# Operational performance of the ATLAS trigger and data acquisition system and its possible evolution

# $\label{eq:Andrea Negri} \begin{array}{c} \mbox{Andrea Negri}^1 \\ \mbox{on behalf of the ATLAS TDAQ collaboration}^2 \end{array}$

<sup>1</sup>University and INFN Pavia

 $^{2} https://cdsweb.cern.ch/record/1386334$ 

May 21, 2012







- Three selection levels
  - $\bullet\,$  Level 1 on custom h/w
  - High Level Triggers (Level 2 & Event Filter) on computer farms





- On Level 1 accept (latency 2.5  $\mu$ s):
  - Data pushed to buffers hosted on ReadOut System PCs (ROS)
  - Region Of Interest sent to L2





- Level 2 (latency  $\sim$ 40 ms)
  - Selection based on Region of Interest concept
  - $\bullet\,$  Only few % of event data pulled via Data Collection network





#### • Event Builder

- Pull data from Data Collection network
- Output full events to Back-End network





- Event Filter (latency  $\sim 1 ext{ s}$ )
  - Full event reconstruction
  - Accepted events sent to Data Logger farm



Andrea Negri (ATLAS TDAQ)

DF Evolution



#### • Data Logger

- Save events in streams (files)
- Files asynchronously transferred to Tier 0



Andrea Negri (ATLAS TDAQ)

Architecture

DF Evolution

- In 2011 some systems running beyond design specification
  - Event Builder
  - Data Logger



Andrea Negri (ATLAS TDAQ)

Architecture

.1 Run 2012

DF Evolutio



#### • 31 weeks of p-p operations • $\sqrt{E} = 7$ TeV

• Continuous luminosity increase

• 
$$\mathcal{L}_{peak} = 3.42 \times 10^{33} cm^{-2} s^{-1}$$
  
•  $\mathcal{L}_{int} = 4.9 \ fb^{-1}$ 

#### • Bunch cross every 50 ns instead of 25

- Higher pile-up
- Overall TDAQ efficiency  $\sim$  94%
  - 2.7 PB of data recorded (5.7M files)



• 
$$\mathcal{L}_{peak} = 5.12 \times 10^{26} cm^{-2} s^{-1}$$
  
•  $\mathcal{L}_{int} = 160 \ \mu b^{-1}$ 

# Data taking 2011





# • $\sqrt{E} = 7$ TeV

Continuous luminosity increase

• 31 weeks of p-p operations

• 
$$\mathcal{L}_{peak} = 3.42 \times 10^{33} cm^{-2} s^{-1}$$
  
•  $\mathcal{L}_{int} = 4.9 \ fb^{-1}$ 

#### Bunch cross every 50 ns instead of 25

- Higher pile-up
- Overall TDAQ efficiency  $\sim 94\%$ 
  - 2.7 PB of data recorded (5.7M files)
- 4 weeks of Pb-Pb operation

• 
$$\mathcal{L}_{peak} = 5.12 \times 10^{26} cm^{-2} s^{-2}$$
  
•  $\mathcal{L}_{int} = 160 \ \mu b^{-1}$ 

# Data taking 2011



ATLAS Online Luminosity





- HLT farm increased with LHC performance
  - 16 new racks (+50%)
- Balance issues promptly addressed
  - To hide h/w heterogeneity, EB-EF system configuration moved from a sliced system to a flat (random) mapping of EF nodes to EB ones
- L2 vs EF rack sharing configurable run-by-run





Andrea Negri (ATLAS TDAQ)

hitecture I

DF Evolutio

clusions



- Preventative maintenance
  - Replaced all Event Builder nodes
  - Rolling replacement of ROS MBs (75/153)
- Major 2011 operational issue
  - Network cards failures in replaced ROS nodes
  - Workaround: installed different network cards
- New functionalities
- E.g.: Missing  $E_T$  at L2
  - Special request for calo ROSes
  - Extracting missing  $E_T$  information directly from front and boards
  - Expensive requests for ROS
- Run control (Details in previous talk)
  - Improved automation of our DAQ monitoring and control system
  - Improved stop-less and automatic recovery procedures

- Preventative maintenance
  - Replaced all Event Builder nodes
  - Rolling replacement of ROS MBs (75/153)
- Major 2011 operational issue
  - Network cards failures in replaced ROS nodes
  - Workaround: installed different network cards
- New functionalities
- E.g.: Missing  $E_T$  at L2
  - Special request for calo ROSes
  - Extracting missing E<sub>T</sub> information directly from front end boards
  - Expensive requests for ROS



- Run control (Details in previous talk)
  - Improved automation of our DAQ monitoring and control system
  - Improved stop-less and automatic recovery procedures



- Preventative maintenance
  - Replaced all Event Builder nodes
  - Rolling replacement of ROS MBs (75/153)
- Major 2011 operational issue
  - Network cards failures in replaced ROS nodes
  - Workaround: installed different network cards
- New functionalities
- E.g.: Missing  $E_T$  at L2
  - Special request for calo ROSes
  - Extracting missing E<sub>T</sub> information directly from front end boards
  - Expensive requests for ROS

- State of the state
- Run control (Details in previous talk)
  - Improved automation of our DAQ monitoring and control system
  - Improved stop-less and automatic recovery procedures

35





#### • ROS rolling replacement continued

#### HLT farm

- 12 new racks replaced 16 old ones
- Now:  $\sim 1600 \text{ nodes}$  (Mother-Boards)
- Most racks (36) configurable as L2 or EF on run by run basis
- Back-End network upgraded
  - Installed second core router for redundancy
  - As for DC network

#### Tests

- At the peak operating conditions expected during 2012
- Predict possible bottlenecks

CPU model	Cores /node	Racks	Nodes	Usage
		11	341	
		14		
X5650			904	



• ROS rolling replacement continued

#### HLT farm

- 12 new racks replaced 16 old ones
- Now:  $\sim 1600~\text{nodes}~(\text{Mother-Boards})$
- Most racks (36) configurable as L2 or EF on run by run basis

#### • Back-End network upgraded

- Installed second core router for redundancy
- As for DC network

#### • Tests

- At the peak operating conditions expected during 2012
- Predict possible bottlenecks

CPU	Cores	Dacks	Nodos	Heare
model	/node	Nacks	Noues	Usage
E5420	8	11	341	L2/EF
E5540	8	14	448	EF
X5650	12	25	904	L2/EF

# 2011/2012 shutdown activities



• ROS rolling replacement continued

#### HLT farm

- 12 new racks replaced 16 old ones
- Now:  $\sim 1600 \text{ nodes}$  (Mother-Boards)
- Most racks (36) configurable as L2 or EF on run by run basis
- Back-End network upgraded
  - Installed second core router for redundancy
  - As for DC network

#### • Tests

- At the peak operating conditions expected during 2012
- Predict possible bottlenecks

CPU	Cores	Dacks	Nodos	Heara
model	/node	Nacks	Noues	Usage
E5420	8	11	341	L2/EF
E5540	8	14	448	EF
X5650	12	25	904	L2/EF



# 2011/2012 shutdown activities



• ROS rolling replacement continued

#### HLT farm

- 12 new racks replaced 16 old ones
- Now:  $\sim 1600 \text{ nodes}$  (Mother-Boards)
- Most racks (36) configurable as L2 or EF on run by run basis
- Back-End network upgraded
  - Installed second core router for redundancy
  - As for DC network
- Tests
  - At the peak operating conditions expected during 2012
  - Predict possible bottlenecks

CPU	Cores	Dacks	Nodos	Heara
model	/node	Nacks	Noues	Usage
E5420	8	11	341	L2/EF
E5540	8	14	448	EF
X5650	12	25	904	L2/EF



# Data taking 2012



- $\sqrt{E} = 8 \text{ TeV}$
- Impressive LHC start-up
  - 80 % of the expected peak luminosity in few weeks

	Done	Max
$\mathcal{L} \ [ imes 10^{33} cm^{-2} s^{-1}]$	5.55	6.68
$\beta^{\star}[m]$	0.6	0.6
Bunches	1082	1331
$p/bunch [ imes 10^{11}]$	1.2	1.65
$<\mu>$	29.8	35

- Overall TDAQ efficiency 93.6%
  - Comparable to last year
- Bunch crossing still 50 ns
  - Pile-up a major concern



# Pile-up 2012: CPU usage



- Processing time linear scaling verified up to  $\langle \mu \rangle {\sim}$  22
- Extrapolating to  $\langle \mu \rangle = 35$ 
  - 25% CPU margin shared across L2&EF
- Extrapolation uncertainties: trigger menu, ROS collection time
  - CPU usage evolution is being surveyed



# Pile-up 2012: CPU usage



- $\bullet$  Processing time linear scaling verified up to  $\langle \mu \rangle {\sim}~22$
- Extrapolating to  $\langle \mu \rangle {=}~35$ 
  - 25% CPU margin shared across L2&EF
- Extrapolation uncertainties: trigger menu, ROS collection time
  - CPU usage evolution is being surveyed





- Pile-up dependency for some detectors (E.g.: Inner)
- Evolution largely linear
  - Future deviations cannot be excluded
- Extrapolation for  $\langle \mu \rangle$  up to 35
  - ullet Event size up to  $\sim$  1.8 MB
- We may face limited operational margins at peak luminosity
- Additional EB capacity to be deployed to meet peak demand
- Data Logger capacity to be increased
  - additional h/w or
  - increase b/w into existing h/w





- Pile-up dependency for some detectors (E.g.: Inner)
- Evolution largely linear
  - Future deviations cannot be excluded
- Extrapolation for  $\langle \mu \rangle$  up to 35
  - ullet Event size up to  $\sim$  1.8 MB
- We may face limited operational margins at peak luminosity
- Additional EB capacity to be deployed to meet peak demand
- Data Logger capacity to be increased
  - additional h/w or
  - increase b/w into existing h/w



11 Run 2012



- Pile-up dependency for some detectors (E.g.: Inner)
- Evolution largely linear
  - Future deviations cannot be excluded
- Extrapolation for  $\langle \mu \rangle$  up to 35
  - Event size up to  $\sim 1.8~{
    m MB}$
- We may face limited operational margins at peak luminosity
- Additional EB capacity to be deployed to meet peak demand
- Data Logger capacity to be increased
  - additional h/w or
  - increase b/w into existing h/w



Run 2012



- Pile-up dependency for some detectors (E.g.: Inner)
- Evolution largely linear
  - Future deviations cannot he excluded
- Extrapolation for  $\langle \mu \rangle$  up to 35
  - Event size up to  $\sim 1.8~{
    m MB}$
- We may face limited operational margins at peak luminosity
- Additional EB capacity to be deployed to meet peak demand
- Data Logger capacity to be increased
  - additional h/w or
  - increase b/w into existing h/w





- ROS performance can be limited by:
  - Access rate
  - Bandwidth
  - Load (not a problem for new h/w)
- ROS parameters are being surveyed
  - Motherboards of Transition Radiation Tracker ROSes recently replaced





- ROS performance can be limited by:
  - Access rate
  - Bandwidth
  - Load (not a problem for new h/w)
- ROS parameters are being surveyed
  - Motherboards of Transition Radiation Tracker ROSes recently replaced





- First occasion for major hardware and software upgrades
- Define a s/w scalable model to be used in 2014 and beyond
- Profit from experience from past and ongoing data-taking
  - Build-in further scalability and flexibility
- Current assumptions for 2014
  - 100 kHz L1 rate
  - $\bullet~1~\rm kHz$  average physics output rate
    - Extension of the Data Logger capacity
    - Provide online data compression for a more efficient use of resources
  - 25 ns bunch crossing
    - But be prepared for 50 ns:
      - learn as much as possible this year on high pile up operation



- First occasion for major hardware and software upgrades
- Define a s/w scalable model to be used in 2014 and beyond
- Profit from experience from past and ongoing data-taking
  - Build-in further scalability and flexibility
- Current assumptions for 2014
  - 100 kHz L1 rate
  - 1 kHz average physics output rate
    - Extension of the Data Logger capacity
    - Provide online data compression for a more efficient use of resources
  - 25 ns bunch crossing
    - But be prepared for 50 ns:
      - learn as much as possible this year on high pile up operation

### Current Architecture

- Data taking confirmed the success of the current design
- ... and stimulated interest to explore possible evolutions
  - Simplify CPU and network resources balancing
  - Reduce complexities
  - Simplify HLT steering



Andrea Negri (ATLAS TDAQ)





#### • Merge L2, EB, EF within a single homogeneous system

- A single farm
- In each node:
  - Rol based processing  $\rightarrow$  event building  $\rightarrow$  full event processing
- Possibility to have a single network



Andrea Negri (ATLAS TDAQ)



- Merge L2, EB, EF within a single homogeneous system
  - A single farm
  - In each node:
    - Rol based processing  $\rightarrow$  event building  $\rightarrow$  full event processing
  - Possibility to have a single network



#### Andrea Negri (ATLAS TDAQ)

cture Run 2011

#### 011 Run 2012

DF Evolution

Conclusio

#### 14 / 19

#### Data Flow Evolution

- A single HLT homogeneous farm
- On each HLT node
  - One Data Collection Manager (DCM) in charge of data collection, caching and integrity
  - Multiple Processing Units (HLTPUs) in charge of event selection
  - Communication via shared memories
- A single SuperVisor (HLTSV) distributes L1 results to HLT nodes
  - Must sustain 100 kHz (otherwise multiple HLTSVs)
  - Possibility to merge HLTSV with a s/w based RoIB under evaluation
  - Data Loggers receive events from DCMs and store them to disk
  - ROS application unchanged







#### Andrea Negri (ATLAS TDAQ)

### Data Flow Evolution

- A single HLT homogeneous farm
- On each HLT node
  - One Data Collection Manager (DCM) in charge of data collection, caching and integrity
  - Multiple Processing Units (HLTPUs) in charge of event selection
  - Communication via shared memories
- A single SuperVisor (HLTSV) distributes L1 results to HLT nodes
  - Must sustain 100 kHz (otherwise multiple HLTSVs)
  - Possibility to merge HLTSV with a s/w based RoIB under evaluation
  - Data Loggers receive events from DCMs and store them to disk
  - ROS application unchanged



14 / 19





#### Andrea Negri (ATLAS TDAQ)

#### Data Flow Evolution

- A single HLT homogeneous farm
- On each HLT node
  - One Data Collection Manager (DCM) in charge of data collection, caching and integrity
  - Multiple Processing Units (HLTPUs) in charge of event selection
  - Communication via shared memories
- A single SuperVisor (HLTSV) distributes L1 results to HLT nodes
  - Must sustain 100 kHz (otherwise multiple HLTSVs)
  - Possibility to merge HLTSV with a s/w based RoIB under evaluation
  - Data Loggers receive events from DCMs and store them to disk
  - ROS application unchanged



14 / 19



- Simpler Data Flow configuration
  - Only 5 application types (were 9)
- Automatic CPU balance on each HLT node
- Automatic HLT system balance
  - $\bullet\,$  No need to pre-determine the L2/EF sharing
- No additional contributions to fragment lifetime inside Read Out Buffers
  - ROS cleared after RoI based processing or EB
- Reduced ROS load
  - All event fragments only requested once from a ROS
  - Less network connections (one per HLT node)
- HLT selection still based on Rol
- A single HLT steering instance





- Simpler Data Flow configuration
  - Only 5 application types (were 9)
- Automatic CPU balance on each HLT node
- Automatic HLT system balance
  - $\bullet\,$  No need to pre-determine the L2/EF sharing
- No additional contributions to fragment lifetime inside Read Out Buffers
  - ROS cleared after RoI based processing or EB
- Reduced ROS load
  - All event fragments only requested once from a ROS
  - Less network connections (one per HLT node)
- HLT selection still based on Rol
- A single HLT steering instance





- Simpler Data Flow configuration
  - Only 5 application types (were 9)
- Automatic CPU balance on each HLT node
- Automatic HLT system balance
  - $\bullet\,$  No need to pre-determine the L2/EF sharing
- No additional contributions to fragment lifetime inside Read Out Buffers
  - ROS cleared after RoI based processing or EB
- Reduced ROS load
  - All event fragments only requested once from a ROS
  - Less network connections (one per HLT node)
- HLT selection still based on Rol
- A single HLT steering instance







- Simpler Data Flow configuration
  - Only 5 application types (were 9)
- Automatic CPU balance on each HLT node
- Automatic HLT system balance
  - $\bullet\,$  No need to pre-determine the L2/EF sharing
- No additional contributions to fragment lifetime inside Read Out Buffers
  - ROS cleared after RoI based processing or EB
- Reduced ROS load
  - All event fragments only requested once from a ROS
  - Less network connections (one per HLT node)
- HLT selection still based on Rol
- A single HLT steering instance









#### • No need to create and transport L2 Result

- ROS access and data unpacking done only once
- Flexibility for HLT strategies and to exploit DF resources
- Different strategies under evaluation (depending on the needs)
  - Minimize L2 latency (giving time to more complex algorithms)
    - Change the chains/steps execution model and re-order the chains
  - Minimize ROS access rate, by optimizing EB request
    - Choose the best time of EB moving algorithms between L2 & EF





- No need to create and transport L2 Result
- ROS access and data unpacking done only once
- Flexibility for HLT strategies and to exploit DF resources
- Different strategies under evaluation (depending on the needs)
  - Minimize L2 latency (giving time to more complex algorithms)
    - Change the chains/steps execution model and re-order the chains
  - Minimize ROS access rate, by optimizing EB request
    - $\bullet\,$  Choose the best time of EB moving algorithms between L2 & EF





- No need to create and transport L2 Result
- ROS access and data unpacking done only once
- Flexibility for HLT strategies and to exploit DF resources
- Different strategies under evaluation (depending on the needs)
  - Minimize L2 latency (giving time to more complex algorithms)
    - Change the chains/steps execution model and re-order the chains
  - Minimize ROS access rate, by optimizing EB request
    - Choose the best time of EB moving algorithms between L2 & EF





- No need to create and transport L2 Result
- ROS access and data unpacking done only once
- Flexibility for HLT strategies and to exploit DF resources
- Different strategies under evaluation (depending on the needs)
  - Minimize L2 latency (giving time to more complex algorithms)
    - Change the chains/steps execution model and re-order the chains
  - Minimize ROS access rate, by optimizing EB request
    - Choose the best time of EB moving algorithms between L2 & EF



#### Design phase ongoing

- First implementation to be ready for the end of the run
- Looking for common solutions, minimizing code duplication
  - A common framework for all the applications
- Different s/w technologies under evaluation
  - Profit from experience
  - But with an open attitude toward new ideas and views
- Prototype available for testing design and spot problems
  - Current applications adapted to the proposed design
  - Developed 2 years ago and integrated in the current release
  - Tested on ATLAS TDAQ system



#### Design phase ongoing

- First implementation to be ready for the end of the run
- Looking for common solutions, minimizing code duplication
  - A common framework for all the applications
- Different s/w technologies under evaluation
  - Profit from experience
  - But with an open attitude toward new ideas and views
- Prototype available for testing design and spot problems
  - Current applications adapted to the proposed design
  - Developed 2 years ago and integrated in the current release
  - Tested on ATLAS TDAQ system



- Design phase ongoing
  - First implementation to be ready for the end of the run
- Looking for common solutions, minimizing code duplication
  - A common framework for all the applications
- Different s/w technologies under evaluation
  - Profit from experience
  - But with an open attitude toward new ideas and views
- Prototype available for testing design and spot problems
  - Current applications adapted to the proposed design
  - Developed 2 years ago and integrated in the current release
  - Tested on ATLAS TDAQ system

#### Data Flow Evolution: measurements @ P1



#### • Scalability validated up to $\sim$ 1200 HLT nodes ( $\sim$ 13k HLTPUs)

- Traffic shaping strategy allows to prevent network congestions
  - In each DCM, limit the number of concurrent requests
  - A similar algorithm is being used in EB nodes of the current system
- A single HLTSV able to sustain more than 100 kHz
  - Overhead of s/w RolB to be evaluated



### Data Flow Evolution: measurements @ P1



- Scalability validated up to  $\sim$  1200 HLT nodes ( $\sim$  13k HLTPUs)
- Traffic shaping strategy allows to prevent network congestions
  - In each DCM, limit the number of concurrent requests
  - A similar algorithm is being used in EB nodes of the current system
- A single HLTSV able to sustain more than 100 kHz
  - Overhead of s/w RolB to be evaluated



### Data Flow Evolution: measurements @ P1



- Scalability validated up to  $\sim$  1200 HLT nodes ( $\sim$  13k HLTPUs)
- Traffic shaping strategy allows to prevent network congestions
  - In each DCM, limit the number of concurrent requests
  - A similar algorithm is being used in EB nodes of the current system
- A single HLTSV able to sustain more than 100 kHz
  - $\bullet\,$  Overhead of s/w RoIB to be evaluated





#### • Data taking 2011

- $\bullet\,$  Smooth TDAQ operation:  $\sim$  94% run efficiency
- Extended HLT farm in course of operations
- Stable and reliable data collection system
- Excellent operational stability of control, configuration and monitoring
- Improved automation: monitoring and recovery procedures
- Data taking 2012
  - Overall smooth and quick start up
  - High pileup effects under control
- Data Flow evolution
  - Merge L2, EB, EF within a single homogeneous system
  - Prototype studies did not spot problems
  - Design phase ongoing
  - To be ready at the beginning of 2013

### Spare: Evolution prototype: Traffic Shaping



- Traffic shaping strategy allows to prevent network congestions
  - In each DCM, limit the number of concurrent requests
  - A similar algorithm is being used in EB nodes of the current system



### Spare: Evolution prototype: load balance



- Automatic load balance inside each node
  - System promptly reacts to operation condition changes
  - System always capable of sharing CPU resources between the L2 and EF algorithms



### Spare: Evolution prototype: fixed L1 rate



- Test in realistic operational conditions: fixed L1 rate
  - As long the CPUs are not saturated the throughput rate is stable with increasing L2 processing time
  - After saturation performance decreases as expected



#### Spare: Evolution prototype: comparison



- Comparison between old and new architecture
  - Same setup: 23 XPU racks to be shared between L2 and EF

