ATLAS Solutions for Phase 2 Storage and Networking

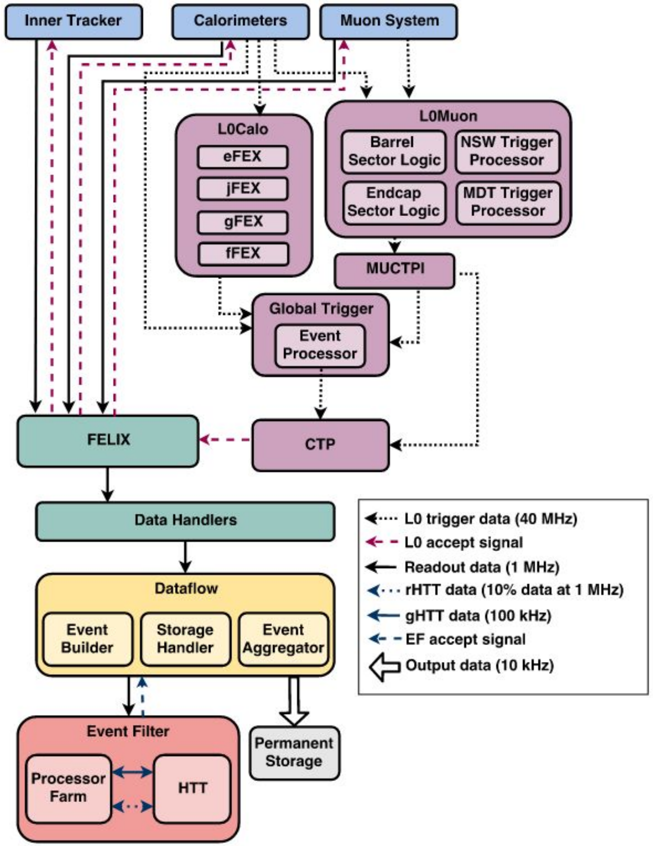
Fabrice Le Goff - 19/11/2019
On behalf of ATLAS TDAQ Collaboration

ATLAS TDAQ[1] for HL-LHC Upgrade

- HL-LHC[2] upgrade
  - Peak luminosity: 7.5 x 10^{34} $cm^{-2}s^{-1}$
  - Collisions per bunch crossing: 200
- LHC to be restarted in 2026

- Detector read-out at **1 MHz**: 10 x more than today's
- Read-out throughput: **5.2 TB/s**: 20 x more than today's

- ⇒ **Major upgrade of all TDAQ** sub-systems



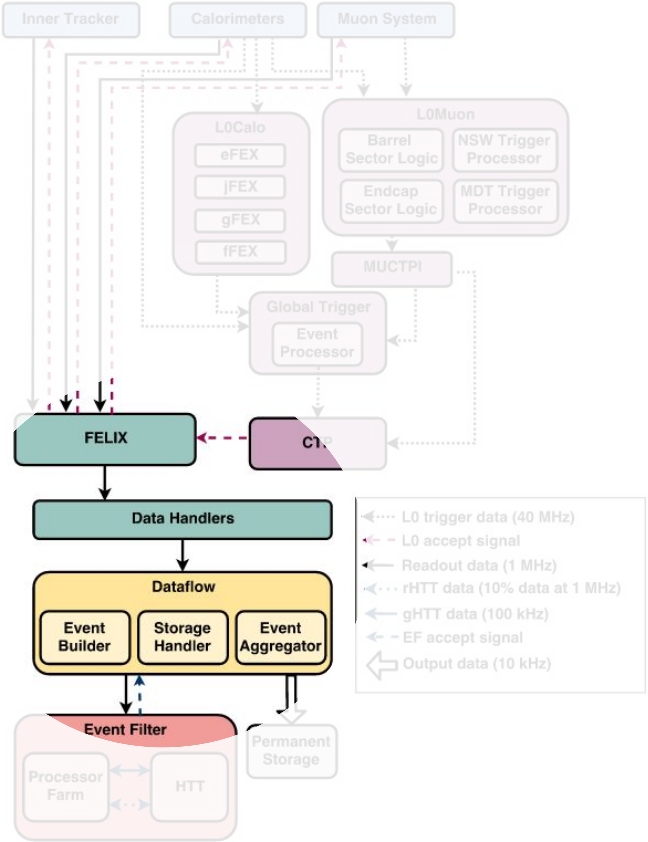
[1] Trigger and Data Acquisition
[2] High-Luminosity Large Hadron Collider

ATLAS TDAQ for HL-LHC Upgrade

- **Networks**
  - **One new network** interconnecting Felix and Data Handlers
  - **One Upgraded network** interconnecting Data Handlers, Dataflow, Event Filter and Offline systems

- **Felix**: common "gateway" between custom electronics and commodity hardware
- **Data Handlers**: detector-specific data processing, formatting, etc.

- **Storage: Dataflow**
  - Data buffering for the Event Filter
  - Event building
  - Persistent storage: decouple detector read-out from event selection
  - Online storage for selected events: decouple online and offline



Network

Networks in Phase 2 TDAQ

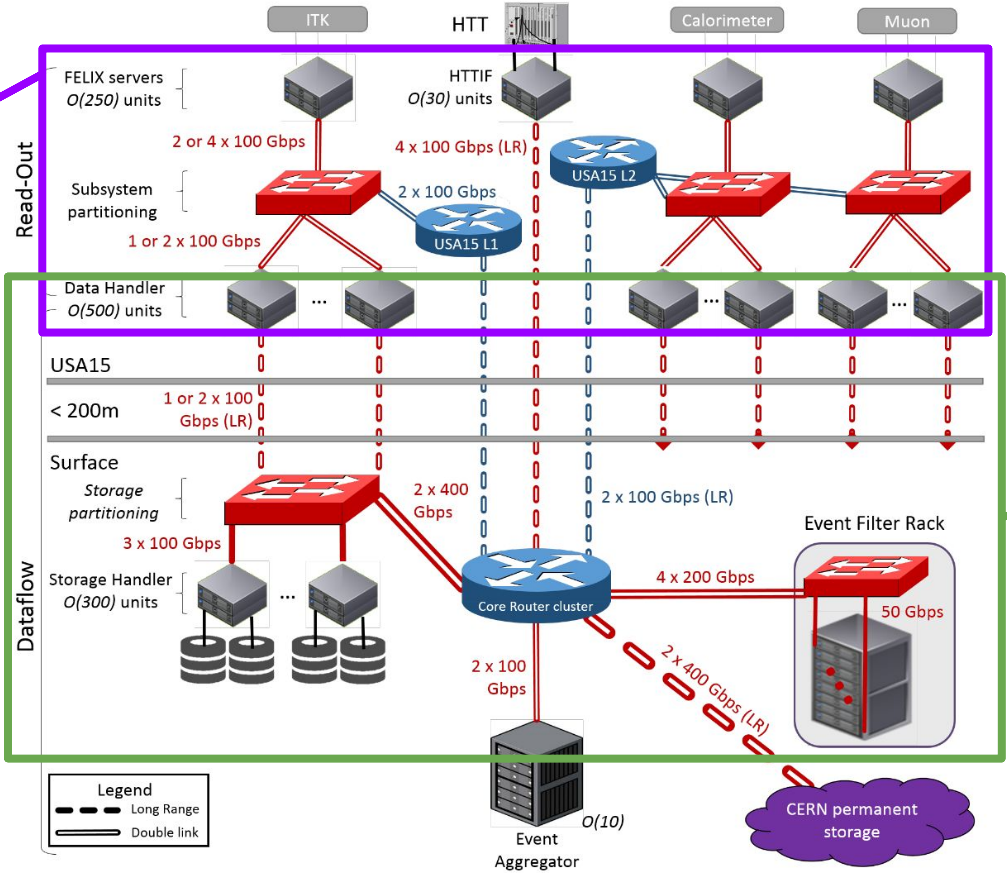
Readout networks (Felix)

High-throughput
Low-latency

Independent slices

Asymmetric traffic:
- 5.2 TB/s of data downstream
- Only detector control and security upstream

O(800) servers



Dataflow network

Interconnected slices

Connects underground experiment to surface data center:
Long-range fibers

8+ TB/s total throughput (writing + reading + extra)

Thousands of nodes (event filter)

Complex core network design: connects all the slices for a many-to-one data pattern

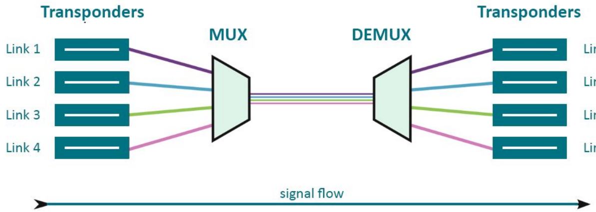
Considerations on Network Technologies

- Commodity technologies
  - Evolution driven by HPC
- DAQ is not HPC
  - Heterogeneous, complex system
  - Real-time processing
  - Network topology aware application
- But HPC hardware technology is interesting for DAQ
  - Low-latency
  - High-throughput
  - RDMA
  - Loss-less

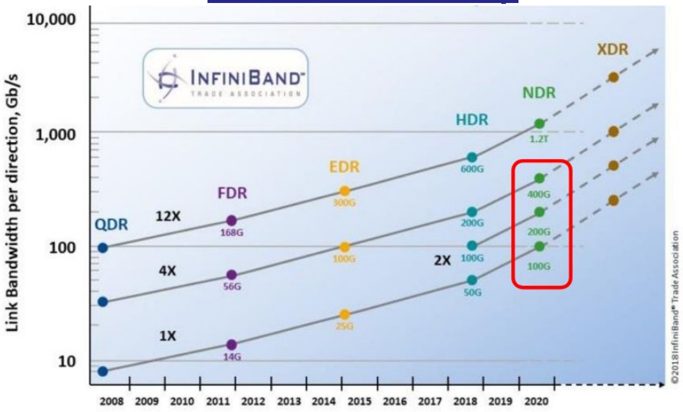
- HPC software not suitable (e.g. MPI)
  - Paradigm is very different (Single Process Multiple Data, heterogeneous system)
  - Developed an in-house network library to take advantage of the technology with our workload
  - NetIO (see Jörn Schumacher's paper for more information)
  - Soon to be open-sourced

Technologies under Evaluation

- Infiniband
  - Forwarding (“routing”) policy is simple
  - Single-speed network
  - Slightly less expensive per port (at given speed)
    - But operations more complex
    - Smaller community
  - ⇒ Considered as a fallback solutions if Ethernet do not reach what we expect
    - For the readout network only, as the Event Filter farm is already equipped with Ethernet
- Ethernet technologies already available (100 and 200 GbE)
  - Now we want: high density, low power, lower price
- WDM (wavelength division multiplexing) to reduce the number of physical long-range fibers
  - Trade-offs like fibers vs. transceivers cost



Infiniband Roadmap



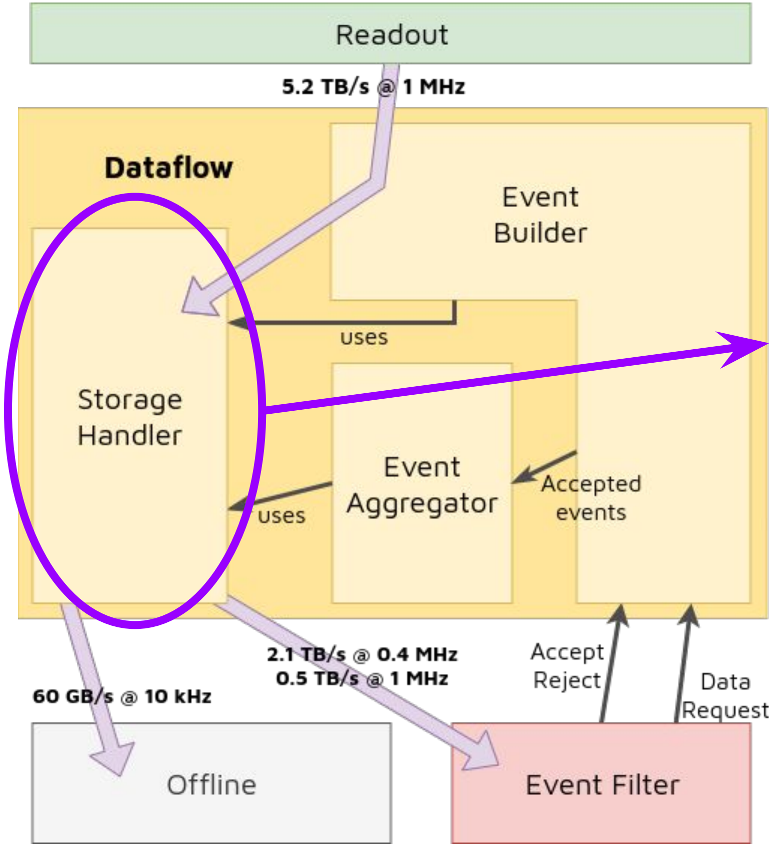
Ethernet Roadmap

Storage

A New Dataflow System

Novel design persistent storage to completely decouple detector readout from event selection

Support:
- recording of all read-out data at **5+ TB/s**
- transfer of read-out data to event filter: **2.5+ TB/s**
- buffering of read-out data for O(10) minutes: **3+ PB**
- recording of selected events at 60 GB/s
- buffering of selected events for 48 hours: **10+ PB**
- Elementary block size: **10 kB**



Challenge

Large-volume
High-throughput
Distributed storage system

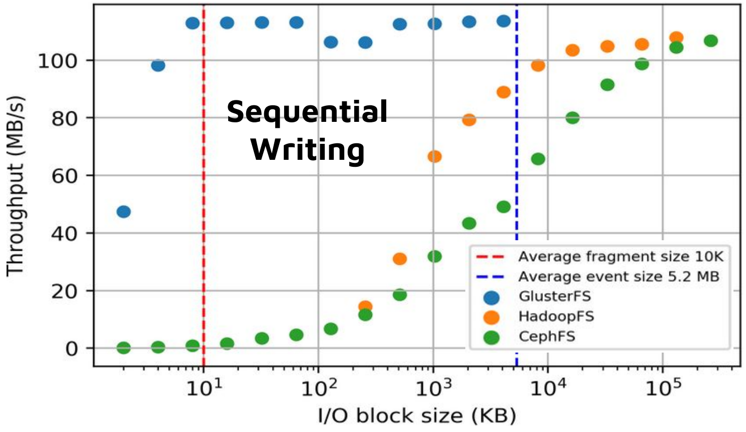
Single "namespace" for all data

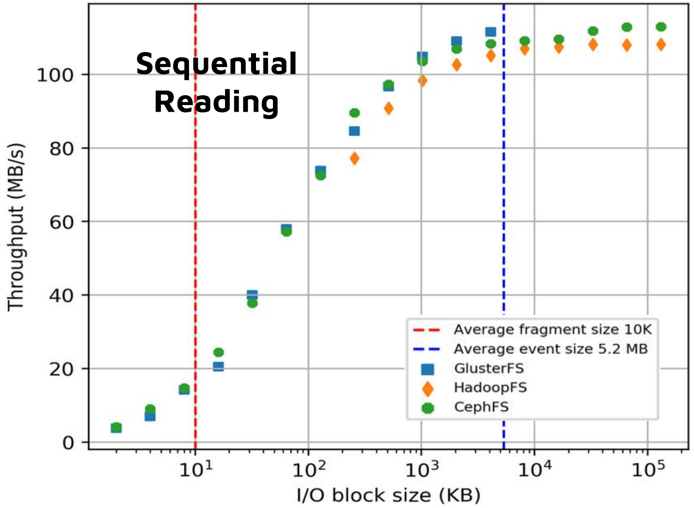
Software Storage Technology Evaluation

- Studies on **Distributed File Systems** (see Adam Abed Abud's paper for more details)
  - Obvious solutions that provide a global namespace
  - Actively developed and heavily used by the industry
  - Comes in many different flavours, with advanced features (load balancing, data redundancy, etc.)

- Tested three DFSs on small scale: GlusterFS, CephFS, HDFS
  - Operation, maintenance, performance quite variable

- **Performance** overhead
  - Especially with small blocks of data





Software Storage Technology Evaluation

- Significant **space** and **network** overhead
  - Order of a few KB per file
  - Same order as our blocks of data

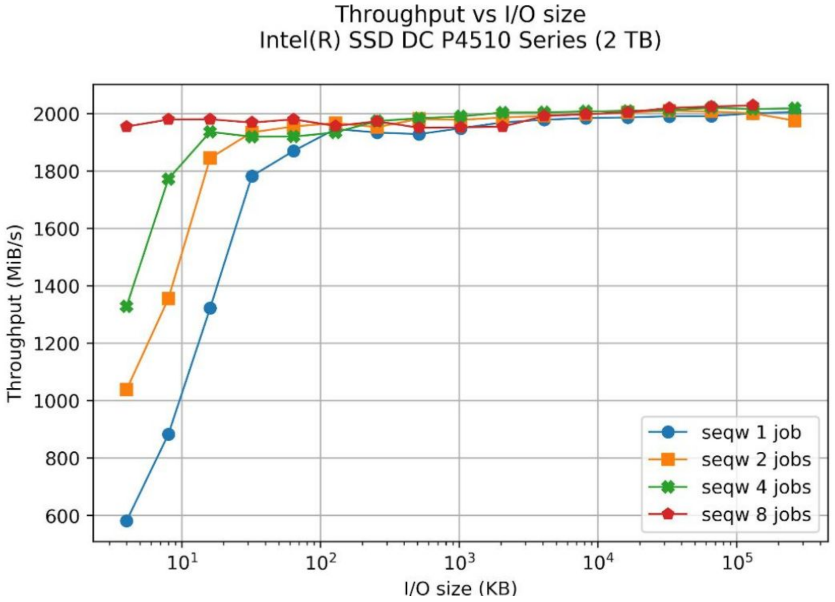
Traffic generated creating empty files

Gluster	Client to Bricks	1.02 KB/File
	Bricks to client	1.02 KB/File
Hadoop	Client to namenode	0.21 KB/File
	Namenode to client	0.08 KB/File
Rados	OSD to MON	2.0 KB/File
	MON to OSD	1.4 KB/File
Ceph	Client to MDS	0.67 KB/File
	MDS to client	1.35 KB/File
	OSD to MDS	0.40 KB/File
	MDS to OSD	5.75 KB/File
	MON to OSD / OSD to MON	Negligible

- **⇒ Unsuited for a direct use in our case**
  - Overheads are too high at our elementary data block size
  - We don’t see DFS improving in this area
- Investigating lower-level solutions like Distributed Hash Tables
  - Considering custom in-house solution in parallel

Hardware Storage Technology Evaluation

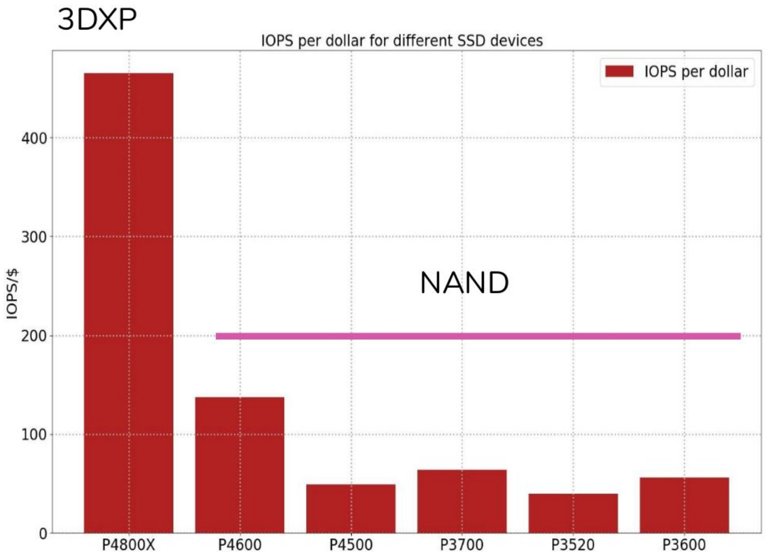
- Projected implementation based on foreseeable technologies
  - O(2000) SSDs distributed in O(300) servers
  - Each SSD: 2.6 GB/s writing, 1.3 GB/s reading
  - Already available on the market: close to expected performance



Hardware Storage Technology Evaluation

- Problem: SSDs wear out when written
- With current NAND SSDs, all SSDs in our system would have to be replaced every year

- 3D-XPoint technologies (Intel, Micron) offers much better endurance: 40 x higher
- IOPS/$ much higher



Conclusions

- HL-LHC upgrade requires a major upgrade of the whole ATLAS TDAQ

- Networking
  - Upgraded and new networks
  - Ethernet is likely to meet our requirements
  - Dedicated software for DAQ’s specific use case
  - Technology evaluation will go on
    - Understand technology, follow its evolution, get hands-on experience
    - Decision in 2023

- Storage
  - New persistent buffer for readout data
  - Commodity hardware, taking advantage of technology evolution
  - Technology evaluation: software and hardware
    - Understand technology
    - Prototype in-house solution
    - Working prototype (small scale) by 2022