Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Running different operating systems on the same worker nodes

Institut für Wissenschaftliches Rechnen
Forschungszentrum Karlsruhe

Institut für Experimentelle Kernphysik
Universität Karlsruhe

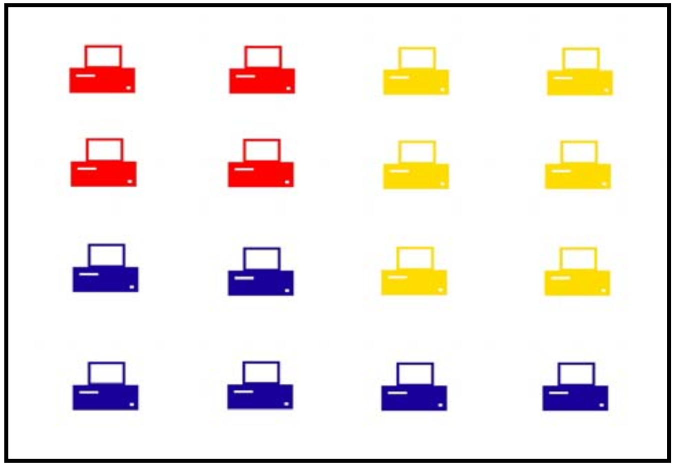
Volker Büge, Yves Kemp, Marcel Kunze,
Oliver Oberst, Günter Quast

Virtualisation of Batch Queues

Basic Ideas:

- Different groups at the same computing centre need different Operating Systems
- Agreement on one OS or no resource sharing
- Virtualisation allows to dynamically partition a cluster with different OS
- Each queue is linked to one type of Virtual Machine

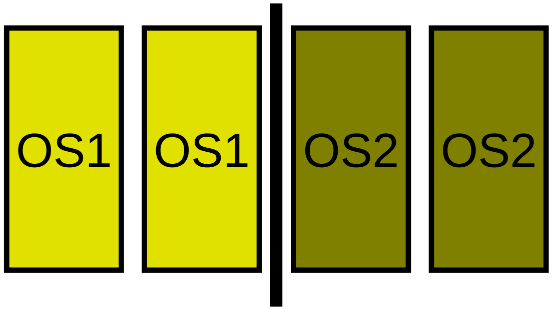
→ Such an approach offers all advantages of a normal batch system combined with the free choice of the OS for the computing centre administration and user groups!



Dynamic Partitioning of a Cluster I

Batch system with 2 different OS required by user groups

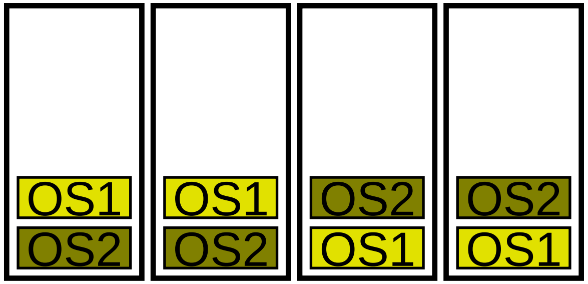
Static batch system:



- No resource sharing possible
- Static partition of the cluster

  → Changing partitioning difficult

Virtualised batch system:

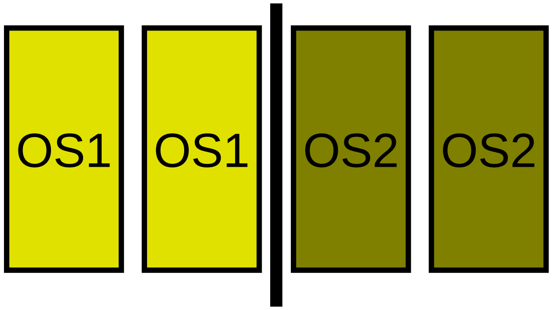


- Each physical Worker Node hosts different VMs
- Memory of unused VMs is lowered to a minimum or the VM is stopped

Dynamic Partitioning of a Cluster II

Batch system with 2 different OS required by user groups

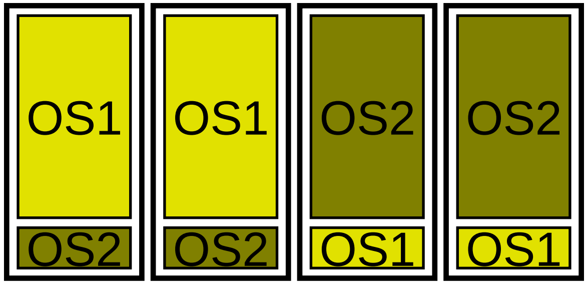
Static batch system:



- No resource sharing possible
- Static partition of the cluster

  → Changing partitioning difficult

Virtualised batch system:



- When a VM on a free WN accepts a job, its memory is increased instantaneous or VM will be started
- Dynamic partitioning of a cluster possible on short timescales

Problem: Different user groups

Typical structure of a High Energy Physics institute, e.g. the IEKP

Group specific requirements:

- CMS
  - Experiment specific software and grid middleware gLite require SLC 3.0.X
- CDF
  - SL Fermi 3.0.X recommended for experiment specific software and grid middleware
  - Also SLC 3.0.X possible
- AMS
  - Can easily recompile their software on different platforms

Current situation:

Compromise possible:

SLC 3.0.6 32bit can be used for all groups …

… but AMS software would benefit from 64bit OS!

The near future:

Diverging needs:

- e.g.: CMS SLC4, CDF SLC3
- e.g.: CMS needs both SLC3 and SLC4
- e.g.: Some need 32bit, other 64bit.
- Sharing with other groups using modern distributions

- Additionally: Security issues!
  - Different user groups, same cluster

Performance Considerations

No noticeable performance loss due to virtualisation:

- Only about 3-4% loss for CMS software application (32 bit) in a XEN VM.

Even performance gain is possible:

- Galprop (AMS main application) runs 22% faster in a virtual 64-bit machine (XEN) than on 32-bit native system! (Same Opteron hardware)

→ An overall performance gain can be possible, at least no drastic performance losses.

EKP - Prototype – Implementation

The batch system:

- MAUI scheduler
- Torque batch server

The cluster:

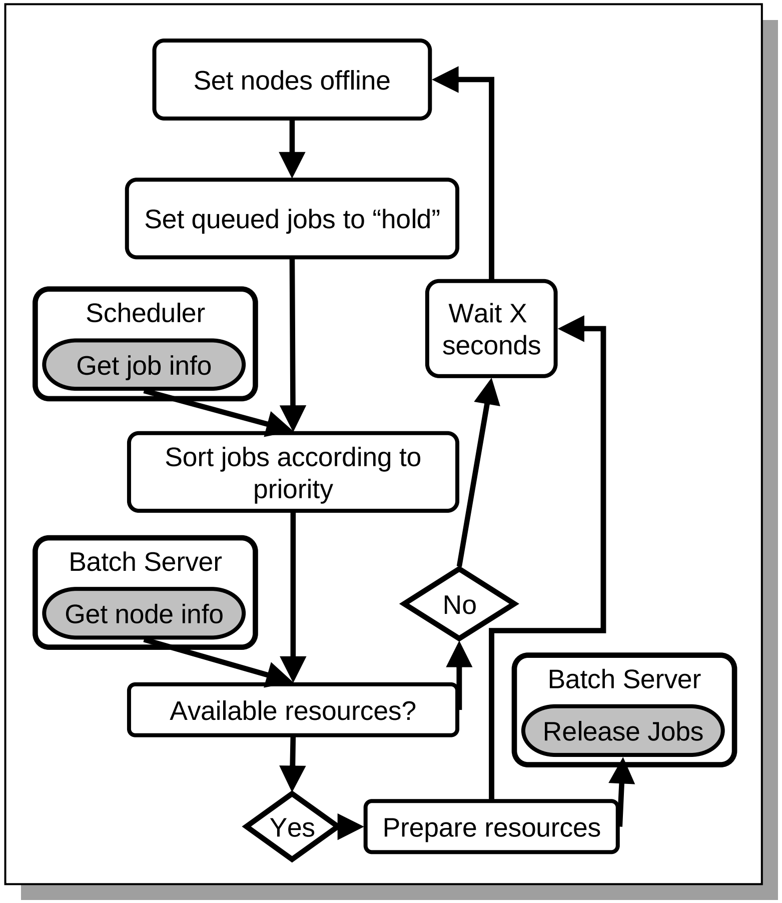
- 30 heterogeneous Worker Nodes
- 2 different OS (32 and 64bit)

Dynamic partitioning is working in our production environment

Paper presented at XHPC06

Virtualizing a Batch Queuing System at a University Grid Center

V. Büge, Y. Kemp, M. Kunze, O. Oberst, G. Quast



Conclusion & Outlook

- Many benefits from Virtualisation of batch systems:
  - Allows to offer optimal OS to each user group
  - Security and privacy through encapsulation
  - Easy deployment and test of new OS without downtimes
- EKP - Prototype – Implementation:
  - Additional daemon works as proof of principle on our production cluster
  - About 500 user jobs executed on different OS (32 and 64 bit)
- Wish list
  - Better to implement such functionality directly in the batch system server and the scheduler
  - Propose a Hardware Mom to organise the usage of VMs on the physical Worker Node
  - Similar to Magrathea by Jirí Denemark et al.

Batch system virtualisation allows an easy setup of a shared cluster