The LHC Computing Challenge: Preparation, Reality and Future Outlook

Tony Cass
Leader, Database Services Group
Information Technology Department

10th November 2010
Outline

• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
• Summary/Conclusion
The fastest racetrack on the planet…

Trillions of protons will race around the 27km ring in opposite directions over 11,000 times a second, travelling at 99.999999991 per cent the speed of light.
The emptiest space in the solar system...

To accelerate protons to almost the speed of light requires a vacuum as empty as interplanetary space. There is 10 times more atmosphere on the moon than there will be in the LHC.
Methodology

One of the coldest places in the universe…

With an operating temperature of about -271 degrees Celsius, just 1.9 degrees above absolute zero, the LHC is colder than outer space.
The hottest spots in the galaxy...

When two beams of protons collide, they will generate temperatures 1000 million times hotter than the heart of the sun, but in a minuscule space.
The biggest most sophisticated detectors ever built…

To sample and record the debris from up to 600 million proton collisions per second, scientists are building gargantuan devices that measure particles with micron precision.
The most extensive computer system in the world…

To analyse the data, tens of thousands of computers around the world are being harnessed in the Grid. The laboratory that gave the world the web, is now taking distributed computing a big step further.
Why?
To push back the frontiers of knowledge…

Newton’s unfinished business… what is mass?
Science’s little embarrassment… what is 96% of the Universe made of?
Nature’s favouritism… why is there no more antimatter?
The secrets of the Big Bang… what was matter like within the first second of the Universe’s life?
**Matter particles**
- **Electron**
  - Responsible for electricity and chemical reactions;
  - it has a charge of -1
- **Electron neutrino**
  - Particle with no electric charge, and possibly no mass;
  - billions fly through your body every second
- **Muon**
  - A heavier relative of the electron; it lives for two-millionths of a second
- **Muon neutrino**
  - Created along with muons when some particles decay
- **Tau**
  - Heavier still; it is extremely unstable. It was discovered in 1975
- **Tau neutrino**
  - Not yet discovered but believed to exist

**Force particles**
- **Gluons**
  - Carriers of the strong force between quarks
  - Felt by: quarks
- **Photons**
  - Particles that make up light; they carry the electromagnetic force
  - Felt by: quarks and charged leptons
- **Intermediate vector bosons**
  - Carriers of the weak force
  - Felt by: quarks and leptons
- **Gavitons**
  - Carriers of gravity
  - Felt by: all particles with mass

**LEPTONS**
- **First Family**
  - Up
  - Has an electric charge of plus two-thirds; protons contain two, neutrons contain one
  - Found in 1974
- **Second Family**
  - Charm
  - A heavier relative of the up; found in 1974
- **Third Family**
  - Top
  - Heavier still
  - Bottom
  - Heavier still; measuring bottom quarks is an important test of electroweak theory

**QUARKS**
- **First Family**
  - Down
  - Has an electric charge of minus one-third; protons contain one, neutrons contain two
  - Found in 1964
- **Second Family**
  - Strange
  - A heavier relative of the down; found in 1964
- **Third Family**
  - Strange
  - A heavier relative of the down; found in 1964

**The explosive release of nuclear energy is the result of the strong force.**

**Electricity, magnetism and chemistry are all the results of electromagnetic force.**

**Some forms of radioactivity are the result of the weak force.**

**All the weight we experience is the result of the gravitational force.**
To push back the frontiers of knowledge…

Newton’s unfinished business… what is mass?
Science’s little embarrassment… what is 96% of the Universe made of?
Nature’s favouritism… why is there no more antimatter?
The secrets of the Big Bang… what was matter like within the first second of the Universe’s life?
To develop new technologies…

Information technology - the Web and the Grid
Medicine - diagnosis and therapy
Security - scanning technologies for harbours and airports
Vacuum - new techniques for flat screen displays or solar energy devices
To unite people from different countries and cultures...

20 Member states
38 Countries with cooperation agreements
111 Nationalities
10000 People
To train the scientists and engineers of tomorrow…

From mini-Einstein workshops for five to sixes, through to professional schools in physics, accelerator science and IT, CERN plays a valuable role in building enthusiasm for science and providing formal training..
“Compact” Detectors!
Large Hadron Collider

ATLAS Detector
The Four LHC Experiments...

**ATLAS**
- General purpose
- Origin of mass
- Supersymmetry
- 2,000 scientists from 34 countries

**CMS**
- General purpose
- Origin of mass
- Supersymmetry
- 1,800 scientists from over 150 institutes

**ALICE**
- Heavy ion collisions, to create quark-gluon plasma
- 50,000 particles in each collision

**LHCb**
- To study the differences between matter and antimatter
- Will detect over 100 million b and b-bar mesons each year
The accelerator generates 40 million particle collisions (events) every second at the centre of each of the four experiments’ detectors.
... generate lots of data ...

Reduced by online computers to a few hundred “good” events per second. Which are recorded on disk and magnetic tape at 100-1,000 MegaBytes/sec → ~15 PetaBytes per year for all four experiments.

- Current forecast ~ 23-25 PB / year, 100-120M files / year
  - ~ 20-25K 1 TB tapes / year
- Archive will need to store 0.1 EB in 2014, ~1Billion files in 2015
which is distributed worldwide

**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
See


For the Google Earth monitoring display
What were the challenges in 2007?
• Introduction to CERN and Experiments
• LHC Computing
• Challenges
  - Capacity Provision
  - Box Management
  - Data Management and Distribution
  - What’s Going On?
• Summary/Conclusion
Outline

• Introduction to CERN and Experiments
• LHC Computing
• Challenges
  - Capacity Provision
  - Box Management
  - Data Management and Distribution
  - What’s Going On?
• Summary/Conclusion
The Grid

- Timely Technology!
- **Deploy** to meet LHC computing needs.
- Challenges for the **Worldwide** LHC Computing Grid Project due to
  - worldwide nature
    - competing middleware...
  - newness of technology
    - competing middleware...
  - scale
  - ...

The GRID
Blueprint for a New Computing Infrastructure
Edited by Ian Foster and Carl Kesselman

2007
Seven Wonders of the IT World

The fastest supercomputer. The most intriguing data center. The constantly changing core at the heart of Linux. Take a tour of the most impressive and most unusual marvels of the IT world.

Leave a comment (65)
By C.G. Lynch

PAGE 4

World's largest scientific grid computing project: The E-science II (EGEE-II) project
Launched: September 2006, for use by scientists around the world.

Helps power: Large-scale scientific research projects in fields from geology to chemistry—for example, will analyze data from CERN's Large Hadron Collider, a particle accelerator being built to help investigate details around the Big Bang and related physics questions.

Amount of work it does: 93,000 jobs a day, more than 1 million per month.

Juggling ability: Runs about 30,000 jobs concurrently, on average.
Remaining Challenges

• Creating a working Grid service across multiple infrastructure is clearly a success, but challenges remain
  - Reliability
  - Ramp-up
  - Collaboration
    • From computer centre empires to a federation
    • consensus rather than control

  ...
Outline

- Introduction to CERN and Experiments
- LHC Computing
- Challenges
  - Capacity Provision
  - Box Management
    - Installation & Configuration
    - Monitoring
    - Workflow
  - Data Management and Distribution
  - What’s Going On?
- Summary/Conclusion
Toolkit developed by CERN in collaboration with many HEP sites and as part of the European DataGrid Project. See http://cern.ch/ELFms
Outline

• Introduction to CERN and Experiments
• LHC Computing
• Challenges
  - Capacity Provision
  - Box Management
  - Data Management and Distribution
  - What’s Going On?
• Summary/Conclusion
Dataflows and rates

Scheduled work only!

700MB/s

420MB/s

700MB/s (1600MB/s)

1120MB/s (2000MB/s)

Averages! Need to be able to support 2x for recovery!

1430MB/s

2007
• 15PB/year. Peak rate to tape >2GB/s
  - 3 full SL8500 robots/year
• Requirement in first 5 years to reread all past data between runs
  - 60PB in 4 months: 6GB/s
• Can run drives at sustained 80MB/s
  - 75 drives flat out merely for controlled access
• Data Volume has interesting impact on choice of technology
  - Media use is advantageous: high-end technology (3592, T10K) favoured over LTO.
Outline

• Introduction to CERN and Experiments
• LHC Computing
• Challenges
  - Capacity Provision
  - Box Management
  - Data Management and Distribution
  - What’s Going On?
• Summary/Conclusion
Site managers understand systems (we hope!).
But do they understand the service?
- and do the users?
- and what about cross site issues?
  • Are things working?
  • If not, just where is the problem?
    - how many different software components, systems and network service providers are involved in a data transfer site X to site Y?
And here’s a couple more...
Energy of a 1TeV Proton
Two nominal beams together can melt ~1,000kg of copper. Current beams: ~100kg of copper.
During high intensity SPS extraction, TT40 damage was observed approximately 110 cm from the entrance.

The beam was a 450 GeV full LHC injection bunch of $3.4 \times 10^{12} p^+$ in 288 bunches, and was extracted from SPS LSS4 with the wrong trajectory, resulting in a yield of $4.4 \times 10^{12}$ particles at 3.5 TeV.
Three accelerator database applications:
- Short term settings and control configuration
  - Considered as “any other active component necessary for beam operation”.
    - No database: no beam
    - Lose database: lose beam (controlled!)
- Short term (7-day) real-time measurement log
- Long term (20 yr+) archive of log subset
This is the **ONLY** element in the LHC that can withstand the impact of the full 7 TeV beam! Nevertheless, the dumped beam must be painted to keep the peak energy densities at a tolerable level!
Accelerator “fly by Oracle”

- Three accelerator database applications:
  - Short term settings and control configuration
  - Considered as “any other active component necessary for beam operation”.
  - No database: no beam
  - Lose database: lose beam (hopefully controlled…)

- Short term (7-day) real-time measurement log
- Long term (20 yr+) archive of log subset
  - ~2,000,000,000,000 rows; ~4,000,000,000/day
Responsibilities & Requirements

A non sleeping 24hr/day 365d/year running system

- **Ensure safe detector operation**
  - anticipating the Detector Safety System (DSS) actions, triggering protection mechanisms on adverse conditions (high temperatures, high humidity, overcurrents, water leaks, electrical trips…)
  - preventing potentially dangerous actions
  - issuing alert notifications (alert screen, SMS, control room voice alerts)

- **Provide efficient detector operation**
  - making sure that voltages are present whenever the accelerator conditions allow for physics data taking
  - guaranteeing that the controlled parameters are stable within their calibrated operating ranges
Control system size

~ $10^6$ control system parameters

<table>
<thead>
<tr>
<th>System Name</th>
<th>Number of PCs</th>
<th>Monitored Parameters</th>
<th>Controlled Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracker</td>
<td>14</td>
<td>350k</td>
<td>20k</td>
</tr>
<tr>
<td>Calorimeter</td>
<td>14</td>
<td>115k</td>
<td>2k</td>
</tr>
<tr>
<td>Muon</td>
<td>30</td>
<td>435k</td>
<td>30k</td>
</tr>
<tr>
<td>Trigger DCS</td>
<td>2</td>
<td>1k</td>
<td>0.5k</td>
</tr>
<tr>
<td>Alignment</td>
<td>3</td>
<td>3k</td>
<td>0.5k</td>
</tr>
<tr>
<td>Services</td>
<td>35</td>
<td>20k</td>
<td>1k</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>98</strong></td>
<td><strong>934k</strong></td>
<td><strong>34k</strong></td>
</tr>
</tbody>
</table>

*PVSS by ETM (now owned by Siemens)*
Main supervisor panel
Main supervisor panel
Main supervisor panel
Main supervisor panel
Oracle RAC (cluster) application scalability

Eric Grancher and Anton Topurov
CERN IT department / openlab

Average Streams Throughput

- **Oracle 10g**
  - Row size = 100B: 40000
  - Row size = 500B: 37000
  - Row size = 1000B: 34000

- **Oracle 11gR2**
  - Row size = 100B: 30000
  - Row size = 500B: 25000

- **Oracle 11g R2 (optimized)**
  - Row size = 100B: 4600
  - Row size = 500B: 2800
  - Row size = 1000B: 1700

PVSS logging to Oracle & Streams Export
Outline

• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
• Summary/Conclusion
Praise for a Sys Admin?

[Silence]
An impressive start

- LHC commissioning process on unprecedented pace
- Experiments showed their readiness for exploitation of the 7 TeV data…
- …ready to follow with more complex triggers due to the increase of luminosity.
- Analyses proceeding very rapidly in all experiments and results being submitted for publication within days

- **Brilliant performances of the WLCG a key factor in the spectacular startup.**
More striking still is the speed with which the raw data are being processed. The freshest batch emerged from the LHC on July 18th and were moulded into meaningful results by July 21st, in time for the Paris conference. Not long ago this process would have taken weeks, says Fabiola Gianotti, the spokeswoman for ATLAS, one of the four main LHC experiments. One reason is the development of the Grid, a computing network CERN hopes will prove a worthy successor to its previous invention, the World Wide Web. The Grid lets centres around the world crunch the numbers as soon as they come out of the machine.
6 months of LHC data

Stored ~ 5 PB this year

Disk Servers (Gbytes/s)

- eth0 in: aver: 2.6G, max: 7.2G, min: 664.6M, curr: 2.1G
- eth0 out: aver: 7.5G, max: 18.3G, min: 1.6G, curr: 7.9G

Tier 0 storage:
- Accepts data at average of 2.6 GB/s; peaks > 7 GB/s
- Serves data at average of 7 GB/s; peaks > 18 GB/s
- CERN Tier 0 moves ~ 1 PB data per day
Large numbers of analysis users

- CMS ~800,
- ATLAS ~1000,
- LHCb/ALICE ~200

WLCG Usage

- Use remains consistently high
  - 1 M jobs/day; >>100k CPU-days/day
  - Actually much more inside pilot jobs

As well as LHC data, large simulation productions ongoing
The 2007 challenges

• Ramp-up

Need foreseen @ TDR for T0+1 CPU and Disk for 1st nominal year

Expected needs in 2011 & 2012
## Hardware ramp-up

### Summary of Computing Resource Requirements

**All experiments - 2008**

*From LCG TDR - June 2005*

<table>
<thead>
<tr>
<th></th>
<th>CERN</th>
<th>All Tier-1s</th>
<th>All Tier-2s</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU (MSPECint2000s)</td>
<td>25</td>
<td>56</td>
<td>61</td>
<td>142</td>
</tr>
<tr>
<td>Disk (PetaBytes)</td>
<td>7</td>
<td>31</td>
<td>19</td>
<td>57</td>
</tr>
<tr>
<td>Tape (PetaBytes)</td>
<td>18</td>
<td>35</td>
<td></td>
<td>53</td>
</tr>
</tbody>
</table>

### Table of Computed Resources

<table>
<thead>
<tr>
<th>Country</th>
<th>Federation</th>
<th>Physical CPU</th>
<th>Logical CPU</th>
<th>HEPSPEC06</th>
<th>CPU Pledge</th>
<th>Total Online Storage (GB)</th>
<th>Disk Pledge</th>
<th>Total Nearline Storage (GB)</th>
<th>Tape Pledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switzerland</td>
<td>CH-CERN</td>
<td>4,496</td>
<td>17,644</td>
<td>197,308</td>
<td>233,400</td>
<td>18,181,259</td>
<td>14,790,000</td>
<td>30,957,137</td>
<td>31,600,000</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>4,496</td>
<td>17,644</td>
<td>197,308</td>
<td>233,400</td>
<td>18,181,259</td>
<td>14,790,000</td>
<td>30,957,137</td>
<td>31,600,000</td>
</tr>
</tbody>
</table>

### Additional Notes

- ~2,500 PCs
- Another ~1
- 500 boxes

4,000HS06 = 1MSPECint2000
The 2007 challenges

- Reliability

I wouldn’t fly on a plane that was 98% reliable!!!! But you probably fly an airline that is...

Punctuality details from flightstats.com
Operational Issues - I

Storage is complicated

- Q3 2010
- Q4 2009
- Q3 2009
- Q2 2009
- Q1 2009
Operational Issues - I

Storage is complicated

Hardware failures are frequent – and cause problems for storage and database systems
Operational Issues - I

- Storage is complicated
- Hardware failures are frequent – and cause problems for storage and database systems
- Infrastructure failures (loss of power or cooling) are a fact of life
Operational Issues - I

Storage is complicated

Hardware failures are frequent – and cause problems for storage and database systems

Infrastructure failures (loss of power or cooling) are a fact of life

Software and Networks seem reliable, surprisingly!
Experiments need to distribute software to sites.

Problems:
- Correct execution of installation task
- Ensuring the software is available on all nodes
- Shared filesystem bottleneck

Solution
- CernVM-FS: Virtual software installation with an HTTP filesystem based on GROW-FS:
  - Shared filesystem bottlenecks

Shared filesystem bottlenecks

CernVM 2 / Standard SL5 Worker Node

SL5 Kernel

CernVM-FS

Fuse

Hierarchy of HTTP Caches

Linux File System Buffers

1 GB CernVM-FS Cache LRU managed

10 GB Single Release (all releases available)
Repositories at CernVM:
ATLAS, CMS, LHCb, ALICE, LCD, NA61, H1, BOSS
HEPSOFT, Grid UI, LCG Externals
Ongoing: ATLAS Nightlies, ATLAS Conditions Database

Shared filesystem bottlenecks

Overall: 600 GB, 18.5 Mio. File System Objects
Repository Core: 97 GB (16%), 3.3 Mio. File System Objects (18%)
(+ 40 GB Archive Data)
• Collaboration
  - From computer centre empires to a federation
  - Consensus rather than control

This remains a challenge in 2010!

We reach consensus on most issues, but
  • Communication is a headache with so many sites and a changing population
  • Site policies can be problematic in certain cases (e.g. installation of setuid software) especially for sites that are not 100% HEP.
  • We reinvent the wheel: quattor & Lemon not widely adopted by Tier1s and Tier2s
    • although adopted after evaluation of various systems by a major financial institution with 10s of thousands of boxes.
Grid sites generally want to maintain a high average CPU utilisation. Easiest to do this if there is a local queue of work to select from when another job ends.

Users are generally interested in turnaround times as well as job throughput. Turnaround is reduced if jobs are held centrally until a processing slot is known to be free at a target site.
Pilot Jobs

Pilot job systems ensure “joblets” are sent to a host that will provide immediate execution.

Pilot job will check for correct s/w environment before loading “joblets”.

They also guarantee experiment control over job execution order. Low priority work can (will...) be pre-empted!

More of the “grid intelligence” is in per-VO software than was imagined at the start of the Grid adventure.
Data Issues

Scheduled work only!

Averages! Need to be able to support 2x for recovery!
• Mass Storage systems have worked well for recording, export and retrieval of “production” data.

• But some features of the CASTOR system developed at CERN are unused or ill-adapted
  - experiments want to manage data availability
  - file sizes, file-placement policies and access patterns interact badly
    • alleviated by experiment management of data transfer between tape and disk...
  - analysis use favours low latency over guaranteed data rates
    • aggravated by experiment management of data; automated replication of busy datasets is disabled.
• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
  - Data Access
  - Virtualisation
• Summary/Conclusion
Outline

• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
  - Data Access
  - Virtualisation
• Summary/Conclusion
LHC Experiments

ANALYSIS

XROOT or mountable file system

Disk Fools

Scalable, secure, accountable, globally accessible, manageable

Allow to choose service level for availability, reliability, performance

decoupled from HW

ASGC
BNL
FNAL
FZK
IN2P3
CNAF
NDGF
NIKHEF
PIC
RAL
TRIUMF

Tier-1s data replication

Castor

Managed repli
tape servers
Data Futures — II

• Address hardware reliability issues in software
  - ... as is done elsewhere...

• Bring back model where storage system maintains multiple replicas of files, but drop disk mirroring
  - CERN switched from parity RAID a few years ago for I/O performance reasons.

• Growing interest in HADOOP at Tier2s.
• Only a small subset of data distributed is actually used

• Experiments don’t know a priori which dataset will be popular
  - CMS has 8 orders magnitude in access between most and least popular

Dynamic data replication: create copies of popular datasets at multiple sites.
Data Futures – IV

- Network capacity is readily available...

MONARC 2000
• Network capacity is readily available...

• ... and it is reliable:

Fibre cut during tests in 2009
Capacity reduced, but alternative links took over
Data Futures — IV

• Network capacity is readily available...
• ... and it is reliable.
• So why not simply copy data from another site
  - rather than recalling from tape?
  - if it not available locally?
Outline

• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
  - Data Access
  - Virtualisation
• Summary/Conclusion
Batch Virtualisation

Integration of virtual machines in the batch system at CERN
S. Goasguen, M. Gujarro, B. Moreira, E. Roche, U. Schwickerath, Ricardo Silva, R. Wartel

CHEP 2010
October 19, 2010

WNoDeS, a tool for integrated Grid/Cloud and computing farm virtualization
Alessandro Italiano, Davide Sandri
INFN-CNAF
CHEP 2010, Taipei
18 - 22 October, 2010

Virtual Batch Worker Nodes

Virtualization of PBS Jobs
Presenter: Pau Tallada Crespi
Authors: Marc Rodriguez, Pau Tallada, Christian Neissner, Manuel Defino

Working schema
Batch Virtualisation

- Virtualisation has a cost for users...

---

Reconstruction

- Job uses large memory (~2GB, with 250MB shared)
- Native is ~9% better than VM.
  - Worse than Simulation, because Reconstruction jobs have more I/O activity and the memory footprint is much larger.
- VM has memory overhead, so 1VM case is swapping with high number of jobs.
- Pages can’t be shared across VMs and 2VMs has more memory overhead. So 2VM case is swapping heavier.
Batch Virtualisation

• Virtualisation has a cost for users...
• ...but efficiency advantages for sites.
Virtualisation has a cost for users...
... but efficiency advantages for sites.
Although multiplication of entities is never a good thing...

Batch Virtualisation

Sharing VM images between sites?
Automatic security updates for small sites?
But trust needed to make remote images acceptable.

– A step to cloud computing?

Can we cut out local workload management systems and dynamically instantiate VM images that connect directly to pilot job frameworks?
Batch Virtualisation

• Virtualisation has a cost for users...
• ... but efficiency advantages for sites.
• Although multiplication of entities is never a good thing...
• ... but maybe users will switch to requesting whole machines, not single processors.
Batch Virtualisation

• Virtualisation has a cost for users...
• ... but efficiency advantages for sites.
• Although multiplication of entities is never a good thing...
• ... but maybe users will switch to requesting whole machines, not single processors.
• **Can we cut out local workload management systems and dynamically instantiate VM images that connect directly to pilot job frameworks?**
  - A step to cloud computing?
• Virtualisation has a cost for users...
• ... but efficiency advantages for sites.
• Although multiplication of entities is never a good thing...
• ... but maybe users will switch to requesting whole machines, not single processors.
• Can we cut out local workload management systems and dynamically instantiate VM images that connect directly to pilot job frameworks?  
  - A step to cloud computing?
• Sharing VM images between sites?  
  - Automatic security updates for small sites?  
  - Trust needed to make remote images acceptable!
• Introduction to CERN, LHC and Experiments
• The LHC Computing Challenge
• Preparation
• Reality
• Future Outlook
• Summary/Conclusion
Conclusions

• Preparation for LHC Computing has been
  - Long
  - Technically challenging
  - Sociologically challenging

• but
  - Successful,
  - Capable of improvements based on experience with real data

• and also
  - An exciting adventure
  - With much more detail than I have been able to give here...