The ATLAS Data Acquisition and High Level Trigger Systems: Experience and Upgrade Plans

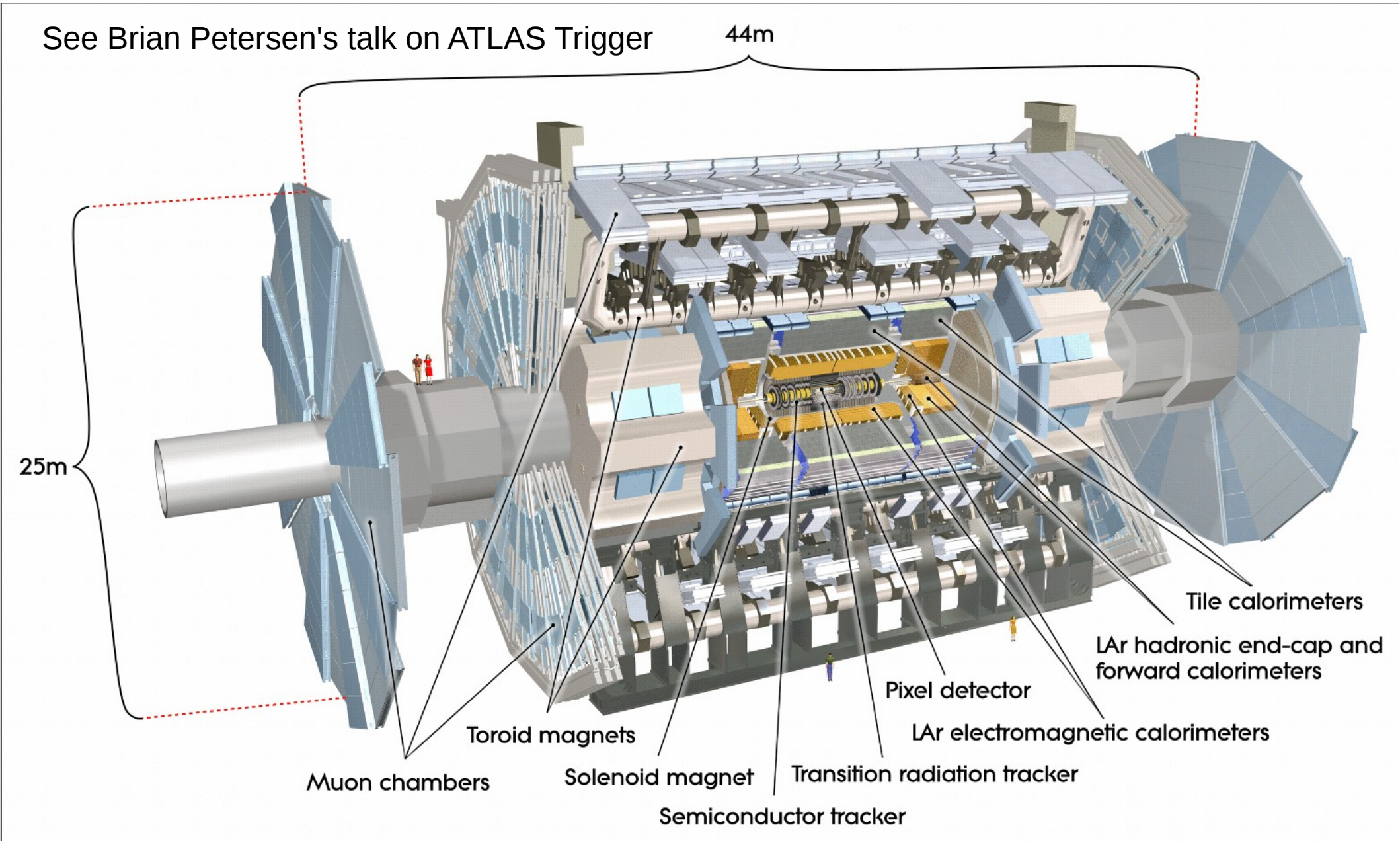
Reiner Hauser
Michigan State University

for the ATLAS Collaboration

36th International Conference on High Energy Physics
July 4th-11th , 2012
Melbourne

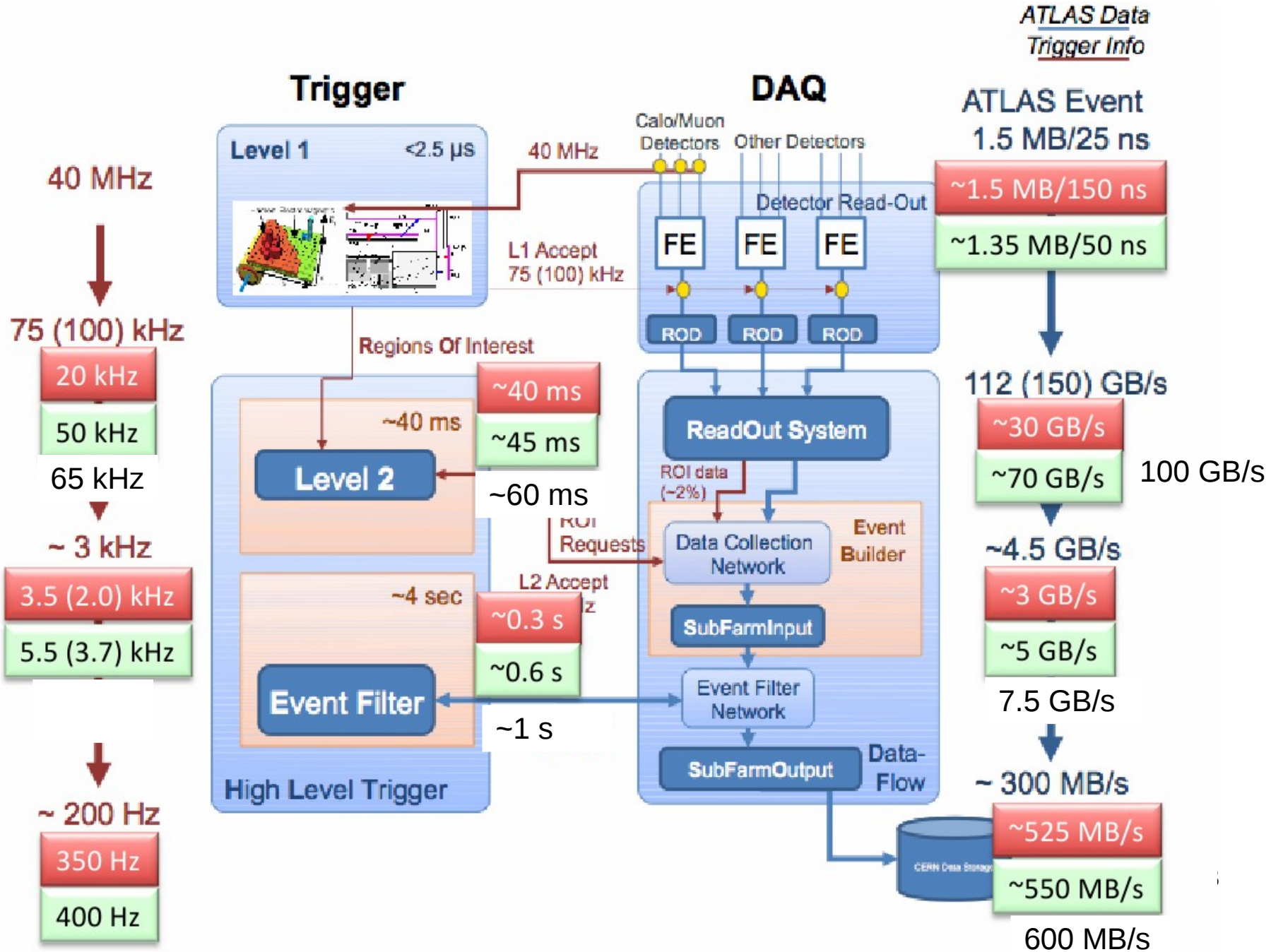
The ATLAS Experiment

See Brian Petersen's talk on ATLAS Trigger



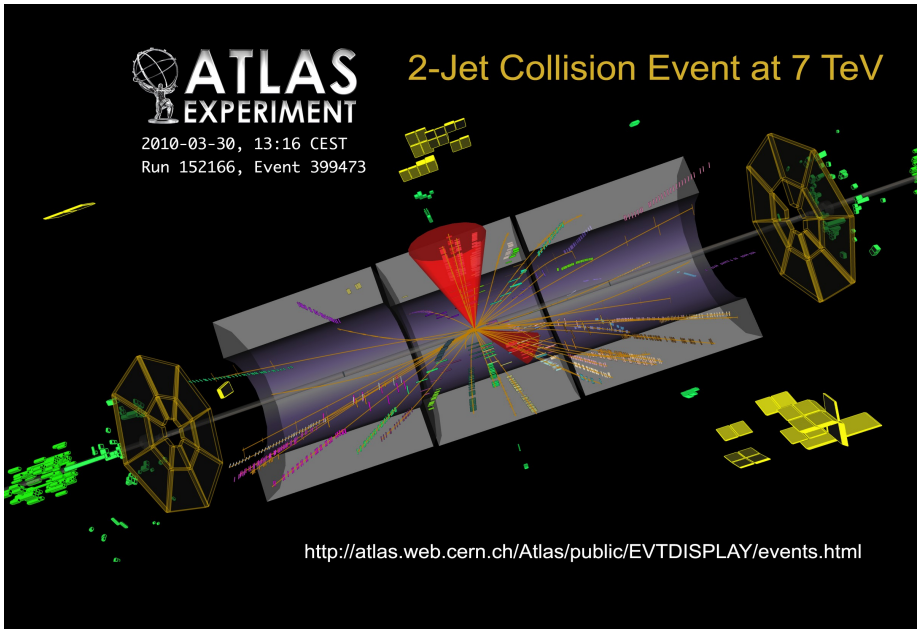
The ATLAS Trigger/DAQ System

Design
2010
2011
2012



Region of Interest

- The Level 1 trigger provides Regions of Interest (RoI) in η/ϕ.
- RoIs are assembled at 100 kHz by RoI Builder (hardware).
- The Level 2 trigger uses these to seed the algorithms and accesses mostly only data related to the RoI.

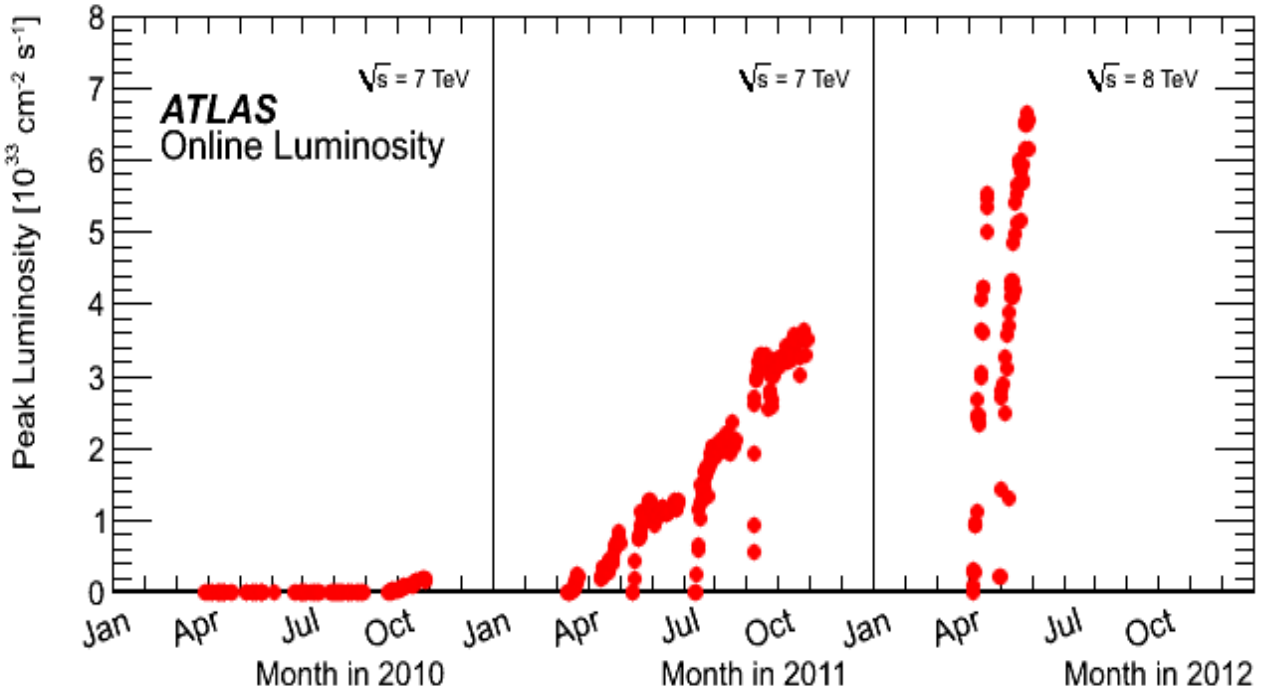


- Only about ~5-10 % of full data is read from Read-out System for rejected events.
- Full scan of a sub-detector is possible, e.g. inner detector at low luminosity.
- Level 2 accepted events are fully built and send to the event filter farm.
- In event filter more “off-line” like algorithms which can use the full event data.
- Accepted events are sent to storage and Tier-0.

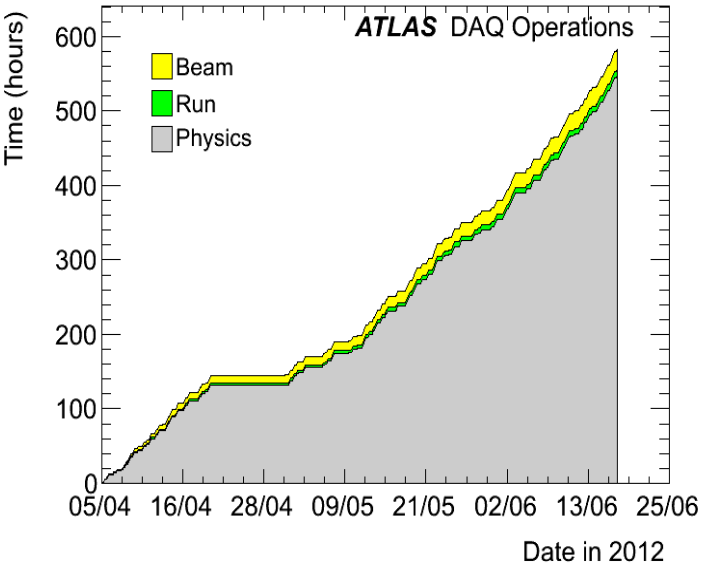
Performance from 2010 to 2012

DAQ Efficiency over 3 years: ~94 %

Includes both p-p and heavy ion runs.



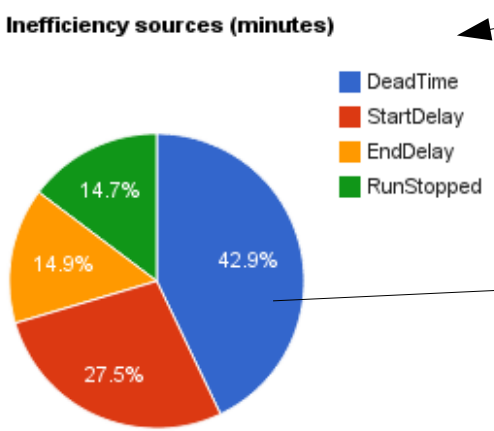
Peak luminosities for 2010-2012



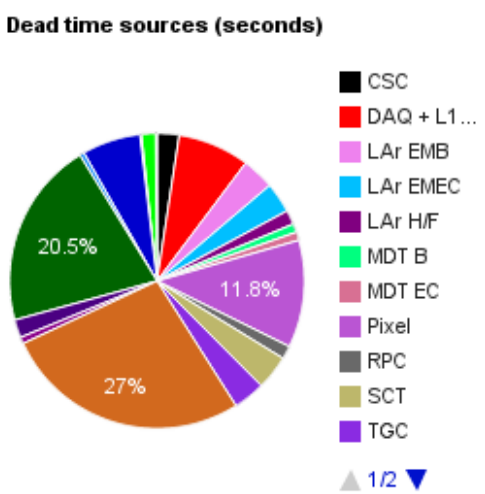
Data taking Efficiency (Time, not luminosity weighted)

Data Taking Inefficiencies

Start Date	End Date	Beam Time (hours)	Beam Time noWarmStop (hours)	Running Time (hours)	Running Time Physics (hours)	Eff (percent)	Eff-Phys (percent)	Eff-Phys-noWarmStop (percent)
Mar 30, 2010	Jun 18, 2012	3214.3	3182.5	3071.4	3000.6	95.6	93.4	94.3



Break down of ~6 % inefficiency



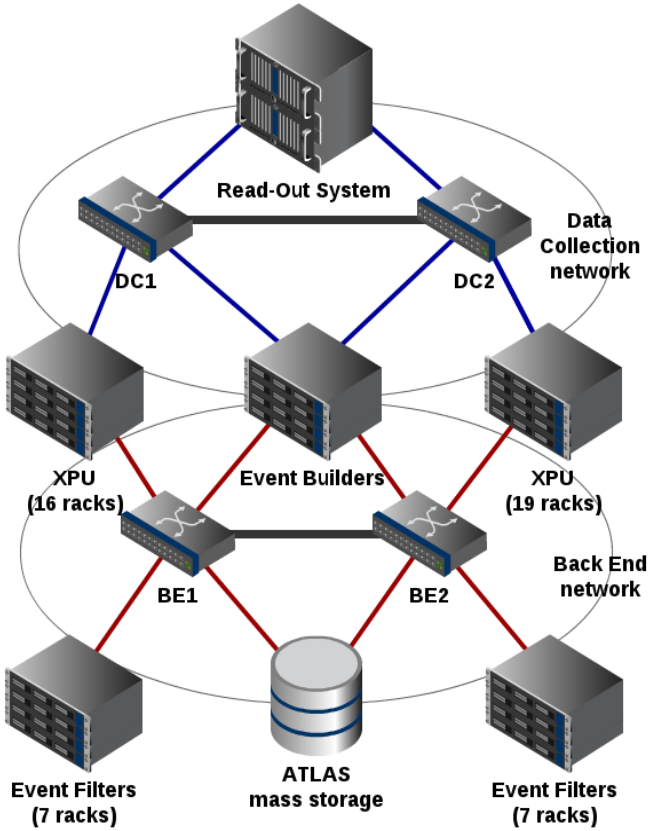
Breakdown of dead time

Source	Seconds	Percent
TRT	130395.3	27
CTPMI_VME	98866.2	20.5
Pixel	56749.4	11.8
DAQ + L1Calo	37832.6	7.8
SIMPLE	30814.9	6.4
SCT	18202.7	3.8
LAr EMB	17076.6	3.5
TGC	16773.7	3.5
LAr EMEC	16060.7	3.3
CSC	11631.5	2.4
Tile LB	9542.2	2
LAr H/F	7555.1	1.6
DAQ	7480.2	1.6

Data taking efficiencies includes all losses due to start/stop delays (ramping of HV), as well As dead time caused by the detector subsystems.

Hardware Changes during the Recent Years

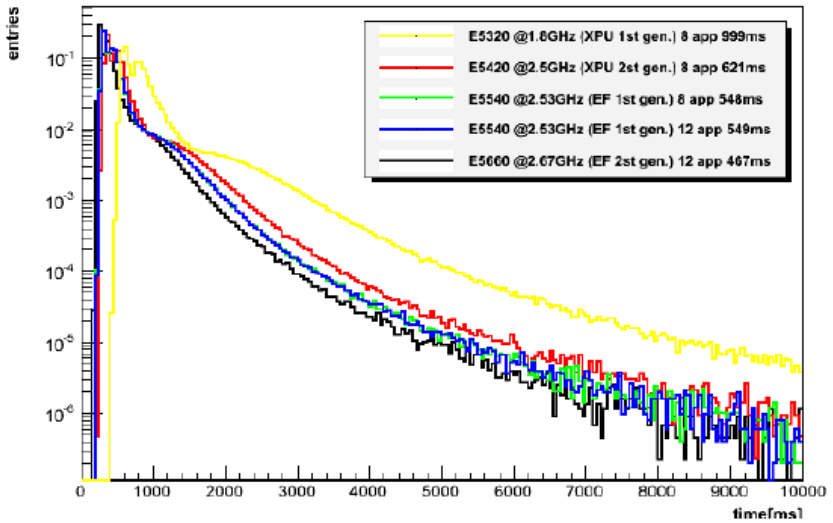
- Increasing the size of HLT farm towards nominal size (4 racks -> 50 racks).
- Regular rolling update of HLT nodes (every 3, now 4 years)
- Complete replacement of event builder nodes.
- Complete replacement of run control, monitoring nodes, and supervisor nodes (~64 + 12)
- Rolling update of Read-out System motherboard (150).
- Data logger update, hardware (5, in 2010) and disks (2011) for additional capacity; doubled network capacity.
- Additional back-end switch for redundancy.

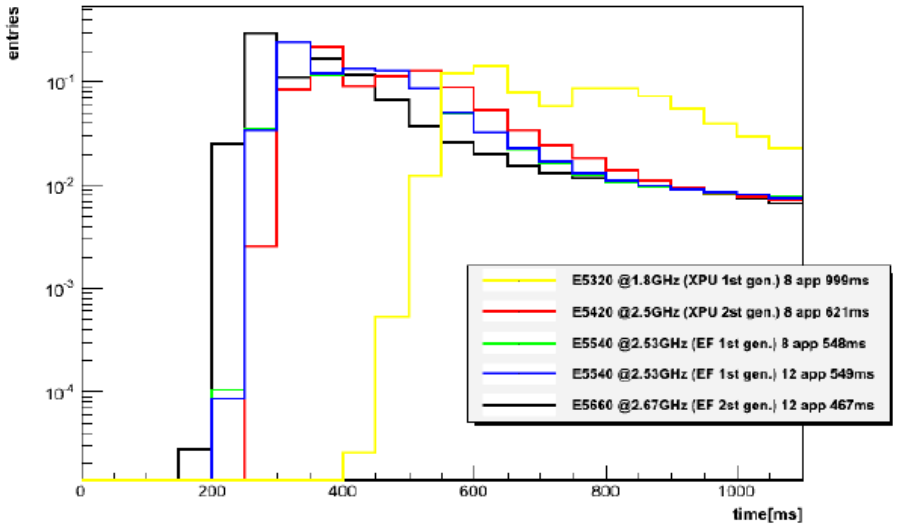


HLT Nodes:
XPU = can be used for LVL2 and EF
EF = can be used only for Event Filter

CPU model	Cores /node	Racks	Nodes	Usage
E5420	8	11	341	XPU
E5540	8	14	448	EF
X5650	12	25	904	XPU

Example: HLT Farm Expansion and Event Filter Performance





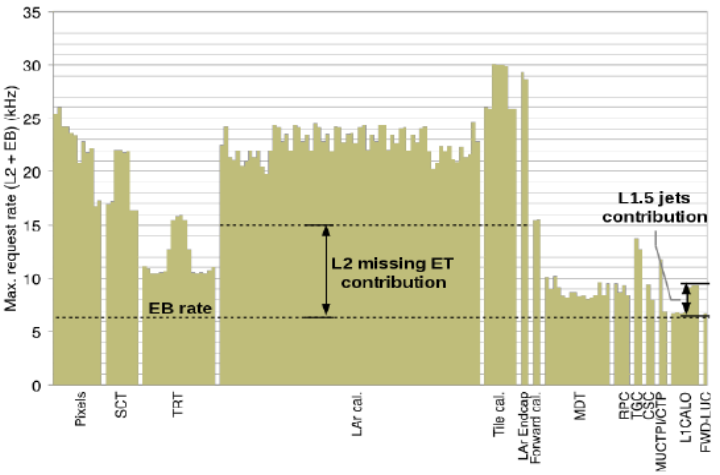
- Different machine generations
  - 1st generation XPUs
    - 2 socket Xeon E5320
    - 4 cores 1.8GHz
    - 16GB RAM
  - 2nd generation XPUs
    - 2 socket Xeon E5420
    - 4 cores 2.5GHz
    - 16 GB RAM
  - 3rd generation EFs
    - 2 socket Xeon E5540
    - 4 cores+HT 2.5GHz
    - 24 GB RAM
  - XPU updates and EF expansion:
    - 2 socket Xeon X56xx 2.6GHz
    - 6 cores+HT
    - 24 GB RAM

Load balancing such a heterogeneous farm efficiently required tuning of our software.

Data Flow Software Changes – too many to list them all...

• Lots of tuning of kernel and network related parameters (interrupts, socket buffers, quality of service).
• Re-implementation of some data flow libraries for more efficient scheduling of HLT nodes.
  • Giving up logical “sub-farms” => handle whole Event Filter as a single farm
• Possibility of per-event compression via zlib.
  • Now done at Tier-0 but possibility to do it online in the future.

- Example: Missing ET trigger for Level 2
  - Challenge: missing ET is a global property, requiring all of the calorimeter data. Lar + Tile calorimeter ~ half of the read-out system in ATLAS.
  - Solution:
    - Front-end boards calculate local E_T.
    - Read-out system extracts information and sends it to Level 2 instead of full event data
  - Required changes from front-end firmware to HLT algorithms ~ 1 year to get in production.



Observations about the current system

- Originally Level 2 was foreseen to have an “online” environment different from Event Filter.
- By ~2005 it became clear that **the same software environment (Athena/Gaudi) can be used in both Level 2 and Event Filter**.
  - **Same framework for configuration, menu steering, algorithm execution for both systems**.
  - Only difference is that Level 2 algorithms know about RoI style access.
- **Both Level 2 and event building access the Read-out system in the same way (request data with list of read-out buffer identifiers)**
- => Use long shutdown to evolve the system to a simpler and more efficient one.
- Assumptions for 2014:
  - L1: 100 kHz, rate to mass storage : 1 kHz

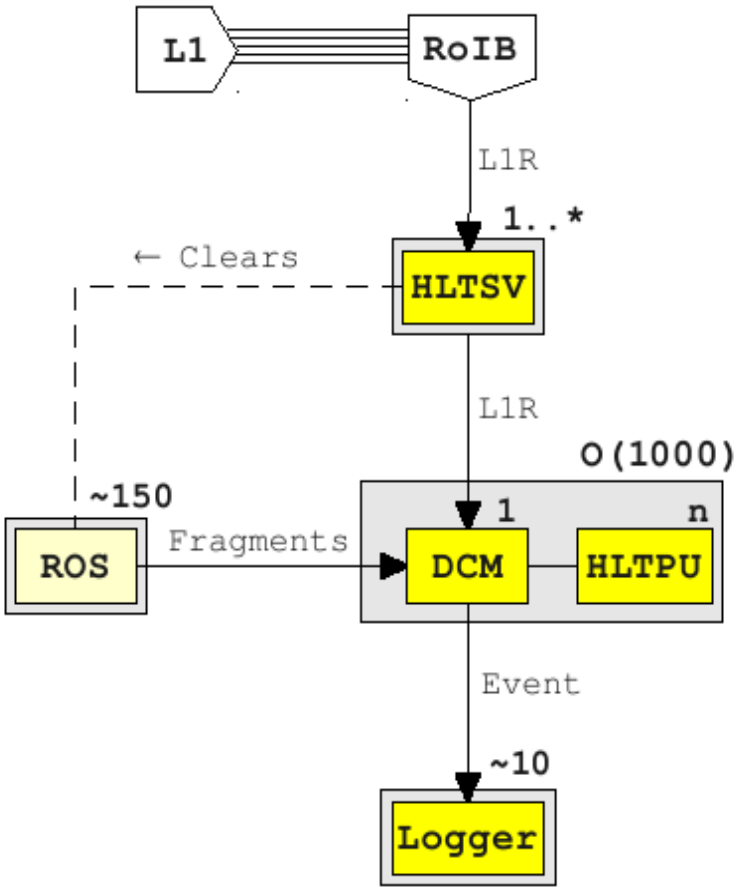
Evolution of the Data Flow Architecture

•Main ingredients:

- Merge Level 2 and Event filter processing into one application (HLTPU).
- Provide RoI based data access and event building through a single application running on the same node: Data Collection Manager (DCM).
- Connect all HLT nodes uniformly to the network.
- To first order, Read-out system and data logger stay the same.

• Reduces number of data flow applications by almost a factor 2.
• Reduces the number of TCP connections to ROS.
• Provides a much simpler system with more flexibility.

- No need to configure system by hand into Level 2 and Event filter – automatic load balancing



Evolution of the RoIBuilder/Supervisor

- Assembles and distributes the RoI information from Level 1 to a set of Level 2 supervisor nodes. The latter assign events to farm nodes.
- Currently the only custom hardware in DAQ/HLT except for Read-out system input.
  - Designed to run at 100 kHz.
  - Running reliably with basically no problems.
  - Coming of age (in terms of components availability etc.)
  - Missing some desired flexibility and error handling/recovery.
- A single supervisor for the new architecture can handle events through the network at > 100 kHz.
- Input (~8 fibers from Level 1) might be possible to handle in software (under study).
  - => Ideally combine RoIB and Supervisor in one software solution.

Evolution of the HLT Application

- Now ~2 Gbyte/HLT application budgeted.
- Move to many-core systems and 64 bit software → not sustainable.
- Common effort with ATLAS offline software framework:
  - Mother process initializes and configures itself.
  - Forks child processes that share a large part of the address space with parent.
  - In case of crash a new child process can be spawned quickly.
- Save memory, save resources during configuration (like database connections).
- Fast restart of new child in case of crash.

Further optimizations in HLT

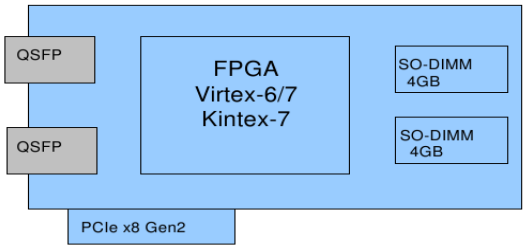
- Avoid duplicate work done in both Level 2 and EF
  - Data unpacking and preparation, like creation of space points for inner detector.
- No hard boundary between Level 2 and EF – can request more data as needed without insisting on building full event.
  - Optimize when to initiate event building.
- Possible optimizations in steering
  - Example: when first trigger chain accepts event can start doing event building, while the rest of the “RoI” based chains still continue.

Evolution of the Network

- Core switches (now 2 for Level 2/Event building, 2 for back-end) will be replaced during long shutdown.
- Re-cabling for new data flow architecture.
  - No distinction between data collection and back-end network.
- Greater use of 10 Gbps Ethernet where appropriate and necessary.
- For the present retain a data flow architecture that keeps RoI concept.
  - Also safe-guard for future ATLAS upgrades going beyond 100 kHz (> 2018).

Evolution of the Read-out System

- Buffers input data on board until requested from Level 2 or event builder.
- Current input board (ROBIN) is PCI based, using a PowerPC CPU and an FPGA.
- Next generation (ROBINNP) is using PCI Express
  - No on-board processor anymore.
  - More input links (8-12 vs. 3).



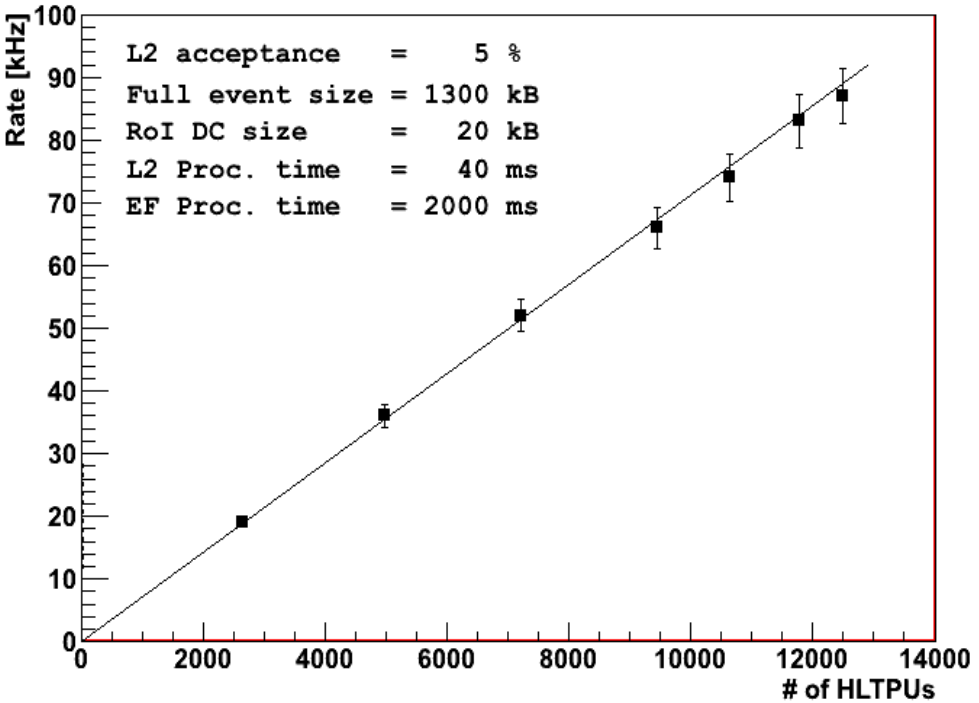
- New design is very close to a board by ALICE that is reaching prototype status.
  - If evaluation successful, might go with this board instead of new development.
- If ATLAS does its own design: probably not deployed before winter shutdown 2015
- Use 10 Gbps links into central core network

Measurements on Prototype

A prototype implementation for the data flow based on the existing software exists and has been measured with up to 13k HLTPU applications.

All data flow applications are available, the HLT processing is only emulated by burning CPU Time. Perfect scaling after adding traffic shaping has been observed.

Remainder of 2012: complete design to have a working software by beginning of long shutdown.

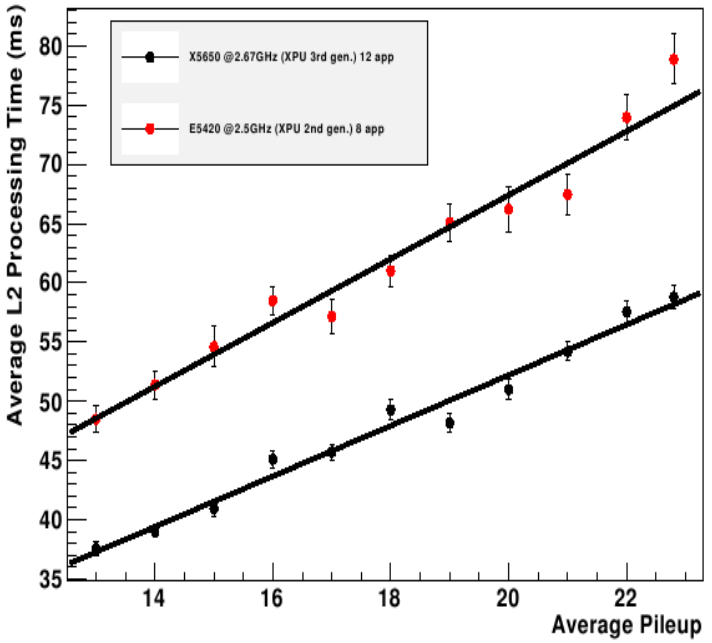


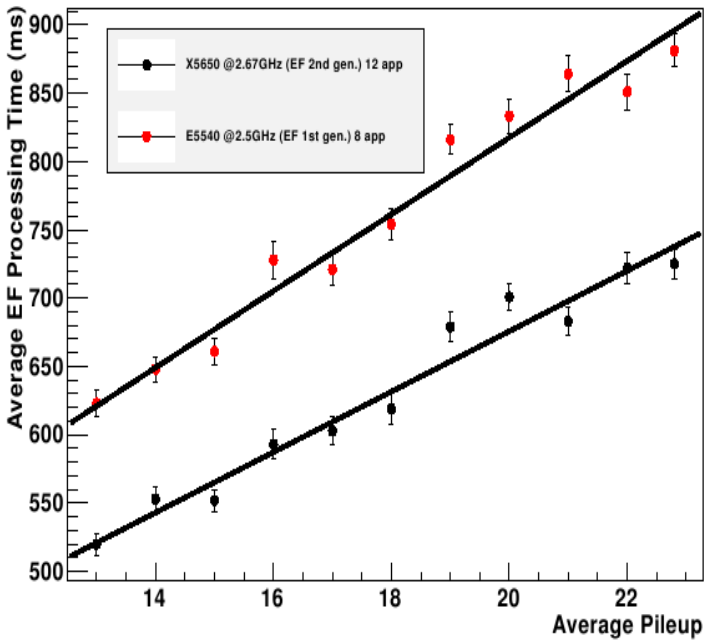
Summary

- The Atlas DAQ/HLT system performed extremely well during the last years of data taking,with an average data taking efficiency of 94%.
  - It handled all expected use cases and adapted quickly to the unexpected ones.
- Continuous software and hardware changes have allowed the system performance to match the requirements.
- The next long shutdown offers a unique opportunity to introduce bigger changes, based on what we have learned in the last years.
- It will lead to a simpler, more performant and more flexible system, ready to handle the challenges of 2014 and beyond.

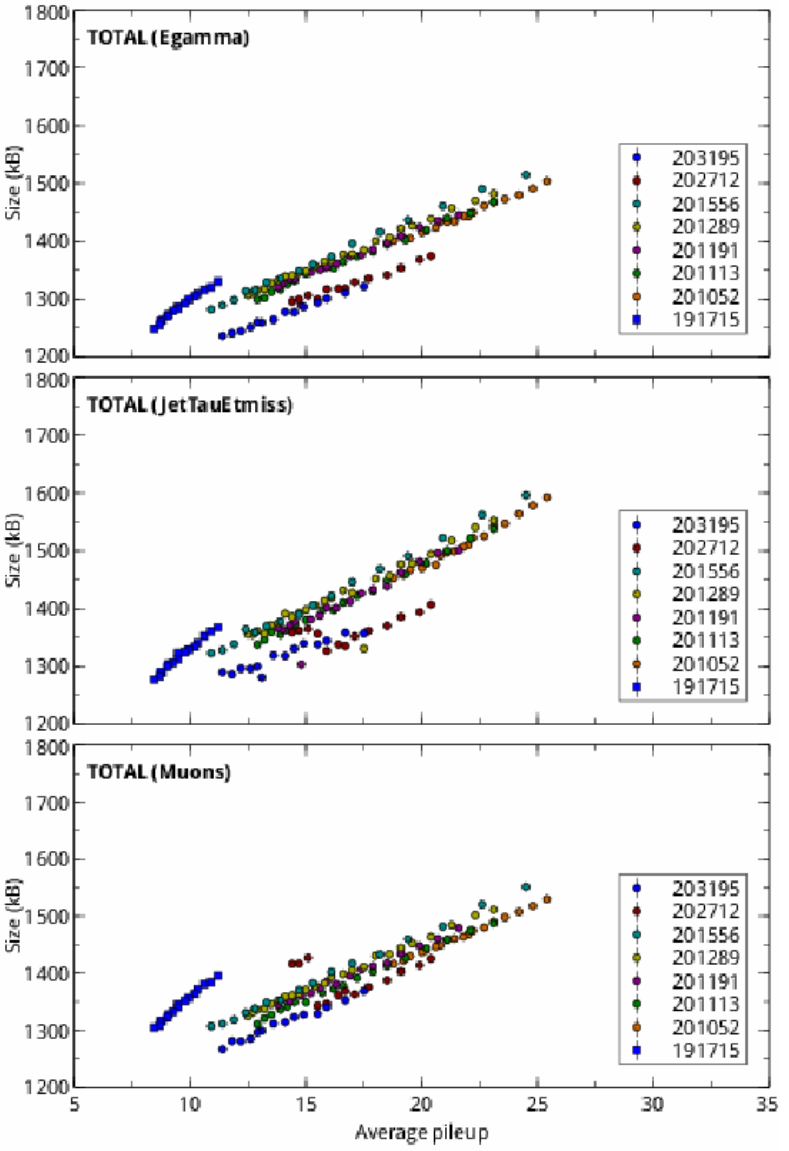
Backup

CPU Usage vs. Pile-Up

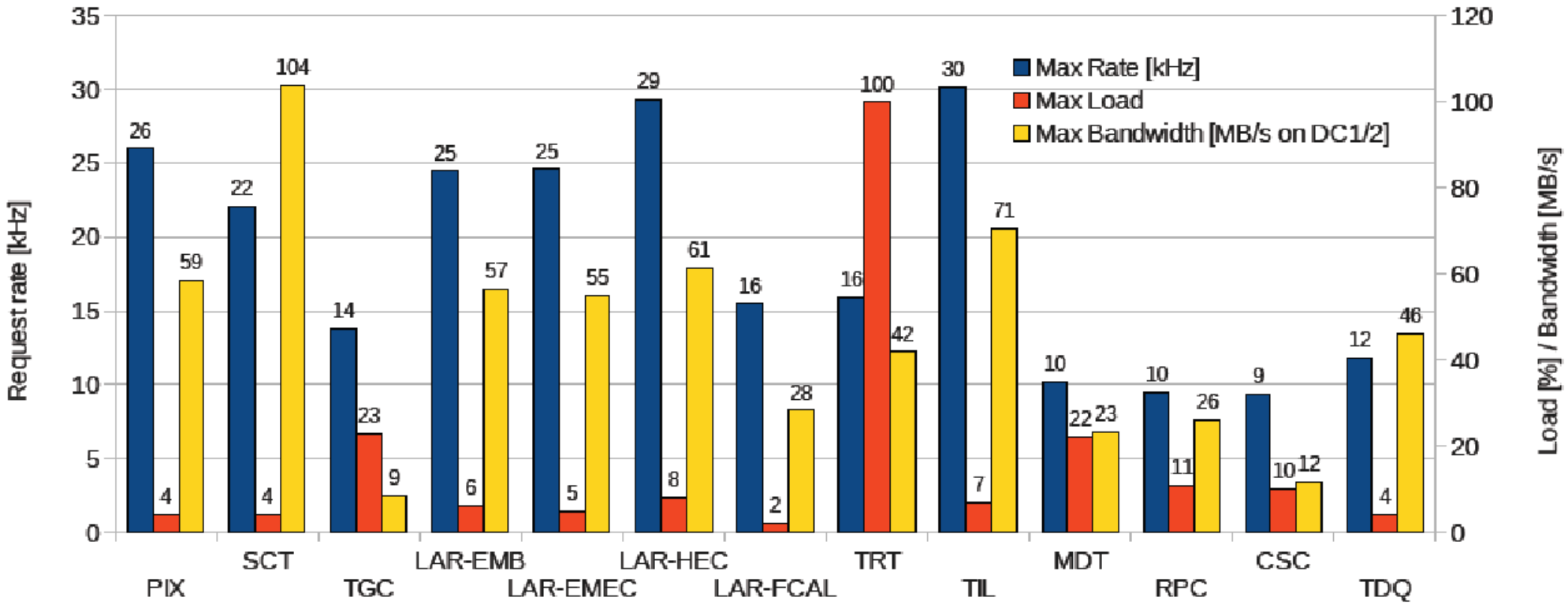




Event Size vs. Pile-Up

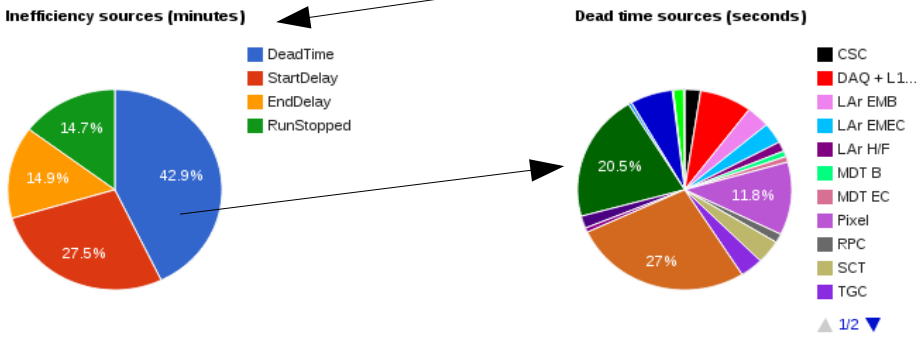


2012 ROS Request Rates



Monitoring of Inefficiencies

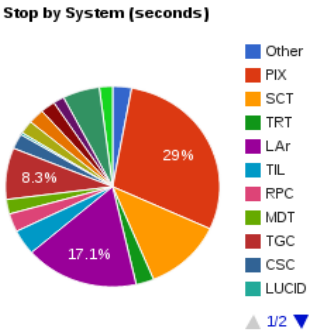
Start Date	End Date	Beam Time (hours)	Beam Time noWarmStop (hours)	Running Time (hours)	Running Time Physics (hours)	Eff (percent)	Eff-Phys (percent)	Eff-Phys-noWarmStop (percent)
Mar 30, 2010	Jun 18, 2012	3214.3	3182.5	3071.4	3000.6	95.6	93.4	94.3



Source	Seconds	Percent
TRT	130395.3	27
CTPMI_VME	98866.2	20.5
Pixel	56749.4	11.8
DAQ + L1Calo	37832.6	7.8
SIMPLE	30814.9	6.4
SCT	18202.7	3.8
LAr EMB	17076.6	3.5
TGC	16773.7	3.5
LAr EMEC	16060.7	3.3
CSC	11631.5	2.4
Tile LB	9542.2	2
LAr H/F	7555.1	1.6
DAQ	7480.2	1.6
RPC	6989.1	1.4
MDT B	4534.8	0.9

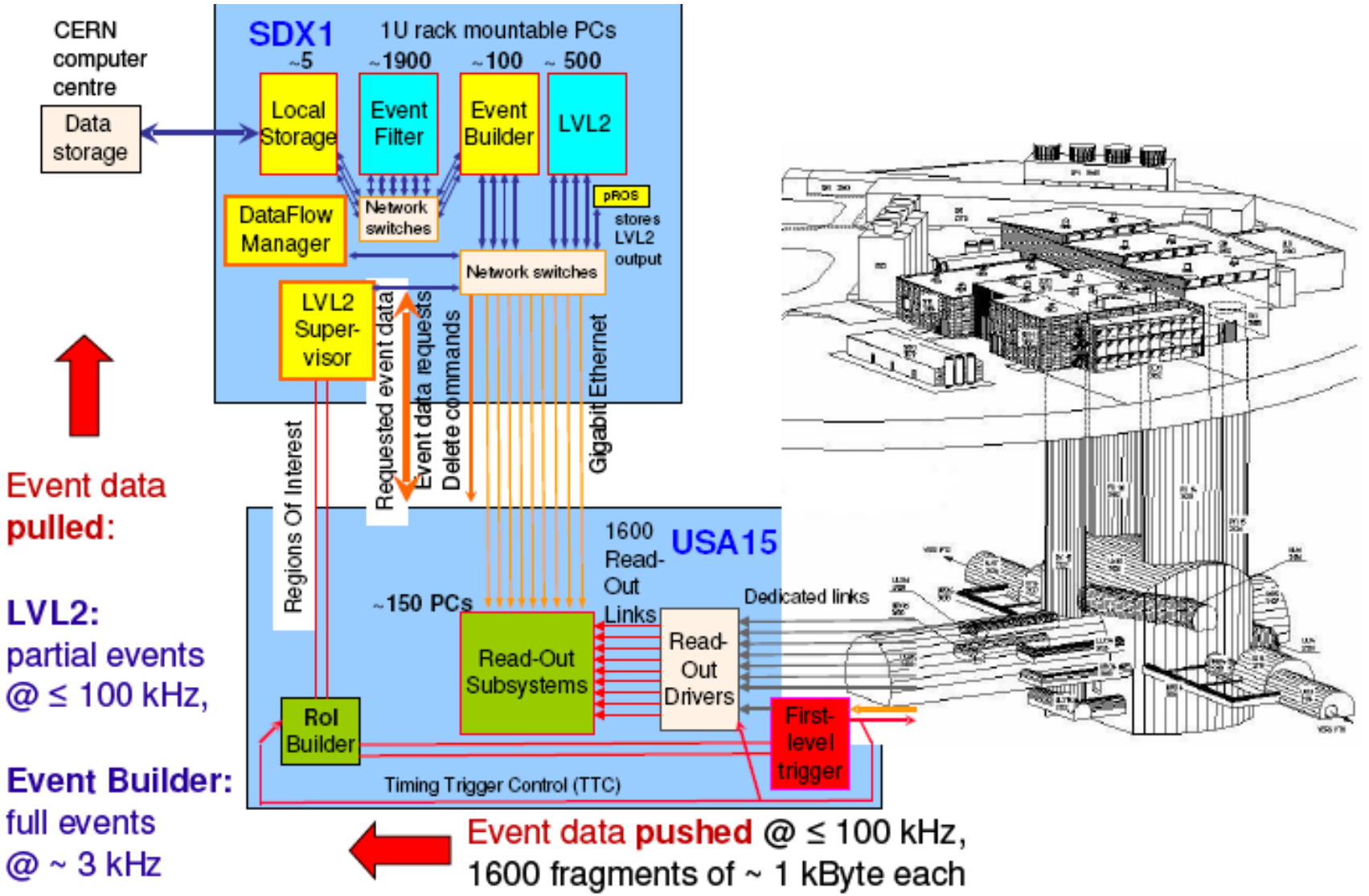
Detailed break down of inefficiencies allows to find problems and address them quickly.

Note: The TRT dead time is almost entirely in empty bunch crossings during the long beam gap and thus does not impact physics efficiency. It comes from a busy asserted to protect the read back of front end information.



System	Seconds	Percent
PIX	35582	29
LAr	21014.8	17.1
SCT	14552.1	11.9
TGC	10182.6	8.3
DAQ	6828.3	5.6
TIL	4947	4
RPC	3619.8	3
Other	3499.5	2.9
TRT	3306.4	2.7
CSC	2963.9	2.4
L1CALO	2956.3	2.4
MDT	2864	2.3
L1CT	2709.1	2.2
ZDC	2591.6	2.1
LHC	2442.9	2

Location of TDAQ Systems



Logical Data Flow