

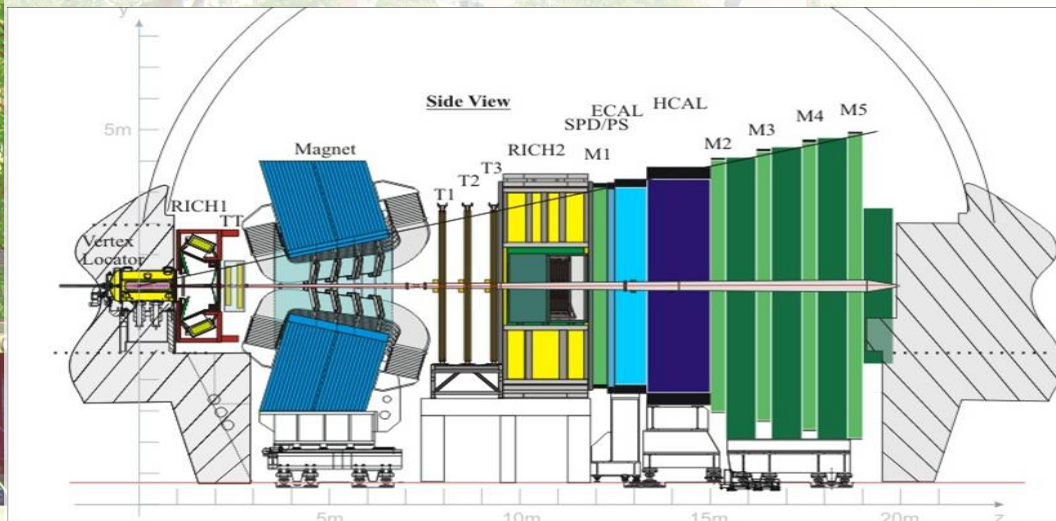
Deferred High Level Trigger in LHCb: A Boost to CPU Resource Utilization

The use of periods without beam for online high level triggers

- Introduction, problem statement
- Realization of the chosen solution
- Conclusions

M.Frank, C.Gaspar, E.v.Herwijnen, B.Jost, N.Neufeld
CERN / LHCb

LHCb Online Computing in Numbers



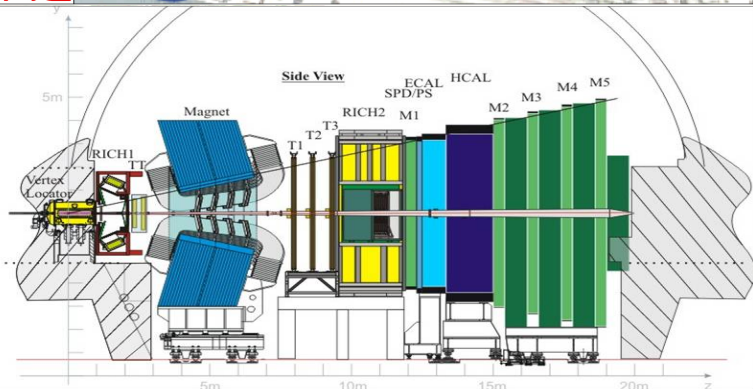
Readout Network

LHCb Online
 Computing Infrastructure

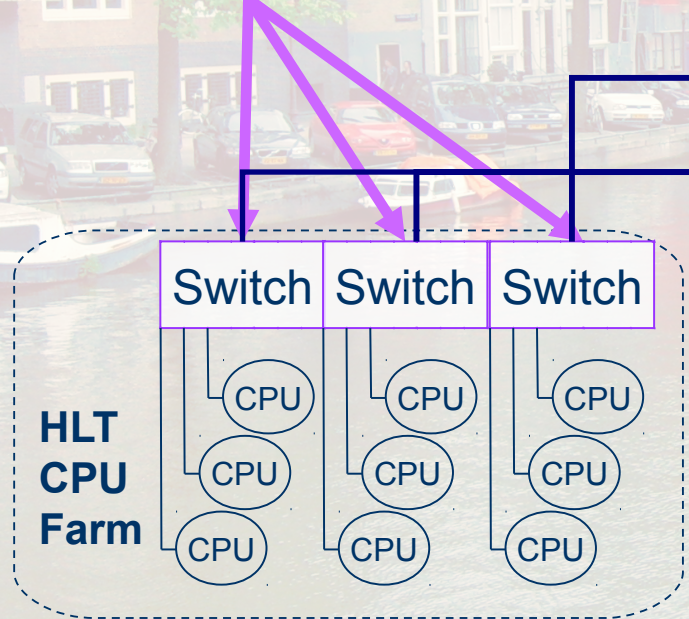
Substantial resources

- Spectrometer for b quark analysis at LHC
- 40 MHz collision rate
- L0 trigger (hardware)
 Accept rate: ~ 1 MHz
 Readout NW: ~ 60 GB/s
- HLT (software)
 Accept rate: ~ 2-8 kHz
- Event size: ~ 50 KB
- Data sources: ~ 350
- Event packing: ~ 13
- 56 Racks
 ~ 1700 Data handling nodes
 ~ 200 Controls nodes
- HLT (expected for 2015):
 ~ 1600 Nodes
 ~ 25000 CPU cores
 ~ 45000 Trigger processes
 ~ 5000 Infrastructure tasks

LHCb Online Computing Hardware



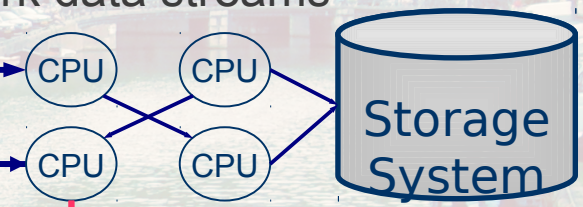
Readout Network



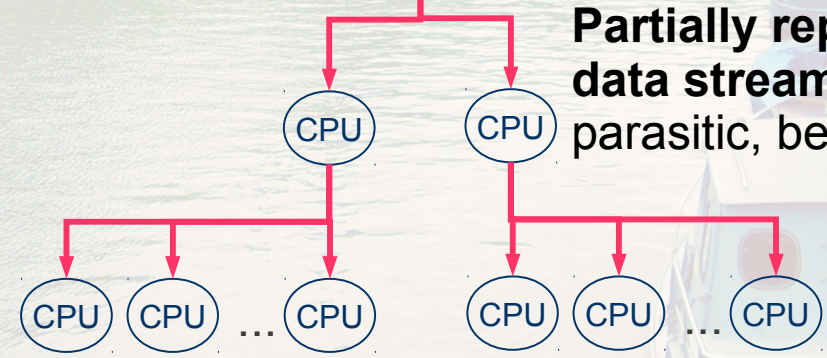
High Level Trigger
 Identify the Good the Bad and the Ugly

Storage Cluster

- File and Stream handling
- Fork data streams



Partially replicated data streams:
 parasitic, best effort



Monitoring Cluster
 Low level monitoring using raw data

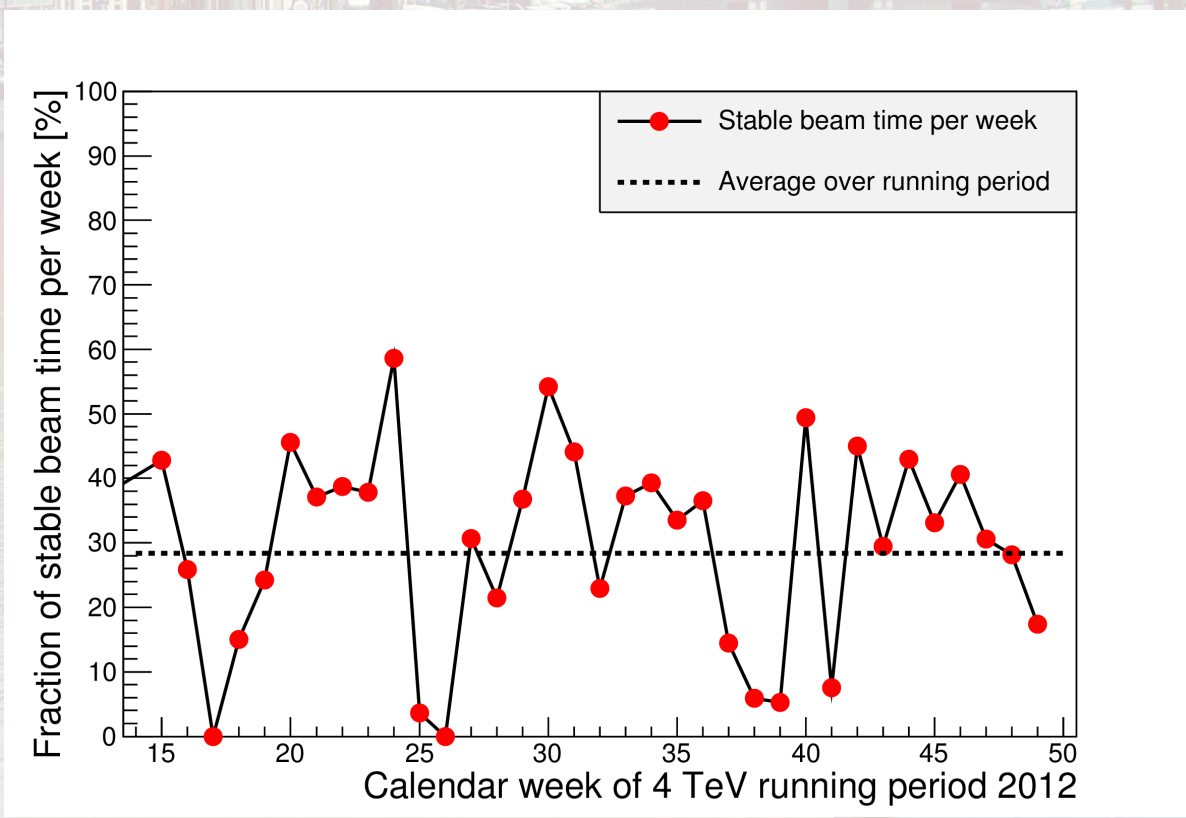
Reconstruction Cluster
 High level monitoring with fully reconstructed events



The Boost: Possible Gain of CPU Time

- LHC delivers roughly during 30% of the running period stable beams to LHCb
- 70% of the time the CPU resources are idle
- Take advantage of the idle-time

- Sophisticated event filtering
- Better selection of 'interesting' events
- Improved physics results

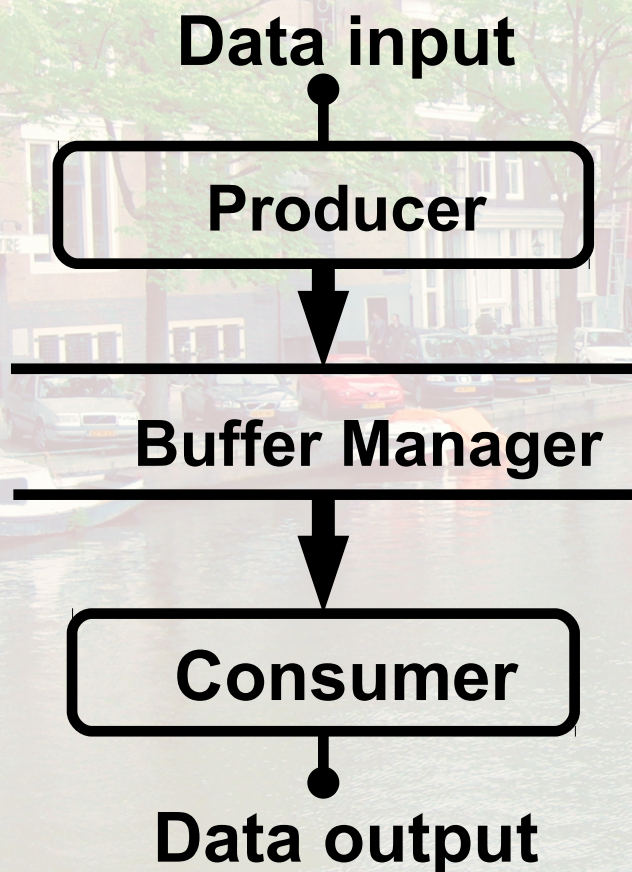


The Roadmap: Benefit from Idle Time

- Try to defer computing needs to time without beam
 - Save events on the local disk of the worker nodes
 - ~ 8-9 hours beam time (~1 day) buffering for 1 TB disks
- Need to split high level trigger program 'Moore'
 - Only save preselected events
 - Rejection factor 6: ~1-2 week of buffering
 - Enough to be busy during MD periods
 - First stage component responsible for pre-selection
 - Second stage component for the final event filtering
- Here I present the supporting infrastructure
 - Not the physics details of this split

Next: Introduction of basic concepts used in the realization

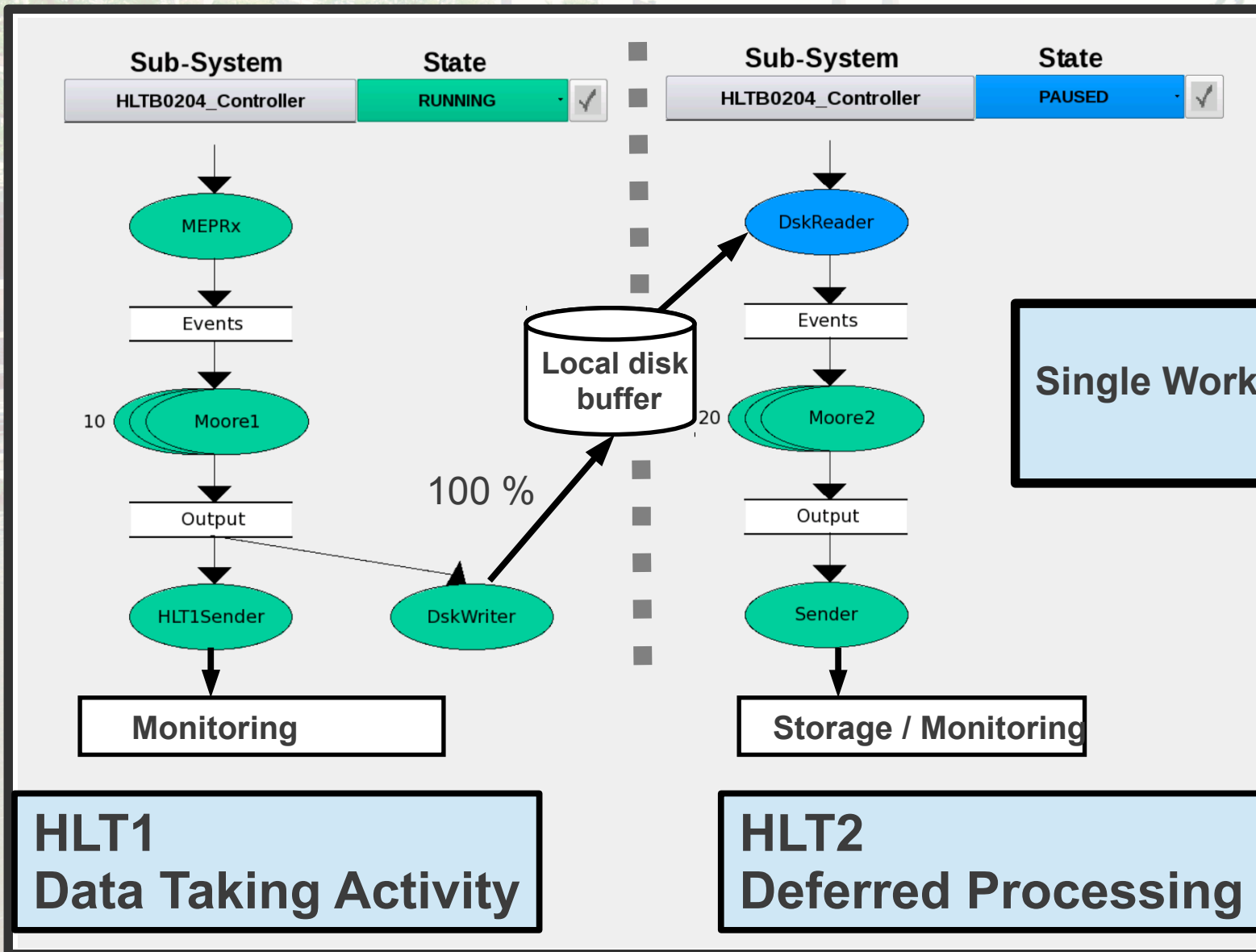
The Basic Pattern: Buffer Manager



- **Managed shared memory**
- **Producers declare events**
- **Consumers subscribe to events**
 - **Receive interrupts when data is present**
- **Pattern used at all stages**
 - **Whenever event data have to be moved**
 - **HLT farm, storage-, monitoring- and reconstruction cluster**

See M.Frank et al., “Data Stream handling in the LHCb experiment”, CHEP 2007, Proceedings, Victoria, BC

The Process Architecture: Worker Node



Worker Nodes: Remarks (1)

- HLT1 and HLT2 activities are entirely asynchronous
 - Loose coupling through local disk cache
 - HLT1 must execute real-time
 - HLT2 executes with lower priority
- HLT2 requires 'offline-quality' calibration
 - Calibration in real-time using fraction of HLT1 accepted events
 - Data monitoring facilities in dedicated farms crucial

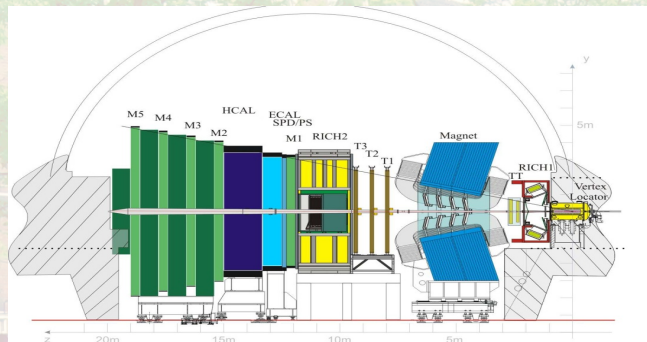
Worker Nodes: Remarks (2)

- We heavily rely on minimizing resource usage
 - Moore processes execution simultaneously on each worker node
- Worker node resources are 'over-committed'
More processes than CPU cores / hyperthreads
 - Memory scarce (2 GB/core) if not addressed
 - CPU and network accesses during configuration
- Resource sharing is mandatory
 - Large benefit from copy-on-write (~70% of memory)
Trigger processes forked after configuration phase
 - Quick application startup using process checkpointing

Worker Nodes: Control

- **All processes of one activity on a worker node**
 - **Need to be started and configured in a well defined order following the states of a finite state machine**
 - **Are controlled by a dedicated process, which reports to the experiment controls system**
- **Consequences for the control of the activities**
 - **Two independent control trees (next slides)**
 - **HLT1 + Experiment**
 - **HLT2 activity**

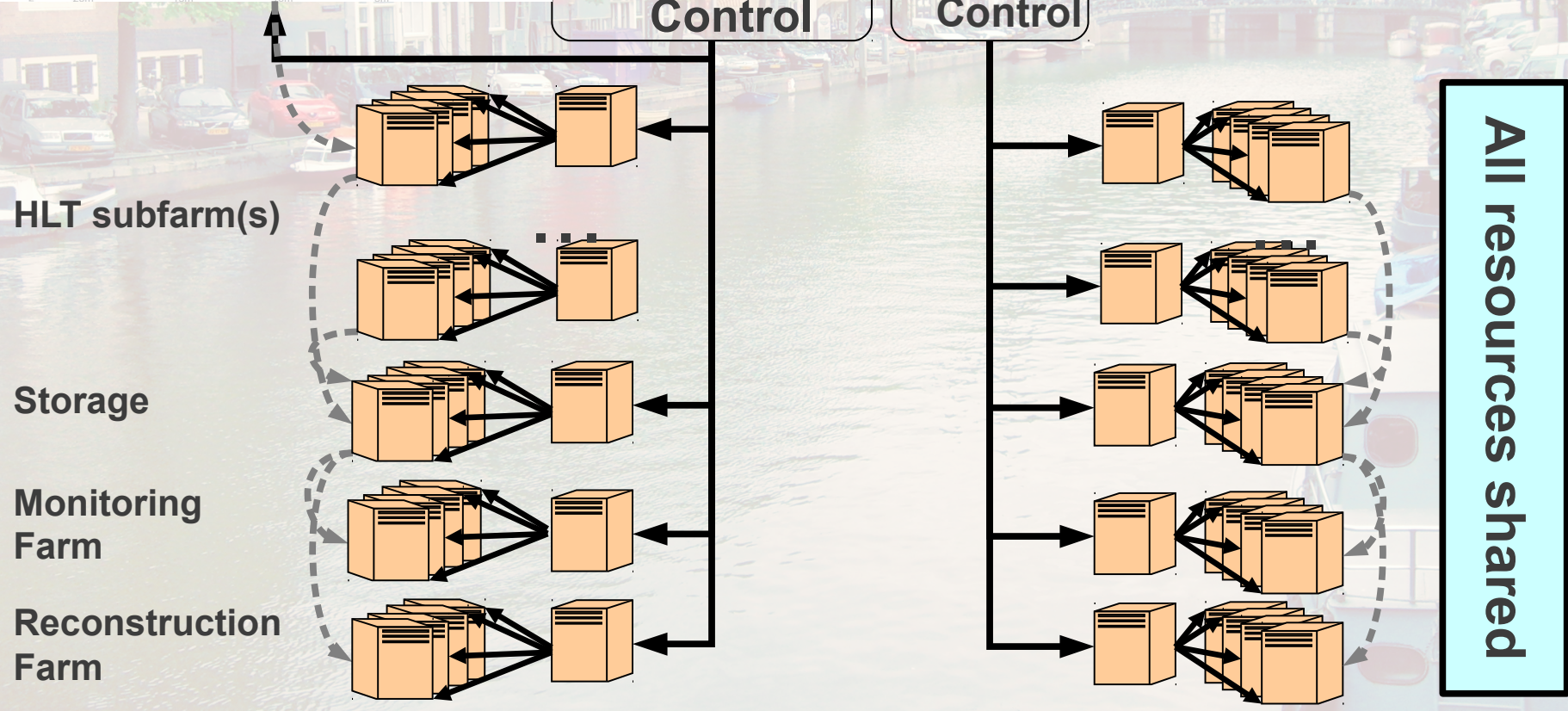
Controls: Two Separated Control Trees



← Control flow
 ← Data flow

Experiment Control

HLT2 Control



Controls Issues

- **Experiment controls system implemented in WinCC**
 - **Commercial SCADA (originally called PVSS)**
 - **Used throughout the experiment**
 - **Hardware configuration (slow control)**
 - **DAQ, Run-Control, Farm operations**
 - **Partitioning concept realized throughout**
 - **Traditionally: Parallel DAQ of independent sub-detectors while no beam**
 - **De facto: Deferred trigger processing = Independent DAQ with data from disk**
- => Presence of partitioning concept eased the implementation of deferred HLT processing**



Controls: Parallel Trees in Reality

LHCb: TOP

System: LHCb State: NOT_READY Auto Pilot: OFF

Sub-System State

DCS	NOT_READY
DAI	NOT_READY
DAQ	NOT_READY
RunInfo	NOT_READY
TFC	NOT_READY
HLT	NOT_READY
Storage	NOT_READY
Monitoring	NOT_ALLOCATED
Reconstruction	NOT_ALLOCATED
Calibration	NOT_READY
HV	NOT_READY

Run Info

Run Number: 140303 Activity: EOF_CALIB

Run Start Time: 23-Sep-2013 18:04:38

Run Duration: 000:15:48

Nr. Events: 0

Step Nr: To Go: 0 1

Deferred HLT Info

LHCb_Deferred NOT_READY

Runs/Files: 15 / 450

Disk Usage 17%

Efficiency Trigger Rates

TFC Control TELL1s LHCb Elog

Sub-Detectors:

TDET	VELOA	VELOC	TT	IT	OTA	OTC	RICH1	RICH2	PRS
NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY

Trigger Components:

L0DU	TCALO	TMUA	TMUC	TPU
NOT_READY	NOT_READY	READY	READY	DEAD

Messages:

23-Sep-2013 18:21:41 - LHCb in state NOT_READY

23-Sep-2013 18:21:41 - LHCb in state ERROR

23-Sep-2013 18:21:41 - LHCb in state NOT_READY

23-Sep-2013 18:21:47 - LHCb executing action

23-Sep-2013 18:21:58 - LHCb in state NOT_READY

**HLT1
Data Taking Activity**

LHCb1: TOP

System: LHCb_Deferred State: NOT_READY Auto Pilot: OFF

Sub-System State

RunInfo	NOT_READY
HLT	NOT_READY
Storage	NOT_READY
Monitoring	NOT_READY

Run Info

Run Number: 140302 Activity: DeferredHLT

Run Start Time: 23-Sep-2013 17:44:02

Run Duration: 000:37:14

Nr. Events: 0

Step Nr: To Go: 0 0

Deferred HLT Info

LHCb_Deferred NOT_READY

Runs/Files: 15 / 450

Disk Usage 17%

Efficiency Trigger Rates

TFC Control TELL1s LHCb1 Elog

Sub-Detectors:

TDET	VELOA	VELOC	TT	IT	OTA	OTC	RICH1	RICH2	PRS
NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY

Trigger Components:

L0DU	TCALO	TMUA	TMUC	TPU
NOT_READY	NOT_READY	READY	READY	DEAD

Messages:

24-Sep-2013 11:12:21 - LHCb1 in state NOT_READY

24-Sep-2013 11:12:28 - LHCb1_Storage executing action INITIALIZE

24-Sep-2013 11:12:38 - LHCb1_Storage executing action RESET

24-Sep-2013 11:12:38 - LHCb1_Storage executing action RESET

24-Sep-2013 11:12:38 - LHCb1_Storage executing action RESET

24-Sep-2013 11:12:38 - LHCb1_Storage executing action RESET

**HLT2
Deferred Processing**



Controls: Parallel Trees in Reality

LHCb: TOP

System: LHCb State: NOT_READY Auto Pilot: OFF

Run Info: Run Number: 140303 Activity: EOF_CALIB

Run Start Time: 23-Sep-2013 18:04:38

Run Duration: 000:15:48

Nr. Events: 0

Step Nr: To Go: 0 1

L0 Rate: 0.00 Hz

HLT Rate: 0.00 Hz

Dead Time: Overflow: 0.00%

Data Destination: Local Data Type: EOF_CALIB13

File: /daqarea/hlcb/data/2013/RAW/FULL/LHCb/EOF_CALIB13/140303

Sub-Detectors:

TDET	VELOA	VELOC	TT	IT	OTA	OTC	RICH1	RICH2	PRS
NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY

Trigger Components:

L0DU	TCALO	TMUA	TMUC	TPU
NOT_READY	NOT_READY	READY	READY	DEAD

Messages:

23-Sep-2013 18:21:41 - LHCb in state NOT_READY
 23-Sep-2013 18:21:41 - LHCb in state ERROR
 23-Sep-2013 18:21:41 - LHCb in state NOT_READY
 23-Sep-2013 18:21:47 - LHCb executing action
 23-Sep-2013 18:21:58 - LHCb in state NOT_READY

**HLT1
Data Taking Activity**

LHCb1: TOP

System: LHCb_Deferred State: NOT_READY Auto Pilot: OFF

Run Info: Run Number: 140302 Activity: DeferredHLT

Run Start Time: 24-Sep-2013 11:12:28

Run Duration: 000:15:48

Nr. Events: 0

Step Nr: To Go: 0 1

L0 Rate: 0.00 Hz

HLT Rate: 0.00 Hz

Dead Time: Overflow: 0.00%

Data Destination: Local Data Type: TEST

File: /daqarea/hlcb/data/2013/RAW/FULL/LHCb1/TEST/140302

Sub-Detectors:

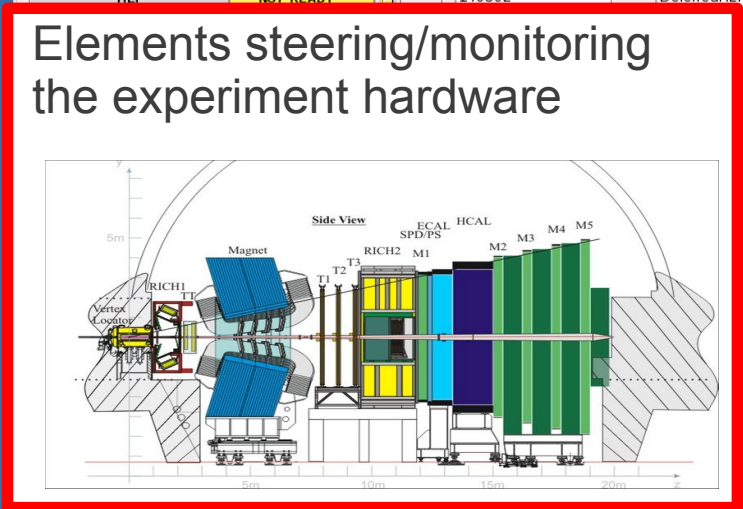
TDET	VELOA	VELOC	TT	IT	OTA	OTC	RICH1	RICH2	PRS
NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY	NOT_READY

Trigger Components:

L0DU	TCALO	TMUA	TMUC	TPU
NOT_READY	NOT_READY	READY	READY	DEAD

Messages:

24-Sep-2013 11:12:21 - LHCb1 in state NOT_READY
 24-Sep-2013 11:12:28 - LHCb1_Storage executing action INITIALIZE
 24-Sep-2013 11:12:38 - LHCb1_Storage executing action RESET
 24-Sep-2013 11:12:47 - LHCb1 in state NOT_READY
 24-Sep-2013 11:12:58 - LHCb1 in state NOT_READY



**HLT2
Deferred Processing**

Conclusions

- We managed a redesign of the high level trigger infrastructure to
 - Benefit from time periods without beam
 - Results in a possible increase of 200% CPU time
 - Gained CPU time to be used to improve event selection in the high level trigger
- The realization was based on two basic concepts
 - Consistent deployment of the Buffer Manager pattern throughout the dataflow
 - The partitioning concept supporting shared computing resources

