Track Summary Online Computing

Sylvain Chapeland
Remigius K Mommsen
Niko Neufeld

Overview

2 sessions devoted to LHC experiments

- 3 talks from ATLAS and CMS
- 2 talks from Alice
- 1 talk from LHCb

Apologies to all presenters whose material cannot be shown in this summary

2 sessions devoted to non-LHC experiments

- 3 neutrino talks (NOνA and DoubleChooz)
- 2 talks reviewing Tevatron DAQ (CDF and DØ)
- 3 talks covering other experiments

51 posters presented (out of 62 submitted)

- Lightning talk: "Online Metadata Collection and Monitoring Framework for the STAR Experiment at RHIC"

LHC
Operation Experiences

Improving Efficiency

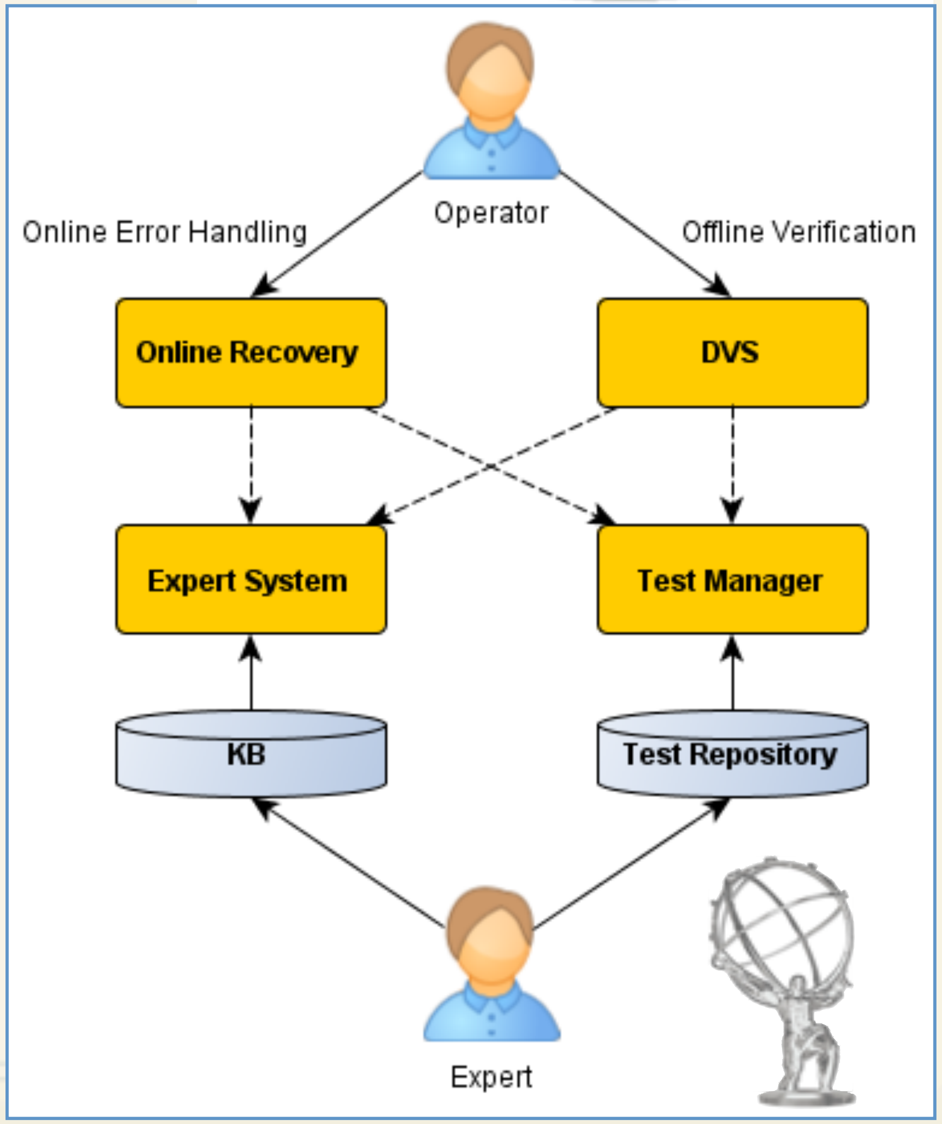
Vasco Barroso, Giuseppe Avolio, Andrea Negri, Hannes Sakulin, Sylvain Chapeland (poster)

Data taking efficiencies > 90% (c.f. CDF ~84%)

- DAQ systems much more reliable (CMS: 99.7%)

Seeking to reduce down-times, while reducing shift crews

- Expert system to assist shifters
  - ATLAS: "DAQ Assistant" based on CLIPS
  - CMS: "DAQ Doctor" (Perl-based)
  - ALICE: "Orthos" alarms integrated with documentation and operations issue tracker
- Automatic recovery & actions
- Formalize expert knowledge (ATLAS)
- Improved accounting of reasons for down-times (ALICE)



Andrea Negri, Hannes Sakulin, Vasco Barroso

Reducing Nb of Run Starts

ATLAS avoids any run starts during data taking

Restarting a run takes > 1 minute (CMS 1'15, ALICE > 4')

Recover from failures without stopping

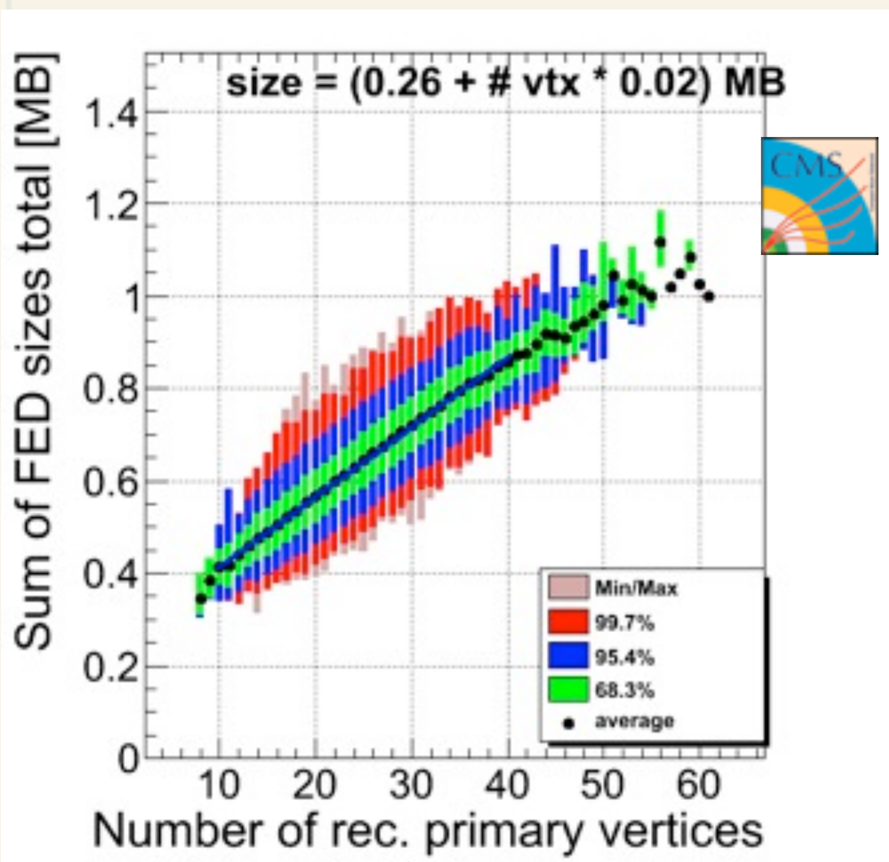
- Pause triggers and allow sub-systems to do recovery action
- Increase fault tolerance for non-critical components
- Restart applications during on-going runs
- Requires vigilance concerning data quality and integrity

Andrea Negri,
Hannes Sakulin

Dealing with Pileup

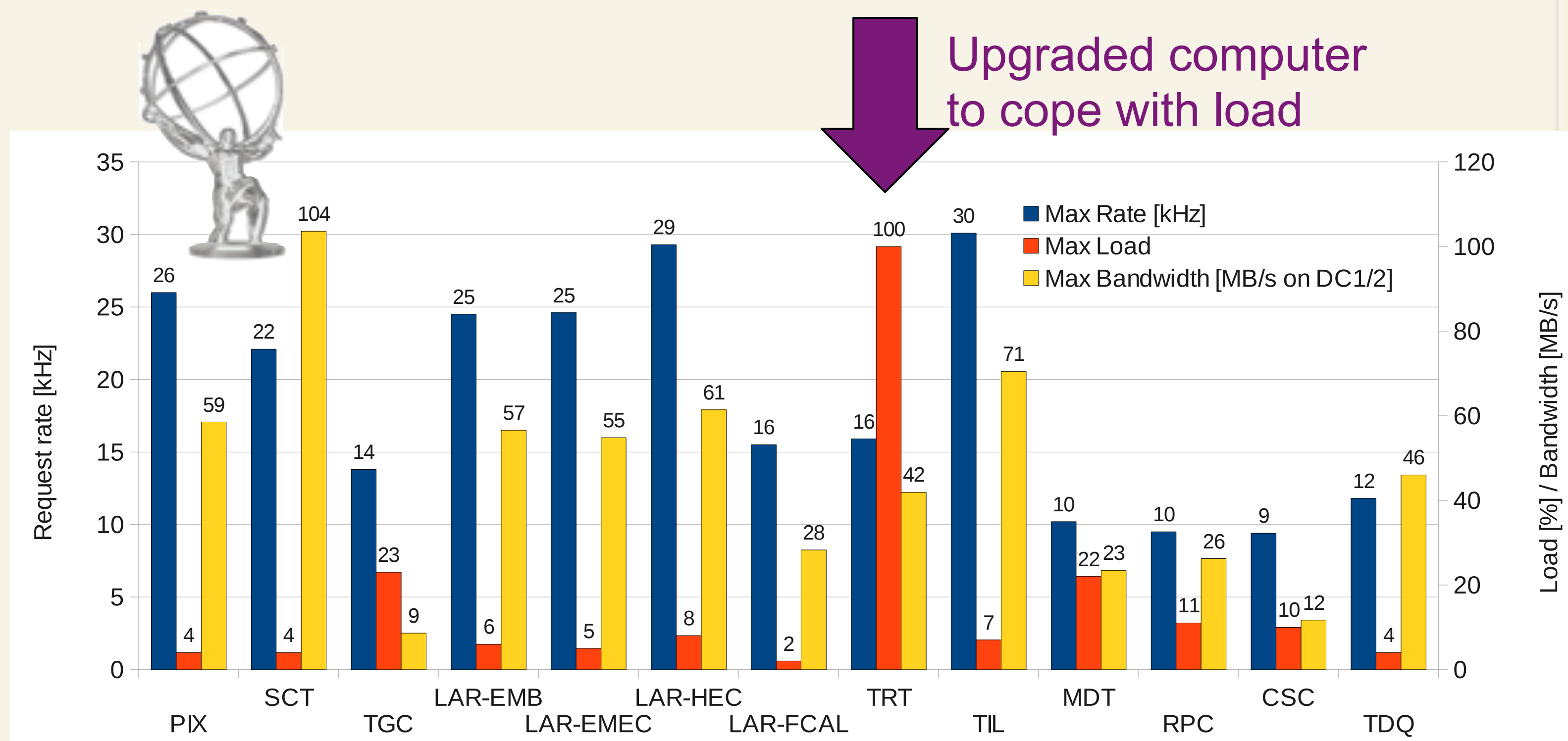
Higher pile-up than DAQ (and detectors) were designed for

- Reroute event fragments in CMS event builder to avoid bottlenecks in the super-fragment builder (1st stage of EvB)
- Invest in h/w to cope with rate, bandwidth & load



Expected 2012

Upgraded computer to cope with load



Hannes Sakulin

Increase HLT CPU Power

2009: 720x

May 2011 add: 72x

May 2012 add: 64x

	Original HLT System Dell Power Edge 1950	2011 extension Dell Power Edge c6100	2012 extension Dell Power Edge c6220
Form factor	1 motherboard in 1U box	4 motherboards in 2U box	4 motherboards in 2U box
CPUs per mother-board	2x 4-core Intel **Xeon E54**30 **Harpertown**, 2.66 GHz, 16GB RAM	2x 6-core Intel **Xeon X5650 Westmere**, 2.66 GHz, hyper-threading, 24 GB RAM	2x 8-core Intel **Xeon E5-2670 Sandy Bridge**, 2.6 GHz, hyper threading, 32 GB RAM
#boxes	720	72 (=288 motherboards)	64 (=256 motherboards)
#cores	5760	3456 (+ hyper-threading)	4096 (+ hyper-threading)
cumulative #cores	5.6k	9.1k	13.2k
cumulative #CMSSW	5k	11k	20k

Per-event CPU budget @ 100 kHz:

2009: ~50 ms / evt

2011: ~100 ms / evt

2012: ~150 ms / evt

(CPU budgets are on 1 core of an Intel Harpertown)

Hannes Sakulin,
Markus Frank (poster)

Forking of HLT Processes

HLT processes need a lot of libraries and conditions data

- Requires a lot of memory for each instance
- Not enough RAM on multicore processors
  - CMS runs 32 CMSSW instances on latest h/w

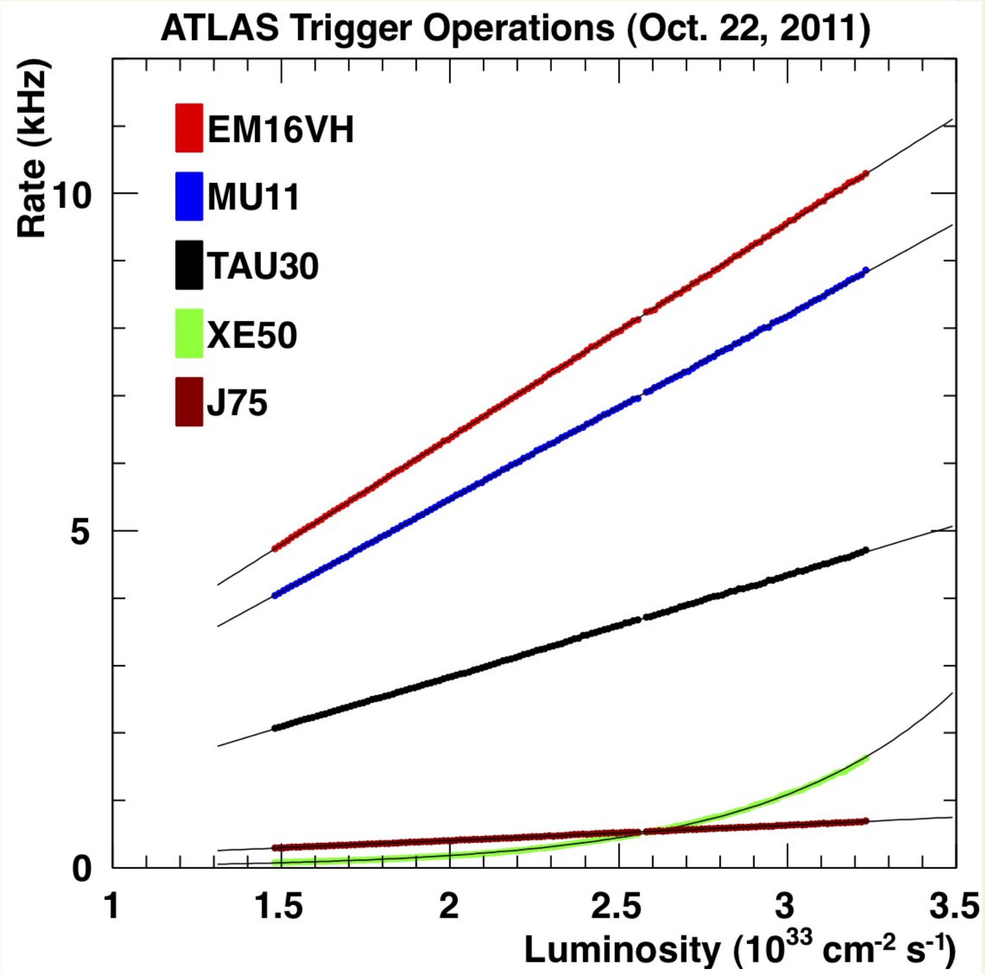
Solution: create a prototype process and then fork children

- Static data is shared between processes
- Copy-on-write takes care of duplicating memory pages when needed
- Difficulties to handle locks & sockets when forking children
  - Children inherit state from parent
- LHCb makes use of check-point file to quickly initialize children
  - Distributed to disk-less worker nodes using BitTorrent protocol

Diego Casadei, Mariusz Witek and many posters

Controlling Trigger Rates

- Raising thresholds no longer possible
  - ~ Cutting into physics phase-space
- More sophisticated HLT algorithms
  - ~ Bonsai Boosted Decision Trees (LHCb)
    - ~ Based on discrete values with fast look-up
  - ~ b-tagging and tau identification
  - ~ Rather complex topological triggers
- MET triggers specially affected by pileup
  - ~ ATLAS is using partial summing on front-ends feeding into L2



David Rohr, Jacob Howard (poster)

GPUs as Trigger Processors

ALICE uses GPUs to find tracklets in TPC

- Almost three times as fast as 6-core processor

Challenge to match physics performance of CPU and GPU due to concurrency

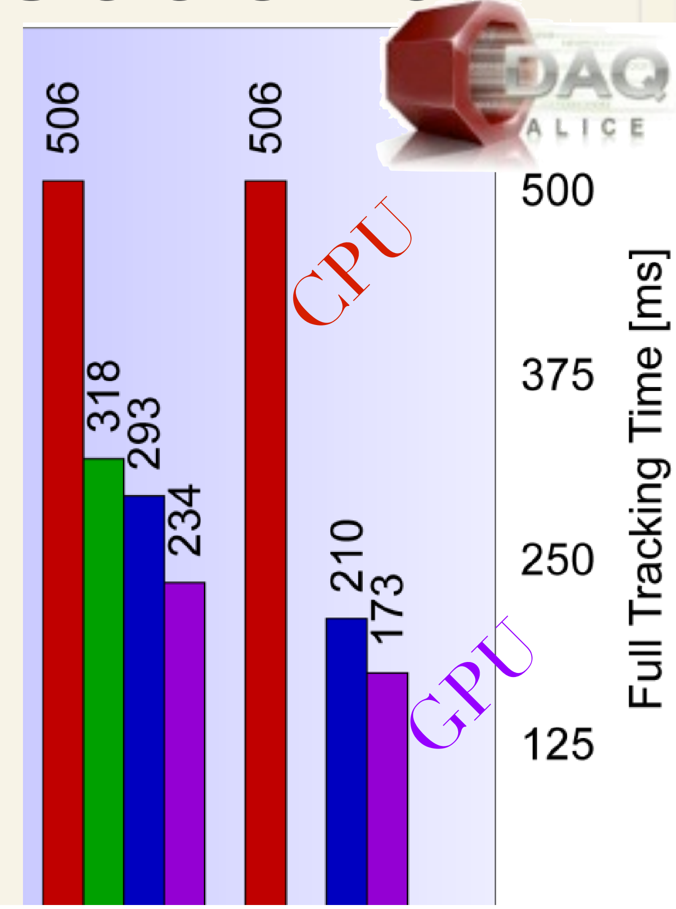
- Residual inconsistency of 0.00024% due to non-associative floating point arithmetic

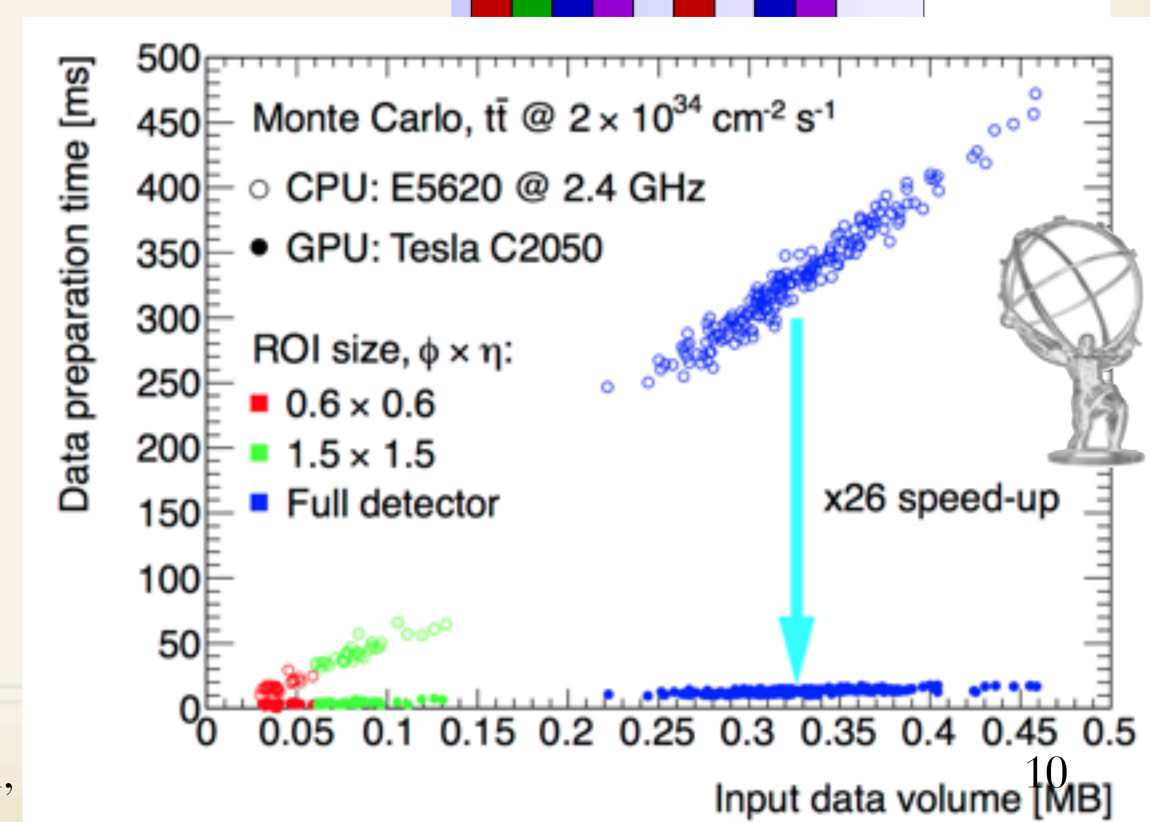
"Tracking for free"

- Most CPU cores useable for other tasks
- GPU cheaper than CPU

ATLAS studies for tracking on L2

- Performance limited by CPU

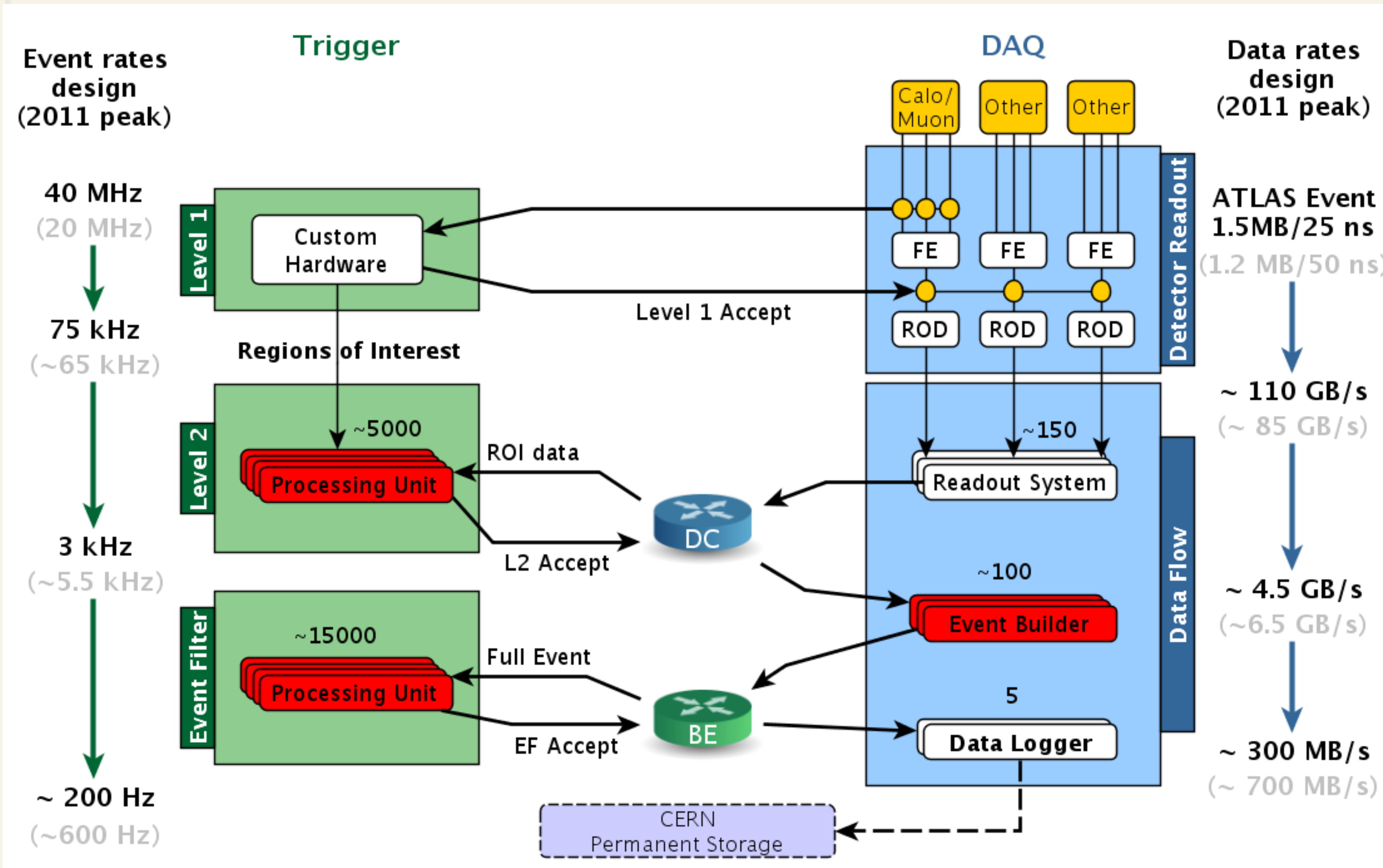




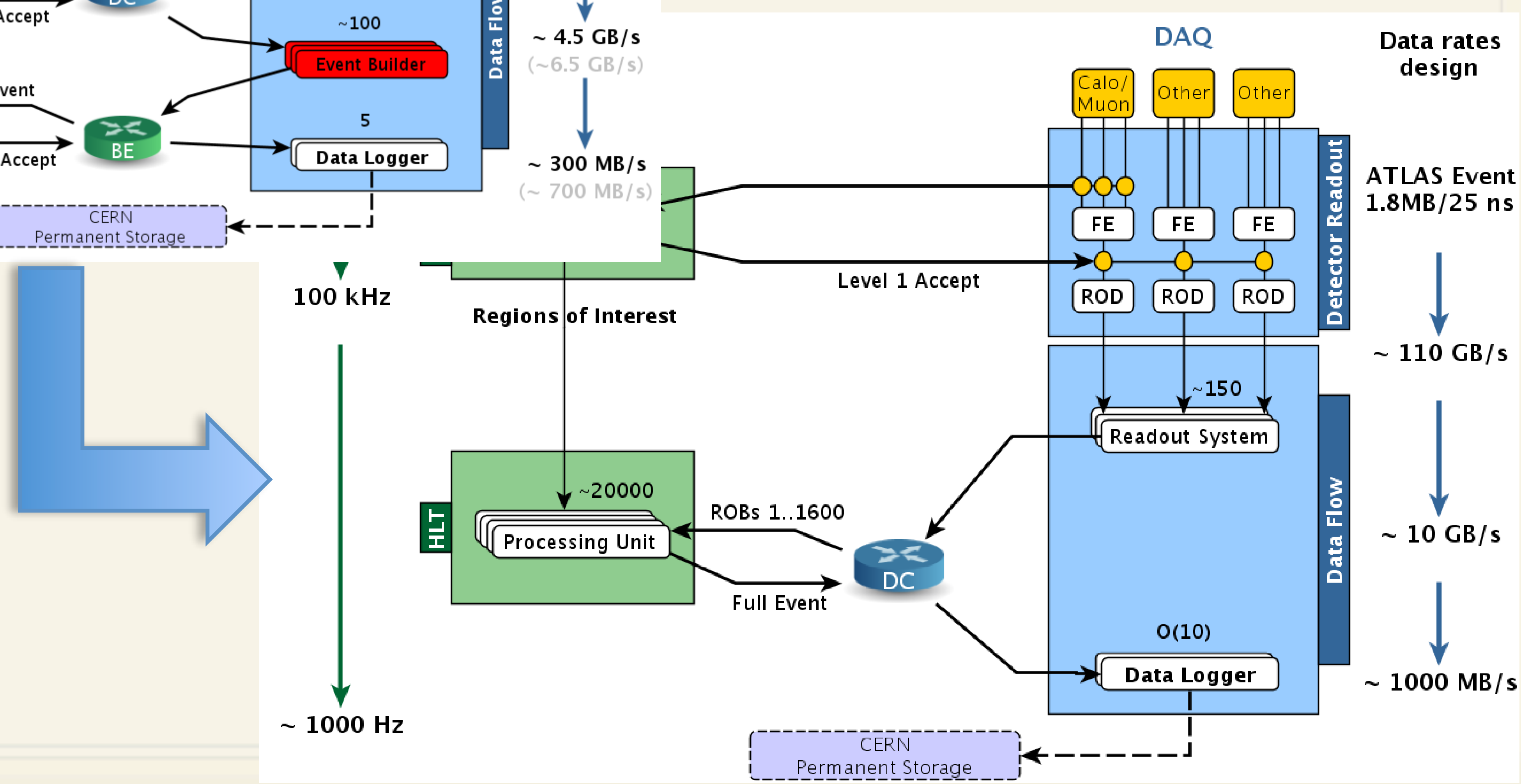
LHC
DAQ Upgrades

Andrea Negri

ATLAS – Simplifying HLT

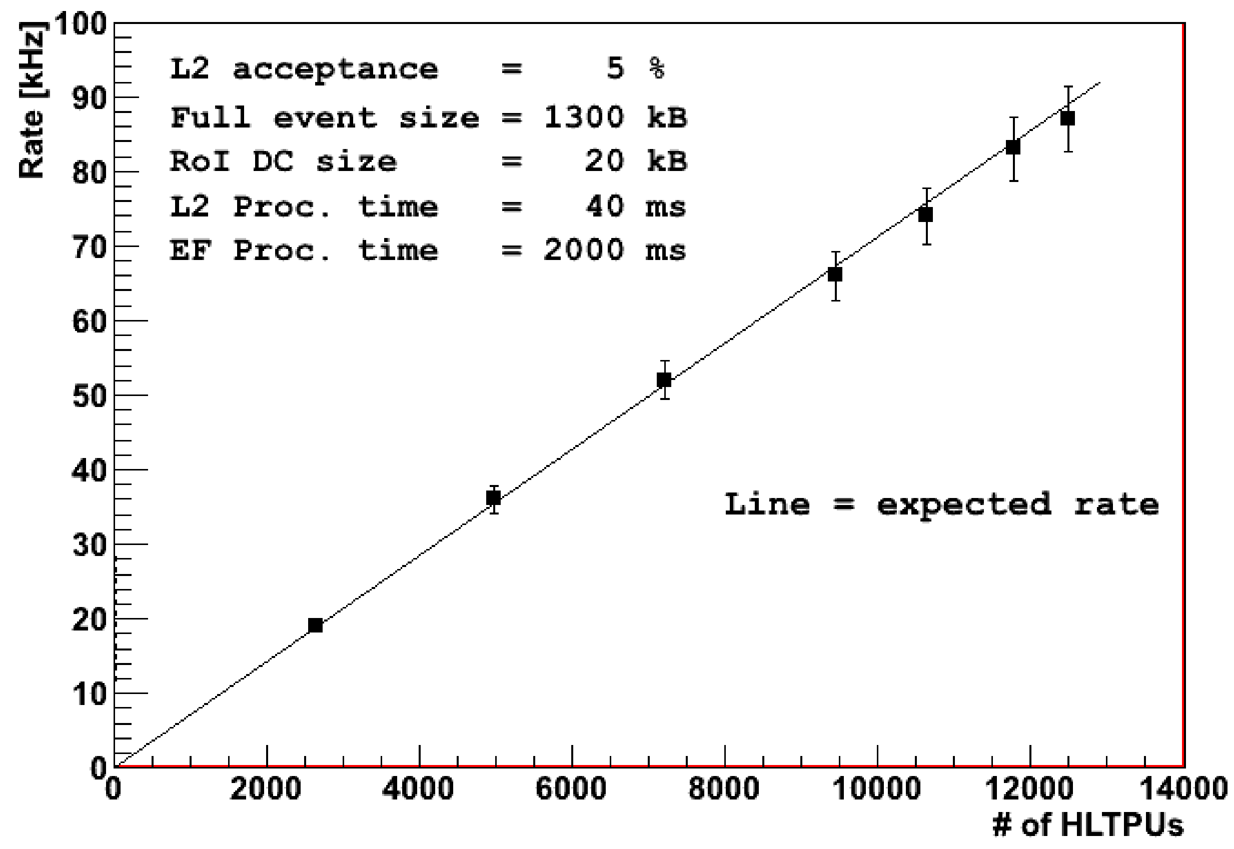


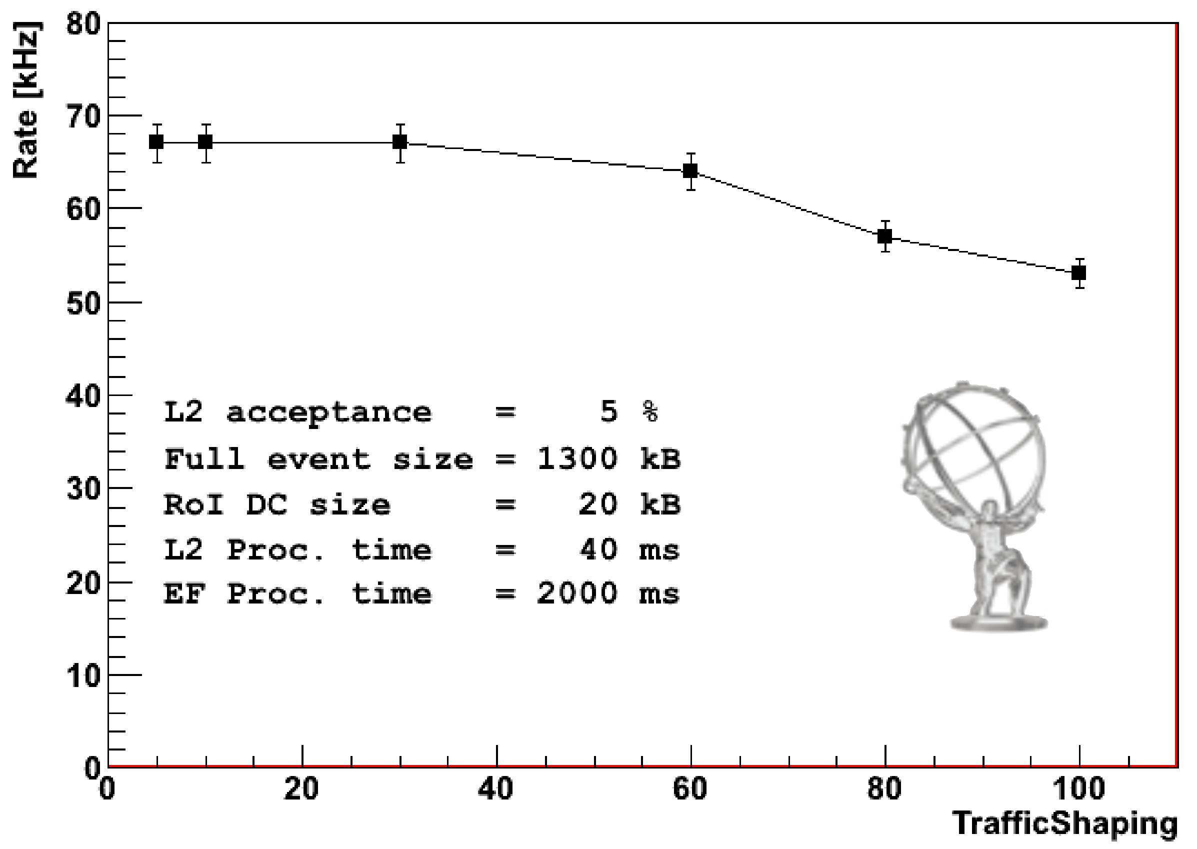
- A single farm
- Each node does RoI processing, event building, and HLT
  - Dynamic balancing btw L2 & EF
  - Less requests to ROS



ATLAS – Measurements

- Scalability validated up to 1200 HLT nodes (with 13k HLTPUs)
- Traffic shaping prevents network congestions
- A single HLT supervisor sustains more then 100 kHz
  - Evaluating merge with a s/w based RoI-Builder

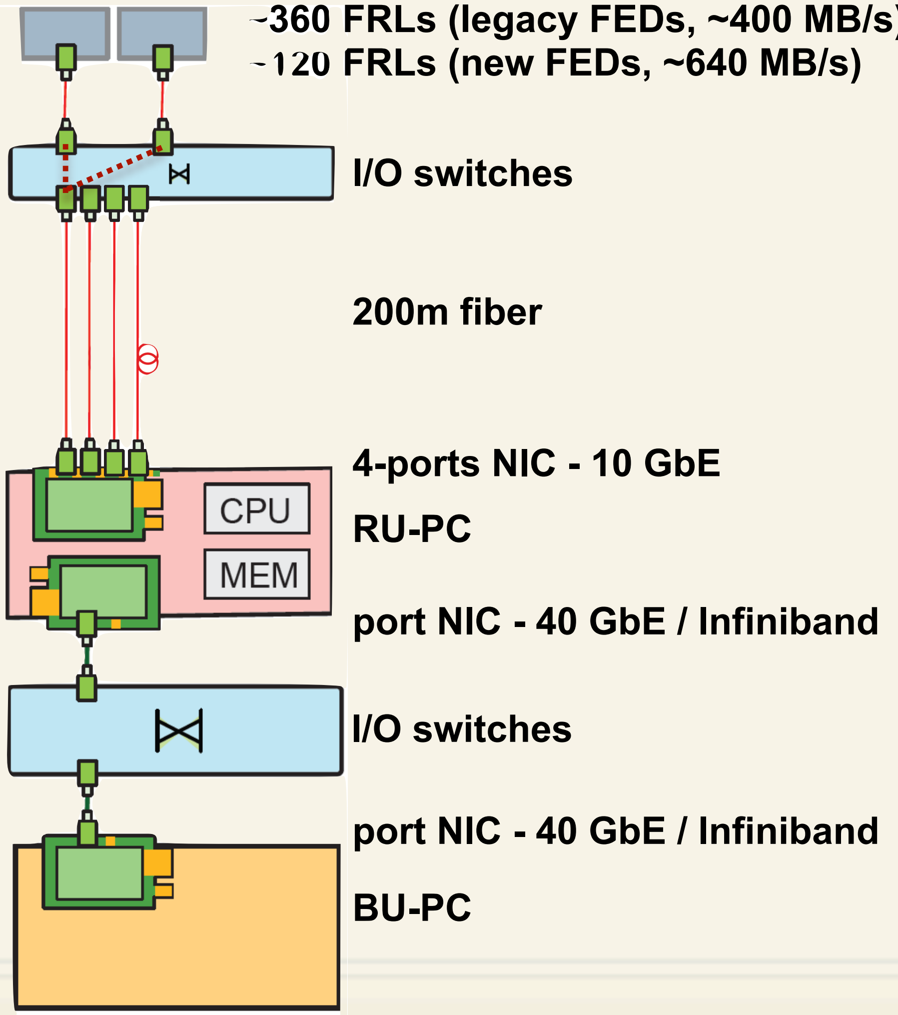




Andrea Petrucci

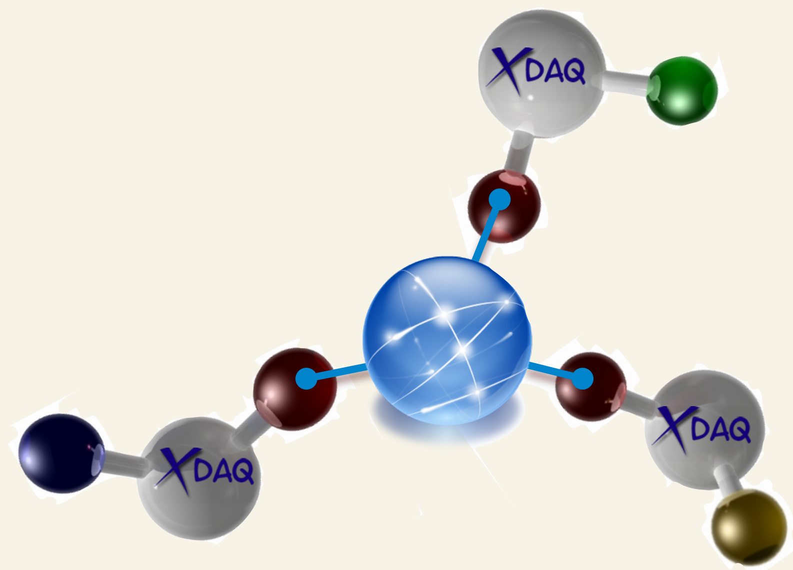
CMS – DAQ Upgrade

- Replace aging h/w (mostly >5 years old)
- Accommodate sub-detectors with new off-detector electronics
  - 37 (TRG, HCAL, HF) + 40 (Pixel - 2 x 10 GbE links) new readout links (maximal fragment size 8 kB vs 2 kB today)
  - Up to 640 MB/s readout per front-end link
- New data to surface network
  - Replace Myrinet (2x2 Gb/s) with 10 Gb Ethernet
  - Readout for legacy and new front-end drivers
- New event builder network
  - Conservative: 10 Gb Ethernet with 300x300 switch
  - Aggressive: 40 Gb Ethernet or Infiniband with 75 x 75 switch



Andrea Petrucci

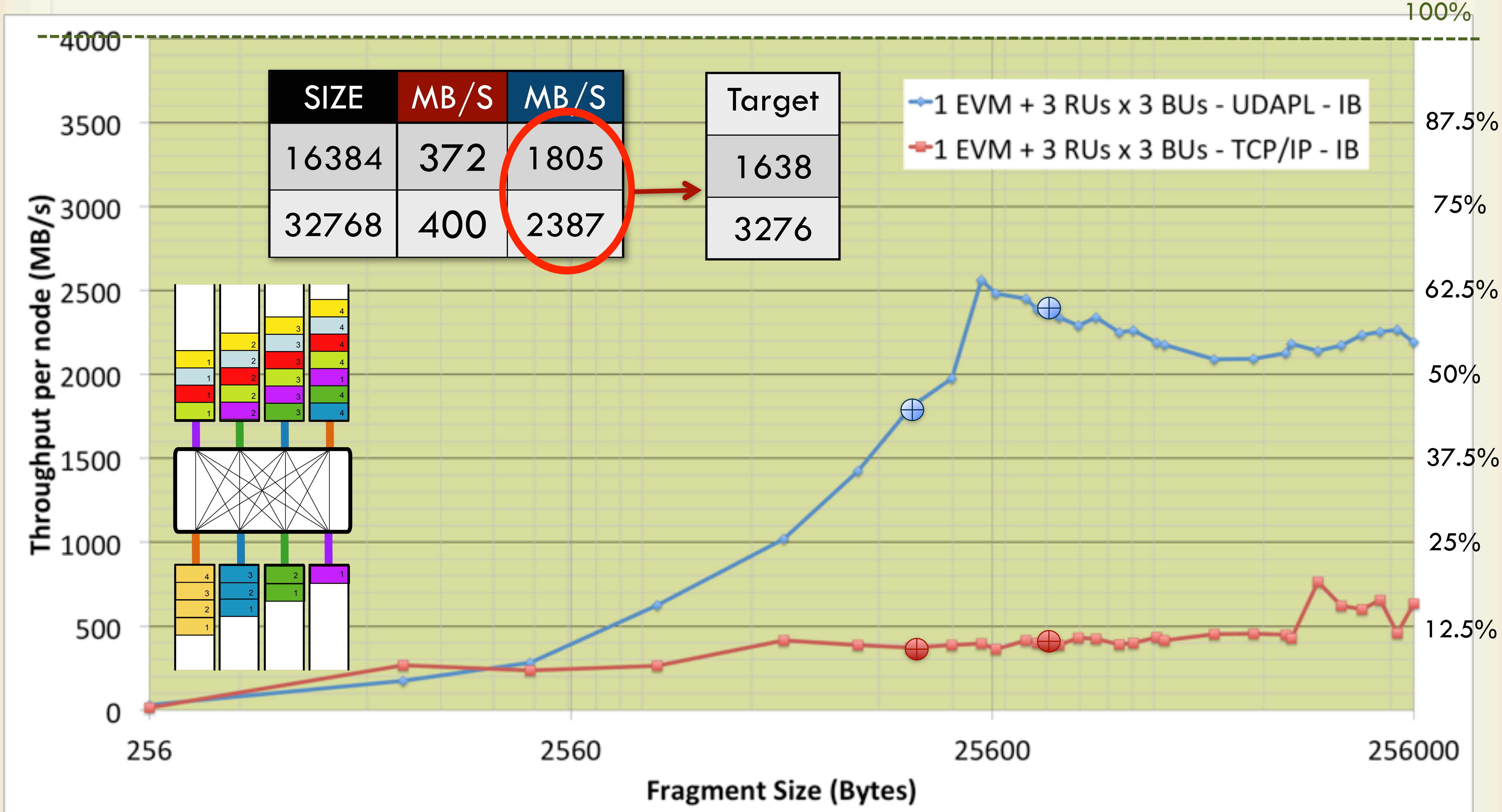
CMS – Feasibility Studies



- CMS online framework (xDAQ)
- Peer-transport as pluggable layer of various networking medium
  - SOAP/HTML: HTTP
  - I2O: TCP, Myrinet, FIFO, etc.
  - NEW: Infiniband and iWarp with zero-copy using DAT library
- Application code does not change when moving from Ethernet to Myrinet or Infiniband
  - Purely change of configuration

Andrea Petrucci

Infiniband: uDAPL vs IPoIB



SIZE	MB/S	MB/S
16384	372	1805
32768	400	2387

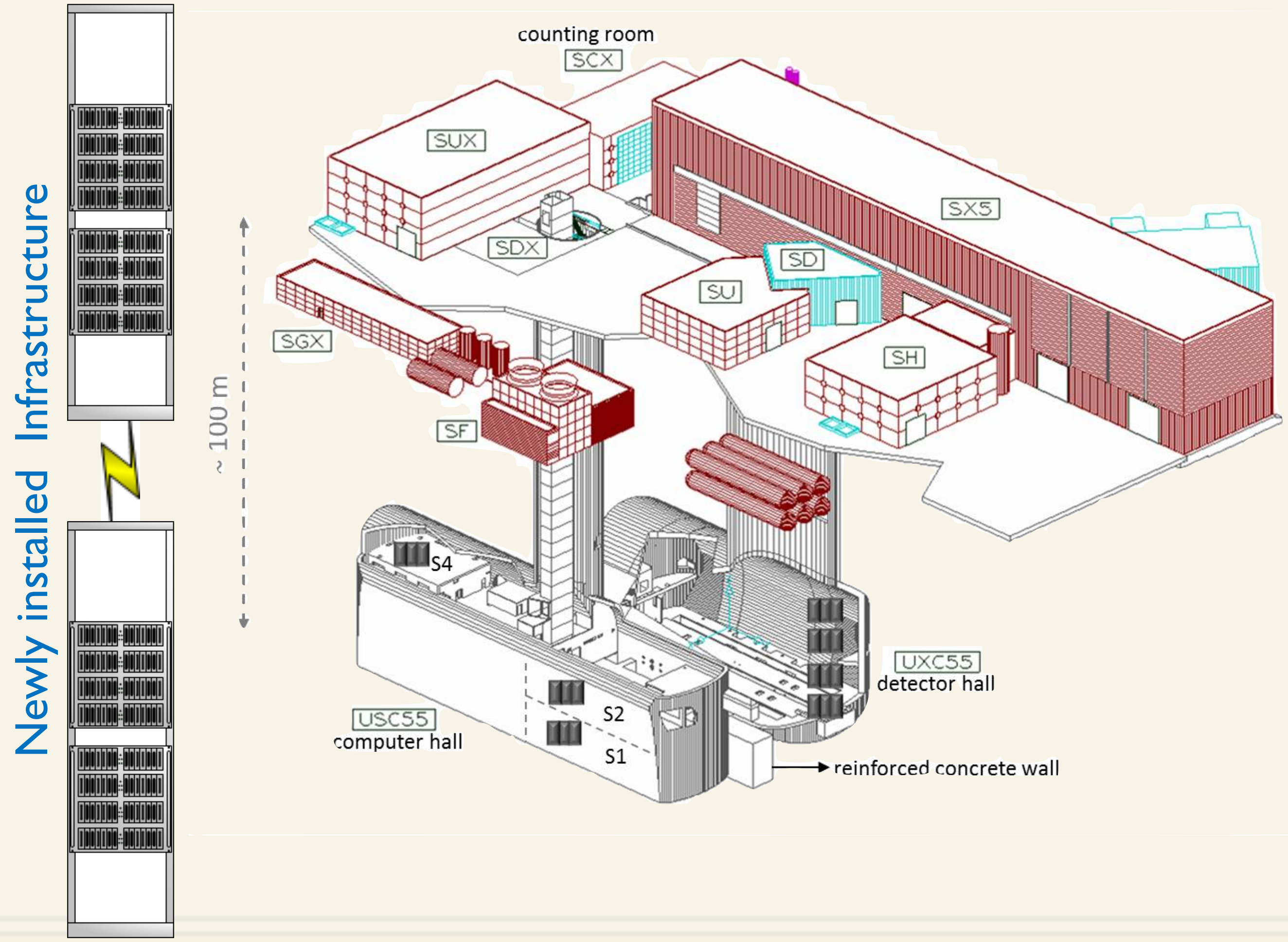
Target
1638
3276

Bobby Gomez-Reino

Detector Control System

CMS DCS upgrades during 2012

- Move from Windows XP to Windows 7 (32 to 64 bits)
- Replace aging PC boxes with fully redundant blade system
- Redundant applications with hot fail-over



Pierre Vande Vyvre (poster)

ALICE – Upgrade for 2017

TPC bus-based readout replaced with point-to-point links

- Continuous readout with ~7800 DDL3 optical links at 10 Gb/s: total 65 Tb/s

Minimum-bias trigger for slow detectors (<50 kHz)

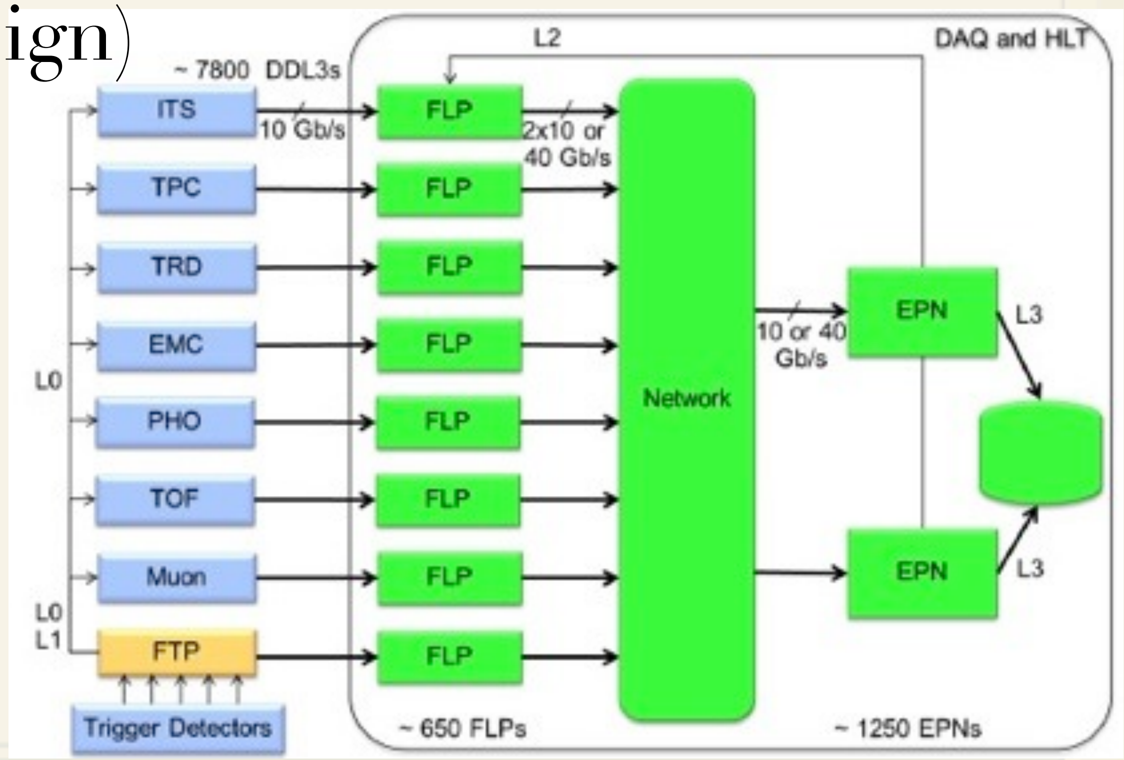
Two-steps HLT to select and compress the data

Bandwidth to mass storage: 20 GB/s (design)

Network for ~1900 nodes with a capacity of 9 Tb/s

2 options considered:

- Infiniband and Ethernet

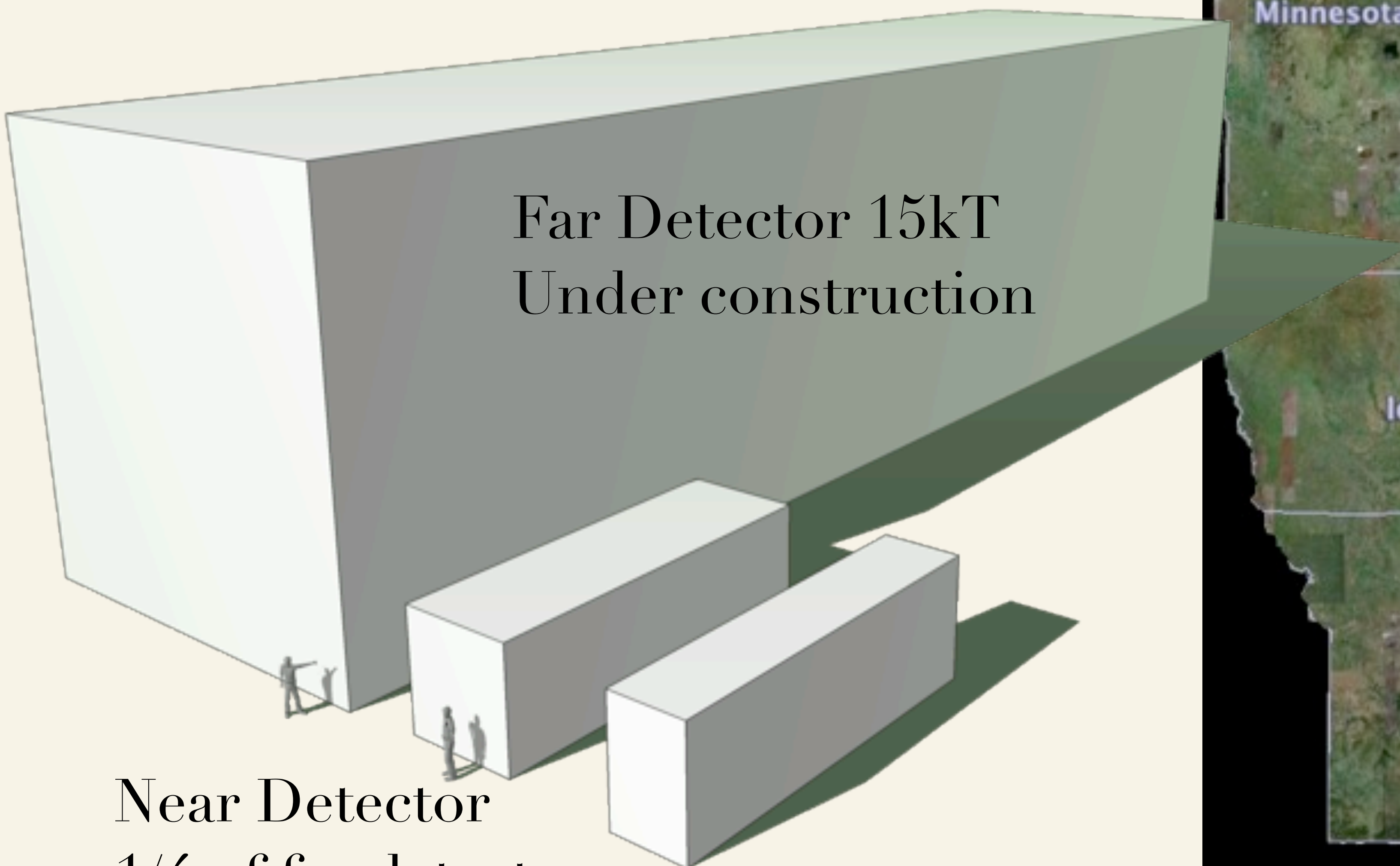


Neutrinos

$\sin^2 2\theta_{13} = 0.086 \pm 0.041 \text{ (stat)} \pm 0.030 \text{ (syst)}$

Andrew Norman

NOνA



Andrew Norman

NOνA DAQ

NOνA has a free running readout

- The electronics are always live, always digitizing
- 386,000 channels are continuously digitized and time-stamped
  - Custom timing system with ARM/PowerPC + FPGAs
  - Sophisticated synchronization scheme
- The whole system is completely deadtime-less.
- The entire raw detector data stream (up to 4.3GB/s) is actively buffered in a large computing farm

Triggering is asynchronous & decoupled from the readout

- Trigger information sent from FNAL over the internet
- Buffered data that has a time overlap with a trigger window is saved to permanent storage

Qiming Lu

NOvA Message Analyzer

A light-weight correlation analysis tool

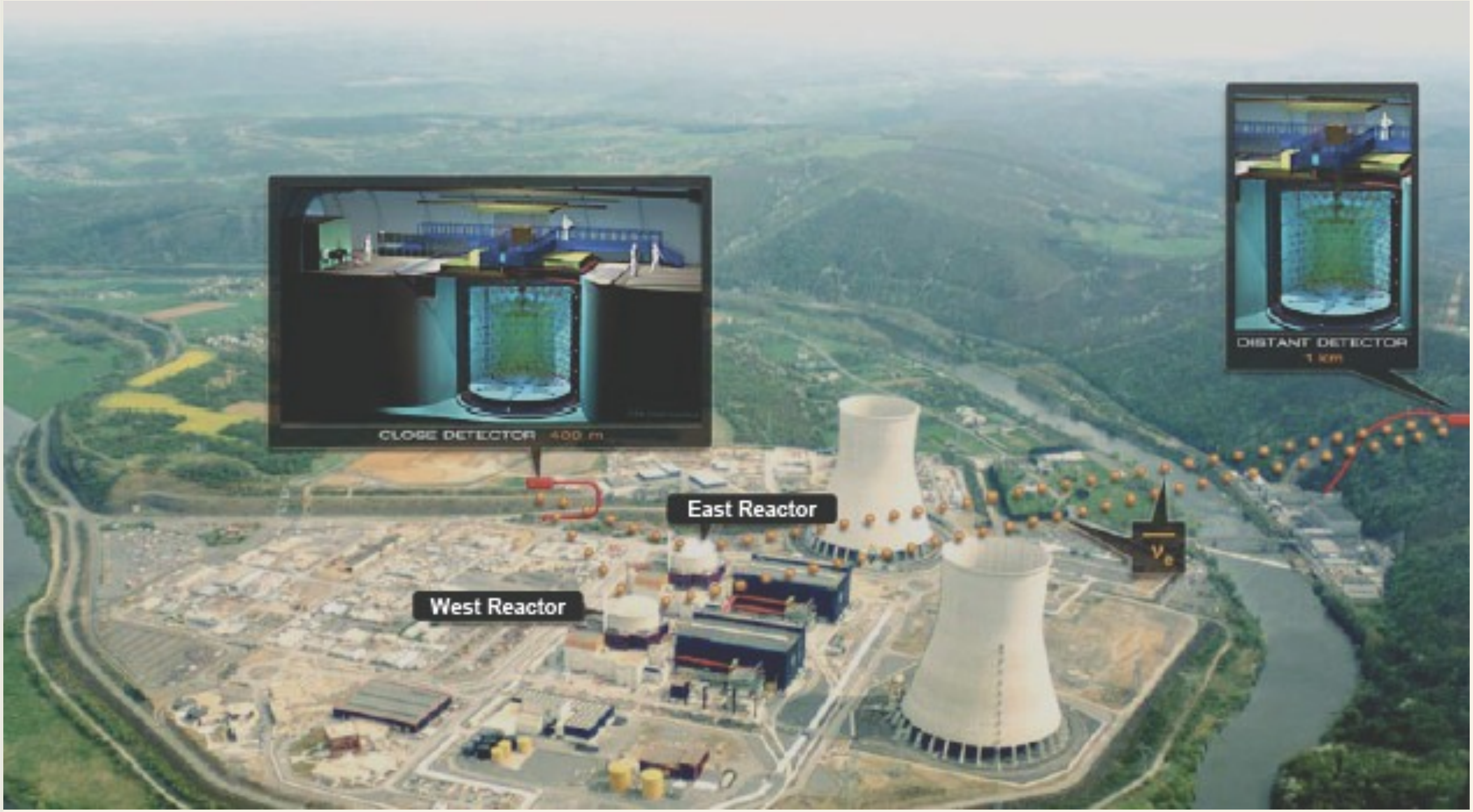
- Extract facts from text messages in log files
- Define logical rules to combine conditions and identify events
  - Formal language
  - User functions in C++

Separation of the system knowledge (rules) from the software implementation

- Easily portable to other experiments

Matt Toups

Double Chooz (France)



Matt Toups

Double Chooz DAQ

Inner and outer detectors readout separately

- Continuous readout of main detector to circular buffer through custom FADCs, VME readout (Ada sw)
  - on trigger → neutrino data
- Outer modules in daisy-chain with fanout/trigger system, readout to PC by custom USB boards
  - outer veto data

ν + outer-veto data are processed independently

- Offline coincidence of ν-DAQ event with outer-veto data

Tevatron Legacy

"The King is Dead! Long Live the King!"

Gordon Watts

Evolution over 10+ Years

Architecture lasted 10 years

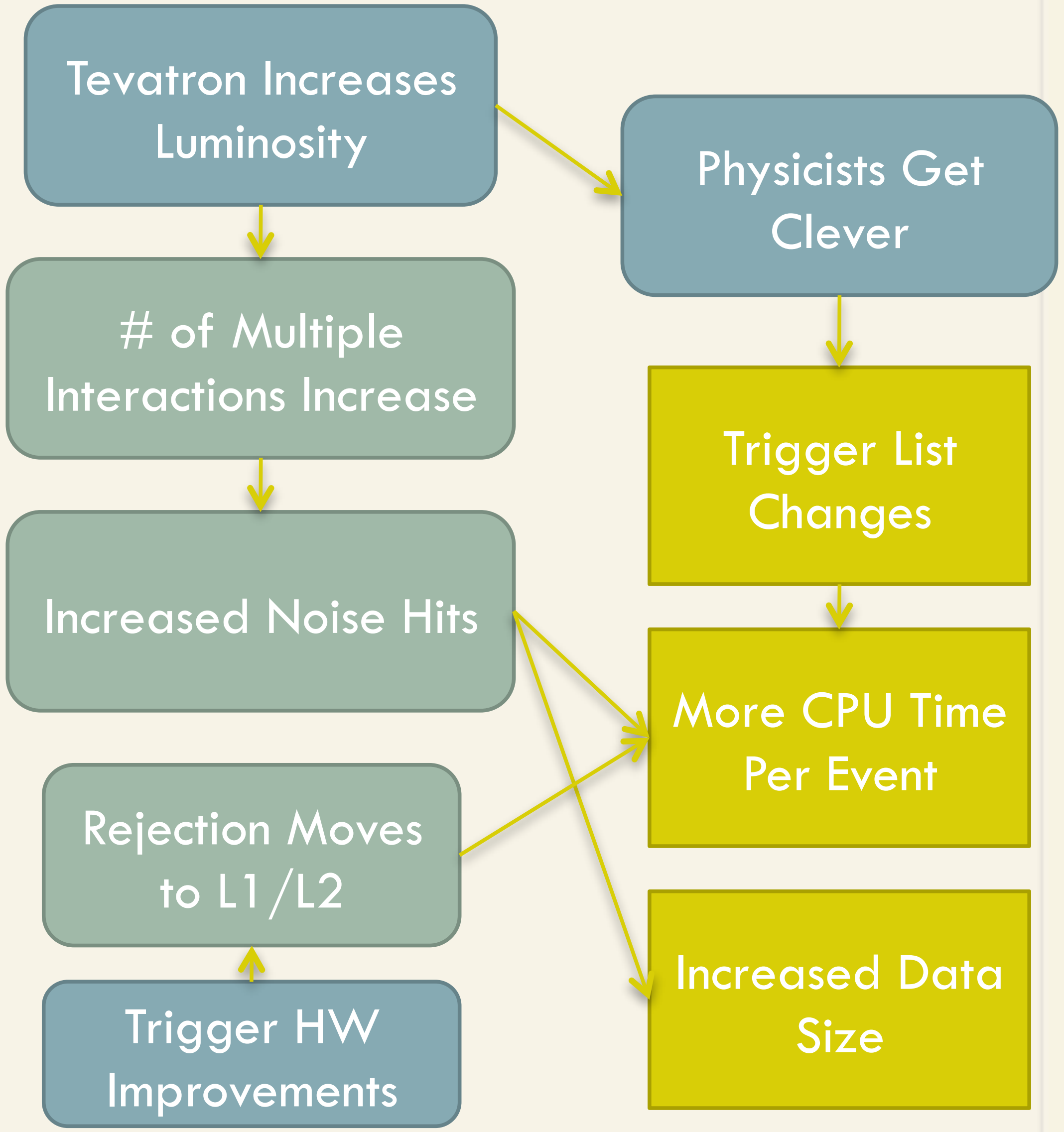
CPU time/event has more than tripled

Continuous upgrades

- Added about 10 new crates
- Started with 90 nodes, ended with almost 200, peak was about 330, all have been replaced
- Single core at start, last purchase was dual 4-core.

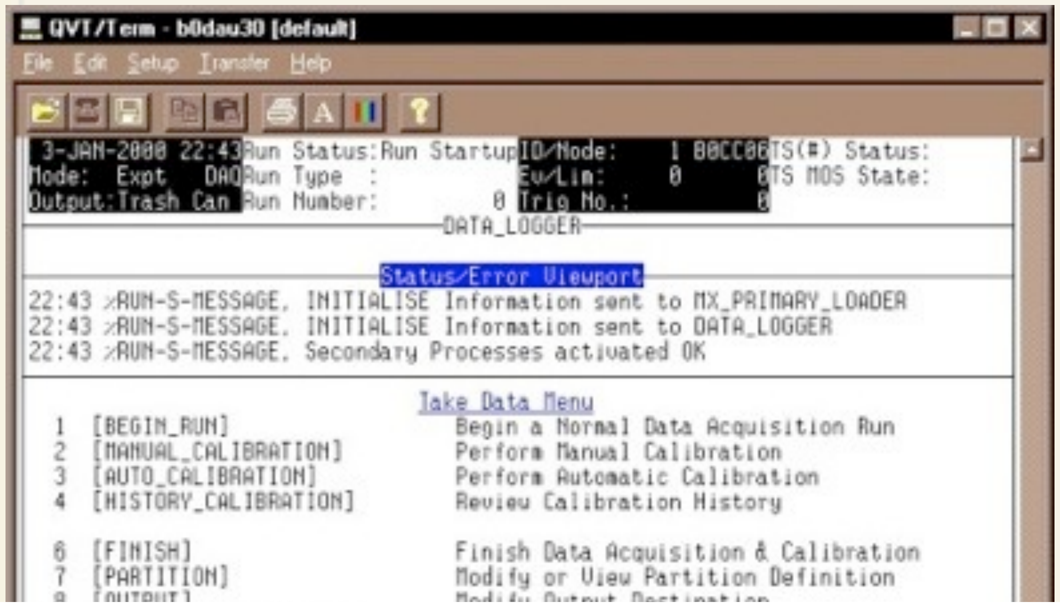
No major unplanned outages

Farm nodes the most unreliable component

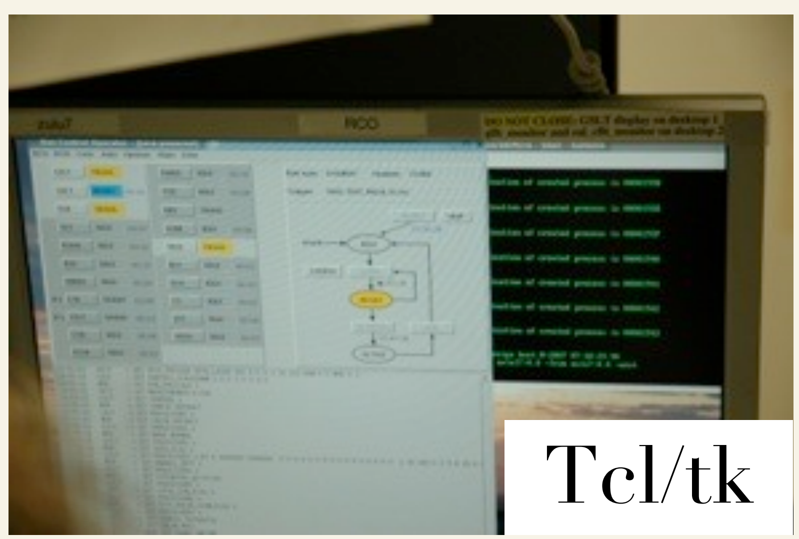


Bill Badgett

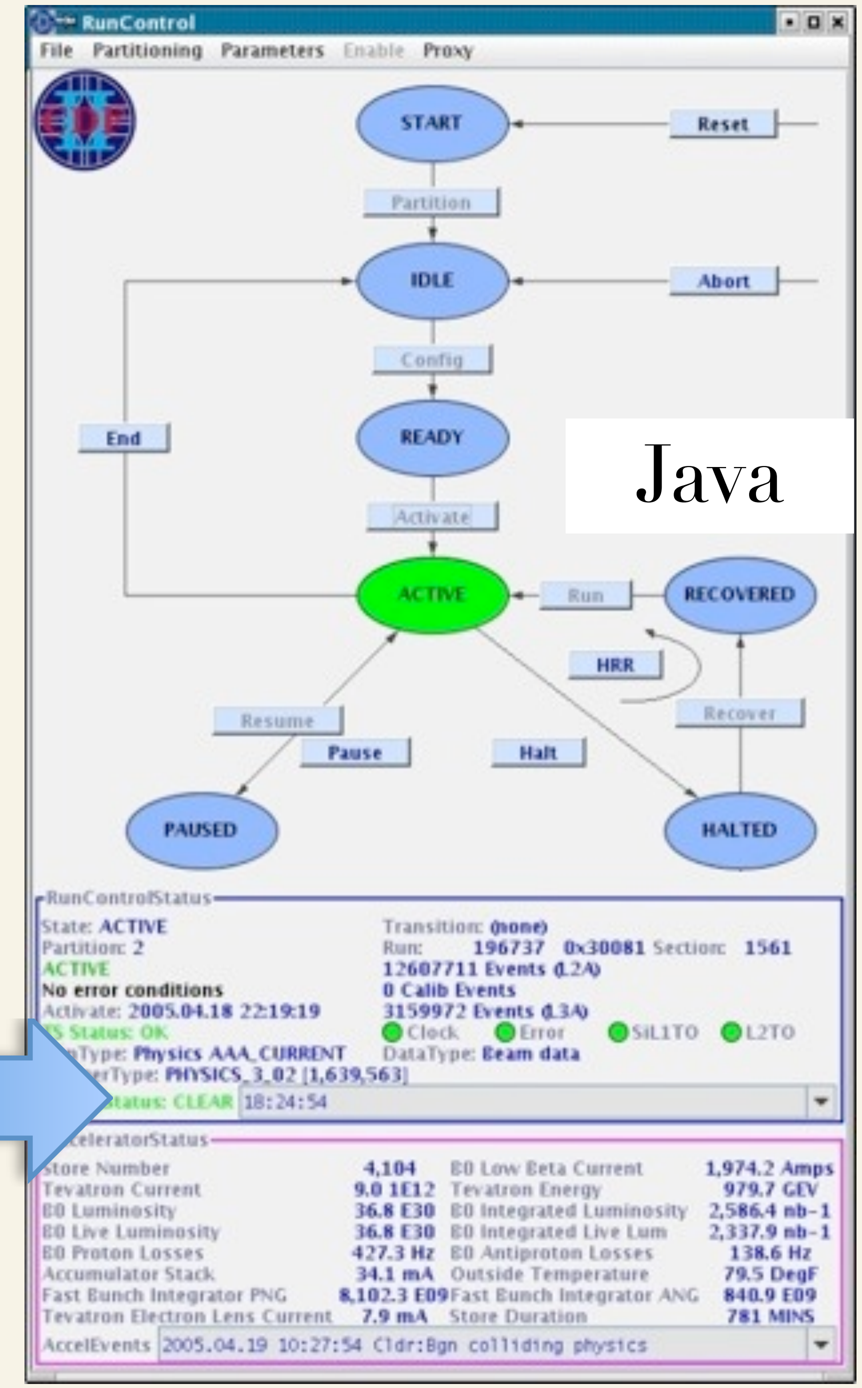
Changing s/w Technologies



Original CDF run-control on VT100 terminal



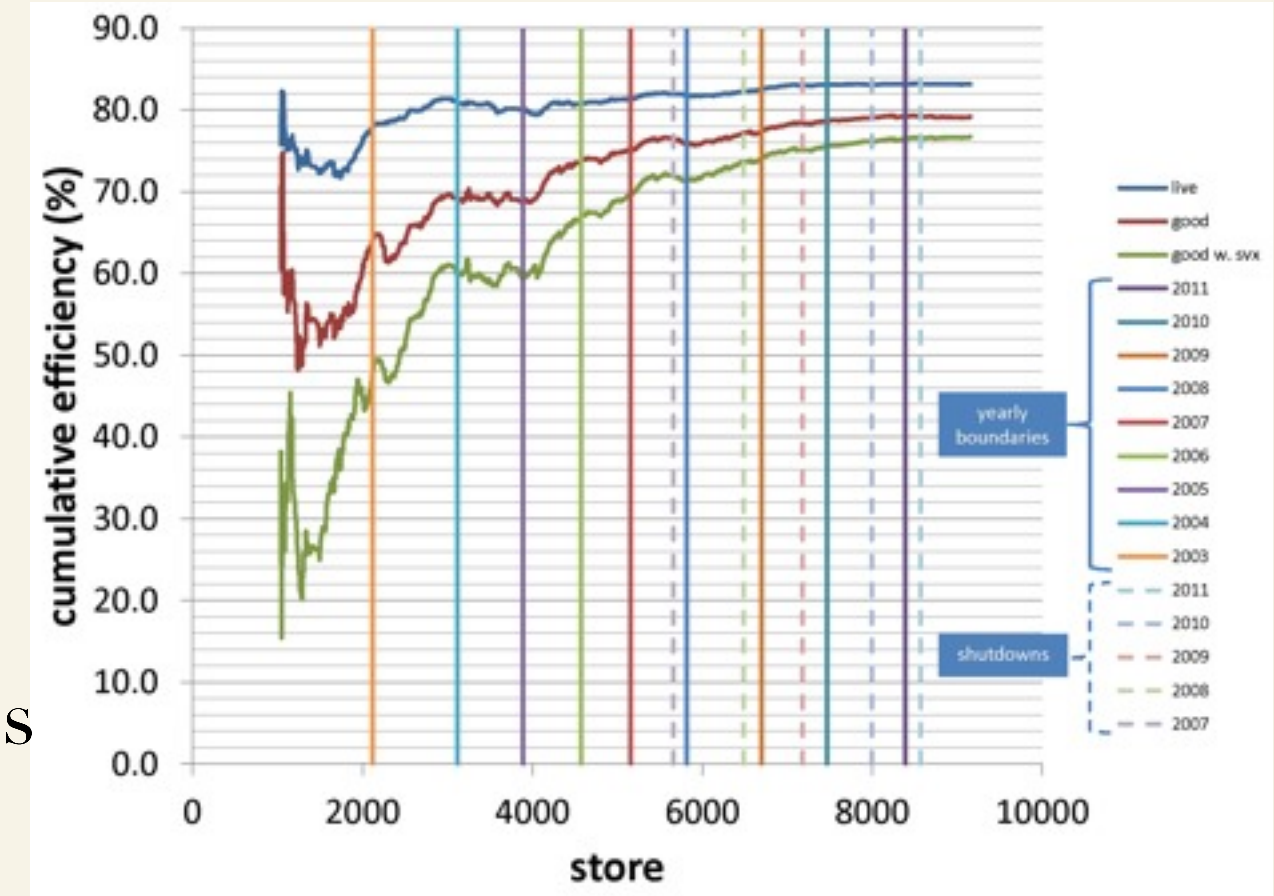
Tcl/tk



Java

Bill Badgett

CDF Data Taking Efficiency



83.2% overall efficiency

Highest lumi saturates system

- Dynamic prescaling to maximize physics reach

Detector HV ramping after beams are stable takes several minutes

Single event upsets requires complete reset of VME crate (~minute)

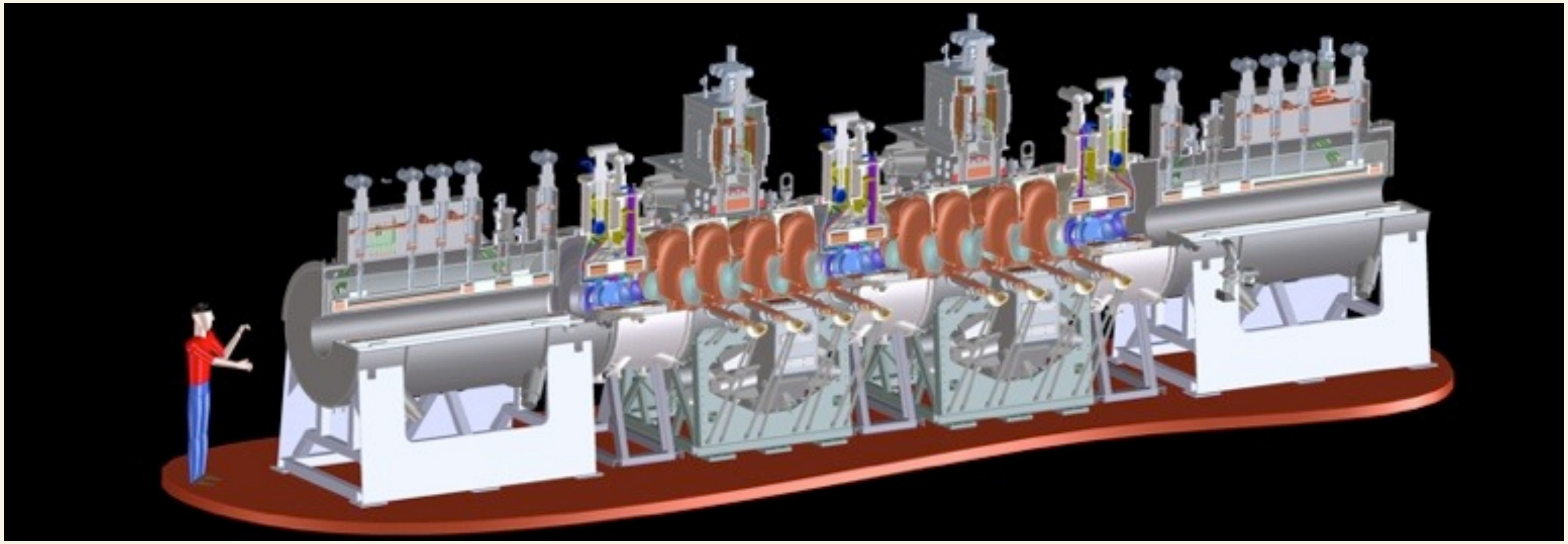
- Don't put complex computing into radiation environment

Over 10 billion physics events served

Other Experiments

Linda Coney

MICE



Muon Ionization Cooling Experiment @ RAL, UK

- Design, build, commission, and operate a realistic section of cooling channel
- Measure its performance in a variety of modes of operation and beam conditions
- Results will be used to optimize Neutrino Factory and Muon Collider designs

Linda Coney

MICE Online systems

EPICS interface for HW control and monitoring

- Configuration database

Front-end electronics VME readout

Data acquisition to PCs with DATE (ALICE DAQ)

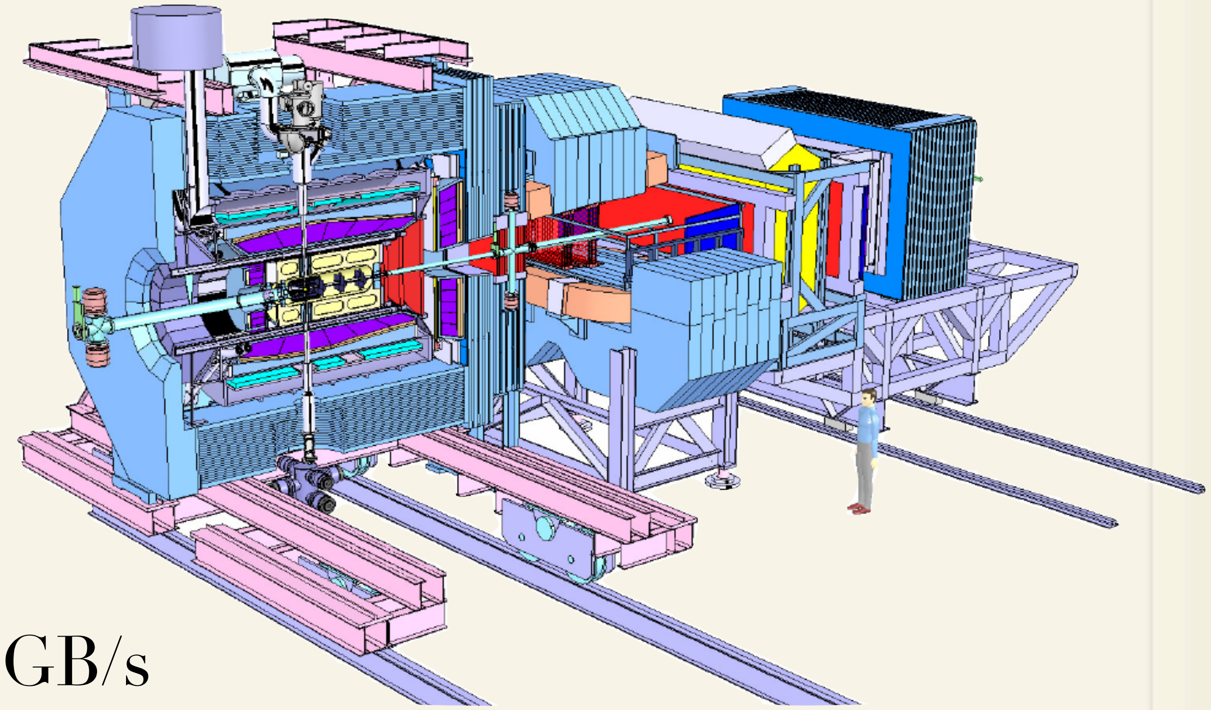
Online monitoring and reconstruction

- First look at physics and detector functionality in real-time

Data transferred to remote storage on the GRID for later analysis once a day

Krzysztof Korcyl

PANDA @ GSI



Experiment at HESR (High Energy Storage Ring) in FAIR (Facility for Antiproton and Ion Research) at GSI, Darmstadt

- 20 MHz interaction rate
- Trigger-less, continues readout at 80 GB/s
- Push-only architecture with time reference of < 20 ps
- Compute node modules with 5 FPGAs, using ATCA standard

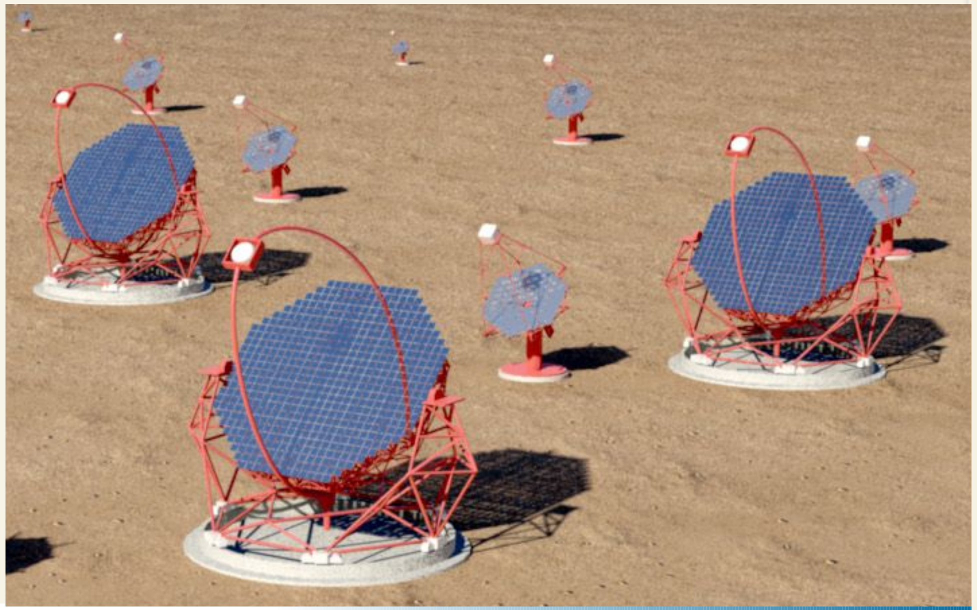
Full scale simulation of DAQ system to demonstrate performance and study the dynamics of the system

- Architecture enables event-building with 100 GB/s throughput
- Run selection algorithms on fully assembled data

Cherenkov Telescope Array

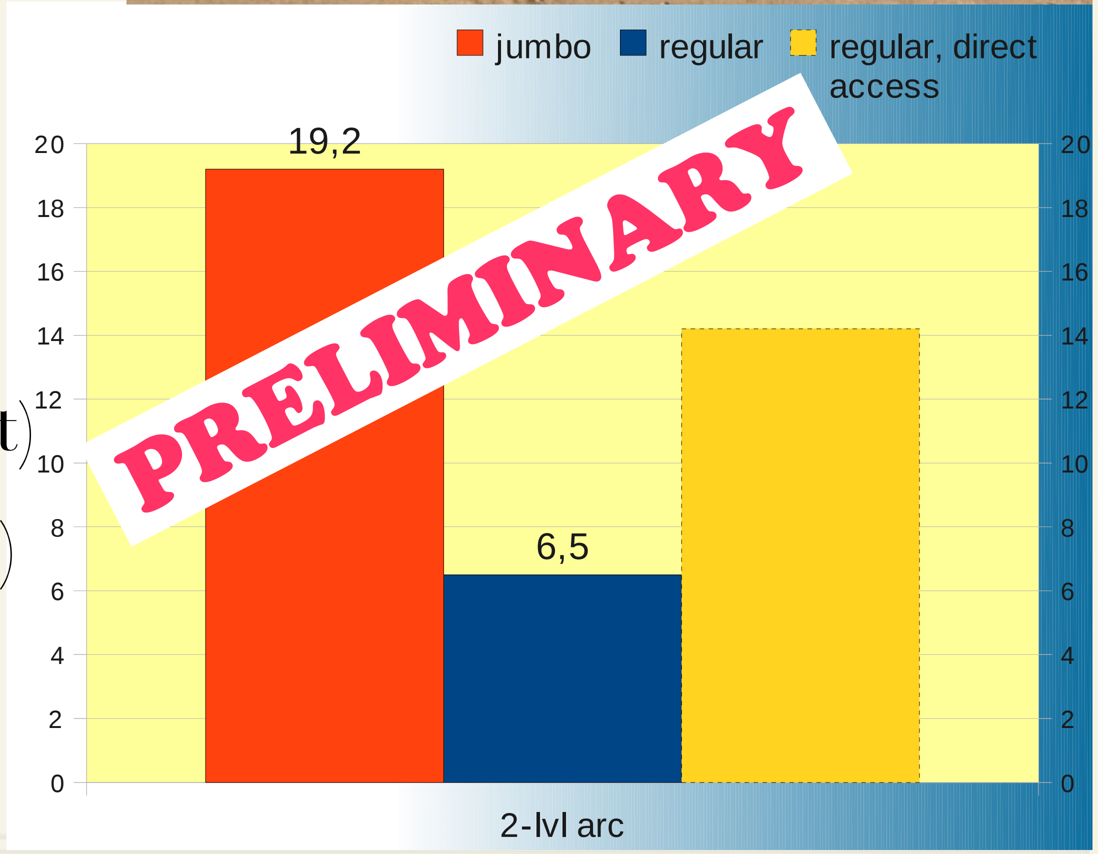
100 Cherenkov telescopes

- Each with 3000 pixel camera with a continuous 20 Gbps data stream
- Investigate several reduction options
- Simple & robust readout
- Optimized cost and industrialization



Initial tests with 10 Gbps network

- Standard ethernet not fast enough
- Use jumbo frames (limited h/w support)
- Use direct h/w access (c.f. CMS studies)



More news at CHEP 2013?

Summary

LHC DAQ systems performing exceptionally

- Also thanks to experiences from Tevatron
- Upgrade-studies underway in ATLAS and CMS for 2013/14 and ALICE for 2017

Trigger-less DAQ systems

- Continuously buffer all data in large computing farms
- Trigger is asynchronous or replaced by offline analysis

Expert-systems become popular

- Improved error-diagnostic and faster reaction times
- Formalize and preserve expert knowledge (long term support)
- Home-grown systems vs open-source tools, e.g. CLIPS