A scalable infrastructure for CMS data analysis based on OpenStack Cloud and GlusterFS

Datacenter Indirection Infrastructure(DII) for High Energy Physics (HEP)

[1]Salman Toor, [2]Lirim Osmani, [1]Paula Eerola, [1]Oscar Kraemer, [1]Tomas Lindén, [2]Sasu Tarkoma and [3]John White
[1]Helsinki Institute of Physics (HIP), CMS Program, [2]Computer Science Department, University of Helsinki, and [3]Helsinki Institute of Physics (HIP), Technology Program
{Salman.Toor, Paula.Eerola, Carl.Kraemer, Tomas.Linden}@helsinki.fi, {Lirim.Osmani, Sasu.Tarkoma}@cs.helsinki.fi, John.White@cern.ch

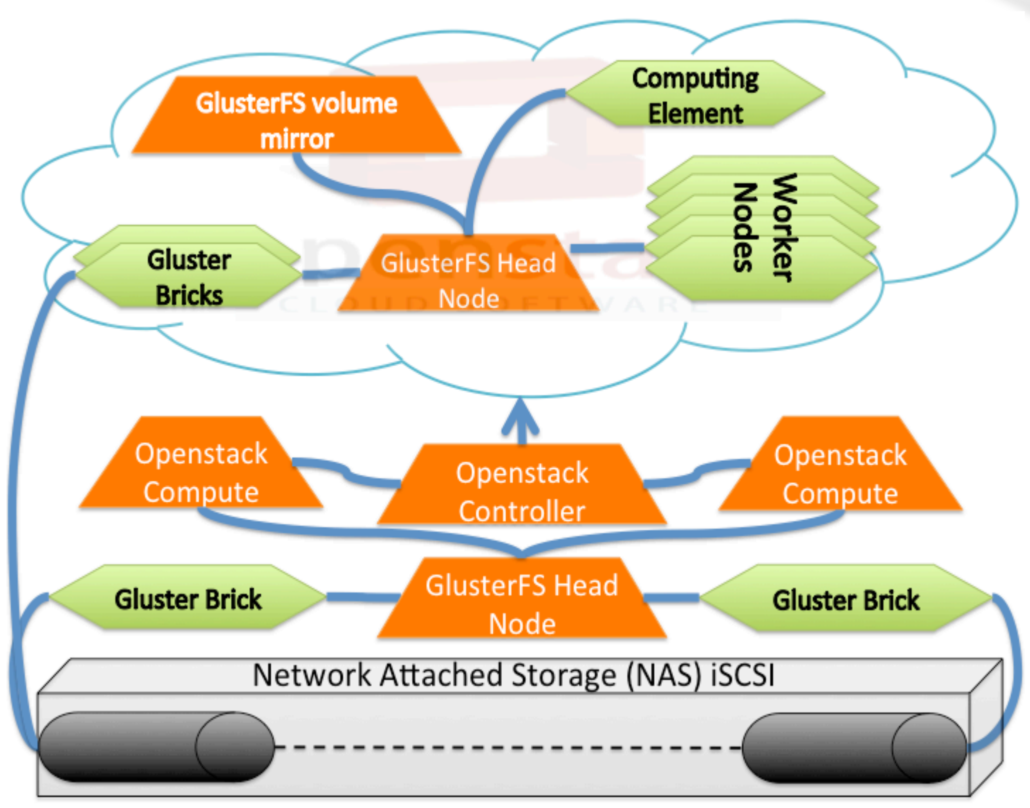
1 Introduction

The aim of this project is to design and implement a scalable and resilient Infrastructure for CERN High Energy Physics (HEP) data analysis. The infrastructure is based on OpenStack components for structuring a private Cloud with Gluster File System. Our test results show that the adopted approach provides a scalable and resilient solution for managing resources.

2 Architecture

- Computational resources
  - Dell PowerEdge, 2 quad core Intel Xeon
  - 32GB (8 x 4GB) RAM
  - 4 x 10GbE Broadcom 57718 network
- Gluster File System servers
  - Dell Power Edge servers
  - 512GB LUN attached to each
  - FCoE (Fibre Channel over Ethernet) protocol

Type	Nodes	Cores	Memory
Virtual Machines (QCOW2 images)			
WN	25	4	14
CE	1	4	14
GlusterFS	2	4	14
Real Machines			
Controller	1	8	32
Compute	19	8	32
GlusterFS	2	8	32



Local and GlusterFS based storage			
	Root local	Ephemeral disk	GlusterFS shared MP
WN	10GB	45GB	900GB
CE	10GB	-	900GB

3 System Components

- OpenStack Cloud (Grizzly release)
- Gluster File System
- Advanced Resource Connector middleware
- CERN Virtual Machine File System

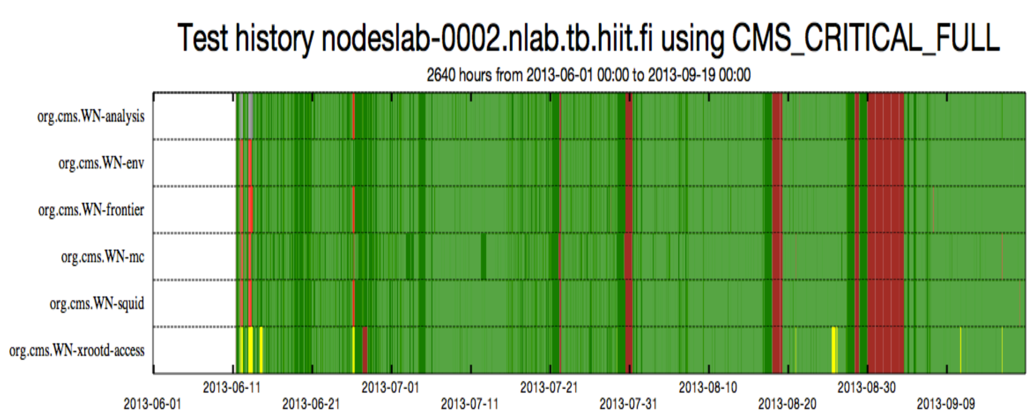
4 GlusterFS IO

- GlusterFS is used
  - inside the Cloud to provide shared area (Brick 1 & 2, 2TB)
  - outside for Glance and Nova repositories (Brick 3 & 4, 1TB)

	Brick-1	Brick-2	Brick-3	Brick-4
Days	20	20	20	20
Total Reads	40GB	41GB	169GB	831GB
Total Writes	38GB	36GB	364GB	1017GB
Average – Maximum Latency (milliseconds) / No. of Calls (in millions)				
Read	0.05-11.6/ 29.8	0.05-10.8/ 31.0	0.29-217.1/ 0.6	0.38-2386/ 0.32
Write	0.08-16.8/ 1.8	0.07-14.0/ 1.9	1.66-5151/ 10.8	1.06-7281/ 7.8

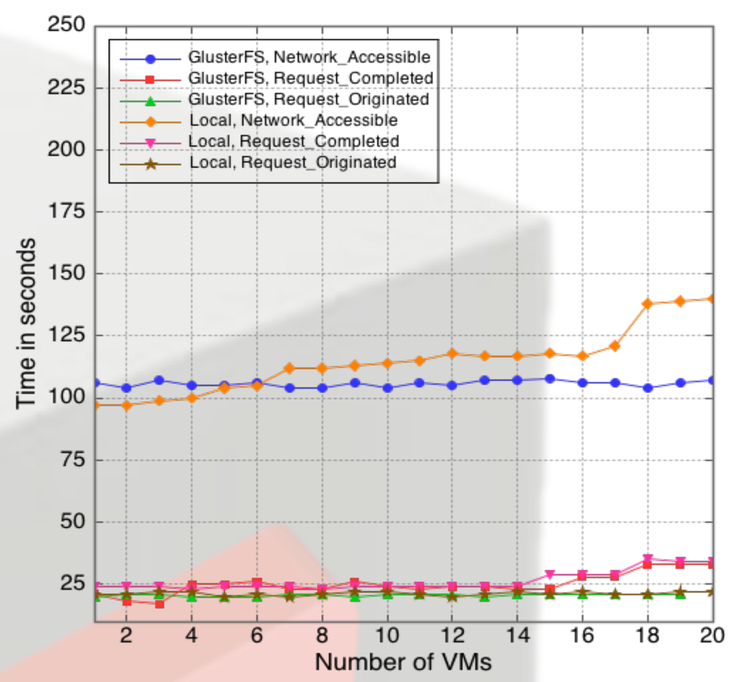
5 System Stability

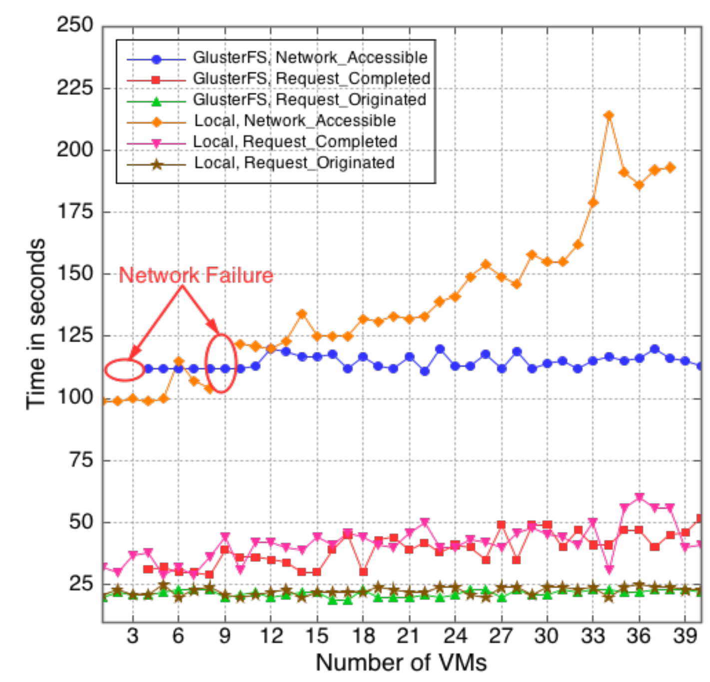
- The evaluation is based on the CMS Dashboard together with CSC and NDGF accounting and monitoring systems
- More than 65k jobs have been processed, including CMS production and analysis jobs
- Example of a specific user
  - Run 400 analysis jobs with 74 walltime days and 85% CPU efficiency



6 Performance Analysis

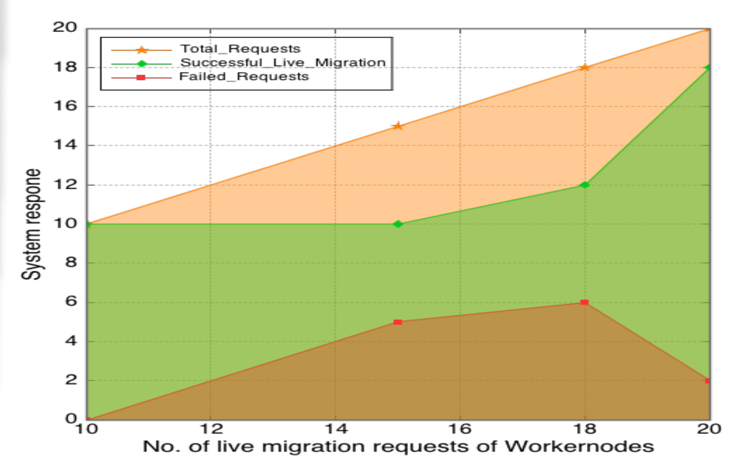
- 4% performance loss evaluated with the HEPSPEC-2006 benchmark
- Burst mode VM boot requests based on Local and GlusterFS based setups
- Uniform boot response with GlusterFS





7 Live Migration

- Experiments with different kinds of instances
  - Minimal VM of Ubuntu m1.small took 6 sec
  - Worker node VM with full configuration took 43 sec
- Experienced random failures in higher number of live migration requests



8 Conclusion

We have demonstrated:

- More flexible system/user management through the virtualized environment;
- Efficient addition/removal of virtual resourses;
- Scalability with an acceptable performance loss;
- A seamless view of our site through ARC middleware.

9 Acknowledgements

- This project is funded by Academy of Finland (AoF)
- Thanks to Ulf Tigerstedt, CSC for help with HEPSPEC tests
- The CMS collaboration for sending production jobs to process