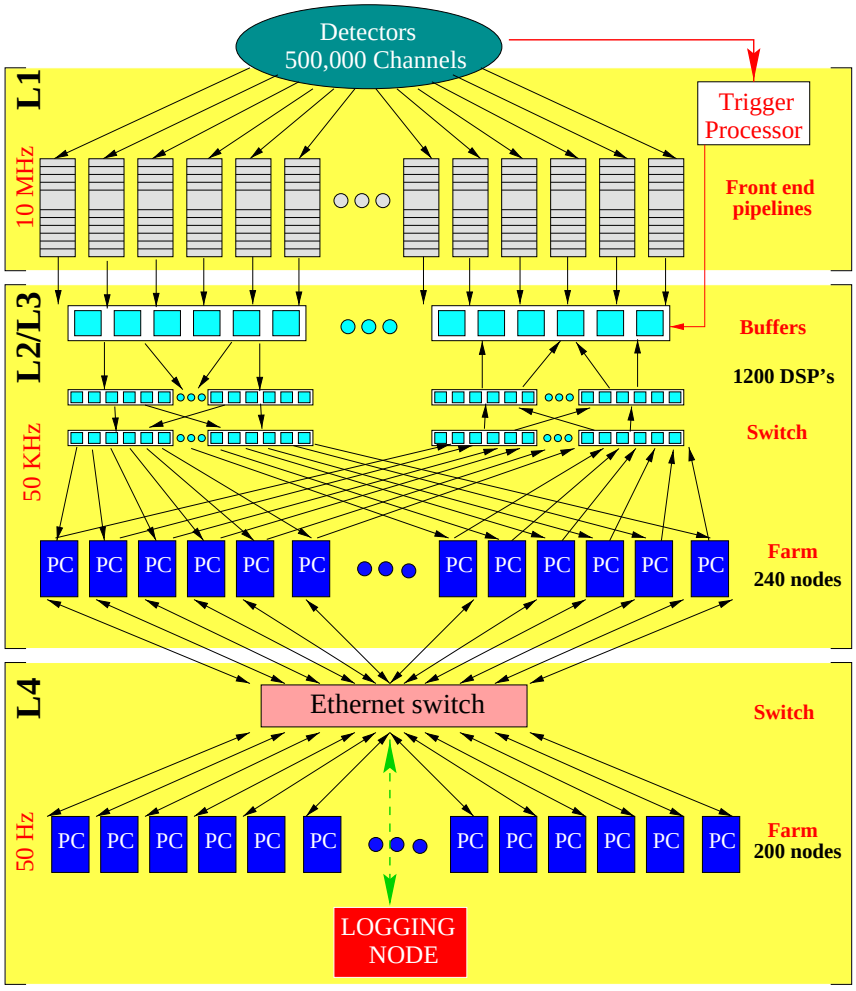
Farming in HERA-B

José Hernández
DESY

for the HERA-B Collaboration

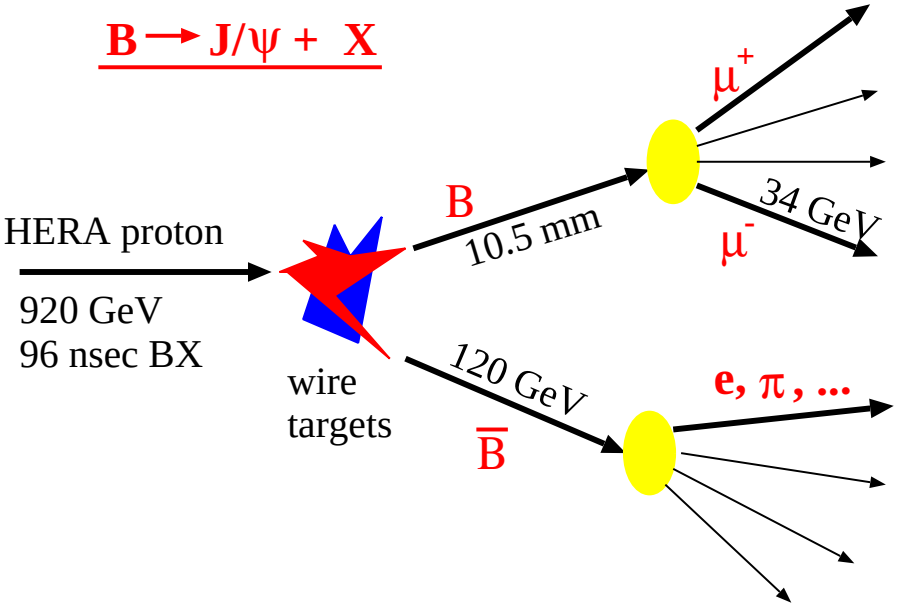
Outline

- The HERA-B DAQ and Trigger
- The Trigger Farm
- The Online reconstruction Farm



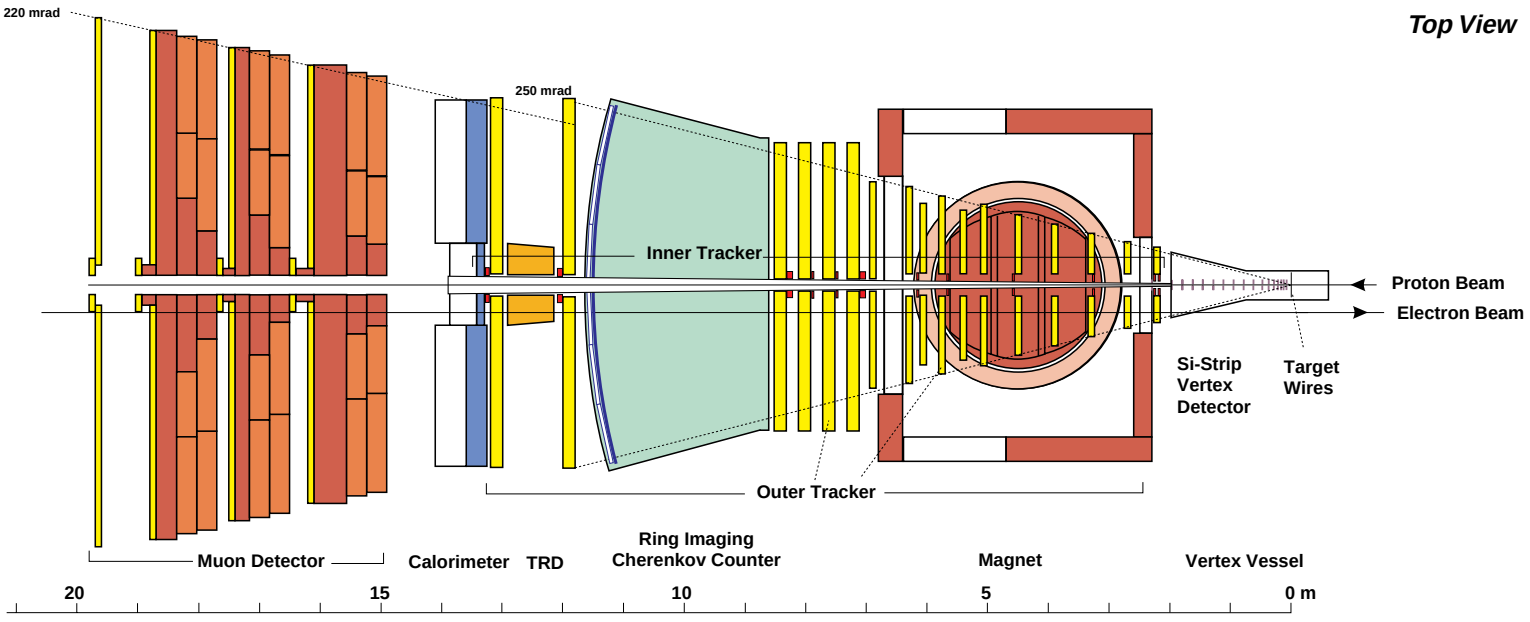
The HERA-B experiment

Fixed Target Hadronic b-factory at HERA



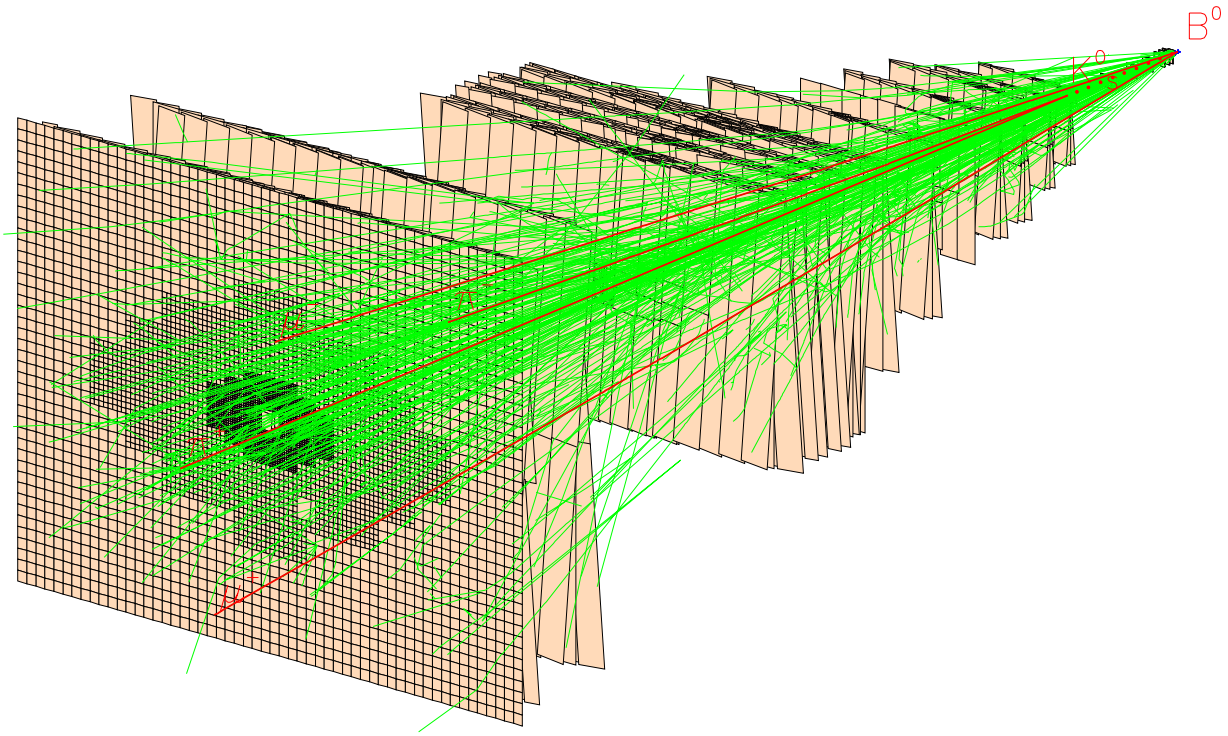
Physics goals:

Measure CP violation in $B^0 \rightarrow J/\psi K_S^0$ and other channels, B_S oscillations, Rare B decays, ...



The Operating Environment

- Signal/Background: $\sim 10^{-9} - 10^{-11}$
- Bunch Crossing: 95 nsec $\sim$ 10 MHz
- Average Number of interactions per BX: 4 - 5
- Rate of interesting signals: $<<$ 1Hz
- Mean track multiplicity per BX: $\sim$ 150
- Detector Occupancy: $\sim$ 20%



The Trigger

- **Requirement:**
  - Select $<<$ 1 Hz of physics from 10 MHz of multi-interaction events
- **Strategy:**
  - High B mass $\Rightarrow$ High p_T decay particles
  - J/Ψ decay product $\Rightarrow$ pair of unlike charge track with J/Ψ mass
  - Long B lifetime $\Rightarrow$ detached vertices
- **Implementation:** Multi-level trigger
  - Pretrigger: Find seeds for Regions of Interest ($e^{\pm}$, $\mu^{\pm}$, High-p_T $K^{\pm}$ or $\pi^{\pm}$)
  - Level 1: Hardware trigger. Simple track finding and J/Ψ di-lepton mass
  - Level 2: Track refit, magnet tracking, vertex finding
  - Level 3: Event building. Particle ID and tracking outside ROI's
  - Level 4: Full online event reconstruction, event classification and selectio

The Scale

- channel count: 550,000
- event size (uncompressed): 500,000 bytes
- event size (compressed): 120,000 bytes

Level	Latency	Input	
		Trigger	Data
	sec	Hz	byte/s
Pre	10^{-6}	$10 \cdot 10^{6}$	$90 \cdot 10^{9}$
1	$12 \cdot 10^{-6}$	$10 \cdot 10^{6}$	$10 \cdot 10^{9}$
2	$7 \cdot 10^{-3}$	$50 \cdot 10^{3}$	$250 \cdot 10^{6}$
3	$100 \cdot 10^{-3}$	$500 \cdot 10^{0}$	$250 \cdot 10^{6}$
4	$4 \cdot 10^{0}$	$50 \cdot 10^{0}$	$6 \cdot 10^{6}$

Data flow (bytes/s)	
$\rightarrow$L1 pipe	$5 \cdot 10^{12}$
$\rightarrow$L2 buffers	$25 \cdot 10^{9}$
$\rightarrow$L3 processors	$250 \cdot 10^{6}$
$\rightarrow$L4 processors	$6 \cdot 10^{6}$
$\rightarrow$tape	$2.4 \cdot 10^{6}$

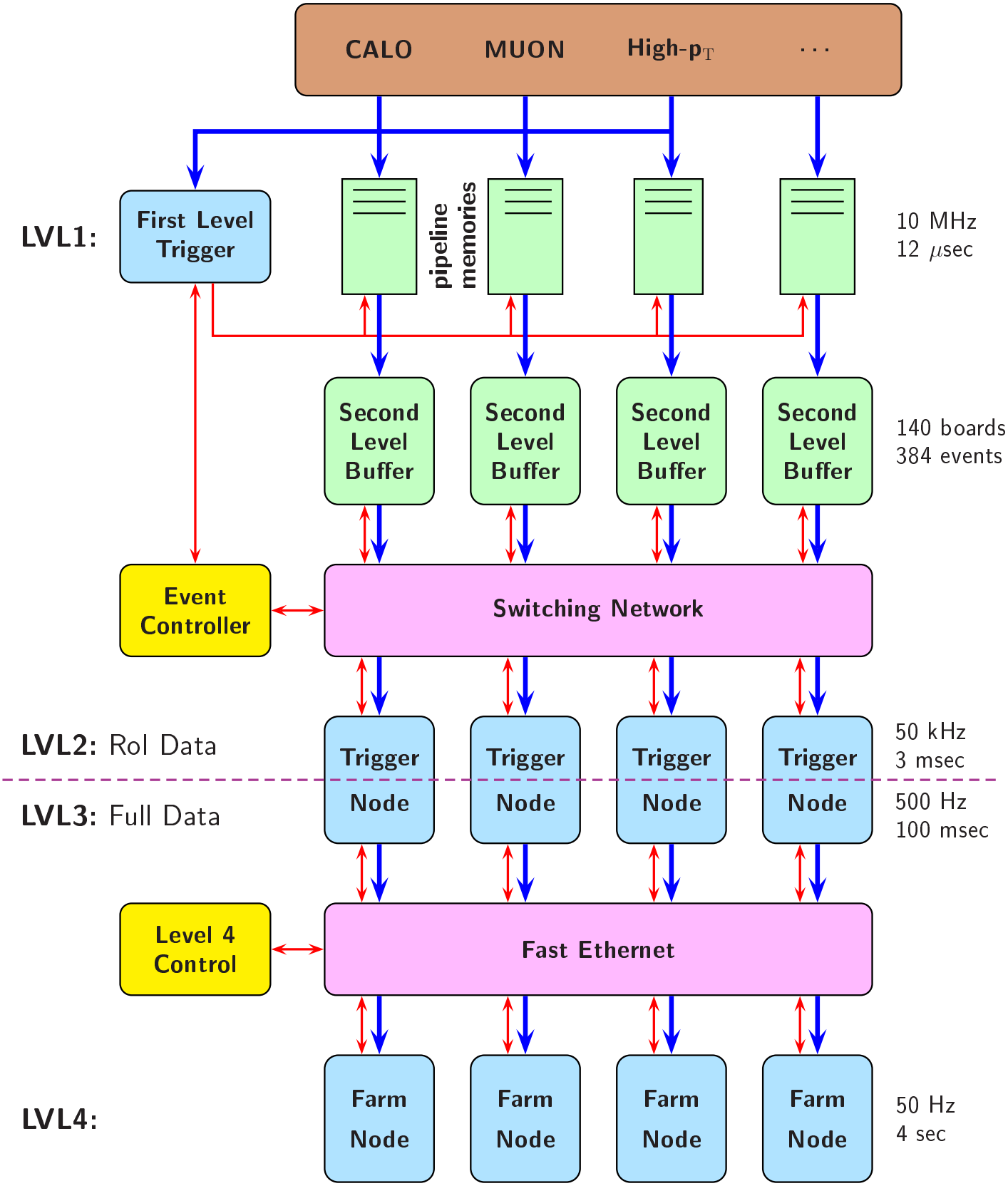
- **Level 2:** Latency and bandwidth dominated

  $\Rightarrow$ ROI based (event fraction O(1%))
- **Level 3:** Still latency and bandwidth dominated

  Full event available (event building)
- **Level 4:** Processing time dominated

  Full event reconstruction

DAQ architecture



Data Path Components

SHARC cluster board

- Level 2 Buffers: 140 boards
- Fast Switch L2 buffers $\leftrightarrow$ L2 processors: 48 boards
- Event Control: 1 board

FARM processors

- L2 Trigger FARM: 240 Pentium II 400 MHz
- Online Reconstruction FARM: 200 Pentium III 500 MHz (100 dual CPU PCs)

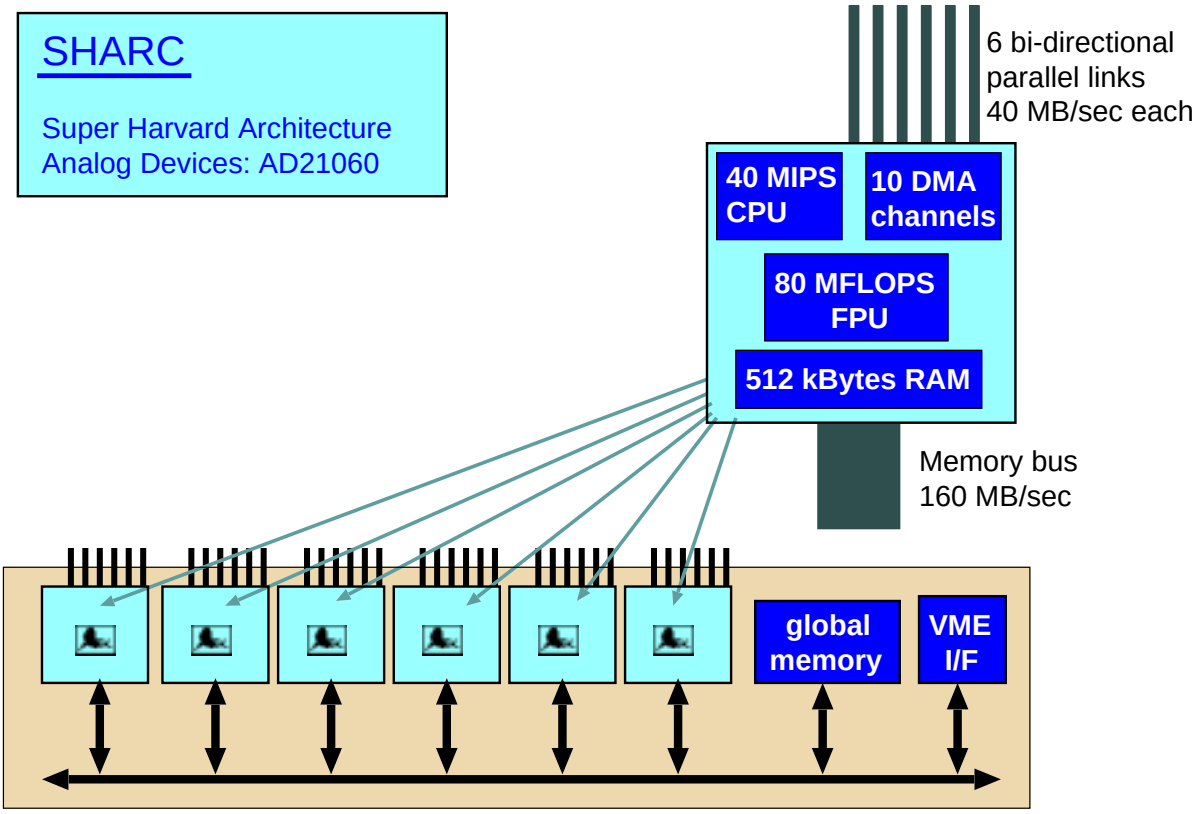
SHARC/PCI interface card

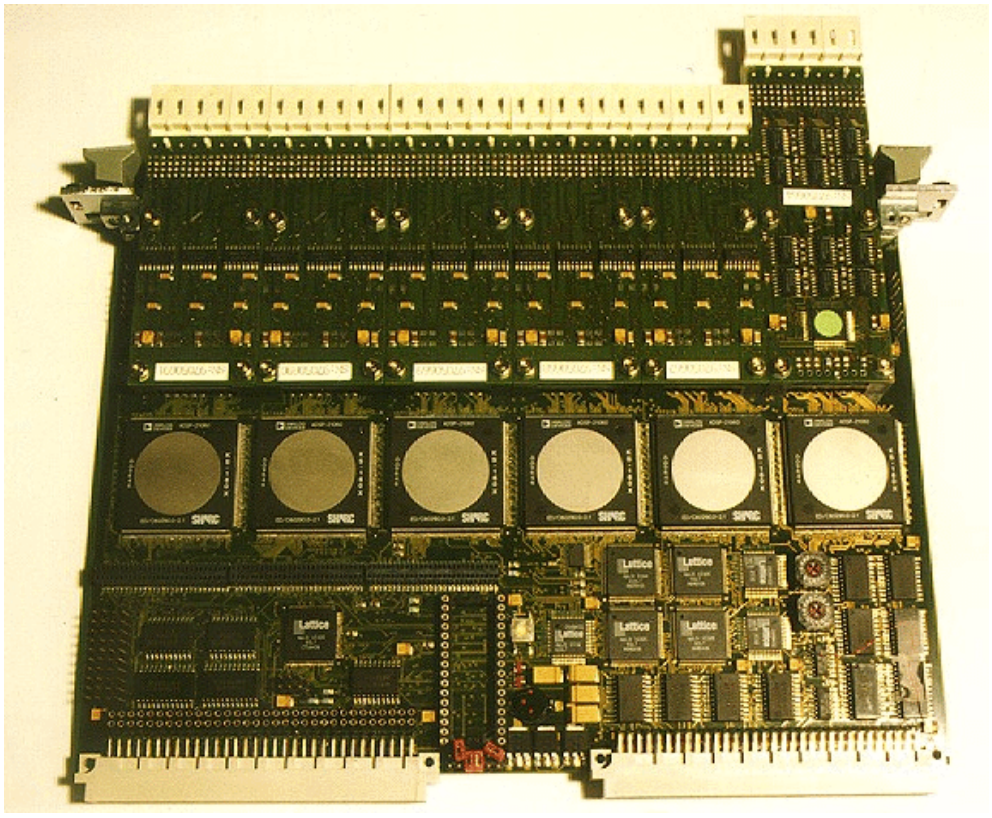
- Access to L2/L3 Trigger processors. 240 units

Fast/Gibabit Ethernet

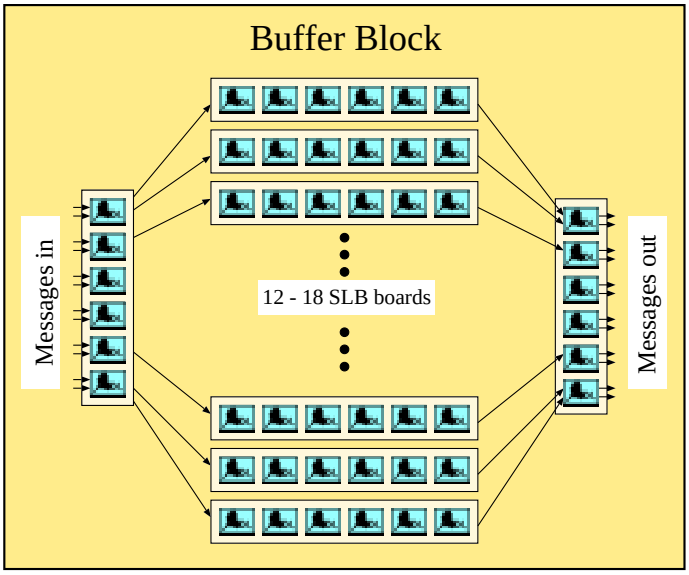
- L3 $\rightarrow$ L4 connection (FE)
- L4 $\rightarrow$ Logger connection (GE)
- Control

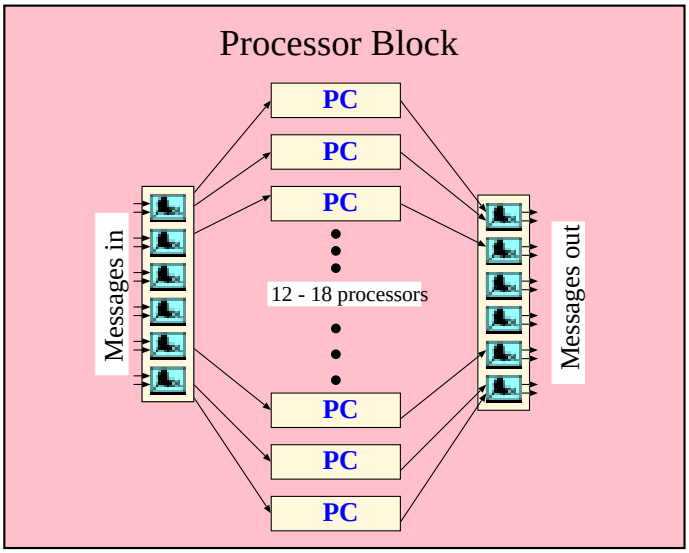
SHARC board

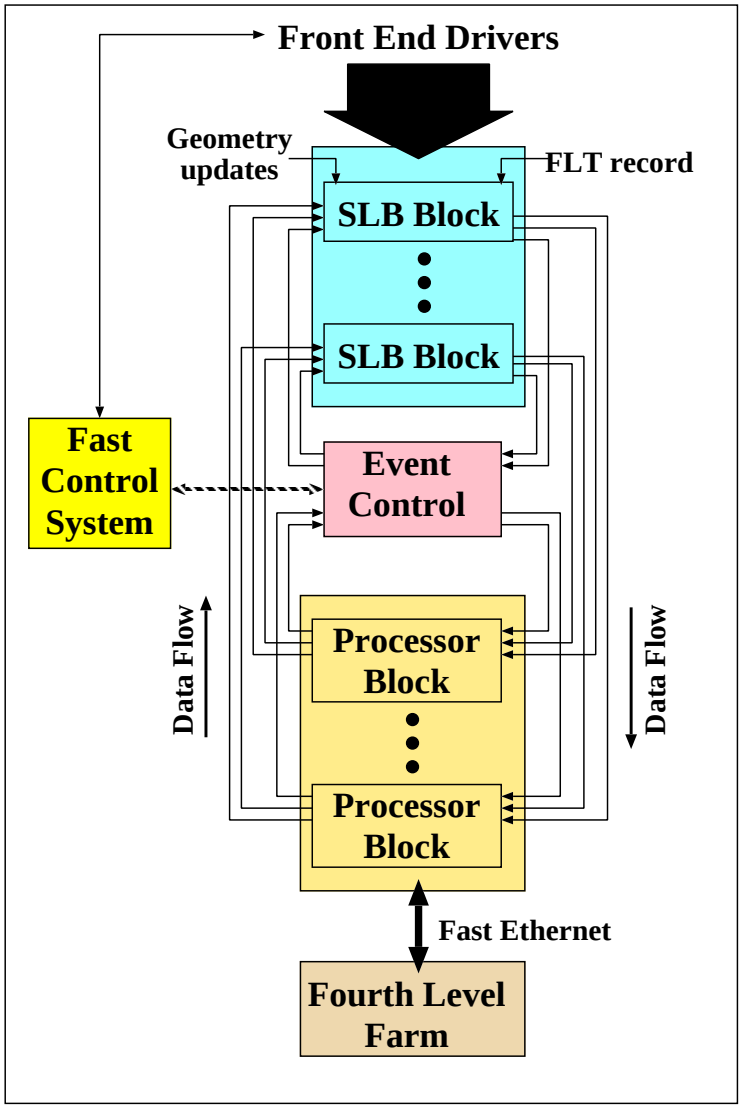

SHARC
Super Harvard Architecture
Analog Devices: AD21060
6 bi-directional
parallel links
40 MB/sec each
40 MIPS
CPU
10 DMA
channels
80 MFLOPS
FPU
512 kBytes RAM
Memory bus
160 MB/sec
global
memory
VME
I/F



DAQ Topology





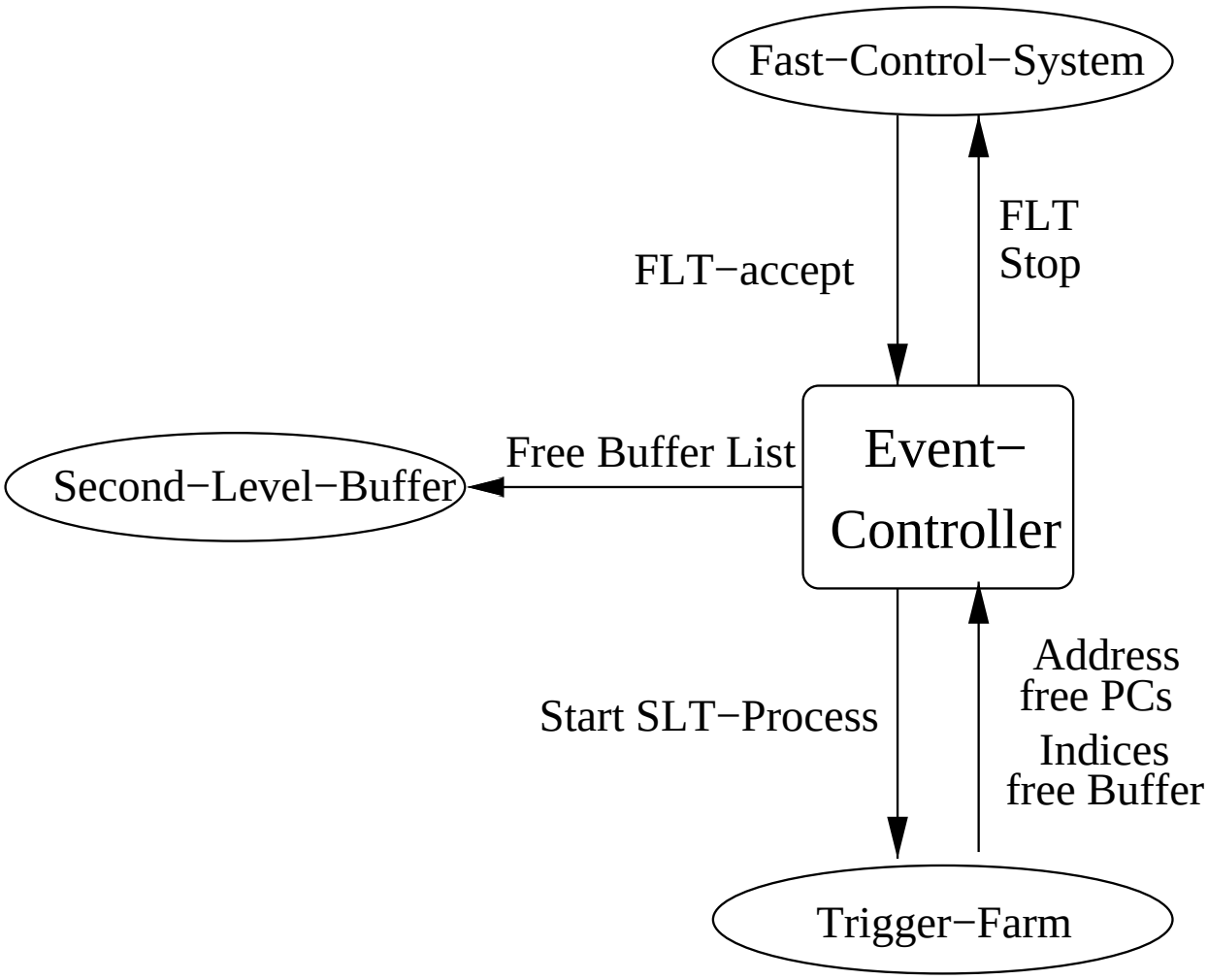


- Full connectivity buffers ⇔ processors
- Independent request and reply lines
- Measured switch performance:

  Bandwitdh: 1 GByte/sec (500 MB/s needed by L2/L3 triggers)

  Message rate: 2.6 MHz (1.8 MHz needed by L2 trigger)

SLB ⇒ SLT data transfer



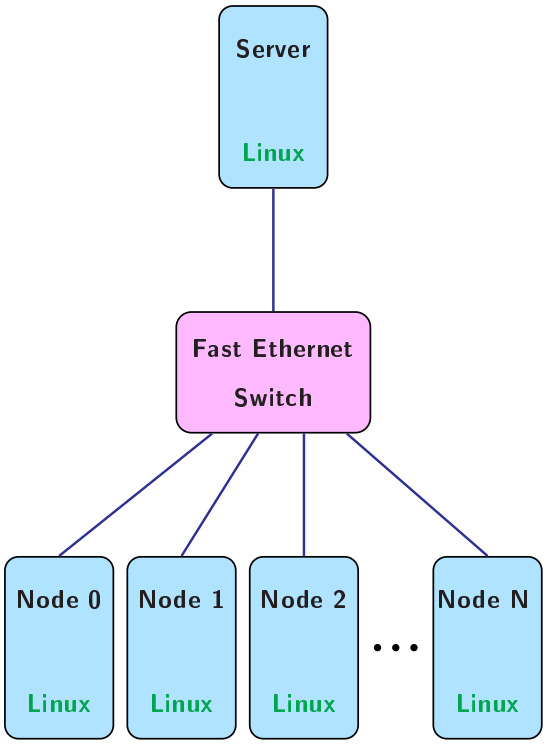
- **Event Controller:**
  - Controls data transfer SLBs ⇒ SLTs
  - Gets FLT accept via the FCS
  - Maintain list of free SLBs and free SLPs
- The ROI data is pulled from the SLBs
- High bandwidth and high transaction rate SHARC switch

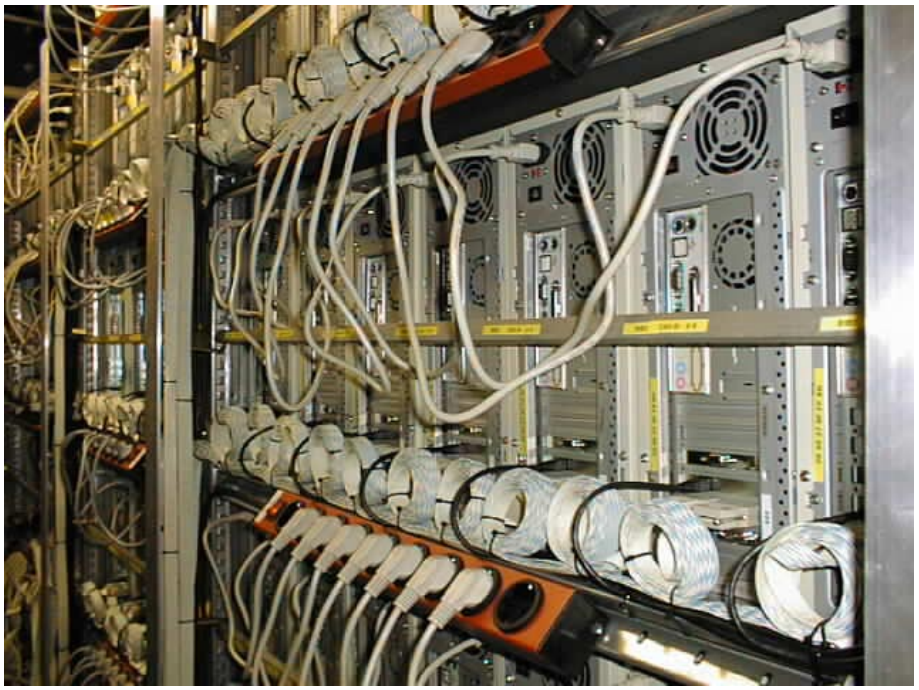
L2/L3 Node

Standard PC equiped with:

- **Off the shelf components:**
  - Pentium II 300 MHz + Pentium III 400 MHz processors
  - NO hard disk (diskless nodes)
  - 64 MB SDRAM
  - Floppy drive and cheap graphics card
  - 100 Mbit/sec Fast Ethernet Adapter
- **Custom made components:**
  - Sharc-to-PCI interface (adapter and driver)
    - Throughput $\sim$ 40 Mbyte/sec
    - Latency $< 1\mu$sec
  - Slow control interface (CAN protocol)
    - Remote power up/down and reset
    - Measure temperature, voltage, fan speed

L2/L3 Farm

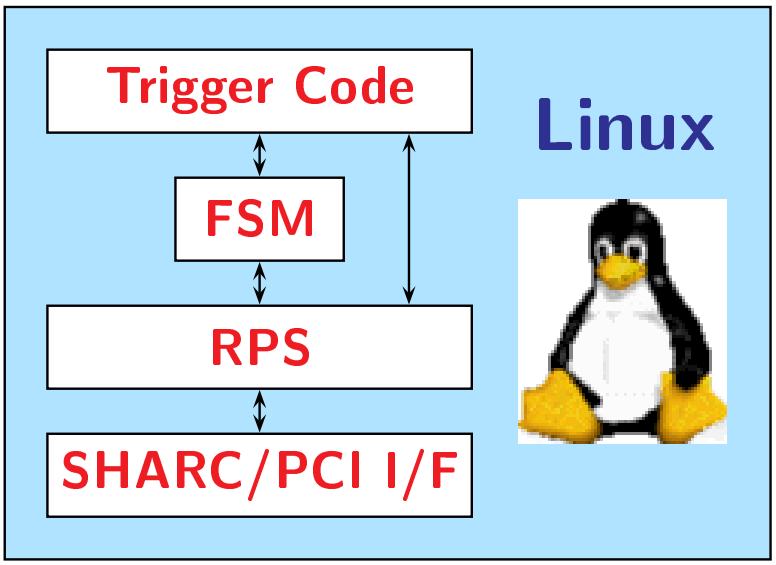




Booting and control:

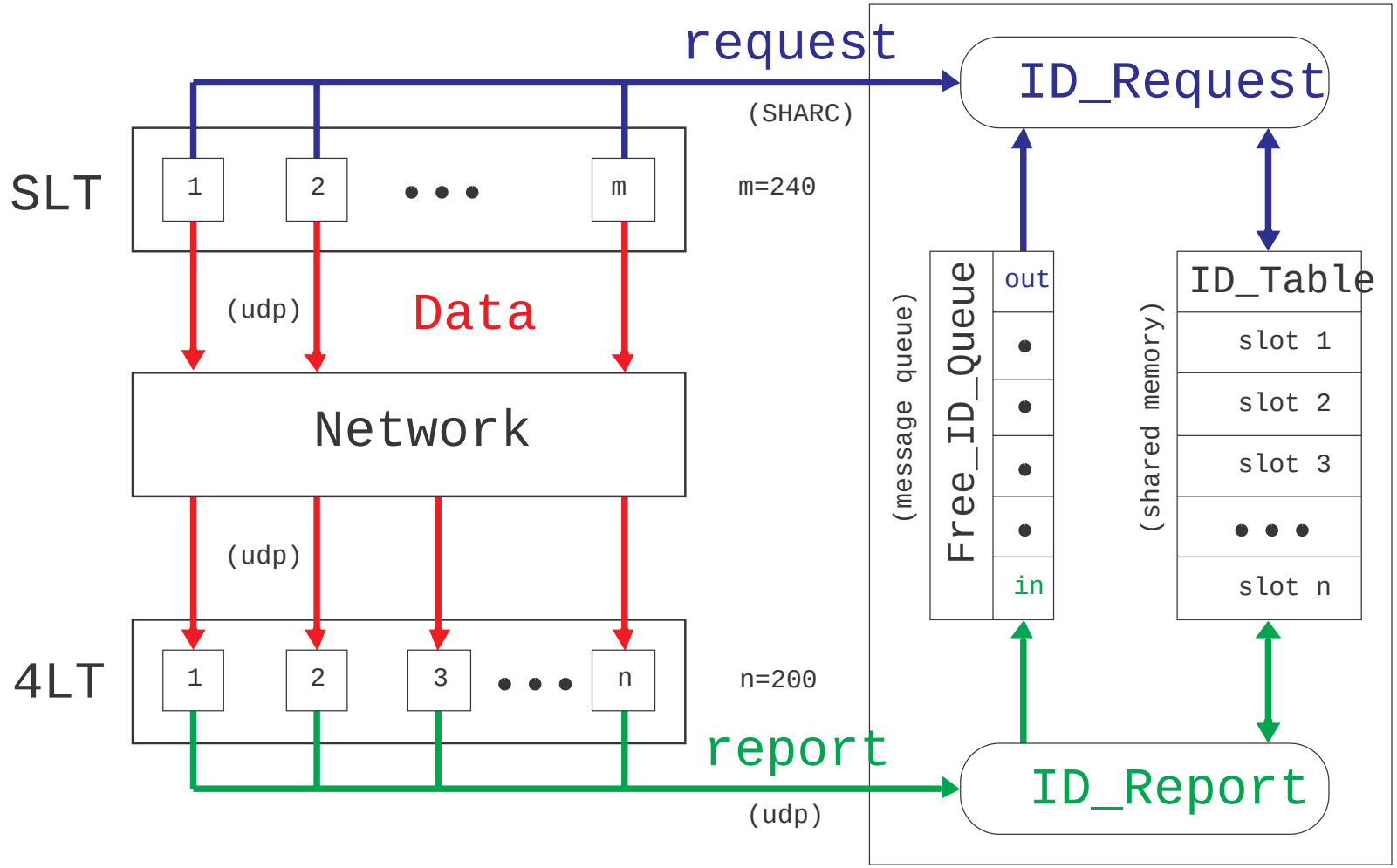
- Linux boot/root from floppy
- IP address from server via bootp
- Executables from server (NFS mount)
- Remote process control via local server process

Environment



- Linux operating system
  Pruned down system: boot from floppy
- Only essential processes run.
  Keep only trigger process active.
  $\Rightarrow$ avoid context switching!
- Message driven trigger: polling protocol.
- Non-standard I/O driver copies data directly between sharc link i/f and user space (overhead $< 1\mu$sec)
- FSM: finite state machine for run control
- RPS: routing service for sharc link.

L2/L3 → L4 Farm Control



L4 Farm Tasks

- **Full Online Event Reconstruction:**
  - Allow immediate physics analysis
  - Avoid relatively slow access to data on tape (20 TB/year)
  - Allow Online Data Quality Monitoring
- **Online Event Classification and Level 4 Event Selection:**
  - Mark events according to physics categories (event directories)
  - L4 trigger step
- **Data Logging:**
  - Add reconstruction info to event and send to logger
- **Online Data Quality Monitoring:**
  - Central gathering of DQM data. High statistics.
- **Produce Data for Online Calibration and Aligment:**
  - Online Reconstruction Requires Online Calibration and Alignment
- **Offline Event Data Reprocessing:**
  - Use data logging protocols for feeding the farm with events and collecting the

Hardware:

- 100 Dual Pentium III 450 MHz
- 256 MB SDRAM
- 100 Mbit/sec Fast Ethernet Adapter
- Network: Fast Ethernet switches with Gigabit Ethernet uplink

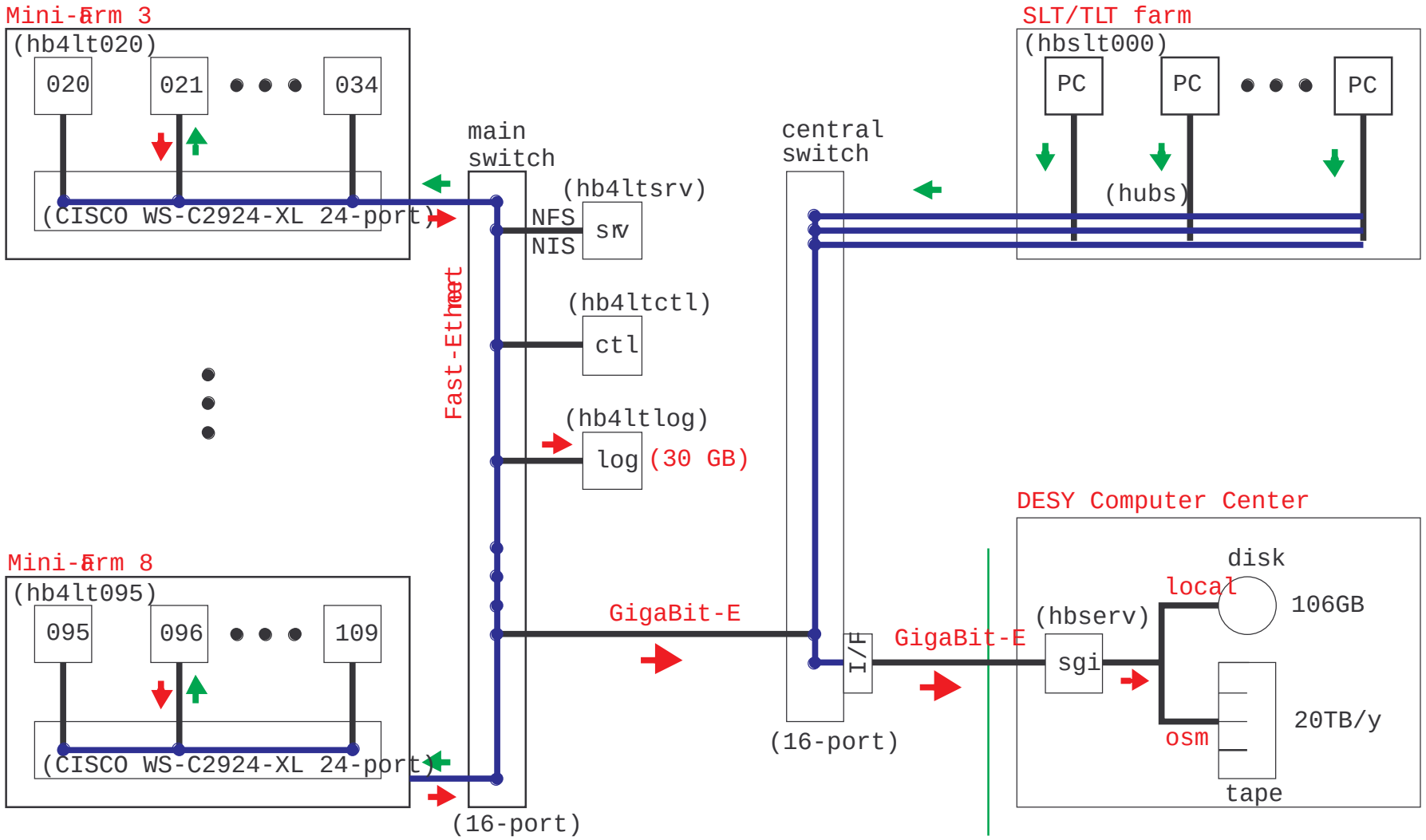
Software Environment:

- Linux Multiprocess environment.
  No real time needs
- Frame Program: ARTE
  - FORTRAN, C and C++ code
  - Data I/O (offline file-based, online memory-based)
  - Event reconstruction, classification, selection
  - Monte Carlo generation and detector simulation
- Same software used Online and Offline

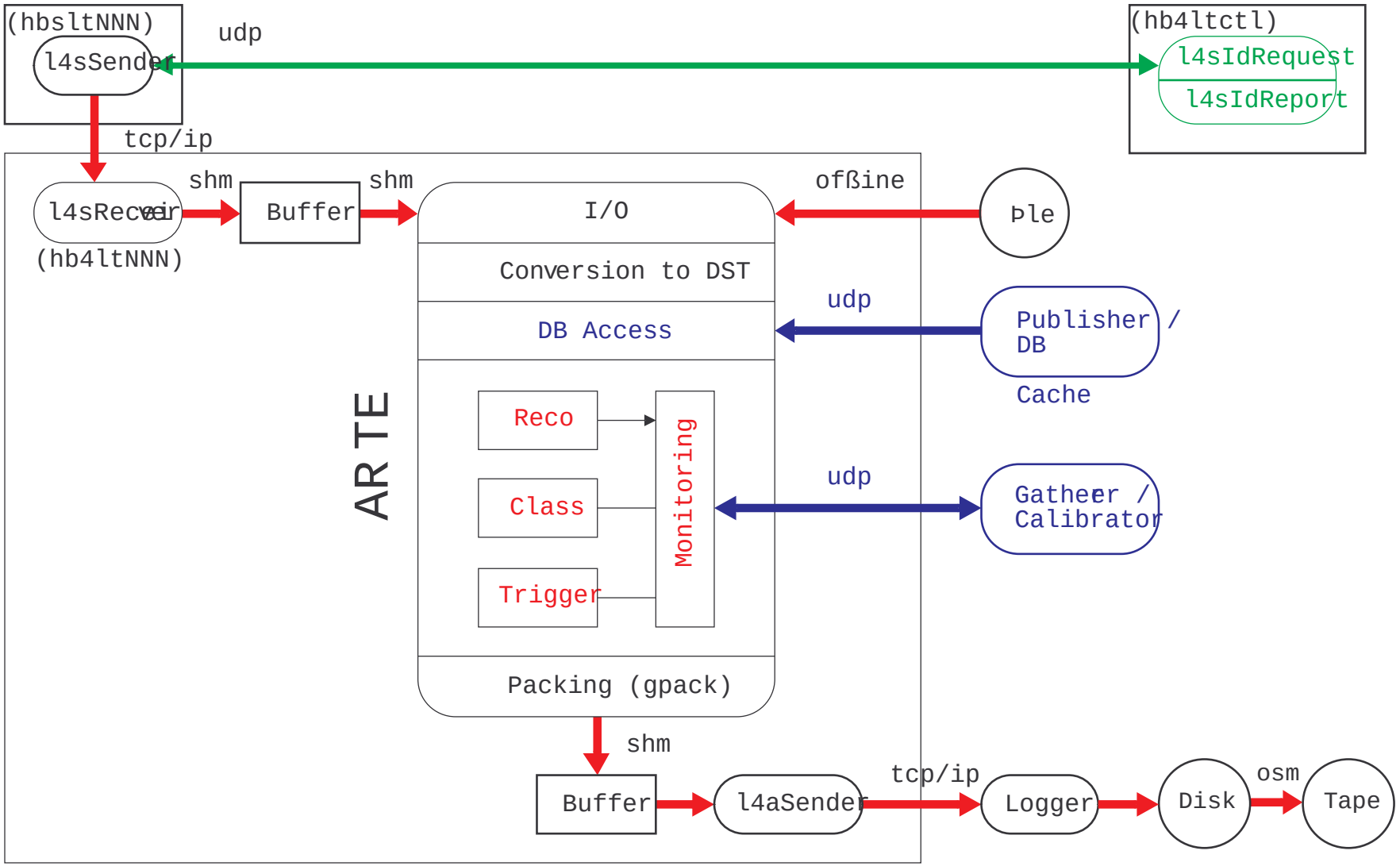
Performance:

- Online Reconstruction:
  - 4 sec/event @ 5 MHz (design 4 sec/event @ 40 MHz)
- Event Data Transfer, logging and archiving:
  - L2 $\rightarrow$ L4: up to 12 MB/sec (design 6 MB/sec)
  - L4 $\rightarrow$ Logger: up to 12 MB/sec (design 2.5 MB/sec)
  - Logger $\rightarrow$ tape: 5 MB/sec. 7 TByte this year

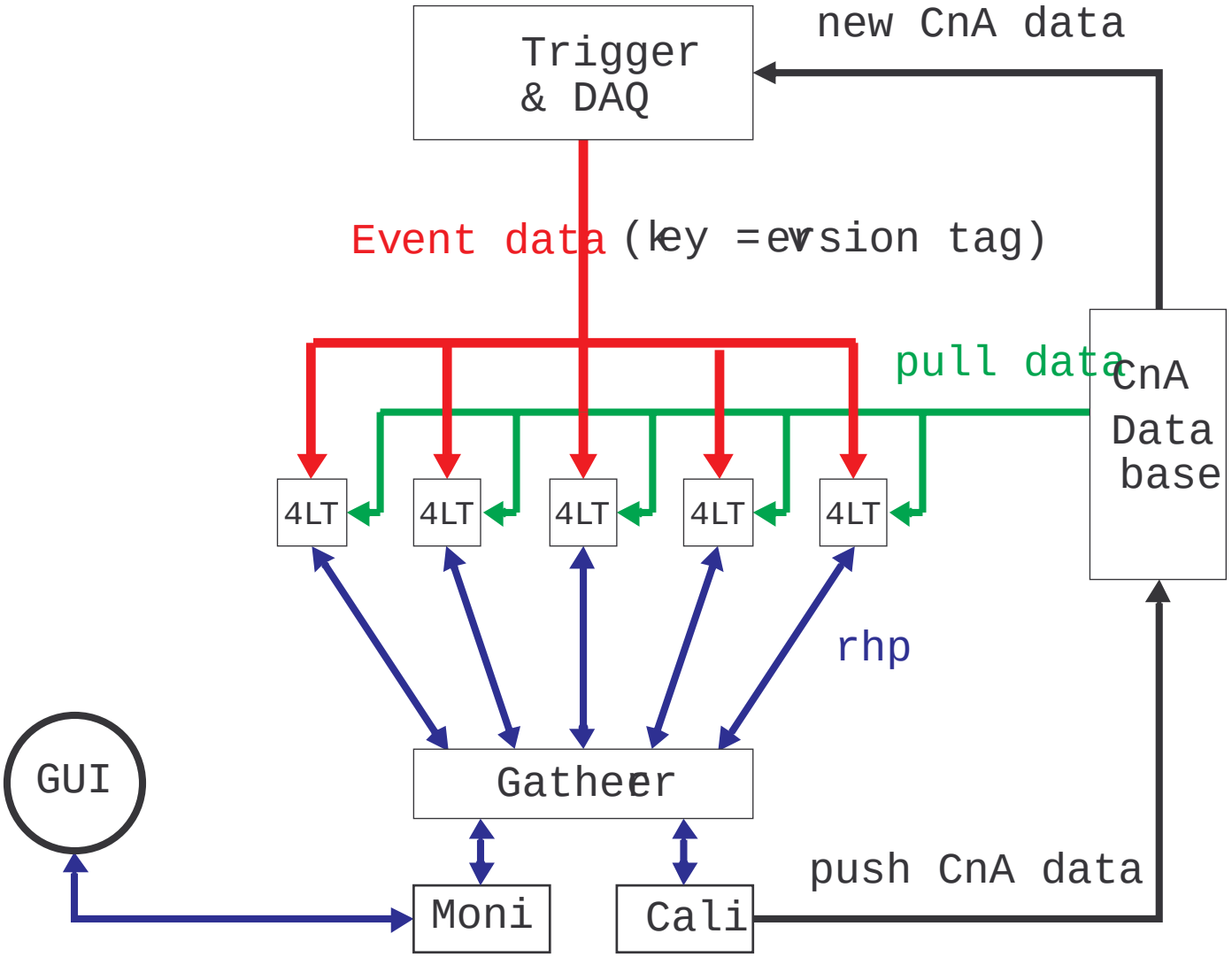
Farm Network



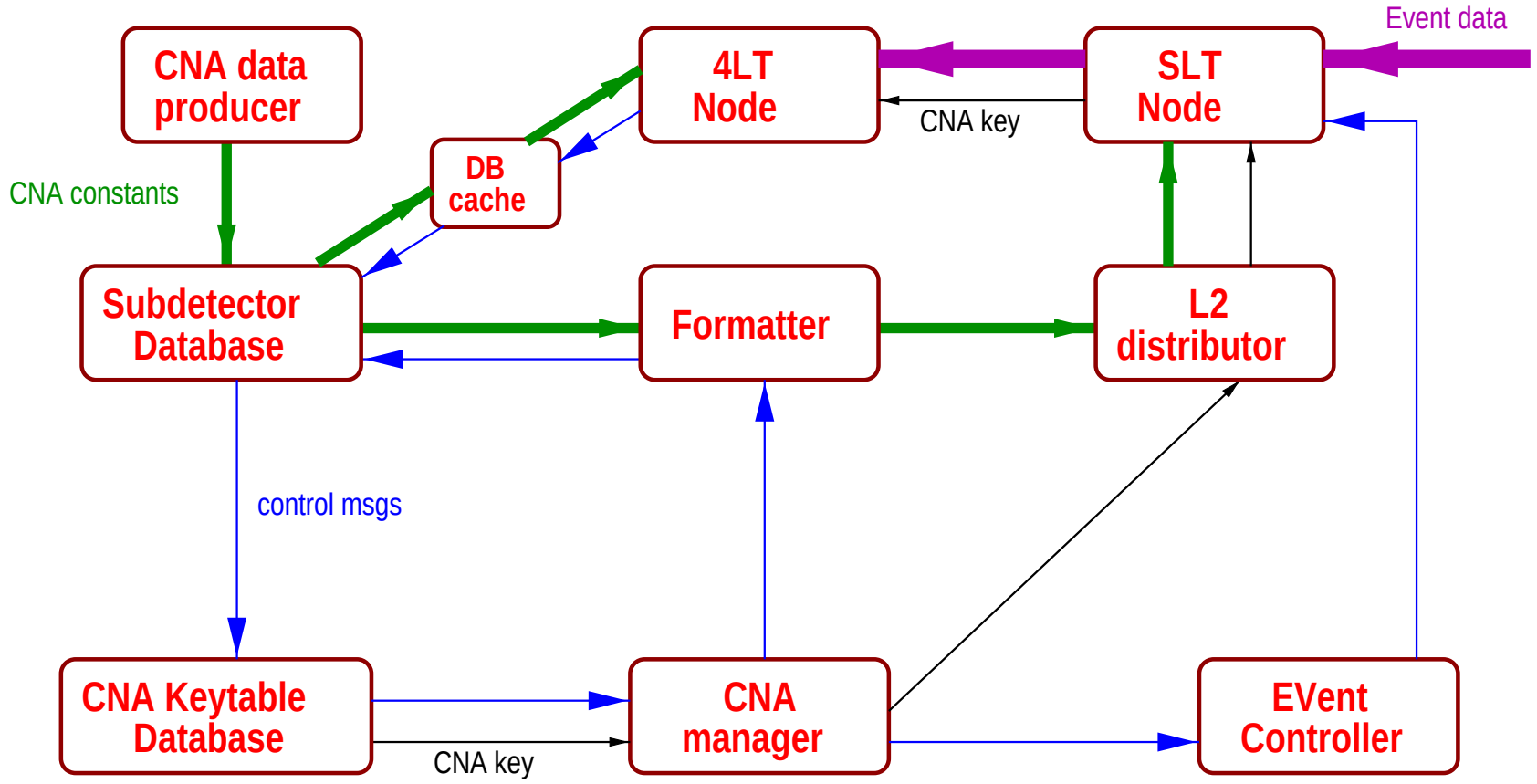
L4 Node Processes



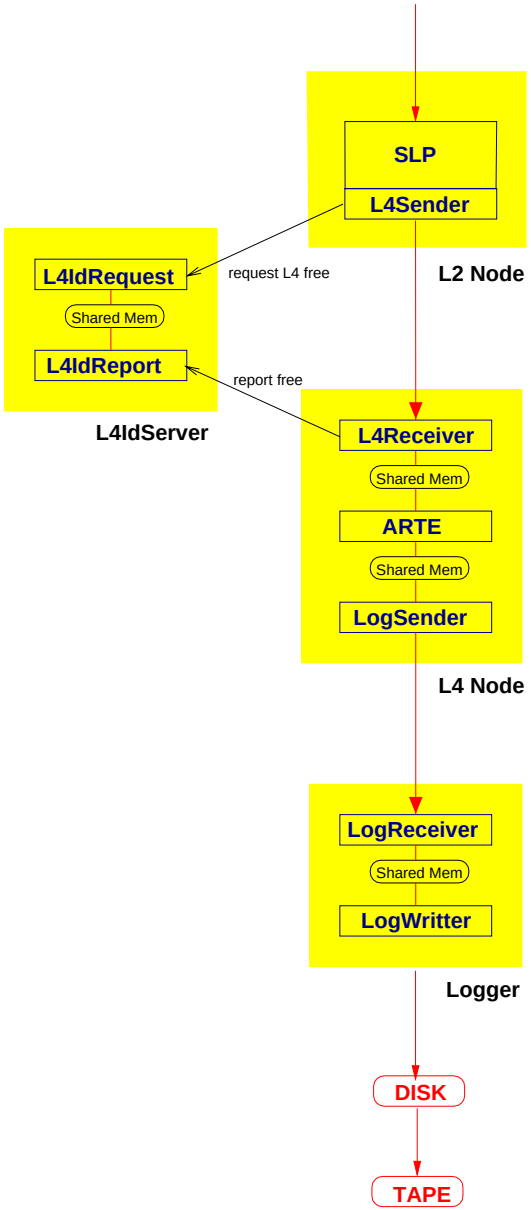
L4 Calibration and Alignment and DQM

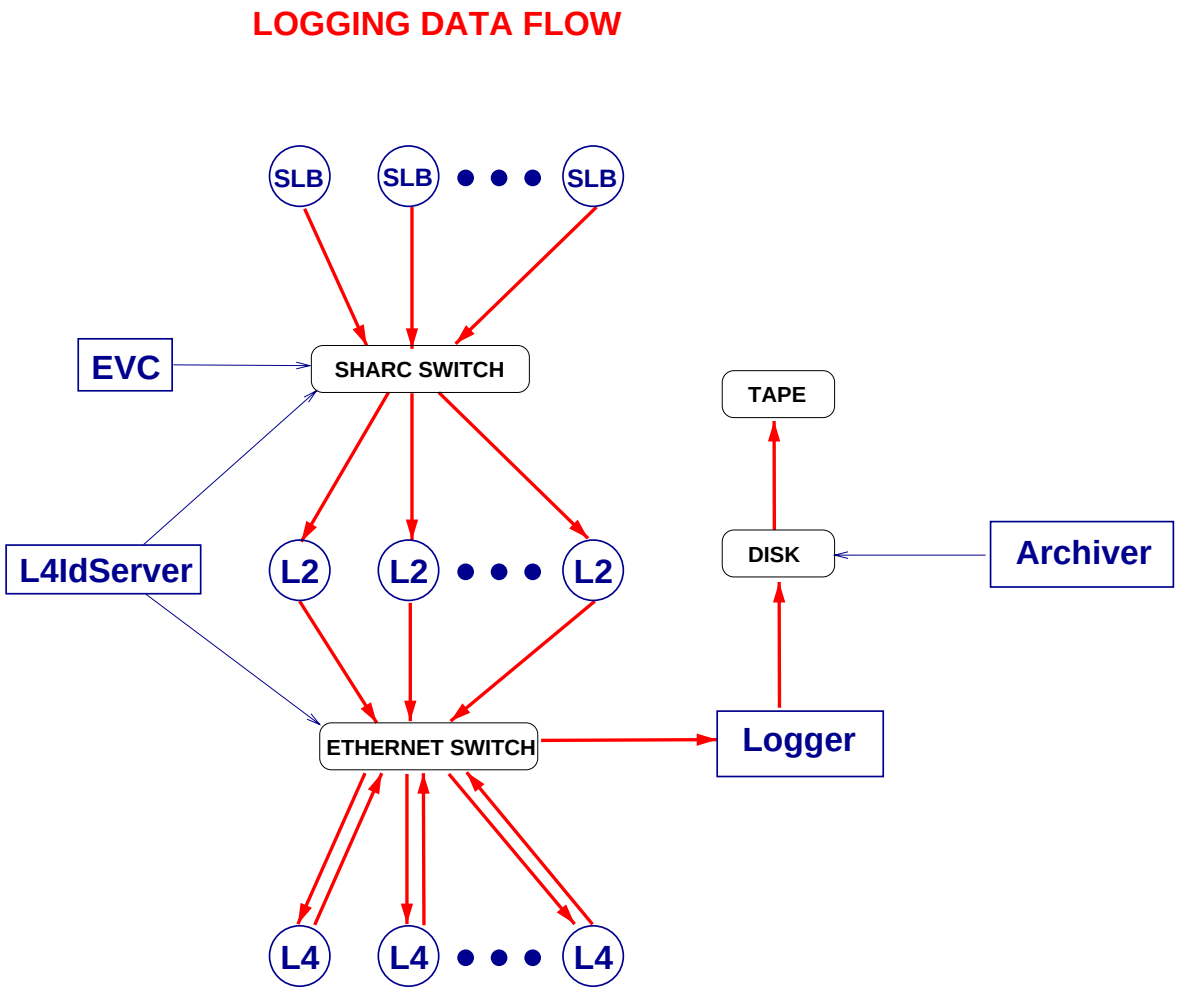


Online Calibration and Alignment

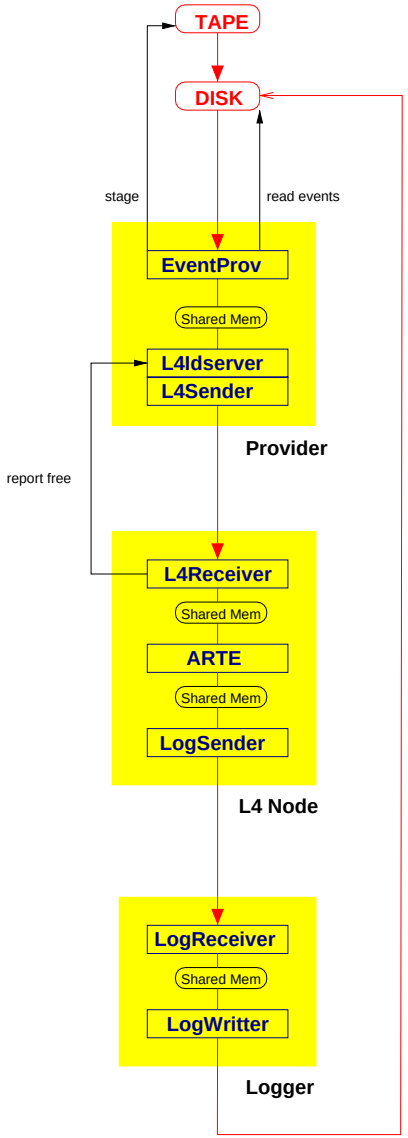


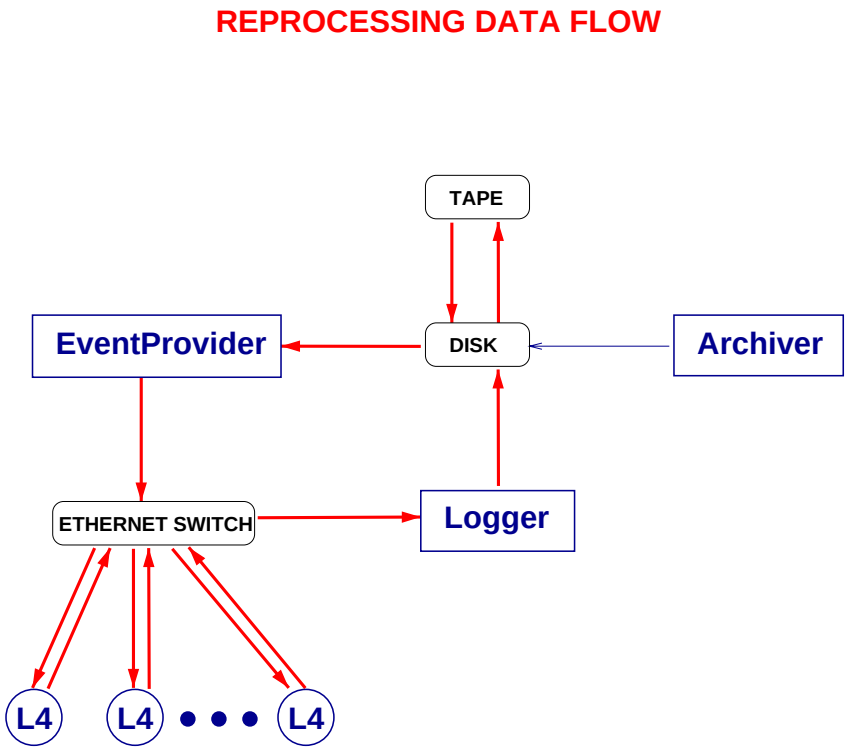
L4 Data Logging and Archiving





L4 Data Reprocessing





Conclusions

- HERA-B has successfully implemented trigger and online reconstruction F farms in the DAQ
- The trigger farm is connected to the previous trigger level via a high bandwic and high transaction rate DSP switch
- Both farms are connected via Fast/Gigabit ethernet switched network
- The online reconstruction farm allows immediate data analysis and high qual data quality monitoring
- Online calibration and alignment is required for online reconstruction
- PC farms running Linux provide a flexible, scalable and low cost solution in HE detectors