VIRTUALIZATION INFRASTRUCTURE WITHIN THE CONTROLS ENVIRONMENT OF THE LIGHT SOURCES AT HZB

D.B. Engel, R. Müller, P. Laux, P. Stange, R. Fleischhauer, HZB, Berlin, Germany

Abstract

The advantages of virtualization techniques and infrastructures with respect to configuration management, high availability and resource management have become obvious also for controls applications. Today a choice of powerful products are easy-to-use and support desirable functionality, performance, usability and maintainability at very matured levels. This paper presents the architecture of the virtual infrastructure and its relations to the hardware based counterpart as it has emerged for BESSY II and MLS controls within the past decade. Successful experiences as well as abandoned attempts and caveats on some intricate troubles are summarized.

GENERAL VIRTUALISATION FEATURES

Pro

- Less general hardware infrastructure
- Less amount of connection cables
- Less cooling and space requirement in the serverroom
- Less hardware maintenance, hardware service contracts
- Good consolidation of services running on old hardware and non portable software
- Very low maintenance time on infrastructure and host Upgrades/Changes
- Central backup and disaster recovery of VMs
- High availability without configuration in the VMs
- Possibility of taking snapshots of a VM
- Clone VMs to suppress update downtime.
- Clone VMs for quick backups

Contra

- Virtual Machines are abstract for many users and administrators
- Problems in the VMs itself are often blamed on the virtualisation hosts
- One defect host potentially crashes many services
- Integration of some specific features of new hardware (like blademanagement) requires upgrade of the whole virtualisation environment.
- More management resources needed at each infrastructure update
- After a power loss waiting time to start the whole infrastructure accumulates (network, storage, blademanagement etc.) before the virtual environment can be restarted.

INSTALLATION AND CONFIGURATIONS OF THE VM INFRASTRUCTURE:

Currently we use two flavours of servers for virtualisation:

Rackclusters

- 6x HP DL380 G5, 64GB RAM, 2x Intel® Xeon® CPU X5160 @ 3.00GHz (4 Cores), 2 fibre channel cards 2x 1GB/s management, 2x 1GB/s VM guest LAN, 2x 1GB/s vMotion®, 438GB local HDD, VMware ESXi 5.0.1 (last supported version for this server), paired as 3 clusters

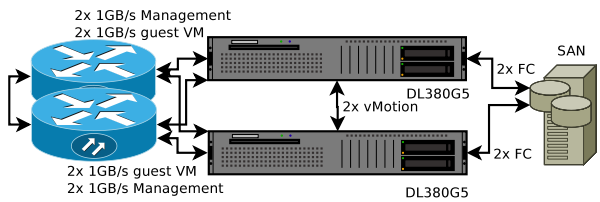

Figure 1: HP DL380G5 cluster

Bladeclusters

- 8x HP BL460G7/G8, 48GB RAM, 2x Intel® Xeon® CPU X5670 @ 2.93GHz (6 Cores), 2 fibre channel cards Flex10® network (0,5GB management, 1,5GB/s vMotion®, 1 GB/s FT®, 7GB VM Guests, 300GB local HDD, VMware ESXi® 5.1 on 6 Blades as one cluster and 2 Blades as one cluster

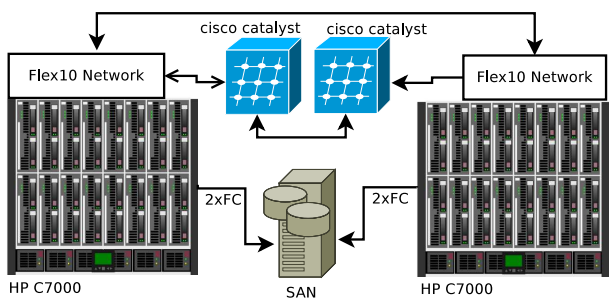

Figure 2: Blade cluster

ISBN 978-3-95450-139-7

Management Host

All our 14 virtualisation hosts are managed by one VMware vSphere® Server. As we started the virtualisation with 4 hosts (with ESX 3.5), we were able to run the management server parallel on our backup host. Since VMware ESX® 4.1 we have to use a dedicated server with 8GB RAM and 64bit architecture. Since releas of ESXi® 5.0 and ESXi® 5.1 we run the management server as a virtual machine on our Bladesystem. This enables high availability for the management without installing a slave management server. Since vSphere® 5.1 a massive increase of resources was required. Now 12GB RAM for the management server including the management database (Microsoft® SQL Server® 2008 R2) and 50 GB disk space is needed.

Virtualisation Hosts

For the core systems, we use VMware vSphere® Enterprise and Enterprise+ [1] for virtualisation. This decision was based on the easy to handle graphical front end, high availability, live-migration, live disaster recovery backup and the available compatible hardware set. In addition KVM is used at BESSYII as virtualisation for the beamline interlock systems with good experience.

The core virtual infrastructure is divided into five main sections. Two production clusters (MLS and BESSYII), two development clusters (one rackcluster and one bladecluster) and one cluster for the experimental infrastructure.

The guest VMs are connected to the network via two bonded network cards. Each card is attached as a network trunk to access all needed VLANs below the trunk. Every VMs network card is bonded to a VLAN on a virtual switch of the trunk.

Each cluster connects to a dedicated SAN storage view. The allocated storage size depends on the number of VMs and the services running in the VMs. Some cluster has 300GB SAN space, another has 3TB SAN space.

The management on each host is also bonded on two network cards and is connected with 2x 1GB/s or 2x 0,5GB/s on Flex10® for connection to the vCenter® Server.

The rack based hosts are connected pair wise crossed the other host for life-migration (vMotion®). On the blade hosts the Flex10® is configured with two separate vMotion® networks. On network failures, like switch misconfiguration, the host are able to detect the cluster partner to prevent a wrong HA decision.

Spanning tree protocol is set to PortFast to enable a fast migration of a VM from host to host, otherwise the migration takes more than one second. With PortFast the network/service downtime stays below half a second.

USED VIRTUAL SERVICES FOR THE LIGHT SOURCE OPERATION

All SoftIOCs and OPCServer (EPICS to PLC communication) are virtualized to benefit of the high availability function. On a server crash we have a typical outage of one control service by 3 minutes and a total maximum downtime of 5 minutes without interaction of an administrator.

Important VMs are in addition cloned onto the local disks of the hosts. In case of an unaccessible SAN, viruses and defect VMs, it is possible to restart the clone. Services are available very soon from this disk backup stage, without long delay to get back the complete backup from tape.

Virtualisation is difficult for high IO processes like databases and multiuser filesystem services. To get high availability on your control system NFS server we have a synchronized hot standby fileserver. On hardware failure it is possible to start a script on the virtual server that resumes the services of the physical server.

The management host (VMware vSphere® server) is also virtualized. This enabled the high availability without a slave management host.

Other services on our virtual infrastructure:

- Read only control system consoles, accessible from any office
- Testinstallations
- Build- and deploy-hosts
- Syslog services
- Archive for legacy software installations
- cold standby for important systems (db management, fileserver)
- Network services like DNS, DHCP, database etc. (no core systems)
- Networkdiagnostic and management

EXPERIENCES WITH VIRTUALISATION

Live Backup

VMware allows to create live backups of our virtual machines by creating a snapshot. The snapshot is read directly from SAN by the backup proxy. After reading, the snapshot is dissolved and the data written to the tape. This reduces the network load dramatically. Since ESX® 4.0 it is possible to use live backups, if user backups are created. To reduce the recovery time, and to save space on the tape, it is possible to enable the change block tracking function. With this function, only changes on the virtual machine are stored on tape and used for recovering.

Unpredictable Powerdown

On power down tasting longer than UPS uptime, the whole computing center shuts down. The automatic host restart after host down should be disabled, because after the power is available the infrastructure is not available completely. The first step is to restore the storage (SAN) and the network. After powering up the management server, the management server automatically restarted all VMs. To prevent a migration of all VMs on one single host, it is important to restart the hosts on one cluster at the same time.

ISBN 978-3-95450-139-7

SAN Breakdown

If the SAN service is not available to the host, the guests are frozen in the attempt to read or write data. The same happens to the virtualized management host. After restoring of SAN, most of the VMs continue running without problems. Some guest VMs need a reboot and some VM are automatically restarted by high availability at the inconsistent on SAN availability on the hosts. If it is needed to run important VMs, it is possible to start local cloned VMs by using the ESXi® command line interface. The management server is also cloned on the local disk and startable if needed.

Time Synchronization Problems

it is possible to distribute the time from the virtualisation host to the VMs, but this include two serious problems. The first problem is misconfigured or forgotten time configuration on the host. In this case all running VMs got the wrong time. VMs store their files with a wrong timestamp, OPC server does not process command or too late if the time shift into the future. The other problem is the kernel tick compensation. Because of scheduling the VMs, not all kernel ticks will be processed and the time went into the past. In this case the VMware tools® put the time back to the host time. But if the time went into the future, because of kernel tick compensation on the VM itself, the time is not corrected by VMware tools® (tested in VMware ESX® 3.5). Solution: To prevent damage from this scenario, we use a NTP client on each VM.

Siemens OPC Server® in Virtual Box®

Virtual Box® does not serve all MS Windows® timer correctly. The system time seems okay. Updated permanent via NTP. But uses a Windows timer witch is not updated correctly by the Hypervisor. With a timer ticking ahead the actions in EPICS, all changes are served with extreme delays. In our case the EPICS timestamp shifted to the future and executed commands up to 20 minutes later. After rebooting the server the servertime is in range, for a few days.

Siemens OPC Server® in VMware Workstation® 8

Changing the virtualisation software for the linac OPC to VMware Workstation® 8 while sticking to the same configuration, the outcome was that the virtual BIOS was wrong interpreted by the OPC server. The OPC server can not setup the network devices correctly and so the CPU rises up to 100% load and did not went down. The workaround is to use an Bios image of Workstation 7 in Workstation 8. Conclusion at this Problem: Because lack of testing times, we decided to install the OPC server directly on OS without virtualisation [2].

VMware Player ®

Nice and Free Tool. Unfortunately with less network configuration options and no possibility to configure more than one bridged network device. For this reason, a test installation of the OPC server was not possible.

Why Not One Big Cluster

- Running VMs can only be migrated on the same processor hardware, if the processor compatibility mode is not active or possible. So if you buy new servers you often build a new cluster
- Less Network IO on each cluster node, not all trunks are needed simultaneously on all hosts.
- Services on different networks, facilities and departments should be separated properly.
- Misconfiguration of resource management on VMs do not affect other important VMs. The effects are more separated.

PLANS

Hardware

Our three HP DL380G5 clusters has the last supported version of VMware vSphere® (5.0.1) for this computer generation installed. The next Step is to replace the hosts to a newer computer architecture. After replacing the DL380G5 Server to a newer architecture we are able to use the VMware Fault Tolerance® technology to increase high availability.

Alternatives

New free virtualisation products on the market are interesting for us as an alternative to our current virtualisation set. Xen®, Proxmox® and others provide a complete virtualisation suit with a good management GUI, live migration, high availability, resource management and backup solutions now also included. The have been tested and certain advantages are visible.

SUMMARY

The ease of installation, configuration, maintenance of server and services within the virtualisation environment is very much appreciated widely used at the HZB controls IT-Infrastructure.

REFERENCES

[1] VMware vSphere Documentation `http://www.vmware.com/vmtn/resources/`

[2] Step 7 on VMware Workstation 8 `http://www.automation.siemens.com/forum/guests/PostShow.aspx?PostID=342902`

ISBN 978-3-95450-139-7