HERA-B’s Online Reconstruction Farm: From a Vague Idea to a Running System

Andreas Gellrich
DESY -IT-

Technical Seminar
Zeuthen

6 March 2001
Introduction:
◆ Phase 0 (1994/1995): Defining the Problem
◆ HERA-B Requirements
◆ Status as of CHEP ‘95
◆ Key Questions

The Farm:
◆ Phase 1 (1995/1996): The Test Farm
◆ Phase 2 (1997/1998): The Prototype Farm
◆ Phase 3 (1999/2000): The 4LT Farm

Summary:
◆ Phase 4 (2001+): Running and Improving
◆ Key Results
◆ Status as of CHEP 2000
◆ Conclusions
Phase 0 (1994/1995)

Defining the problem.

1994/1995
The Golden Decay
A Typical Event

prism

evt inp=/net/sgi/herab/rm/rot/gtest-rot.zeb
date=06/03/96
time=11.25.16
event=2
picture=4

g eo=/net/sgi/herab/rm/rot/hbxgeo-rot.zeb
meta=events.ps
hbook=tf.hist
nvert/mult=5/191
HERA-B’s Requirements (1995)

◆ physics program starts in 1998 (serious constraint!)

◆ event building after SLT
◆ trigger task on TLT
◆ full online event reconstruction and final selection on 4LT

◆ TLT/4LT multi-processor system (farm)

◆ raw event size: 100 kB
◆ SLT $\rightarrow$ TLT : 1-2 kHz @100 - 200 MB/s with low latency
◆ TLT $\rightarrow$ 4LT : 50 Hz @ 5 MB/s
◆ 4LT $\rightarrow$ tape: 20 Hz @ 2MB/s
◆ yearly volume: 20 TB (200 M events)
◆ TLT processing time: 30 ms
◆ 4LT processing time: 2s
◆ farm nodes: 100

◆ Software: FORTRAN, ZEBRA, CERN-libs
Running Systems:

- clear separation between online and offline:
  - real-time environment in DAQ (VME)
  - offline calibration and alignment on main-frame
  - offline event reconstruction on main-frame

- DAQ output rate to tape 10 Hz (H1)
- 1.5 TB/y raw data (H1)
- 5 TB/y reconstructed data (H1)
- 12 M events/y (H1)
- reconstruction 150 ms/event (MIPS R4400) (ZEUS)

- L4 online farm based on CES VME-68K processors (H1)
- L5 offline event reconstruction on SGI Challenge XL (H1/ZEUS)
CHEP ‘95 (Computing in the Next Millenium):
“Unix and VxWorks operating systems.”
“... pseudo-real-time reconstruction pass with a latency of minutes.”
“A Reconstruction (...) Farm based on multiple Unix processors ...”
“... the impact of Windows-NT should be evaluated.”
“... primarily, the C++ programming language.”
(D. Quarrie, BaBar)

“... systems based on parallel commercial systems ...”
“... PCs represent a continuously growing commodity market ...”
“... speakers tended to go for Microsoft products (...) for PC (...) for LHC.”
“The motivation was cost.”
(M. Fischler, Fermilab)

“... trend for the second level triggers (...) on conventional processor farms.“
“... data storage and access problems (...) frightening by todays standards ...
“τ_{experiment} >> τ_{change}”
(J. N. Butler, Fermilab)
Key Questions

How can we get the computing power at a reasonable price?
   How can a big system be maintained?
   Does the system scale?
   Is the system flexible?
   Do we really know the requirements?

How can we get the data in and out?
   What are the requirements?
   Do we need real-time capabilities?

How do we get the software running?
   Do we need an operating system at all?
   How do we structure the software?

Is it just the farm?
   How do we store the data?
   How do we access the data afterwards?
The Test Farm System

1995/1996
Test Farm (1995/1996)

Prerequisites:
- SHARC-DSPs in DAQ system and SLB
- SLB/SLT switch unclear
- pipelined SLT to work on regions of interests (RoI)
- TLT/4LT on same hardware
- real-time requirements to TLT/4LT (bandwidth, latency)

Test Farm System:
- multi-processor system for TLT/4LT
- DS-links in conjunction with C104-switch for event building (SLB → SLT)
- VME-based PowerPC processors (CES, Cetia, Motorola)
- real-time operating system (VxWorks, embedded-AIX)

Questions:
- Do we fit into HERA-B’s DAQ system?
- Do we provide enough processing power?
- Does the operating system work for the HERA-B software (ARTE)?
- Do we have enough manpower to fulfill all tasks?
Data-Strobe Link (1996)

Ulli Schwendicke (Hardware), Ioji Legrand (Software)
Data-Strobe-to-PCI Interface Card (1996)

Kalle Sulanke (Hardware), Rainer Dippel (Software)

PCI-Controller
S 5933 (AMCC)

Tx-FIFO
16k x 8

Tx-Port
C 101
Data-Strobe-Link Adapter
Out

Rx-Port
In

Reg.-Port
Control

FIFO-Ctrl.

Control Logic

PCI-Ctrl.

C101-Ctrl.

Kalle Sulanke (Hardware), Rainer Dippel (Software)
Test Farm System (1996)

- PowerPC
- AIX
- Workstation

- 604e / 200 / 64
- 604 / 100 / 32
- PCI

- VME

- C104 Switch

- DS-Interface

- Ethernet

- NFS

- Sub-Net

- Disk

- Xterm

- Data-Strobe Links 10 MB/s

- SHARC link

Towards a Prototype Farm System

Problems:

◆ TLT and 4LT in same system:
  TLT: small times, high bandwidth, low latency (real-time-like) (latency driven)
  4LT: large times, moderate bandwidth (offline-like) (processing power driven)

Conceptual Changes:

◆ SLT/TLT on same nodes (Linux-PCs)
◆ WS/PC-based 4LT nodes
◆ Unix operating system (AIX, Linux)
◆ standard network (Fast-Ethernet)

Consequences:

◆ running ARTE as is on 4LT under Unix (Linux!)
◆ choice of farm nodes similar to SLT
◆ cheap and scalable solution
◆ consideration of growing Intel-CPU processing power
Alternative Solutions (1997)

VME system:
- standard online solution
- custom hardware
- embedded (diskless) system
- RISC (PowerPC)
- (VxWorks), AIX
- dedicated switch

Workstation/PC:
- commodity solution
- custom hardware
- RISC (PowerPC, Pentium Pro/II/III)
- AIX; Windows-NT, Linux
- (switched) standard network
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mainframes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hera-b (DESY)</td>
<td>SGI-PowerChallenge</td>
<td>IRIX 6.2 mips4</td>
<td>R10000</td>
<td>200</td>
<td></td>
<td>2000</td>
<td>2.7</td>
<td>85</td>
</tr>
<tr>
<td>aisat09 (IfH)</td>
<td>IBM-SP2 RS/6000</td>
<td>AIX 3.2.5</td>
<td>Power2</td>
<td>66</td>
<td>64</td>
<td>1000</td>
<td>4.7</td>
<td>49</td>
</tr>
<tr>
<td><strong>Workstations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hbwgs01e (DESY)</td>
<td>IBM-73P RS/6000</td>
<td>AIX 4.1</td>
<td>PPC 604</td>
<td>133</td>
<td>144</td>
<td>512</td>
<td>3.8</td>
<td>61</td>
</tr>
<tr>
<td>neptun (IfH)</td>
<td>Motorola RS/6000</td>
<td>AIX 4.1.4</td>
<td>PPC 604e</td>
<td>200</td>
<td>64</td>
<td>256</td>
<td>2.3</td>
<td>100</td>
</tr>
<tr>
<td><strong>Personal Computers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hb-con03 (DESY)</td>
<td>PC</td>
<td>Linux</td>
<td>P5</td>
<td>60</td>
<td></td>
<td>256</td>
<td>14.6</td>
<td>16</td>
</tr>
<tr>
<td>hb-con02 (DESY)</td>
<td>PC</td>
<td>Linux</td>
<td>P5</td>
<td>90</td>
<td></td>
<td>256</td>
<td>11.3</td>
<td>20</td>
</tr>
<tr>
<td>elbe1 (DESY)</td>
<td>PC</td>
<td>Linux</td>
<td>P5</td>
<td>133</td>
<td></td>
<td>256</td>
<td>6.9</td>
<td>33</td>
</tr>
<tr>
<td>hb-linux (DESY)</td>
<td>PC</td>
<td>Linux</td>
<td>P5</td>
<td>200</td>
<td>128</td>
<td>512</td>
<td>5.1</td>
<td>45</td>
</tr>
<tr>
<td>hb-con09 (DESY)</td>
<td>PC</td>
<td>Linux</td>
<td>PPro</td>
<td>200</td>
<td></td>
<td></td>
<td>3.1</td>
<td>74</td>
</tr>
</tbody>
</table>
DAQ & Trigger

![Diagram showing the HERA Detector Fast Control Flow.](image)

**Rate**
- Detector: 10MHz
- Pre FLT
  - ECAL, MUON, HPT
  - simple tracking, pt-cuts, di-lepton mass
- SLT
  - 50kHz
  - track refit, magnet, vertexing
- TLT
  - 500Hz
  - SVD tracking, particle-ID
- 4LT
  - 50Hz
  - full reconstruction, physics selection
- Tape
  - 20Hz
  - 2.5

**Time**
- 10µs
- 5ms
- ~100ms
- ~4s
4LT Prototype Farm (1998)

Processor Nodes:
- 14 single-PII/400MHz
- 1 single-PII/400MHz SCSI-server
- 64 MB RAM, 4GB disks

Network:
- 12-port 3COM hub

Services:
- NFS/NIS service (executables, files)

Operating System:
- Linux (S.u.S.E.)
Towards the 4LT Farm

(1999/2000)
Purpose and Tasks

Full Online Event Reconstruction:

- 50 Hz * 4 s = 200 nodes
  → multi-processor farm

- run offline developed software online
  → provide appropriate software environment
  → make offline developments online-compliant (I/O)

Event Classification:

- mark events due to there physical contents
- to be used in event directories

Final Event Selection:

- 4LT trigger step

Data Logging:

- add reconstruction information to event
- send events to logger
Purpose and Tasks (cont’d)

Data Quality Monitoring:
◆ use availability of data
◆ use high statistics
  → central collection (gathering) of histograms

Preparation of Data for Calibration and Alignment:
◆ use availability of data
◆ use high statistics
  → central collection (gathering) of data
  → feedback system for database constants

Event Data Reprocessing in Shutdown Periods:
◆ use vast processing power of the farm
  → exploit online infrastructure
## 4LT Requirements Overview (1999)

<table>
<thead>
<tr>
<th>Item</th>
<th>TDR(^a)</th>
<th>Design</th>
<th>Status</th>
<th>Plan(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event size [kB]</td>
<td>50/70</td>
<td>100</td>
<td>10 - 120</td>
<td>10 - 200</td>
</tr>
<tr>
<td>Input rate [Hz]</td>
<td>50</td>
<td>50</td>
<td>25 - 250</td>
<td>25 - 250</td>
</tr>
<tr>
<td>Input bandwidth [MB/sec]</td>
<td>2.5</td>
<td>5</td>
<td>2 - 4</td>
<td>2 - 5</td>
</tr>
<tr>
<td>Processing time [sec/event]</td>
<td>2</td>
<td>4</td>
<td>0 - 2</td>
<td>0 - 4</td>
</tr>
<tr>
<td>4LT nodes [CPU]</td>
<td>100</td>
<td>200</td>
<td>~30</td>
<td>200</td>
</tr>
<tr>
<td>Output rate [Hz]</td>
<td>5</td>
<td>20</td>
<td>25 - 250</td>
<td>10 - 250</td>
</tr>
<tr>
<td>Output bandwidth [MB/sec]</td>
<td>0.35</td>
<td>2</td>
<td>2 - 4</td>
<td>2 - 4</td>
</tr>
<tr>
<td>Data volume [TB/year]</td>
<td>3.5</td>
<td>20</td>
<td>~4</td>
<td>20</td>
</tr>
</tbody>
</table>

\(^a\) RAW data (L3 output) are kept for limited time; PGT and MINI data sets are archived.
\(^b\) Readiness for 2000 running.
\(^c\) No OTR yet in common DAQ running.
\(^d\) No OTR yet; dominated by stand-alone RICH ring search.
\(^e\) For current standard running only a handful of nodes are in use.
\(^f\) 3 TB in spring + 1 TB after July.
4LT Farm (mid 1999)

Infrastructure:
- 4 shelves a 30 PCs
- air cooling
- power supply

Services:
- hb4ltsrv (NFS/NIS)
- hb4ltctl (installation)
- hb4ltlog (local logging: 18 + 8 GB)

Nodes:
- 17 single-PII / 450 MHz, 64 MB
- 6 dual-PII / 450 MHz, 256 MB

Network:
- Fast-Ethernet
- 8 CISCO 2924-XL switches
- 1 CISCO 2916-XL-M switch
4LT Farm (end 1999)

Processor Nodes:
- 93 dual-PIII/500MHz
- some single-CPU machines
- 256 MB SDRAM, 13 GB disks

Network:
- 8 switched Fast-Ethernet mini-farms
- CISCO-switches
- Gigabit-Ethernet uplink

Services:
- NFS/NIS service (executables, files)
- slow control (http)
- local logging (36 GB, DLT7000)

Operating System:
- Linux (S.u.S.E.)
4LT Farm Hardware
4LT Farm Network

Central Switch: (16-port)
- 8 Mini-Farms (hb4ltsw1-8)
- server (hb4ltsrv) [NFS/YP]
- controller (hb4ltctl) [http]
- logger (hb4ltlog / hb4ltout)
4LT Farm Services

Services: NFS/YP-server, control node (http), local logging (disk/tape)
4LT Farm Network

Mini-Farm 3 (hb4lt020)

020 021 034

(CISCO WS-C2924-XL 24-port)

Mini-Farm 8 (hb4lt095)

095 096 109

(CISCO WS-C2924-XL 24-port)

Fast-Ethernet

main switch

(hb4ltsrv)

srv

NFS

NIS

(hb4ltctl)

ctl

(hb4ltlog)

log (30 GB)

central switch

GigaBit-E

(12-port)

GigaBit-E

(16-port)

DESY Computer Center

(hbslt000)

PC

PC

PC

(hubs)

local

106GB

sgi

osm

tape

20TB/y

Disk

SLT/TLT farm

095 096 109

(CISCO WS-C2924-XL 24-port)

Zeuthen, 6 March 2001

Andreas Gellrich, DESY -IT-
4LT Farm Event Control

(Fang Sun, Peter Wegner)

**S/TLT**

1 2 ... m

**Network**

Data

1 2 3 ... n

**4LT**

ID_Request

ID_Table

ID_Report

(ID_Request)

Request

(SHARC)

Free_ID_Queue

(out)

in

Slot 1

Slot 2

Slot 3

Slot n

Network

Data

4LT Farm Event Control

(Fang Sun, Peter Wegner)
4LT Calibration & Alignment

(A. Gellrich, Ulli Schwanke)

DAQ & Trigger

new CnA data

new key

Event data (key = version tag)

pull data

CnA Data base

L-4 L-4 L-4 L-4 L-4

rhp

push CnA data

Gatherer

Moni Cali

GUI
4LT Farm Slow Control

(Andreas Gellrich, Fang Sun, Ulli Schwanke)

Requirements:
- shift crew usage
- remote access
- status control
- temperature control
- monitoring

Implementation:
- one process per node
- one file per node
- sysinfo
- lm_sensors.o
- /proc/sensors
- http-service

Alternatives:
- CAN-bus
- common slow control
Data Management

Physics:
- yearly volume: 20 Hz * 150 kB * $10^7$ s = **30 TB** (8 TB in 2000)
- physics rate: ~1 Hz (Golden Decay: O(1000)/y)

Environment:
- logging to disk measured up to 12 MB/s
- archiving/mining to/from tape library (OSM) measured at 5 MB/s

Data Sets:
- raw data plus reconstruction output (**DST**) on tape (30 TB/y)
- reconstruction output only (**MINI**) on disk (O(1 TB/y))
  - standard analysis based on MINIs
- only O(1-10%) of selected DST on disk

Tools:
- event index files / event directories
- automatic pseudo-online event selection based on classification
- staging (common disk pool for copies of tape files)
A 200 processor Linux-farm is in operation since beginning of 2000.

The system runs stably and reliably.

Full event reconstruction is performed online; re-processing has started.

Compared to pseudo-online or offline approaches
reconstructed events are available for analysis immediately,
event classification and a final event selection can be done online,
a sophisticated data quality monitoring system is running online,
HERA-B is moving towards online calibration & alignment.

System performance exceeded design values considerably
with respect to processing power,
with respect to in- and output bandwidths.

(As expected) most of the work is still needed in the algorithmic part.
Phase 4 (2001+)

Running and Improving

2001+
HERA-B’s Farm History

From a vague idea ...

1994: Multi-processor farm.
1995: VME-boards with PowerPC and AIX.  
1996: Workstations with PowerPC and AIX.  
1997: SLT/TLT and 4LT.  
    ARTE under Linux  
    SLT Linux-PC farm.  
    4LT Linux-PC farm.  
    4LT in data path.  
1999: Switched Network with 100 dual-CPU PCs.  
    Completion of 4LT Farm.  
2000: Complete 4LT Farm runs in the 2000 run.  
    Full online event reconstruction in summer.  

... to a running system.
Key Results

Use SLT farm also for a potential TLT task because,
they both require real-time capabilities,
they allow to explicitly exploit commodity components on 4LT.

Use a multi-processor farm, because
it is cheaper than a mainframe,
no massive communication between farm nodes is needed,
it is highly scalable.

Use commodity computers, because
PCs get cheaper and cheaper,
PCs get faster and faster.

Use standard network components, because
the market can be followed (10 Mbit/s -> 100 MB/s -> 1 GB/s)

Use Linux, because
Linux becomes main platform in HEP.
Status as of CHEP 2000

Hardware:
- WS/PC farms
- Unix (Linux, Solaris), little Windows-NT
- Fast/GigaBit-Ethernet

Software:
- C/C++
- ROOT
- Objectivity (BaBar: even for event data)

Examples:
- BaBar: 79+221 Solaris SPARC (prompt reconstruction)
- D0: L3: 50 dual Windows-NT-PCs (online)
  100 dual-PentiumIII Linux-PCs (offline)
- NA48: L3: 42 dual-PentiumII Linux-PCs (online)
- BLAST: 50 Linux PC farem (online event reconstruction)
- HERA-B: L4: 100 dual-PentiumIII Linux-PCs (online event reconstruction)
Personal Remarks

Has been a long way, including ...

Developing, Building, Testing, Implementing, Commissioning, Running,
Thinking, Discussing, Debatting, Working, Presenting

... and we finally hit the goal.

Many thanx to ALL colleagues at

DESY Zeuthen and the Humboldt University,
delegated by Hermann Kolanoski and Thomas Lohse
Personal Remarks (2nd)

For ALL of us joining the Farm group was a success story!

Non-permanents prepared for life: (almost all were or became computer nerds!)

Andreas Gellrich [1995-2001] DESY Hamburg -IT-
(Axel Koehler) [1994-1998] Sun Microsystems
(Rainer Kowalik) [1994-1996] Siemens
Ulli Schwanke [1997-2001] UC San Diego (CMS)
Jose Hernandez [1999+] DESY Zeuthen

Permanents prospered their knowledge:

Ulli Gensch
Kolan Hermannoski
Holger Leich
Ulli Schwendicke
Kalle Sulanke
Peter Wegner
What are we doing next?
Personal Remarks (4th)

Drinking Champagne!