ATLAS Grid Compute Cluster with virtualized service nodes

J Mejia\textsuperscript{1}, S Stonjek\textsuperscript{2} and S Kluth\textsuperscript{2}
\textsuperscript{1} Rechenzentrum Garching der Max-Planck-Gesellschaft, Garching bei München, Germany
\textsuperscript{2} Max-Planck-Institut für Physik, München, Germany

Abstract. The ATLAS Computing Grid consists of several hundred compute clusters distributed around the world as part of the Worldwide LHC Computing Grid (WLCG). The Grid middleware and the ATLAS software which has to be installed on each site, often require a certain Linux distribution and sometimes even specific version thereof.

On the other hand, mostly due to maintenance reasons, computer centres install the same operating system and version on all computers. This might lead to problems with the Grid middleware if the local version is different from the one for which it has been developed.

At RZG we partly solved this conflict by using virtualization technology for the service nodes. We will present the setup used at RZG and show how it helped to solve the problems described above. In addition we will illustrate the additional advantages gained by the above setup.

1. Introduction
To participate in the ATLAS worldwide computing effort, the involved computer centers have to install Grid middleware of the LHC flavour. This includes at least one compute element (CE), at least one storage element (SE), one BDII information server and one monitoring box. A failure of any of those machines will most likely create problems for the affected Grid center and the jobs running there or those jobs which are already scheduled to run on that particular center.

Therefore it is advisable to run the central middleware services of a Grid computer center on hardware with increased reliability. The increase in reliability is often achieved by redundant hardware components (power supply etc.). This results in higher prices for this type of machine. Some of the Grid services require only limited compute power. Therefore a separate server grade machine for those services is, from an economical point of view, not a viable solution. On the other hand it facilitates the administration of the different services, if each server runs on a separate operating system installation.

In addition some of the Grid middleware is closely tied to the operating system. Therefore the Tier-1/2/3 centers are sometimes forced to install single machines with an operating system which is different form the local standard. Most sites use a hardware setup which was tested with their operating system of choice in mind.

This can cause interoperability problems between the local hardware and the operating system required by some Grid middleware services.
2. Virtualization

With virtualization it is possible to install more than one virtual machine on a single real machine. It is therefore possible to keep every Grid middleware service in a separate operating system installation but run them on a small number of virtualization hosts.

Virtual machines provide an abstraction layer on top of the real hardware. This abstraction layer does not suffer from the frequent changes in the hardware world. This abstraction layer typically simulates hardware which is already well supported by different operating systems. It is therefore possible to install old versions of operating systems on virtual machines which are installed on new hardware. This way on can disentangle the purchase of new hardware from the requirements of old operating system versions.

2.1. Types of Virtualization

In general Virtualization is the technique to run an operating system on a physical machine with a layer in between to decouple the hardware form the OS. There are several approaches which include:

- **full virtualization**: The host system generates a virtual environment in which the guest operating system can not determine whether it is executed on a real or a virtual machine. As a result any operating system can run unchanged on such a virtual machine. [7]

- **para virtualization**: The host operating system generates a virtual environment in which only some parts of a real machine are simulated. The remaining parts can be controlled by a special interface. This requires adaptations in the guest OS but is generally faster than full virtualization.[?] 

- **hardware-assisted virtualization**: The command set of modern CPUs contain special commands for fast context switching and memory separation. This helps the host OS to increase the performance of the guest operating systems.

Several software solutions exits which implement one ore more virtualization techniques. The most relevant ones are XEN[1], KVM[2], VirtualBox[4] and VMware[3].

3. Installation at Rechenzentrum Garching (RZG)

When we started our virtualization effort, KVM and VirtualBox were not mature enough. VMware is proprietary software but XEN is included in the Linux distribution installed on the host system. As a result we choose XEN as virtualization software. To keep the performance loss minimal we are using para virtualization. Since we have full access to the source code of the operating system on the guest machines, para virtualization does not create problems.

In our current setup, we are using two server grade machines as XEN hosts:

- **Host-1**: 8 CPU cores, 12GB RAM, Ubuntu 7.10
- **Host-2**: 4 CPU cores, 8GB RAM, Ubuntu 7.10

These two physical machines host 5 virtual machines:

- **Compute Element**: 3 CPUs, 2GB RAM, Scientific Linux 4.6 64bit
- **Storage Element**: 3 CPUs, 6GB RAM, Scientific Linux 4.6 64bit
- **Site BDII**: 1 CPU, 2GB RAM, Scientific Linux 4.6 32bit
- **Monitoring Box**: 1 CPU, 2GB RAM, Scientific Linux 4.6 32bit
- **Nagios Box**: 1 CPU, 2GB RAM, Scientific Linux 4.6 32bit
All virtual machines are configured to use a flat file on the host system as their virtual disk drive. The Compute Element and the Storage Element are both located on Host-1. The Site BDII, the Monitoring Box and the Nagios Box are installed on Host-2. The remaining CPUs are used for test machines and test installations. And some CPU power is required by the host operating system as well.

As shown in figure 2 and figure 3 the Compute Element and the Storage Element produce a significant load on their respective virtual machine. The other virtual machines are nearly idle most of the time.

**Figure 1.** Number of jobs administered by the Compute element at RZG

**Figure 2.** Load average of the virtualized machine containing the Compute Element (same weeks as in figure 1)

**Figure 3.** Load average of the virtualized machine containing the Storage Element (same weeks as in figure 1)
4. Lessons learned
The installation described above is in operation for several months already. So far our initial assumptions were confirmed and we found further advantages of this setup.

As expected, the two physical machines are able to host five services in independent virtual machines. This way we were able to reduce the costs for hardware and operation. We were also able to install an operating system on new hardware without the need to adjust the operating system to the hardware since Scientific Linux includes an optional kernel capable to run as a XEN guest.

Virtualized machines are more flexible than physical ones. It is relatively easy to add CPUs or memory to a virtual machine, as long as one stays within the limits of the host machine.

It is also very simple to backup a virtual machine. Since the whole filesystem of a virtual machine is stored as a single file on the host system, a copy of that file constitutes a backup.

Updates of the Grid middleware sometimes lead to situations in which one wants to roll back to the setup just before the update. This can also be solved by copying the file containing the virtual disk prior to the update. If the update creates problems, the rollback can be done by using the copy of the virtual disk. The virtual disk which contains the update can be deleted.

Since the virtual disk images are not bound to any particular hardware, one can easily migrate a virtual machine from one host machine to another host. This feature can be used if one host machine fails. But this is also useful if one purchases new physical machines and want to migrate some virtual machines to the new hardware.

So far our worker nodes are not virtualized and we have no plans to do so. Since many user jobs are CPU-intensive, the performance penalty by virtualization outweighs the ease of administration.

5. Summary and Outlook
The Grid services at the ATLAS Tier-2 center at Rechenzentrum Garching are installed on virtual machines. This facilitates maintenance and administration. The performance of virtual machines is sufficient for the server nodes. With the increasing support for virtualization in hardware, virtualized worker nodes might become an option in the future.

6. References
[1] http://www.xen.org/
[2] http://www.linux-kvm.org/
[3] http://www.vmware.com/
[4] http://www.virtualbox.org/
[5] http://en.wikipedia.org/wiki/Full_virtualization
[6] http://en.wikipedia.org/wiki/Paravirtualization
[7] http://en.wikipedia.org/wiki/Hardware-assisted_virtualization