ATLAS TDAQ system: current status and performance

Sergio Ballestrero
University of Johannesburg
& CERN PH/ADT

on behalf of the
ATLAS Collaboration
Outline

• Introduction
  - LHC&ATLAS
  - TDAQ Design, scheme
  - 2010 status
• Evolution from 2010 to 2011
  - Luminosity
  - Trigger and data rates
  - Event Filter expansion
  - ROS optimisation and updates
• Status in 2011
  - Efficiency
  - TDAQ & Controls Software
• Conclusion
• Outlook
Nominal working conditions:

**p-p beams:** $\sqrt{s}=14$ TeV; $\mathcal{L}=10^{34}$ cm$^{-2}$s$^{-1}$; Bunch Cross every 25 ns

**Pb-Pb beams:** $\sqrt{s}=5.5$ TeV; $\mathcal{L}=10^{27}$ cm$^{-2}$s$^{-1}$

### 2010
- $\sqrt{s}=7$ TeV
- up to 233 colliding bunches in ATLAS
- Peak $\mathcal{L} \sim 10^{32}$ cm$^{-2}$s$^{-1}$ (October)
- Pb ion run (November)

### 2011
- $\sqrt{s}=7$ TeV
- May: 1092 bunches, $\mathcal{L}_{\text{peak}} \sim 1.26 \cdot 10^{33}$ cm$^{-2}$s$^{-1}$
- goal: 1400 bunches, $\mathcal{L} \sim 1.6 \cdot 10^{33}$ cm$^{-2}$s$^{-1}$
- possible higher $\mathcal{L}$ with beam optimisations
ATLAS TDAQ Design

Trigger

- **Level 1**
  - <2.5 μs
  - Regions Of Interest
  - L1 Accept 75 (100) kHz

- **Level 2**
  - ~40 ms
  - L2 Accept ~3 kHz

- **Event Filter**
  - ~4 sec
  - EF Accept ~200 Hz

- **High Level Trigger**
  - ~3 kHz

- **Event Filter**
  - ~200 Hz

DAQ

- **Detector Read-Out**
  - FE
  - ROD

- **Data Collection Network**
  - Event Filter Network
  - SubFarmOutput

- **Event Filter**
  - ~3 kHz

- **Event Builder**
  - 112 (150) GB/s

- **SubFarmInput**
  - Data-Flow
  - ~4.5 GB/s

- **CERN Data Storage**
  - ~300 MB/s

ATLAS Event
1.5 MB/25 ns

ATLAS Data
Trigger Info

ATLAS Data

Trigger Info

[~90M channels]

~98% fully operational in 2010

Surface

Underground

Region Of Interest (ROI)

~90M channels

~98% fully operational

in 2010

ATLAS TDAQ Implementation

(also to Design)
### TDAQ Farm status: 2010 & early 2011

<table>
<thead>
<tr>
<th>Component</th>
<th>Installed</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online &amp; Monitoring</td>
<td>100%</td>
<td>~60 nodes</td>
</tr>
<tr>
<td>ROSes</td>
<td>100%</td>
<td>~150 nodes</td>
</tr>
<tr>
<td>ROIB &amp; L2SVs</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>HLT (L2+EF)</td>
<td>~50%</td>
<td>~800 xpu nodes; ~300 EF nodes (since October 2010)</td>
</tr>
<tr>
<td>Event Builder</td>
<td>100%</td>
<td>~60 nodes (exploiting multi-core)</td>
</tr>
<tr>
<td>SFO</td>
<td>100%</td>
<td>Headroom for high instantaneous throughput</td>
</tr>
<tr>
<td>Networking</td>
<td>100%</td>
<td>Redundancy deployed in critical areas</td>
</tr>
</tbody>
</table>

- **27 XPU racks ~800 XPU nodes**
  - XPU = L2 or EF Processing Unit (connected to both EF & L2 networks) on a “run by run” basis can be configured to run either as L2 or EF, allows high flexibility to meet the trigger needs
  - Functional role assignment not automated
- **10 EF racks ~300 EF nodes**
  - EF nodes are dedicated (connected to EF network only)
**TDAQ rates 2010**

**Trigger**
- Level 1: <2.5 μs
- Level 2: ~40 ms
- Event Filter: ~4 sec
- High Level Trigger: ~300 ms

**DAQ**
- Detector Read-Out
- Event Filter
- Event Builder
- Data Collection Network
- ROD

**Network**
- SubFarmInput
- SubFarmOutput
- Event Filter Network
- ROD

**Event Builder Info**
- 3.5 kHz
- 30 GB/s
- ~500 MB/s

**ReadOut System**
- 40 MHz
- ~1 MHz
- 20 kHz

**Regions Of Interest**
- L1 Accept: 75 (100) kHz
- ROI data (~2%)
LHC from 2010 to 2011

LHC has delivered excellent luminosity since the beginning of 2011

\[
\mathcal{L}_{\text{peak}} = 1.26 \cdot 10^{33} \text{ cm}^{-2}\text{s}^{-1}
\]

\[
\mathcal{L}_{\text{int}} = 539 \, 768 \text{ pb}^{-1}
\]

so good that we can't put the 2010 plots to scale!
ATLAS Data Calo/Muon Detectors

**Data Flow**

- **Event Filter**
  - High Level Trigger
  - Level 2
  - Event Filter
  - L2 Accept: ~3 kHz
  - EF Accept: ~200 Hz

- **ROI Requests**
  - L1 Accept: 75 (100) kHz

- **ROD**
  - Detector Read-Out
  - FE

- **Data Collection Network**
  - Data Flow
  - SubFarmInput
  - Event Filter Network
  - SubFarmOutput
  - Event Builder

- **ReadOut System**
  - Data Flow
  - CERN Data Storage

- **ATLAS Event**
  - 1.5 MB/50 ns
  - 1.5 MB/150 ns
  - 1.5 MB/25 ns

- **TDAQ rates 2010 → 2011**
  - 2010
    - 5 kHz
    - ~300 Hz
    - ~0.6 s
    - ~45 ms
    - ~400 ms
  - 2011
    - 10 kHz
    - ~3 kHz
    - ~5 kHz
    - ~350 Hz
    - ~300 Hz
    - ~0.6 s
    - ~45 ms
    - ~400 ms

**See Trigger talk by S. Rajagopalan**

**ATLAS Data Trigger Info**

- **CERN Data Storage**
  - ~450 MB/s
  - ~7.5 GB/s
  - 75 GB/s
  - ~3.5 kHz
  - ~20 kHz
  - ~350 Hz
  - ~300 Hz
  - ~300 Hz
  - ~5 kHz
  - ~0.6 s
  - ~45 ms
  - ~400 ms

- **ATLAS Event**
  - 1.5 MB/50 ns
  - 1.5 MB/150 ns
  - 1.5 MB/25 ns

- **Event Filter**
  - Data Collection Network
  - Event Builder
  - SubFarmInput
  - Event Filter Network
  - SubFarmOutput
  - Data Flow
  - CERN Data Storage
Luminosity vs EF CPU load

- **EF CPU usage**
  - $\sim$ linear with $\angle$
  - $> \text{linear with } \mu$
  (beam optim. will increase pileup)

- **Balance XPU assignment**
  between L2 and EF
  given other constraints as
  - EF processing time
  - ROS load
  - XPU network B/W
  - Tier0 storage B/W
# TDAQ Farm 2011: HLT expansion

<table>
<thead>
<tr>
<th>Component</th>
<th>Installed</th>
<th>Comments</th>
<th>Mid 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online &amp; Monitoring</td>
<td>100%</td>
<td>~60 nodes</td>
<td>100%</td>
</tr>
<tr>
<td>ROSes</td>
<td>100%</td>
<td>~150 nodes</td>
<td>100%</td>
</tr>
<tr>
<td>ROIB &amp; L2SVs</td>
<td>100%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>HLT (L2+EF)</td>
<td>~50%</td>
<td>~300 EF nodes -&gt; 750</td>
<td>~70% of design #</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~800 xpu nodes -&gt; assign more to L2</td>
<td></td>
</tr>
<tr>
<td>Event Builder</td>
<td>100%</td>
<td>~60 nodes (exploiting multi-core)</td>
<td>100%</td>
</tr>
<tr>
<td>SFO</td>
<td>100%</td>
<td>Headroom for high instantaneous throughput</td>
<td>100%</td>
</tr>
<tr>
<td>Networking</td>
<td>100%</td>
<td>Redundancy deployed in critical areas</td>
<td>100%</td>
</tr>
</tbody>
</table>

- Using full CPU capacity on EF racks
- Network limit for EF on XPU racks (beyond design rates)
  2 Gbit links/rack → maximum possible EF bandwidth with XPU racks ~6 GB/s

**Conclusion:**
- it's time to install more EF racks, according to design plans, and reassign XPU to L2
- also rolling update of XPU racks to current CPUs
Pixel ROS mapping optimisation

- RODs for adjacent detector sections connected to different ROSes
- for a single ROI requires requests to >1 ROS, increases total requests
- optimal ROD-ROS links allow rates further above design
- implemented January 2011

More details in talk by R. Ospanov
ROS L2 vs EB load

- **ROS provide L2 & EB**
  - L2: ROI fragments
  - EB: all fragments
- **Total nr. of requests limited by CPU load**
- **5KHz @EB above design (3kHz)**
  → too little headroom for L2 requests

![Graph showing event building rate vs. ROI requests per event](image)

- 75 kHz L1A rate, fragment size: 400 words
- **Current CPUs**
- **New CPUs** *
- Design rate
- * not optimised, does reach network B/W limits
## TDAQ Farm 2011: update ROS & SFO

<table>
<thead>
<tr>
<th>Component</th>
<th>Installed</th>
<th>Comments</th>
<th>Mid 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online &amp; Monitoring</td>
<td>100%</td>
<td>~60 nodes</td>
<td>100%</td>
</tr>
<tr>
<td>ROSes</td>
<td>100%</td>
<td>~150 nodes -&gt; update each</td>
<td>100%</td>
</tr>
<tr>
<td>ROIB &amp; L2SVs</td>
<td>100%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>HLT (L2+EF)</td>
<td>~50%</td>
<td>~300 EF nodes -&gt; 750</td>
<td>~70%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>~800 xpu nodes -&gt; assign more to L2</td>
<td></td>
</tr>
<tr>
<td>Event Builder</td>
<td>100%</td>
<td>~60 nodes (exploiting multi-core)</td>
<td>100%</td>
</tr>
<tr>
<td>SFO</td>
<td>100%</td>
<td>Increase disks for 400Hz sustained</td>
<td>100%</td>
</tr>
<tr>
<td>Networking</td>
<td>100%</td>
<td>Redundancy deployed in critical areas</td>
<td>100%</td>
</tr>
</tbody>
</table>

- **ROS request rates are CPU limited**  
  - a limited update provides headroom beyond design  
  - update ROS PCs with more performant CPUs  
  - sufficient for 7 TeV LHC run (until Long Shutdown of 2013)  
- **SFO disk capacity increase for 2-day buffer at 400Hz sustained**  
  - rate capability proven with (burst) up to 1.3kHz in 2010
TDAQ Efficiency

- Run efficiency tools matured in 2010
  - Identify & keep dead-time sources under control
- Stop-less & automatic recovery procedures
  - minimise dead time, avoid full stop-start cycles
  - supported by an extensible Expert System
TDAQ & Controls Software

- New TDAQ SW Release
  - consolidate platform
  - additional scalability and fault tolerance
- Support Missing $E_T$ trigger at Level2
  - now precise $E_T^{miss}$ trigger requires a full Event Build
  - L2 will use instead ROI + $E_T$ information from front-ends
  - more rate headroom at ROSEs
  - ready for deployment
- Data compression at Tier0
  - soon in production

- DAQ Shifter Assistant
  - collect information streams from multiple sources
  - Knowledge Base rules provide alerts and suggestions to shifter
  - ease shift crew reduction

Diagram:
- Information gathering
- Information processing
- Result distribution
- Expert instructions
- Alert
- Web
- Knowledge Base
- DAQ Assistant
- TDAQ
- LHC
- Sysadmin
Conclusions

- ATLAS DAQ coped well with the 2010 operation, from low rates to beyond design
  - detector commissioning tested dataflow bandwidth & peak storage rates far beyond design specifications

- 2011 is higher luminosity and fully physics-oriented
  - now L2 and EF capacity will be increased as planned, to keep up with fast LHC progress

- The high efficiency (~95%) of 2010 has been maintained in these first months of 2011 along a luminosity growth of 6 orders of magnitude
Bonus slides
Outlook on Upgrades

- **Network**
  Upgrade of network cores proposed for 2013. Basic option maintains current architecture, improving scalability and fault tolerance. Investigating a new network topology that would unify L2 and EF networks. More flexible in balancing L2 and EF, and possibly better fitting one proposed evolution of the DataFlow software, but could be more difficult to scale.

- **MultiCore CPU**
  CPUs did not follow the clock speed increase path expected at the time of design, multi-core prevailed instead. Redesigns of certain SW components will allow better efficiency and new features, e.g. data compression at the SFOs.
HLT Farm Systems / CPUs

- **1st generation XPU**
  - 2 socket Xeon E5320 4 cores 1.8GHz “Clovertown”
  - 16GB RAM

- **2nd generation XPU**
  - 2 socket Xeon E5420 4 cores 2.5GHz “Harpertown”
  - 16 GB RAM

- **3rd generation EF**
  - 2 socket Xeon E5540 4 cores+HT 2.5GHz “Gainestown”
  - 24 GB RAM

- **Coming XPU updates and EF expansion:**
  - 2 socket Xeon X56xx 2.6GHz 4 cores+HT “Gulftown”
  - 24 GB RAM
TDAQ Phase diagrams

- TDAQ model includes constraints on EF/XPU allocation, network bandwidth, rates, rejection etc
- chosen one set of “given” parameters, e.g. EF latency, explore impact of resource allocation options
- provide input to Trigger group