CERN openlab
Technical Achievements and Challenges

Workshop on DAQ@LHC

Maria Girone
CERN openlab CTO

April 14th 2016
Introduction

› CERN openlab has been created to support the computing and data management goals set by the LHC
  ▪ 15 years of innovative projects between CERN and leading IT companies

› In its phase V, CERN openlab is working to solve some of the key technical challenges facing the LHC in Run3 and Run4
  ▪ Mutual benefit for industry and research communities

› Ever-increasing interest in CERN openlab
  ▪ well established mechanism of partnership between industry and research communities
  ▪ a path to common developments for future challenges

This talk gives a general project overview, highlighting some (but not all) of the achievements. For more details, please refer to the Technical Workshop, 5-6 November 2015 at https://indico.cern.ch/event/452614/ and to specific project reports
CERN openlab in a nutshell

- A unique science – industry partnership to drive R&D and innovation with over a decade of success
- Evaluate state-of-the-art technologies in a challenging environment and improve them
- Test in a research environment today what will be used in many business sectors tomorrow
- Train next generation of engineers/employees
- Disseminate results and outreach to new audiences
LHC Run3 and Run4 Scale and Challenges

Raw data volume for LHC increases exponentially
- And with it processing and analysis load
- Current estimate by Run4 for technology improvements for flat budget is an increase of a factor 8-10

LHCb and ALICE have big upgrades in Run3
- Event rate x 40-100 and factor 10 in volume

ATLAS and CMS upgrade for Run4
- Event rate x 10 and big increase in volume

Maria Girone – CERN openlab CTO
Run3 and Run4 Scale and Challenges

- The increased data volume is combined with an increase of event complexity
  - Resulting in a huge processing challenge
    - Example from CMS, but other experiments are similar

<table>
<thead>
<tr>
<th>Detector</th>
<th>HLT output rate (kHz)</th>
<th>Data Reco.</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase-I</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Phase-II (140)</td>
<td>5</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>Phase-II (200)</td>
<td>7.5</td>
<td>340</td>
<td>7.5</td>
</tr>
</tbody>
</table>

https://cds.cern.ch/record/2020886

- Total computing needs go up by a factor of 65-200 (wrt Run2)
  - Technology improvements only solve a factor of 10
  - **Code optimization** and **technology revolutions** are needed
CERN openlab and WLCG

Recently WLCG presented some goals for solving the gap

Goal

- Assume we need to save factor 10 in cost over what we may expect from Moore’s law
- 1/3 from reducing infrastructure cost
- 1/3 from software performance (better use of clock cycles, accelerators, etc. etc)
- 1/3 from more intelligence – write less data, move processing closer to experiment (keep less) - writing lots of data is not a goal

Some CERN openlab projects directly contribute to the goal

- Computing Management and Provisioning
- Computing Platforms and Code Optimization
- Data Analytics

Others are more directly linked to experiments activities and IT services

- Data Acquisition
- Networks and Connectivity
- Data Storage

Maria Girone – CERN openlab CTO
Information Technology Research Areas

- **Data acquisition and filtering**
  - Collecting data

- **Networks and connectivity**
  - Connecting resources

- **Data storage architectures**
  - Storing and serving data

- **Compute management and provisioning (cloud)**
  - Managing resources for processing

- **Computing platforms, data analysis, simulation**
  - Improving processing and code efficiency

- **Medical applications**

- **Data analytics**
  - Extracting information

---

Maria Girone – CERN openlab CTO
Information Technology Research Areas

Data acquisition and filtering
Collecting data

Networks and connectivity
Connecting resources

Data storage architectures
Storing and serving data

Compute management and provisioning (cloud)
Managing resources for processing

Computing platforms, data analysis, simulation
Improving processing and code efficiency

Data analytics
Extracting information

Maria Girone – CERN openlab CTO
The next runs at LHC represent technical challenges in real time filtering, data movement and networking, high level trigger (event selection) and partial reconstruction

- The High Throughput Computing Collaboration investigates the use of Intel technologies in trigger and data acquisition (TDAQ) systems
  - Investigate benefits of Xeon/FPGA, Omni-Path interconnect, Xeon Phi (KNL)

<table>
<thead>
<tr>
<th>L1 Trigger</th>
<th>DAQ</th>
<th>High-Level Trigger</th>
</tr>
</thead>
<tbody>
<tr>
<td>High efficiency despite overlapping collisions add tracking information</td>
<td>Collision data spread over 10’000 pieces</td>
<td>large software infrastructure</td>
</tr>
<tr>
<td>Flexible, robust and easy to reproduce</td>
<td>Data gathered onto one of 1000s compute units</td>
<td>flat time profile</td>
</tr>
<tr>
<td>Algorithms must decide in O(10) ms</td>
<td>Compute units run complex filter algorithms</td>
<td>complex and costly algorithms for reconstruction</td>
</tr>
<tr>
<td>~10000x</td>
<td>~1000x</td>
<td>~3000x</td>
</tr>
</tbody>
</table>

Maria Girone – CERN openlab CTO
HTCC - Selected Results

Real time filtering calculations
Xeon FPGA computing accelerator results
- Sorting: a factor 50 faster
- Mandelbrot: a factor 12 faster
- Cubic root: a factor 35 faster

Accelerating the HLT using many cores
- Use Intel next-gen Xeon Phi to speed up big consumers of computing cycles
  - Pattern recognition & tracking
  - Particle identification
- Pattern recognition and tracking with OpenMP and Intel Thread Building Blocks (TBB) code (Velopixel demonstrator)

Parallel event sorting and building results
- Benchmark for many core parallel collision event grouping
  - On KNC, up to 26GB/s bandwidth measured

Benchmarks of Intel Omni-Path and Infiniband EDR
- Already observing 75 Gb/s on EDR

More than a factor 10
HTCC Next Steps

• Xeon FPGA
  • Implement and test the acceleration of other high-level trigger parts, e.g. tracking, Kalman Filter
  • Test the use of platform for rawdata encoding for calorimeter upgrade.

• Xeon Phi
  • Perform benchmarks on next-gen hardware for high-level trigger and event sorting codes
  • Implement other parts of high-level trigger using OpenMP or TBB

• DAQPIPE
  • Scalability tests (on Gallileo at INFN and Curie for a 500 node scale test).
  • Some short studies on failure recovery

Maria Girone – CERN openlab CTO
Data Analytics use case

› Evaluate throughput using low latency and low power RapidIO interconnect
  - IT infrastructure monitoring and logging data
  - Exploring direct reads from ROOT

› Finalize use-cases for analytics and start use-cases for DAQ
Information Technology Research Areas

Data acquisition and filtering
Collecting data

Networks and connectivity
Connecting resources

Data storage architectures
Storing and serving data

Compute management and provisioning (cloud)
Managing resources for processing

Computing platforms, data analysis, simulation
Improving processing and code efficiency

Data analytics
Extracting information

Maria Girone – CERN openlab CTO
Data Storage Architectures

Concluded Huawei/S3 storage project in 2015

- performed and documented comparison of ROOT workloads with different IO patterns
- confirmed importance of S3 vector reads for sparse / random analysis workload
- implemented S3 backend to CVMFS systems and evaluated performance and stability in a prototype setup for the LHCb experiment

In 2016 LHC has 300PB of disk storage and 400PB of tape

- Increasing selection rate in Run3 and Run4 pushes this exponentially
  - Looking at expansion ideas and new architectures
Storage Technology R&D

Object-disks (Seagate/Kinetic) are now available and offer:

- Semantics matched to shingled recording (required for volume growth)
- An open API (supported by all major vendors - Seagate, Toshiba, WD)
- Since March 2015:
  - Transparent integration with EOS system achieved
- Planned for 2016:
  - Evaluate TCO gain within a EOS prototype system and Active Disk with ROOT

The Storage Technology team is also looking at Non-Volatile Memory:

- NVRAM is available now - may allow to solve some persistent meta-data problems at DRAM speed
- Evaluating gains for EOS catalogue
- Comparison to lower performance alternatives like Flash/SSD

Maria Girone, CERN openlab CTO
Information Technology Research Areas

Data acquisition and filtering
Collecting data

Networks and connectivity
Connecting resources

Data storage architectures
Storing and serving data

Compute management and provisioning (cloud)
Managing resources for processing

Computing platforms, data analysis, simulation
Improving processing and code efficiency

Data analytics
Extracting information
WLCG has more than half a million processor cores
- OpenStack is heavily used at CERN
- The community is looking at expanding to dynamically provisioned resources

Rackspace Collaboration – 2H 2015
- Collaboration on cloud federation with the upstream OpenStack community
- Keystone-to-Keystone bursting capabilities
  - Allows multiple OpenStack clouds to trust each others’ identity management which simplifies configuration
  - Service provider filtering allows only certain cloud services to be exposed to the collaborating clouds such as only exposing the object store and not the compute resources
- Shadow users
  - Unify the local and federated users so authentication tokens are identical and auditing/billing is consistent

Maria Girone – CERN openlab CTO
Rackspace Collaboration - 1H 2016 Plans

› Federation functionalities are now evolving smoothly with new releases
› Focus now shifting to enhancing container support in OpenStack for scientific computing
  ▪ Magnum is a recently started project integrating docker, kubernetes, mesos into OpenStack
  ▪ Uses existing OpenStack components for provisioning, security, metering, networking and storage
  ▪ Follows the same upstream first model as for Federation with the plans to be defined during the Austin OpenStack summit in April
Information Technology Research Areas

- Data acquisition and filtering
  - Collecting data

- Networks and connectivity
  - Connecting resources

- Data storage architectures
  - Storing and serving data

- Compute management and provisioning (cloud)
  - Managing resources for processing

- Computing platforms, data analysis, simulation
  - Improving processing and code efficiency

- Data analytics
  - Extracting information
Alice O² Upgrade

By Run 3 most of the ALICE detectors and its computing system are expected to inspect and read-out all the interactions up to a rate of \(6\text{TB/s}\) and storing up to \(90\text{GB/s}\).

This is a major processing and data handling challenge and unprecedented in a Heavy Ion detector.

FLPs (First Level Processor) receive data from the detector readout, preprocess it (FPGA), chop it into manageable pieces (time frames) and send it out to EPNs.

EPNs (Event Processing Node) collect sub-time frames from all FLPs to build a full time frame for reconstruction (on the EPN nodes).

Goal: Reduce the I/O latency, avoid memory copy, context switching and system calls
Investigating impact of removing kernel mediation from the data path on distributed applications

Implement support for Cisco’s usnic in nanomsg framework

- 40Gbps throughput, 2µs p2p latency, low CPU load

Measure impact on O2’s applications in the ALFA framework
The increasing need for computing has prompted an effort to optimise scientific codes for the new computing architectures

- Possible to achieve enormous improvements in code performance using modern techniques
- One of the few areas with enough potential for improvement to close the resource gaps in the upgrade program

The Code Modernization Project is an umbrella for addressing several use cases in different disciplines

- Possible extensions currently under discussion
**The GeantV project**

**Rethinking particle transport**

### HOW

- Intel expertise and tools through a IPCC program
- Rethink particle transport in detector simulations
- R&D on the vertical scalability to profit from multiple levels of parallelism and use of accelerators.
- Code improvements are applicable to all fields that use this simulation framework
  - Medical applications

### WHY

- Detector simulation is one of the most CPU intensive tasks in modern HEP
  - 50% of the WLCG cycles are used by simulation
  - Improving the simulation by factors would allow for customized samples, more simulation, and more potential for discovery
GeantV results on Xeon Phi

X-Ray benchmark

The X-Ray benchmark tests geometry navigation in a real detector geometry.

- X-Ray scans a module with virtual rays in a grid corresponding to pixels on the final image.
- Probed the vectorized geometry elements + global navigation as task.

Geometry is 30-40% of the total CPU time in Geant4.

A library of vectorized geometry algorithms to take maximum advantage of SIMD architectures.

<table>
<thead>
<tr>
<th></th>
<th>16 particles</th>
<th>1024 particles</th>
<th>SIMD max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Ivy-Bridge (AVX)</td>
<td>~2.8x</td>
<td>~4x</td>
<td>4x</td>
</tr>
<tr>
<td>Intel Haswell (AVX2)</td>
<td>~3x</td>
<td>~5x</td>
<td>4x</td>
</tr>
<tr>
<td>Intel Xeon Phi (AVX-512)</td>
<td>~4.1</td>
<td>~4.8</td>
<td>8x</td>
</tr>
</tbody>
</table>

Overall performance for a simplified detector vs. scalar ROOT/5.34.17

Saturation beyond 2x cores

Balanced model better

Basketized performance approaching ideal (N times same) case

Vector/scalar profit maximized for nthreads≈core_count

Maria Girone, CERN openlab CTO
As part of their upgrade ALICE is collaborating with FAIR (an ION accelerator in Germany). ALICE-FAIR project aiming to massive data volume reduction by (partial) online reconstruction and compression

- tighter coupling between online and offline reconstruction software

ALFA constitutes a framework that contains:
- Transport layer (FairMQ, based on: ZeroMQ, nanomsg)
- Configuration tools
- Management and monitoring tools
- A data-flow based model (Message Queues based multi-processing)
- Provide unified access to configuration parameters and databases

FairRoot is a common system for reconstruction and simulation

FairRoot/ALFA allows the use of hardware accelerators to improve performance by treating tasks separately
- GPUs, Phi’s, etc.
- Work is ongoing to improve the transport performance to/from the Phi with FairMQ

https://fairroot.gsi.de
Community access to CERN openlab Technology

- CERN IT has mechanisms to test off-the-shelf hardware

- However, we are aware of community interest in evaluating next-generation hardware and software tool with their use cases

- Several vendors expressed an interest in having their next generation equipment tested in a lightweight project structure

- We are investigating how to best make this equipment available and how to handle any required NDA’s, as well as how to share the results with the broader community
Information Technology Research Areas

- **Data acquisition and filtering**
  - Collecting data

- **Networks and connectivity**
  - Connecting resources

- **Data storage architectures**
  - Storing and serving data

- **Compute management and provisioning (cloud)**
  - Managing resources for processing

- **Computing platforms, data analysis, simulation**
  - Improving processing and code efficiency

- **Data analytics**
  - Extracting information
How to make more effective use of the data collected is critical to maximise scientific discovery and close the resource gap

- There are currently ongoing projects in
  - System controls
  - Data Storage and quality optimizations

- Organising projects on
  - Data reduction
  - Optimized formats
  - Investigations for machine learning for analysis and event categorization
The LHC is the largest piece of scientific apparatus ever built

- There is a tremendous amount of real time monitoring information to assess health and diagnose faults
- The volume and diversity of information makes this an interesting application of big data analytics.

Designed and developed algorithms for several use-cases to improve the robustness and performance of control systems

### Online monitoring for operational support
- Detection of cryogenics valve oscillation
- Anomaly detection by sensors data mining

### Fault diagnosis support
- Root cause analysis for system alarms
- Discovery of fault sensors measurements by a rule/model-based approach

### Engineering & design support
- Automatic evaluation of PID supervision
Data Analytics for Storage and Data Quality Optimization

Data Storage Optimization
- Developed and tested interpretable algorithm that allows saving up to 40% of disk storage
  - Data placement based on popularity

Data quality management (anomalies detection)
- Developed algorithm for unmanned anomalies detection for rare anomalies
  - Tested on 2015 data

General issues for the experiments and sites

Maria Girone – CERN openlab CTO
Physics Data Analytics

After the upgrade LHC will collect large datasets. Investigating ways to more efficiently select events from the stream of data using “big data” techniques

- Traditional methods in HEP for deriving a rich data sample from a big data sample have not changed much in high energy physics computing

- Need to reduce multi-petabyte datasets by a factor of 1000 based on physics selection criteria
  - Performance, reproducibly, and completeness are all important
Workshop on Machine Learning and Data Analytics

› Will gather together experiments and industry for a day of discussions on April 29th (Intel, Siemens, IBM, Microsoft, Yandex, google, Oracle, cloudera and NVIDIA

▪ https://indico.cern.ch/event/514434/

› Opportunity to discuss on challenges and agree on projects of common interest between our community and industry

Maria Girone – CERN openlab CTO
Looking Forward

CERN openlab V has completed its first year
- Consolidation of the ongoing projects, while ensuring innovation and technology evolution are key

Run3 and Run4 represent significant technical challenges
- Big gap in processing need vs what can be procured with flat budgets and expected technology improvements. Solutions will need to be found.
  - New architectures and fabrics show potential for big gain
  - Software parallelization and vectorization can dramatically improve performance
  - Dynamically provisioned resources and improved virtualization grow computing resources
  - Better use of the data through improved analysis using big data analytics techniques

Maria Girone, CERN openlab CTO
EXECUTIVE CONTACT
Alberto Di Meglio, CERN openlab Head
alberto.di.meglio@cern.ch

TECHNICAL CONTACT
Maria Girone, CERN openlab CTO
maria.girone@cern.ch

Fons Rademakers, CERN openlab CRO
Fons.rademakers@cern.ch

COMMUNICATION CONTACT
Andrew Purcell, CERN openlab Communication Officer
andrew.purcell@cern.ch

Mélissa Gaillard, CERN IT Communication Officer
melissa.gaillard@cern.ch

ADMIN CONTACT
Kristina Gunne, CERN openlab Administration Officer
kristina.gunne@cern.ch