CHEP06 - Mumbai / INDIA

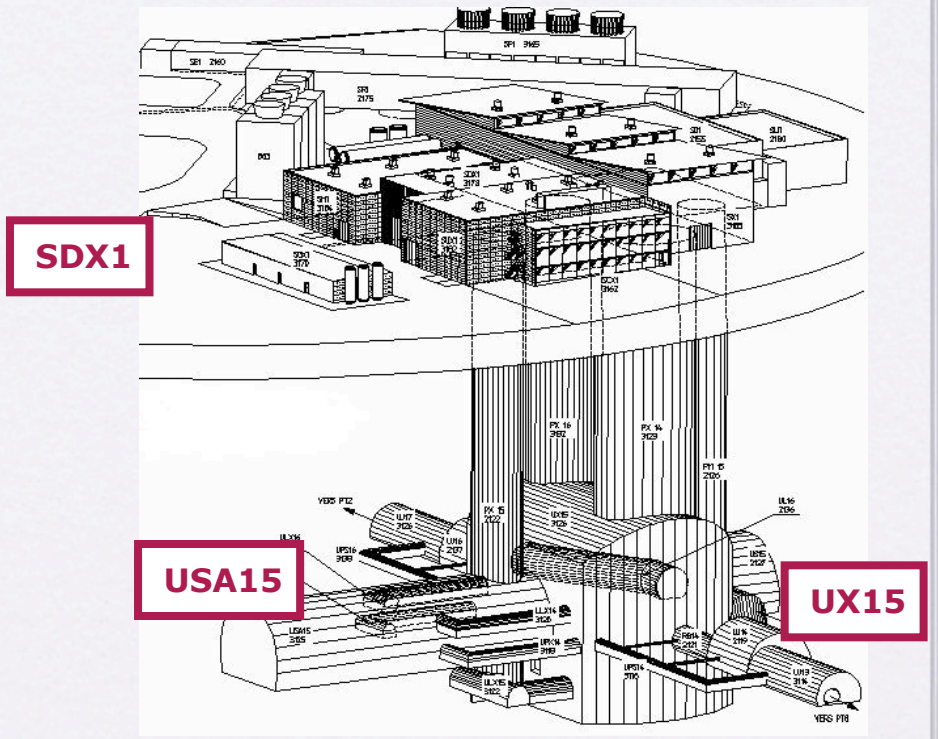
Studies with the ATLAS Trigger & Data Acquisition "pre-series" setup

N. Gökhan Ünel (UCI & CERN)

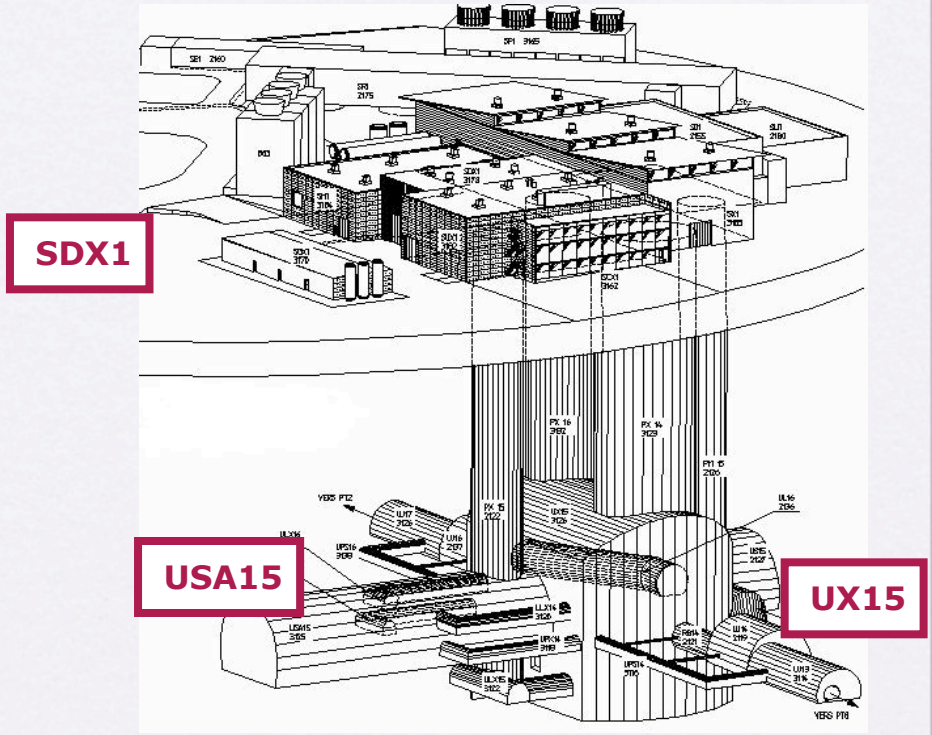
on behalf of

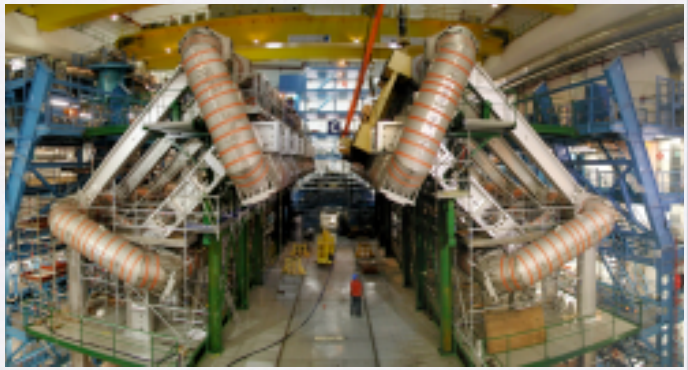
ATLAS TDAQ community

ATLAS Trigger / DAQ

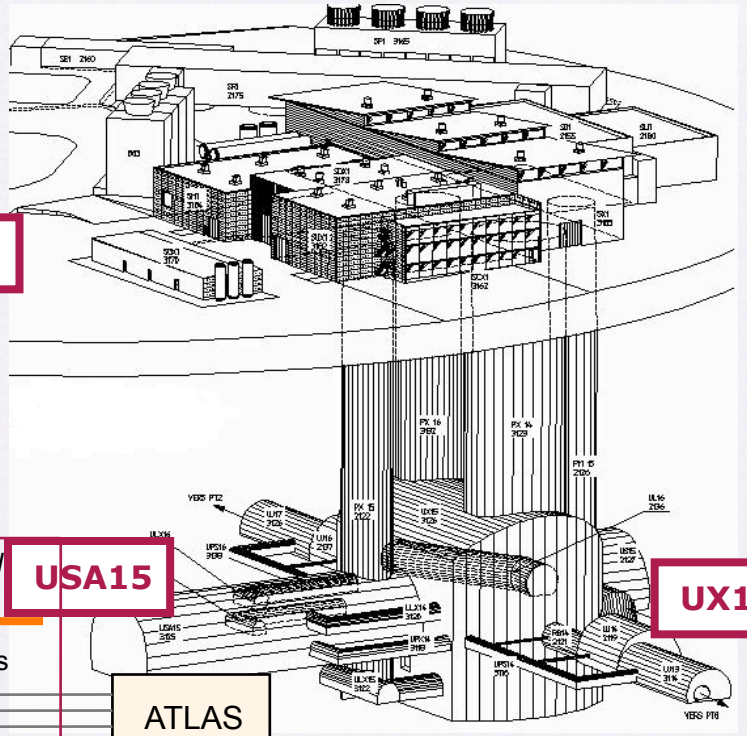


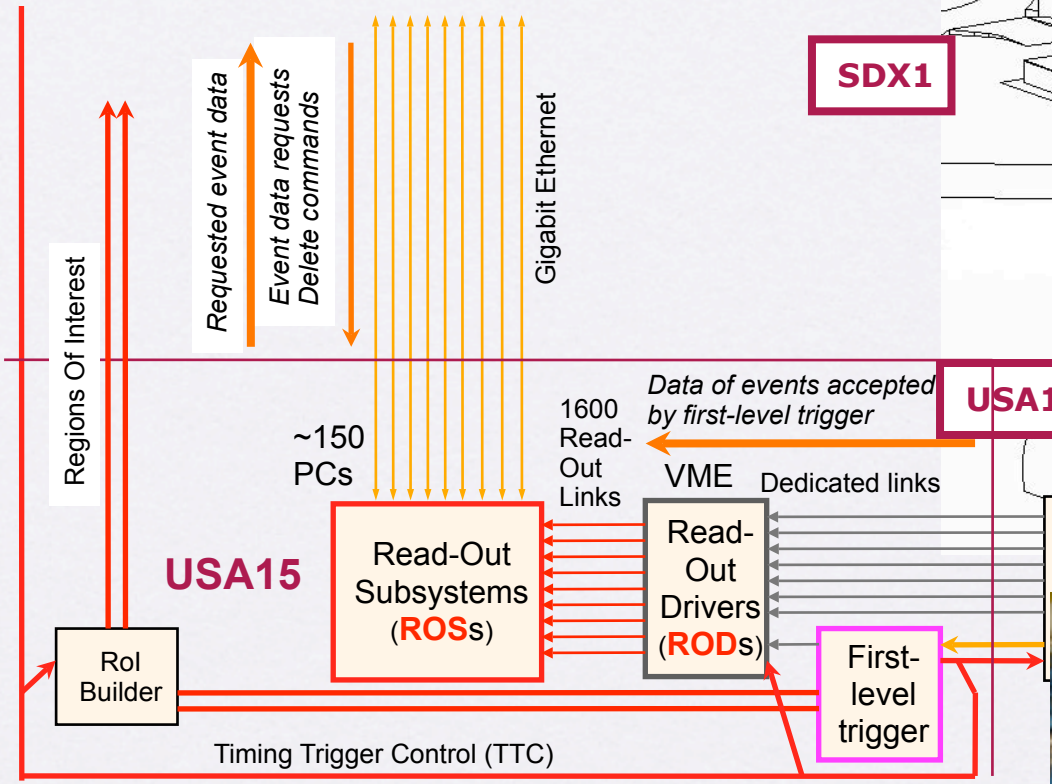
ATLAS Trigger / DAQ

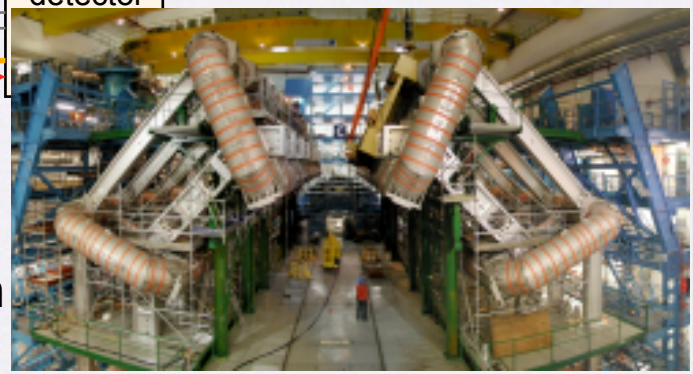




ATLAS Trigger / DAQ

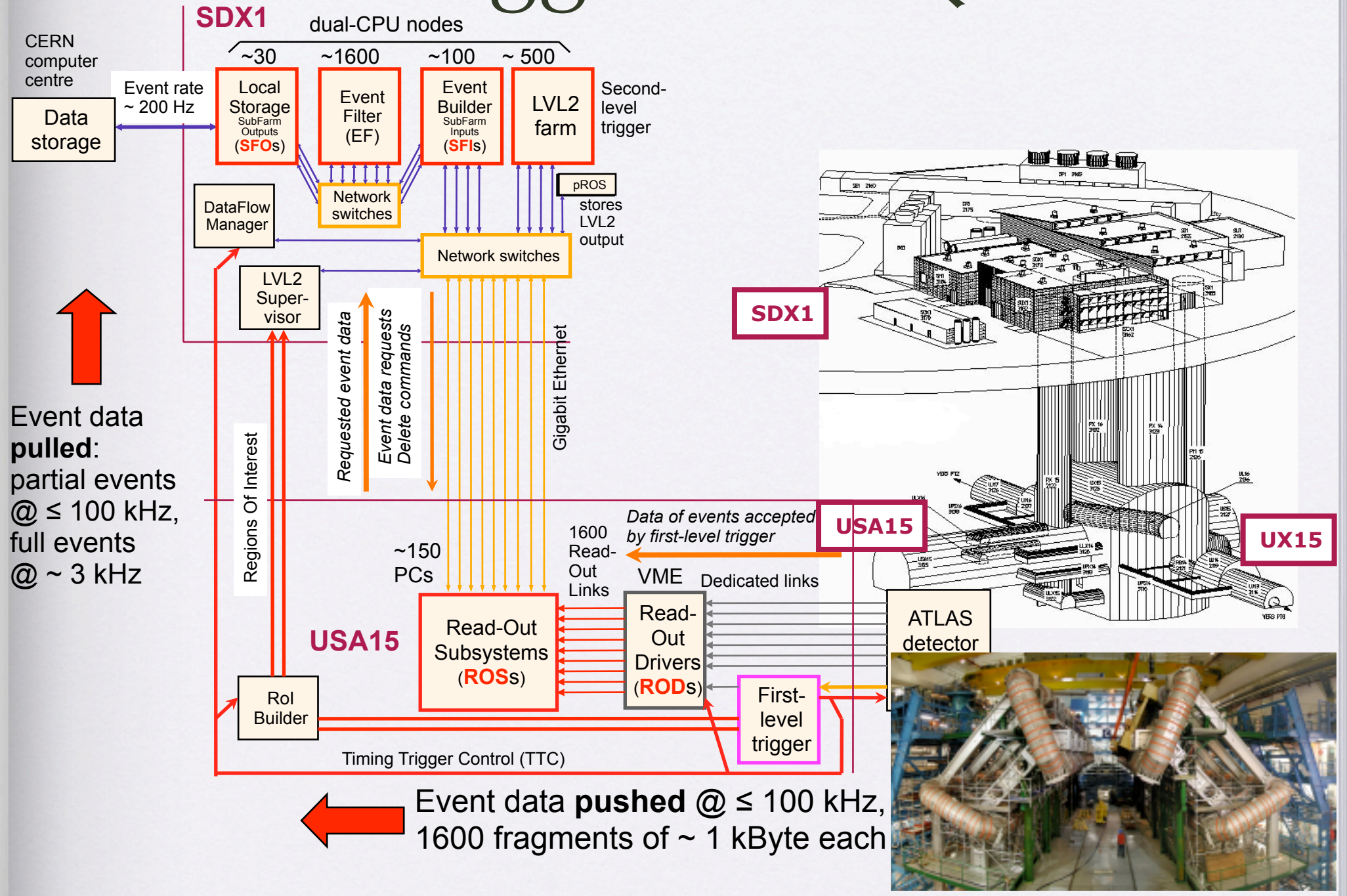






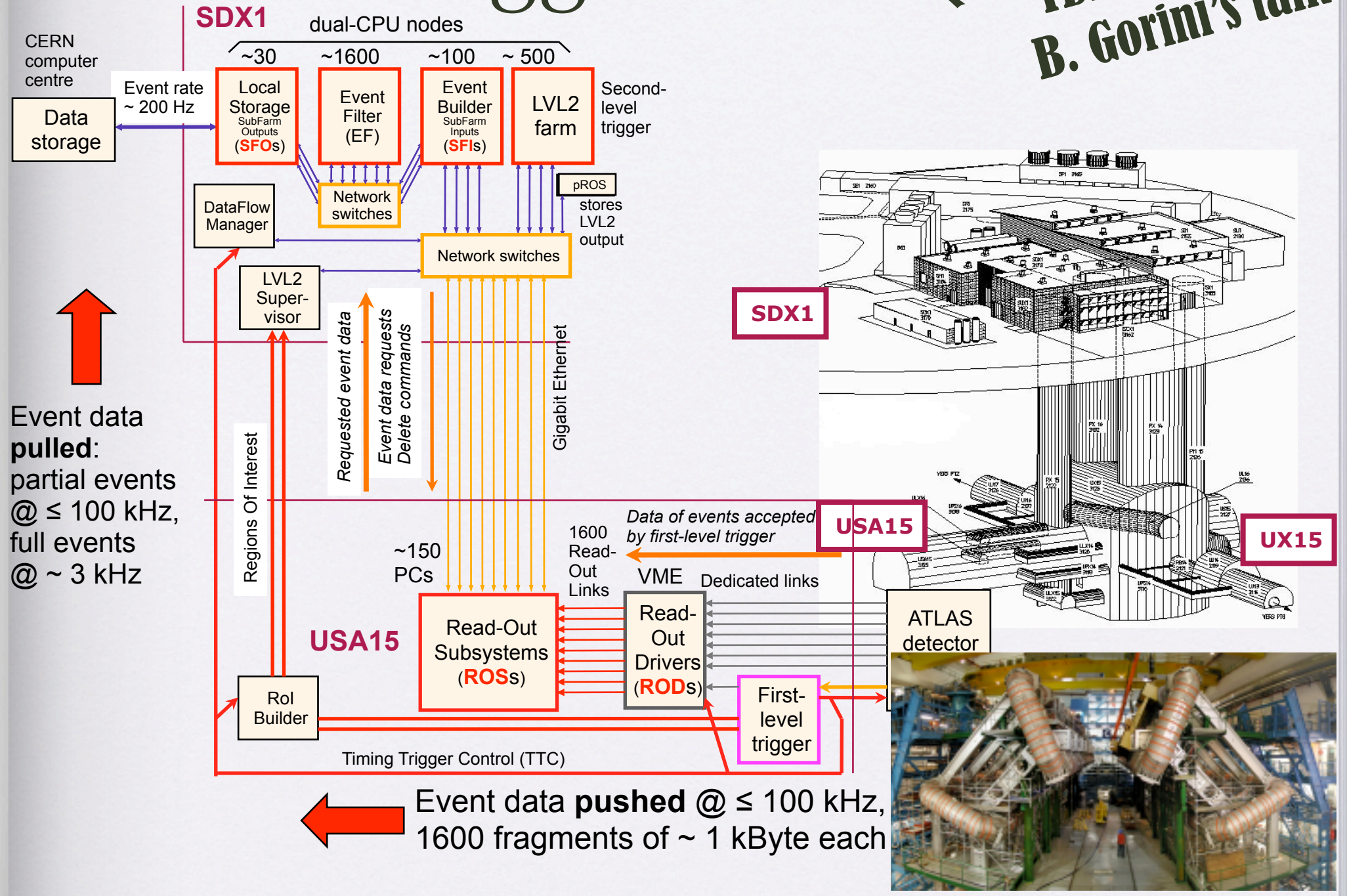
Event data **pushed** @ ≤ 100 kHz,
1600 fragments of ~ 1 kByte each

ATLAS Trigger / DAQ



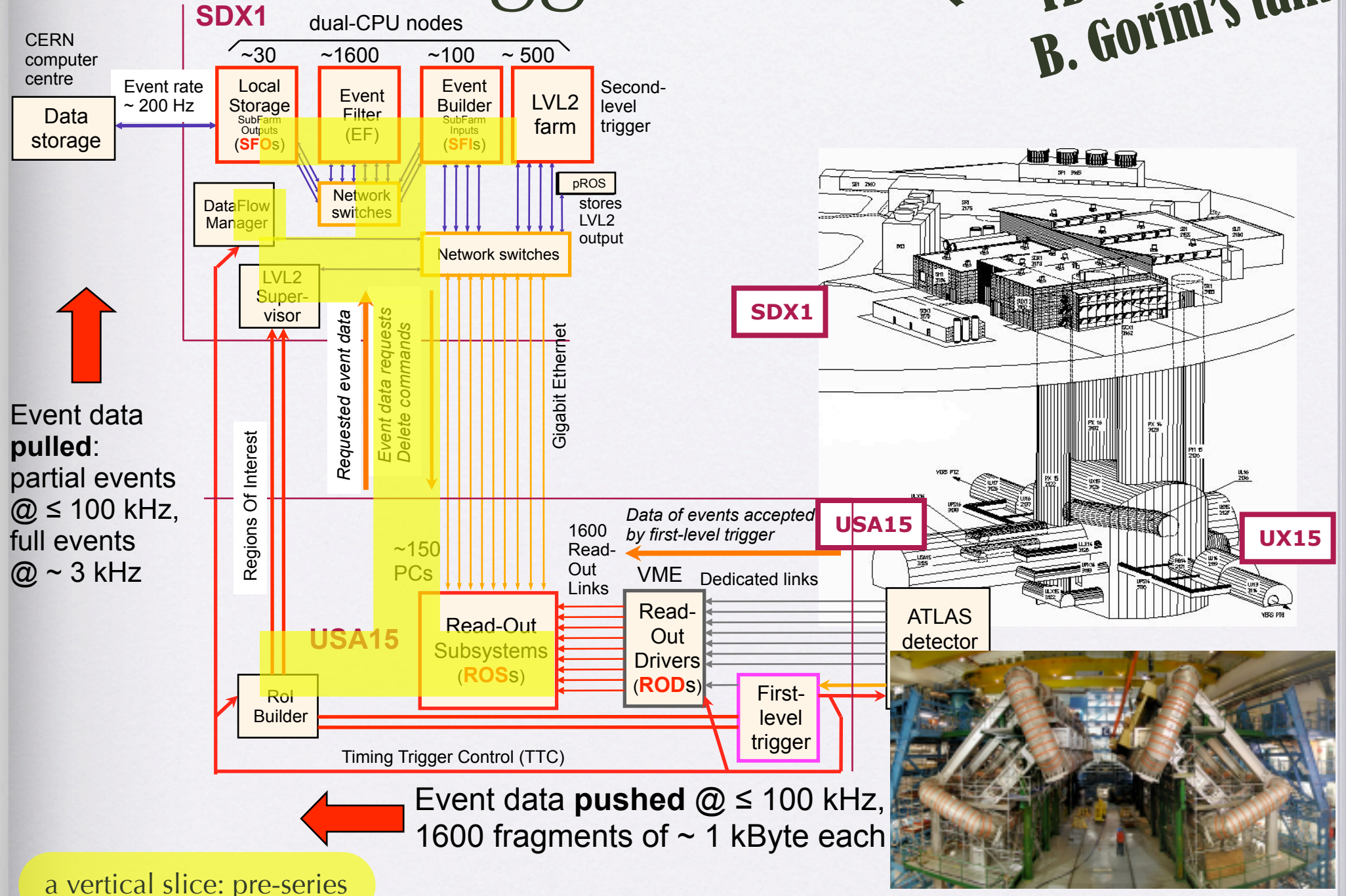
ATLAS Trigger / DAQ

TDAQ Details: B. Gorini's talk



ATLAS Trigger / DAQ

TDAQ Details: B. Gorini's talk



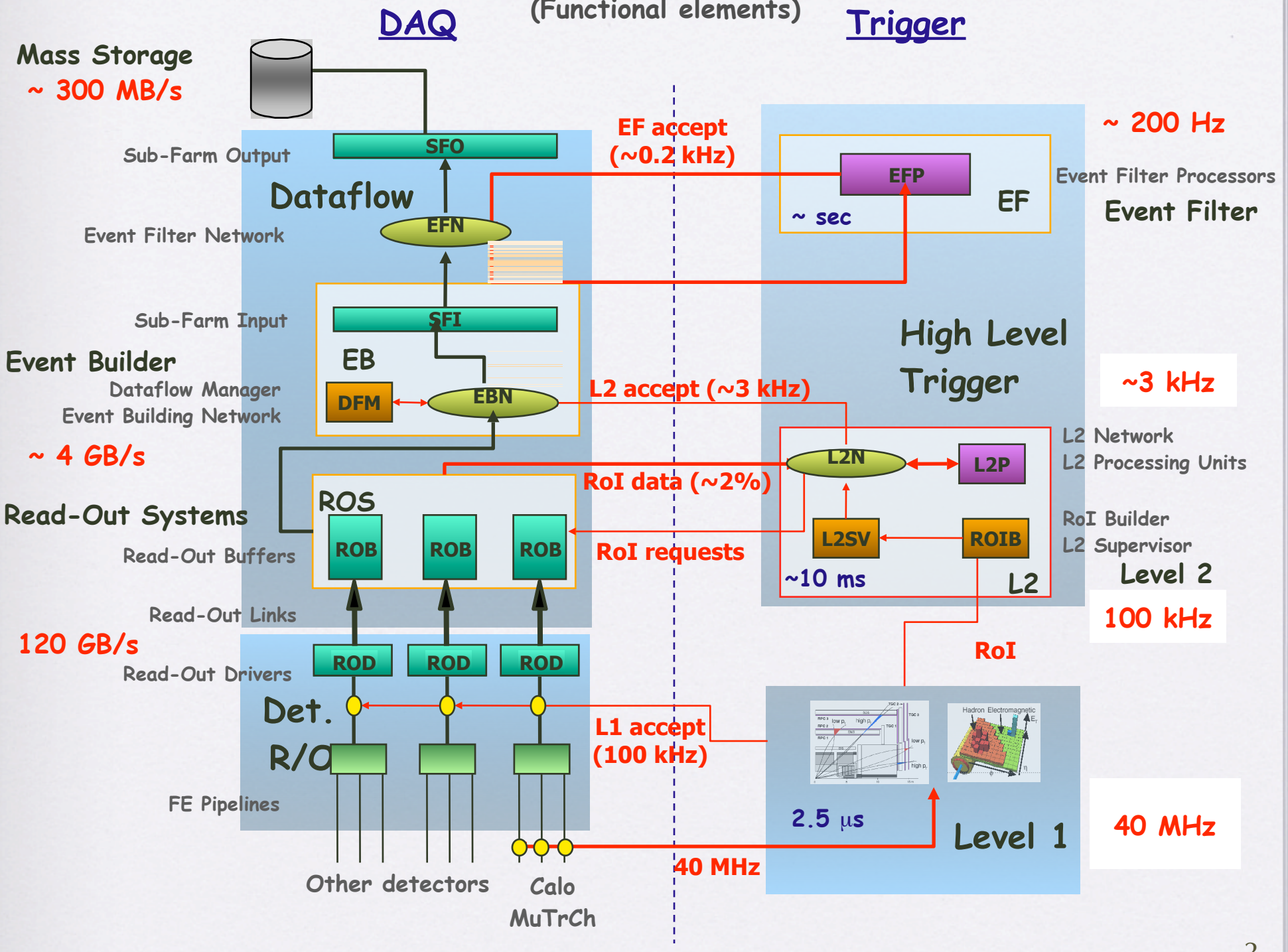
Event data **pulled**:
partial events
@ ≤ 100 kHz,
full events
@ ~ 3 kHz

Event data **pushed** @ ≤ 100 kHz,
1600 fragments of ~ 1 kByte each

a vertical slice: pre-series

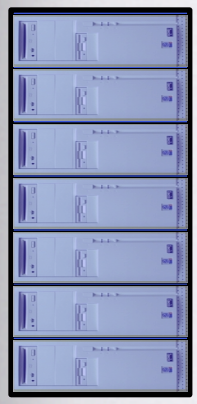
ARCHITECTURE

(Functional elements)

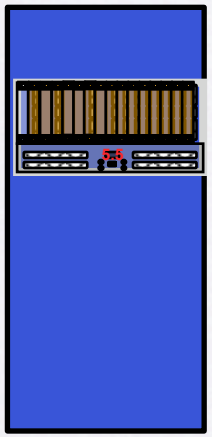


Pre-series test bed

- 8 racks at Point-1 (10% of final dataflow)
- to exercise the TDAQ before detector integration
- to estimate the quantity and characteristics of TDAQ components



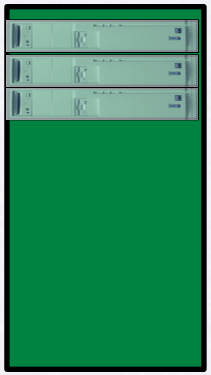
One ROS rack
-
TC rack + horiz. Cooling
-
12 ROS
48 ROBINs



RoIB rack
-
TC rack + horiz. cooling
-
50% of RoIB

underground : **USA15**

One Full L2 rack
-
TDAQ rack
-
30 HLT PCs



Partial Superv'r rack
-
TDAQ rack
-
3 HE PCs

One Switch rack
-
TDAQ rack
-
128-port GEth for L2 +EB

Partial EFIO rack
-
TDAQ rack
-
10 HE PC (6 SFI - 2 SFO - 2 DFM)

Partial EF rack
-
TDAQ rack
-
12 HLT PCs

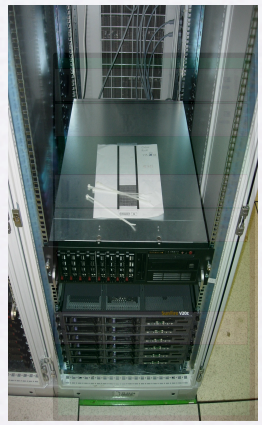
Partial ONLINE rack
-
TDAQ rack
-
4 HLT PC (monitoring)
2 LE PC (control)
1 Central FileServer

surface: **SDX1**

- **ROS, L2, EFIO and EF racks:** one Local File Server, one or more Local Switches
- **Machine Park**: Dual Opteron and Xeon nodes, ROS nodes uniprocessor
- **OS issues:** Net booted and diskless nodes (localdisks as scratch), runing Scientific Linux, Cern v3.
- **Trigger** : Free trigger from L2SV or frequency manually set using LTP

Pre-series test bed

- 8 racks at Point-1 (10% of final dataflow)
- to exercise the TDAQ before detector integration
- to estimate the quantity and characteristics of TDAQ components



One ROS rack
-
TC rack + horiz. Cooling
-
12 ROS
48 ROBINs

RoIB rack
-
TC rack + horiz. cooling
-
50% of RoIB

One Full L2 rack
-
TDAQ rack
-
30 HLT PCs

Partial Superv'r rack
-
TDAQ rack
-
3 HE PCs

One Switch rack
-
TDAQ rack
-
128-port GEth for L2 +EB

Partial EFIO rack
-
TDAQ rack
-
10 HE PC (6 SFI - 2 SFO - 2 DFM)

Partial EF rack
-
TDAQ rack
-
12 HLT PCs

Partial ONLINE rack
-
TDAQ rack
-
4 HLT PC (monitoring)
2 LE PC (control)
1 Central FileServer

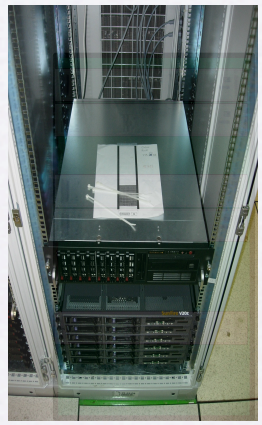
***underground* : USA15**

***surface*: SDX1**

- **ROS, L2, EFIO and EF racks:** one Local File Server, one or more Local Switches
- **Machine Park**: Dual Opteron and Xeon nodes, ROS nodes uniprocessor
- **OS issues:** Net booted and diskless nodes (localdisks as scratch), runing Scientific Linux, Cern v3.
- **Trigger** : Free trigger from L2SV or frequency manually set using LTP

Pre-series test bed

- 8 racks at Point-1 (10% of final dataflow)
- to exercise the TDAQ before detector integration
- to estimate the quantity and characteristics of TDAQ components



One ROS rack
-
TC rack + horiz. Cooling
-
12 ROS
48 ROBINs

RoIB rack
-
TC rack + horiz. cooling
-
50% of RoIB

One Full L2 rack
-
TDAQ rack
-
30 HLT PCs

Partial Superv'r rack
-
TDAQ rack
-
3 HE PCs

One Switch rack
-
TDAQ rack
-
128-port GEth for L2 +EB

Partial EFIO rack
-
TDAQ rack
-
10 HE PC (6 SFI - 2 SFO - 2 DFM)

Partial EF rack
-
TDAQ rack
-
12 HLT PCs

Partial ONLINE rack
-
TDAQ rack
-
4 HLT PC (monitoring)
2 LE PC (control)
1 Central FileServer

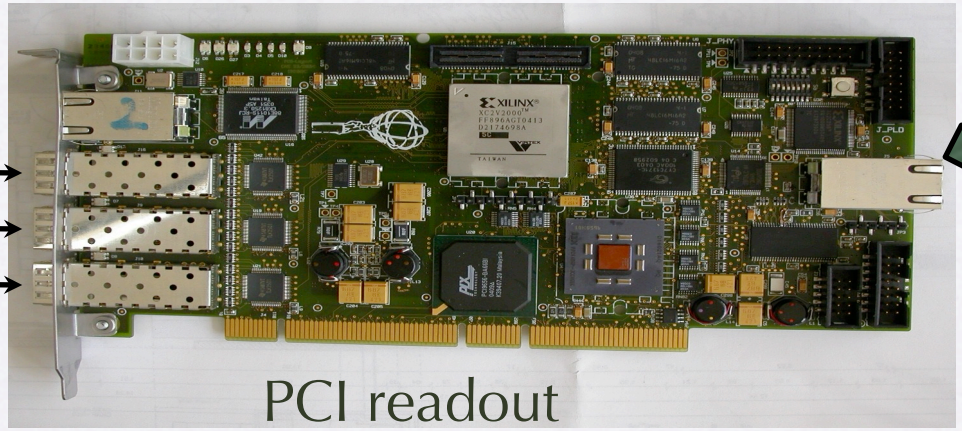
underground **: USA15**

surface**: SDX1**

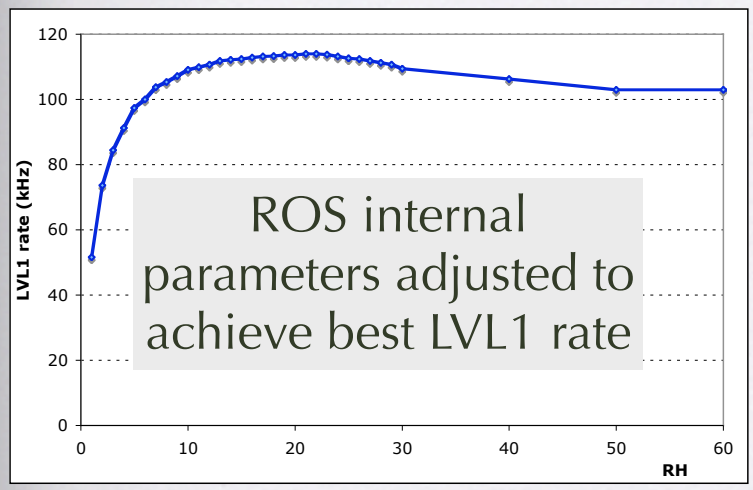
Farm management details: M Dobson's talk

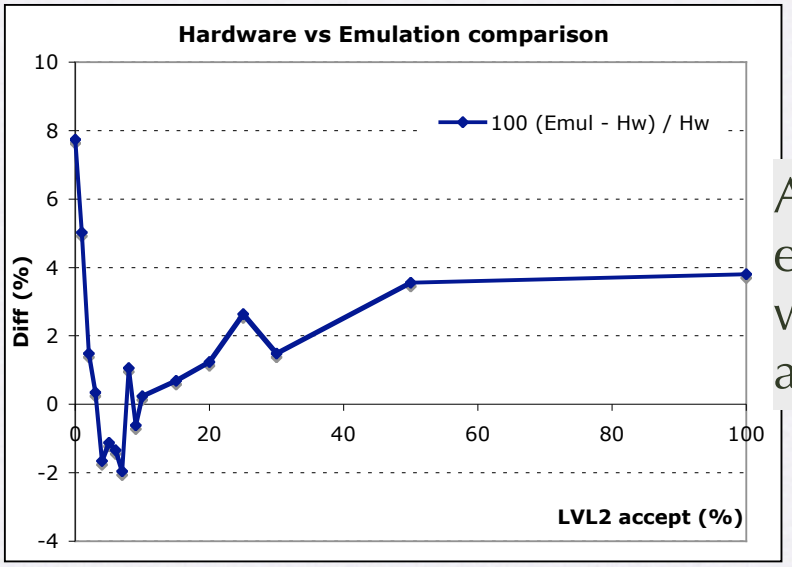
- **ROS, L2, EFIO and EF racks:** one Local File Server [illegible] more Local Switches
- **Machine Park**: Dual Opteron and Xeon nodes, [illegible] nodes uniprocessor
- **OS issues:** Net booted and diskless nodes (localdisks as scratch), runing Scientific Linux, Cern v3.
- **Trigger** : Free trigger from L2SV or frequency manually set using LTP

ROS & ROBIN basics



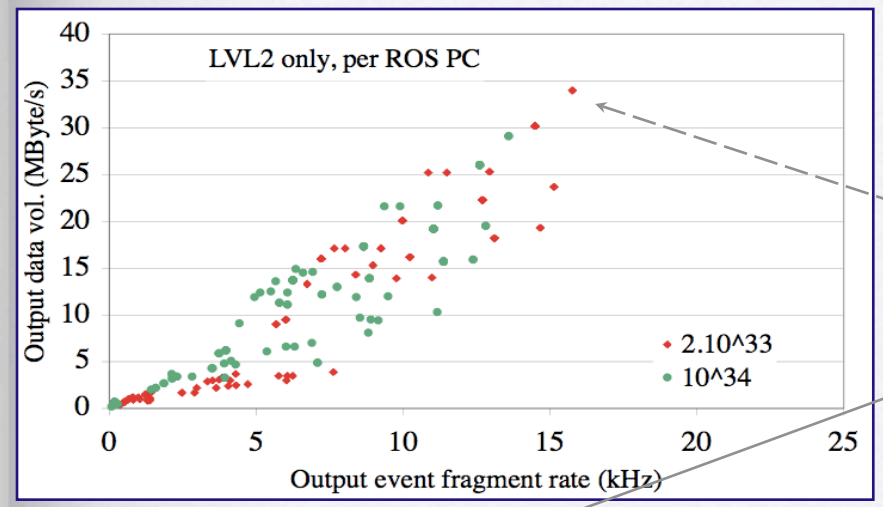
- ROBIN cards with 3 s-link fiber ports is the basic readout system (ROS) input unit. A typical ROS will house 4 such cards (4x3 =12 input channels).
  - ATLAS has ~1600 fiber links, ~150 ROS nodes.
  - ROS sw ensures the parallel processing of 12 input channels





A faithful ROBIN emulation is used when hw not available

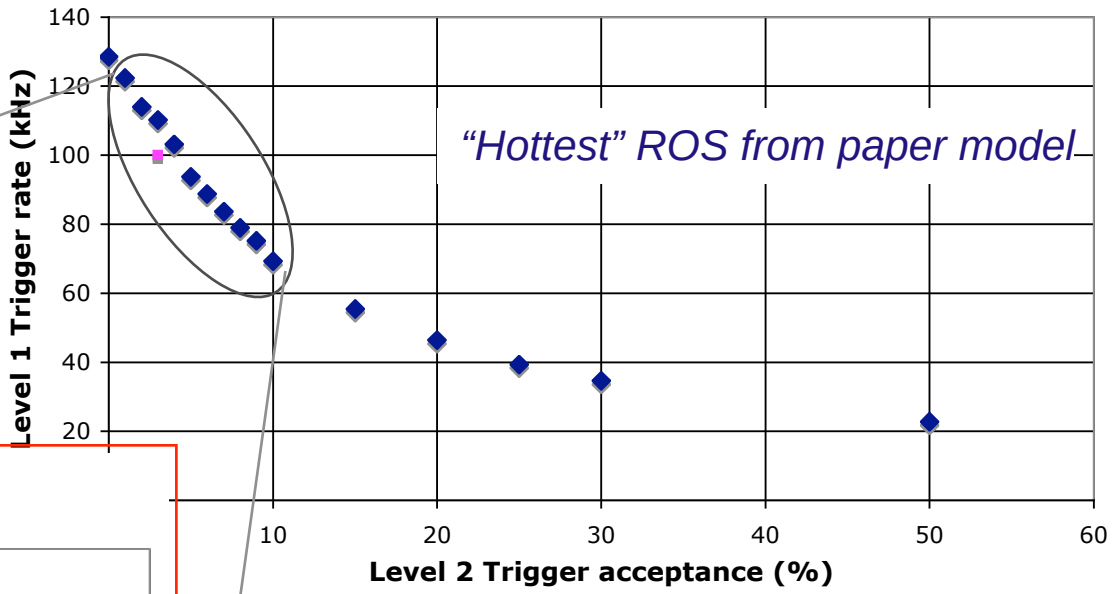
ROS studies



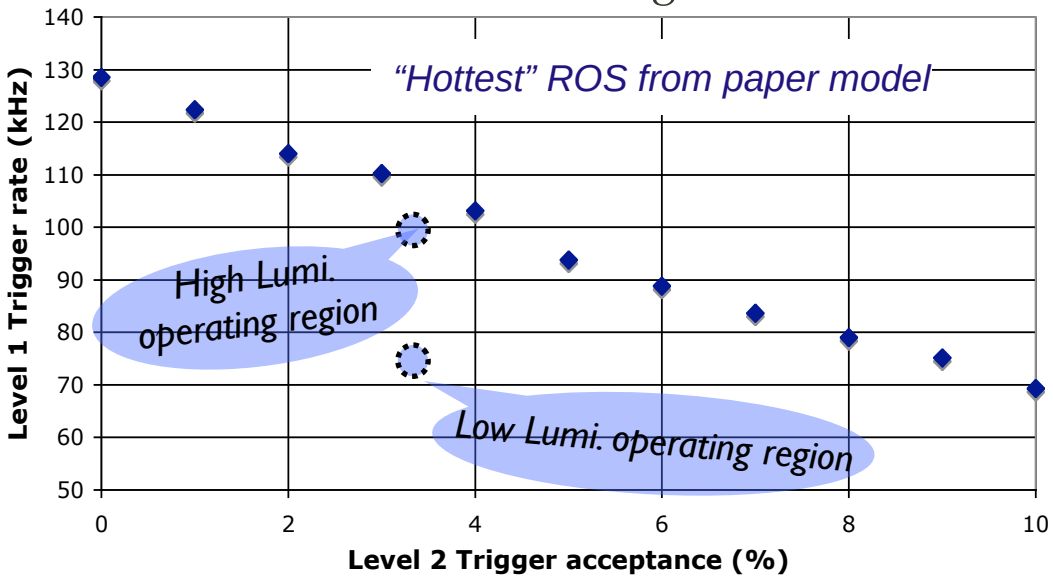
1. A paper model is used to estimate ROS requirements

Final ROS HW used
UDP as n/w protocol
ROS with 12 ROLs
ROL size=1kB

2. max LVL1 rate measured on final hardware



3. Zoom in to "ATLAS region"



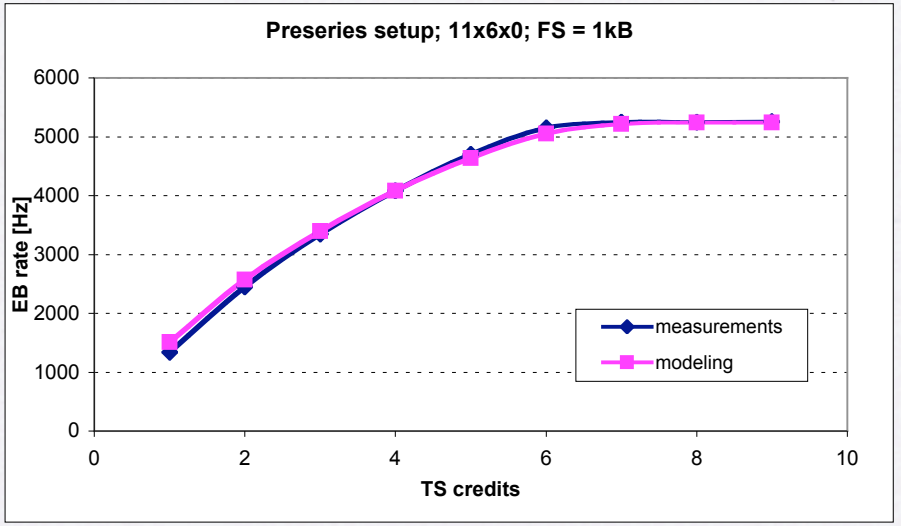
Performance of final ROS (PC+ROBIN) is already above requirements.

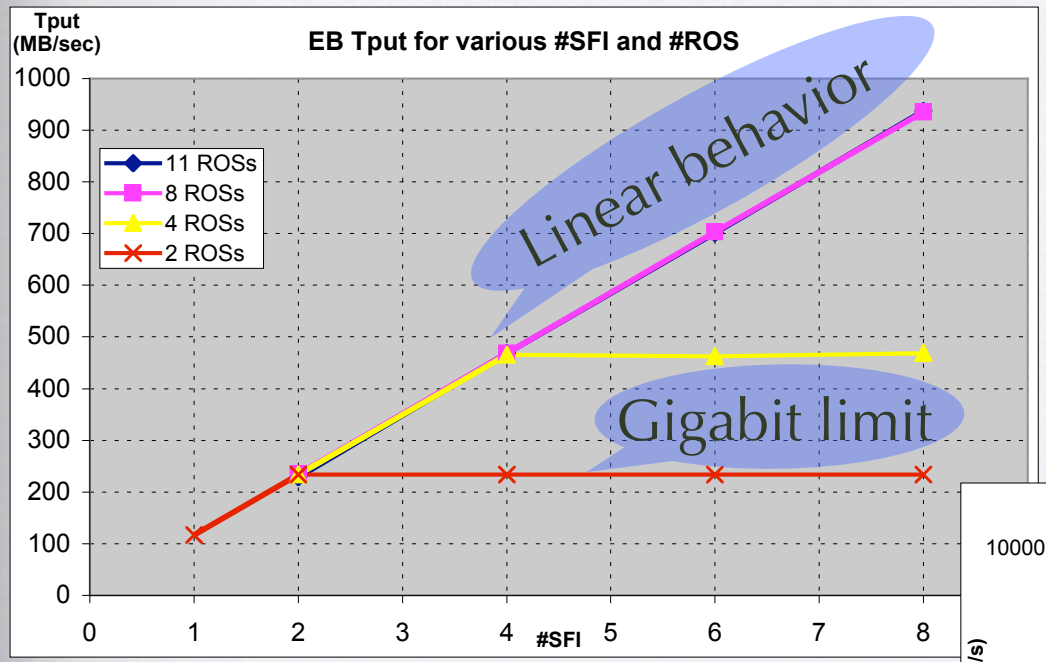
ROD-ROS mapping optimization would further reduce ROS requirements.

EB studies

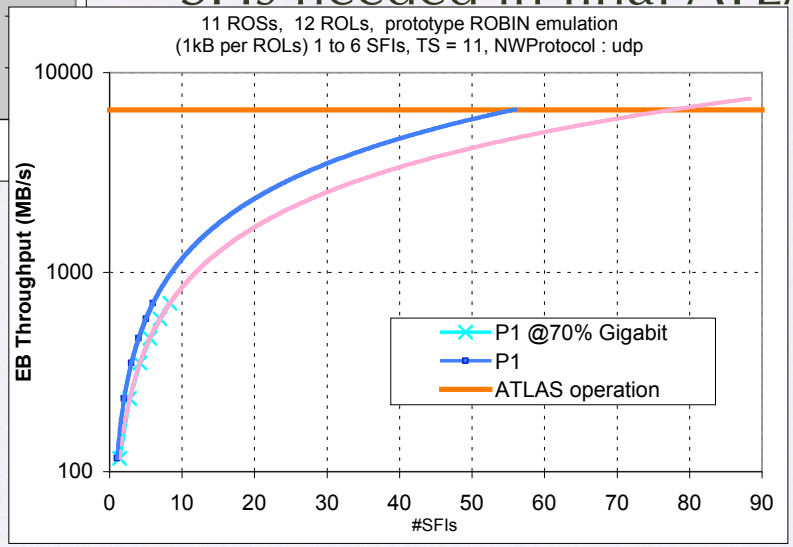
★EB subsystem parameters optimized using measurements on hardware.

★Discrete event simulation modeling faithfully reproduces hardware measurements.



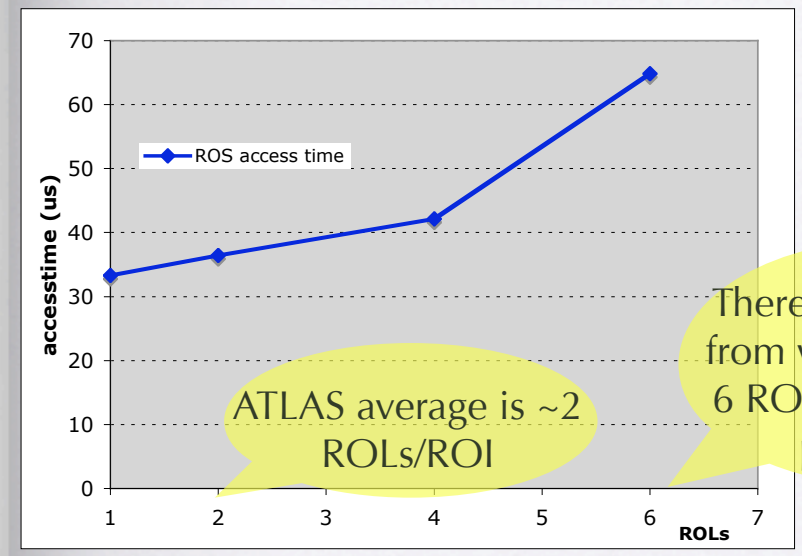


- EB performance is understood in terms of Gigabit line speed and SFI performance. (No SFI output)
- The used to predict the number of SFIs needed in final ATLAS



We estimate final ATLAS would need **~80SFIs** when 70% of the input bandwidth is utilized.

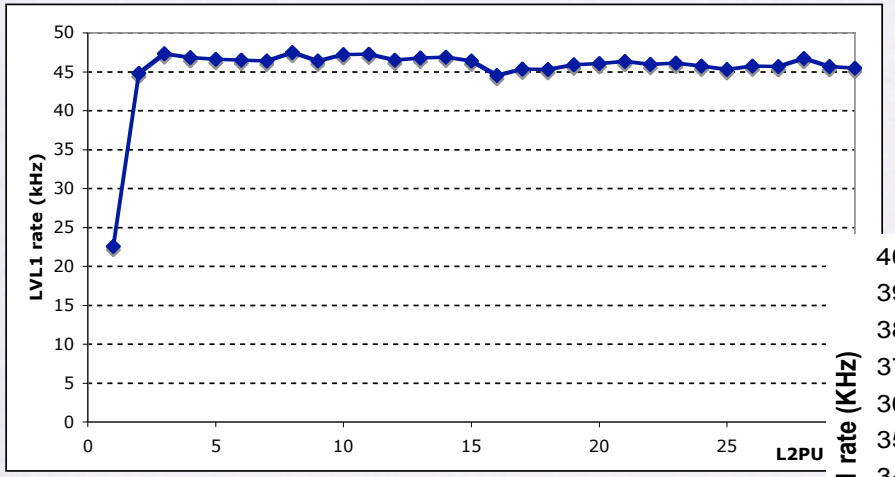
LVL2 studies



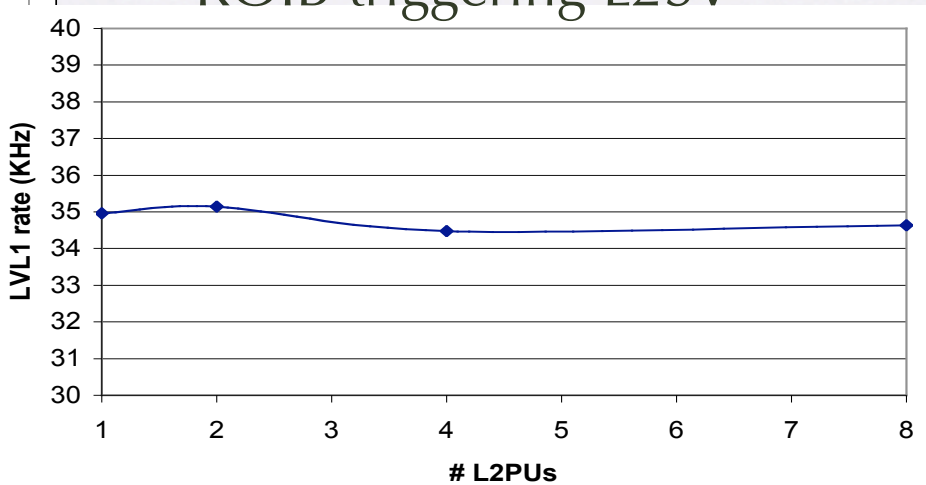
LVL2 has 10ms/event for accept/reject decision

Even for the worst case, the data retrieval time is less than 1% of allocated time.

Self triggering L2SV

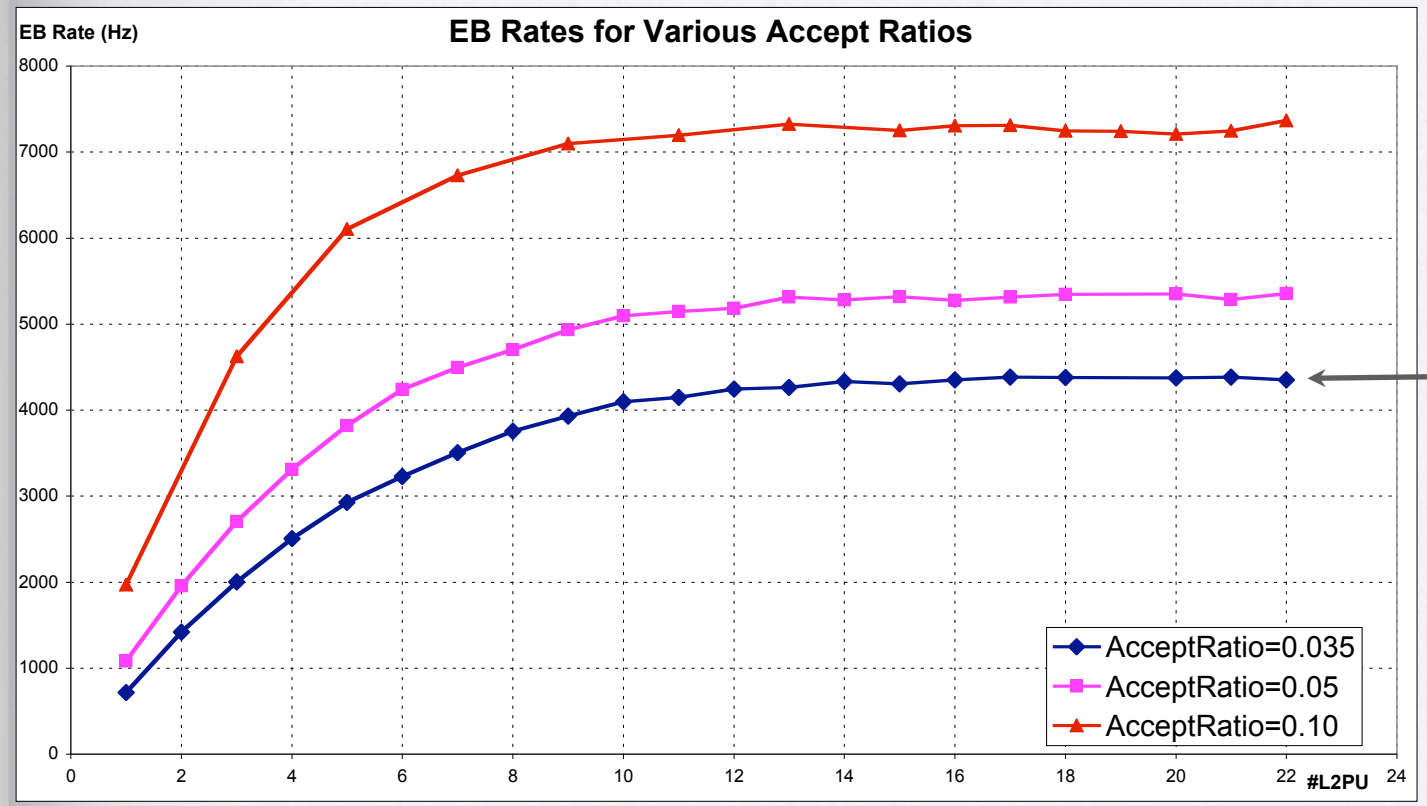


ROIB triggering L2SV



- Single L2SV can handle 35 kHz of LVL1
- We estimate *4 L2SVs* can handle the load of full ATLAS.

Combined system -1



8 ROS, 8 SFI .. 22 L2PU

ATLAS runs at 3.5% LVL2 accept ratio

- The triggers are generated by the L2SVs driving the system as fast as possible. Stable operation observed even for overdriven conditions.

$$\frac{TS \,.\, N_{SFI}}{WT \,.\, N_{L2}} > a \,.\, N_{ROS}$$

stability condition for TDAQ configuration parameters

TS, WT: strength coefficients for EB and LVL2 systems
a: LVL2 acceptance

- L2PUs run without algorithms, with multiple parallel threads
  - ➡ each L2PU represents an LVL2 subfarm

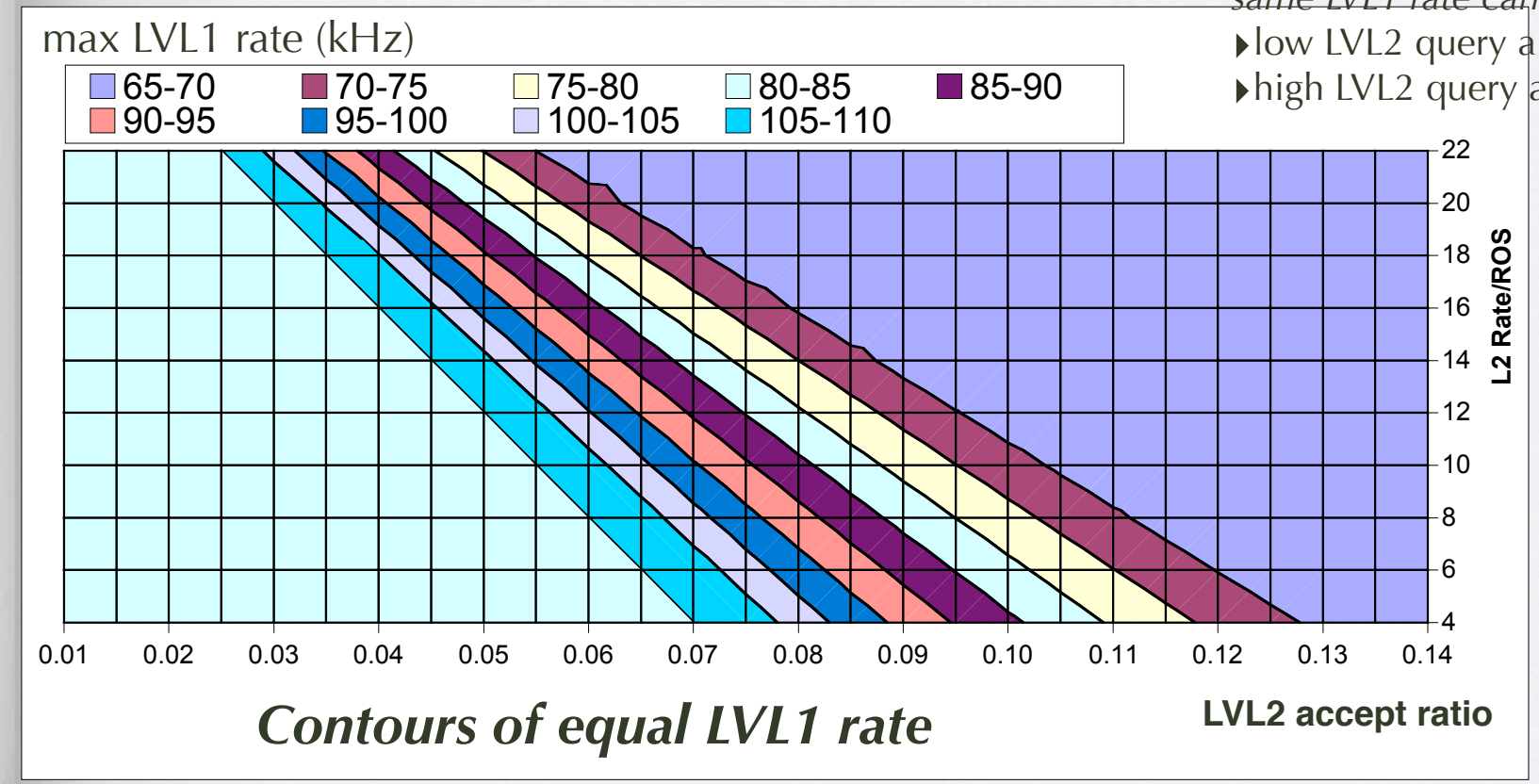
Combined system -2

- Except ROS, multiple instances of TDAQ applications run in parallel for max readout rate; each ROS is responsible from a section of the detector
- ROS runs multiple tasks, and its performance can affect the LVL1 rate
- $CPU_{ROS} = R_{EB} \times CPU^{EB} + R_{L2} \times CPU^{L2} + R_{L1} \times CPU^{Cl}$

- **CPU^{EB}** is the CPU power spent by a ROS on 1 kHz of Event Building task
- **CPU^{L2}** is the CPU power spent by a ROS on 1 kHz of LVL2 ROI collection task
- **CPU^{Cl}** is the CPU power spent by a ROS on 1 kHz of Event Clear task

same LVL1 rate can be achieved either:

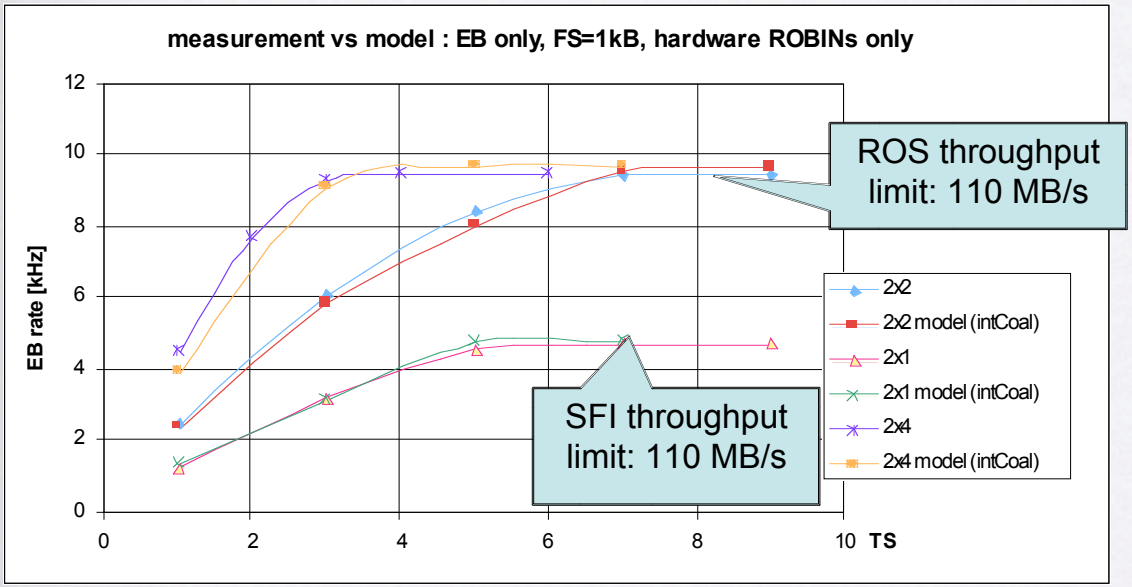
- low LVL2 query and high EB rate
- high LVL2 query and low EB rate



Contours of equal LVL1 rate

Modeling preseries

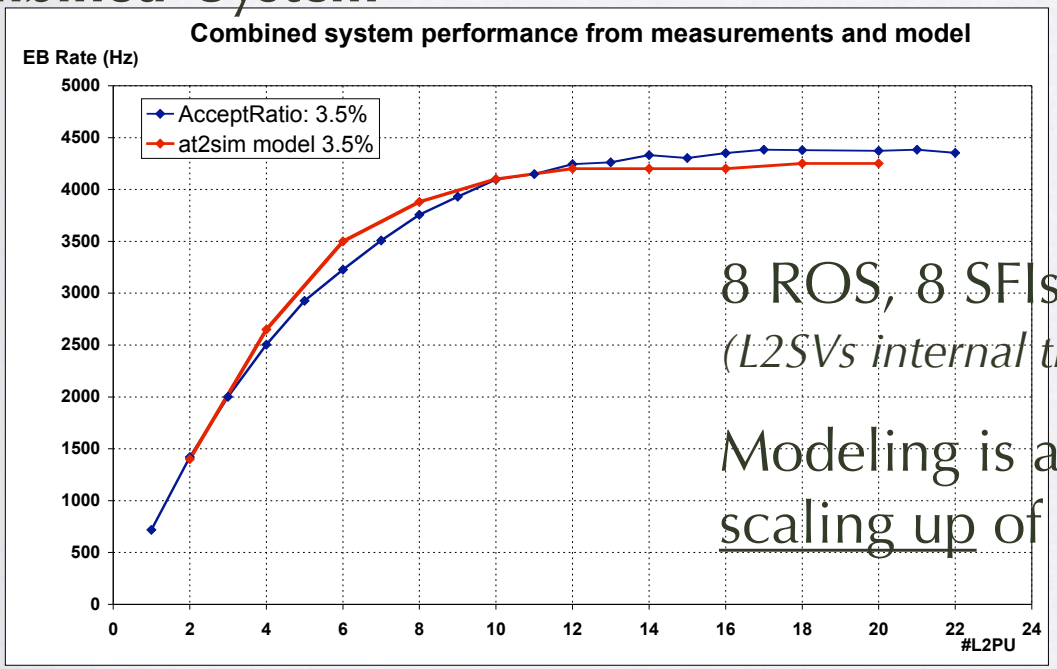
Event building only



2 ROS and 1, 2 and 4 SFIs
(DFM internal trigger)

Modeling is able to reproduce the impact of configuration parameters and the limitations coming from various TDAQ components.

Combined System

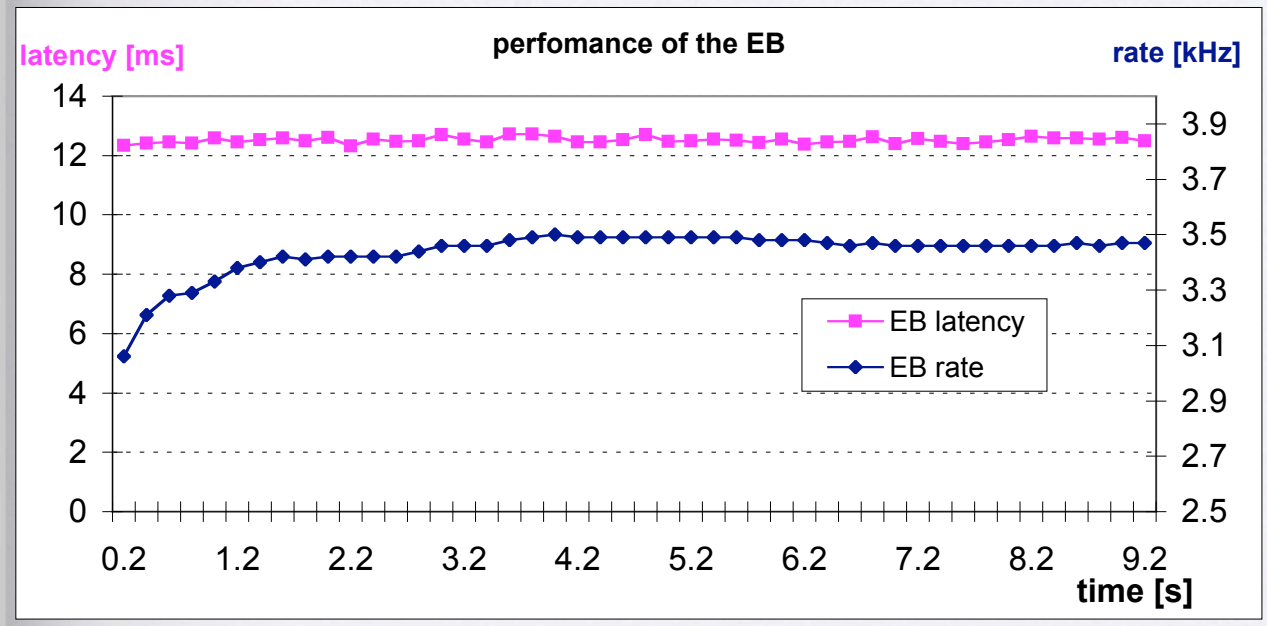


8 ROS, 8 SFIs and up to 22 L2PUs
(L2SVs internal trigger)

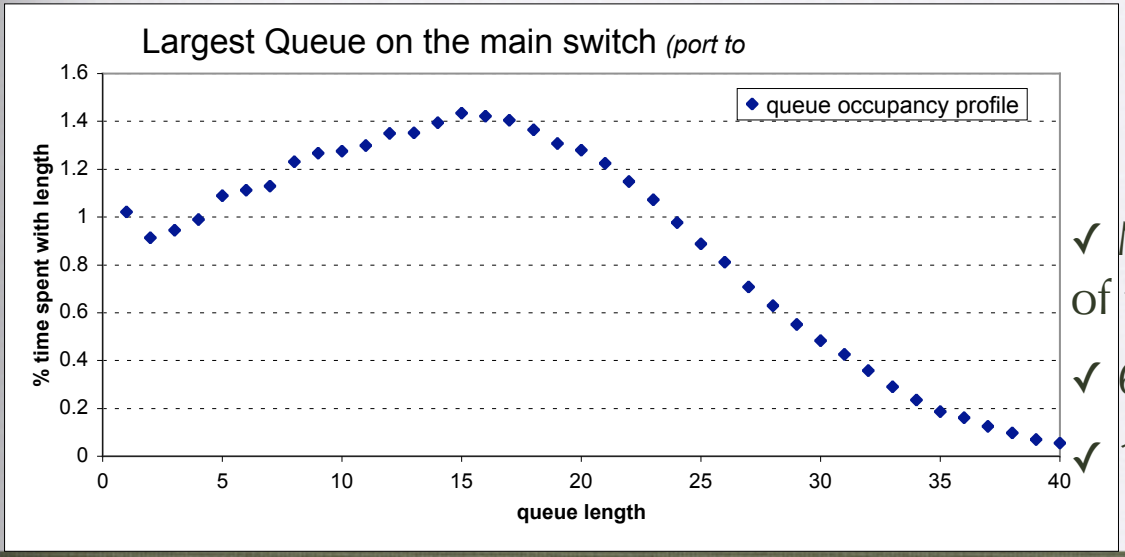
Modeling is able to reproduce the impact of scaling up of the preseries test bed size.

Modeling final ATLAS

- Simulations from preseries test bed was scaled up to final ATLAS size: 131 ROS, 110 SFIs, 504 L2PUs
  - out of 150 ROS, only 131 accessed by LVL2 are simulated, no algorithms in LVL2

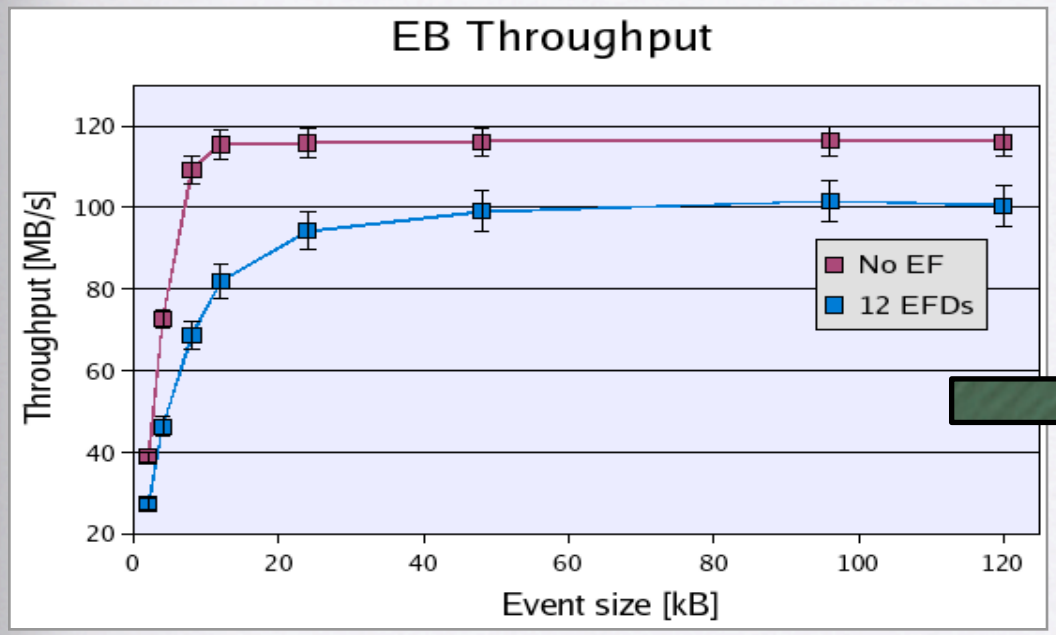


- ✓ ~1M events simulated
- ✓ EB latency stable at ~12ms
- ✓ Stable operation at 100kHz of LVL1 rate giving 3.5 kHz of EB as required

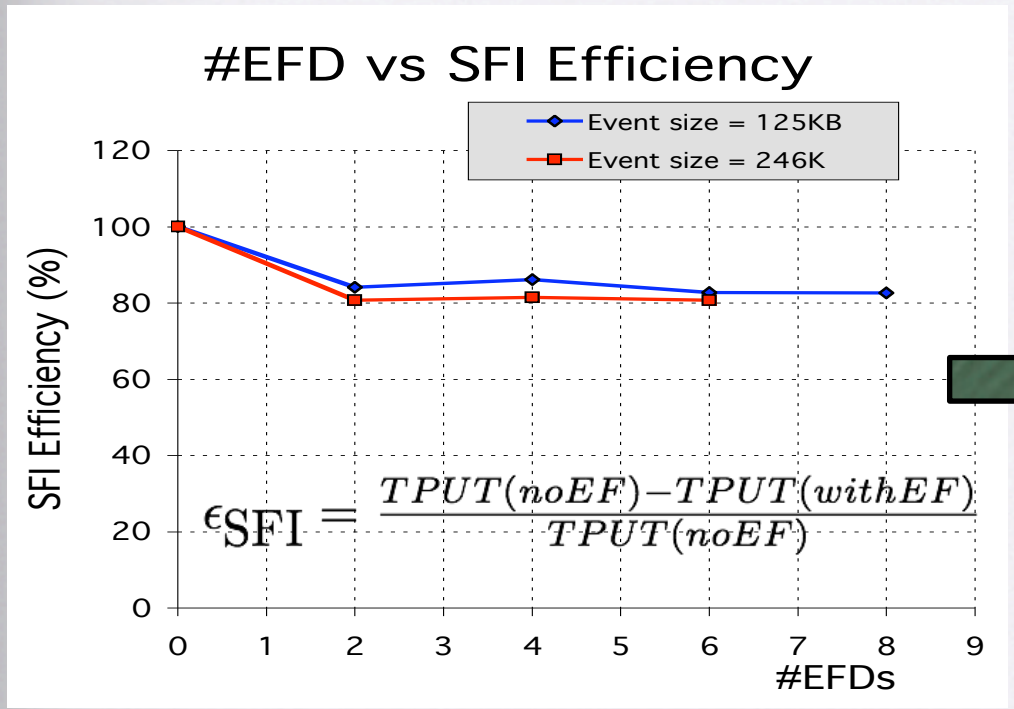


- ✓ Modeling helps to understand the internals of the network switches (max queue 64 packets)
- ✓ 60% of the time queue is empty
- ✓ 1.4% of the time queue is 1/3 occupied

Adding Event Filter

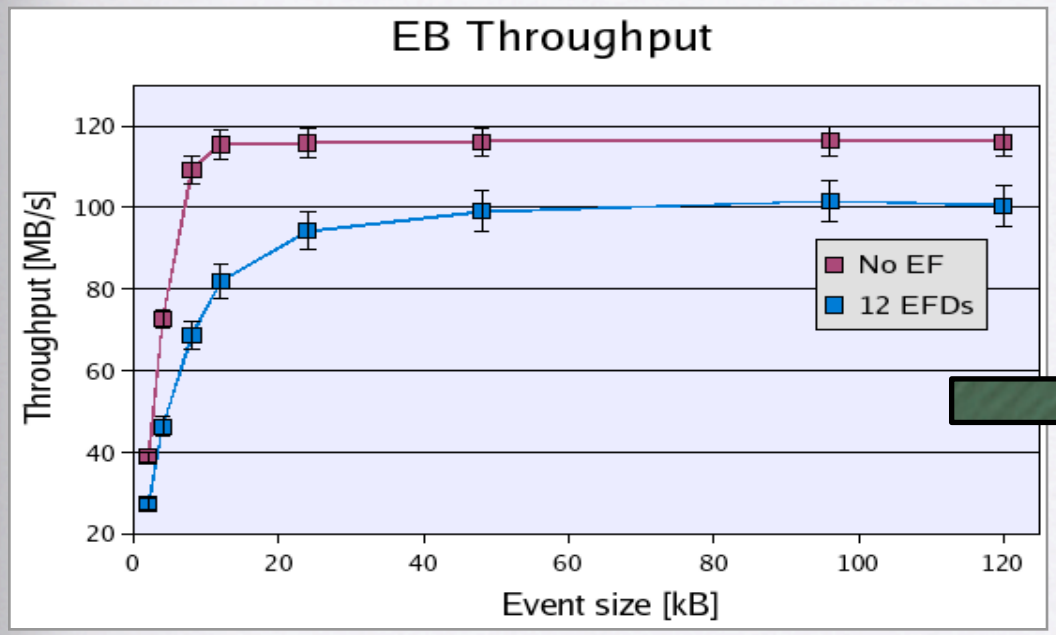


- Output from SFI to EF nodes decreases maximum SFI throughput.
- The throughput with EF saturates to about 80% of the maximum for large events. (ATLAS events ~1.5MB)
  - 80 / 0.80 = 100 SFIs needed for final ATLAS



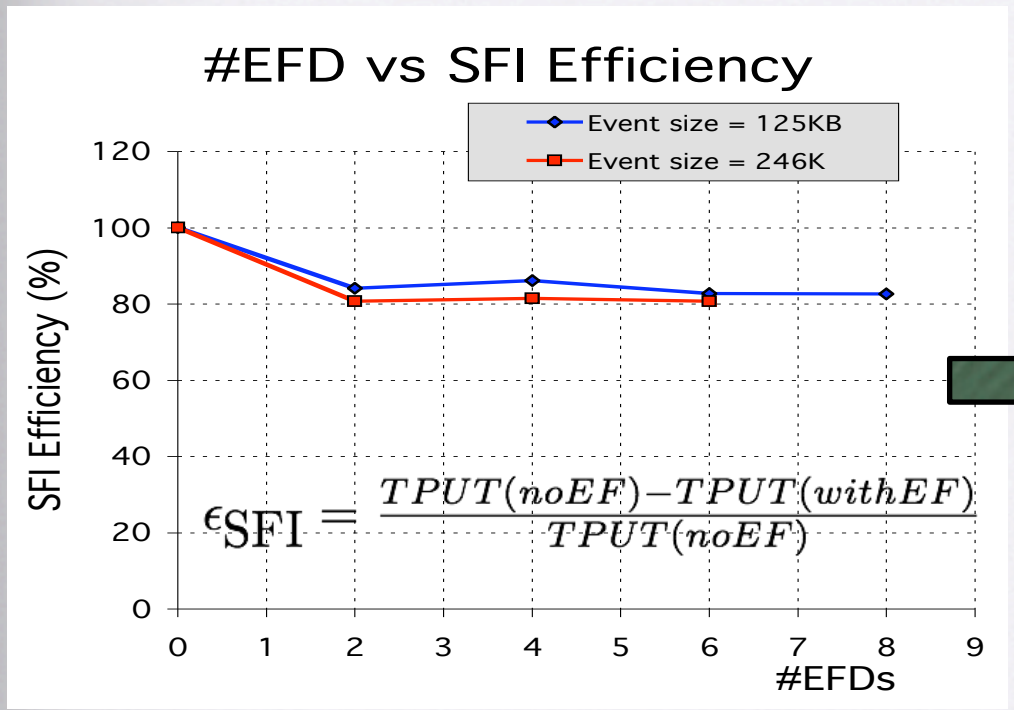
- 2EF nodes (w/o) algorithms are enough to saturate 1 SFI
  - 6 x2=12 EF nodes needed to drive the pre-series.

Adding Event Filter



Event Filter details: K. Kordas' talk

- Output from SFI to EF nodes decreases maximum SFI throughput.
- The throughput with EF saturates to about 80% of the maximum for large events. (ATLAS events ~1.5MB)
  - 80 / 0.80 = 100 SFIs needed for final ATLAS
- 2EF nodes (w/o) algorithms are enough to saturate 1 SFI
  - 6 x2=12 EF nodes needed to drive the pre-series.



Adding SFO

- ATLAS writes 200Hz for events ~1.5MB => 300 MB/s in total
  - TDR: 15MB/s of throughput on single disk: 20 .. 30 PCs up to 450 MB/s
- ATLAS requires ~24 hour independent running => ~25 TB total disk space
  - TDR: ~30 1U Nodes with 1TB of HD each
- Faster disk I/O and larger disks could mean less SFOs to maintain
- RAID can bring protection against failures, but:
  - which raid level? (Raid1, Raid5, Raid50), which filesystem, 64-bit?

0.5 .. 1.5MB events, linux ext2 filesystem, 372GB disks

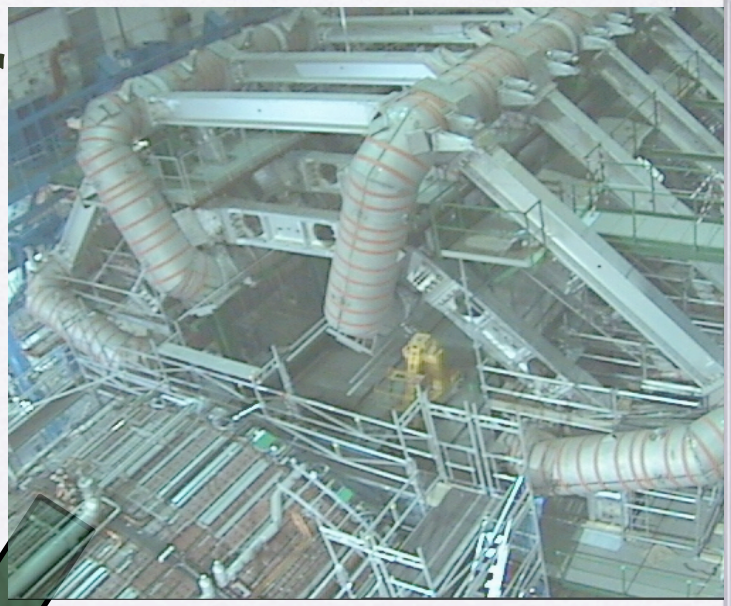
writing speed (MB/s)	1u-sw	1u-hw	3u-hw
raid1 (2HD)	44	48	51
raid5 (3HD)	53	73	73*

* 93 MB/s with 6HD

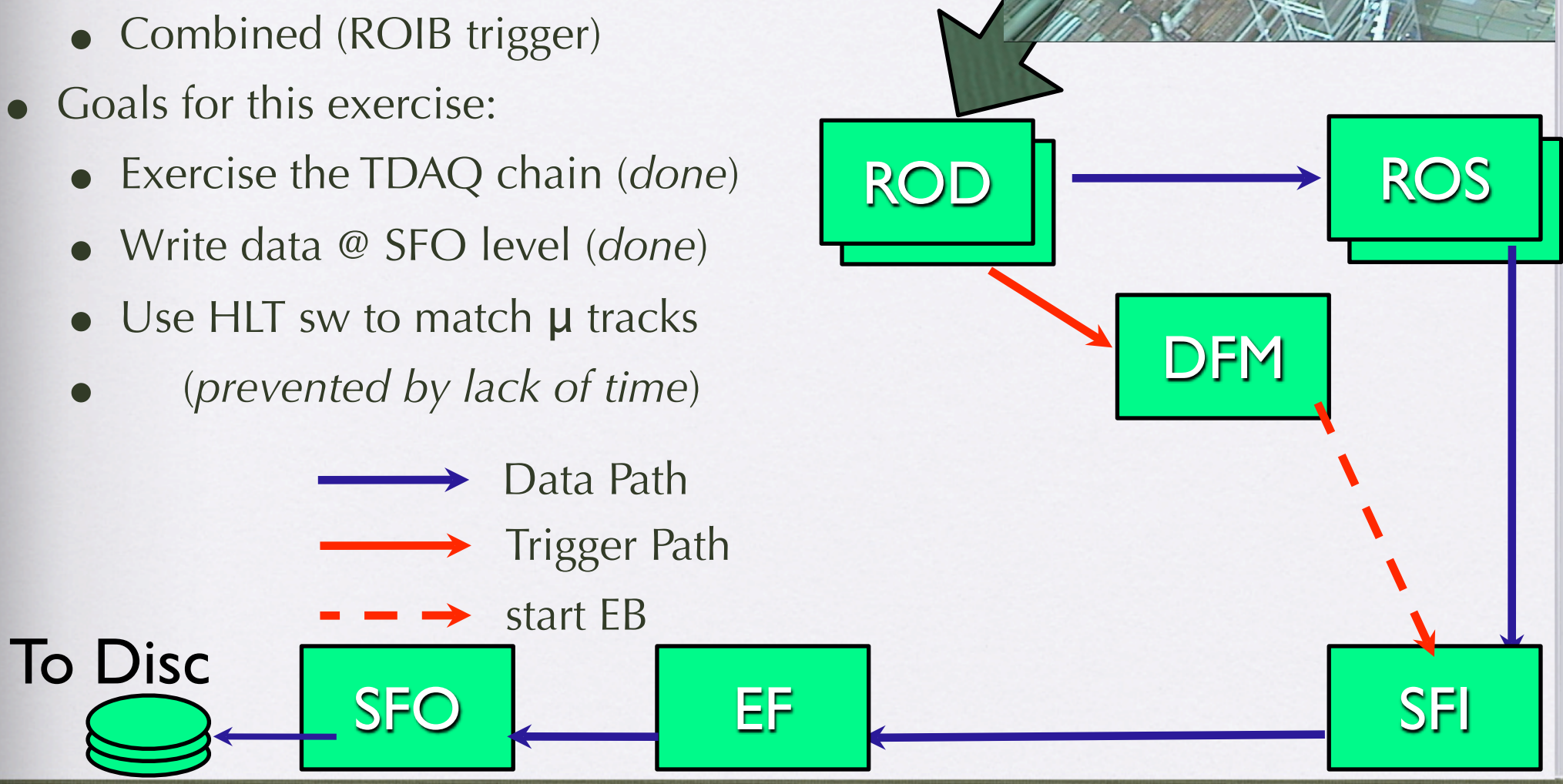
can house 16 HD, yielding 5.2 TB Raid5 or Raid50

✦ 5 such node (+1 hot spare) can match the speed and capacity requirements
✦ less nodes: less space, less work, less problems!

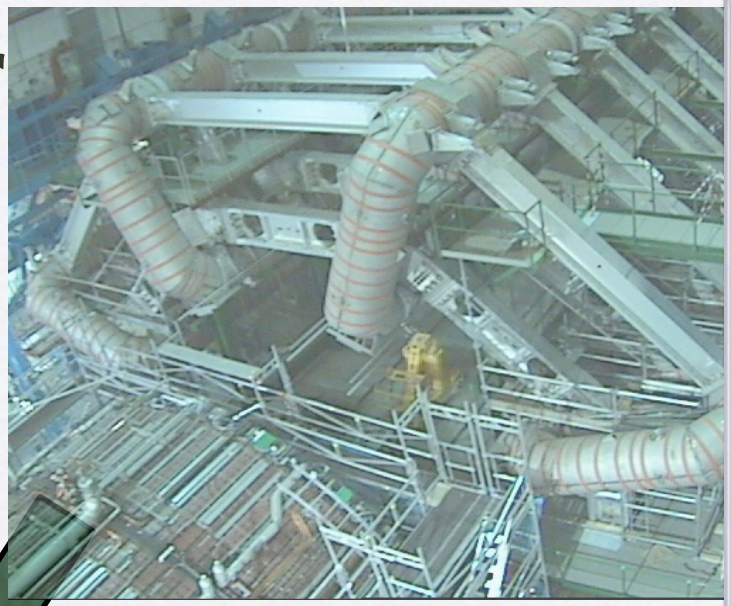
Adding a detector



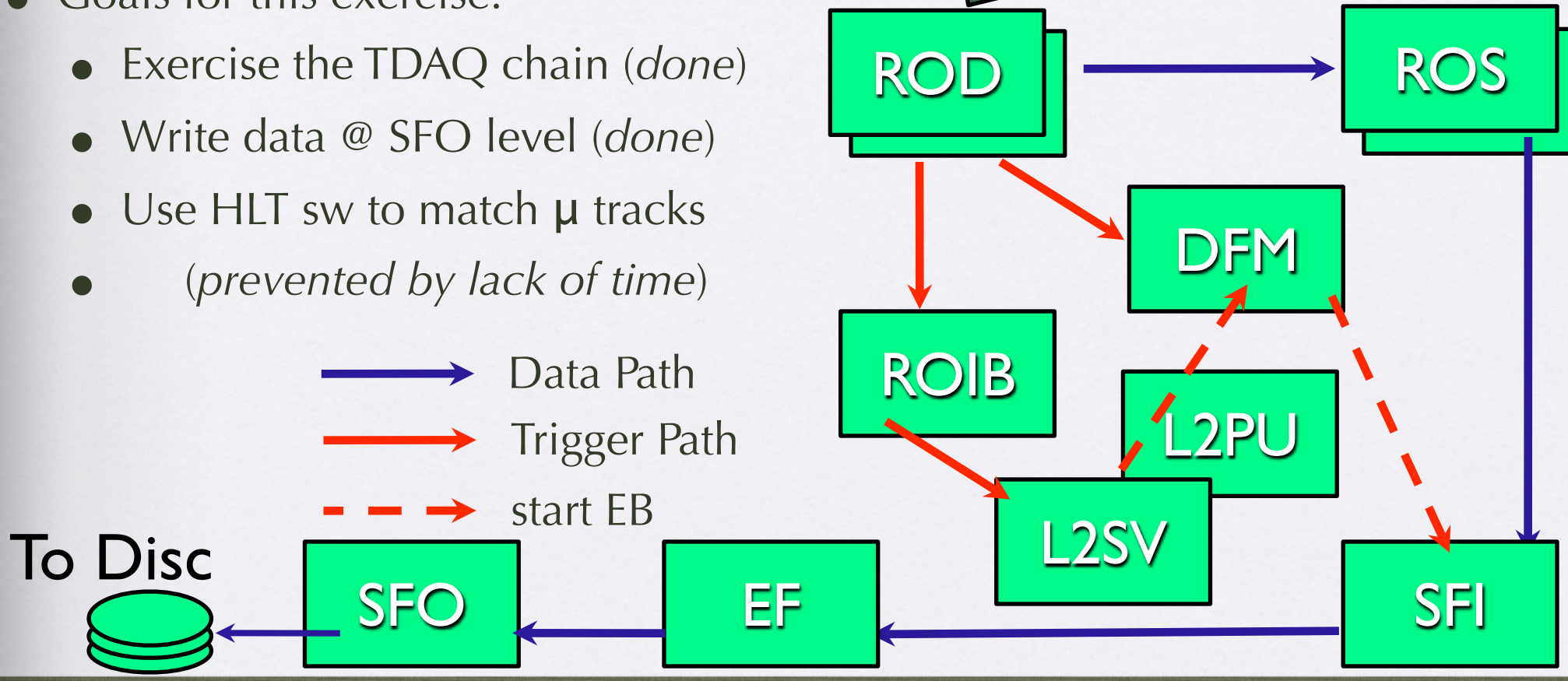
- Using pre-series with Tile Calorimeter Barrel setup at the pit (16 ROLs in 8 ROBINs, 2 ROSs)
- Possible Setups
  - Simple EB (DFM self trigger)
  - Combined (ROIB trigger)
- Goals for this exercise:
  - Exercise the TDAQ chain (*done*)
  - Write data @ SFO level (*done*)
  - Use HLT sw to match μ tracks
  - (*prevented by lack of time*)



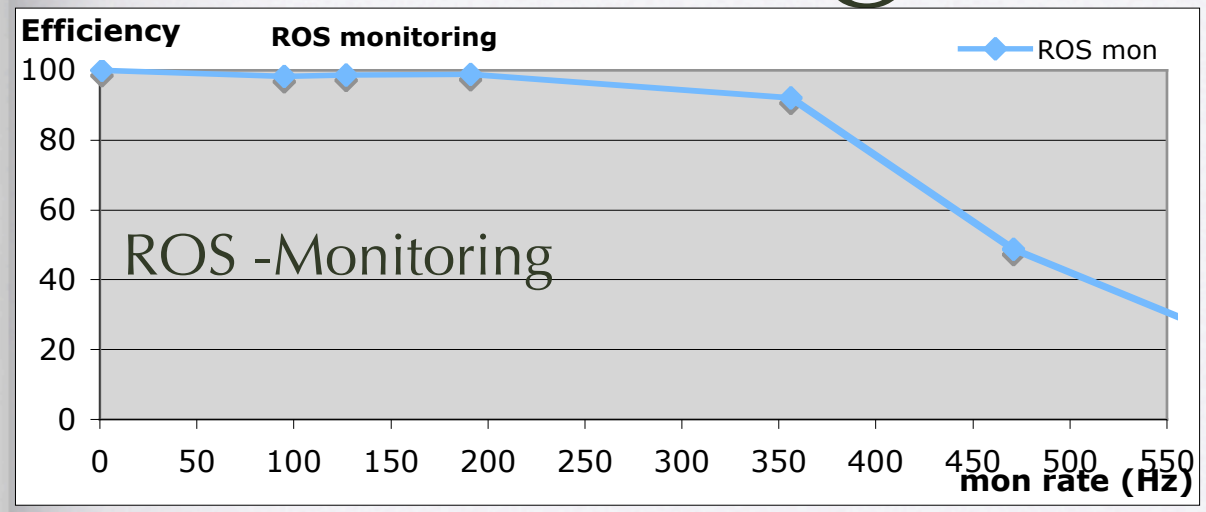
Adding a detector



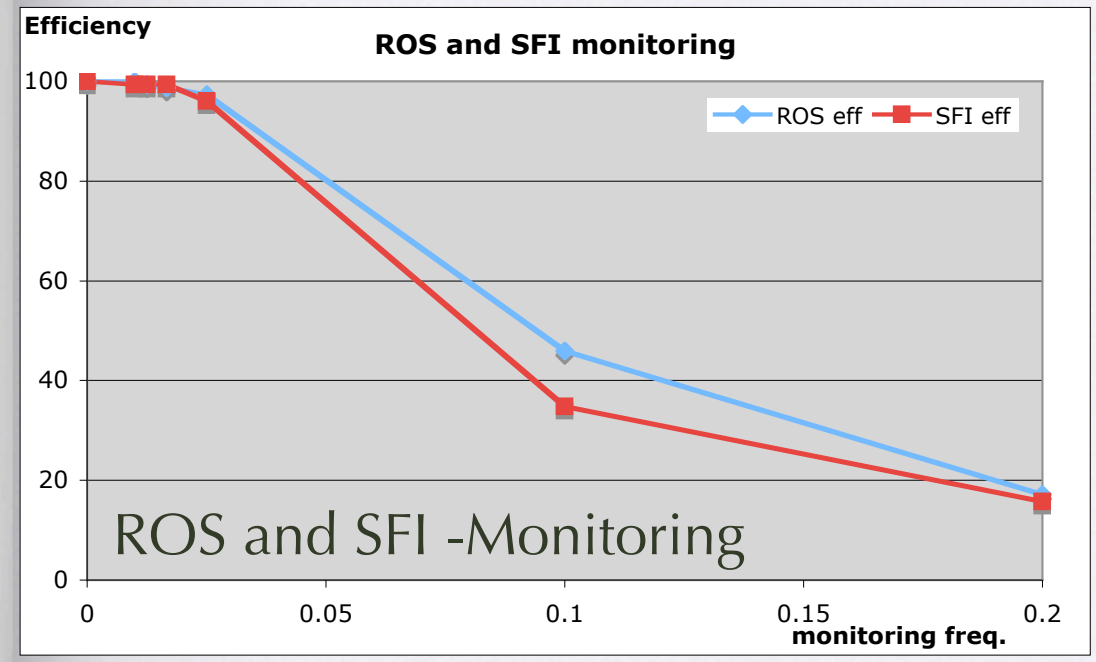
- Using pre-series with Tile Calorimeter Barrel setup at the pit (16 ROLs in 8 ROBINs, 2 ROSs)
- Possible Setups
  - Simple EB (DFM self trigger)
  - Combined (ROIB trigger)
- Goals for this exercise:
  - Exercise the TDAQ chain (*done*)
  - Write data @ SFO level (*done*)
  - Use HLT sw to match μ tracks
  - (*prevented by lack of time*)



Adding Monitoring



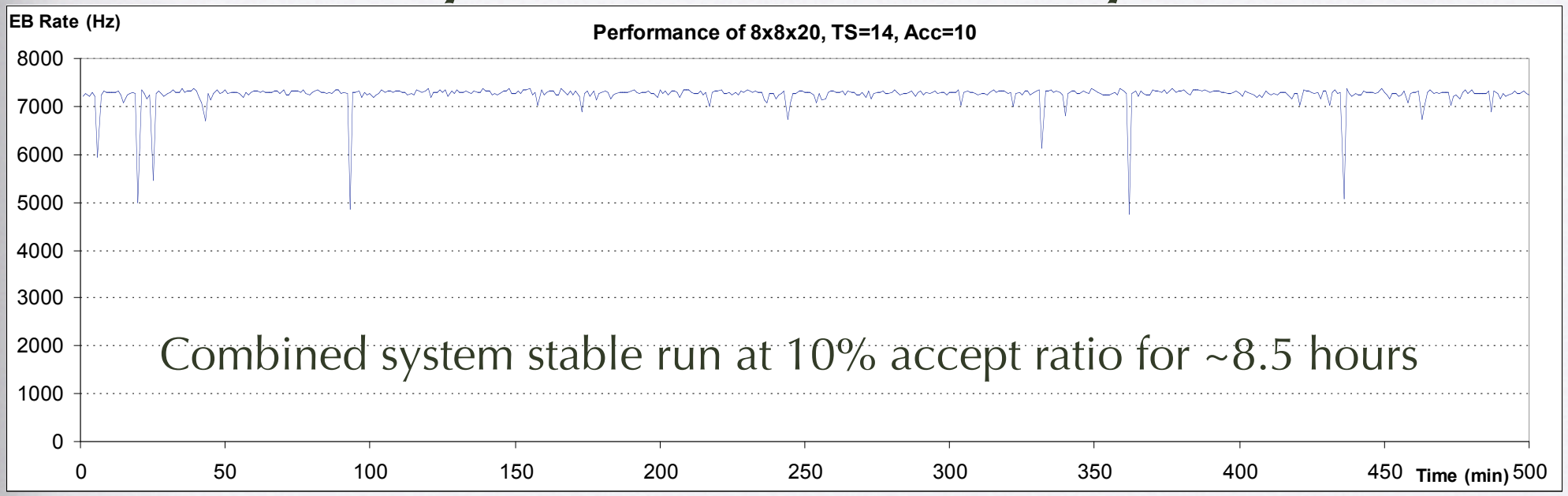
- 8 ROS x 8 SFI system, EB only, Gigabit limited
- Each node running 1 sampler
- As monitoring rate increases, EB (& monitoring) rate drops.



Efficiency is obtained by comparing rates with and without monitoring.

For both applications, up to 3% of the maximum input rate can be send to monitoring without affecting the readout rate.

Stability & Recovery issues



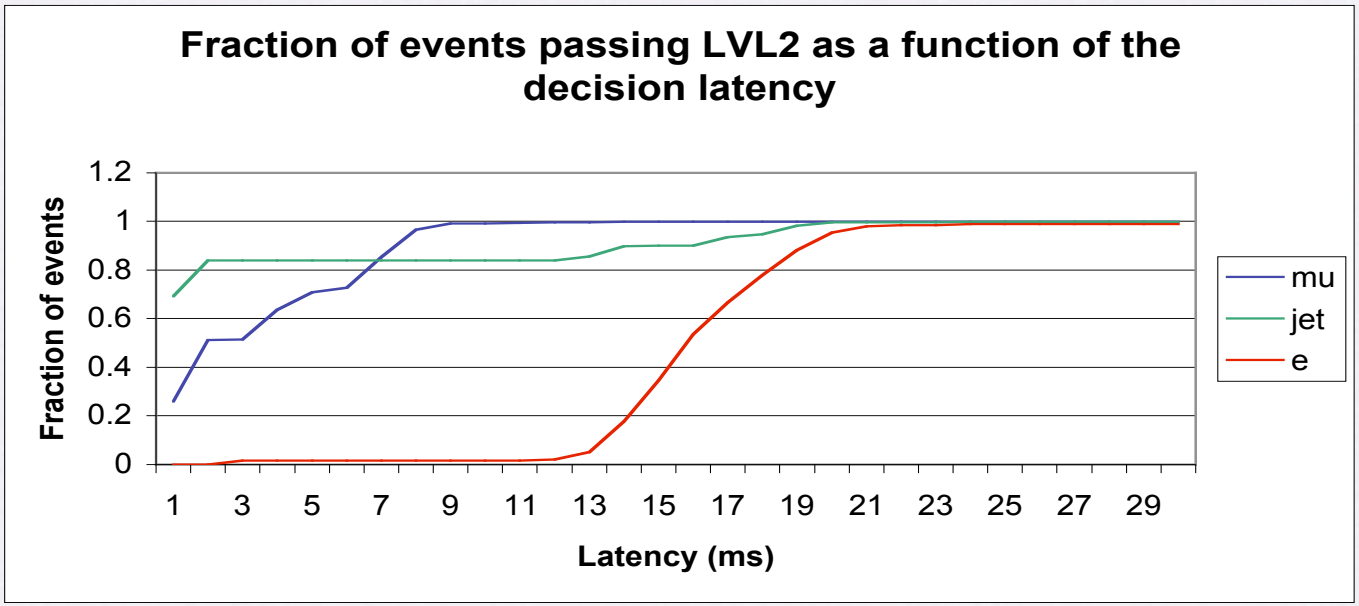
Fault tolerance and Recovery tests

- Force systematic failures in hardware and in software, check the system recovery
  - Almost all infrastructure application failures can be recovered
  - Apart from ROS, all DataFlow applications can be recovered. (ROS needs hardware handshake with detector side)
  - For hardware failures, Process Manager will be part of Unix services to reintegrate a dead node into TDAQ chain.

Replaying physics data

- Event data files from simulation studies for different physics objects (muons, e, jet) were preloaded into ROS & ROIB
- The whole TDAQ chain was triggered by the ROIB
- Various physics algorithms were run at the LVL2 farm nodes to select events matching the required triggers
- Selected events were assembled from all ROS and send to event filter farm

For the first time we have realistic measurements on various algorithm execution times.

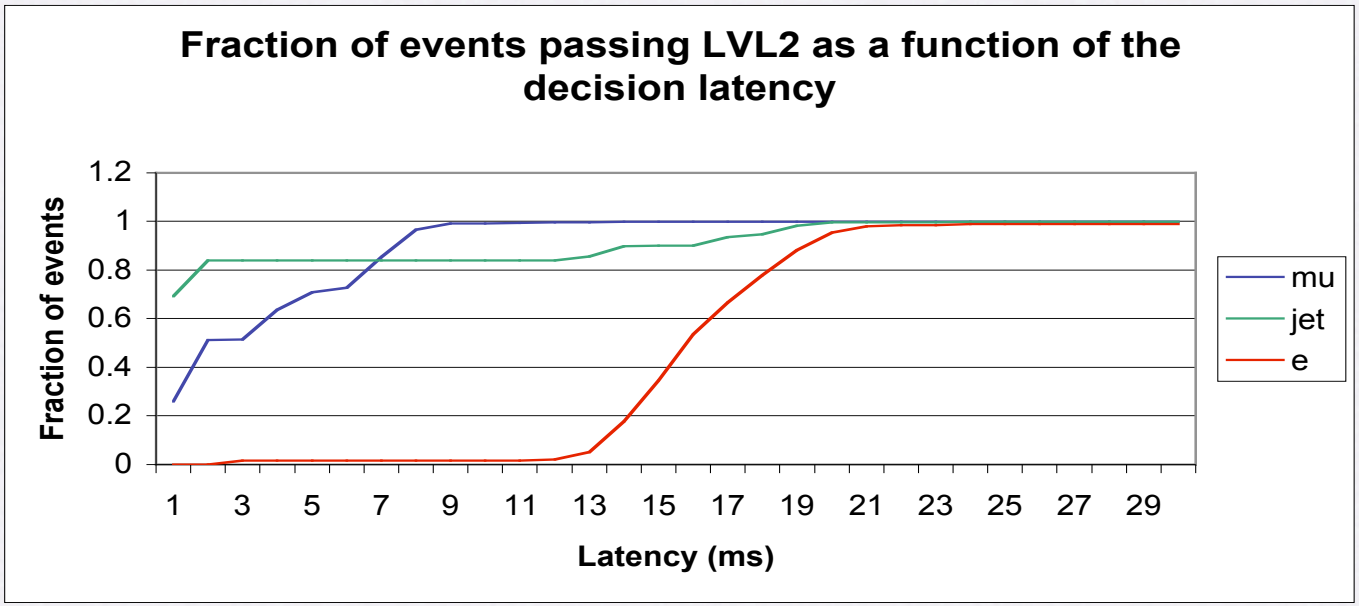


Replaying physics data

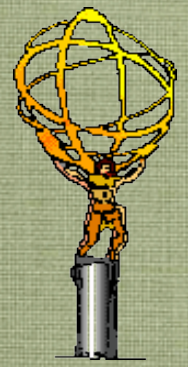
- Event data files from simulation studies for different physics objects (muons, e, jet) were preloaded into ROS & ROIB
- The whole TDAQ chain was triggered by the ROIB
- Various physics algorithms were run at the LVL2 farm nodes to select events matching the required triggers
- Selected events were assembled from all ROS and send to event filter farm

Algorithm details: K. Kordas' talk

For the first time we have realistic measurements on various algorithm execution times.



Next steps



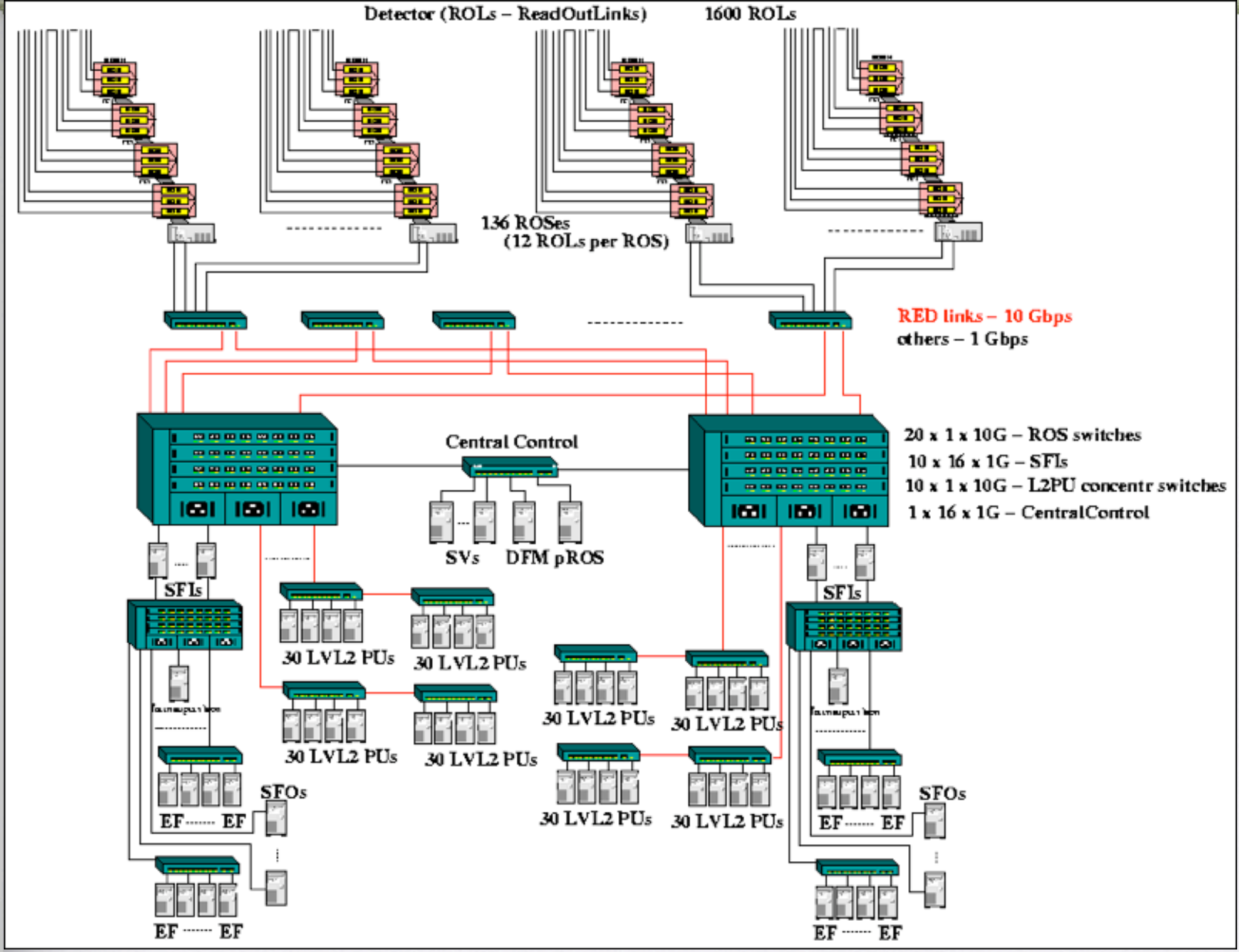
- Exploitation with algorithms
- Garbage collection
- State transition timing measurements
- Preseries as a release testing environment

Conclusions

✓ Pre-series system has shown that ATLAS TDAQ operates reliably and up to the requirements.

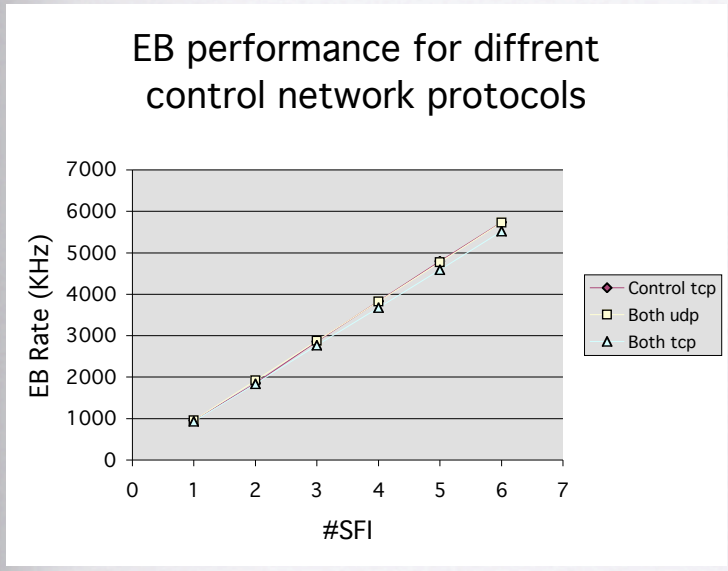
✓ We will be ready in time to match the LHC challenge.

- BACK UP!

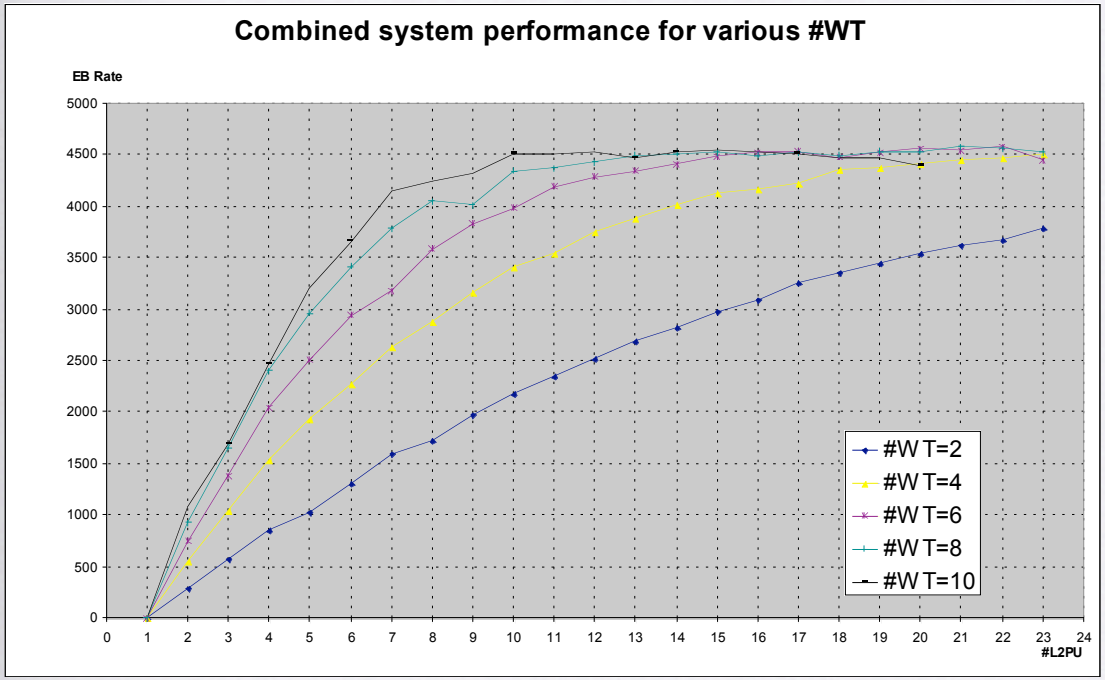


- The final ATLAS architecture as modeled.

details



- small size system doesn't see any difference



the number of worker threads is set to 6

machine park

- **ROS x 11**
  - CPU : Intel Xeon, 3.2 GHz UP
  - Memory : 512 MB
- **SFI x 6**
  - CPU : Intel Xeon, 3.2 GHz SMP
  - Memory : 512 MB
- **L2PU x 30**
  - CPU : AMD Opteron P250, 2.4 GHz SMP
  - Memory : 4 GB
- **L2SV x 2**
  - CPU : Intel Xeon, 3.2 GHz SMP
  - Memory : 512 MB
- **DFM x 2**
  - CPU : Intel Xeon, 3.2 GHz SMP
  - Memory : 512 MB
- **SFO x 2**
  - CPU : Intel Xeon, 3.2 GHz SMP
  - Memory : 4 GB
- **EF x 12**
  - CPU : AMD Opteron P250, 2.4 GHz SMP
  - Memory : 4 GB

- **Switches**
  - L2/EB switch :
    - 24 GE fibre ports
    - 35 GE copper ports (36 minus one management port)
    - 2 10GE fibre ports (only one used)
  - L2PU concentration switch :
    - 48 GE copper ports
    - 2 10 GE fibre ports (only one used)
  - BackEnd (EF central) switch:
    - 24 GE copper ports
  - EF switch (rack concentrator) :
    - 24 GE copper ports
- **DOLARs** (used to emulate output from ROD)
  - installed in pc-preseries-ros-01
- **ROBINs**
  - installed in pc-preseries-ros-02
  - pc-preseries-ros-03
  - pc-preseries-ros-04
  - pc-preseries-ros-05