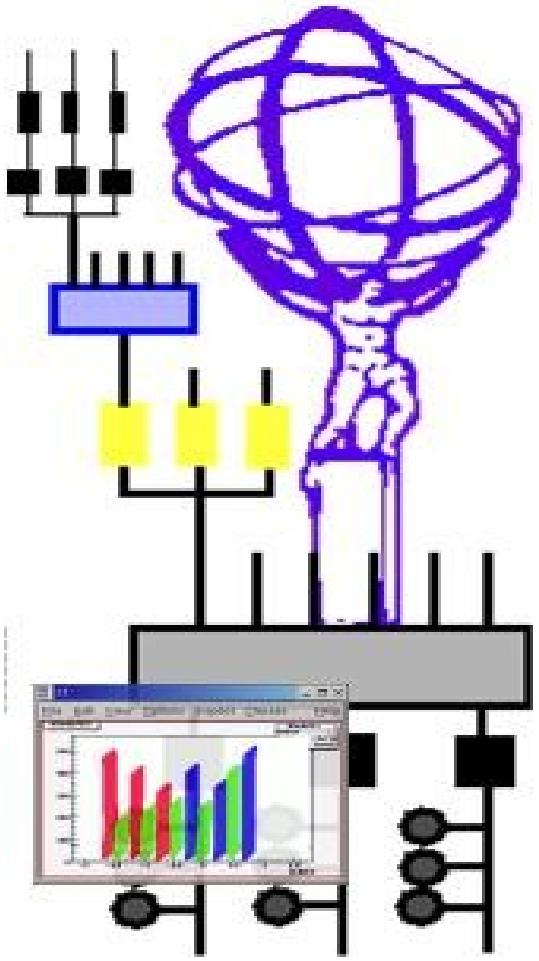
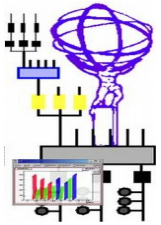


# Evolution of the ReadOut System for the ATLAS Experiment

W.Vandelli – CERN Physics Department/ATD

on behalf of
ATLAS TDAQ ReadOut Team

A. Borga (Nikhef), G. Crone (UCL), B. Green (RHUL), A. Kugel (Heidelberg), M. Joos (CERN), W. Panduro Vazquez (RHUL), J. Schumacher (CERN & Paderborn University), P. Teixeira-Dias (RHUL), L. Tremblet (CERN), W. Vandelli (CERN), J. Vermeulen (Nikhef), P. Werner (CERN), F. Wickens (Rutherford Lab.)

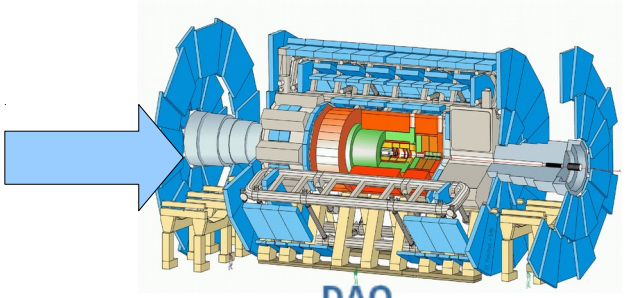


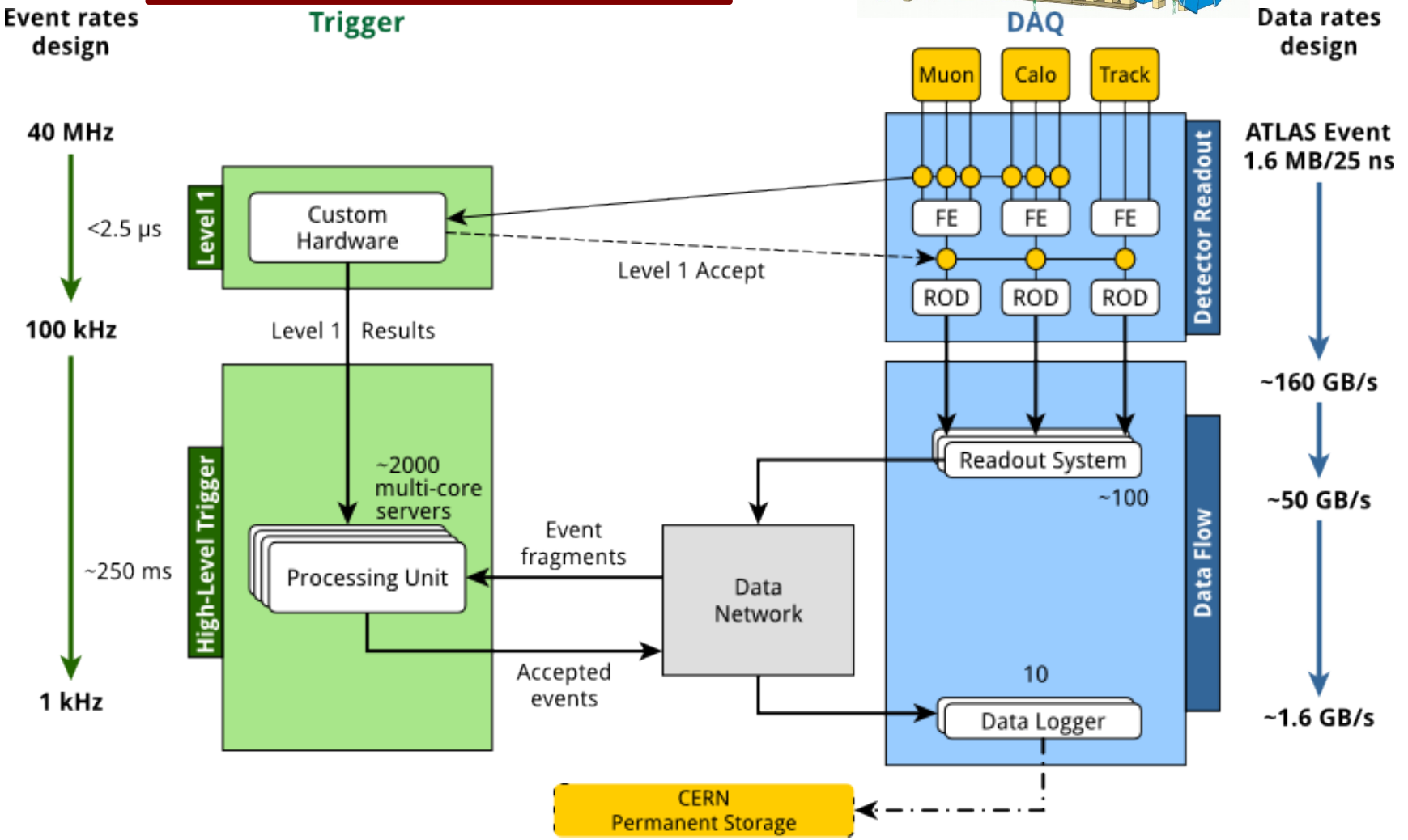
# Outline


- ATLAS Data-Acquisition and ReadOut System
- **ReadOut System Functions**
- **Requirements from Run2** DAQ evolution
- Third generation **ReadOut System design**
- Computer **Architectures and Performance**

# ATLAS TDAQ in Run2

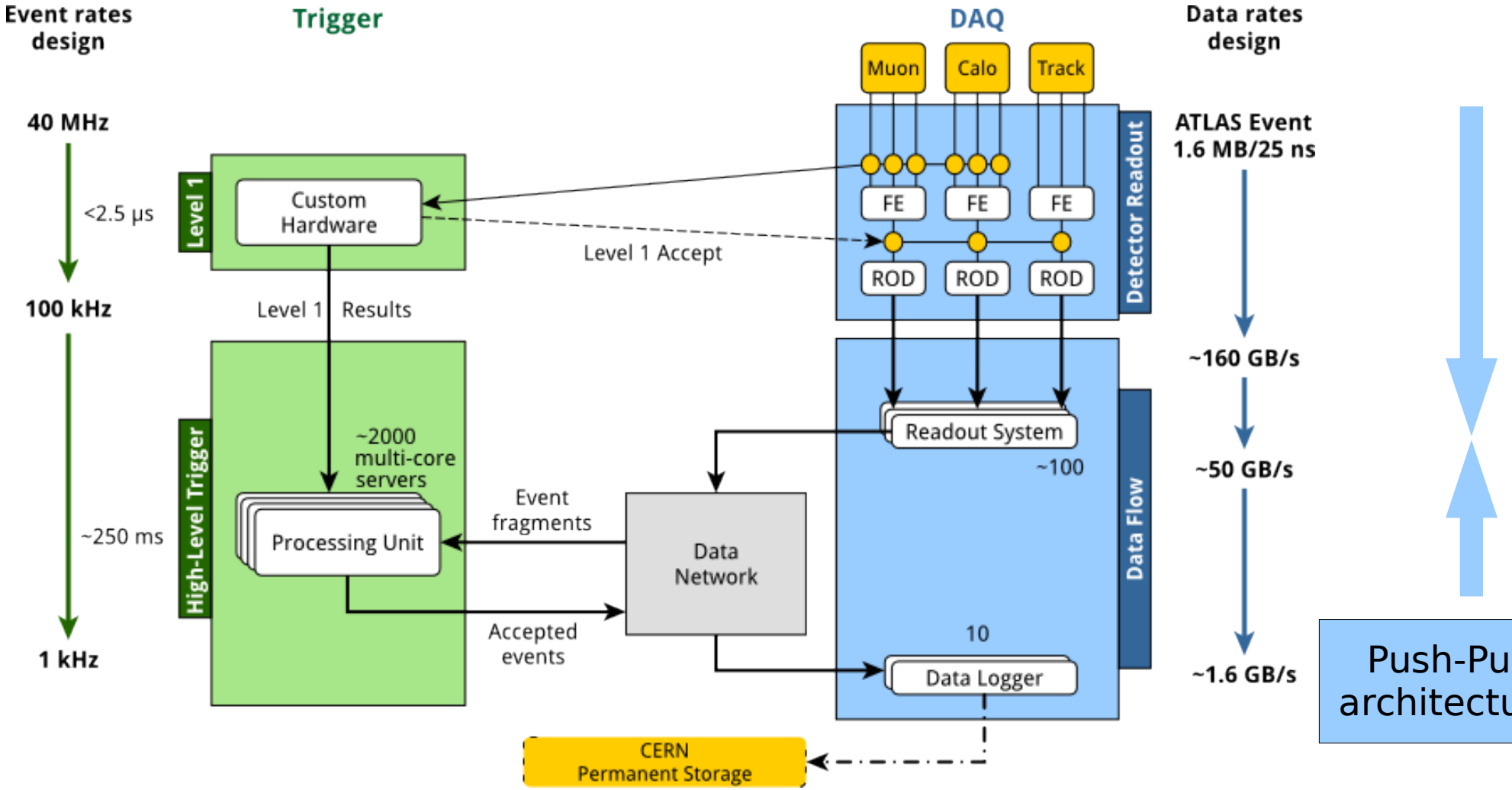
ATLAS@LHC - CERN
General purpose detector
Wide physics search goals
46m long, 22m high, 7000 tons
140M channels





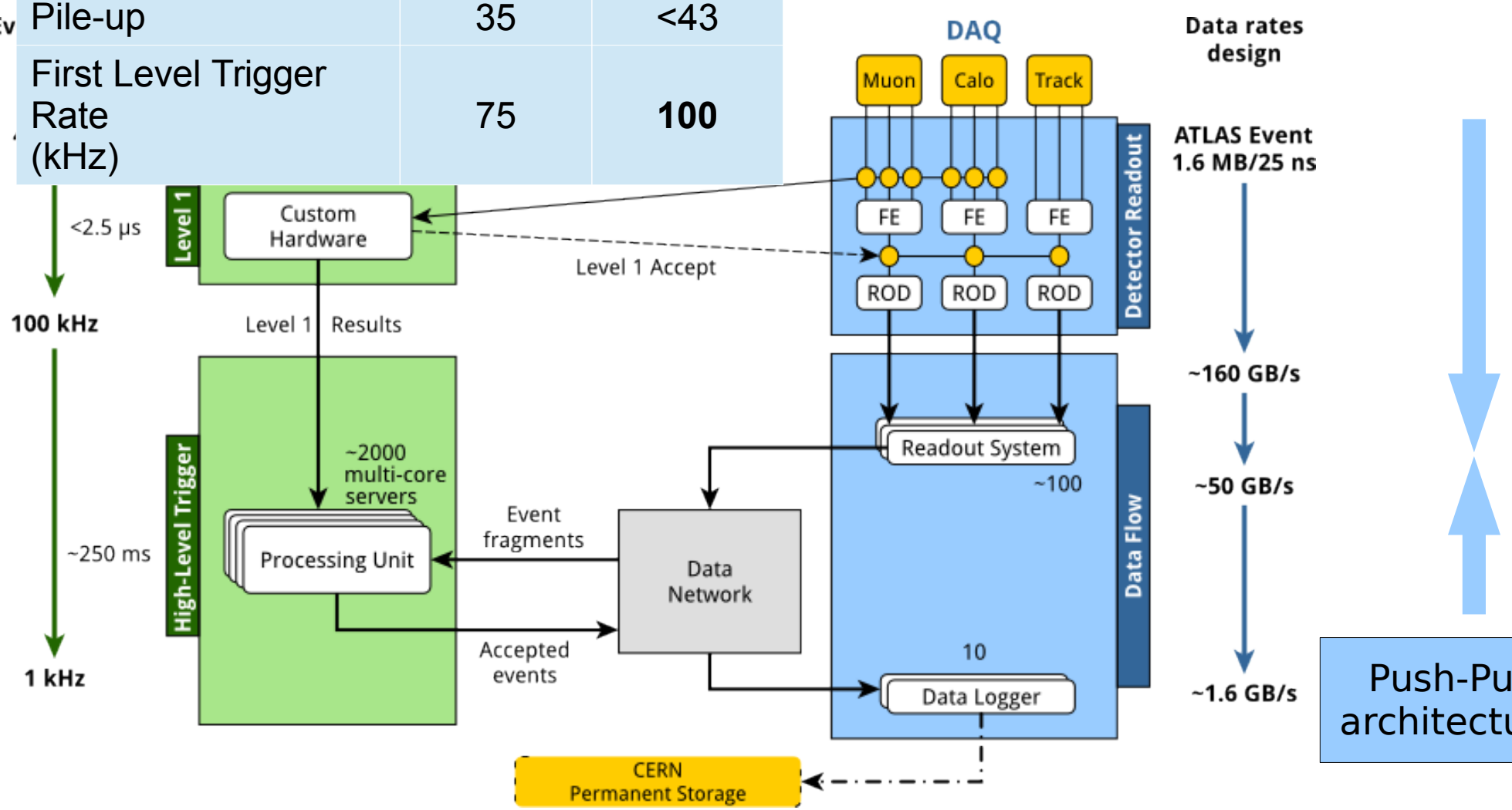
# ATLAS TDAQ in Run2

- **Incremental data-collection and processing** in the High-Level Trigger driven by Level1-tagged features (Region of Interest)
- **Event selection based on partial event data**: not all events are fully assembled

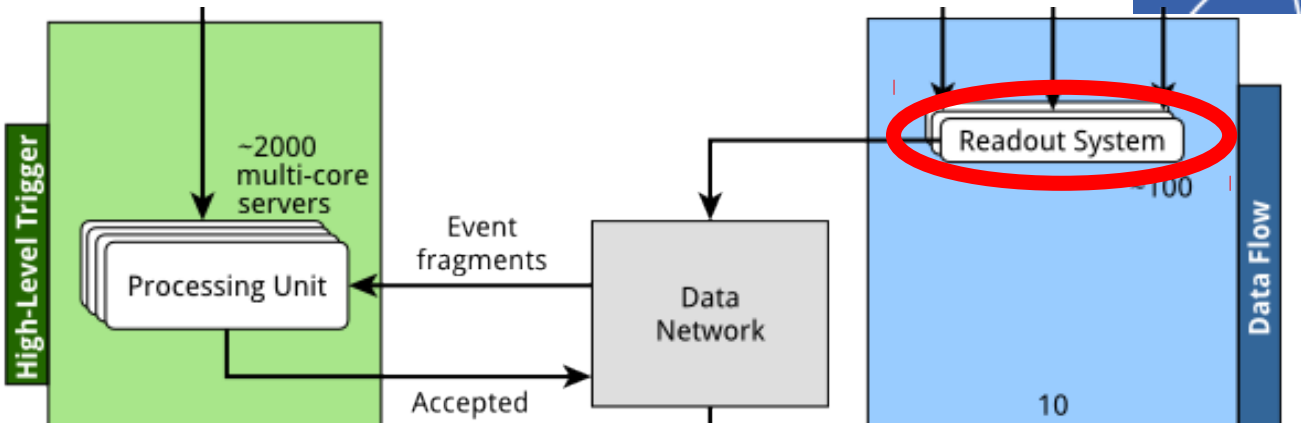


# ATLAS TDAQ in Run2

| | Run1 | Run2 |
|---|---|---|
| Luminosity ($10^{34}$ $cm^{-2}s^{-1}$) | 0.7 | **<1.6** |
| Pile-up | 35 | <43 |
| First Level Trigger Rate (kHz) | 75 | **100** |



# ReadOut System (ROS)



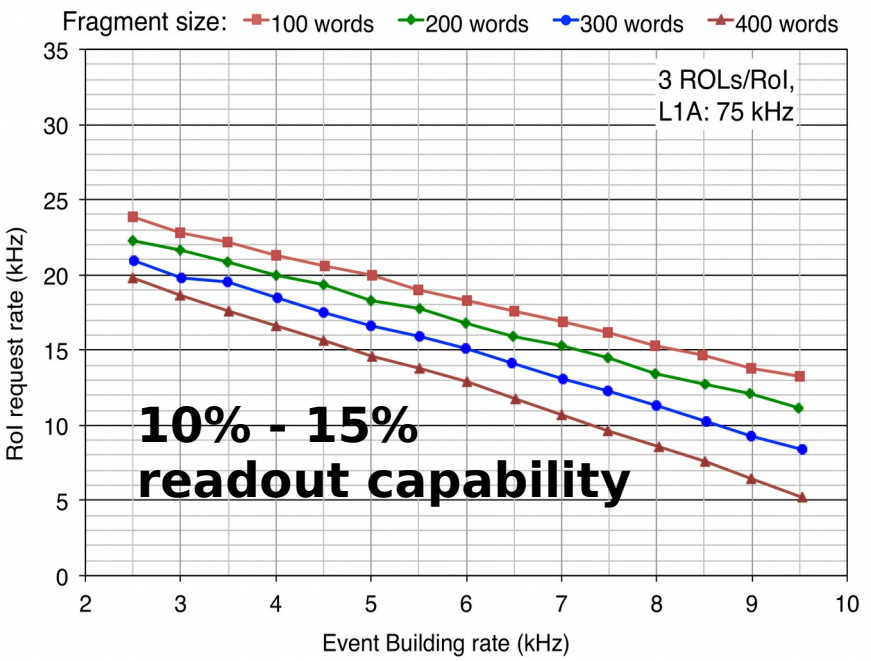

- Functions
  - Interfaces custom optical links (Readout Links - ROLs) from off-detector electronics
  - Buffers event data fragments until rejection or event building
  - Upon request, serves data fragments to High-Level Trigger processors over an Ethernet network

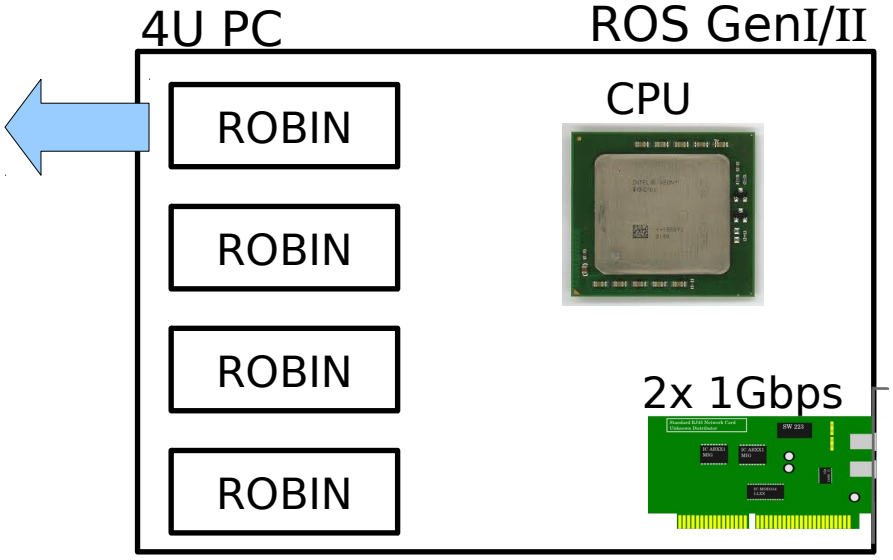
Heavy I/O workload with very limited computational requirements

| | Run1 |
|---|---|
| Number of ROLs | ~1600 |
| Number of ROS PCs | ~150 |
| Number of 1 Gbps ports | ~300 |

# ReadOut System in Run1

- 4U PCs equipped with
  - 2x 1Gbps Ethernet ports for data transfer
  - four (five) FPGA-based custom receiver and buffer cards (ROBIN)
- ROBIN
  - PCI interface (2.1 Gbps)
  - 3 optical inputs compatible with S-link (2 Gbps – nominal bandwidth 160 MB/s)
  - 64 MB/link buffer memory
  - on-board PPC processor for data and request management



# TDAQ evolution & ReadOut System

ReadOut System functions remain unchanged in Run2

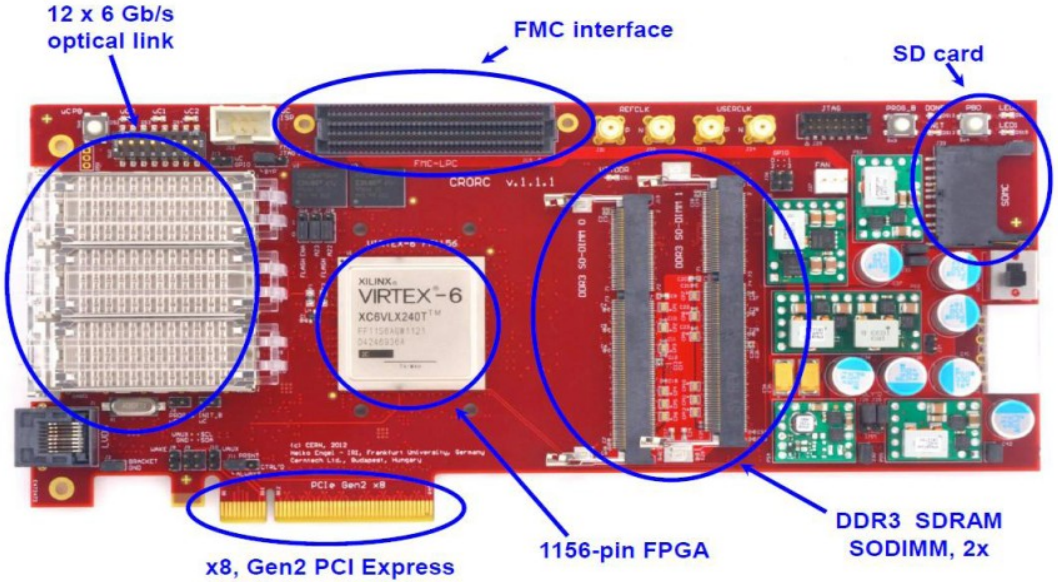
- Changes on the detector side reflected by increased number of ROLs (+16%) → **denser solution** in terms of ROL/rack space
- Denser solution implies higher throughput per node → **move from 1 Gbps copper to 10 Gbps optical** Ethernet
- PCI is ageing technology, not very common in current COTS → **prefer a PCIe**-based solution
- Size of memory buffer limits the average processing time and ultimately the HLT farm size → **larger buffer memory**
- Higher luminosity and higher first-level trigger rate → **capable of 50% readout**
- Future compatibility with newer generations of faster ROL → **new, faster optical receivers**

**ROS upgrade project** → GenIII

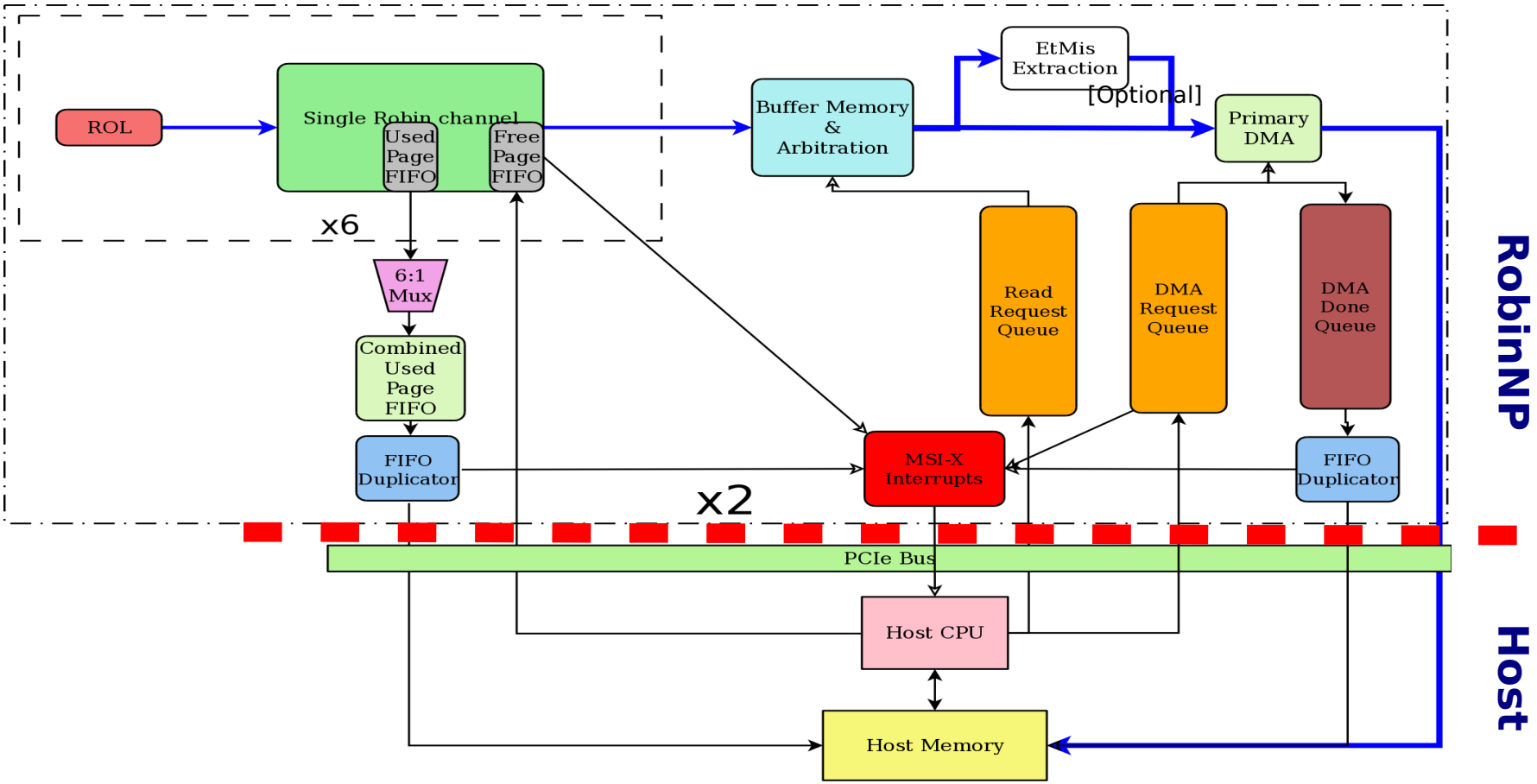
# C-RORC & RobinNP

- ALICE Common ReadOut Receiver Card (C-RORC)
  - 3 QSFP: 12 serial optical links, up to 6 Gbps each
  - Xilinx Virtex-6 FPGA
  - 4-8 GB on-board SODIMM RAM (up to DDR3-1066)
  - PCIe interface, up to Gen2 x8 (32 Gbps)

- RobinNP (No Processor)
  - ATLAS-specific firmware for C-RORC cards
  - based upon Robin firmware, but offload tasks to the host CPU
    - direct access to future CPU performance improvements

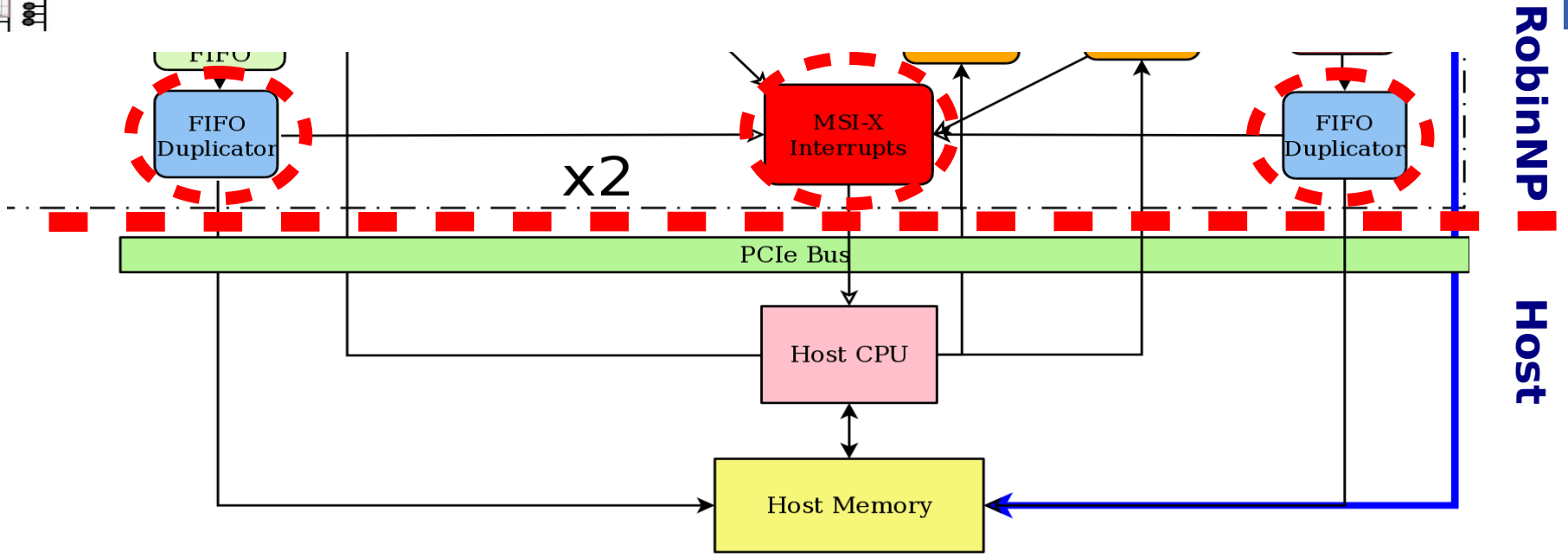


# RobinNP Firmware and Software



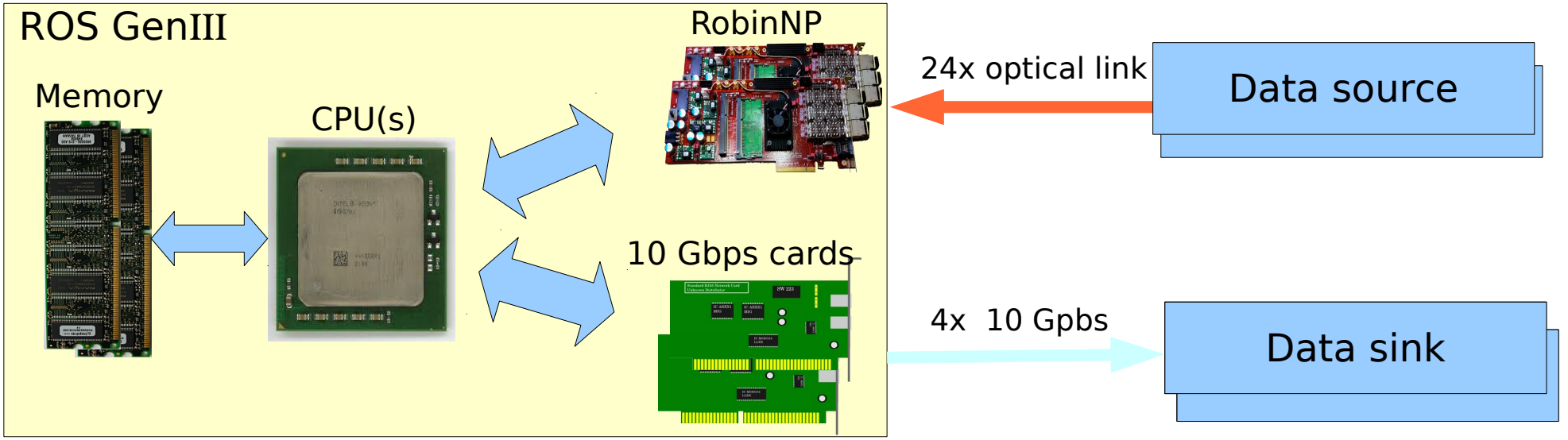

- Key innovations at the **boundary between the RobinNP and the host**

# RobinNP Firmware and Software




- **FIFO duplicator**: automatic transfer of hardware FIFO contents into host ring buffers
  - avoid high latency paths using PCIe write cycles only
- **MSI-X interrupts**: actively inform host software of DMA completion → blocking calls instead of polling
  - coalescing scheme to limit interrupt rates
- New multi-threaded SW stack
  - keeps track of incoming data fragments and handles data requests instantiating DMA transactions
  - interfaces with the TDAQ network communication library

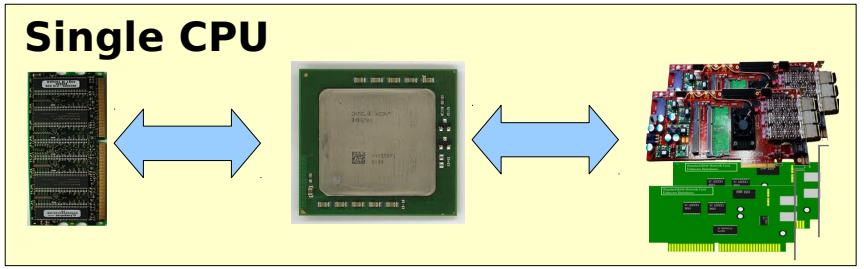
# ROS GenIII Architecture

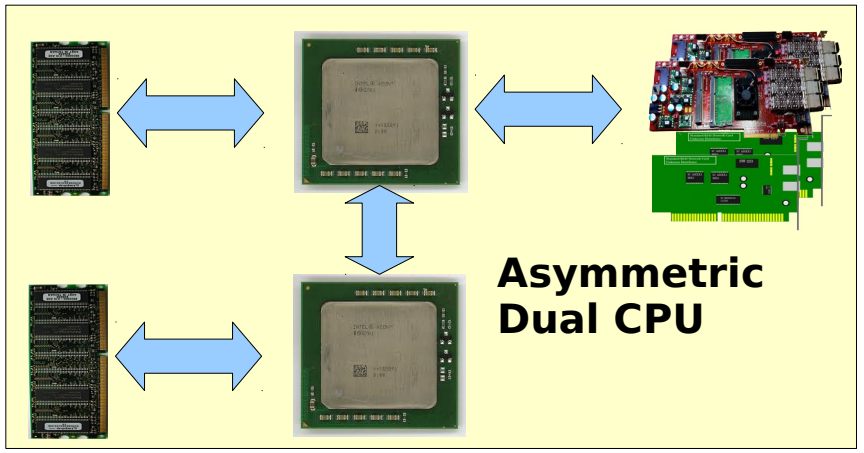


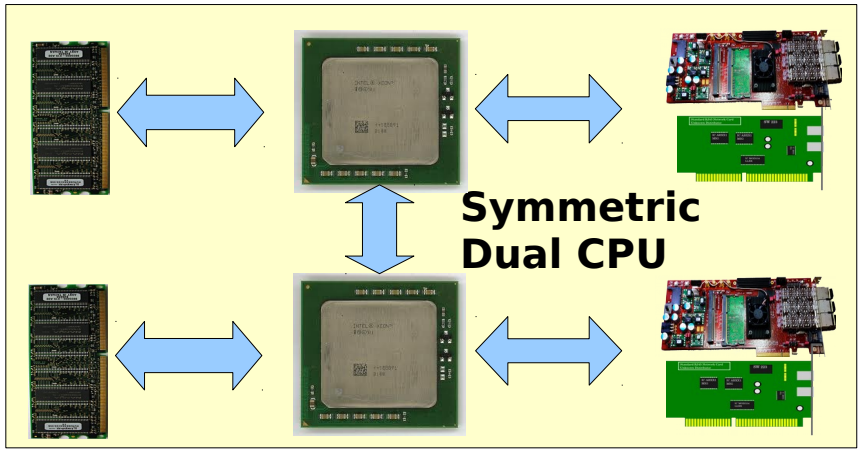

- Baseline ROS GenIII configuration includes
  - 2x RobinNP → 24 input optical links
  - 4x 10 Gbps Ethernet ports
    - redundancy
- Performance and development studies in lab setup
  - with/without optical data sources
    - RobinNP can internally generate test data
  - data-sink units equipped with 10 Gbps connectivity

At 50% readout → up to **~16 Gbps**
- RobinNPs to memory
- memory to network cards

# Computer Architecture



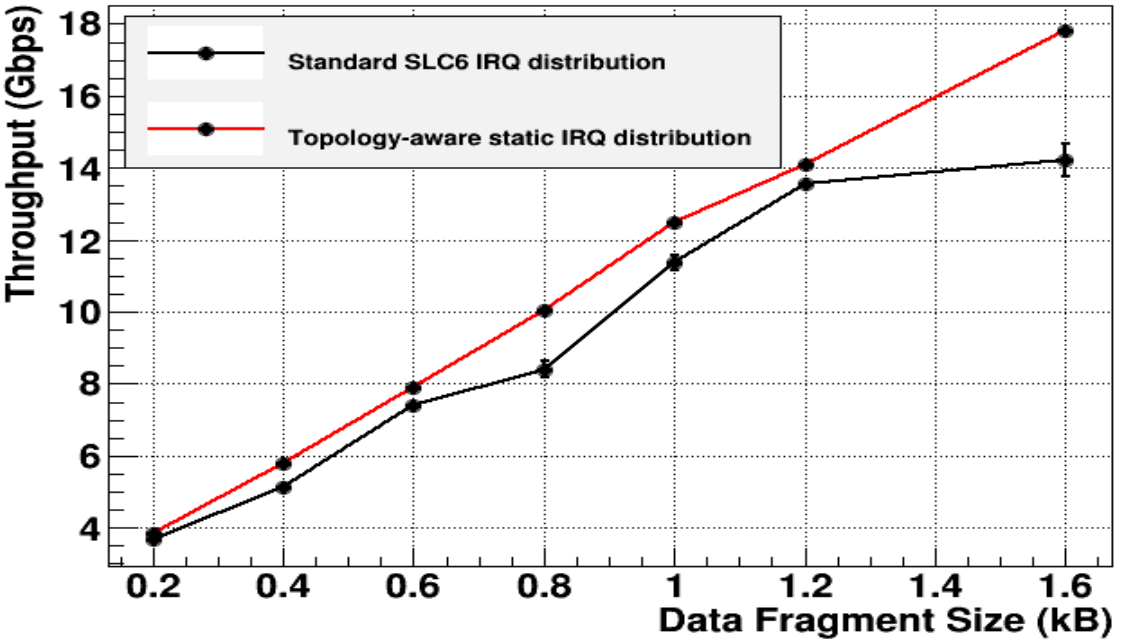



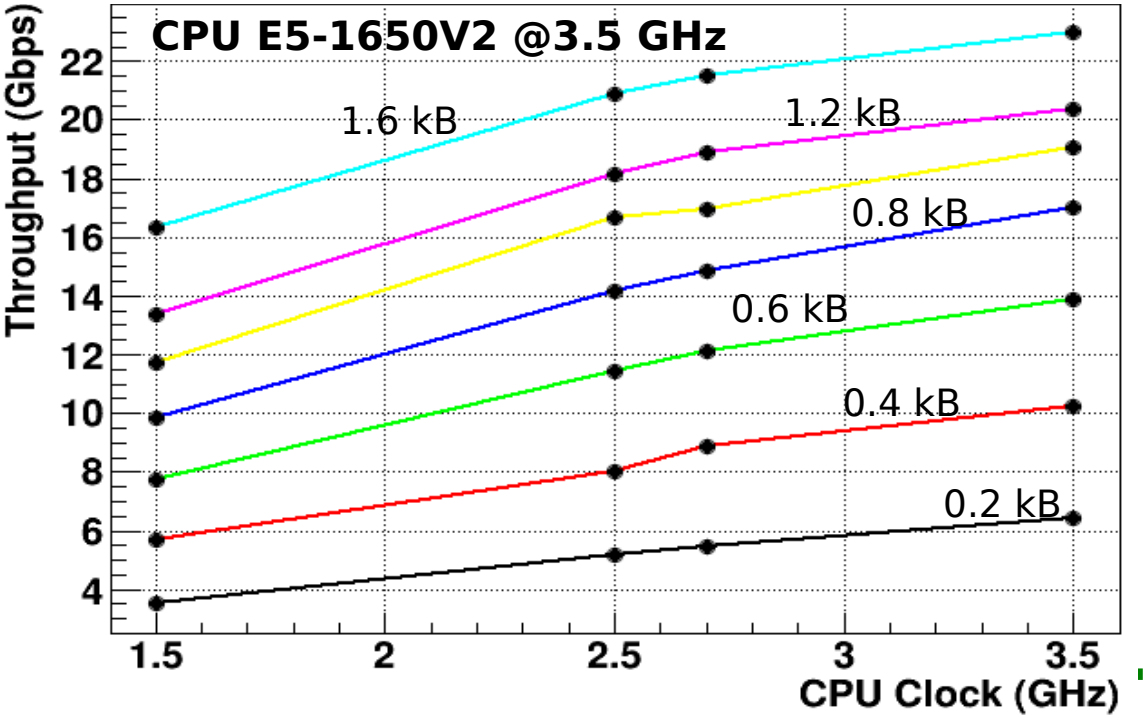




- Modern CPUs embed controllers for memory and PCIe
- Multi-CPU server architectures have non-uniform access patterns
  - memory banks per CPU package (NUMA)
  - PCIe lanes per CPU package (NUIOA)
- What is the best configuration for a ROS GenIII PC?
  - how modern PC architectures deal with heavy I/O workload?
- Possible configurations are limited by the available commercial motherboards
- Single socket CPUs typically are cheaper and reach higher clocks than similar dual-socket CPUs

# Interrupts and CPU Clock Speed

- Scientific Linux 6 (SLC6) uses a **interrupt balancing daemon***
  - dynamically associates interrupt lines to core based on heuristics
- Discouraged by most 10 Gbps card manufacturers
  - suggest **topology-aware** static interrupt distribution
- Observed balancing daemon leading to inferior, unstable, poorly reproducible results
- Performance **almost linear** CPU clock dependency
  - probed using dynamic frequency scaling
  - ~linearity allows to compare result from different PCs

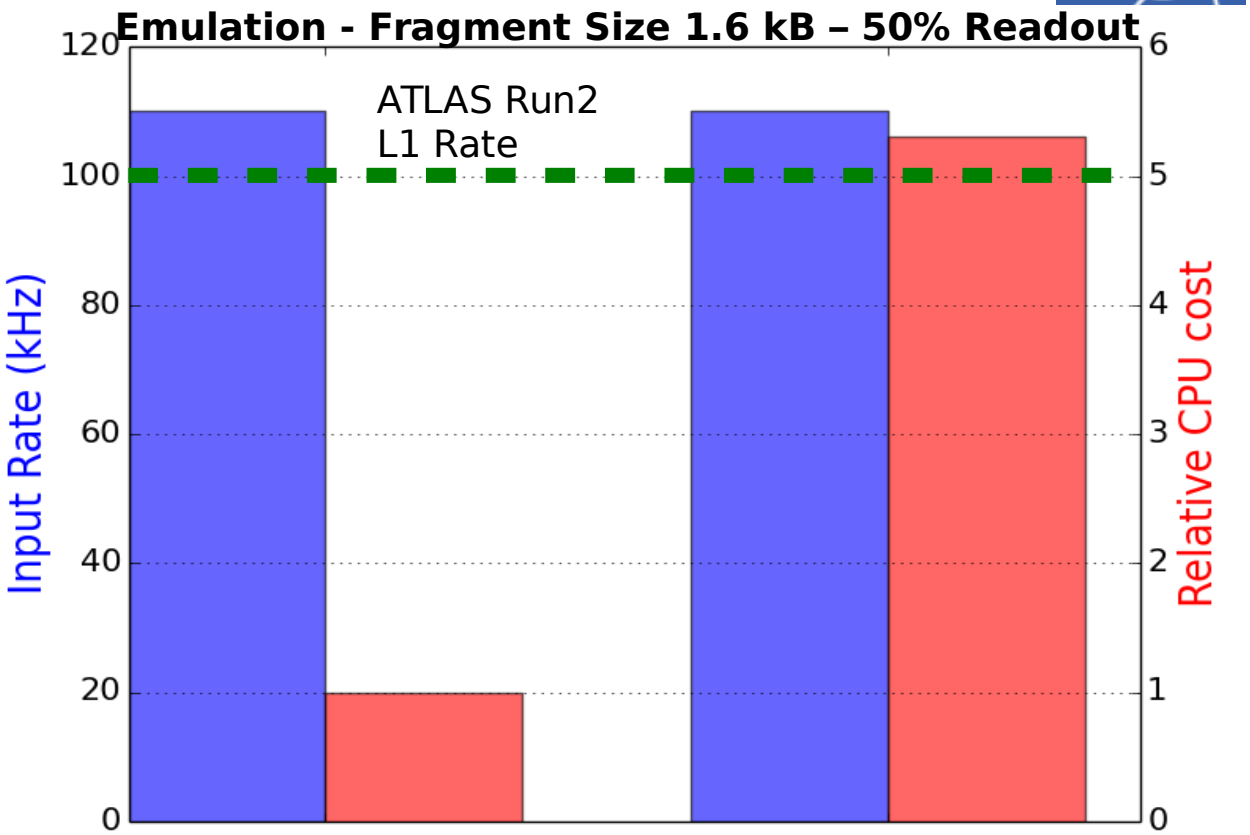


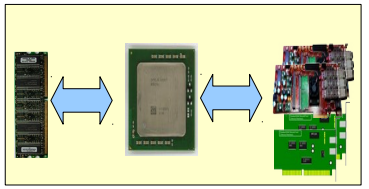




* https://github.com/Irqbalance/irqbalance

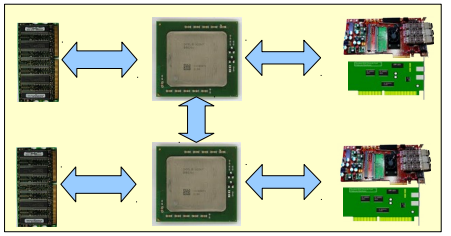
# Performance vs Architecture

- **A single CPU configuration can satisfy the performance requirements**
- A symmetrical dual-CPU configuration, with similar CPU figures, does not provide significant improvements
  - operating a single application
  - potential advantages outweighed by cost difference



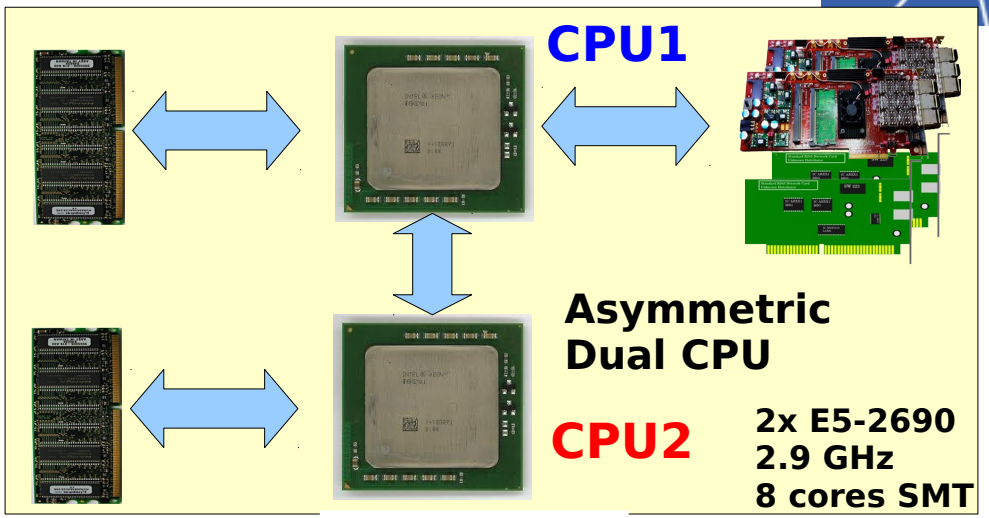
**1x E5-1650V2**
**3.5 GHz**
**6 cores SMT**

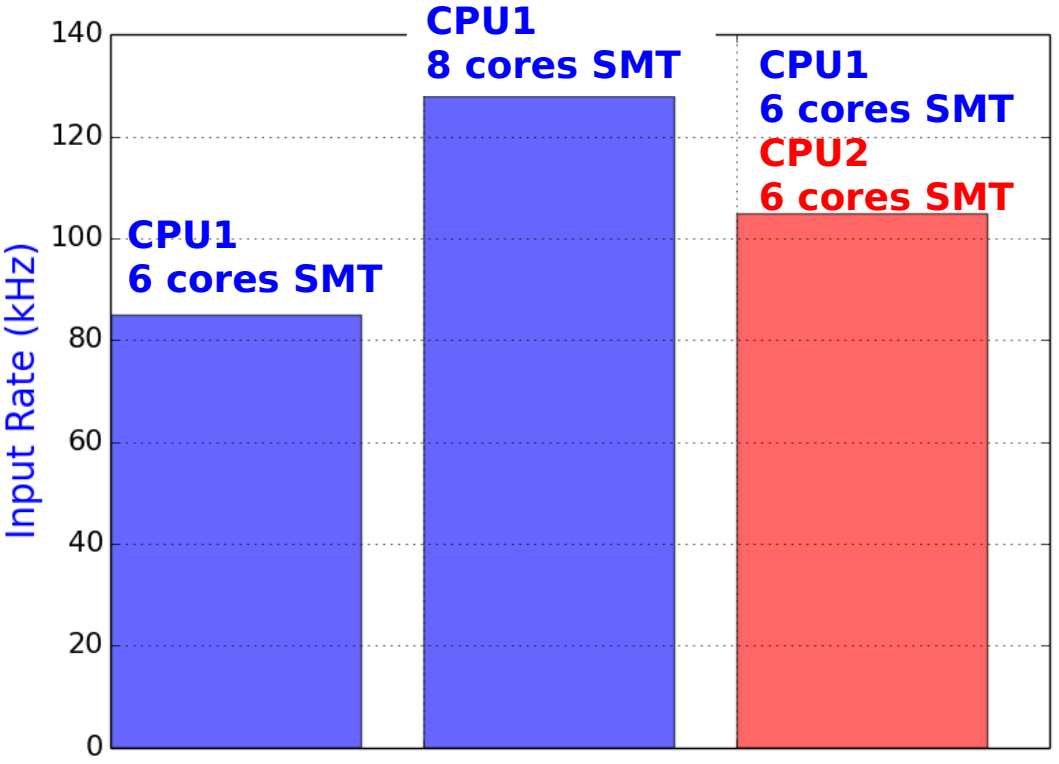


**2x E5-2643**
**3.3 GHz**
**4 cores SMT**

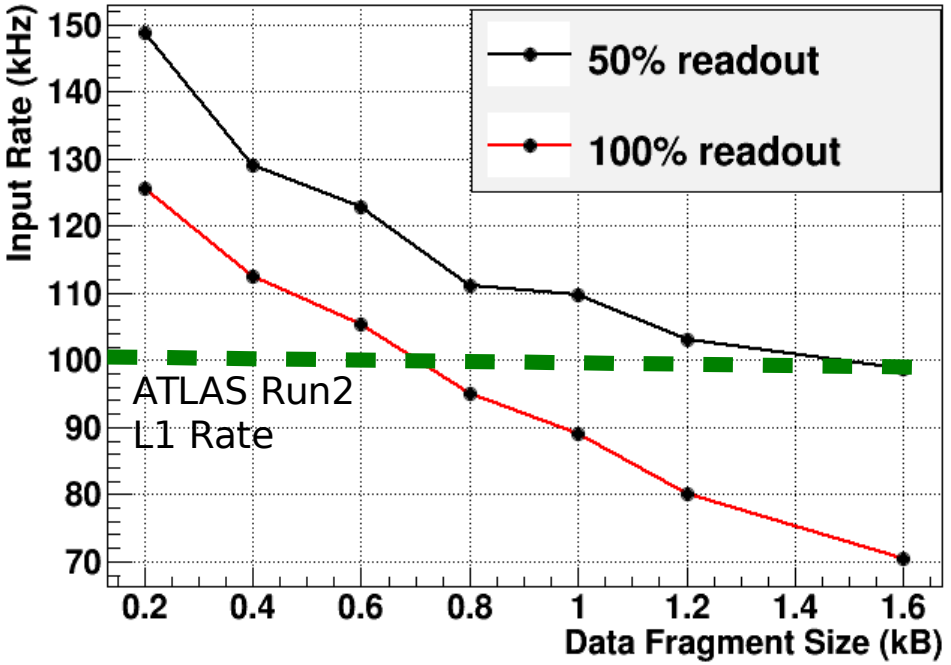
# Performance vs Architecture

- Linux CPU hotplug allows to enable/disable individual cores at run time
- Study in-situ the benefits of a second satellite CPU
  - as well as performance versus number of cores
- For different workloads
  - performances strongly depend on the number of cores enabled in CPU1
  - enabling extra cores in CPU2 yields marginal improvements
    - performance loss in some cases



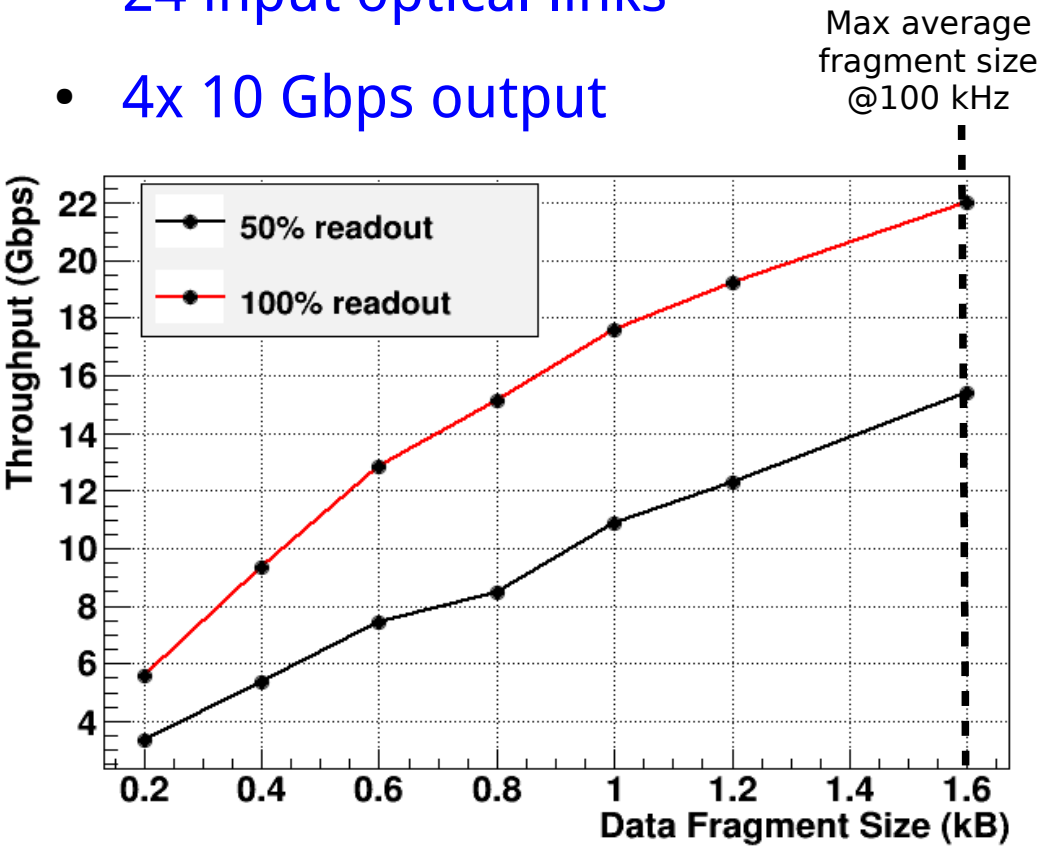



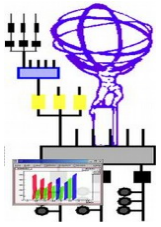
# Baseline Performance of ROS GenIII




- ROS firmware and software exceed ATLAS Run2 requirements
  - ongoing optimizations will yield performance improvements
- Full readout (100%) possible for small fragments or fewer channels
  - beneficial for specific (trigger) detectors

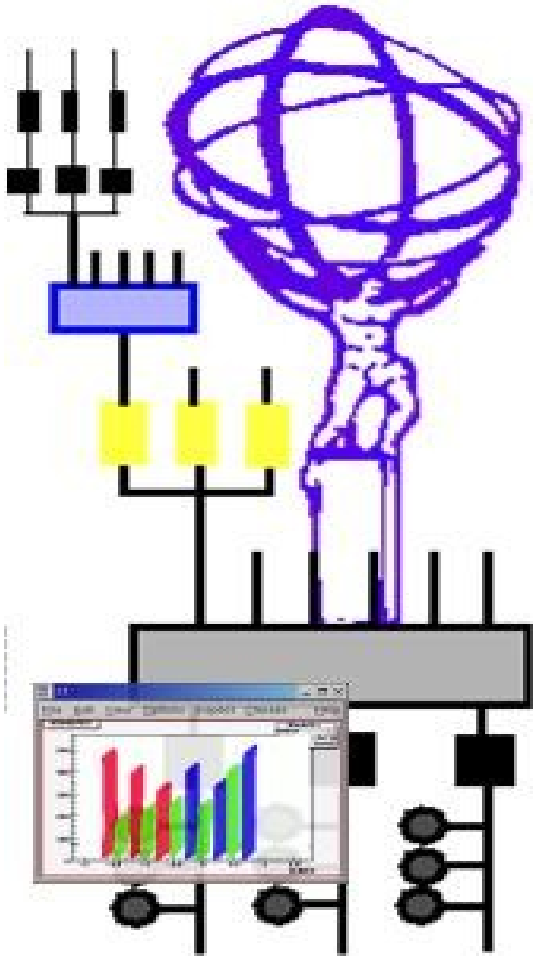
- ROS host configuration
  - Single CPU
    - Intel E5-1650V2 @3.5 GHz
    - 6 cores SMT
- 24 input optical links
- 4x 10 Gbps output





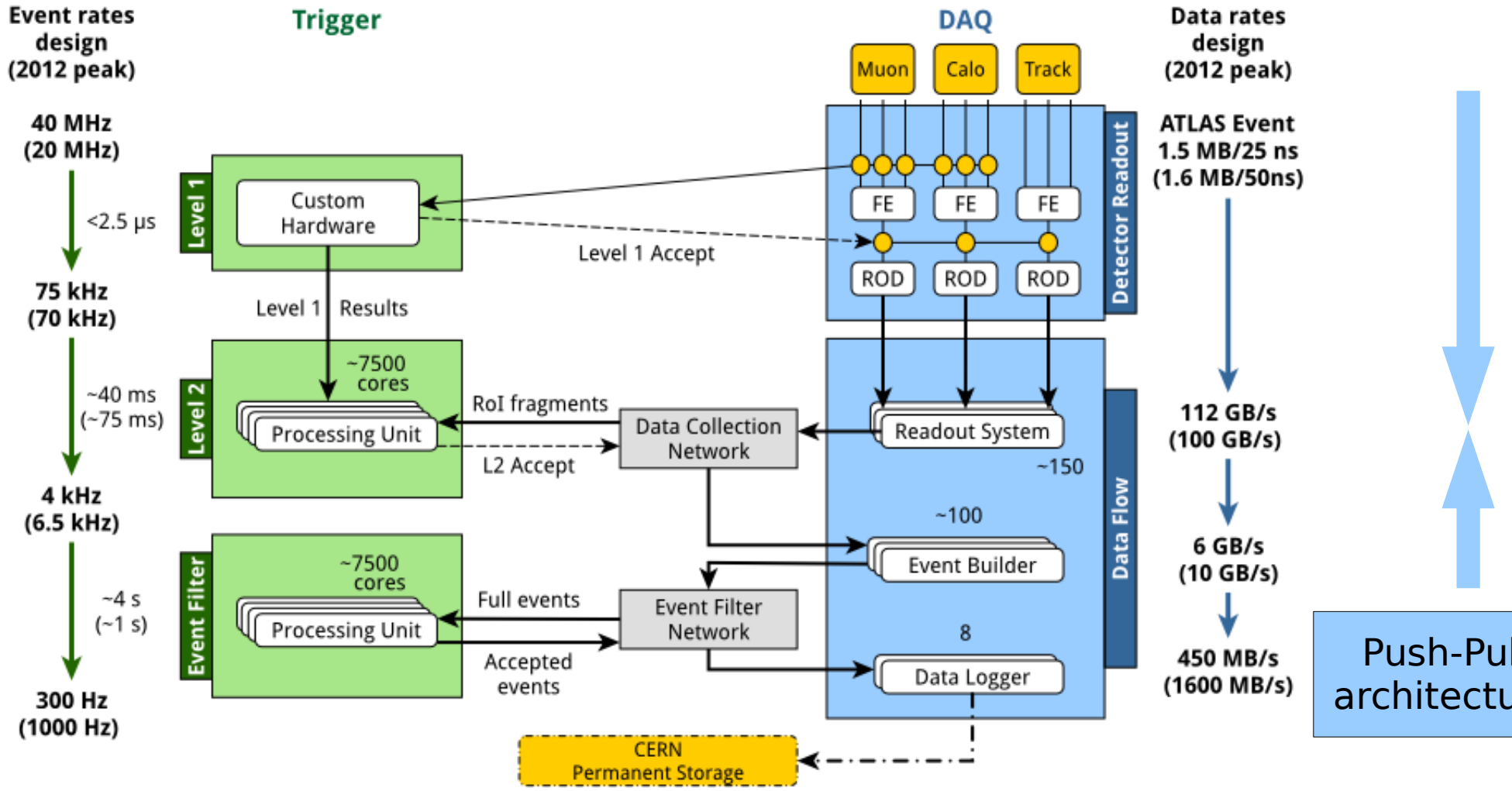
# Outlook

- ATLAS ReadOut System Generation III (r)evolution is shifting from development to deployment phase
  - same functions as in Run1
  - new technology landscape at every level: optical interfaces, networking, firmware and software architectures, ...
- Current development firmware and software already exceed requirements
  - more than a **factor three performance** improvement with respect to the present ROS: **10-15%** readout fraction → **>50%** readout fraction
    - **factor six performance** improvement for the individual node
- A single CPU solution is sufficient for the needed performance goals
- Gained insights on handling and optimizing heavy I/O workload on non-uniform computer architecture
- **CPU hotplugging and power management** are **powerful tools** for performance and bottleneck investigations



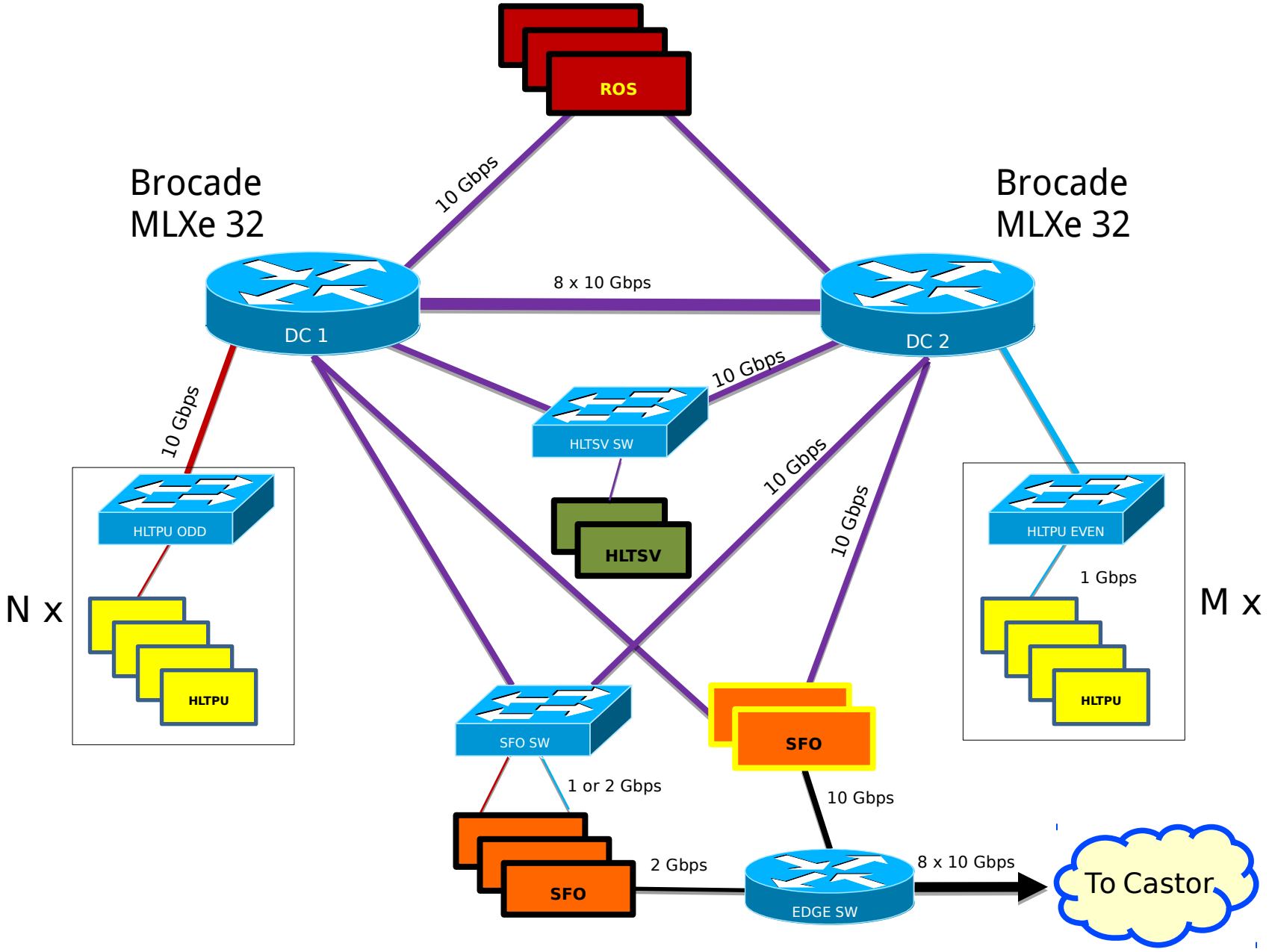
# Bonus

# ATLAS TDAQ in Run1

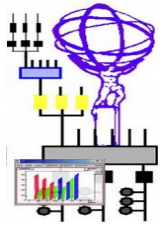



Second Level Trigger (L2) implemented in SW: partial event reconstruction driven by L1-tagged features.
Followed by explicit Event Building step, decoupling L2 and Event Filter

# ATLAS TDAQ Data Network in Run2



# Software Architecture




Descriptor Flow
Page Flow
DMA & DMA Control
HLT Traffic and Data Collection

**RobinNPReadoutModule**

Request Processor Thread
(1 per ROBgroup)

***Queue DMA***

RobinNP Hardware

***Memory Pages (FIFO + Interrupts)***

***Request Queue (Semaphore)***

***Receive DMA (FIFO + Interrupts)***

DMA Monitoring Thread
(1 per ROBgroup)

RobinNP Indexer Thread
(1 per ROBgroup)

***Processed Descriptor Queue (Semaphore)***

Async IO Threads

Data Collector Thread

***HLT Traffic***

***Free Descriptor Queue (Semaphore)***

**TriggerIn**

# Effect of IRQ affinity

Interrupt rate per core for different data fragment sizes with:
- standard SLC6 interrupt distribution
- manually tweaked interrupt affinity

In the later case, the interrupt load is more uniformly distributed → better, stabler and reproducible perfomance

**CPU 2x E-2643 @3.3 GHz**