

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Presented at NLIT, June 16, 2011

Vail, Colorado

David Cortijo

Brookhaven National Laboratory

dcortijo@bnl.gov

Notice: This presentation was authored by employees of Brookhaven National Laboratory, under Contract No. DE-AC02-98CH10886 with the U.S. Department of Energy. The United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this presentation, or allow others to do so, for United States Government purposes.

BROOKHAVEN
NATIONAL LABORATORY

a passion for discovery



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

- Historical perspective of infrastructure management at BNL
- Decision points for Virtualization platform
- Hardware and software requirements for RHEV implementation
- Brief overview of RHEV features – current and future
- The story so far...
- Path forward at BNL
- Caveats and potential problems

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Historical perspective

- Like many datacenters, BNL was using bare-metal servers to provide nearly all services
- Due to power, space, and cooling constraints within the datacenter, potential growth of service offerings was slowed immensely
- Hardware purchase delays caused new service implementations to take several months before preliminary testing could really begin

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Historical perspective – some details

- Dozens of lightweight services running on bare-metal servers
- Multiple services often shared hardware out of necessity (cost, space, etc)
- Some examples:
 - 14 DNS servers, several also serving DHCP
 - 3 dedicated DHCP servers
 - Web servers hosting dozens of virtual hosts – some internal only and others with external access on the same machine
 - Hardware was found to be underutilized to an untenable degree – 10 machines were doing the work that one could do in some cases

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

The path to Virtualization

- Decision was made to virtualize in order to address the multitude of concerns and constraints presented
- Initial work done with Xen 3.0.3 embedded into Red Hat Enterprise Linux 5, with Linux HA/Heartbeat and custom scripts to provide redundancy
- Few resources beyond hardware and manpower existed – no money for licensing at the time
- Many problems – in particular lack of VLAN support – caused unnecessary physical server sprawl
- Xen project was scrapped in favor of a better supported solution

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Decision points on Virtualization

- Various factors contributed to the decision to move to RHEV
 - Cost
 - Best support of Linux platforms – in particular RHEL
 - Visibility into host – not looking for bare-metal hypervisor implementations as a requirement
 - Live migration of VMs
 - 802.1q/VLAN tag support
- Cost proved to be the largest determinant, but not the only one

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Why RHEV?

- Cost was 1/6 that of VMWare in our sample implementation pricing
- RHEL Server acting as host platform included unlimited guest licensing for RHEL
- 802.1q support worked out of the box without needing complicated configuration
- Storage Live Migration (the only VMWare feature that RHEV did/does not have) was not viewed as a strict requirement – more on this later

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Glossary of RHEV-related Terms

- KVM – Kernel-based Virtual Machine
- Host – Physical machine that VMs run on
- Data Center – Set of Hosts with shared storage and network definitions
- Cluster – Subset of a Data Center; must share identical networks between Hosts
- LVM – Logical Volume Manager
- Storage Domains – Analogous to LVM Volume Groups; set of disks shared to RHEV Hosts
- Live Migration – moving a running VM from one Host to another without interrupting service

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Glossary of RHEV-related Terms

- Bare-Metal Hypervisor – lightweight OS that allows hardware to run VMs without a full-blown OS installation
- RHEV-H – Red Hat's Bare-Metal Hypervisor, which is a stripped down RHEL implementation
- RHEV-M – RHEV Manager software, resides on a separate server
- vdsmd – Virtual Desktop Server Manager daemon, which allows RHEV-M to manage and monitor VMs and send commands to Hosts
- Virtual Guest, or Guest – another term for a Virtual Machine

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Data Center configuration

- Note the “type” constraint



New Data Center

Name: TestDatacenter

Description: Datacenter for Demo

Type: NFS

Compatibility Version: NFS

FCP

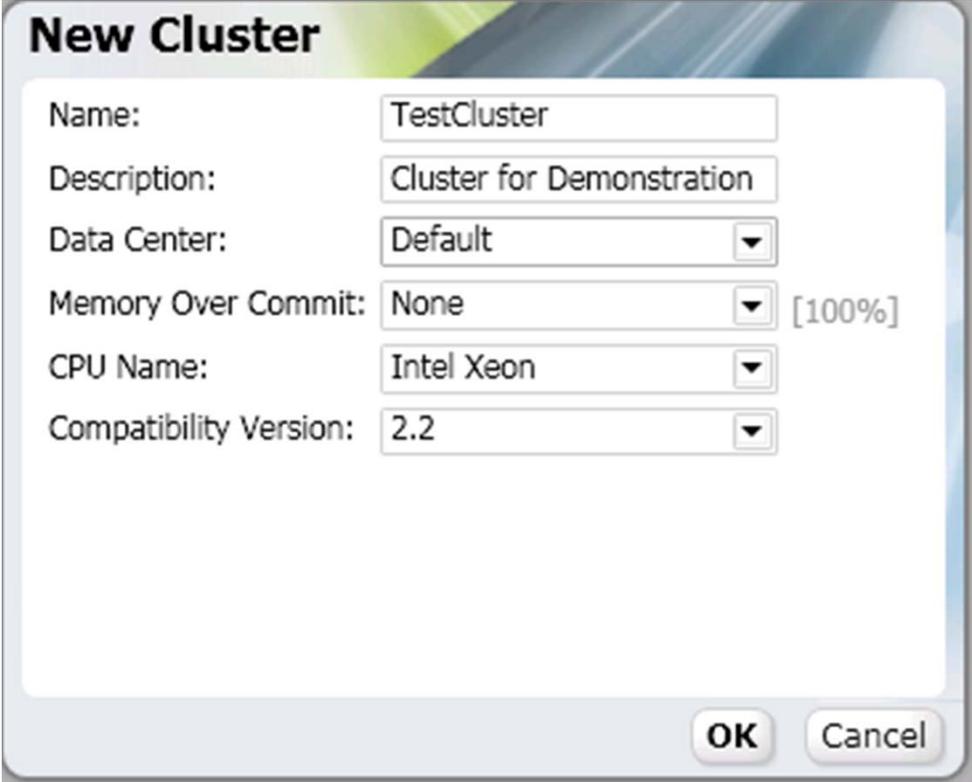
iSCSI

OK Cancel

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Cluster configuration

- Cluster is bound to a Data Center on creation – cannot be changed later



New Cluster

Name:	<input type="text" value="TestCluster"/>
Description:	<input type="text" value="Cluster for Demonstration"/>
Data Center:	<input type="text" value="Default"/> ▼
Memory Over Commit:	<input type="text" value="None"/> ▼ [100%]
CPU Name:	<input type="text" value="Intel Xeon"/> ▼
Compatibility Version:	<input type="text" value="2.2"/> ▼

OK Cancel

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Host configuration

- Hosts can be moved between Clusters/ Data Centers, but must have the appropriate Logical Networks or they will not work properly

Edit Host

Name:

Address:

Port:

Host Cluster:

Enable Power Management

Address:

User Name:

Password:

Type:

Options:

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Hardware and Software Requirements

- RHEV requires at least 2 Hosts per Cluster (and Data Center) to properly operate
- Hosts must have AMD-V or Intel VT hardware virtualization support and Intel 64 or AMD64 CPU extensions
- All members of the Data Center must have access to the same shared storage via iSCSI, Fiber Channel, or NFS
- Sufficient RAM and CPU to run virtual machines
- Network Connectivity to all networks assigned to the Cluster that the Host is a member of
- Dedicated RHEV-M machine (currently required to be a Windows Server platform – RHEL in next major release)

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

RHEV features

- Hosts can be integrated into RHEV with remote installation direct from the interface
- Storage domains integrate seamlessly into RHEL hosts using LVM
- Quick, push-button migration of Guests between Hosts in the Cluster
- Guests can be easily installed/kickstarted through shared ISOs directly from the UI

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

RHEV features

- Automatic balancing of load within Clusters based on parameters defined by the admin
- Ability to snapshot or move Guest data of a downed VM to different storage domains – cannot be done live
- Ability to fence unresponsive hosts and restart VMs automatically
- Data Centers/Clusters scale horizontally very easily

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

BNL implementation – the story so far

■ Clusters/Data Centers

- Two RHEV Data Centers – one within a load balanced environment and another on the main campus network
- Three clusters – one within the load balancer, main campus split into two
- Each cluster has 2-3 Hosts
- Access to multiple networks via dual ethernet and 802.1q trunks

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

BNL implementation – the story so far

■ Hosts

- Seven Dell m610 blades
 - Dual Intel Xeon E5530 w/ Hyper -Threading enabled (16 simultaneous threads)
 - 48 GB RAM
 - Dual port Qlogic HBA expansion card in each blade
- Blades currently reside in a single m1000e chassis – second is being prepared for production
- M1000e contains Brocade 4424 FC switch for SAN connectivity

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

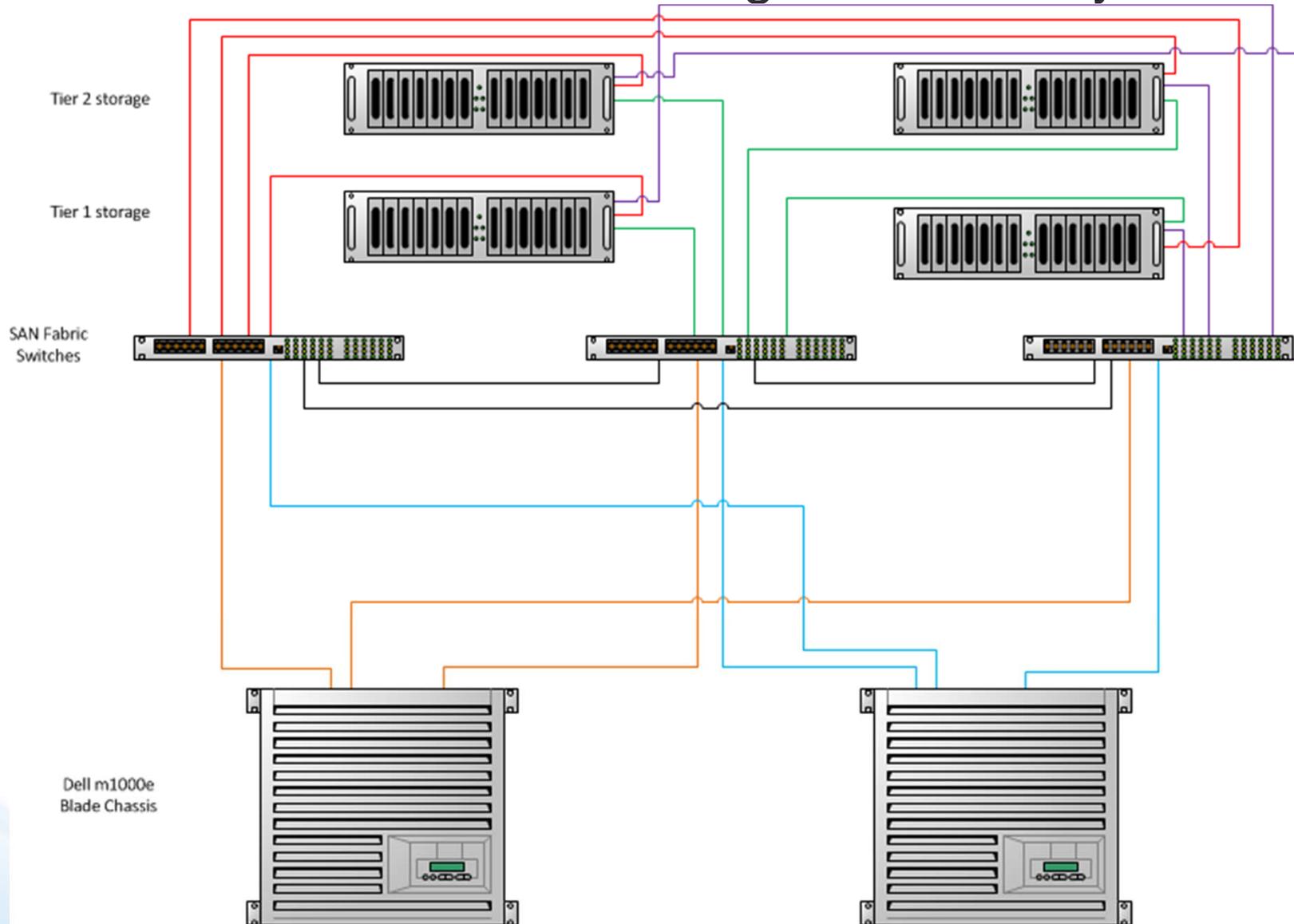
BNL implementation – the story so far

■ Storage

- Four RAID units – two Tier 1 and two Tier 2 – connected via redundant Fiber Channel SAN
- Roughly 7 TB of storage shared out via multiple storage domains to the appropriate RHEV Data Centers
- Storage for Guests that provide redundant services (i.e. paired DNS servers) lives in Storage Domains provided by different physical RAID devices (manual determination/implementation)

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Brief look at redundant storage connectivity:



Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Impact on service offerings

- Many core services are now virtual and provide far better reliability than ever before. For example:
 - DNS
 - DHCP
 - SSH gateways
 - Second-tier mail relays
 - NTP
- Problematic webservers and other services have been properly segregated (internal vs. external access)
- Guests for testing or new builds are now available within minutes, rather than hunting down hardware

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Performance and Reliability thus far

- With the exception of a few bugs early on, RHEV has proven to be an extremely stable platform
- In the event of Host failure or loss of connectivity to Host, all VMs set for High Availability are restarted on other available hosts in the cluster in under 5 minutes (from time of failure, not detection)
- Live migration between Hosts takes moments, even for those with significant I/O, allowing manual and automatic balancing of load as well as Host maintenance to be transparent to users
- Storage has proven to be our largest vulnerability

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Power, space, and cooling impact

- Virtualization has thus far allowed us to eliminate dozens of servers from the datacenter
 - Over 60 bare-metal servers decommissioned
 - Eliminated over 2 full racks of servers and supporting hardware
- Net power and cooling consumption dropped dramatically
 - Peak power change > 30 kW
 - Approximate heat delta $> 100,000$ BTU/hr reduction
 - 20% reduction of peak load on PDUs

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Storage woes

- Two storage failures in recent months
 - First one was a critical failure of a RAID volume
 - Unavoidable
 - Recovery from backup was as per standard operating procedures
 - Second left RAID unit in a semi-working state
 - Device was unmanageable
 - Forced choice between availability of VMs or maintenance of array hardware
- Both issues could have been addressed with better storage management in RHEV

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

More details on storage woes

- Were live snapshots available in RHEV, recovery from initial failure may have been cut to a fraction of the time
- Snapshots in RHEV are done to the same disk as the Guest's storage – not helpful in the event of failure of the underlying disk
- Even given the problems created by the first failure, time to recover was no more than that of a bare-metal server's reinstall/recovery from backup

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Some more woes

- In the second instance, live migration functions would have allowed us to move running VMs off of the array, rather than a series of short outages for a large number of systems
- Due to switch resets, Dell Brocade 4424 took over as “principal switch”, which caused path failures
- Lack of hardware redundancy on the blade servers and internal switches made full restoration take much longer than expected

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Path forward at BNL

- Additional Dell m1000e chassis to balance blade load
- 4 Additional blades to expand existing clusters to 3+ nodes
- Upgrade storage for increased performance and to spread out impact of potential failures
- Additional in-chassis Fiber Channel switches to protect against failure and allow for switch maintenance

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Path forward at BNL

- Increase reliability by merging the 2-Cluster Data Center into a single Cluster
 - Longer term work
 - Requires collaboration with Networking team to re-architect service-specific subnets
 - End state of 2 Clusters (1/Data Center) with 4+ Hosts
- Increase Virtualized service offerings to Lab-wide community
 - Replace bare-metal service offerings and support of disparate machines
 - Move more services into the central datacenter

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Path forward at BNL

- Investigate VDI (Virtual Desktop Infrastructure) solutions
 - Software comes built in to RHEV
 - Additional licensing costs are minimal
- Continue to identify targets for virtual servers, eliminating unnecessary hardware wherever possible
- Make better use of power management and load balancing functionality within RHEV itself (not currently used to its full potential)

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Caveats and potential problems

■ Storage

- Extremely important to understand how your underlying storage is configured both physically and logically to identify potential issues
- Redundant storage fabric and physical devices is the largest single concern we've had

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Caveats and potential problems

■ Networking

- Work with your Networking experts to ensure all layers that can be are fully redundant
- Where possible, connect members of the same Cluster to different physical switches to ensure availability in case of failure
- Even if you think you might not need them at first, start from day one with 802.1q trunks – flipping the switch later is quite the headache
- Decide which network you will use for management (Logical Network called “rhevm”) before you begin

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Caveats and potential problems

- Host hardware and software
 - Everything relies on your Hosts – make sure they're under warranty, you have spare parts, etc
 - Make sure your Host systems are patched – should go without saying
 - If possible, segregate them physically
 - Avoids the Fiber switch problems we've seen
 - Allows for chassis, switch, or rack maintenance where necessary

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

In closing...

- RHEV has provided huge benefits in terms of reliability and reducing administrative overhead on hardware overall, as well as cleaning up and “greening” the datacenter
- Storage concerns are biggest Achilles Heel, but impact will seem exaggerated given the 18 months since implementation saw few problems
- If we had to start from scratch, RHEV would still be the product of choice – we'd simply invest more in storage and SAN support from the beginning

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Additional Resources

- <http://www.redhat.com/virtualization/rhev/server/>
 - Red Hat's page on RHEV for servers
 - Contains links to various whitepapers
- https://access.redhat.com/knowledge/docs/Red_Hat_Enterprise_Virtualization_for_Servers/
 - RHEV Documentation
- <http://rcritical.blogspot.com/>
 - RHEV related blog
 - Lots of useful notes, tips, and tricks for RHEV

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

In closing...

- RHEV has provided huge benefits in terms of reliability and reducing administrative overhead on hardware overall, as well as cleaning up and “greening” the datacenter
- Storage concerns are biggest Achilles Heel, but impact will seem exaggerated given the 18 months since implementation saw few problems
- If we had to start from scratch, RHEV would still be the product of choice – we'd simply invest more in storage and SAN support from the beginning

Red Hat Enterprise Virtualization - KVM-based infrastructure services at BNL

Q & A (time permitting)

If you have any further questions, feel free to contact me at:

David Cortijo

ITD Unix Services

Brookhaven National Laboratory

dcortijo@bnl.gov

631-344-2053