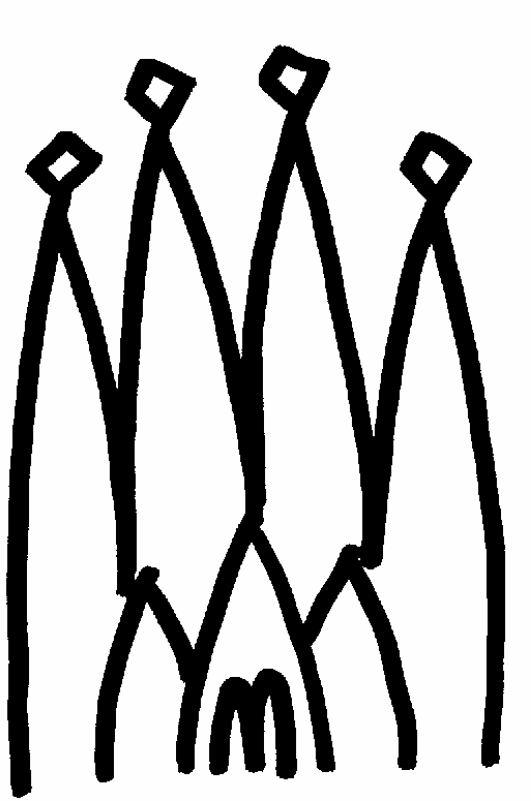
LHCb plans for 2007 until first beam

WLCG Workshop, January 22-26 2007

Ph.Charpentier, CERN - LHCb



QuickTime™ et un
décompresseur TIFF (non compressé)
sont requis pour visionner cette image.

01101001110101
10101010001101
01010110100
Boole

QuickTime™ et un
décompresseur TIFF (LZW)
sont requis pour visionner cette image.

LHCb Computing model (1)

- DAQ recording rate: ~2000 evts/s
  - Small events : 35 kB
  - ~ 2GB files : 1 file every 30 seconds at nominal rate
  - 70 MB/s recording rate
    - including machine duty cycle: 35 MB/s
- Data files replicated to Tier0 (Castor)
  - Dedicated storage class (Tape1Disk0)
- Reconstruction
  - Files processed at Tier0 and Tier1s
    - IN2P3, CNAF, RAL, GridKa, NIKHEF, PIC
    - sharing according to the local CPU power available to LHCb
  - Files replicated to one of the 6 LCHb Tier1s
    - CERN's share also replicated for resilience
  - Storage class needed at all Tier1's: Tape1Disk0
    - Files pinned until processed

LHCb Computing Model (2)

- Reconstruction (cont'd)
  - CPU resources goal: 2.4 kSI2K.s (currently ~7 kSI2K.s)
    - Reconstructing one file would take 44 hours on a Xeon 3 GHz
- Reconstructed data (rDST)
  - Size: 20 kB/evt (rDST size ~ □Raw size)
  - Stored locally at each Tier1 + Tier0 (file size ~ 1.2 GB)
  - Storage class: Tape1Disk0
  - rDST pinned until physics pre-selection done (stripping)
- Physics pre-selection (stripping)
  - Data streaming according to physics channels
    - ~10 streams (grouping similar channels)
    - Each stream ~ 1% of the initial sample (factor 10 total reduction)
  - CPU resources goal: 0.2 kSI2K.s
    - Stripping one file would take 4 hours on a Xeon 3GHz
    - most probably strip ~10 files at once

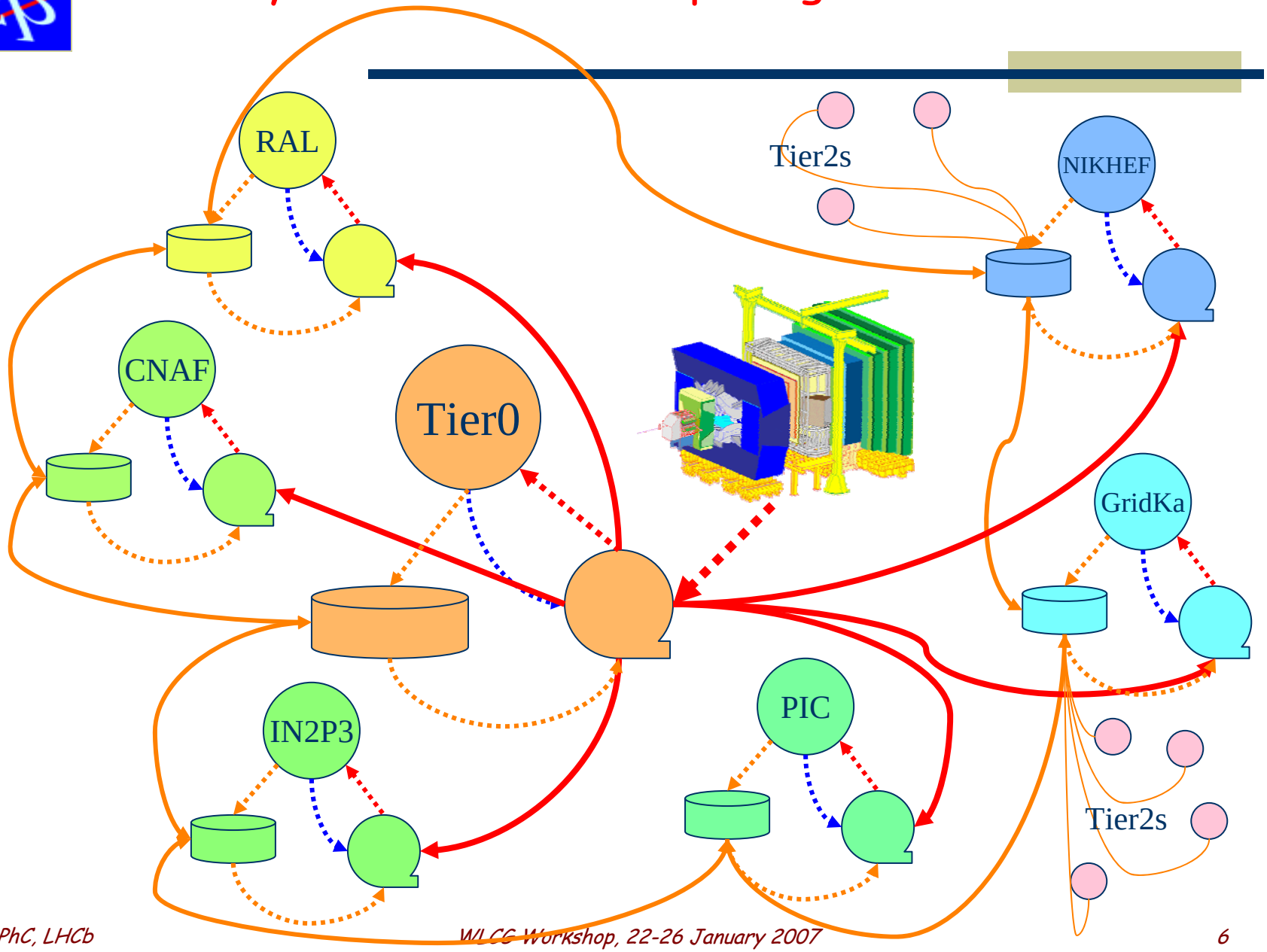
LHCb Computing model (3)

- Streamed data (DST)
  - Size: 100 kB/evt
  - DST sizes: ~ 600 MB/stream
  - DSTs replicated at ALL Tier1s
  - Storage class: Tape1Disk1 at production site, Tape0Disk1 at all other sites
    - Transition Tape1Disk1 -> Tape1Disk0 foreseen when the dataset becomes obsolete
- Simulation
  - Simulation + reconstruction running at all sites (low priority at Tier1s)
  - Producing DSTs (400 kB/evt) stored at associated Tier1s
    - Tape1Disk1 at the primary Tier1
    - One copy as Tape0Disk1 at another Tier1

LHCb Computing Model (4)

- Analysis
  - Performed on stripped DSTs (i.e. on disk)
    - Small data samples
  - Performed at all Tier1s + CAF (+ a few large Tier2s with enough disk capacity)
  - Fully distributed analysis until final microDST / NTuple
  - Final analysis (interactive) on desktop or interactive cluster
  - Part of the analysis doesn't include data processing
    - large "toy-MC" studies
    - data fitting procedures
    - can easily be performed at any site using the Grid
- Alignment / calibration
  - Mainly at CAF and LHCb-dedicated cluster (part of the pit infrastructure)

Summary of the LHCb Computing model



DIRAC Workload Management System

- Central task queue
  - Optimisers (data, job type, CPU requirements…)
  - Job queues for each set of requirements
    - Fast matching
- Pilot agents
  - Submission driven by jobs in the queues
  - Runs with user's credentials
  - Submitted to LCG-RB or gLite-WMS
  - On WN: matches job in the central task queue
- Job wrapper
  - Installs software (if needed)
  - Executes workflow
  - Uploads output data, log files, output sandbox
  - Submits additional data transfer requests

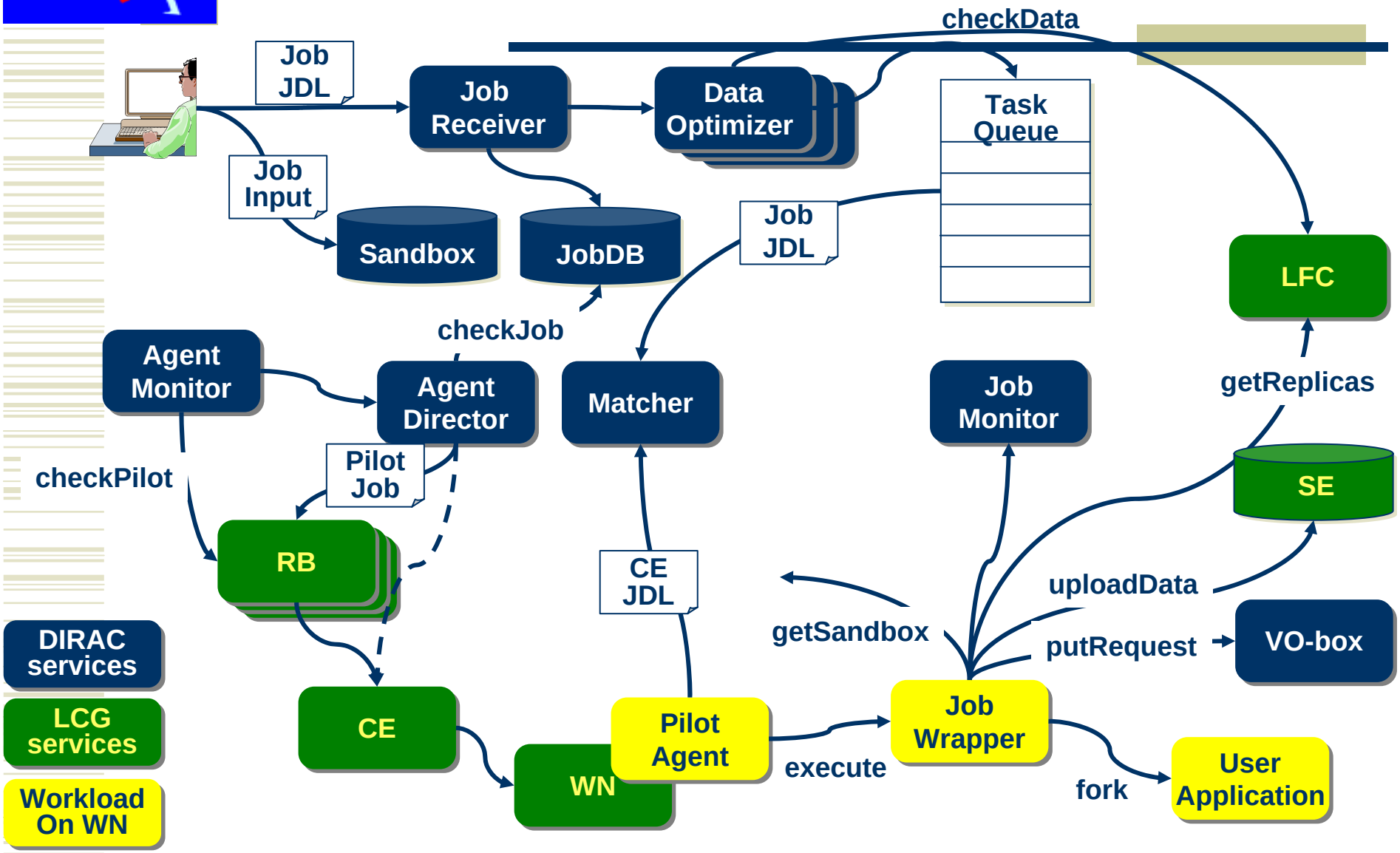
DIRAC Production Management tools

- Preparation of (complex) workflows
- Instantiation of production templates
  - Specializations of workflow parameters
    - SW versions, event types, number of events
    - Input data
- Submission of productions to the DIRAC WMS
  - set of jobs sharing the workflow but with incremental parameter (event number, dataset…)
- Definition of data transformations
  - Acts on specific datasets (from a FC-like database)
  - Instantiates production jobs
  - Allows automatic job submission (input data driven)
    - Input data can be defined manually or by completing jobs

DIRAC Data Management System

- Replica Manager
  - Takes care of replicating data to a defined SE
  - Retries until success
  - For initial upload: failover mechanism
    - File uploaded to another SE and a movement request is queued in a VOBox
- Transfer Agent
  - Replicates datasets between SEs
  - Uses third-party transfers (lcg-cp or equivalent)
- FTS transfers
  - Driven by a Data-WMS
  - Bulk transfer requests (possibly generated by a transformation request)
  - Used for transfers involving Tier0 and Tier1s
    - For the time being, single file replication handled by transfer agent

DIRAC workload management



Software installation

- Handled by the LHCb-specific SAM test jobs
- Uses the VO_LHCB_SW_DIR shared area
  - Reliability is a must as all jobs refer to it
- Checks that active versions of the LHCb SW are installed
  - if not, installs them and runs a short simulation+reconstruction job (few events)
  - installs LCG-AA software as well as OS- and compiler-dependent libraries
    - allows to run on non-supported platforms
- On sites without shared area
  - each job installs the SW in the working area
  - same if the required version is not installed (per project check)
- Typical size of one version (simulation+reconstruction+Gaudi+LCG)
  - 1 GByte

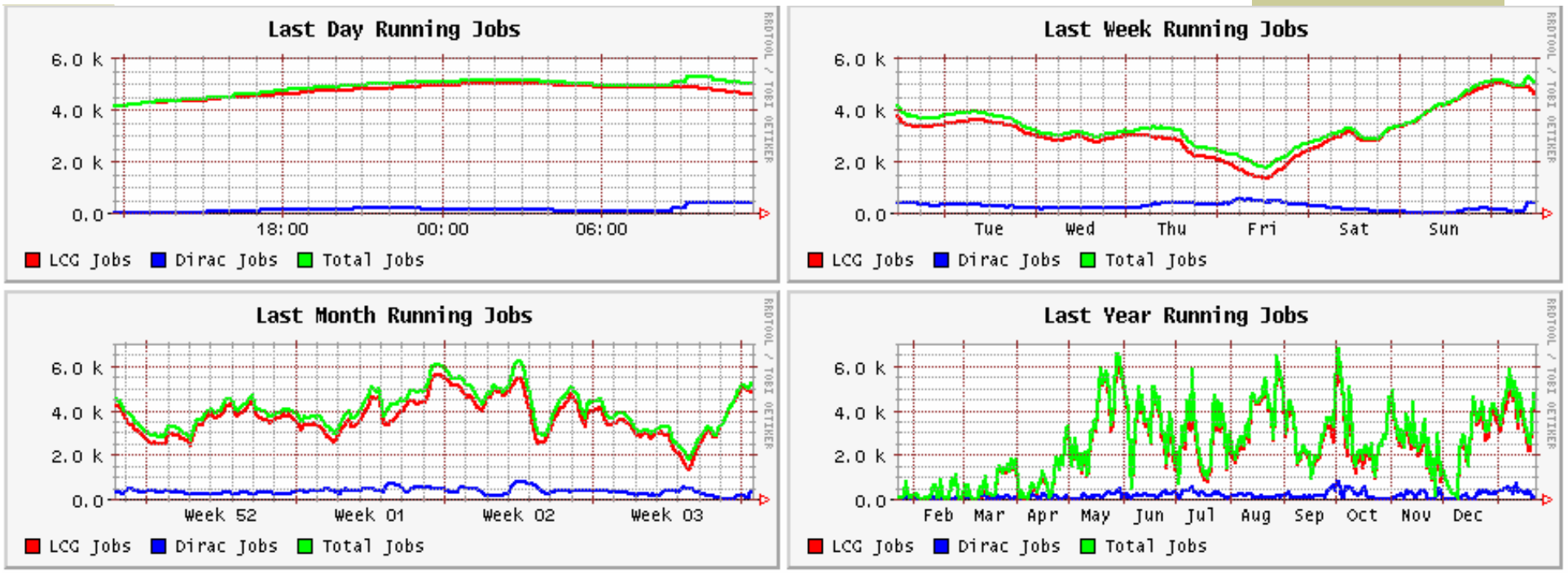
Data access from the WN

- Two possibilities
  - Copy input data to the working directory
    - caveat: not necessarily enough space
    - waiting for staging
  - Access data from the SE
    - protocols: rootd, rfio, dcap, gsidcap
    - scales to large datasets
    - parallel staging strategies can be implemented
- LHCb's baseline is to access data from the SE
  - Input files declared as LFNs
  - tURL obtained using lcg-utils (lcg-lr and lcg-gt)
    - Currently pre-staging before starting the application (all data are on disk, no problem)
  - Fully relying on SRM (SURLs registered in LFC)

LHCb activities in late 2006

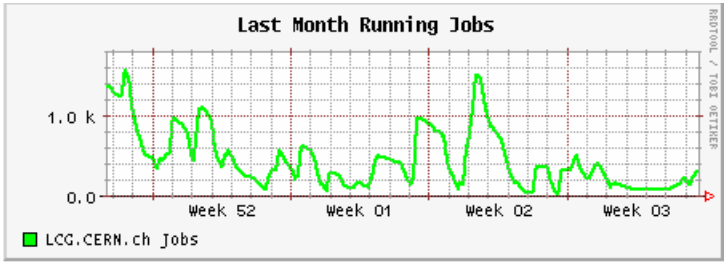
- DC06 in 2006
  - production of 140 Mevts at all sites
  - DIGI files (raw data + MV truth) all at CERN
  - Distribution of DIGI to Tier1s
    - share according to pledges
    - test of automatic reconstruction submission (with temporary recons)
    - many developments around Data Management
      - transfer error recovery (failover mechanism)
      - data access problems (gsidcap usage + NIKHEF/SARA access, SE configuration, new versions of Castor and dCache)
  - many small productions for validation of the "physics quality" reconstruction
- Conditions Database replication
  - tested at RAL, IN2P3, GridKa (CNAF joining, waiting for NIKHEF and PIC)
- LFC replication tests
  - successful between CERN and CNAF

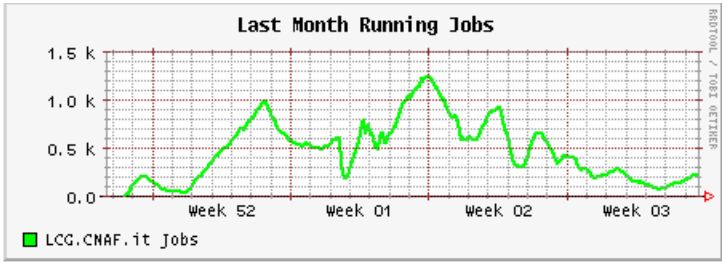
CPU usage for C06

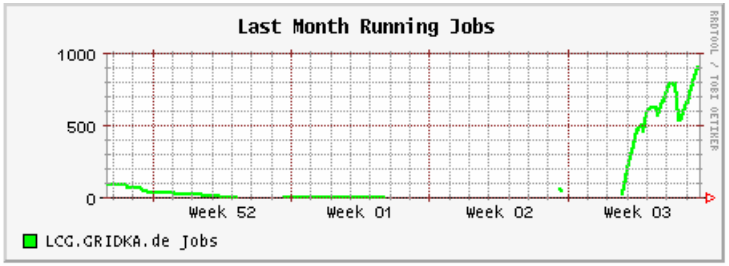


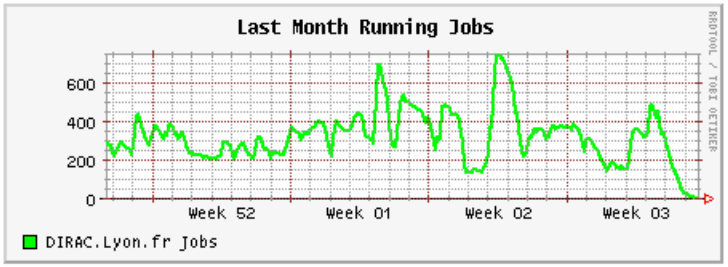
- On average 3000 jobs running since May '06
- Fluctuations
  - partly due to LHCb' activity
  - largely due to "fair share" effects

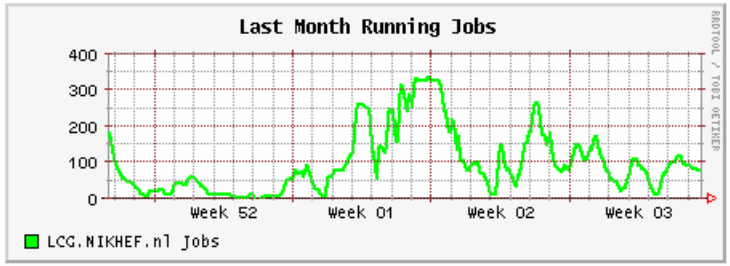
Tier1s in DC06

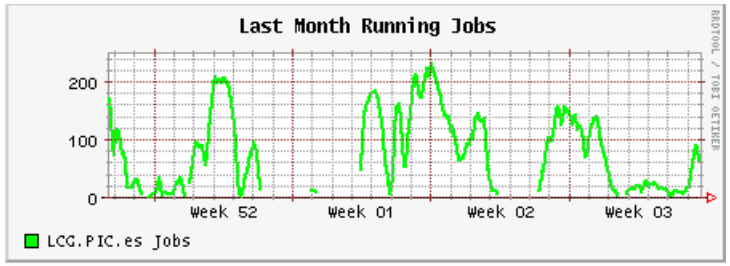


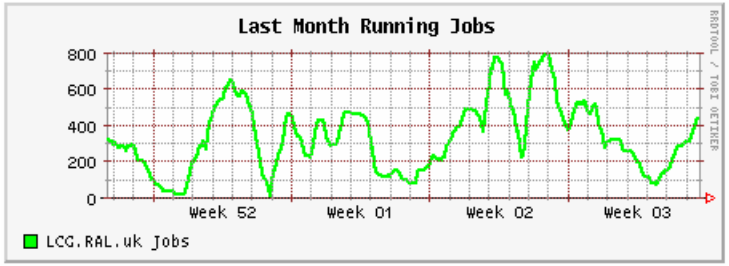




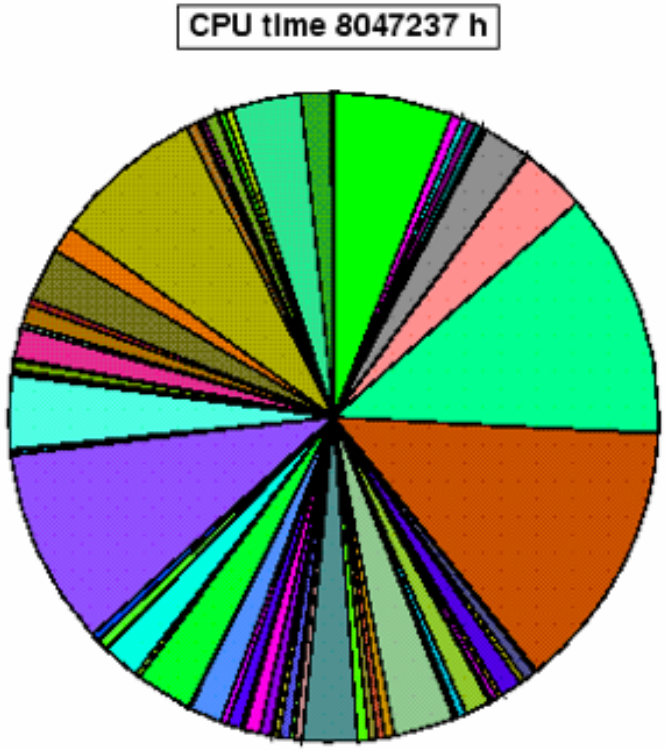








Sites in DC06



DIRAC.LHCBONLINE.ch	0.00%
DIRAC.Lyon.fr	5.92%
DIRAC.NIPNE.ro	0.00%
DIRAC.UCD.ie	0.00%
DIRAC.Zurich-MH.ch	0.54%
DIRAC.Zurich.ch	0.33%
DIRAC.joel.ch	0.00%
LCG.ACAD.bg	0.40%
LCG.AUVER.fr	0.35%
LCG.BHAM-HEP.uk	0.14%
LCG.BIFI.es	0.10%
LCG.Barcelona.es	2.33%
LCG.Bari.it	0.06%
LCG.Bologna.it	0.01%
LCG.Brunel.uk	3.30%
LCG.CERN.ch	12.30%
LCG.CESGA.es	0.00%
LCG.CGG.fr	0.01%
LCG.CNAF-GRIDIT.it	0.04%
LCG.CNAF.it	13.41%
LCG.CNB.es	0.14%
LCG.CPPM.fr	0.57%
LCG.CSCS.ch	0.28%
LCG.CY01.cy	0.04%
LCG.Cagliari.it	0.00%
LCG.Cambridge.uk	0.00%
LCG.Catania.it	0.00%
LCG.Dortmund.de	1.07%
LCG.Durham.uk	0.24%
LCG.EELA-CIEMAT.es	0.30%
LCG.EELA-UFRJ.br	0.01%
LCG.ETF-RTH.lv	0.10%
LCG.Edinburgh.uk	0.05%
LCG.FESB.hr	0.00%
LCG.FORTH.gr	1.24%
LCG.Ferrara.it	0.03%
LCG.Firenze.it	0.39%
LCG.GOG.sg	0.09%
LCG.GR-01.gr	0.01%
LCG.GR-03.gr	0.14%
LCG.GR-04.gr	0.01%
LCG.GR-05.gr	0.00%
LCG.GRIDKA.de	3.01%
LCG.GRNET.gr	0.50%
LCG.Glasgow.uk	0.38%
LCG.HG-02.gr	0.40%
LCG.HG-04.gr	0.56%
LCG.HG-06.gr	2.71%
LCG.HPC2N.se	0.01%
LCG.HellasGrid.gr	0.38%
LCG.ICI.ro	0.01%
LCG.IHEP.su	0.22%
LCG.IN2P3.fr	0.52%

LCG.INR.ru	0.26%
LCG.IPP.bg	0.05%
LCG.IPSL-IPGP.fr	0.03%
LCG.IRB.hr	0.08%
LCG.ITEP.ru	0.56%
LCG.ITWM.de	0.01%
LCG.Imperial.uk	0.77%
LCG.Iowa.us	0.00%
LCG.JINR.ru	0.07%
LCG.KFKI.hu	0.12%
LCG.KIAE.ru	0.66%
LCG.KIAM.ru	0.36%
LCG.Krakow.pl	1.64%
LCG.LAL.fr	0.10%
LCG.LAPP.fr	0.01%
LCG.LISA.nl	0.00%
LCG.LPC.fr	2.80%
LCG.LPN-fails.fr	0.03%
LCG.LPN.fr	0.24%
LCG.Lancashire.uk	1.83%
LCG.LeSC.uk	0.05%
LCG.Legnaro.it	0.56%
LCG.Liverpool.uk	0.40%
LCG.Manchester.uk	9.81%
LCG.Milano.it	0.19%
LCG.NCP.pk	0.01%
LCG.NCU.tw	0.01%
LCG.NIKHEF-save.nl	0.08%
LCG.NIKHEF.nl	3.65%
LCG.NIPNE.ro	0.09%
LCG.Napoli-Atlas.it	0.09%
LCG.Napoli.it	0.02%
LCG.OU.il	0.08%
LCG.Oxford.uk	0.47%
LCG.PAKGRID.pk	0.01%
LCG.PDC.se	0.12%
LCG.PIC.es	1.51%
LCG.PNPI.ru	0.23%
LCG.Padova.it	0.92%
LCG.Pisa.it	0.39%
LCG.QMUL.uk	2.60%
LCG.RAL-HEP.uk	1.38%
LCG.RAL.uk	7.55%
LCG.RHUL.uk	0.58%
LCG.SARA.nl	0.18%
LCG.SINP.ru	0.25%
LCG.SRCE.hr	0.00%
LCG.Sheffield.uk	0.62%
LCG.Sofia.bg	0.00%
LCG.TAU.il	0.09%
LCG.TCD.ie	0.01%
LCG.Torino.it	0.32%
LCG.UCL-CCC.uk	0.00%
LCG.ULAKBIM.tr	0.39%
LCG.USC.es	3.42%
LCG.WARSAW.pl	1.51%
LCG.WCSS.pl	0.04%
LCG.WEIZMANN.il	0.07%

~ 100 sites

DC06 in 2007

- Reconstruction of 140 M b-events, 100 M min. bias
  - already replicated on Tier1s (as part of DC06 transfer)
  - 20 input files, rDST stored locally at Tier1
- Simulation and reconstruction of 74 Mevts
  - at all sites
  - 500 events
  - DST replicated at 7 sites
- Stripping of reconstructed b-events and min. bias
  - at Tier1s, reading rDST and DIGI
  - 10 input files, DST replicated at 7 sites
- Analysis for the LHCb "physics book"
  - at Tier1s, using Ganga submitting jobs to DIRAC
  - still some analysis using "traditional" batch analysis (LXBATCH)

January	CPU (kSI2k. months)	New disk (TB)	New tape (TB)	DC06 accumul-ated disk (TB)	DC06 accumul-ated tape (TB)
CERN	385.6	12.1	4.2	23.8	55.7
Lyon	416.5	12.1	4.4	23.8	16.7
FZK	234.8	12.1	3.2	23.8	12
CNAF	219.4	12.1	3.1	23.8	11.7
NIKHEF/ SARA	385.6	12.1	4.2	23.8	13.8
PIC	115	12.1	2.4	23.8	11.9
RAL	234.8	12.1	3.2	23.8	13.9
“other”	1933.1				

CPU:

40M signal evts simul & recons
33M bb-incl recons
50M bb-incl strip

Disk:

DST o/p of signal recons
DST o/p of bb-incl stripping

Tape:

DST o/p of bb-incl
DIGI o/p of signal sim

February	CPU (kSI2k. months)	New disk (TB)	New tape (TB)	DC06 accumul-ated disk (TB)	DC06 accumul-ated tape (TB)
CERN	385.6	12.1	4.2	35.9	59.9
Lyon	416.5	12.1	4.4	35.9	21.1
FZK	234.8	12.1	3.2	35.9	15.2
CNAF	219.4	12.1	3.1	35.9	14.8
NIKHEF/ SARA	385.6	12.1	4.2	35.9	18
PIC	115	12.1	2.4	35.9	14.3
RAL	234.8	12.1	3.2	35.9	17.1
“other”	1933.1				

CPU:

40M signal evts simul & recons
33M bb-incl recons
50M bb-incl strip

Disk:

DST o/p of signal recons
DST o/p of bb-incl stripping

Tape:

DST o/p of bb-incl
DIGI o/p of signal sim

March	CPU (kSI2k. months)	New disk (TB)	New tape (TB)	DC06 accumul-ated disk (TB)	DC06 accumul-ated tape (TB)
CERN	377.0	10.3	2.5	46.2	62.4
Lyon	407.9	10.3	2.7	46.2	23.8
FZK	226.2	10.3	1.5	46.2	16.7
CNAF	210.8	10.3	1.4	46.2	16.2
NIKHEF/ SARA	377.0	10.3	2.5	46.2	20.5
PIC	106.4	10.3	0.7	46.2	15
RAL	226.2	10.3	1.5	46.2	18.6
“other”	1933.1				

CPU:

40M signal evts simul & recons

Disk:

DST o/p of signal recons

Tape:

DIGI o/p of signal sim

April	CPU (kSI2k. months)	New disk (TB)	New tape (TB)	DC06 accumul-ated disk (TB)	DC06 accumul-ated tape (TB)
CERN	377.0	10.3	2.5	56.5	64.9
Lyon	407.9	10.3	2.7	56.5	26.5
FZK	226.2	10.3	1.5	56.5	18.2
CNAF	210.8	10.3	1.4	56.5	16.2
NIKHEF/ SARA	377.0	10.3	2.5	56.5	20.5
PIC	106.4	10.3	0.7	56.5	15
RAL	226.2	10.3	1.5	56.5	18.6
“other”	1933.1				

CPU:

40M signal evts simul & recons

Disk:

DST o/p of signal recons

Tape:

DIGI o/p of signal sim

Changes w.r.t. initial plans

- Reconstruction and signal simulation
  - proceeding as foreseen, using all resources available
  - started just before Christmas break
- Stripping
  - delayed to March
  - late availability of high performance pre-selection SW
  - priority given to getting reconstruction running at all Tier1s
- Overall same resource needs as planned
  - signal simulation earlier
  - stripping later

Current challenges: data access

- Procedure (ex. of reconstruction job)
  - at start of job: 20 "lcg-gt" (all input files)
    - if needed, stages files (all should still be on disk however)
    - creates a POOL XML catalog
    - protocol precedence: dcap, gsidcap, root (castor), rfio
  - application processes 20 files
- Issues encountered
  - castor2 configuration at CNAF
    - very good response from site, work-around found, waiting for new version of Castor2 (dead tiem in job submission)
  - gsidcap at NIKHEF
    - needed dCache 1.7
    - available in November (deployed at all sites)
  - dCache instabilities at sites
  - only last week 7 sites working simultaneously

Current challenges: data replication

- Procedure
  - in job: replication to local/associated Tier1 (WN to SE)
  - replication request queued in VO-box
    - transfer agent on VO-box: "lcg-cp" (3rd party transfer)
- Issues
  - SE availability…
- Work-around
  - in job
    - temporary replication to a fail-over disk SE (all Tier1s)
    - replication to final destination queued in VO-box
  - in VO-box
    - retry until transfer is successful (transfer re-queued if failed)
  - extremely reliable
    - caveat: many transfer requests accumulating in VO-box
    - solution: multi-threaded transfer agent (just released)

Data Management issues to come

- RAL
  - migration from dCache to Castor2
    - needs to be thoroughly tests (cf CNAF)
- CNAF
  - deployment of new version of Castor2
    - should avoid problems with LSF queues
    - to be tested, remove patch on job submission dead time
- PIC
  - migration to Castor2?
- All sites
  - SRM v2.2 migration
  - make best usage of Tape0(1)Disk1 storage
    - avoid disk-to-disk copy
  - Lack of bulk data deletion (SRM + LFC)

Other coming issues

- SLC4 migration
  - see panel discussion tomorrow
  - OK for LHCb applications
  - OK for DIRAC
  - Issues
    - unavailability of middleware
    - only using middleware in compatibility mode (lcg-utils)
    - gfal, lfc missing (due to globus/VDT)
    - impossible to deploy applications depending directly on middleware
      - SRM-enabled Gaudi only on slc3
- gLite migration
  - moving from LCG-RB to gLite WMS (now)
  - testing gLite-CE when available (beware instability!)

Other coming issues (2)

- Improvements in DM tools
  - migration to SRM v2.2
  - additional functionality in lcg-utils (bulk requests)
    - LHCb can contribute in defining and implementing
- Testing middleware
  - need for frequent client release, deployable by VOs
    - needed for testing fast (bug fixes, new required features…)
  - discussion ongoing with GD
- Need to better debug what is going on
  - interactive access to WNs at Tier1s for a few users
  - access to log files (RB, CE…)

Analysis

- GANGA
  - LHCb-ATLAS joint development
  - sponsored primarily by GridPP and ARDA
  - now many developers
  - frequent tutorial sessions (ATLAS and LHCb)
  - more and more users
- LHCb usage
  - Ganga submitting to DIRAC
  - separate DIRAC-WMS instance
  - pilot agents submitted with original user's credentials
    - caveat: pilot agents competing with production ones
    - current solution: use shorter queues
      - thanks to sites who set them up
  - main issue: data access…

Job priorities

- Issue
  - need for prioritization between analysis / reconstruction / simulation
  - currently only effective for production jobs (recons. / simul.)
    - jobs in the same WMS / task queue
- Request
  - benefit from the pilot agent + central taks queue strategy
  - generic pilot agents (specific credentials)
    - selects highest priority matching job
    - logs change of identity, delegates execution to original user's credentials
    - can possibly execute more than one payload if allowed time permits
  - glexec on WN being prototyped at Lyon, GridKa
    - second priority activity…

Databases

- See Marco Clemencic's and Barbara Martinelli's talks
  - DB BOF session (next to come)
- ConditionDB
  - COOL-based
  - needed at all Tier1s (scalability)
    - for reconstruction and analysis (local usage, with fail-over possibility)
  - for simulation
    - use SQLite DB slice, installed with SW
  - master DB at CERN
  - replication using 3D's ORACLE streaming
- LFC
  - scalability and fail-over, at Tier1s (read-only)
  - using 3D's streaming
  - currently at CNAF, to be deployed at all Tier1s
    - improved reliability of LFC access

Alignment challenge

- Goal
  - simulation with mis-aligned detector (few 100 k events)
  - alignment procedure used on subset to extract parameters
  - alignment inserted into the CondiionDB (master at CERN)
  - tested on remaining data (replicated at all Tier1s)
    - test conditional submission of jobs (only when DB updated)
- Time-table
  - starting in April 2007
  - ORACLE streaming needs to be tested in February on all sites
    - currently missing: NIKHEF and PIC

Dress-rehearsal

- Testing full chain
  - from DAQ buffer to Tier0
  - data distribution to Tier1s
  - reconstruction + stripping at Tier0 and Tier1s
- Time-table
  - DAQ to Tier0
    - throughput tests possible as from March-April
      - goal: 70 MB/s needed (passed), aim at 200 MB/s (success)
    - data replication using the DIRAC DMS
      - LFC registration, entering files into the Processing DB
      - bookkeeping information
  - DAQ to Tier0 to Tier1
    - tests in June / July
    - use existing simulated data (merged raw data files of 2 GBytes)

Other activities

- Re-reconstruction of b- and min. bias- events
  - latest reconstruction
  - use of ConditionsDB
  - September? (low C.L.)
- Final analysis on the Grid
  - no-processing jobs
  - physics parameter extraction
  - needs tuning of DIRAC / Ganga
- November 2007
  - first data coming
  - most probably some learning phase
    - DAQ, trigger
    - calibration & alignment (at CAF)
    - multiple reconstruction at Tier0
    - getting ready for 2008 first physics run

Conclusions

- LHCb have set up an operational set of tools
  - DIRAC WMS using pilot agents paradigm (late binding, resource reservation)
  - DIRAC DMS (VO-boxes, transfer agents, automatic transfer DB)
  - Ganga: analysis on the Grid
- Many challenges are ahead of us all in 2007
  - **Stability and robustness** are the main issue, mainly SE
  - Many changes in 2007, careful deployment, but should be exposed to experiments ASAP
  - Expect the ramp up of resources at Tier1s as pledged
- LHCb will test the full chain during summer 2007
- LHCb will be ready for the pilot run in November 2007
- Looking forward to physics data in 2008