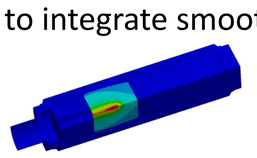
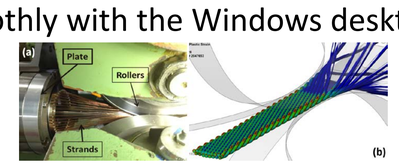
HPC in a HEP lab: lessons learned from setting up cost-effective HPC clusters

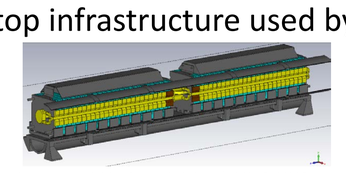
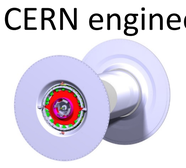
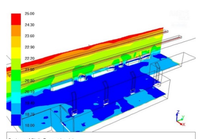
Ioannis AGTZIDIS, Pierre BAEHLER, Tadeusz DUL, John EVANS, Nils HØIMYR, Michal HUSEJKO, Helge MEINHARD
CERN, Switzerland

Introduction

Until recently powerful engineering workstations were sufficient to solve most of the engineering simulation and analysis problems encountered at CERN; however, we saw a growing number of enquiries from users seeking for more computing power, more memory and storage for analysing more complex models or for more detailed analysis. In this paper we searched for solutions which use as much as possible standard CERN Computer Center hardware, CERN IT services with out of the box configured commercial engineering applications used at CERN. In cases where CERN standard hardware/software components were not adequate, we propose economic solutions for enhancing the performance.

In our CHEP 2013 paper [1] we presented our approach building low cost Linux based HPC clusters – in this paper we present our approach building Windows based HPC clusters which have the advantage to integrate smoothly with the Windows desktop infrastructure used by CERN engineers.

[1] “Self-service for software development projects and HPC activities”, HUSEJKO, M. et al. CHEP 2013

Main requirements:

- Smooth integration with Windows workstation environment
- Low Total Cost of Ownership (which imposes alignment with standard CERN CC infrastructure)
- Support for multiple HPC enabled engineering applications and solvers: ANSYS, FLUENT, CST, COMSOL, HFSS, LS-DYNA …

Constraints:

- Limited number of licenses (<1000 cores)
- Limited number of users (<100)
- Limited number of concurrent jobs (<50)
- Soft constraint: Build on top of CERN's standard High Throughput Computing and Computing Center infrastructure

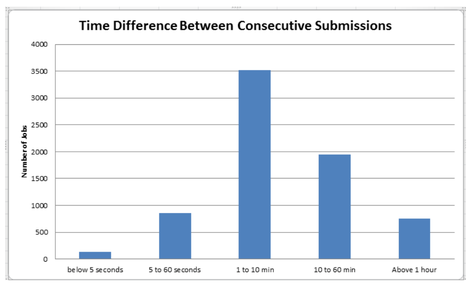
Architecture optimization:

- The standard CERN CC components are:
  - Compute node: Dual Socket, SandyBridge CPU, up to 16 physical cores, 64 GB RAM per node
  - Interconnect: 1Gb Ethernet, no RDMA, no low latency MPI
  - Storage: DFS over 1 Gb Ethernet, no RDMA offload for SMB
  - Monitoring: ElasticSearch/Kibana based system
- To make HPC applications running at optimum speed, some of the above components have to be replaced:
  - Compute node: Quad Socket, 32 physical cores, 512 GB RAM – some applications and solvers can not scale beyond single node, yet they require a big amount of RAM. Moreover we favor in-core simulations (in RAM) than complicated storage systems (costly IO).
  - Interconnect: 10 Gb low latency with support for iWARP (MPI) and RDMA (SMB storage) – some other solvers can profit from multi node distributed computing, and some solvers produce large result files
  - Storage: DFS with 10 Gb Ethernet with RDMA (SMB)
  - Monitoring: Develop bridge between MS HPC monitoring/accounting database and ElasticSearch/Kibana monitoring system.

	Compute Node	RAM per core (node)	Interconnect	DFS Storage interface	Monitoring
CERN CC Standart	16 cores	4 GB (64 GB)	1 Gb Ethernet	1Gb Ethernet	Elasticsearch Kibana
HPC	32 cores	16 GB (512 GB)	10 Gb iWARP/RDMA	10 Gb RDMA (SMB)	Elasticsearch Kibana

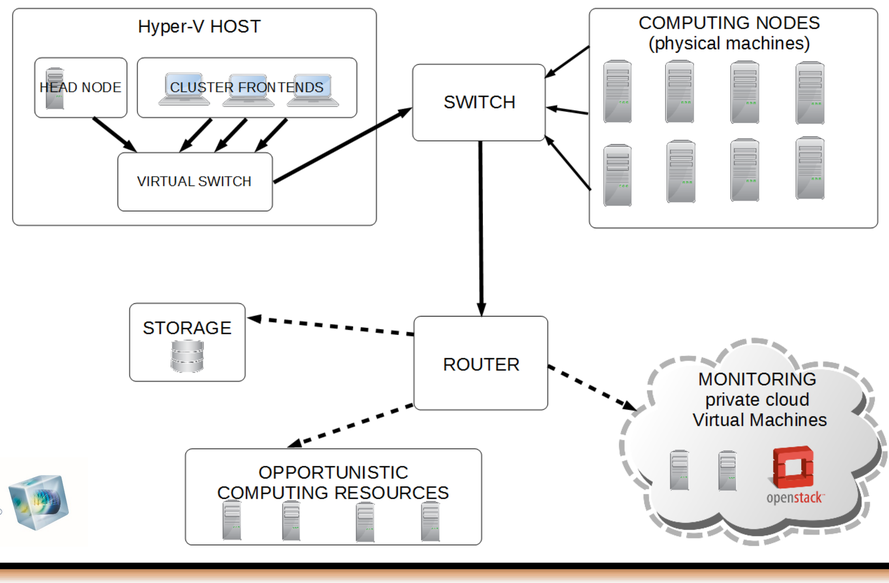
Performance measurements:

- The last year jobs’ histogram presenting time difference between consecutive submission is presented on the diagram below
- With job submission up to 4 jobs/second we see no impact on job scheduling – which is much more than required.

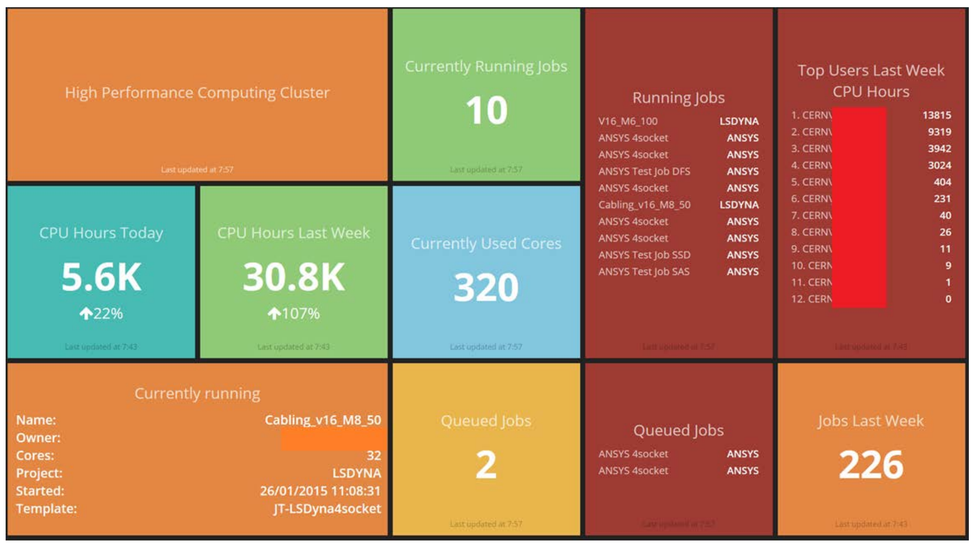


Windows HPC cluster architecture

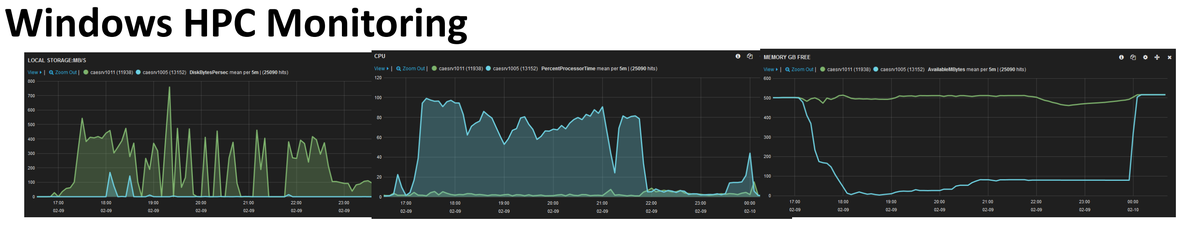
- Windows Server 2012 R2 (bare metal compute nodes, Hyper-V VMs for cluster’s head node and front-end machines)
- Chelsio T520-LL-CR low latency Ethernet NIC for MPI (iWARP), SMB (RDMA) and Hyper-V virtual switch.
- HP 5900 low latency 10Gb Ethernet switch
- Windows HPC Pack 2012 R2
- DFS storage



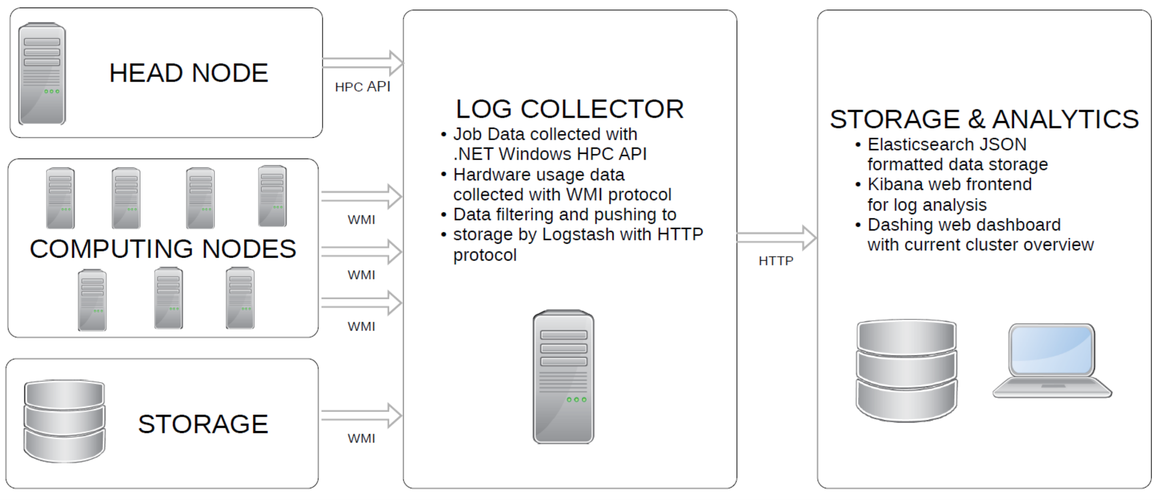
Service overview dashboard



Windows HPC Monitoring



Microsoft logstash elasticsearch. kibana



Conclusions

- 10 Gb low latency Ethernet is a “good enough” interconnect to build low scale HPC clusters
- High appreciation of user community for smooth integration with Windows workstations
- With minor development work the Windows HPC can be smoothly connected to a standard CERN monitoring infrastructure

Future work

- Head node failover (split between bare metal and VM on CERN OpenStack)
- Software defined storage (StorageSpaces) for HPC (available under the DFS)
- License aware scheduling