The ATLAS Data Flow system for the Second LHC Run

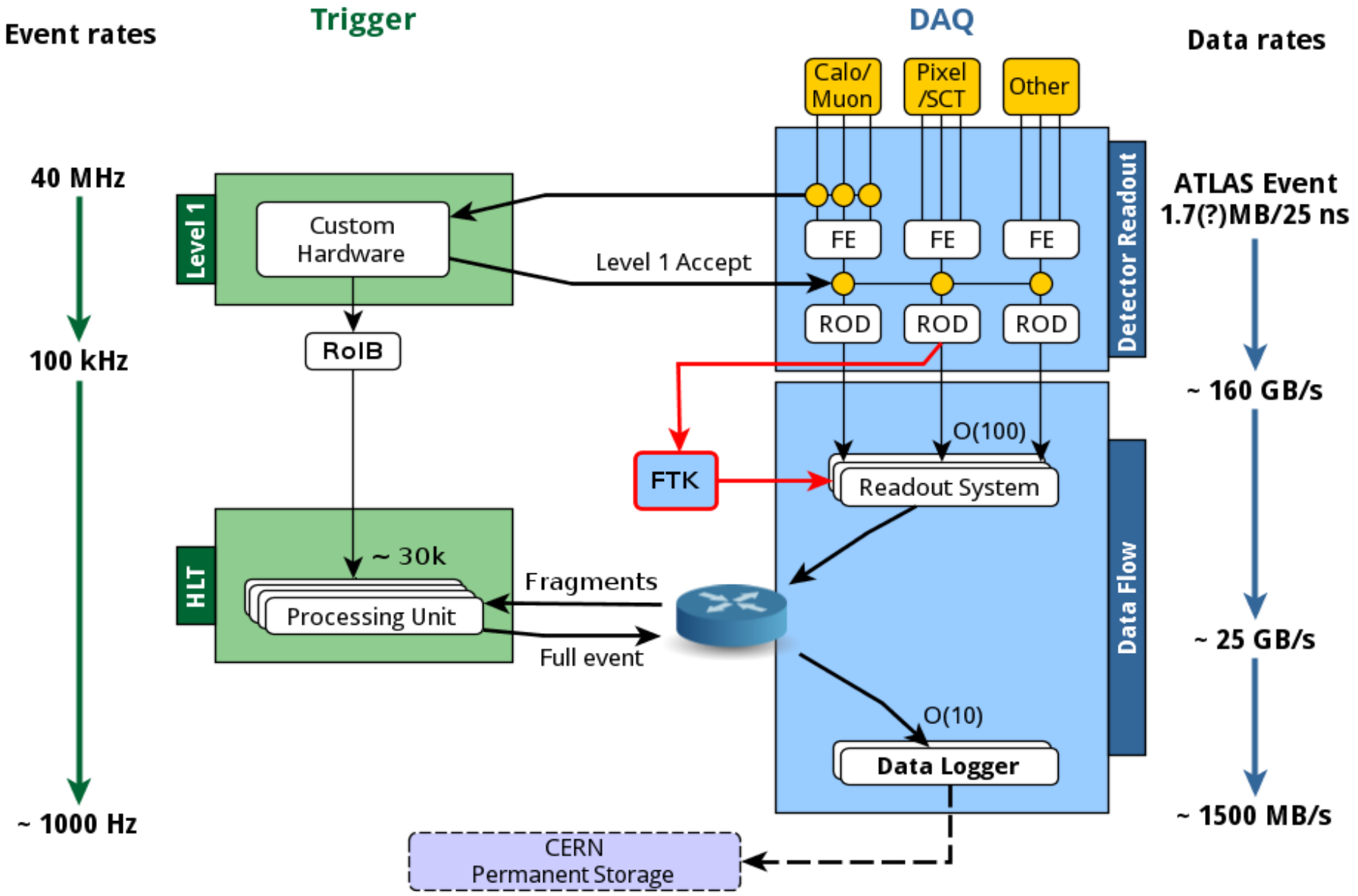
Reiner Hauser
Michigan State University

On behalf of the ATLAS Collaboration

Outline

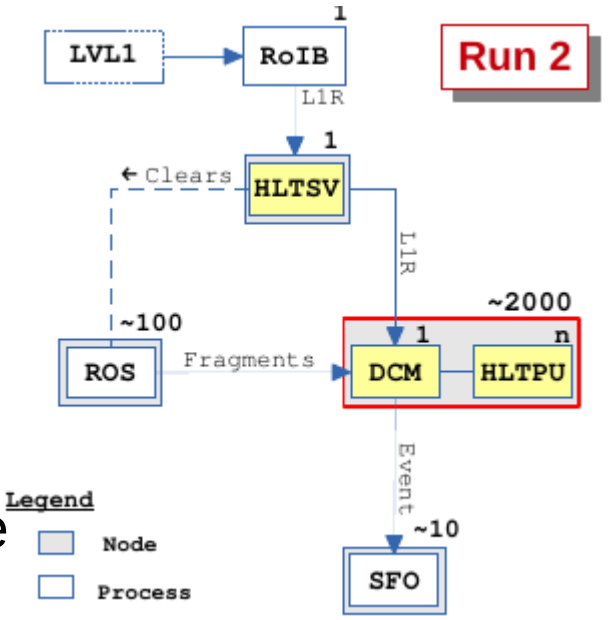
- Overview of the ATLAS Trigger and DAQ System
- Network
- Readout System
- Region of Interest Builder and Planned Changes
- HLT Supervisor
- Data Collection Manager
- HLT Processing Unit
- Data Logger

The ATLAS Trigger and DAQ System



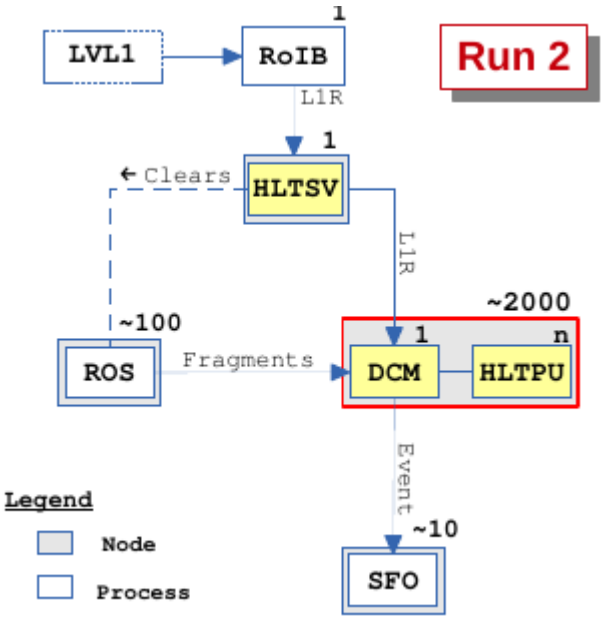
Components of the ATLAS Data Flow

- For Run 2 the former Level 2 and Event Filter farms have been merged into a single HLT farm.
  - A big simplification in terms of applications and network.
- The Region of Interest (RoI) concept has been kept, and processing and data collection proceeds in stages, beginning with fast algorithms based on RoIs. The decision when to build the event is flexible, and afterwards more off-line style algorithms have access to the full event.
  - The **Readout System (ROS)** buffers front-end data from the detectors and provides a standard interface to the DAQ.
  - The **Region of Interest Builder (RoIB)** receives L1 trigger information and RoIs and combines the information for the HLT Supervisor.
  - The **HLT supervisor (HLTSV)** schedules events to the HLT farm and handles eventual time-outs.



Components of the ATLAS Data Flow (2)

- Components (cont'd)
  - The **Data Collection Manager (DCM)** handles all I/O on the HLT nodes, including RoI requests from the HLT and full event building.
  - The **HLT processing tasks** are forked from a single mother process to maximize memory sharing, and run the ATLAS Athena/Gaudi framework in a special online mode. All I/O is done via the DCM.
  - The **data loggers (SFO)** are responsible for saving accepted events to disk, and send the files to EOS.



Network

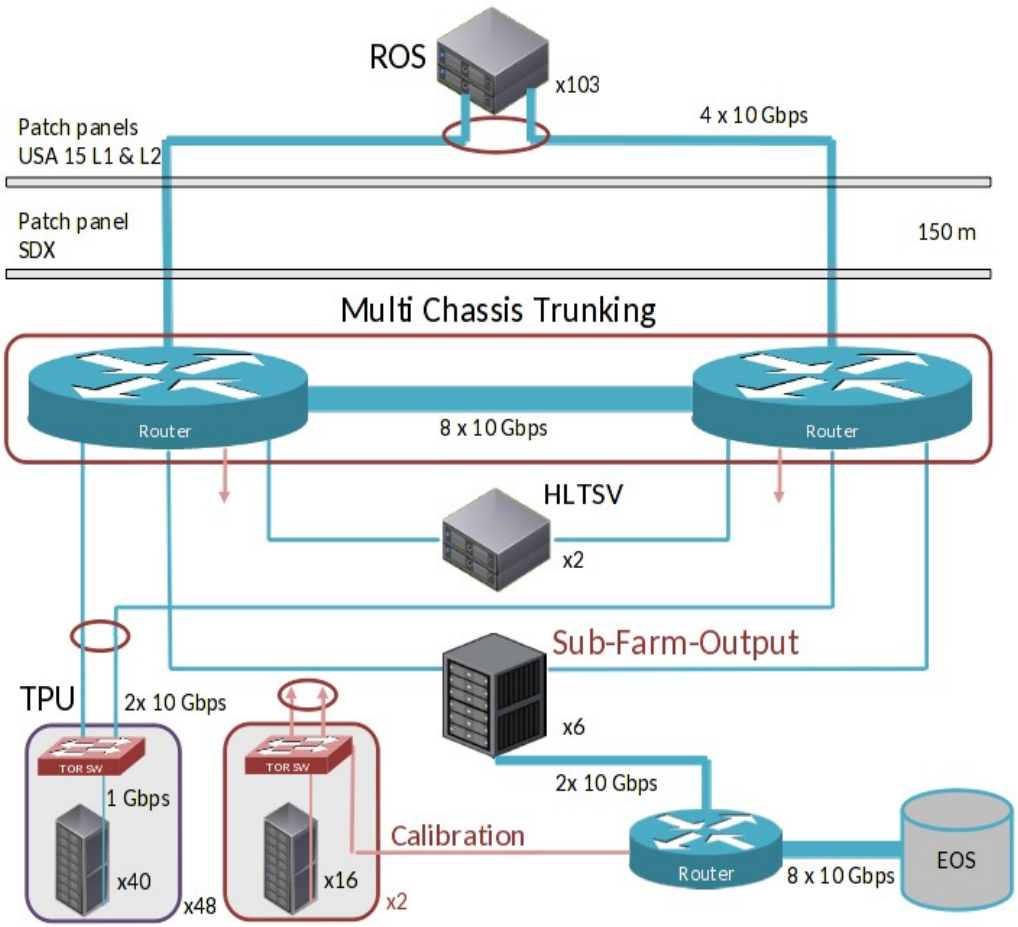
The data flow network has seen a significant upgrade and simplification.

Multi chassis trunking of the core routers provides load balancing and link redundancy to the network.

Virtual output queue mechanism avoids head of line blocking and allows the routers to run in non-blocking mode.

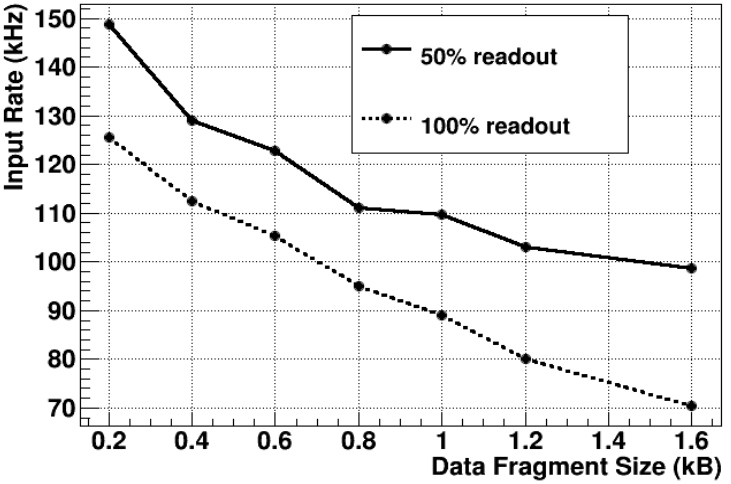
Rack concentrator switches have been identified and are being purchased.

In addition to the data flow network the control network has been made more redundant with active backup solutions for all important components.



Readout System (ROS)

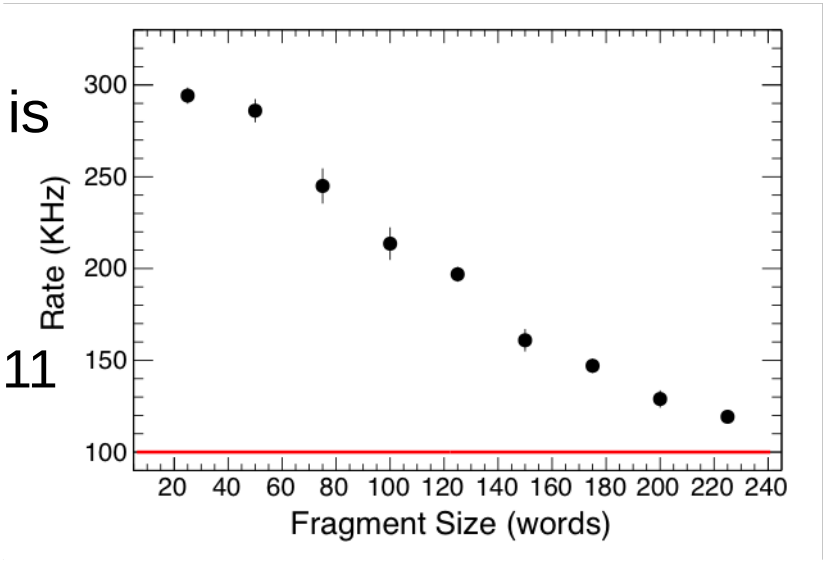
- The S-Link input and buffer hardware (Robin) has been upgraded to a new board (RobinNP) based on the Alice C-RORC card with ATLAS specific firmware.
  - Switch from PCI-X to PCI Express.
  - Higher density of optical link connectors:
    - 12 per card, 2 cards per PC.
  - Larger memory buffer.
  - New set of ROS PCs:
    - 2U form factor instead of 4U.
  - 4 x 10 Gbit/s Ethernet per ROS PC.
    - (was 2x1 Gbit/s)



- A fully connected ROS (24 links) sustains the required RoIB request rate plus ~50 kHz of event building rate (lab measurement).
  - A ROS with fewer input links and/or small enough fragments can run at 100 kHz.

Region of Interest Builder (RoIB)

- The RoIB is a 9U VME based custom hardware solution, consisting of multiple cards.
  - The original hardware from Run 1 is the current baseline.
  - Larger input fragments for the Run 2 upgrade show that the hardware is close to its limits.
- A replacement based on the RobinNP board is being developed.
  - Common hardware between ROS and RoIB.
  - A single PCI Express board will suffice for all 11 inputs.
  - The board will directly integrate into the HLT supervisor.
  - The combining of the input fragments will be done in software.
  - First test results in the lab are very promising, far exceeding the requirements.



HLT Supervisor (HLTSV)

- A single HLT supervisor replaces the set of L2 supervisors used in Run 1.
- It uses a heavily multi-threaded design using the Boost ASIO library for communication and Intel Thread Building Blocks for concurrent data structures.
- 2 x 10 Gbit/s Ethernet to the data flow network.
- A single application can handle the input from the RoIB and manage the HLT farm of ~1500 machines at ~115 kHz under realistic ATLAS conditions.
  - A new, faster machine will be deployed before collisions start.
- In the future the HLTSV will be integrated with the new RoI builder.
  - Deployment of this solution probably only during next winter shutdown.

Data Collection Manager (DCM)

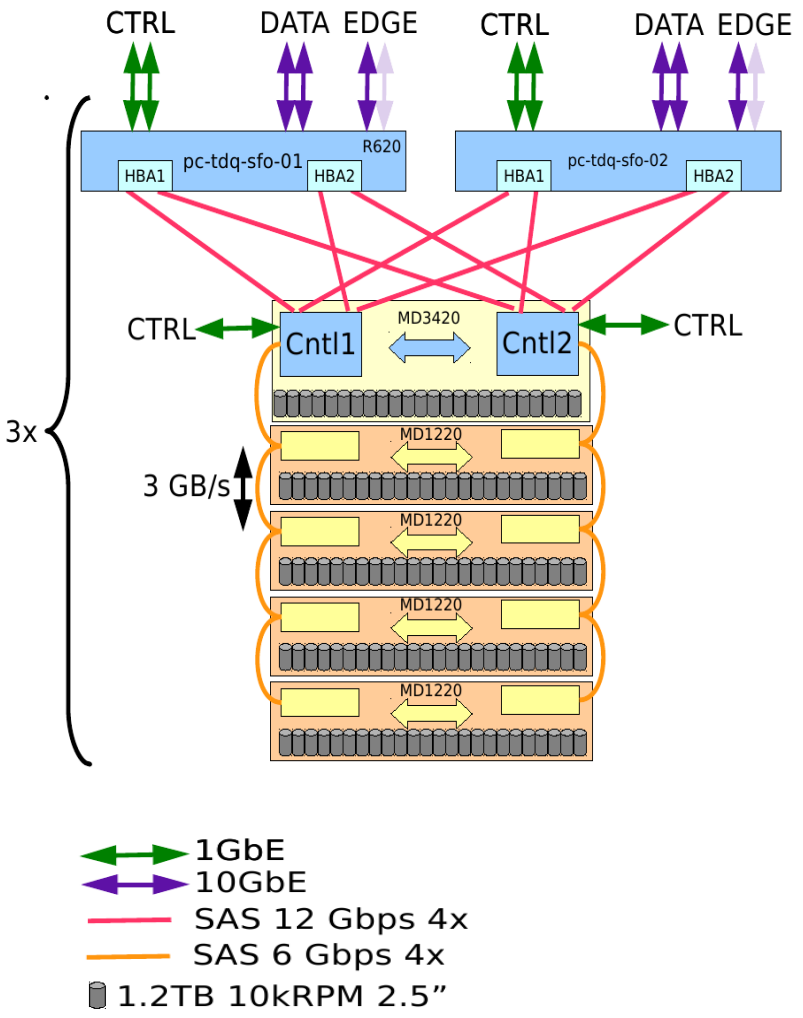
- The DCM is a single application per HLT node that deals with all data requests from the multiple HLT processing talks.
  - It handles both Region of Interest request and full event building.
- It communicates to the HLT tasks via sockets and shared memory.
- Its design is essentially single-threaded based on non-blocking I/O using the Boost ASIO library.
  - A credit based traffic shaping mechanism is used to prevent overloading the incoming network link.
- For accepted events the DCM also handles the preparation for an event going to multiple output streams (e.g. for calibration purposes).
- Finally it compresses the event payload before sending it to the data logger.

HLT Processing Unit (HLTMPPU)

- The HLT processing unit encapsulates the Athena framework that is running the actual HLT algorithms.
- It communicates with the DCM for I/O requests and provides the trigger decision for each event.
- It takes care of publishing monitoring information like histograms into the online system for data quality purposes.
- On each node a mother process is started first and goes through all the configuration process. A set of child processes is forked when the run starts, maximizing the memory sharing.
  - Crashed HLT applications can be quickly replaced by forking another child instance.
  - Tests with the full 2012 trigger menu show a memory consumption of ~ 1.8 GByte + N x 700 MByte (N = 8)

Data Logger

- In Run 1 a data logger was a PC with 3 internal Raid5 raid arrays of 8 disks each.
- For Run 2 a direct attached storage unit is used, with multiple front-ends and redundant data paths for fault tolerance and resilience.
- First measurements on test system in lab show more than adequate performance.
  - Final numbers will depend on trigger menu, number of streams, etc.
- Background jobs copy the files to permanent storage, deleting them on the local disk only when they are safely on tape.



120 disks per node
340 TB total effectively for 3 systems.

Summary

- The ATLAS data flow has seen a considerable simplification compared to Run 1.
- Every component has been either upgraded hardware wise, or rewritten to take advantage of modern designs.
- The new data flow system was in place beginning of 2014 and has been used for all integration and cosmic runs taken by ATLAS since then.
  - In addition the TDAQ software has been used for the commissioning of sub-detectors during the long shutdown.
- Final touches are being applied, but we see no major bottlenecks at this point and have a lot of headroom in several areas.
- ATLAS TDAQ is ready for Run 2.