FROM REAL TO VIRTUAL - HOW TO PROVIDE A HIGH-AVALIBLITY COMPUTER SERVER INFRASTRUCTURE

R. Kapeller, R. Krempaska, C. Higgs, H. Lutz
Paul Scherrer Institute, 5232 Villigen PSI, Switzerland

Abstract

During the commissioning phase of the Swiss Light Source (SLS) at the Paul Scherrer Institute (PSI) we decided in 2000 for a strategy to separate individual services for the control system. The reason was to prevent interruptions due to network congestion, mis-directed control, and other causes between different service contexts. This concept proved to be reliable over the years. Today, each accelerator facility and beamline of PSI resides on a separated subnet and uses its dedicated set of service computers. As the number of beamlines and accelerators grew, the variety of services and their quantity rapidly increased. Fortunately, about the time when the SLS announced its first beam, VMware introduced its VMware Virtual Platform for Intel IA_32 architecture. This was a great opportunity for us to start with the virtualization of the controls services. Currently, we have about 200 such systems. In this presentation we discuss the way how we achieved the high-level-virtualization controls infrastructure, as well as how we will proceed in the future.

ASSIGNMENT AND ISOLATION OF SERVICES

The core of the control system for our large scale research facilities and related test sites, consists of a large number of Input Output Controllers (IOCs) running VxWorks operating system on VME hardware using the EPICS (Experimental Physics and Industrial Control System) [1] software. Additionally, there is an increasing number of Linux based hosts, which provide a wide range of different services. The services listed in Fig.1 are all tightly linked to the control system core and high reliability is indispensable.

Even though today's computer performance would allow consolidation of multiple services on a single host, we gave preference to dedication and isolation of services for the following reasons:

Easier management

The semi-automatic installation and configuration of hosts has been adapted to its primary functionality (service).

Proportioned security

Often different services desire different levels of security. For example, specific hosts must prevent remote user access, such as 'ssh' or 'telnet', whereas other hosts allow user logins for service configuration purposes.

No negative service interaction

Running multiple service processes on a single host may also become a resource (CPU, RAM, IO, ..) problem in situations where processes getting out of control (Runaway Processes).

Reboot without fear

With the complexity of an accelerator control system and a 24x7 day operation schedule, experts are not always on site. The single-service/single-host concept allows operators to reboot a system without the fear of interrupting other areas.

Service	Total
SSH Gateway	23
Channel Access GW	40
Softioc	37
Port Server Host	23
IOC Boot PC	32
WWW, Elog, Test, ..	35
Total	190

Figure 1: Virtual Machines (VMs) listed by services as of Q3 2013.

FROM REAL TO VIRTUAL

Of the 500 Linux computers in the control system, almost 200 are virtual machines running on two VMware vSphere clusters. The remaining systems, such as User Consoles, Camera Servers, NFS File Servers, EPICS Channel Archivers or central Login Servers, are still running on dedicated hardware.

Dedicated hardware, but on a smaller scale, was also used a decade ago, when the Swiss Synchrotron Light Source was brought towards stable user operation. Small footprint Linux PCs were used to setup central services. This space consuming solution was no longer practical, as the number of beamlines started to rise on a monthly basis. Fortunately mid 2006, VMware released its VMware Server 1.0 [2] at no cost. After a short but successful test installation on a legacy Rack Server, we decided to fully convert to computer virtualization. Being short on rack space, we wanted to squeeze as many virtual systems in as little space as possible. The relatively new C7000 Blade Enclosure from HP [3] was the answer to our problem. Figure 2 shows blades, racked inside a C7000 blade enclosure, which supplies them with power, cooling and networking.

The HP Proliant BL465, shown in Fig. 2, is equipped with a single Dual Core AMD processor, 2 GB RAM and a local Raid1 disk is sufficient for 5 VMs running on VMware Server. However, VMware Server 1.0 still needed an underlying host OS (Scientific Linux [4] in our case), which had to be installed and maintained. This was definitely a big step in the right direction, but stability,

Copyright © 2014 CC-BY-3.0 and by the respective authors

ISBN 978-3-95450-139-7

manageability and efficiency of this blade solution still offered room for improvement.

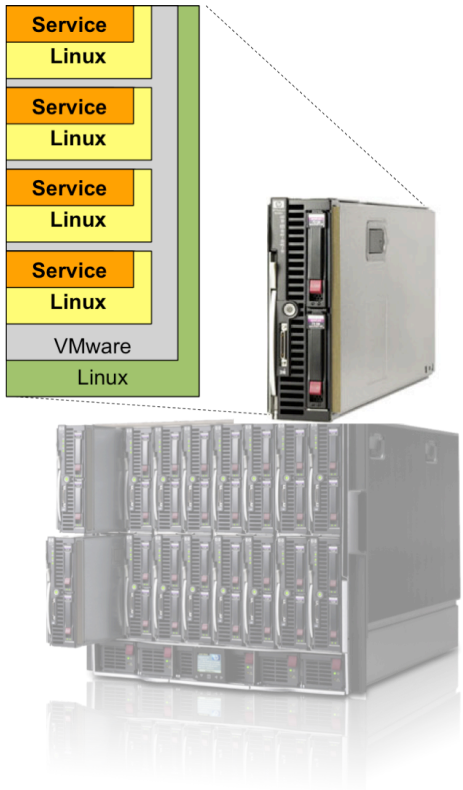

Figure 2: VMware still required a hosting OS (eg.Linux) on each HP Each Blade Server.

After almost 3 years of operation, the next virtualization step was foreseeable. With VMwares ESXi [5], a bare metal embedded hypervisor became available. ESXi directly ran on the host server hardware and no longer required an additional underlying operating system.

Based on a decent Intel server (Dell PowerEdge 2640) and a directly attached FC disk array (both decommissioned hardware from another project), the first ESXi server was installed late 2008 and VMs were gradually migrated. Soon after a second ESXi was installed on similar hardware.

Even though the new hypervisor based ESXi proved to be a highly stable virtualisation host, we finally decided to invest in a state of the art Storage Area Network (SAN) to replace the directly attached FC disk arrays. Fig.3 shows one of the two clusters, each consisting of 3 HP Intel rack servers (Proliant DL380 G6), a dual controller Sun Disk Array (STK6140) and two Qlogig Fibre Channel (FC) switches. All FC connections are with full redundancy.

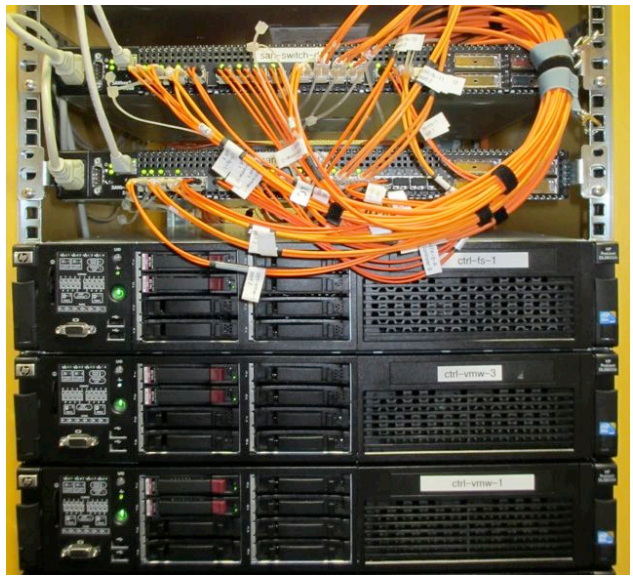

Figure 3: Cluster with 3 HP Intel servers and full redundant FC connectivity across two Qlogic switches.

GOING ENTERPRISE

Until then all of the VMware software was still for free. However, having built a state of the art ESXi hardware Cluster, we started looking for a number of features, only available with VMware's enterprise licensed products.

In particular we were interested in the following features:

- **vMotion -** Eliminates application downtime due to planned server maintenance by migrating live virtual machines between hosts.
- **DRS -** The Distributed Resource Scheduler aligns compute resources with business priorities through automatic load balancing across hosts. Optimizes power consumption by turning off hosts during lower load periods
- **Replication -** Eliminates third-party replication costs through built-in vSphere replication.
- **HA -** The High Availability guarantees, that during a node failure, all affected VMs immediately start on an other cluster node.
- **Update Manager -** Reduces time spent on routine remediation by automating the tracking, patching and updating of vSphere hosts, applications and operating systems.
- **vCenter Server -** Provides centralized management for the vSphere platform.

Fortunately our central computing department had already built a full VMware enterprise infrastructure, including a vCenter Management Server [6]. This was the chance for us to participate, rather than duplicating anything similar. After purchasing the necessary licenses at the end of 2011, the integration of our clusters into vCenter went surprisingly smoothly. Today we use personalized logins with exact defined roles within vCenter. Thus we can

ISBN 978-3-95450-139-7

Copyright © 2014 CC-BY-3.0 and by the respective authors

delegate management tasks for individuals to specific VMs or/and hosts.

The concept of server virtualization within the IT infrastructure for our accelerator- and beamline control, has so far been a very successful story. Operation has proved to be very stable since the beginning of 2010, with no major outage since. One incident in Q3 2012, where the mainboard of a cluster node completely failed, was unrecognized by operation, as all affected VMs were immediately started on the remaining nodes.

STORAGE VIRTUALISATION.

Current plans are to replace the existing cluster in Q1 2015, mainly because hardware is becoming outdated and overloaded.

Most likely VMware ESXi will remain the product of choice, whereas the current FC-based SAN is going to be replaced by a NFS-NAS solution, as shown in Fig. 5. A possible candidate for a future NAS solution may be a filer from NetApp [7]. A Netapp FAS2240-2 has been successfully in operation in another project for Controls-IT since Q1 2013.

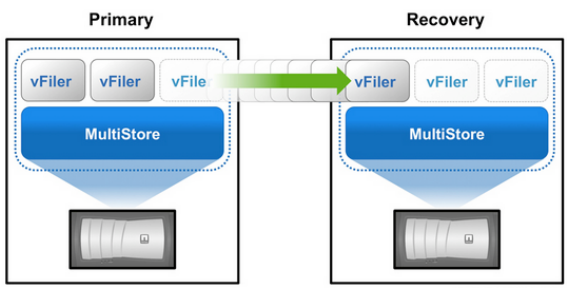

Figure 4: MultiStore enables you to migrate or back up data from one storage system to another without extensive reconfiguration on the destination storage system.

Multistore which is better known as vFilers in the world of NetApp, opens the door to virtual storage servers. An easy explanation of a vFiler is like a virtual NetApp within a physical NetApp. The vFiler "acts" and "feels" just like an actual filer but has less functionality. It owns one or more volumes (disk space), has its own IP configuration (address, netmask, gateway, DNS, ..) and a network file service like NFS, or CIFS. In a dual controller, high-available filer like the FAS2240, a vFiler belongs either to controller 'A', or controller 'B'. In case of a controller loss, all affected vFilers including their resources will automatically move to the other controller. If one ever runs into a disk capacity, or CPU load shortage, a vFiler and its resources can be migrated to a different NetApp, without an extensive setup process, as shown in Fig. 4.

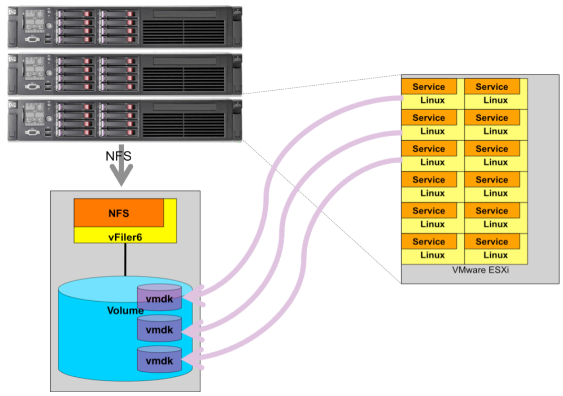

Figure 5: Using a virtual NFS server to store virtual disk images for virtual machines.

FAST REPRODUCTION IN FAVOR OF A COSTLY BACKUP SOLUTION

Using Red Hat Enterprise Linux (RHEL) kickstart [8], in combination with Puppet [9], a central configuration management tool and a slim-host class scheme, any production VM must be re-installable within 1-2 hour(s). Following this concept, we have not scheduled regular backups for our VMs.

However, once or twice a year, or on request, we use a script named 'ghettoVCB.sh' [10] to write a copy of each VM to a dedicated NFS Server. The author, William Lam, made the script public to the community years ago and it has been continuously updated to work with the latest ESXi versions.

The script takes snapshots of a live running VM, backs up the master VMDK and then upon completion, deletes the snapshot until the next backup.

The backup script can save multiple generations and provides auto retention. In case of an emergency, any previously saved VM can be started right off the NFS store.

A backup solution at no cost, that requires SSH access to the ESXi servers and some basic Linux skills!

ACKNOWLEDGMENTS

We want to acknowledge our colleagues P. Hüsser and H. Billich from the PSI AIT group, for their contribution in commissioning the SAN infrastructure and integrating our ESXi hosts into the VMware vCenter.

REFERENCES

[1] http://www.aps.anl.gov/epics/
[2] http://www.vmware.com
[3] http://www.hp.com
[4] http://www.scientificlinux.org
[5] http://en.wikipedia.org/wiki/VMware_ESX
[6] http://www.vmware.com/products/vcenter-server/
[7] http://www.netapp.com
[8] http://en.wikipedia.org/wiki/Kickstart_(Linux)
[9] http://www.puppetlabs.com
[10] https://communities.vmware.com/docs/DOC-876

Copyright © 2014 CC-BY-3.0 and by the respective authors

ISBN 978-3-95450-139-7