From raw data to new fundamental particles: The data management lifecycle at the Large Hadron Collider

Andrew Washbrook
School of Physics and Astronomy
University of Edinburgh
Dealing with Data Conference
31st August 2015

With thanks to: Wahid Bhimji, Simone Campana, Roger Jones, Borut Kersevan and Paul Laycock for providing material
The Large Hadron Collider

- Large Hadron Collider (LHC) based at CERN, Geneva
- 27 km in circumference
- Protons collide every 25 ns travelling at 99.9999991% of speed of light
- 1,232 superconducting magnets cooled below 2K (-271C)
- Beam size 0.2mm
- Studying interactions at the scale of $10^{-16}$ m

- 8 TeV collision energy during Run 1 (2010-2013)
- Run 2 started earlier this year at an energy of 13 TeV

✓ Higgs Boson

Supersymmetry

Dark Matter
The LHC Detectors

- Proton collisions are detected and recorded by multi-purpose detectors surrounding the four collision points.

The ATLAS experiment
- 44m x 25m in size
- 7,000 tonnes
- 100 million electronic channels
- A collaboration of 2,900 physicists at 176 institutions in 38 countries

I will use ATLAS as an example to highlight the techniques used to handle LHC data.
Data Collection

Trigger Systems
- Rapidly decide which collision events to keep for further analysis
- Selection criteria based upon early identification of "interesting" events (such as the decays of rare particles)
  - e.g. One event in a billion has direct evidence of the Higgs Boson
- Ultimately limited by hardware resources - we collect as much as we can to optimise discovery potential
- Events selected by the Trigger are passed for offline storage for further processing
- 99.99% of the data collected are discarded at source

**Level 1 Trigger**
- Fixed latency of 2.5 μs
- Identifies regions of interest in the detector
- Custom hardware (FPGAs)

**High Level Trigger**
- Accesses full detector resolution
- Customised fast software on commercial CPUs

ATLAS Event Display
Detector Simulation

The raw data collected from the LHC is only part of the bigger data picture

- Data-driven analysis depends upon Monte-Carlo simulation to model the data as accurately as possible
- Translate theoretical models into detector observations
- Proper treatment of background estimation and sources of systematic errors
- Expect multiple interactions per crossing (pile-up)
- 10 billion events simulated by ATLAS to date

Comparable storage requirements to raw data in addition to a significant processing overhead
Data Volume

The LHC has already delivered billions of recorded collision events

ATLAS

- Trigger has 100 million channels and reads out 40 million events per second =\~ 1 PB/s
- Over 10 PB/year of raw data stored for analysis
- Data volume is expected to double over the next two years

LHC

- Over 100 PB of data recorded
- Several 100 PB more storage needed for data replication, simulation and analysis derivation
Data Management
Distributed Data Management

Before the start of LHC operations there was no ready made solution to our "Big Data" requirements

Worldwide LHC Computing Grid (WLCG)
- Seamless access to resources at hundreds of computing centres built upon Grid technologies
- Proved to be a successful resource during Run 1 operations
- Provides over ~200 PB of disk space and a further 200PB of tape
- It was hoped that this would be adopted as a general computing utility solution - before Clouds appeared on the horizon

LHC experiments rely on distributed computing resources arranged in a Tier structure

ATLAS Tiered computing model

<table>
<thead>
<tr>
<th>Tier</th>
<th>Sites</th>
<th>Role</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>Central Facility for data processing</td>
<td>CERN</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>Regional computing centres with high quality of service, 24/7 operations, large storage and compute resources</td>
<td>RAL</td>
</tr>
<tr>
<td>2</td>
<td>140</td>
<td>Computing centres within a regional cloud used primarily for data analysis and simulation</td>
<td>Edinburgh (ECDF)</td>
</tr>
</tbody>
</table>
Data Management Systems

**Distributed Data management (DDM) systems** developed by each LHC experiment enable data placement, replication and access control across WLCG computing sites.

**Rucio**
- Framework to manage all ATLAS data on the Grid
- Discover, transfer and delete data across all registered computing sites
- Ensure data consistency
- Client tools for end users to locate, create and delete datasets on the Grid

<table>
<thead>
<tr>
<th>Experiment</th>
<th>DDM Solution</th>
<th>Workload Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATLAS</td>
<td>Rucio (formerly DQ2)</td>
<td>PanDA</td>
</tr>
<tr>
<td>LHCb</td>
<td></td>
<td>Dirac</td>
</tr>
<tr>
<td>CMS</td>
<td>CRAB</td>
<td>PhedEX</td>
</tr>
<tr>
<td>ALICE</td>
<td></td>
<td>AliEN</td>
</tr>
</tbody>
</table>

Data and Workload management systems by LHC experiment

**X509-based certificate authentication and VO authorisation**

**List of dataset replicas**

**Retrieve dataset**

**DDM Client Tools**
- `$ voms-proxy-init -voms atlas`
- `$ dq2-ls -r DATASETNAME`
- `...
- COMPLETE: BNL-OSG2_DATADISK, SARA-MATRIX_DATADISK, TAIWAN-LCG2_DATADISK, TRIUMF-LCG2_DATADISK`
- `$ dq2-get DATASETNAME`

**Other Data Management Activities**
- **Database caching**: Squid proxy installed at each computing centre to efficiently access central databases
  - e.g. Retrieving detector conditions information for event reconstruction
- **Software access**: CVMFS - http read-only file system for synchronised software access
Data Transfer and Replication

**LHC Experiments continually transfer a lot of data between storage endpoints**

**Data Transfer**
- Use multi-stream transfer protocols: `GridFTP`, `xrootd`, `http`
- Orchestrated through a File Transfer Service (FTS)

**Data Replication**
- Set of rules defined in DDM systems to automate data distribution
- Define minimum number of replicas of a dataset stored on the Grid
- Interface for users to request datasets to be replicated at a particular location (or staged from tape)
- No optimal solution to follow for data placement to ensure minimal transfer and access latency
  - Policy evolves with changes in technology and the needs of researchers
  - Recent efforts to survey data popularity using analytic tools

We (approximately) delete as much as we write

**ATLAS worldwide transfer volume per day**

**ATLAS file deletion activity**
On rare occasions the data volume transferred between computing sites can be extraordinary.

Recent redistribution exercise transferred hundreds of TB of experiment data to storage in Edinburgh.

Mea Culpa! A transfer cap of 5 Gbps is now in place as a safeguard against future rate increases.

### IS Alert

**EaStMAN Network external link congestion**

**Unplanned; Complete**

**Overview**

- External internet connection for EaStMAN
- Degraded Service Time: 10:10 AM, 10-Aug-15 — 12:00 PM, 10-Aug-15

Unusually high levels of network traffic on the EaStMAN network link to Janet are causing a degradation of network connectivity to the outside world. This is being investigated with ECDF systems located at the ACF Bush. UPDATE 12:00: jobs on the research systems generating the unusually high traffic levels have been stopped, and traffic levels have returned to normal. Options will be investigated for preventing a repeat of this incident.

### ATLAS transfer volume to Edinburgh Grid storage (by day)

<table>
<thead>
<tr>
<th>Site</th>
<th>Transfer Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKK-LT2-Brunel</td>
<td>3.1 Gbps</td>
</tr>
<tr>
<td>UKK-LT2-IC-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-LT2-QMUL</td>
<td>3.2 Gbps</td>
</tr>
<tr>
<td>UKI-LT2-RHUL</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-LT2-UCL-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-NORTHGRID-LANCSC-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-NORTHGRID-LIV-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-NORTHGRID-MAN-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-NORTHGRID-SHEF-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SCOTGRID-DURHAM</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SCOTGRID-ECDF</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SCOTGRID-GLASGOW</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-BHAM-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-BRIS-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-CAM-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-OX-HEP</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-RALPH</td>
<td>3.4 Gbps</td>
</tr>
<tr>
<td>UKI-SOUTHGRID-SUSX</td>
<td>3.4 Gbps</td>
</tr>
</tbody>
</table>

### Network Traffic history from JANET to EaStMAN

- **UK.**
- **UKN.**
- **UKM.**
- **UKH.**
- **UKS.**

**Transfer Volume**

- **2015-08-01 00:00 to 2015-08-22 00:00 UTC**

**Destinations:**
- UK. UKN. UKM. UKH. UKS.
Data Processing

Over 350,000 CPU cores available to LHC experiments through the WLCG enabled computing centres

- HEP data is ideal for batch processing
- Events can be processed independently without inter-process communication (e.g. MPI)

- Detector simulation is CPU intensive - approximately 10 minutes per event
- Event reconstruction roughly ~20s per event but requires over 3GB of memory

PanDA
- Uses a pilot model to pull jobs from central queue once a suitable resource found
  - Late binding of jobs to payload
- Pilot factories continually submit jobs to available computing resources
  - Self-regulating workload

ATLAS total slots of running jobs by job type
Data Placement

How do our batch jobs access datasets distributed across hundreds of computing sites?

“Jobs to the Data”
- Traditional approach
- Dataset replicas are transferred ahead of time to computing sites before processing
- Storage and compute resources have to be co-located and balanced

Federated Storage Model
- A flexible, unified data federation presents a common namespace across remote sites
- Exploit improved network capabilities by accessing datasets from an application over the WAN
- Allows jobs to be run at sites with available CPUs without the dataset being stored locally

FAX data request flow
Edinburgh Involvement

The University of Edinburgh provides Tier-2 computing resources to LHC experiments through the Edinburgh Compute and Data Facility (ECDF)

- One of only a few sites globally that successfully pledges resources to LHC through a shared computing facility
- Pilot model is effective in soaking up opportunistic resources on the ECDF cluster
- Expanding processing and storage capabilities to use other University resources
  - Storage interface to ~150 TB of storage at Research Data Facility
  - LHC event simulation can be run on Archer supercomputer

Edinburgh Resources Delivered to LHC
- 57 million kS12K CPU hours over the last 5 years
- 1 PB of dedicated Grid storage

Edinburgh PPE Group
- Member of the ATLAS and LHCb collaborations
- 40 staff and students
- Leading roles on data analysis, trigger systems, distributed computing and software frameworks

Edinburgh Tier-2 Infrastructure
The Future (and the past)
Future Requirements

- LHC experiments will continue to collect data for many more years at higher energies and higher data rates
- Event complexity will increase by an order of magnitude
  - Event data size and processing time will be higher
  - Data simulation requirements will be much larger

Future Data Taking
- Run 2 (now - 2019)
  - Double the dataset size
- High-Luminosity LHC (2024-)
  - 10x increase in event rate
  - ~200 PB produced per year

WLCG Disk Projection to 2017
(Ian Bird - WLCG Collaboration workshop)

ATLAS Disk Projection to 2023

More resources needed on a flat budget!
Computing Model Evolution

LHC computing models are being re-evaluated to meet constraints on resources

- More flexibility in the model to open up opportunistic resources
  - e.g. HPC facilities, commercial clouds, volunteer computing
- Responsiveness to changes in computing and storage technology
- Optimisation of data analysis workflows
- Limit avoidable resource consumption

Tier-less Model

- Tiered model is effectively obsolete - some Tier 2s are now equivalent in size to Tier 1 facilities
- Network bandwidth has increased more than anticipated
  - Data can reside anywhere
  - Better use of storage resources
Data Preservation

Analysis Preservation
- Goal is to reproduce data analyses many years after their initial publication
- Cross experiment pilot prototyped using Invenio digital library platform

Open Data Initiatives
- Disseminate selected datasets cleared for public release
- Allow any member of the public to study experiment data
- Four levels of data access defined by HEP community

There is a distinction between preservation (and sharing) effort for internal use and for public use
Conclusions

- The management and processing of LHC data to produce timely physics results was a big success in Run 1

- The discovery of the Higgs Boson would not have been possible without the coordinated access to resources across hundreds of computing sites

- The LHC faces big computing challenges ahead to avoid constraining science output

- Lots more excitement to come in LHC Run 2 and beyond