PetaByte-scale data management at CERN

Università degli Studi di Trieste
Corso di Laurea in Fisica

4th of May 2012
The LHC Computing Grid - February 2012

Service information

- full name: Computer Centre By Numbers
- short name: CCBYNUM
- group: IT-CF-FPP
- site: CERN
- email: imre.szebenyi@cern.ch
- manager: Imre Szebenyi

Service availability (more)

- availability: 100%
- status: available
- last update: 11:01:02, 19 Feb 2012 (6 minutes ago)
- expires after: 1440 minutes

Additional service information (more)

- Number of processors: 14,972
- Number of cores: 64,623
- Memory capacity (TiB): 165
- Memory modules: 55,729
- Raw HDD capacity (TiB): 62,660
- Number of HDD's: 62,023
- Number of systems: 7,975
- Number of RAID controllers: 3,607
- Number of enclosures: 1,554
- SPEC CPU2006: 503,637
- Number of racks: 1,070
- Number of virtual machines: 3,908
- Number of Fibre channel ports: 742
- Number of 1G ports: 16,773
- Number of 10G ports: 622
- Current power consumption (kW): 2,186
- Current power consumption (kVA): 2,305

- 24x7 operator and system admin support
- Management and Automation framework for large scale Linux clusters
- Hardware installation & retirement
- ~7,000 hardware movements/year; ~1000 disk failures/year
What are all these computers/disks/networks... for?

(Detector) simulation

Data acquisition and processing

Data distribution

Data analysis
From Physics to Raw Data

Basic physics

Fragmentation, Decay

Interaction with detector material
Multiple scattering, interactions

Detector response
Noise, pile-up, cross-talk, inefficiency, ambiguity, resolution, response function, alignment, temperature

Raw data (Bytes)
Read-out addresses, ADC, TDC values, Bit patterns
From Raw Data to Physics

Raw data
Convert to physics quantities

Detector response
apply calibration, alignment,

Interaction with detector material
Pattern, recognition, Particle identification

Fragmentation, Basic physics
Decay
Physics analysis
Results

Reconstruction

Simulation (Monte-Carlo)

Analysis
Analysis flow (user view)

Data selection (quality, type, config...)

Raw data (Exp. Data)

Simulation

Production System

RECONSTRUCTION

ANALYSIS

But how this is done in practice? Of course we need CPUs, disks, networks etc.. The problem is orchestrating hardware resources, software and humans :)

“All” data are stored in files (aggregated as “datasets” = collections of files). Only a small fraction of data in real DBs (e.g calibrations). This is one characteristics of HEP computing.
The role of the CERN Computer Centre

The LHC Computing Grid - April 2011
ATLAS: 7,000 tons, 150 million sensors generating data 40 millions times per second i.e. a petabyte/s (1 million GB/s)

ATLAS is around than 3,000 collaborators
From 169 universities from 37 countries
~1000 students!!!
One of CERN main responsibilities
(LHC scientific programme – LHC data management)

Tier 0 function

- Experimental data to tape (RAW)
- Data distribution (Tier0 → Tier1s)
- Data reconstruction and redistribution
  - RAW → ESD/DST,AOD.. 
- Archival data (e.g. simulation data generated in other centres)

Analysis centre

- Each LHC physicist has access to CERN!
- CERN-based analysis groups
CERN Computer Centre: Storage, Distribution and Processing (Reconstruction and Analysis)

Géant: the pan-european Research and Education Network

LHCOPN: dedicated links with major computer centres worldwide
The HEP Data Challenge

- LHC will run for 20 years
- Experiments are producing about **15 Million Gigabytes** of data each year (about 20 million CDs!)
- LHC data analysis requires a computing power equivalent to ~100,000 of today's fastest PC processors
- Requires many cooperating computer centres, as CERN can only provide ~20% of the capacity

A challenge for physics...

... and a challenge for technology research and industry as well
Big data non HEP specific (… any longer)
Moore's law

Technology helps you... IF you have good ideas AND you can master the global complexity.
**CERN celebrates 20 years of a free, open web**

30 Apr 2013

Geneva, 30 April 2013 - Twenty years ago CERN¹ published a statement that made the World Wide Web ("W3", or simply "the web") technology available on a royalty-free basis. By making the software required to run a web server freely available, along with a basic browser and a library of code, the web was allowed to flourish.

The technology, invented in 1989 at CERN by Tim Berners-Lee, was originally conceived and developed to meet the demand for information sharing between physicists in universities and institutes around the world.

Other information retrieval systems that used the Internet - such as WAIS and Gopher - were available at the time, but the web’s simplicity along with the fact that the tech adoption and development.

“There is no sector of society that has not been transformed by the web”, says Rolf Hanoj, CERN Director-General. “From how we communicate, work, innovate and everything in between - it has benefited humankind.”

The first website at CERN - and in the world – was dedicated to the inference engine. This was set up on Berners-Lee’s NeXT computer. The website did not store any documents, but rather provided a way to set up your own: the original web server - was still at CERN, sadly the world’s first web address.

To mark the anniversary of the publication of the document that to use, CERN is starting a project to restore the first website and associated with the birth of the web. To learn more about the project: http://info.cern.ch

---

**1990s:**

The web was invented at CERN! The machine used by Tim Berners-Lee in 1990 to develop and run the first WWW server, multi-media browser and web editor.
(Data) challenge

Semi inclusive → reconstruct “the rest” of the deep inelastic event

- 35 MB/s
- 300 TB/year

Use a Linux PC farm

- Instead of the “usual super computers”

C++

- Instead of good ol' Fortran IV

All data in a data base (Objectivity/DB)

- Instead of “plain” files
- Object store (noSQL)

First experiment using CASTOR

- Instead of writing tapes “by ourself”

More than a prototype!!!

- 3 times less data 14 years before LHC (LHC experiment) (7 Moore’s cycles hence ~2^7)
- 40x factor *against* us (2^{7/3})
Close collaboration with CERN IT

COMPASS was the first/among the first to:

- Migration to Linux PC farms (conclusions of CHEP2000)
  - CCF at CERN
  - ACID in Trieste
- Use of Objectivity/DB at the $10^2$ TB scale
- First production user of CASTOR
- Deliver and operate CORAL (C++ framework) for data processing
  - And the algorithmic parts, notably the RICH reconstruction
Run 2010 - Collected data to CASTOR

Design Value (35 MB/s)
Data management today

- CASTOR for Tier0
  - “RAW” data, tape access
- EOS for analysis
  - Analysis data, disk only
- Optimised for different use cases (T0 and analysis)

Total installed disk capacity:
13.0 PB + 22.4PB

Over 100 PB in the CERN store (disk+tape)

NB: 1 week of LHC (2012) = 1 PB on tape!
## DSS/FDO weekly report (03 May 2013)

### ACRON service

<table>
<thead>
<tr>
<th>Instance</th>
<th>Jobs</th>
<th>Efficiency (%)</th>
<th>Unique user/host</th>
</tr>
</thead>
<tbody>
<tr>
<td>@cern</td>
<td>859.6 K</td>
<td>94.8</td>
<td>659.0</td>
</tr>
</tbody>
</table>

### AFS service

<table>
<thead>
<tr>
<th>Instance</th>
<th>Capacity</th>
<th>Files</th>
<th>Δ</th>
<th>Size</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>@cern</td>
<td>501.7 TB</td>
<td>1.7 G</td>
<td>20.3 M</td>
<td>198.8 TB</td>
<td>3.2 TB</td>
</tr>
</tbody>
</table>

### FILER service

<table>
<thead>
<tr>
<th>Instance</th>
<th>Capacity</th>
<th>Files</th>
<th>Δ</th>
<th>Size</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>@cern</td>
<td>192.0 TB</td>
<td>110.4 M</td>
<td>0.7 M</td>
<td>23.7 TB</td>
<td>0.1 TB</td>
</tr>
</tbody>
</table>

### CASTOR service

<table>
<thead>
<tr>
<th>Instance</th>
<th>Files</th>
<th>Δ</th>
<th>Size</th>
<th>Δ</th>
<th>OnTape</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>@cern</td>
<td>311.2 M</td>
<td>-0.2 M</td>
<td>86.3 PB</td>
<td>132.0 TB</td>
<td>74.6 PB</td>
<td>121.0 TB</td>
</tr>
</tbody>
</table>

### EOS service

<table>
<thead>
<tr>
<th>Instance</th>
<th>Files</th>
<th>Δ</th>
<th>Size</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>@cern</td>
<td>134.5 M</td>
<td>-1.7 M</td>
<td>15.7 PB</td>
<td>462.1 TB</td>
</tr>
<tr>
<td>alice</td>
<td>96.8 M</td>
<td>-2.4 M</td>
<td>4.1 PB</td>
<td>-86.4 TB</td>
</tr>
<tr>
<td>atlas</td>
<td>29.9 M</td>
<td>0.5 M</td>
<td>4.2 PB</td>
<td>132.0 TB</td>
</tr>
<tr>
<td>cms</td>
<td>6.0 M</td>
<td>0.2 M</td>
<td>4.0 PB</td>
<td>400.0 TB</td>
</tr>
<tr>
<td>lhcb</td>
<td>1.0 M</td>
<td>0.1 M</td>
<td>2.5 PB</td>
<td>15.6 TB</td>
</tr>
</tbody>
</table>
## Data durability on a disk farm

### Command and Output

```bash
% df -H
Filesystem       Size  Used  Avail  Use% Mounted on
...               
/dev/sdb          1.9T   2.2G  1.9T   1% /srv/castor/01
/dev/sdc          2.0T   2.2G  2.0T   1% /srv/castor/02
/dev/sdd          2.0T   2.2G  2.0T   1% /srv/castor/03
/dev/sde          2.0T   2.2G  2.0T   1% /srv/castor/04
/dev/sdf          2.0T   2.2G  2.0T   1% /srv/castor/05
/dev/sdg          2.0T   2.2G  2.0T   1% /srv/castor/06
/dev/sdh          2.0T   2.2G  2.0T   1% /srv/castor/07
/dev/sdi          2.0T   2.2G  2.0T   1% /srv/castor/08
/dev/sdj          2.0T   2.2G  2.0T   1% /srv/castor/09
/dev/sdk          2.0T   2.2G  2.0T   1% /srv/castor/10
/dev/sdl          2.0T   2.2G  2.0T   1% /srv/castor/11
```

### Analysis

- 11 RAID1 pairs
- 2 TB * 11 filesystem (* 2 copies)
- ~ 44 TB raw storage in one box
- 1000 boxes → 2000 disks (JBOD)
- + compact (~1 PB in 4 boxes)
- – potentially “too little spindles” (analysis)
- – potentially “too small network I/F” (analysis)
- – data placement problems?
Different setups (mirror)

• **Durability (data on HD1)**
  - \( \approx 1 - p(\text{HD1}) \times p(\text{HD1}') \)
  - First approximation dominated by the number of replicas
  - If anything, in setup 1 there is a correlation (“positive”) which accounts for correlated failures in HD1 and HD1'

• **Availability (data on HD1)**
  - Setup 1
    - \( 1 - p(\text{host1}) \)
  - Setup 2
    - \( \approx 1 - p(\text{host1}) \times p(\text{host2}) \)

As today:
- **CASTOR**: mainly Hardware RAID (RAID1)
- **EOS**: tunable number of full copies (n>1)
Reliability... Good

- File loss is not nice but unavoidable with a certain probability
  - RAID-1 does not protect against controller or machine problem, filesystem corruption and finger trouble
  - Typically important files can be recovered from offsite
- In case of backup (CASTOR) the tape reliability is helping the disk one

Files lost per million

Not lost (because of tape copy)

Real loss (but not primary data)
Network IO for file creations with 3 replicas:

- 1 GB/s output on eth0 of all disk servers (write out of FS1 0.5 GB/s twice)
- 1.5 GB/s input on eth0 of all disk servers (0.5 GB/s x 3 copies)

Plain (no replica)
Replica (here 3 replicas)
More sophisticated redundant storage (RAID5, RAID6, LDPC)

500 MB/s injection result in

1. return code
2. write(offset, len)
3. 2
4. 3
5. 5
6. write(offset, len)
Client **RW** reopen of an existing file triggers
- creation of a new replica
- dropping of offline replica

This is at the heart of our model of operations (asynchronous operations)
Bringing all together...

The **World Wide Web** provides seamless access to information that is stored in many millions of different geographical locations.

The **Grid** is an infrastructure that provides seamless access to computing power and data storage capacity distributed over the globe.
WLCG Tiers Organization

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
How does it work?

ATLAS
Not substantially different for the other HEP experiments
Heavily simplified...
What do we want to achieve
The user wants to specify a subset of the data and run applications on it (chain of programs reading intermediate outputs)
Only at the end of the chain data sizes and computational complexity this can be (possibly) done on a laptop
1000+ of physicists worldwide after the same data
Behind the scenes...

User sitting “anywhere”

Dataset → \{f_1, f_2, \ldots, f_n\}

Dataset → \{site_1, site_2, \ldots, site_k\}

myprogram fj

Dataset content

DataSet subscription

Site services

FTS

Payload transfer

Local access protocol

Monitoring (Dashboard)

Simplified!

Dataset → \{f_1, f_2, \ldots, f_n\}

myprogram fj

Payload transfer

CE+Batch system

fj → /////fj.dat

Physical filenames

Storage

User sitting “anywhere”

Dataset → \{f_1, f_2, \ldots, f_n\}

myprogram fj

Dataset content

DataSet replica

PanDA

Job executor

Computer

Data distribution (asynchronous)
More info:
INFN (Istituto Nazionale Fisica Nucleare): http://www.infn.it
IGI (Italian Grid Initiative): http://www.italiangrid.org/
Outlook

• **CERN data for LHC:**
  - Solid foundation (CASTOR)
  - Successful new product (EOS)
  - CASTOR and EOS coexist (complementary)

• **Goals**
  - Stability with low operations costs
  - Open to adapt to new ideas (also from non-HEP areas)

• **Open directions ahead:**
  - Data management at the heart of our activity
    - Critical for the success of physics research
    - An exciting field of study by itself
    - New stuff coming
      » Is it any good?
Questions?

\( \sqrt{s} = 7 \text{ TeV} \int L dt = 0.05 \text{ fb}^{-1} \)  

**ATLAS Preliminary**  

H→ZZ\(^{(r)}\)→4l channel

- Signal \((m_H = 125 \text{ GeV})\)
- Background ZZ\(^{(r)}\)
- Background Z+jets, t\(\bar{t}\)
- Data
CERN options for students

University level
(BS/Master)
- Summer student
- OpenLab summer students

Master thesis
- Technical student (non physicist)

PhD students
- Doctoral students

Young scientists/engineers
- Fellowship
- Other programmes