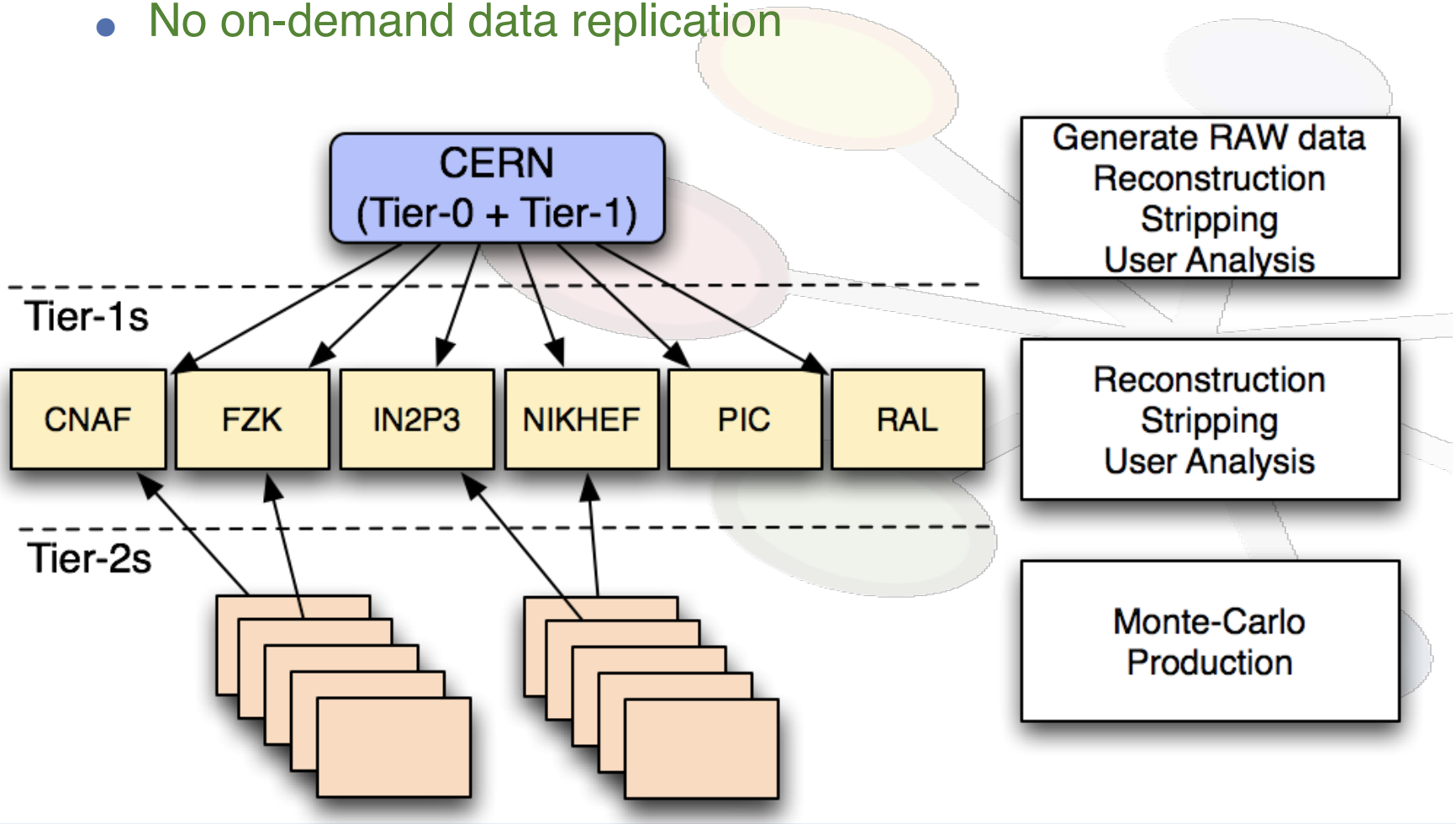
Performance of Combined Production and Analysis WMS in DIRAC

S. K. Paterson (CERN)

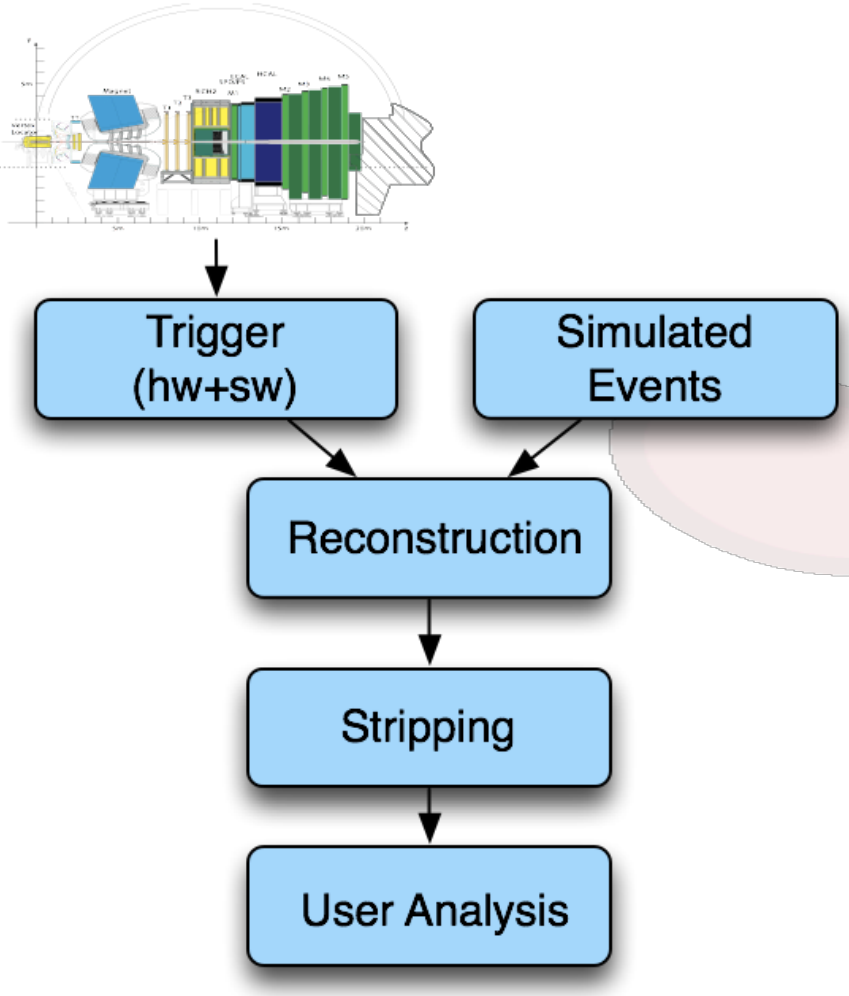
On behalf of the LHCb DIRAC Team

LHCb Computing Model

- LHCb distributed data analysis will occur at the Tier-1 sites
  - CERN (Tier-0) also acts as a Tier-1 site
- Jobs are sent where replicas of the data are
  - No on-demand data replication



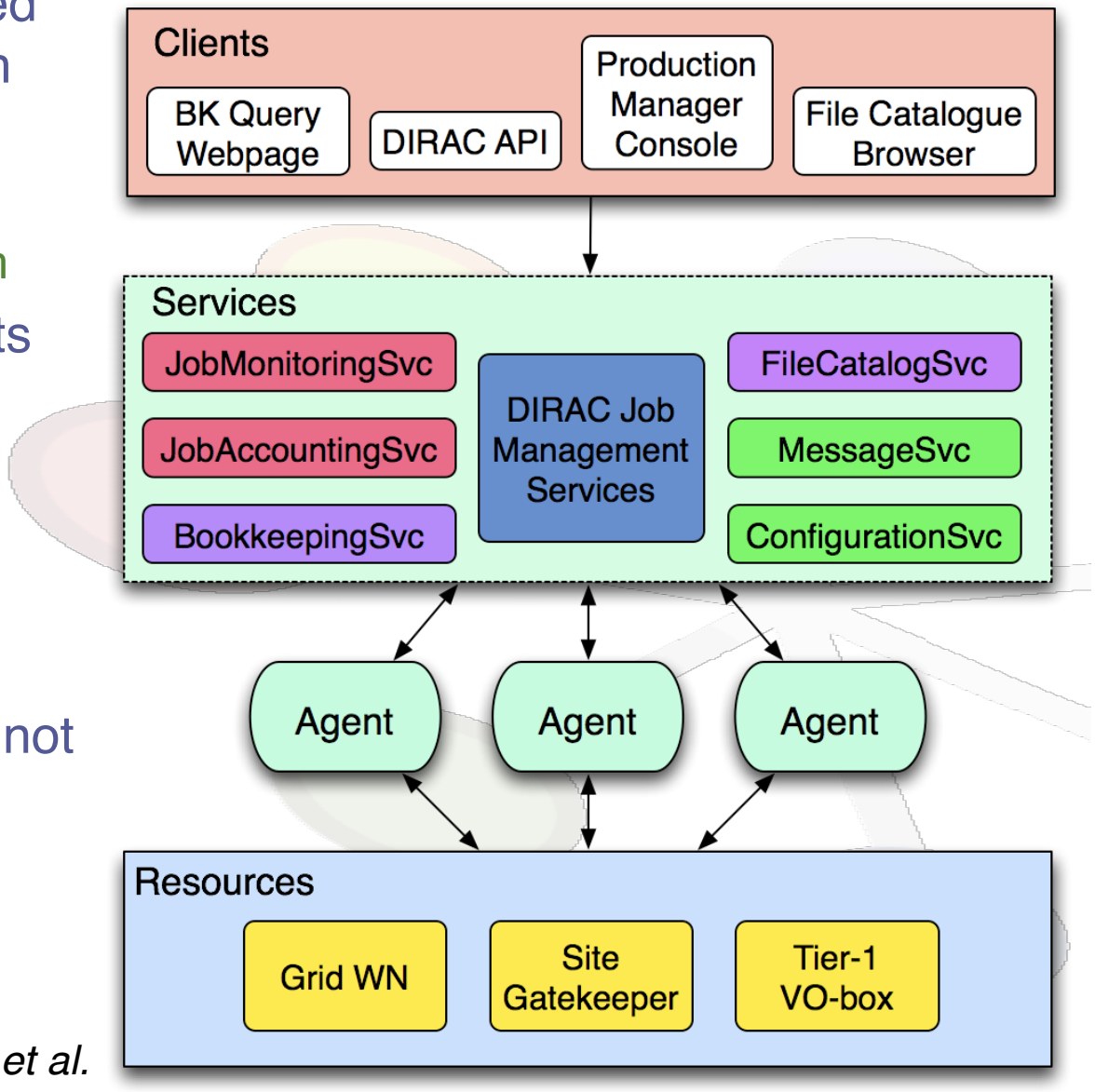
LHCb Data Processing Chain



- Stripping is a centrally managed analysis run as a production
  - Reduces analysis datasets to a size of 10^6 to 10^7 events per year
- User analysis is performed on stripped data throughout the year
  - All stripped data will be disk resident and replicated to all Tier-1 sites
- LHCb distributed analysis jobs are run on the Grid

DIRAC Architecture

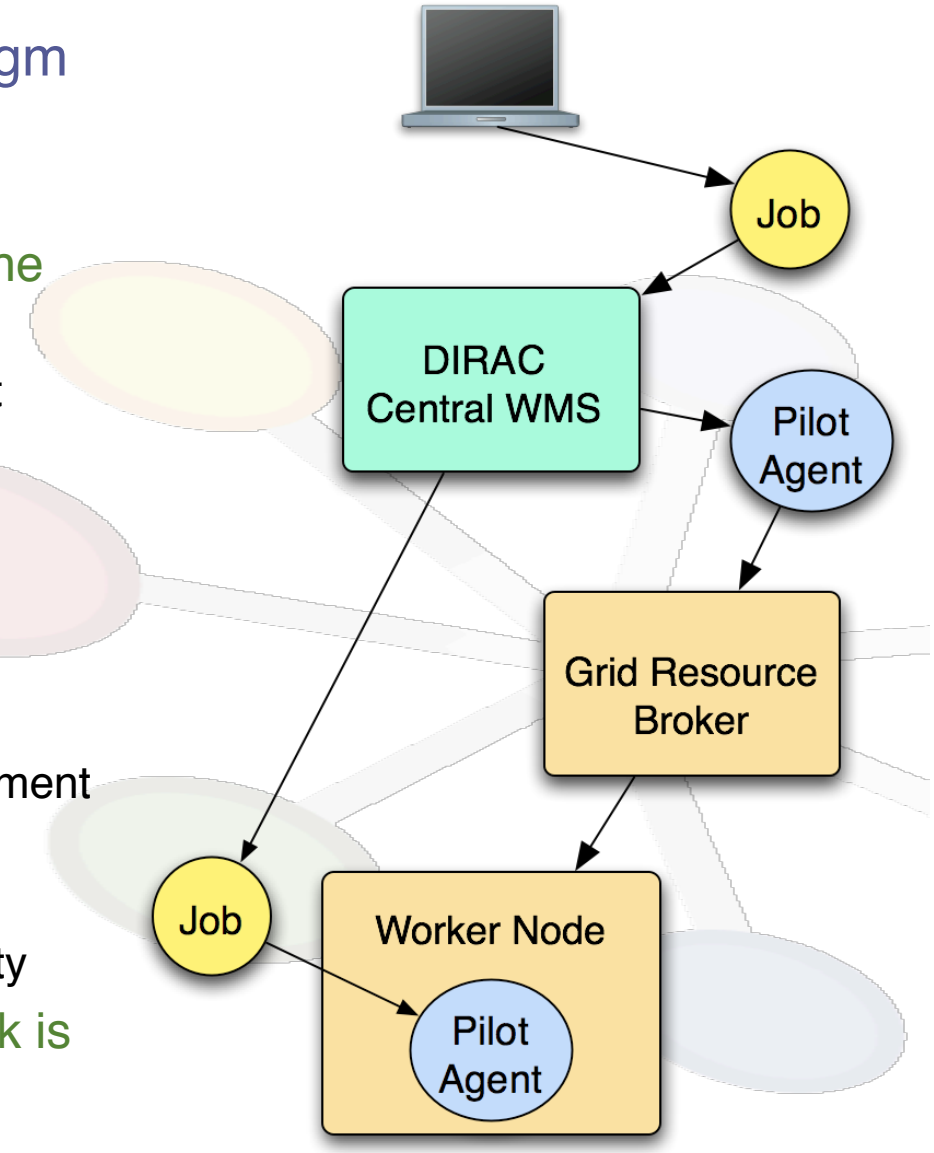
- DIRAC can be viewed as a very large batch system
  - Accounting
  - Priority mechanism
- The core components of the system are
  - Clients
  - Services
  - Agents
  - Resources
- The DIRAC WMS is not LHCb specific
  - GSI authentication
  - Standard JDL



See [427] – A.Tsaregorodtsev et al.

Pilot Agent Paradigm

- The DIRAC Pilot Agent paradigm provides a reliable means to deploy Agents to the Grid
  - wLCG jobs are Pilot jobs for the DIRAC WMS
    - Actual workload management performed by DIRAC
- Pilot Agents are submitted on demand from the WMS
  - Reserve the resource for immediate use
    - Allow checking of the environment before job scheduling
    - Bidirectional network traffic
    - Only unidirectional connectivity
  - Terminate gracefully if no work is available



Generic & Private Pilot Jobs

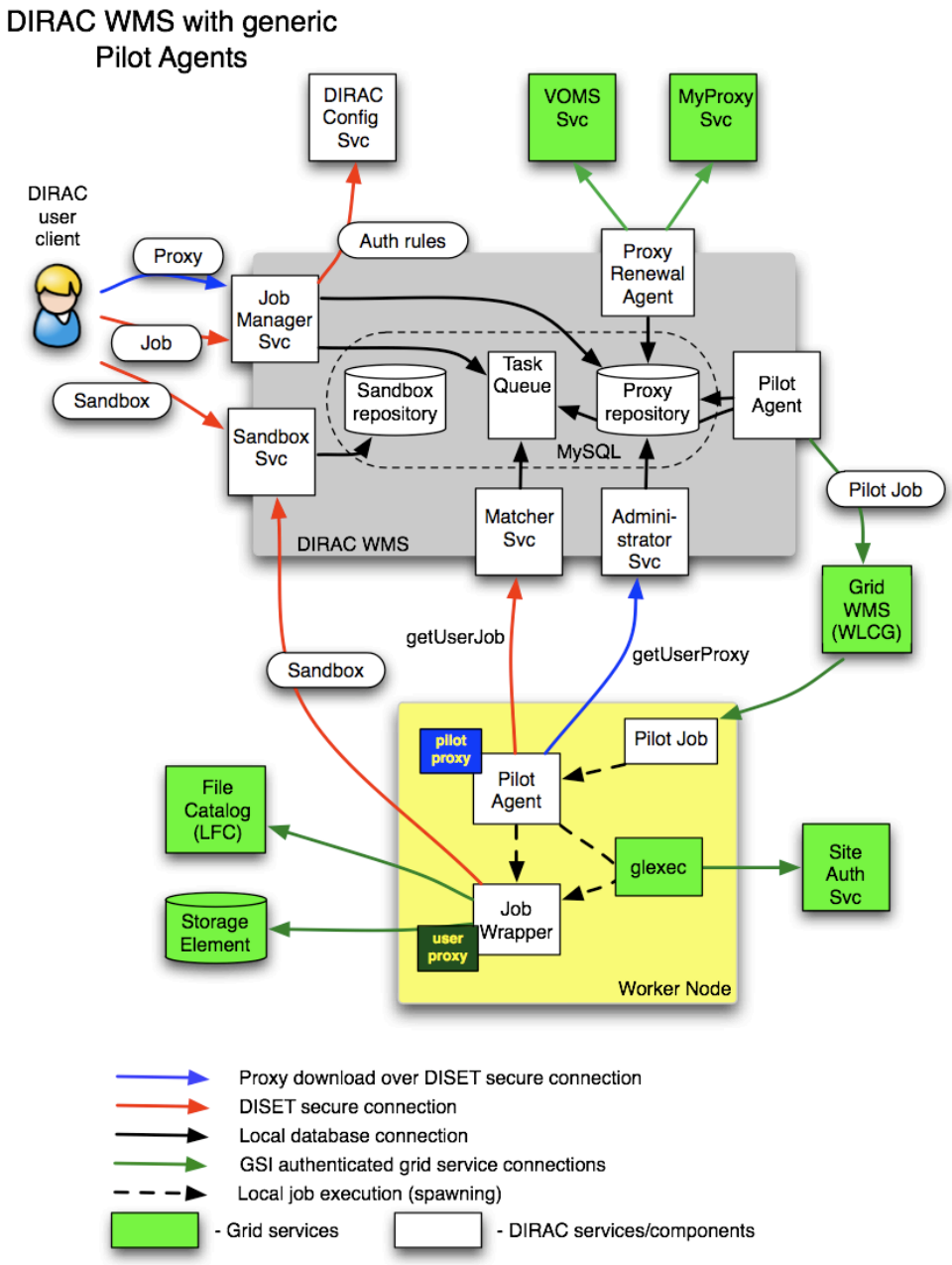
- Private pilots are submitted on behalf of the owner of the job
  - Same FQAN as the submitter
- Generic pilots are submitted on behalf of the VO (with 'pilot' role)
  - Allows the VO to schedule the most important task for the community to the available slots
    - User or production jobs
- Even with gLExec deployed we should cope with a mixture of generic and private pilots
  - DIRAC SAM jobs (lcgadmin) are a persistent use-case for private pilots e.g. to take advantage of priority queues set up at some sites

See [108] – R. Graciani Diaz et al.

DIRAC WMS & Proxy Renewal

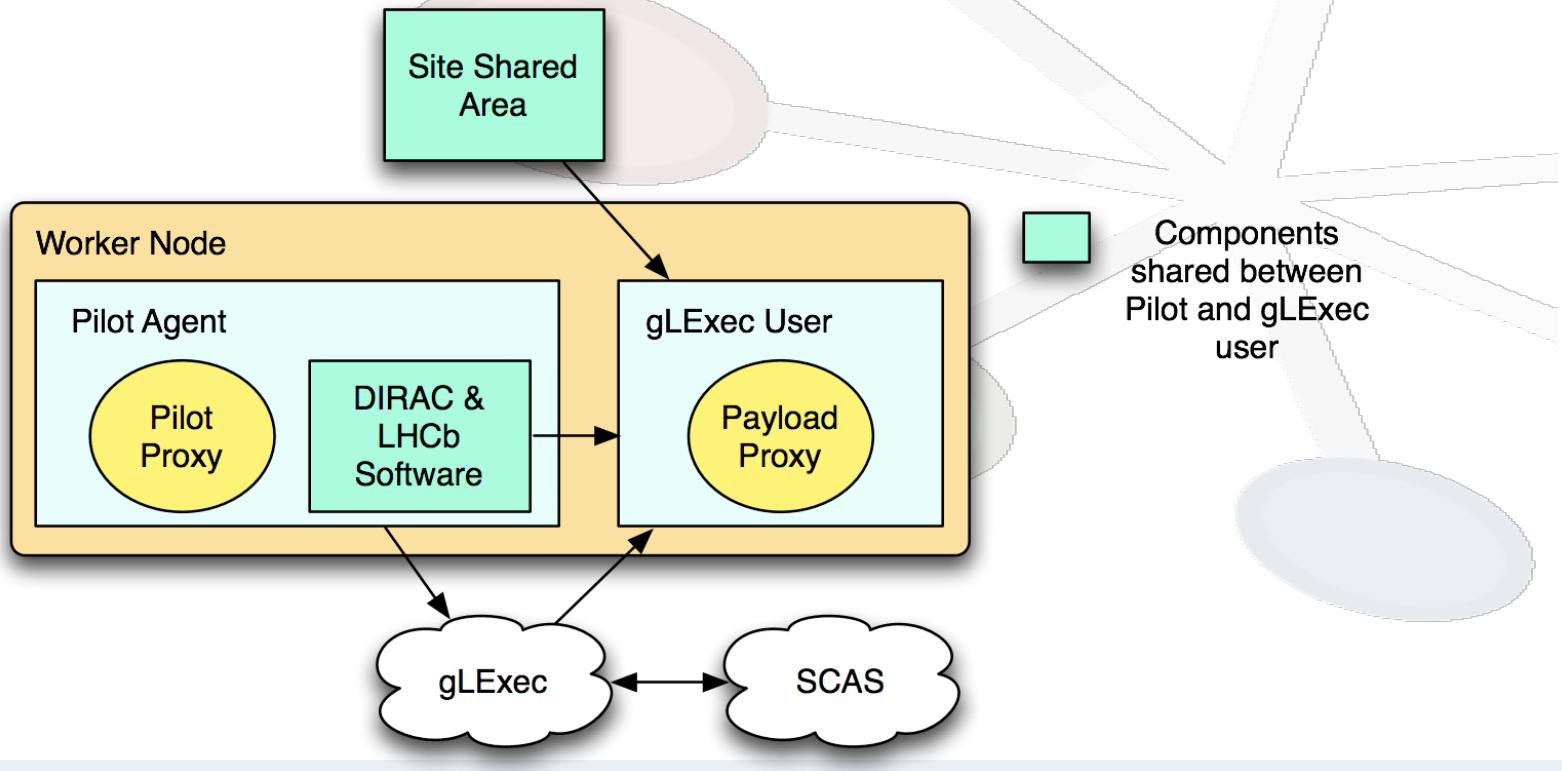
- LHCb WMS already optimized for generic pilots
  - Single WMS for all activities
  - Less freedom to optimize scheduling without generic pilots
- Pilot proxy never extended by DIRAC
  - Rely on gLite renewal
- Payload (limited) proxy is extended using the pilot proxy to contact DIRAC
  - Amount of extensions is restricted

See [90] – A. Casajus Ramo et al.



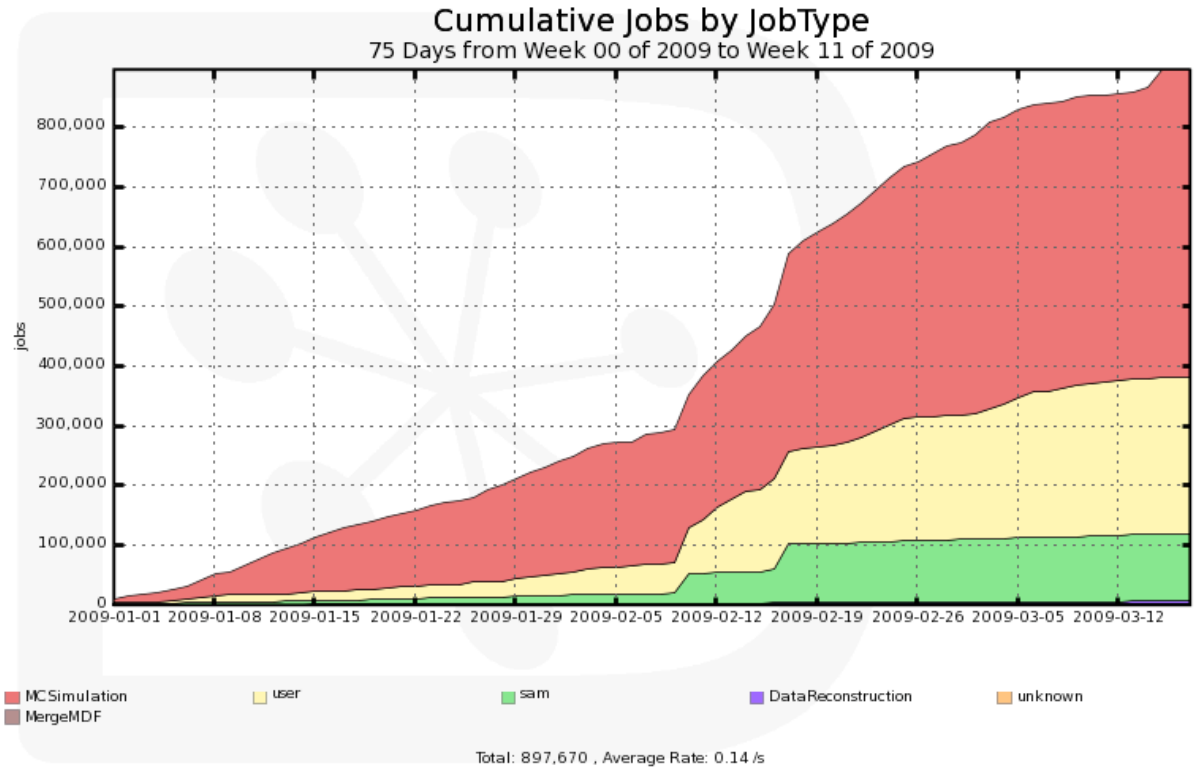
Multi-User Pilot Jobs In DIRAC

- LHCb VOMS roles are lcgadmin, production, user and pilot
  - Pilot role is strictly controlled
- If a site supports the pilot role LHCb generic pilots are submitted there
  - gLExec is imposed by the site at the user level to provide traceability

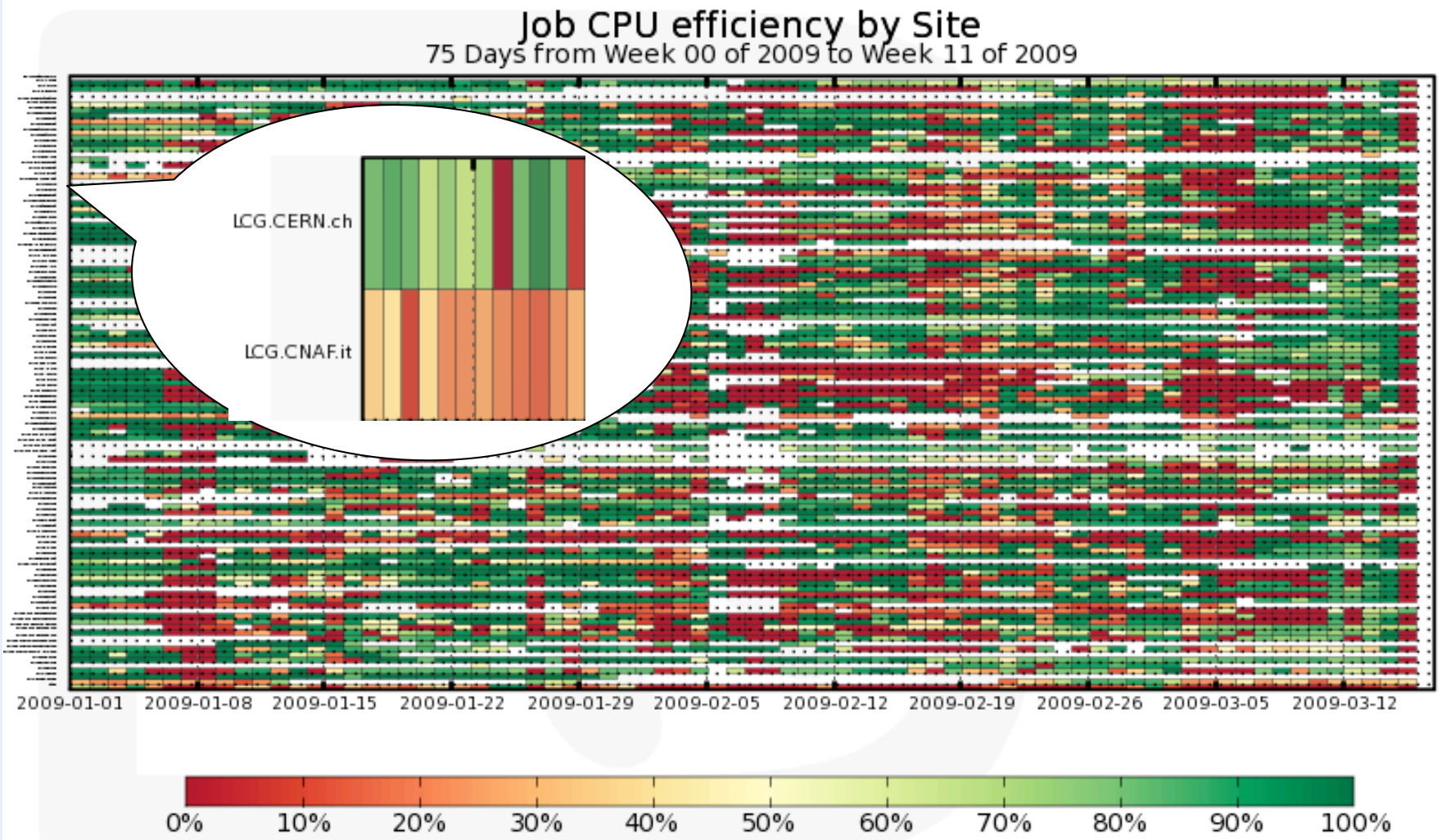


LHCb Jobs In 2009

- ~1M LHCb jobs were submitted to DIRAC so far in 2009
  - 62% MC simulation – generally low priority (24hrs)
  - 28% Data processing – high priority (variable length)
    - Dominated by user processing jobs
  - 10% SAM jobs – high priority (short jobs)



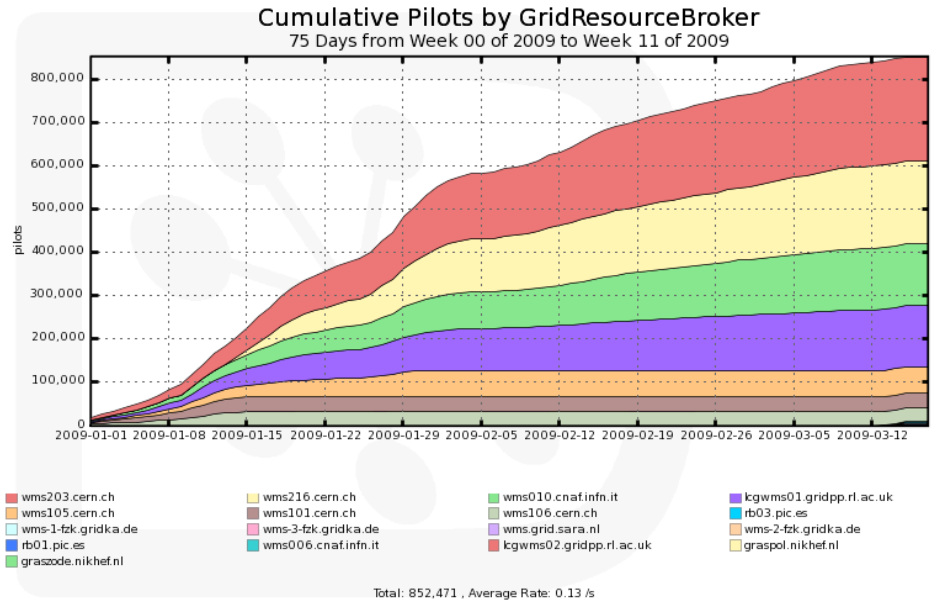
Operations can be a challenge…

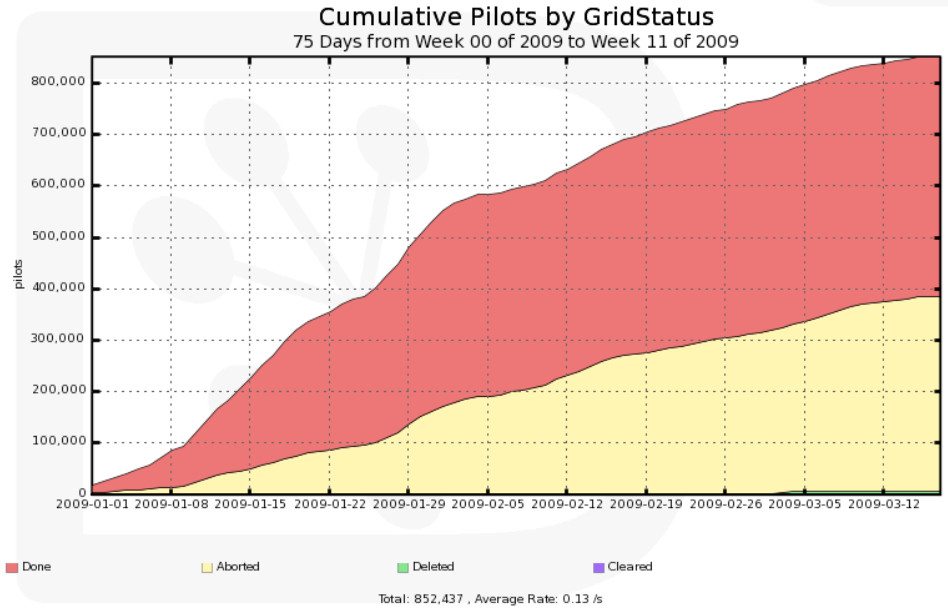


- ~120 sites regularly used this year

DIRAC Pilot Monitoring

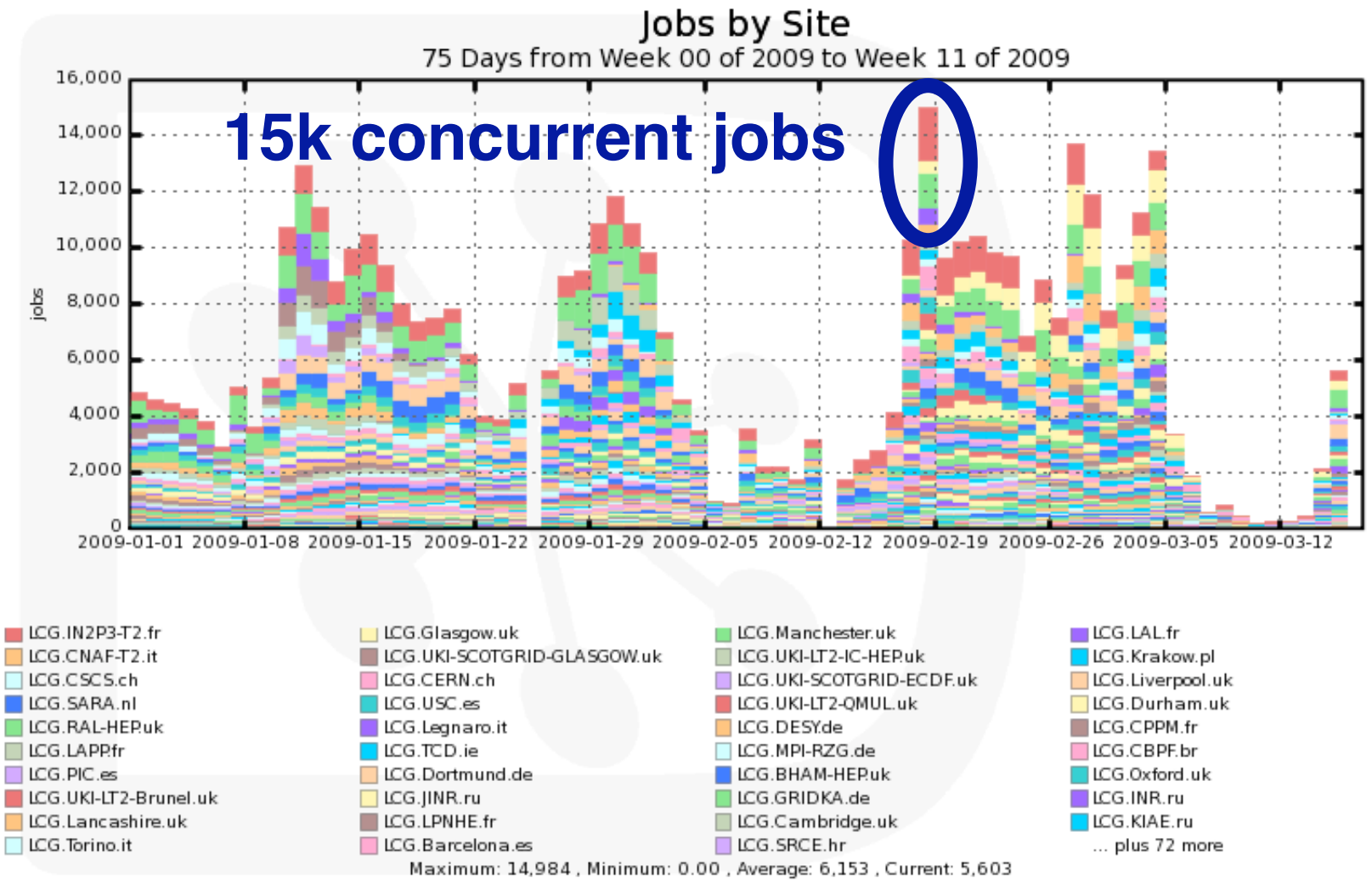
- Submitted pilots are tracked by DIRAC
  - ~1M pilots sent in 2009 to 17 gLite WMS instances



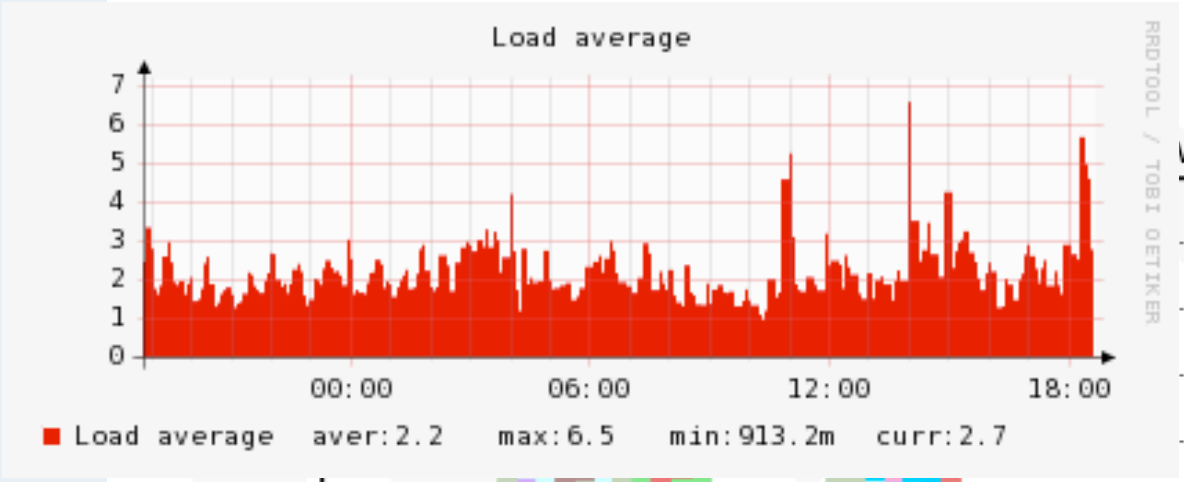


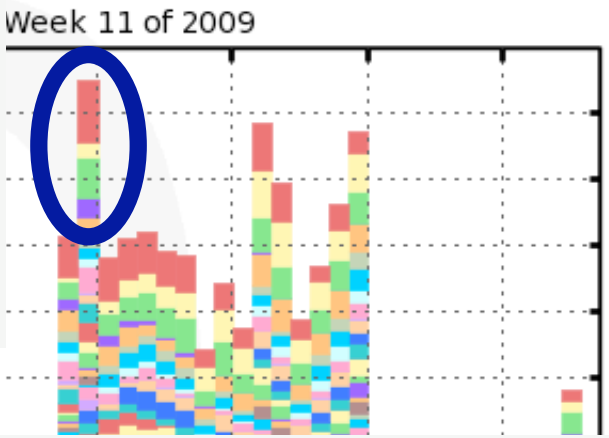
- ~40% of pilots were 'Aborted' in 2009
  - No retry at WMS level
  - Includes pilots with indeterminable status

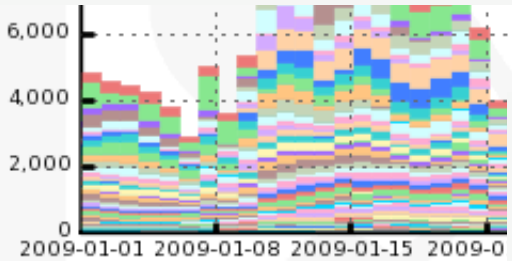
Concurrently Running Jobs

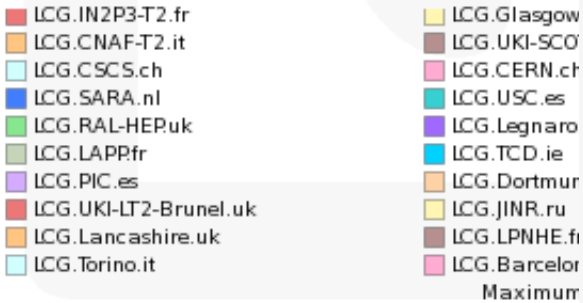


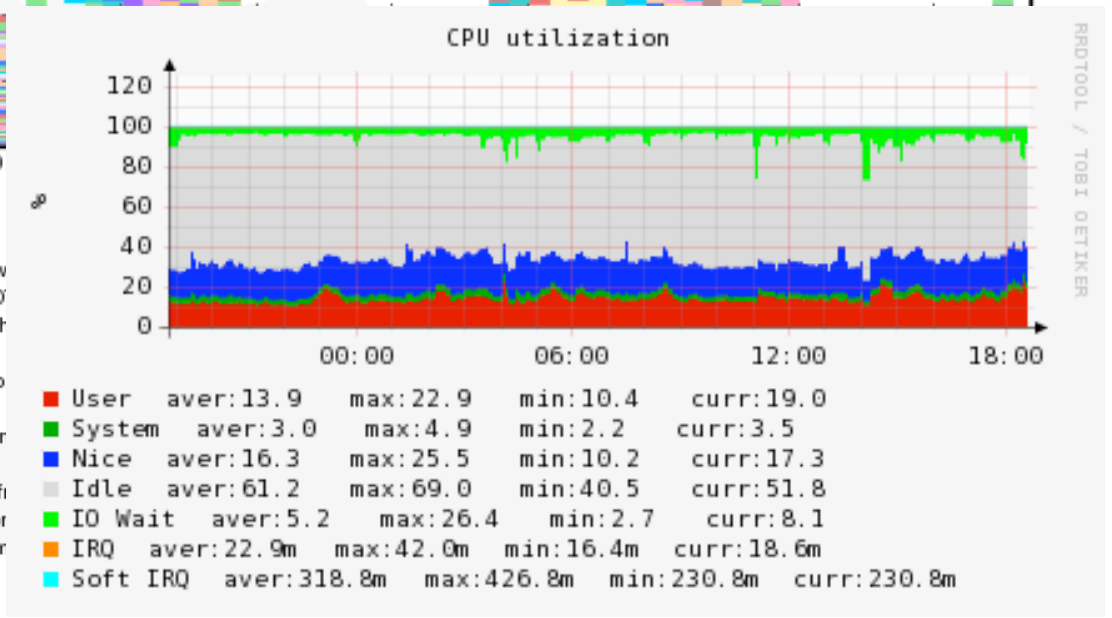
Concurrently Running Jobs





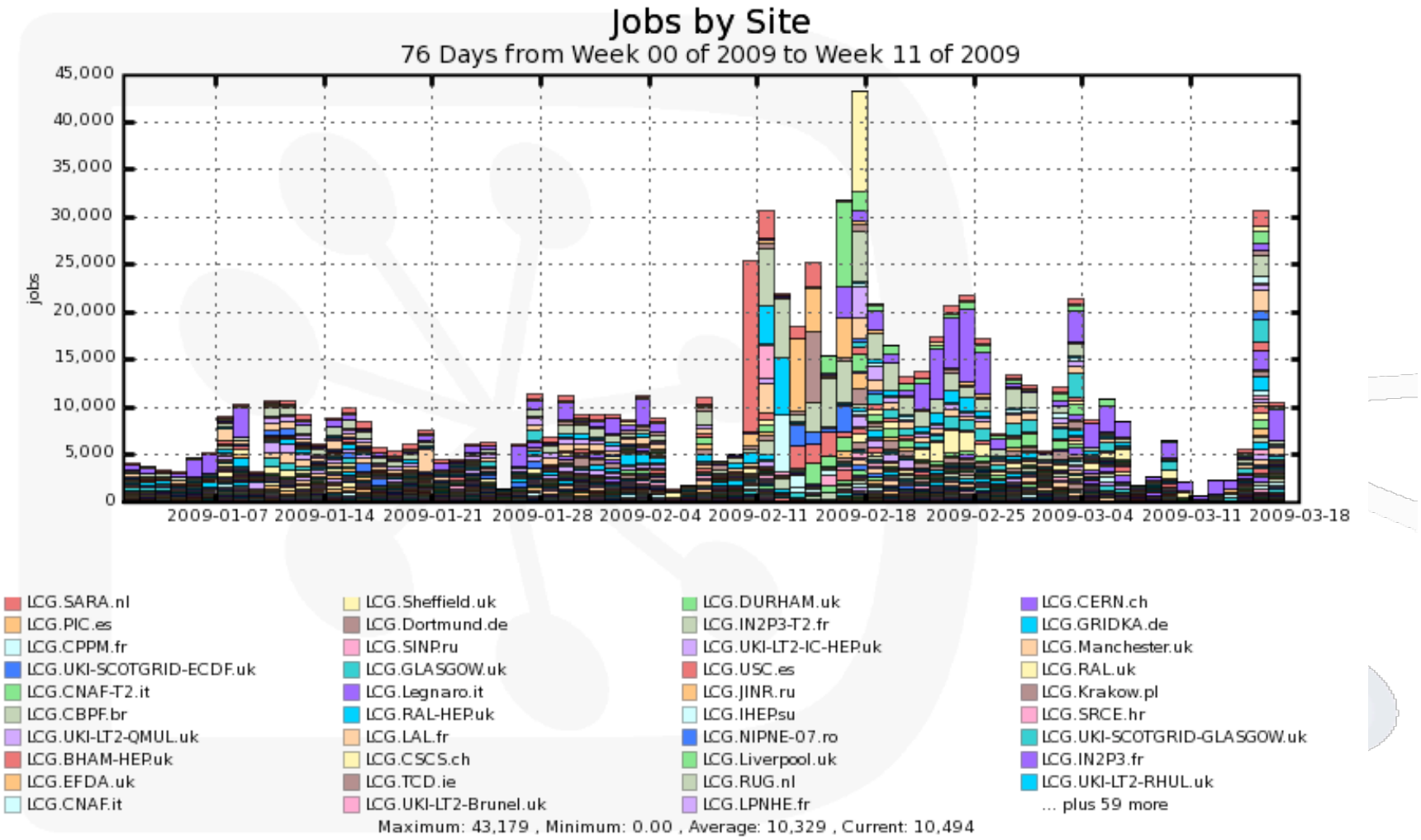




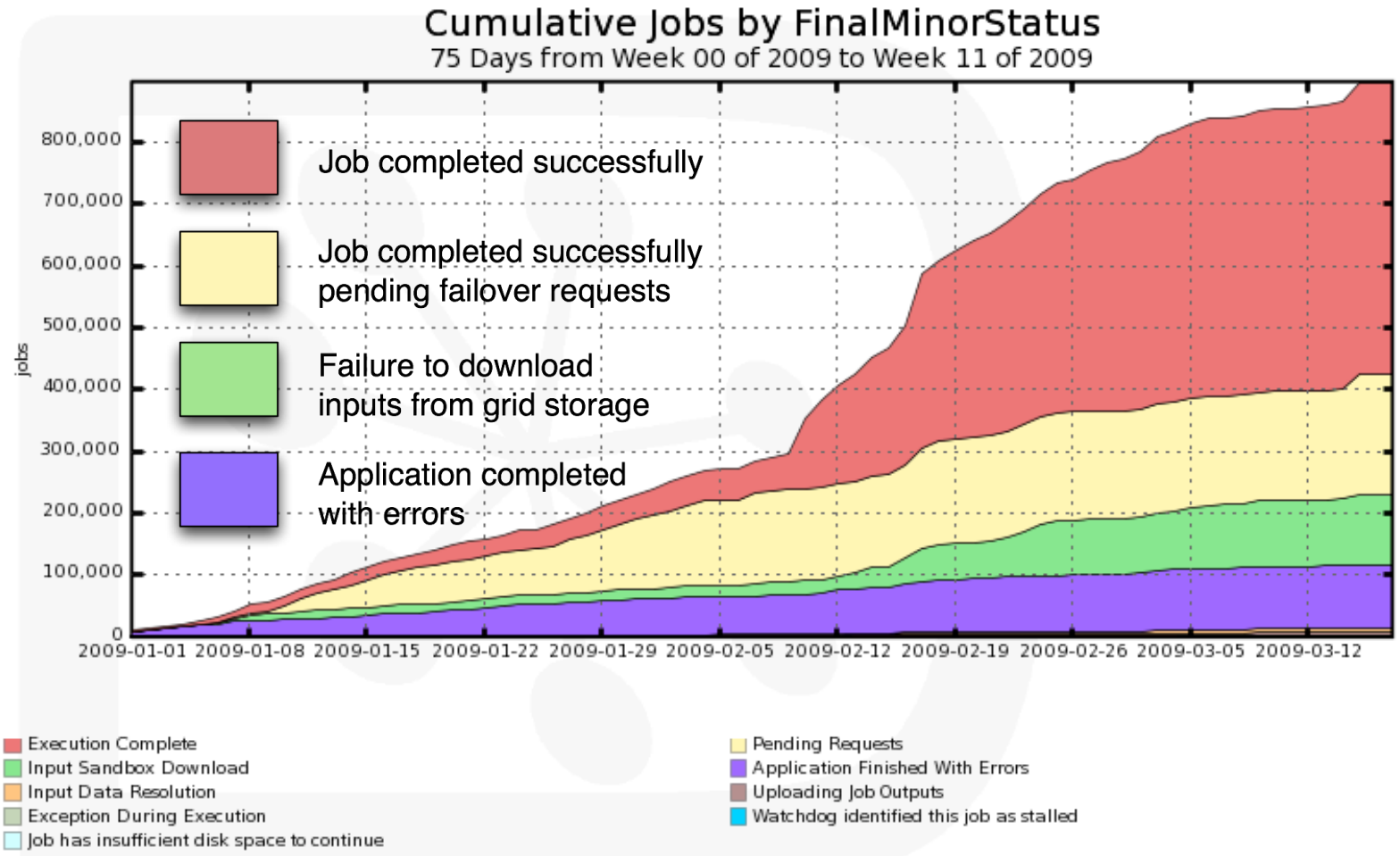


Throughput

- At peak running we achieved > 40k jobs per day



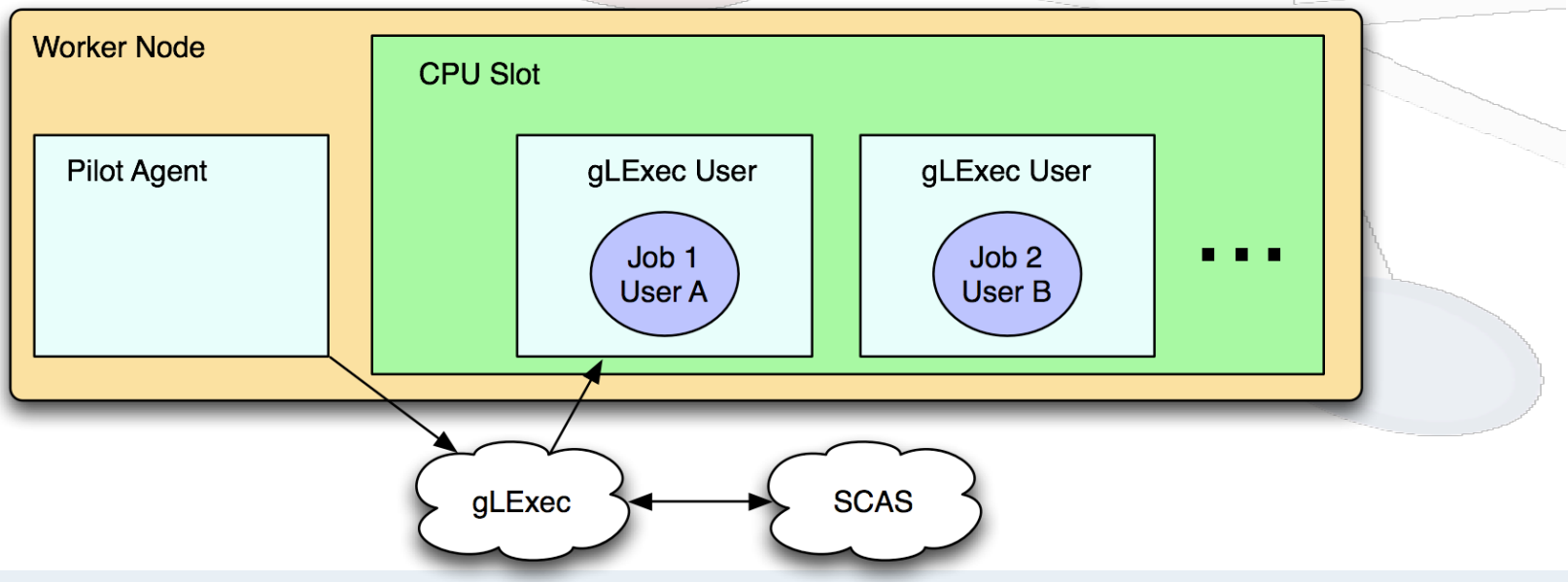
2009 Job Status Breakdown



- Points to note –
  - 20% (yellow) of jobs relied on DIRAC failover mechanisms
  - 10% of jobs relying on Grid storage for input sandbox failed

Coming Soon - Filling mode

- The use of generic pilot agents provides an elegant solution to the job prioritization problem
  - Agents are sent on behalf of the VO
    - Eligible to run the tasks of any VO-member
    - Job priority applied in the central Task Queue
- Agents work in an optimized 'Filling Mode'
  - Multiple jobs can run in the same CPU slot
  - Significant performance gains for short, high priority tasks
    - Also reduces load on LCG since fewer pilots need to be submitted



CPU Benchmarking

- In order to get the most out of filling mode need a benchmarking factor for available CPUs
- Timeleft utility in DIRAC retrieves information from the local batch system to perform scaling
  - Inconsistencies were found between different instances of the same batch system software
- Requirements to Grid sites then simplified
  - Can request long (e.g. 24hrs) queues everywhere
  - Masks local batch queue waiting times

See [473] – R. Graciani Diaz

Conclusions

- DIRAC is a complete community Grid solution
- Use of the PULL scheduling coupled with the Pilot Agent paradigm offers an elegant approach to VO workload management
  - VO-policy can be applied in one central location
  - Eagerly anticipate gLExec deployment
    - Leading to filling mode by default
- DIRAC meets the requirements of the LHCb VO
  - Could also be of use to other user communities
    - e-mail: lhcb-dirac@cern.ch
    - http://dirac.cern.ch