**eGee**

# The Grid Observatory

**Grid Observatory**

**www.grid-observatory.org**

**www.eu-egee.org**

e-infrastructure

q

- *Integrate the collection of data on the behaviour of the EGEE grid and users with the development of models and of an ontology for the domain knowledge*

- **A cluster of EGEE (NA4): LRI (CNRS), Universita Piemonte Orientale, London Imperial College, ASGC**

- **With the collaboration of: Digiteo Labs, CNRS, LAL, MIS**

- **Goals**
  - Data collection and publication
    - Public access to *a posteriori* traces of EGEE activity
  - Analysis
    - Propose realistic methods to answer real operational issues
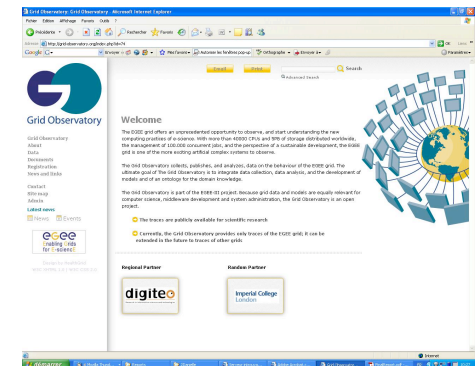  - Contribute to interactions with computer science research

- **Autonomic Computing:** *"Computing systems that manage themselves in accordance with high-level objectives from humans".* Kephart & Chess  A vision of Autonomic Computing, IEEE Computer 2003

    - Self-*: configuration, optimization, healing, protection

    - On open non steady state virtual systems

      *Data required*

- **Statistical analysis, machine learning, data mining**

- **Promote a quantitative approach**

    - Optimize for the most frequent case

    - Compare solutions on consensual datasets aka Benchmarks

**Enabling Grids for E-sciencE**

- **Acquisition, consolidation, long-term conservation of traces of EGEE activities**

- **Since October 2008: the grid observatory portal**

  ## www.grid-observatory.org

  - Heap of data « as it is »
    - Get understanding of the complexities of the data
    - Start analysis and bootstrap interactions
  - Next step (September 09): filtered data

- **Long-term goal: permanent storage of reliable, exhaustive, filtered information**

  - Reliable: data curation and provenance
  - Exhaustive: Added value in snapshots of the inputs and grid state e.g. workload and available services during a relevant time range
  - Filtered: remove redundancy

**Enabling Grids for E-sciencE**

- **Traces from**
  - The Information System (BDII)
    - LDIF format, CE and SE information, limited information about services
    - One master file each day, plus diff each 15 minutes
  - The Logging and Bookkeeping service
    - From the GRIF (Grille de Recherche Ile de France)/LAL site
    - ascii dump of the SQL tables events, short_fields and long_fields: all the events in the lifcycle of a job, tagged
  - Batch system(s)
    - From the GRIF/LAL site
    - Job controller traces allow to recover the grid identifier
  - WMS internals (condorG, wm_proxy, etc.)
    - From the GRIF/LAL site
    - Possible application: disappearing jobs
- **RTM data**
    - Summary of the lifecycle of jobs from the Real Time Monitor project
- **Next: Storage traffic – open issue: how to relate them to user files**
- **Outside the scope: external traffic on shared resources**

- **Clustering the jobs timestamps by StrAP, an on-line version of the Affinity Propagation algorithm**
- **Clusters described by real jobs, the exemplars**
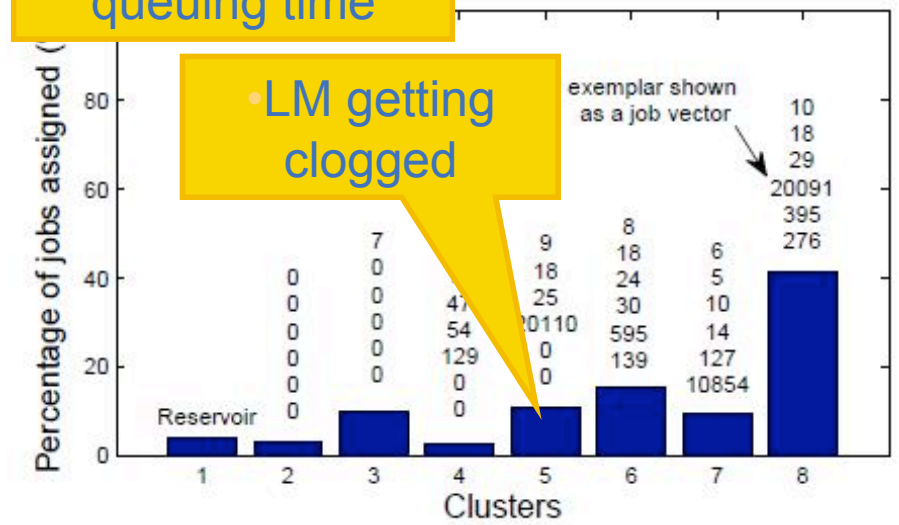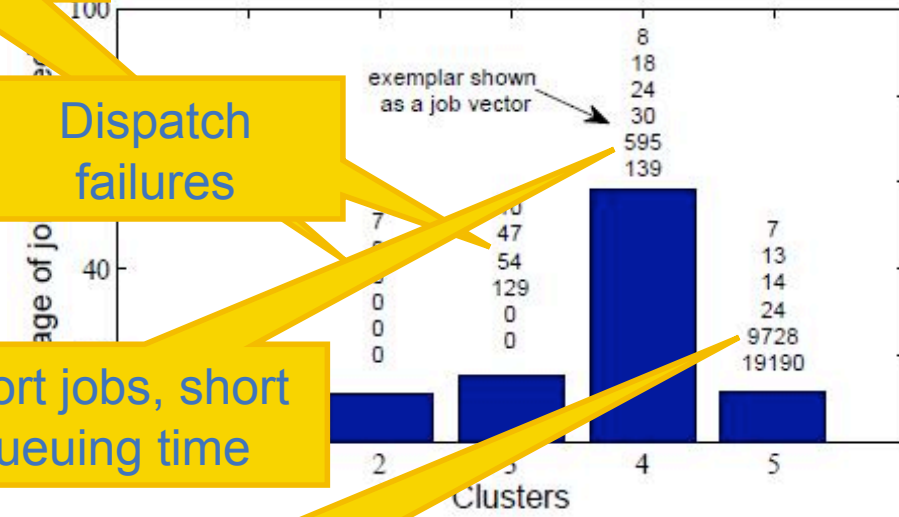
Submission failures

Dispatch failures

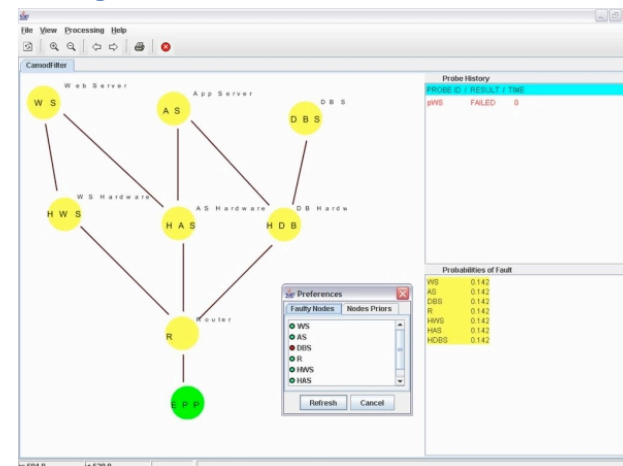Short jobs, short queuing time

Long jobs, long queuing time

LM getting clogged

- **Data acquisition**
  - Tagged data
- **Post-mortem analysis**
- **"Smart" monitoring**
  - e.g. intelligent and dynamic troubleshooting as IBM RAIL
  - Stream mining
  - …
  - new with EGEE: scaling issues

IBM Real-Time Active Inference and Learning (RAIL) based on I. Rish reseacrh work



- **Towards a knowledge base of "interesting" use cases**
  - https://savannah.cern.ch/projects/goreq/
- **In practice for next year: internships**