# *Running CE and SE*

# *in a XEN virtualized environment*

*S.Chechelnitskiy*

*Simon Fraser University*

*CHEP 2007*

*September 6th 2007*

ATLAS on






***Simon Fraser***

# *Running CE and SE in a XEN virtualized environment*

- **Motivation**
- **Proposed architecture**
- **Required software**
- **XEN performance tests**
- **Test suite and results**
- **Implementation schedule**
- **Running SE in XEN**
- **Conclusion**

ATLAS on




Simon Fraser

# Motivation

- SFU is a Tier-2 site
- SFU has to serve its researchers
- Funding is done through WestGrid only

- WestGrid is a consortia of Canadian Universities. It operates a high performance computing (HPC), collaboration and visualization infrastructure across Western Canada.
- WestGrid SFU cluster must be an Atlas Tier-2 CE.
- CE must run on a WestGrid facility

# Motivation (cont.)

Current WestGrid SFU cluster load

- 70% of jobs are serial
- 30% of jobs are parallel (MPICH-2)
- 10% of serial jobs are preemptible

Scheduling policies

- No walltime limit for serial jobs
- 1 MPI job per user with Ncpu = 1/4 of Ncpu_total
- Minimize waiting time through the use of preemption
- Extensive backfill policies

# Motivation (cont.)

**Atlas and local jobs must run on the same hardware.**

Atlas Cluster Requirements:

- Particular operating system
- LCG software (+experimental software, updates, etc.)
- Connectivity to the Internet

WestGrid (local) Cluster Requirements:

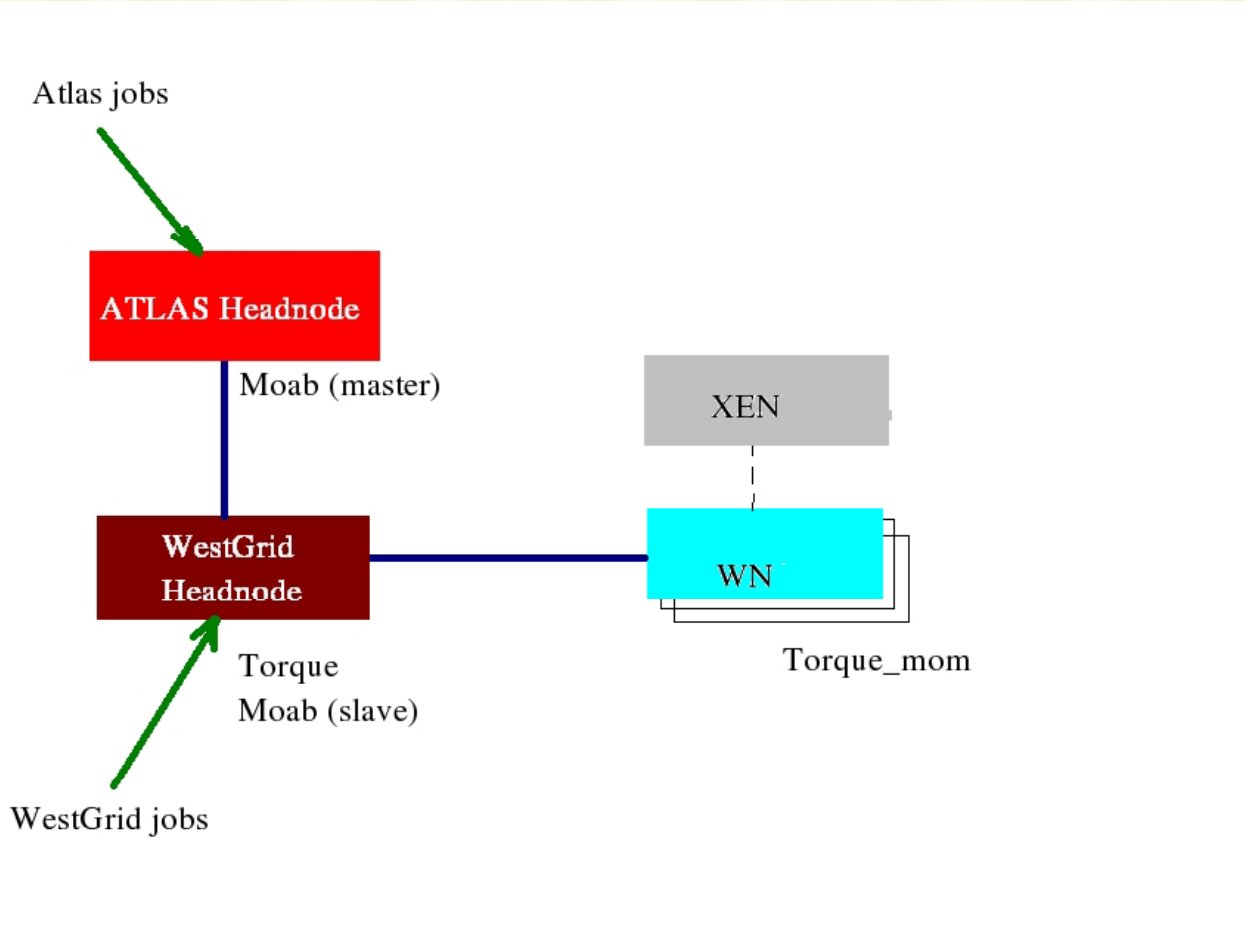
- No connectivity to the Internet
- Lots of different software
- Recent operating systems
- Low latency for MPI jobs

# Proposed solution

3 Different Environments for 3 Jobs Types:

- WNs run a recent operating system with XEN capability (openSUSe-10.2)
- **MPI jobs** run in the non-virtualized environment
- For each **serial job (Atlas or local)** a separate virtual machine (XEN) is started on a WN (max number of XENS == Number of cores)
- **Local serial jobs** start XEN with openSUSe-10.2 and local collection of software
- **Atlas jobs** start XEN with SC-4 and LCG software
- 2 separate hardware headnodes for the local and Atlas clusters

# Proposed architecture






Simon Fraser

# Software Requirements

- OS with XEN capability
- Recent Torque version >= 2.0
- Moab cluster manager, version >= 5.0
- Some modifications to LCG software is necessary

# XEN Performance Tests

- Performance for serial jobs in XEN is better than 95% of the performance of the same job running in a non-virtualized environment and the same OS
- Network bandwidth is still a problem for XEN but it is not required by serial jobs (neither Atlas, nor local)
- Memory usage for one XEN virtual machine is about 230-250 MB
- XEN startup time is less than a minute

# Test Suite and Results

Test Suite:

- One SUN X2100 as an Atlas headnode, software installed: LCG-CE, Moab (master), mounts /var/spool/pbs from the local headnode for accounting purposes
- One SUN V20Z as a local headnode, software installed: Moab (slave), Torque
- Two SUN X4100 (dual core) as WNs, software installed: torque_mom, XEN-kernel
- Images repository is located on the headnodes and mounted to WNs

# Test Results (continued)

Test Scenario:

- One 4 cpu MPI job running on 2 different WNs, two local jobs and two Atlas jobs starting their XENs
- Play with priorities, queues
- Play with different images

Test Results:

- Everything works, ready for implementation
- Preliminary results: migration of serial jobs (from node to node) also works
- Images are successfully stored/migrated/deleted

# Test Results (continued)

Advantages of this Configuration:

- Great flexibility in memory usage (Moab can handle various scenarios)
- Very efficient hardware usage
- Clusters can be configured/updated/upgraded almost independently
- Great flexibility: can migrate all serial jobs between nodes to free the whole WN(s) for MPI jobs

# Implementation Schedule

- October 2007 - Setup a small prototype of the future cluster: two final separate headnodes, 2 to 5 worker nodes
- November 2007 - Test the prototype, optimize scheduling policies, verify the changes to LCG software
- November - December 2007 - convert the prototype into a production 28+ nodes blade center cluster, run real WestGrid and Atlas jobs, replace the current SFU Atlas cluster
- March 2008 - expand the small cluster into the final WestGrid/Atlas cluster (2000+ cores)

# Running SE-dcache in XEN

## Motivation

- Reliability.
  - In case of a hardware failure you can restart the virtual machine on another box in 5 minutes.
  - Run XEN on openSolaris with ZFS and use ZFS advantages (improved mirroring, fast snapshots) to increase the stability and improve the emergency repair time.

# Running SE-dcache in XEN (cont.)

Limitations:

- Due to XEN's poor networking properties only the core dCache services can run in a virtualized environment (no pools, no grdftp or dcap doors)

Test Result:

- A base dCache-1.8 with Chimera namespace provider was installed in a XEN running on openSolaris-10.3 (guest OS was openSUSe-10.2)
- Same dCache installations were made on the same hardware running openSolaris and openSUSe (to compare the performance)
- Pools and doors were started on different hardware

# Running SE-dcache in XEN (cont.)

Test Result:

- All functional tests were successful
- Performance was up to 95% of the same dCache configuration performance running at the same OS in a non-virtualized environment
- Benefit: run different dCache core services in different virtual machines and balance the load on the fly

- Non-XEN dCache runs approximately 10-15% faster on Solaris than on Linux (AMD box). Better java optimization, memory usage ?

# Benefits of a Virtualized Environment

- Low Risk Scalability: Running jobs in a virtualized environment is possible and has a lot of advantages. Disadvantages are minor
- Lower Costs: Buy a general purpose cluster and use it for many different purposes
- Stability: Setup your site specific scheduling policies without affecting your Atlas cluster
- Flexibility: Run different LCG services in a virtualized environment to improve the stability and flexibility of your site

# Acknowledgments!


- Thanks to Cluster Resources Inc. for their help in setting up the specific configuration
- And thanks also to SUN Microsystems Inc. for providing the hardware and software for our tests


# Questions?