LHCb Online event processing and filtering

Niko Neufeld, CERN/PH

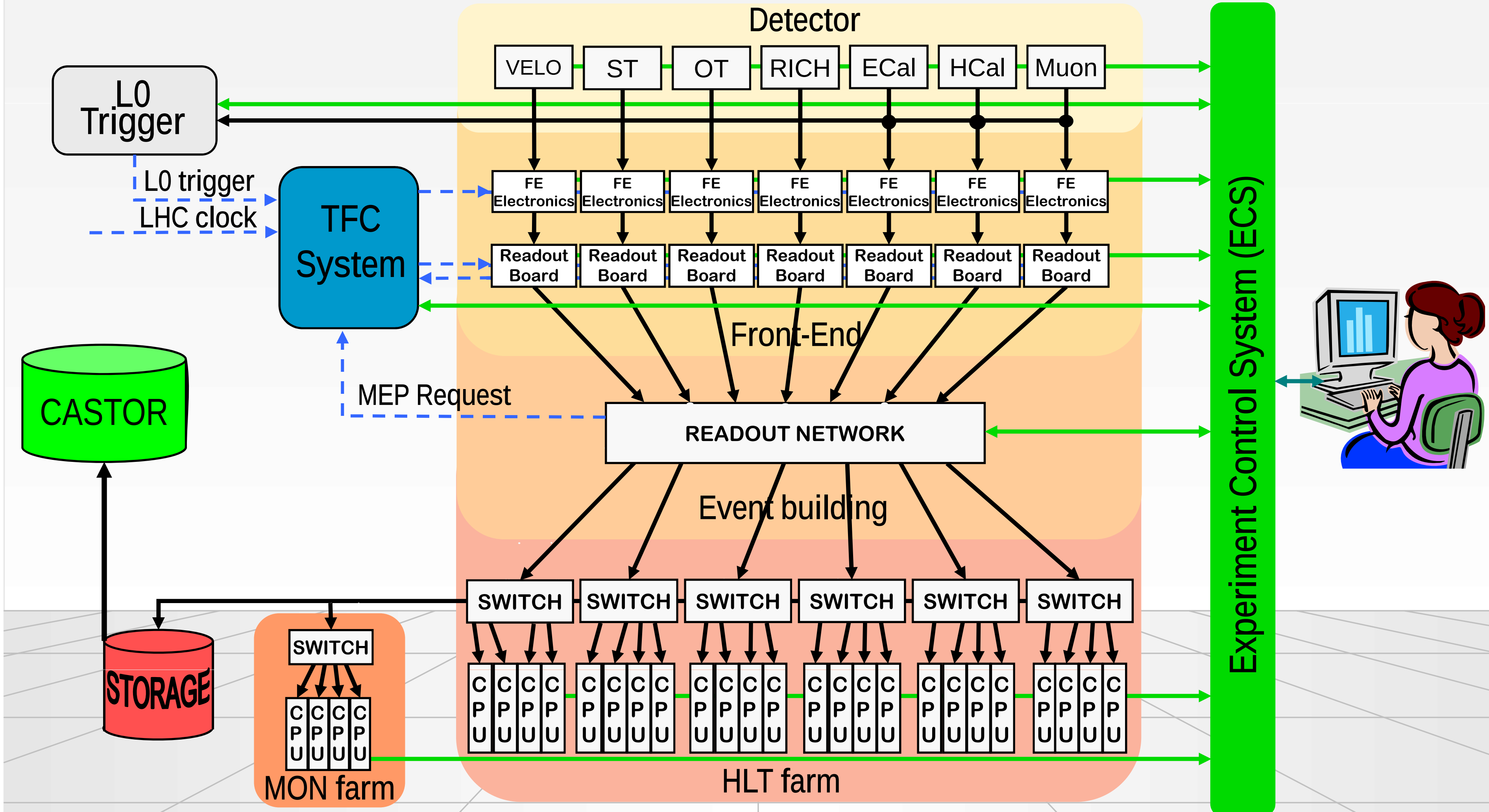
on behalf of the LHCb Online Team

CHEP'07 VICTORIA, BC

International Conference on Computing in High Energy and Nuclear Physics

2-7 Sept 2007 Victoria BC Canada

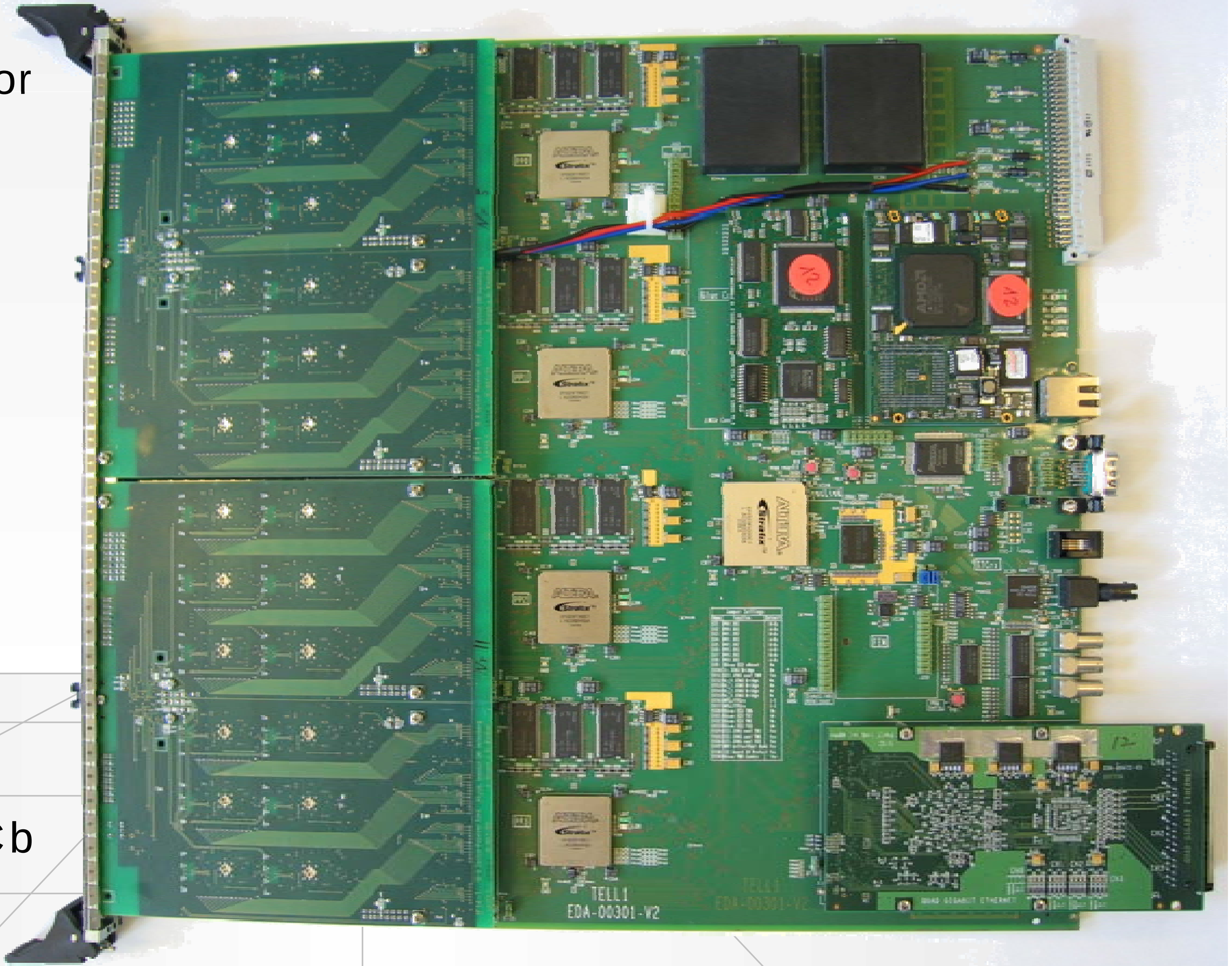
LHCb Online system



Event data
Timing and Fast Control Signals
Control and Monitoring data

Readout Board

- 24 inputs (optical or analogue) @1.6 Gbit/s
- Parallelised FPGA processing
- Power-only backplane
- 4 x 1 Gigabit Ethernet output
- Control via embedded i486 Microcontroller on separate LAN for control and monitoring (counters, histograms)
- ~ 320 for all of LHCb

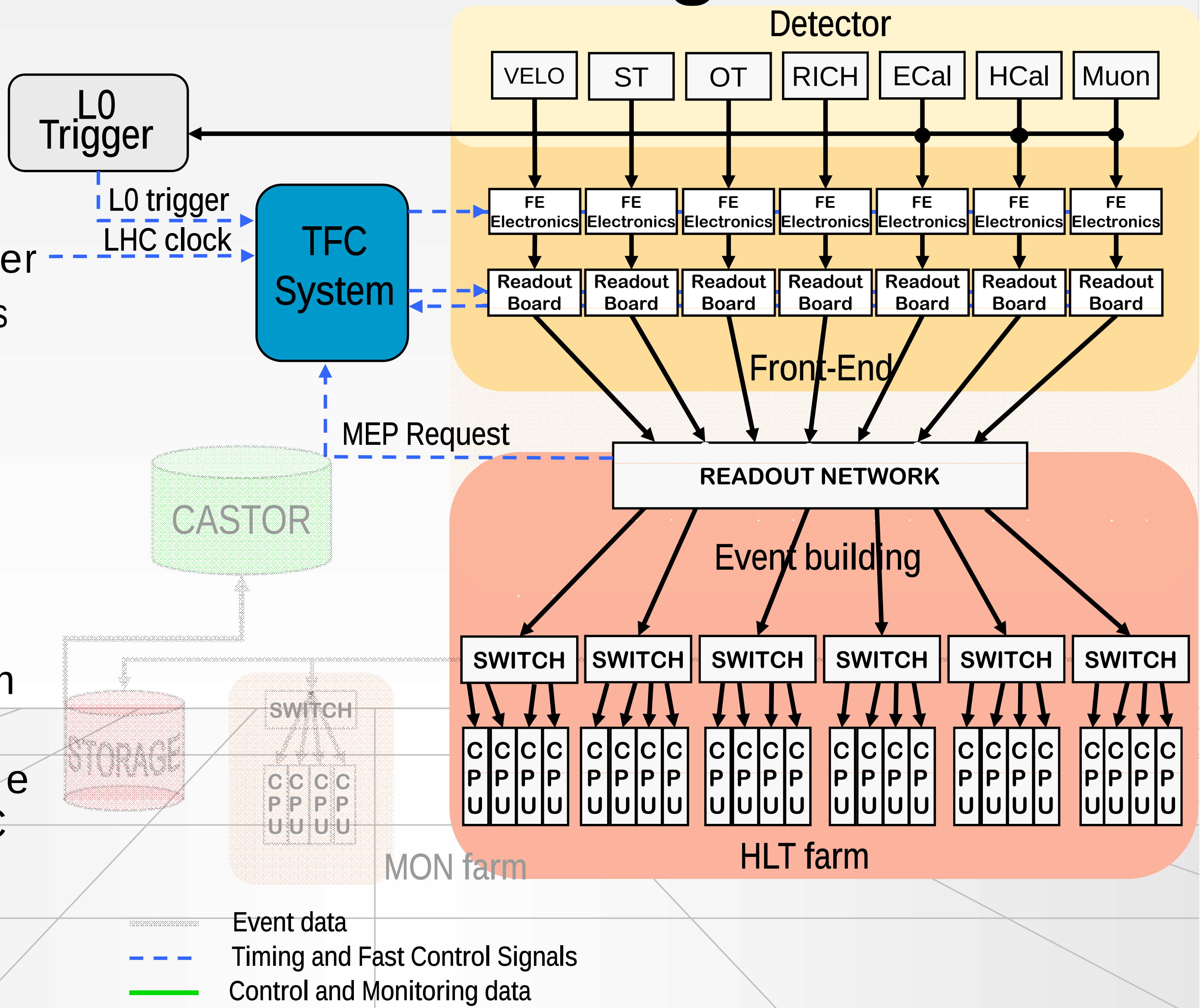


Protocols & Architecture

- Design principle: simplicity – limit the number of protocols and technologies
- FPGA based Readout Boards are directly attached to Gigabit Ethernet DAQ network
  - simple datagram protocol: MEP over IPv4
    - well suited for implementation in FPGAs
      - no re-transmission
      - no buffering
    - no back-traffic
      - can use over-commited router-ports (to some extent)
- MEP: collects several triggers into one packet → reduces overhead (average event fragment ~ 120 Bytes only / overhead on IP/Ethernet is 58 Bytes
- TCP/IP for all data out of the event filter farm
- DIM for information exchange in controls and monitoring (histograms, counters)
- GRIDftp for data transfer to Tier0 (used in LHCb via DIRAC)

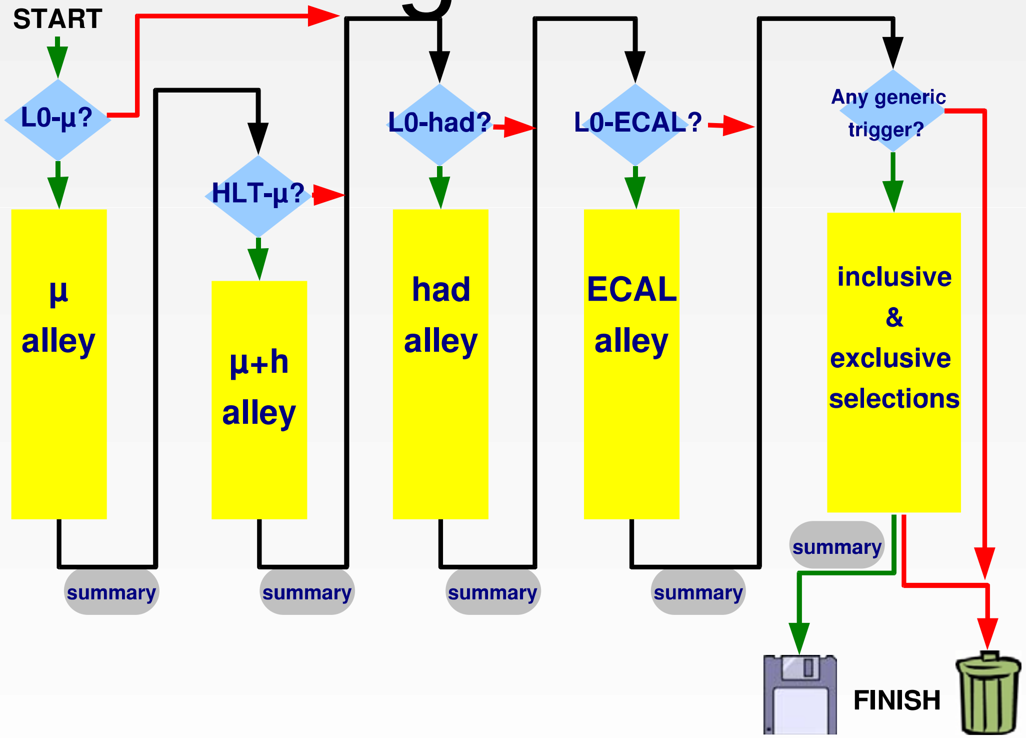
Event building

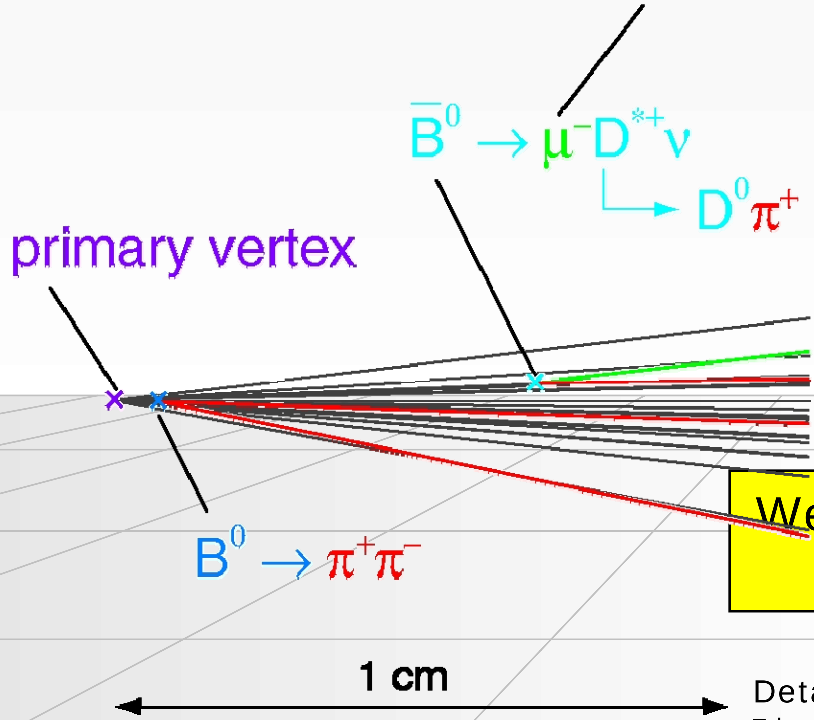
- TFC system broadcasts IP address of destination server
- Readout Boards send ("push") data to destination
- Destination collects all fragments and hands over to trigger algorithm
- Destination announces to be available to TFC ("MEP request")



Event Filtering

Want to run ***this*** @ 1 MHz



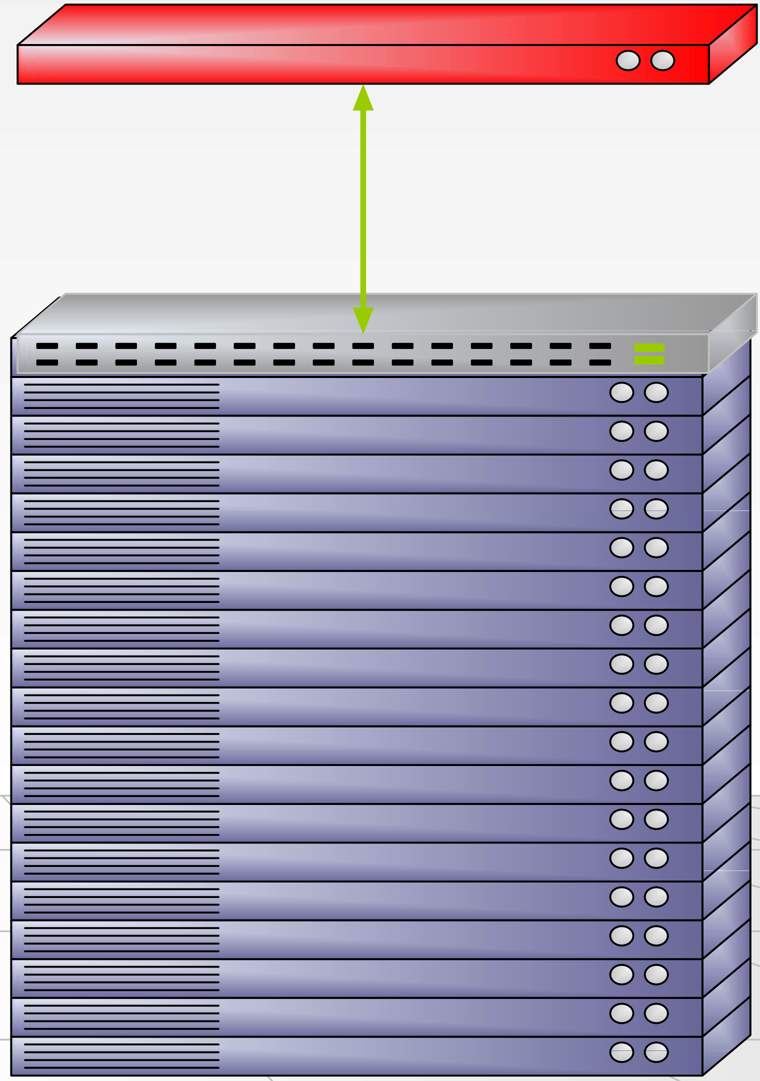


We need ***O(1000)*** servers

Details in OC139 “The LHCb High Level Trigger Software Framework”

Event Filter Farm architecture

- The farm is physically and logically split in to 50 sub-farms
- Each sub-farm is booted, controlled and monitored by a Controls PC
- The Controls PC coalesces counters, histograms, logs, commands
- All farm nodes are identical and run the same application. They are diskless and run via NFS from the Controls PC



DAQ in numbers

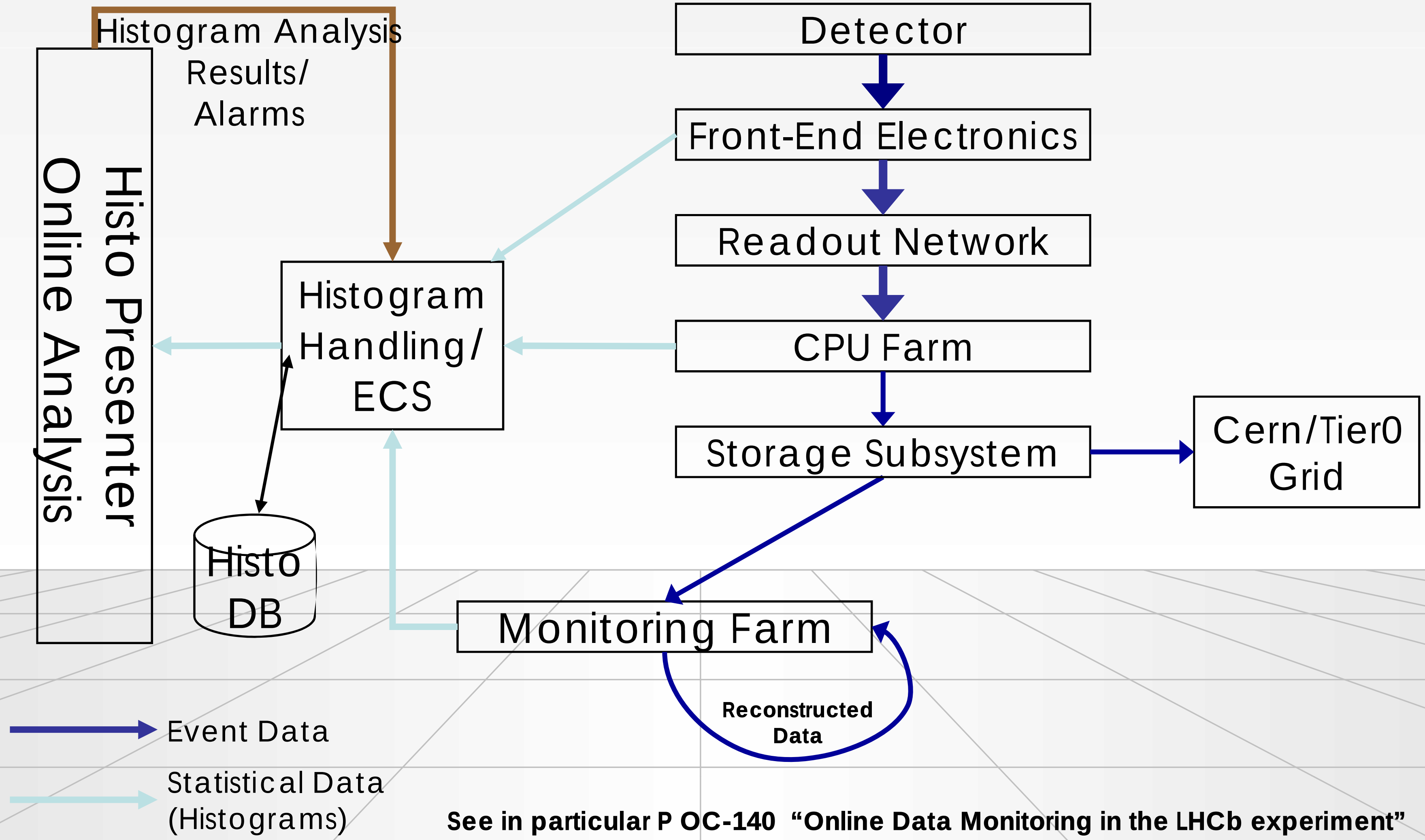
2007/2008

- Can sustain 200 kHz
- 600 router ports installed
- 8 sub-farms of 4 nodes each installed
  - 200 more nodes for early next year
- 50 TB installed, fully commissioned

Final System

- 1 MHz of events – 35 GB/s capacity
- ~ 3000 Gigabit Ethernet router ports
- 50 sub-farms of up to 44 nodes
- 50 TB of local storage up to 800 MB/s total I/O

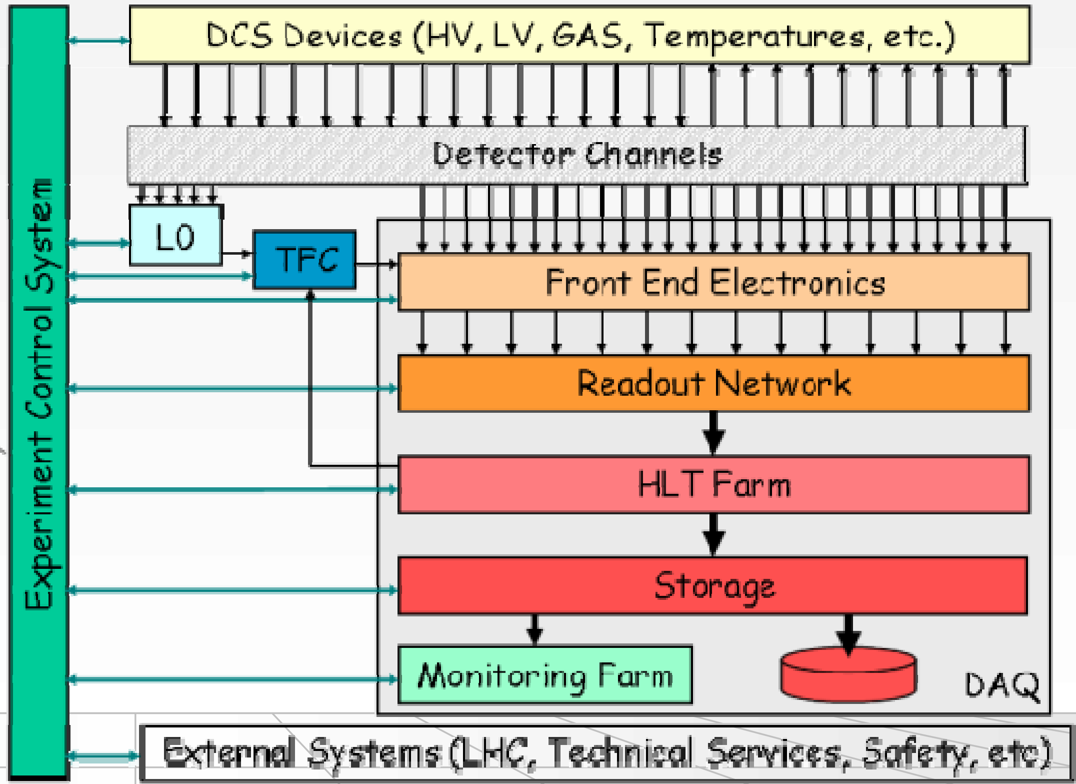
Monitoring Infrastructure



See in particular P OC-140 "Online Data Monitoring in the LHCb experiment"

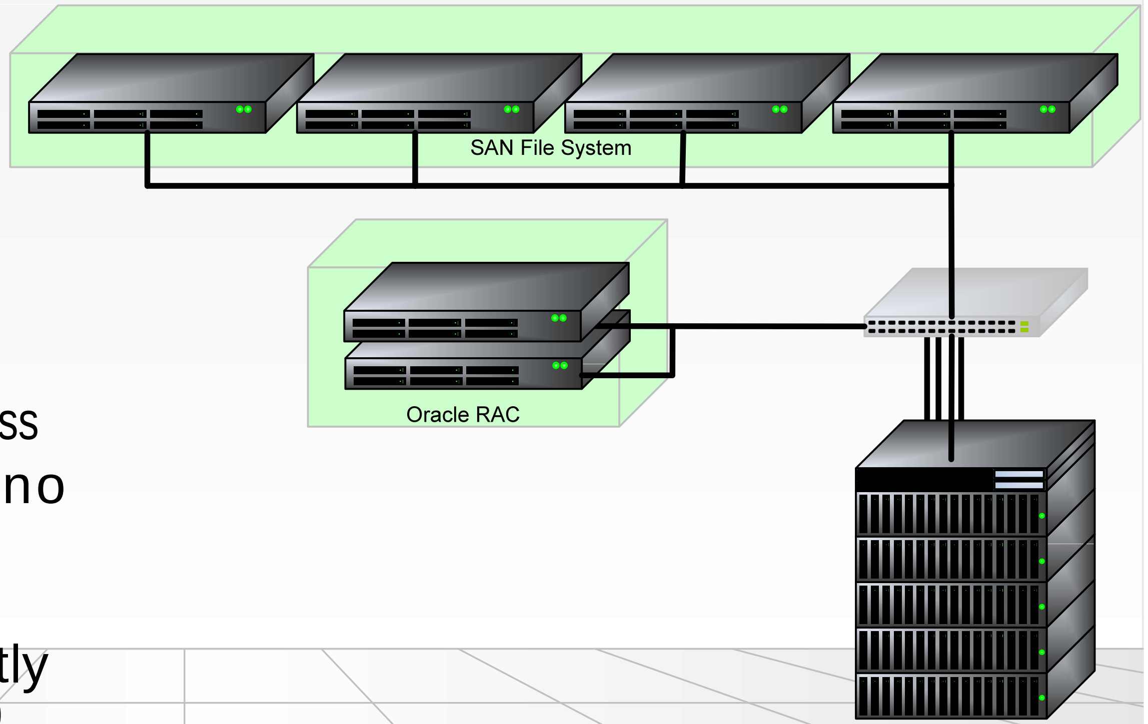
Control System

- Reuse: LHCb uses the same software infrastructure for the detector control (slow controls) and run control
- Independent paths for control and data everywhere
- Based on commercial SCADA software (PVSS) and a common software framework developed by all 4 LHC collaborations (JCOP)



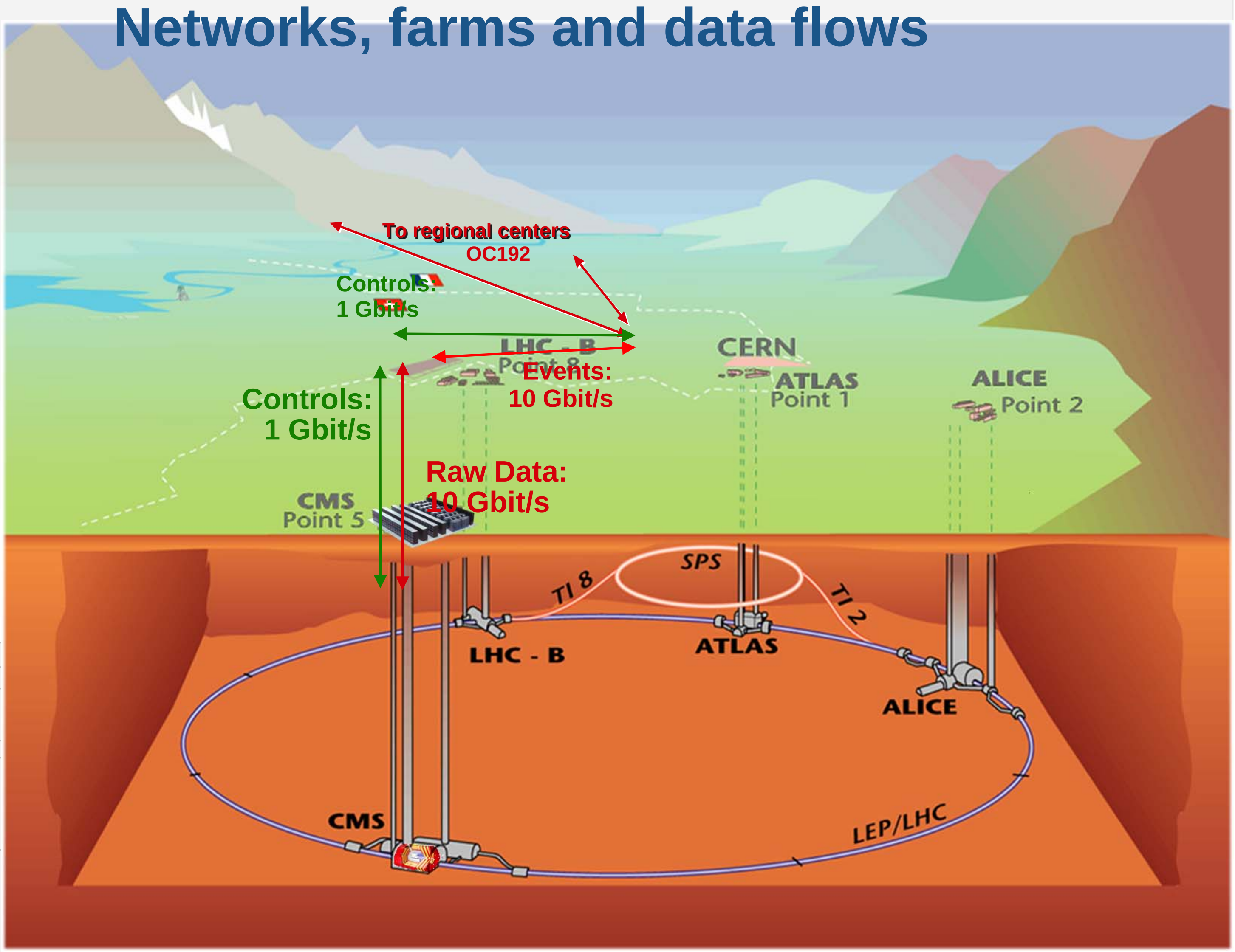
Storage

- Central SAN provides storage for
  - data bases
  - software
  - home directories
  - raw data
- Single sign-on, uniform file-system everywhere
- Most systems are diskless
- SAN is fully redundant (no single point of failure)
- Local buffer of ~ 40 TB
- Data are copied directly from SAN to CERN Tier0 via dedicated fibre (next slide)



From Point 8 to the Grid

- Data files are transferred to Tier0 (CASTOR) using LHCb's DIRAC
- DIRAC takes care of registration in LCG
- During shutdown periods the Event Filter Farm will be operated as a Tier1 centre (using DIRAC)



Conclusion

- Event processing in LHCb is done mostly on commercial commodity hardware (IA32, Gigabit Ethernet)
- Monitoring is implemented at all stages using a dedicated infrastructure, reusing as much as possible existing frameworks
- The architecture is scalable – information and control is coalesced at every layer
- Standard tools / technologies from other fields are used wherever possible: DIRAC/Grid for data transfer, PVSS/SCADA for run-control
- Base system is deployed and running: remaining hardware will be bought and installed in sync with the LHC schedule