First Experience with LCG Operation and the future ...

CERN openlab
Board of Sponsors
3rd April 2009

Ian Bird
LCG Project Leader
The LHC Computing Challenge

- Signal/Noise: $10^{-9}$
- Data volume
  - High rate * large number of channels * 4 experiments
  - $15$ PetaBytes of new data each year
- Compute power
  - Event complexity * Nb. events * thousands users
  - $100$ k of (today's) fastest CPUs
  - $45$ PB of disk storage
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis everywhere
  - GRID technology
WLCG – what and why?

- A distributed computing infrastructure to provide the production and analysis environments for the LHC experiments
- Managed and operated by a worldwide collaboration between the experiments and the participating computer centres

- The resources are distributed – for funding and sociological reasons

- Our task is to make use of the resources available to us – no matter where they are located
  - We know it would be simpler to put all the resources in 1 or 2 large centres
  - This is not an option ... today

Ian.Bird@cern.ch
Tier 0 at CERN: Acquisition, First pass processing

Storage & Distribution

CERN Computer Centre

LHCb ~ 50 MB/sec

ATLAS ~ 320 MB/sec

ALICE ~ 100 MB/sec

CMS ~ 220 MB/sec

1.25 GB/sec (ions)
Tier 0 – Tier 1 – Tier 2

Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

Tier-1 (11 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (~130 centres):
- Simulation
- End-user analysis
CERN + Tier 1 accounting - 2008

CPU Time Delivered

Ratio of CPU : Wall_clock Times

Disk Storage Used

Tape Storage Used

Ian.Bird@cern.ch
Average data transfer volume

Expected burst level in the first year

Terabytes per day

Aggregate Data Movement From Mar 2005 To Mar 2008
VO-wise Data Transfer From All Sites To All Sites

May 6th 2008
LHCC referees: CMS - Computing
10M files Test @ ATLAS

Number of Files per Day

(From S. Campana)

M.C. Vetterli – LHCC review, CERN; Feb.’09 – #13
Number of jobs/month

Main outstanding issues related to service/site reliability

<table>
<thead>
<tr>
<th></th>
<th>Alice</th>
<th>ATLAS</th>
<th>CMS</th>
<th>LHCb</th>
<th>Total</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier-1s</td>
<td>6.24</td>
<td>32.03</td>
<td>30.73</td>
<td>2.50</td>
<td>71.50</td>
<td>34.3%</td>
</tr>
<tr>
<td>Tier-2s</td>
<td>9.61</td>
<td>52.23</td>
<td>55.04</td>
<td>20.14</td>
<td>137.02</td>
<td>65.7%</td>
</tr>
<tr>
<td>Total</td>
<td>15.85</td>
<td>84.26</td>
<td>85.77</td>
<td>22.64</td>
<td>208.52</td>
<td></td>
</tr>
</tbody>
</table>

From APEL accounting portal for Aug.'08 to Jan.'09; #s in MSI2k

Number of Users: Unique Users Number of unique users: 1490
The 3-D project is now finished – runs as production service

- In full production
  - Several GB/day user data can be sustained to all Tier 1s
- ~100 DB nodes at CERN and several 10’s of nodes at Tier 1 sites
  - Very large distributed database deployment
- Used for several applications
  - Experiment calibration data; replicating (central, read-only) file catalogues

ian.Bird@cern.ch
Improvement during CCRC and later is encouraging
- Tests do not show full picture – e.g. Hide experiment-specific issues,
- “OR” of service instances probably too simplistic

a) publish VO-specific tests regularly;
b) rethink algorithm for combining service instances
Improving Reliability

- Testing
- Task forces/challenges
- Monitoring
  - Appropriate
  - Followed up
challenges

- **Multi(many)-core:**
  - Better memory use efficiency
  - Co-scheduling many similar processes onto a single box
  - Parallelizing (multi-thread, MPI, ...) the applications

- **New technologies:**
  - Clouds + virtualisation
    - Use as “overflow” resources for peak periods – demonstrated
    - Running our facilities as “clouds” – use of virtualisation, management tools, etc.
    - Buying resources directly ... needs education of the funding agencies
    - Lessons: simplicity of interfaces, usage patterns,...
    - Virtualisation – helps improve service reliability; simplifies facility management (tbc) and leaves apps to deal with dependencies
    - Grid → grid of “cloud-like” objects

- **Filesystems**
  - Lustre, Hadoop, NFS4.1, etc
  - Can we use these to improve our service reliability? Usability?

- **Messaging systems**
  - Use for integrating systems – Web services across languages etc did not deliver
What are the limitations & possible solutions?
Messing Software Ecosystem Examples

- MRG Grid provides low latency scheduling via messaging
  - Useful pattern for other systems
- MRG/Qpid provides features people often build on top of messaging
  - XML Exchange, LVQ, Ring Queue, TTI, Federation, Management, etc.
- Open Source projects are building on AMQP Messaging
  - OpenIPA project is using AMQP Messaging for management and monitoring of Identity, Policy, Audit systems
  - LibVirt project is using AMQP messaging for management and monitoring
  - Wireshark supports AMQP
Challenges cont...

- Simplification of data management
  - Clouds don’t help much here
  - Abstraction – SRM – has added complexity
  - How much is required? How can we simplify?
  - What are the lessons to learn?
  - Database access – grid authn/authz would help – ...

- New Tier 0 centre
  - We will run out of power, new centre planned, will it be ready when we need it???

- Moving from EGEE to a European sustainable grid infrastructure
  - Whilst maintaining a solid service
EGEE \(\Rightarrow\) EGI+NGIs

- EGI blueprint published in December; endorsed in January by 20 NGIs
- March – policy board has selected Amsterdam as the location of EGI.org (body resp. for managing EGI)
- Initiation of transition process to create an EGI council
  - MoU to be prepared as an interim measure to identify NGIs prepared to commit as described in the blueprint (start with Letters of Intent)
  - Anticipate 1st council meeting in May
- Task force to be established for preparation of EGI proposals, for EC calls anticipated to close in November
- EGEE has outlined a fairly detailed transition plan for the final year of the project
  - But can only go so far
EGI and WLCG

- WLCG cannot take the risk of assuming EGI will be in place at the end of EGEE-III
- We plan to ensure that services provided to use to day by EGEE are assured by our Tier 1 sites
  - Support the formation of a gLite consortium to support the middleware
- In parallel we work with EGI_DS and EC to try and ensure that EGI and the NGIs will deliver what we need
Conclusions

- We have built a working system that will be used for first data taking
  - But it has taken a lot longer than anticipated ... and was a lot harder ... and the reality of grids does not quite match the hype ...

- We now have an opportunity to rethink how we want this to develop in the future
  - Clearer ideas of what is needed
  - And must consider the risks, maintainability, reliability, and complexity

- Change of funding model and new technologies provide opportunities
- Challenges: data management and reliability, reliability, ...

- Should remember ... Our goal is to enable the experiments’ computing, not necessarily to develop computer science (unless we have to ...)
WLCG timeline 2009-2010

2009
- CCRC’09
  Tests to be scheduled
- Switch to SL5/64bit completed?

2009 Capacity commissioned

2010
- 2010 Capacity commissioned

2010 Capacity commissioned

Deployment of glexec/SCAS; CREAM; SRM upgrades; SL5 WN