Software based readout driver evolution towards 1 MHz readout as part of the ATLAS HL-LHC upgrade

Serguei Kolos, University of California Irvine, USA

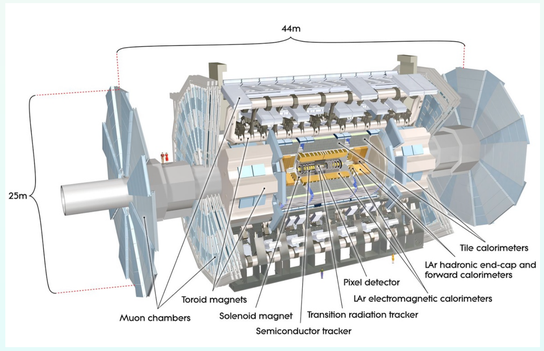
on behalf of the ATLAS TDAQ Collaboration

1. LHC Performance and ATLAS Evolution

ATLAS is one of the four major LHC experiments.

ATLAS is the largest detector ever constructed for a particle collider: 44 meters long and 25 meters in diameter

More than 100 million sensitive electronics channels are used to record the particles produced by LHC collisions.

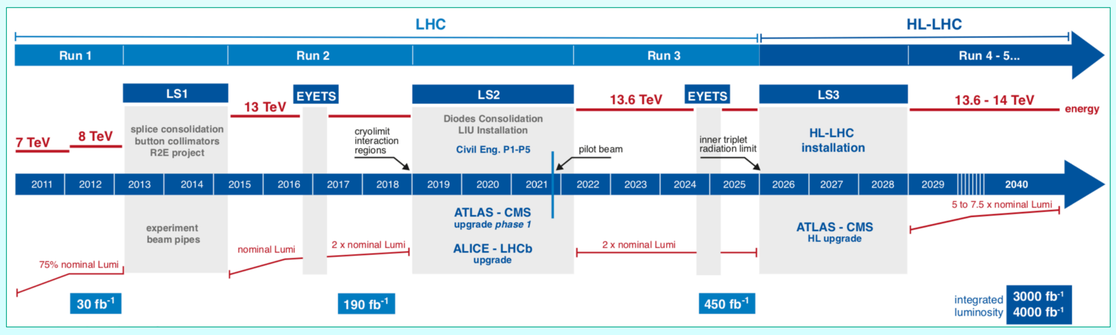


LHC Performance evolution

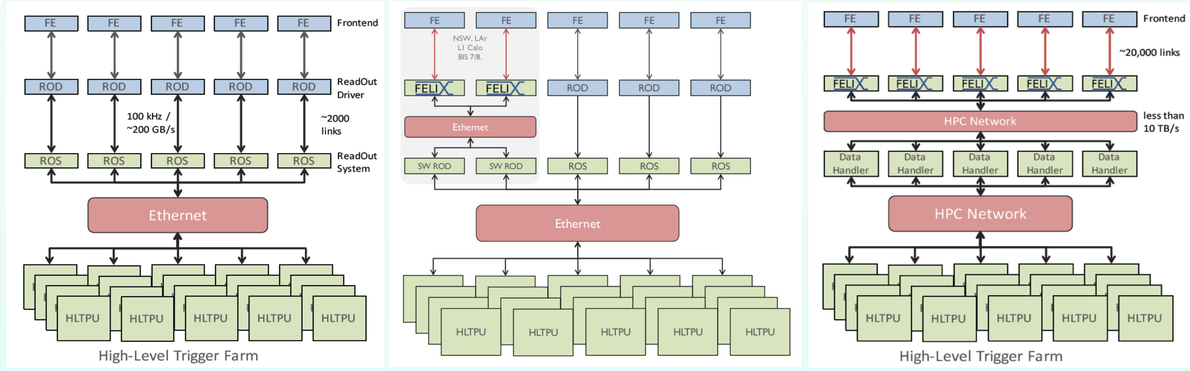
	Period	Energy [TeV]	Peak Lumi [10^{34} cm^{-2}s^{-1}]	Peak Pileup
Run 1	2009 - 2013	7 - 8	0.7	35
Run 2	2015 - 2018	13	2	60
Run 3	2022 - 2025	13.6	2	60
Run 4+	2029 -	13.6 - 14	5 - 7.5	140 - 200

ATLAS Trigger/DAQ system evolution mainly driven by the evolution of LHC performance.

High Luminosity LHC upgrade after Run 3 will require a major upgrade of the ATLAS TDAQ system



2. ATLAS Trigger/DAQ System Evolution



Run 1 & 2

Readout Drivers (RODs) provide interface between Front-End (FE) and DAQ:

- VME boards developed and maintained by detectors
- Connected via point-to-point optical link to a custom to custom PCI/PCIe I/O cards (ROBIN/ RobinNP)
- I/O cards are hosted by Readout System (ROS) commodity computers
- ROSes transfer data to the High-Level Trigger (HLT) farm via a commodity switched network

Run 3

ATLAS uses a mixture of the legacy and new **FELIX**-based readout systems

- **FELIX** is used to read out the Muon New Small Wheel detector, upgraded Barrel RPCs; new Liquid Argon calorimeter digital readout and Level 1 calorimeter trigger.
- A new component, known as the **Software Readout Driver (SW ROD)** has been developed:
- Receives data from FELIX
- Supports the legacy HLT interface

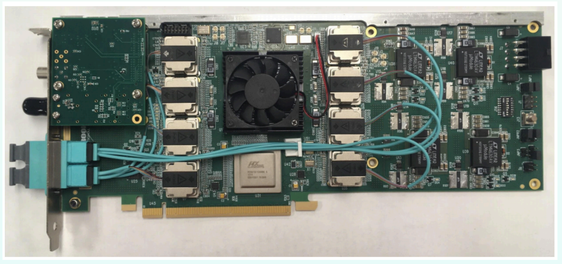
Run 4

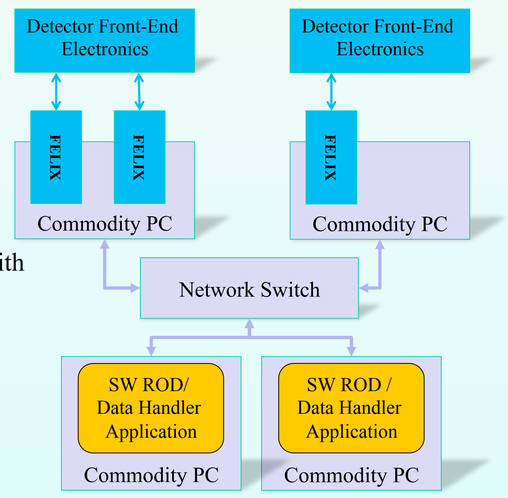
New readout architecture is based on the **FELIX** system:

- New **Data Handler** is an evolution of the **SW ROD**
- **Data Handler** has the same functional requirements as **SW ROD**
- Performance requirements are substantially higher than for Run 3:
  - **1 MHz** L1 rate (10x)
  - **4.6 TB/s** data readout rate (20x)

3. FELIX & SW ROD Readout for Run 3 & 4

New Readout system is based on a custom PCIe card called FELIX





Run 3 version of the FELIX I/O card is a custom PCIe board with Gen 3 x 16 interface installed into a commodity computer:

- Up to 48 optical input links

Can be operated in several modes:

GBT Mode:
- 4.8 Gb/s per link input rate
- Each link can be split into multiple logical sub-links (E-Links)
- Up to 192 virtual E-Links per card

FULL Mode:
- 12 links at full speed for Run 3
- 24 links at full speed for Run 4
- Up to 9.6 Gb/s per link input rate
- No virtual link subdivision

Run 4 version of the FELIX I/O card will support higher throughput as well as additional link protocols:

- **lpGBT** is a new protocol for Low Power Gigabit Transceiver device that can transfer data at 10.24 Gb/s input rate
- **Interlaken** is a point-to-point protocol that support 25 Gb/s input rate

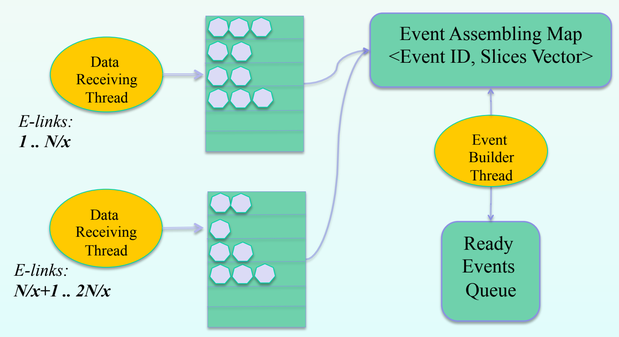
Run 3 Performance Requirements

	Packet Size (B)	Packet Rate per Link (kHz)	Links per FELIX Card	Packet Rate per card (MHz)	Data Rate (Gb/s)
GBT Mode	40	100	192	19.2	6
FULL Mode	5000	100	12	2.4	50

Run 4 Performance Requirements

	Packet Size (B)	Packet Rate per Link (kHz)	Links per FELIX Card	Packet Rate per card (MHz)	Data Rate (Gb/s)
GBT Mode	64	1000	384	384	196
FULL Mode	1024	1000	24	24	192

4. SW ROD Event Building Algorithm Performance for Run 3

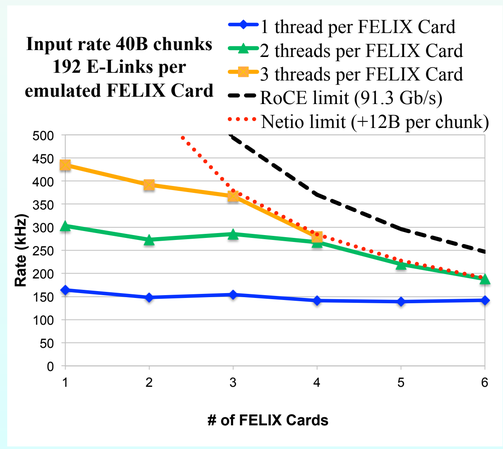


Data Receiving Thread:

- Receives data from a subset of E-Links
- Builds an **Event slice** by copying data into a pre allocated memory buffer

Data Receiving Threads are almost independent:

- Interaction happens when completed **slices** are inserted into the Event Assembly Map, through which complete **Events** are built



Event building rate scales almost linearly with the number of Data Receiving Threads

5. Run 4 Performance Test Setup

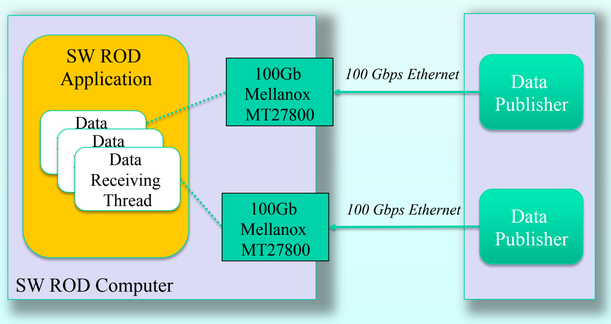
To verify how the Run 3 implementation of the SW ROD scales towards Run 4 requirements a dedicated testbed has been set up. Two server models have been tested:

Option #1: Run 3 SW ROD Computer	Option #2
Dual Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz (16x2 cores) • L1d cache: 32K, L1i cache: 32K • L2 cache: 1024K • L3 cache: 22528K 96 GB of RAM	AMD Epyc 7313P @ 3GHz (16 cores): • L1d cache: 32K, L1i cache: 32K • L2 cache: 512K • L3 cache: 32768K 128 GB of RAM

Data Receiving Threads perform event building:

- Aggregate incoming data chunks into events

The number of **Data Receiving Threads** in the SW ROD Application varied from **2** to **12**



Data Publishers simulate input from FELIX:

- Publish data chunks of the given size for the given number of input links

Number of links:

- 24, 48, 96, 192, 384

6. Run 4 Performance Test Results

Packet sizes used in tests calculated with the following equation:

$$PacketSize(B) = \frac{1.9 * 10^{11}\left(\frac{Gb}{s}\right)}{10^{6}\,(Hz)} : 8 : N_{E-Links} - 12$$

Accounts for the overhead introduced by:

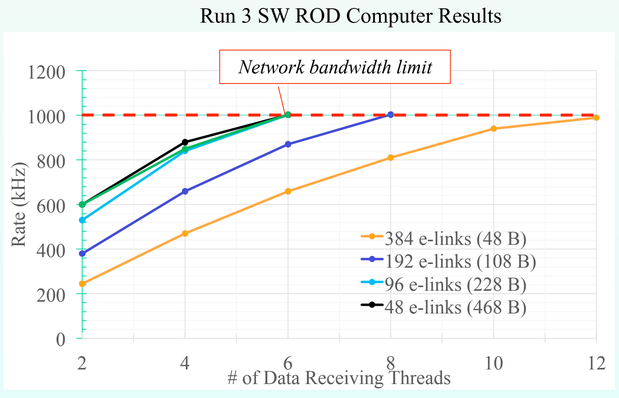
- 12 bytes-per-packet transport overhead:
- 190 Gb/s real bandwidth of the test network

E-links per GBT link	$N_{E-Links}$	PacketSize (B)
1	24	**946**
1	48	**468**
2	96	**228**
4	192	**108**
8	384	**48**

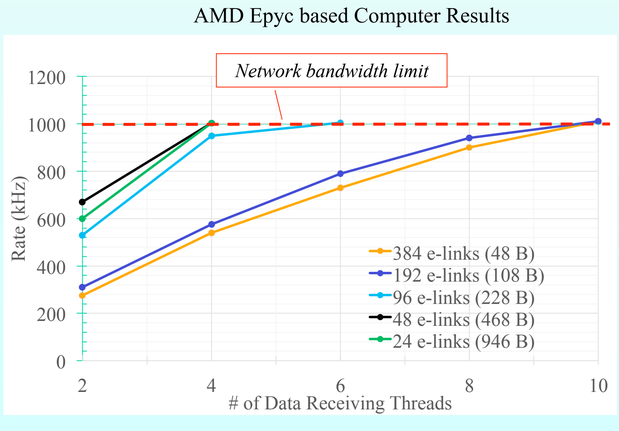
Event builder performance scales well with the number of E-Links and packet sizes:

- The number of required Data Receiving Threads increases proportionally to the number of E-Links
- The overhead produced by thread synchronization is insignificant

Run 3 SW ROD Computer Results



AMD Epyc based Computer Results



7. Conclusion

The High-Luminosity Large Hadron Collider (HL-LHC), expected to enter in operation in 2029, aims to increase LHC luminosity by a factor of 10 beyond its original design.

The new Readout system for the ATLAS experiment is based on the Front-End LInk eXchange (FELIX), introduced for some detectors in Run 3. A new component, called the SW ROD, has been developed to receive data from FELIX.

The Data Handler component of the Run 4 DAQ system will be an evolution of the SW ROD, that will support the same functional requirements but must be able to operate at an input rate of 1 MHz to cope with the HL-LHC luminosity.

Performance testing to date demonstrates that the Run 3 SW ROD application is able to process data at 1 MHz rate for realistic Run 4 input configurations.

It is expected that single CPU core performance should increase by at least 50% in the next 5 years, which will provide extra computing power and decrease overall system cost.