Big Data al CERN: come gestire i dati prodotti dagli esperimenti?

Giuseppe Lo Presti
CERN IT Department

Italian Teachers Programme 2019 - Academy
Agenda

• The Big Picture
  • Computing and Data Management at CERN
• Big Data at CERN and outside
• Future Prospects
  • The “High-Luminosity” LHC and its challenges
Ground-breaking ceremony for the High-Luminosity LHC

The earthmovers are at work on the ATLAS site in Meyrin and at CMS in Cessy, digging the new shafts for the High-Luminosity LHC (HL-LHC). The start of the work for this new phase of the project was marked by a ceremony held on 15 June, which was attended by VIP guests including the President of the State Council of the Republic and Canton of Geneva, the Prefect of the Rhône-Alpes-Auvergne region, the Mayor of Meyrin, the Deputy Mayor of Cessy and representatives of CERN’s Member and Associate Member States.

“All the chapters of CERN’s history have begun with a shovel of earth, and each chapter has begun with the promise of great progress in fundamental knowledge, new technologies that benefit society, and collaboration on a European and now a global scale. This was true of the Large Hadron Collider (LHC) and its experiments and it is true of the project for which we are gathered here today,” said Fabiola Gianotti, CERN Director-General.

It would be impossible for anyone to conceive of carrying out a particle-physics experiment today without the use of computers and software. Since the 1960s, high-energy physicists have pioneered the use of computers for data acquisition, simulation and analysis.

This hasn’t just accelerated progress in the field, but driven computing technology generally—from the development of the World Wide Web at CERN to the massive distributed resources of the Worldwide LHC Computing Grid (WLCG) that supports the LHC experiments. For many years these developments and the increasing complexity of data analysis code a wave of hardware improvements that saw computers get faster every year. However, those blissful days of reliance on Moore’s law are now well behind us (see panel overleaf), and this has major ramifications for our field.

The high-luminosity upgrade of the LHC (HL-LHC), due to enter operation in the mid-2020s, will push the boundaries of accelerator and detector technology, bringing enormous challenges to software and computing (CERN Courier October 2017 p5). The scale of the HL-LHC data challenge is staggering: the machine will collect almost 25 times more data than the LHC has produced up to now, and the total LHC dataset (which already stands at almost 1 exabyte) will grow many times larger. If the LHC’s ATLAS and CMS experiments project their current computing models to Run 4 of the LHC in 2026, the CPU and disk space required will jump by between a factor of 20 to 40 (figures 1 and 2).

Even with optimistic projections of technological improvements there would be a huge shortfall in computing resources. The WLCG hardware budget is already around 100 million Swiss francs per year and, given the changing nature of computing hardware and slowing technological gains, it is not the question to simply throw in...
CERN Headquarters, Geneva, Friday 14 July 2017 – SKA Organisation and CERN, the European Laboratory for Particle Physics, yesterday signed an agreement formalising their growing collaboration in the area of extreme-scale computing.

The agreement establishes a framework for collaborative projects that addresses joint challenges in approaching Exascale computing and data storage, and comes as the LHC will generate even more data in the coming decade and SKA is preparing to collect a vast amount of scientific data as well.

This year CERN’s data centre broke its own record, when it collected more data than ever before. During October 2017, the data centre stored the colossal amount of 12.3 petabytes of data. To put this in context, one petabyte is equivalent to the storage capacity of around 15,000 64GB smartphones. Most of this data come from the Large Hadron Collider’s experiments, so this record is a direct result of the outstanding LHC performance, the rest is made up of data from other experiments and backups.

“For the last ten years, the data volume stored on tape at CERN has been growing at an almost exponential rate. By the end of June we had already passed a data storage milestone, with a total of 200 petabytes of data permanently archived on tape,” explains German Cancio, who leads the tape, archive & backups storage section in CERN’s IT department.
From the Hit to the Bit: DAQ

100 million channels
40 million pictures a second
Synchronised signals from all detector parts
From the Hit to the Bit: event filtering

L1: 40 million events per second
   Fast, simple information
   **Hardware** trigger in a few micro seconds

L2: 100,000 events per second
   Fast algorithms in local computer farm
   **Software** trigger in <1 second

EF: Few 1000s per second recorded for offline analysis
   By each experiment!
Data Processing

- Experiments send over 10 PB of data per month
  - 115 PB from all experiments in 2018
- The LHC data is aggregated at the CERN data centre to be stored, processed, and distributed
The CERN Data Centre
The CERN Data Centre

- Built in the 70s on the main CERN site
  - 3.5 MW for equipment
- Nowadays, hardware generally based on commodity
  - \(~15,000\) servers, providing \(230,000\) processor cores
  - \(~130,000\) disk drives providing \(280\text{PB}\) of disk space
    - SSDs increasingly finding their applications (fast caches, metadata journals, tape “repack”, etc.), but still too expensive for massive deployments
  - \(~30,000\) tapes, providing \(0.5\text{EB}\) capacity
- Typical issues
  - PUE (Power Usage Effectiveness) and Green-IT
CERN CC: an ordinary week in numbers
CERN CC: largest scientific data repository
The Worldwide LHC Computing Grid

- The Worldwide LHC Computing Grid (WLCG) is a global collaboration of more than 170 data centres around the world, in 42 countries.

- The CERN data centre (Tier-0) distributes the LHC data worldwide to the other WLCG sites (Tier-1 and Tier-2).

- WLCG provides global computing resources to store, distribute and analyse the LHC data:
  - CERN = only 15% of CPU resources

- The resources are distributed – for funding and sociological reasons.
Software Platforms

- How did we get there?
  - Home made solutions vs. integrating software systems from the market
  - Moving towards the latter as industry grew in front of us!
Take-away #1

- LHC data rates range from the PB/sec at the detector to the GB/sec after filtering
- Scientific data towards the Exabyte scale
  - +100% of LHC data in 2018 vs 2017
- Data centres run on commodity hardware
- Commercial providers are (much) larger
  - CERN remains the world-largest scientific repository
- …Is this really “Big Data”?
Big Data and what’s coming next
Big Data

- *Big data* is a field that treats of ways to analyse [...] or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software (*Wikipedia*)
  - **Moving target** by definition!
- From *structured* data, relational DBs, centralized processing...
- To *unstructured* data and decentralized (i.e. parallel and loosely-coupled) processing, more adapted to the Cloud
  - E.g. trend analysis, pattern recognition, image segmentation, natural language interpretation/translation, …
Big Data out there

- Increasing interest in Big Data analysis
  - **The Power of Data:** Neural Networks are well known since the 1990s, but it’s only now with very large and easily accessible data sets that they become effective!
  - Lots of software frameworks for *Deep Machine Learning* with NNs coming up

[Image of a neural network diagram]
Big Data at CERN

- Experiments have long used Machine Learning (once called Multi-Variate Analysis) techniques
  - Track reconstruction ~ pattern matching
  - Deep Neural Networks coming to help?

- HiggsML and TrackML Challenges
  - 2018 edition: best results obtained with pure parallel processing, **without** ML!
Big Data at CERN

- More recently, LHC Beams Control Logging
  - Data migrated from Oracle DB to Hadoop
    - Explosion of data from (connected) sensors
  - Extract trends and detect/predict failures

- In general, ML techniques are getting attention in contexts where analytical approaches are inapplicable/unpractical
  - Security forensics, system analysis/profiling, etc.
    - Typically boiling down to log analysis
HL-LHC: a computing challenge

LHC / HL-LHC Plan

HL-LHC: x10 luminosity

Magnet splice update
Phase I upgrades (injectors)
Phase II upgrades (final focus)
HL-LHC: a computing challenge

400 PB/year in 2023 (estimated)

50x today’s levels
10x over what technology will provide
HL-LHC and friends

- High Luminosity LHC is not alone in the current arena of large scientific collaborations

- **New Big Science experiments** coming up:
  - Square Kilometer Array (**SKA**)
  - Cherenkov Telescope Array (**CTA**)
  - Deep Underground Neutrino Experiment (**DUNE**): prototype at CERN, full sized experiment in USA

- Time for R&D, opportunity for new **synergies**
  - Typical trend: migrating the 1st Level Trigger from FPGAs to GPUs
  - Increasing role of ML techniques, in particular in other sciences
    - LIGO: GW signal detection
CERN-IT: pushing boundaries

• CERN-IT impact on society through computing:
  • Need for collaboration of computing resources for the Global LHC led to adopt **Grid Computing** and first concept of **Computing Clouds**

• Open access to science
  • Need for sharing the results had led CERN to pave to way to open access to documents and now data: **LHC@home** and **CERN Opendata Portal**

• Openlab
  • “CERN openlab is a unique public-private partnership that accelerates the development of cutting-edge solutions for the worldwide LHC community and wider scientific research”
    • Testing software and hardware
    • **Important student internship program**

• EU projects:
  • HNSciCloud (cloud computing resources), EOSC (data infrastructures)
Explore more than 1 petabyte of open data from particle physics!

Search examples: collision datasets, keywords.education, energy:7TeV

Explore
- datasets
- software
- environments
- documentation

Focus on
- ATLAS
- ALICE
- CMS
- LHCb
- OPERA
PHASE V R&D PROJECTS

High-bandwidth fabrics, accelerated platforms

Data Acquisition (LHCb, IT-CF)

Data Analytics, Machine Learning (many)

Code modernization (EP-SFT, IT-CF)

Simulation, HPC on the Cloud, benchmarking

Cloud infra (IT-CM)

Cloud federations, containers, scalability

Networks (IT-CS)

Data Storage (IT-ST, IT-DB)

Storage architectures, scalability, monitoring

Control Systems (BE-ICS)

Software Defined Networks, Security

Predictive/proactive maintenance and operations

Data quality monitoring, anomaly detection, physics data reduction, benchmarking/scalability, systems biology and large-scale multi-disciplinary platforms

Maria Girone – CERN openlab CTO
To conclude...
From CERN to the world

- Fundamental Science always pushed technology boundaries, with large returns on investments
- For computing, CERN R&D led for instance to:
  - Invention of the Web (1989, cf. #Web30)
    - Key contribution to the Internet infrastructure
    - 80% of the total European Internet traffic going through CERN in the late 1980s
  - Touch screens (1972)
    - Super Proton Synchrotron control system required complex controls and developed capacitive touch screen
    - It was based on open standards and moved into industry

...mmm... web + touch-screen: what do you have in your pocket?
Take-away #2

- **Fundamental** Science continues to be main inspiration for revolutionary ideas, due to revolutionary needs
  - Industry has well defined offer and demand. We do not. This is the key for **innovation**.

- **IT** industry has **globally** evolved **beyond our scale**
  - Big Data analysis techniques gaining more and more momentum
    - But there’s no silver bullet!
  - Hard to invent the ‘next Web’, but plenty of room for collaboration and – again – innovation
    - Openlab, EU projects, etc.
Thanks! (More) Questions?

If time's not up, there's something more…
Bonus:
An interesting real-life case of data corruption

- 11 June 2015: one router line card was discovered to have been corrupting packets overnight, leading to checksum errors and packet drops over a period of around 8 hours.

- Following the incident, a number of corrupted files were detected by some storage applications, including CERNBox and ATLAS DDM
### Table 1. adler32 checksums of detected file corruptions.

<table>
<thead>
<tr>
<th>Filename</th>
<th>adler32 (src)</th>
<th>adler32 (dst)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19a57ba25e9cc5f15d7d27ea49ecb44</td>
<td>4a70cc869</td>
<td>55d4c869</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0290...</td>
<td>4e2f92c2</td>
<td>0cc292c2</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0465...</td>
<td>343db07b</td>
<td>0276a06b</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0490...</td>
<td>3ec61e55b</td>
<td>964a1661</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0498...</td>
<td>67fbb90f</td>
<td>8911b910</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0550...</td>
<td>70243376</td>
<td>73b5336d</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0605...</td>
<td>039e784d</td>
<td>9bc7784d</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0617...</td>
<td>6e6f5a68</td>
<td>b4dd5a66</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0758...</td>
<td>6de8b9fb</td>
<td>d458b9fb</td>
</tr>
<tr>
<td>data13.2p76TeV...ib0761...</td>
<td>e85fa3e0</td>
<td>7b4aa3e3</td>
</tr>
</tbody>
</table>
Examples of corruptions

Table 1. adler32 checksums of detected file corruptions.

<table>
<thead>
<tr>
<th>Filename</th>
<th>adler32 (src)</th>
<th>adler32 (dst)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19a57ba25e9cc5f15d7d27ea49eecb44</td>
<td>a8c869c869</td>
<td>55c869c869</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0290...</td>
<td>4e292c2</td>
<td>0c292c2</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0465...</td>
<td>3e2ab07</td>
<td>027ab06</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0490...</td>
<td>3ec1558</td>
<td>96e1661</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0498...</td>
<td>679b901</td>
<td>89b910</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0550...</td>
<td>7023376</td>
<td>73f33ed</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0605...</td>
<td>038784d</td>
<td>9be784d</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0617...</td>
<td>66e5a68</td>
<td>b4e5a66</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0758...</td>
<td>6debd9f1b</td>
<td>d4eb9fb</td>
</tr>
<tr>
<td>data13.2p76TeV....lb0761...</td>
<td>a8ea3e3</td>
<td>7bea3e3</td>
</tr>
</tbody>
</table>

Half of the adler-32 checksum is almost identical in every case.

Two pairs of corruptions, each 560 bytes apart.

Ferrara, 12/05/2016
G. Lo Presti - Distributed Storage for the LHC experiments
Comparing Adler-32 and TCP checksums

- **Adler-32** is the concatenation of two rolling 16-bit sums, $S_1$ & $S_2$.
  
  for each byte $b$:
  
  $S_1 += b$
  
  $S_2 += S_1$
  
  adler-32 = $S_2 \ll 16 + S_1$ with $S_1, S_2 \mod 65521$

- The **TCP checksum** is a 16-bit sum of all 16-bit words in the TCP packet (1-complemented).
  
  for each word $w$:
  
  $S += (65536-w) \mod 65536$
  
  TCP-cksum = $\lnot S$

  These are the same computation! Adler-32 is partly **redundant** with TCP CRC-16!
When something may go wrong...

- In February 2016, ATLAS reported suspicious file corruptions
  - Jobs crashing when accessing some RAW data at CERN
  - Quickly correlated with the router incident…

- Campaign to identify and stage in from T1s and from CERN tapes all potentially concerned files
  - 9K files, 11 TB of data in the time window of the incident + 2h before and after
  - About 20 TB of data were actually written to the Tier0 in that time window

- Result: out of ~7800 potentially affected files
  - ~10% were detected as corrupted by Adler-32 and retransmitted
  - For cases we have looked at, the S1 parts of the Adler-32 checksums were exactly matching
  - 17 cases (~2%) were false negatives, leaving the corrupted file undetected!
Time distribution of the observed corruptions

- The undetected corruptions are strongly correlated with the traffic as expected.

- The detected corruptions are delayed by ~1h as the log signature we identified corresponds to the attempted migration to tape.
Bonus Take-away

- Silent data corruption may and does happen, and it’s **expensive** if not **impossible** to detect!

- …This is (part of) the fun we have when dealing with large-scale computing and networking

- That’s really all folks!