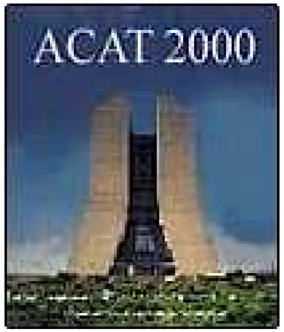
A Large Linux-PC Farm for Online Event Reconstruction at HERA-B

Andreas Gellrich
Humboldt-University Berlin, Germany
18 October 2000

(The project is sponsored by the German Ministry for Education and Science BMBF.)



Parallel Session: Very Large Scale Computing

Contents

- Introduction
    - HERA-B
    - DAQ & Trigger
- Purpose and Tasks
- Implementation
    - Hardware
    - Software
- Data Management
- Experiences
- Summary

Introduction: HERA-B

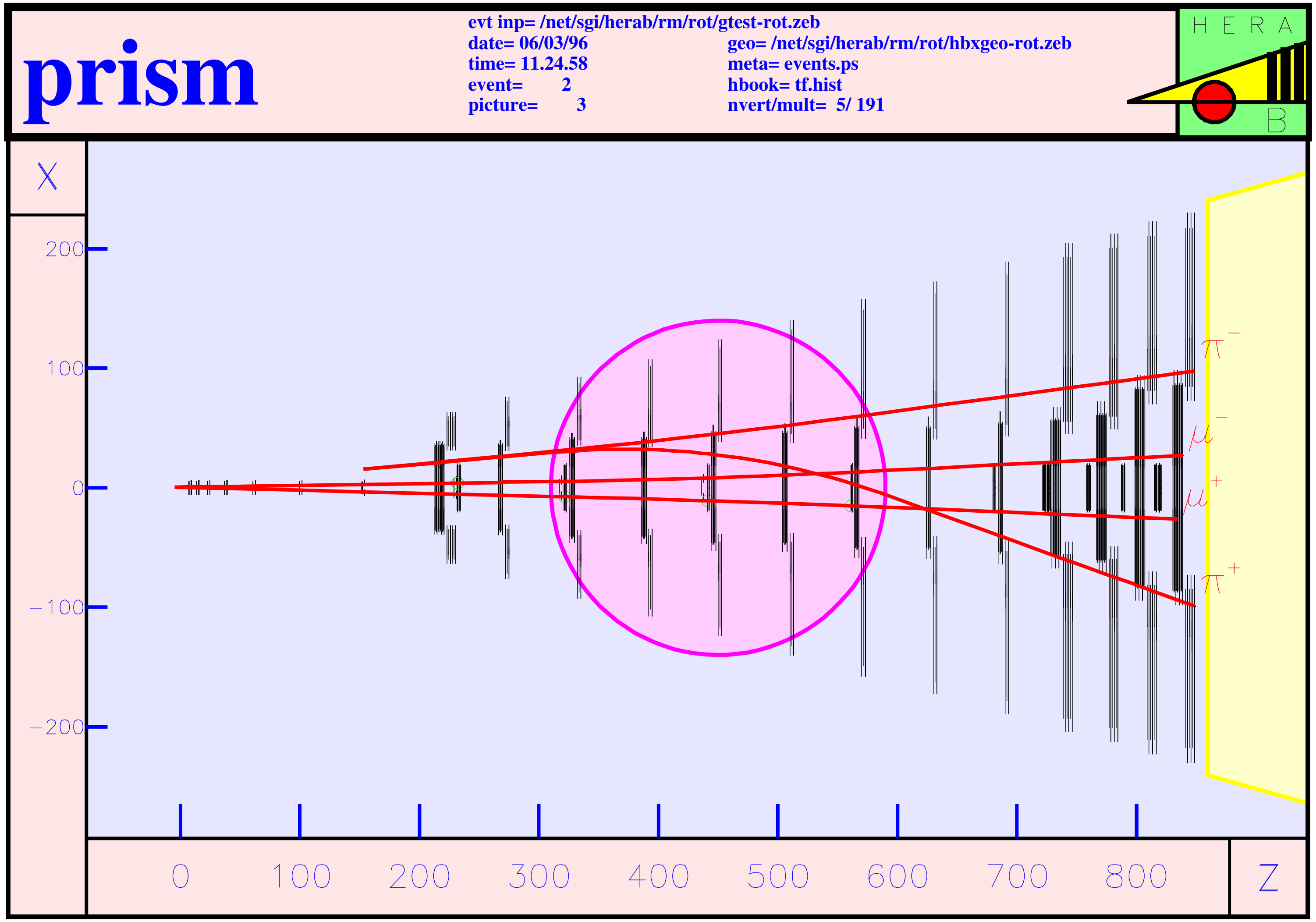
- **Physics:** → **study of B-decays**
  - **CP violation in $B^0 \rightarrow J/\psi\ K^0_s$**
  - **needs O(1000) reconstructed B^0**
  - $\sigma_{b\bar{b}} / \sigma_{inel}$ **= 12 nb / 13 mb =** 10^{-6}

- **Target:** → **wire target in HERA proton beam halo**
  - **10 MHz event rate @ 4 interactions**
  - **~1 Hz interesting physics**

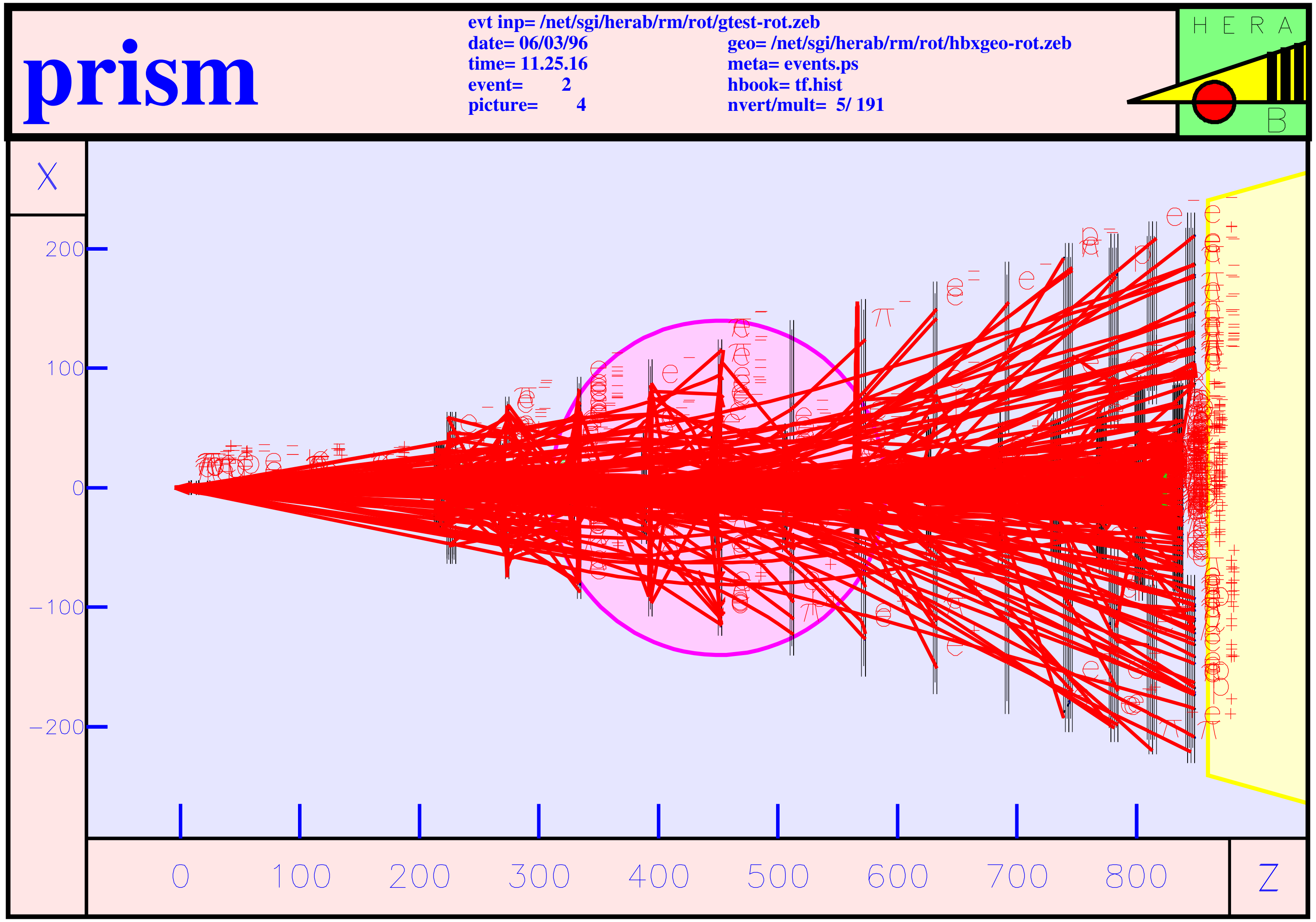
- **Detector:** → **read-out channels: ~540000**
  - **tracks / event: ~120**
  - **occupancy: <=20%**

- **Analysis:** → **full online event reconstruction**
  - **processing time: 4 s/event**
  - **logging rate: 20 Hz * 150 kB = 3 MB/s**
  - **data volume: 30 TB/y**

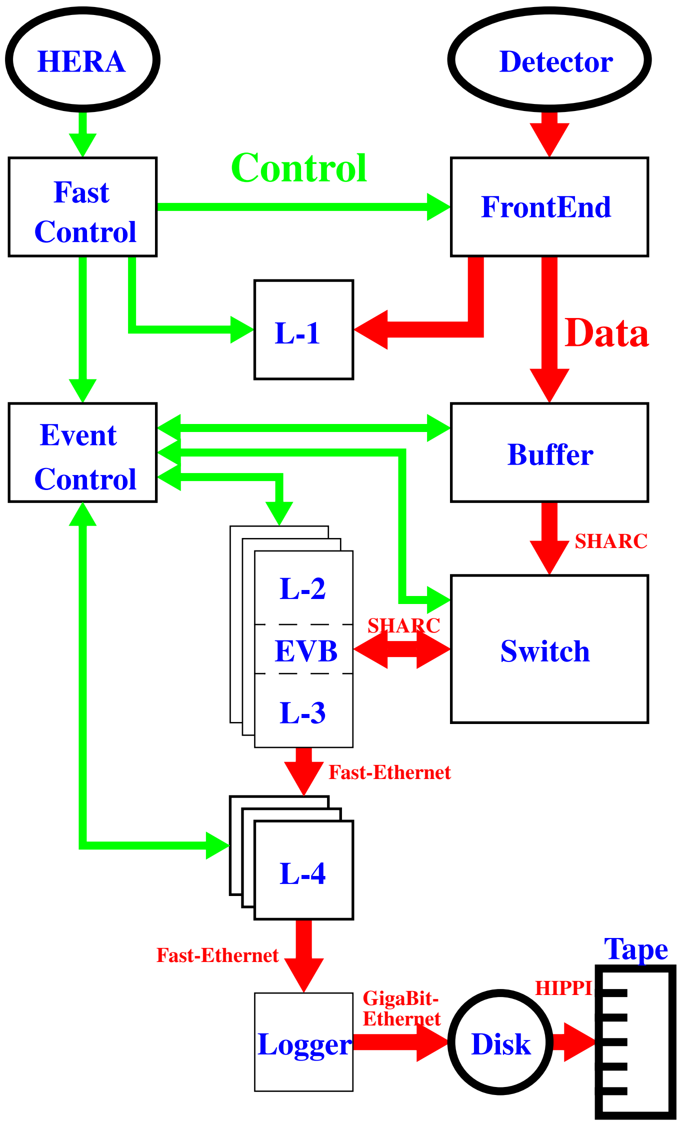
Introduction: Golden Decay

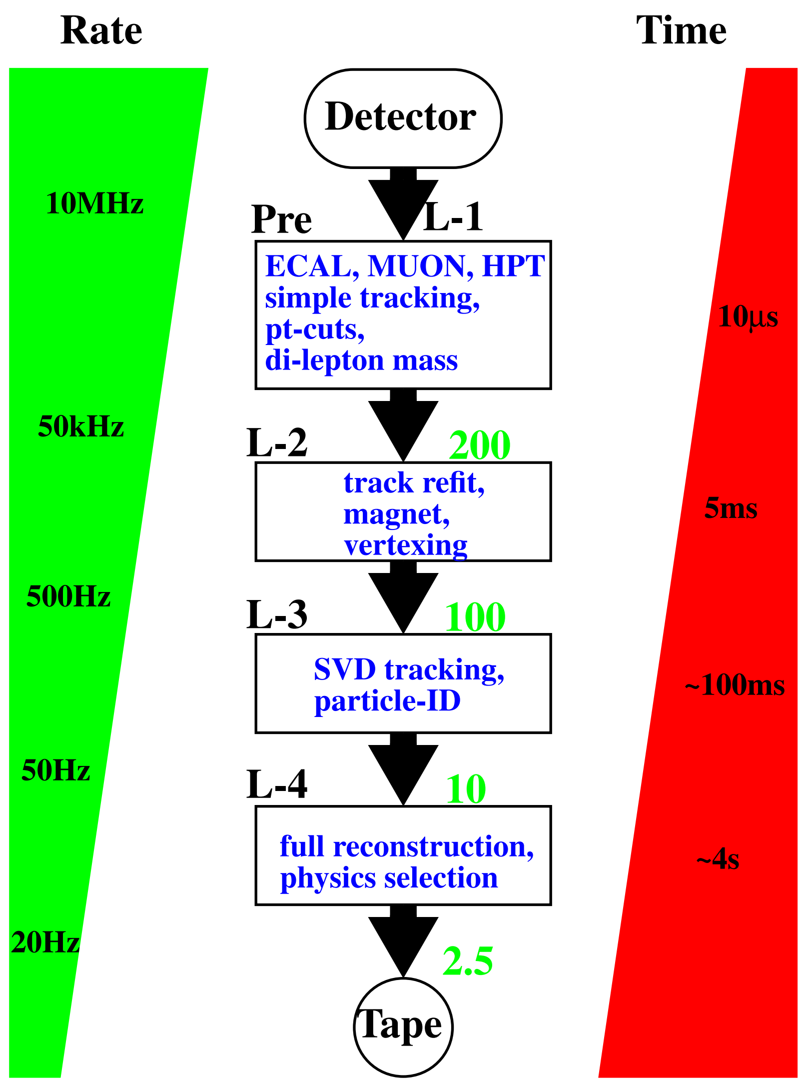


Introduction: Typical Event



Introduction: DAQ & Trigger





Concept: 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:**
  - L-4 trigger step
- **Data Logging:**
  - add reconstruction information to event
  - send events to logger

Concept: Purpose and Tasks (cont'd)

- **Data Quality Monitoring:**
  - use availability of data
  - use high statistics
    - ➔ central collection (gathering) of histograms

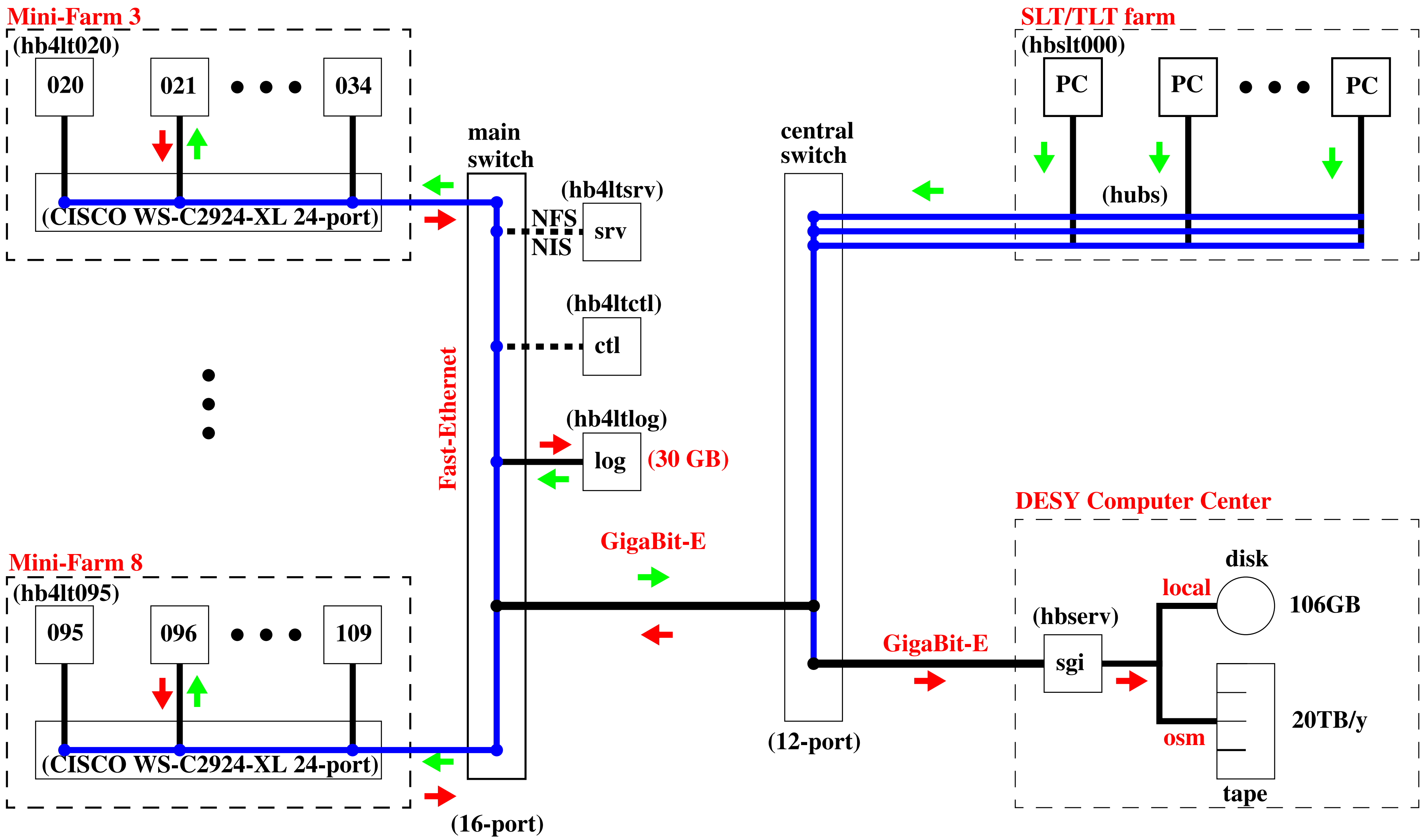
- **Preparation of Data for Calibration and Alignment:**
  - use availability of data
  - use high statistics
    - ➔ central collection (gathering) of data
    - ➔ feedback system for database constants

- **Event Data Re-processing in Shutdown Periods:**
  - use vast processing power of the farm
    - ➔ exploit online infrastructure

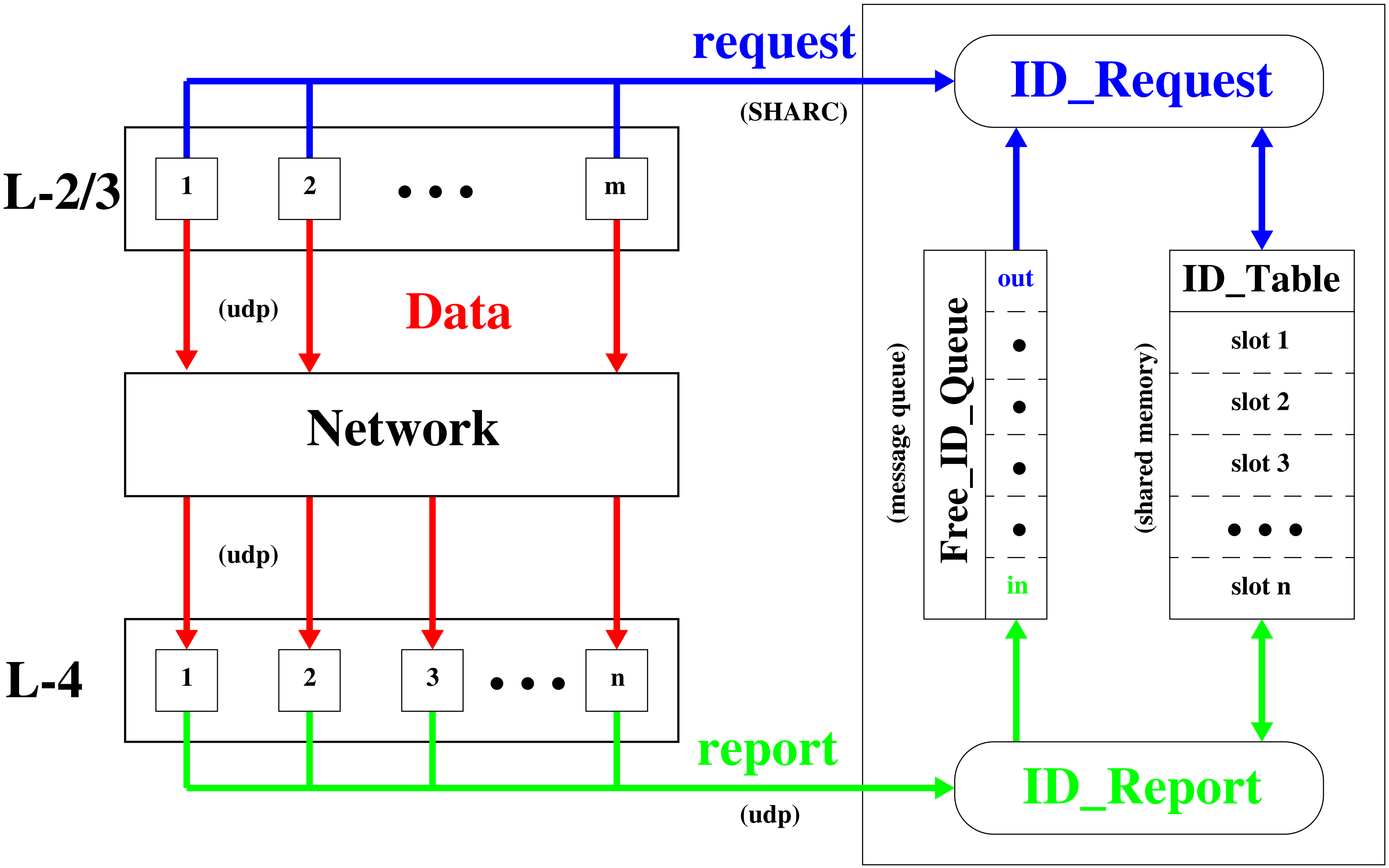
Implementation: Farm Nodes

- **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.)

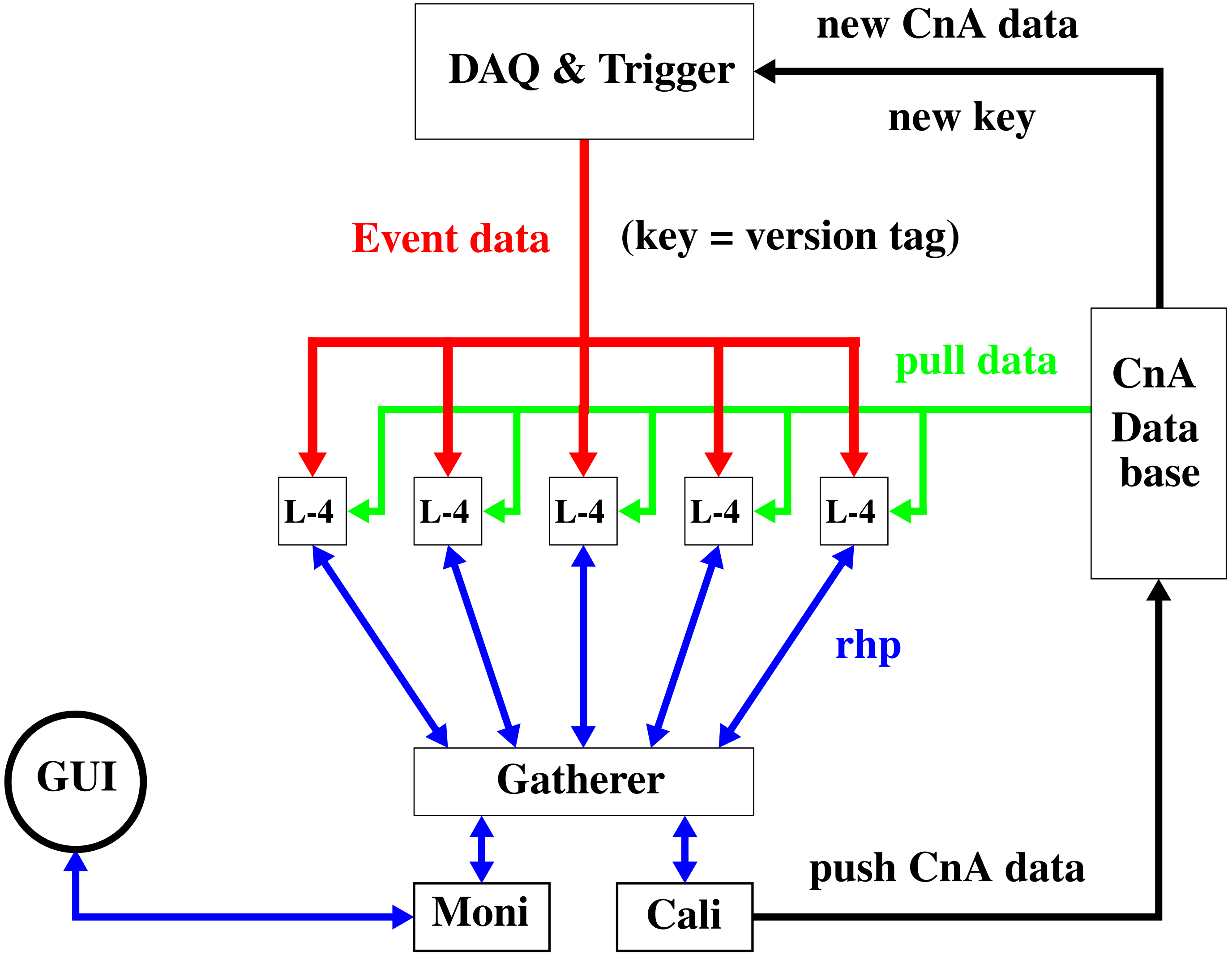
Implementation: Farm Network



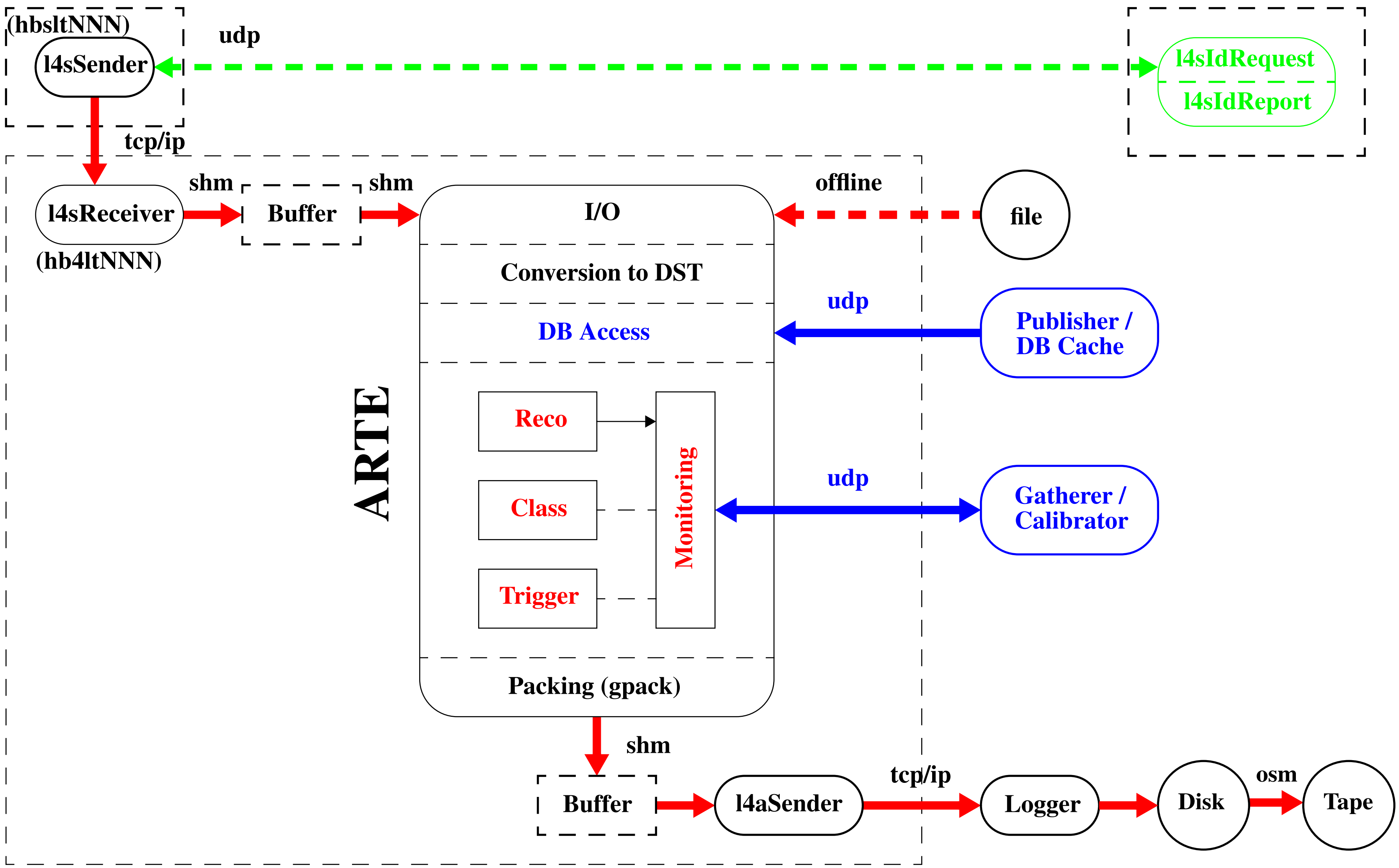
Implementation: Farm Event Control



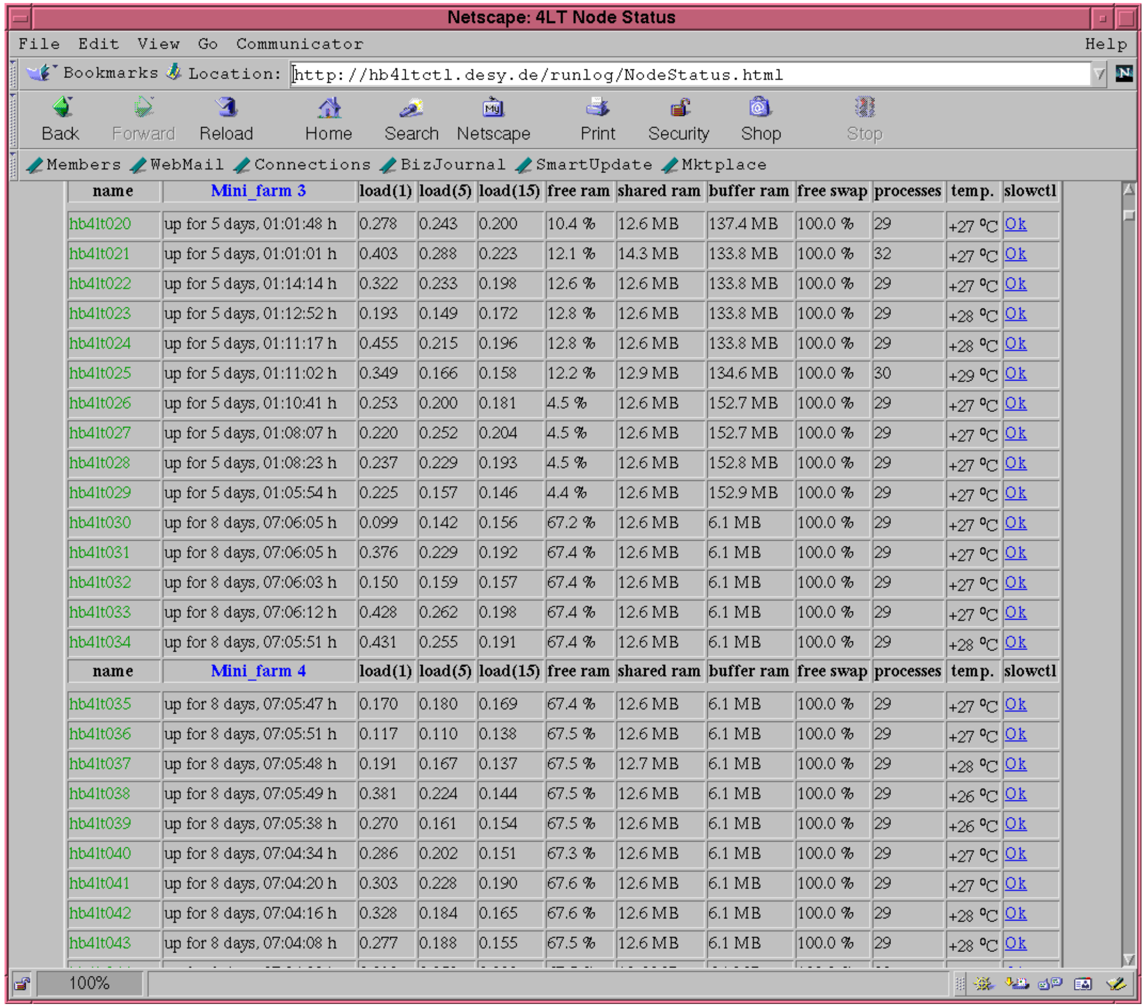
Implementation: Calibration & Alignment



Implementation: Node Processes



Implementation: Farm Slow Control



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)

Experiences

- **Concept:**
  Using of-the-shelf components allowed to build a scalable and flexible system which came well below the estimated costs.
- **Farm nodes:**
  PCs running Linux over sufficient computing power. Bottlenecks in the timing are due to the algorithmic part of the reconstruction.
- **Network:**
  Fast-Ethernet and Gigabit-Ethernet uplinks are standards.
- **Event control:**
  The tcp/ip and IPC-based package were easy to implement and exceeded bandwidth/rate requirements considerably.
- **Calibration and alignment:**
  The framework is set up but needs further developments.
- **Node processes:**
  The modular structure helped to commission the system.
- **Slow control:**
  A simple web-based scheme ran successfully. Integration into the common HERA-B scheme is pending.

Summary

- 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.