ALICE
Trigger and DAQ

LHC Days in Split
October-2004

P.Vande Vyvre - CERN/PH

for the ALICE DAQ project
(Birmingham, Budapest, CERN, Mexico, Split, Zagreb)
Outline

• ALICE DAQ
  – Trigger/DAQ logical model - Requirements
  – Trigger/DAQ at LHC
  – ALICE Trigger/DAQ Architecture

• Hardware components

• Software packages

• Towards the final ALICE

• Conclusions
Outline

• ALICE DAQ
  – Trigger/DAQ logical model - Requirements
  – Trigger/DAQ at LHC
  – ALICE Trigger/DAQ Architecture

• Hardware components
• Software packages
• Towards the final ALICE
• Conclusions
Trigger & DAQ logical model

- Low-Level Trigger System
- Detector Electronics
- High-Level Trigger System
- DAQ System
ALICE Physics requirements

Pb-Pb beam
- Central 20 Hz 86.0 MB
- MB 20 Hz 20.0 MB
- Dimuon 1600 Hz 0.5 MB
- Dielectron 200 Hz 9.0 MB

pp beam
MB 100 Hz 2.5 MB

25 GB/s

2.50 GB/s

1.25 GB/s
### DAQ @ LHC

<table>
<thead>
<tr>
<th>Collaboration</th>
<th>Beam</th>
<th>Event Size (Byte)</th>
<th>Readout (HLT input) (Events/s.)</th>
<th>Readout (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALICE</td>
<td>Pb-Pb</td>
<td>$4 \times 10^7$</td>
<td>$2 \times 10^3$</td>
<td>25</td>
</tr>
<tr>
<td>ATLAS</td>
<td>pp</td>
<td>$10^6$</td>
<td>$2 \times 10^3$</td>
<td>10</td>
</tr>
<tr>
<td>CMS</td>
<td>pp</td>
<td>$10^6$</td>
<td>$10^5$</td>
<td>100</td>
</tr>
<tr>
<td>LHCb</td>
<td>pp</td>
<td>$2 \times 10^5$</td>
<td>$40 \times 10^4$</td>
<td>4</td>
</tr>
<tr>
<td>Experiment</td>
<td>Interaction</td>
<td>Events/s.</td>
<td>MB/s</td>
<td>PBytes</td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-----------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>ALICE</td>
<td>Pb-Pb</td>
<td>200</td>
<td>1250</td>
<td>2.3</td>
</tr>
<tr>
<td>ATLAS</td>
<td>pp</td>
<td>100</td>
<td>100</td>
<td>6.0</td>
</tr>
<tr>
<td>CMS</td>
<td>pp</td>
<td>100</td>
<td>100</td>
<td>3.0</td>
</tr>
<tr>
<td>LHCb</td>
<td>pp</td>
<td>200</td>
<td>40</td>
<td>1.0</td>
</tr>
</tbody>
</table>
ALICE TRG/DAQ architecture

Event Building Network

Storage Network

File

Event

Event Fragment

Load Bal.

Sub-event

EDM

Rare/All

BUSY

CTP

L0, L1a, L2

LTU

L0, L1a, L2

TTC

FERO

FERO

LDC

123 DDLs

175 Detector LDC

262 DDLs

329 D-RORC

50 GDC

25 TDS

10 D-RORC

10 HLT LDC

10 DDLs

10 D-RORC

10 HLT LDC

DDL

H-RORC

HLT Farm

FEP

FEP

FEP

EDM

GDC

GDC

GDC

GDC

50 GDC

25 TDS

DSS

DSS

5 DSS

PDS

TDS

TDS
Outline

• ALICE DAQ
  – Trigger/DAQ logical model - Requirements
  – Trigger/DAQ at LHC
  – ALICE Trigger/DAQ Architecture

• Hardware components

• Software packages

• Towards the final ALICE

• Conclusions
Local Trigger Unit

◆ Common interface between Central Trigger Processor to TTC tree of each detector
Trigger Control Software
Detector Data Links

- Standard and stable interface detector/DAQ
- Point-to-point full-duplex digital data link
- Massive parallelism (100’s)
- Integrated with SDD, TPC, TRD, TOF, Muon, HMPID
Link performance

- HEP development based on commodity components:
  - Fiber Channel or Gig. Ethernet: 2.125 Gb/s
  - Optical transceiver 850 nm VCSEL
  - Flash-based FPGA (Radiation tolerant)

200 MB/s sustained
Lots of bw available.
Major fraction available to end application.
Links Adapters

- Adapter for 1 or a few links to PC I/O bus
- A few-to-one multiplexer
- Massive parallelism (100’s)
Link and adapter performance

- Currently PCI and PCI-X busses. PCI-XP (PCI Express) in the future.
- No large local memory. Fast transfer to PC memory
- PCI interface
  - IP core (VHDL code synthesized in FPGA)
  - PCI 64 bit 66 MHz. Master enabled.

200 MB/s sustained

Total PCI load: 92 %
Data transfer PCI load: 83 %

Lots of bw available.
Major fraction available to end application
Subevent & event buffer

- Baseline:
  - Fast dual-port memories
  - Electronics racks are over
    Extensive use of dual-CPU PCs

- Key parameters:
  - Cost/performance
  - Performance: I/O and memory bandwidth
Readout System Performance

- Supermicro server motherboard with dual Xeon CPUs @ 2.4 GHz
  - In the future: multicore CPUs from Intel and AMD
- Six PCI-X slots, 4 bus segments (3+1+1+1)
- Linux OS
- ALICE Data-Acquisition software (DATE)
Performance: 6 D-RORCs

- Testing the fully populated PC using data source internal to PCI interface
  - Interoperability test
  - Measure the maximal input bandwidth

<table>
<thead>
<tr>
<th>Controller #1</th>
<th>PCI #1</th>
<th>PCI #2</th>
<th>PCI #3</th>
<th>PCI #4</th>
<th>PCI #5</th>
<th>PCI #6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment #1</td>
<td>1 Ch</td>
<td>1 Ch</td>
<td>1 Ch</td>
<td>1 Ch</td>
<td>1 Ch</td>
<td>1 Ch</td>
</tr>
<tr>
<td></td>
<td>264</td>
<td>232</td>
<td>141.3</td>
<td>264</td>
<td>264</td>
<td>261.3</td>
</tr>
</tbody>
</table>

Bandwidth [MB/s]
Normalized Bandwidth [MB/s/Ch]
Event Building Network

- **Baseline:**
  - Adopt broadly exploited standards
    - Switched Ethernet and TCP/IP

- **Motivations for switched Ethernet:**
  - Performance of Gigabit Ethernet switches already adequate: 2 Tbit/s of aggregate bandwidth
  - Use of commodity items: network switches and interfaces
  - Easy (re)configuration and reallocation of resources

- **Network Interface Card (NIC):**
  - TCP/IP Offload Engine (TOE)
    - Dedicated processor to execute IP stack
  - 10 Gbit Ethernet in the PCs
2003 ALICE Data Challenge (ADC V)

3COM 4900
16 x Gbit

Enterasys E1 OAS
12 Gbit, 1 x 10 Gbit

4 x GE
10GE

4 x 7 Disk servers
2 x 2.0 GHz Xeon
1 GB RAM
Intel 82544GC

32 1A64 HP-rx2600 Servers
2 x 1 GHz Itanium-2
2 GB RAM
Broadcom NetXtrem BCM5701 (tg3)
RedHat Advanced Workstation 2.1
6.4 GB/s to memory, 4.0 GB/s to I/O

3COM 4900

10 Tape Server
4 x GE

3COM 4900

10GE

Enterasys ER16
16 slots
4/8 x Gbit or 1 x 10 Gbit/slot

3COM 4900

16 x Gbit

~ 80 CPU servers 2 x 2.4 GHz Xeon, 1 GB RAM, Intel 8254EM Gigabit in PCI-X 133 (Intel PRO/1000), CERN Linux 7.3.3

P. Vande Vyvre - CERN/PH

~ 80 CPU servers 2 x 2.4 GHz Xeon, 1 GB RAM, Intel 8254EM Gigabit in PCI-X 133 (Intel PRO/1000), CERN Linux 7.3.3

GDCs

LDCs
ADC V Trunking

Trunk of 4 x Gb Eth

<table>
<thead>
<tr>
<th># LDCs</th>
<th>MB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>6</td>
<td>240</td>
</tr>
<tr>
<td>7</td>
<td>280</td>
</tr>
<tr>
<td>8</td>
<td>320</td>
</tr>
<tr>
<td>9</td>
<td>360</td>
</tr>
</tbody>
</table>
Scalability of network-based event building

- Reliability and scalability of the whole system
- Throughput limited by data sources
Mass Storage System

- Logical model: distributed file-system
- Software implementation
  - CASTOR system developed by CERN/IT
  - Accessibility via the GRID
- Baseline hardware implementation
  - Transient Data Storage
    - Located at the experimental area
    - Capacity: a few hours of autonomous data taking
    - Before archiving to tertiary storage, if any
  - Permanent Data Storage
    - Located in the computing centre
    - Infinite capacity, very low cost
    - Single write and a few read
    - Sufficient performance to achieve performances with reasonable number of parallel streams and media operations
      - 1 GByte/s: 40 active streams at ~30 MB/s
      - 1 media operation every 2'30" with 200 GB/volume
- Open to new implementations thanks to software model
Possible Transient Storage

- TDS: local GDC storage
- Storage attachment: internal IDE
- TDS: storage array
- Storage attachment: Fibre Channel
Transient Data Storage

- Transient data storage at the experimental area
- Baseline
  - Storage arrays of commodity disks
  - Box attachment: Fibre Channel
  - Disk attachment: IDE or serial-ATA
  - RAID-level
- Partnering with industry for test of equipment
- Key selection criteria:
  - Cost/performance
  - Bandwidth/box
  - Robustness
Storage Arrays

- dotHILL SANnet II 200 FC
  - 12 fiber channel disk slots
  - 1 GB cache
  - 1 x 2Gbit fiber host channel

- Infortrend IFT-6330
  - 12 IDE drive slots
  - 128 MB cache
  - 2 x 2Gbit fiber host channels

- Infortrend EonStor A16F-G1A2
  (INFN CASPUR Storage Lab)
  - 16 SATA drive slots
  - 1GB cache
  - 2 x 2Gbit fiber host channel
Storage Arrays Performance

- Aggregate throughput measured for
  - Set of 5 disks configured as RAID 5
  - Filesizes of 100, 300, 1024 and 2048 MB
  - Recl=8 kB, 32 kB, 128 kB, 512 kB, 2 MB, 8 MB, 32 MB and 128 MB
Permanent Data Storage (1)

- Permanent data storage in the computing centre
- Baseline solution
  - Magnetic tape
- Critical areas
  - High Energy Physics peculiar use of tapes
  - Infrastructure hidden by a hierarchical storage management sw
  - Limited market, different application
  - Limited competition
  - No demonstrated alternative yet
- Demonstrated solution for LHC
  - 15 parallel streams
Performance of Permanent Data Storage

Data to tape
Outline

• ALICE DAQ
  – Trigger/DAQ logical model - Requirements
  – Trigger/DAQ at LHC
  – ALICE Trigger/DAQ Architecture

• Hardware components

• Software packages

• Towards the final ALICE

• Conclusions
ALICE DAQ Software

Dataflow (DATE)

Control (DATE)
Monitoring (AFFAIR, MOOD)

ROOT I/O CASTOR

DDL and D-RORC software

CTP
LTU
TTC
FERO
FERO
LTU
TTC
FERO
FERO

BSY
Rare/All
L0, L1a, L2
L0, L1a, L2
BSY

Detector LDC

DDL
H-RORC
FEP
FEP

DDL
D-RORC
HLT LDC
LDC
LDC

L0, L1a, L2

TDS
PDS

EDM

FEP
FEP

DDC

262 DDLs
D-RORC

Event Fragment Sub-event Event File

Load Bal.

HLT Farm

Monitoring (AFFAIR, MOOD)

Control (DATE)

ROOT I/O CASTOR

Dataflow (DATE)

DDL and D-RORC software

CTP
LTU
TTC
FERO
FERO
LTU
TTC
FERO
FERO

BSY
Rare/All
L0, L1a, L2
L0, L1a, L2
BSY

Detector LDC

DDL
H-RORC
FEP
FEP

DDL
D-RORC
HLT LDC
LDC
LDC

L0, L1a, L2

TDS
PDS

EDM

FEP
FEP

DDC

262 DDLs
D-RORC

Event Fragment Sub-event Event File

Load Bal.

HLT Farm

Monitoring (AFFAIR, MOOD)

Control (DATE)

ROOT I/O CASTOR

Dataflow (DATE)

DDL and D-RORC software

CTP
LTU
TTC
FERO
FERO
LTU
TTC
FERO
FERO

BSY
Rare/All
L0, L1a, L2
L0, L1a, L2
BSY

Detector LDC

DDL
H-RORC
FEP
FEP

DDL
D-RORC
HLT LDC
LDC
LDC

L0, L1a, L2

TDS
PDS

EDM

FEP
FEP

DDC

262 DDLs
D-RORC

Event Fragment Sub-event Event File

Load Bal.

HLT Farm

Monitoring (AFFAIR, MOOD)

Control (DATE)

ROOT I/O CASTOR
DAQ Software Framework

- DAQ Software Framework
  - Common interfaces for detector-depandant applications
  - Address all configurations and all development phases

- DAQ Software
  - Complete ALICE DAQ software framework in 3 packages:
    - DATE:
      - Data-flow: detector readout, event building
      - System configuration, control (1000’s of programs to start, stop, synchronize)
    - AFFAIR: Performance monitoring
    - MOOD: Data quality monitoring
  - Production-quality releases
  - Evolving with requirements and technology ⇒ home-development

- Key issues
  - Scalability (1 to 1000, demonstrate it)
  - Support and documentation
DATE software
- Operator console
- State machines
- Control of distributed system

Home-made development based on free software
DATE Scalability in ADC V
DAQ Configuration Database

- DATE Configuration Database
  - Operator console
  - State machines
- Home-made development based on free software (MySQL)
Web access to Configuration Database

Roles

<table>
<thead>
<tr>
<th>Name</th>
<th>Hostname</th>
<th>Description</th>
<th>Role</th>
<th>Id</th>
<th>Toplevel</th>
<th>Madeof</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>gdc1</td>
<td>host8</td>
<td>GDC #1</td>
<td>GDC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>gdc2</td>
<td>host9</td>
<td>GDC #2</td>
<td>GDC</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetOneLdc1</td>
<td>host1</td>
<td>DetOne LDC #1</td>
<td>LDC</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetOneLdc2</td>
<td>host2</td>
<td>DetOne LDC #2</td>
<td>LDC</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetOneLdc3</td>
<td>host3</td>
<td>DetOne LDC #3</td>
<td>LDC</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetTwoLdc1</td>
<td>pcald37</td>
<td>DetTwo LDC #1</td>
<td>LDC</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetTwoLdc2</td>
<td>host5</td>
<td>DetTwo LDC #2</td>
<td>LDC</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>AloneLdc</td>
<td>host7</td>
<td>Single LDC</td>
<td>LDC</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetThreeLdc</td>
<td>host6</td>
<td>DetThree LDC</td>
<td>LDC</td>
<td>160</td>
<td>0</td>
<td>1</td>
<td>Undefined</td>
</tr>
<tr>
<td>DetOne</td>
<td></td>
<td>Detector 1</td>
<td>Detector</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Subdetector</td>
</tr>
<tr>
<td>DetTwo</td>
<td></td>
<td>Detector 2</td>
<td>Detector</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>LDC</td>
</tr>
</tbody>
</table>

Connected successfully to DATE database
Performance Monitoring - AFFAIR

- Home-made development
- Based on free software

Fabric monitoring
DATE performances
ROOT I/O performances
CASTOR performances
Data quality monitoring - MOOD

- MOOD framework
  - Interfaces to detector code
  - Software development in all institutes

- Applications:
  - Raw data integrity
  - Detector performance
Data quality monitoring - MOOD
Outline

• ALICE DAQ
  – Trigger/DAQ logical model - Requirements
  – Trigger/DAQ at LHC
  – ALICE Trigger/DAQ Architecture

• Hardware components

• Software packages

• Towards the final ALICE

• Conclusions
Combined ITS test beam

Most major detectors (SDD, TPC, TRD, TOF, Muon, HMPID) integrated with the ALICE DAQ system
DAQ for ITS Test Beam

LDC SDD
LDC SPD

GDC:
Seil 2 Xeon 2.4 GHz 1GB
RH linux 7.3.2 – DATE 4.6

CASTOR

1 equipment:
• RORC

LDC VME:
CCT VP CP1
RH linux 7.3.2 – DATE 4.6

3 equipments:
• Trigger (HW/SW)
• Microstrip readout
• TDC

• Event builder
• Run control
• Monitoring
• NFS server

Fast ethernet
Control for ITS Test Beam

- ITS Partition Control Agent
  - PCA
  - Waiting for commands
  - FERO READY: TRUE
  - Access rights granted to the PCA
  - DCS ..., DAQ ..., TRIGGER ...
  - PCA info: 16:16:19 :{}

- ITS Individual Detectors operations
  - SDD, SPD, SSD
  - Waiting for commands: STANDALONE_RUN

- ITS TRG details
  - TPA, LTU-SDD, LTU-SPD, LTU-SSD
  - STOPPED, STANDALONE_STOPPED, START, STOPPED, STANDALONE_STOPPED
Simulation

- Complete model of the ALICE TRG/DAQ
- See next 2 talks:
  - Tome Anticic
    ALICE Trigger and DAQ – Simulation
  - Linda Vickovic
    ALICE Mass Storage System Simulation
Conclusions

- ALICE TRG DAQ system
  - All major hw and sw components released
  - Production phase
- Towards the final system
  - Integrated with most major detectors
  - First combined test beam
  - Simulation
- System evolution
  - New technology: PCI-XP, multicore CPU, 10 Gbit Eth
  - Isolate from technology obsolescence by logical interfaces (DDL or Mass Storage System)
- Start installation 1Q 2005. Startup 1Q 2007