To: OSG Council
From: Sebastien Goasguen, Jerome Lauret
Date: November 5th 2010

Memo on the feasibility of running virtual machines on OSG sites using existing OSG job submission mechanisms.

Background:
Over Summer 2010 Clemson CI research group worked intensely with the STAR VO to make use of the cloud capabilities on the Clemson Palmetto Cluster. This resulted in one of the largest PYTHIA run for the STAR VO. The Clemson/STAR collaboration is on-going partially through the OSG Satellite project EXTENCI. The Clemson Cluster which is available to OSG was used but the OSG mechanisms were bypassed. This memo reports on new tests carried on to assess the feasibility of using regular CE mechanisms to start VM on OSG sites.

Setup:
The Clemson cluster is a regular centos based linux cluster with ~15,000 cores. PBS/Maui is used, but Condor backills the cluster. An OSG CE uses the condor job-manager to access Palmetto cycles. The KVM modules were installed via regular rpm updates. KVM is fast becoming the de-facto virtualization technology since RedHat announced its full support. The big advantage is that VMs can be started by regular unix users and that VMs appear has regular unix processes. In that sense a VM can be seen as a regular job running on the cluster. A KVM unix group was created to manage authorization to KVM. OSG VO accounts for STAR and Engage were added to this KVM group. To avoid any issues with networking, each virtual machine gets its own NAT, no public addresses are used. This has the main drawback that VMs started on Palmetto do not have any inbound network connectivity. The main advantage is that it is no headache for the IT networking group.

Tests:
In addition to the summer 2010 tests that have been reported at SciDAC and CHEP by Jerome Lauret, we carried out straightforward tests early this month.
Using the Engage job submission framework created by Mats Rynge, we created a simple job which sole purpose was to start a VM using the KVM binary. The VM image itself was treated as input data to the job. The Engage framework worked seamlessly. The job executable (a bash script) and the VM image were tarred up and zip, then transferred to the Clemson CE (osg-gw.clemson.edu). There the condor job manager found a slot to start the VM. The VM has now been running for 48 hours, and these hours will be reported to Gratia. This process is the exact Engage job submission process.

Caveats:
- In our tests, the VM is fairly small (i.e ~250MB), the STAR and CMS VM (being worked on via the Extenci project are several GB). Data staging processes will need to be used to make the images available at sites. But these processes should not be different from any current data staging processes.

- Clemson does not advertize its KVM capability via CEMon, so the Clemson site was hardcoded in the Engage submission scripts. This can be easily solved. Ideally however, agreement will need to be reached on how to advertize virtualization capabilities from sites and this should be reported via CEMon.
- As mentioned in our setup, there is no in-bound network connectivity to the started VMs. This is where techniques such as glide-ins and the Clemson Kestrel system come into play. There is no doubt however that some sites will offer a different network setup for the VMs, and that running jobs within them could be achieved for instance by joining the local condor pool.
- Obviously we trust our own VMs, but certainly site admins may not be inclined to trust everyone's VMs. We would like to point to the HEPiX working group on virtualization. This group is developing policies to trust virtual machine images among VOs and sites.

Recommendations:
Based on these tests it is our strong belief that there is nothing currently preventing OSG sites from running virtual machines.

- Sites should be recommended to setup KVM on their cluster and gain technical knowledge about it.
- We need to agree on how to publish virtualization capabilities via CEmon, once this is done Engage and STAR could seamlessly start running on VMs across OSG using the trusted job submission mechanisms.
- The OSG security group should work with the HEPiX Virtualization working group to participate int the discussions on policy for trusted virtual machines. If anything, it is the software being created for managing trusted VMs that will need to be put in VDT.