

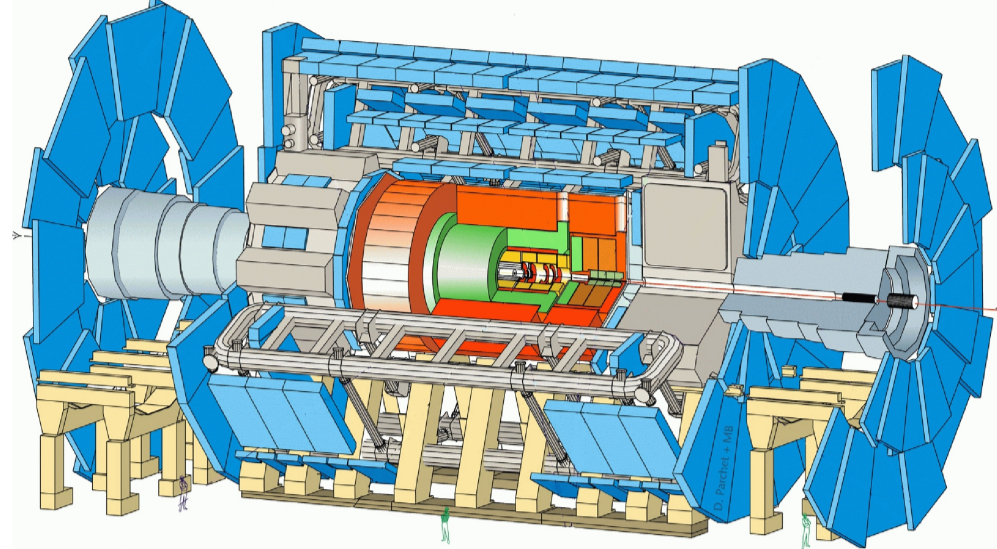
# ATLAS DataFlow Infrastructure: recent results from ATLAS cosmic and first-beam data-taking



Wainer Vandelli\* (wainer.vandelli@cern.ch), CERN, Geneva, Switzerland  
on behalf of the ATLAS Collaboration

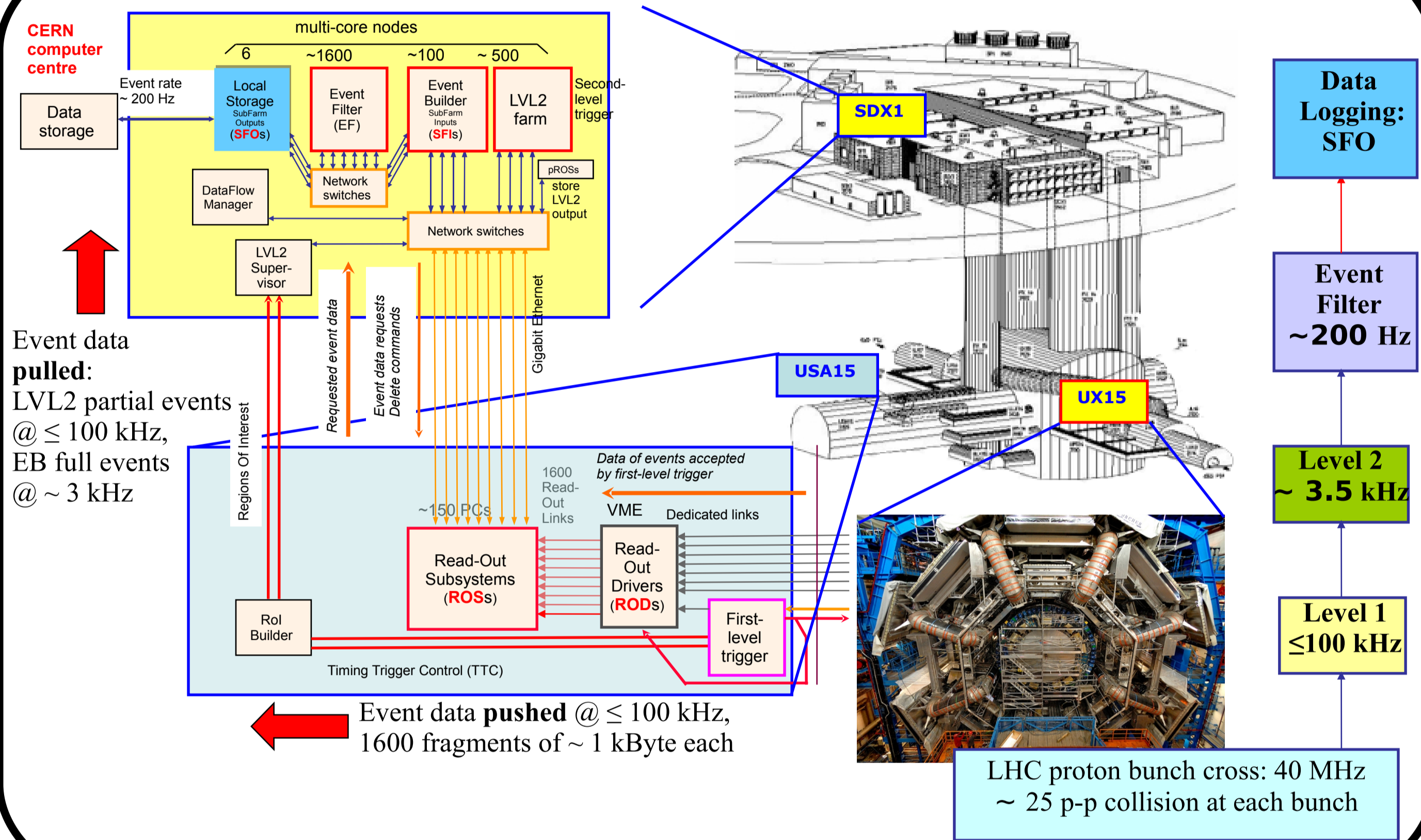
## ATLAS Trigger and Data Acquisition (TDAQ)

ATLAS is one of the four experiments installed at the LHC, CERN, Geneva, Switzerland. The ATLAS Trigger and Data Acquisition (TDAQ) system is responsible for the selection and the conveyance of interesting physics data, reducing the initial LHC frequency of 40 MHz to a rate of stored events of ~200 Hz. In its final configuration, the TDAQ system will include O(20k) applications running on roughly 2000 nodes interconnected by a multi-stage Gigabit Ethernet network.



The ATLAS TDAQ is organized in a three-level selection scheme, including a hardware-based first-level trigger and software-based second and third level triggers. In particular, the second-level trigger operates over limited regions of the detector, the so-called Region-of-Interest (RoI). The last selection step, the Event Filter, instead deals with completely built events. The TDAQ system is based on in-house designed multi-threaded software, mostly written in C++ and Java and running on a Linux operating system.

## Architecture



## ATLAS DataFlow Infrastructure

The ATLAS DataFlow infrastructure is responsible for the collection and the conveyance of event data from the detector front-end electronics to the mass storage, while serving the trigger processors. This purpose is fulfilled by several dedicated applications, which can be classified based on the two different networking domains that form the TDAQ data networking infrastructure.

In the first domain, the so-called "Data-Collection" network, the system is based on a push-pull architecture. The Read-Out System (ROS) in fact buffers and serves over the network the data fragments received via ~1600 fibers from the front-end. The ROS is composed by roughly 150 PCs housing custom PCI boards. The ROS clients are the Event Builder (EB) and the second level trigger (LVL2). In the final running conditions, the LVL2 farm, which will include up to 500 nodes, will analyse partial event data fetched from the ROS PCs, reducing the initial trigger rate of 75 (100) kHz to ~3.5 kHz. The ROS PCs will experience a LVL2 request rate as big as 12 kHz and an aggregated throughput of O(3 GB/s). Upon the LVL2 acceptance, the events are fully built by the Event Builder. Given the predicted average event size of 1.5 MB, the EB, which includes ~100 building applications, must sustain a total throughput of ~5 GB/s.

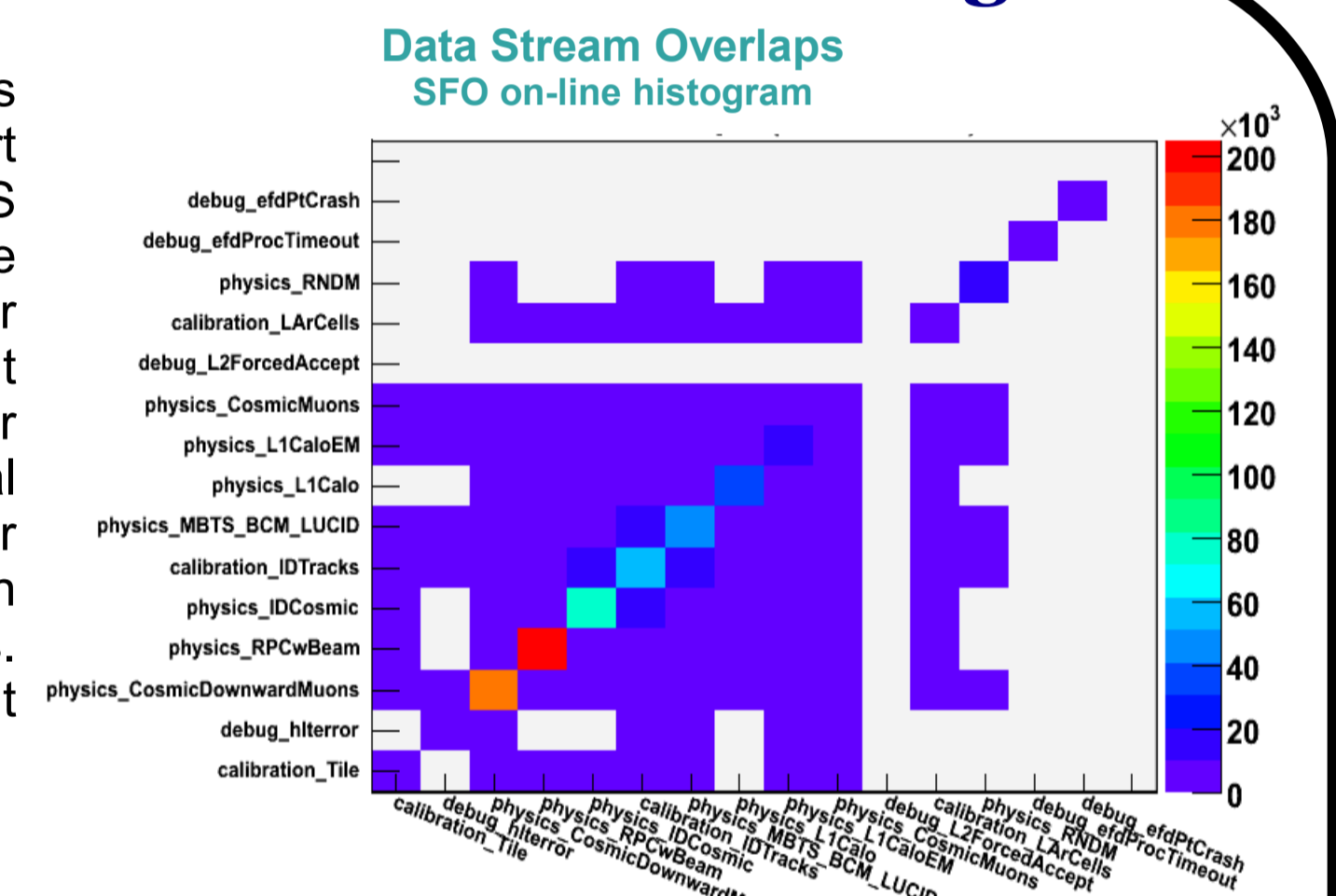
The EB moreover decouples the Data-Collection network from the Event Filter network, the second networking domain of the DataFlow infrastructure. The building applications in fact buffer the fully built events and serve them to the third triggering level, the Event Filter (EF). On each EF node a DataFlow application is responsible for the data handling and for the distribution to the selection processes. Since the EF farm will include up to 1600 computing nodes, the connection between the EB and EF is organized in sub-farms, containing subsets of the building applications and processing racks. Such a configuration allows for flexibility and redundancy in the usage of the available resources. The last elements of ATLAS TDAQ system are the data-logging nodes, called Sub-Farm Outputs (SFOs), where the accepted events are temporary stored on local disks while waiting for the transmission to the mass storage. The SFO farm must be able to handle a I/O rate of 300 MB/s and to buffer up to 24 hours of data-taking results, in order to decouple the on-line system from the off-line mass storage.

	Installed nodes (March 2009)	Final farm size
Read-Out System	149	149
LVL2	850*	500
Event Builder	63**	63
Event Filter	850*	1600
SFO	5	6

\* Shared between LVL2 and EF  
\*\* Running 94 building applications

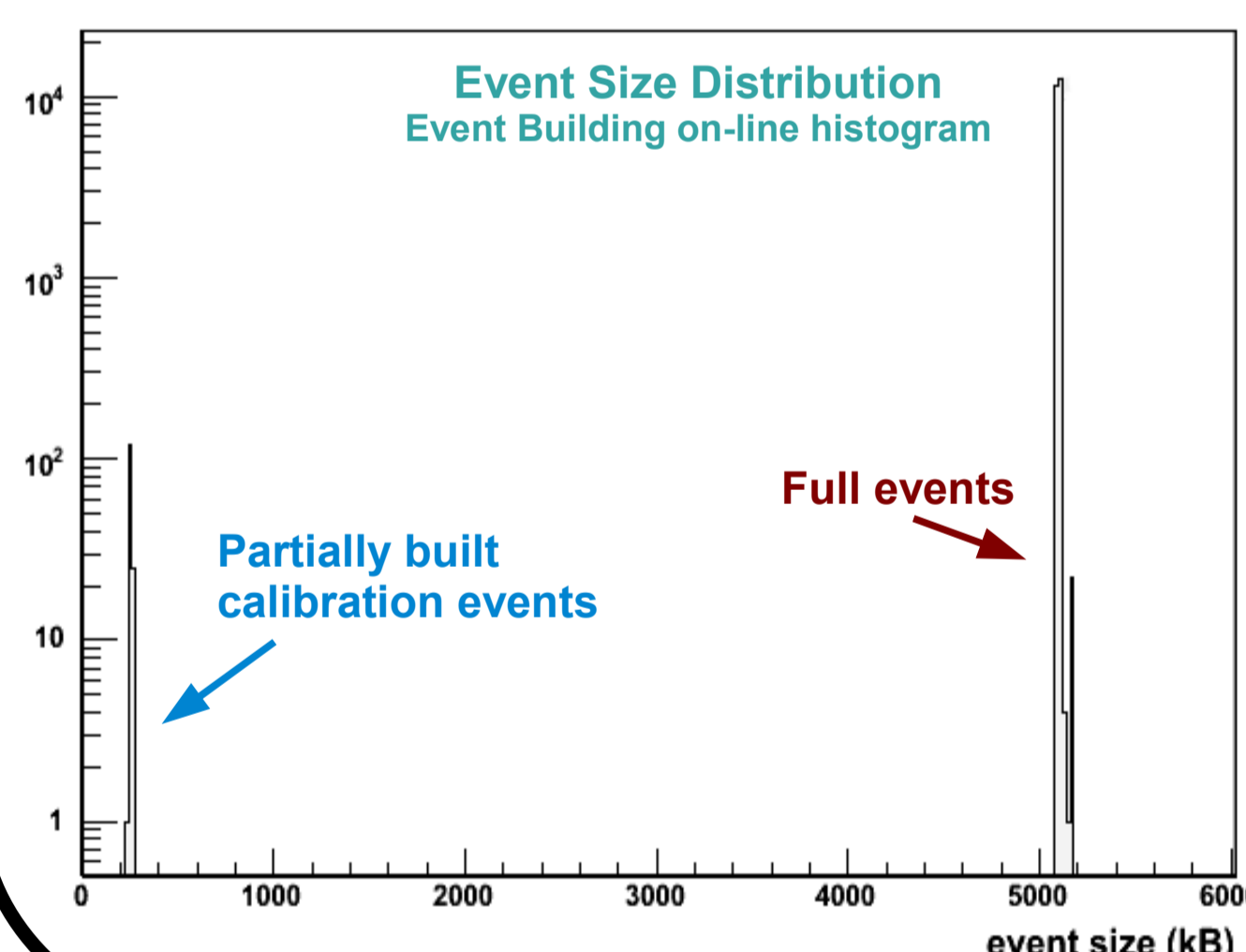
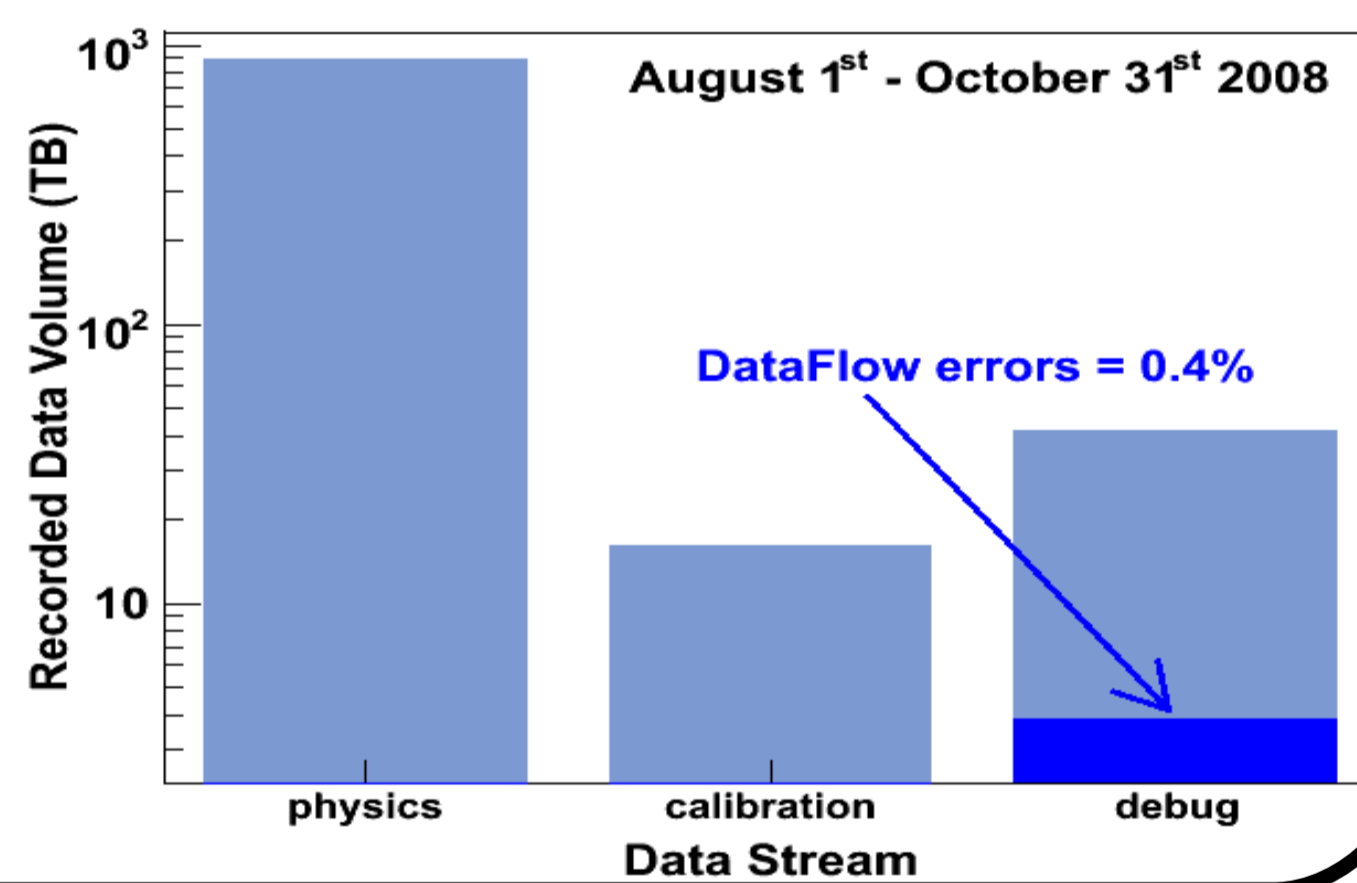
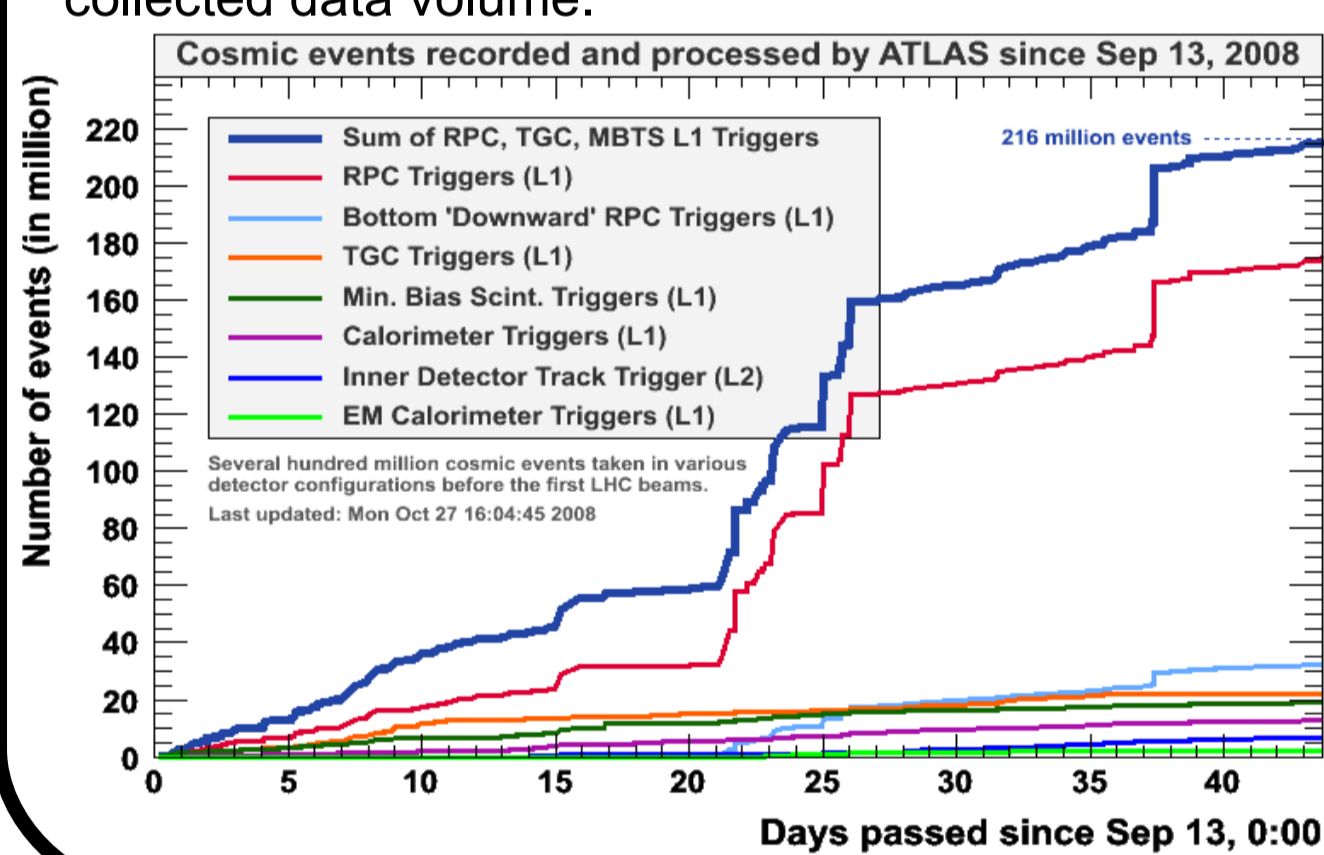
## Routing, Streaming and Partial Event Building

The ATLAS DataFlow infrastructure also provides routing and streaming capabilities as well as support for optimized handling of calibration data. In the ATLAS TDAQ framework, streaming is defined as the on-line classification of raw events based either on their physics content or processing results. The event classification is normally performed by the trigger software, while the SFOs are responsible for the actual streaming of the data into different data files. The other DataFlow applications are anyway stream-aware in order to implement a correct routing of the events. Routing in fact enables the optimization of the event paths in the on-line system, allowing resource savings.



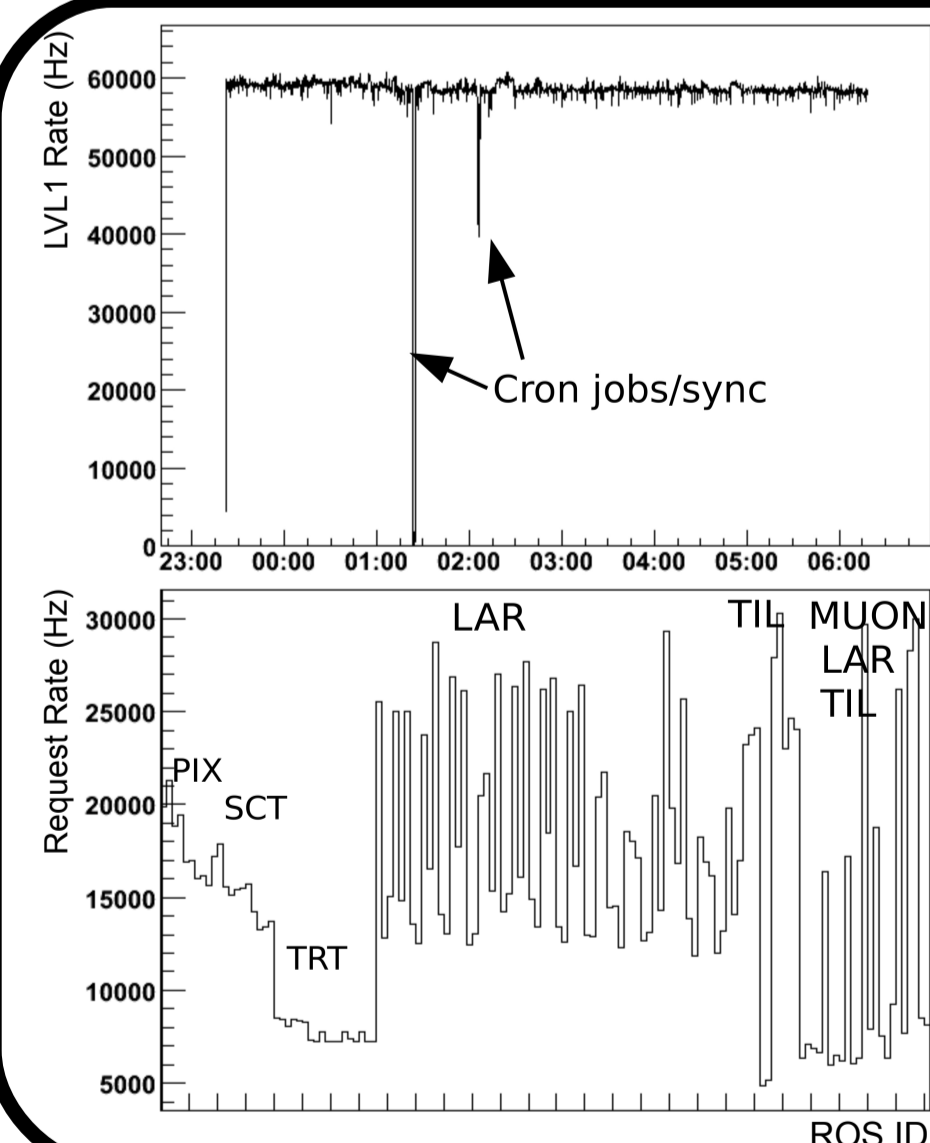
## Cosmic data-taking

In preparation for the first LHC beam in September 2008, the ATLAS experiment went in continuous cosmic data-taking data on August 1<sup>st</sup>. This first extensive data-taking period provided an invaluable feedback to the ATLAS DataFlow in terms of functionality, stability and efficiency. Over this period, ~550 millions of events, corresponding to roughly 1 PB of recorded data, have been collected and handed over the off-line facilities by the ATLAS DataFlow infrastructure. DataFlow errors, like application crashes and communication timeouts, only affected 0.4% (of which 63% is due to a single major accident) of the collected data volume.



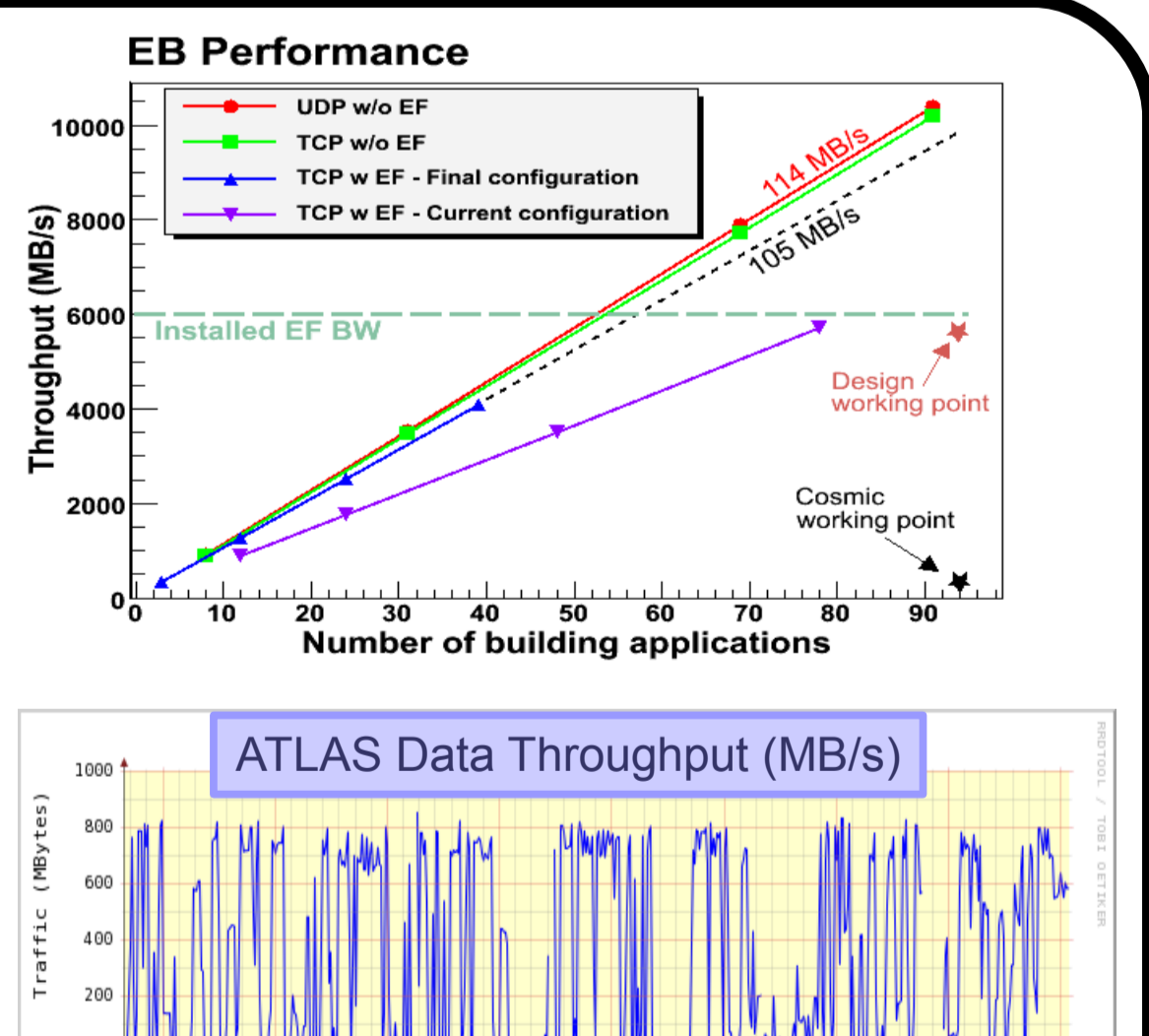
A major feature provided by the DataFlow infrastructure, combining the routing and streaming capabilities, is the so-called "Partial Event Building" (PEB). For detector calibrations often only a subset of the full event data is needed. Being able to collect and transport only such a subset allows to sustain higher calibration rates and to reduce the on-line and off-line bandwidth and storage volume requirements. In the ATLAS TDAQ system, dedicated LVL2 calibration algorithms select events interesting for detector calibrations specifying the needed data subsets. The EB is then allowed to build a partial event which is forwarded to the data-logging nodes, skipping further processing steps. The streaming and routing capabilities as well as the PEB functionality have been largely used during the ATLAS cosmic data taking period.

## DataFlow Performance



Regular TDAQ tests are performed in order to assess the system performance, beyond the cosmic working point, and to evaluate new software releases. The present system do not allow to probe the final working point conditions, however the system size already allows for the evaluation of scalability and performance. Recently the trigger menu dedicated to the initial LHC luminosity of  $10^{31} \text{ cm}^{-2} \text{ s}^{-1}$  has been tested loading a representative simulated data sample into the ROS PCs. The system has been able reach a LVL1 trigger rate of 60 kHz, with a trigger menu optimized for 10 kHz only, limited by the ROS processing power. The ROS PCs in fact were experiencing request rates up to ~30 kHz, where in the final running conditions at most 12 kHz are foreseen. The EB was driven by the LVL2 at a rate of 4.2 kHz corresponding to throughput of 3 GB/s due to the small event size of the data sample (800 kB).

The results of the trigger menu test are backed-up by independent tests of the EB-EF scaling properties. The EB, when not sending data to the Event Filter processors is able to almost double the throughput required in the design working point. When the EF is connected to the Event Builder, the additional data handling only introduces a performance penalty of ~5%, at least up to the presently installed EF bandwidth. The SFO farm is instead the only component which is regularly used at the design working point and even beyond. The full farm is in fact able to sustain an aggregated I/O rate of 550 MB/s. Roughly 1 PB of data, distributed over 650 thousand files, have been handled by the farm during the ATLAS cosmic data-taking.



\* This research project has been supported by a Marie Curie Early Stage Research Training Fellowship of the European Community's Sixth Framework Programme under contract number (MRTN-CT-2006-035606)