The Base-line DataFlow system of the ATLAS Trigger & DAQ

Hanspeter Beck
LHEP / Universität Bern
On behalf of the Atlas TDAQ DataFlow Group

13th IEEE-NPSS Real Time Conference 2003
Montréal, Canada
May 18-23 2003

Outline

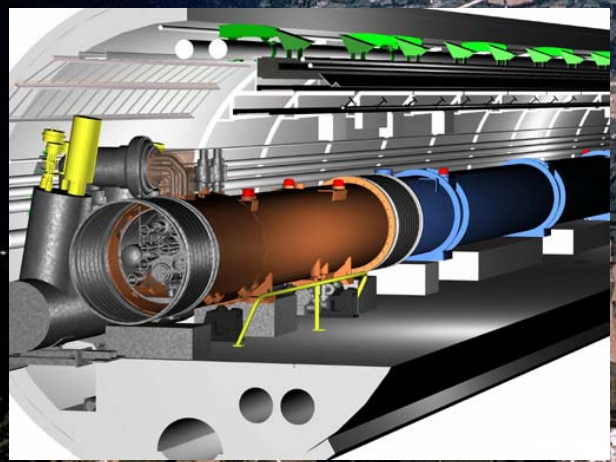
- ATLAS
  - Interaction rates and event sizes
  - The Trigger/DAQ architecture
- The DataFlow
  - ReadOut Link
  - ReadOut System
  - Region of Interest Builder
  - DataCollection
    - RoI DataCollection
    - Event Builder
- Conclusions & Outlook

The Large Hadron Collider

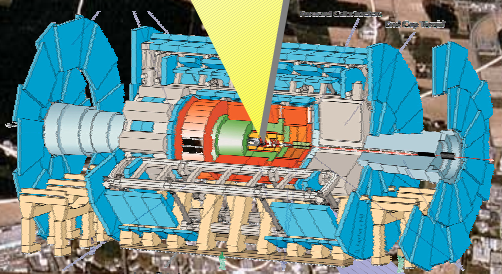
proton-proton collider at $\sqrt{s}$=14 TeV and 10^{34} $cm^{-2}s^{-1}$

heavy ions collider at 5.5 TeV/nucleon and 10^{27} $cm^{-2}s^{-1}$

CMS



Colliding particles:	protons
Center of mass Energy:	14 TeV
Bunch crossing rate:	40 MHz
Interaction rate:	10^9 Hz
Event size:	1-2 Mbytes

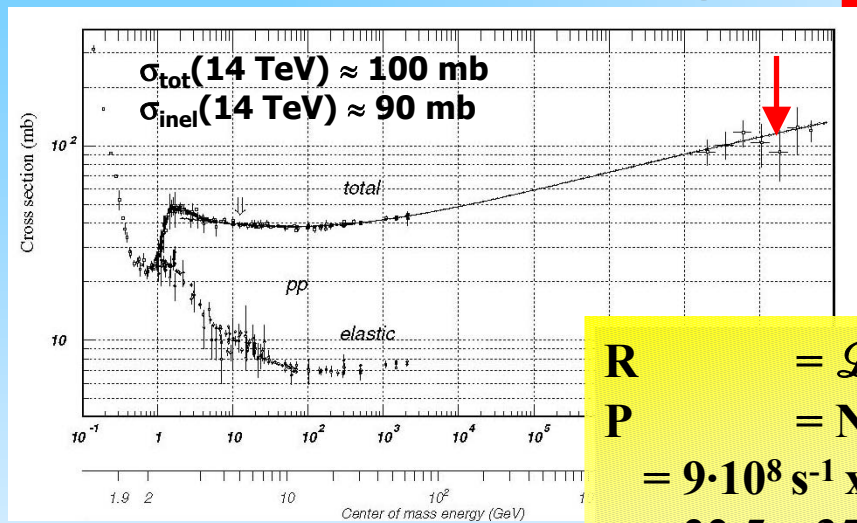


ALICE

ATLAS

LHCb

Event rate and particle multiplicity



R = Event rate
$\mathscr{L}$ = luminosity = 10^{34} cm^{-2} s^{-1}
σ_{inel} = inel. cross-section = 90 mb
N = no. interactions / bunch crossing
Δt = bunch crossing interval = 25 ns

$R = \mathscr{L} \times \sigma_{inel} = 10^{34}\ cm^{-2}\ s^{-1} \times 90mb = 9 \cdot 10^{8}$ Hz

$P = N / \Delta t$

$= 9 \cdot 10^{8}\ s^{-1} \times 25 \cdot 10^{-9}\ s = 22.5$ (not all bunches are filled)

$= 22.5 \times 3564 / 2835$

ossing (Pileup events)

With every bunch crossing

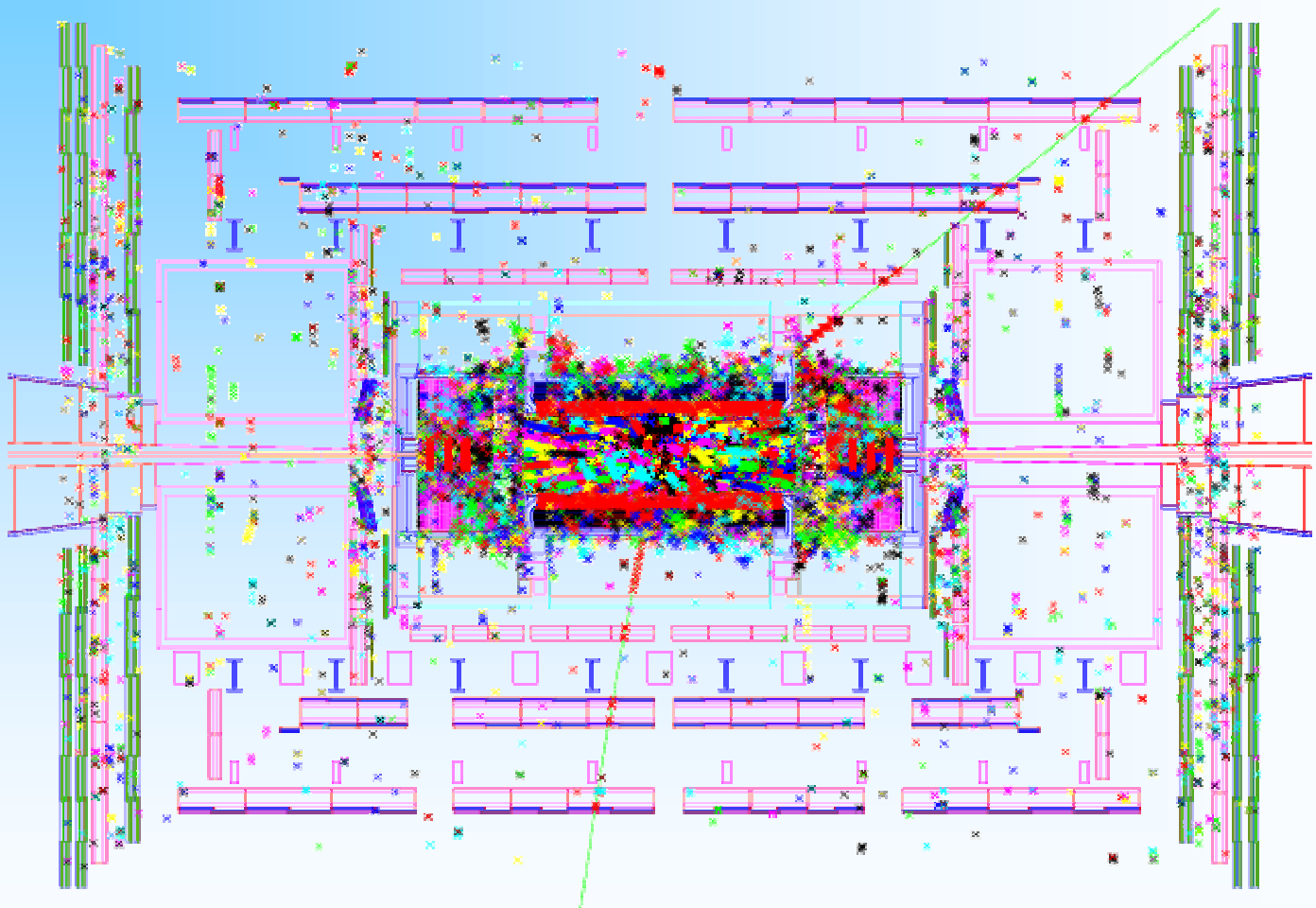
28 min bias interactions

producing 2100 particles into Atlas

ged particles / interaction
charged particles / BC
particles / BC

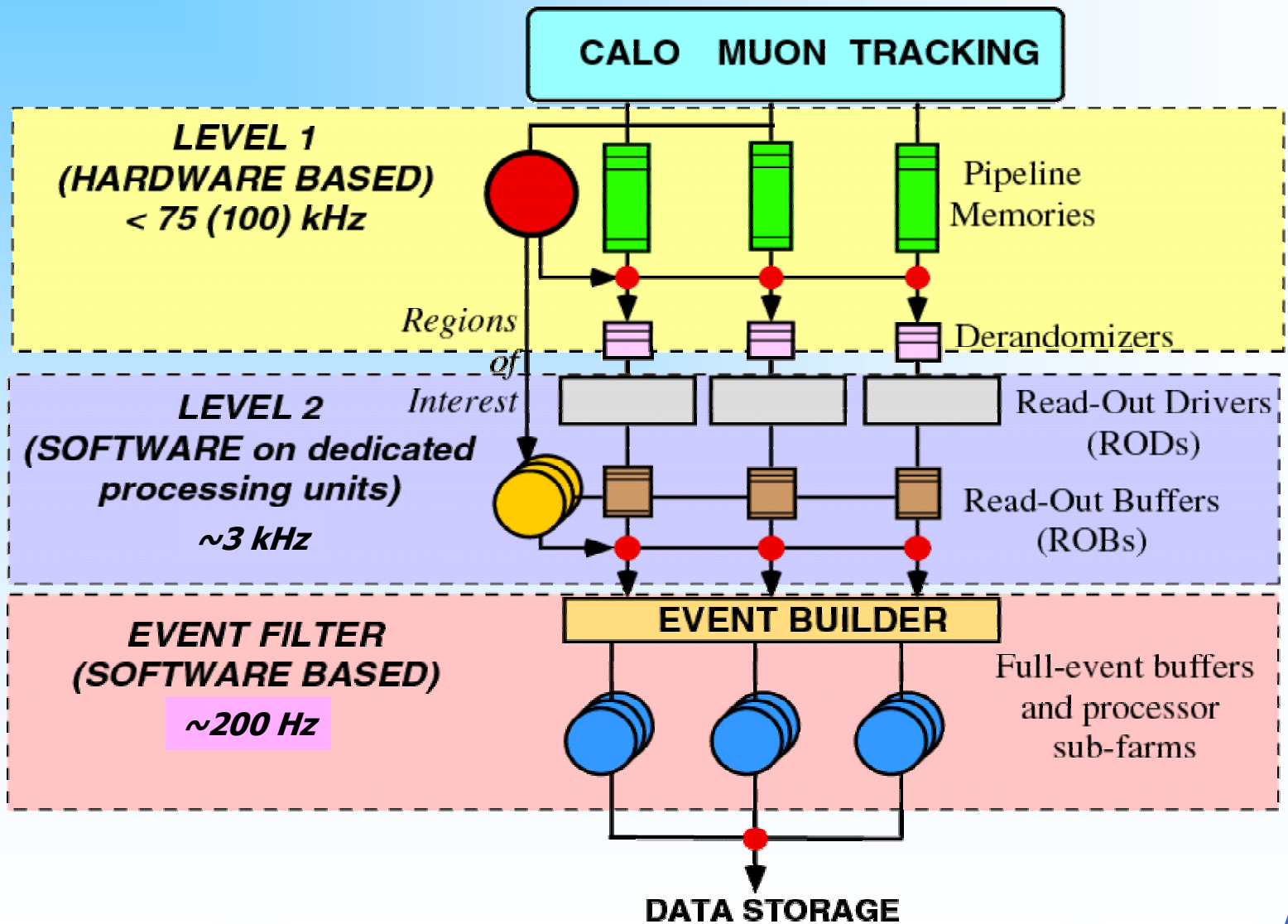
$n_{ch} \approx 50$
$N_{ch} = n_{ch} \times 28 = \sim 1400$
$N_{to} = N_{ch} \times 1.5 = \sim 2100$

Looking for Interesting Physics

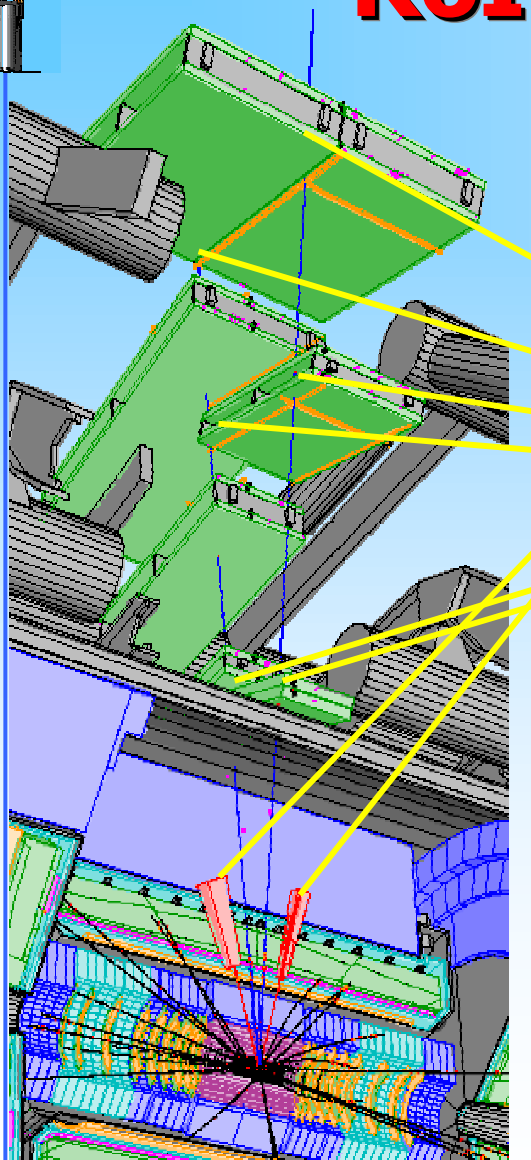


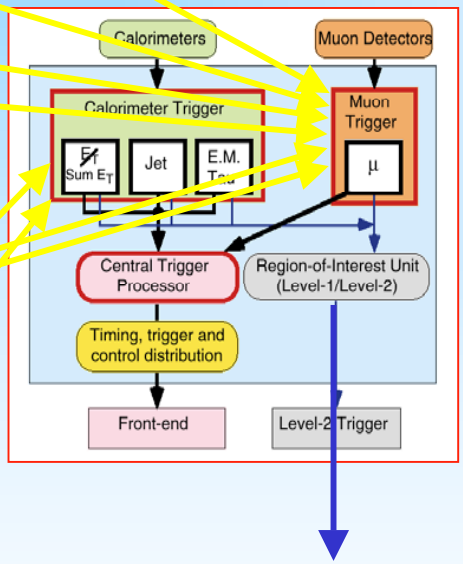
28 min bias events

Three Trigger Levels



RoI - Implementation





- **A Level-2 processor receives η-φ addresses for each RoI**
  - **Level-2 decides which ROBs to access**
  - **Request-reply mechanism**
  - **Sequential processing at Level-2**
- **About 2% of an events data volume makes up an RoI**

⇒ **one order of magnitude smaller ReadOut network**

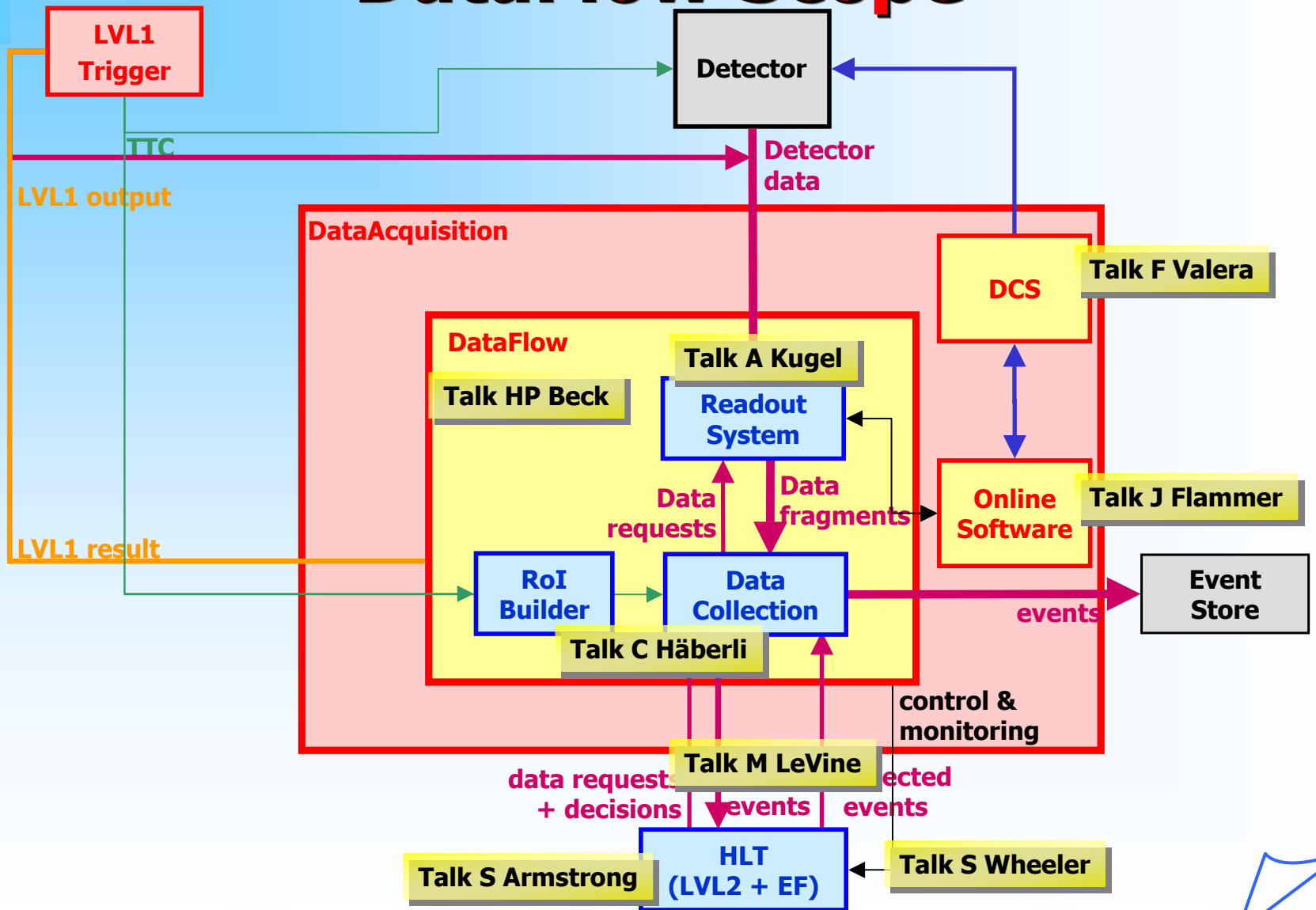
⇒ **higher system complexity**

⇒ **RoI data requests integrated in DataFlow system**

2 μ-RoIs + 2 e-RoIs

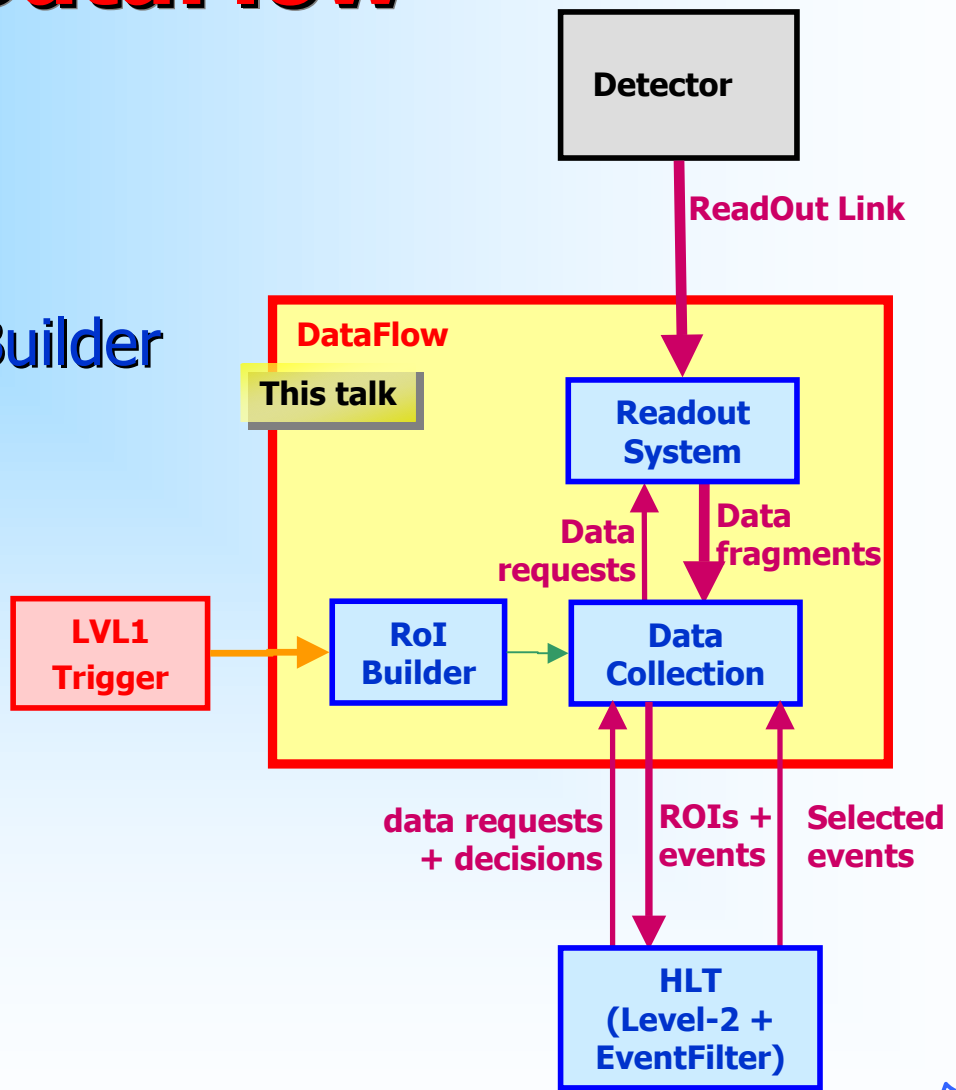
η-φ addresses sent to Level-2

DataFlow Scope



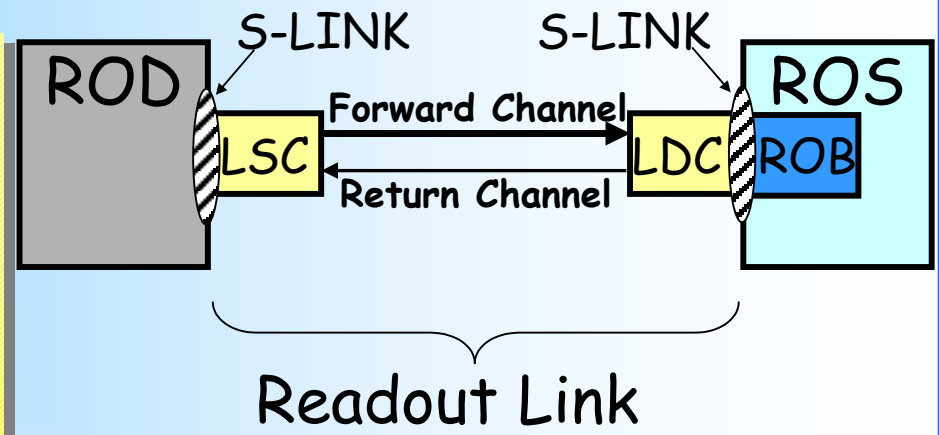
DataFlow

- ReadOut Link
- ReadOut System
- Region of Interest Builder
- DataCollection
  - RoI DataCollection
  - Event Builder



ReadOut Link

- ~1600 ROLs connect detector ReadOut Drivers [ROD] with ReadOut Buffers [ROB] organized within ReadOut Systems [ROS]
- All Level-1 accepted data is moved from RODs to ROSs at up to 160 MByte/s per ROL
  - Up to 256 GByte/s capability of Atlas Level-1 accepted data
- The S-Link protocol allows for backpressure via a independent return channel
  - Xon/Xoff
- Fibre and copper based implementations of LinkSource- [LSC] and LinkDestination Cards [LDC]
  - Final decision pending

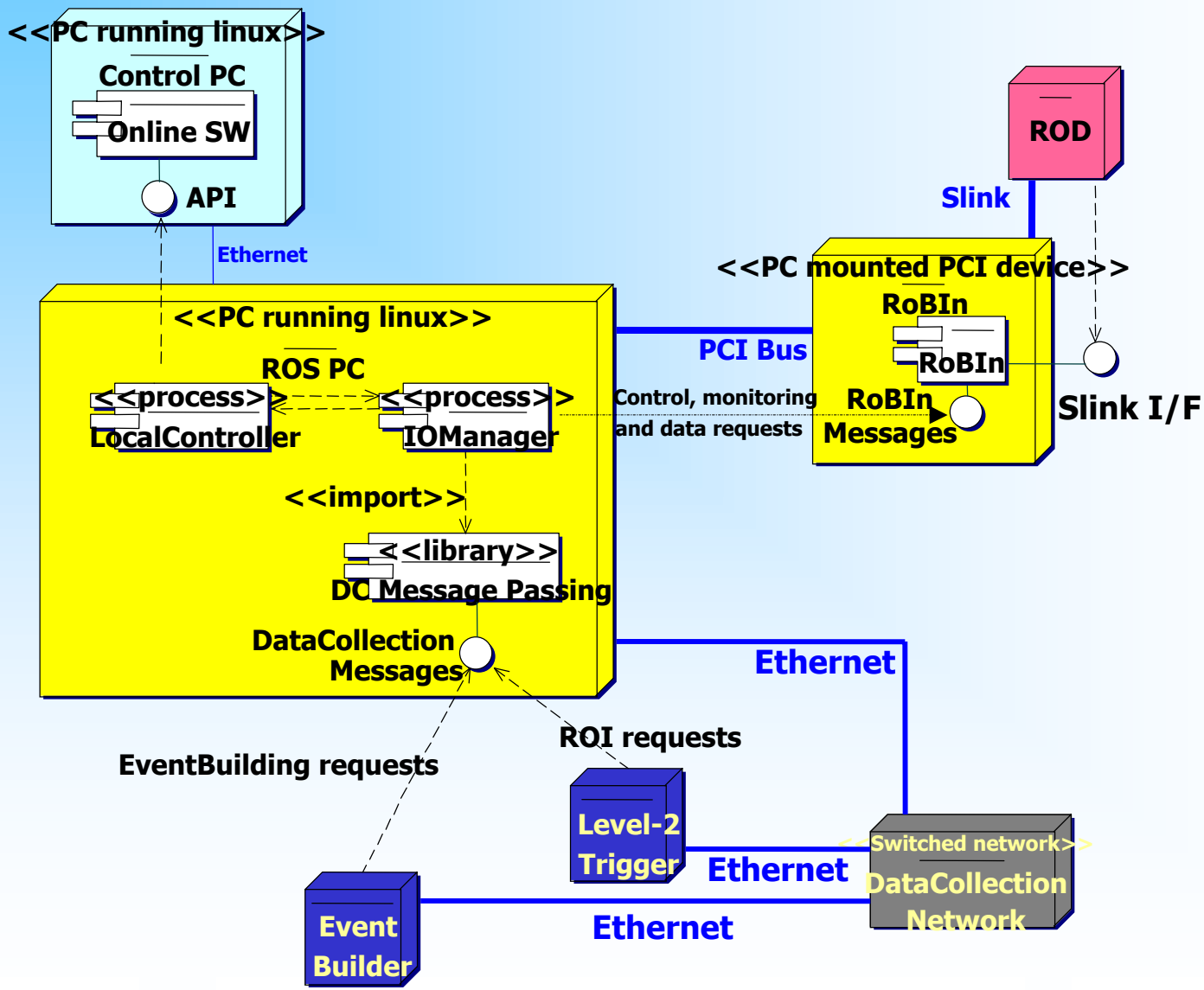


- ROD sub-detector specific
- ROB and ROS common for all of TDAQ
  - between detector and TDAQ world

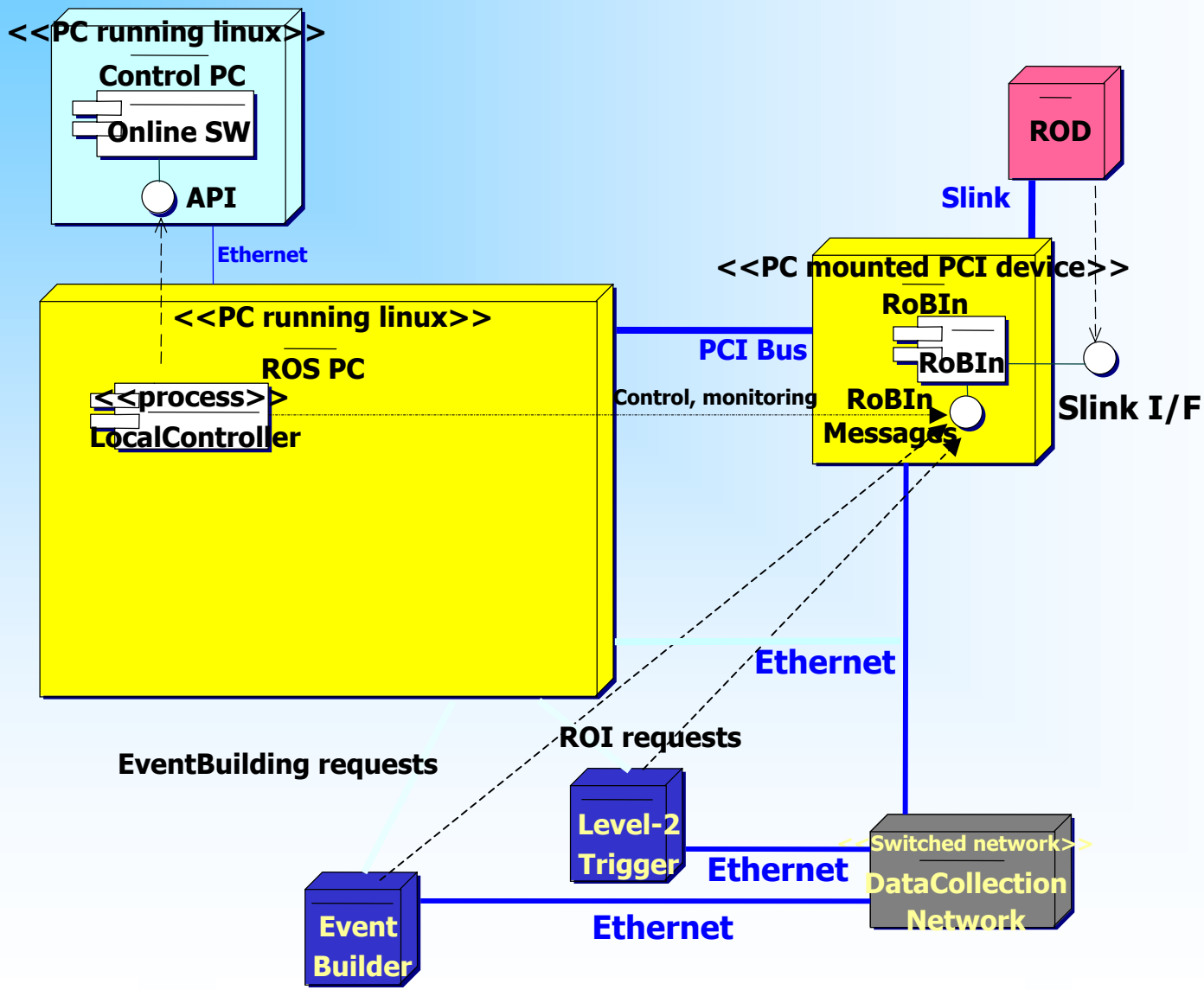
ReadOutSystem

- **Receive & buffer event fragments from the ~1600 detector ROLs**
  - Up to 160 MByte/s per ROL
  - ~1600 ReadOutBuffers [ROBs] **implemented by RoBIns holding 2-4 ROBs each**
- **Send event fragments on request**
  - **ROI requests:** high rate, low data volume
    - Rate: LVL1 rate **75 kHz**, volume: **~2%** of ROLs: **~30 kB**
  - **EventBuilding requests:** low rate, high data volume
    - Rate: **~4%** of LVL1 rate: **~3 kHz,** volume: complete event data: **1-2 MByte**
- **Two deployment scenarios possible**
  - **Bus-based ROS:** ROS concatenates detector data if more than one ROB data fragment is requested from a ROS unit
    - **RoBIns read out via PCI bus**
      - Less ports in network
      - Less data request messages in DataFlow system
  - **Switch-based ROS:** RoBIns accessed directly from Level-2 and EventBuilding
    - **RoBIns read out via Gigabit Ethernet**
      - Flexible scaling (ROB request rates and EB rates)
      - No dependency of PCI bus and CPU in data-path

ROS: Deployment Views



ROS: Deployment Views

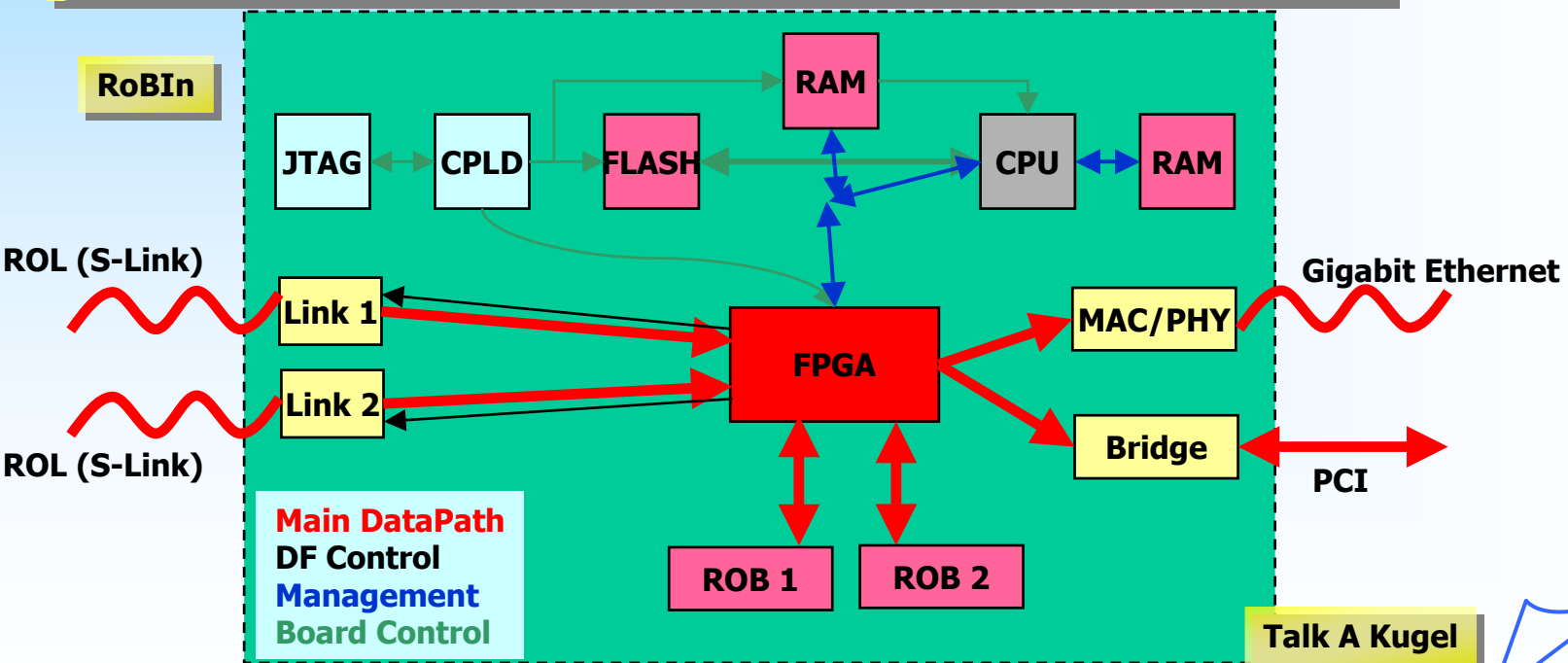


ReadOut Buffer

Prototype implementation of RoBIn existing

- **Receive and buffer detector da**

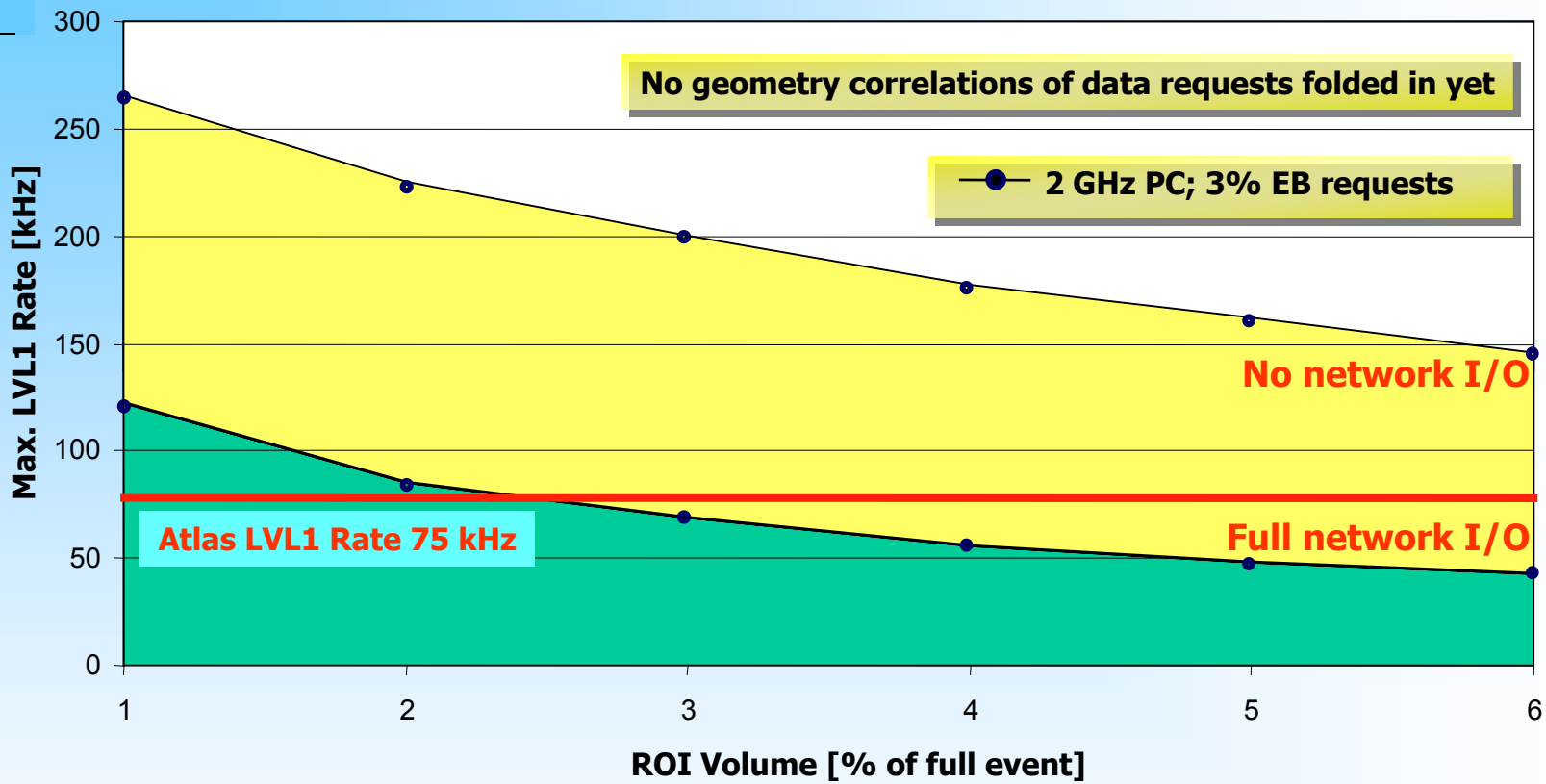
– decision pending



Test Setup: ROS performance

- **Bus-based ROS implemented on a 2 GHz PC**
  - **with 4 PCI busses (64 bit/66 MHz)**
- **3 RoBIn emulators on PCI-bus**
  - **On-board "local" bus limited to 266 MByte/s**
  - **Each emulates 4 input links ⇒ 12 ROLs per ROS-PC**
- **Level-2 & event-building emulators**
  - **Linux PCs connected to the ROS via Gigabit Ethernet**
  - **Sends RoI/EventBuilding requests and ROB clear messages to the ROS**
  - **Receive event data shipped back from ROS-PC**
  - **Using TCP/IP as communication protocol**
    - **Raw Ethernet and UDP/IP also possible**
    - **TCP/IP unveils worst case scenario**

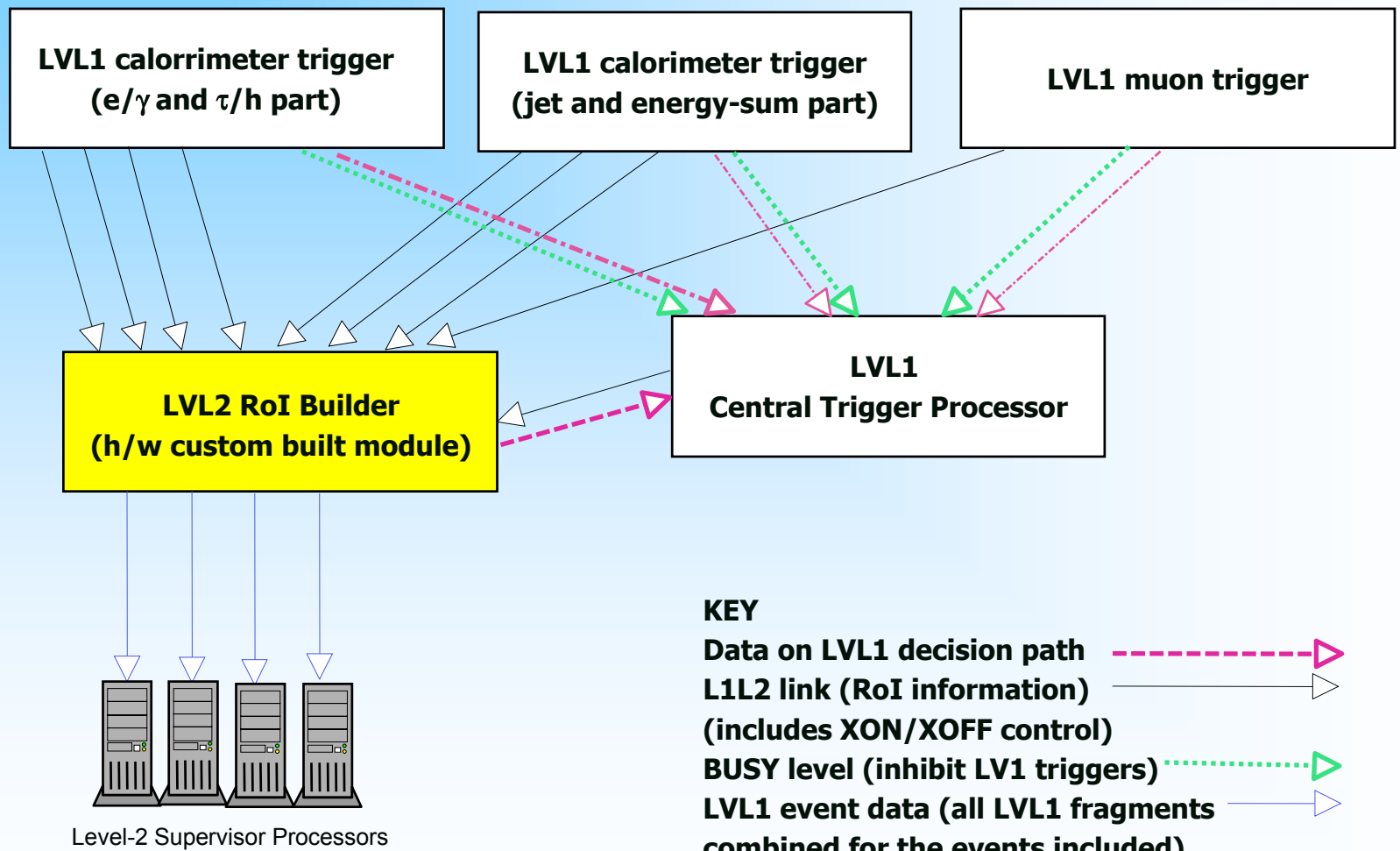
ROS Performance



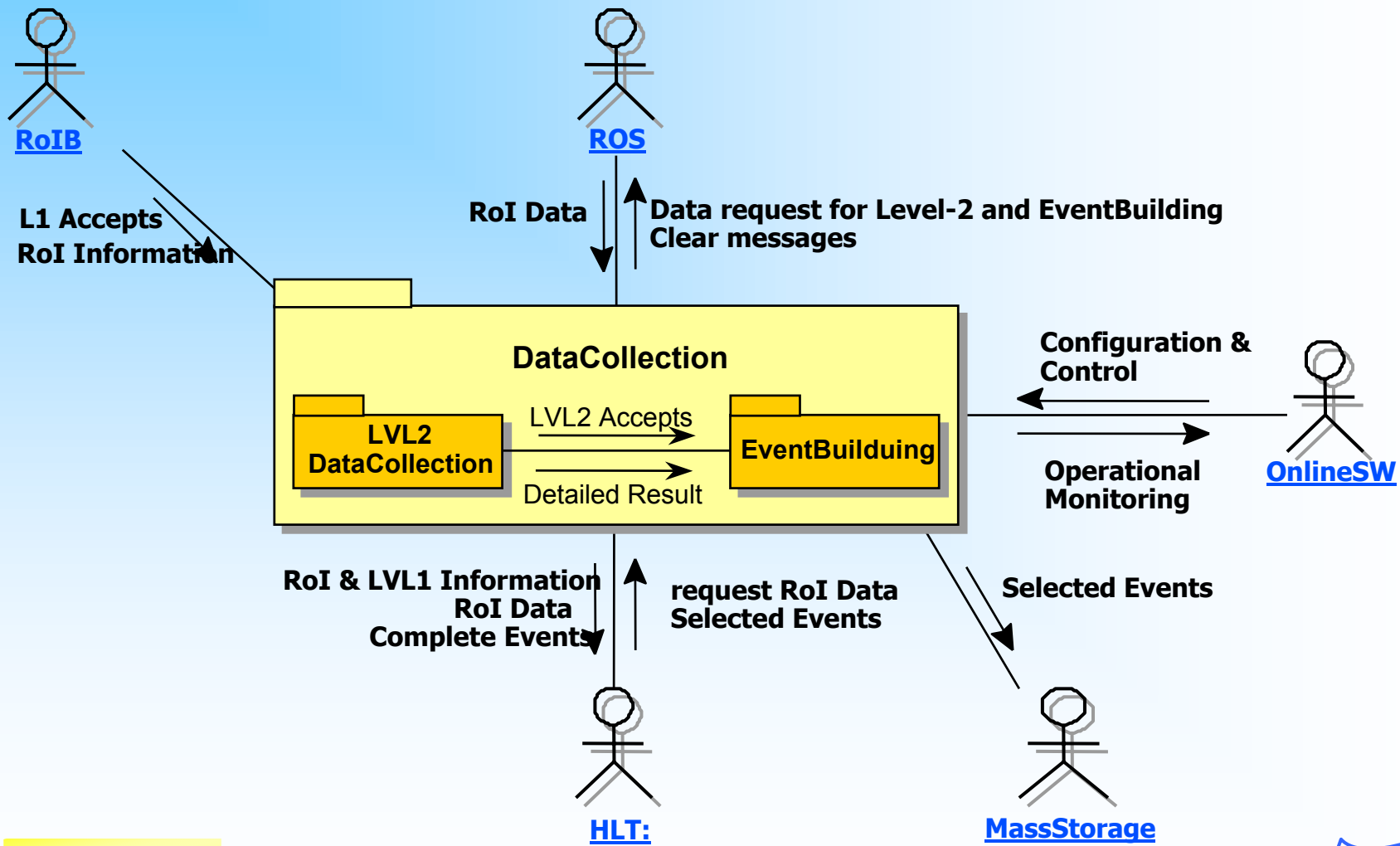
⇨ A ROS housing 12 RO

Higher RoI Volumes can be reached with a better decoupling of output task from ROS internal tasks; i.e. SMP and/or faster CPU

RoI Builder


LVL1 calorrimeter trigger
(e/γ and τ/h part)
LVL1 calorimeter trigger
(jet and energy-sum part)
LVL1 muon trigger
LVL2 RoI Builder
(h/w custom built module)
LVL1
Central Trigger Processor
Level-2 Supervisor Processors
KEY
Data on LVL1 decision path
L1L2 link (RoI information)
(includes XON/XOFF control)
BUSY level (inhibit LV1 triggers)
LVL1 event data (all LVL1 fragments
combined for the events included)

DataCollection

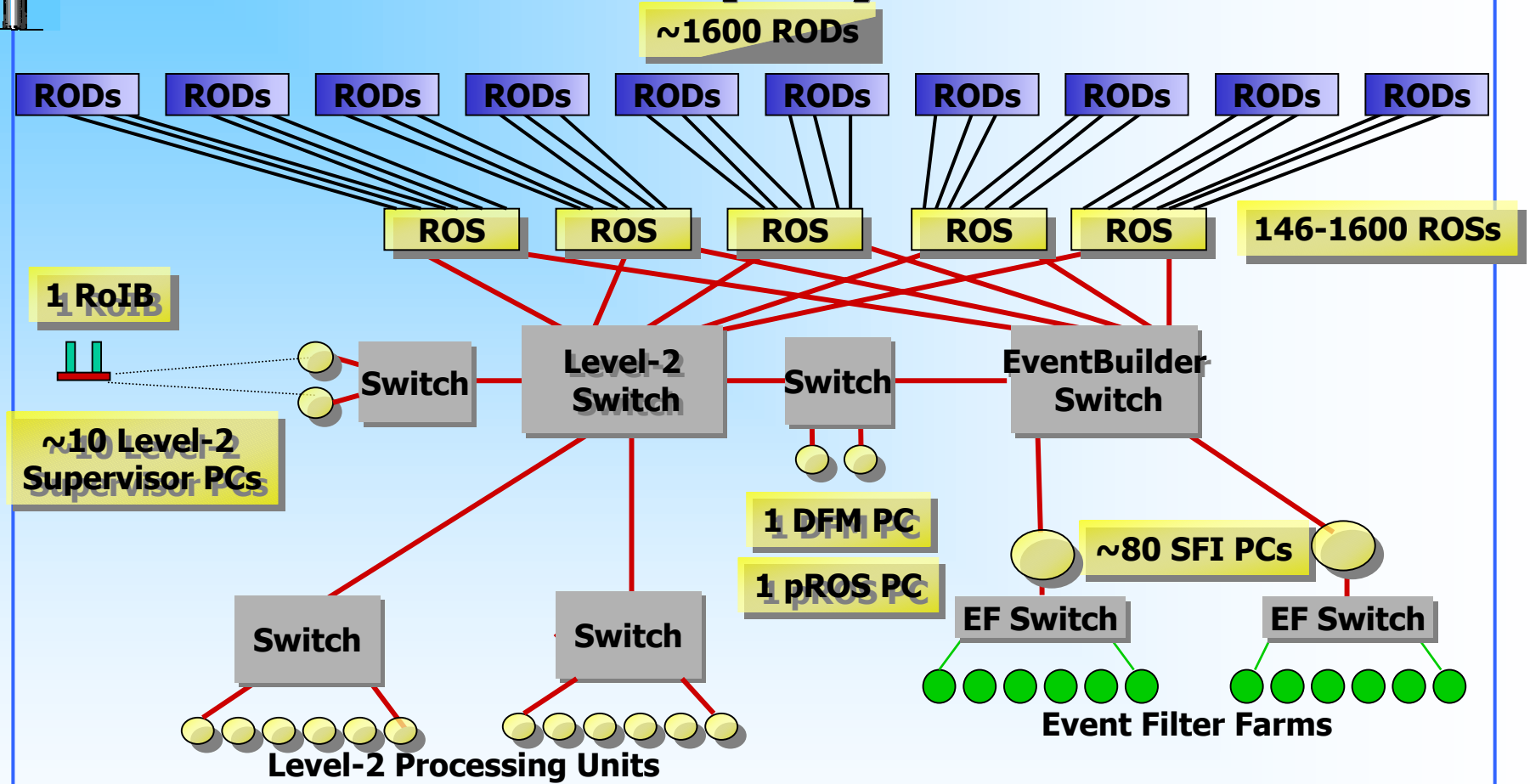


Talk C Häberli

DataCollection Components

- **DataCollection Components**
  - S/w process written in C++ run on Linux PCs
  - communication via Gigabit Ethernet: TCP/IP, UDP/IP, Raw
  - Level-2 SuperVisor — load-balancing of Level-2 farm
  - Level-2 Processing Unit — Level-2 comunication layers i.e. no trigger algorithms added
  - DataFlowManager — load-balancing of event-building SFIs
  - Pseudo ROS — Level-2 ROS: allows to add a Level-2 record into the event stream
  - SubFarmInput — building the full event
  - SubFarmOutput — event data available for MassStorage

DataFlow Deployment view



DataFlow Prototype

DataFlow Prototype to demonstrate functionality, performance and scalability of the proposed Atlas DataFlow

~40 Xeon 2.0-2.4 GHz dual CPU Rack-mountable PCs
Fully interconnected with Gigabit Ethernet

Prototype setup at Cern

4 types of ROS emulators

Talk M LeVine

- **128 FPGA based network tester**
- **16 Alteon NIC firmware re-programmed**
- **s/w ROS emulator – PCs**
- **Fully functional ROS prototype - PCs**

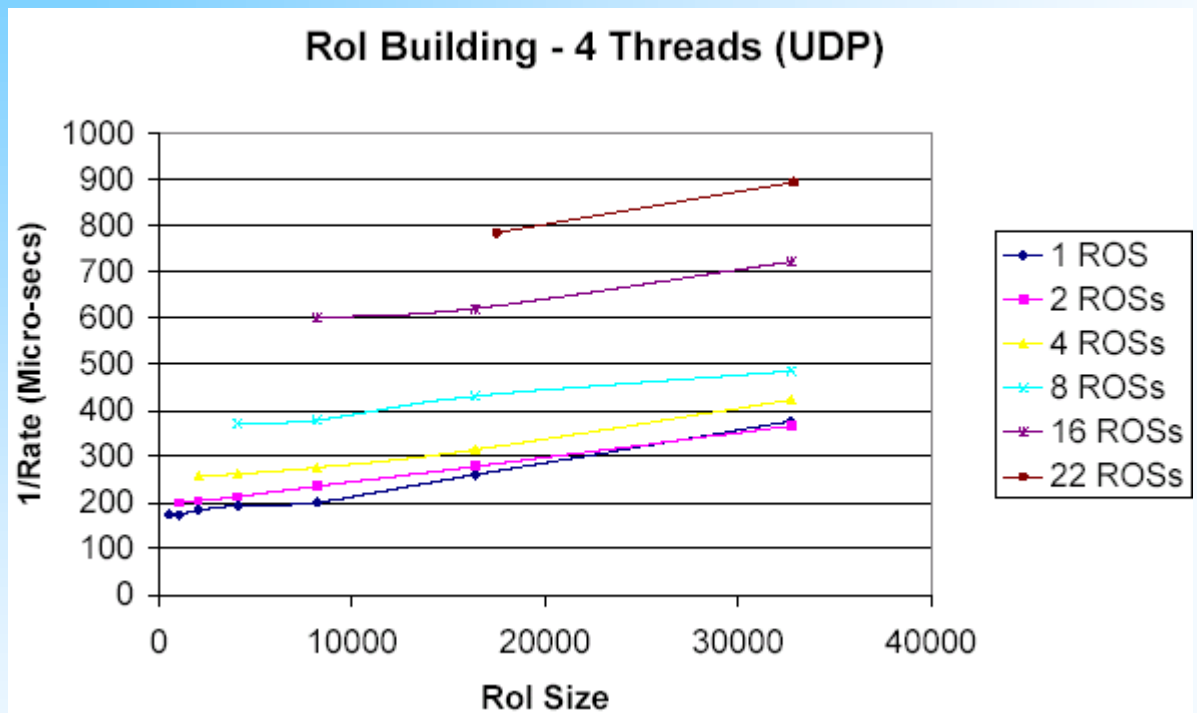
Capable of testing ~10% of final Atlas dataflow

The TDAQ Technical Design Report will be submitted end of June 2003

Performances

- **Performance of RoI Builder**
  - Custom built 12U VME prototype has achieved required performance
- **Performance of Level-2 supervisor**
  - 1 Level-2 Supervisor absorbs up to ~30 kHz of Level-1 rate on a 2.4 GHz dual CPU PC ⇒ a few PCs needed for final system
  - Is insensitive to the number of Level-2 Processing Units
- **Performance of Pseudo-ROS**
  - Not a demanding application
  - Requirement to receive <10 kB of data at Level-2 accept rate (~3 kHz) and forward them to the EB is largely satisfied

DataFlow Performance of a Level-2 Processing-Unit



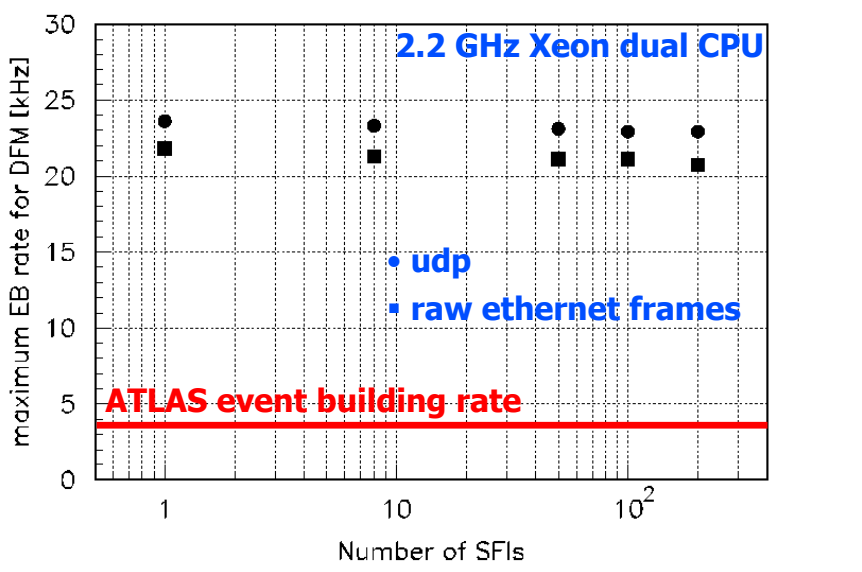
total time-budget for Level-2 of 10 ms

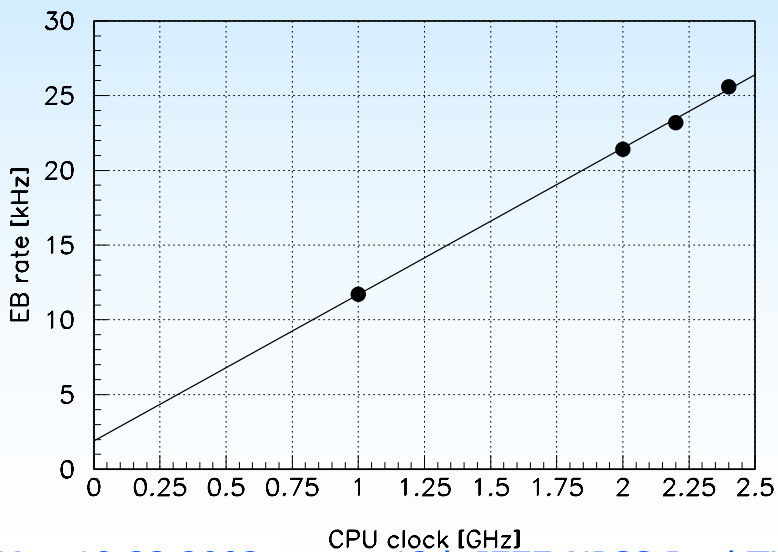
i.e. less than 10% for I/O and 90% of CPU time available for processing

Talk S Armstrong

In this test setup, the total RoI size is distributed evenly over a number of ROSs from which the Level-2 Processing-Unit needs to request RoI data from

DFM Performance





DataFlow Manager

Load-balancing SFIs (event building nodes)

Dedicated set-up exposes DFM to full I/O rates and bandwidth using special tester application

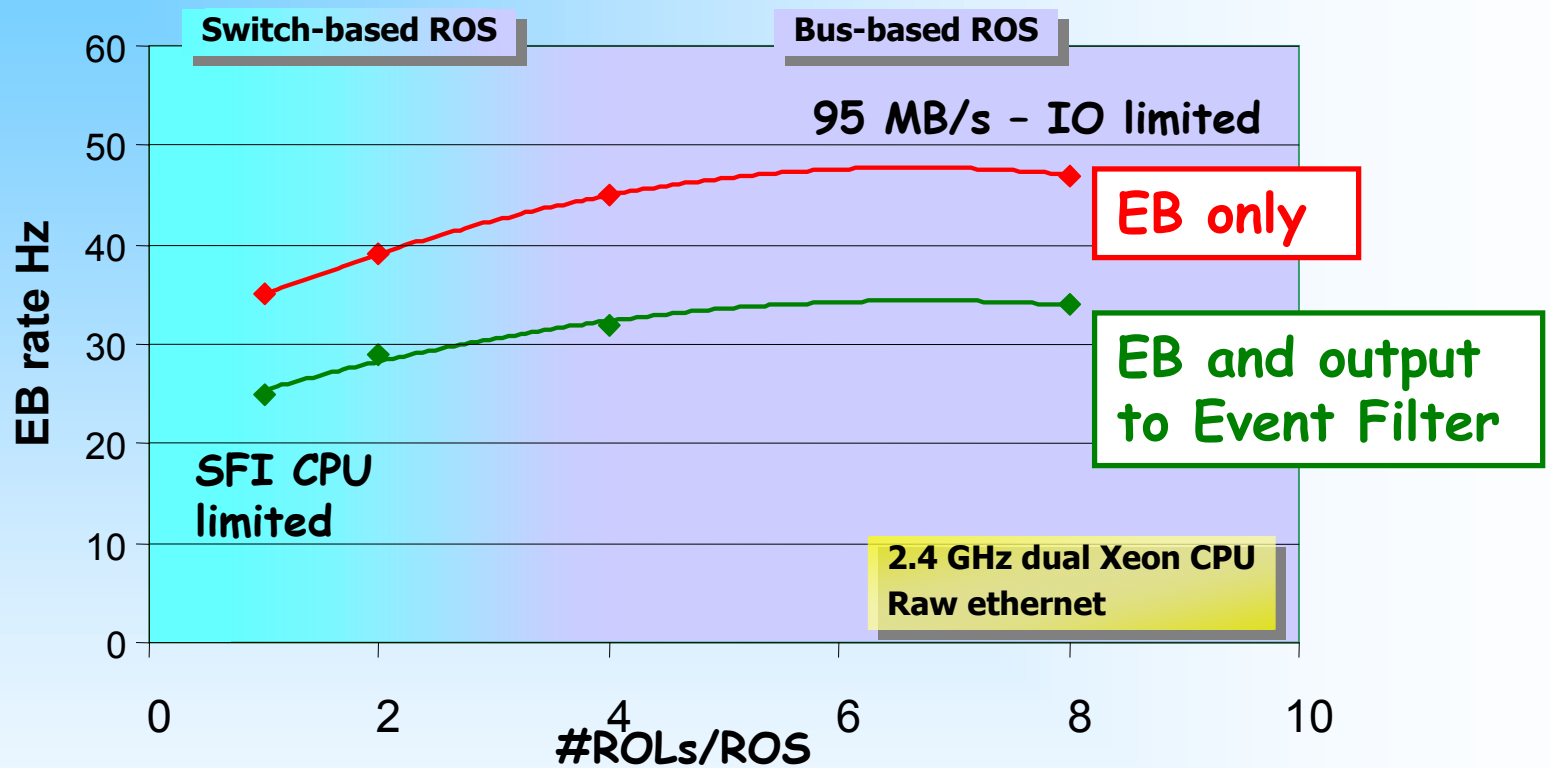
The DFM performs an order of magnitude above the Atlas baseline requirements

23 kHz Event Building rate could be sustained by the DFM

The DFM performance is insensitive to the number of SFIs (load-balancing)

The DFM performance scales linearly with CPU clock speed

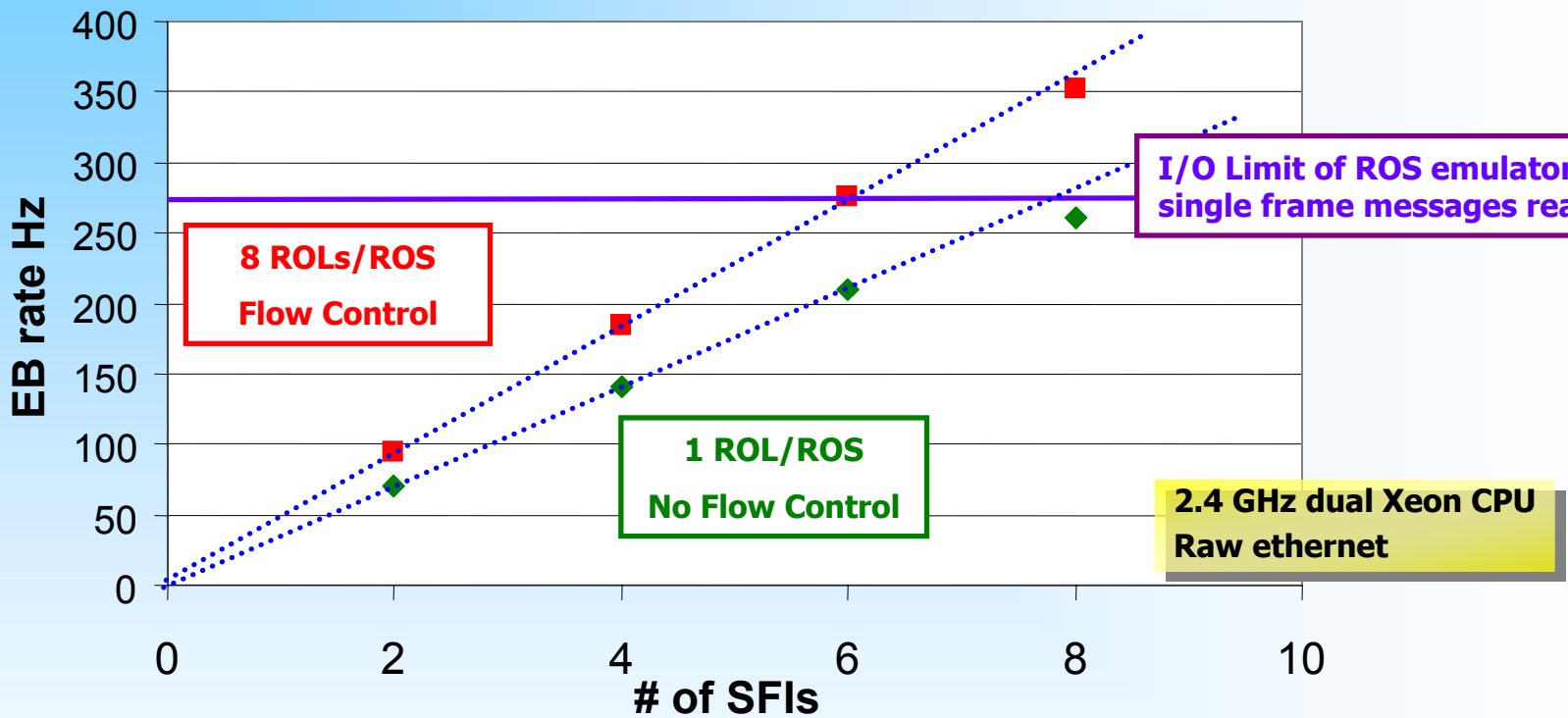
SFI Performance



One SFI is capable of building

g event data from 1600 ROSs

EventBuilding Performance



byte/s (15% of Atlas EB b/w)

Conclusions & Outlook

- DataFlow system for ATLAS TDAQ defined
- All components well specified
  - Mostly commodity components
    - Linux PCs, Gigabit Ethernet
  - ROL, RoBIn and RoI builder custom built
- Few areas in the design with outstanding decisions
  - Number of ROLs per RoBIn
  - Bus-based ROS vs switch-based ROS
- Testbed capable of providing 10% of the final expected Atlas throughput
  - ~40 Linux PCs: 2.0 – 2.4 dual Xeon CPU
  - Dedicated network testers emulating large number of ROSs
- Performance figures promising and sometimes above requirements
- Now writing ATLAS TDAQ TDR; due by end of June 2003
- DataFlow system being explored on testbeam setups at CERN
  - further insight on stability
  - get the detector community acquainted
- TDAQ is well on track for ATLAS

Backup Slides

LHC - DESIGN PARAMETERS

LHC = Two colliding proton synchrotrons (26.7 km circumference)

Injection Energy	0.45 TeV
Collision Energy	7 TeV
Dipole field at 7 TeV	8.33 T
Design Luminosity	10^{34} cm^{-2} s^{-1}
Luminosity Lifetime	10 h
Protons per bunch	10^{11}
Bunches per beam	2808 (filled) 3546 (total)
Bunch spacing	25 ns
DC Beam Current	0.56 A

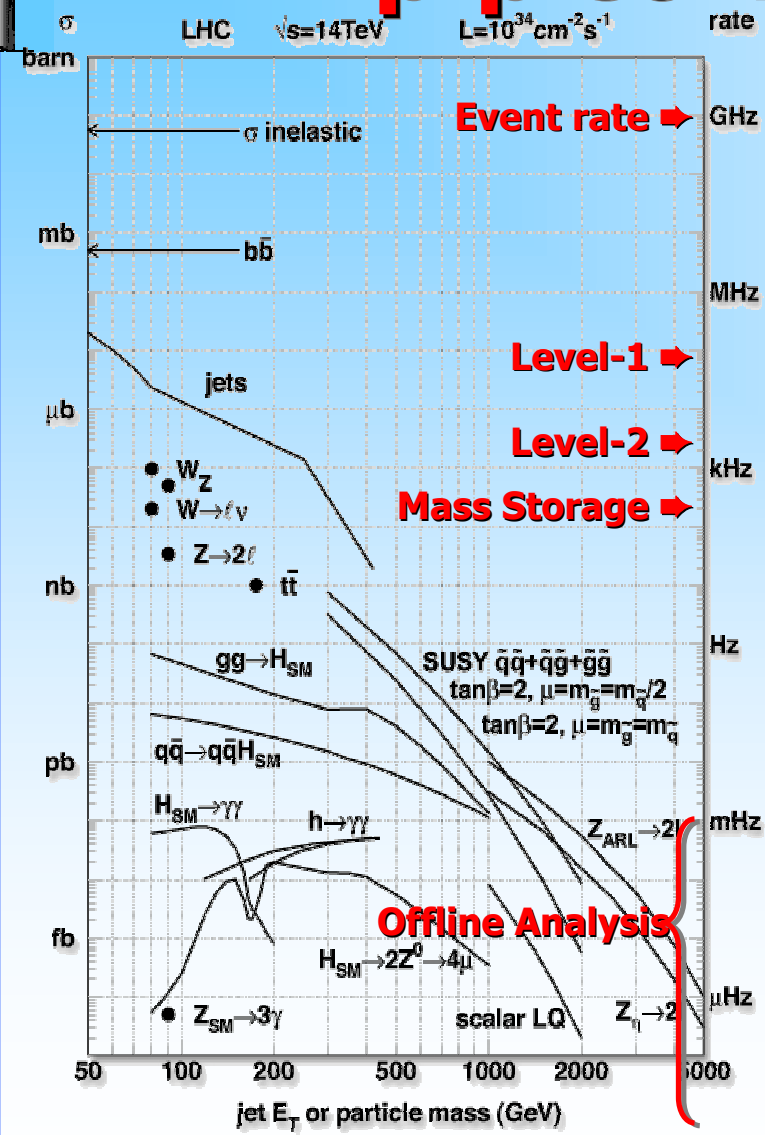
Extreme demands on detectors:

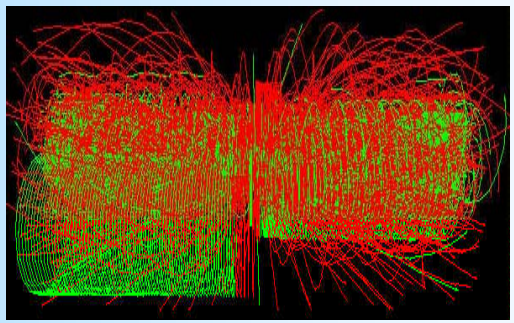
- **high granularity**
- **high data-taking rate**
- **high radiation environment**

Initial Luminosity $\mathcal{L}$=2·10^{33} cm^{-2} s^{-1} with goal of $\int \mathcal{L}\,dt \approx$ 20 fb^{-1} /year for 3 years

High Luminosity $\mathcal{L}$=10^{34} cm^{-2} s^{-1} with goal of $\int \mathcal{L}\,dt \approx$ 100 fb^{-1} /year for 10+ years

p-p Collisions at LHC





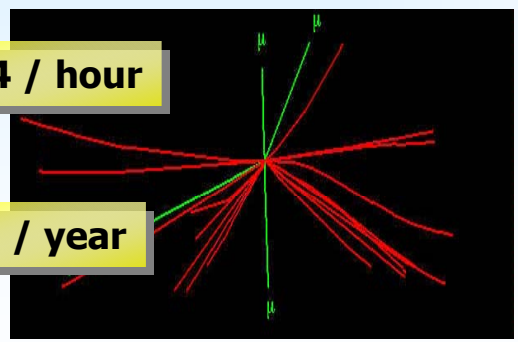
Event Rate	1 GHz
Level-1 Output	75 kHz
Level-2 Output	~ 3 kHz
Mass storage	~ 200 Hz
New Physics	~ μHz - mHz

1 : 10^7

Total selection power needed for Higgs finding: 1:10^{13}

3-4 / hour

10 / year



H→4μ

Atlas Event Size

Inner Detector	Channels	Fragment size - kB
Pixels	1.4×10^8	60
SCT	6.2×10^6	110
TRT	3.7×10^5	307

Muon Spectrometer	Channels	Fragment size - kB
MDT	3.7×10^5	154
CSC	6.7×10^4	256
RPC	3.5×10^5	12
TGC	4.4×10^5	6

Calorimetry	Channels	Fragment size - kB
LAr	1.8×10^5	576
Tile	10^4	48

Trigger	Channels	Fragment size - kB
LVL1		28

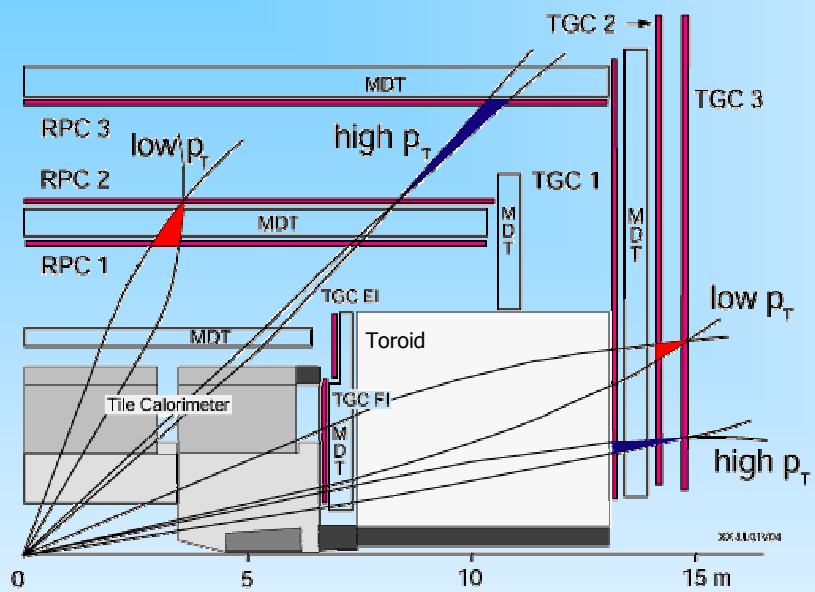
Atlas total event size: 1.5 Mbytes
140 Mio Channels
organized into ~1600 Readout Links

Mass Storage:

300 MBytes/sec

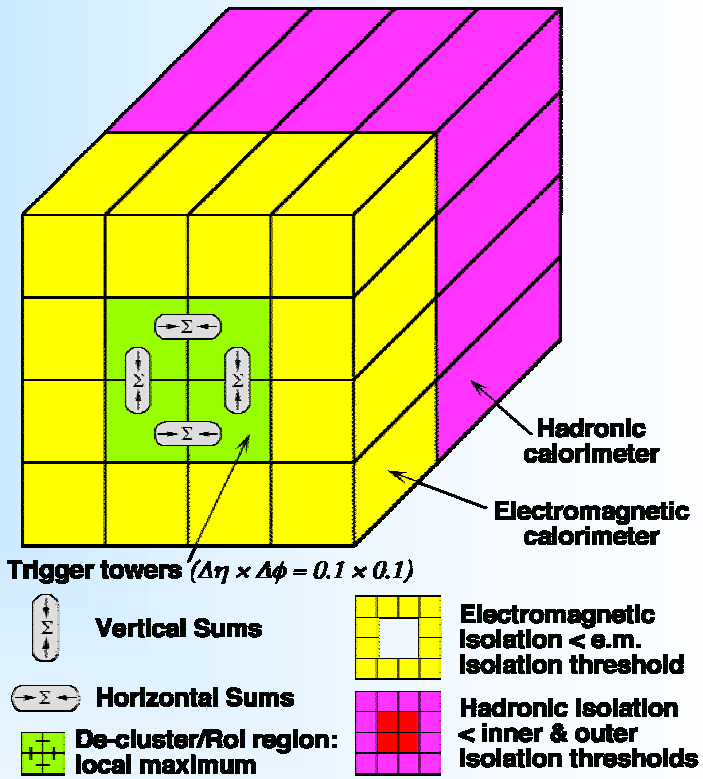
☛ 3 PetaBytes/year for offline analysis

LVL1 - Muons & Calorimetry



Muon Trigger looking for coincidences in muon trigger chambers
2 out of 3 (low-p_T; >6 GeV) and 3 out of 3 (high-p_T; > 20 GeV)

Triger efficiency 99% (low-p_T) and 98% (high-p_T)



Calorimetry Trigger looking for $e/\gamma/\tau + jets$

- **Various combinations of cluster sums and isolation criteria**
- **$\Sigma E_T^{em,had}$, E_T^{miss}**

LVL1 Trigger Rates

Selection		$2*10^{33}$ cm^{-2}s^{-1}	10^{34} cm^{-2}s^{-1}
MU20	(20)	0.8	4.0
2MU6		0.2	1.0
EM25I	(30)	12.0	22.0
2EM15I	(20)	4.0	5.0
J200	(290)	0.2	0.2
3J90	(130)	0.2	0.2
4J65	(90)	0.2	0.2
J60 + xE60	(100+100)	0.4	0.5
TAU25 + xE30	(60+60)	2.0	1.0
MU10 + EM15I		0.1	0.4
Others (pre-scales, calibration, ...)		5.0	5.0
Total		~ 25	~ 40

- Rates given in kHz — No safety factor included!
  - → E_T thresholds imply 95% efficiency values

The LVL1 rate is dominated by EM cluster triggers

Region of Interest - Why?

- At hadron colliders, the most severe background comes from Physics
  - QCD production of jet-jet events
    - QCD production (qq-gg-qg) is orders of magnitudes higher than interesting Physics signals
    - Quarks and gluons "materialise" into jets of particles of variable multiplicities (-> π's, K's, etc.)
- The LVL1 trigger rejects a large fraction of it
  - from a crude profile analysis of calorimetric energy deposition
    - But the full identification of an electron against $\pi^0 \rightarrow \gamma\gamma$ requires:
      - high calorimeter granularity
      - association track-energy
- Because of interconnectivity, LVL1 has
  - Poor calorimeter granularity
  - No access to tracking information

LVL2 - RoI mechanism

Level-1 triggers on high p_T objects

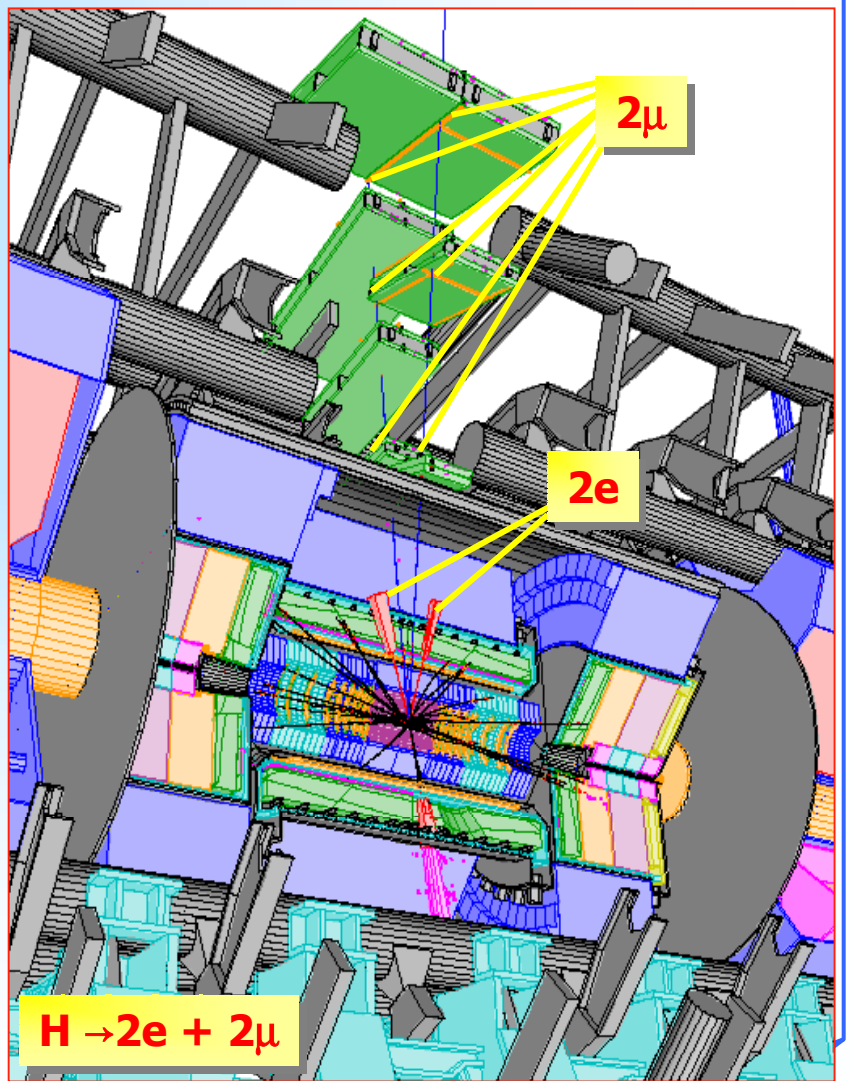
- Caloriemeter cells and muon chambers to find e/γ/τ-jet-μ candidates above thresholds

Level-2 uses Regions of Interest as identified by Level-1

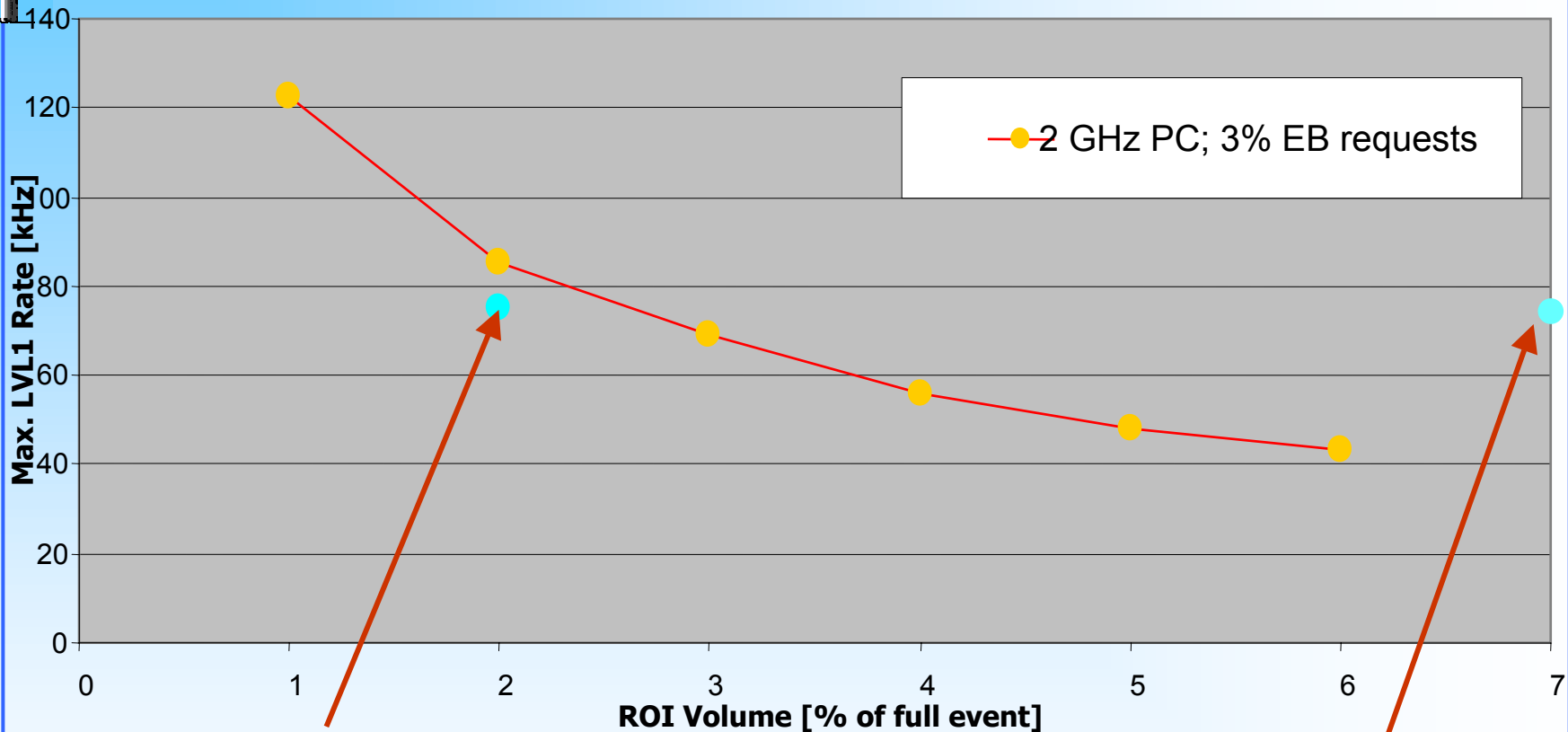
- Local data reconstruction, analysis, and sub-detector matching of RoI data

The total amount of RoI data is minimal

- ~2% of the Level-1 throughput but it has to be extracted from the rest at 75 kHz



ROS Performance



ATLAS baseline conditions

(from paper model: contains safety factor 4 with respect to physics simulation – for typical ROS)

ROS in high |η|-region

ATLAS commissioning

Phase A
System at ROD level.
Systems for LVL1, DCS and DAQ.
Check cable connections.
Infrastructure.
Some system tests.

Phase B
Calibration runs on local systems.

Phase C
Systems/Trigger/DAQ combined.

Phase D
Global commissioning.
Cosmic ray runs.
Initial off-line software.
Initial physics runs.

The discussions and the planning for the commissioning phases of the experiment have started in the Collaboration at many levels