Virtualization in High Performance Computing

ATLAS on

S.Chechelnitskiy

Simon Fraser University

WestGrid

Global Sun Virtualization Chalk

October 4th 2007

Motivation

- One cluster should run two very different kinds of jobs:
  - WestGrid jobs (serving all the University's researchers)
  - Atlas jobs (serving only the High Energy physics - HEP community)

WestGrid is a consortia of Canadian Universities. It operates a high performance computing (HPC), collaboration and visualization infrastructure across Western Canada.

Simon Fraser

Motivation (cont.)

Atlas Cluster Requirements:

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

WestGrid Cluster Requirements:

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

Proposed solution

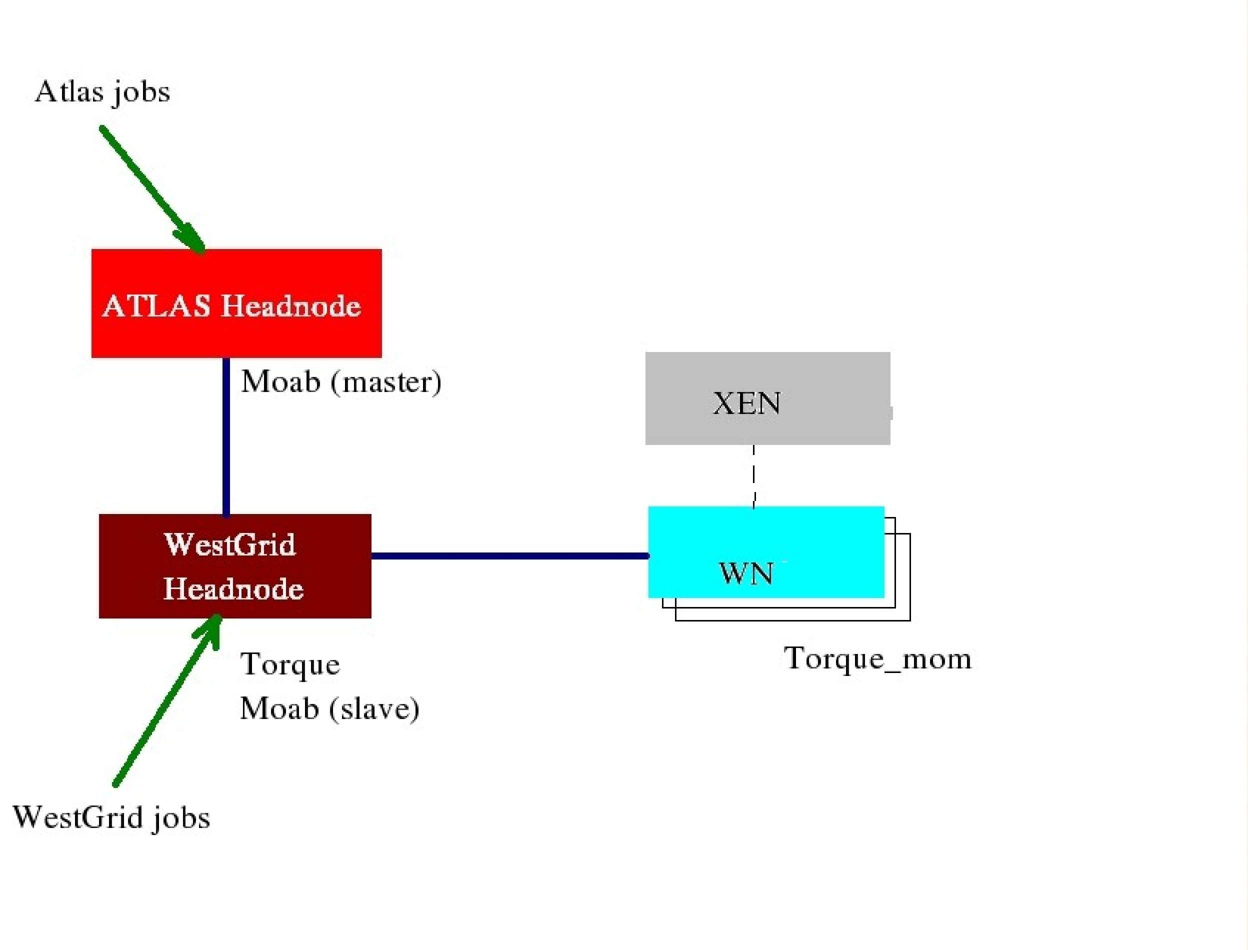
3 Different Environments for 3 Jobs Types:

- Worker Nodes run a recent operating system with XEN capability
- For each **serial job (Atlas or WestGrid)** a separate virtual machine (XEN) is started on a WN
  (max number of XENs = Number of cores)
- **WestGrid serial jobs** start XEN with openSUSe-10.2 and local collection of software
- **Atlas jobs** start XEN with SC-5 and LCG software
- **MPI jobs** run in the non-virtualized environment

WestGrid

Simon Fraser

Proposed architecture



Simon Fraser

XEN Performance Tests

- Run the same program or application in Dom0 (Linux or Solaris) and DomU (Linux only) and compare the performances:
  - Simple C program (pure integer or pure floating point) - 99%
  - Multi-threaded Java application - 95%
  - Multi-threaded Java application with a database backend and high networking activity - 50-70%

Test Suite

Test Suite:

- One SunFire X2100 as an Atlas headnode
- One SunFire V20Z as a local headnode
- Two SunFire X4100 (dual core) as WNs
- Images repository is located on the headnodes and mounted to WNs

Test Scenario and Results

Test Scenario:

- One 4 cpu MPI job running on 2 different WNs, two WestGrid 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

Virtualization advantages

- Flexibility
  - Great flexibility in memory usage (the scheduler can handle various scenarios)
  - Can migrate all serial jobs between nodes to free the whole WN(s) for MPI jobs
  - Clusters can be configured/updated/upgraded almost independently
  - Setup your site specific policies without affecting other cluster(s)
  - Reorganize your cluster (add/remove/upgrade nodes, change virtual hardware configurations) on the fly

Virtualization advantages

- Reliability
  - Run XEN on Solaris with ZFS and use ZFS advantages (improved mirroring, fast snapshots) to increase the stability and improve the emergency repair time.
  - In case of a hardware failure you can restart the virtual machine on another box in 5 minutes.
  - Run different software components in different VMs but within exactly the same environment, add/remove/mirror VMs on the fly

Virtualization advantages

- Security
  - Setup licensed software/licensing server in a virtual environment with specific networking/firewall settings
  - Restrict access to your Dom0 environment and provide a user access only to your virtual environment
  - Provide different user credential - user account mappings in different environments