DIRAC Workload Management System

egee User Forum

Manchester, 10th May 2007
Outline

- Introduction to DIRAC
- Workload Management System
  - Interaction with EGEE Grid components
- Performance for LHCb workflows
- Plans for DIRAC3 and outlook
Introduction to LHCb

- LHCb is a forward single arm spectrometer
  - Designed to study CP violation in the b-quark sector in the LHC
- LHCb has a tiered computing model (6 Tier-1s + CERN)

LHCb is expected to generate 1 Petabyte of data per year
  - Must exploit distributed compute resources to cope with this
- No direct submission to Grid for LHCb jobs - instead submit via DIRAC
The LHCb Computing Model prescribes that all jobs run where the data are:
- Raw data shipped in real time to Tier-0
- All processing up to final histogram or Tuple is distributed to LHCb Tier-1 sites

For more information on LHCb data management see the following presentations in the Data Management track:
- A.C. Smith - DIRAC Data Management System
- M. Bargiotti - Data Management in LHCb
The DIRAC Workload & Data Management System (WMS) is made up of Central Services and Distributed Agents.

The main aims of DIRAC are:
- To integrate all of the heterogeneous compute resources available to LHCb
- Minimize the human intervention at sites
- Use wLCG services wherever possible

DIRAC realizes these goals via:
- Pilot Agent paradigm
- Overlay Network paradigm
DIRAC is a *PULL* scheduling system

- Agents first occupy a resource and then request jobs from a central task queue
- This ‘late binding’ allows execution environment to be checked in advance

- Pilot Agents are sent to the gLite Resource Broker as normal jobs
  - Facilitate *PULL* approach on *PUSH* system

- LCG jobs are Pilot jobs in the context of the DIRAC WMS
  - Actual workload management performed by DIRAC
LHCb resources comprise of individual PCs, site clusters and the Grid.
DIRAC Overlay Network Paradigm

- DIRAC Agents are deployed close to resources

Computing Resources

Site Clusters

Grid

PCs
DIRAC Overlay Network Paradigm

- Forms an overlay network of Agents masking the underlying diversity of the available compute resources

Computing Resources

- PCs
- Site Clusters
- Grid
- Agents
Services interact with Agents
Heterogeneous groupings of resources such as clusters / Grids become homogeneous via DIRAC.

DIRAC can therefore be viewed as a (very) large batch system:
- Accounting
- Priority Mechanism
- Fair share

The possibility to reuse standard batch system components also exists:
- e.g. schedulers
Overview of the DIRAC WMS
<table>
<thead>
<tr>
<th>Job ID</th>
<th>Status</th>
<th>Message</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Date/Time</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>4062</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 14:00:20</td>
<td>kritmert</td>
</tr>
<tr>
<td>4063</td>
<td>failed</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>Gauss_1</td>
<td>2006-01-11 14:00:22</td>
<td>jdluciano</td>
</tr>
<tr>
<td>4064</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 14:00:22</td>
<td>pkoffeb</td>
</tr>
<tr>
<td>4065</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 14:02:34</td>
<td>kritmert</td>
</tr>
<tr>
<td>4066</td>
<td>failed</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 14:54:03</td>
<td>lesanoff</td>
</tr>
<tr>
<td>4067</td>
<td>failed</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>Gauss_1</td>
<td>2006-01-11 14:57:13</td>
<td>sean</td>
</tr>
<tr>
<td>4068</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 14:54:12</td>
<td>powell</td>
</tr>
<tr>
<td>4069</td>
<td>failed</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>Gauss_1</td>
<td>2006-01-11 15:02:43</td>
<td>sean</td>
</tr>
<tr>
<td>4071</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:14:30</td>
<td>paterson</td>
</tr>
<tr>
<td>4072</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:08:37</td>
<td>paterson</td>
</tr>
<tr>
<td>4073</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:07:35</td>
<td>paterson</td>
</tr>
<tr>
<td>4074</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:07:58</td>
<td>paterson</td>
</tr>
<tr>
<td>4075</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:08:42</td>
<td>paterson</td>
</tr>
<tr>
<td>4076</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:02:19</td>
<td>pkoffeb</td>
</tr>
<tr>
<td>4077</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:07:30</td>
<td>mstiles</td>
</tr>
<tr>
<td>4078</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:18:38</td>
<td>soler</td>
</tr>
<tr>
<td>4079</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:14:59</td>
<td>soler</td>
</tr>
<tr>
<td>4080</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Pic.es</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:33:38</td>
<td>nnja</td>
</tr>
<tr>
<td>4081</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Pic.es</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:31:13</td>
<td>nnja</td>
</tr>
<tr>
<td>4082</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.Lyon.fr</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:12:28</td>
<td>nnja</td>
</tr>
<tr>
<td>4083</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:07:54</td>
<td>pkoffeb</td>
</tr>
<tr>
<td>4084</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:22:02</td>
<td>pkoffeb</td>
</tr>
<tr>
<td>4085</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:41:45</td>
<td>uedgede</td>
</tr>
<tr>
<td>4086</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:33:16</td>
<td>powell</td>
</tr>
<tr>
<td>4087</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 15:55:49</td>
<td>uedgede</td>
</tr>
<tr>
<td>4088</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 18:08:33</td>
<td>soler</td>
</tr>
<tr>
<td>4089</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CERN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 18:03:43</td>
<td>soler</td>
</tr>
<tr>
<td>4090</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.RAL.uk</td>
<td>DaVinci_1</td>
<td>2006-01-11 17:48:03</td>
<td>soler</td>
</tr>
<tr>
<td>4091</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.RAL.uk</td>
<td>DaVinci_1</td>
<td>2006-01-11 17:46:25</td>
<td>soler</td>
</tr>
<tr>
<td>4092</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.RAL.uk</td>
<td>DaVinci_1</td>
<td>2006-01-11 17:43:18</td>
<td>uedgede</td>
</tr>
<tr>
<td>4093</td>
<td>outpre</td>
<td>Job finished successfully</td>
<td>LCG.CFRN.ch</td>
<td>DaVinci_1</td>
<td>2006-01-11 17:31:36</td>
<td>paterson</td>
</tr>
</tbody>
</table>
Details of job:

name: hcb01.plt.es

Parameter Name:

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Job Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApplicationStatus</td>
<td>Job finished successfully</td>
</tr>
</tbody>
</table>

JDL:

```text
Requirements = | member("CMAP_Castor",other.Location)
Arguments = "jobDescription.xml"
JobName = "DaVinci_1"
parameters =
    |
    STEPS = "1";
    STEP_1_NAME = "0_D_1"
    |
    SoftwarePackages =
        |
        "DaVinci.v12r14"
    |
    JobType = "user"
Executable = "$LHCBDIR/DIRAC/scripts/jobexec"
CFTime = "500"
StdOutput = "std.out"
Owner = "paterson"
outputSandbox =
    |
    "pool.xml_catalog.xml",
    "std.out",
    "std.err",
    "DAVncples.hook",
    "DaVinciRun.csh"
```
LHCb computing activities on the Grid include:
- Monte Carlo (MC) Simulation (Tier-1 & Tier-2 sites)
- Reconstruction and Stripping (Tier-1 sites)
- User Distributed Analysis (mainly Tier-1 sites)

LHCb software is deployed to all Tier-1 sites via the SAM site testing mechanism
- Jobs can still install software if necessary
**LHCb Simulation Results**

- Typical MC Production job lasts 24hrs
- Recently achieved 10K concurrent production jobs
  - Throughput only limited by available capacity of LCG
- ~80 distinct sites accessed via Grid or DIRAC directly

**Sustained resource usage over extended periods of time**
- System is stable for simulation
Providing Efficient Access to Data

- Workload management was the first big challenge for DIRAC
  - The next major task is Data Management
    - DIRAC Stager illustrates this
- LHCb Reconstruction jobs require access to data stored on tape at the LHCb Tier-1 sites
  - DIRAC Stager remotely pre-stages data automatically
  - Jobs only allowed to run once all data is staged

- Currently no generic way to stage files across LHCb Tier-1s (dCache / Castor sites)
  - Will be possible in SRM v2…

Stuart K. Paterson
Providing Efficient Access to Data

- Workload management was the first big challenge for DIRAC
  - The next major task is Data Management
    - DIRAC Stager illustrates this
- LHCb Reconstruction jobs require access to data on tape storage at the LHCb Tier-1 sites
  - DIRAC Stager remotely pre-stages data automatically
  - Jobs only allowed to run once all data is staged

Currently no generic way to stage files across LHCb Tier-1s (dCache / Castor sites)
  - Will be possible in SRM v2…

Stuart K. Paterson
**LHCb Reconstruction Results**

- **Total Running Jobs:** 441
  - **DIRAC:** 0.23%
  - **LCG:** 99.77%

- **Last Week Running Jobs**

- **April 2007, reconstruction jobs successfully running at all LHCb Tier-1 sites**
  - CERN (Switzerland)
  - IN2P3 (France)
  - GridKa (Germany)
  - CNAF (Italy)
  - NIKHEF (Netherlands)
  - PIC (Spain)
  - RAL (U.K.)

- **Reconstruction effort is ongoing**
  - Current issues include: SRM instability; VOMS mappings; site downtime; tape failures…

*Stuart K. Paterson*  
*EGEE User Forum 2 - DIRAC WMS*
PULL scheduling and the Pilot Agent paradigm can allow multiple jobs to run in the same CPU slot.

LHCb distributed data analysis tasks currently run in this mode:
- Significant performance gains for short, high priority tasks
- Also reduces load on LCG WMS
  - Fewer Pilot Agents submitted

Requirements from Grid sites are simple:
- Can request long queues everywhere
Sample of ~4.5K jobs from January and February 2007

Jobs can either be submitted to DIRAC directly or through Ganga

DIRAC Performance ~3K Jobs (January-February 2007)

- Successful: 74.7%
- Input data not available: 16.1%
- Stalled: 9.2%
- Waiting: 2.0%
- Software installation failure: 1.3%
- Exception during execution: 1.1%
- Failed to upload output data: 0.5%
- Failed to resolve input data: 0.5%
- Failed to resolve output data: 3.8%

DIRAC masks any inefficiencies of the underlying Grid

- Users need not be concerned with where jobs run

Also have the possibility to submit to other Grids
Matcher time is the ‘overhead’ associated with the PULL paradigm

- Time taken to schedule job
- Biggest test for DIRAC is analysis jobs with varying requirements

Sample of ~3K real user analysis jobs ran in Jan/Feb 2007
Plans for DIRAC3

- Better support for high load production tasks
  - Such as bulk job submission and job progress monitoring
- Improved treatment of DAGs
  - Splitting is one case
- Interactive Monitoring pages
  - Authenticate via user certificate in browser
    - Automatically present information relevant to that user
- Optimized ‘Filling’ mode
  - Possibility to make our Pilot Agents ‘generic’ via glexec
    - Optimization of workloads at the VO-level would guarantee a minimal response time for high priority tasks
- Job Priority Mechanism
Performance results with DIRAC are encouraging

- DIRAC Stager has ensured high efficiency for LHCb Reconstruction jobs

The DIRAC optimized ‘Filling’ mode allows further performance gains

- Could ensure a minimal job start time for high priority tasks
- Also reduces load on LCG

DIRAC3 development is underway…

- System will become a generic Grid ‘toolkit’ with VO-specific plugins
Backup Slides
The DIRAC Optimized Filling mode is currently restricted to jobs with the same owner:
- Pilots run with the owner’s credentials

`glexec` is the solution to generic pilots:
- Can switch identity on the WN

Generic Pilots would further improve the performance:
- Optimization of workloads at the VO-level instead of the individual users

DIRAC Central Task Queue easily allows VO policies to be applied:
- Elegant solution to job prioritization problem