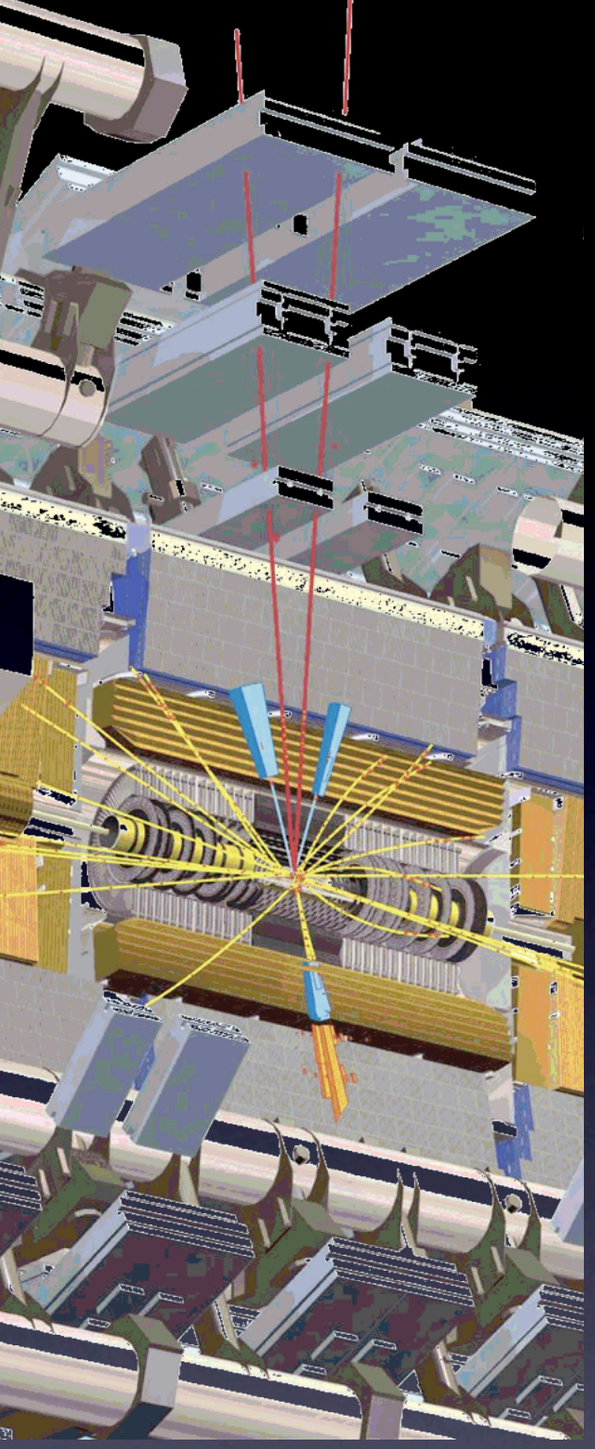


The ATLAS DataFlow System: Present Implementation, Performance and Future Evolution

G. Crone[2], D. Della Volpe[3,4] , B. Gorini[4], B. Green[1], M. Joos[4], G. Kieft[5], K. Kordas[8], A. Kugel[6], A. Misiejuk[1], N. Schroer[6], P. Teixeira-Dias[1], L. Tremblet[4], W. Vandelli[4] , J. Vermeulen[5], P. Werner[4], F. Wickens[7]

[1]Royal Holloway University of London, [2]University College London, [3]Università di Napoli, [4]CERN, [5]Nikhef, Amsterdam, [6]Ruprecht-Karls-Universitaet Heidelberg, [7]Rutherford Appleton Laboratory, [8]University Bern

D. della Volpe

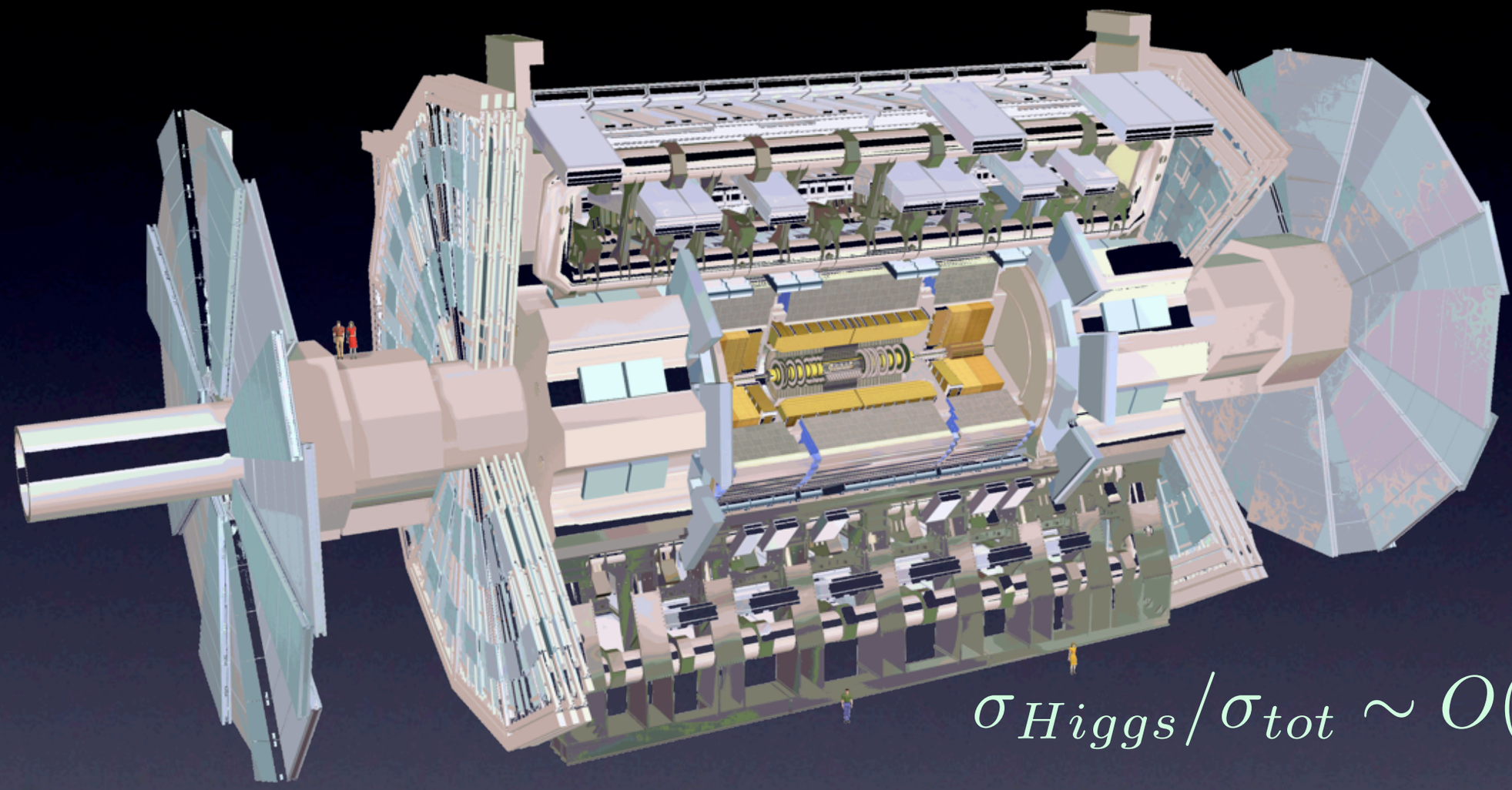
†Università "FEDERICO II" di Napoli & CERN - Physics Department/ADT

RT '09 Bejing -May 2009

- This talk is focused on the Read Out System of the ATLAS DataFlow.
- It is not intended to be a complete and exhaustive overview of the DataFlow and its performance.
- The TDAQ system and its performance are presented in the talk by J. Zhang at this conference
  (TDA2-4, ATLAS DAQ and Control)
- Here I will mention just what is relevant for the talk .

At full LHC Luminosity, the ATLAS detector can produce about

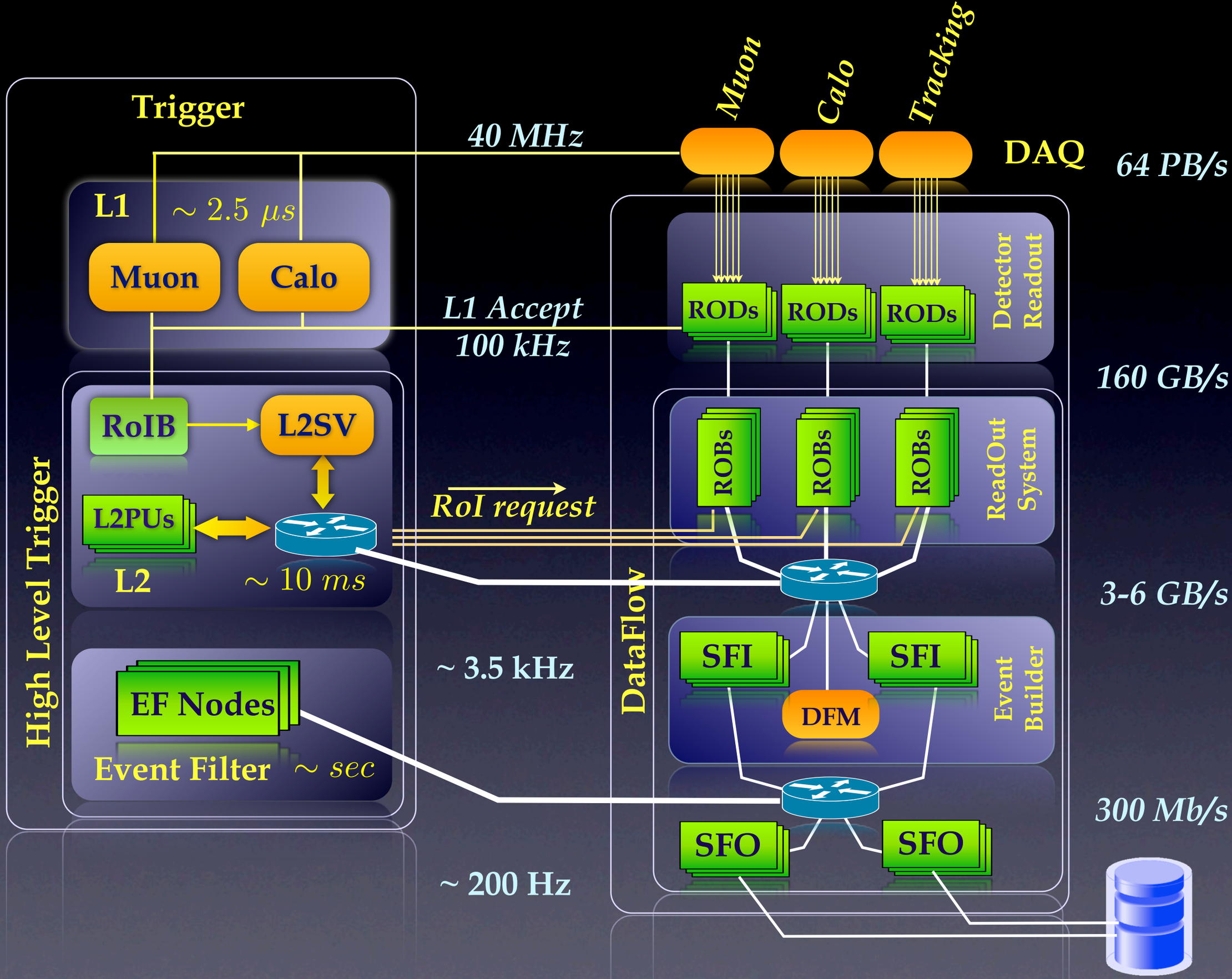
1.6 MB/event *40 MHz = 64 PB/s

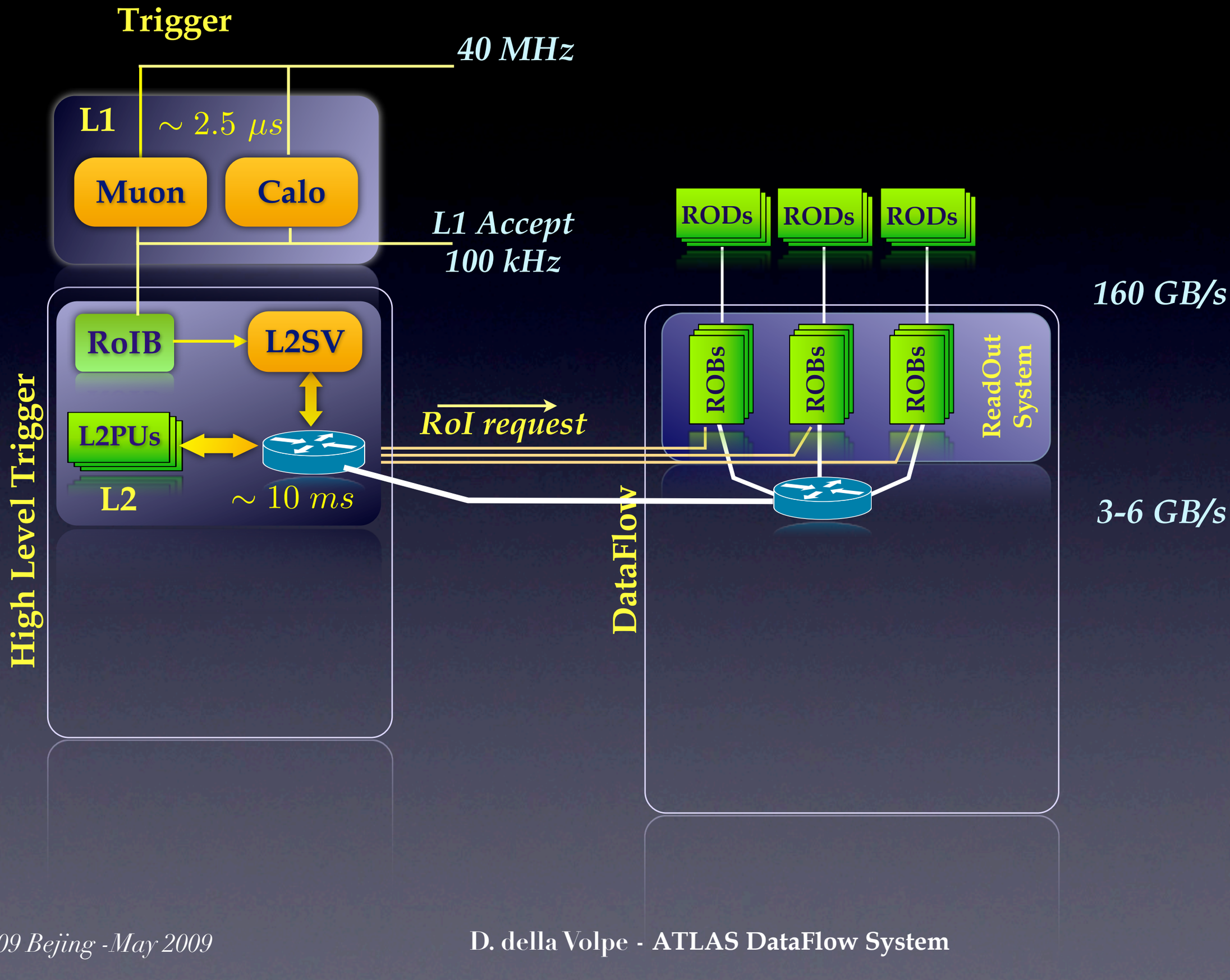


$$\sigma_{Higgs}/\sigma_{tot} \sim O(10^{-13})$$

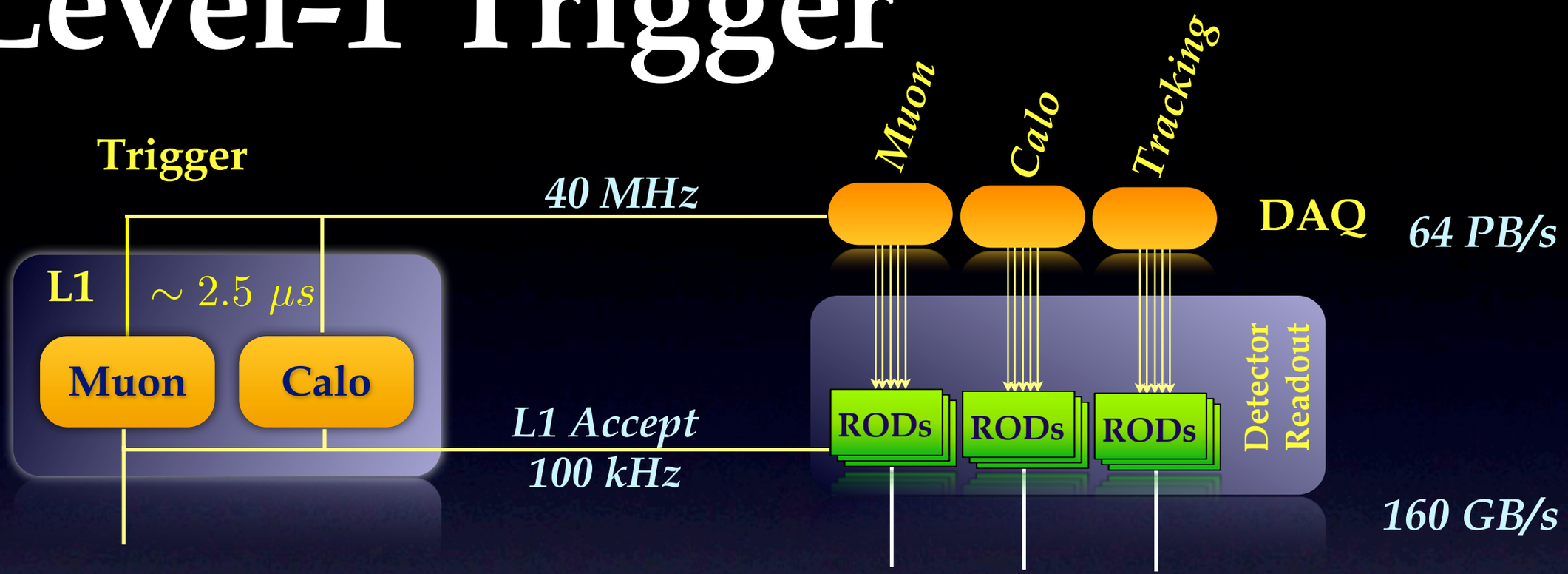
To reduce the amount of data to write to disk and of unecessary data transfer

- *Three -level Trigger Architecture*
- *"Region of Interest" mechanism*


Trigger
40 MHz
Muon
Calo
Tracking
DAQ
64 PB/s
L1
~ 2.5 μs
Muon
Calo
L1 Accept
100 kHz
RODs
RODs
RODs
Detector Readout
160 GB/s
RoIB
L2SV
ROBs
ROBs
ROBs
ReadOut System
RoI request
L2PUs
L2
~ 10 ms
3-6 GB/s
SFI
SFI
DFM
Event Builder
~ 3.5 kHz
DataFlow
High Level Trigger
EF Nodes
Event Filter
~ sec
300 Mb/s
SFO
SFO
~ 200 Hz


Trigger
40 MHz
L1
~ 2.5 μs
Muon
Calo
L1 Accept
100 kHz
RoIB
L2SV
L2PUs
L2
~ 10 ms
High Level Trigger
RoI request
RODs
RODs
RODs
ROBs
ROBs
ROBs
ReadOut System
160 GB/s
DataFlow
3-6 GB/s

Level-1 Trigger



- Hardware based
- It has a maximum latency of 2.5 μs
- It has 3 main components: Calorimeter Trigger, Muon Trigger and Central Trigger
- The trigger deals with detector information from coarse region in the eta-phi plane (Region of interest (RoI))

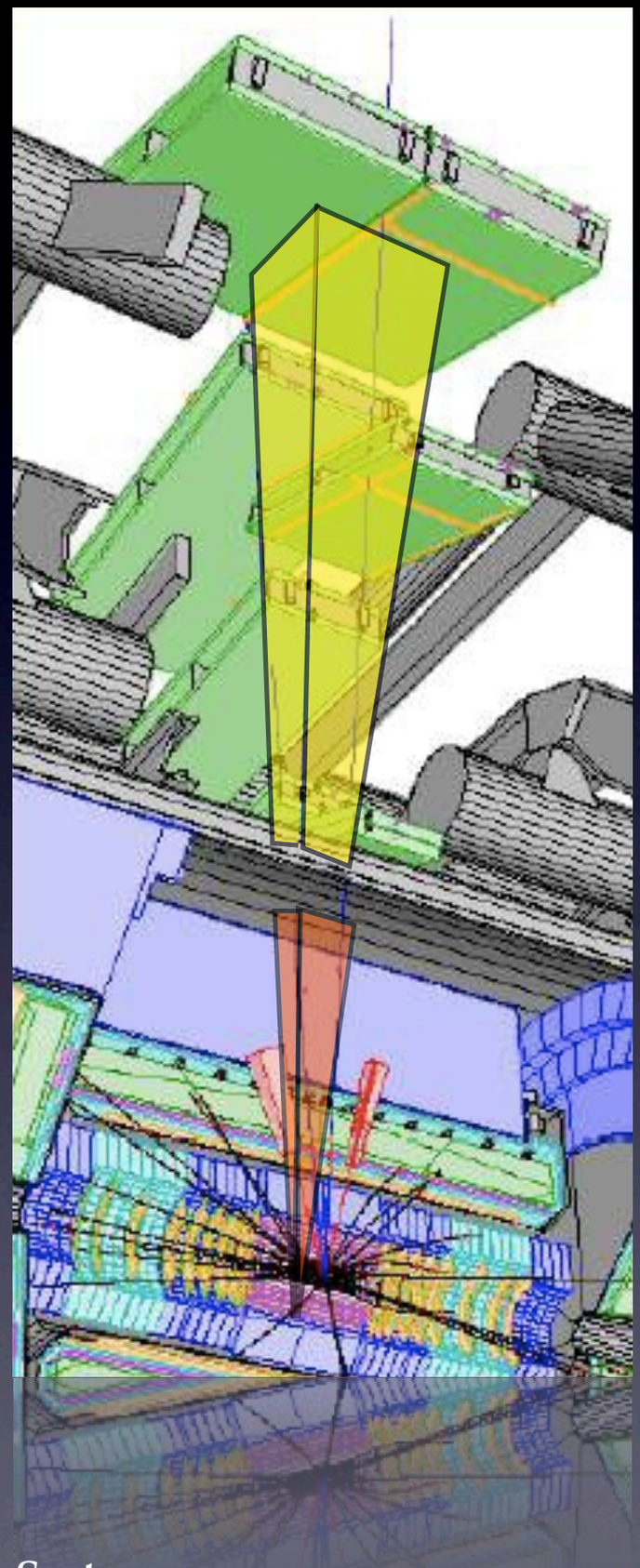
The RoI

- The Level 1 Trigger elements (jet, electromagnetic, muon candidate, etc.) determine Regions of Interest (RoIs) that seed further trigger decisions

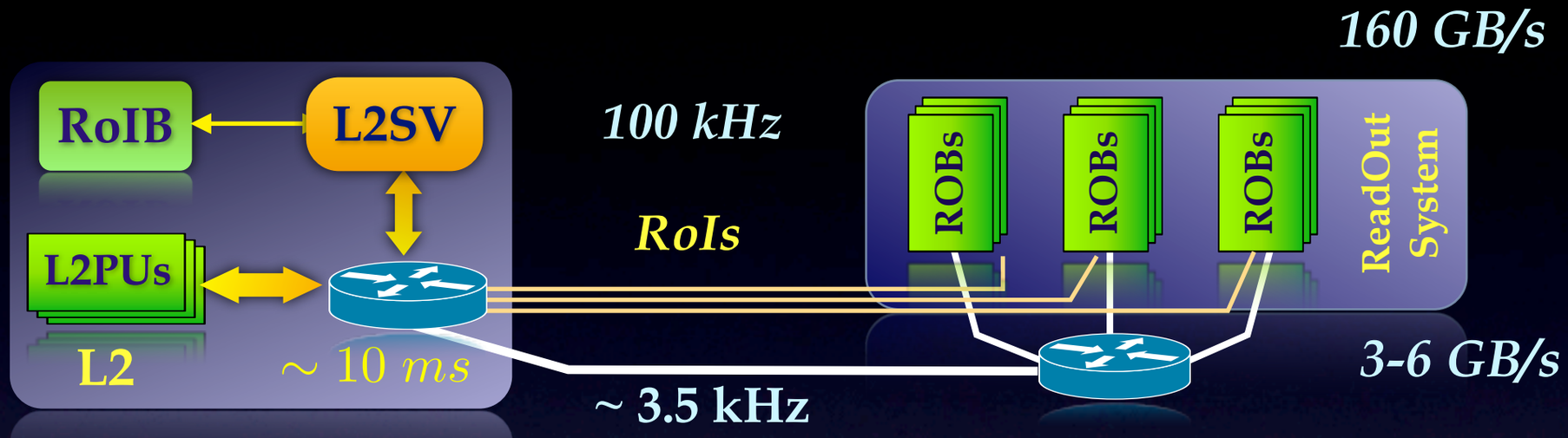
- A typically RoI size is $0.1\ \Delta\eta \times\ 0.1\ \Delta\phi$ (larger for jets)
  Based on coarse, fast information

- Identify Regions based on local areas of activity at L1 and pass on to HLT

- Only RoI data readout to first part of HLT ⇒ Reduces stress on DataFlow



The Level 2 Trigger



- The data are buffered into the ReadOut System
- The L2 requests RoI data from ROS and refines the L1 result
  - only 1-2 % of data are collected by L2
  - relaxes request on ROS system and network bandwidth

RoI-based L2 trigger ⟹ much smaller ReadOut network ⟹ higher control traffic

The ROS PC

The ReadOut System (ROS) PC

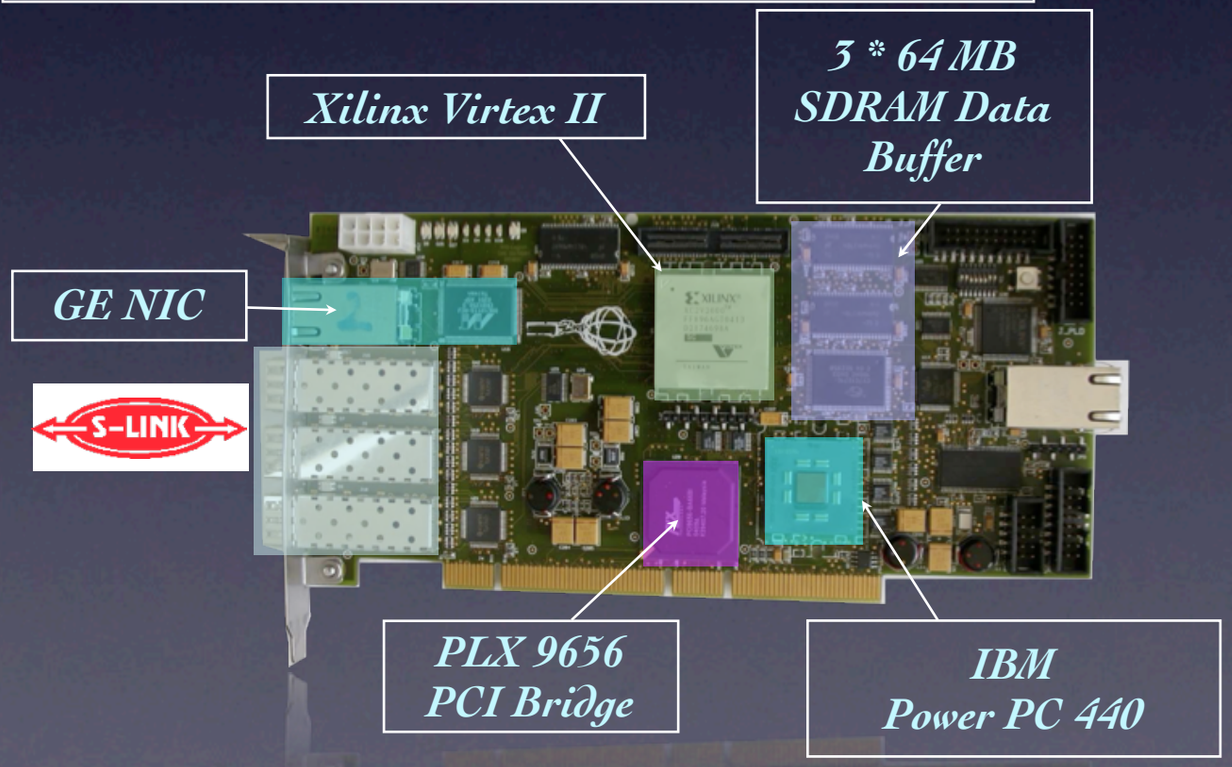
Houses 1 to 5 Robin cards (typically 4 cards)

Configures and controls the Robins

Reads data from the Robins and provides it to the Second-Level Trigger and to the Event Builder

Receives clear requests for event fragments and forwards them to the Robins

Interfaces to the operational and physics monitoring systems



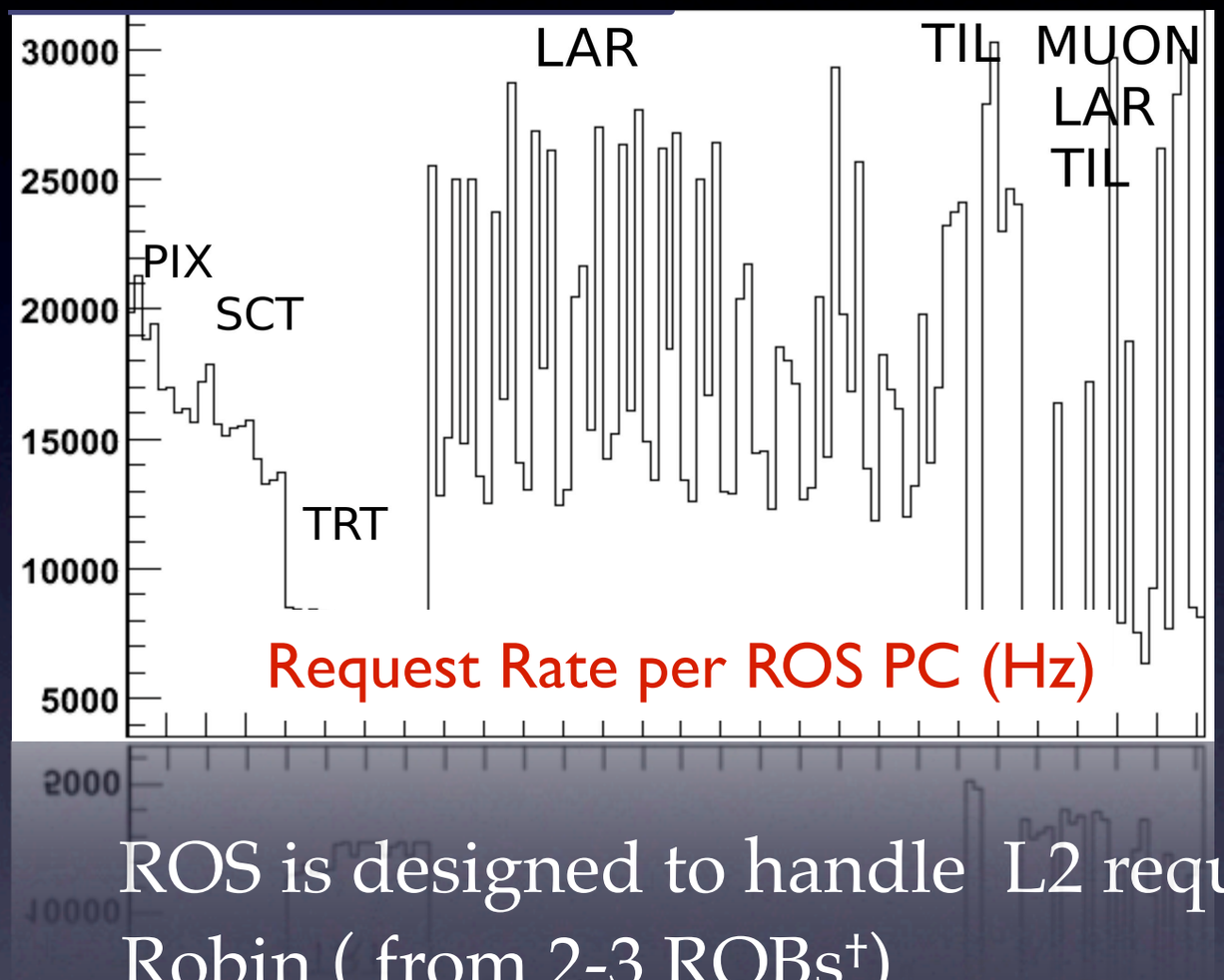
The Robin PCI card

Receives event fragments from sub-detector specific front-end electronics (RODs) via 3 optical links (ROL) (1 ROL= 1 ROB)

Buffers events during the decision latency of the Second-Level Trigger and the time required for building of events accepted by L2

The optical link is based on the S-LINK interface: 32 bit @ 40MHz = 160 MB/s

Readout Performance



Notice: small event size (800kB)

~2GB/s ROS → L2

Testing 10^{31} trigger menu with simulated data uploaded into the ROS

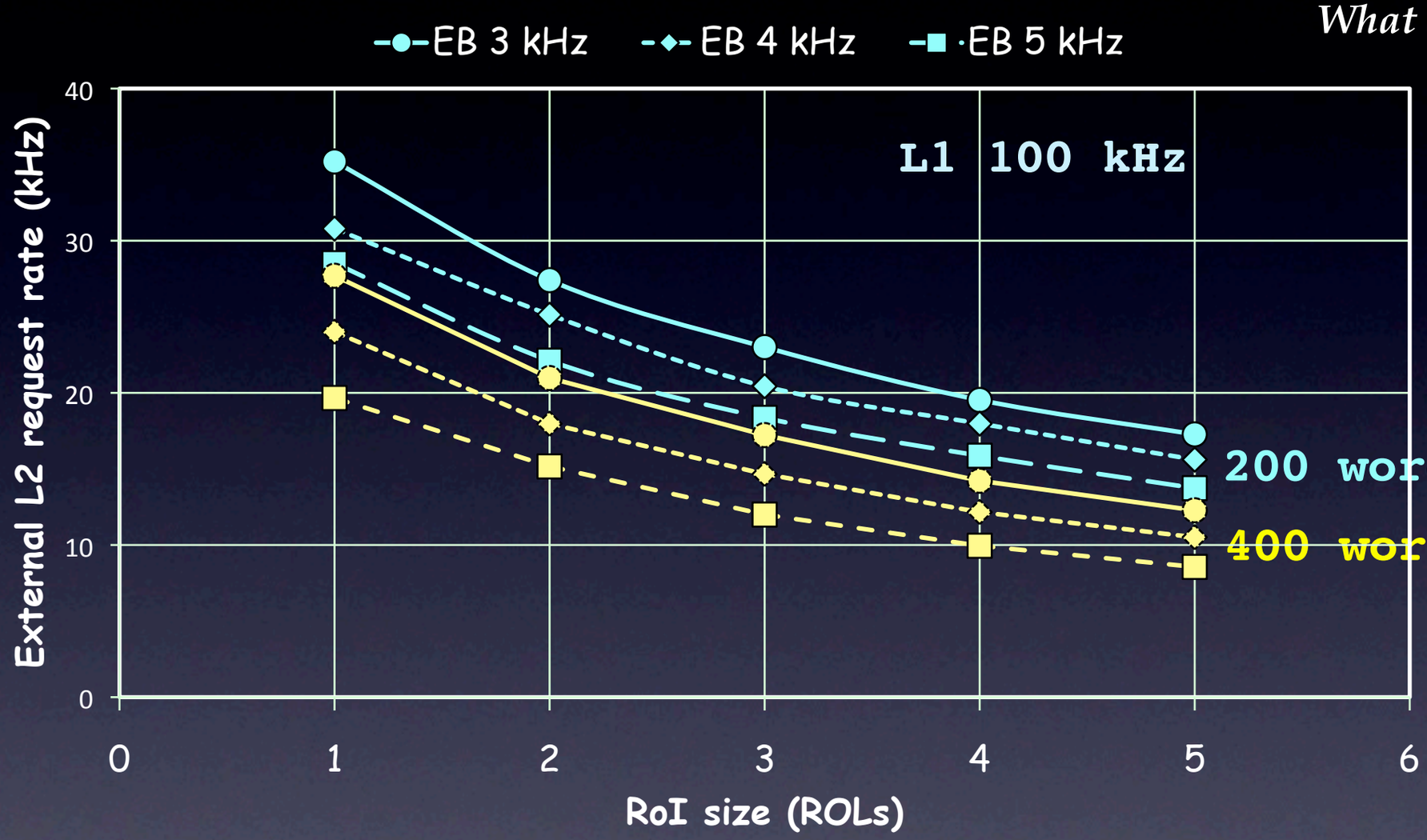
ROS is designed to handle L2 request rates up to ~ 21 kHz per Robin (from 2-3 ROBs†)

- In recent tests some "Hot" ROS PCs hit 30 kHz
- The request rate depends on L2 threshold and algorithms

† Each Robin has 3 ROBs

ROS PC Performance

What determines the performance?



- PCI bandwidth
- Robin latency
- NIC Load
- Request Rate (EB+L2)
- Event Size
- L1 Trigger Rate

Robin **27 kHz request rate per ROL**
(1 kByte fragment size @ 100 kHz L1 rate)

ROS Performance

ATLAS upgrade phases.

- Phase 0 (until 2013)
  $\mathcal{L} \approx 10^{34} cm^{-2} s^{-1}$

- Phase 1 (2013 – 2017)
  $\mathcal{L} \approx 3 * 10^{34} cm^{-2} s^{-1}$

- Phase 2 (from 2018 on)
  $\mathcal{L} \approx 10 * 10^{34} cm^{-2} s^{-1}$

ROS performance improvements

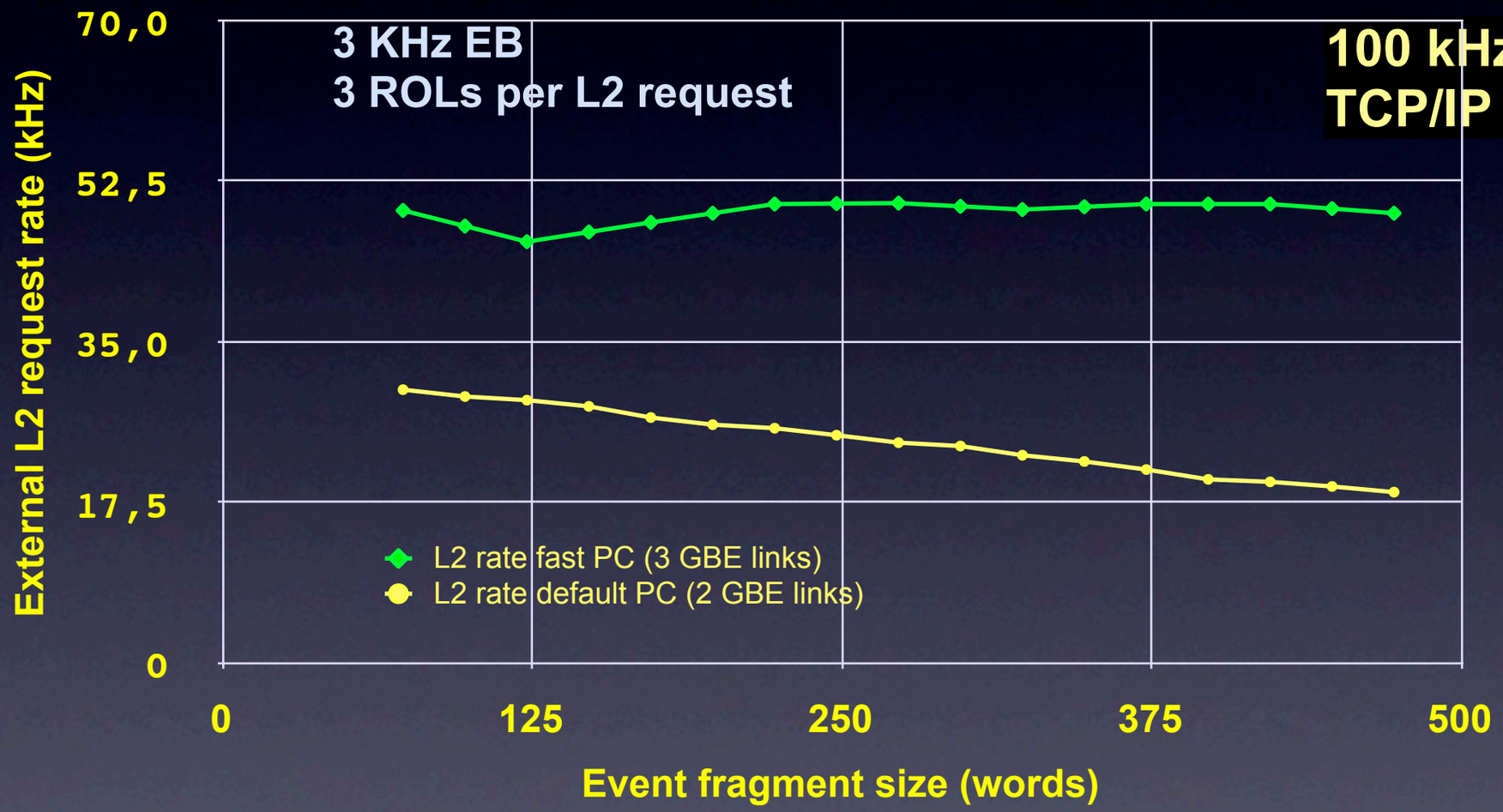
- Higher L2 request rate.
  modified thresholds of the L2 trigger
  additional bandwidth-demanding triggers.
  - Inner detector full scan for b-physics
  - Calorimeter full scan for missing ET
- Higher data rates due to increased luminosity
  Still use (current) ROS PCs & Robins
  Requires more network bandwidth (switches, Robins & ROS)
- Much higher data rates
  Replace ROS system

ROS future scenarios

Improve PC

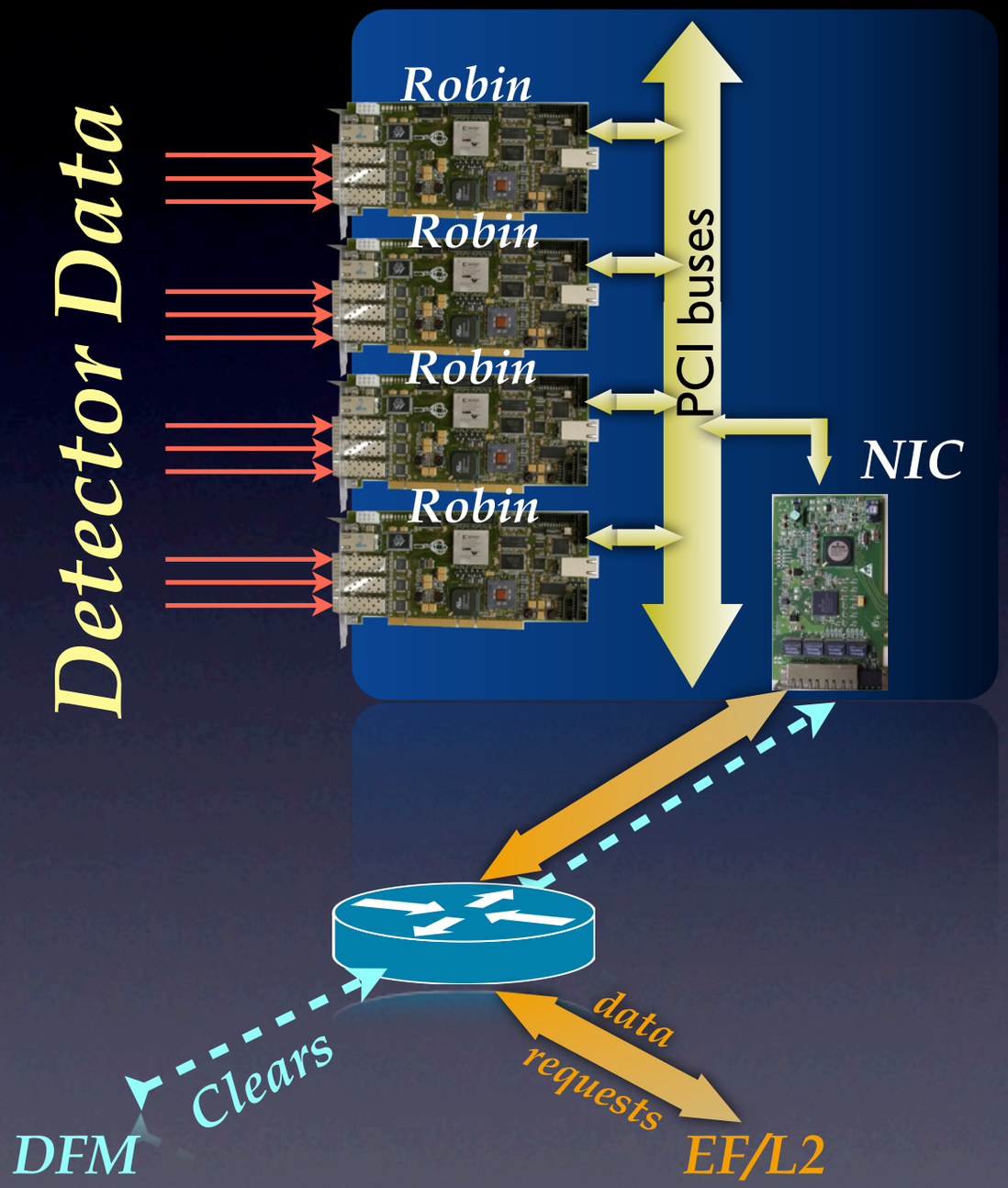
Current ROS hardware was procured in 2005/2006.

Measurement using current generation hardware showing a gain in performance.



- L2 request rate (almost) fragment size independent
- L2 request rate increases by 50% to 150%
- (expensive) Robin cards can be reused

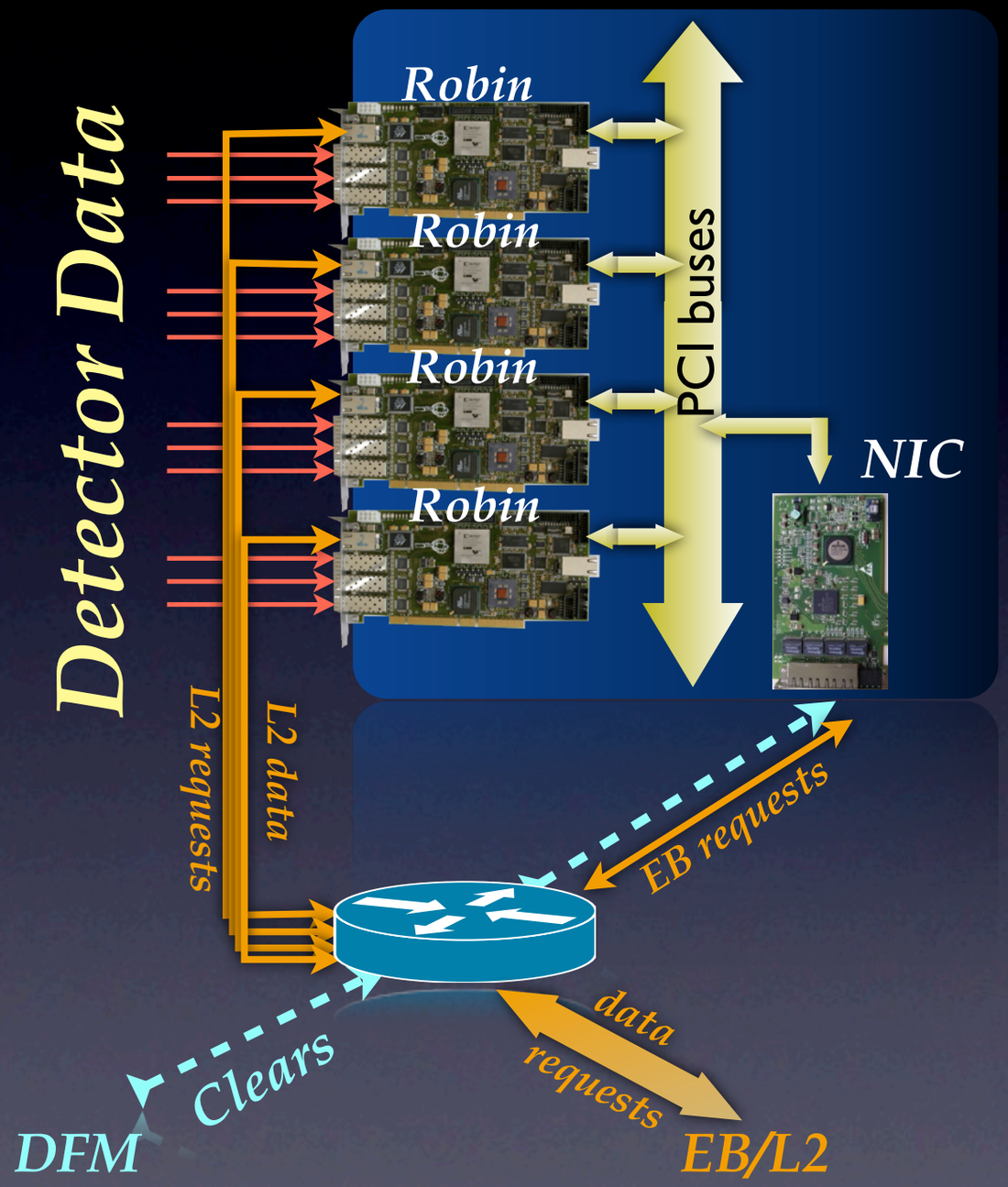
Current ROS



In the current configuration

- Data from the detector are collected in the Robin
- Both data requests and clears are routed via the NIC and the PCI bus

Switch-based ROS



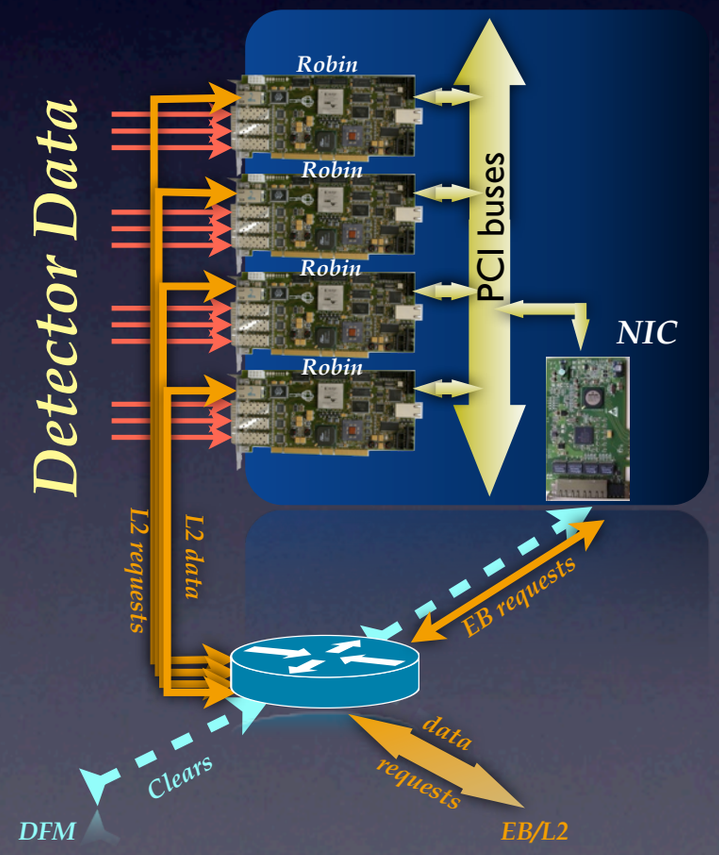
In the switch-based ROS

- Data from the detector are collected in the Robin
- Clears are routed via the NIC and the PCI bus
- Requests from and data to L2 routed via the on board network interface of the Robin (UDP)
- Requests and data to and from EB through the NIC (TCP/IP)

New hardware needed: network switches, cables, etc

Switch-based ROS

- A L2/EB + ROS slice has been set-up
- The production software has been modified to adapt to the new architecture
- Measurement are being performed



Conclusions

- The Dataflow system is being operated in running condition stably since last year
- Based on today's understanding of the ATLAS TDAQ (HLT rejection factor and algorithms) as well as the planned upgrades of ATLAS and LHC the current ROS architecture fulfills the requirements of phase 0 & 1 of ATLAS
- On the other hand enhancing ROS requests rate would also allow to run more demanding algorithm for interesting Physics (B-Physics, Et-miss, ..)
- Short term upgrades could be deployed to improve the performance of this few "hot" ROS PCs just using current generation hardware (50-150% gain in performance)
- The switch-based ROS has the potential to improve significantly the ROS PC performance.