ATLAS DAQ for Run4

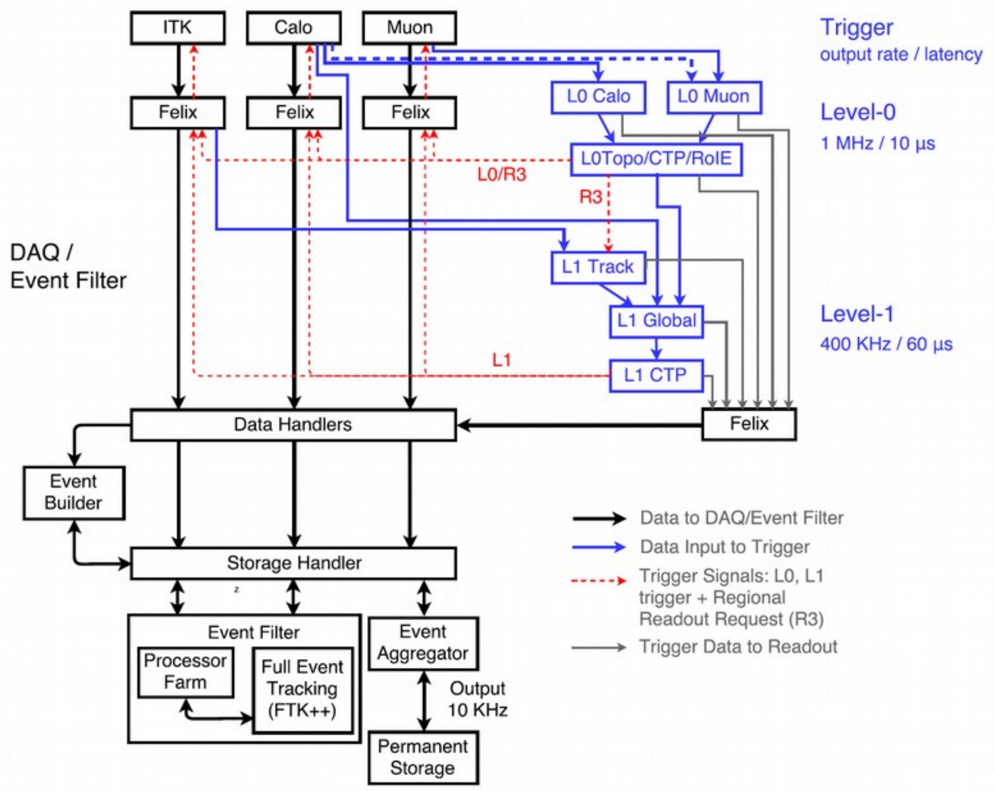
Outline

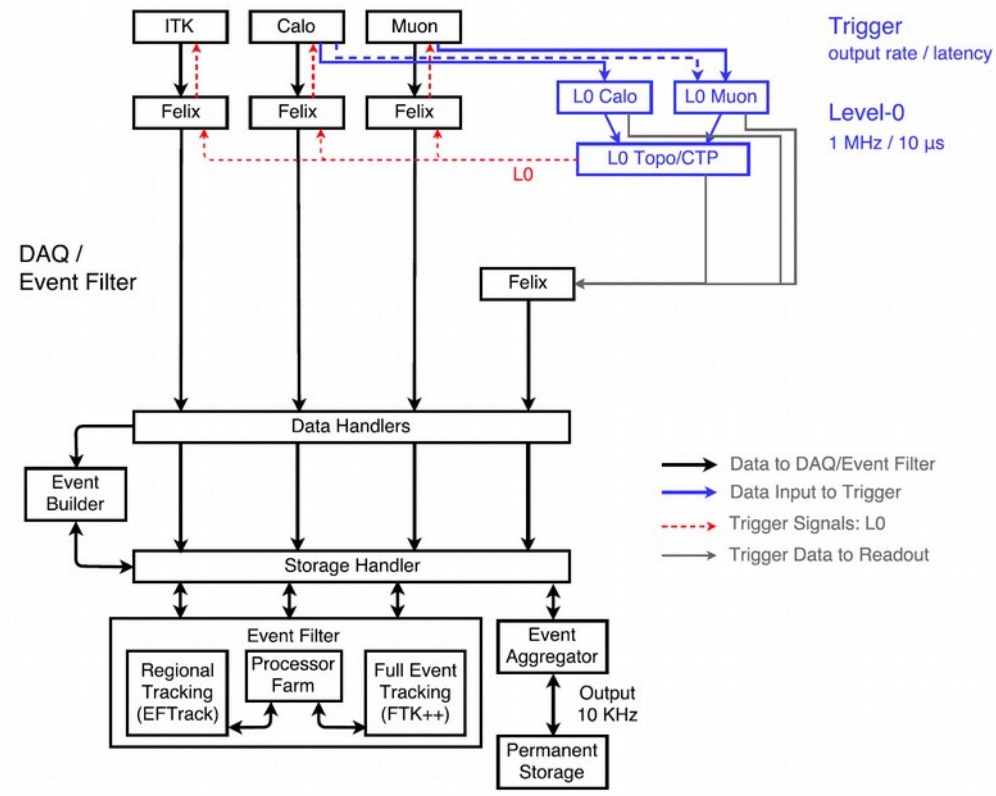
- ATLAS in Run4
- Overall TDAQ Architecture and Hardware Tracking
- DAQ Requirements and Design
- Storage and Event Filter
- Outlook

ATLAS Detector in Run4

- Significant detector upgrades
  - new all-silicon inner detector (ITK)
  - new forward calorimeter being considered
- Complete transition to GBT (-like) detector interface
  - electronic replacement for all legacy detectors (muon spectrometer and calorimeters)
    - *partially implemented in Run3*
- New Trigger, Timing and Control infrastructure
  - passive optical network tree
  - configurable/customizable edge element (Local Trigger Interface LTI)
- New Trigger and Data Acquisition

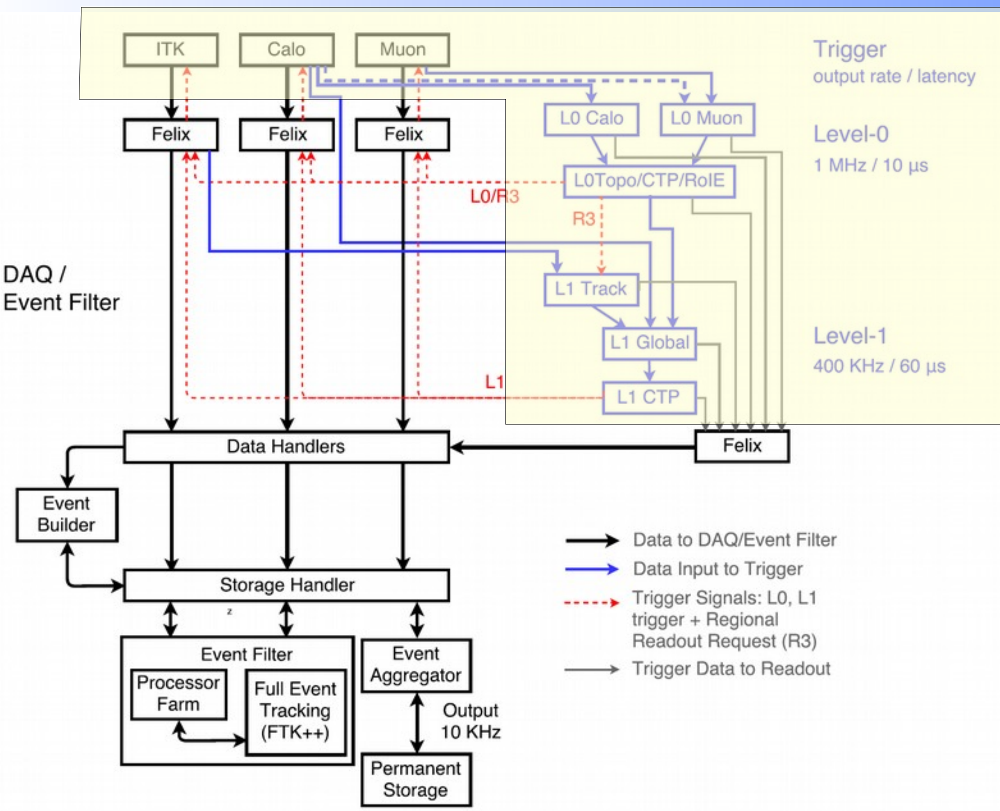
Overall TDAQ architectures and scenarios

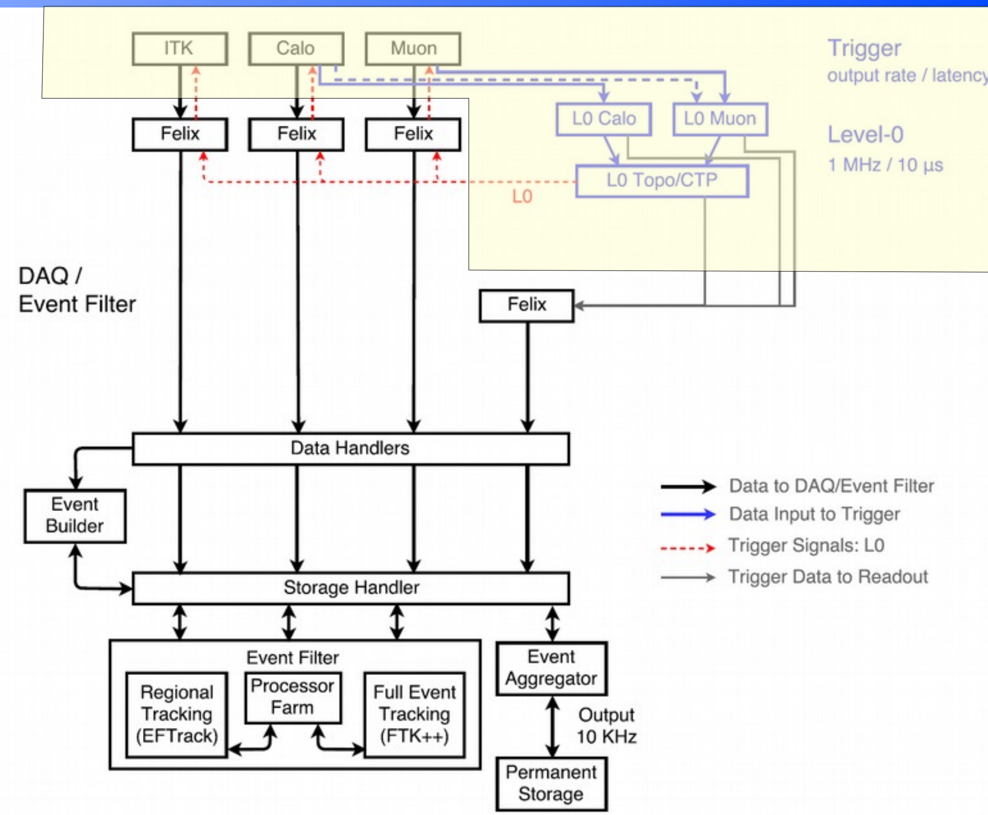




- Two main architectures being considered
  - driven by the detector readout capabilities
  - different hardware trigger organization
- L0/L1
  - two hardware trigger levels → 1 MHz – 400 kHz
- L0
  - one hardware trigger level → 1 MHz

Overall TDAQ architectures and scenarios

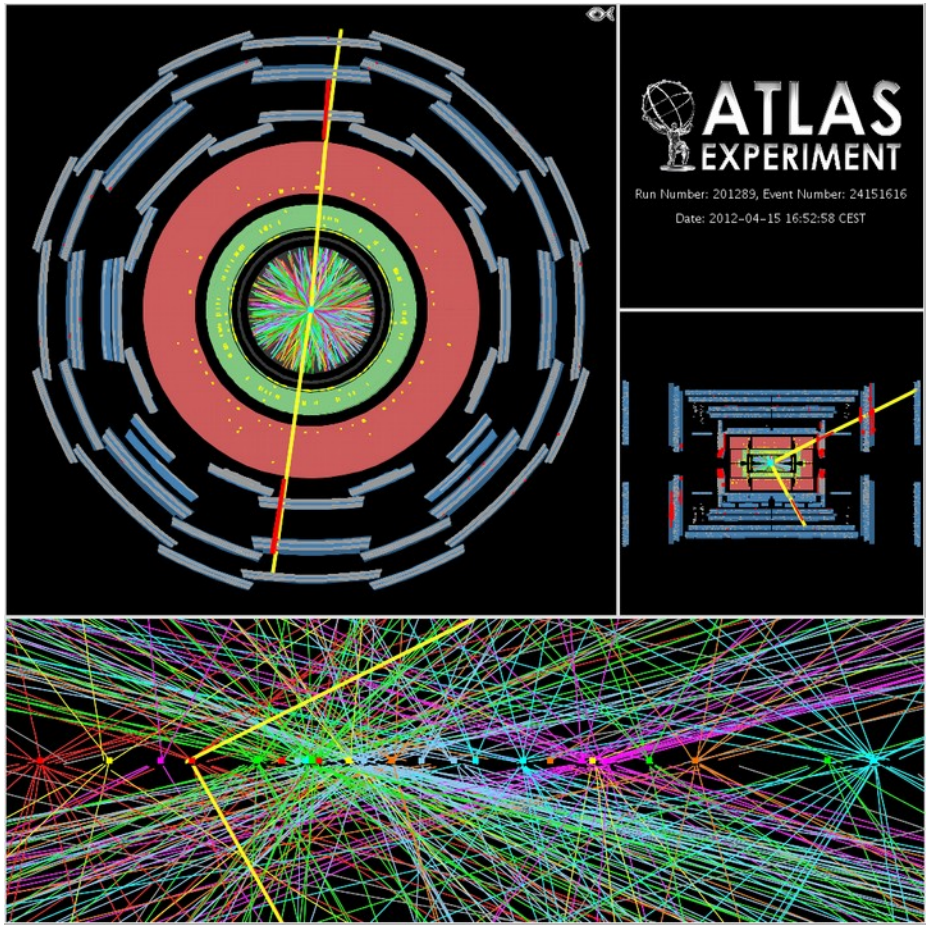


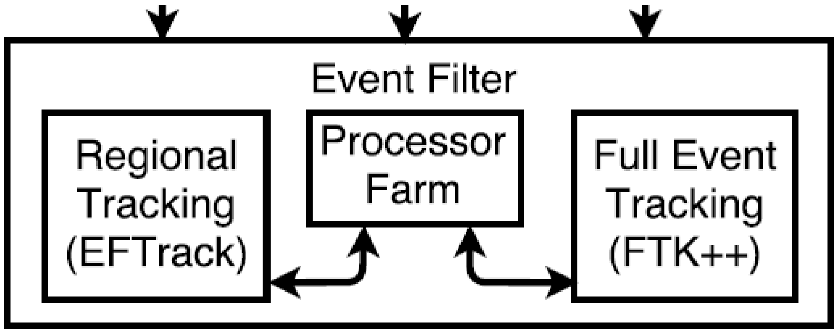


- No major differences in the DAQ architecture
- Readout and Filtering rate
  - L0/L1 → 400 kHz
    - *but some detector may readout 1 MHz*
  - L0 → 1 MHz

Hardware Tracking

- Tracking major component of processing time and key tool in high-pile environment
- Specialized device being deployed for Run2/3
  - Fast Tracker (FTK) based on track patterns stored in associative memory (AM) banks (talk from S. Veneziano)
- In Run4 expect dedicated tracking devices in Event Filter
  - full tracking for a fraction of the input rate (100 kHz)
  - in L0 scenario, RoI-based tracking at 1 MHz
    - *part of L1 hardware trigger in L0/L1*
- Tracking devices based on the same technology
  - AM banks, FPGA, ...





DAQ: operation point and design principles

Parameter	L0/L1	L0	Run2/3
Input Rate	400 kHz/ 1 MHz (*)	1 MHz	100 kHz
Filtering Rate	400 kHz	1 MHz	100 kHz
Output Rate	10 kHz	10 kHz	1 kHz
Event Size	5 MB	5 MB	1.5 MB

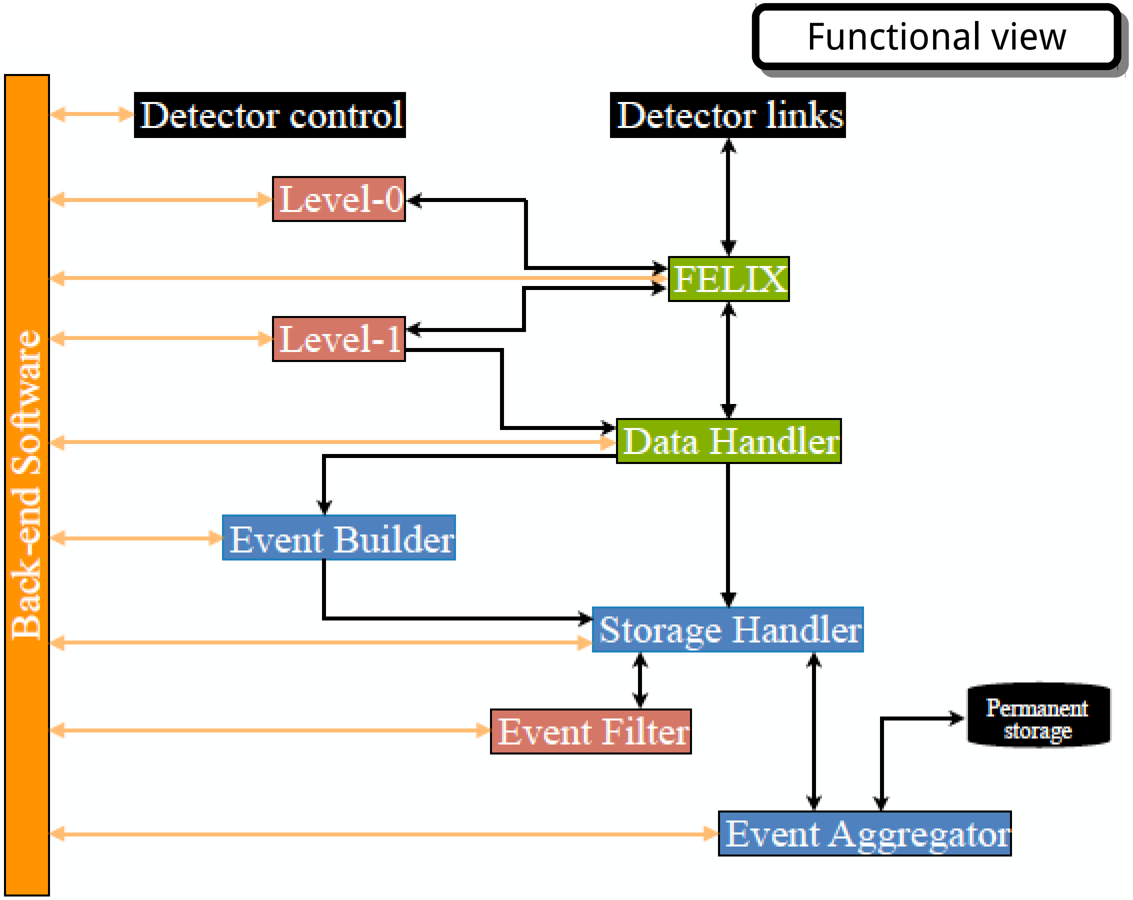
*some detectors planning readout at L0 rate

- Significant operation envelop change
  - 10-fold in readout rate and output rate (L0-scenario)
- Increase decoupling between data movement and data processing
  - at the implementation and operation level
  - expose a well defined, common interface allowing heterogeneous data processing infrastructure
    - *servers, (custom) tracking devices, accelerators, remote resources, ...*
- Rely on "COTS" as much as possible
  - hardware and software

DAQ Architecture

- At high-level, classic architecture
  - Readout infrastructure to transport data out of the detector
  - Dataflow infrastructure to build events and buffer during filtering time

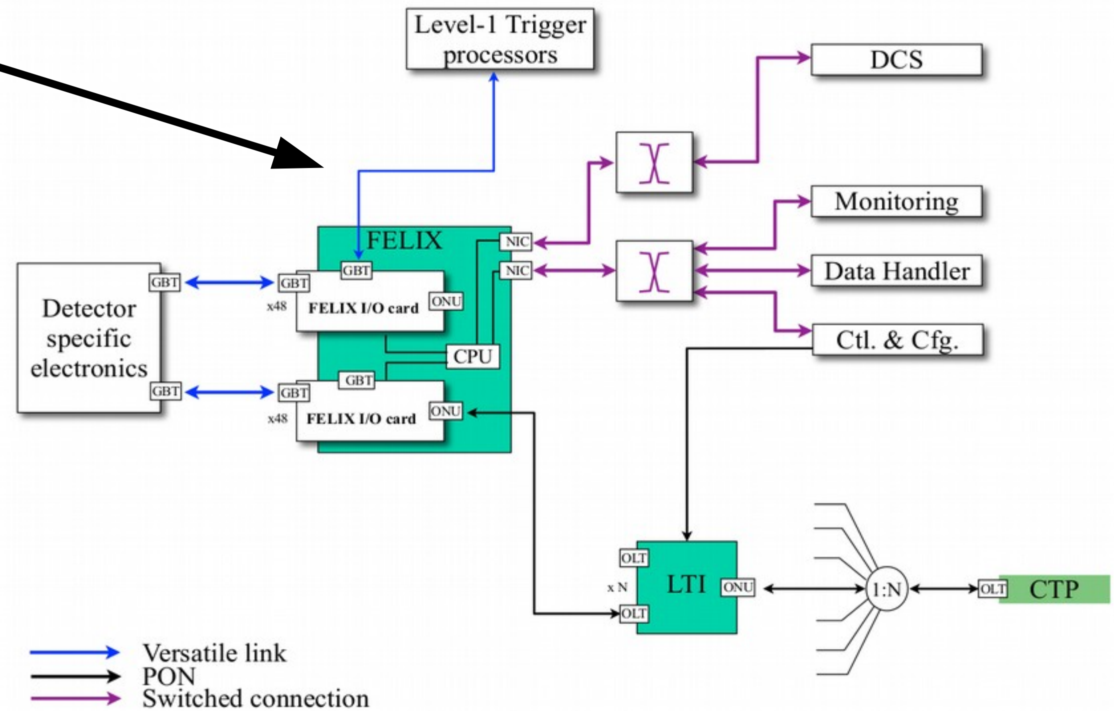
- Introduce a large storage area before filtering
  - high-level interface between dataflow and filtering
    - *allow for a heterogeneous farm (accelerator , tracking devices, ...)*
  - decouple filtering operation from LHC cycle
    - *take advantage of interfill periods → best use of compute resources*
    - *as pioneered by LHCb and soon ALICE*



Detector Interfacing and Readout

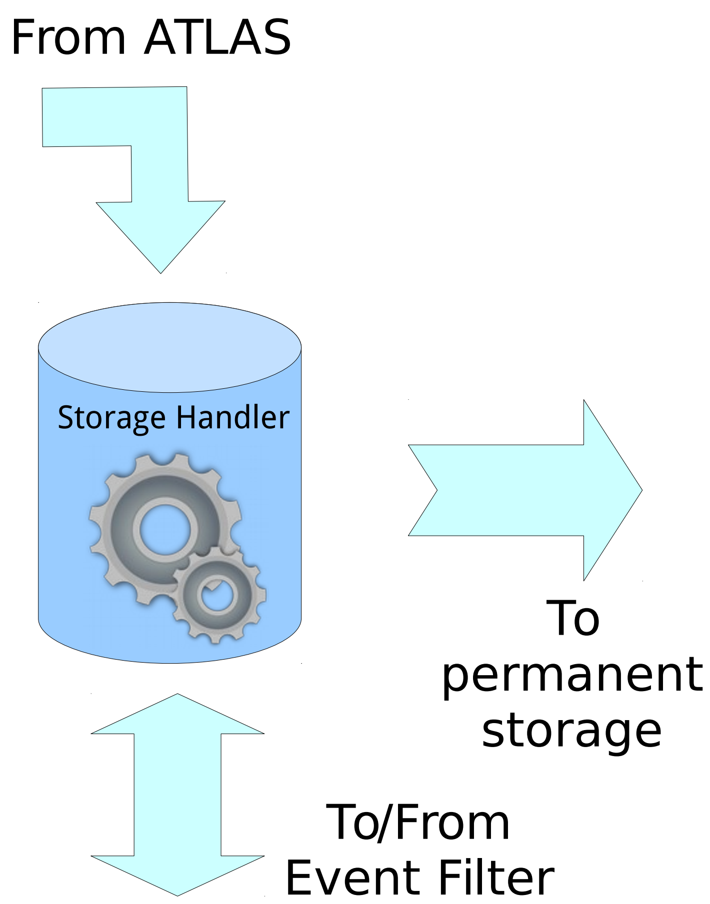
- FELIX extended to the full ATLAS
  - principles and functionalities as described by J.Zhang
    - *heterogeneous router*
    - *unique detector interface*
- New hardware/software implementation
  - faster interface, denser solution, new TTC technology
- Low-latency links towards tracking device
  - RoI data duplication into serial output links
  - RoI map decoding and data-tagging
- Considering needs for detector-specific firmware
  - for time-critical functionalities
- Data Handler implements detector-specific data processing
  - aka SW ROD in Run3

	two-level architecture
Links from detectors	11000
FELIX I/O card	∼450
FELIX servers	∼230
Low latency links	∼900
FELIX NICs (100 Gbps)	∼250
Data Handler PCs	∼400
Data Handler NICs (100 Gbps)	∼500



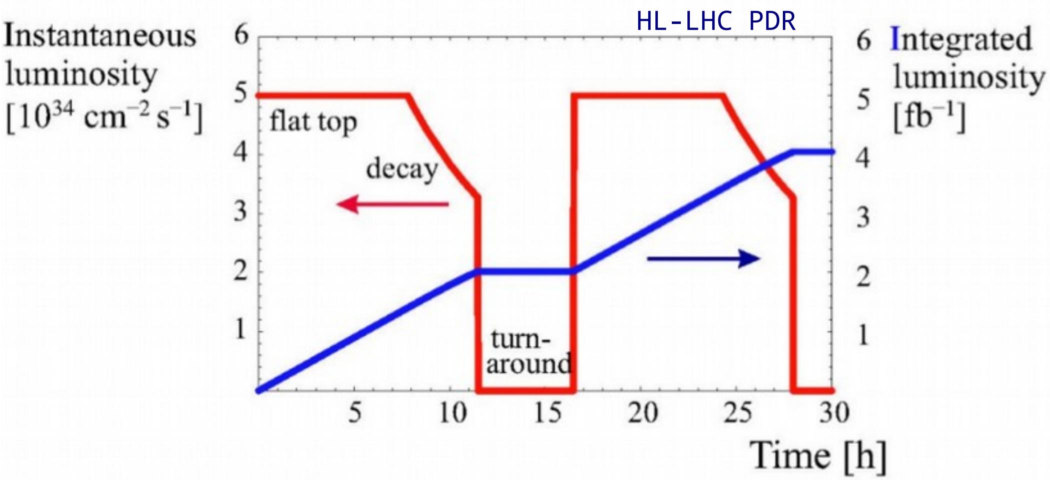
Storage Handler

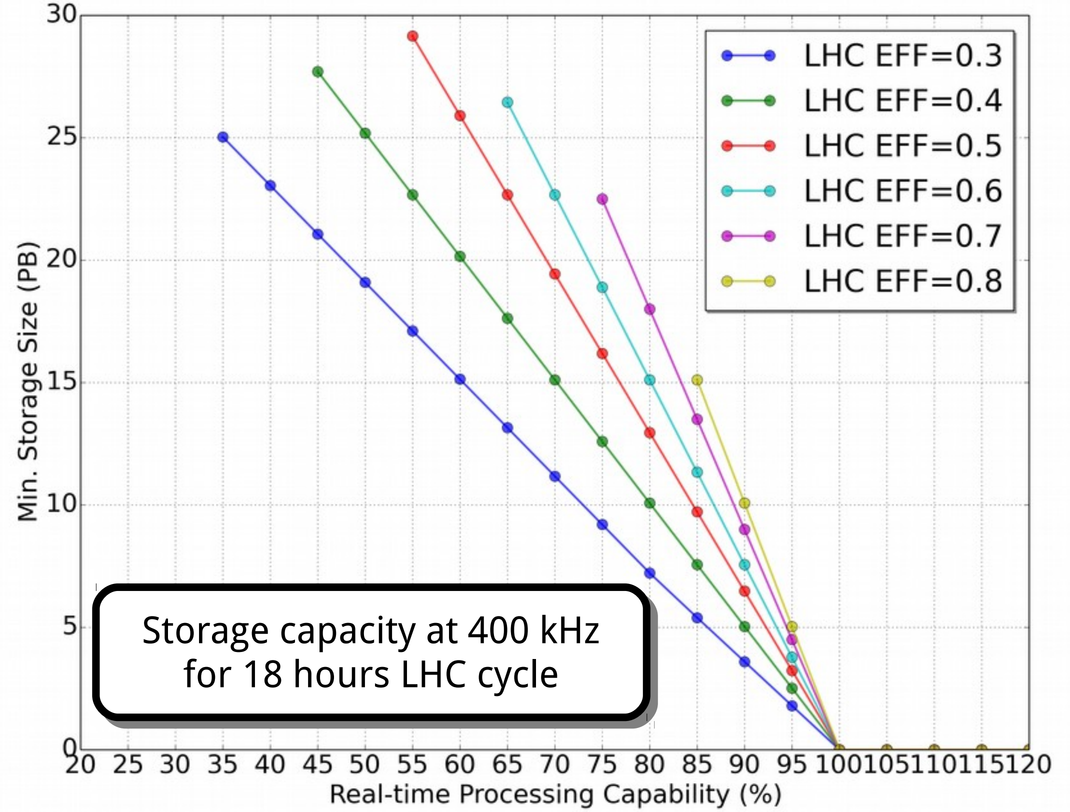
- Storage Handler is core data-flow infrastructure
- Large buffer area, decoupling DAQ and filtering operation
- Offload data movements to distributed file system infrastructure
- Still need to provide
  - data bookkeeping
  - event assignment
  - load balancing
- Do not need dedicated storage for accepted events
  - “Event Aggregator” fetch and aggregates events on their way to permanent storage
    - *e.g. EOS can be mounted as a local filesystem*



Storage Requirements: Capacity

- Storage capacity is a trade-off
  - volume vs level of asynchronous processing
- Depends on
  - typical LHC cycle and efficiency
  - considered timescale
- Several tens of PB for a single cycle
  - 20 – 60 PB





Storage Requirements: Throughput

Parameter	L0/L1	L0
Input from detector	2 TB/s	5 TB/s
Output to tracking devices	<0.5 TB/s	<1 TB/s
Output to farm	<2 TB/s	<2 TB/s
Output to off-line	50 GB/s	50 GB/s

- Throughput is real challenge
  - especially if considering spinning hard-drives
- Exacerbated by evolution of drive characteristics
  - capacity growing much faster than I/O capabilities
- Assume 10 TB/drive
  - 50 PB → 5000 drives
- Assume (optimistic) 100 MB/s/drive
  - 5 TB/s → **50000 drives**

Real world example: Backblaze

- Backblaze: on-line backup founded in 2007
  - open policies: share hardware designs and operational data
    - *hard-drive failure database (including SMART data)*
    - *yearly statistical report on drive failures*
  - by end of 2015 their data centre operated ~56000 spinning drives
    - *~200 PB deployed capacity*

- Backup is a very specific workload
  - capacity problem with rare data readout, no associated processing (HLT farm)
  - results may not apply directly to our case

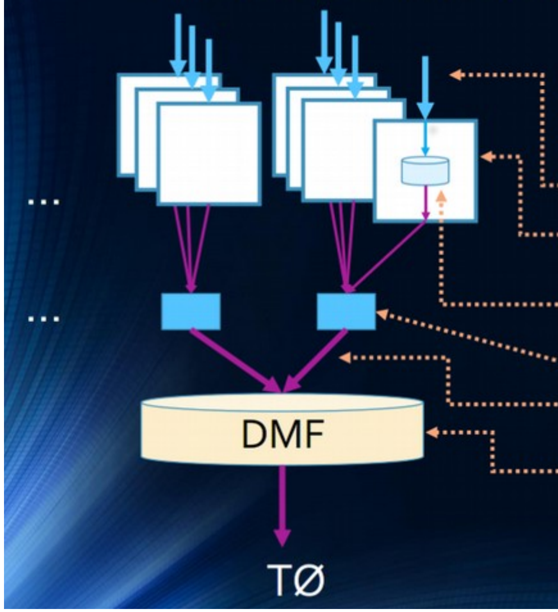
- Average yearly failure rate <3%
  - i.e. 4 drive/day

- 45-drives “Storage Pod” → ~8000 USD
  - ~8 MUSD for ~50000 drives

From www.backblaze.com

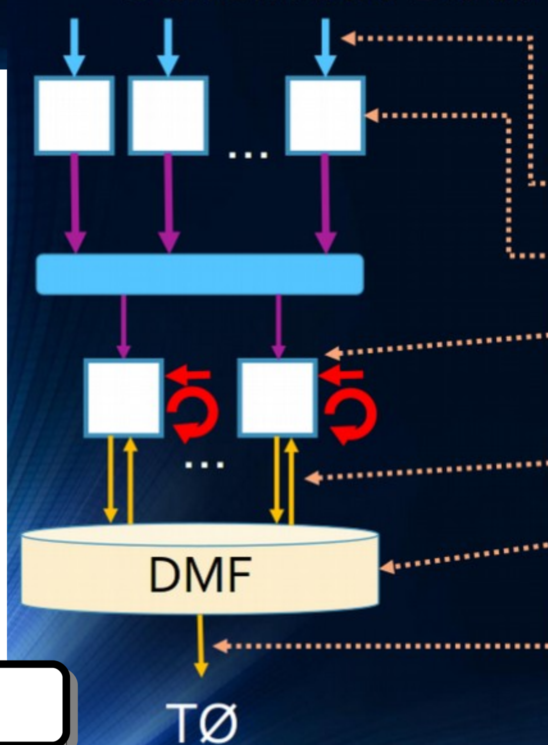
Learn from the future

Schematic Data Flow: LHCb



	Run 2	Run 3
Total Data Input	60 GBps	40 TBps
Readout Servers	2000 30 MBps/server	8000 5 GBps/server
Internal Storage	2-4 TB 7 MBps x2 (rw)	14-32 TB 27 MBps x2 (rw)
I/O aggregation	4 Proxy Servers	Proxy Servers
Total I/O	Max 13 GBps	> 30 GBps
Data Mgmt Facility	340 TB (1 wk data) avg. (1w+2r) GBps	4 PB (1 wk data) avg. (10w+20r) GBps

Schematic Data Flow: ALICE

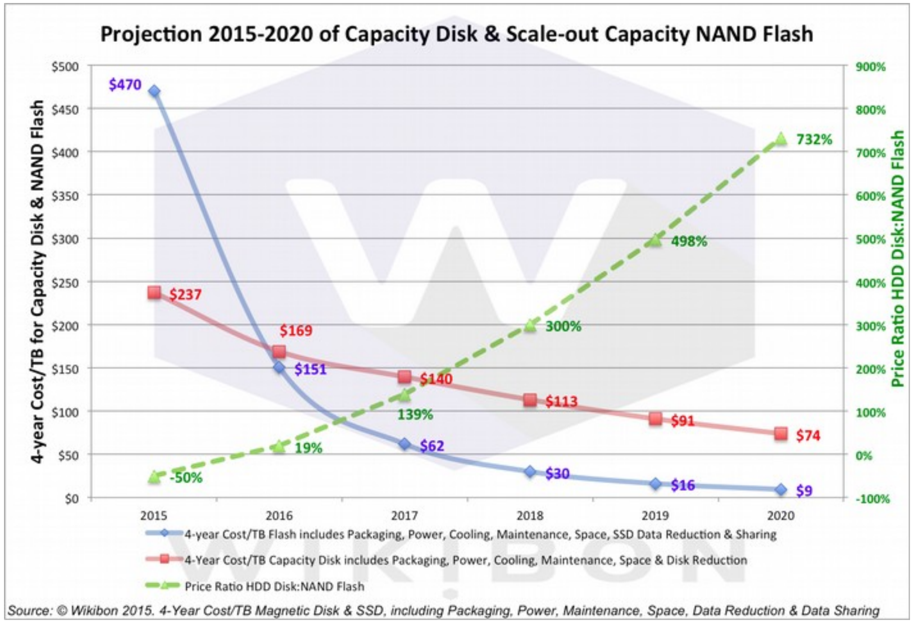


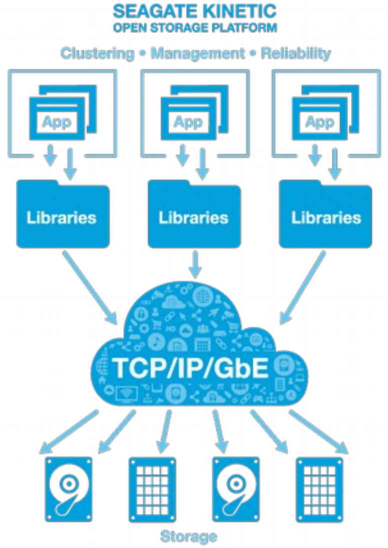
	Run 2	Run 3
Total Data Input	60 Gbps	> 5 Tbps
Readout Servers	200 0.1-10 Gbps/server	250 24 Gbps/server
Event Builder	20 4-5 Gbps/server	1500 3-5 Gbps/server
	L2: External source	EBuild./FTrack./Calib.
Event Builder Out	Max 700-1000 MBps	> 80 GBps
Data Mgmt Facility	1 PB (1.5 days data) avg. (12w+10r) GBps 1200 disks	~100 PB (1y of event and analysis data) avg. (80w+60r) GBps
Data Mover	10 servers max. 1 GBps / server	1/3 of event data / all analysis copied to To

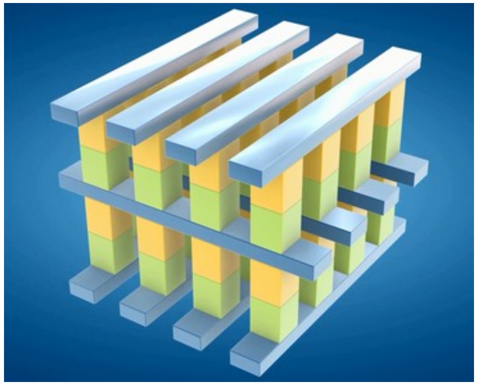
U.Fuchs

Storage Technology Evolution

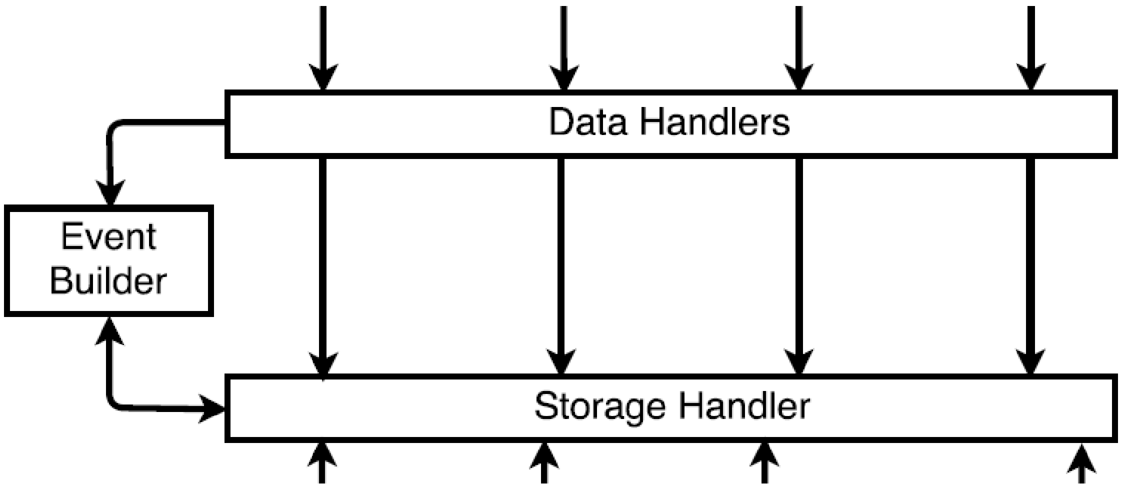
- Looking at storage technologies 10 years from now
- Evolution of existing technologies
  - ~this year consumer NAND cheaper than spinning drive
  - Lustre and GPFS
- New technologies
  - waiting to see 3D XPoint
- Innovations in the storage stack
  - Seagate Kinetic, ...

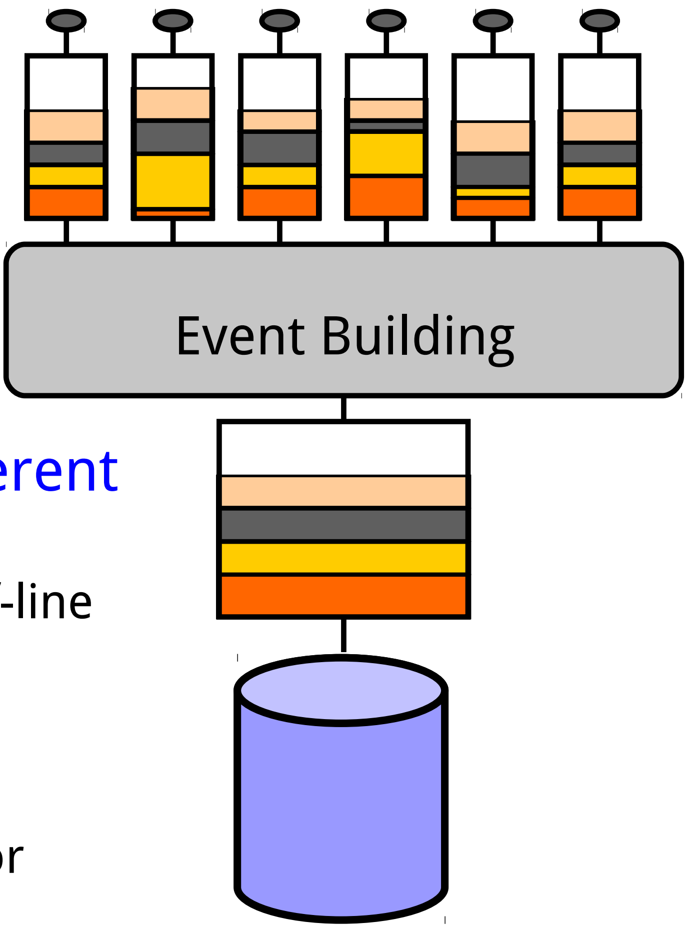






Event Building

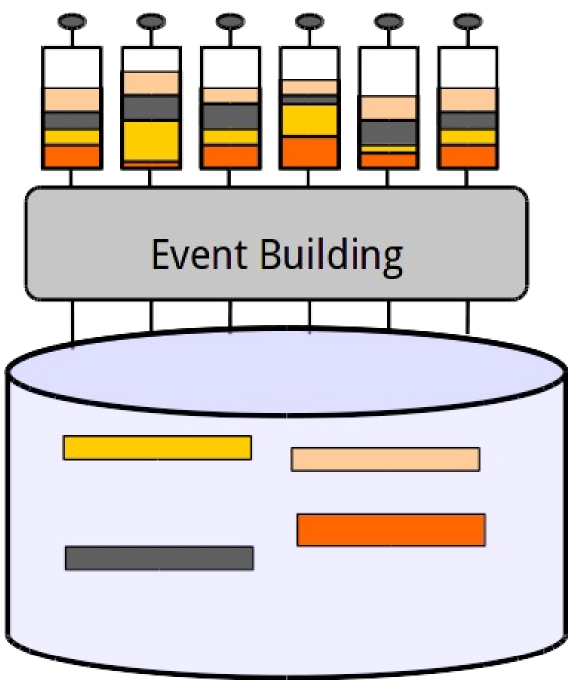




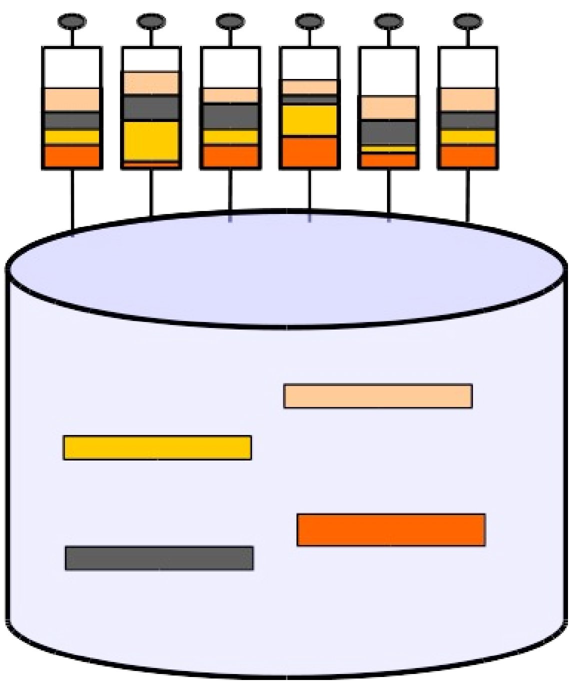
- Aggregating partial data fragments into a coherent unit
  - convenient format for filtering and necessary for off-line transmission
- Really need to gather all pieces together?
  - effectively in Run2 event building takes place only for accepted events
  - need a recipe to access or discard any piece
- "Physical" and "Logical" event building

Event Building

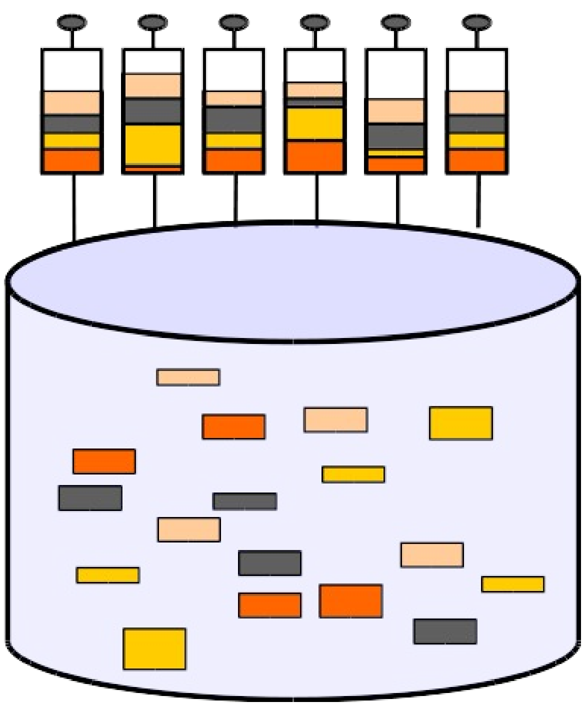
- Physical EB with dedicated resources
  - if needed, isolate EB specific network challenges
  - event-level data compression
    - *ATLAS Run2 events 50% compression ratio*



- Physical EB offloaded to storage
  - possible optimisation
  - depends on storage performance, implementation, ...



- Logical EB
  - only aggregate information on fragment location
  - physical data still fragmented
  - key-value database



Event Filter Implementation

- Expect the Event Filter to include different technologies
  - Run1/2/3 “homogeneous” processor farm
  - Run4 processor farm aided by accelerators
    - *at least for tracking*
    - *do not want to exclude other arising technologies*

- Clear interface allowing various processing implementations
  - what better than files and events?
    - *or object storage*
  - in Run1/2, HLT
    - *requests data-fragment from Readout system → has knowledge the DAQ cabling and partitioning*
    - *use offline software with a software layer mating it to the DAQ environment*

- Expect event processing to be RoI-based
  - reduced data-access for promptly rejected events

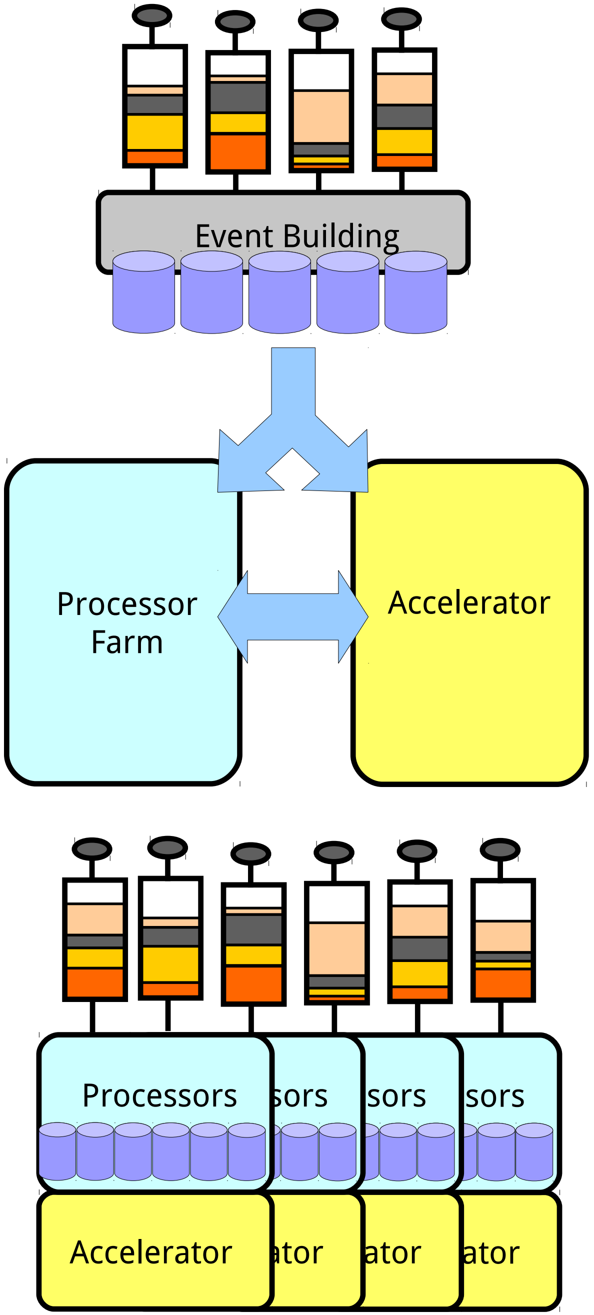
Parameter	L0/L1	L0	Run2
Filtering Rate	400 kHz	1 MHz	100 kHz
Overall Compute Power	11 MHS06	>11 MHS06	0.8 MHS06
Computer Power excluding tracking	5 MHS06	5 MHS06	–

based on projections from Run1

Storage Alternatives: Plan B

- Current design relies on a distributed storage infrastructure
  - scalable
  - affordable, reliable
  - does it all for us
- Better have a plan B ...

- Fallback on a more classical architecture
  - full-event building at input rate
    - *keep the interface simple: just an event*
  - buffering on the building or filtering nodes
    - *buffering on filtering nodes scales with farm size*
    - *monolithic accelerators (tracking) may require upstream buffering*
  - storage for selected events before permanent storage



Outlook

- Initial design for ATLAS data-acquisition in Run4
  - two main scenarios related to detector capabilities evolution

- Full deployment of FELIX and software-based detector-specific processing
  - expect new implementation wrt Run3
    - *new TTC interface*
    - *low latency output link for L1 trigger*
    - *interested in modular firmware experiences*

- DataFlow and Event Building centered around a large storage system
  - decouple filtering from LHC operation
    - *best use of deployed compute resources*
  - expose well-defined stable interface
    - *enable heterogeneous Event Filter implementation*
  - offload data transport to distributed file system engine

- Challenging (but possible?) implementation today
  - confident in the next 10 years of storage technology evolutions
  - learn from (work with) the ALICE and LHCb experience in Run3

Extras

Detailed requirements

DAQ and Event Filter requirements

	Level-0 / Level-1	Level-0
Input rate	400 kHz	1 MHz
FELIX input links	11000	11000
FELIX output links (100 Gbps)	250	700
FELIX input rate:		
Inputs to Level-1:		
ITk	100 kHz	NA
Detector readout:		
ITk-pixel	1* MHz	1 MHz
ITk-strip	1* MHz	1 MHz
Calorimeters	1 MHz	1 MHz
Muon	TBD	TBD
CPU "power" data handler	TBD	TBD
Storage input rate	400 kHz	1 MHz
Event building rate	400 kHz	400 kHz or 1 MHz
<Event Filter processing time>	xxx ms	xxx ms
Output	10 kHz	10 kHz

* Current baseline (but 400 kHz not excluded)