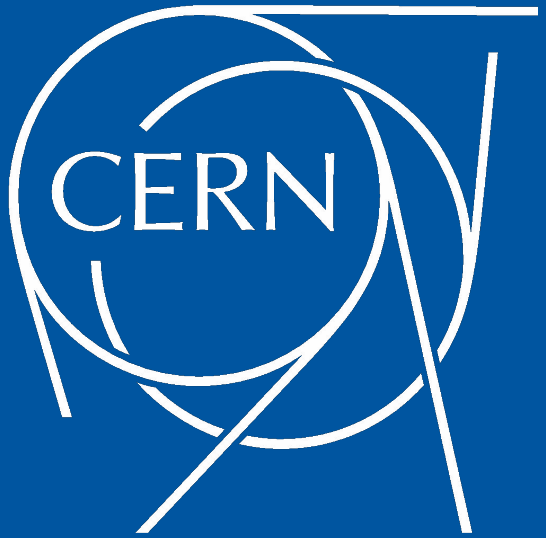
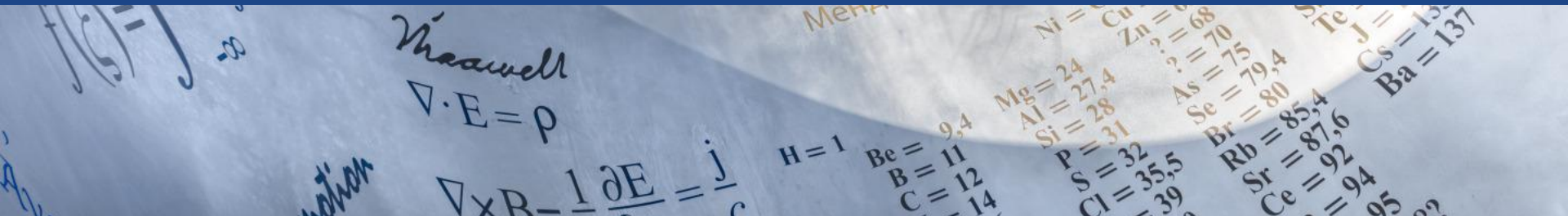

CERN

Optimizing OpenStack Nova for Scientific Workloads

CHEP 2018 - Bulgaria, 2018

Belmiro Moreira

belmiro.moreira@cern.ch @belmiromoreira



Outline

- CERN Cloud Infrastructure
- Virtualization Overhead challenge
- Resource Utilization challenge
- CERN/SKA collaboration
  - How to Remove Virtualization Overhead
  - How to Increase Resource Utilization

CERN Cloud Infrastructure

- Production service since July 2013
- OpenStack based
  - Offering more than 15 OpenStack projects
  - Keystone, Nova, Neutron, Glance, Cinder, Magnum, Ironic, …
- Running the last OpenStack release (Queens)
- 2 Data Centres (Geneva and Budapest)
  - One Region
  - 5 Availability Zones
  - Nova Cells (>70 Cells)
- More details about the OpenStack services provided by CERN Cloud
  - (Tue@11:15 - T7, S3) Advanced features of the CERN OpenStack Cloud

Used	Available	Used	Available	Used	Available
314.5 K cores	317.4 K cores	866.2 TiB RAM	928.7 TiB RAM	10.3 PiB disk	15.3 PiB disk

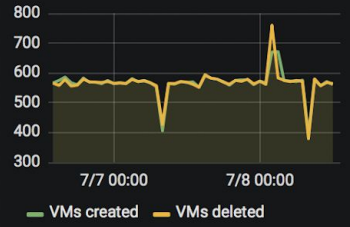
Openstack services stats

Users	Projects	VMs	Magnum clusters	Hypervisors	Images	Baremetal nodes
3356	4351	39090	292	9117	3040	1247

Volumes	Volume size	Fileshares	Fileshares size
5844	1.81 PiB	126	210 TiB

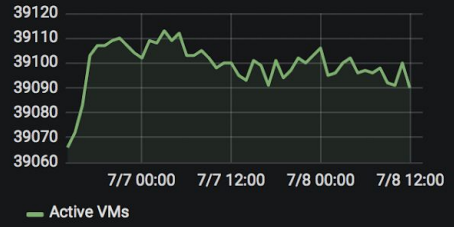
Resource overview by time

VMs created/deleted

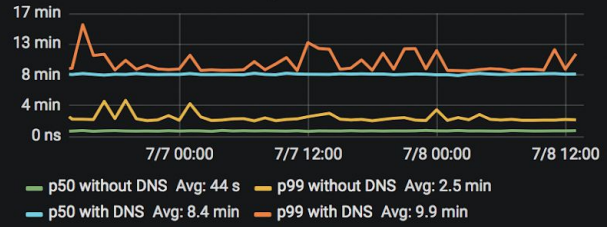


VM changes

Total VMs



Average VM boot time

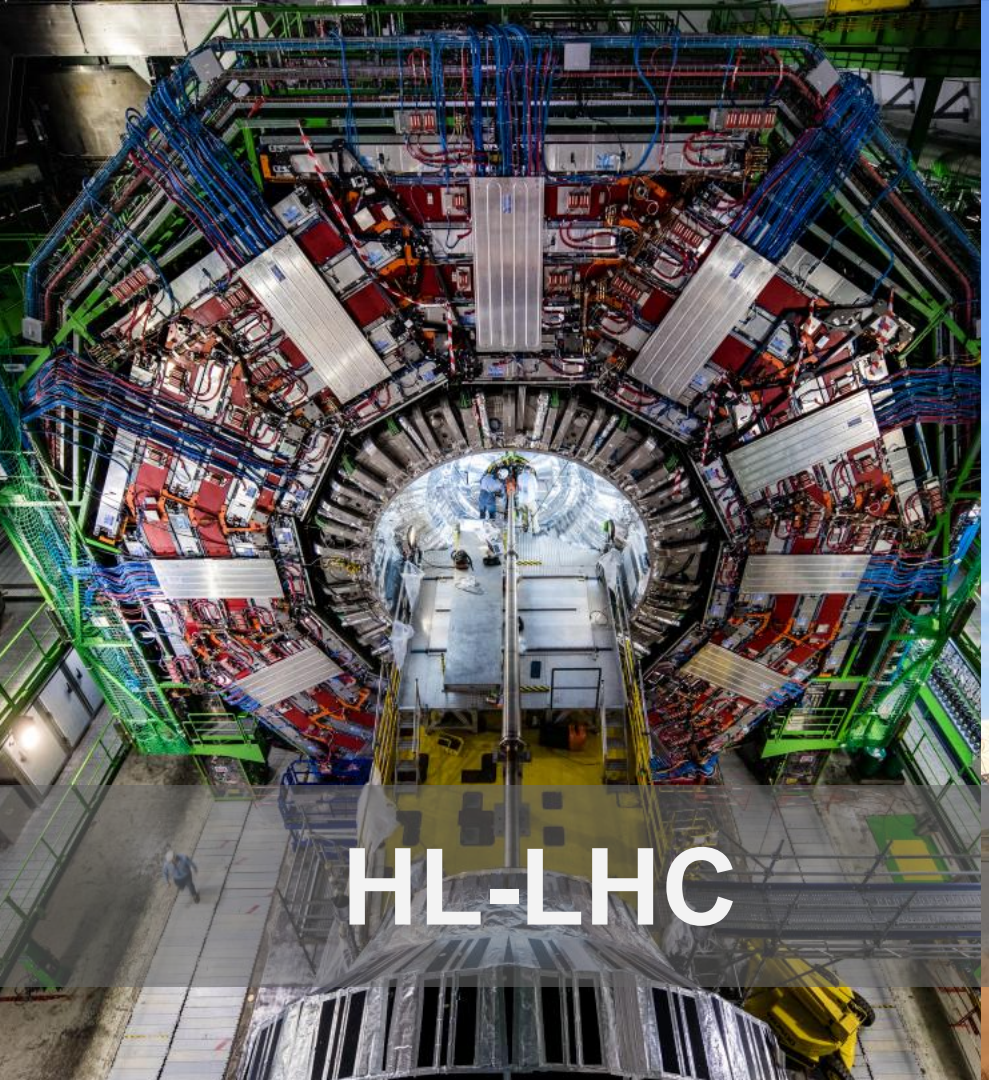


Virtualization Overhead/Resource Utilization

- >80% of the CERN Cloud resources are dedicated for Batch processing
- CPU Performance Optimizations deployed in the CERN Cloud
  - NUMA + CPU pinning
  - 2MB Huge Pages
  - EPT (enabled)
  - KSM (enabled)

Current Virtualization CPU Overhead : **~3%**

- Virtualization Overhead depends on VM size
  - 4 VMs (8 cores each) the Virtualizition overhead reduced from 7.8% to 3.3%
  - 2 VMs (16 cores each) the Virtualization overhead reduced from 16% to 4.6%
- Quotas are per Project
  - No Overcommit. Unused quota = Unused resources


HL-LHC


SKA

How to remove the Virtualization Overhead?

- Containers provide a lightweight alternative to VMs
  - Small footprint
  - Increased performance
- OpenStack Magnum
  - Container Orchestration Engine
  - Easy to deploy Kubernetes/Swarm clusters
  - Clusters are per tenant
- Clusters usually deployed using VMs
  - Virtualization Overhead!
  - VMs are used for Security
  - Flexibility VS Performance

Containers on Baremetal

- Containers deployed directly on Baremetal
- OpenStack Ironic
  - Baremetal provisioning using OpenStack Nova APIs
- Magnum to deploy clusters directly in Baremetal
  - CERN and SKA working to add Baremetal support into Magnum
  - Several assumptions required
    - network, location, timeouts, ...
  - Ironic as provision engine
  - Kubernetes and Swarm supported

Containers on Baremetal

- Containers on Baremetal performance
  - Similar to native Baremetal
- SKA testing Containers on Baremetal in ALaSKA prototype
  - Performance and Fast context switching
- CERN evaluating how to run Batch in Containers on Baremetal
  - More RAM available for Jobs
  - More CPU time for Jobs throughput
- Containers on Baremetal can benefit other workloads
  - (Wed@12:00 - T7, S5) Apache Spark usage and deployment models for scientific computing

How to improve Resource 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

Preemptible/Spot Instances

- 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 - OpenStack Nova

- Preemptible Instances Workflow in OpenStack Nova
  - 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"
- CERN and SKA are working with OpenStack Nova team to implement Preemptible Instances
- CERN is deploying a Preemptible Instances prototype
  - Expected to be ready by the end of Q3/2018

Summary

- HL-LHC and SKA will produce an unprecedented amount of data to analyse
- Small compute inefficiencies in large infrastructures translate in a huge number resources underutilized
- Flexibility vs Efficiency
- CERN and SKA working together to build a High Efficient Infrastructure
- How to Remove Virtualization Overhead?
  - Containers on Baremetal
- How to Improve Resource Utilization?
  - Preemptible Instances

belmiro.moreira@cern.ch
@belmiromoreira

www.cern.ch