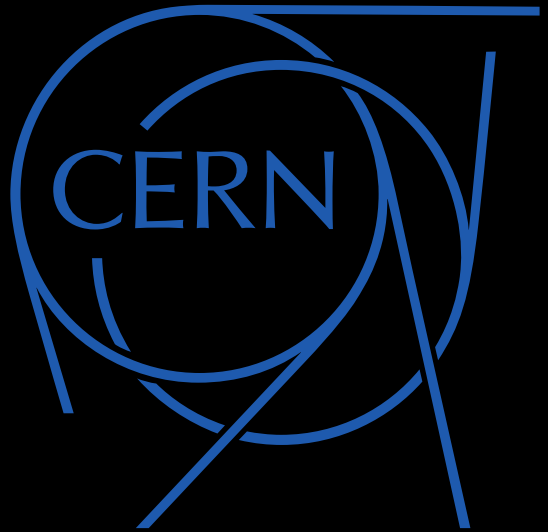
Opportunistic usage of the CMS Online cluster using a cloud overlay



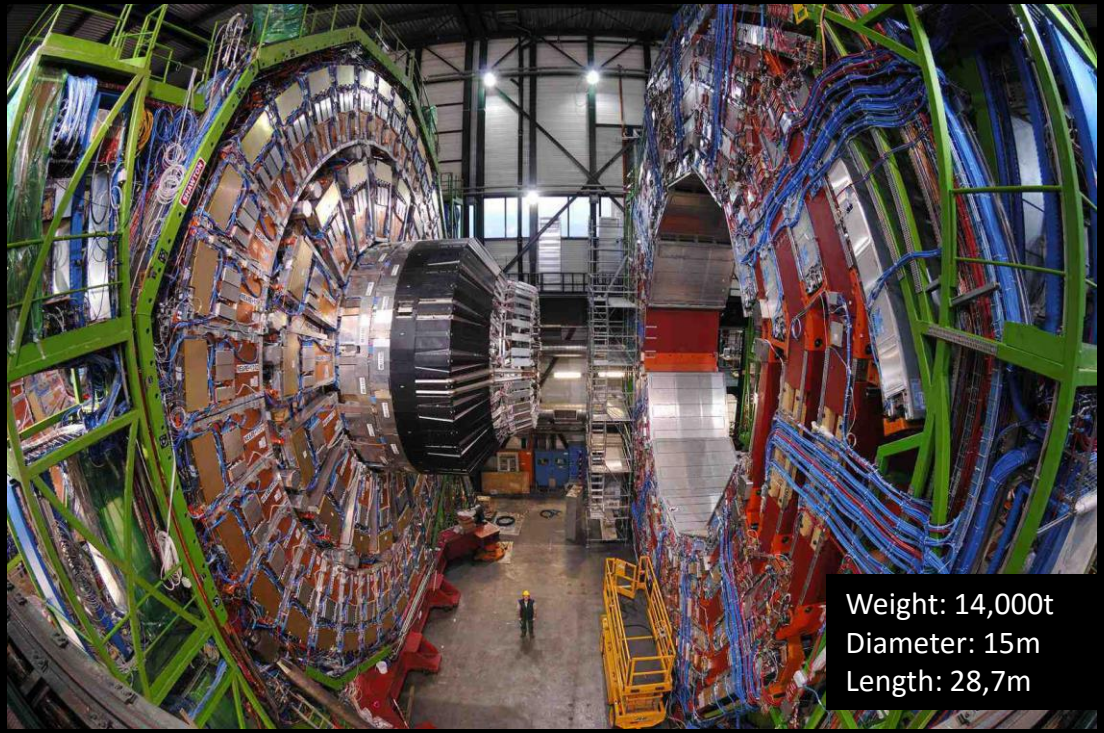
Olivier Chaze
CERN Geneva, Switzerland
for the CMS Data Acquisition group (DAQ)

ISGC 2016, 13th – 18th March
Academia Sinica – Taipei, Taiwan

CERN LHC (Large Hadron Collider)

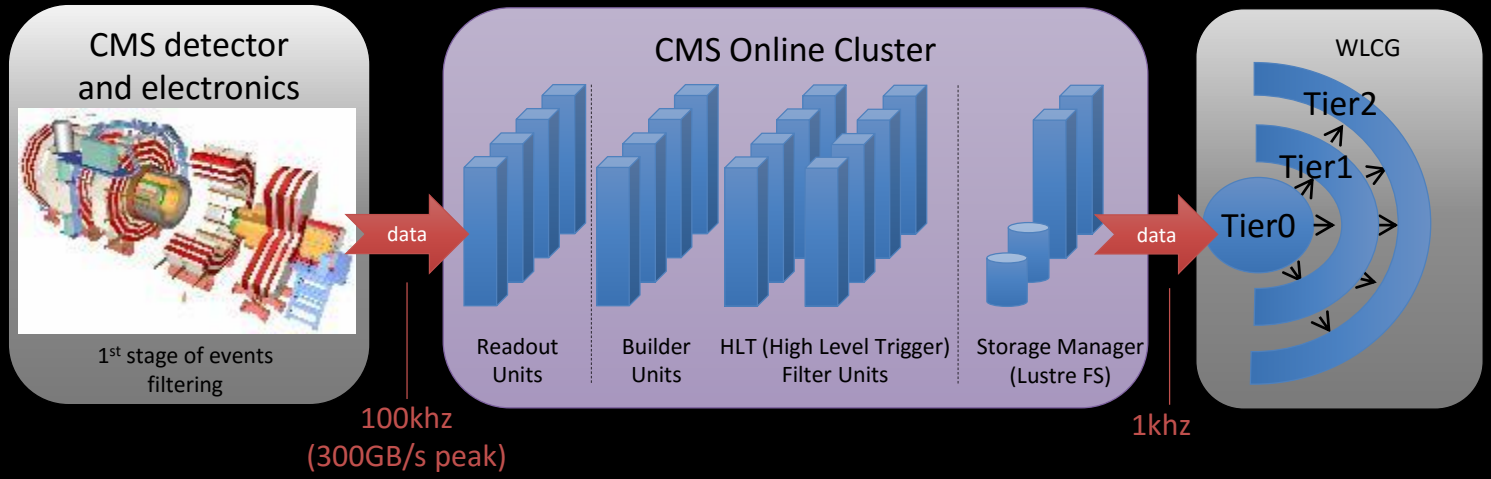
CMS: one of the 2 general purpose physics experiments at the LHC
LHC: biggest proton-proton particle accelerator in the world (27km cir)

The CMS (Compact Muon Solenoid) Detector



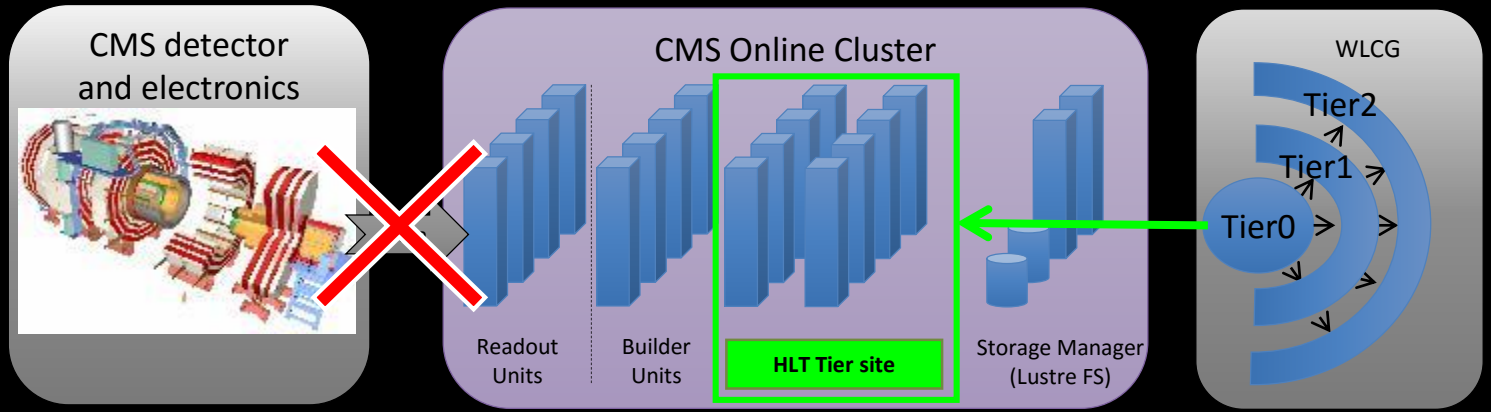
- 40 MHz (40 million collisions per second)
- Data from each collision is ~1 MegaByte
- 40 Terabytes per second

Flow of data



- The CMS Online Cluster is used to **filter interesting events** before shipping them to Tier0
- The data is then analysed offline by the Worldwide LHC Computing Grid (WLCG)

Cloud Purpose: Provide the computing power to the WLCG

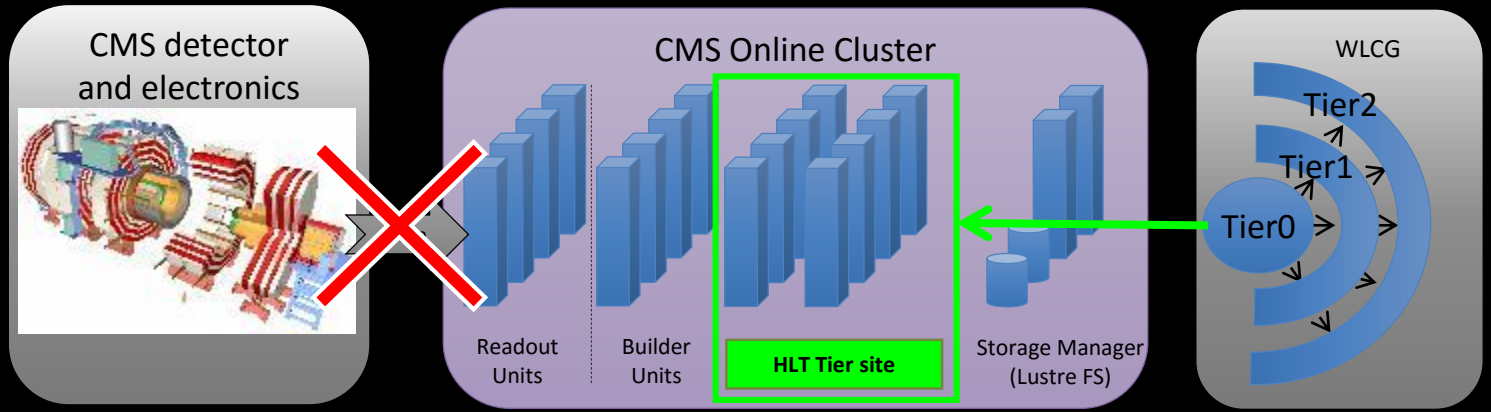


The cluster is idle during several weeks a year

Year	Technical stops	Machine development	Total
2015	116 days	15 days	131 days
2016	86 days	22 days	108 days

CMS HLT computing power

Nodes	Cores	RAM GB	Disk TB
904	16192	38144	450

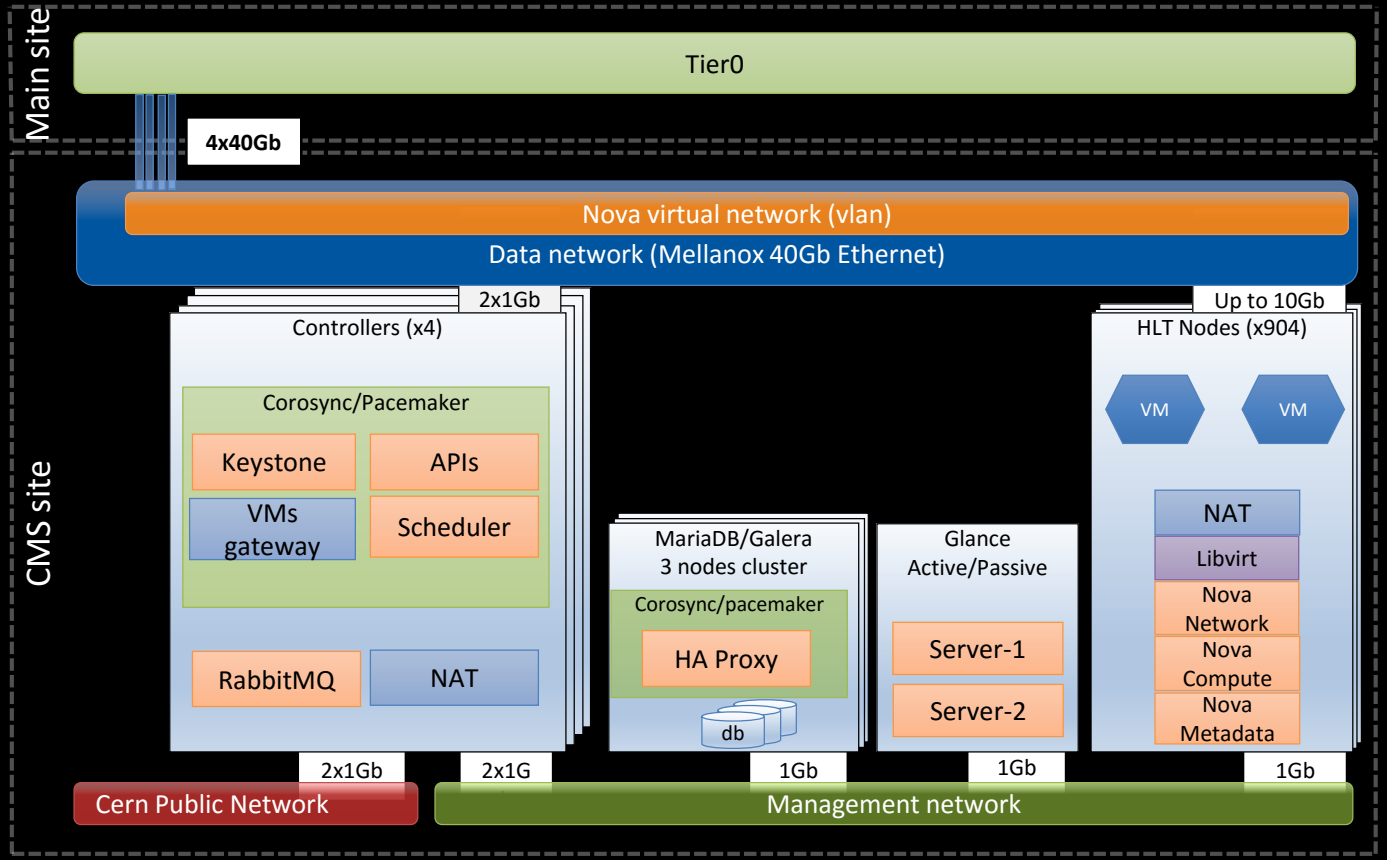


Farm size in 2015 (HEP-SPEC06)	HLT	Tier0	All CMS Tier1 sites
CMS	350K (~500k in 2016)	300K	~300K (>500k in 2016)

The CMS HLT Cloud

- Openstack Grizzly
- Only core services
(nova, glance, keystone, APIs)
- Corosync/Pacemaker
- RabbitMQ
- MariaDB/Galera cluster

Openstack architecture



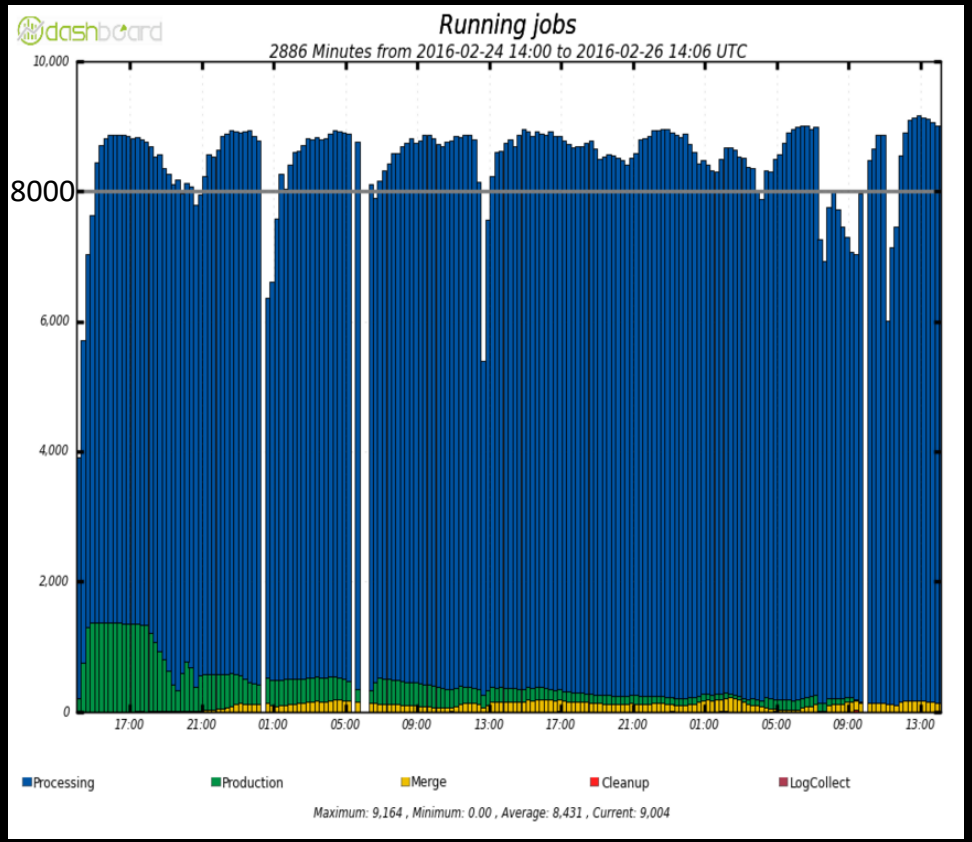
HLT Cloud specifications

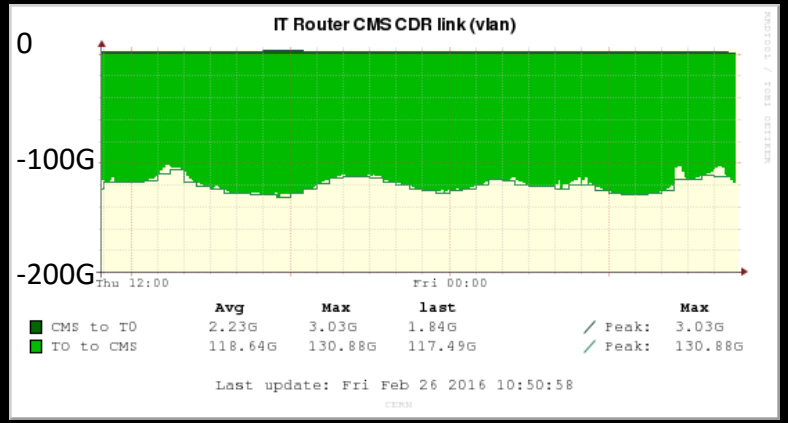
- Typical WLCG jobs specifications:
  - Single core
  - 2.5 GB of memory per job
  - 8 hours length job

HLT Compute nodes			Virtual machines			WLCG jobs
Nodes	**Cores**	**RAM GB**	**VMs per node**	**VM Cores**	**VM RAM GB**	**Jobs per VM**
288 x dual X5650	12	24	1	12	18	7
256 x dual E5-2670	16	32	2	8	13	5
360 x dual E5-2680v3	24	64	3	8	19	7
Total: 904	**16192**	38144	**1880**	16192	32360	**12136**

- Same VM image everywhere

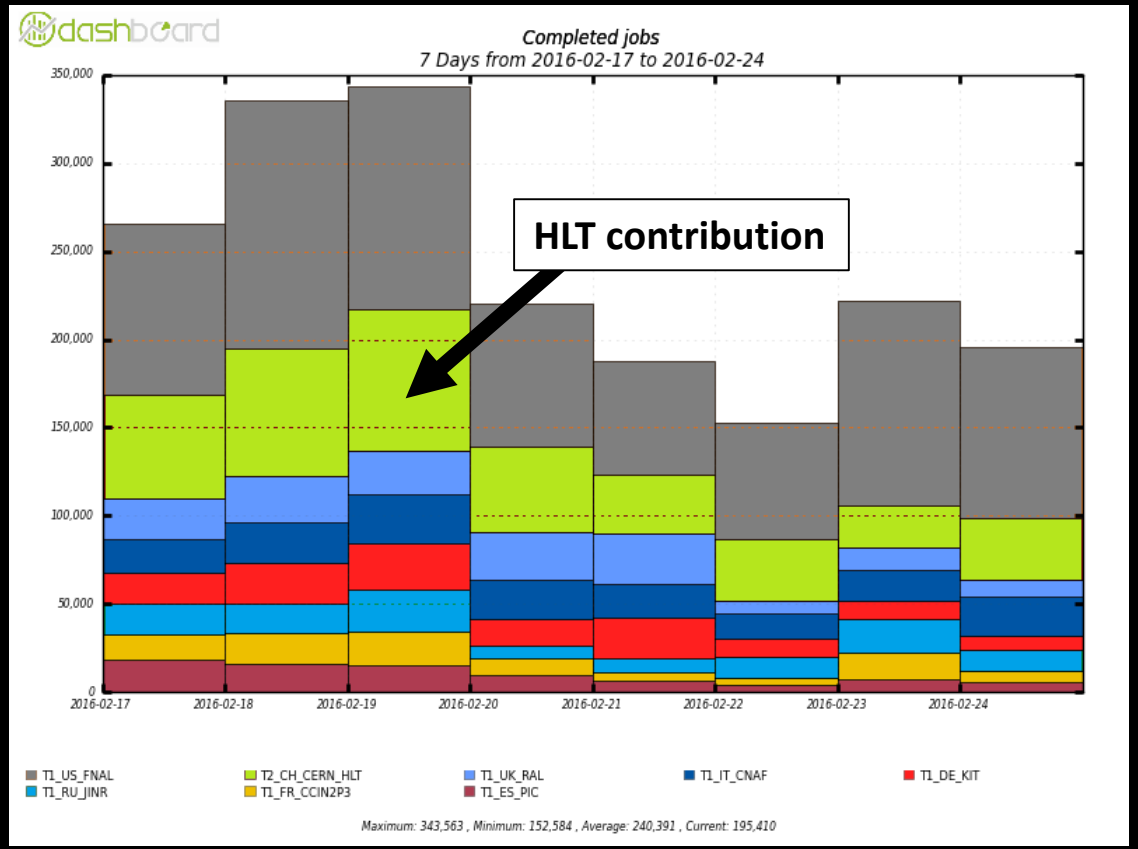
Jobs running in the Cloud with ~70% of the HLT farm





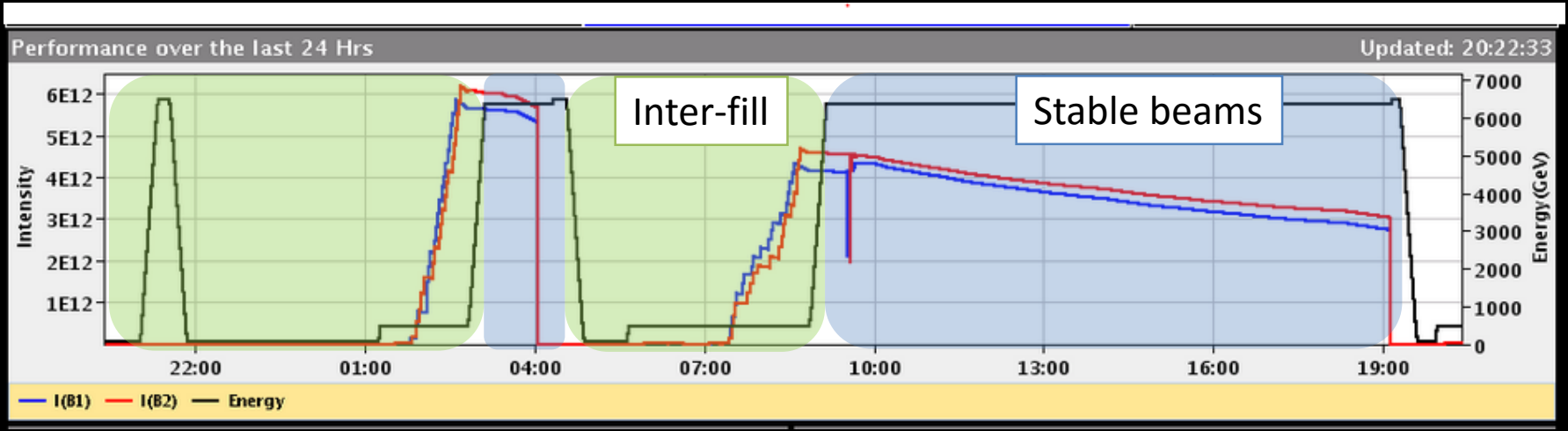
- Even network intensive jobs can run in the HLT

HLT Cloud participation vs Tier1 sites (~70% of the HLT farm used)



Challenge for 2016

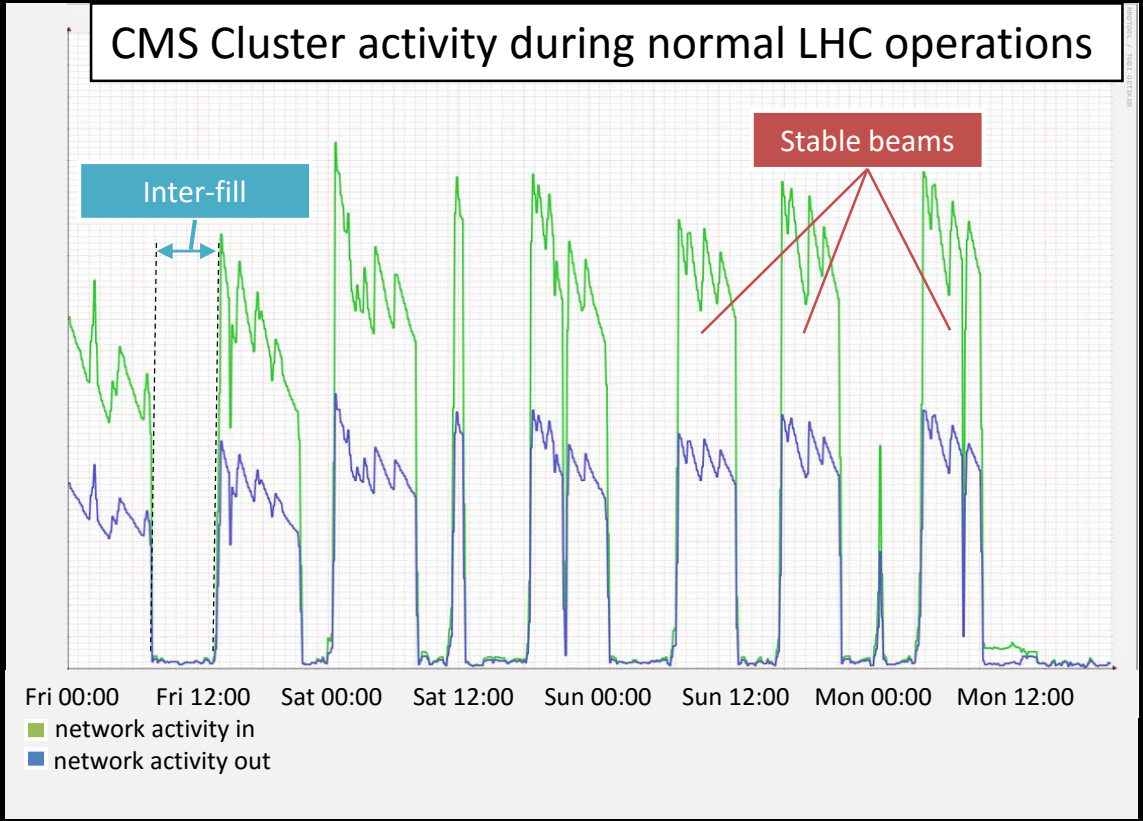
Inter-fill periods (3h to 6h in average)



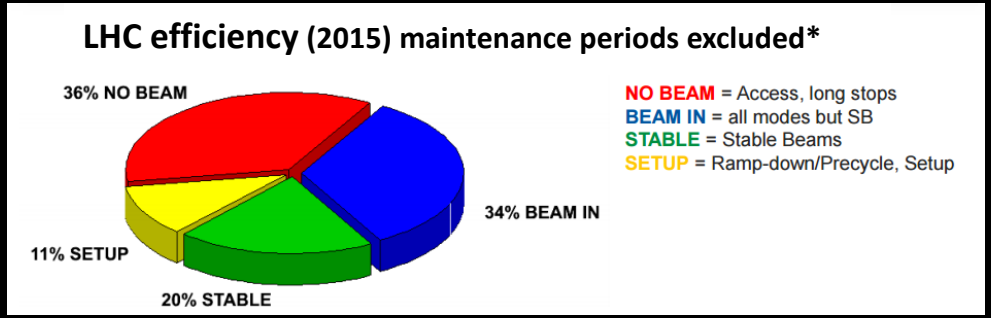
- Inter-fill period = Period between two stable beams
  - Stable beams = particles collisions ongoing
  - Inter-fill = no particles collisions ongoing

Challenge for 2016

Inter-fill periods (3h to 6h in average)



Inter-fill periods in number of days



Period	Number of days
maintenance periods	108
Inter-fill (2015 stats)	(365-108)x0.8 = **205**

*M. Solfaroli, "LHC operation and efficiency in 2015", LHC Performance Workshop, Chamonix, France, 25-28 January 2016

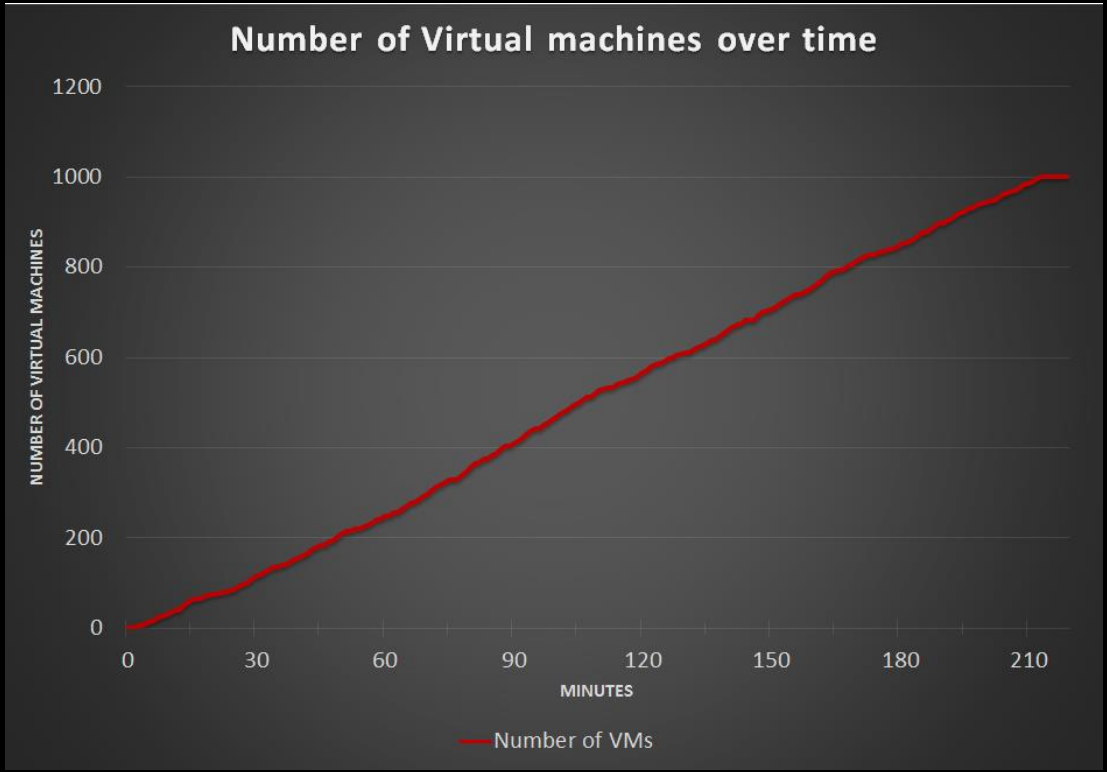
Inter-fill periods constraints

- Unpredictable and short periods of few hours

What's needed:

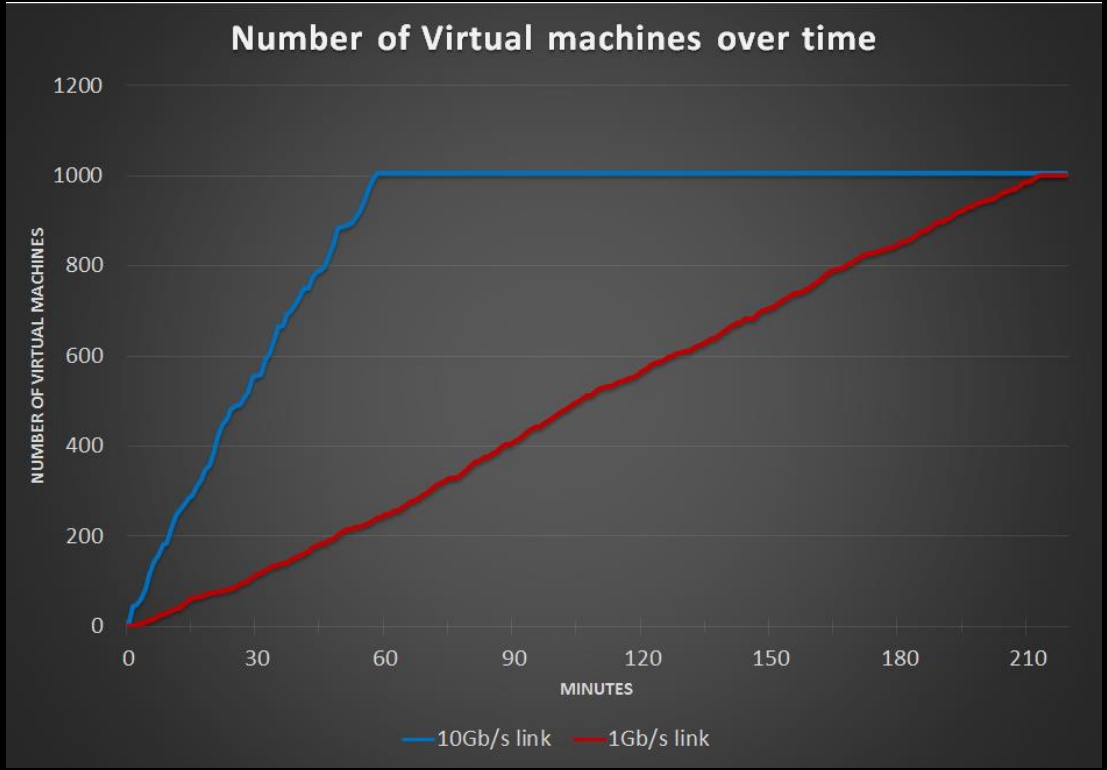
1. Start/Stop the Cloud quickly
2. Detect when begin/end the inter-fill periods
3. Select which nodes can be allocated to the cloud
4. Start/Stop the Cloud automatically
5. Need specific grid jobs length for better efficiency

1.1 The HLT OpenStack Cloud performance without tuning



- The network is the bottleneck (nova-glance)
- The image is served via a 1Gb/s

1.2 The HLT OpenStack Cloud performance with a 10Gb/s link



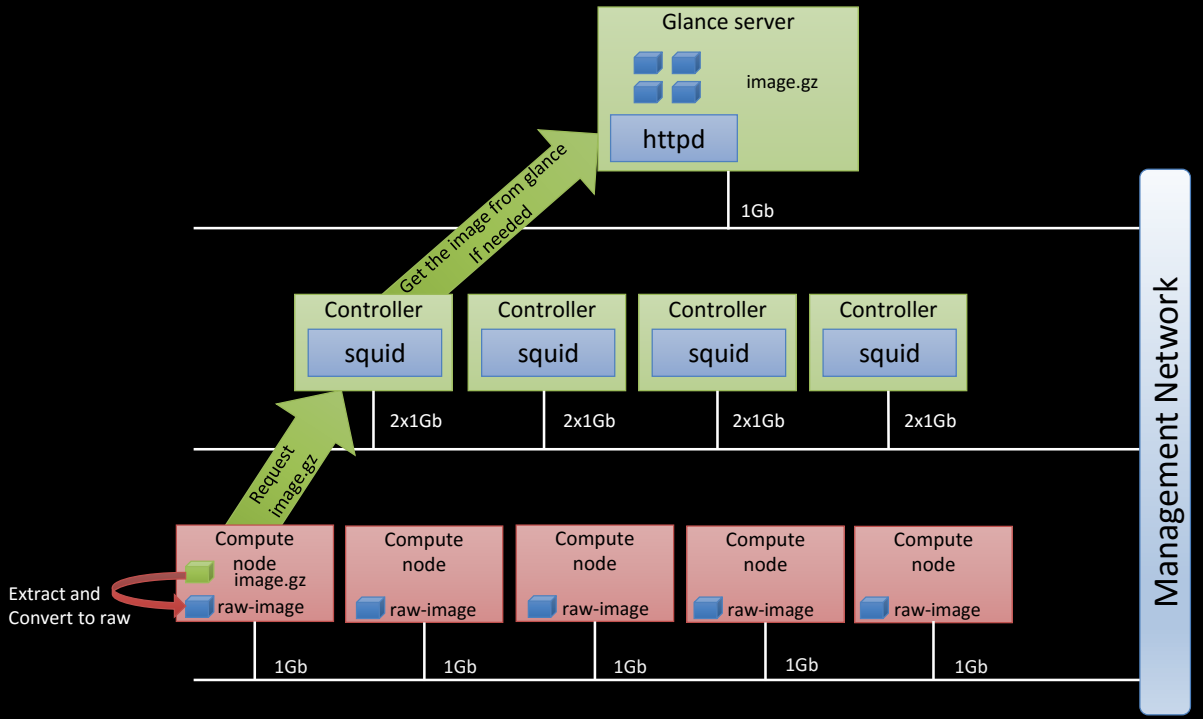
- From 4 hours to 1 hour
- nova-glance and nova-scheduler are the bottleneck

1.3 Speed up virtual machine image distribution

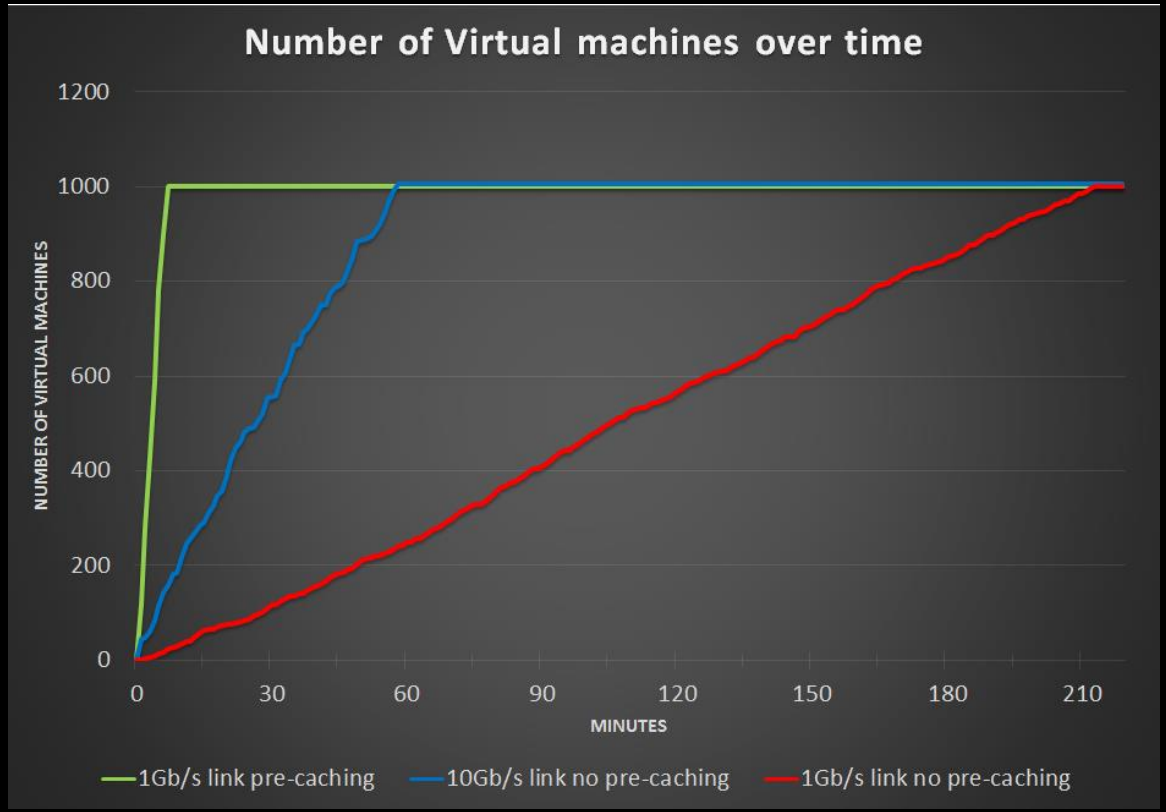
- Push image into nova cache
- Image compression
- Proxies

1.3 Speed up virtual machine image distribution

- Image compression
- Proxies
- Push image into nova cache

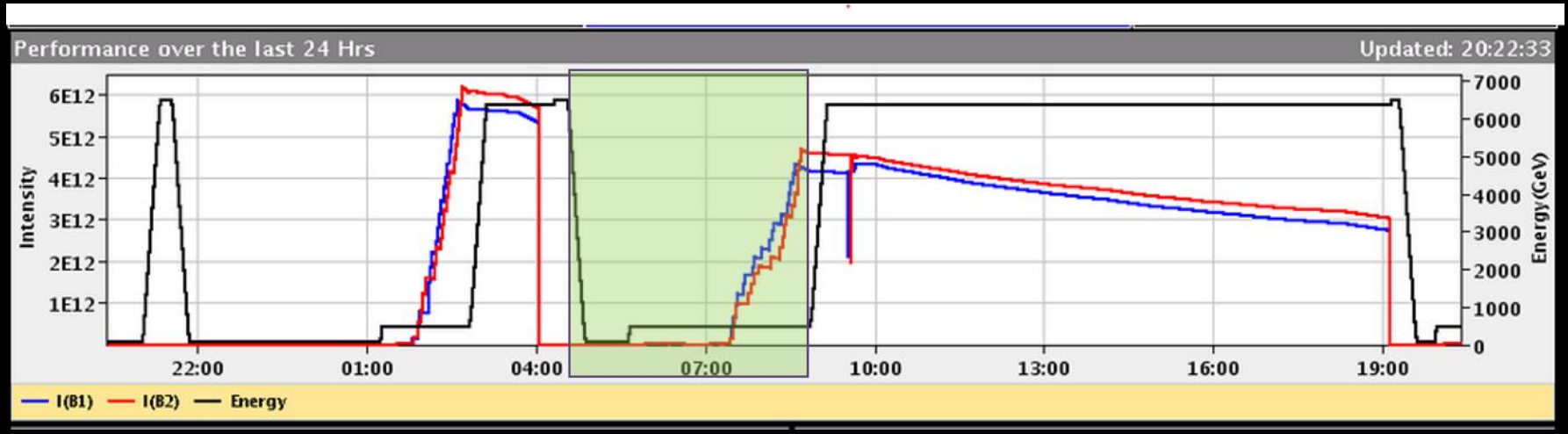


1.3 The HLT OpenStack Cloud performance with pre-caching



- 8 minutes to start 1000 virtual machines

2. Detect the beginning/end of the inter-fill periods



- Detect LHC states (Ramp Down and Ramp)
  - Ramp Down : Start the HLT Cloud
  - Ramp : Stop the HLT Cloud

- LHC states are available via a SOAP interface

3. Select which nodes can be allocated to the Cloud

- During inter-fill periods, a minimal amount of nodes have to be left for test runs
- The HLT configuration is stored in a database
- A script generates a list of machines that can be allocated to the Cloud based on the information stored in the database

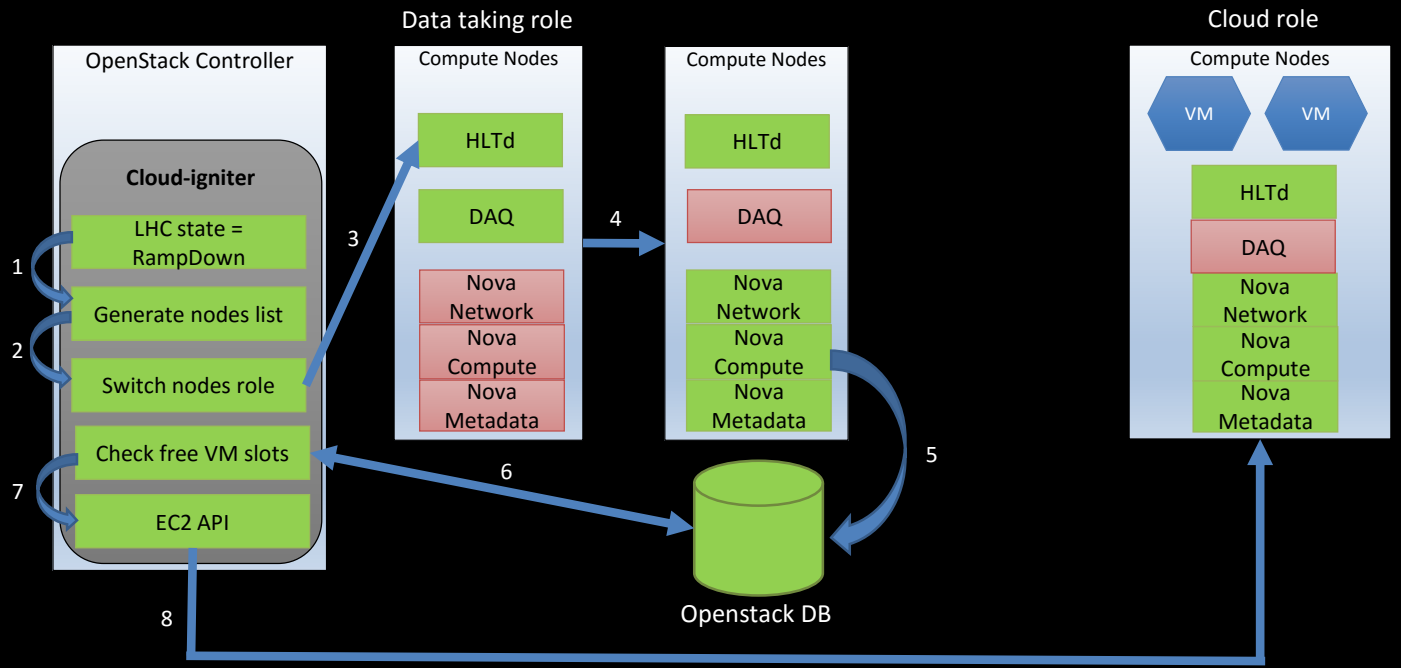
4.1 Start/Stop the OpenStack services automatically (Compute node level)

- A local script manages the OpenStack services
  - Starts them using the SytemV init scripts
  - Stop them using the SystemV init scripts
    and Kill the virtual machines
    and Files cleanup
    and OpenStack database content update
    (OpenStack APIs are too slow and not reliable to stop VMs)

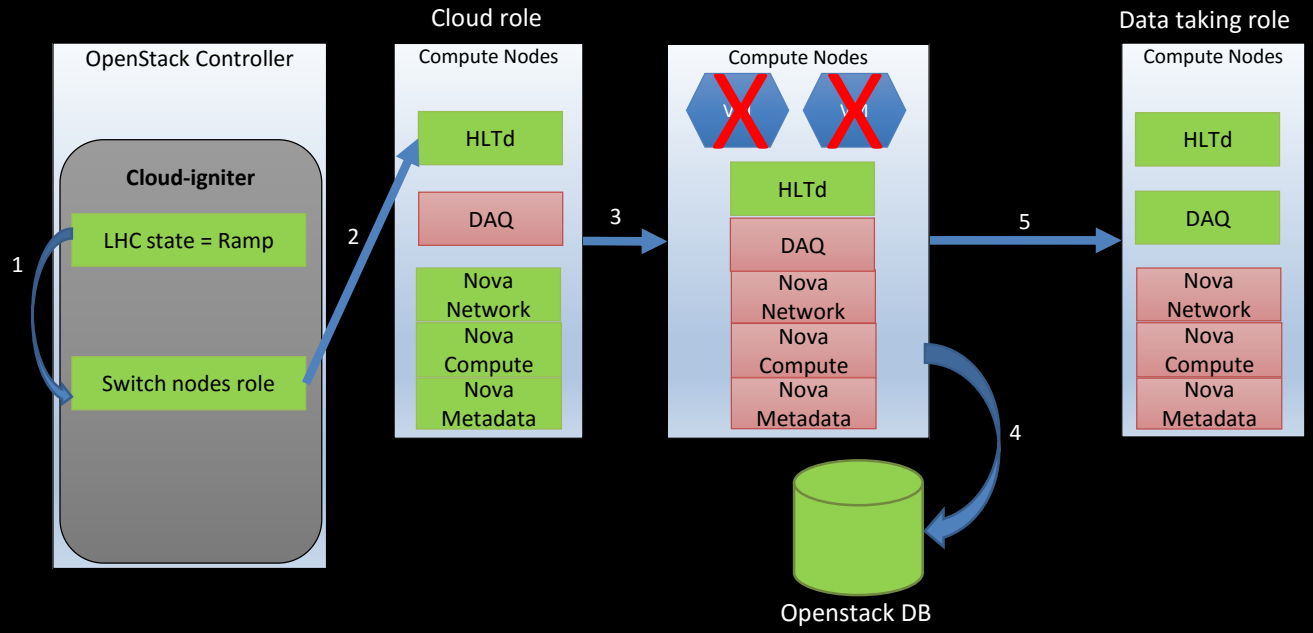
4.2 Start/Stop the OpenStack services automatically (Cluster level)

- A daemon (HLTd) runs on all compute nodes
  - Developed initially to control the DAQ applications for data taking
  - Provides an API
  - New feature added : Call an external script when the DAQ applications are off

Cloud start fully automatized: Cloud-igniter script

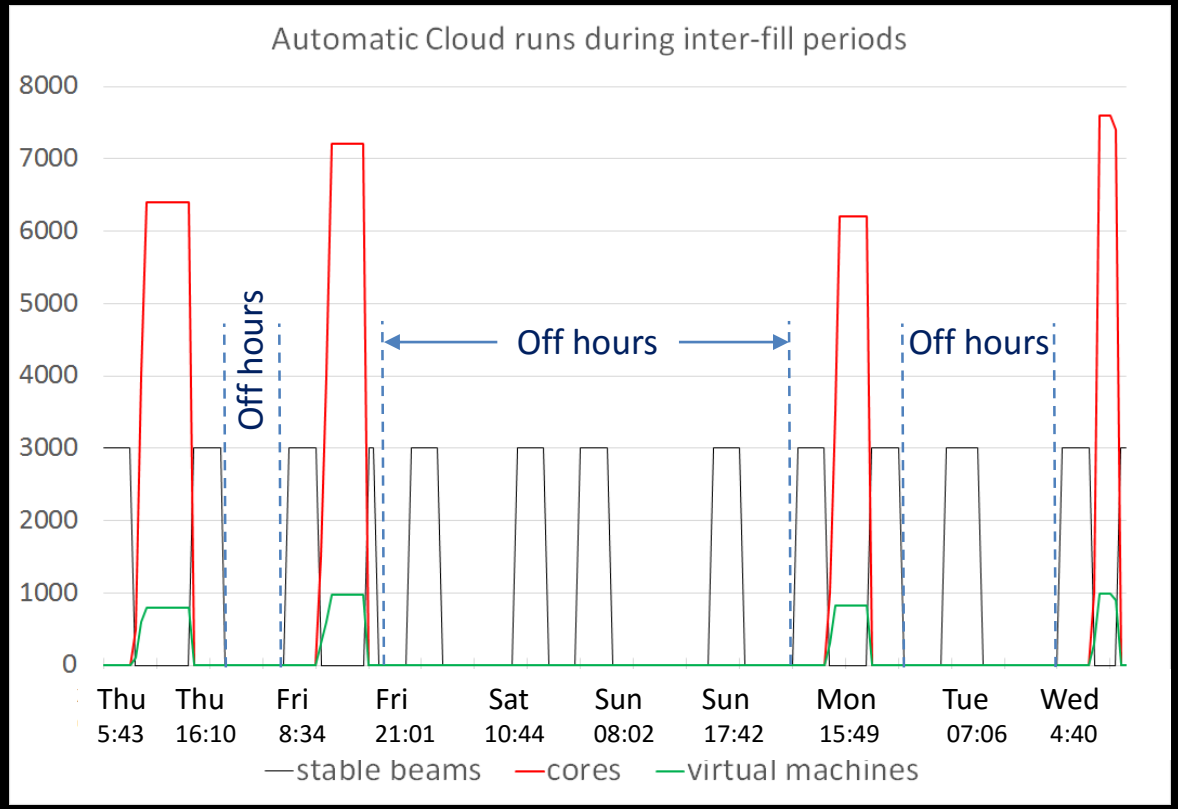


Cloud stop fully automatized: Cloud-igniter script



Proof of concept: Inter-fill Cloud runs

HLT Cloud started/stopped automatically (work-hours)



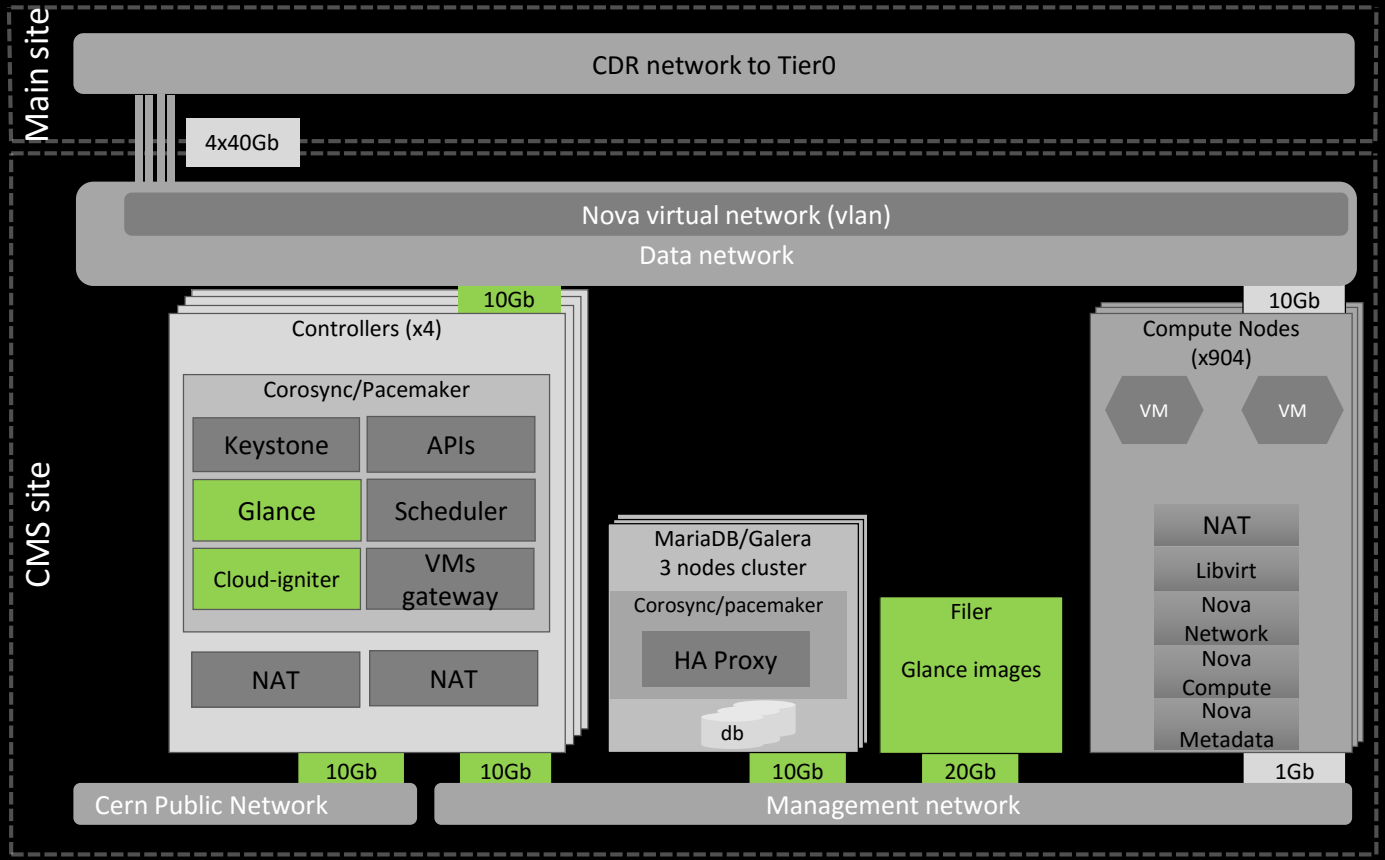
Inter-fill Cloud runs efficiency ?

- Requires shorter jobs during inter-fill periods (**2 hours instead of 8 hours**)

Cloud mode	Number of days allowed to run the HLT Cloud (2015)	Efficiency
Static (maintenance periods)	108 days	**100%**
Inter-fill (ramp down -> ramp)	192 days	**50%** in worst case scenario as some jobs will be killed before they are completed
Total	**300 Days**	

*M. Solfaroli, "LHC operation and efficiency in 2015", LHC Performance Workshop, Chamonix, France, 25-28 January 2016

HLT OpenStack Architecture: Recent upgrades



Conclusion

- Multi-role cluster
  - Data taking
  - Common Tier site (Maintenance periods)
  - Opportunistic Tier site (Inter-fill periods)

- HLT Cluster computing resources usage is optimized

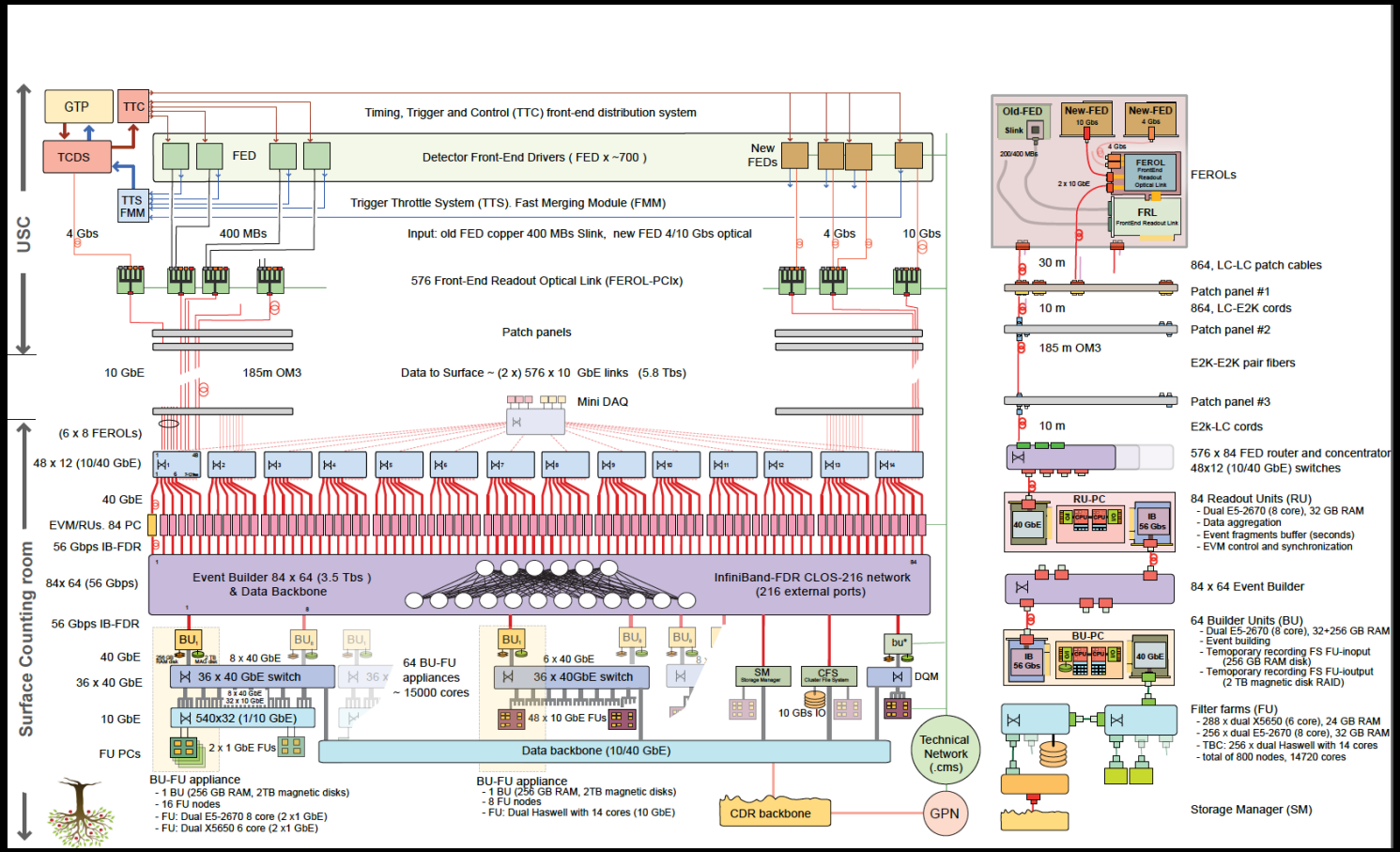
- Ongoing work: Ensure the cloud manager is fully fault tolerant

- Next ?

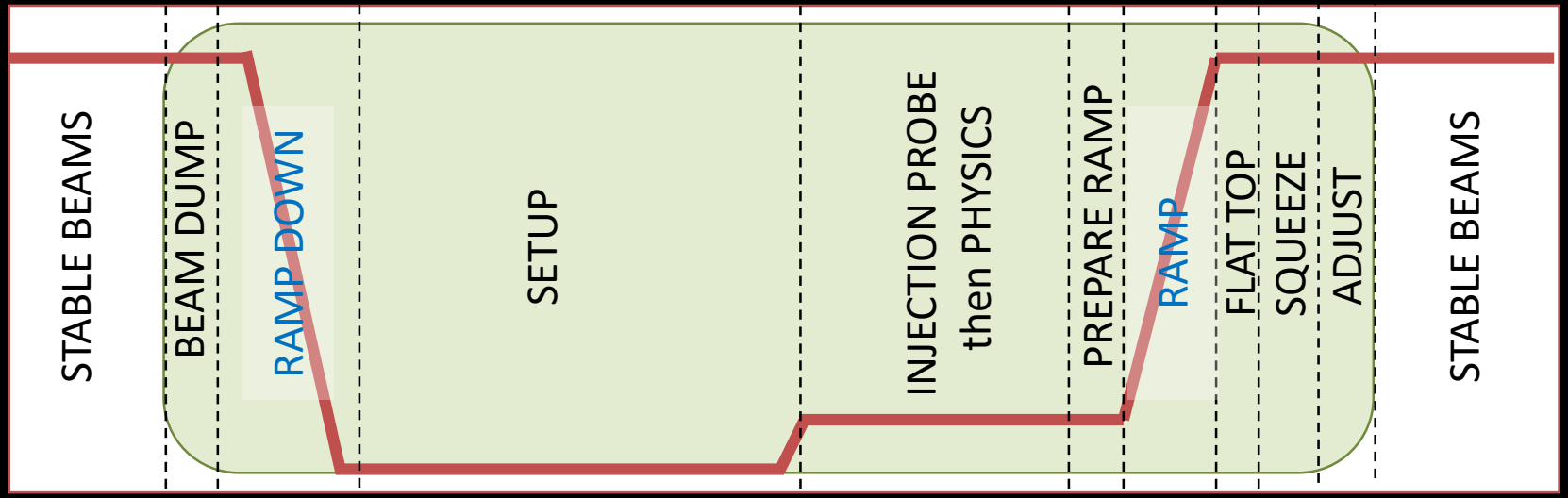
Extend the HLT Cloud periods by starting as soon as the rate of data taken from the detector and the computing power needed decrease.

Thank you.

CMS Online Cluster

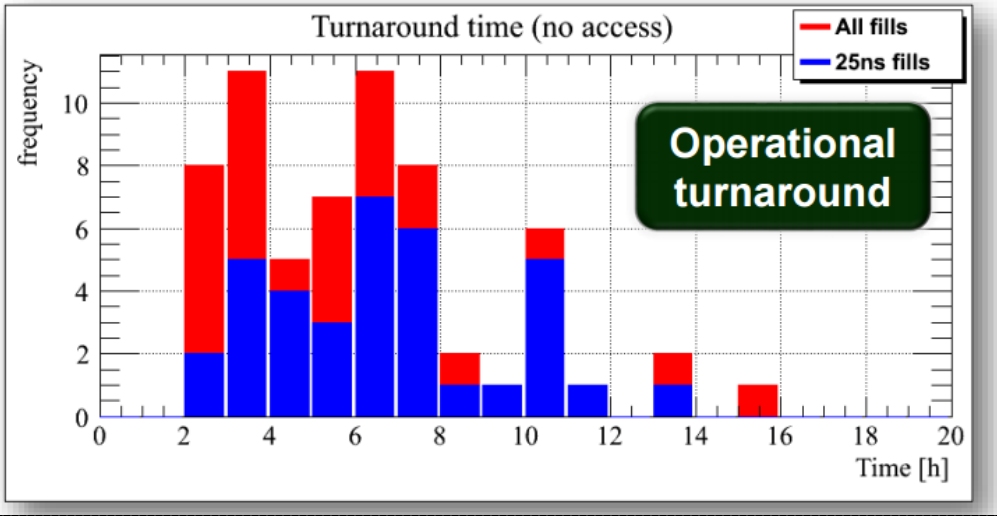


LHC cycles


STABLE BEAMS
BEAM DUMP
RAMP DOWN
SETUP
INJECTION PROBE
then PHYSICS
PREPARE RAMP
RAMP
FLAT TOP
SQUEEZE
ADJUST
STABLE BEAMS

Inter-fill periods length in hours*

	Mean	Min
Total	6.3 h	2.2 h
25 ns	6.8 h	2.8 h



*M. Solfaroli, "LHC operation and efficiency in 2015", LHC Performance Workshop, Chamonix, France, 25-28 January 2016