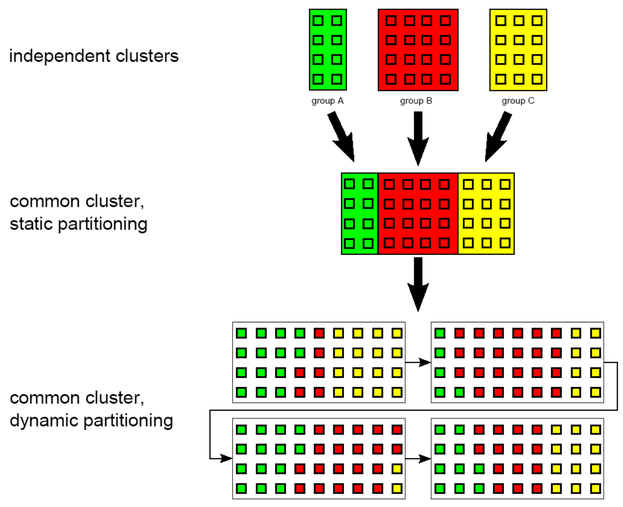
Dynamic Virtualization of Worker Nodes

Poster Session, Mid-Term Review 30.11&1.12.2009

V. Büge, H. Hessling , Y. Kemp, B. Klein, P. Krauß, M. Kunze, O. Oberst, G. Quast, A. Scheurer and Owen Synge



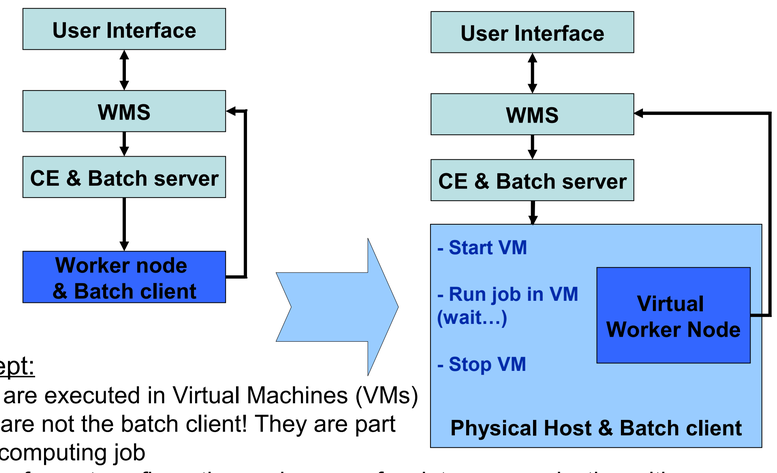
Goals
- Share a computing resource between different user groups
- Provide multiple computing environments
- Isolate group partitions

Possible solutions:
- Common computing cluster, static paritioned
  - No load-balancing between user group partitions
- Common computing cluster, dynamic partitioned
  - Load-balancing possible

  → Provides opportunistic use of the computing resources

Current Implementation:
- Same concept two implementations
  - vmimagemanager at DESY
  - vBatch at KIT

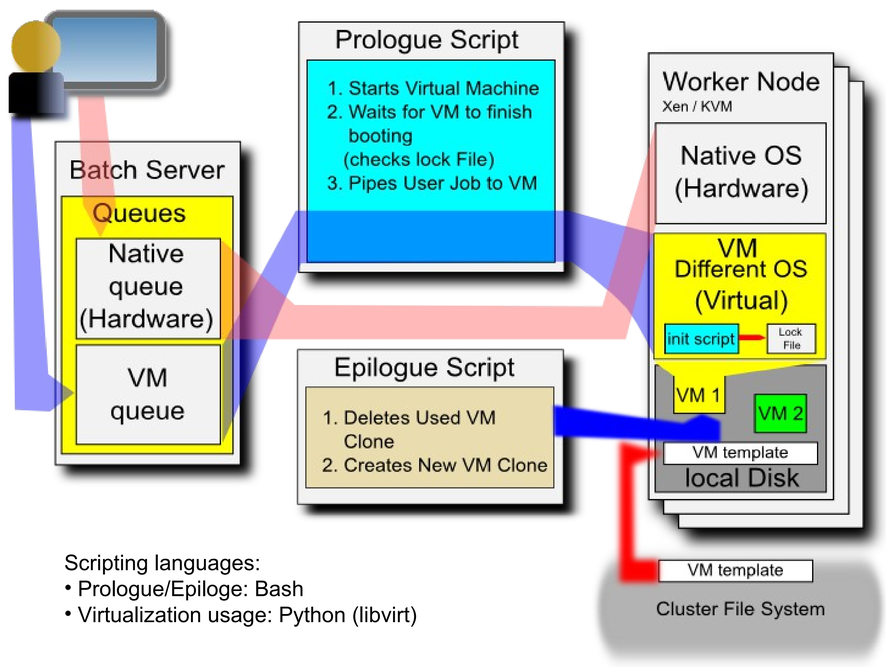
Several postdocs, PhD and diploma students are involved



Concept:
- Jobs are executed in Virtual Machines (VMs)
- VMs are not the batch client! They are part of the computing job
- Issue of smart configuration and usage of scripts communicating with virtualization layer
  - → Independent from used batch system
  - → Independent of used virtualization solution by using libvirt API
- Virtualization techniques hidden from the user

The Testbed: IC1 Cluster at SCC

- Shared between nine different KIT department
  - 200 compute nodes
  - 2 x Intel Quadcore Xeon
  - 17.5 TFlop peak performance
  - Lustre cluster file system
  - 70 TB home space
  - 350 TB work space
- Two user group partitions:
  - HPC partition: parallel computing (MPI)
  - HEP (High Energy Physics) Partition: Seriell computing, High Throughput Computing (HTC)
- Operating Systems:
  - Suse Linux Enterprise 10.0Sp2 on the hardware machines
  - Scientific Linux 5 in virtualized worker nodes
- Virtualization technique: Kernel Virtual Machine (KVM)
- Batch system: MAUI/Torque (PBS)

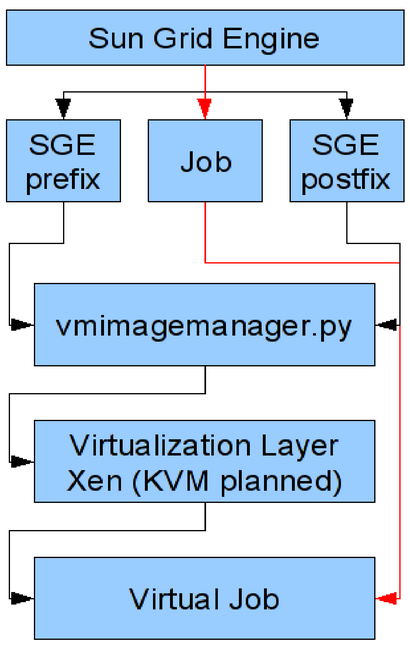


Scripting languages:
- Prologue/Epiloge: Bash
- Virtualization usage: Python (libvirt)

Karlsruhe Institute of Technology

Realization of the virtualization system:

- Wrapper script (Prologue/Epilogue) around the actual computing job inside the batch system
  - Virtual machines are deployed on the hardware node disc
  - Wrapper script starts VM and checks status of the VM
  - Computing job is piped to the VM via ssh
  - Job runs on VM
  - After job execution VM is destroyed and image is deleted
  - New images are prepared for following jobs

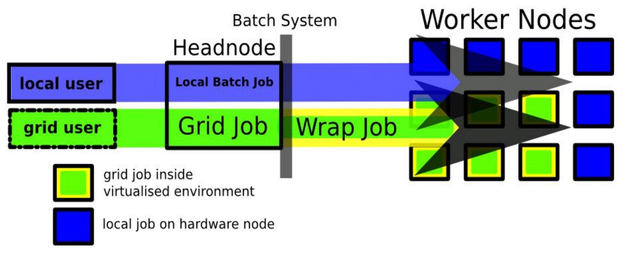


DESY Solution:

- Main goal: Bringing virtualization to Grid/gLite users
- Using the Sun Grid Engine (SGE) as back-end
  - SGE support by the gLite Grid middleware
- lcg-CE (CE) as Grid front-end
- Operating Systems:
  - Scientific Linux 5.1 on the hardware machines
    - No middleware installed, runs SGE execd and shepherd
  - SL 4.7 in virtualized worker nodes
    - Has middleware installed, but no batch client

Implementation Details:

- Short prologue, epilogue and started scripts
  - Basically run vmimagemanager
  - Execution of job payload via ssh on VM
- Proof of principle implementation: 1 CE, 1 SGE master, 1 host, 2 VMs
  - Stress test with > 100 jobs submitted simultaneously via WMS successful!
- No changes to CE, minor SGE configuration changes



Performance Considerations

- Typicall High Energy Physics (HEP) analyses showed a near native performance (3-6%) per core for Xen and KVM (with „virtio“ paravirtualization support).

Educational usage of Virtualization

- Virtualization tutorials at GridKA School 2007, 2008 and 2009
- GridKA School computing environment partially virtualized to provide school service and hands-on machines

Conclusion

- Dynamic Virtualization:
  - Opportunistic use of computing resources
  - Avoids limitation of the compute environment (Operation System, architecture, worker node setup)
- Status:
  - Development and testing at KIT and DESY

HELMHOLTZ ASSOCIATION

DESY
KIT - Institut für Experimentelle Kernphysik
KIT - Steinbuch Centre for Computing
External Partner: HTW Berlin

Publications:

- V. Büge, H. Hessling, Y. Kemp, M. Kunze, O. Oberst, G. Quast, A. Scheurer, and O. Synge *Integration of Virtualized Worker Nodes in Standard-Batch-Systems,* CHEP09 Proceedings – to be published
- V. Büge, *Virtualization of Grid Resources and Prospects of the Measurement of Z Boson Production in Association with Jets at the LHC*, PhD Thesis – IEKP-KA/2008-18
- B. Klein, *Application of Virtualization Techniques to Grid Resources and Reconstruction of Heavy Resonances Decaying to Quark and Gluon Pairs with the CMS Detector at the LHC,* Diploma Thesis – IEKP-KA/2008-23

Ongoing Theses: O. Oberst (PhD Thesis), P. Klein (Diploma Thesis)