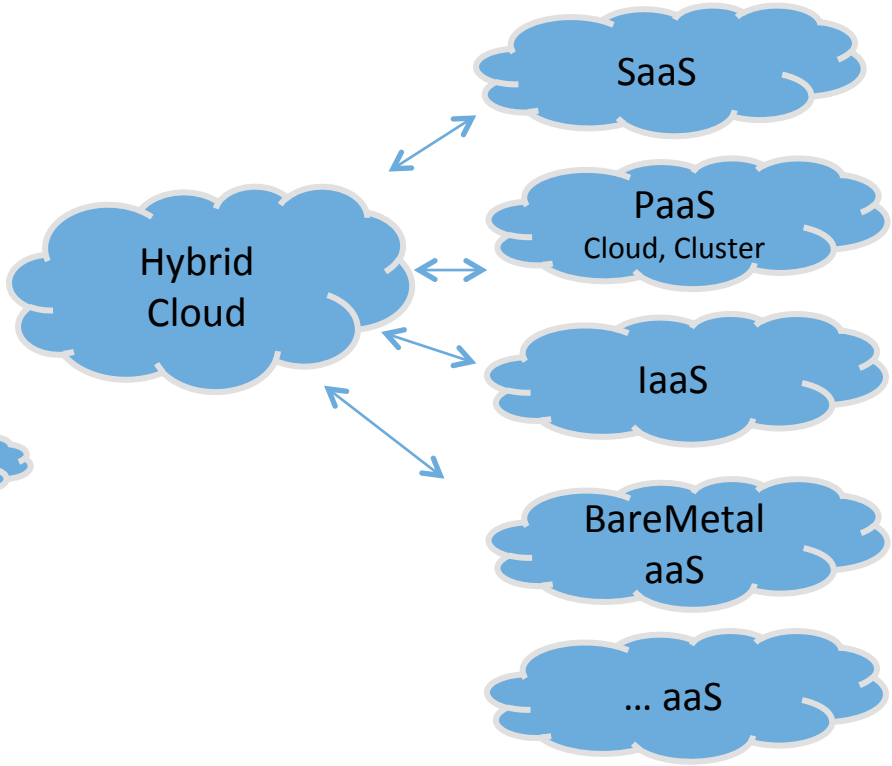
http://cloudmesh.github.io/client/

Integrated Clouds

Capacity Integration

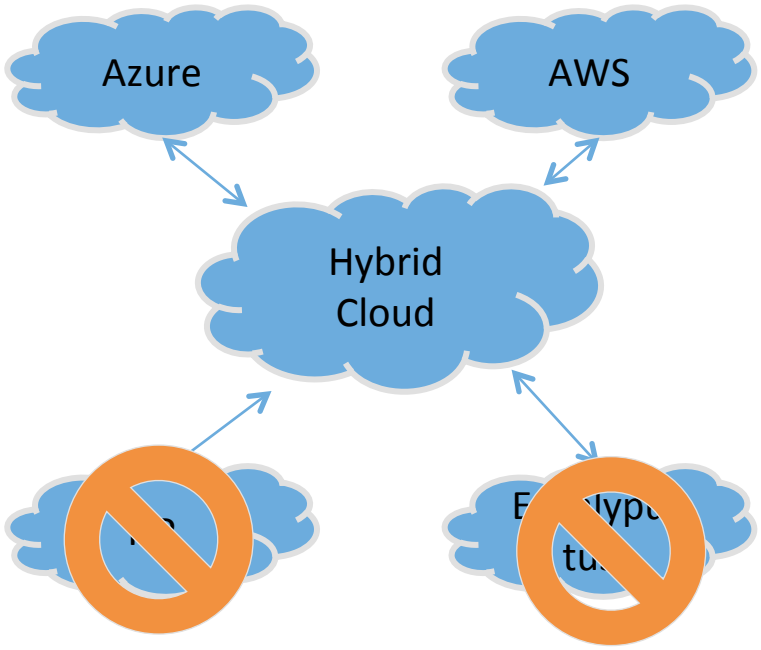


Technology Integration

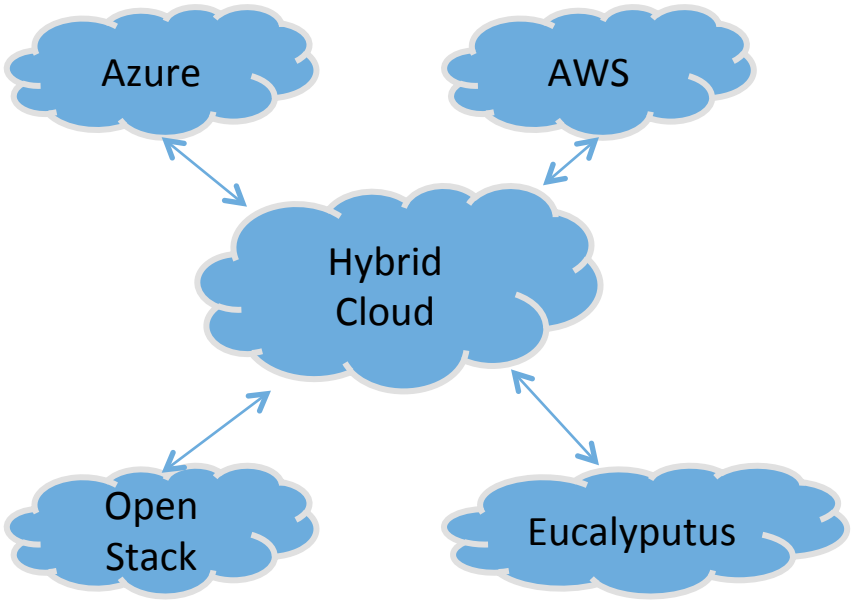


Why are Integrated Multi-Clouds Useful …

Technology Robustness



Technology Federation



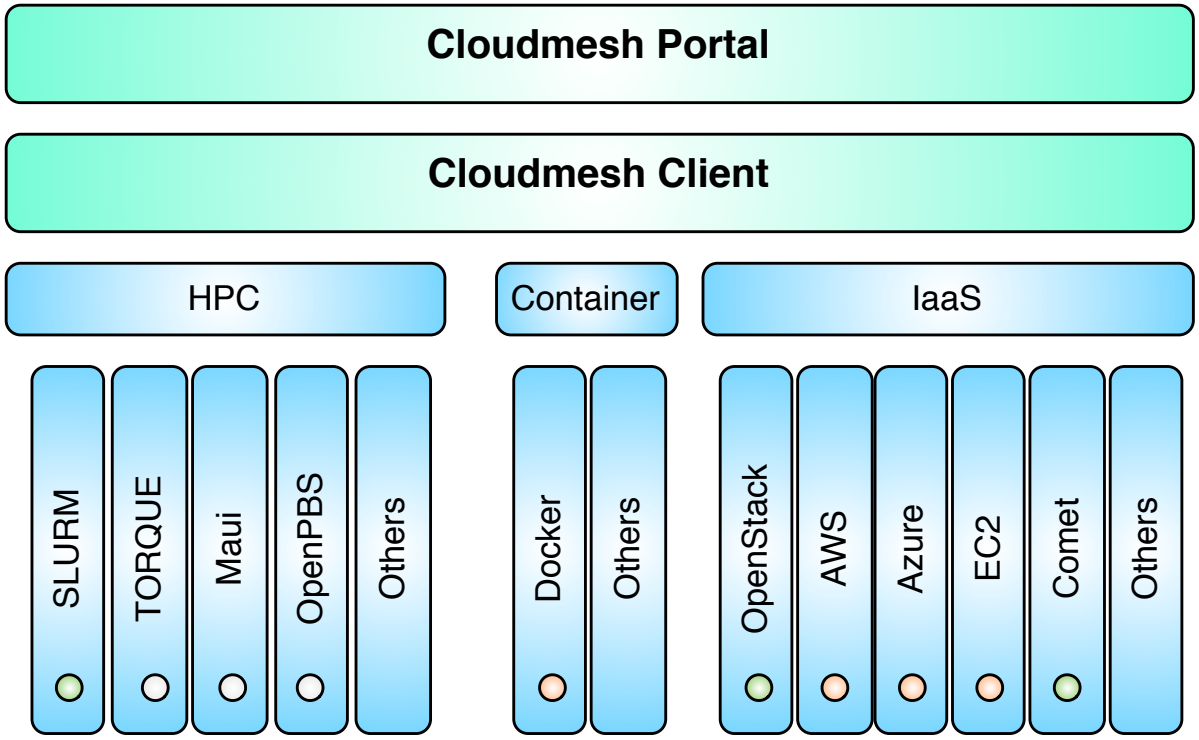
Cloudmesh Fills the Gap

- Matches user needs with multiple provider's services.
- Researchers can use one platform to manage their clouds.
- Orchestrates provisioning and allocation of cloud resources.
- Local copy of your cloud data is created, so jobs and VMs are traceable across clouds
- New clouds with similar configurations can be created easily.
- Default attributes allow easy control of cloud artefacts.
- Users can switch easily between clouds.
- Users can switch easily between HPC systems

- `cm default cloud=comet`
- `cm vm boot`
- `cm default cloud=chameleon`
- `cm vm boot`

Cloudmesh Client and Portal

- Simplify access
  - Queueing Systems
  - Containers
  - IaaS
- Parallelism
  - Task
  - Compute
  - Data
- (GPU)
- XSEDE
  - HPC
  - Clouds
    - Comet
    - Jetstream
    - Bridges

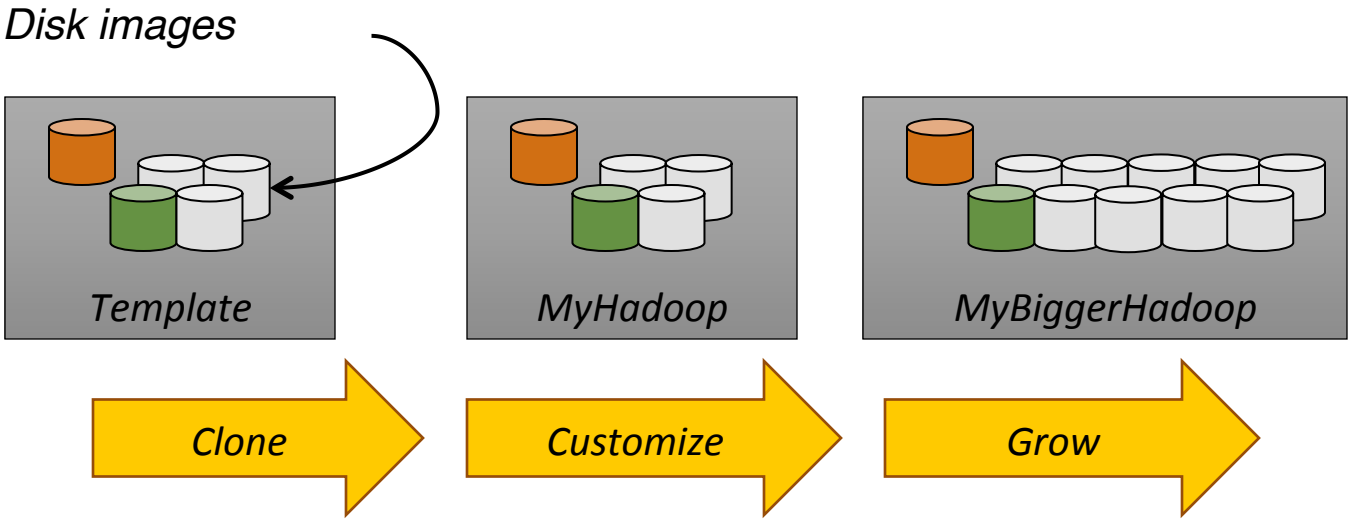


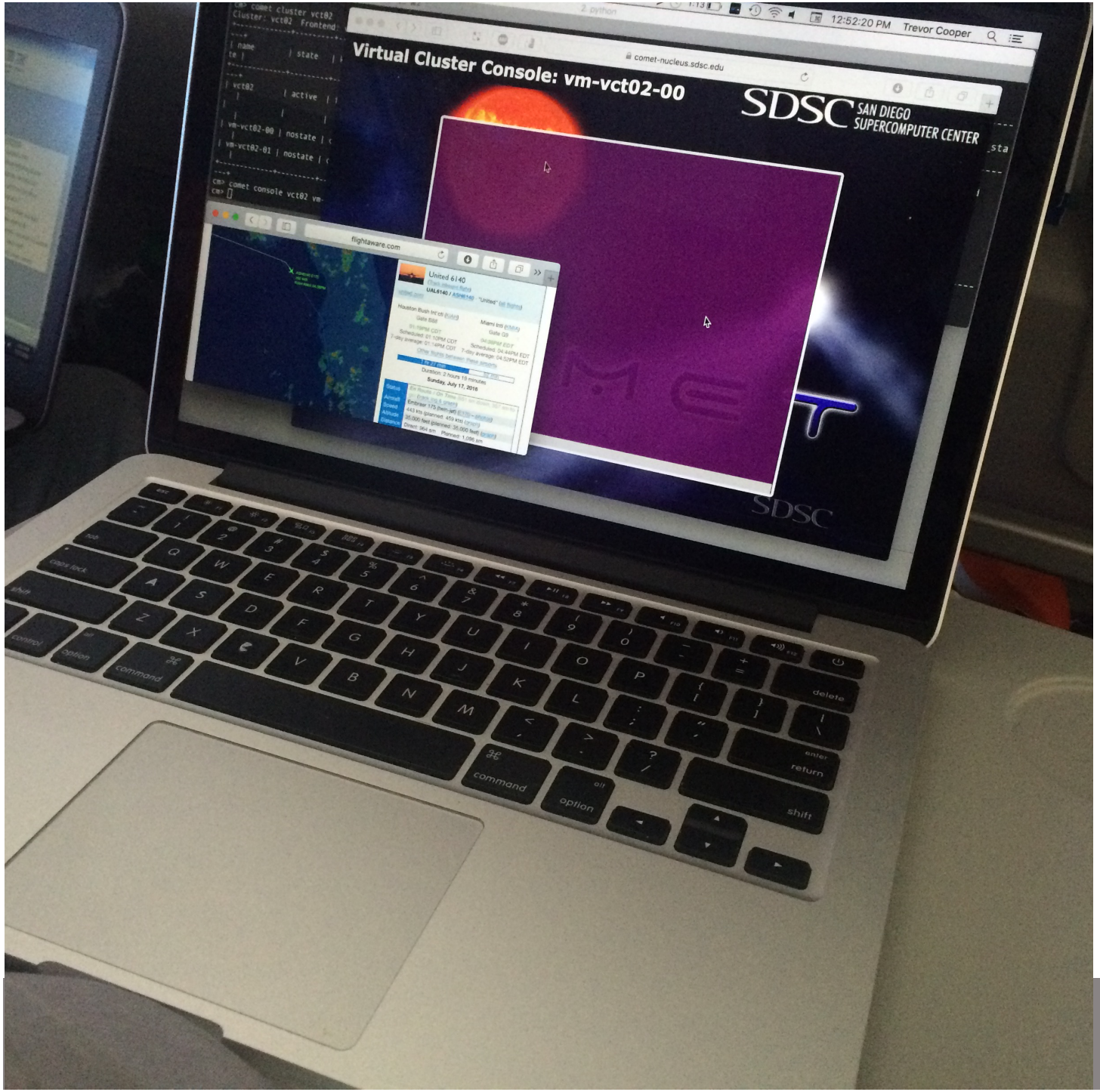
Future: Comet Cloudmesh Platform Launchers

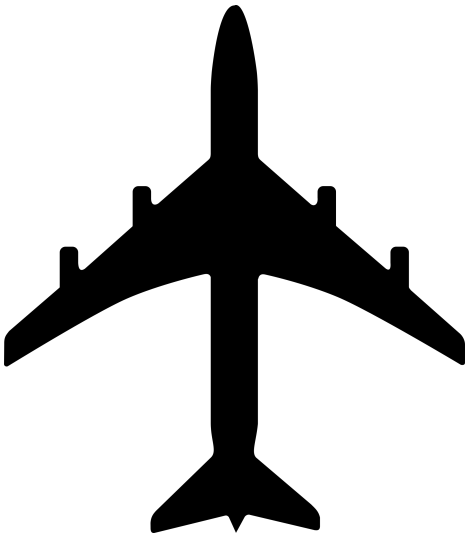
- Customizable launchers
- Launchers available through commandline or browser

Example: Hadoop

```
$cm hadoop -n 10 -group=myHadoop
```



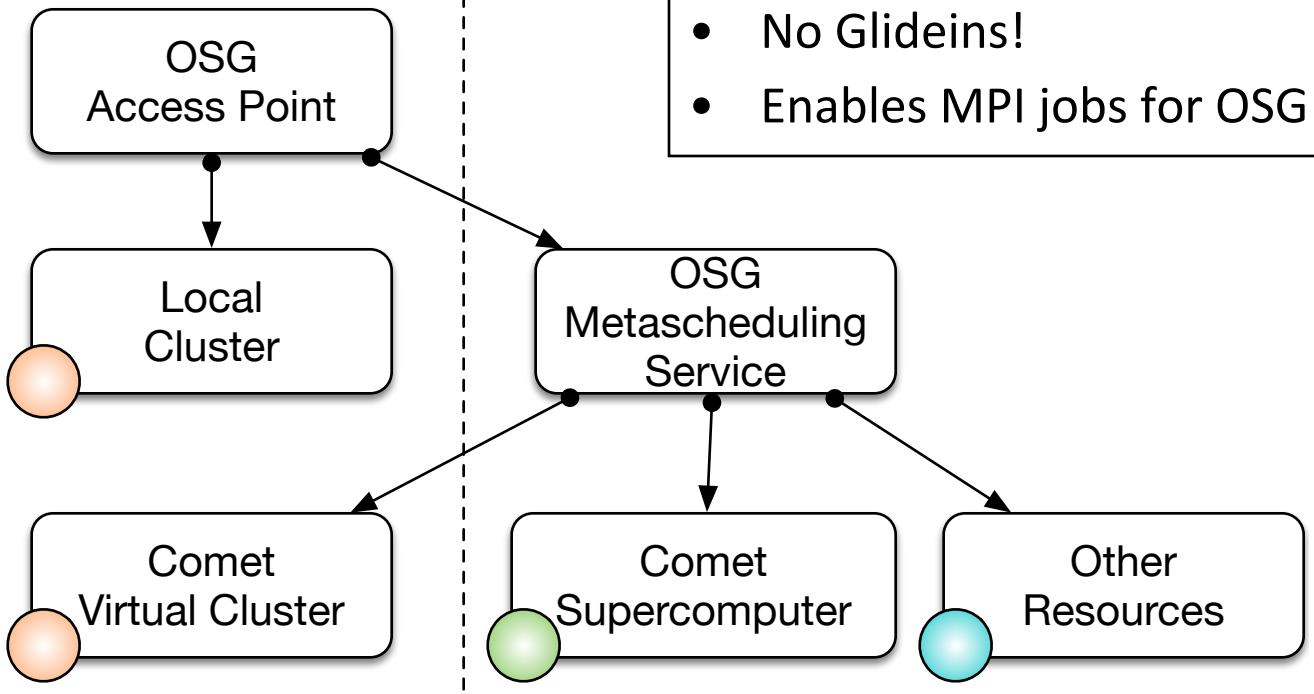

Virtual Cluster Console: vm-vct02-00
SDSC
SAN DIEGO
SUPERCOMPUTER CENTER
comet-nucleus.sdsc.edu
flightaware.com
12:52:20 PM
Trevor Cooper



UC San Diego

Early Success: Open Science Grid

- Native HTCondor cluster inside of Comet!
- No Glideins!
- Enables MPI jobs for OSG



Goal to use the same images/setup

Different setup

New Ideas & Changes

Control & Responsibility

- Shared responsibility
  - SDSC: hosting environment
  - Cluster admin & PI: virtual machine stack
- Different from current HPC roles
- Considering an explicit "Acknowledgement of Responsibility" signed by PI & cluster admin
  - Based on SDSC's "Outback Network" agreement
  - Outback is a separate VLAN & IP subnet for user-managed systems

Innovation in Utilization

- VMs co-scheduled within the batch system
- Works for automated workflows bringing up virtual compute nodes when needed
- What about responsiveness?
  - Can tune batch policies based on need
  - Reserve some physical nodes for fast launch to some scale
- Reuse existing allocations & accounting process
- Likewise, existing science gateway policies: e.g., community account for cluster

New Use Case: Training

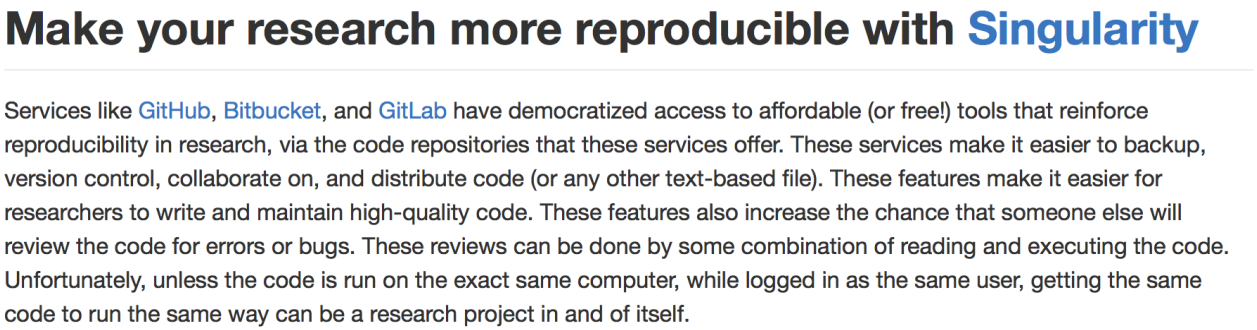
- Training
  - Nearly full stack: OS, networking (IP & InfiniBand), applications
  - Any type of cluster: HPC; N-tier web framework; big data
  - Limits: no layer 1 (VLANs, etc.) networking

New Use Case: Campus Bursting

- Custom HPC clusters can help campuses extend a familiar environment
- How can a campus get a large compute allocation?
- Need to justify science and compute time
- Single PI proposals from large users?
- Historical campus cluster utilizations
- Don't want to burn out Campus Champion allocations

And in Other News

- Check out Singularity: http://singularity.lbl.gov/
- User space containers for HPC
- Deployed on Comet, Gordon, and TSCC (campus cluster)
- Impromptu BoF Wednesday afternoon (location TBD)
- https://github.com/cjprybol/reproducibility-via-singularity



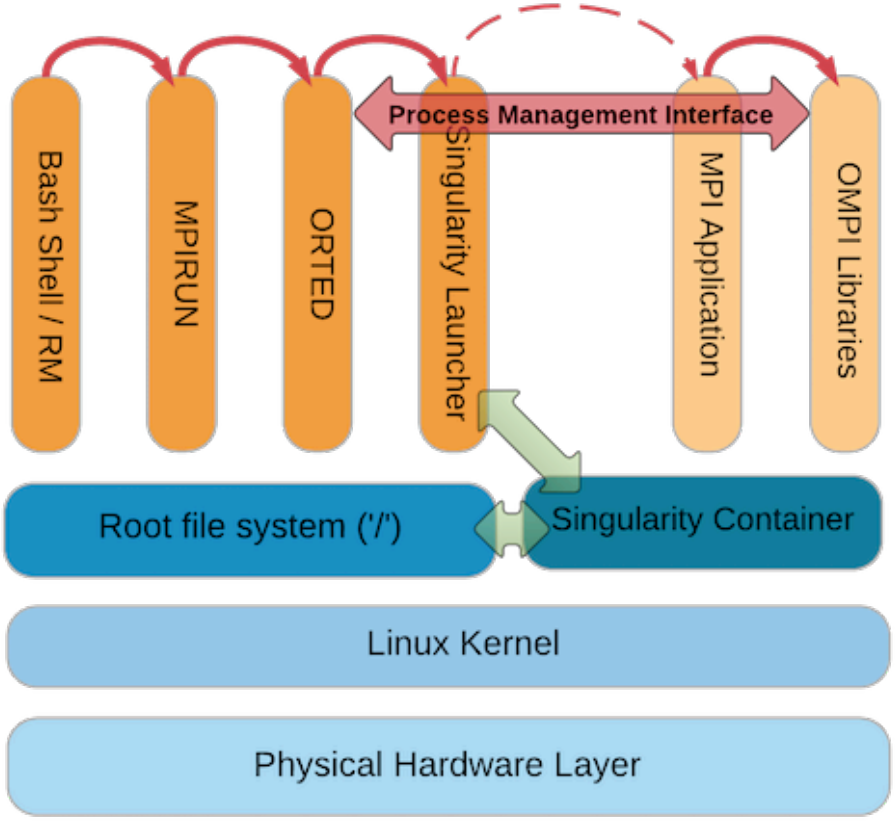
Make your research more reproducible with Singularity

Services like GitHub, Bitbucket, and GitLab have democratized access to affordable (or free!) tools that reinforce reproducibility in research, via the code repositories that these services offer. These services make it easier to backup, version control, collaborate on, and distribute code (or any other text-based file). These features make it easier for researchers to write and maintain high-quality code. These features also increase the chance that someone else will review the code for errors or bugs. These reviews can be done by some combination of reading and executing the code. Unfortunately, unless the code is run on the exact same computer, while logged in as the same user, getting the same code to run the same way can be a research project in and of itself.

Singularity

https://github.com/gmkurtzer/singularity

https://github.com/singularityware



Singularity & Open MPI