Replacement virtualisation service

Alastair Scobie *

24th January 2010

*with input from Chris Cooke and others

1 Introduction

A simple server virtualisation service was introduced in the spring of 2009. As we urgently required a virtualisation service, the technology is developing quickly and we wished to avoid wasting effort duplicating work which might be going on elsewhere in the University, it was decided that, instead of expending substantial effort developing a fully featured service, that we should develop an interim solution for our short term needs.

After a short project to evaluate current offerings, the freely available VMware Server 2.0 was chosen. Alternatives considered were VMware ESXi, Citrix Xenserver and the version of Xen shipped with SL5.

Two years on, a number of significant problems encountered with VMware Server 2.0 necessitate the replacement of this technology with an alternative in a fairly short time-frame.

IS will imminently introduce a virtualisation service based on the enterprise edition of VMware known as vSphere. This service provides guest migration, load balancing and will soon support site-to-site replication of guests. IS have been using this service for a large number of their services for over a year and have, presumably, made great efforts to ensure that it is reliable and performant. In the current financial climate it is politic that we make use of this service unless there are sound reasons for not doing so. Investing any significant effort in developing an in-house service without very strong reasons would be inadvisable.

As from SL5.5, Scientific Linux ships with KVM (Kernel-based Virtual Machine) integrated into the Linux kernel. Fedora 13 (and shortly SL6) ship with significantly improved tools (both script and GUI based) for managing KVM guests. This very much reduces the work in deploying a simple virtualisation infrastructure.

This report takes the form :-

1. Description of existing system and problems
2. Enumeration of requirements for new system
3. Virtualisation technologies
4. Evaluation of IS against requirements
5. Evaluation of KVM against requirements
6. Proposal

2 Existing virtualisation service

Description

Our current virtualisation service is based on the free VMware server 2.0 hosted on Scientific Linux.

There are now three hosts (central, bakerloo and metropolitan) based in the Forum and one host (northern) based at KB. Each site has a spare host: district in the Forum, piccadilly at KB.

Until early 2010, the VM hosts shared the Dothill EVO disk arrays with other servers (eg. AFS file servers). As the load on the VM hosts increased with the growing number of VM guests, sharing this storage became untenable as the disk arrays became overloaded. The problem was exacerbated by the choice of RAID 5 for the disk array volumes. A two disk RAID 5 failure forced the issue, with bakerloo's guests being migrated to a faster RAID 10 EVO volume and central's guests being migrated to a RAID 10 volume on an elderly, but serviceable, Nexsan ATABeast. An IBM DS3524 disk array was purchased in summer 2010 to be dedicated solely for the VM hosts: this array has substantially better performance than the previous arrays used.

Problems encountered

Usability

VMware Server 2.0 is managed either by a Windows only application or by a web based application. As few of the computing team have access to the Windows platform, the web based application has been the primary means of managing guests, both for configuration and serial console access. Unfortunately, however, this application is reliant on a proprietary browser plugin which is not supported by recent versions of Firefox (> Firefox 3.5).

Reliability

For the most part, VM Server 2.0 has proved to be fairly reliable. However, there have been a number of incidents of VM guest corruption. The majority of these occurred before the storage restructure in early 2010, but there have been reports of incidents since then. It is probably reasonable to state that VMware Server has stability issues when using overloaded storage.

The web based management application has proved to be frustratingly unreliable, requiring frequent restarts of the back-end service on the VM hosts.

Reliability has not been sufficiently good enough for CEG to feel that it could mandate use of the existing service over alternatives (eg Virtualbox and physical hardware).

Performance

Performance has been shown to be very heavily related to underlying disk storage performance, particularly to spindle count. Whilst the shared Dothill EVO arrays were struggling to service almost idle VMs, in one recent test with the new IBM array, ten DICE VM guests were installed simultaneously in thirty minutes onto one VM host.

There is some evidence that guests consume an increasing number of CPU cycles over their lifetime - eg an idle guest may consume 2-3% of a CPU after boot, but over 10% after a month.

Maintenance

VMware server 2.0 has not been updated since October 2009. The kernel module required on the VM hosts does not compile under kernels later than that shipped with SL5.3. SL5.3 is unlikely to be supported past the near future : this has obvious implications for maintaining adequate security of the hosts.

3 Requirements

Must

1. Be affordable
2. Be manageable from a variety of platforms and locations
3. Host platform must support VLANs, bonded ethernet, multipath fibre
4. Support Informatics VLANs
5. Performance should be adequate for services which are obvious candidates for virtualisation (eg with low usage and disk I/O needs)
6. Be stable
7. Shutdown VMs cleanly on host shutdown
8. Support DICE server guests
9. Timely creation of new guests (by end of next working day)
10. Timely reconfiguration of existing guests (**Max ?? hours**)

Highly desirable

1. Support live migration
2. Management via a platform agnostic technology
3. Management from command line
4. Serial console support from command line
5. Nagios monitoring of the host platform
6. Delegated reconfiguration of existing guests
7. Management interface to be available on a VLAN other than those in use by guest machine
8. Support suspend and resume to disk

Desirable

1. Potential to integrate console support with existing console infrastructure
2. Support cross site replication for failover
3. Automatic load balancing
4. Direct access to SAN from guest

Some comments

1. **George**: some thought should be given to the network bandwidth implications of putting machines outside our usual network edge, and whether there are any resultant network charging implications.
2. **Graham**: management from command line + serial console support from command line should both be ”must”
3. **Iain**: must support periodic snapshots

4 Virtualisation technologies

KVM

Kernel-based Virtual Machine (KVM) is a loadable kernel module that converts the Linux kernel into a bare metal hypervisor. The KVM hypervisor in Redhat and Fedora distributions is managed with the libvirt API and tools built for libvirt - virt-manager and virsh. Introduced as a feature preview in Redhat 5.4 (and derived distributions), it became a main-stream feature from Fedora 13 and Redhat 6.

KVM is evaluated later in this document.

Xen

Until very recently, Xen was the leading open-source virtualisation technology. It has shipped with Redhat and Fedora distributions for a number of years. However, it is expected to take a back seat to KVM; certainly, RedHat have stated that they will be concentrating on KVM in the future.

Citrix XenServer

XenServer is a commercial offering based on Xen, adding enterprise level features such as live migration, high availability and central management. Unfortunately, however, the management tools are Windows based.

VirtualBox

Oracle VirtualBox is a product which is available in both a proprietary ”personal use and evaluation” form and a less-featured open-source form. It has been widely used within the School for some time. It is mainly targeted at desktop virtualisation and is considered unsuitable for anything other than lightweight server use. There have been some concerns expressed with respect to Oracle’s future intentions for VirtualBox.

VMware vSphere

VMware vSphere is VMware’s enterprise virtualisation product. It is the current market leader

Substantial cost

VMware ESXi

Free, but very much cut down with no enterprise features at all.

Cloud services, eg EC2

Whilst cloud services such as EC2 may be suitable for some virtualisation requirements, a number of concerns preclude consideration of these for provisioning front-line services :-

1. bandwidth limitations for installing and maintaining machines
2. bureaucracy associated with charging for individual machines
3. DP concerns re outwith EU
4. long term availability worries

5 Evaluation of IS virtualisation service

IS provide a virtualisation service based on VMware's vSphere product.

A DICE SL5 guest, trailblazer, was installed on the IS service for evaluation. The guest was installed, using the DICE PXE installer, on the DICE aliens VLAN. The install proved to be remarkably pain-free. The guest was configured as an LCFG slave to provide a realistic workload.

Evaluation against requirements

Be affordable

The current rental charge for a VM guest (with 2GB RAM, 100GB disk and 25% of an E5540 core) on the IS service is £365 per year. The equivalent, excluding any software costs, on a school based service would be around £150 per year. **See appendix**

Neither of these figures include energy consumption. The IS charge includes staff costs associated with running the service.

Be manageable from a variety of platforms and from locations

vSphere is managed from a Windows only application. This application is used to configure and control guests, and is the only route to access a guest's console.

IS are willing to install a Windows 2008 server with the vSphere client application installed, to which users can connect via the RDP protocol. This would provide access from a variety of platforms and locations - RDP clients are available for Windows, Linux, MacOS and iPhone.

Host platform must support VLANs, bonded ethernet, multipath fibre

IS confirm that these are all supported. It's worth noting that IS use RAID5, but with a relatively small number of disks per RAID set to minimise RAID rebuild time. Presumably this partitioning limits the number of spindles available to a VM at any one time, but the flip side is that it reduces the impact of a rogue disk greedy VM.

Support Informatics VLANs

Tested and confirmed to work. Normal DICE PXE installs work fine.

Be stable

Anecdotally the service is reliable. No problems were encountered during the three week test.

Shutdown VMs cleanly on host shutdown

Tested and confirmed to work.

Support DICE server guests

Yes.

Timely creation of new guests (by end of next working day)

Awaiting SLA.

Timely reconfiguration of existing guests (Max ?? hours)

Awaiting SLA.

Support live migration

This is supported and routinely performed for load balancing. It is effectively invisible to a VM's manager and service operation.

Management via a platform agnostic technology

Although clients for RDP are available for many platforms, the protocol is proprietary and so there is a small risk that future versions of the protocol may not be supported by non Microsoft clients.

Management from command line

Not possible. Only graphical access via RDP is possible.

Serial console support from command line

Not possible. Only graphical access via RDP is possible.

Adequate performance for likely virtualisation candidates

During the test, it was found that LCFG compilation was faster on the virtualised guest than on our service LCFG slaves.

Nagios monitoring of the host platform

Presumably IS monitor using their own monitoring framework, but it is unlikely that we could integrate any monitoring with our Nagios setup. It is not clear how much use this would be as there is limited action we could take. We could still monitor guests using Nagios: this is probably sufficient.

Delegated reconfiguration of existing guests

Almost no configuration of existing guests is delegated : just boot device order and uploading of ISO boot images. This is unsurprising given that most changes of configuration would result in a change to the rental charge.

Management interface to be on VLAN other than those in use by guest machine

Yes, this is supported and is default.

Support suspend and resume to disk

Tested and confirmed to work.

Potential to integrate console support with existing console infrastructure

This is not possible.

Support cross site replication for failover

The vSphere technology allows this, but IS have not yet deployed this functionality. They intend to deploy this in 2011.

Automatic load balancing

Yes.

Direct access to SAN from guest

It would be possible to have direct access to a volume on the IS SAN, but not a volume on the Informatics SAN. There is a charging structure in place for use of the IS SAN, but it is significantly more expensive than storage in the Informatics SAN.

Conclusion

This, admittedly short, test showed no problems. The service worked as advertised with good performance.

A significant issue with the IS service is the inability to manage VMs, and access VMs' consoles, from the command line. This would have a serious effect on the computing staff's ability to manage services when working remotely.

The per guest cost of the IS service is roughly double that of an equivalent School service: the high cost of the VMware software and IS staff time associated with providing the service would account for this.

We would want some kind of guarantee from IS that if other units' guests start consuming lots of disk I/O that IS will throttle those guests.

More thought would need to go into the specific network requirements before we deployed any particular "services which require high availability" on the IS service. For example, which site (Forum, AT, JCMB) would you want your machine to be virtually part of? Would you care that our bridged VLAN connectivity doesn't automatically fail over? Is your bandwidth requirement such that we should really factor in EdLAN charges? I think we would also be wary of speaking OSPF to any of those VMs, so services which require ":n" interfaces would probably be out (George).

6 Evaluation of KVM

For the purposes of this report, KVM was tested under Fedora 13 on two Dell desktop hosts with both Fedora 13 and SL5.5 VM guests.

Evaluation against requirements will assume a service hosted within Informatics.

Evaluation against requirements

Be affordable

KVM, and the open source tools to manage it, are shipped as part of Fedora and Redhat derived distributions. A simple service (ie no live migration) based on KVM would be relatively cheap to deploy. Supporting features such as live migration would involve significant development cost.

Be manageable from a variety of platforms and from locations

KVM is managed via the libvirt API. Both GUI and command line tools are available to configure and manage guests.

Host platform must support VLANs, bonded ethernet, multipath fibre

As KVM runs on top of a regular Fedora or Redhat platform these are all supported. Some minor work to the existing LCFG network component will be required to support bridged interfaces on top of VLAN interfaces.

Support Informatics VLANs

Yes.

Be stable

No problems encountered so far. Have had a number of guests running for several weeks.

Shutdown VMs cleanly on host shutdown

Yes. Guests can either be suspended or shutdown on VM host shutdown.

Support DICE server guests

Yes.

Timely creation of new guests (by end of next working day)

Internal School service, so up to us.

Timely reconfiguration of existing guests (Max ?? hours)

Internal School service, so up to us.

Support live migration

Live migration requires that VM guests are located in disk storage shared amongst several VM hosts. This shared storage is typically NFS, iSCSI or FC SAN based. Whilst configuring NFS is simple, configuring shared filesystem storage (eg using GFS2) over iSCSI or FC SAN) is relatively complex. However NFS is not particularly performant and would introduce an extra dependency (an NFS file server).

Management via a platform agnostic technology

Yes, using the libvirt API.

Management from command line

Yes, eg. by using virsh.

Serial console support from command line

Yes.

Adequate performance for likely virtualisation candidates

not yet measured

Nagios monitoring of the host platform

Yes.

Delegated reconfiguration of existing guests

Internal School service, so up to us.

Management interface to be on VLAN other than those in use by guest machine

This is possible.

Support suspend and resume to disk

This is supported and has been tested. However, there does not appear to be a way for a guest to be informed that it has been resumed - this is likely to be necessary to fix things that have broken over the suspend/resume cycle- eg time.

Potential to integrate console support with existing console infrastructure

In principle, but not attempted.

Support cross site replication for failover

The work to support this would be an extension to the support required to support live migration.

Automatic load balancing

This should be achievable, but would depend on live migration support.

Direct access to SAN from guest

KVM supports N_Port ID Virtualisation (NPIV) so it should be possible to allocate unique WWNs to individual VM guests so that LUN masking applies to individual guests rather than to the VM host and all its guests. However, this support depends on FC HBAs (and crucially their drivers) also supporting NPIV.

Conclusion

Deploying a simple service using KVM would involve little development work, particularly if we stick to using the stock virt-manager GUI and virsh command line tools. Slightly more work would be required to make use of the LCFG components already written by the School of Physics - these would allow LCFG to manage guest creation and configuration. However, a KVM deployment would require either an Fedora 13 or Redhat 6 based hosting platform : we have yet to release a server version of Fedora 13.

Performance (particularly disk) is still unknown, but initial indications are that CPU load (particularly for Fedora 13 guests) is significantly less than for VMware.

A substantial amount of work would be required to support live migration using anything other than NFS as shared storage. It is likely that configuring high performance shared filesystem storage systems will become easier in the future as such systems become more main-stream.

Redhat sell a KVM based virtualisation product. This is currently managed using windows based tools, but these tools are being ported to Linux with an expected release of sometime in 2011. This may be worth investigating in the future.

7 Proposal

Given that virtualisation, particularly open-source variants, is still a relatively immature and fast moving field, we propose that we once again opt for a solution which involves minimal development investment.

We propose a hybrid solution :-

Services which require high availability

are deployed on the IS service. This option provides us with a high availability technology with minimal development cost. The only significant problem is that management must be performed using a graphical interface over RDP - this may be an issue for managing VMs over a broadband connection.

An alternative solution would be to use KVM with VM guests residing on shared storage using NFS. However, this is not a performant solution and introduces a dependency on another service, which itself would need to be high availability.

All other services

are deployed on DICE VM servers using KVM with VM guests residing on FC attached storage.

Very little development work (mainly documentation) would be required to deploy a service based on the stock Redhat tools. Slightly more work would be required to allow VM creation and configuration to be managed by LCFG.

If KVM proves to be stable after, say, 12 months we consider investing the effort in deploying a shared storage subsystem for the VM guests allowing us to support features such as VM migration etc.

There will be some low-end requirements which aren't suitable for virtualisation (eg due to security concerns) - these can be satisfied by Atom based kit.