# CONTROL SYSTEM VIRTUALIZATION FOR THE LHCB ONLINE SYSTEM

E. Bonaccorsi, L. Granado Cardoso, N. Neufeld CERN, Geneva, Switzerland
F. Sborzacchi, INFN/LNF, Frascati, Roma, Italy

*Abstract*

Virtualization provides many benefits such as more efficiency in resource utilization, less power consumption, better management by centralized control and higher availability. It can also save time for IT projects by eliminating dedicated hardware procurement and providing standard software configurations. In view of this virtualization is very attractive for mission-critical projects like the Experiment Control-System (ECS) of the large LHCb experiment at CERN. This paper describes our implementation of the control system infrastructure on a general purpose server-hardware based on Linux and the RHEV enterprise clustering platform. The paper describes the methods used, our experiences and the knowledge acquired in evaluating the performance of the setup using test systems, constraints and limitations we encountered. We compare these with parameters measured under typical load conditions in a real production system. We also present the specific measures taken to guarantee optimal performance for the SCADA system (WinCC OA), which is the backbone of our control system.

## INTRODUCTION

LHCb is a dedicated heavy-flavour physics experiment designed to perform precise measurements of CP violation as well as rare decays of B hadrons in the Large Hadron Collider (LHC) [1]. The experiment is located at point 8 of the LHC particle accelerator.

The LHCb online system has been designed to run completely isolated and independent, as an autonomous system. It consists of ~2000 physical servers and embedded systems interconnected through three main high density routers and ~100 distribution switches. The only connection to CERN networks and the Internet is through the boundary network.

Hosts in the system are grouped as Experiment Control System (ECS) [2] hosts, Data Acquisition (DAQ) hosts and general infrastructure hosts. LHCb's ECS is in charge of the configuration, control and monitoring of all the components of the online system. This includes all devices in the following areas: data acquisition, detector control, trigger, timing and the interaction with the outside world.

The servers in the ECS network are common data centre infrastructure servers (DNS, DHCP, *etc.*) and control PCs that run the standard LHC SCADA system, PVSS/WinCC, on top of Linux or Windows.

It has proven, at least in our experience, advantageous to distribute different control application over different computer minimizing the number of different functions of a single server, typically assigning one SCADA project per control PC.

Virtualization solves the problem of the increasing number of single role machines, reassigning hardware resources on demand. Because of redundancy, operation, consolidation and economical reasons many control PCs of the LHCb experiment are already migrated to the virtual infrastructure.

## CURRENT VIRTUALIZATION INFRASTRUCTURE

### *Servers*

The first implementation deployed on a single blade chassis consisting of ten blade servers (Dell Poweredge M610) based on Intel Xeon E5530 has been extended to twenty servers distributed across two independent blade chassis. The specification of the memory and the I/O cards of the servers are summarized in Table 1

Table 1: Server Hardware Specifications

| CPU | Memory | I/O cards |
|---|---|---|
| x E5530 @ 2.4GHz (8 real cores + Hyper Threading) | 3 x 16GB = 48GB RAM | 2 x 10Gb network interfaces ( for VLAN sharing, 1 linked to LHCb)<br>2 X 8Gb Fiber channel switches (linked to two isolated fabrics) |

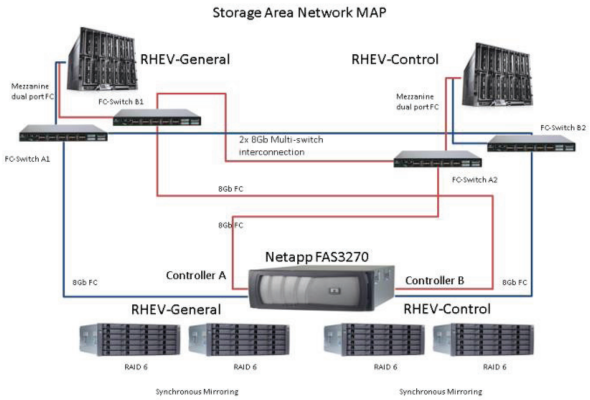
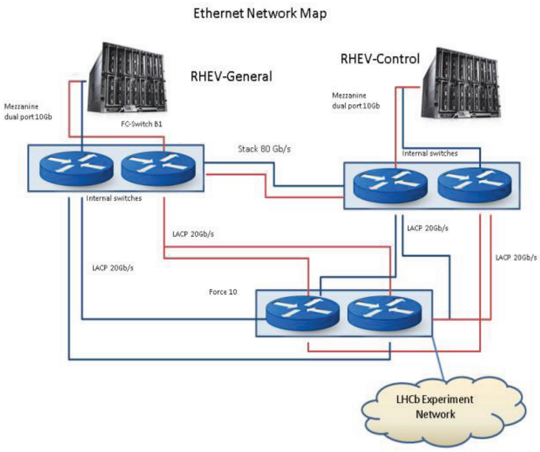


Figure 1. Storage Area Network map.

ISBN 978-3-95450-139-7

### Ethernet Network and Storage Area Network

The network and storage infrastructures based on Dell and Broadcom have been upgraded. Using additional stacking modules an active/active fault tolerant configuration has been put into production. Every server is now connected via Link Aggregation Control Protocol (LACP) and Fiber Channel interfaces to two fiber channel fabrics. Storage Area Network and Ethernet network are illustrated respectively in Figure 1 and Figure 2



.

Figure 2. Ethernet network map.

### Clusters

Two independent clusters based on Red Hat Enterprise Virtualization (RHEV), which is a management software on top of the open source Kernel Virtual Machine (KVM) hypervisor provided by Red Hat [3], have been configured on top of the hardware infrastructure:

- The General Purpose (GP) cluster is mainly used for web, development and infrastructure servers
- The ECS cluster is used exclusively for mission critical control PCs needed by the experiment control system.

The separation in two independent clusters avoids any interference between the mission critical ECS virtual machines and GP virtual machines that are less critical for data taking.

Each cluster uses a dedicated storage controller and an isolated storage space. This setup allows the two clusters to not interfere with each other in terms of storage load.

### Shared Storage and Test Environment

One of the main challenges of the control system virtualization is the shared storage. Its reliability and efficiency are crucial aspects.

Instead of basing our decision to abstract row numbers provided by storage vendors, after an extensive test of the previous existing storage – which turned out to be inadequate for virtualization because of its internal architecture – we took a lot of effort to prepare a test environment able to emulate the storage I/O load as the one produced by a fully virtualized control system.

The test environment uses multiple servers connected via Fiber Channel (FC) to the storage test unit. The test environment consists of running 300 VMs at the same time, each one executing a properly configured I/O program-tool called IORATE [4]. The block unit selected for the input/output operations are 2 KB and 4 KB random, tested under six different circumstances:

- Only reading – Using a small part of the virtual drives in order to always stay inside the internal cache of the storage unit candidate
- Only reading – Using a big part of the virtual drives in order to understand the behaviour of the candidate storage system with its local cache invalidated by the use of a bigger dataset than the size of the internal cache
- Only writing – Using a small part of the drives
- Only writing – Using a large part of the virtual drives
- Writing and reading at the same time using a small part of the drives
- Writing and reading at the same time using a big part of virtual drive

All the operations on disk were Random I/O using two working threads.

An inadequate shared storage system that does not satisfy a minimum of ~40 random IOPS per VM will break, rendering the entire infrastructure operationally not stable. This will be exposed by our test.

This environment has been used to test several shared storages and it brought us to the choice of the NetApp 3270 with hybrid shelves (SSD + SATA)[5].

An additional test scenario for the long-term reliability has been prepared. Using the production SCADA system, with an artificial highly pessimistic workload, we obtain good results by running for several weeks over 150 VMs.

## BENCHMARKS

The most common storage performance characteristics are sequential and random write operations per second. In a virtualized scenario both operations will likely access locations on the storage device in a non-contiguous manner because the placement of the virtual block of data is stored according to the position in which was previously dynamically created during the deployment of the VMs.

In these conditions the use of a smart displacement of the blocks of data can help in improving performance. The storage system itself can improve response-time during intense disk operations: the mixing of SSD drives with HDD on the same storage pool is becoming a common solution across many vendors.

Several units have been tested of which three will be described in this paper.

The first unit tested is a NetApp 3240, connected via FC on a storage pool of 18 SATA 2 x 1 TB 7.2 k RPM and 6 SSD 100 GB drives, both configured in RAID 6. The unit came already configured for the aggregation of hybrid data displacement over the entire storage pool.

ISBN 978-3-95450-139-7

The unit from vendor 2 uses the iSCSI protocol for data export and a storage pool of 48 SAS 300 GB 10k RPM disks, 2 SSD 100 GB and 1 SSD 8 GB.

The unit has been configured by the vendor with 6 Volumes of 8 disks in RAID 6 and exported via iSCSI as a single striping volume.

The unit from vendor 3 uses 10 x SSD drives of 140 GB each. Benchmark results are summarized in Table 2.

Table 2: Storage Benchmarks

| NetApp 3240 | Vendor 2 | Vendor 3 |
|---|---|---|
| 4K reading: 307 IOPS | 4K reading: 450 IOPS | 4K reading:<br>26 IOPS |
| 4K writing:<br>153 IOPS in cache and 150 out of cache | 4K writing: 10 IOPS in cache and 10 out of cache | 4k Writing:<br>13 IOPS |
| Mixed 4K writing + 4K reading:<br>212 IOPS in cache and 30 outside cache | Mixed 4K writing + 4K reading:<br>38 IOPS in cache and 10 IOPS outside cache | Mixed 4K writing + 4K reading:<br>20 IOPS |

## OPTIMIZING RESOURCES

High performance commercial storage systems are expensive; this is why the virtualization infrastructure has been tuned for an optimal usage of resources exploiting the maximum application features of all network, hypervisor and storage layers.

Thin provisioning is an excellent way to reduce disk space from the hypervisor layer but we measured degradation of performances, in particular because of the additional overhead in terms of CPUs on the servers.

Commercial storage offers a way to achieve about the same disk space saving: deduplication [6], a technique that eliminates duplicate copies of repeating data. The virtualization infrastructure is profiting from this technique and it is currently saving about 67% of storage space as shown in Figure 3.

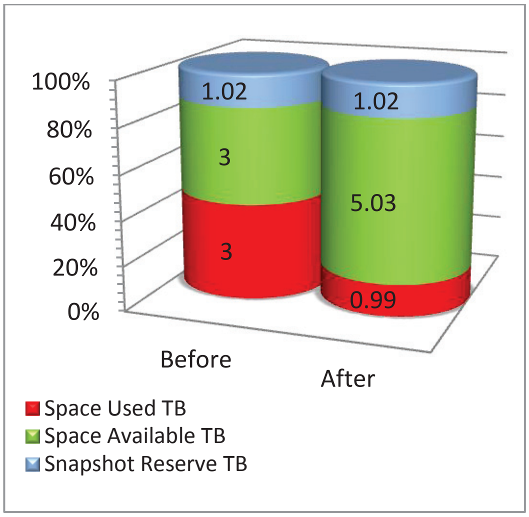


Figure 3. Storage efficiency deduplication saving 67%.

Kernel Shared Memory allows the two independent clusters to maximize the usage of memory, merging the same pages allowing overcommitting in terms of memory.

## CONCLUSIONS

In our initial studies we have identified the shared storage as the key element for consistent and reliable performance. In order to select a suitable system or our experiment we have developed an emulator, which realistically models the I/O patterns created by the LHCb SCADA systems. We have found this emulator indispensable in the evaluation of the many possible commercial storage systems. Based on these results we have found the best price performance in the NetApp.

Further we show in this paper how we have attempted to optimize the resource usage of the hypervisors themselves, notably in terms of disk-space, I/O and RAM usage. We have shown deduplication provides the best IOPS performance out of the several possible methods for saving disk-space.

Together with management scripts developed by us we have started the migration of the LHCb SCADA systems in July 2013 and it is completed to about 40% now (September 2013). The migration will be finished by the end of the year. No difficulties have been encountered and the robustness of the system has already been verified on several occasions.

## REFERENCES

[1] LHCb Trigger System TDR, LHCb TDR 10, CERN/LHCC/2003-31, 2003.

[2] C. Gaspar et al., "An Integrated Experiment Control System, Architecture, and Benefits: The LHCb Approach," IEEE Transactions on Nuclear Science Vol. 51, No. 3 (June 2004), pp. 513-520; DOI:10.1109/TNS.2004.828878.

[3] http://www.redhat.com/products/cloud-computing/virtualization/

[4] http://www.iorate.org

[5] http://www.netapp.com/

[6] http://www.netapp.com/us/products/platform-os/dedupe.aspx