PRACE 5IP-T6.2.5: The deployment of containers into HPC infrastructures

NorduGrid 2018

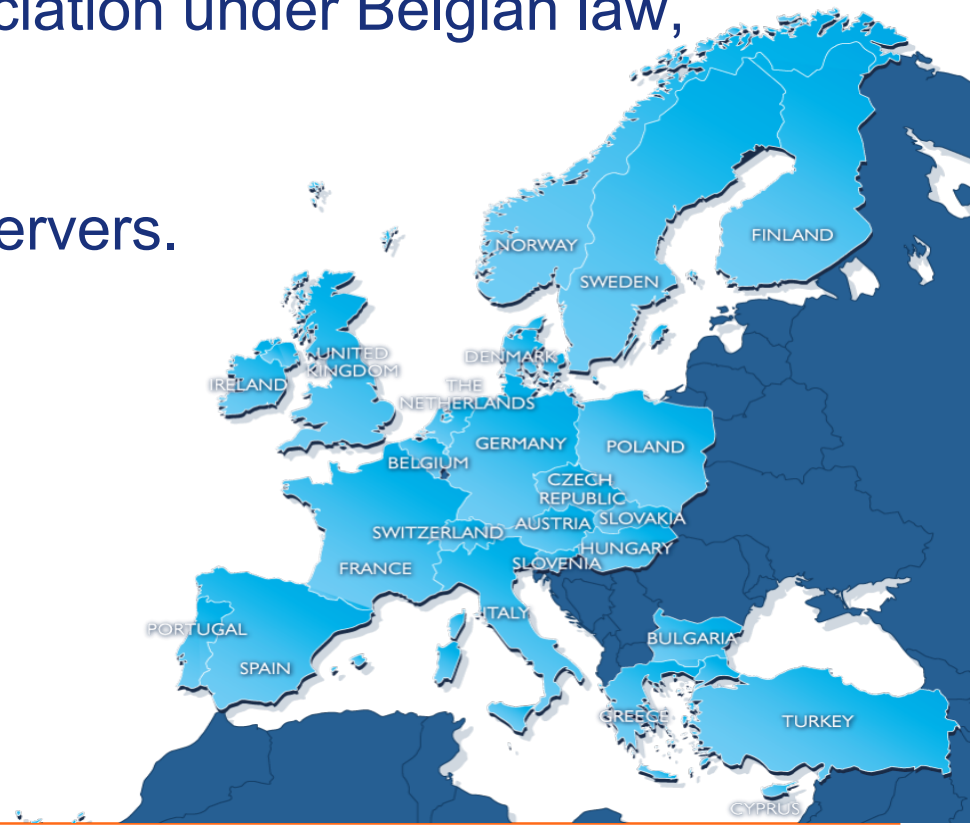
Abdulrahman Azab

Dept. of Research Computing

University of Oslo, Norway

Partnership for Advanced Computing in Europe (**PRACE**)

- International not-for-profit association under Belgian law, with its seat in Brussels.
- Counts 25 members and 2 observers.



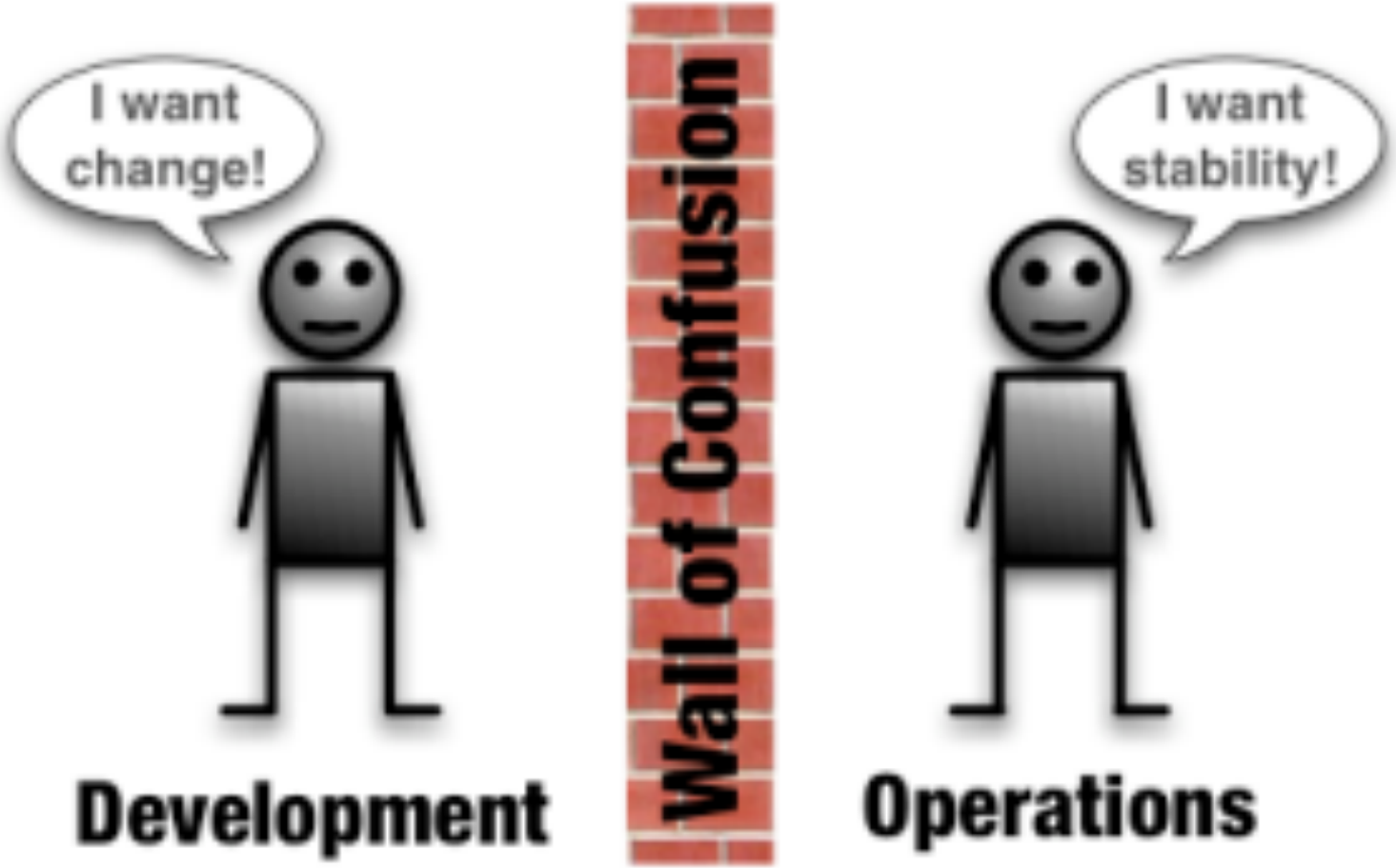
Outline

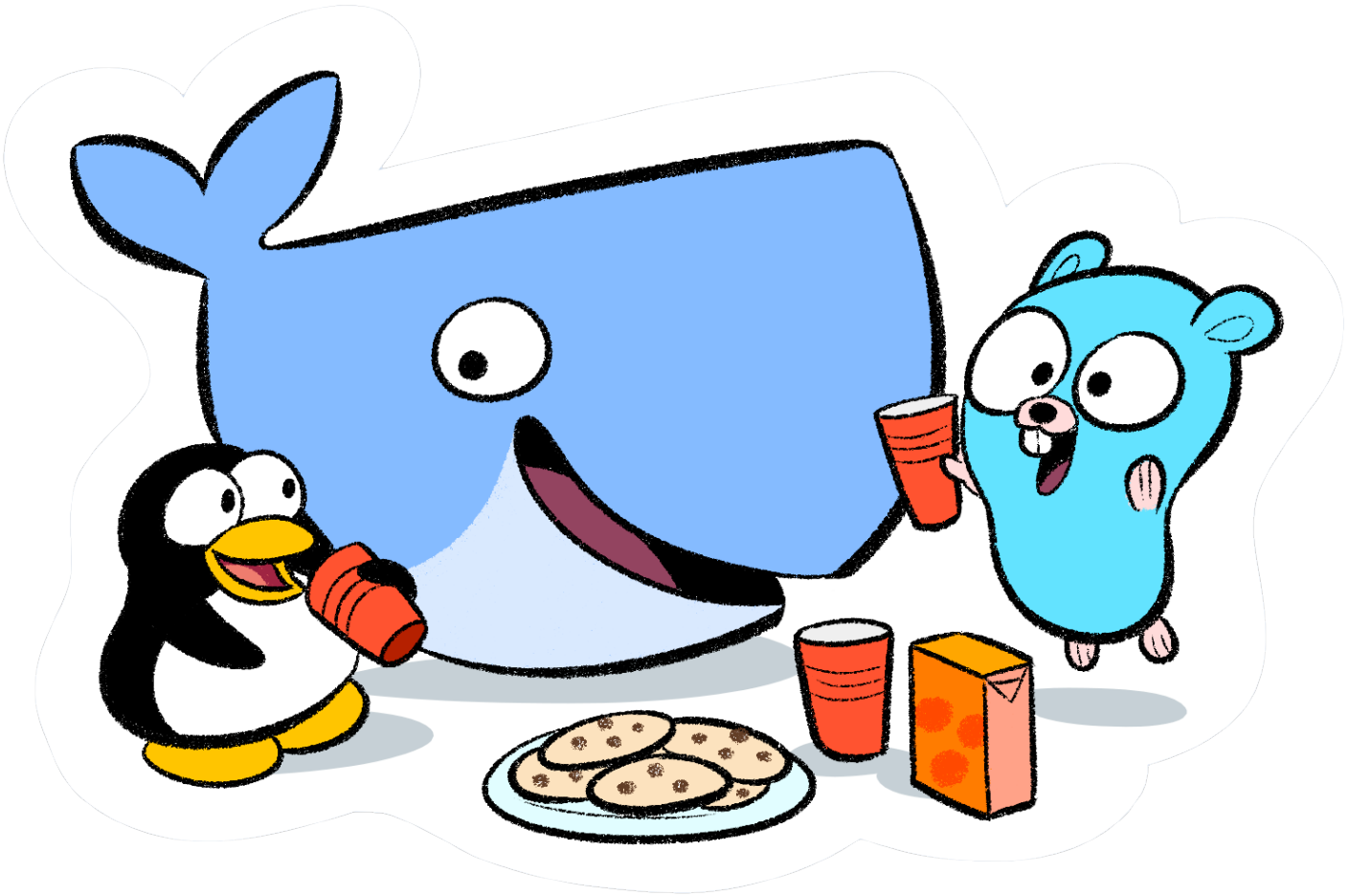
- Introduction
- Prototypes
- Use cases

Introduction

Developer: “PaaS is so easy, who needs sys admins anyway”?

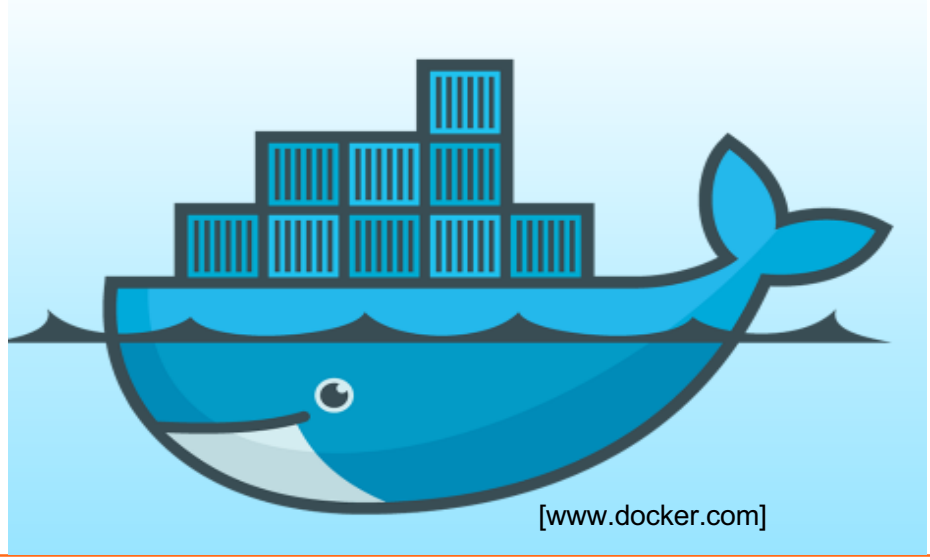
Sys admin: “PaaS is just giant blackbox toy that I can’t really use for real-world app”


I want change!
Wall of Confusion
I want stability!
Development
Operations



Docker

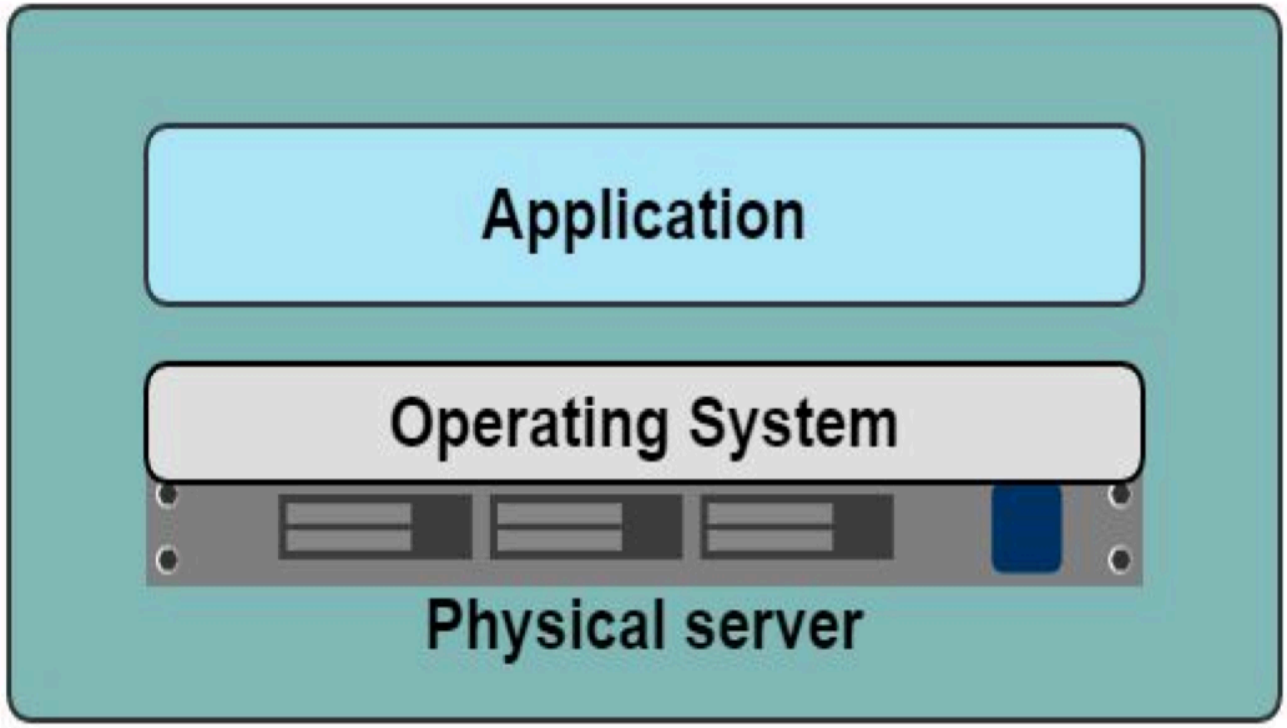
Docker is an open-source project that automates the ***deployment of applications inside software containers****, by providing an additional layer of abstraction and automation of* ***operating system–level virtualization*** *on Linux.*



[www.docker.com]

Application-Server

- Slow deployment
- Huge cost
- Wasted resources
- Difficult to Scale
- Difficult to Migrate



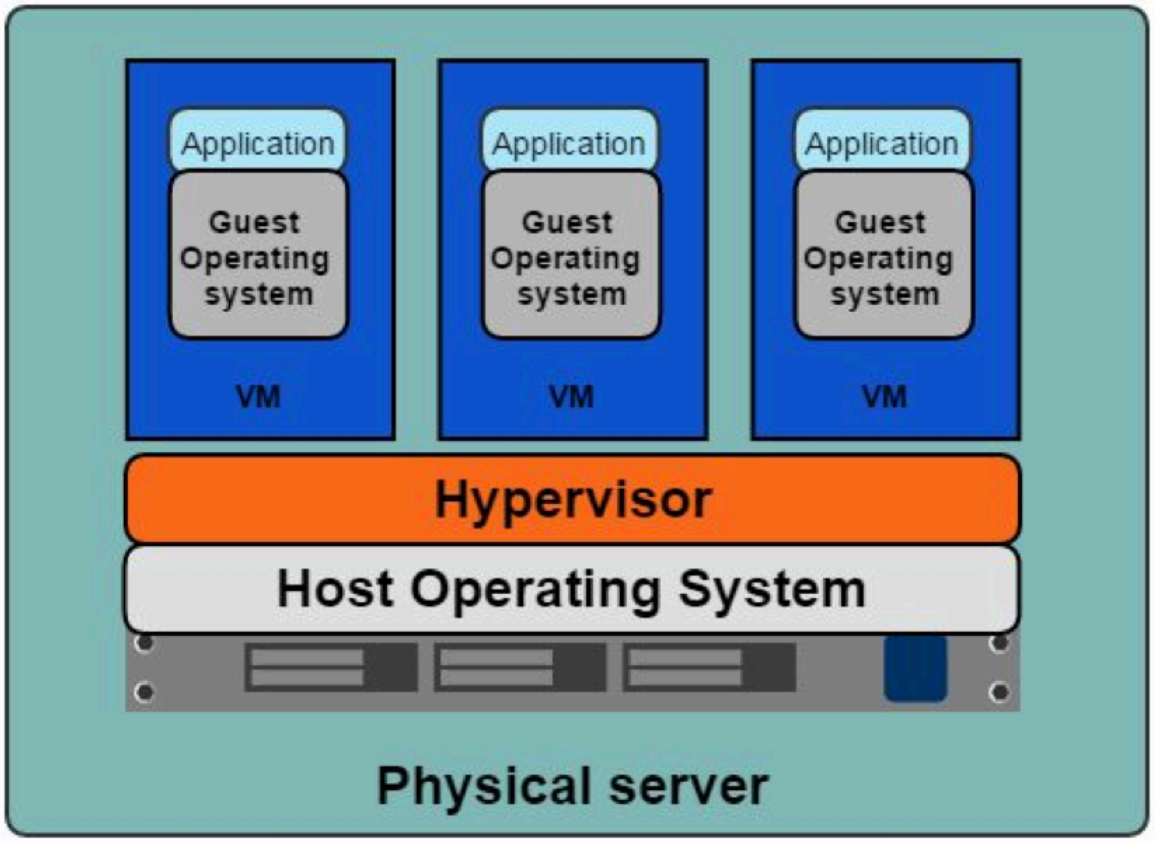
VMs

Benefits

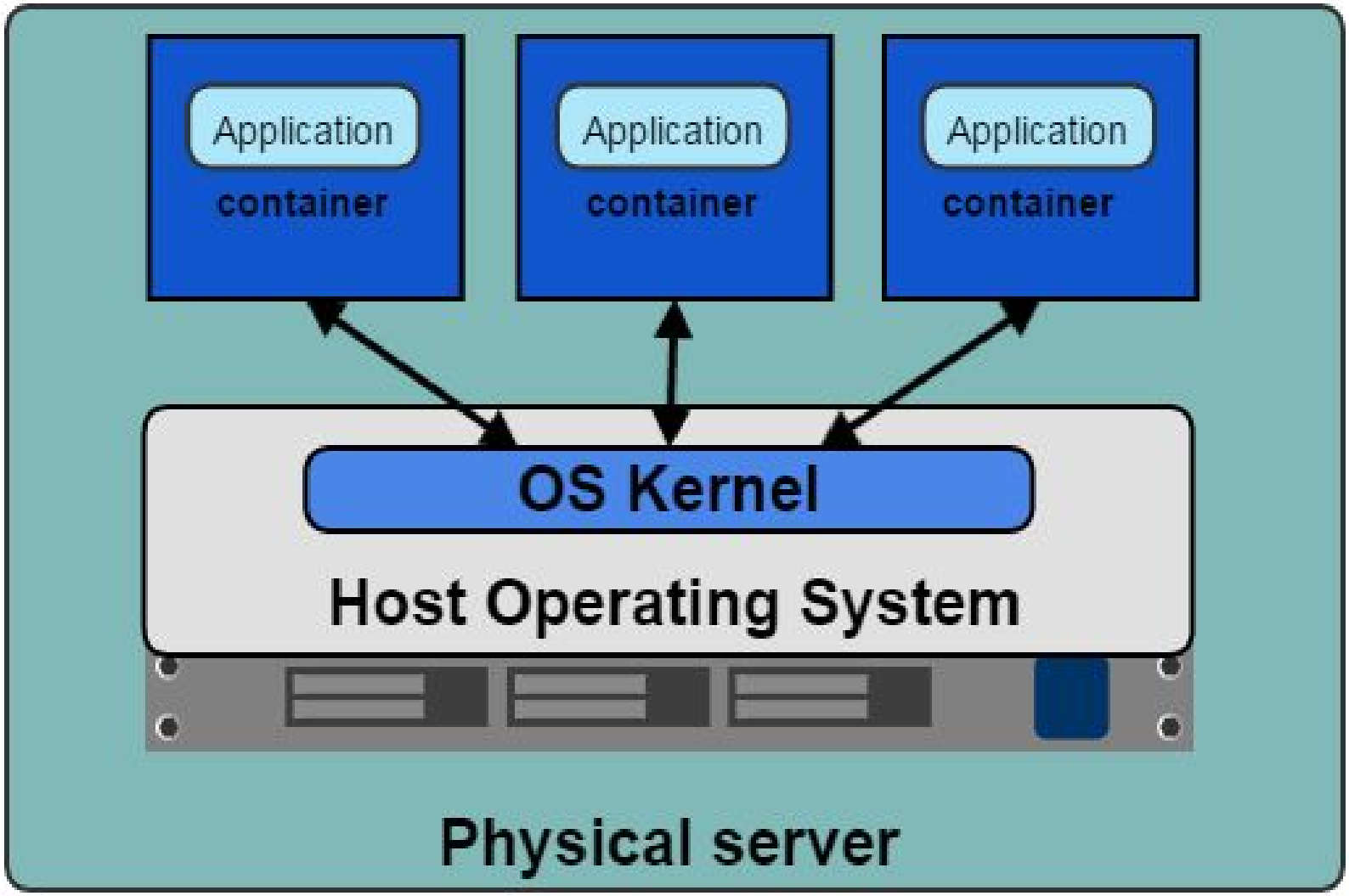
- Better resource pooling
- Easier to scale
- VM's on the cloud.

Limitations

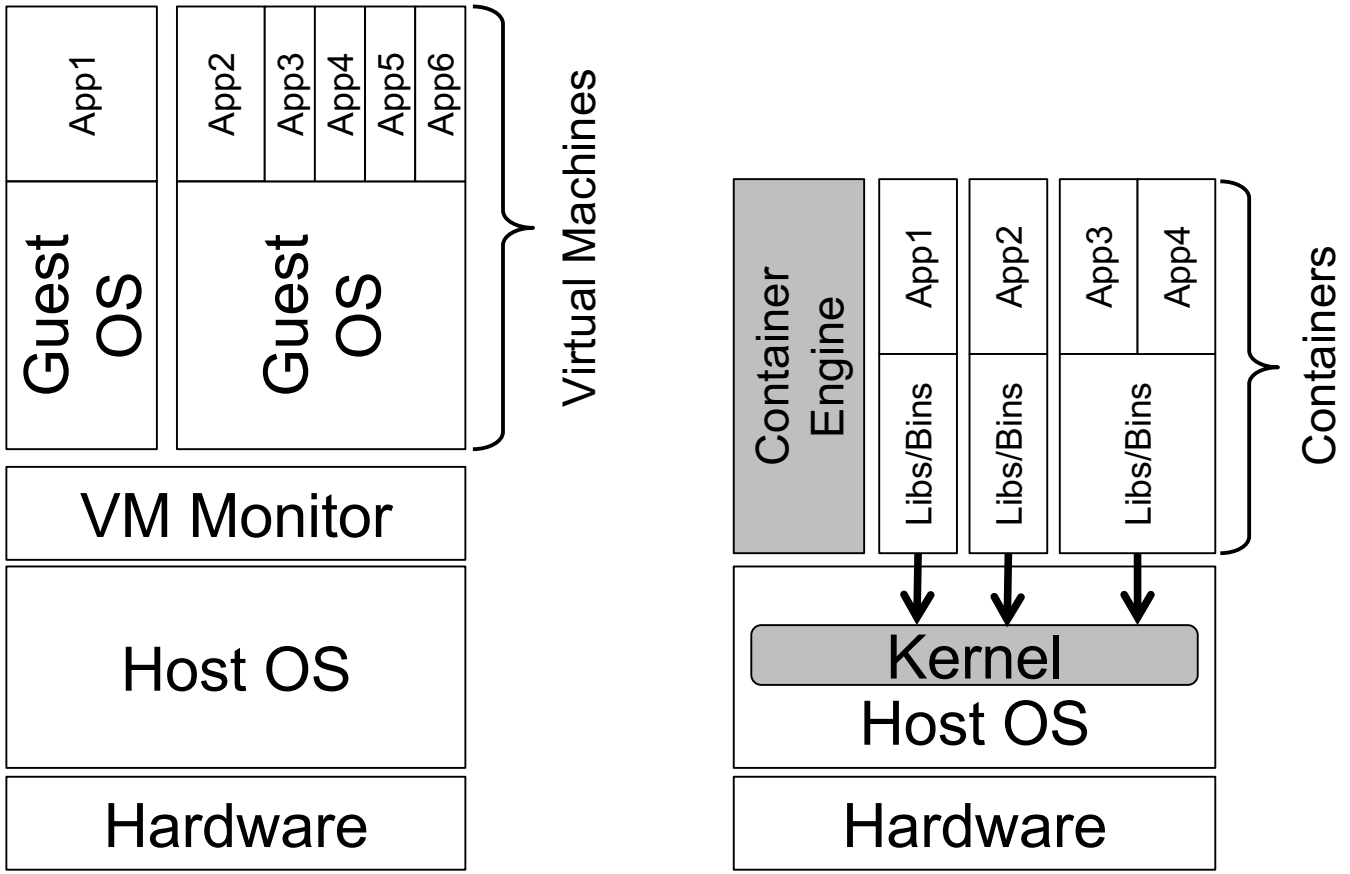
- Dedicated resources for each VM (more VM = more resources).
- Guest VM = Wasted resources.



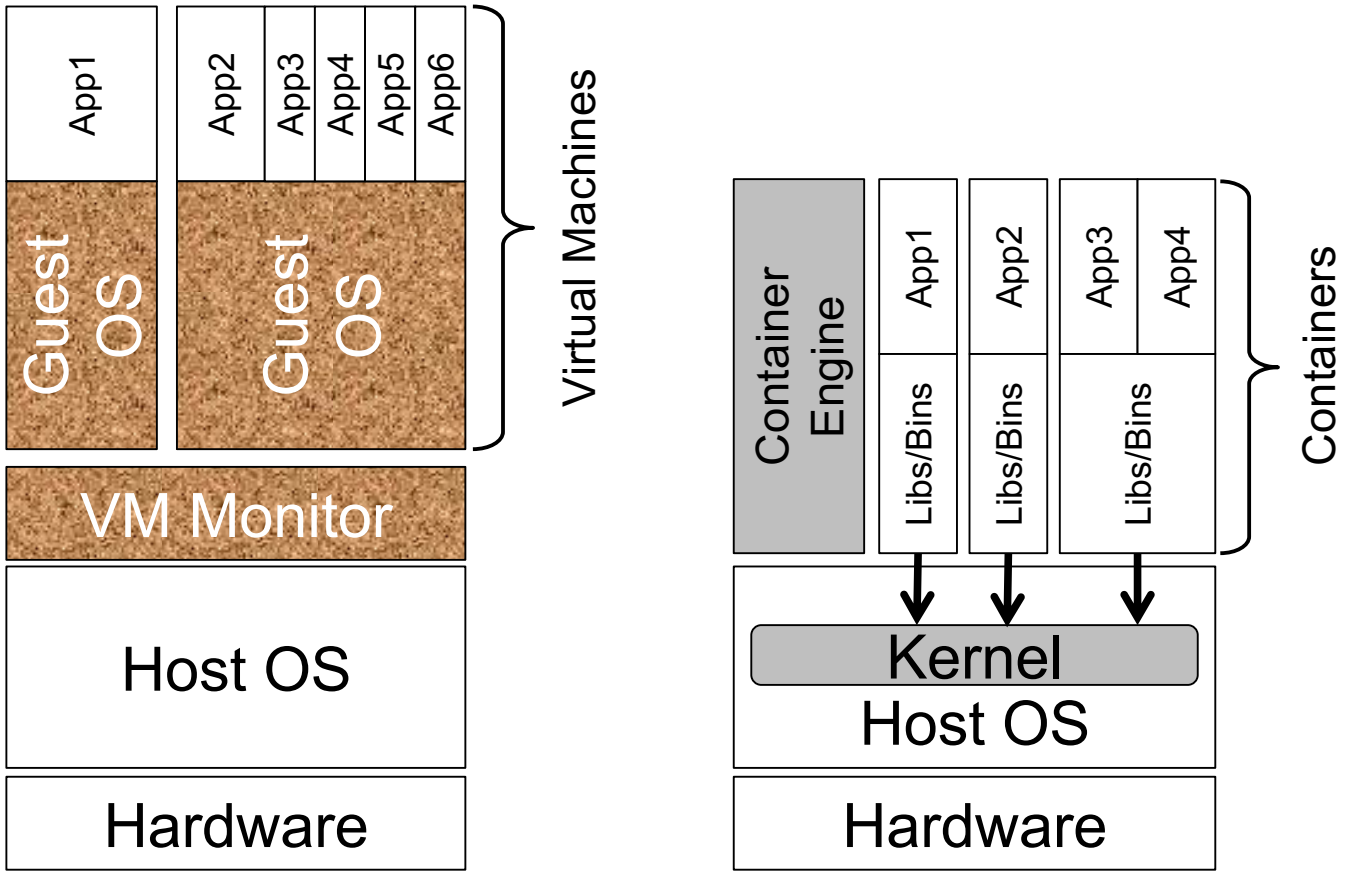
Containers


Application
container
Application
container
Application
container
OS Kernel
Host Operating System
Physical server

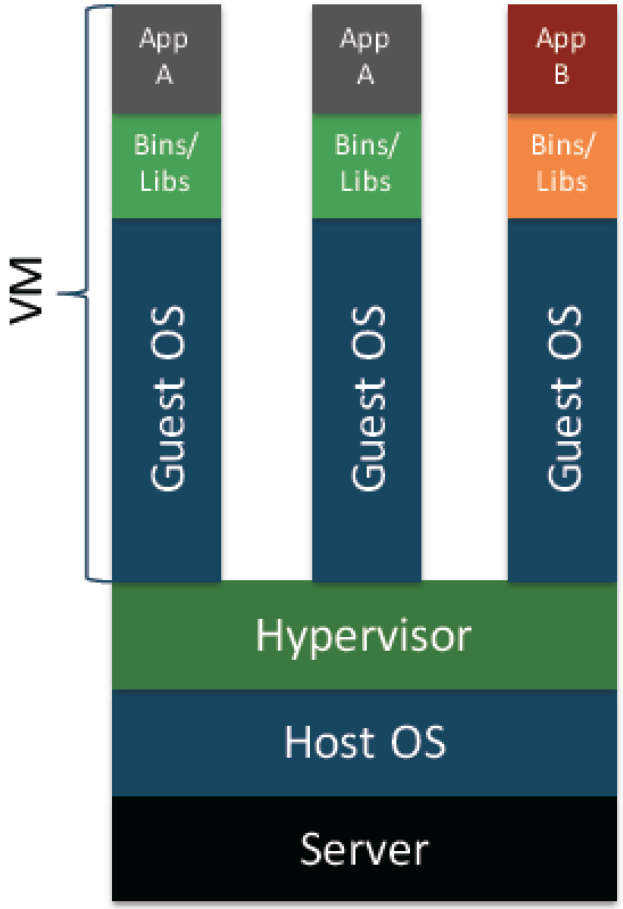
Containers and VMs



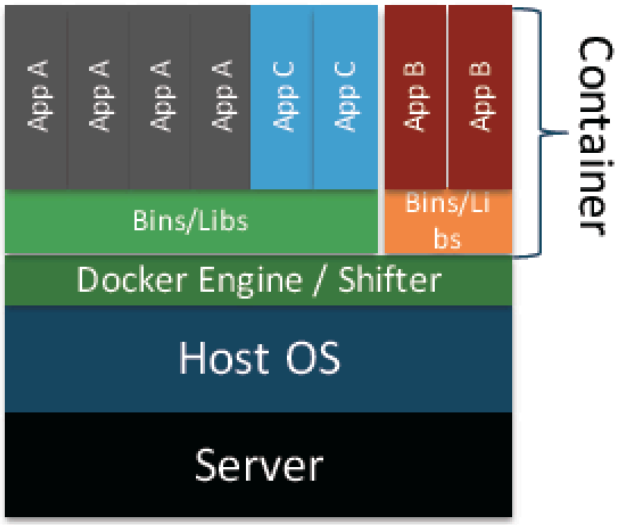
Containers and VMs



Linux Containers vs. Virtual Machines



Containers are isolated, but share kernel and, where appropriate, bins/libraries



Containers provide close to native performance

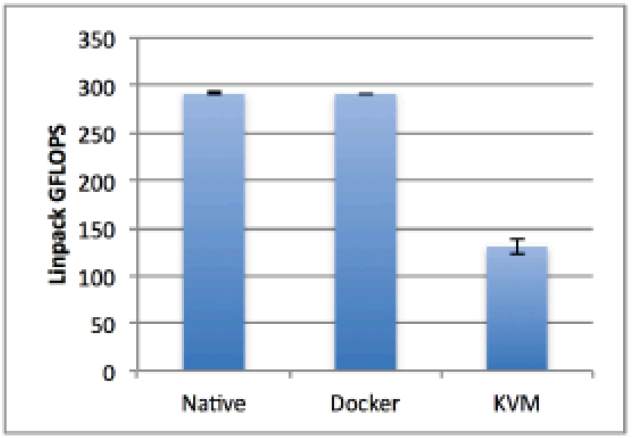
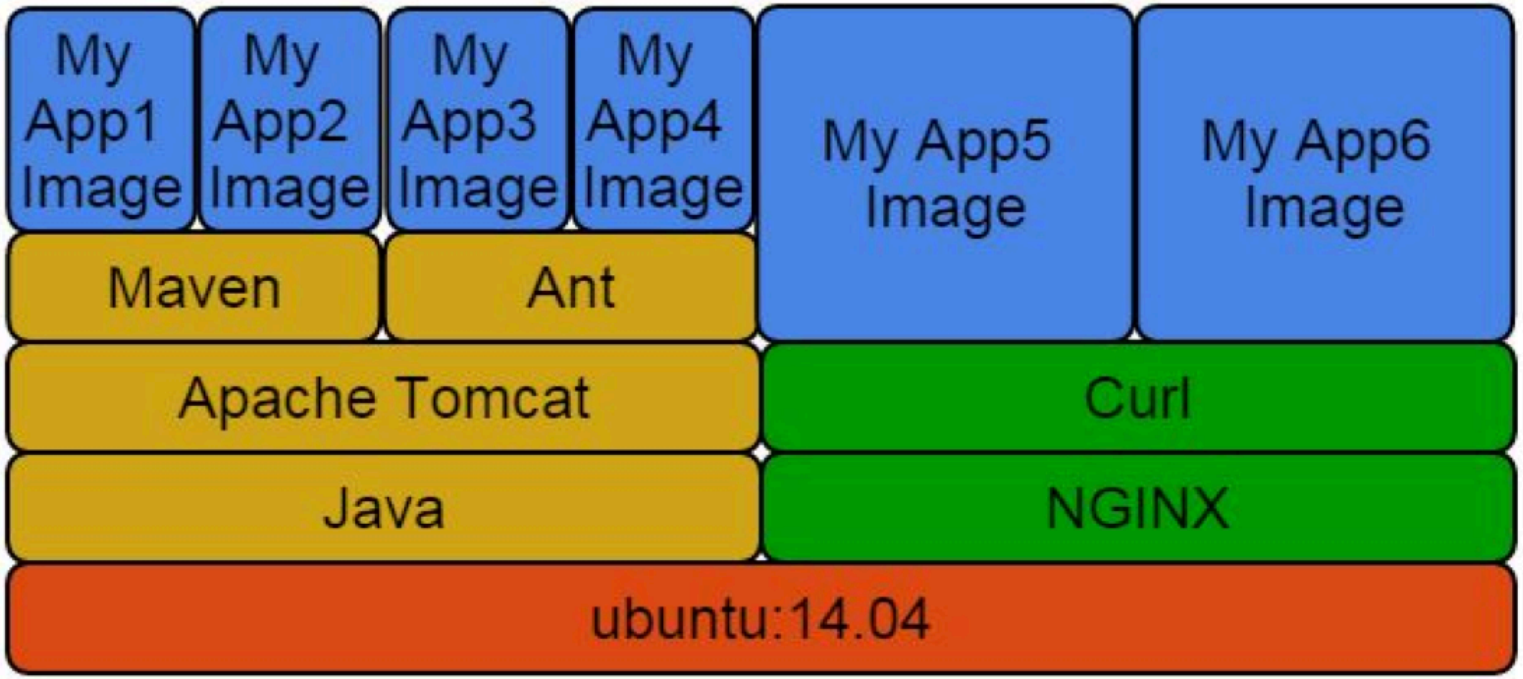


Figure 1. Linpack performance on two sockets (16 cores). Each data point is the arithmetic mean obtained from ten runs. Error bars indicate the standard deviation obtained over all runs.

Source: IBM Research Report (RC25482)

A "container" delivers an application with all the libraries, environment, and dependencies needed to run.

Docker Layers



Docker Layers

ubuntu : 200 Mb
ubuntu + R : 250 Mb
ubuntu + matlab : 250 Mb

All three: 300 Mb

Why containers?

- Without containers:

"I need software X, and here is the installation guide, please install it!"

- With containers:

"I need software X, here is the name of its Docker image, please pull it"

- Very little performance degradation compared to native
- Security: SeLinux, Capability whitelist, syscall whitelist, and user namespaces

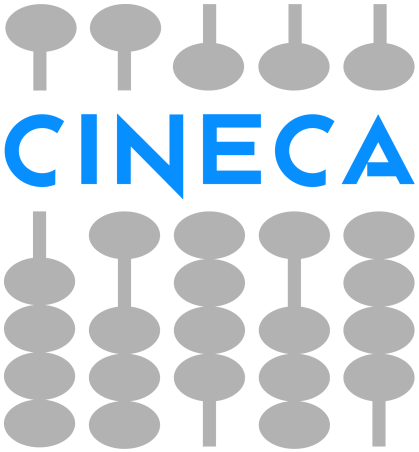
What containers don’t solve

- Hardware architecture and kernel incompatibility.
- Operational maintenance mess (e.g. two different versions of MPI).
- Containers are not for huge software packages, e.g. Bio-Linux. To package those in one unit, VMs are more suitable

Why VMs?

- VM jobs are useful in cases of too old kernel on compute nodes.
- VMs are also effective in cases where a specific Linux kernel or Windows OS or OSX is needed.

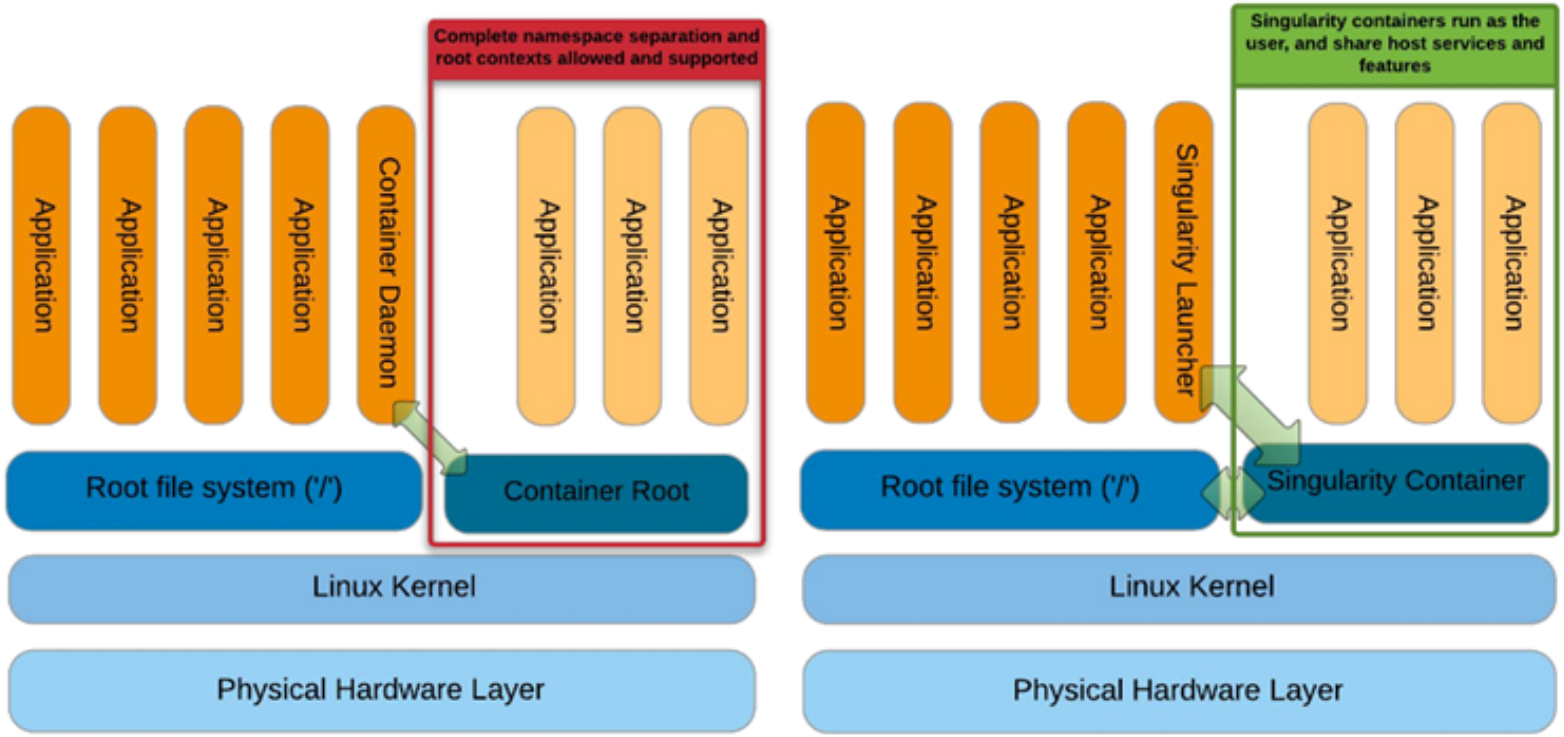
Contributors



Prototypes

- Singularity
- Shifter
- Socker
- HTCondor VM and Docker universes
- Galaxy
- aCT

Singularity: Unprivileged containers for HPC



Singularity: Unprivileged containers for HPC

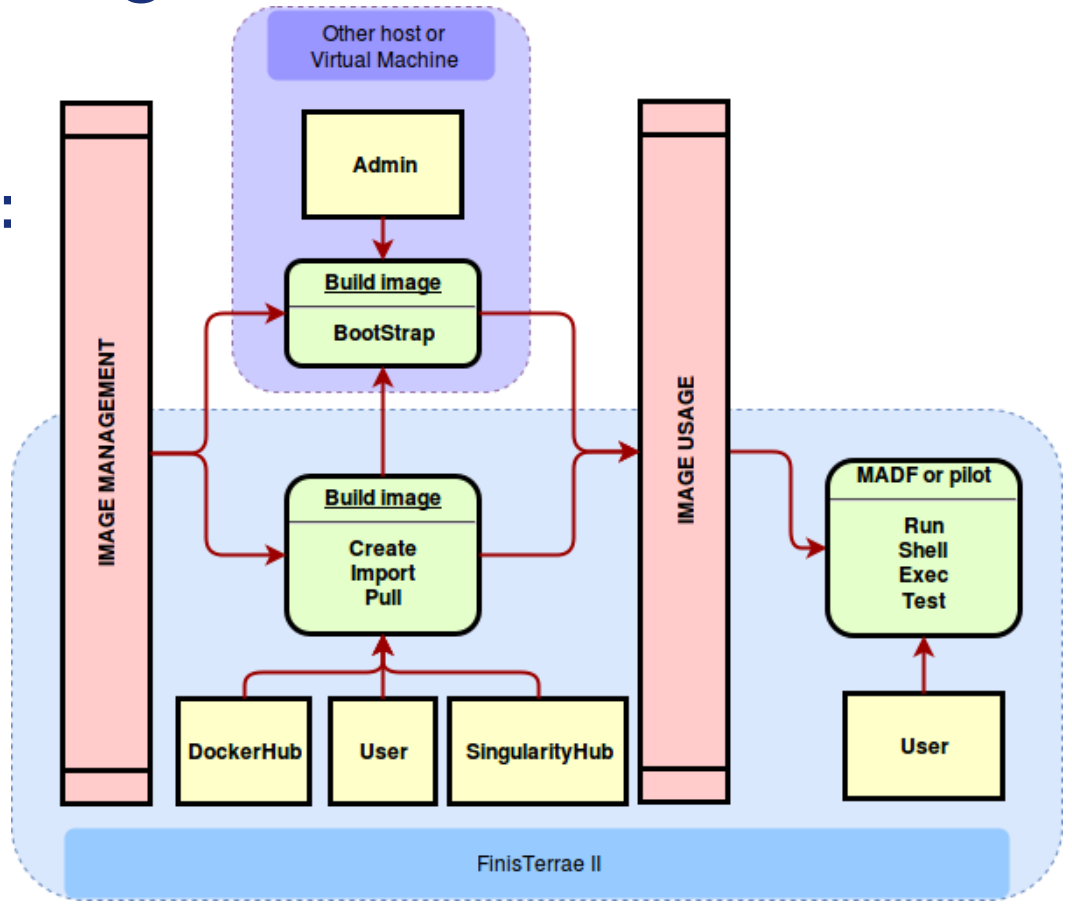
- **MPI support:** build-in support for MPI (OpenMPI, MPICH, IntelMPI)
- **Data analysis example:** Computing principal components for the first 10,000 variants from the 1000 Genomes Project chromosomes 21 and 22:

```
wget https://…/chr21.head.vcf.gz
wget https://…/chr22.head.vcf.gz
LANG=C CHUNKSIZE=10000000 mpirun -x LANG -x
CHUNKSIZE -np 2 singularity run -H $(pwd)
variant_pca.img
```

Singularity: Unprivileged containers for HPC

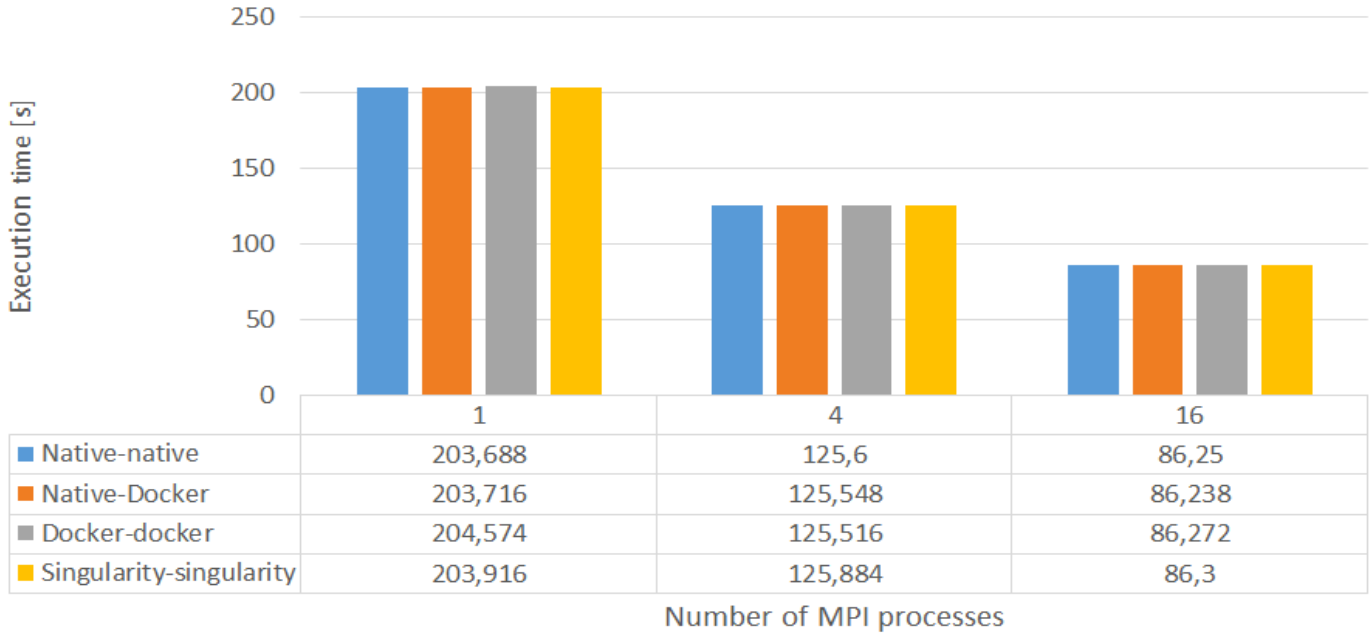
- **MPI testing (**MSO4SC **):**

Architecture



Singularity: Unprivileged containers for HPC

- **MPI testing (**MSO4SC **):** Feel++ Lid driven cavity 2D simulation benchmark



	1	4	16
Native-native	203,688	125,6	86,25
Native-Docker	203,716	125,548	86,238
Docker-docker	204,574	125,516	86,272
Singularity-singularity	203,916	125,884	86,3

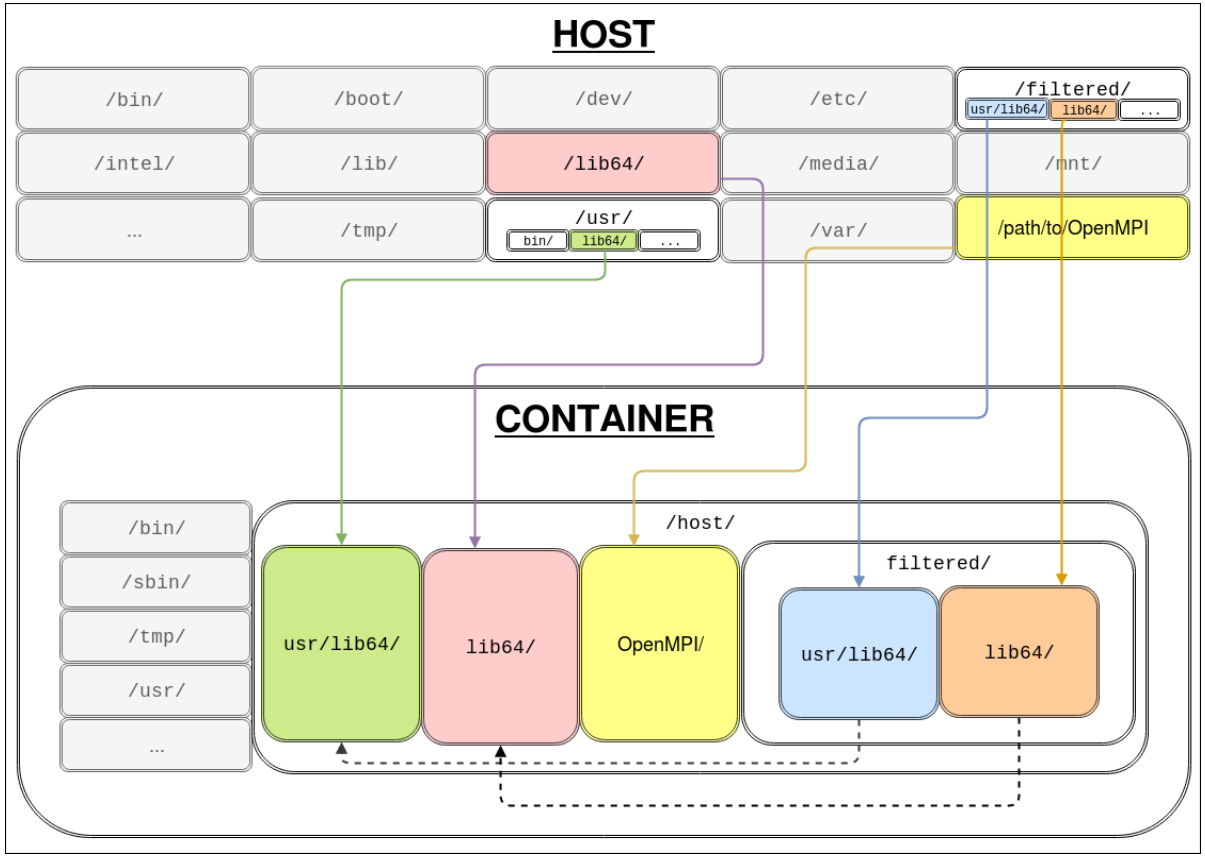
Portability

➢ Singularity & OpenMPI: Exact matching version

Container OpenMPI	Host OpenMPI				
	1.10.2	2.0.0	2.0.1	2.0.3	2.1.1
1.10.2	☑	☒	☒	☒	☒
2.0.0	☒	☑	☒	☒	☒
2.0.1	☒	☒	☑	☒	☒
2.0.3	☒	☒	☒	☑	☒
2.1.1	☒	☒	☒	☒	☑

Portability

- Singularity & OpenMPI
  Bind-mount host MPI



Portability

➢ Singularity & OpenMPI: Bind-mount host MPI

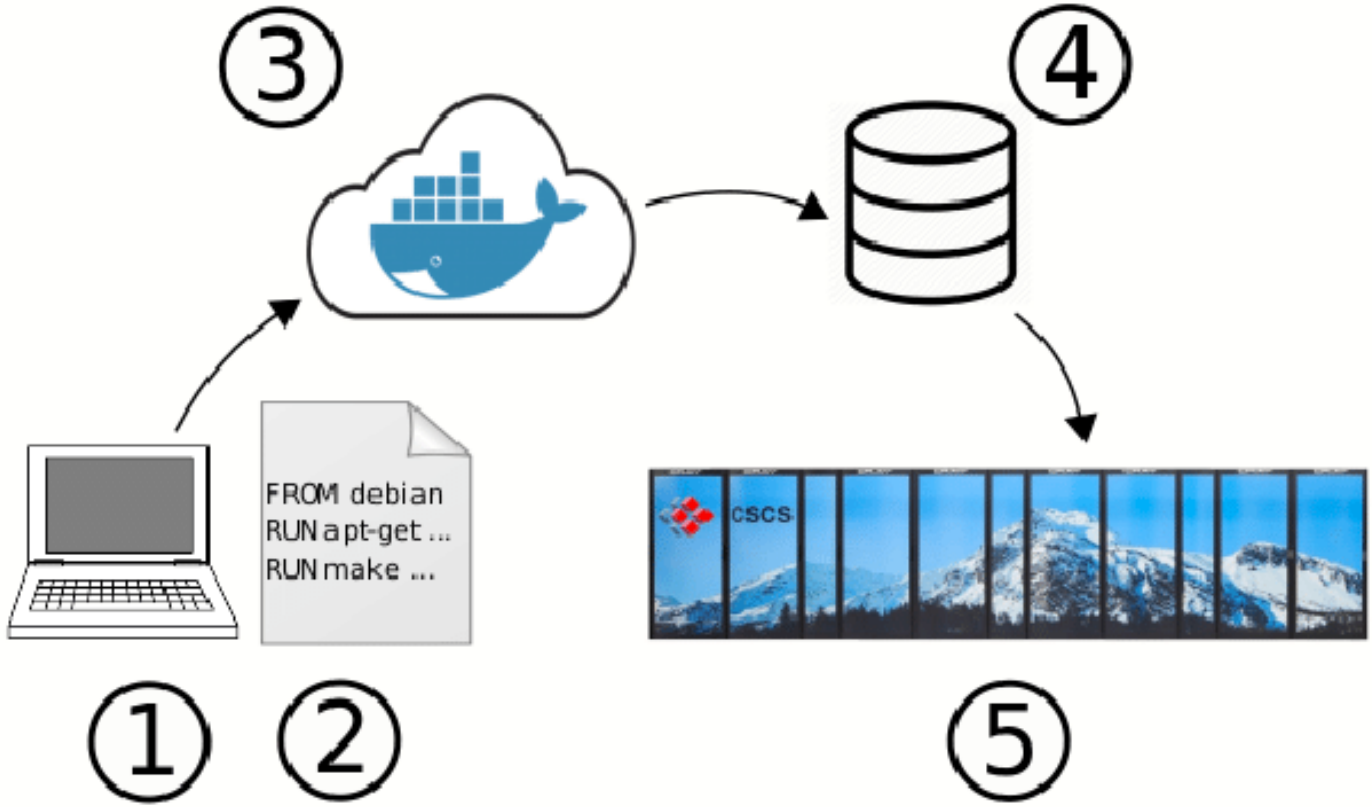
Container OpenMPI	Host OpenMPI				
	1.10.2	2.0.0	2.0.1	2.0.3	2.1.1
1.10.2	☑	☑!	☑!	☑!	☑!
2.0.0	☑	☑	☑	☑	☑
2.0.1	☑	☑	☑	☑	☑
2.0.3	☑	☑	☑	☑	☑
2.1.1	☑	☑	☑	☑	☑

! Symbol size warning

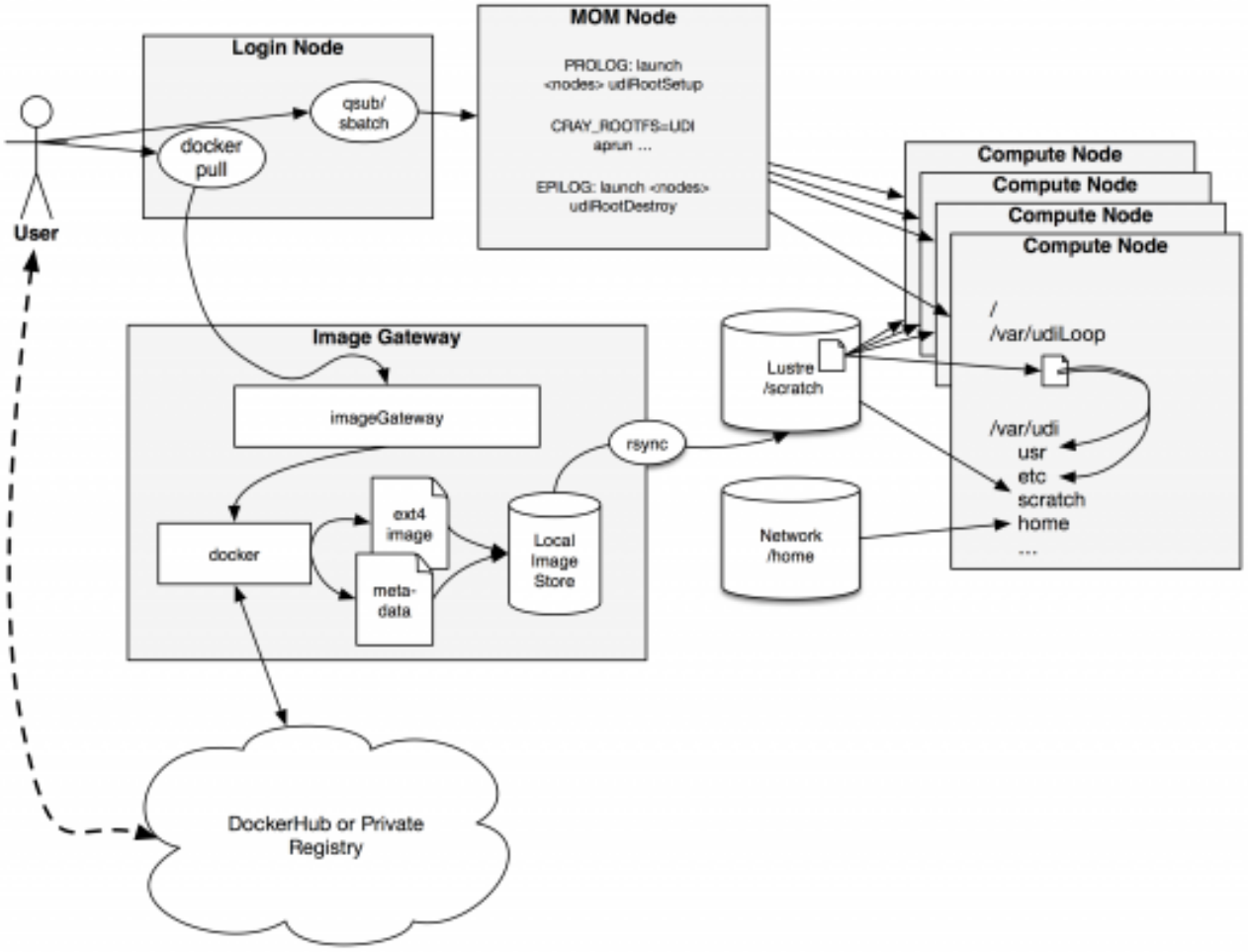
Prototypes

- Singularity
- Shifter
- Socker
- HTCondor VM and Docker universes
- Galaxy
- aCT

Shifter: Enabling docker containers for HPC



Shifter: Enabling docker containers for HPC



Shifter: Enabling docker containers for HPC

Pull the image from docker hub

```
$ module load shifter
$ shifter pull docker:<image-name>
```

In the Slurm script: Run the container

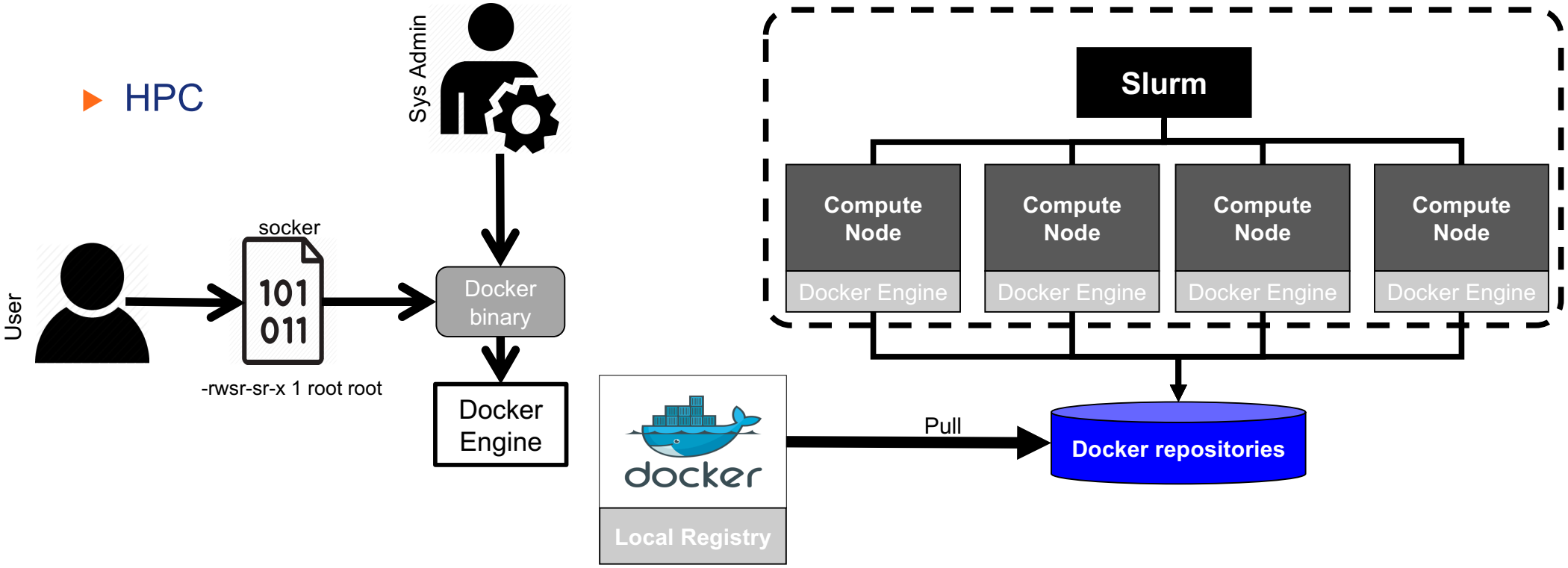
```
#!/bin/bash
#SBATCH --image=docker:<image-name>
#SBATCH --nodes=1
#SBATCH --partition=regular

module load shifter
srun -n 32 shifter <command>
```

Prototypes

- Singularity
- Shifter
- Socker
- HTCondor VM and Docker universes
- Galaxy
- aCT

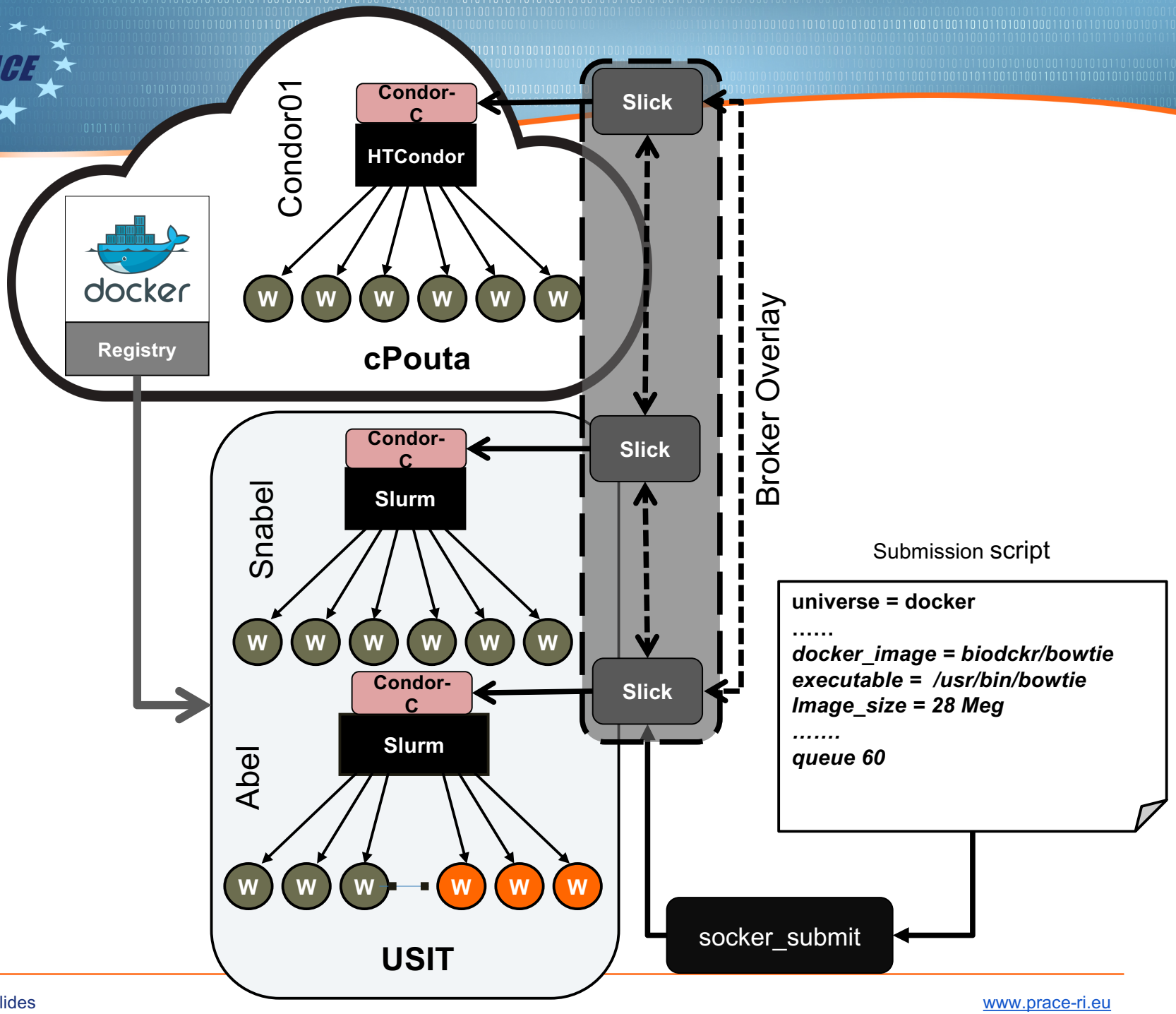
Socker: Secure Docker containers on HPC and MTC

- HPC



Socker

- Many-Task Computing



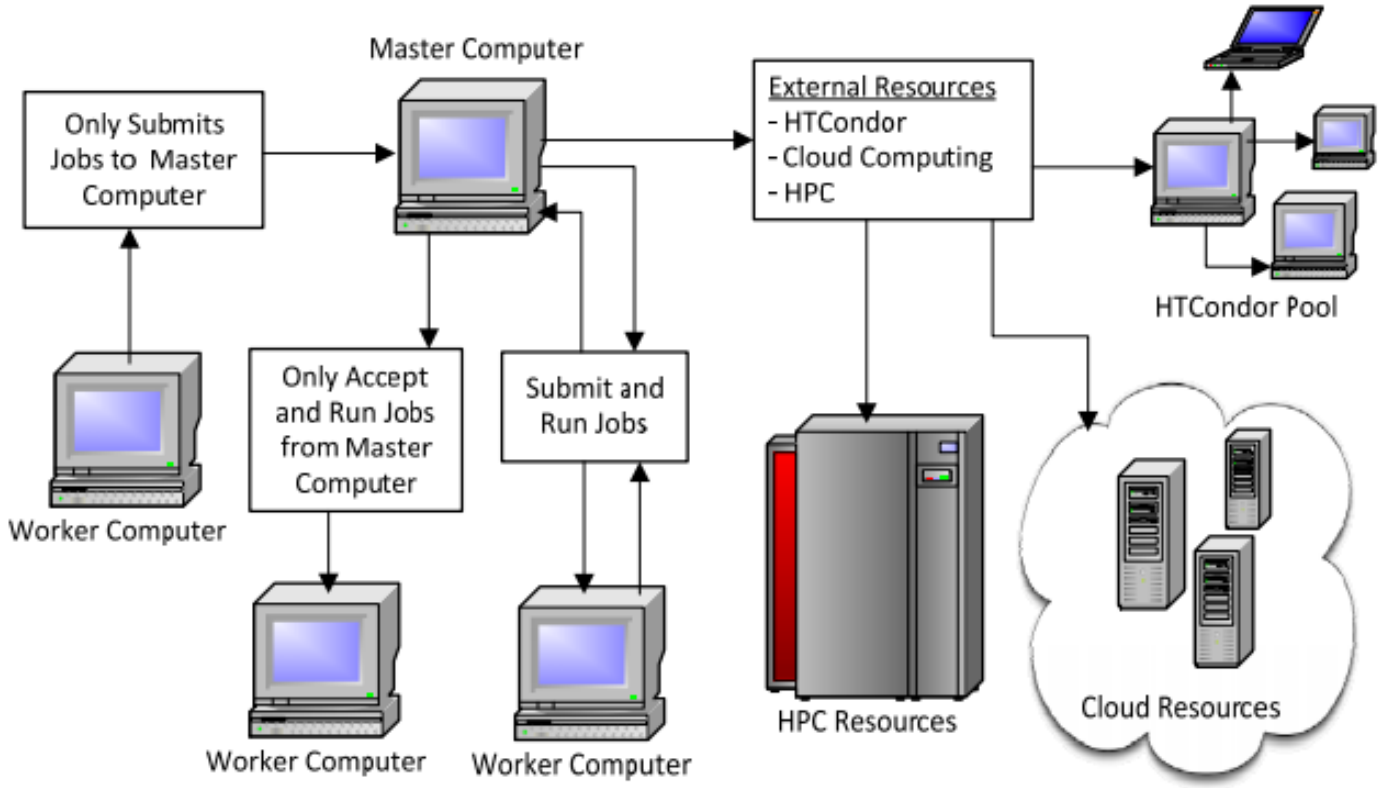
Socker: Secure Docker containers on HPC and MTC

- Publication: A. Azab, Enabling Docker Containers for High-Performance and Many-Task Computing, in 2017 IEEE International Conference on Cloud Engineering (IC2E). IEEE Computer Society, p 279 – 285
- Future
  - MPI support: MPICH

Prototypes

- Singularity
- Shifter
- Socker
- HTCondor VM and Docker universes
- Galaxy
- aCT

HTCondor


Master Computer
Only Submits Jobs to Master Computer
External Resources
- HTCondor
- Cloud Computing
- HPC
HTCondor Pool
Worker Computer
Only Accept and Run Jobs from Master Computer
Submit and Run Jobs
HPC Resources
Cloud Resources
Worker Computer
Worker Computer

HTCondor VM universe

- VM universe:

```
universe = vm
executable = vmware_sample_job
log = simple.vm.log.txt
vm_type = vmware
vm_memory = 64
vmware_dir = C:\condor-test
vm_checkpoint = true
queue
```

HTCondor Docker universe

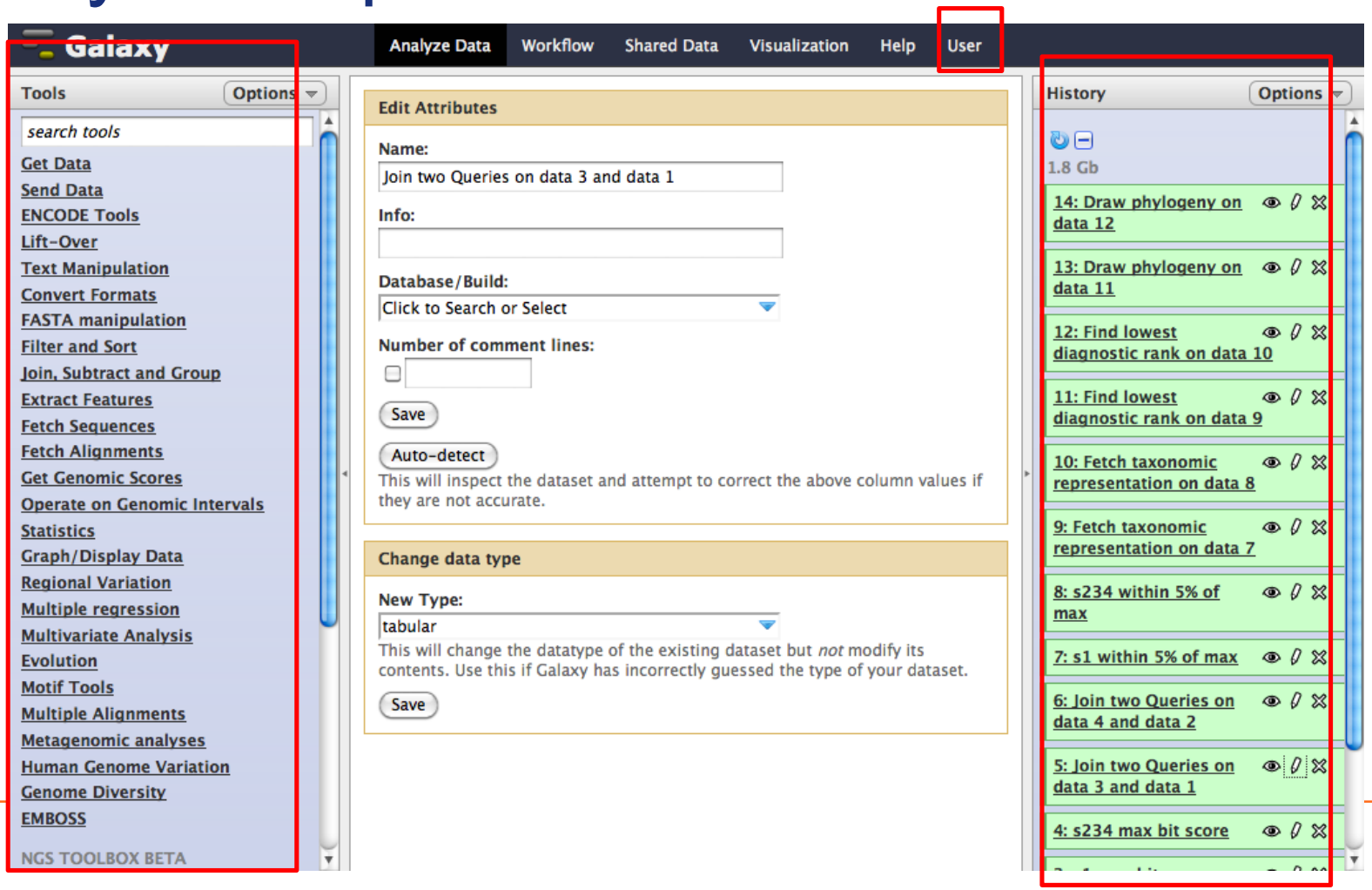
- Docker universe:

```
universe = docker
docker_image = debian
executable = /bin/cat
arguments = /etc/hosts
output = out.$(Process)
error = err.$(Process)
request_memory = 100M
queue 10
```

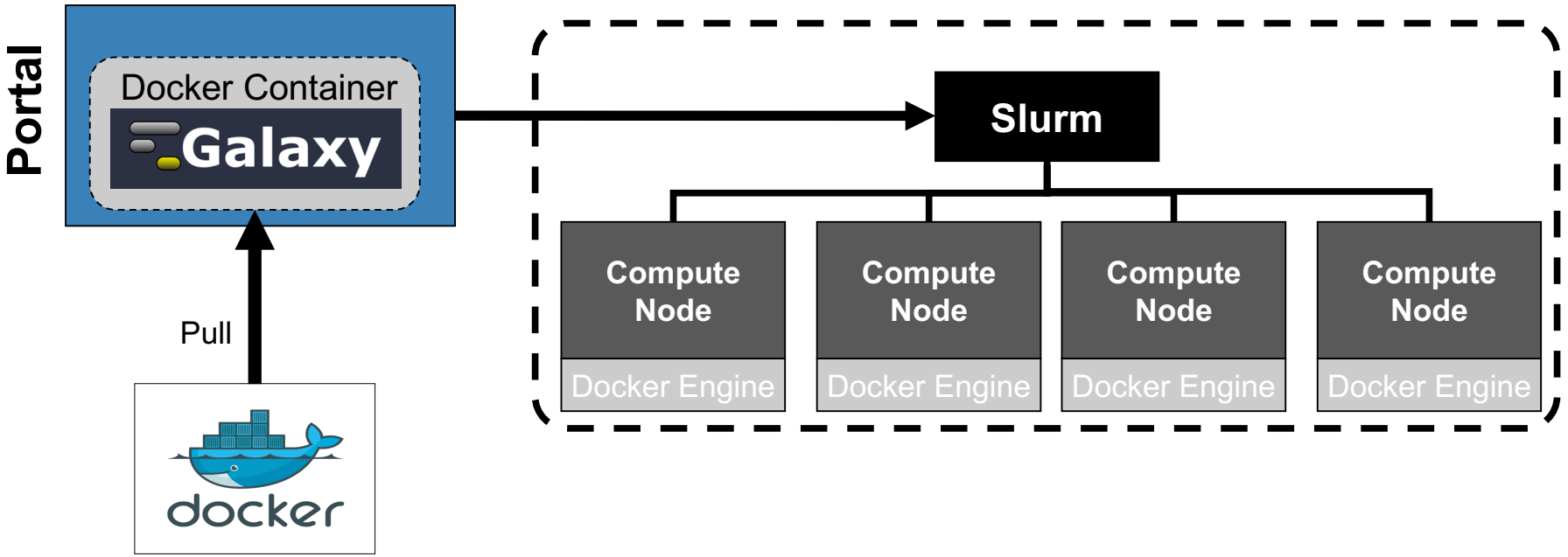
Prototypes

- Singularity
- Shifter
- Socker
- HTCondor VM and Docker universes
- Galaxy
- aCT

Galaxy: HPC portal for container tools



Galaxy: HPC portal for container tools

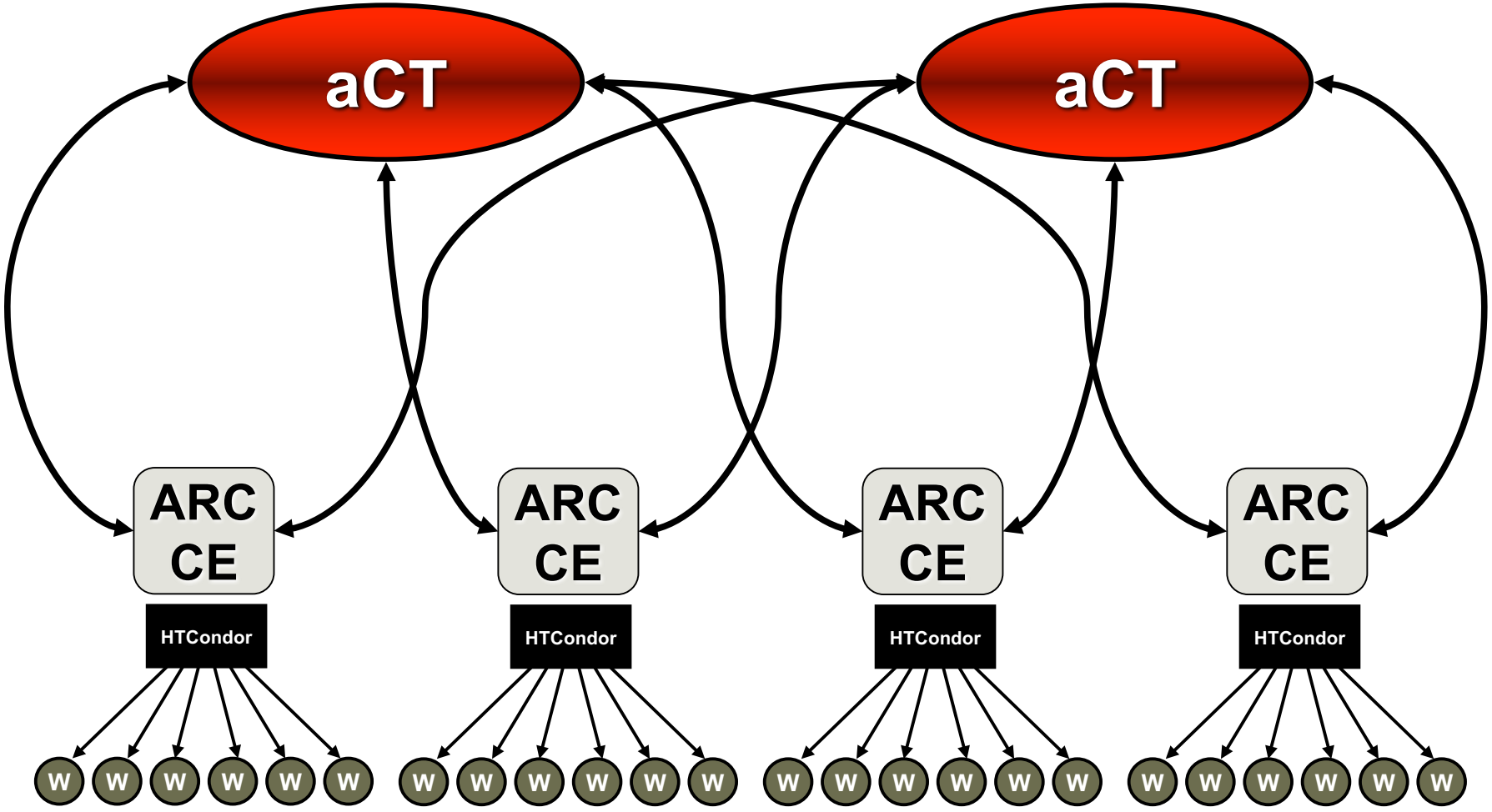


Galaxy: HPC portal for container tools

- Done:
  - In production at UiO: https://lifeportal.uio.no
  - Production support for singularity containers

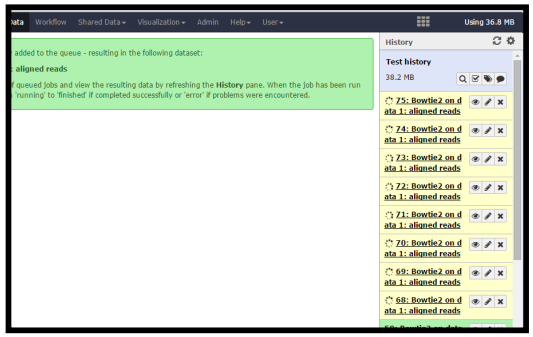
Prototypes

- Singularity
- Shifter
- Socker
- LSF/Docker
- HTCondor VM and Docker universes
- Galaxy
- aCT


aCT
aCT
ARC CE
ARC CE
ARC CE
ARC CE
HTCondor
HTCondor
HTCondor
HTCondor
W W W W W W
W W W W W W
W W W W W W
W W W W W W

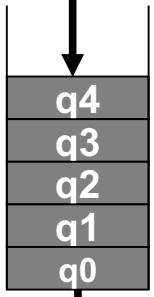
Galaxy and ARC

Users



Galaxy web Portal

Core Galaxy



Galaxy Job Queue

Job DB

Use cases

Use cases

- Service 5 has published a web-form for collecting research use cases for containers and VMs in HPC.
- **The web-form: https://skjema.uio.no/prace-containers**

Use cases: Submission report

Name of the software *

A single item or a comma separated list of software that should be packaged in a single VM or single container

- caffe
- ROS,gazebo
- FEniCS
- mriqc, freesurfer, heudiconv, everything from docker hub.
- upc,upc++,gasnet
- intel PCM, performance counter monitor
- GAMBIT
- FEniCS
- ARMplusplus, anvi'o
- GAMBIT
- Ubuntu
- matlab

Purpose of the software *

Answer	Number of	Percentage
Data analytics	7	43.8%
Virtualization	3	18.8%
Deep learning	4	25%
Machine learning	1	6.2%
Other	6	37.5%

Other purpos(es) *

- Computational Fluid Dynamics
- Pure computer science, applied math, physics, geophysics, cardiac modeling, etc.
- energy and performance measurement
- Numerical Simulation
- Running arbitrary software inside.
- calculation of wind field

What type of packaging *

Answer	Number of	Percentage
Virtual Machine	5	31.2%
Container	11	68.8%

The container already exists? *

e.g. on Docker hub or Singularity hub

Answer	Number of	Percentage
Yes	8	66.7%
No	4	33.3%

The virtual machine already exists? *

On a public server

Answer	Number of	Percentage
Yes	1	50%
No	1	50%

Does the software support parallelization? *

Answer	Number of	Percentage
Yes	13	81.2%
No	3	18.8%

What kind of parallelization? *

Answer	Number of	Percentage
Shared memory (e.g. OpenMP)	9	56.2%
Destributed memory (e.g. MPI)	9	56.2%

Approximately how many researchers will use this software? *

Answer	Number of	Percentage
Less than 5	2	12.5%
Between 5 and 20	10	62.5%
More than 20	4	25%

Use cases: Submission report

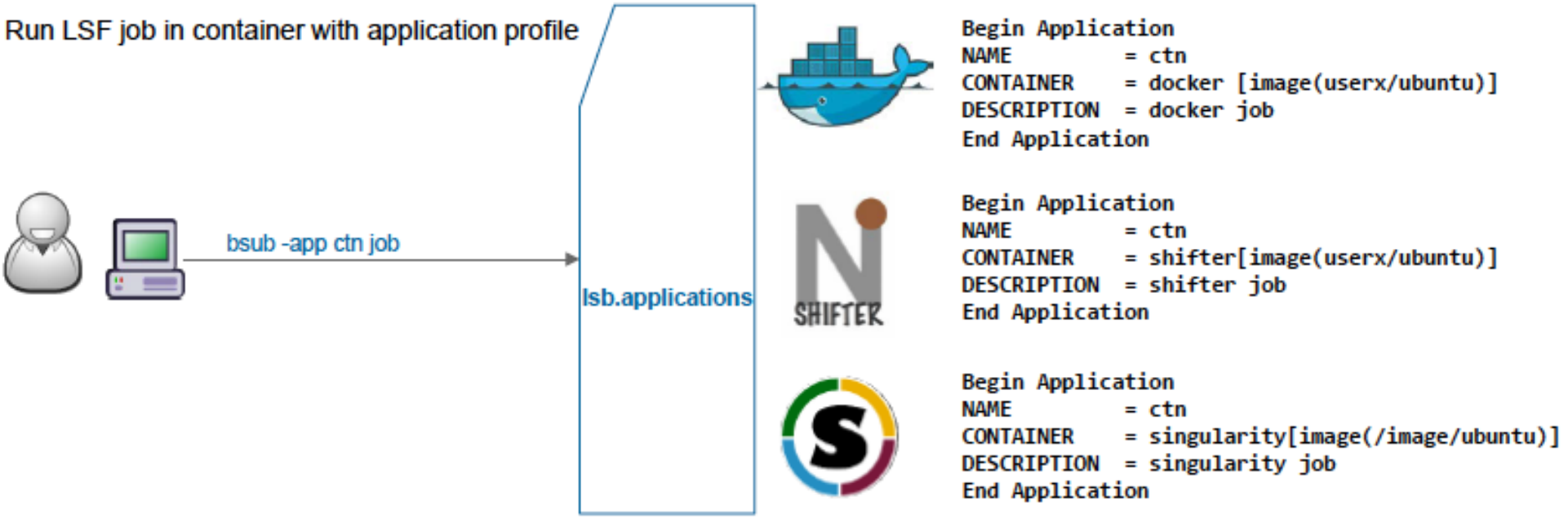
- The majority of use cases are for containers.
- Most use cases are for software that supports parallelisation.
- Most of the requested containers are publicly available


Q&A

Prototypes

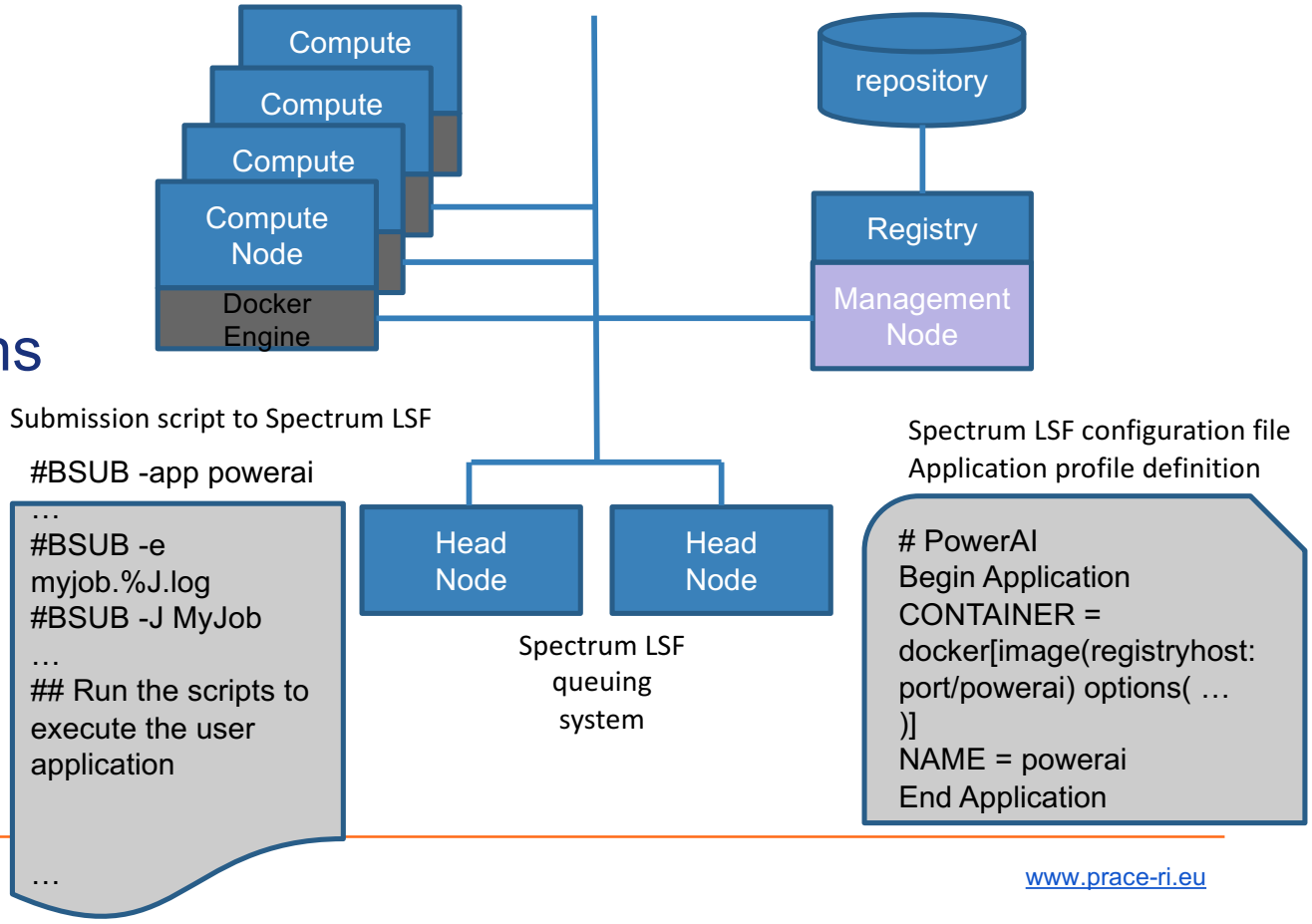
- Singularity
- Shifter
- Socker
- LSF/Docker
- HTCondor VM and Docker universes
- Galaxy
- aCT

LSF/Docker: IBM Spectrum LSF jobs in Docker containers



LSF/Docker: IBM Spectrum LSF jobs in Docker containers

- IDRIS Installation:
  - Power8 machine
  - More user restrictions
  - Better Security



LSF/Docker: IBM Spectrum LSF jobs in Docker containers

- Ongoing security tests:
  - Secure Docker installation on the Linux hosts
  - Docker daemon and registry configuration and image management
  - How LSF specifications are forwarded to the Docker environment.

LSF/Docker: IBM Spectrum LSF jobs in Docker containers

- Done:
  - Docker support is deployed in the LSF cluster at IDRIS. Tests are ongoing
- Future:
  - Collaborate with IBM to improve the configuration to match the security restrictions and the user needs.
  - Complete the test and evaluation of the platform, and produce the report