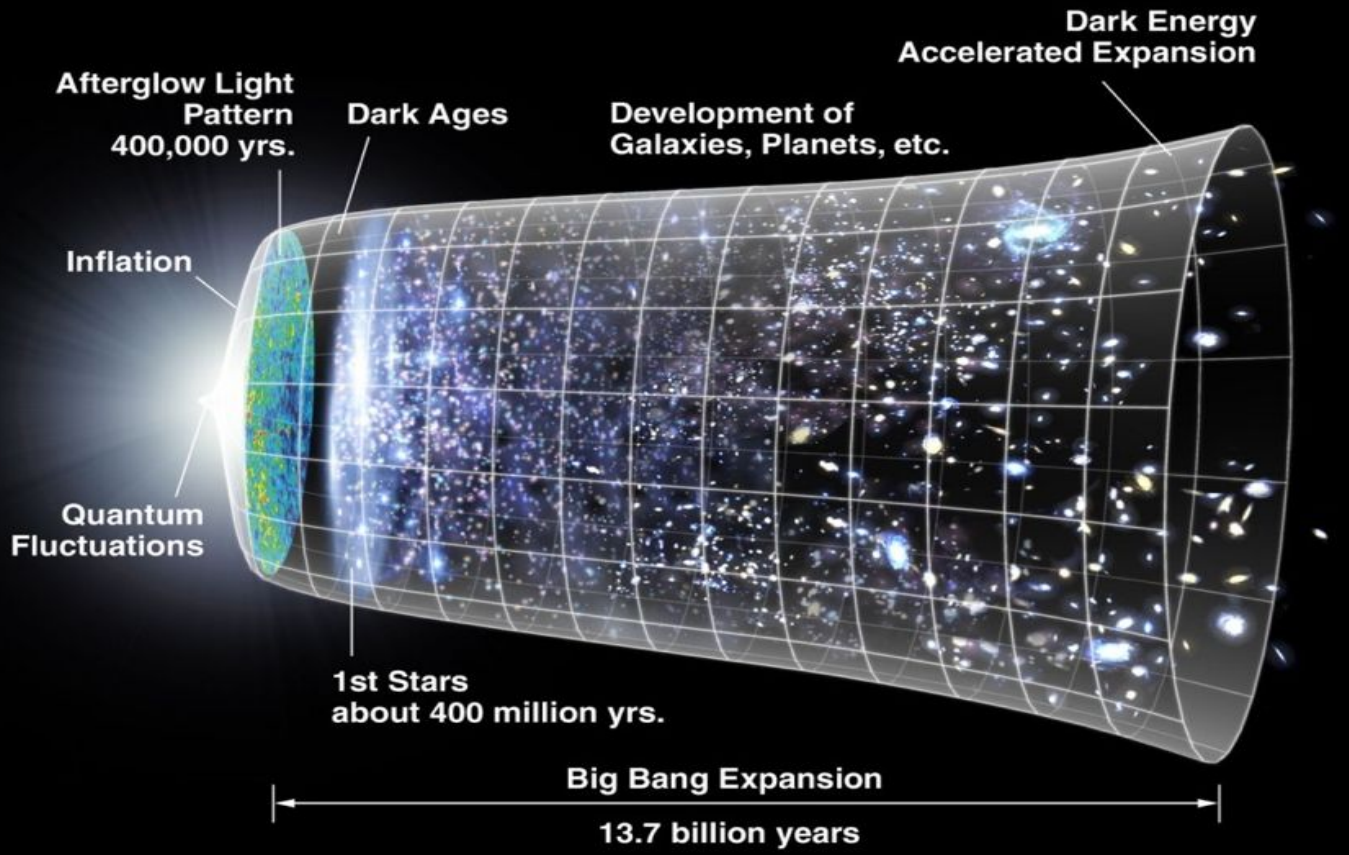
# Containers on Baremetal And Preemptible Servers

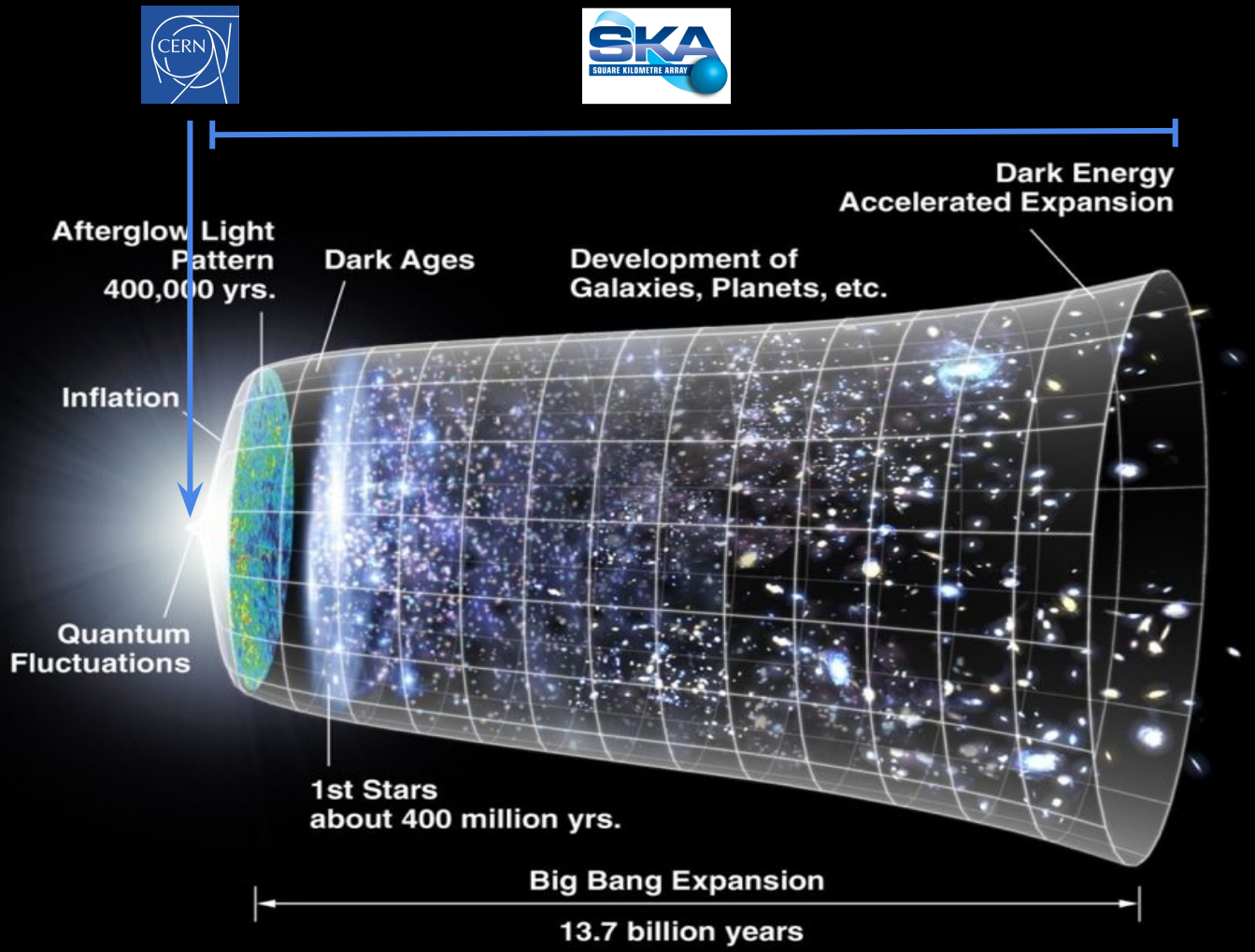
At CERN and SKA

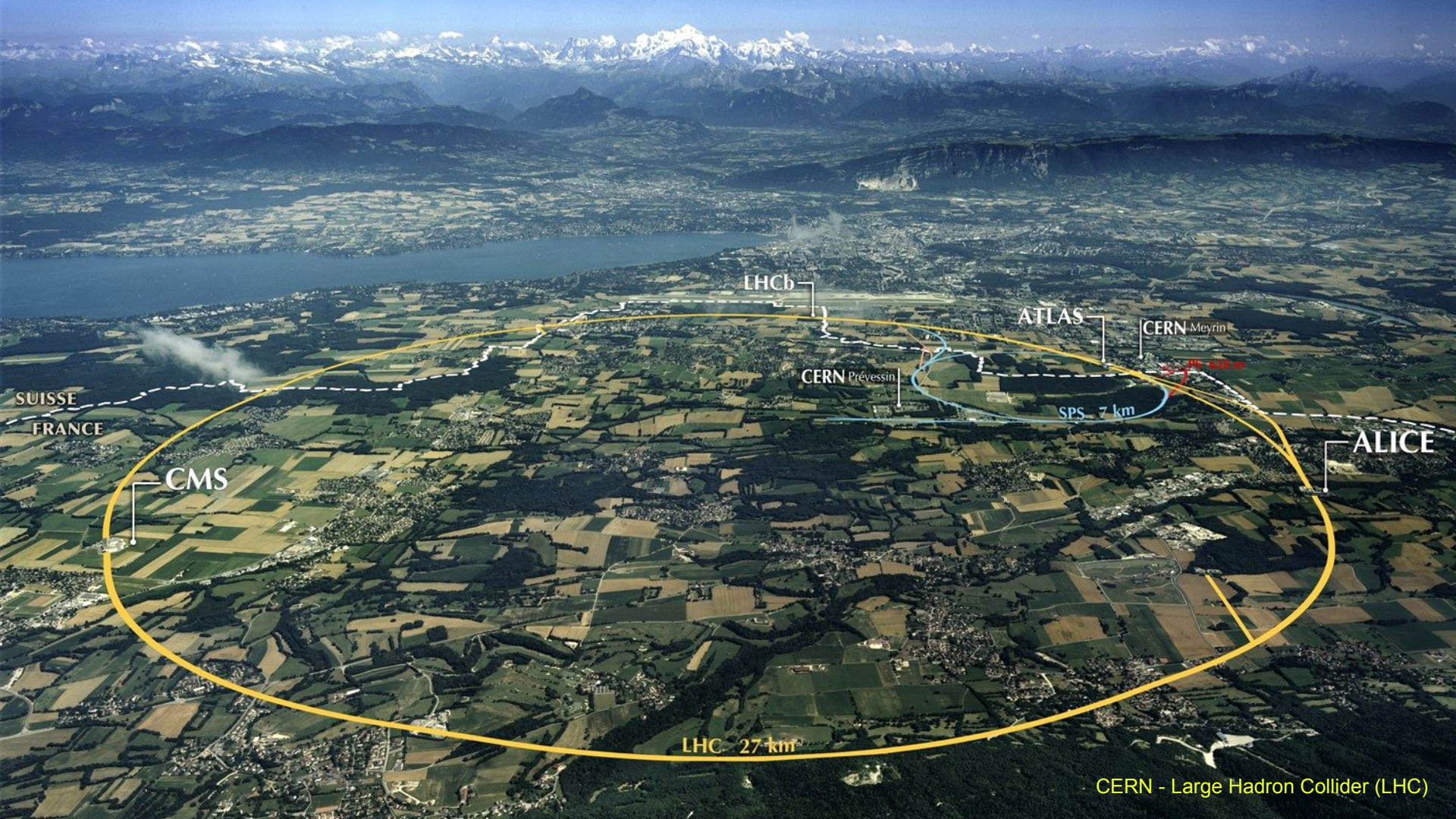
Belmiro Moreira, CERN
@belmoreira

John Garbutt, StackHPC
@johnthetubaguy

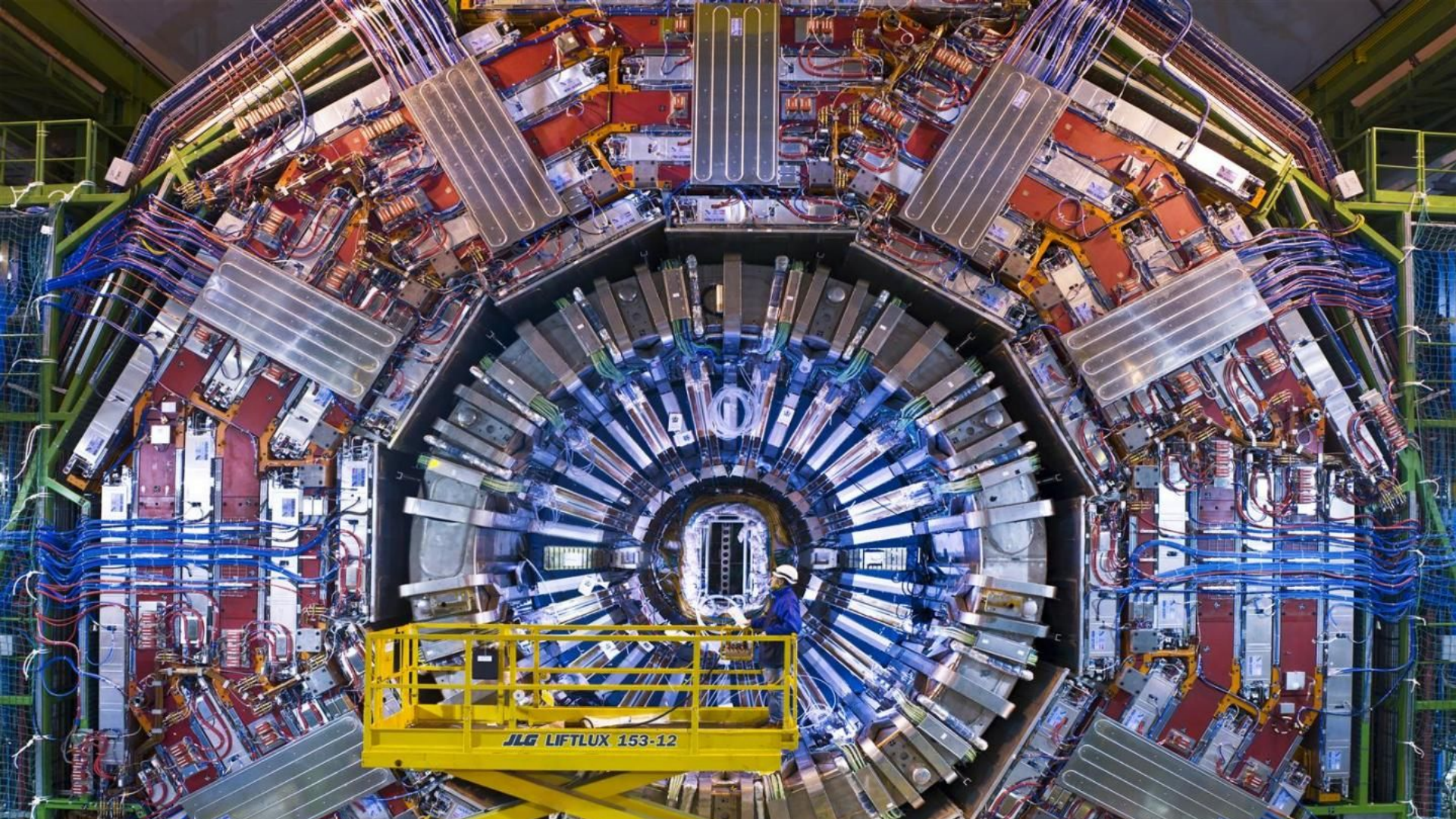

CERN
SKA
SQUARE KILOMETRE ARRAY


Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
400,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
Quantum
Fluctuations
1st Stars
about 400 million yrs.
Big Bang Expansion
13.7 billion years


CERN
SKA
SQUARE KILOMETRE ARRAY
Dark Energy
Accelerated Expansion
Afterglow Light
Pattern
400,000 yrs.
Dark Ages
Development of
Galaxies, Planets, etc.
Inflation
Quantum
Fluctuations
1st Stars
about 400 million yrs.
Big Bang Expansion
13.7 billion years




CERN - Large Hadron Collider (LHC)


JLG LIFTLUX 153-12

## Cloud resources

| Used | Available | Used | Available | Used | Available |
|---|---|---|---|---|---|
| 299.9 K cores | 316.6 K cores | 829.9 TiB RAM | 924.6 TiB RAM | 10.0 PiB disk | 15.3 PiB disk |

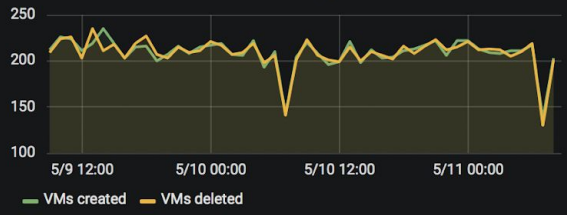
## Openstack services stats

| Users | Projects | VMs | Magnum clusters | Hypervisors | Images | Baremetal nodes |
|---|---|---|---|---|---|---|
| 3264 | 4296 | 37067 | 260 | 9113 | 2843 | 1147 |

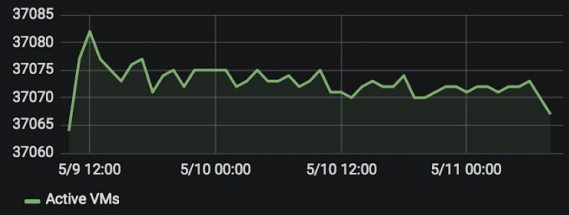
| Volumes | Volume size | Fileshares | Fileshares size |
|---|---|---|---|
| 5729 | 1.77 PiB | 103 | 191 TiB |

## Resource overview by time

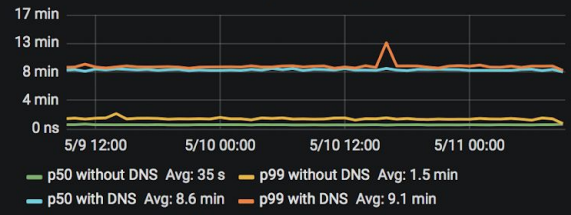
### VM changes



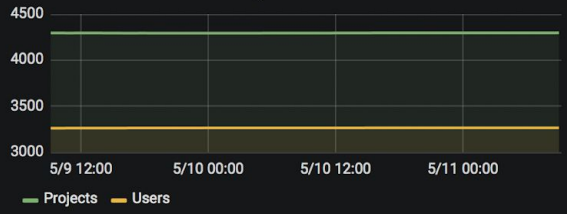

### Total VMs



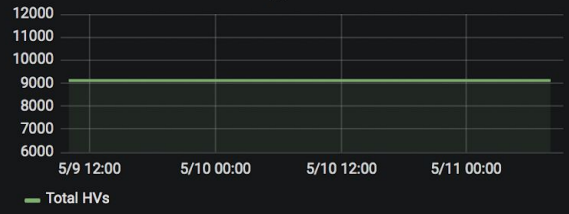

### Average VM boot time



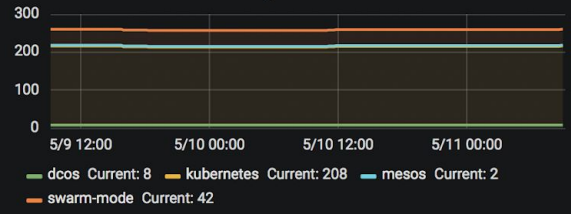

### Projects and users




### Hypervisors




### Magnum clusters




CERN - Cloud resources status board - 11/05/2018@09:23

# What is SKA?

Image courtesy of CSIRO

**Antennas**

**Digital Signal Processing (DSP)**

Transfer antennas to DSP
2020: 5,000 PBytes/day
2030: 100,000 PBytes/day

Over 10's to 1000's kms

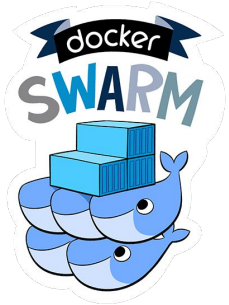
To Process in HPC
2020: 50 PBytes/day
2030: 10,000 PBytes/day

Over 10's to 1000's kms

**High Performance Computing Facility (HPC)**

**HPC Processing**
**2023: 250 PFlop**
**2033: 25 EFlop**

# Containers on Baremetal



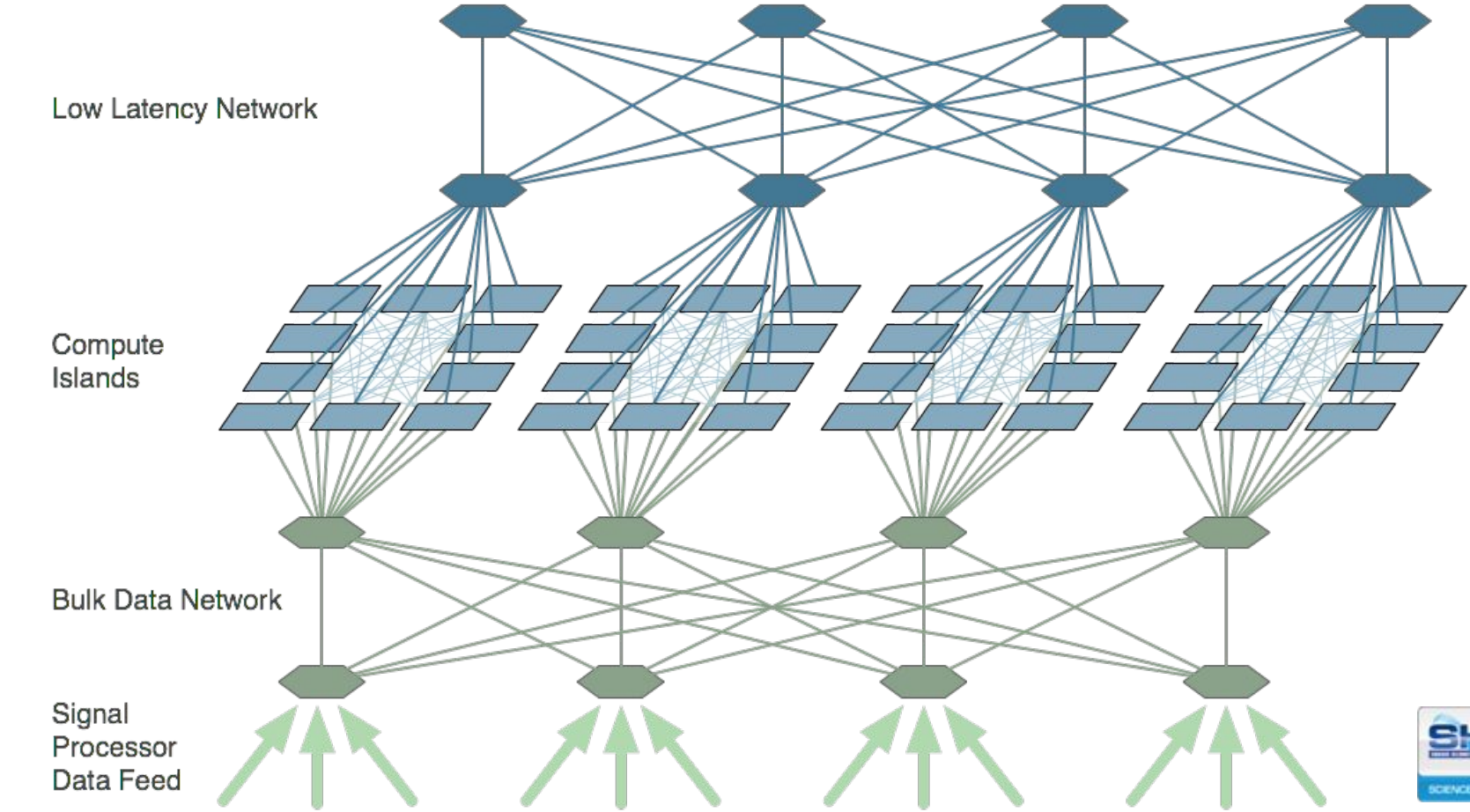

MAGNUM

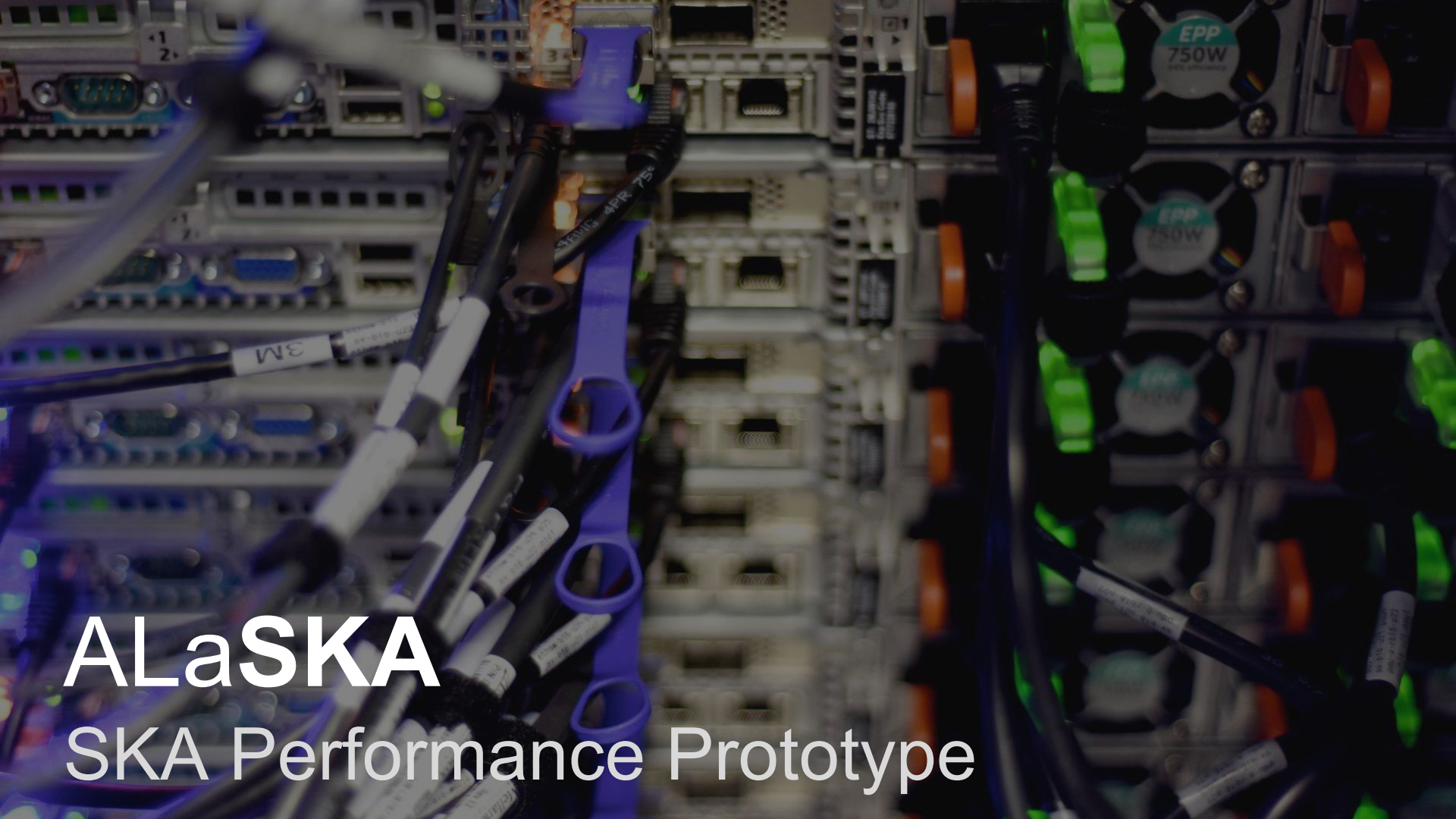
an OpenStack Community Project

IRONIC

an OpenStack Community Project

# SKA's Science Data Processor

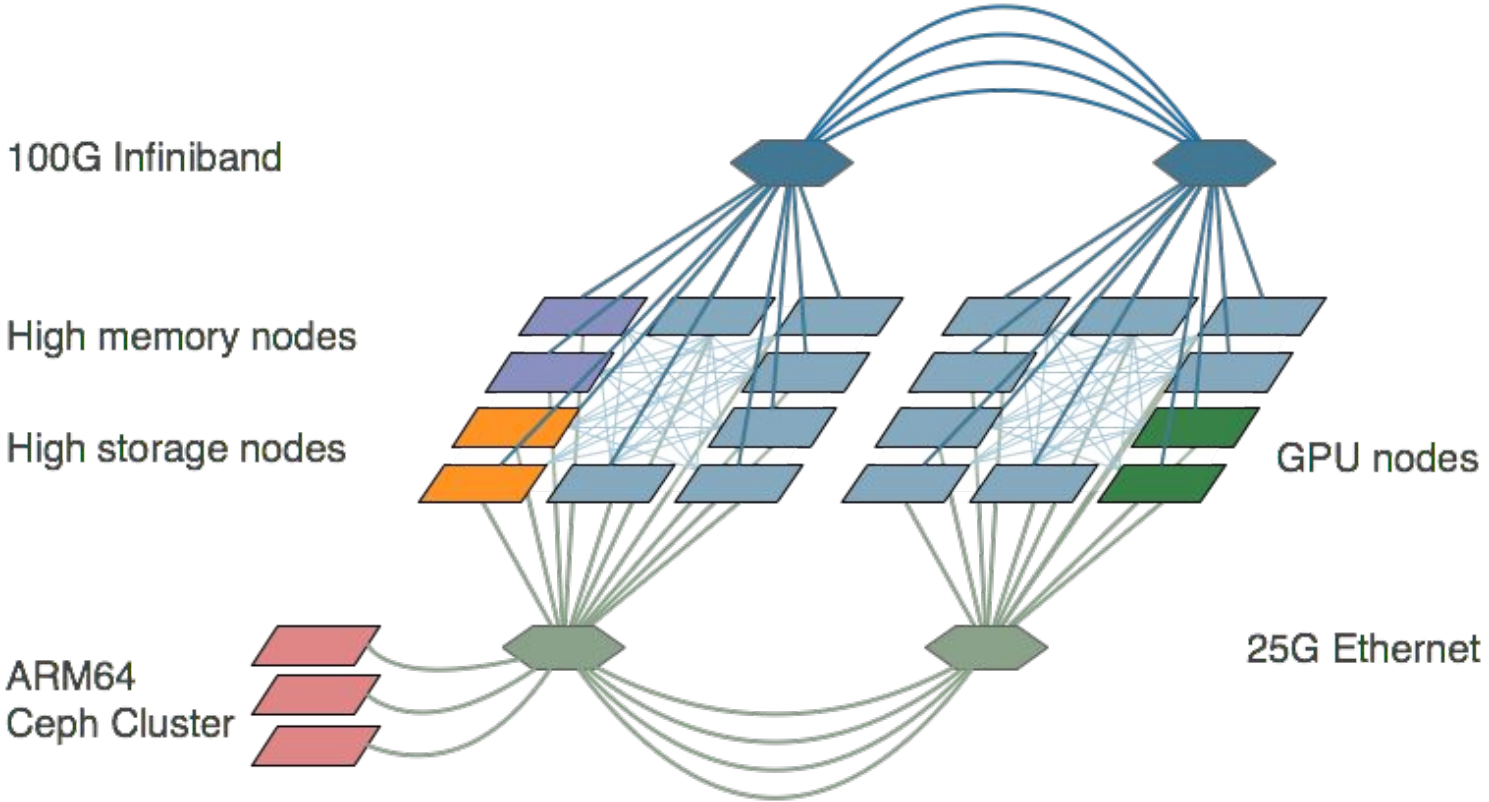





# ALa**SKA**

## SKA Performance Prototype

# AlaSKA


100G Infiniband
High memory nodes
High storage nodes
GPU nodes
25G Ethernet
ARM64
Ceph Cluster

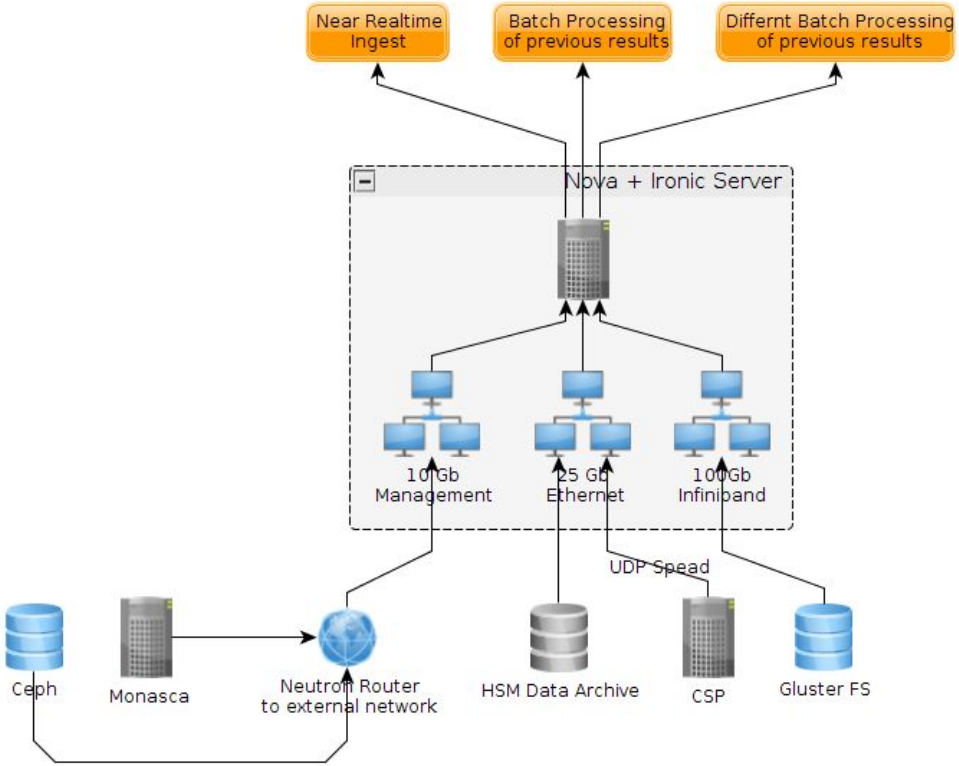
# Why Baremetal? Why Containers?

- Single security zone
- No need for virt, so target baremetal
- 30 seconds to switch ingest to Supernova, Fast radio burst, …
- Easier development and deployment cycles

# Magnum with Ironic


- Magnum used extensively at CERN
- Docker Swarm and Kubernetes are supported
- Historically a separate driver for baremetal, badly maintained
- Queens moves to using Fedora Atomic for VM and baremetal

# System Integration Prototype



# Lessons learned

- Extra network ports added after initial setup
- Updated Docker version in Fedora Atomic 27
- Updating Atomic image with RDMA drivers was tricky
- Root disk wasn’t resized by cloud-init

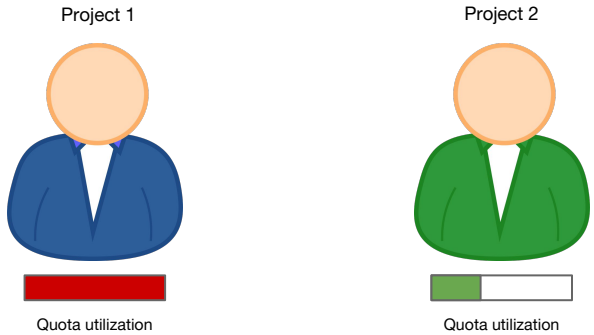
http://www.stackhpc.com/magnum-queens.html

# Preemptible Instances

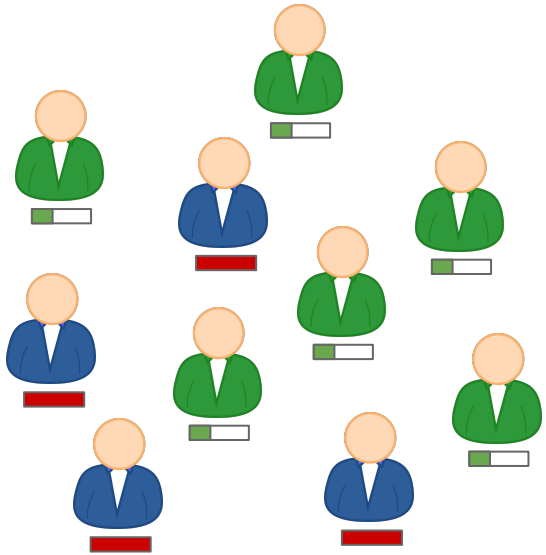
# Resources Utilization

- Public Clouds give the illusion of infinite capacity
  - Users pay for resources that they use
- Private Clouds
  - Resource management usually is based in project quotas
  - Prevent resources being exhausted
  - Prevent “over-committing” resources/quota
  - Manage individual projects requirements
  - Reserve resources for operations with higher priority
  - Scientific Clouds
    - Projects have different funding models
    - They expect a predefined number of resources available
    - But not always these resources are used full time

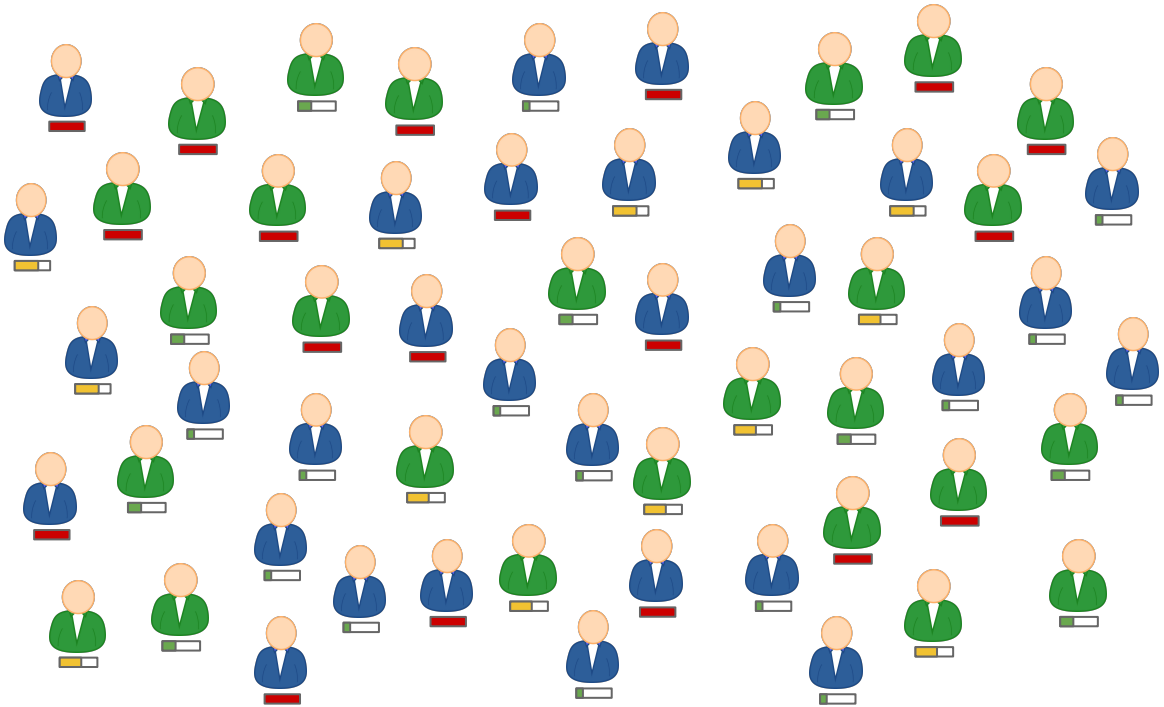
# Idle Resources with quotas



# Idle Resources with quotas



# Idle Resources with quotas



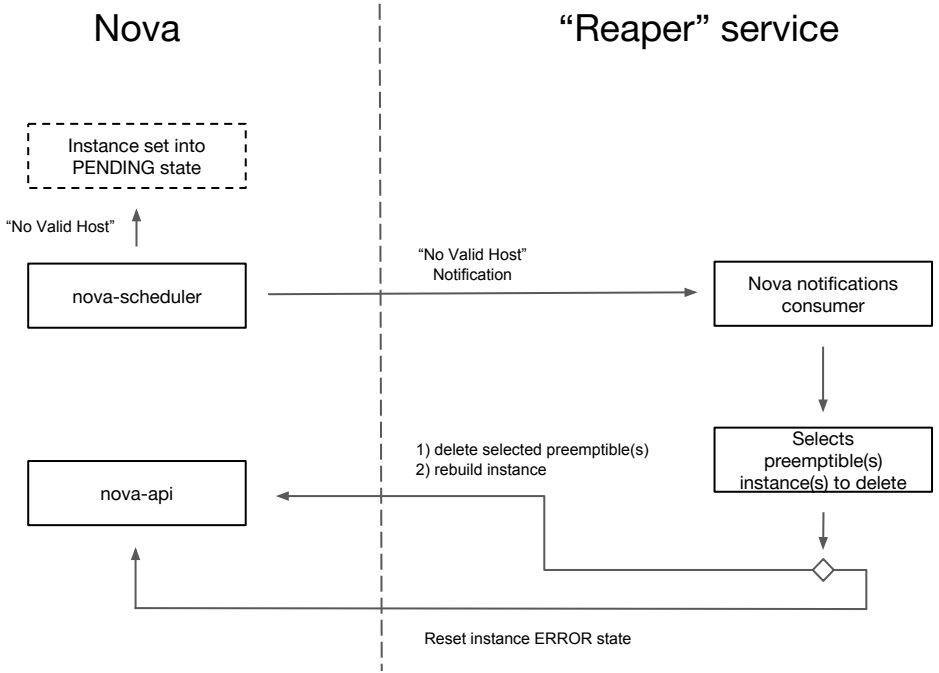
# Maximize Resource Utilization

- Public Clouds
  - Based on different pricing/SLA considering resource availability
  - Reserved instances vs spot-market
- Private Clouds
  - Quotas are hard limits. Leads to a reduction in resource utilization
  - Preemptible instances
    - Projects that exhausted their quota can continue to create instances
      - Opportunistic workloads
      - Low SLA

# Preemptible Instances

- Proposal to implement Preemptible Instances into OpenStack
  - Build a prototype
  - Minimise changes required in OpenStack nova
- Starting simple
  - Use dedicated projects for Preemptible Instances
    - Avoids tagging individual instances
  - Introduce a “Reaper” service
    - Orchestrator to manage preemptible instances
      - Removes preemptible instances when resources are required for non preemptible instances
      - Applies a maximum TTL to preemptible instances

# Workflow


Nova
“Reaper” service
Instance set into PENDING state
“No Valid Host”
nova-scheduler
“No Valid Host” Notification
Nova notifications consumer
Selects preemptible(s) instance(s) to delete
1) delete selected preemptible(s)
2) rebuild instance
nova-api
Reset instance ERROR state

# Workflow

- The creation of a non preemptible VM fails because there aren’t available resources
- Instances that fail with “Nova Valid Host”, go to “PENDING” state instead of “ERROR”
- The Reaper service is notified and it tries to free the requested resources
  - Rebuild the instance
  - Or change instance state to “ERROR”

# Current work in Preemptible Instances

- Add instance state PENDING (spec)
  - https://review.openstack.org/#/c/554212/
- Allow rebuild instances in cell0 (spec)
  - https://review.openstack.org/#/c/554218/
- Add scheduling notification
  - https://review.openstack.org/#/c/566470/
- Implement instance state PENDING
  - https://review.openstack.org/#/c/566473/
- Reaper prototype:
  - https://gitlab.cern.ch/ttsiouts/ReaperServicePrototype

Join the Scientific SIG and...

# Get involved!

https://www.openstack.org/science/

Belmiro Moreira, CERN
@belmoreira

John Garbutt, StackHPC
@johnthetubaguy

Join the Scientific SIG and...

# Get involved!

https://www.openstack.org/science/