Challenge and Future of Job Distribution at a Multi-VO Grid Site

Andreas Gellrich, Birgit Lewendel (DESY)

2015-04-13

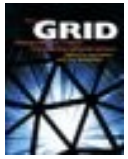
CHEP2015, Okinawa, Japan
Track 6, Abstract 370

HELMHOLTZ ASSOCIATION

DESY

Introduction: A short reminder of the Grid idea

- Grid Blueprint by Kesselmann/Foster (1999)
  - Collaborative problem solving → VOs

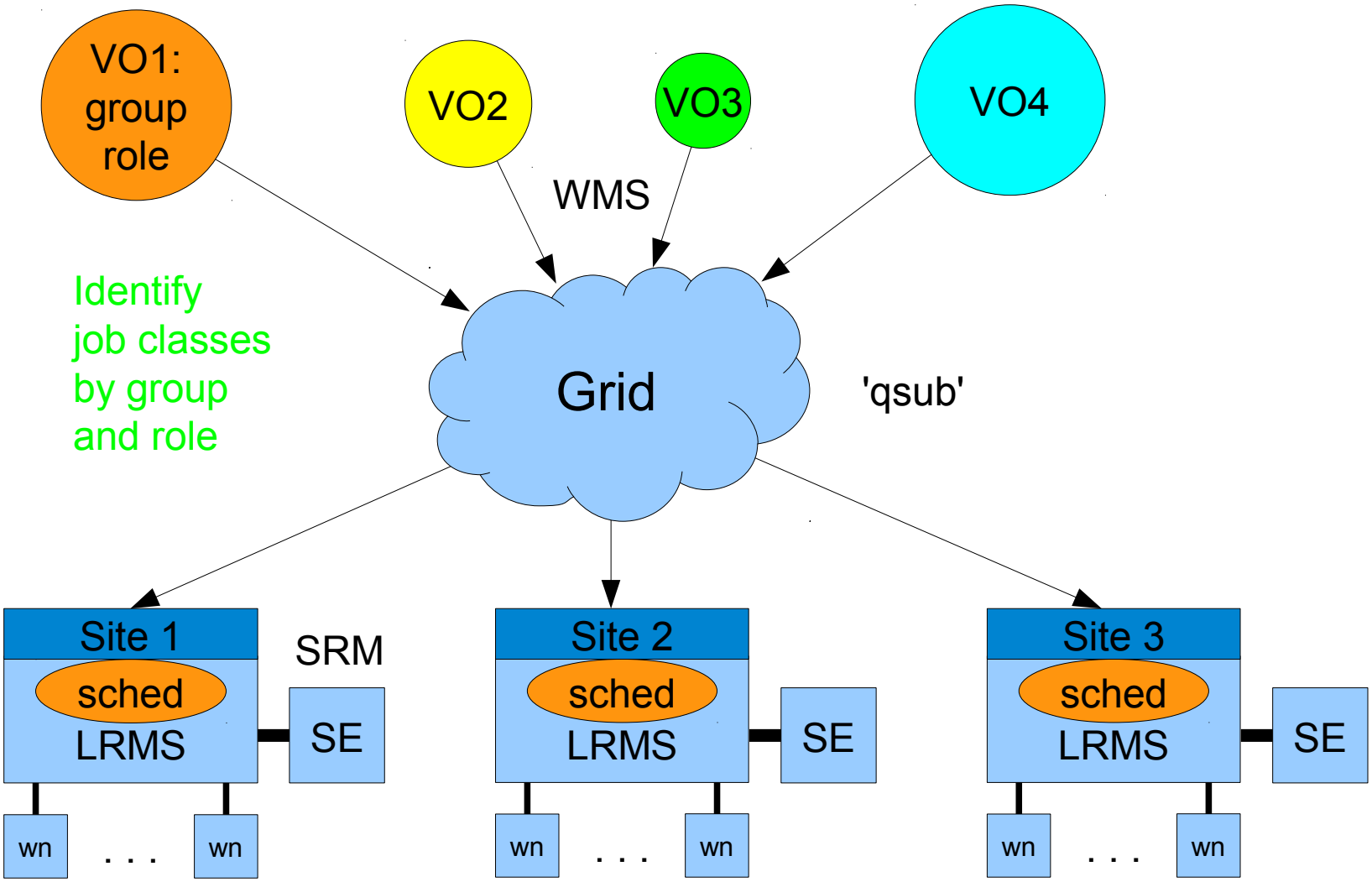


- "Three Point Checklist" Foster (2000)
  - No centralized controls (NOT like distributed computing)
  - Standard, open, general-purpose protocols
  - Non-trivial quality of services (NOT like web)

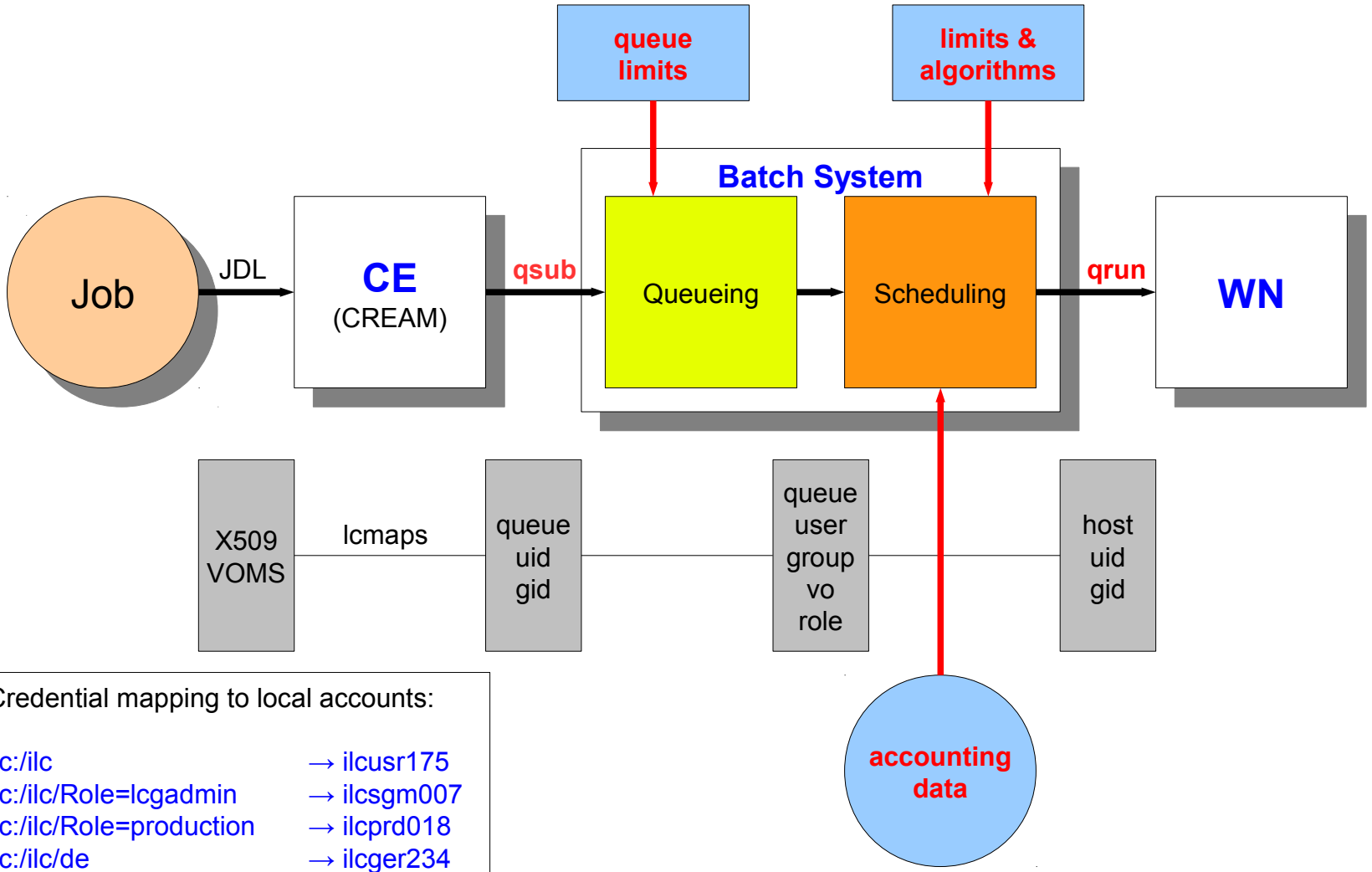
- EDG → LCG → gLite → EMI, NorduGrid → …

- Virtual Organization (VO) w/ individual groups and roles to identify job classes

The Grid: Initial Concept



The Grid: Local Resource Management System (LRMS)



Credential mapping to local accounts:

ilc:/ilc	→ ilcusr175
ilc:/ilc/Role=lcgadmin	→ ilcsgm007
ilc:/ilc/Role=production	→ ilcprd018
ilc:/ilc/de	→ ilcger234
ilc:/ilc/Role=pilot	→ ilcplt079

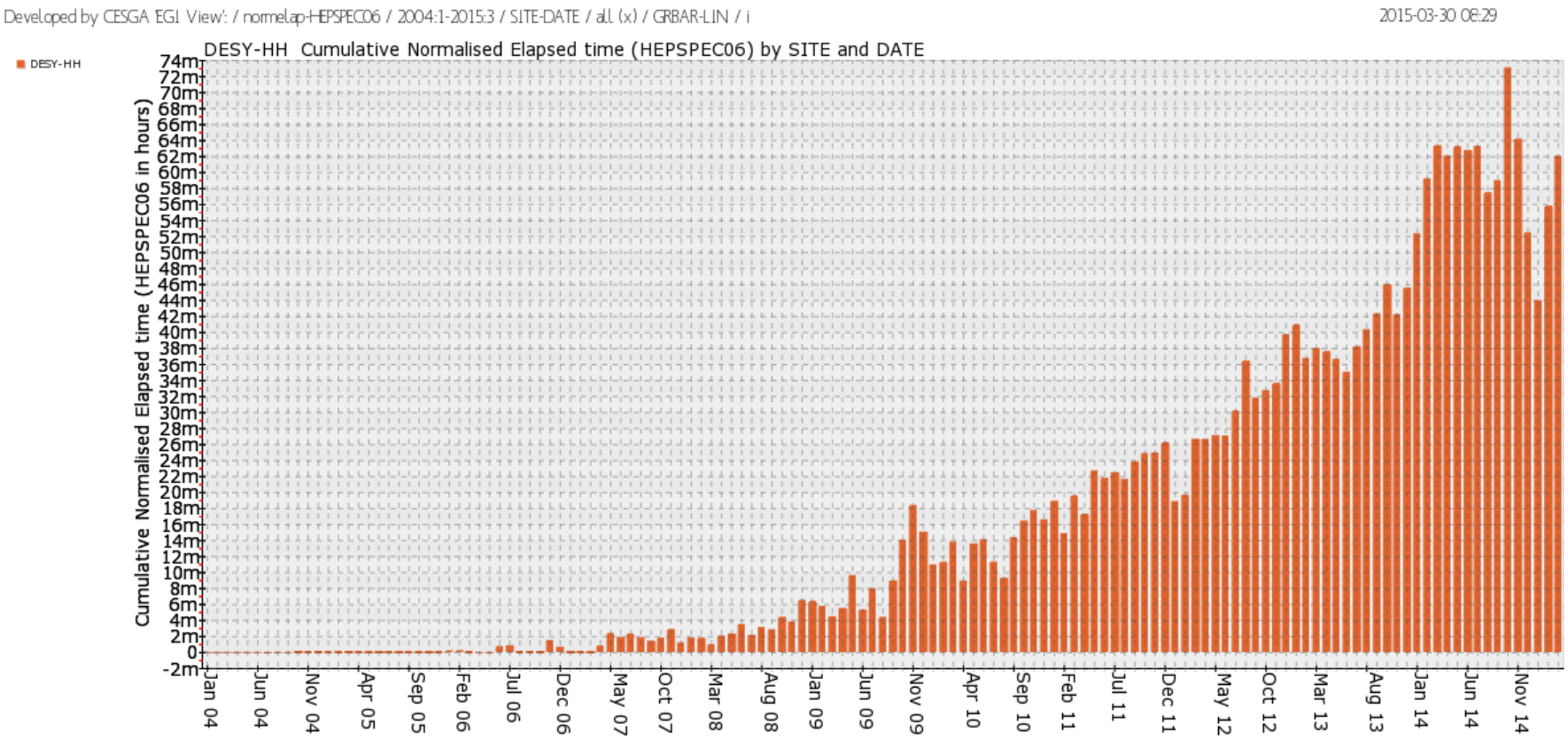
DESY-HH: History

- Most sites based their installations on the initial concepts

- DESY is an example for a big multi-VO site as Tier-2 w/ many disciplines
  - 2003: First look into Grid computing as key technology to access resources
  - 2004: LCG_2-1 Grid infrastructure H1 and ZEUS, IceCube, ILC, ILDG
  - 2004:: EU-Project EGEE(2/3), EGI
  - 2004: Tier-2 for ATLAS, CMS, LHCb
  - 2005: D-Grid (DGI(2) and HEPCG)
  - 2011: DESY is world-wide biggest Tier-2 center for CMS
  - 2012: scheduler studies because of scaling problems
  - 2013: BELLE2 added; major contributions to MC campaings
  - 2015: multi-core support started

DESY-HH: Accounting 2004-2015 (APEL)

~11,000 cores in 2015



DESY-HH: Site Set-up

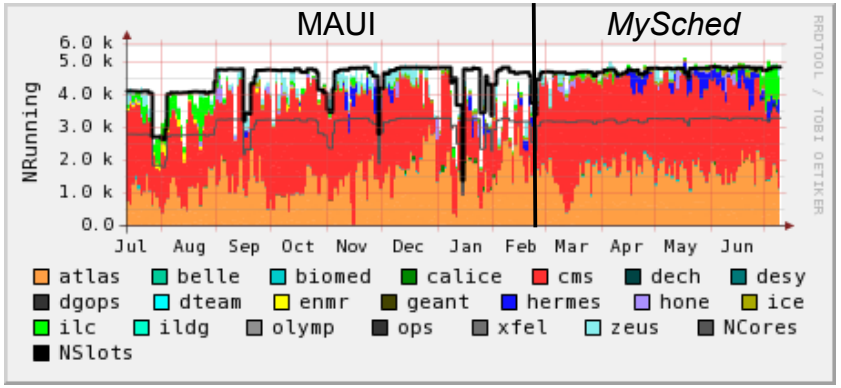
- The Grid infrastructure reflects DESY's manifold scientific programme:
  - DESY is the *home* of 10 VOs (6 global), incl. non-HEP
  - Tier-2 for ATLAS, CMS, LHCb, BELLE2 in Germany (Tier-1: GridKa), LHCone
  - Tier-0/1 for ILC VOs incl. Testbeams w/ tape back-end

- One *complete generic* Grid infrastructure for *all* VOs
  - All necessary Grid services (VOMS, LFC, WMS, PX, CVMFS-stratum0/1 ...)
  - *Federated* resources w/ *opportunistic* usage ("*everybody profits*")
  - Roughly 2/3 of the resources for WLCG and 1/3 for BELLE2 and ILC

- Operational aspects
  - Stable operations (no crashes, no local resource exhaustion)
  - VO requirements (MoUs, shares)
  - Utilization of resources (cpu, memory, disk, network)

DESY-HH: Local Resource Management System (LRMS)

- Initial standard set-up TORQUE / MAUI didn't scale beyond ~10k jobs
  - Instabilities
  - Blocking of submissions
  - Low occupancy
  - Configuration problems
  - Exhaustion of resources



- Heterogeneous Worker Nodes
  - 8 – 64 cores (Intel, AMD) (partially w/ HT)
  - 2-4 GB/core RAM
  - 20 GB/core disk
  - 11k cores == 12 kHS06 (2015-04-01)

myS(c)hed: A Scheduler Study

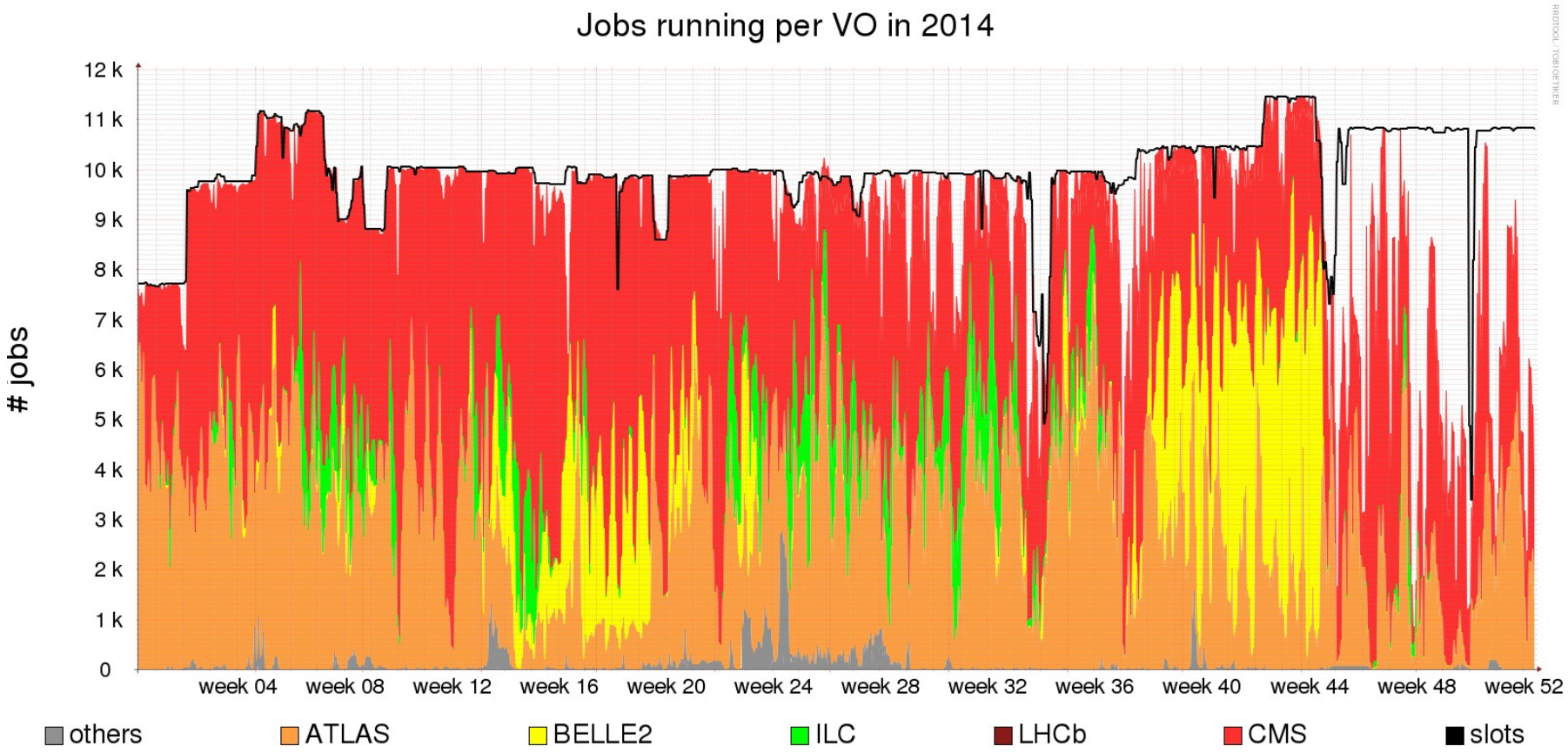
- Tailored to HEP (job-parallel)
- Scalable (number of jobs; number of slots)
- Optimize resource utilization (maximize diversity per node)
- Configurable (config file) with shares and rules
- Based on the torque C-API (libtorque.so.2)
- Light-weighted (CPU and memory usage)

- *myS(c)hed* algorithm:
  - Re-order job list according to shares (multi-core jobs first)
  - Find suitable node for job

Re-order jobs by share

DESY

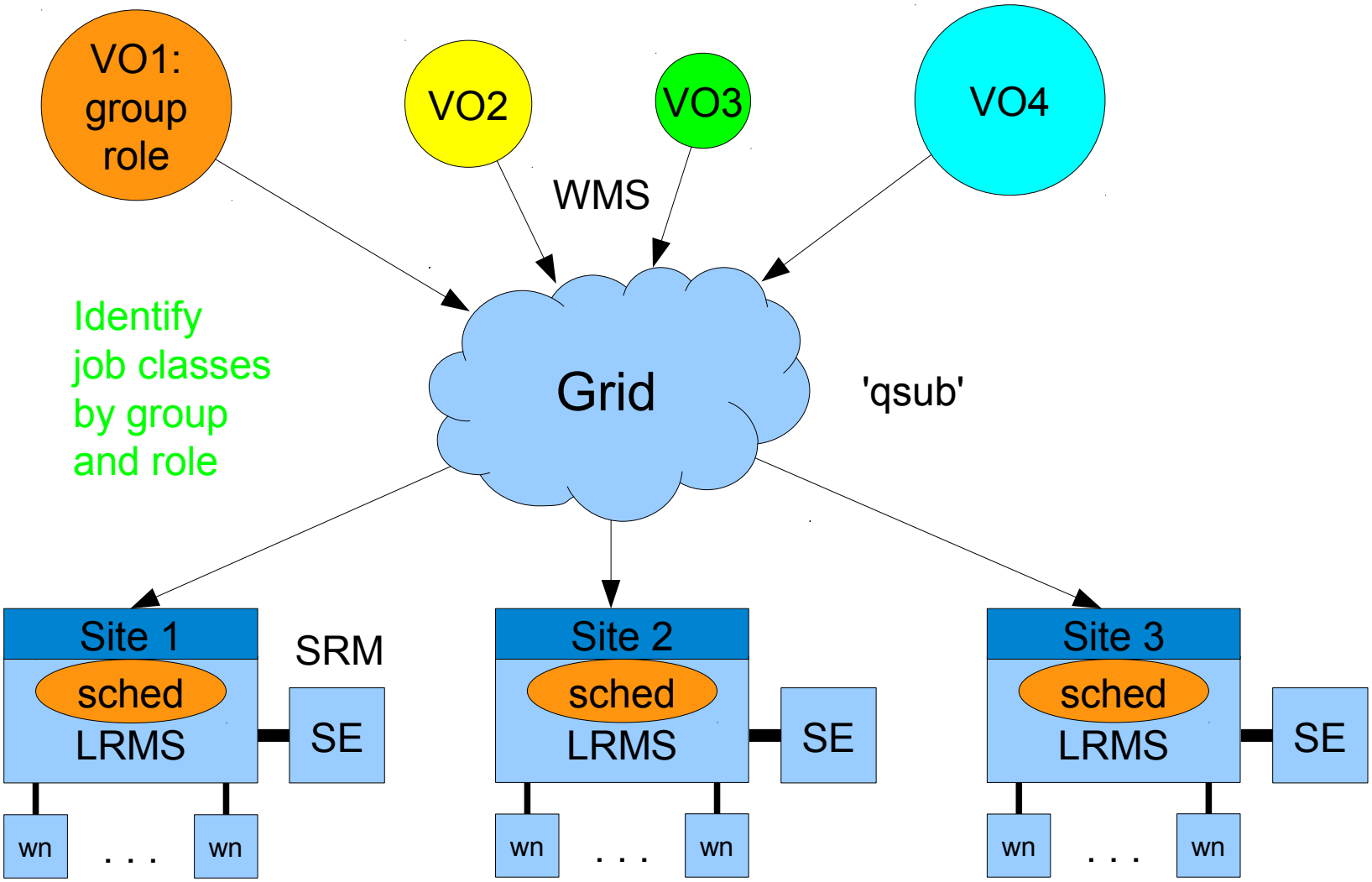
DESY-HH: Jobs 2014 (~120 kHS06)



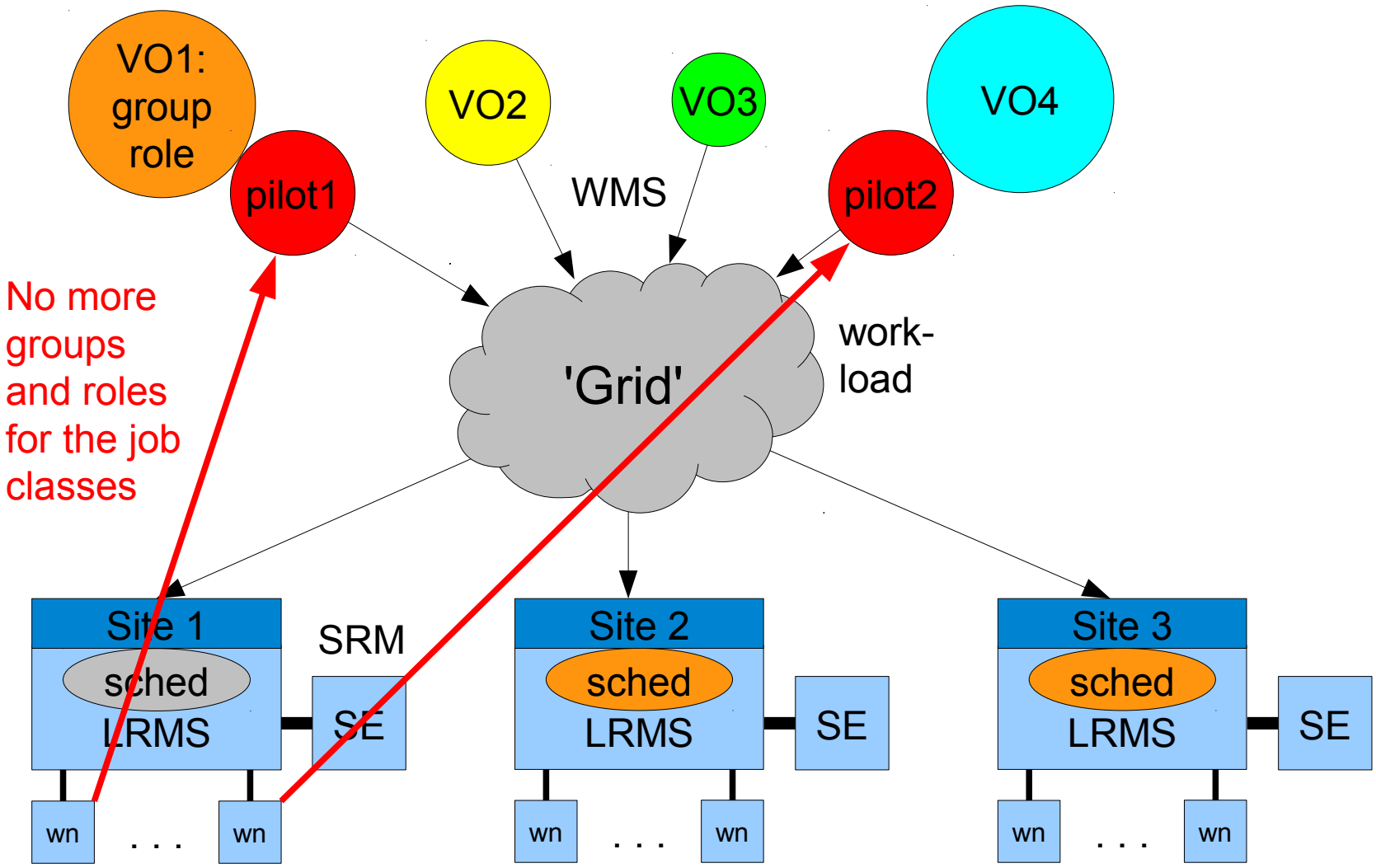
The Grid: Current Situation

- Main VOs use pilot factories (major impact on sites)
  - The Grid Information system is not used
  - Scheduling is done outside sites (by pilot factory)
  - Job classes (MC, analysis, etc.) can not be identified (anymore)
  - Very few (one) user per VO
  - The classical (local) scheduling approaches don't work
- Small sites vanish
  - Middleware and operational support decreasing
  - Know-how and awareness of computing leaves
- Clouds?

The Grid: Initial Concept

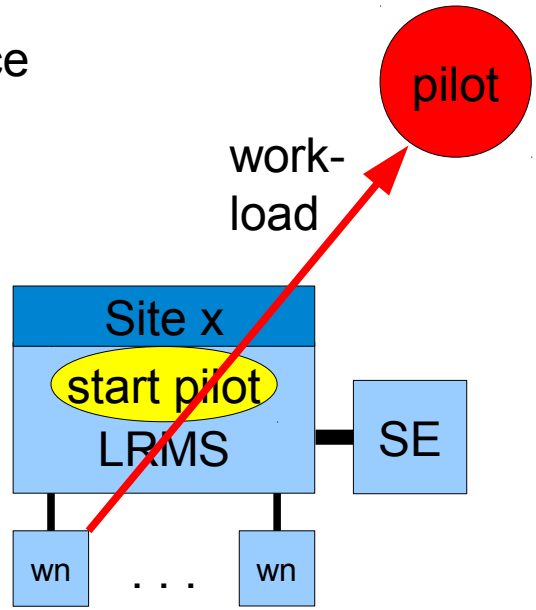


The Grid: Recent developments



Future: From Grid to Cloud?

- Clouds
  - Non-Grid approach
  - Sites just may provide their infrastructure as a service
- *Vacuum* approach
  - Pilots are started directly by the site
  - Save Grid overhead (WMS, CE)
  - Simple resource allocation model



- Small scale classical Grid sites for minor VOs?
  - Still demands
  - Even major VOs have regional groups

Summary and Conclusions

- The Grid has proven to be a key technology for WLCG, in HEP, and elsewhere
- Sites were set up based on the ideas of Kesselman and Foster ~10 years ago

- Introduction of pilot factories and multi-core jobs required massive changes
  - Focus on big sites as small sites vanish
  - Classical LRMS approach reaches limits

- Future of computing resource provisioning seems questionable
  - Give up on Grid as is and focus clouds instead?
  - What about small VOs?
  - How does this fit to scientific program of sites