LHCb Roadmap 2009-10

DIRAC

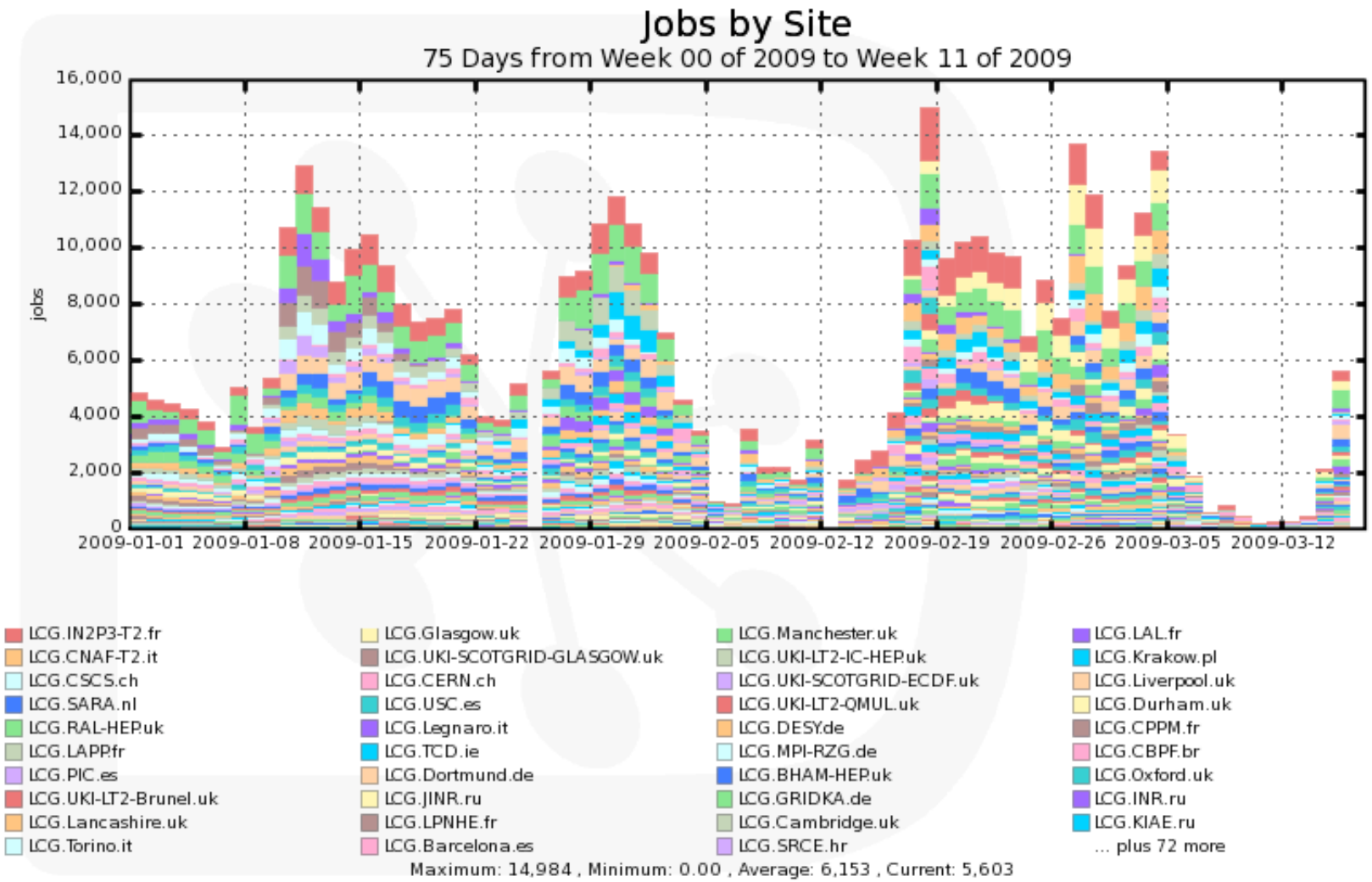
COMMUNITY GRID SOLUTION ™

2008: DIRAC3 put in production

- Production activities
  - Started in July
  - Simulation, reconstruction, stripping
    - Includes file distribution strategy, failover mechanism
    - File access using local access protocol (rootd, rfio, (gsi)dcap, xrootd)
    - Commissioned alternative method: copy to local disk
      - Drawback: non-guaranteed space, less CPU efficiency, additional network traffic (possibly copied from remote site)
  - Failover using VOBOXes
    - File transfers (delegated to FTS)
    - LFC registration
    - Internal DIRAC operations (bookkeeping, job monitoring…)
- Analysis
  - Started in September
  - Ganga available for DIRAC3 in November
  - DIRAC2 de-commissioned on January 12th

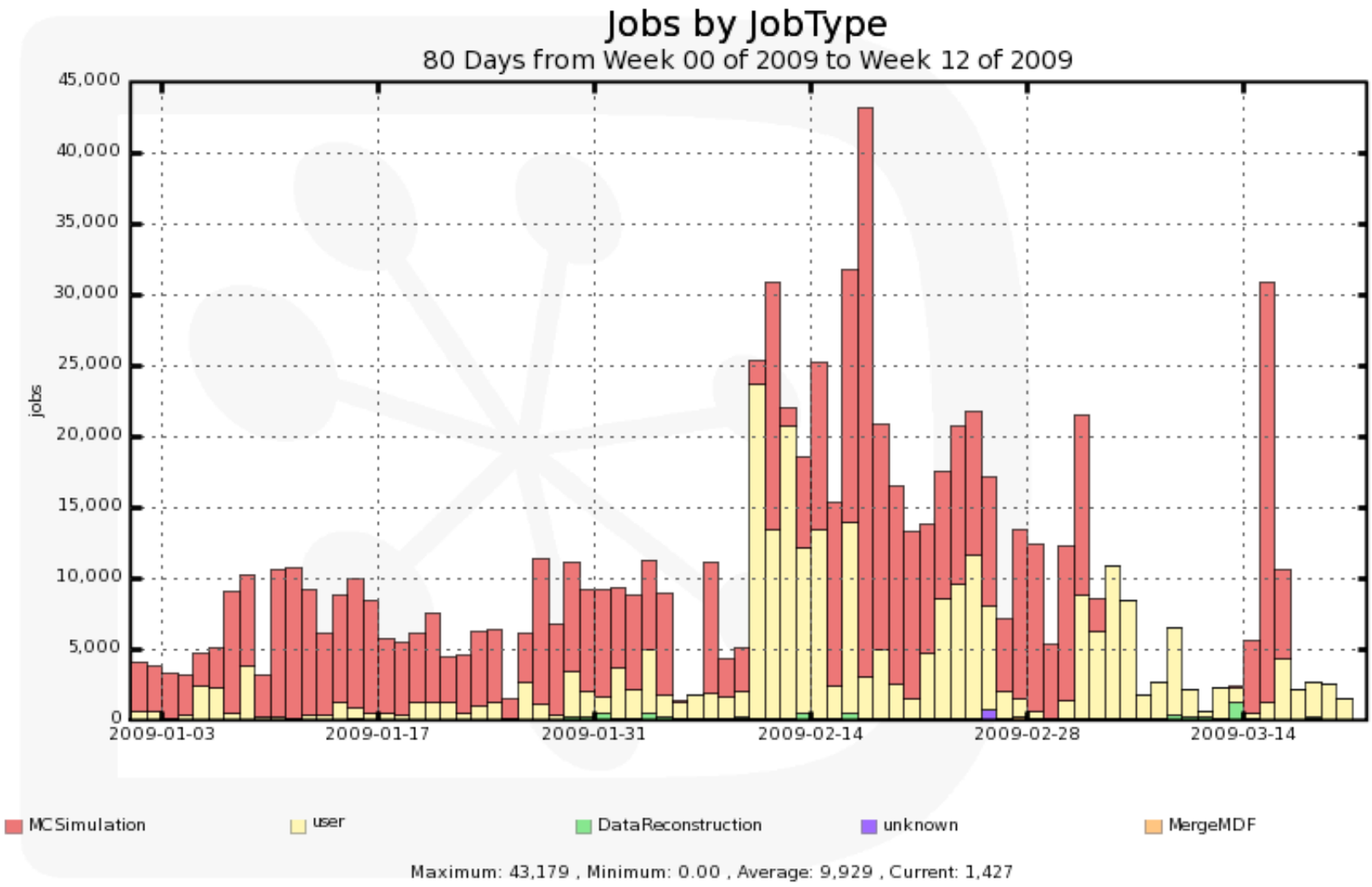
Call me DIRAC now….

2009 DIRAC concurrent jobs



111 sites

DIRAC jobs per day



LHCb Computing Operations

- Production manager
  - Schedules production work, sets up and checks workflows, reports to LHCb operations
- Computing shifters
  - Computing Operations shifter (pool of ~12 shifters)
    - Covers 14h/day, 7 days / week
    - Computing Control room (2-R-014)
  - Data Quality shifter
    - Covers 8h/day, 7 days / week
  - Both are in the LHCb Computing Control room (2-R-014)
- Daily DQ and Operations meetings
  - Week days (twice a week during shutdowns)
- Grid Expert on-call
  - On duty for a week
  - Runs the operations meetings
- Grid Team (~6 FTEs needed, ~2 missing)
  - Shared responsibilities (WMS, DMS, SAM, Bookkeeping…)

Activities in 2008

- Completion of MC simulation called "DC06"
  - Additional channels
  - Re-reconstruction (at Tier1s)
    - Involves a lot of pre-staging (2 years old files)
  - Stripping (at Tier1s)
- User analysis of "DC06"
  - At Tier1s, using ganga and DIRAC (2, then 3)
    - Access to D1 data (some files are 2 years old)
- Commissioning for 2008 data taking
  - CCRC'08 (February, May)
    - Managed to distribute data at nominal rate
    - Automatic job submission to Tier1s
    - Re-processing of data still on disk
  - Very few cosmics data (only saved at Tier0, analysed online)
  - First beam data
    - Very few events (rate: 1 event / 48 seconds…)

Plans for 2009

- Simulation… and its analysis in 2009
  - Tuning stripping and HLT for 2010 (DC09)
    - 4/5 TeV, 50 ns (no spillover), 10^{32} $cm^{-1}s^{-1}$
    - Benchmark channels for first physics studies (100 Mevts)
      - $B \rightarrow \mu\mu$, Γ_s, $B \rightarrow Dh$, $B_s \rightarrow J/\psi\phi$, $B \rightarrow K^*\mu\mu$ …
    - Large minimum bias samples (~ 1mn of LHC running, 10^9 events)
    - Stripping performance required: ~ 50 Hz for benchmark channels
    - Tune HLT: efficiency vs retention, optimisation
  - Replacing DC06 datasets (DC09-2)
    - Signal and background samples (~500 Mevts)
    - Minimum bias for L0, HLT and stripping commissioning (~100 Mevts)
    - Used for CP-violation performance studies
    - Nominal LHC settings (7 TeV, 25 ns, 2 10^{32} $cm^{-2}s^{-1}$)
  - Preparation for very first physics (MC-2009)
    - 2 TeV, low luminosity
    - Large minimum bias sample (10^9 events, part used for FEST'09)
- Commissioning for 2009-10 data taking (FEST'09)
  - See next slides

- Aim
  - Replace the non-existing 2008 beam data with MC
  - Learn on how to desal with "real" data
    - HLT strategy: from 1 MHz to 2 kHz
      - First data (loose trigger)
      - Higher lumi/energy data (b-physics trigger)
    - Online detector monitoring
      - Based on event selection from HLT e.g. J/Psi events
      - Automatic detector problems detection
    - Online Data streaming
      - Physics stream (all triggers) and calibration stream (subset of triggers, typically 5 Hz)
    - Alignment and calibration loop
      - Trigger re-alignment
      - Run alignment processes
      - Validate new alignment (based on calibration stream)
    - Feedback of calibration to reconstruction
    - Stripping, streaming, data merging and distribution
    - Physics Analysis (group analysis, end-user…)

FEST'09 preparation (2)

- Online developments
  - Event injector
    - Read MC files with emulated L0 trigger
    - Creates Multi-Event Packets (MEP as font-end does)
    - Send MEP to an HLT farm node
  - Event injector control system
    - Emulation of the standard Run Control
    - Simulates a regular run, but using event injector as source
  - Multiple online streams
    - Using HLT classification as criterion
      - Was not needed for 2008 run, hence was delayed
  - Status
    - Tests in December, operational in January
    - First FEST week: 26 January
      - Mainly online commissioning, limited data transfers
    - Second FEST week: 2 March
      - Data Quality commissioning, feedback to reconstruction

Resources (preliminary)

- Consider 2009-10 as a whole (new LHC schedule)
  - Real data
    - Split year in two parts:
      - 0.5 10^6 s at low lumi – LHC-phase1
      - 3 to 6 10^6 s at higher lumi (1 10^{32}) – LHC phase2
    - Trigger rate independent on lumi and energy: 2 kHz
  - Simulation: 2 10^9 events (nominal year) in 2010
- New assumptions for (re-)processing and analysis
  - More re-processings during LHC-phase1
  - Add calibration checks (done at CERN)
  - Envision more analysis at CERN with first data
    - Increase from 25% (TDR) to 50% (phase1) and 35% (phase2)
    - Include SW development and testing (LXBATCH)
  - Adjust event sizes and CPU needs to current estimates
    - Important effort to reduce data size (packed format for rDST, DST, μDST…)
    - Use new HEP-SPEC06 benchmarking

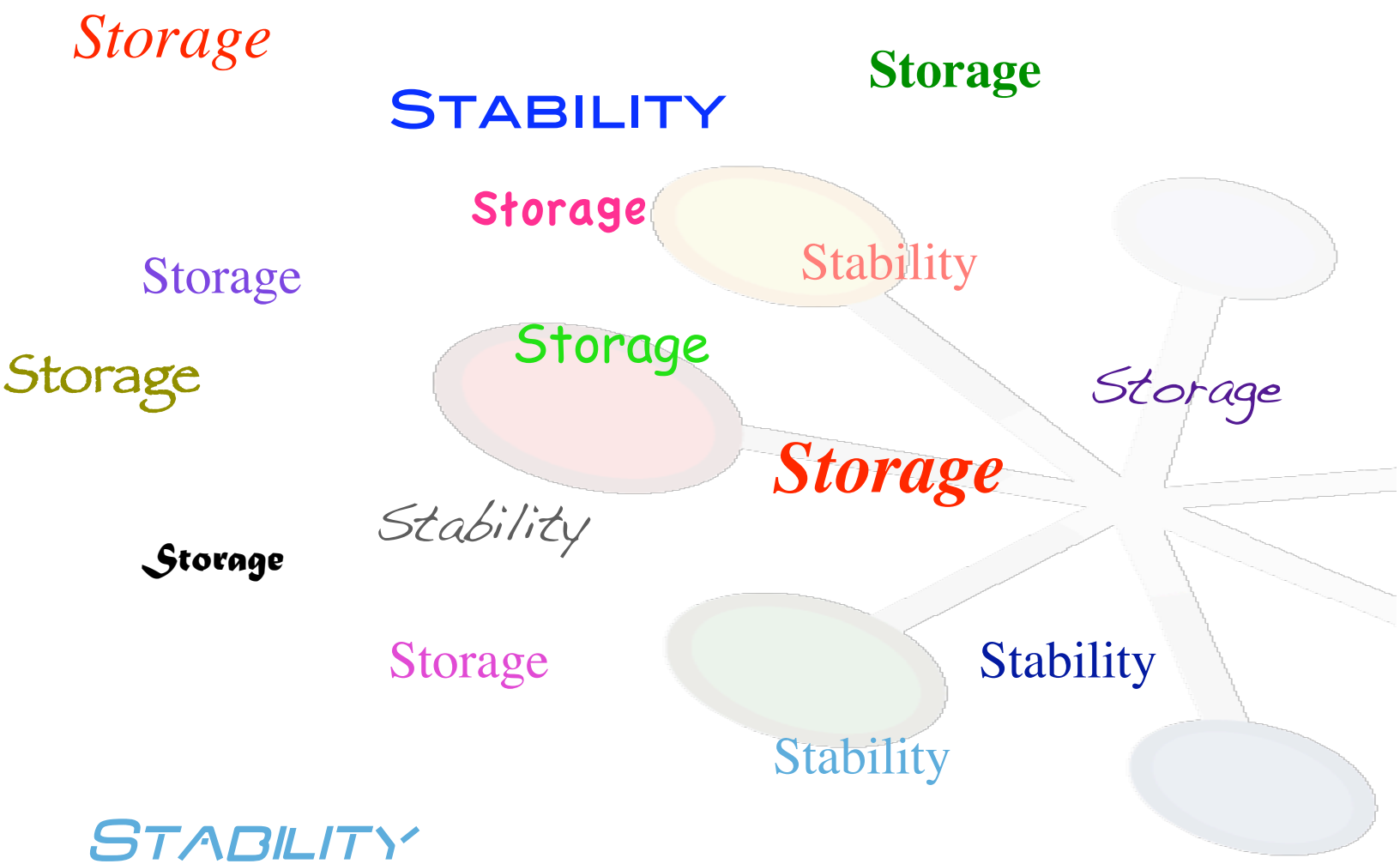
Resources (cont'd)

- CERN usage
  - Tier0:
    - Real data recording, export to Tier1s
    - First pass reconstruction of ~85% of raw data
    - Reprocessing (in future foresee to use also the Online HLT farm)
  - CAF ("Calibration and Alignment Facility")
    - Dedicated LXBATCH resources
    - Detector studies, alignment and calibration
  - CAF ("CERN Analysis Facility")
    - Part of Grid distributed analysis facilities (estimate 40% in 2009-10)
    - Histograms and interactive analysis (lxplus, desk/lap-tops)
- Tier1 usage
  - Re[-re]construction
    - First pass during data taking, reprocessing
  - Analysis facilities
    - Grid distributed analysis
    - Local storage for users' data (LHCb_USER SRM space)
  - Simulation in 2009 (background activity)

Resource requirements trends

- Numbers being finalised for MB meeting and C-RRB
- Trends are:
  - Shift in tape requirements due to LHC schedule
  - Increase in CERN CPU requirements
    - Change in assumptions in the computing model
  - Tier1s:
    - CPU requirements lower in 2009 but similar in 2010
      - More real data re-processings in 2010
    - Decrease in disk requirements
  - Tier2s:
    - CPU decrease due to less MC simulation requests in 2009
- Anyway:
  - All this is full of many unknowns!
    - LHC running time
    - Machine background
    - Number of re-processings (how fast can we calibrate?)
  - More than anything hard to predict needed power and space as function of time! Only integrated CPU, final storage estimates

What are the remaining issues?



Storage and data access

- 3 years after Mumbai
  - Definition of storage classes
  - Roadmap to SRM v2.2
- Where are we?
  - Many scalability issues
    - We do use and only use SRM
    - Data access from storage (no local copy)
  - Instabilities of storage-ware (and DM tools)
    - Delay in coping in changes (inconsistent tools)
  - Data disappearance….
    - Tapes damaged
    - Disk servers offline
  - Still no unified RFIO library between Castor and DPM…
- What can be done?
  - Regular meetings between experiments DM experts, sites and storage-ware developers
    - Pre-GDB resurrected?
    - Should be technical, not political

Storage and Data Access (2)

- Reliability of data access?
  - We (experiments) cannot design sites storage
  - If more hardware is needed, should be evaluated by sites
    - Flexible to changes
    - Number of tape drives, size of disk caches, cache configuration…
    - Examples:
      - Write pools different from read pools:
        - Is it a good choice?
        - How large pools should be
      - Scale number of tape drives to disk cache and staging policy
- Consistency of storage with catalogs
  - Unaccessible data (tape or disk)
  - Job matching based on Catalog
    - For T1D0 data, we use pre-staging: ensures availability of data
      - Spot lost files
    - For D1 data, we assume it is available
      - We can query SRM, but will collapse
      - Will SRM reply the truth, i.e. "UNAVAILABLE"?
      - We often can get a tURL, but opening file just hangs…

Software repository and deployment

- Very important service:
  - Can make a site unusable!
  - Should scale with number of WNs
  - Use proper technology
    - Example: at CERN LHCb has 1 write AFS server and 4 read-only AFS servers
  - Of course proper permissions should be set…
    - Write to lcg-admin (a.k.a. sgm accounts)
    - Read-only to all others
    - Make your choice: pool accounts and separate groups or single accounts
  - Intermittent outages can kill all jobs on a site!
- Middleware client
  - We do need support for multiplatform
    - Libraries linked to applications (LFC, gfal, castor, dCache…)
  - Therefore we must distribute it
    - LCG-AA distribution is primordial

Workload Management

- Stability and reliability of gLite WMS
  - Mega-patch is not a great experience…
  - In most cases we don't need brokering
    - Next step is direct CE submission (CREAM)
      - Need a reliable CEMON information
- Job matching to WNs: shopping list
  - MaxCPUTime matching: which units?
    - Is it guaranteed?
  - Memory usage
    - We are very modest memory consumers, but…
    - Jobs are often killed by batch systems due to excessive memory (virtual)
    - There is no queue parameter allowing a JDL requirement
      - Only indication on WN memory
    - Some sites have linked memory to CPU!!!
      - Seem strange… short jobs all fail…
    - Limits should be increased
      - Can bias physics results (e.g. large number of particles in Geant4)
    - CPUs with (really) many cores are almost here…

SAM jobs and reports

- Need to report on usability by the experiments
  - Tests reproduce standard use cases
  - Should run as normal jobs, i.e. not on special clean environment
- Reserve lcg-admin for software installation
  - Needs dedicated mapping for permissions to repository
- Use normal accounts for running tests
  - Running as "Ultimate Priority" DIRAC jobs
  - Matched by the first pilot job that starts
    - Scans the WN domain
      - Often see WN-dependent problems (bad config)
    - Regular environment
  - Should allow for longer periods without report
    - Queues may be full (which is actually good sign) but then no new job can start!

Conclusions

- 2008
  - CCRC very useful for LHCb (although irrelevant to be simultaneous due to low throughput)
  - DIRAC3 fully commissioned
    - Production in July
    - Analysis in November
    - As of now, called DIRAC
  - Last processing on DC06
    - Analysis will continue in 2009
  - Commission simulation and reconstruction for real data
- 2009-10
  - Large simulation requests for replacing DC06, preparing 2009-10
  - FEST'09: ~1 week a month and 1 day a week
  - Resource requirements being prepared for C-RRB in April

Services are not stable enough yet!