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

Andreas Gellrich
DESY -IT-

Technical Seminar
Zeuthen

6 March 2001

Contents

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

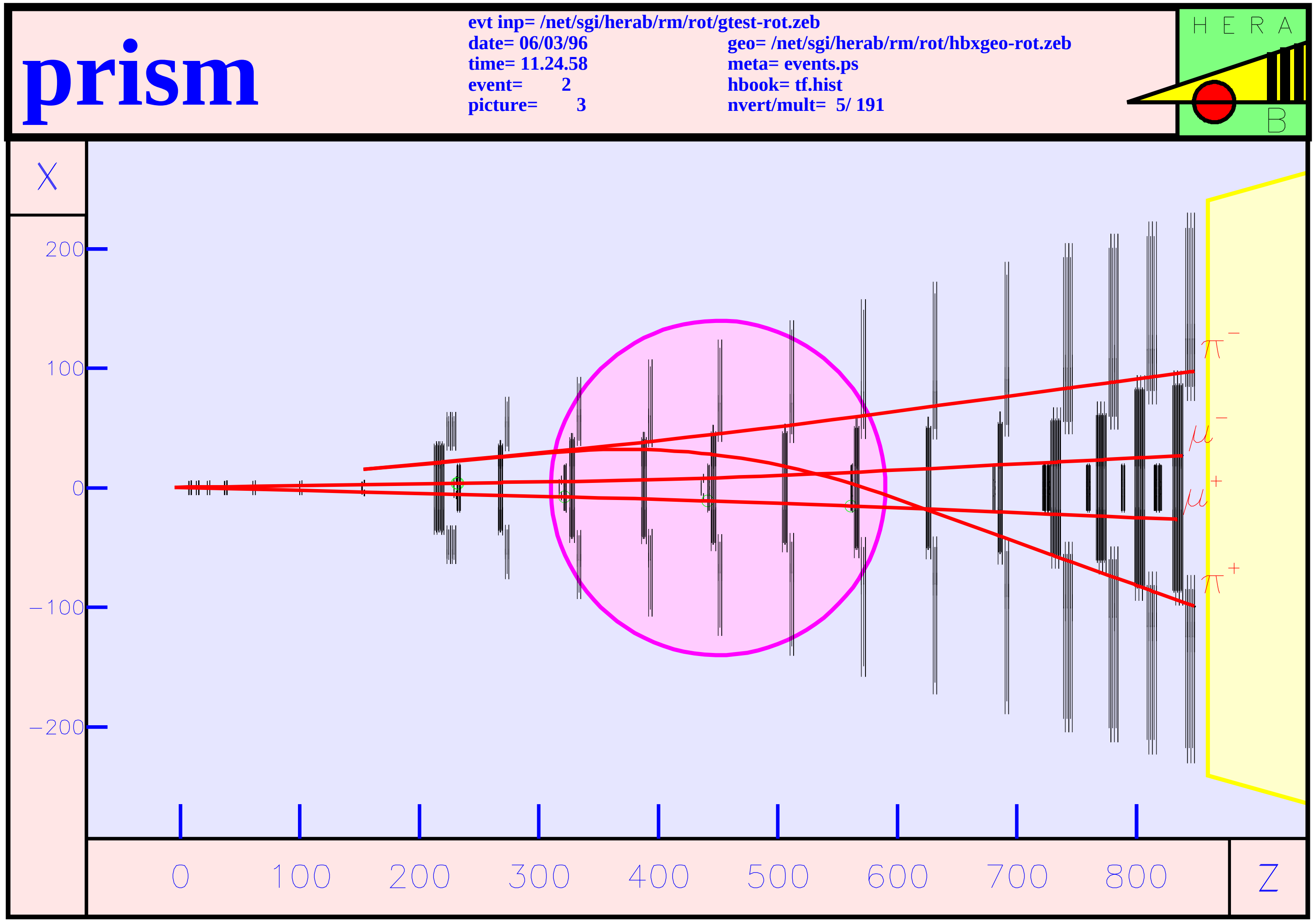
Summary:
- Phase 4 (2001+): Running and Improving
- Key Results
- Status as of CHEP 2000
- Conclusions

Phase 0 (1994/1995)

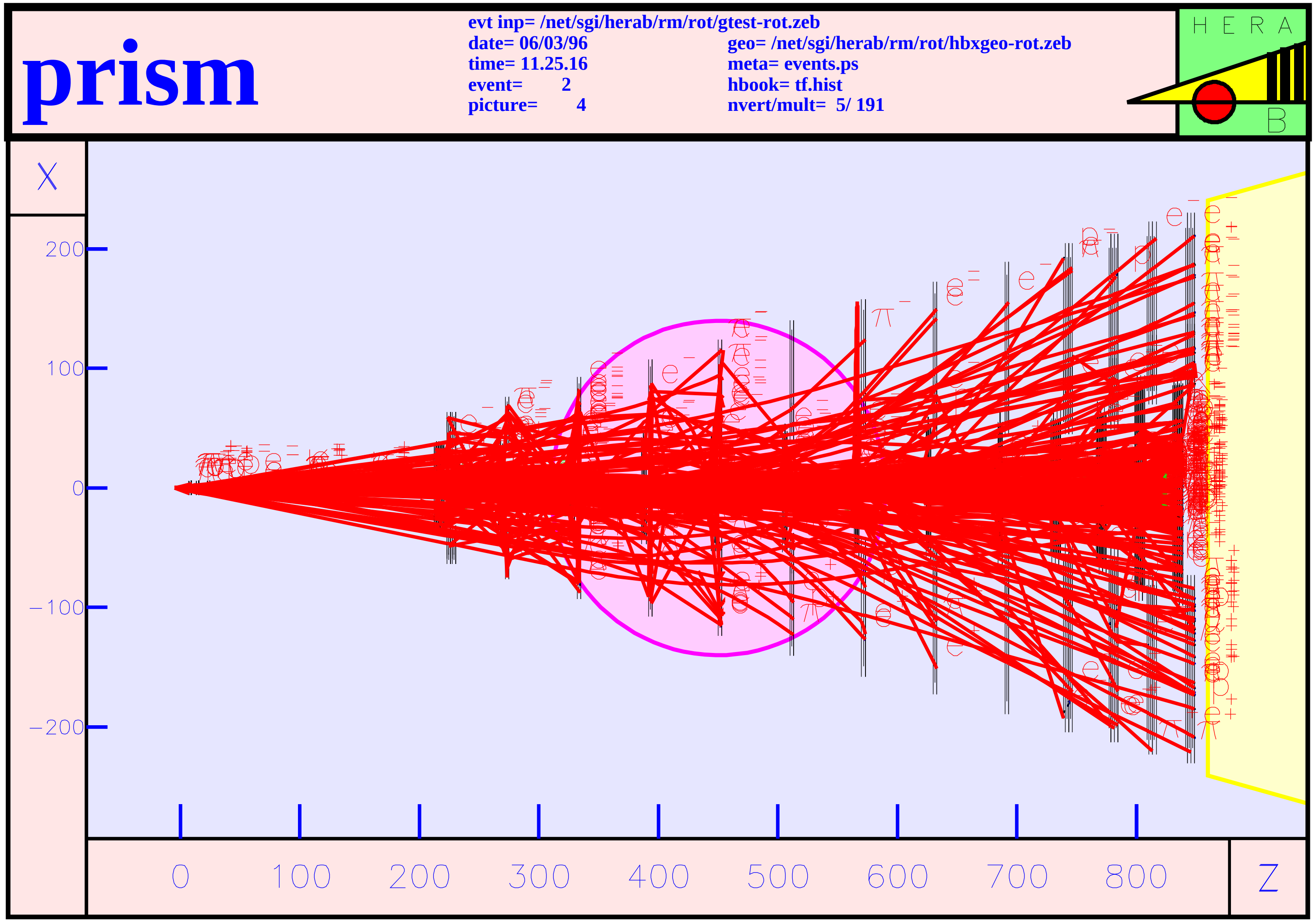
Defining the problem.

1994/1995

The Golden Decay



A Typical Event



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 → TLT :	1-2 kHz @100 - 200 MB/s with low latency
TLT → 4LT :	50 Hz @ 5 MB/s
4LT → 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

Status as of CHEP ‘95

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)

Status as of CHEP ‘95 (cont’d)

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 ...“
“$\tau_{experiment} >> \tau_{change}$”
(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:

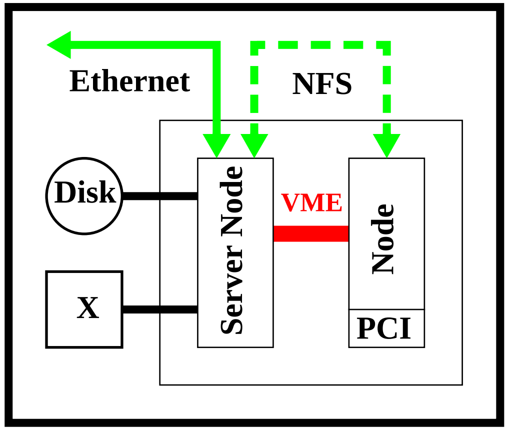
- multi-processor system for TLT/4LT
- DS-links in conjunction with C104-switch for event building (SLB → SLT)
- VME-based PowerPC processors (CES, Cetia, Motorola)
- real-time operating system (VxWorks, embedded-AIX)

Questions:

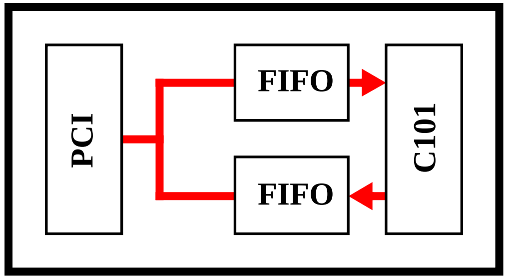
Do we fit into HERA-B's DAQ system?
Do we provide enough processing power?
Does the operating system work for the HERA-B software (ARTE)?
Do we have enough manpower to fulfill all tasks?

Project Overview (1996)

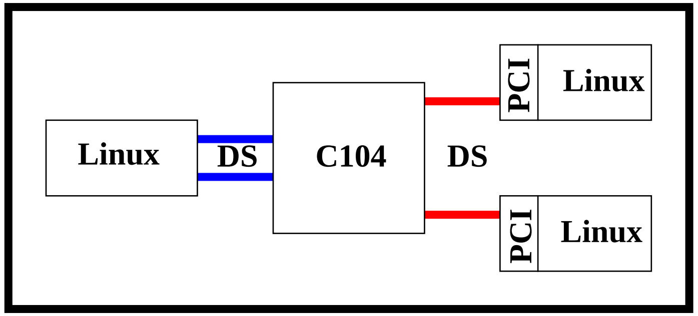
Embedded AIX



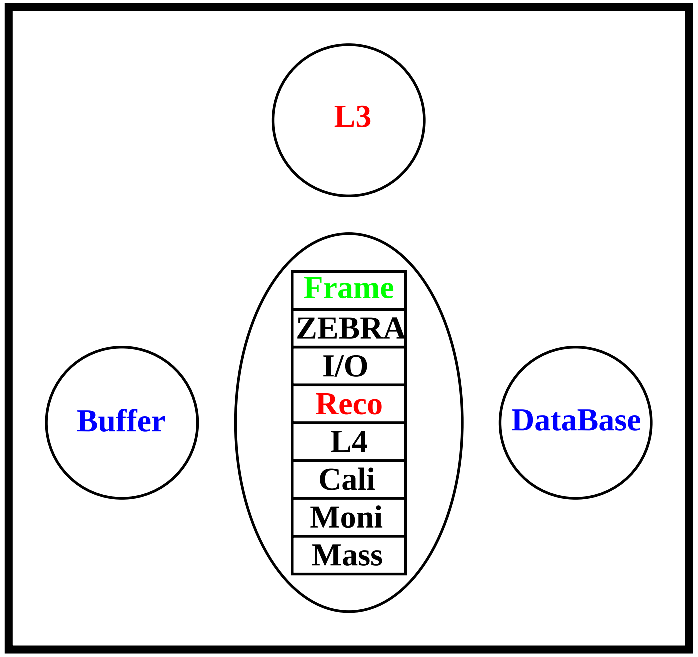
DS-to-PCI



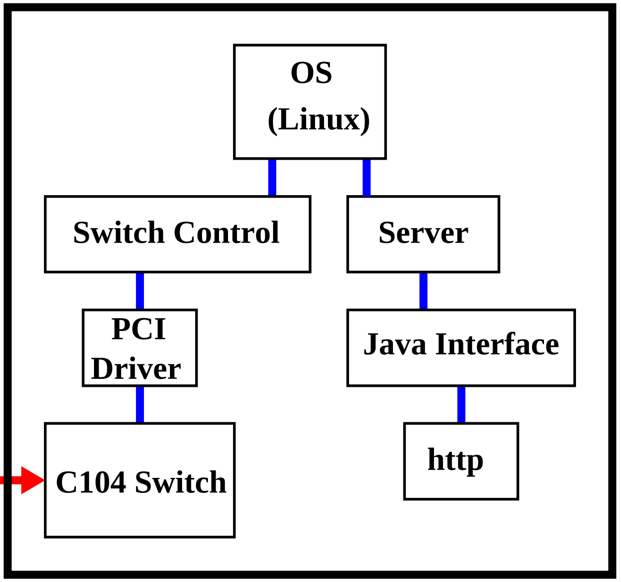
C104-Switch



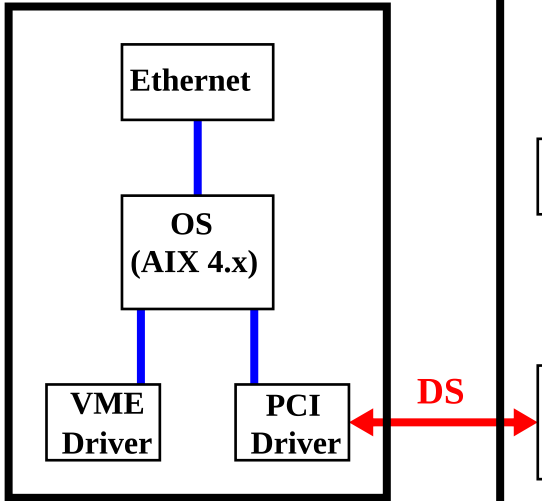
Applications



Switch Control

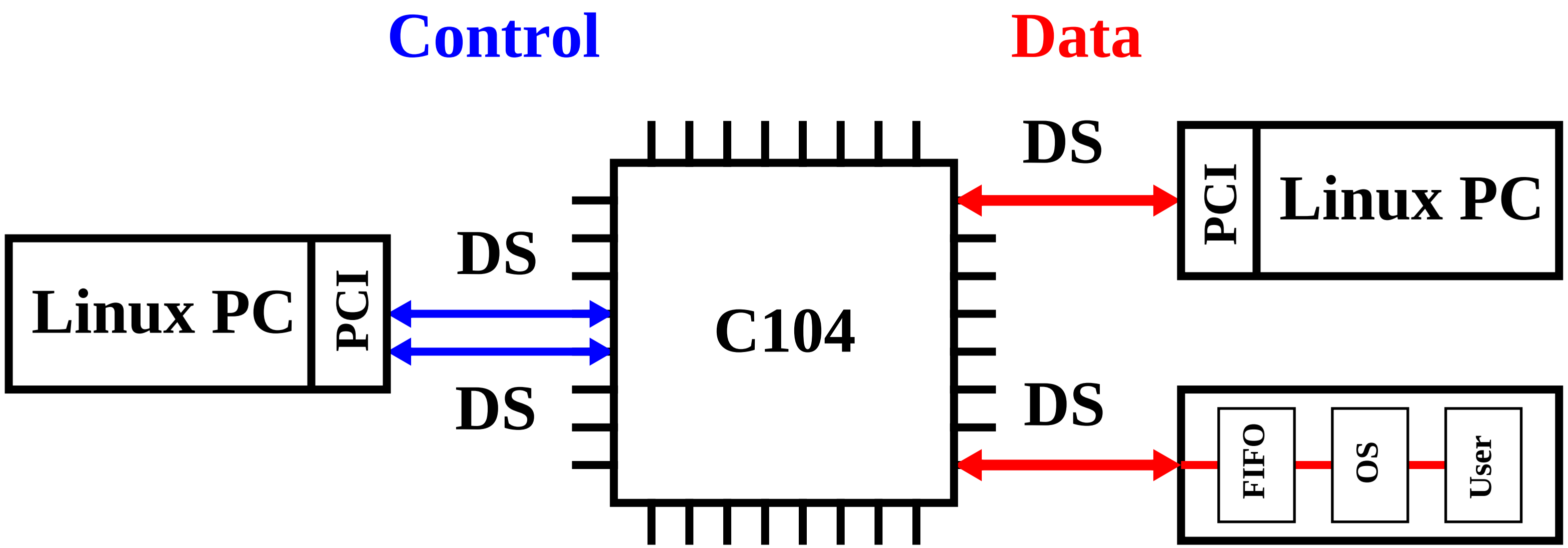


System Software



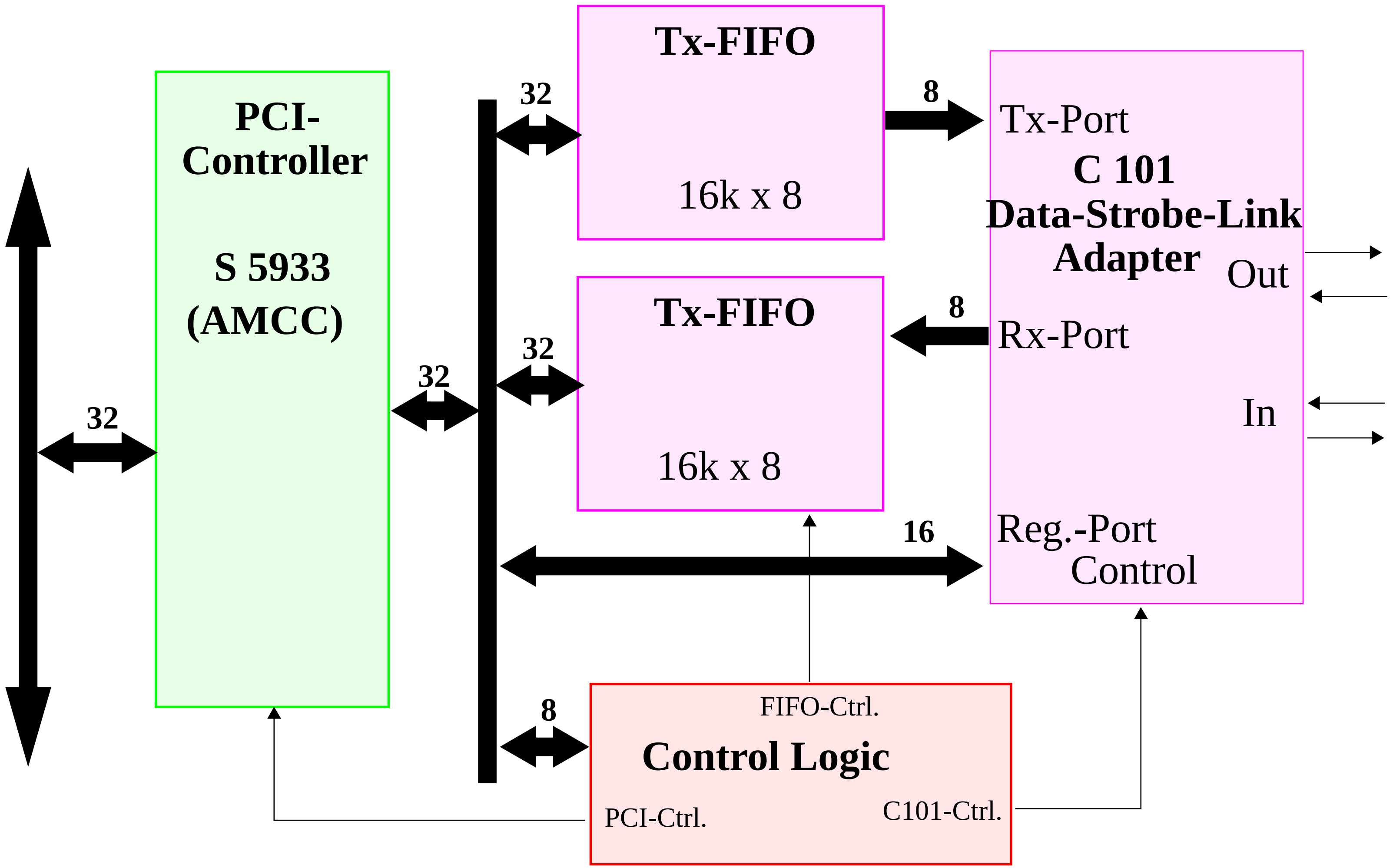
Data-Strobe Link (1996)

Ulli Schwendicke (Hardware), Ioji Legrand (Software)

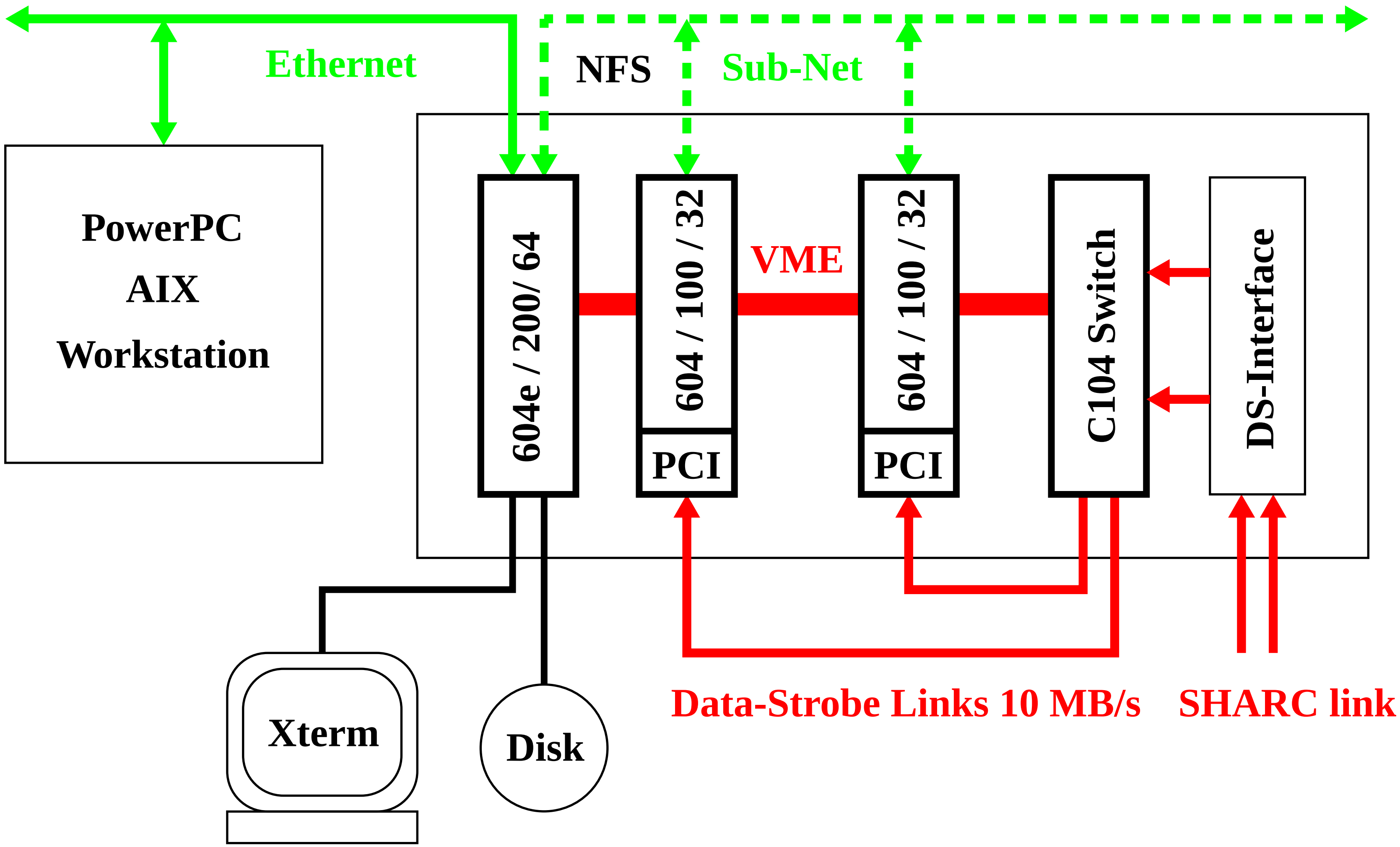


Data-Strobe-to-PCI Interface Card (1996)

Kalle Sulanke (Hardware), Rainer Dippel (Software)



Test Farm System (1996)



Phase 2 (1997/1998)

Towards a Prototype Farm System

(1997/1998)

Evaluation of Test Farm Project (1997)

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)

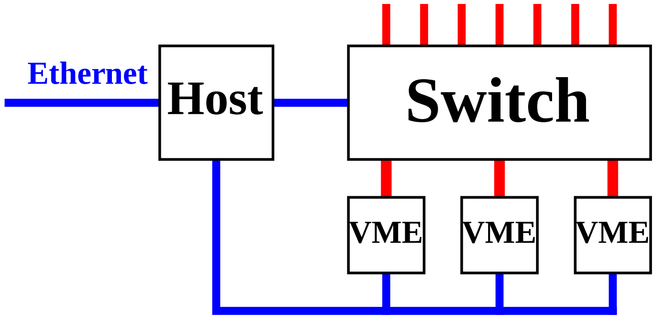
Consequences:

- 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)

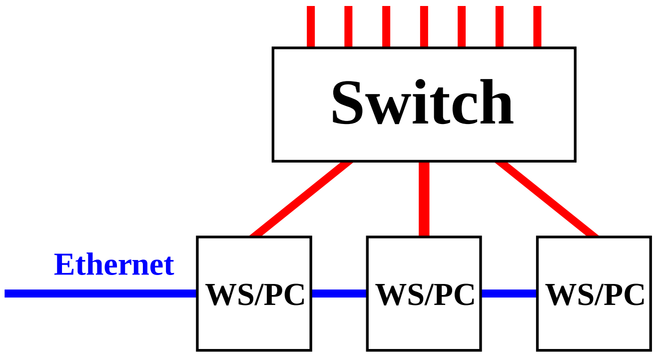
VME system:

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



Workstation/PC:

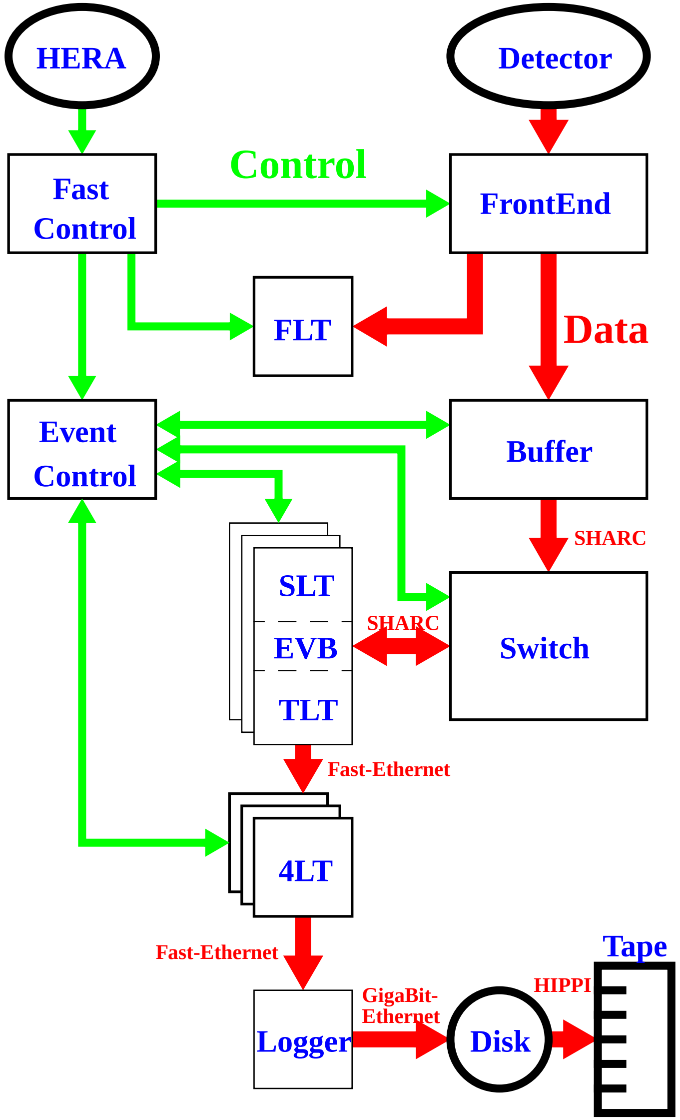
- commodity solution
- custom hardware
- RISC (PowerPC, Pentium Pro/II/III)
- AIX; Windows-NT, Linux
- (switched) standard network

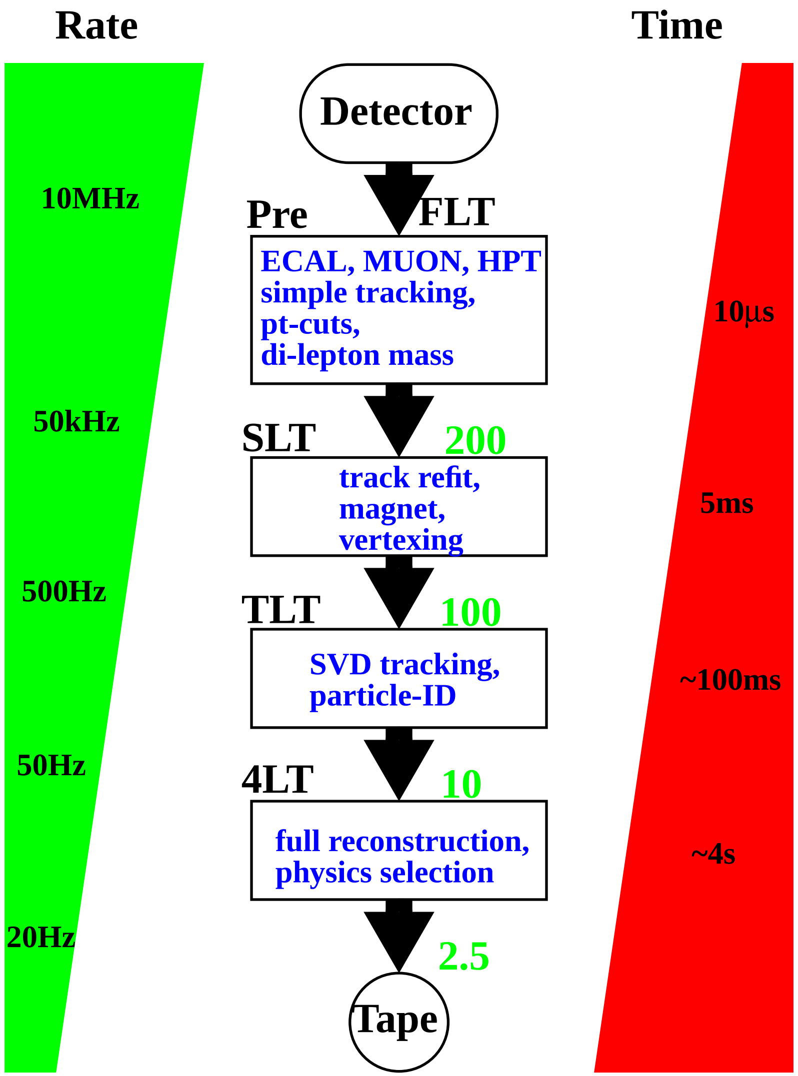


ARTE-Benchmarks (1997)

Machine	System	Operating System	CPU	Clock [MHz]	RAM [MB]	L2 [kB]	Time [s]	Performance [%]
Mainframes								
hera-b (DESY)	SGI-PowerChallenge	IRIX 6.2 mips4	R10000	200		2000	2.7	85
aisa09 (IfH)	IBM-SP2 RS/6000	AIX 3.2.5	Power2	66	64	1000	4.7	49
Workstations								
hbwgs01e (DESY)	IBM-73P RS/6000	AIX 4.1	PPC 604	133	144	512	3.8	61
neptun (IfH)	Motorola RS/6000	AIX 4.1.4	PPC 604e	200	64	256	2.3	100
Personal Computers								
hb-con03 (DESY)	PC	Linux	P5	60		256	14.6	16
hb-con02 (DESY)	PC	Linux	P5	90		256	11.3	20
elbe1 (DESY)	PC	Linux	P5	133		256	6.9	33
hb-linux (DESY)	PC	Linux	P5	200	128	512	5.1	45
hb-con09 (DESY)	PC	Linux	PPro	200			3.1	74

DAQ & Trigger


HERA
Detector
Fast Control
Control
FrontEnd
FLT
Data
Event Control
Buffer
SHARC
Switch
SLT
EVB
TLT
SHARC
Fast-Ethernet
4LT
Fast-Ethernet
Tape
GigaBit-Ethernet
HIPPI
Logger
Disk


Rate
Time
10MHz
50kHz
500Hz
50Hz
20Hz
Detector
Pre
FLT
ECAL, MUON, HPT
simple tracking,
pt-cuts,
di-lepton mass
SLT
200
track refit,
magnet,
vertexing
TLT
100
SVD tracking,
particle-ID
4LT
10
full reconstruction,
physics selection
2.5
Tape
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)

Item	TDR[a]	Design	Status	Plan[b]
Event size [kB]	50/70	100	10 - 120 [c]	10 - 200
Input rate [Hz]	50	50	25 - 250	25 - 250
Input bandwidth [MB/sec]	2.5	5	2 - 4	2 - 5
Processing time [sec/event]	2	4	0 - 2 [d]	0 - 4
4LT nodes [CPU]	100	200	~30 [e]	200
Output rate [Hz]	5	20	25 - 250	10 - 250
Output bandwidth [MB/sec]	0.35	2	2 - 4	2 - 4
Data volume [TB/year]	3.5	20	~4 [f]	20

a. RAW data (L3 output) are kept for limited time; POT 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)

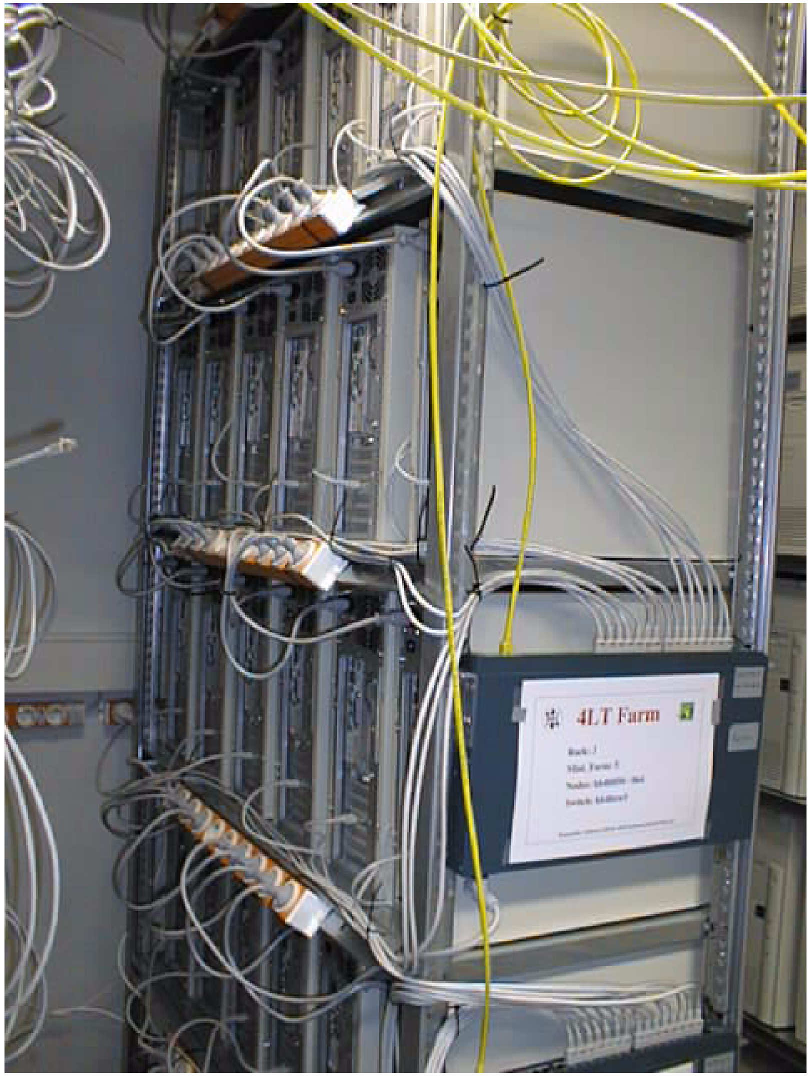
Processor Nodes:
- 93 dual-PIII/500MHz
- some single-CPU machines
- 256 MB SDRAM, 13 GB disks

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.)

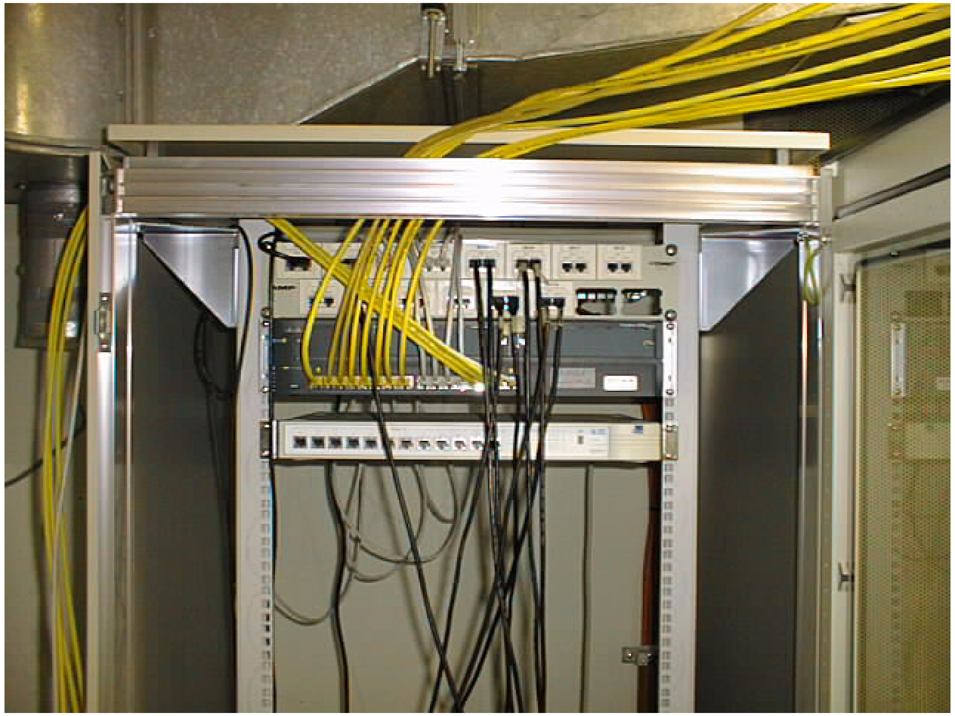
4LT Farm Hardware



4LT Farm Network

Central Switch: (16-port)

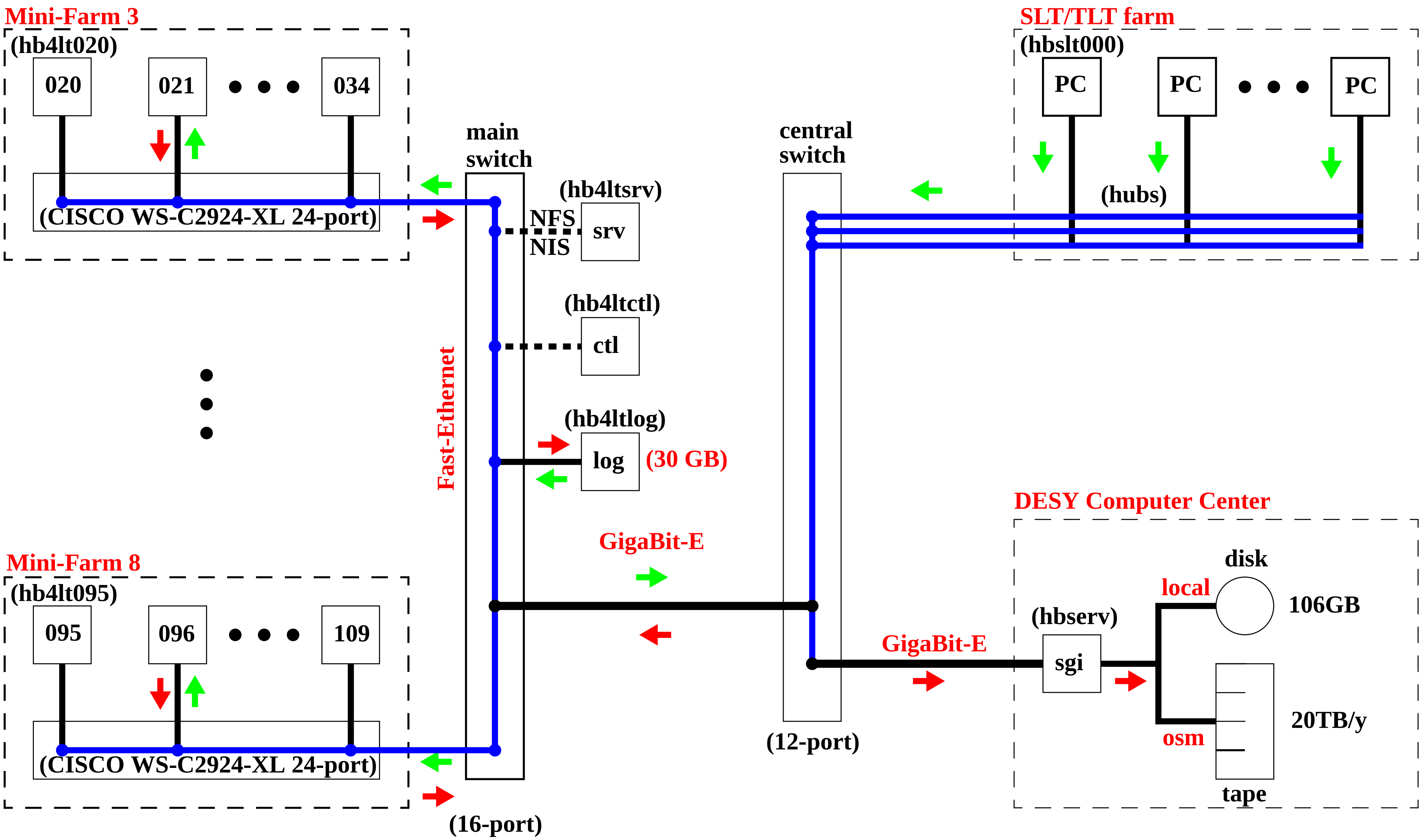
- 8 Mini-Farms (hb4ltsw1-8)
- server (hb4ltsrv) [NFS/YP]
- controller (hb4ltctl) [http]
- logger (hb4ltlog / hb4ltout)



4LT Farm Services

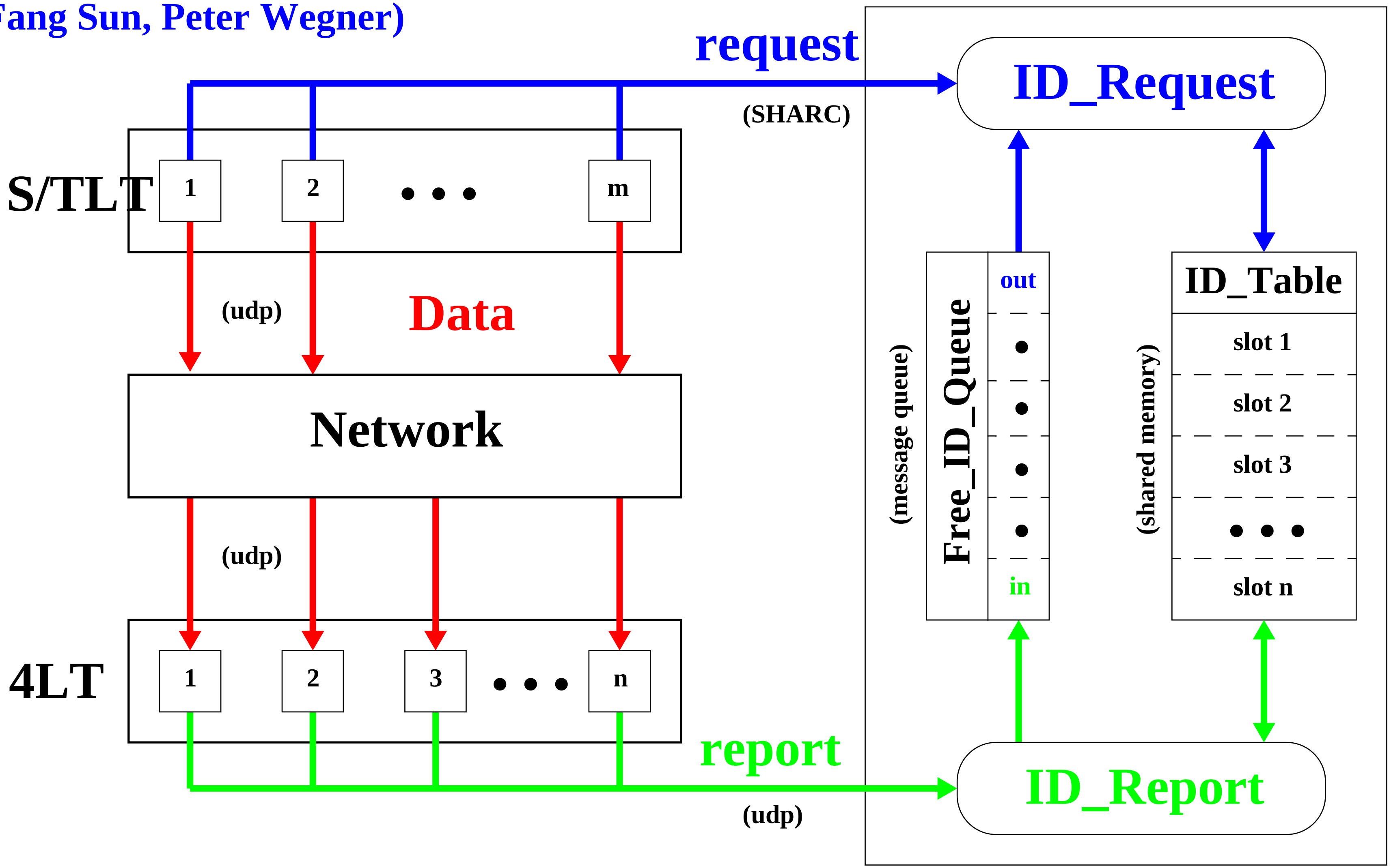
Services: NFS/YP-server, control node (http), local logging (disk/tape)

4LT Farm Network



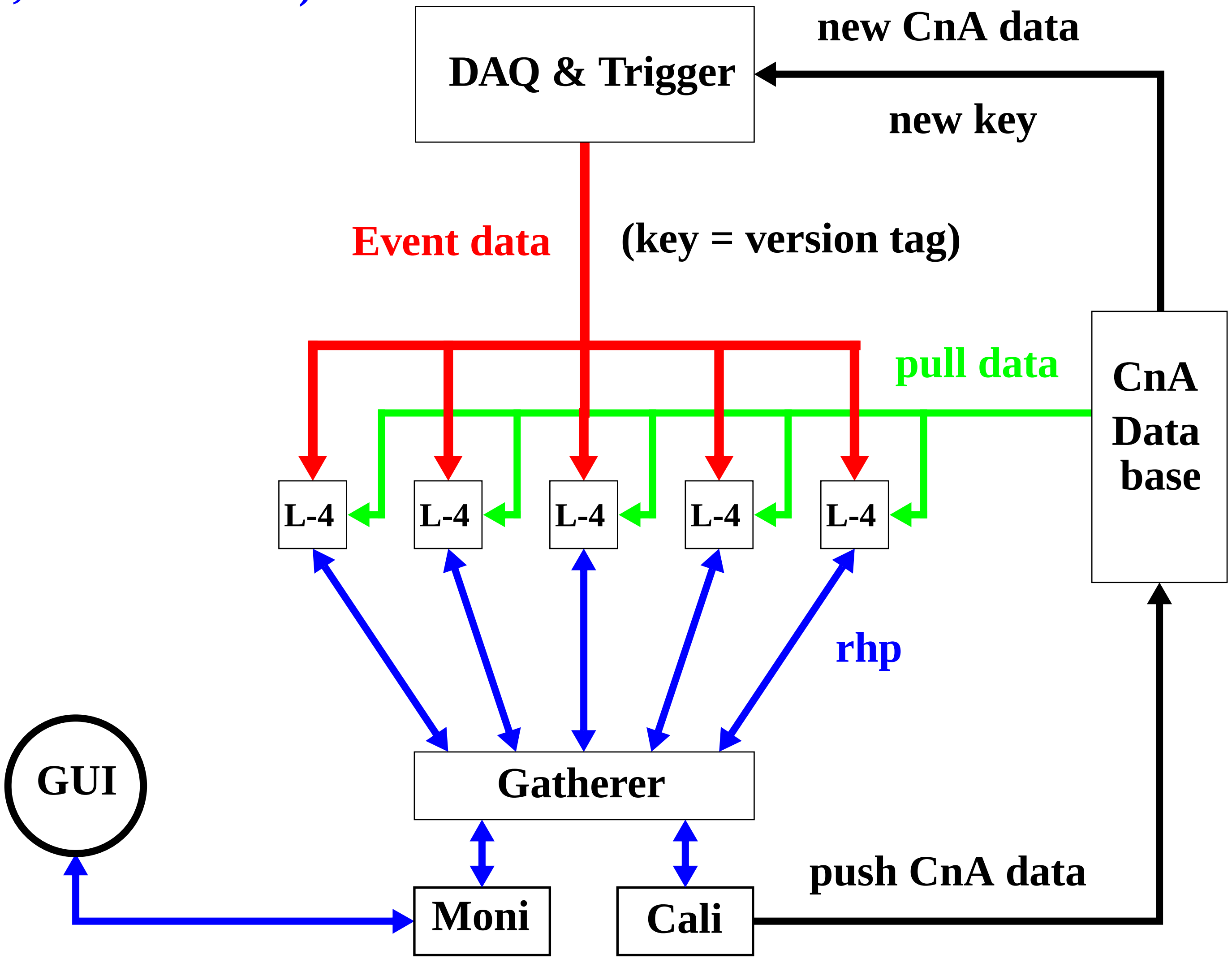
4LT Farm Event Control

(Fang Sun, Peter Wegner)

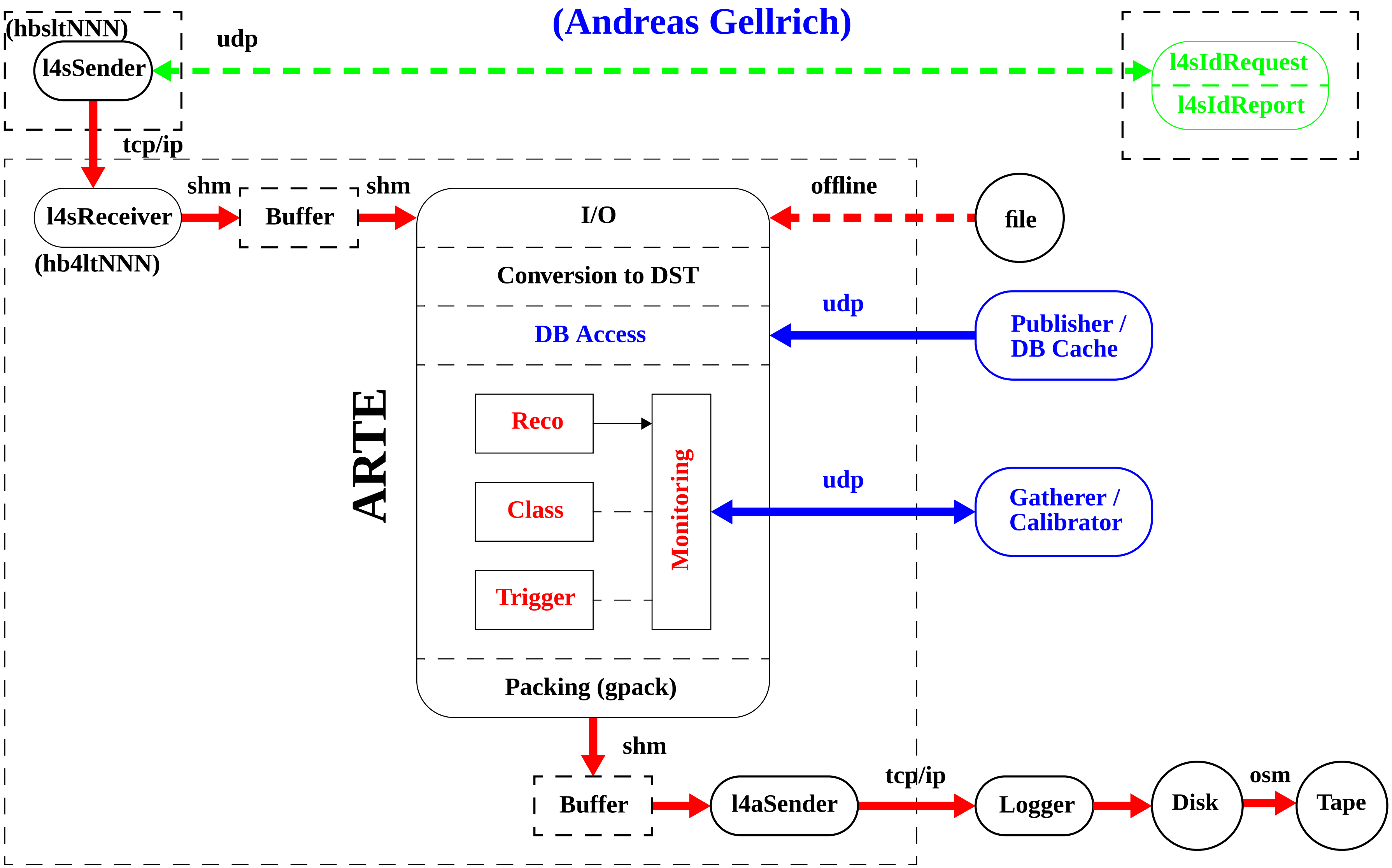


4LT Calibration & Alignment

(A. Gellrich, Ulli Schwanke)

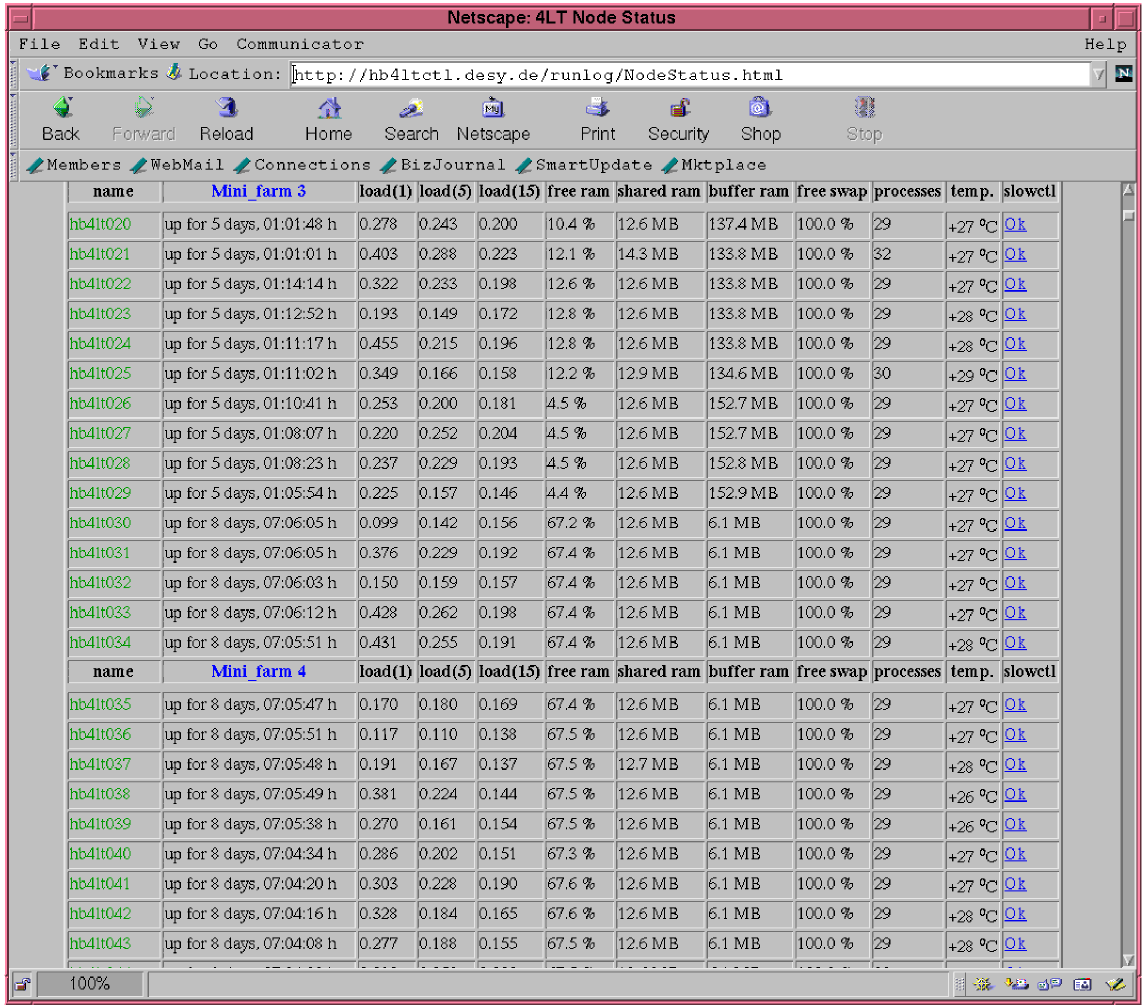


4LT Node Processes



4LT Farm Slow Control

(Andreas Gellrich, Fang Sun, Ulli Schwanke)



name	Mini_farm 3	load(1)	load(5)	load(15)	free ram	shared ram	buffer ram	free swap	processes	temp.	slowctl
hb4lt020	up for 5 days, 01:01:48 h	0.278	0.243	0.200	10.4 %	12.6 MB	137.4 MB	100.0 %	29	+27 °C	Ok
hb4lt021	up for 5 days, 01:01:01 h	0.403	0.288	0.223	12.1 %	14.3 MB	133.8 MB	100.0 %	32	+27 °C	Ok
hb4lt022	up for 5 days, 01:14:14 h	0.322	0.233	0.198	12.6 %	12.6 MB	133.8 MB	100.0 %	29	+27 °C	Ok
hb4lt023	up for 5 days, 01:12:52 h	0.193	0.149	0.172	12.8 %	12.6 MB	133.8 MB	100.0 %	29	+28 °C	Ok
hb4lt024	up for 5 days, 01:11:17 h	0.455	0.215	0.196	12.8 %	12.6 MB	133.8 MB	100.0 %	29	+28 °C	Ok
hb4lt025	up for 5 days, 01:11:02 h	0.349	0.166	0.158	12.2 %	12.9 MB	134.6 MB	100.0 %	30	+29 °C	Ok
hb4lt026	up for 5 days, 01:10:41 h	0.253	0.200	0.181	4.5 %	12.6 MB	152.7 MB	100.0 %	29	+27 °C	Ok
hb4lt027	up for 5 days, 01:08:07 h	0.220	0.252	0.204	4.5 %	12.6 MB	152.7 MB	100.0 %	29	+27 °C	Ok
hb4lt028	up for 5 days, 01:08:23 h	0.237	0.229	0.193	4.5 %	12.6 MB	152.8 MB	100.0 %	29	+27 °C	Ok
hb4lt029	up for 5 days, 01:05:54 h	0.225	0.157	0.146	4.4 %	12.6 MB	152.9 MB	100.0 %	29	+27 °C	Ok
hb4lt030	up for 8 days, 07:06:05 h	0.099	0.142	0.156	67.2 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt031	up for 8 days, 07:06:05 h	0.376	0.229	0.192	67.4 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt032	up for 8 days, 07:06:03 h	0.150	0.159	0.157	67.4 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt033	up for 8 days, 07:06:12 h	0.428	0.262	0.198	67.4 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt034	up for 8 days, 07:05:51 h	0.431	0.255	0.191	67.4 %	12.6 MB	6.1 MB	100.0 %	29	+28 °C	Ok
name	Mini_farm 4	load(1)	load(5)	load(15)	free ram	shared ram	buffer ram	free swap	processes	temp.	slowctl
hb4lt035	up for 8 days, 07:05:47 h	0.170	0.180	0.169	67.4 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt036	up for 8 days, 07:05:51 h	0.117	0.110	0.138	67.5 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt037	up for 8 days, 07:05:48 h	0.191	0.167	0.137	67.5 %	12.7 MB	6.1 MB	100.0 %	29	+28 °C	Ok
hb4lt038	up for 8 days, 07:05:49 h	0.381	0.224	0.144	67.5 %	12.6 MB	6.1 MB	100.0 %	29	+26 °C	Ok
hb4lt039	up for 8 days, 07:05:38 h	0.270	0.161	0.154	67.5 %	12.6 MB	6.1 MB	100.0 %	29	+26 °C	Ok
hb4lt040	up for 8 days, 07:04:34 h	0.286	0.202	0.151	67.3 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt041	up for 8 days, 07:04:20 h	0.303	0.228	0.190	67.6 %	12.6 MB	6.1 MB	100.0 %	29	+27 °C	Ok
hb4lt042	up for 8 days, 07:04:16 h	0.328	0.184	0.165	67.6 %	12.6 MB	6.1 MB	100.0 %	29	+28 °C	Ok
hb4lt043	up for 8 days, 07:04:08 h	0.277	0.188	0.155	67.5 %	12.6 MB	6.1 MB	100.0 %	29	+28 °C	Ok

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)

Status (2000)

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.	CHEP ‘95, Rio
1996:	Workstations with PowerPC and AIX.	DAQ 96, Osaka
1997:	SLT/TLT and 4LT.	CHEP ‘97, Berlin
	ARTE under Linux	
	SLT Linux-PC farm.	RT ‘97, Beaune
	4LT Linux-PC farm.	Beauty ‘97, Los Angeles
1998:	20 single-CPU nodes.	CHEP ‘98, Chicago
	4LT in data path.	
1999:	Switched Network with 100 dual-CPU PCs.	RT ‘99, Santa Fe
	Completetion of 4LT Farm.	
2000:	Complete 4LT Farm runs in the 2000 run.	CHEP 2000, Padova
	Full online event reconstruction in summer.	ACAT 2000, Fermilab
2001:	Full reprocessing of 2000 data.	

... 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!)

Rainer Dippel	[1994-1997]	Siemens
Andreas Gellrich	[1995-2001]	DESY Hamburg -IT-
(Axel Koehler)	[1994-1998]	Sun Microsystems
(Rainer Kowalik)	[1994-1996]	Siemens
Ioji Legrand	[1994-1998]	CERN/Caltech (MONARC)
Ulli Schwanke	[1997-2001]	UC San Diego (CMS)
Fang Sun	[1996-2000]	Nortel/Canada
Jose Hernandez	[1999+]	DESY Zeuthen

Permanents prospered their knowledge:

- Ulli Gensch
- Kolan Hermannoski
- Holger Leich
- Ulli Schwendicke
- Kalle Sulanke
- Peter Wegner

Personal Remarks (3rd)

What are we doing next?

Personal Remarks (4th)

Drinking Champagne!