Virtualization

A very short summary
by Owen Synge

Outline

- What is Virtulization?
  - What's virtulization good for?
  - What's virtualisation bad for?
- We had a workshop.
  - What was presented?
  - What did we do with the rest of our time?
- So whats happened since?
- Collaboration and progress.

What is virtualization?

- from Enterprise Management Associates:
  - technique for hiding physical characteristics of computing resources from the way in which other systems, applications, or end users interact with those resources.
  - making a single physical resource appear to function as multiple logical resources.
  - making multiple physical resources appear as a single logical resource.

Yves Kemp found this was like Gid for his GrudKa talk on grid virtualisation

Whats Virtualization Good for?

- Consolidating services (Not todays talk)
- Test systems (Quickly)
  - Build Nodes
  - Deployment testing
- Security / Management (Main Focus of Today)
  - Isolation of concerns
- Scheduling (Maybe if Time)
  - Better management of resources

What Virtualizations Bad For ?

- Depends on implementation.
  - Xen
    - Latency
      - (MPI on the worker node may not be ready need more specs)
  - UML
    - Disk/Network IO
  - Vserver/Chroot
    - Isolation
    - Security
- Clear Xen is Currently most popular amongst the HEP community.

We had a workshop at DESY

- 16-17 January 2007
- Bias toward LCG
- 70:30 ratio of Presentations:Group.
- Attendants
  - System administrators dominate
    - wLCG deployment team
  - Solution Providers / Users
- Group Session
  - Many Admins and deployment experts

Presentations

- Deployment
  - Trinity Collage Dublin (Irland)
  - CERN (Swis)
  - DESY (Germany)

- Worker Nodes
  - Metacenter (Czech)
  - Karlsruhe (Germany)
  - Luebeck (Germany)
  - Globus (USA)
  - Masaryk (Czech)

Virtualization Users workshop Group Discussion

- Focused just on worker Node
- 5 models compared.
  1. (1)Persistent VM 1 OS
  2. (2)Persistent Vm Many OS
  3. (3)Non Persistent VM's
  4. (4)Non Persistent VM's OS Library
  5. (5)Non Persistent VM Dynamic OS
- Root on the worker node

One persistent virtual machine with a single OS images.

- Benefit
  - Job is isolated from man. operating system
  - Security of base OS image
  - Possibly use able to suspend jobs
  - Consistent DOM0 image (No user access)
  - Easy to restore images/Maintain
  - Eases hardware abstraction (SL3/SL4 EXAMPLE)
  - Technology is already available.

- Cost
  - Performance (Slight)
  - Maintenance of Two OS's
  - Lots of different systems used need to learn how to integrate them
  - management tools are not available.

Multiple/2 persistent virtual machines with multiple/2 OS's.

- Benefit
  - Useful for parallel Jobs and back filling
    - Increasing cluster utilization when queue draining would normally be required
  - Job is isolated from management operating system
  - Security of base OS image
  - Possibly use able to suspend jobs
  - Consistent DOM0 image (No user access)
  - Easy to restore images/Maintain

- Cost
  - Performance (Slight)
  - Maintenance of Two OS's
  - Lots of different systems used need to learn how to integrate them
  - management tools are not available.
  - Memory is not shared so does not scale to N images.

Running non persistent virtual machine images.

- Benefit
  - Security is greatly enhanced
    - Worker node cleaning and job deamoization fears are eliminated
  - We believe Minimal modification to batch system required as job submission epilogue and prologue can hide virtual machines details.
  - We believe that non persistent virtual machines will be the future of the worker node
  - Memory available to a job is clearly split so providing better job isolation.
- Cost
  - Cost of restarting Virtual machines.
  - Virtual machines is not just a process, its a set of requests to hypervisor from schedulers perspective.
  - An Adapter will be needed.
  - Batch system needs information on CPU used ability for stopping VM's from scheduler work
  - Memory issues may cause problems as clearly split.
  - Logging issues, will we be able to store logs without further development.

Running non persistent virtual machine images.

- Benefit
  - Security is greatly enhanced
    - Worker node cleaning and job deamoization fears are eliminated
  - We believe Minimal modification to batch system required as job submission epilogue and prologue can hide virtual machines details.
  - We believe that non persistent virtual machines will be the future of the worker node
  - Memory available to a job is clearly split so providing better job isolation.
- Cost
  - Cost of restarting Virtual machines.
  - Virtual machines is not just a process, its a set of requests to hypervisor from schedulers perspective.
  - An Adapter will be needed.
  - Batch system needs information on CPU used ability for stopping VM's from scheduler work
  - Memory issues may cause problems as clearly split.
  - Logging issues, will we be able to store logs without further development.

Non persistent virtual machines image from a library of images.

- Benefits
  - Great flexibility of run time environment
  - Multiple environments
  - Security
  - No need to balance resources.
  - Experiments don't have to agree on SL3/4
  - All benefits shown for other models.

- Cost
  - Additional integration with VM and batch system
  - Information System changes.
  - JDL should describe Image to be used.
  - CE needs all of JDL information, unfortunately this is lost in current LCG grid.
  - Scheduler has needs to be greater aware of VM
  - Concerns over batch system Independence.
  - We would not get all batch systems integrated before OFFICIAL switch on even with dedicated funding.

User defined images running on non persistent virtual machines.

- Benefits
  - Experiments don't have to agree on SL3/4
  - All benefits shown for other models.

- Cost
  - New infrastructure needed
  - Need to define Image generation
  - User supplied images may scare admins

Globus Virtual Workspaces Summary

- Web service controlling VM
  - Model 5
- PBS plugins
- rpath based deployment
- Starting with production deployments
  - STAR community
  - Cliemate Community (CCSM)

Simon Fraser University

- Virtualized serial jobs
- Native MPI jobs
  - Due to xen latency / bandwidth issues issues
- Implementation
  - Solaris Xen management domain
  - Torque >= 2.0 + MOAB >= 5.0
  - Migration to production of 2000 nodes by March

University of Magdiburg

- Model 5 Worker nodes
  - User defined per job VM's
- Click and build Image
  - Debian and Ubuntu based images
- Supporting Torque Batch Queue
  - Planning to support Sun Grid Engine
- Research project only.

CERN's Glite testing frame work

- In production
- For sumulating grids
  - For testing deployment
    - Basis of Certification for Glite CERN
  - For training perposes
- Uses Cluster management techniques
- Web interface

Trinity Collage Dublin

- Simulates Grid Irland
  - Used to deploy nodes before shipping to site
- Provides full Quattor cluster management
- pygrub
  - pxeBoot mechanism for VM
  - management of master OS
  - Can then use existing boot env

Desy Build and development

- xen-image-manager.py
  - 45 seconds to reinstall an OS and boot it
  - Going on source forge in next few days.
- Build service
  - Not yet ported to SL4
- Vm based
  - Development hosts
  - Release management / Certification
- XEN SL5 + AFS plus pygrub going int production for service consolidation

Conclusions from work shop

- We believe Virtualization will be adopted on the worker node
  - in an incremental fashion
  - Both Karlsruhe and the MetaCentre
    - have already reached beyond the simplest model of a persistent virtual image, due to local demands on resources.
- Non persistent virtual machines image from a library of images.
  - No more needed for wLHC at moment.

Conclusions from work shop

- We expect that sites will adopt Virtualization in stages due to the benefits of abstracting the hardware from the operating system.
  - Experiments wanting to use older OS.
  - Site specific differences in deployment
    - easing use of the Grid for the scientific community.
      - Eg French or German version of Perl
- Root on the worker node possible but not keen.
  - Insecure network services.
  - Poor practice would become established.

Conclusions from work shop

- The LHC computing community demands that all users use a consistent operating system.
  - Challenged as we upgrade from SL 3 -> 4.
  - Not all grid user communities happy with OS driven centrally.
- Resolved by running within virtual machines.
  - New potential grid communities who may have standardized on other operating system environments.

Conclusions from work shop

- The benefits of running non persistent VM's.
  - Takes about 45 seconds.
    - xen-image-manger.py
  - The Virtual host is reinstalled per job.
  - Management and security Advantages.
- Many members of the workshop believe that this is the compelling reason for running virtual machines.

Conclusions from work shop

- The grid will ultimately demand a heterogeneous model of OS's.
  - Virtualization seems a way to mange this.
- We expect that there will be some opposition to immediate adoption of Virtualization
  - communities who typically have a performance focus and low security requirements.
  - Inter VM Latency concerns

Conclusions from work shop

- Usage of VM will evolve and change as these technologies are still young in the commodity computing sector.
- Still need integration between batch queue and VM.
- Worker nodes will all be a VM in next 5 years
  - Near unanimous perspective.