Automated workflow for an SGE environment Grid Computing at DESY

Andreas Haupt
<andreas.haupt@desy.de>

Agenda

- Short introduction to DESY
- Grid Computing at DESY
- Automated SGE installation & configuration
- Our attempt of a tight qrsh/ssh integration

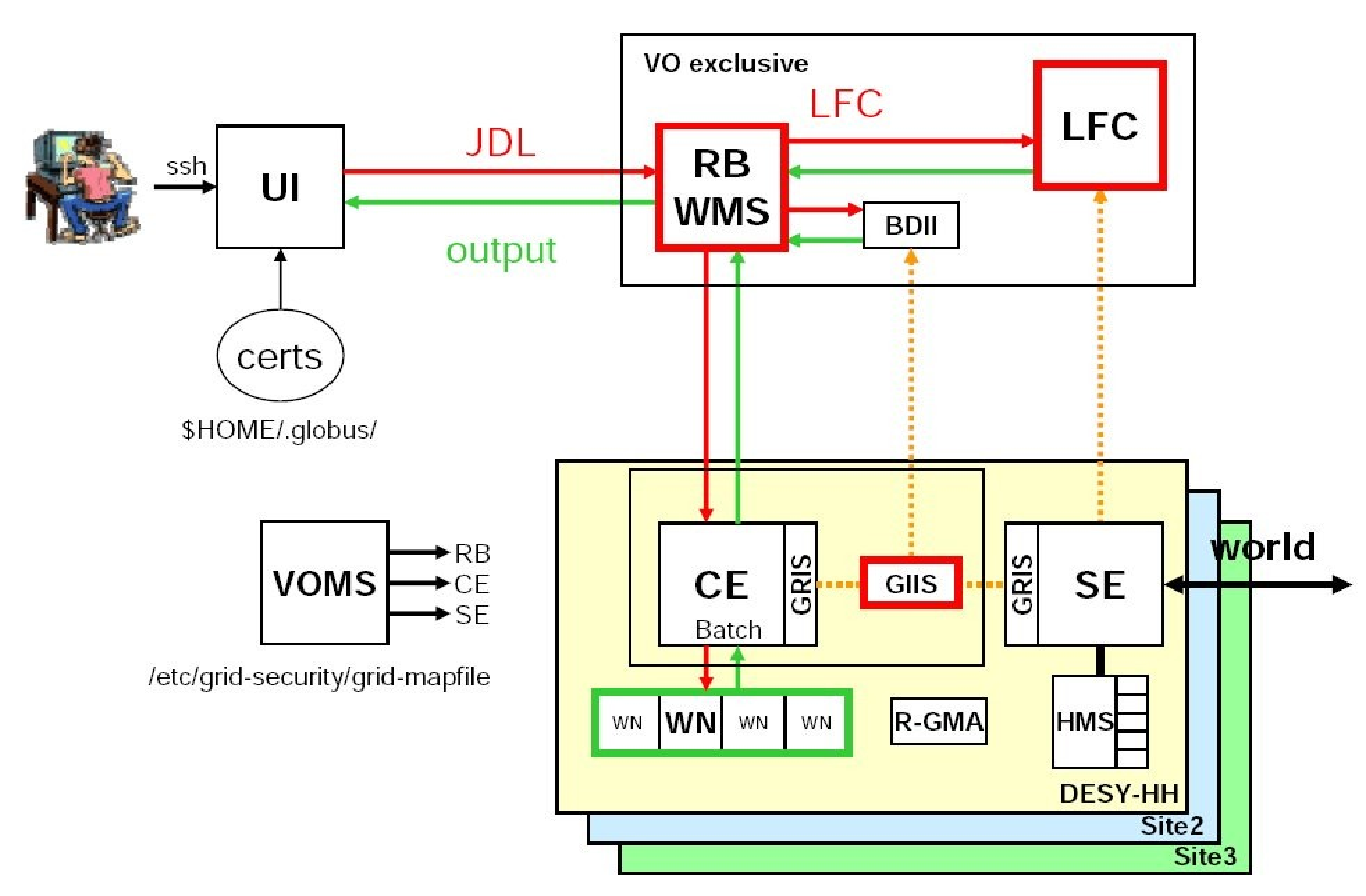
DESY at a glance

- national research center supported by public funds
  - Internationally used but nationally funded
  - Particle physics (H1, Zeus, Atlas, CMS)
  - Research with photons (Flash, Pitz, XFEL soon)
  - In Zeuthen also astro physics (Amanda/Icecube)
- member of the Helmholtz Association
- locations in Hamburg and Zeuthen
  - ~1800 staff members in Hamburg & Zeuthen

Grid Computing at DESY

- DESY takes part in the LHC Computing Grid (LCG)
- It is a Tier2 centre for the LHC experiments ATLAS, CMS and LHCb
- gLite middleware
  - partly Globus-based
- As LRMS mostly Torque/Maui is used – but SGE, Condor, LSF work as well

Grid Computing at DESY



gLite – SGE integration test

- Local batch system behind a Computing Element (CE)
- Batch system support plugin-based:
  - must provide the basic batch system operations
    - submit, query, cancel, hold, release
- Information provider reporting LRMS status
- SGE was integrated successfully and CE is still running as proof of concept
  - globe-ce1.ifh.de

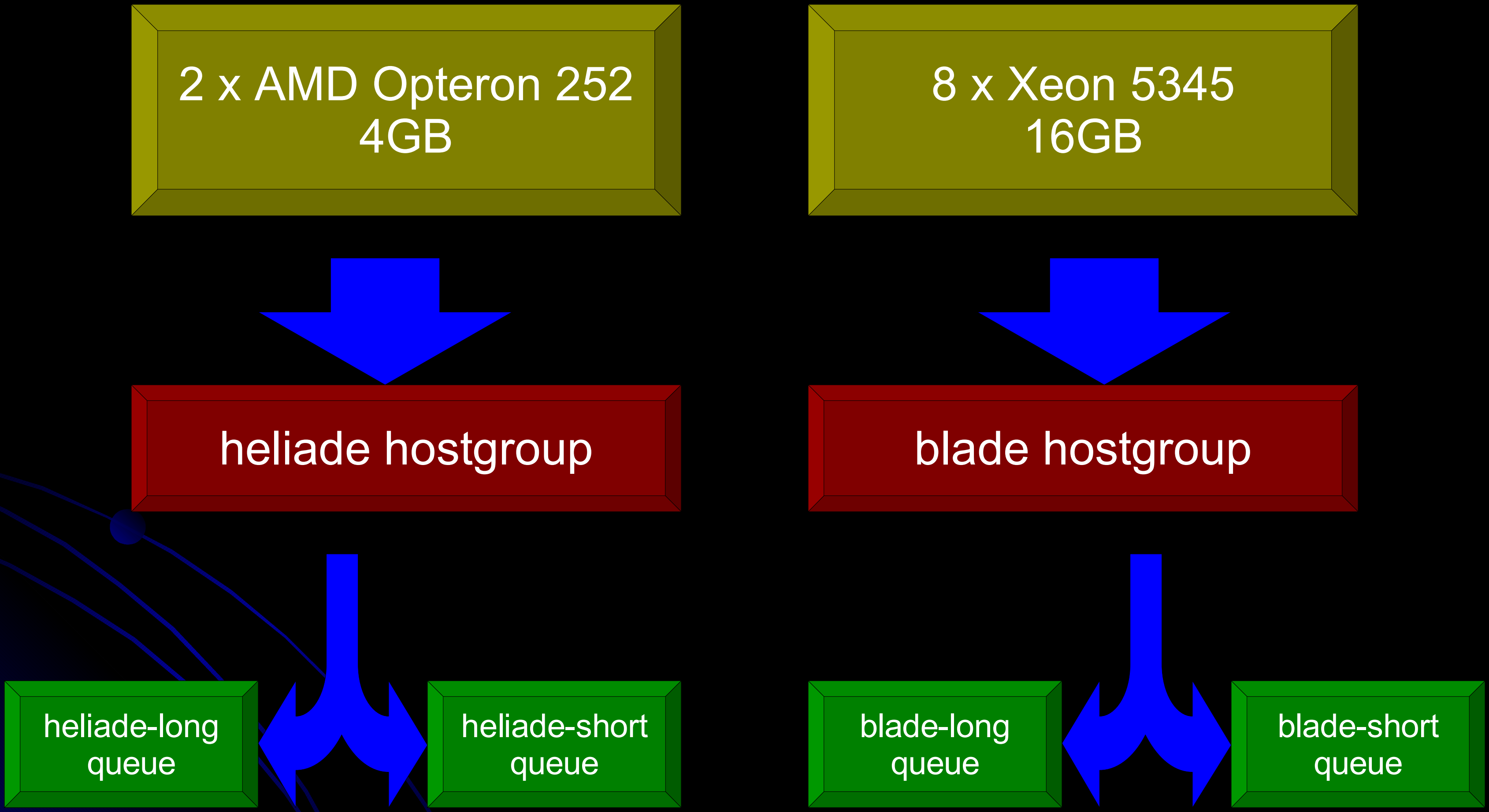
SGE installation in Zeuthen

- Used Codine since 1993
- Now running GE 6.0u9
  - self-built from sources
  - Plan to upgrade to 6.1 in autumn this year
- Now hosting ~450 cores (~150 hosts)
  - Scientific Linux (RHEL clone) 3 & 5, x86_64
  - Tight MPI-Integration on ~75 cores
- K5/AFS integration
  - see also Wolfgang Friebel's talk tomorrow

Coming up soon...

- DESY is building up a National Analysis Facility for the German LHC experiments
  - Starting in December 2007
  - It has been decided SGE will be the LRMS
    - Will also provide services for the LCG
  - Available cpu cores within the NAF will be higher than the available cores in all farms in Hamburg and Zeuthen

SGE host configuration (1)



2 x AMD Opteron 252
4GB

8 x Xeon 5345
16GB

heliade hostgroup

blade hostgroup

heliade-long queue

heliade-short queue

blade-long queue

blade-short queue

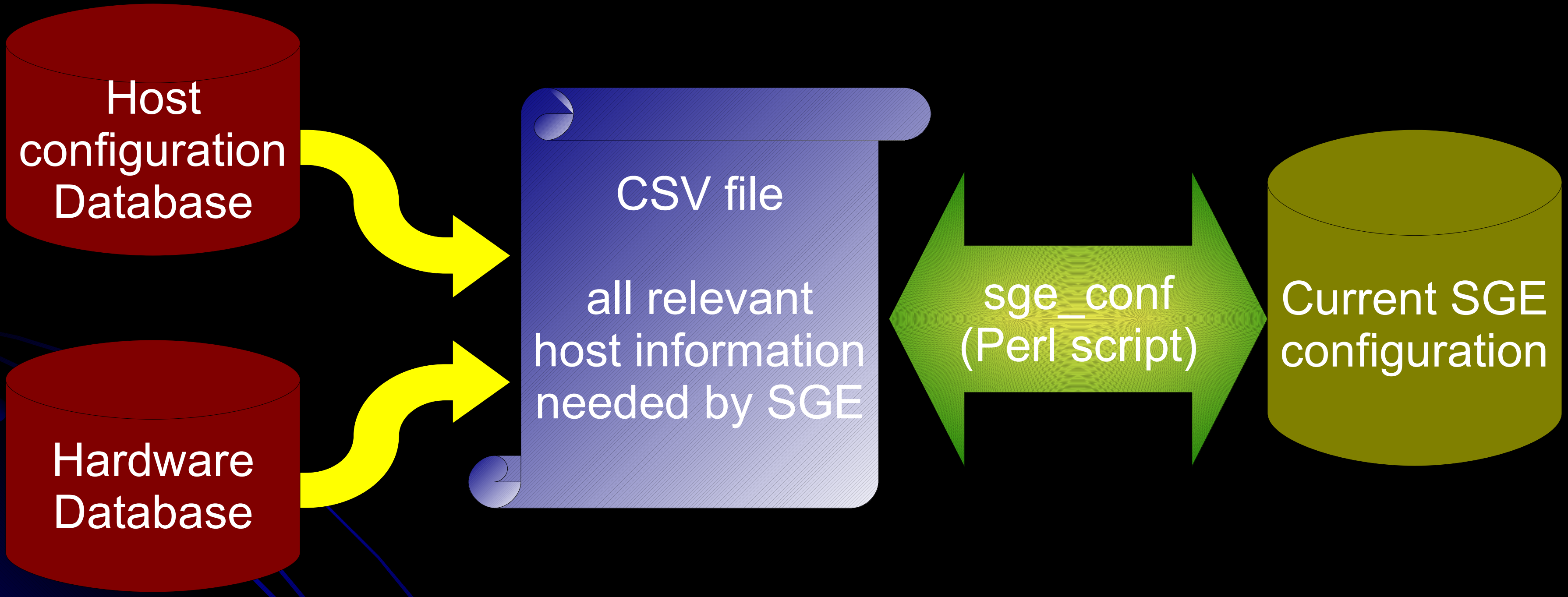
SGE host configuration (2)

- Long queues for jobs needing up to 48 hours
- Short queues for jobs needing up to 30 minutes
  - Allow for a temporary oversubscription (more jobs than cores) of hosts
- Users don't specify the queue but the actual job's requirements
  - SGE finds the matching queue on the fastest available host

SGE host configuration (3)

- Additional/modified complexes
  - Default arch complex modified (amd64, ia32 - instead of e.g. lx26-amd64)
  - os: operating system (sl3, sl4, sl5)
  - tmp_free: free disk space on $TMPDIR
    - Determined by a load sensor
  - h_vmem as consumable complex set to the host's amount of RAM initially

Automated host configuration (1)


Host configuration Database

Hardware Database

CSV file

all relevant host information needed by SGE

sge_conf (Perl script)

Current SGE configuration

Automated host configuration (2)

- On changes of the CSV file sge_conf is called automatically:
  - Compares host data in CSV file with current SGE configuration (qconf queries)
  - Detects changes and executes qconf statements if changes are needed
  - Host attributes:
    - Type: submit, administration, execution
    - For execution hosts:
      - arch, os, h_vmem, num_proc, slots, usage scaling, ...

Automated host configuration (3)

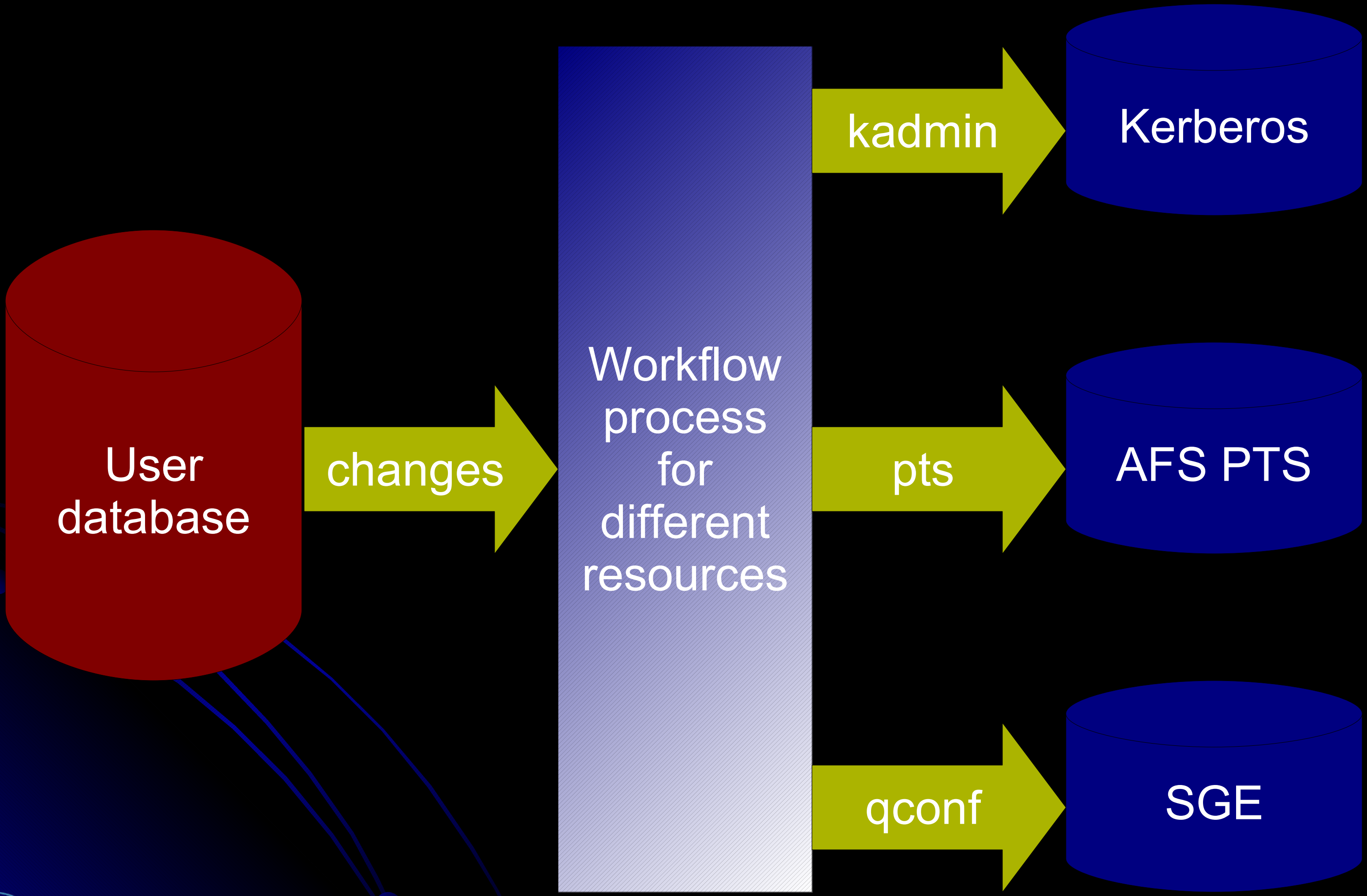
- Example host configuration:

```
[oreade38] ~ % qconf -se blade00
hostname       blade00.ifh.de
load_scaling   NONE
complex_values arch=amd64,num_proc=8,slots=12,os=sl5,h_vmem=16184M, \
               mem_total=16184M,virtual_total=16184M
[...]
processors     8
usage_scaling  cpu=2.327000
```

Automated user configuration (1)

- All users are members of different SGE projects
  - The configured SGE projects are a subset of our available unix groups (flag in a database)
  - 1 : 1 mapping userset – project
  - User's default project is identical to her/his primary unix group
  - User's additional unix groups lead to additional userset/project membership in SGE

Automated user configuration (2)



User database

changes

Workflow process for different resources

kadmin

Kerberos

pts

AFS PTS

qconf

SGE

qrsh / ssh integration (1)

- Problem:
  - you can configure qrsh to use ssh
  - but the execd is not able to monitor/setup the session correctly:
    - accounting missing
    - resource overusage not penalised
    - SGE environment not set within the session
  - To get the first two problems solved, the ssh session actually just needs the additional group id set – indicates the job's membership of a process

qrsh / ssh integration (2)

- Default approach: patch OpenSSH
  - Ron Chen's tight SGE-SSH integration
- Our approach: use PAM
  - pam_sge-qrsh-setup.so
  - Module sources the job's environment file
    - Setup job env, add additional group id
  - /etc/pam.d/sshd:

```
auth    required        pam_sge-qrsh-setup.so
auth    include         system-auth
...
```

qrsh / ssh integration (3)

- qrsh qconf settings:
  - rshd-wrapper stores information in a file for later usage in PAM module

```
rsh_daemon        /opt/products/gridengine/6.0u9/util/rshd-wrapper
rsh_command       ssh -tt -o GSSAPIDelegateCredentials=no
qlogin_daemon     /opt/products/gridengine/6.0u9/util/rshd-wrapper
qlogin_command    ssh -tt -o GSSAPIDelegateCredentials=no
rlogin_daemon     /opt/products/gridengine/6.0u9/util/rshd-wrapper
rlogin_command    ssh -tt -o GSSAPIDelegateCredentials=no
```

qrsh / ssh integration

Patching OpenSSH

- Cannot use vendor version
- In case of security problems in OpenSSH you must hope the patch also applies to the new version – or wait for a new patch

PAM module

- Need to have a PAM aware and enabled system

Thanks for your attention

- qrsh/ssh integration:
  - http://www-zeuthen.desy.de/~ahaupt/downloads/sge-sshd-control-1.2-1.src.rpm
- SGE/gLite integration:
  - https://twiki.cern.ch/twiki/bin/view/LCG/GenericInstallGuide310#The_SGE_batch_system