Fachhochschule für Technik
und Wirtschaft Berlin
University of Applied Sciences

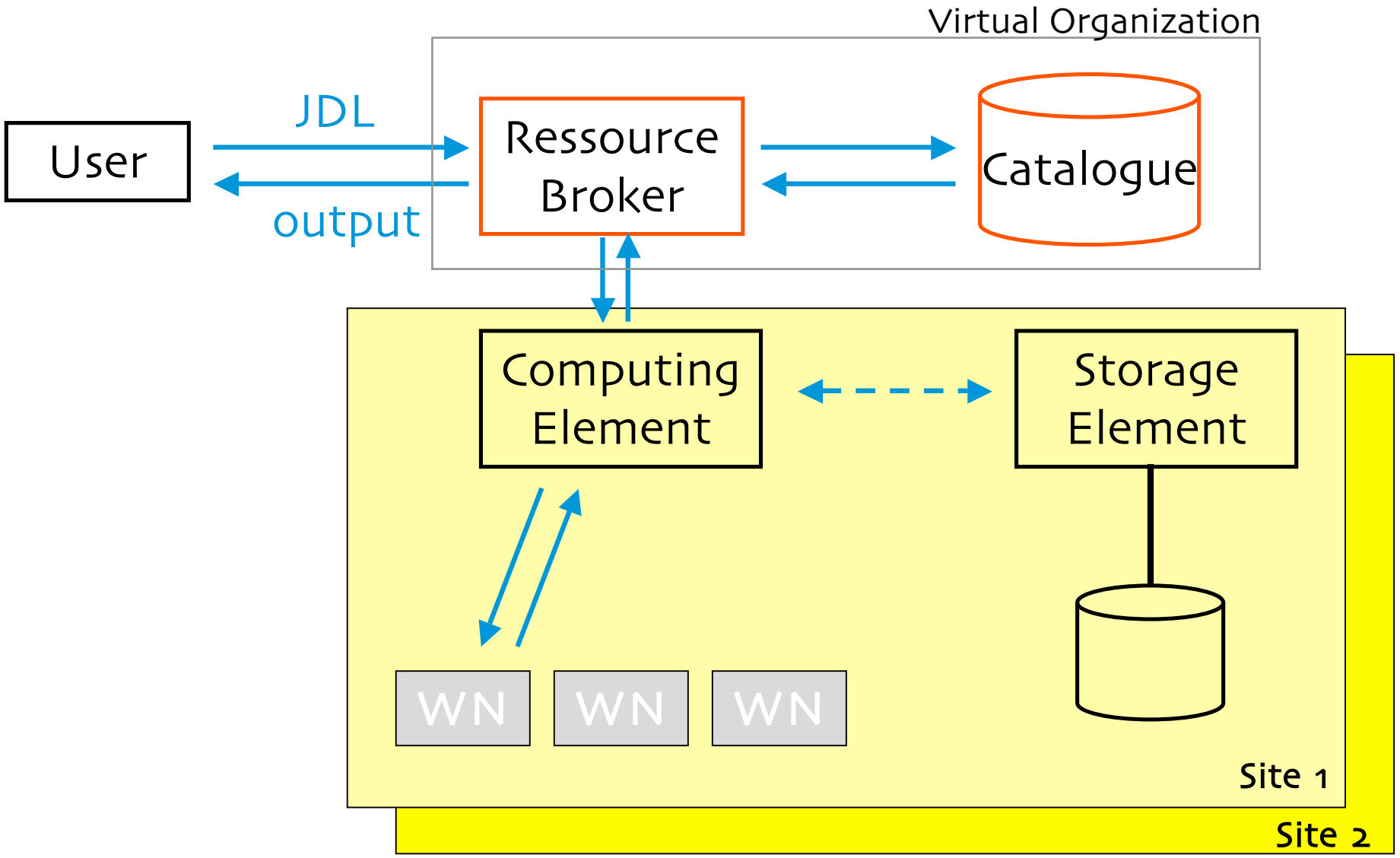
Virtualization of Worker Nodes in the Grid

Hermann Heßling

26. 5. 2008
DESY Computing Seminar
University of Hamburg

Content

1. Grid Infrastructure



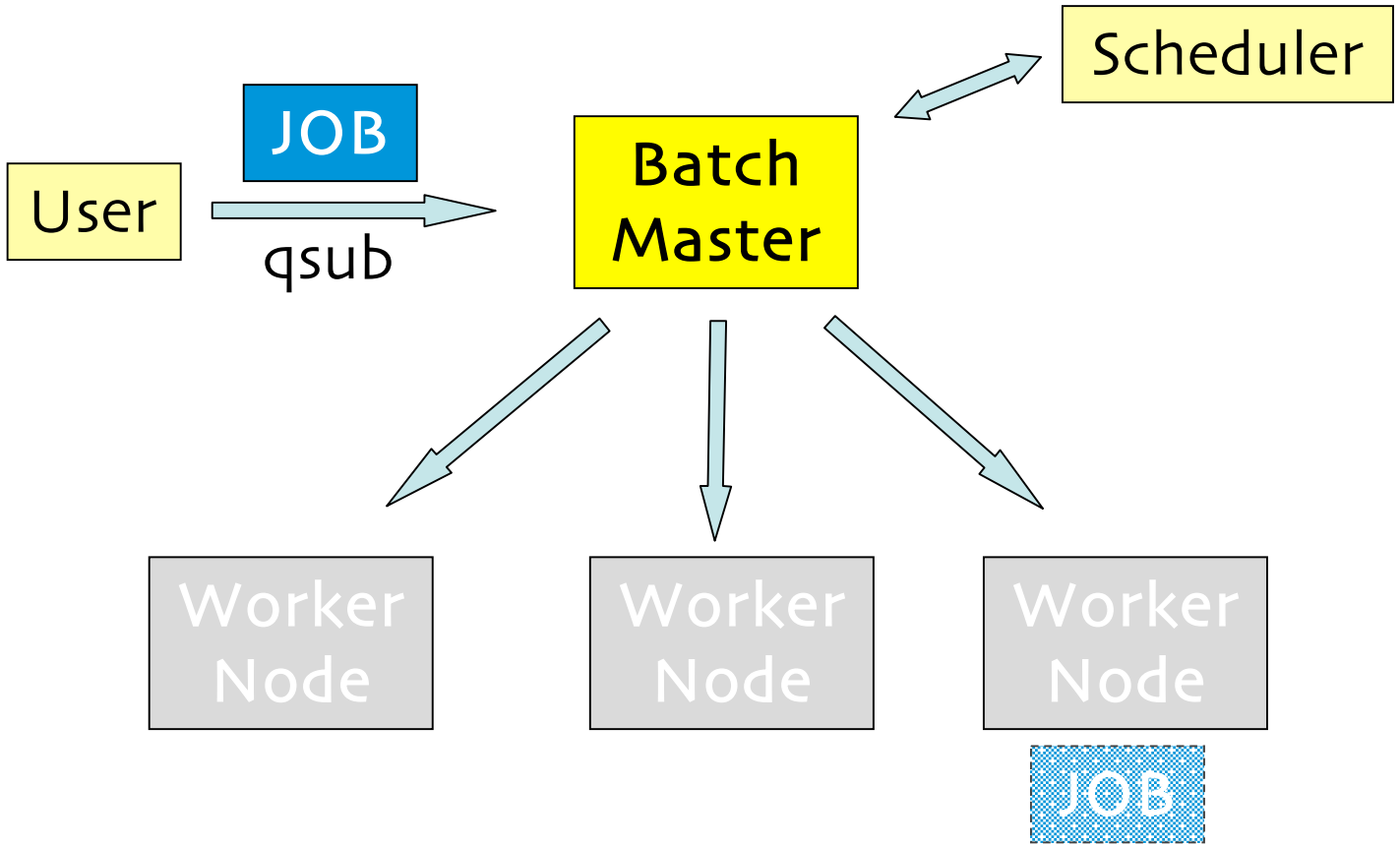
1. Grid Infrastructure

Batch Queueing

- PBS/Torque + Maui
  - Used at DESY to run jobs from LCG (LHC Grid)
- Sun Grid Engine (SGE)
  - NAF/Desy, Zeuthen
- ...

1. Grid Infrastructure

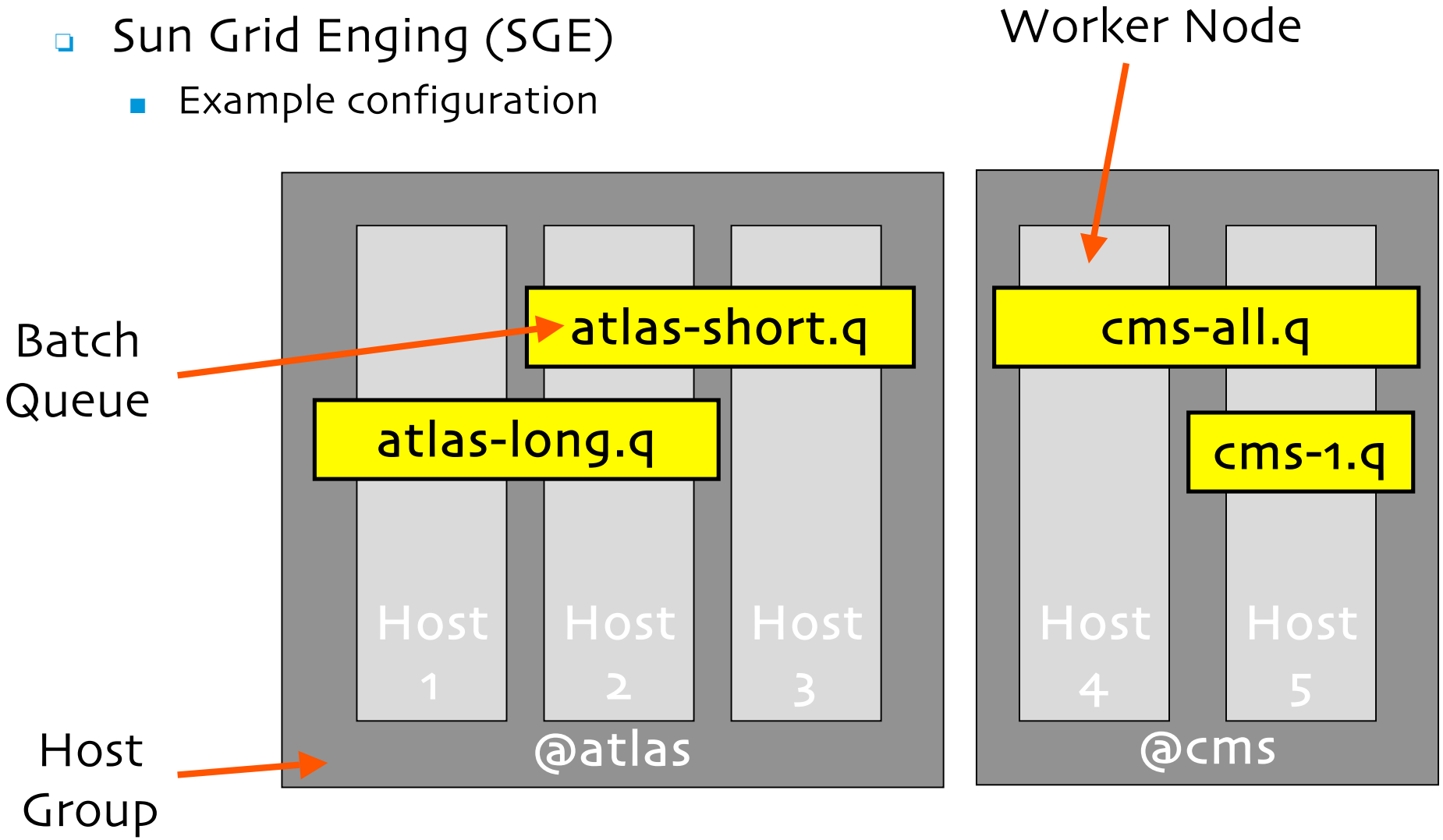
Batch Queueing



1. Grid Infrastructure

Batch Queueing

- Sun Grid Enging (SGE)
  - Example configuration



Content

2. Virtualization

Products (selection)

Xen™

The Xen™ virtual machine monitor

vmware®
AN EMC COMPANY

Microsoft
Microsoft Virtual Server

parallels

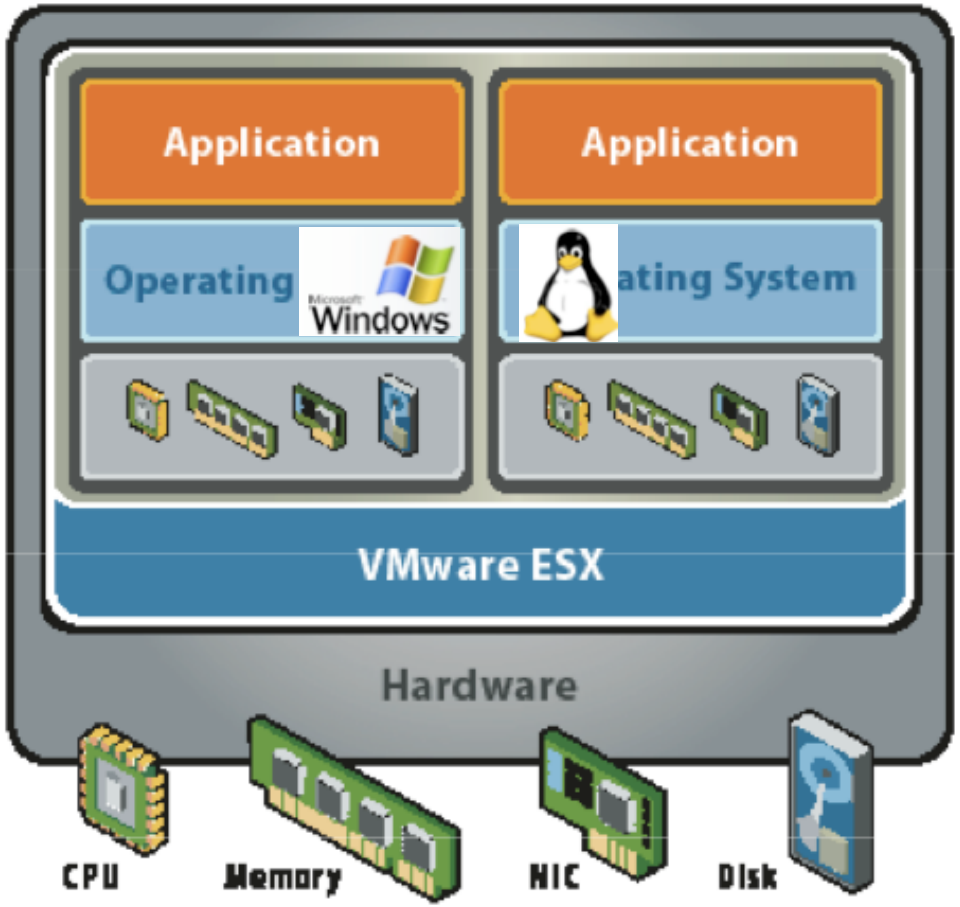
Linux VServer

OpenVZ

Win4Lin
A Product of Virtual Bridges

2. Virtualization

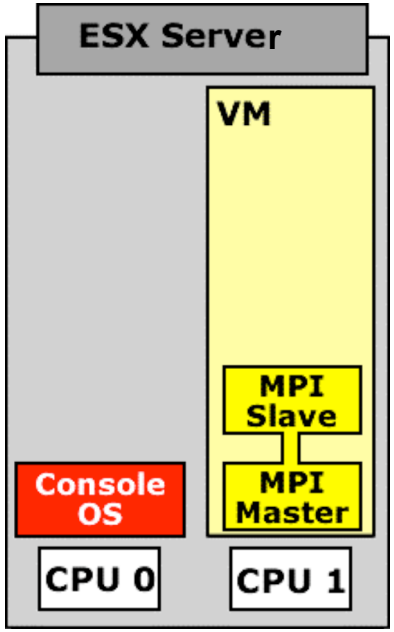
VMWare ESX 3.0



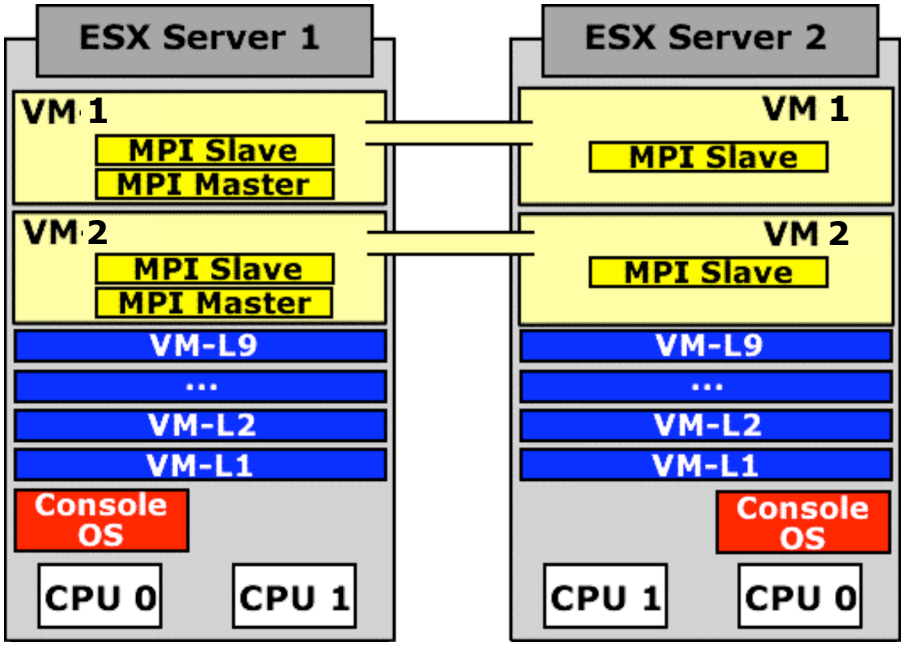
- Inserts a virtualization layer between the hardware and the operating system.
- Each virtual machine represents a complete sytem (processors, memory, networking, storage …).
- Able to run and manage almost any x86 OS.

2. Virtualization

VMWare ESX 2.5



One virtual machine (VM) running exclusively on CPU 1.



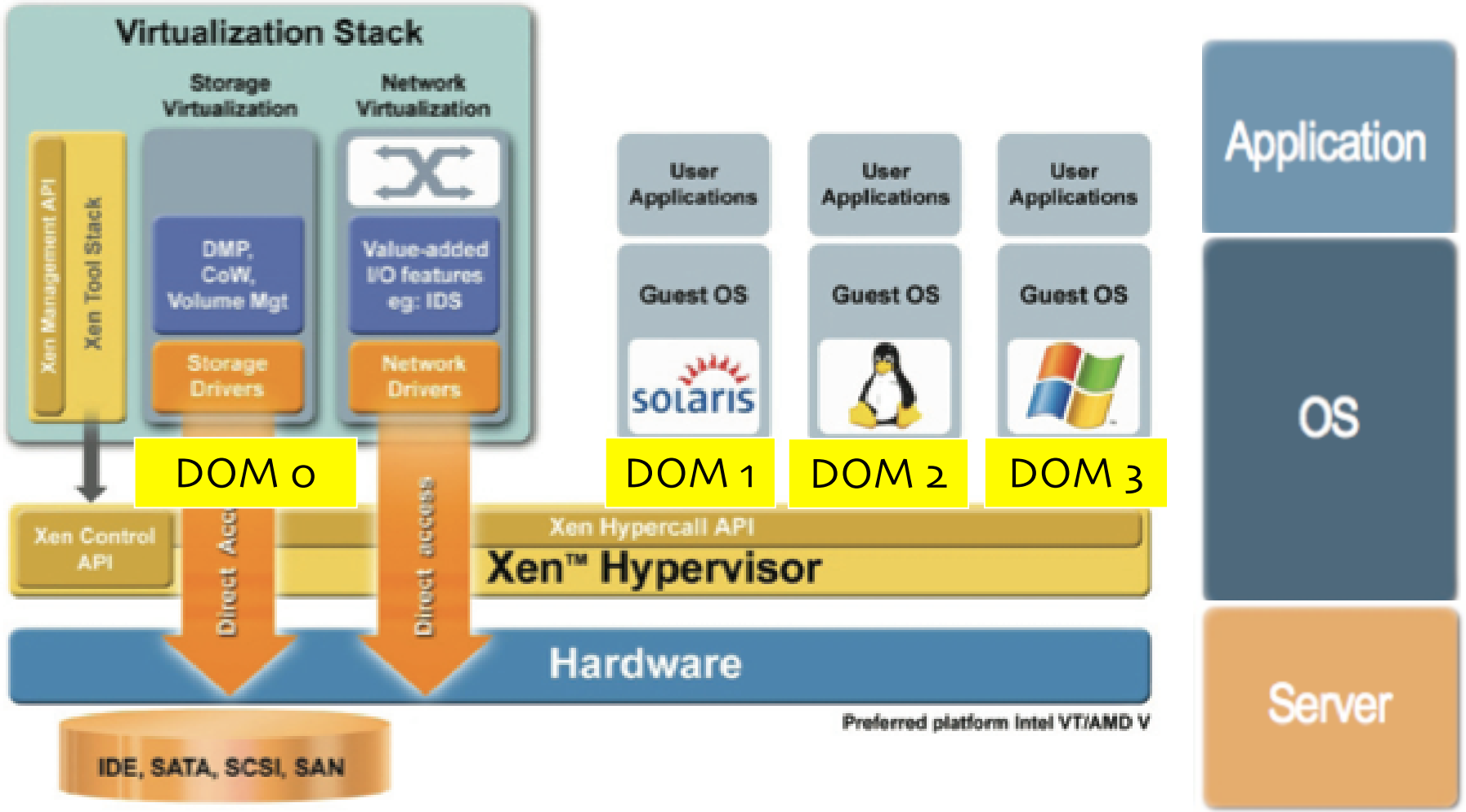
Two VMs distributed over two servers.

Per server several VMs are distributed over both CPUs.

For each VM a fraction of the CPUs can be reserved.

2. Virtualization

Citrix Xen 3.0

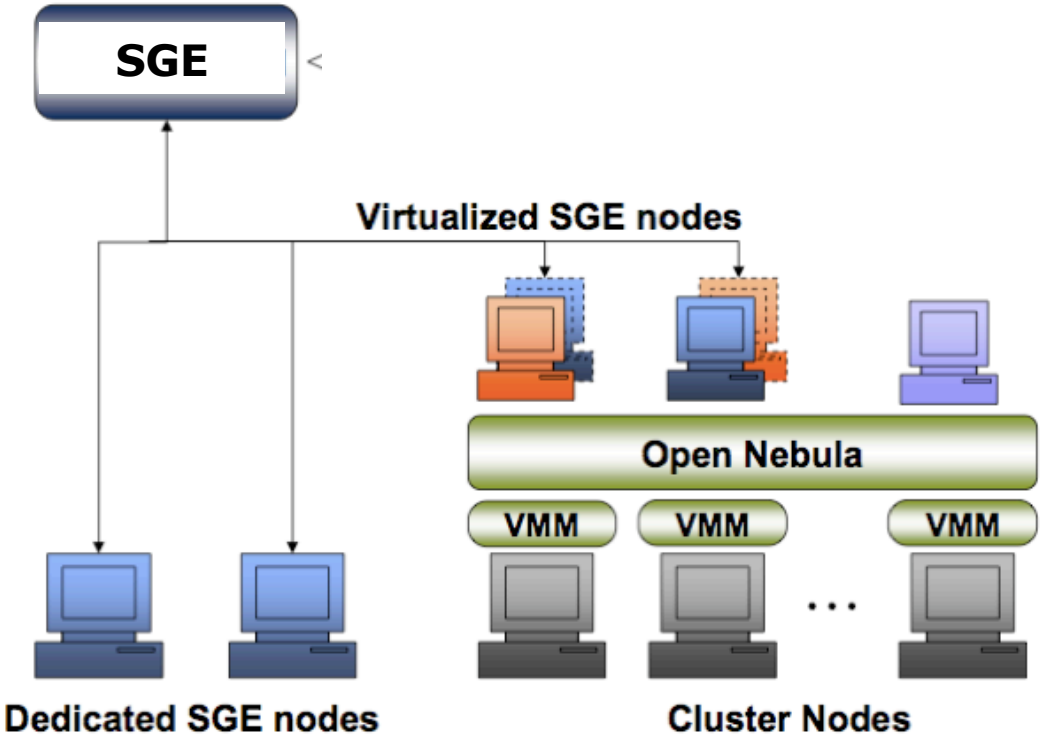


Content

3. Virtualization in Batch Systems

Open NEbula (Univ. of Madrid)

- Dynamic deployment and re-allocation of virtual machines (⇒ Xen) on a pool of physical resources by introducing a distributed virtualization layer
- Consolidation: multiple worker nodes in a single physical host

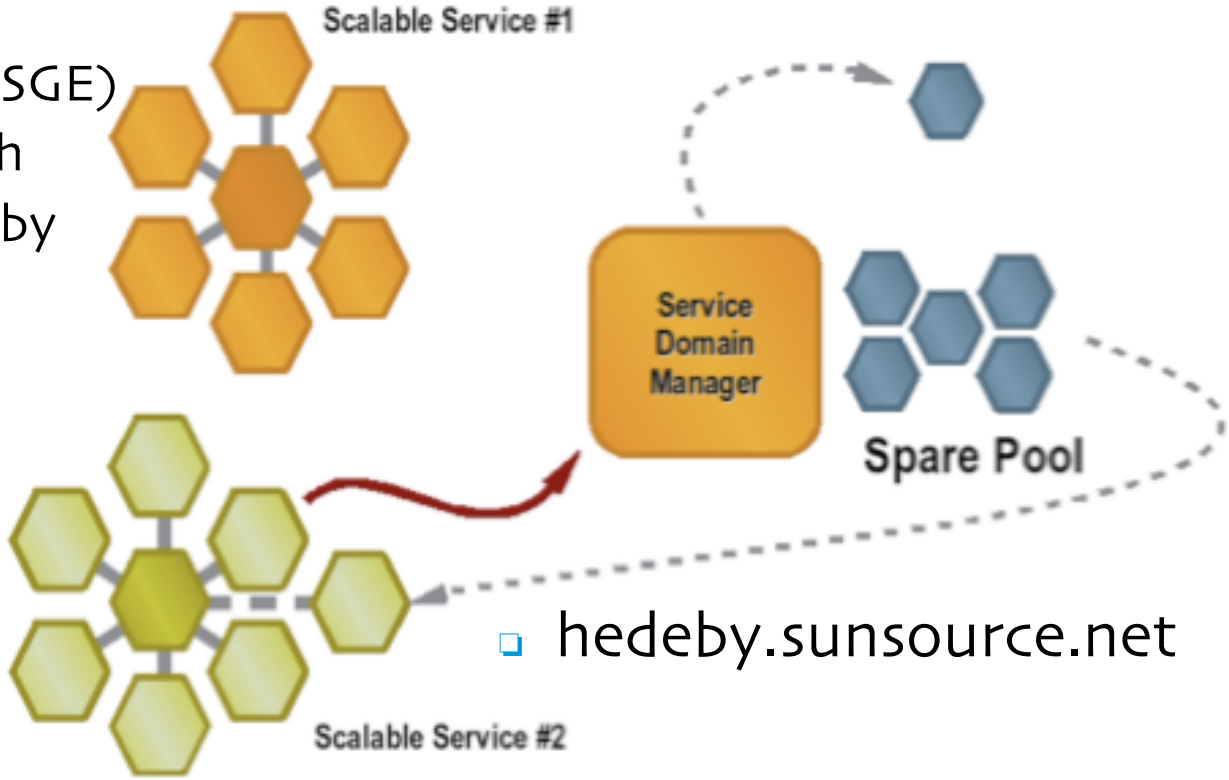


- On-demand provisioning of resources
  - SGE knows the virtual nodes
- OpenNEbula part of new EU project RESERVOIR (17 M €)
  - Deployment and management of IT services across different administrative domains, IT platforms and geographies
  - Cloud Computing (IBM)
    - shared infrastructures
  - IBM, SAP, Telefonica, …

3. Virtualization in Batch Systems

Hedeby (german: Haithabu, largest Nordic city during the Viking Age)

- Service management system
  - If a service needs more resources, Hedeby is able to provide them.
  - Resources are shifted/shared between services on-demand.
    - Spare pool: Hedeby is able to switches nodes on/off for power saving.
- Services
  - Sun Grid Engine (SGE)
  - Any software with interface to Hedeby
    - Database, …
- Resources
  - Hosts
  - CPUs
  - Memory
  - OS

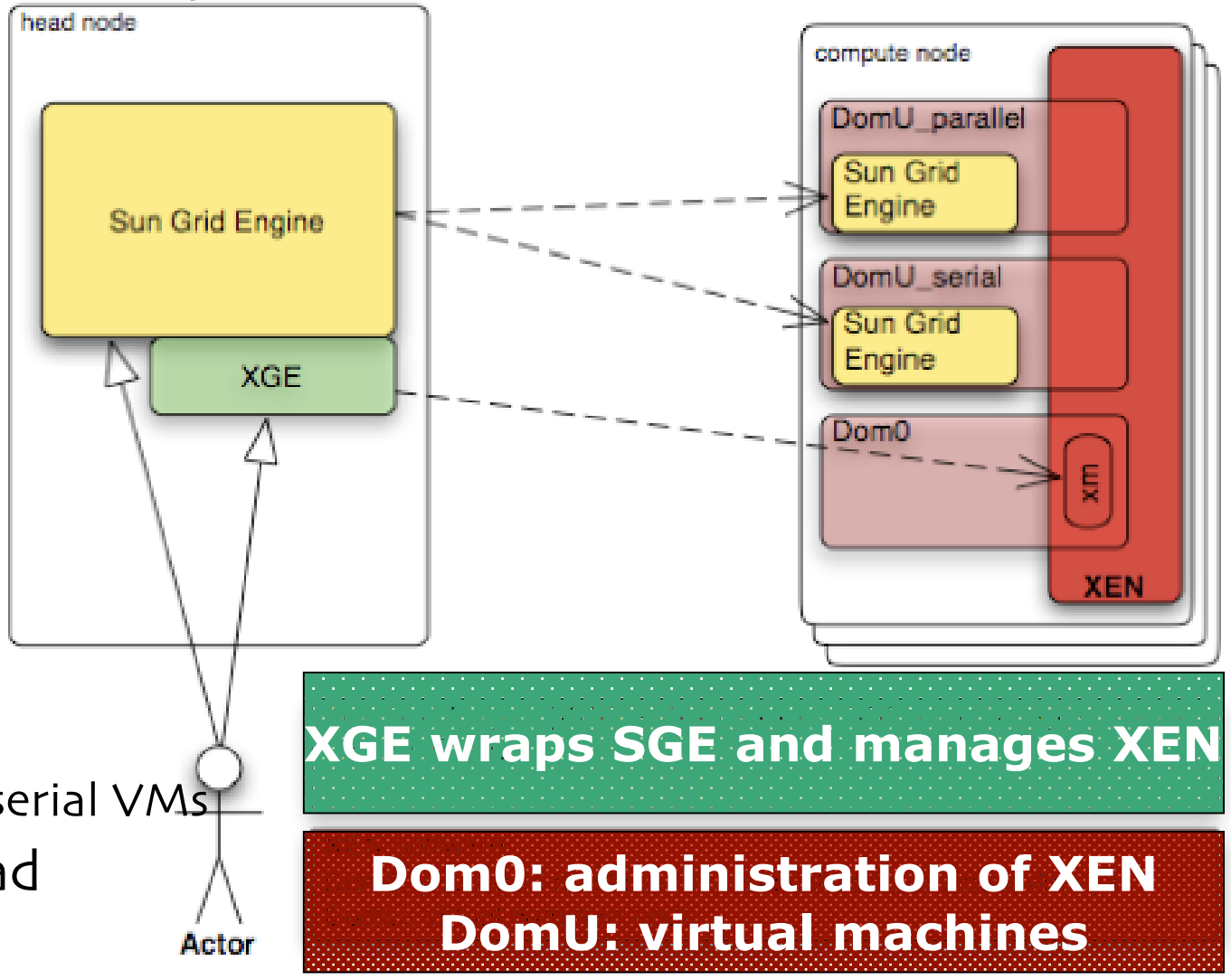


- hedeby.sunsource.net

3. Virtualization in Batch Systems

XGE (Univ. of Marburg)

- Execution of short parallel and serial jobs in parallel
- SGE's qsub is wrapped
- dynamic virtual cluster partitions
  - supending serial VMs
- XGE overhead negligible



Content

4. Virtualization in Grid Computing

Using VM-based worker nodes in Grid Computing

- Evercome static partition of resources
  - nodes with flexible OS (e. g. different Scientific Linux for EGEE)
- Enhance security
  - isolate users from each other and from the hardware
- Flexible customization
  - Individual environments for different classes of users

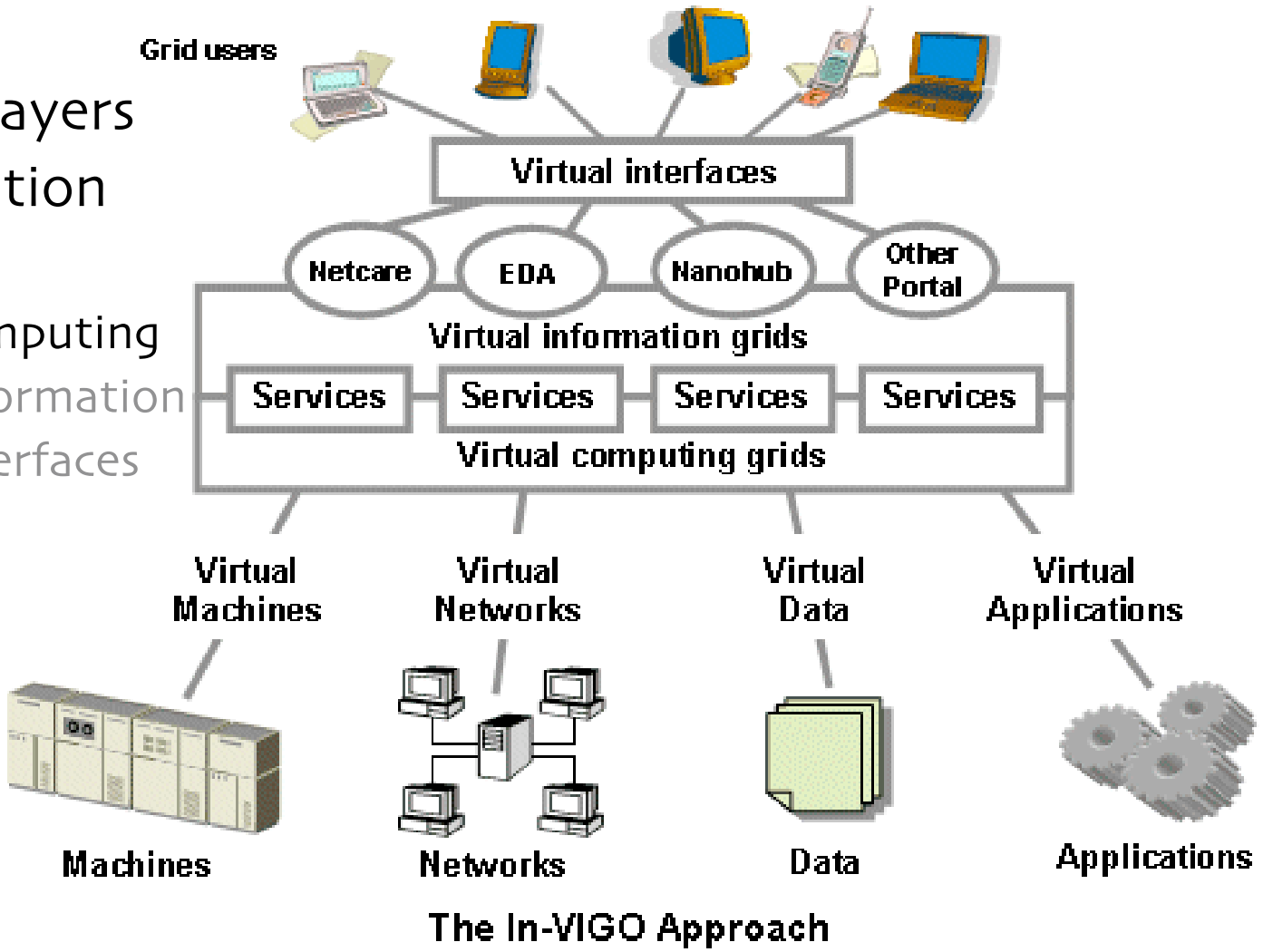
General Problems

- Integrate low level VM mechanisms into existing Grid architectures
- Convince users and **site admins** to accept VMs as the primary execution environment

4. Virtualization in Grid Computing

In-VIGO

- Adds three layers of virtualization to the Grid
  - Virtual computing
  - Virtual information
  - Virtual interfaces



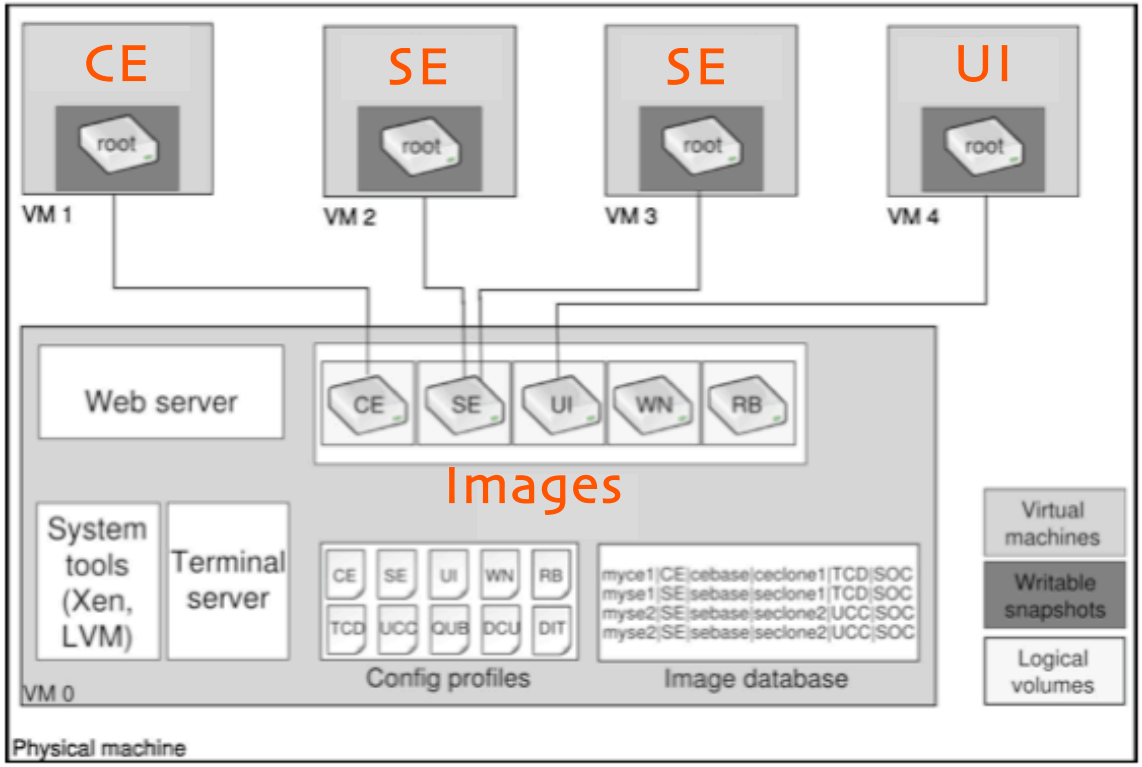
The In-VIGO Approach

Fortes, et al., 2003

4. Virtualization in Grid Computing

Grid-Ireland

- Grid-Builder
  - Creating and managing XEN virtual machines
  - Managing the images
  - Automatic configuration of the Grid middleware
  - Local Grid services running in VMs in one physical machine



Childs et al., 2007

- Central Management of 17 Grid sites in Ireland
  - Local **site admins manage** only their **worker nodes**
  - Devolopers have common testbeds

4. Virtualization in Grid Computing

vGrid

- Virtualization in gLite certification
  - Integrating software of different Middle-ware providers
  - Testing & certifying of several patches at the same time
- Portal to manage a XEN virtualized cluster
  - Deploying/terminating VMs
  - Generation / storing / managing VM images
- Unterkircher et al. (2007)
  - http://vgrid.web.cern.ch

4. Virtualization in Grid Computing

Virtualization users workshop

16-17 January 2007 DESY

- Main use cases for virtualization
  - Server consolidation
    - Already an effective use case for virtualization
  - Deployment testing
    - The work of several communities (Ireland, CERN, and DESY) relies already strongly on virtualization
  - Virtualization of worker nodes
    - Better integration in the batch systems and the grid environment needed
- General statements
  - High availability computing will use VM technologies
  - Any worker node will be a VM in the next 5 years

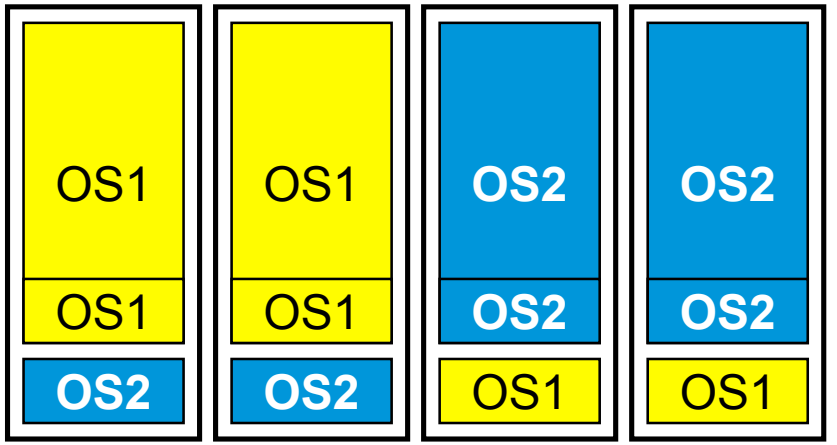
Content

5. Virtual Worker Nodes in the Grid

Virtualized Worker Nodes (Karlsruhe)

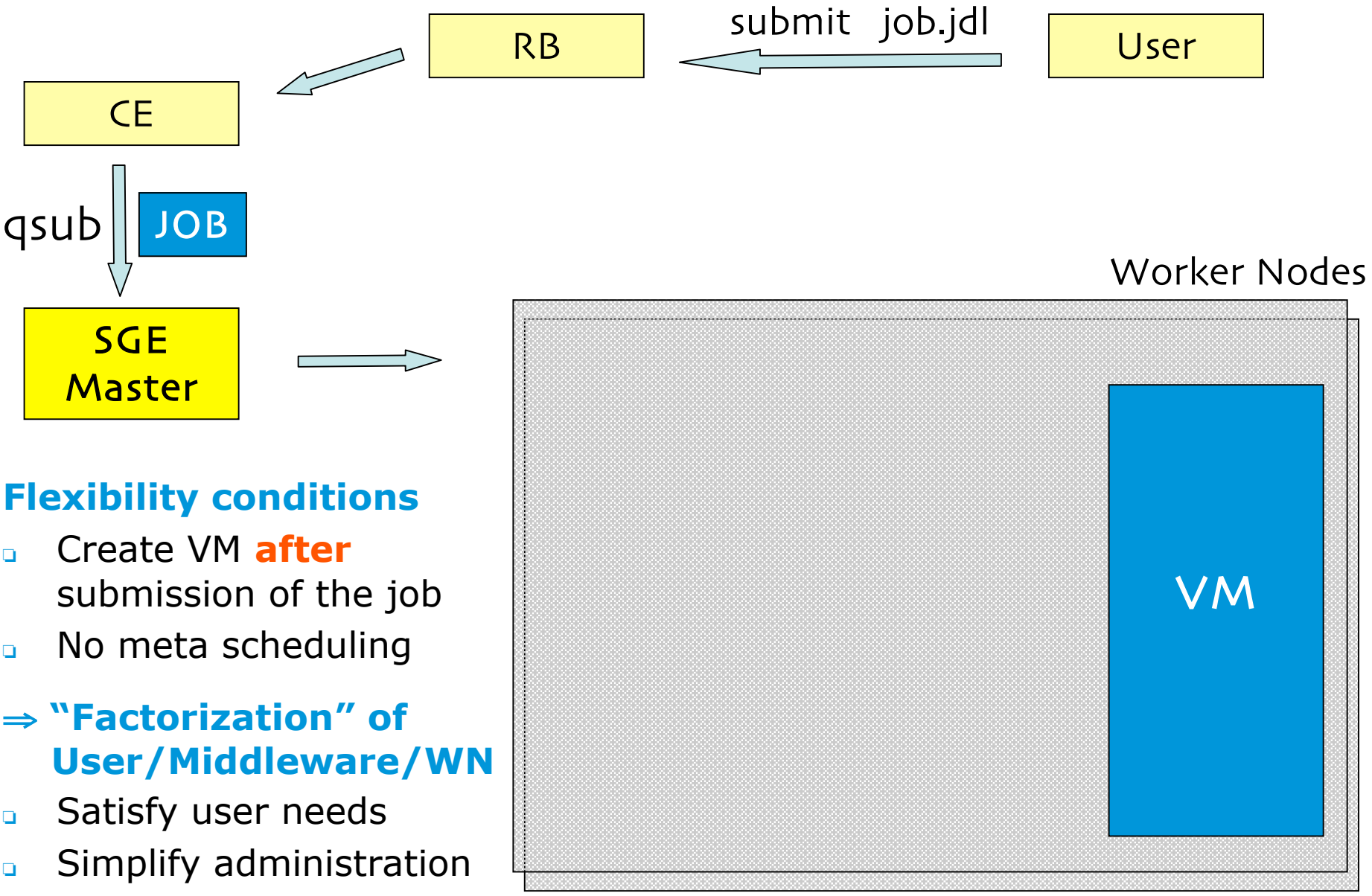
- Integration of virtual worker nodes (XEN) in the Grid with in-house means
  - no change noticeable to user
  - no changes of batch system
- Integration into batch queueing systems Maui/Torque by a daemon
  - Meta scheduling problem: daemon leads to a reimplementation of parts of the batch system
- Similar solution: Magrathea

Büge, Kemp et al., XHPC 06



- Non-active domains are only provided with minimal resources (CPU, memory)
- An active VM is provided with "all" physical resources

5. Virtual Worker Nodes in the Grid



Flexibility conditions

- Create VM **after** submission of the job
- No meta scheduling

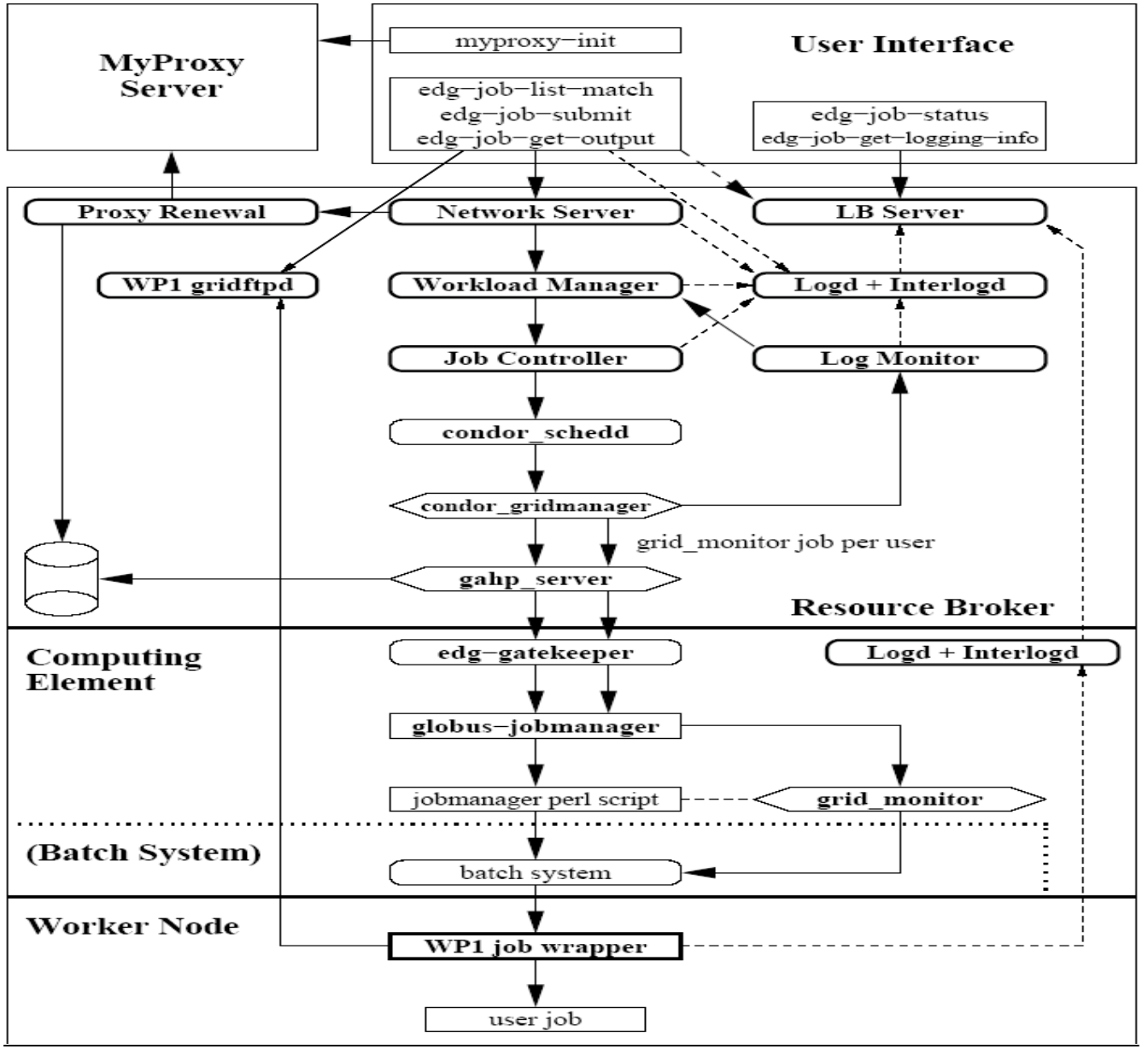
⇒ **"Factorization" of User/Middleware/WN**

- Satisfy user needs
- Simplify administration

5. Virtual Worker Nodes in the Grid

Simone Campana (CERN IT) "What gets submitted to the WN is a perl script

- Sets up a bit of environment
- Fetches via gridftp a tarball from the CE
- …
- Stuff … stuff … stuff …
- Runs the Job Wrapper"



5. Virtual Worker Nodes in the Grid

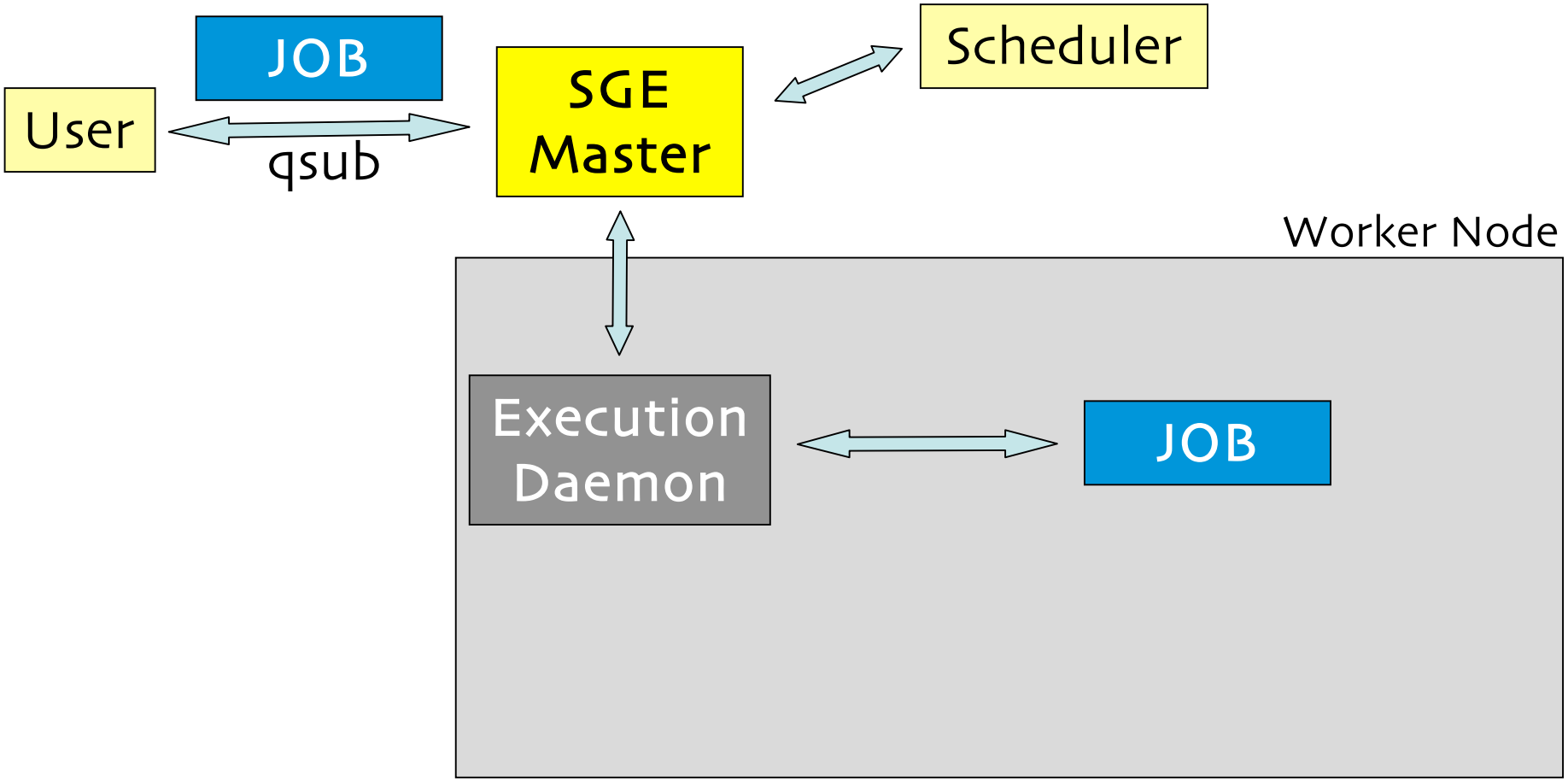
```
USER     PID   PPID  TIME     COMMAND

root     17417 16184 00:01:09 sge_shepherd-87 -bg
43733    17420 17417 00:01:09 -sh /usr/local/sge/pro/default/spool/tb014/job_scripts/87
43733    17444 17420 00:01:09 /usr/bin/perl -w /tmp/bootstrap.t17441 /home/desyusr033/ grid-ce2.desy.de
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/x509_up X509SELF /dev/null
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/stdout stdoutftp /home/desyusr033/.globus/job/grid-
ce2.desy.de/31481.1207259801/stderr stderrftp /home/desyusr033/.lcgjm/globus-cache-export.N31758 https://grid-
ce2.desy.de:20001/31481/1207259801/ /home/desyusr033/ NONE /home/desyusr033/.globus/.gass_cache/local/md5/68/
98f6cf61daf9753e34ebe30d71fc46/md5/68/6c166e13e7044da8197c5f73d6b43e/data UI=000003:NS=0000000003:
WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17448 17444 00:01:09 /usr/bin/perl -w /tmp/bootstrap.t17441 /home/desyusr033/ grid-ce2.desy.de
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/x509_up X509SELF /dev/null
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/stdout stdoutftp /home/desyusr033/.globus/job/grid-
ce2.desy.de/31481.1207259801/stderr stderrftp /home/desyusr033/.lcgjm/globus-cache-export.N31758 https://grid-
ce2.desy.de:20001/31481/1207259801/ /home/desyusr033/ NONE /home/desyusr033/.globus/.gass_cache/local/md5/68/
98f6cf61daf9753e34ebe30d71fc46/md5/68/6c166e13e7044da8197c5f73d6b43e/data UI=000003:NS=0000000003:WM=000004:
BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17533 17444 00:01:10 /usr/bin/perl -w /tmp/bootstrap.t17441 /home/desyusr033/ grid-ce2.desy.de
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/x509_up X509SELF /dev/null
/home/desyusr033/.globus/job/grid-ce2.desy.de/31481.1207259801/stdout stdoutftp /home/desyusr033/.globus/job/grid-
ce2.desy.de/31481.1207259801/stderr stderrftp /home/desyusr033/.lcgjm/globus-cache-export.N31758 https://grid-
ce2.desy.de:20001/31481/1207259801/ /home/desyusr033/ NONE /home/desyusr033/.globus/.gass_cache/local/md5/68/
98f6cf61daf9753e34ebe30d71fc46/md5/68/6c166e13e7044da8197c5f73d6b43e/data UI=000003:NS=0000000003:
WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17551 17444 00:01:10 sh -c if [ -x ${LCG_LOCATION:-/opt/lcg}/libexec/jobwrapper ]; then ${LCG_LOCATION:-
/opt/lcg}/libexec/jobwrapper /home/desyusr033/globus-tmp.tb014.17444.0/globus-tmp.tb014.17444.2
UI=000003:NS=0000000003:WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000; else /home/desyusr033/
globus-tmp.tb014.17444.0/globus-tmp.tb014.17444.2 UI=000003:NS=0000000003:WM=000004:BH=0000000000:JSS=000003
:LM=000000:LRMS=000000:APP=000000; fi
43733    17552 17551 00:01:10 /bin/sh /opt/lcg/libexec/jobwrapper /home/desyusr033/globus-tmp.tb014.17444.0/globus-
tmp.tb014.17444.2 UI=000003:NS=0000000003:WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17553 17552 00:01:10 bash /home/desyusr033/globus-tmp.tb014.17444.0/globus-tmp.tb014.17444.2 UI=000003:
NS=0000000003:WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17599 17553 00:01:11 bash /home/desyusr033/globus-tmp.tb014.17444.0/globus-tmp.tb014.17444.2 UI=000003:
NS=0000000003:WM=000004:BH=0000000000:JSS=000003:LM=000000:LRMS=000000:APP=000000
43733    17600 17599 00:01:11 /bin/sh ./my_sleep.sh
```

5. Virtual Worker Nodes in the Grid

Sun Grid Engine (SGE)

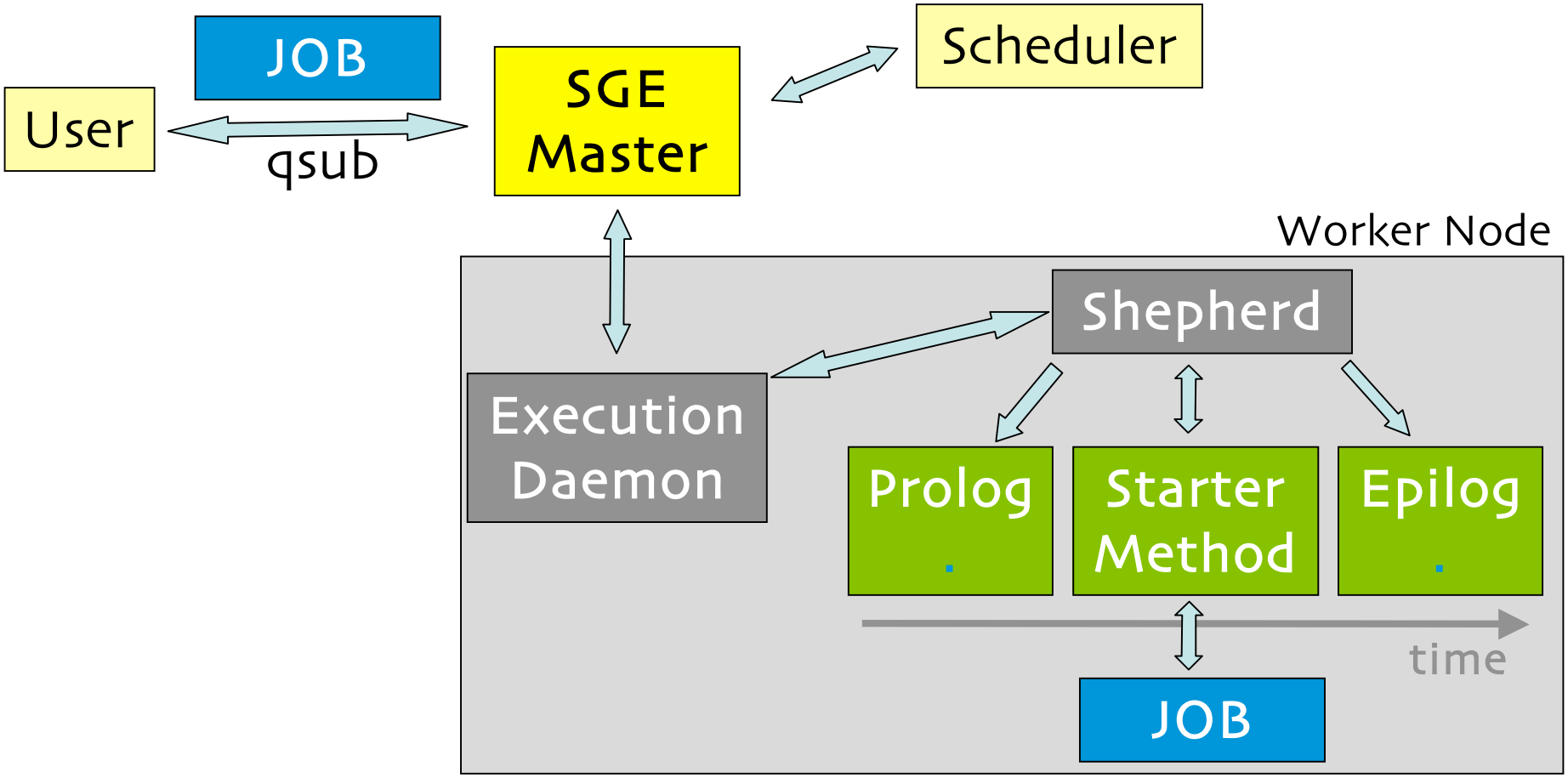
- Running a job



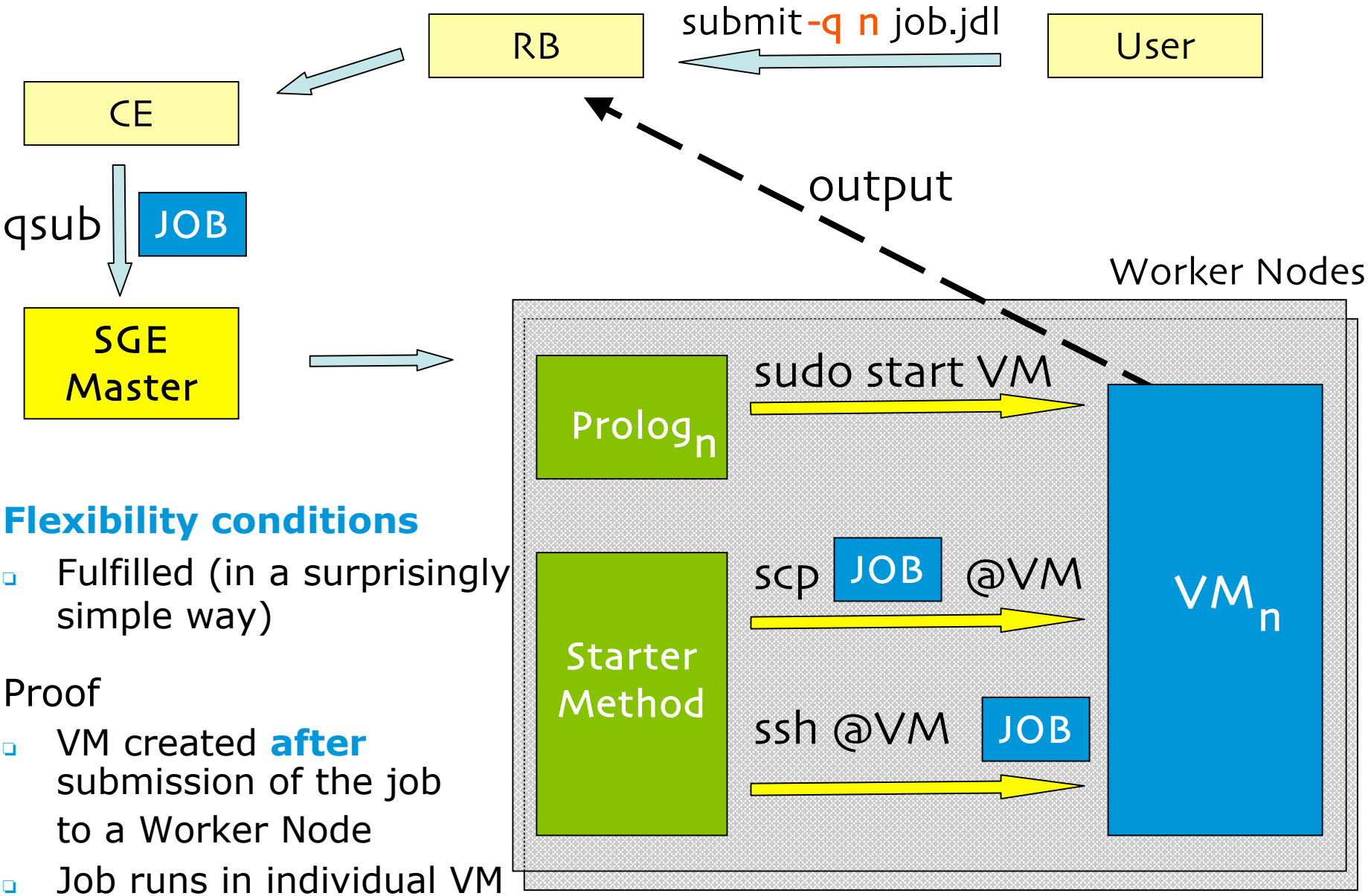
5. Virtual Worker Nodes in the Grid

Sun Grid Engine (SGE)

- Running a job: a closer look



5. Virtual Worker Nodes in the Grid



Flexibility conditions

- Fulfilled (in a surprisingly simple way)

Proof

- VM created **after** submission of the job to a Worker Node
- Job runs in individual VM

5. Virtual Worker Nodes in the Grid

vLite

- Middleware
  - gLite: no modifications
- Batch Queueing System
  - Prolog/epilog scripts: start/stop of a queue-specific VM
  - Starter method script: copy a job to the VM and start it there
  - No meta scheduling
    - Jobs submitted to physical worker nodes
    - Rules per physical worker node
      - #jobs ≤ #cores
      - Any job runs exclusively in a Virtual Machine
- Virtualization: VMImageManager
  - Managing a cluster of XEN virtual machines on a single host
  - Snapshot and image deployment
    - Reinstalling of an image takes ~ 45 - 90 seconds
  - Small and simple python script (developed by Owen Synge)

Content

6. Summary and Outlook

Current status of vLite

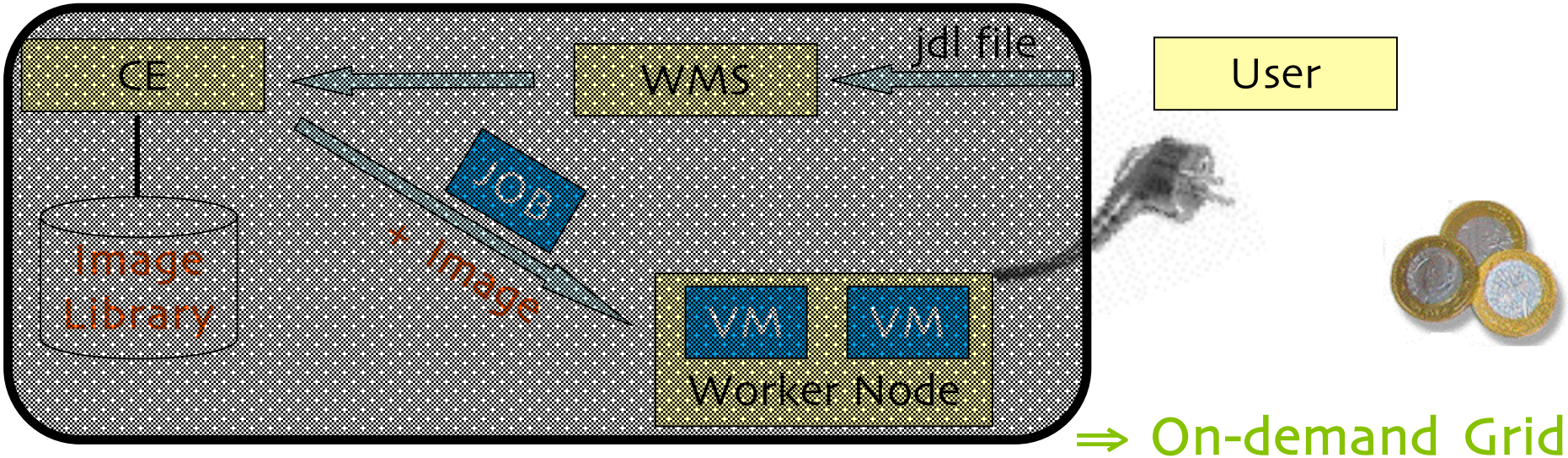
- Proof of concept successful

Future work

- Experience with production-like conditions
  - Simplify vLite installation (⇒ tar ball)
- Support of parallel jobs
  - SGE offers script interfaces for parallel computing
- VMImageManager
  - LVM (Logical Volume Manager) snap shotting
    - Copy-on-write snapshots of LVM volumes in Linux 2.6.8
  - Support of KVM (Kernel-based Virtual Machine)
    - Hardware-based virtualization: KVM driver in Linux 2.6.20

6. Summary and Outlook

- **CE Integration:** CREAM-CE (successor of LCG-CE)
  - 3 additional parameters (GLUE Schema 2)
    - OSFLAVOR — SL3, SL4, SL5, SL4(H1), ...
    - ARCHITECTURE — x86-32, x86-64
    - IMAGE_LFN — user specific image + location



⇒ On-demand Grid

Many thanks to the physics computing team at Desy IT
- especially to Yves Kemp and Owen Synge (⇒ vLite)