Status and Prospects of LHC Experiments Data Acquisition

Niko Neufeld, CERN/PH

Computing in High Energy and Nuclear Physics

Prague | Czech Republic | 21 - 27 March 2009

Acknowledgements & Disclaimer

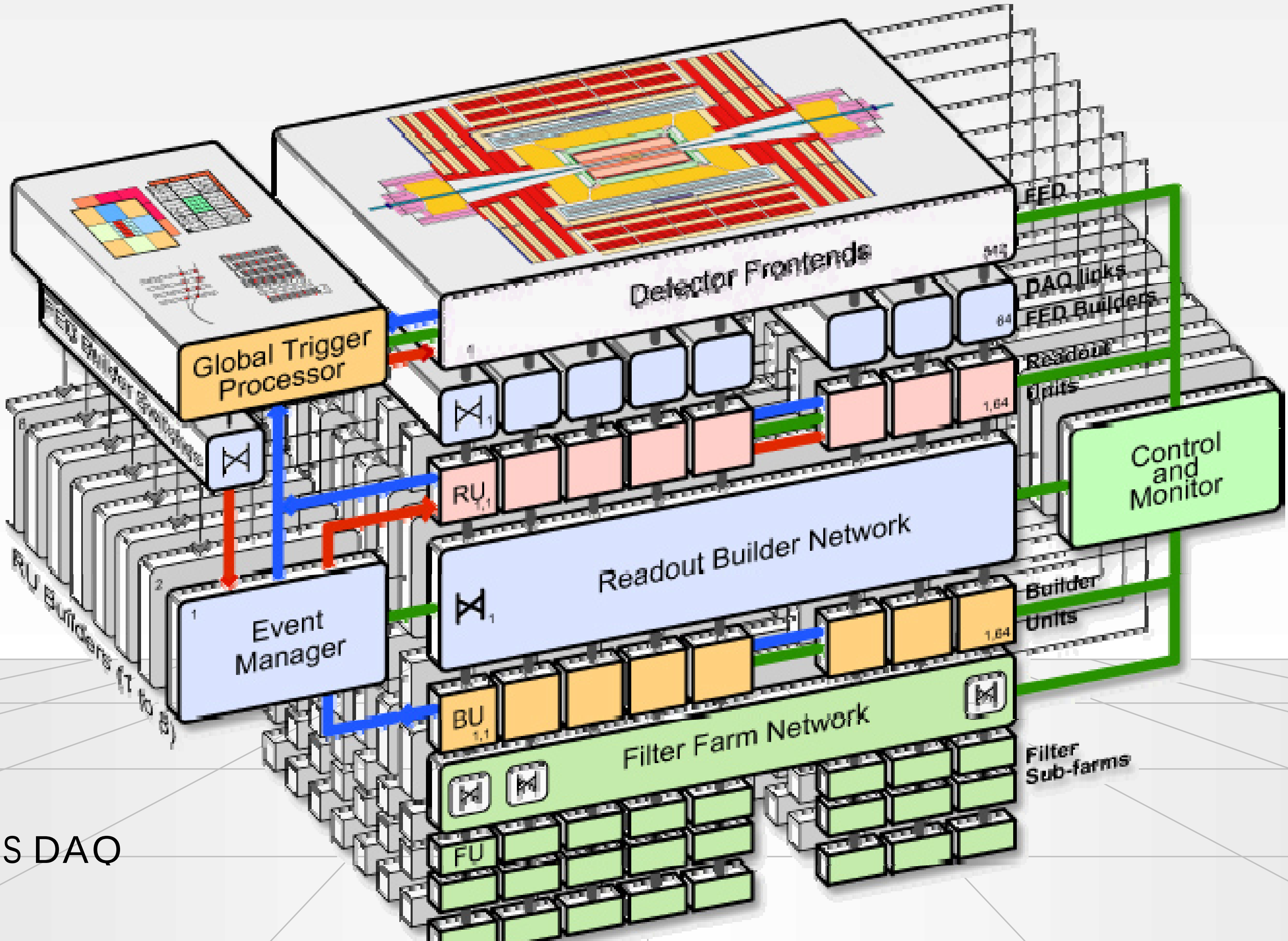
- I would like to thank Bernd Panzer, Pierre Vande Vyvre, David Francis, John-Erik Sloper, Frans Meijers, Christoph Schwick and of course my friends and colleagues in LHCb for answering my questions and sharing their ideas

- Any misrepresentations, misunderstandings are my fault
- Any value-statements are my personal opinion

Outline

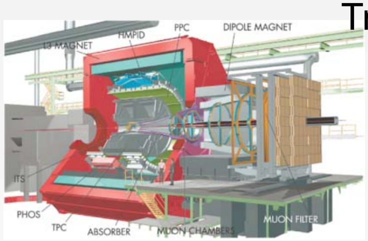
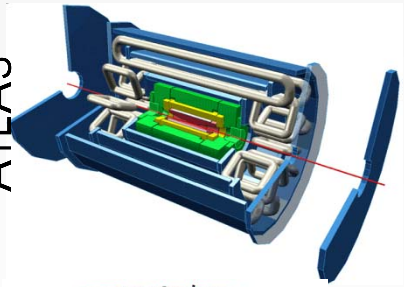
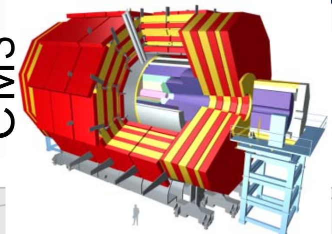
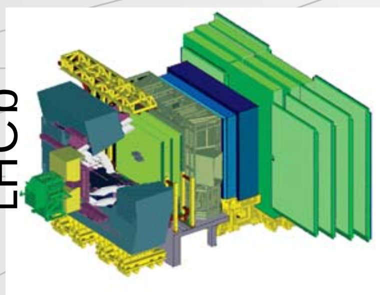
- Readout & Architecture
- Online Farms
- Run Control & Commissioning
- Outlook & Status

Readout Architectures

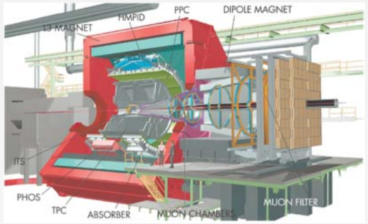
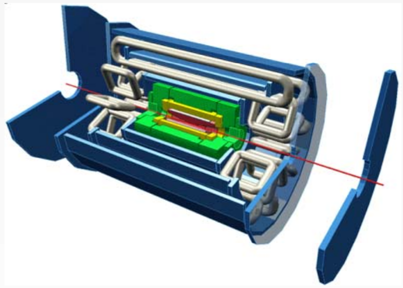
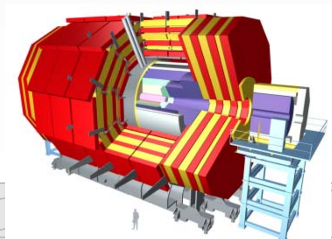
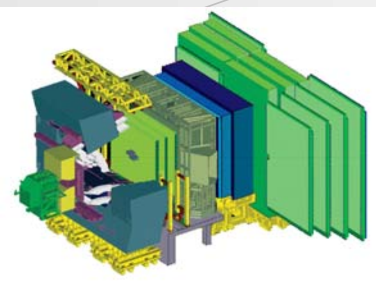


CMS DAQ

Trigger/DAQ parameters

	No.Levels Trigger	Level-0,1,2 Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	HLT Out MB/s (Event/s)
ALICE 	**4**	Pb-Pb **500** p-p **10^3**	**$5x10^7$** **$2x10^6$**	**25**	**1250** (10^2) **200** (10^2)
ATLAS 	**3**	LV-1 **10^5** LV-2 **$3x10^3$**	**$1.5x10^6$**	**4.5**	**300** ($2x10^2$)
CMS 	**2**	LV-1 **10^5**	**10^6**	**100**	**~1000** (10^2)
LHCb 	**2**	LV-0 **10^6**	**$3.5x10^4$**	**35**	**70** ($2x10^3$)

Readout Links of LHC Experiments

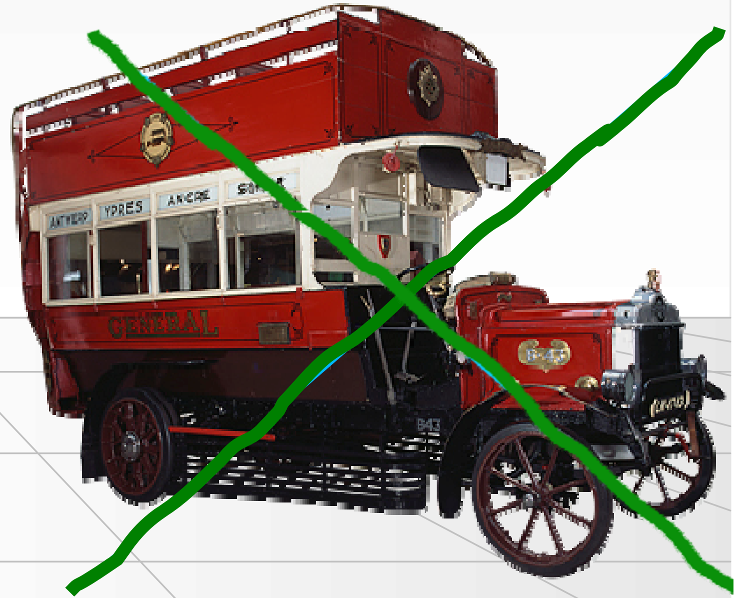
			Flow Control
	DDL	Optical 200 MB/s ≈ 400 links Full duplex: Controls FE (commands, Pedestals, Calibration data) Receiver card interfaces to PC	yes
	SLINK	Optical: 160 MB/s ≈ 1600 Links Receiver card interfaces to PC.	yes
	SLINK 64	LVDS: 200 MB/s (max. 15m) ≈ 500 links Peak throughput 400 MB/s to absorb fluctuations Receiver card interfaces to commercial NIC (Myrinet)	yes
	Glink (GOL)	Optical 200 MB/s ≈ 400 links Receiver card interfaces to custom-built Ethernet NIC (4 x 1 Gbit/s over copper)	(no)

Readout Architecture Yesterday's discussions

- Partial vs. Full readout
  - LHCb & CMS readout everything on a first-level trigger
  - ALICE has an (optional) sparse readout
  - ATLAS has a partial, on-demand, readout (Level-2) seeded by the Region of Interest (ROI) followed by a full readout
- Pull vs. Push
  - Push is used by everybody from the front-end (with backpressure except LHCb)
    - ATLAS & CMS pull in the global event-building
    - ALICE pushes over TCP/IP (implicit rate-control)
    - LHCb uses push throughout (with a global backpressure signal and central control of FE buffers)

"Point to point" the demise of buses

- All readout is on point-to-point links in Local Area Networks
  - except the sub-event building in "readout-unit" PC-servers (the last stand of the buses)
- Many have been called forward, few have been chosen: SCI, Myrinet, ATM, Ethernet (100 Mbit), **Ethernet (1000 Mbit)**, InfiniBand



Convergence

- Readout links of quite similar characteristics: can the new GBT become a universal standard?
- COTS hardware (as much as possible)
- LAN technology (actually all core Ethernet in the LHC DAQs comes from the same vendor)
- Standard protocols: Ethernet, UDP, (TCP)/IP
- Message coalescing: message rate needs to be controlled (for LHCb this is an issue even for the data packets)

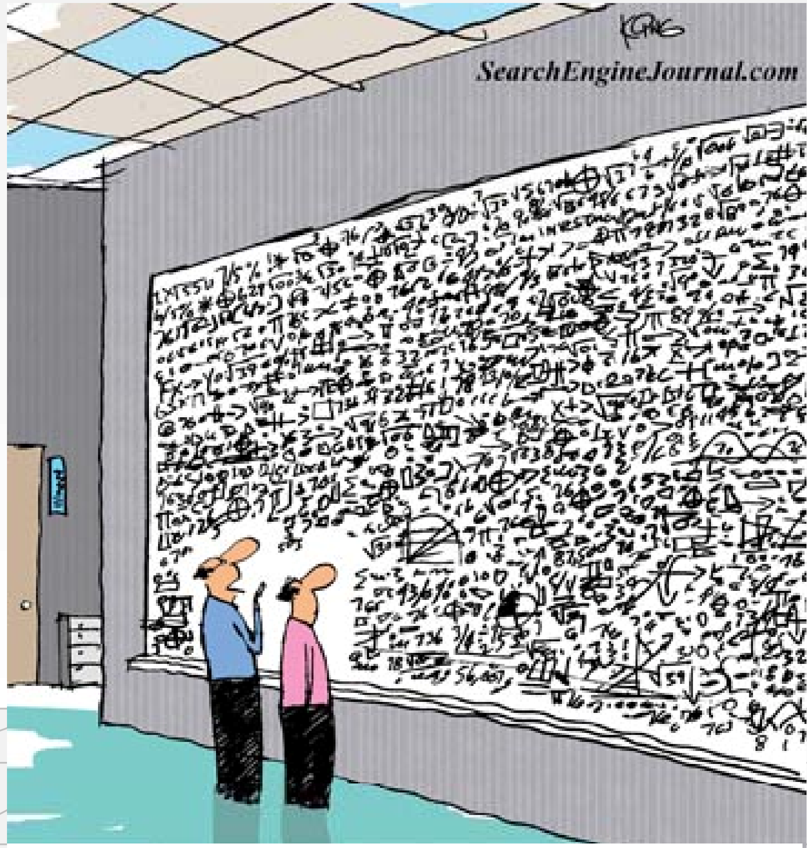
Heresy

- We have seen a lot of similar technologies & ideas
- 4 scalable systems
- Could all 4 experiments have used the same DAQ system?
- I think the answer is probably: Yes
  - Up to the output of the filter-farms
  - with suitable adaptations
- On the other hand:
  - by doing it differently we can learn from each other
  - independent teams are the best to cater to the needs of the individual experiments

A **personal** hit-list

- ALICE
  - The most versatile, universal link. A comprehensive, easy to install, well documented software stack.
- ATLAS
  - the most economical system – using the physics signature (RoI) to read out a huge detector with a relatively small LAN
- CMS
  - the optimum in scalability and elegance
- LHCb
  - The leanest. The minimum number of different components, the lightest protocol

High Level Trigger Farms



And that, in simple terms, is what we do in the High Level Trigger

Online Trigger Farms 2009

	ALICE	ATLAS	CMS	LHCb	CERN IT
# servers	81(1)	837	900	550	5700
# cores	324	~ 6400	7200	4400	~ 34600
total available power (kW)		~ 2000(2)	~ 1000	550	2.9 MW
currently used power (kW)		~ 250	450(3)	~ 145	2.0 MW
total available cooling power	~ 500	~ 820	800 (currently)	525	2.9 MW
total available rack-space (Us)	~ 2000	2449	~ 3600	2200	n/a
CPU type(s)	AMD Opteron	Intel Hapertown	Intel (mostly) Harpertown	Intel Harpertown	Mixed (Intel)

(1) 4-U servers with powerful FPGA preprocessor cards H-RORC
(2) Available from transformer (3) PSU rating

Technologies

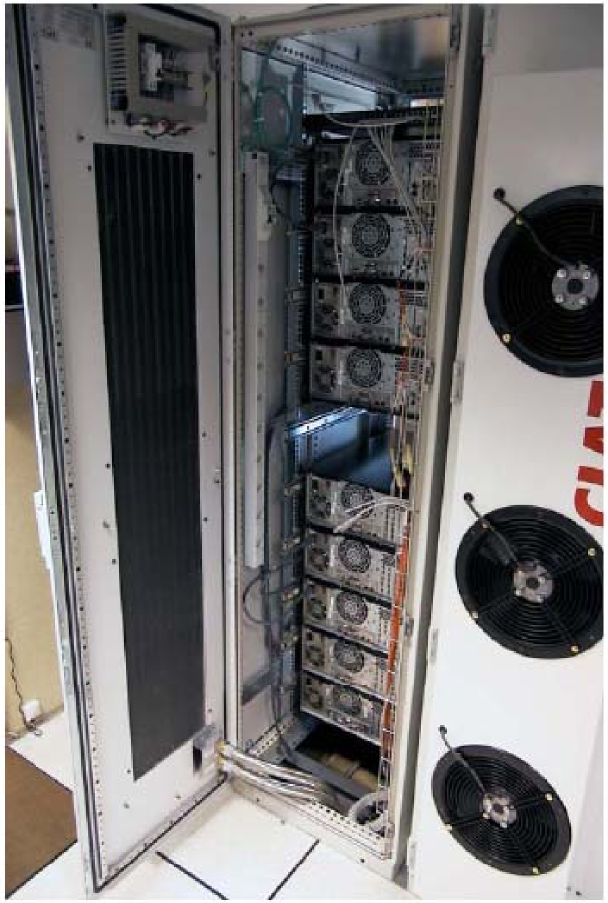
- Operating System: Linux (SLC4 and SLC5) 32-bit and 64-bits: standard kernels, no (hard) real-time. (Windows is used only in parts of the detector control-systems)
- Hardware:
  - PC-server (Intel and AMD): rack-mount and blades
  - Network (Core-routers and aggregation switches)

Managing Online farms

- How to manage the software: Quattor (CMS & LHCb) RPMs + scripts (ALICE & ATLAS)
- We all *love* IPMI. In particular if it comes with console redirection!
- How to monitor the fabric: Lemon, FMC/PVSS, Nagios, …
- Run them disk-less (ATLAS, LHCb) or with local OS installation (ALICE, CMS)
- How to use them during shutdowns: Online use only (ALICE, ATLAS, CMS), use as a “Tier2” (LHCb)

Online farms
Old problems & some "new" solutions

- The problems are always the same:
  - power, space & cooling
- Space:
  - E.g. Twin-mainboard server (Supermicro) bring 2 x 2 x 4 = 16 cores + up to 64 GB of memory on 1 U
  - Blades (typically ~ 13 cores /U)
- Power: in-rush currents, harmonic distortions
- Cooling: all experiments use heat-exchangers mounted to the back of the racks ("rack-cooler doors") instead of room air-conditioning. A co-development of all 4 experiments with support from the CERN PH department



Networks

- Large Ethernet networks
- Thousands of Gigabit ports & Hundreds of 10 Gigabit ports (e.g. ATLAS 200)
- 100es of switches
- Several separated but (partly) connected networks:
  - Experiment Technical Network
  - CERN Technical Network
  - CERN General Purpose Network
  - Experiment DAQ network
- DAQ networks are of course a critical part of the data-flow:
  - lots of monitoring: Nagios, (custom applications using) SNMP, PVSS, Spectrum
- ALICE and LHCb have dedicated Storage Area Networks (SAN) based on FibreChannel. 200 FC4 ports (ALICE), 64 FC4 / 8 FC8 (LHCb)

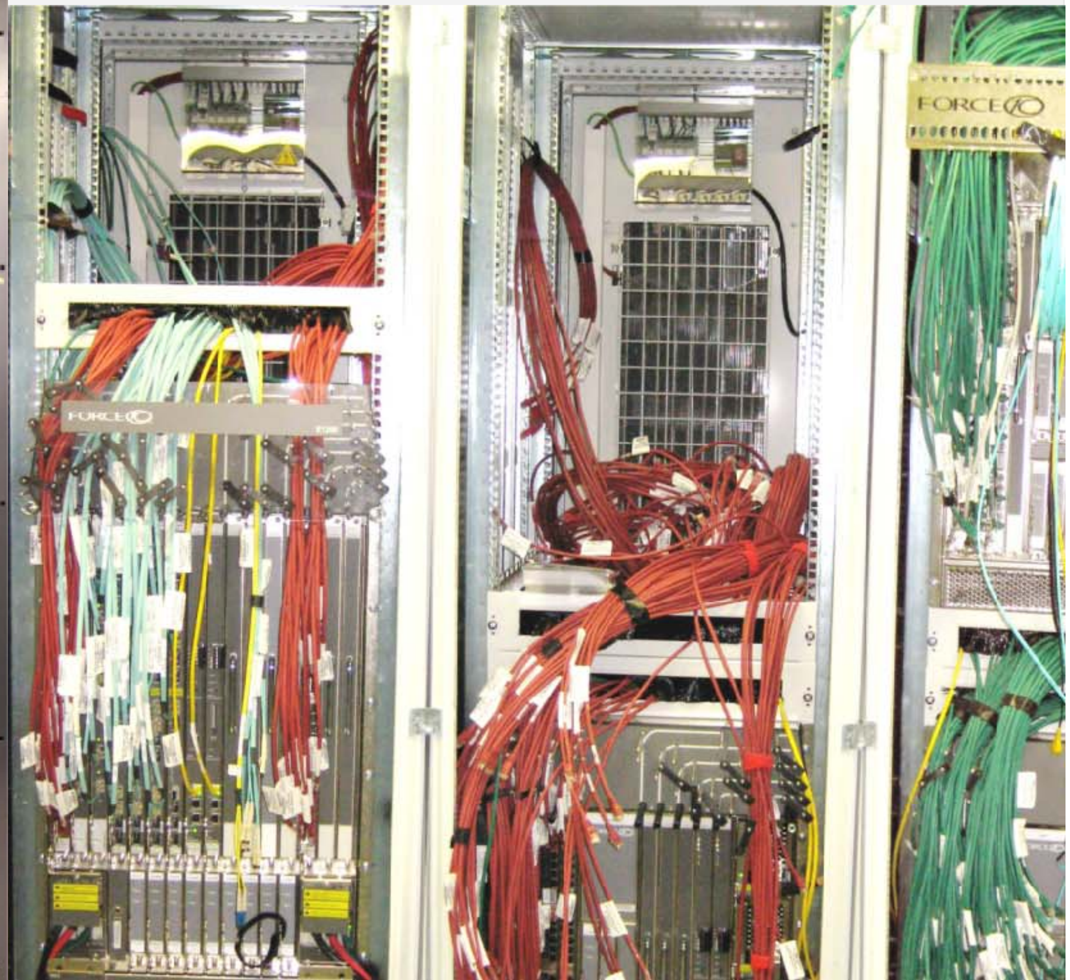
Problems

- Quality of commodity hardware:
  - memory, mainboards, disks, power-supplies, switch-ports, riser-cards
- Software stack: firmware issues (in BMCs, switches, routers), OS (e.g. Ethernet device numbering)
- Hardware obsolescence: PCI-X cards, KVM
- Heterogeneity of the hardware
  - Purchasing rules lead to many different vendors /warranty contracts over the years → manifold procedures, support-contacts

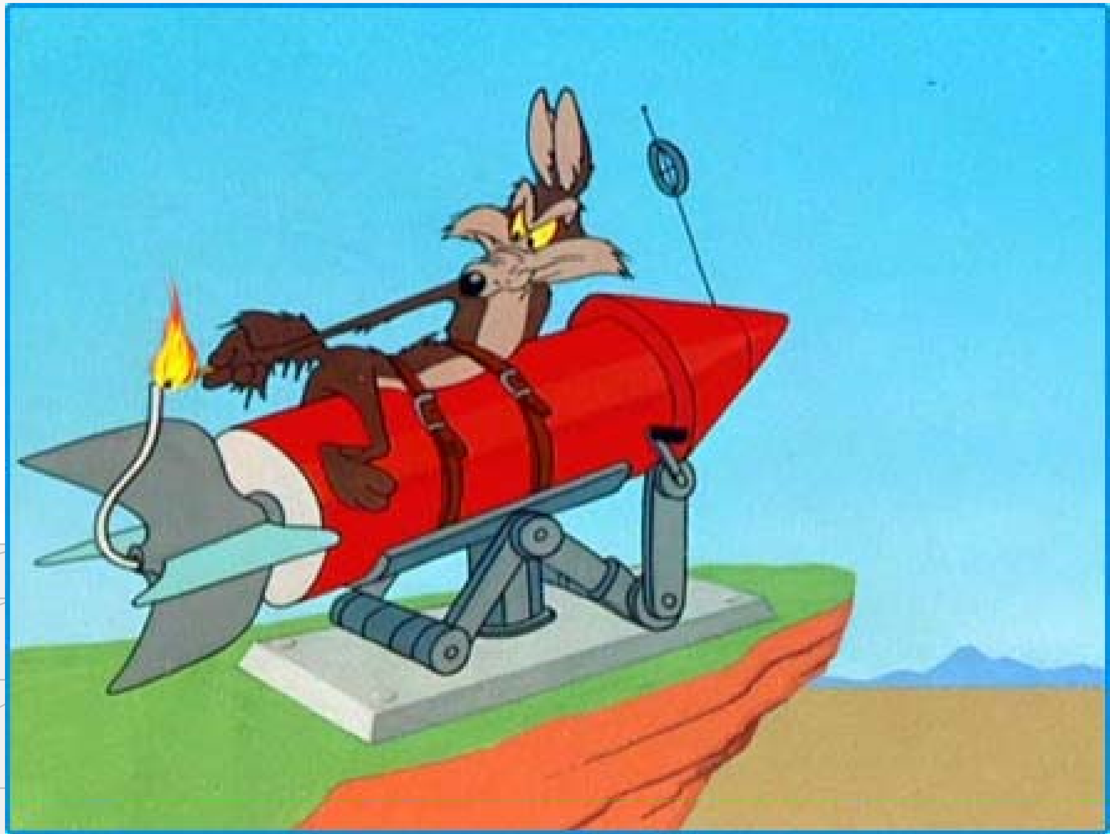
Gallery

ALICE Storage System

Online Network Infrastructure



Runcontrol



© Warner Bros.

Runcontrol challenges

- Start, configure and control O(10000) processes on farms of several 1000 nodes
- Configure and monitor O(10000) front-end elements
- Fast data-base access, caching, pre-loading, parallelization and all this 100% reliable!

Runcontrol technologies

- Communication:
  - CORBA (ATLAS)
  - HTTP/SOAP (CMS)
  - DIM (LHCb, ALICE)
- Behavior & Automatisation:
  - SMI++ (Alice)
  - CLIPS (ATLAS)
  - RCMS (CMS)
  - SMI++ (in PVSS) (used also in the DCS)
- Job/Process control:
  - Based on XDAQ, CORBA, ...
  - FMC/PVSS (LHCb, does also fabric monitoring)
- Logging:
  - log4C, log4j, syslog, FMC (again), ...

How fast can we start it?

"Starting" a run here means bringing the DAQ from the "Unconfigured" state to the "Running" state. This will typically imply:

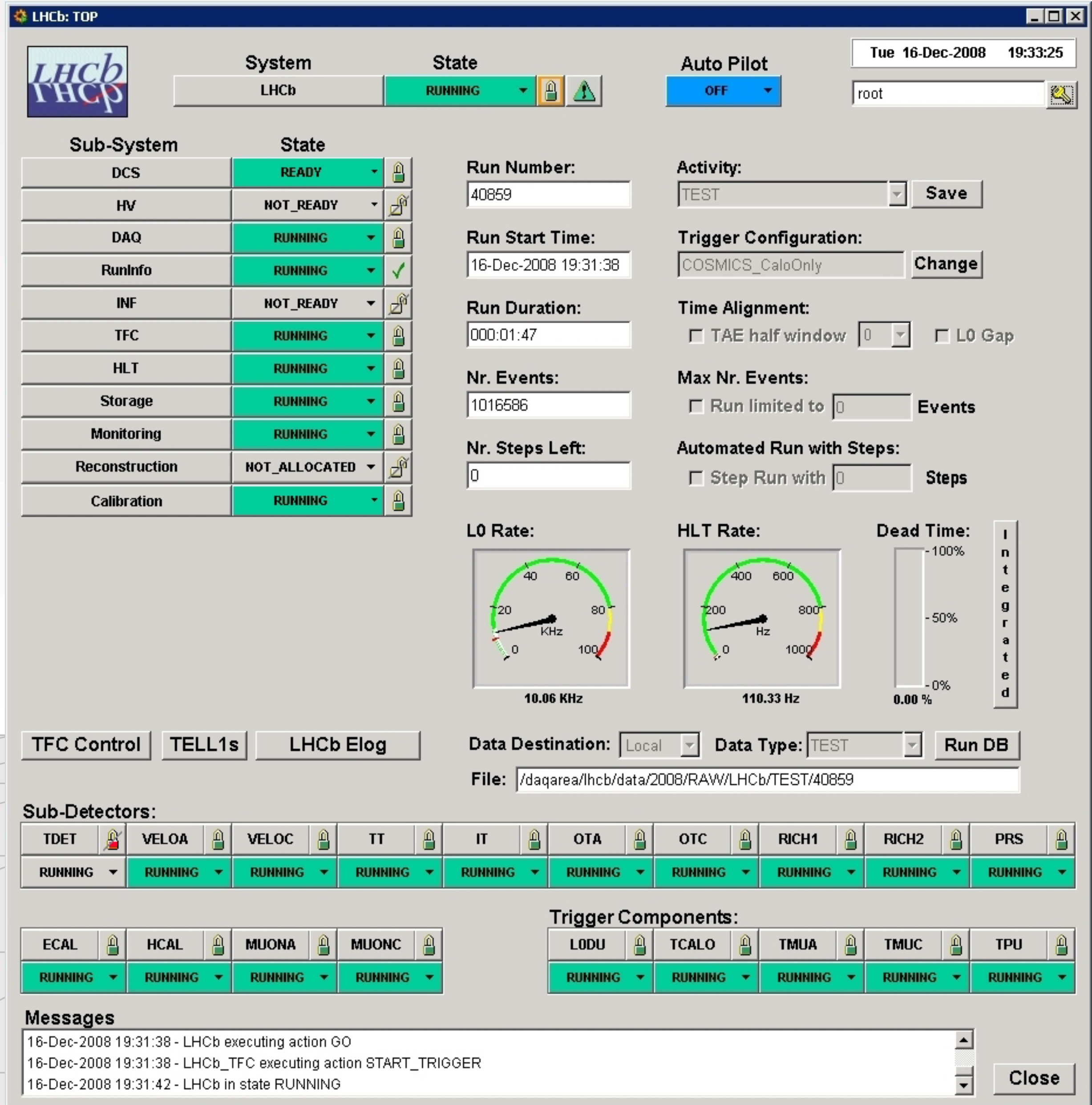
- Configuring the detector front-ends
- Loading and/or configuring the trigger processes in the HLT farms
- Configuring the L1 trigger

	Warm start	Limited by
ALICE	~ 5 min	detector FE config
ATLAS	~ 7 min(*)	detector FE config
CMS	~ 1 1/2 min (central DAQ) + 2 min	One subdetector
LHCb	~ 4 min	One subdetector

All experiments are working hard to reduce this time.
These times hold for the "good case": i.e. all goes well (**Y.M.M.V.**)

(*)measured 10/09/08

Run Control GUI



Main panel of the LHCb run-control (PVSS II)

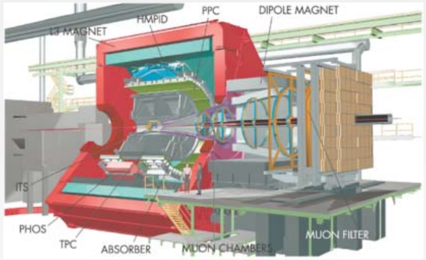
Databases

- The Online systems use a lot of data-bases:
  - Run database, Configuration DB, Conditions DB, DB for logs, for logbooks, histogram DB, inventory DB, …
  - Not to forget: the archiving of data collected from PVSS (used by all detector control systems)
- All experiments run Oracle RAC infrastructures, some use in addition MySQL, object data-base for ATLAS Configuration (OKS)
- Administration of Oracle DBs is largely outsourced to our good friends in the CERN IT/DM group
- Exchange of conditions between offline and online uses Oracle streaming (like replication to Tier1s)

Upgrades

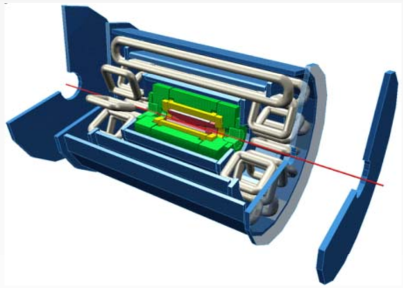
- Farm upgrades transparent within power, space and cooling budgets
- Higher L1 rate: general feeling is that it is too early:
  - first wait for data and see
- All systems are scalable and will work a long way up
- How much do we loose in the high p_t trigger?
  - extreme case LHCb: about 50% → read out entire detector at collision rate (trigger-free DAQ)
- Any upgrade in speed beyond the maximal L1 rate will require new front-end electronics and readout-links
- Upgrade considerations will start from the readout-link and TTC developments (GBT)

Are we ready?

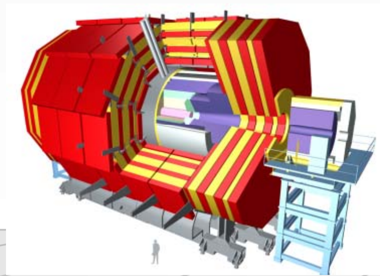


2008: 10000 stable runs, 3 PB of data readout, 350 TB data recorded, 515 days of data taking

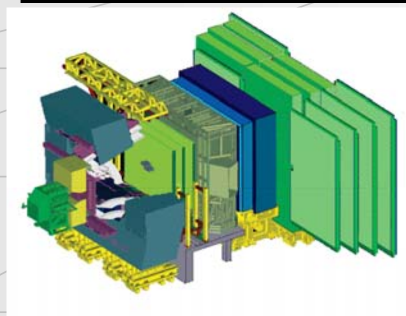
---



Since 09/12:
400 k files of Cosmics, 216 millions of events, 453 TB



no BField ~300 M cosmic events
nominal BField 3.8T ~300 M cosmic events,
~100 TB of raw data



Cosmics since Spring 2008: 1138 runs, 2459 files, 469041 events, 3.16 TB

Status & Summary

We are ready

LHC DAQ / Online talks

in depth coverage of topics in this talk

- [40] Commissioning the ALICE Experiment P. V. Vyvre
- [3] CMS Data Acquisition System Software J. Gutleber
- [150] The ATLAS Online High Level Trigger Framework: Experience reusing Offline Software Components in the ATLAS Trigger W. Wiedenmann
- [38] The ALICE Online Data Storage System R. Divia
- [313] The LHCb Run Control C. Gaspar
- [540] SMI++ Object Oriented Framework used for automation and error recovery in the LHC experiments B. Franek (poster)
- [138] Dynamic configuration of the CMS Data Acquisition cluster H. Sakulin
- [461] The ALICE Online-Offline Framework for the Extraction of Conditions Data C. Zampolli
- [178] The CMS Online Cluster: IT for a Large Data Acquisition and Control Cluster J. A. Coarasa Perez
- [47] Event reconstruction in the LHCb Online cluster A. Puig Navarro
- [94] Commissioning of the ATLAS High Level Trigger with Single Beam and Cosmic Rays A. Di Mattia