

Data Acquisition and Analysis at Belle – past, now and future –

Ryosuke Itoh
KEK

for Belle DAQ, Computing and DST subgroups

CHEP01, Beijing, September 4, 2001

Outline

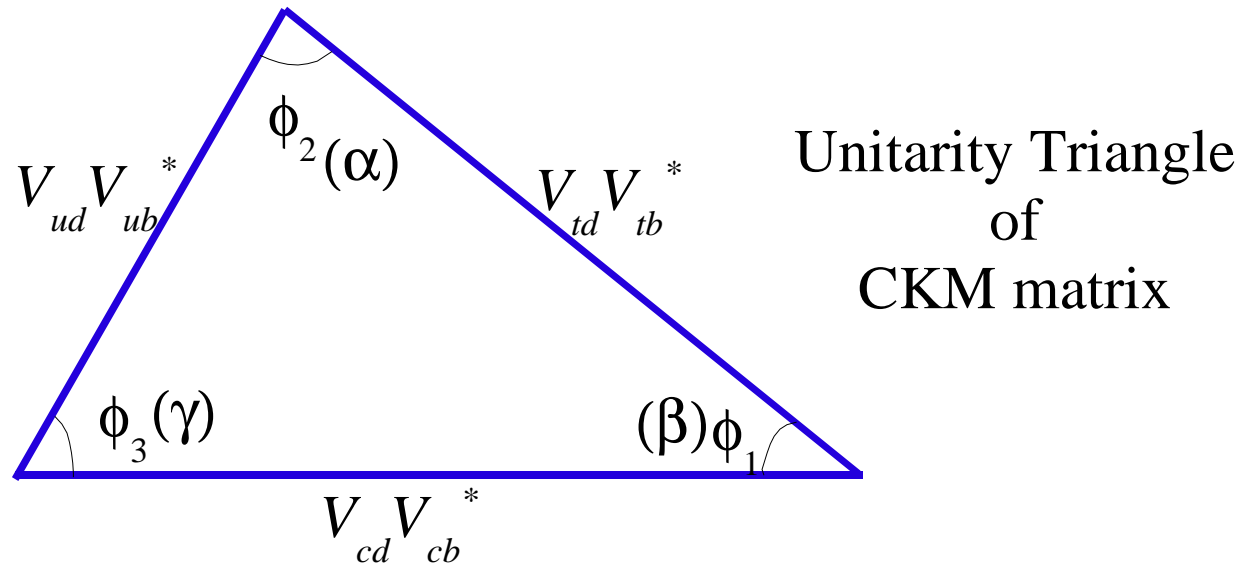


1. Introduction
2. Requirements for DAQ and Analysis
3. Data Acquisition
4. Data Analysis
5. Future of Belle
6. Conclusion

1. Introduction

Belle : **B**–factory experiment at KEK

Purpose : study of CP violation in B meson system



CP violation \equiv non-zero angle

CP Asymmetry :

$$A(\Delta t) = \frac{[\Gamma(B_d^0 \rightarrow J/\psi K_s) - \Gamma(\overline{B}_d^0 \rightarrow J/\psi K_s)]}{[\Gamma(B_d^0 \rightarrow J/\psi K_s) + \Gamma(\overline{B}_d^0 \rightarrow J/\psi K_s)]} = \sin 2\phi_1 \sin \Delta m \Delta t$$

Δm : mass difference between 2 B^0 mass eigenstates (B_1 and B_2)

$\Delta t = t(B^0 \text{ decay}) - t(\overline{B}^0 \text{ decay})$

Requirements to experiment

1. Fully reconstructed CP decays of B mesons with a high statistics
 $\text{Br}(B \rightarrow \text{CP decay}) \sim 10^{-4-5} \rightarrow \text{needs } 10^{7-8} B\overline{B} \text{ events}$
2. Need to measure decay time difference of B^0 and \overline{B}^0
Time integrated asymmetry becomes 0.
 \rightarrow Asymmetric Collision

• How to measure the decay time of B^0 ?

– At B factory, B^0 is produced from
$$e^+ e^- \rightarrow Y(4S) \rightarrow B^0 \bar{B}^0$$

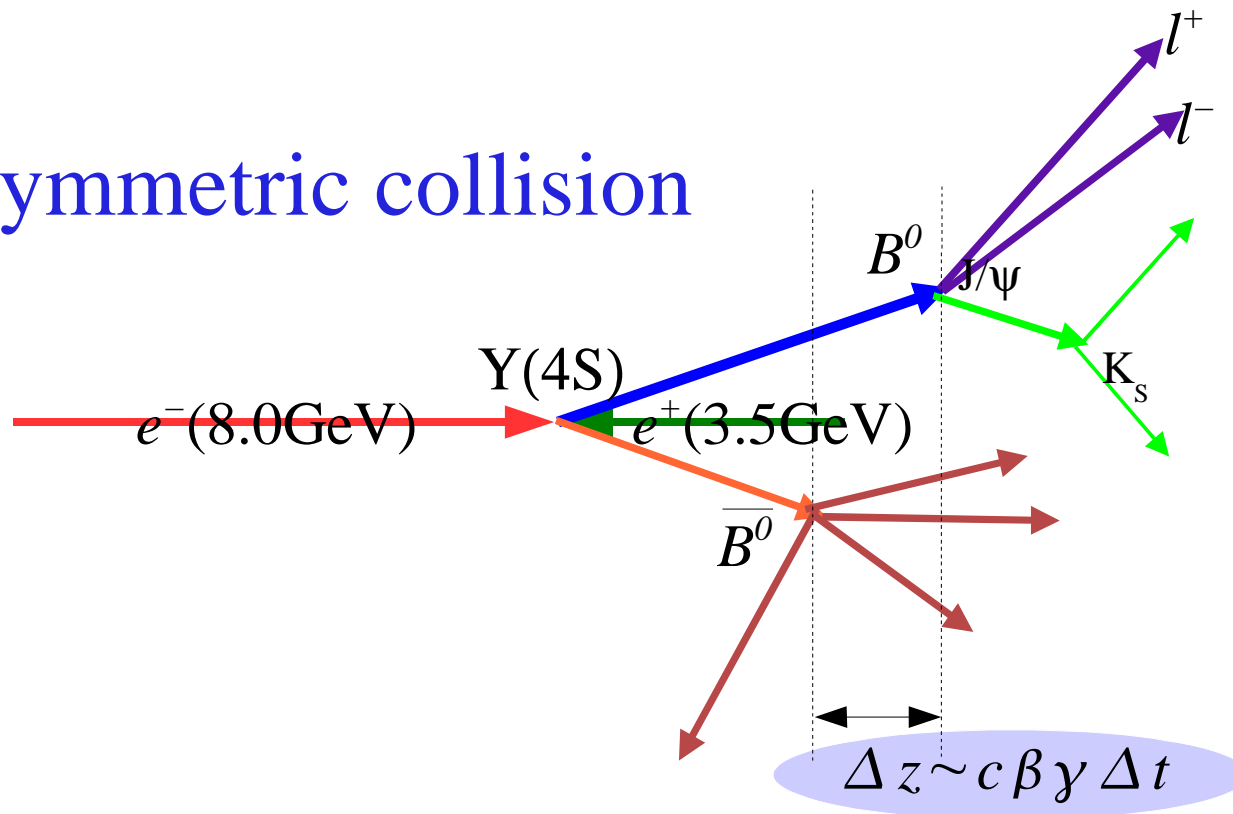
★ Decay time of B^0 is measured by looking for the decay vertex.

★ B^0 is produced almost at rest.

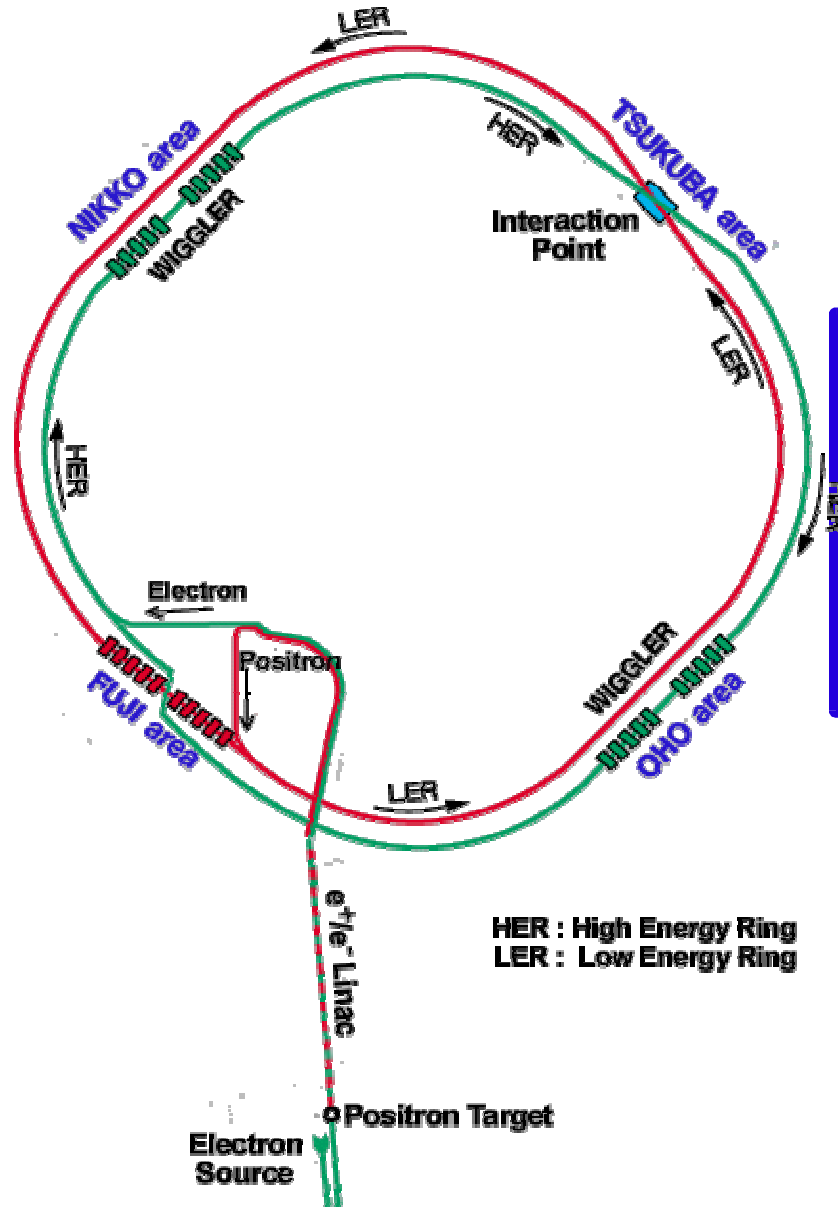
★ The lifetime of B^0 is $\sim 1.5\text{ps}$.

→ *very hard to measure decay vertex.*

Asymmetric collision



KEKB asymmetric e^+e^- collider



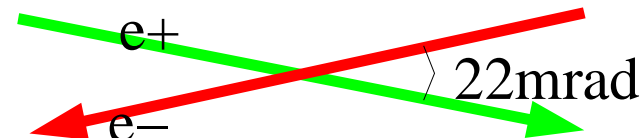
Two separate rings for
electrons (HER) : 8.0 GeV
positrons (LER) : 3.5 GeV
 $\rightarrow E_{\text{CM}} = Y(4S)$

Luminosity: World Record!!

$$L_{\text{peak}} = 4.49 \times 10^{33} \text{ cm}^{-2} \text{ sec}^{-1}$$

(Design = $10^{34} \text{ cm}^{-2} \text{ sec}^{-1}$)

Crossing Angle : $\pm 11 \text{ mrad}$

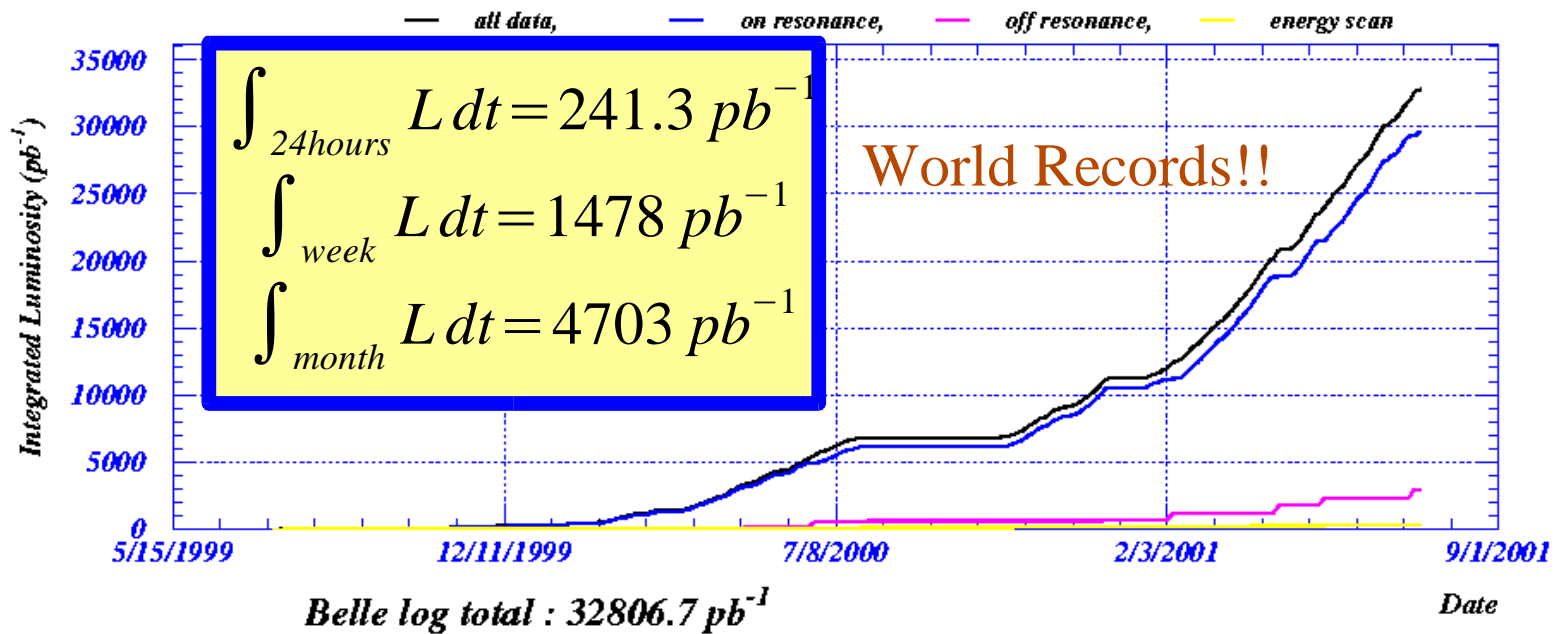
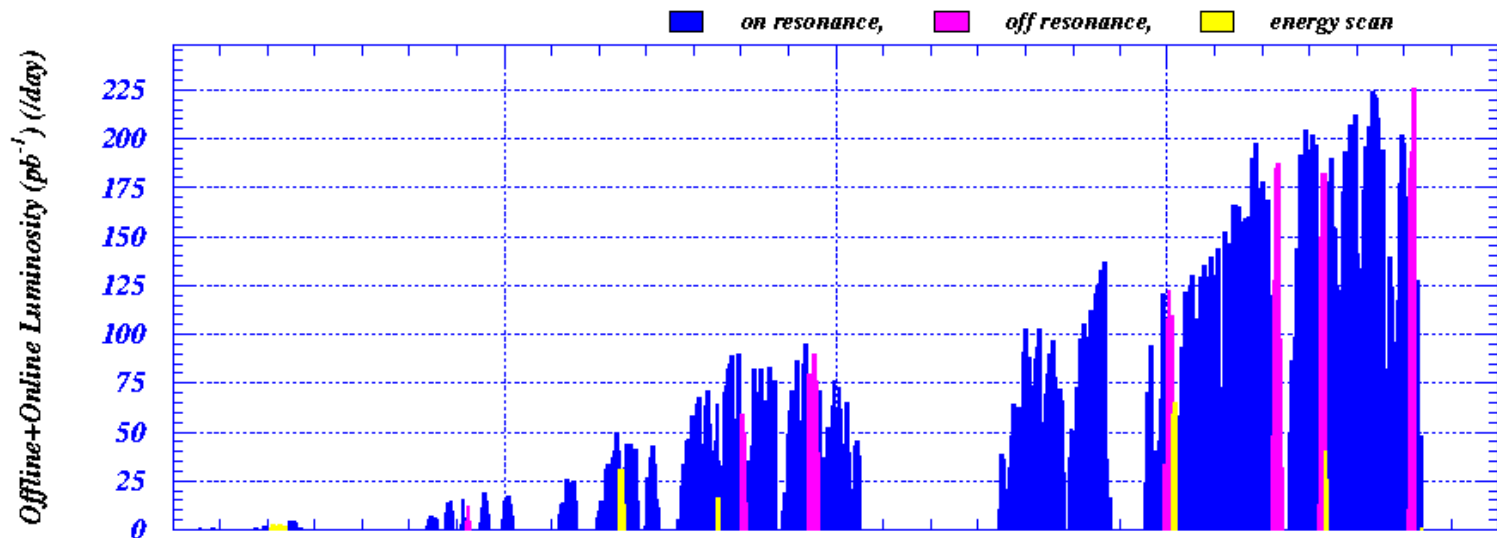


- no bending magnet
- background reduction in detector

Beam Size : $\sigma_x \sim 100 \mu\text{m}$; $\sigma_y \sim 3 \mu\text{m}$

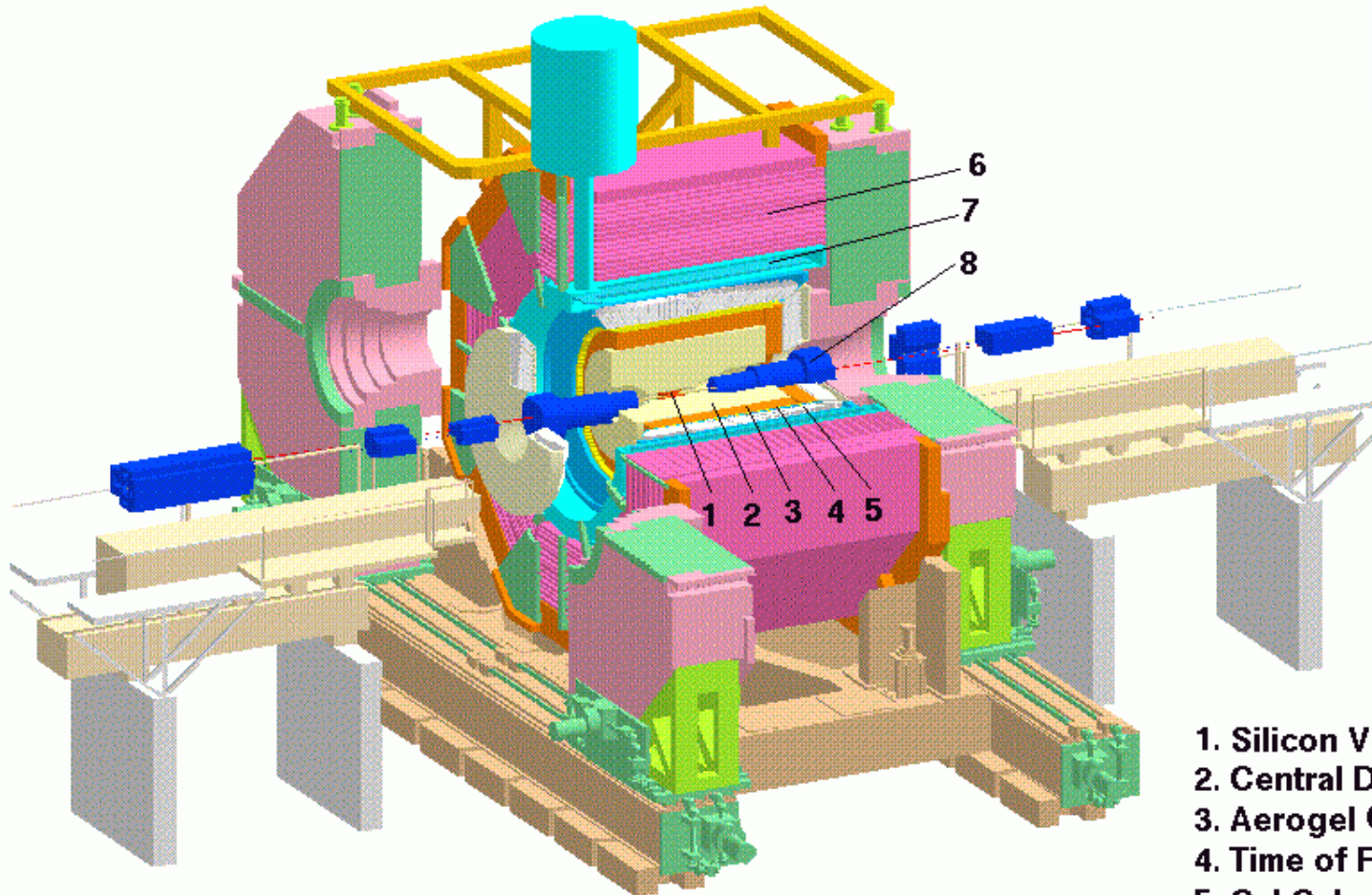
Offline+Online Luminosity (pb^{-1}) (/day)

2001/07/23 14.17



runinfo ver.1.41 Exp3 Run 1 - Exp13 Run 1640 BELLE LEVEL latest

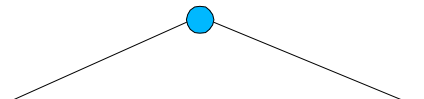
Belle Detector



1. Silicon Vertex Detector
2. Central Drift Chamber
3. Aerogel Cherenkov Counter
4. Time of Flight Counter
5. CsI Calorimeter
6. KLM Detector
7. Superconducting Solenoid
8. Superconducting Final Focussing System

Detector Performance

- SVD : **occupancy** < 4%
 $\sigma \sim 55\mu\text{m}$ for 1GeV/c track @ 90°
- CDC : **inner layer occupancy** < 10%
 $\sigma_{p_t/p_t} = (0.19p_t \oplus 0.3)\%$ ($\sim 0.35\%$ @ 1GeV/c)
 $\sigma_{\pi}(dE/dx) \sim 7.0\%$
- ECL : **pedestal spread : endcap** < 1MeV; **barrel** < 500 keV
 $\sigma_E/E = (1.3 \oplus 0.07/E \oplus 0.8/E^{1/4})\%$ ($\sim 1.8\%$ @ 1GeV)
- Hadron ID : TOF + ACC + dE/dx
Kaon ID : efficiency > 75%, fake rate < 5.5% for $0 < p < 5.0\text{GeV}$
- Lepton ID: electron (E/p, dE/dx, etc), muon (KLM,ECL)
Electron ID: Eff > 90%, 0.3% fake rate at 1GeV/c
Muon ID : Eff > 90%, 2% fake rate > 1GeV/c
- DAQ: **Trigger rate** $\sim 200\text{--}300\text{Hz}$, **Deadtime** < 5%, **Record Speed** $\sim 5\text{MB/s}$

- 
- Recently we measured the value of $\sin 2\phi_1$ in the data with 29.1 fb^{-1} ($\sim 31.3\text{M } B\bar{B}$ events) and found that the value is very large!

$$\sin 2\phi_1 = 0.99 \pm 0.14(\text{stat}) \pm 0.06(\text{sys})$$



CP is violated in B meson system!!
> 6 σ effect

2. Requirements for DAQ and Analysis

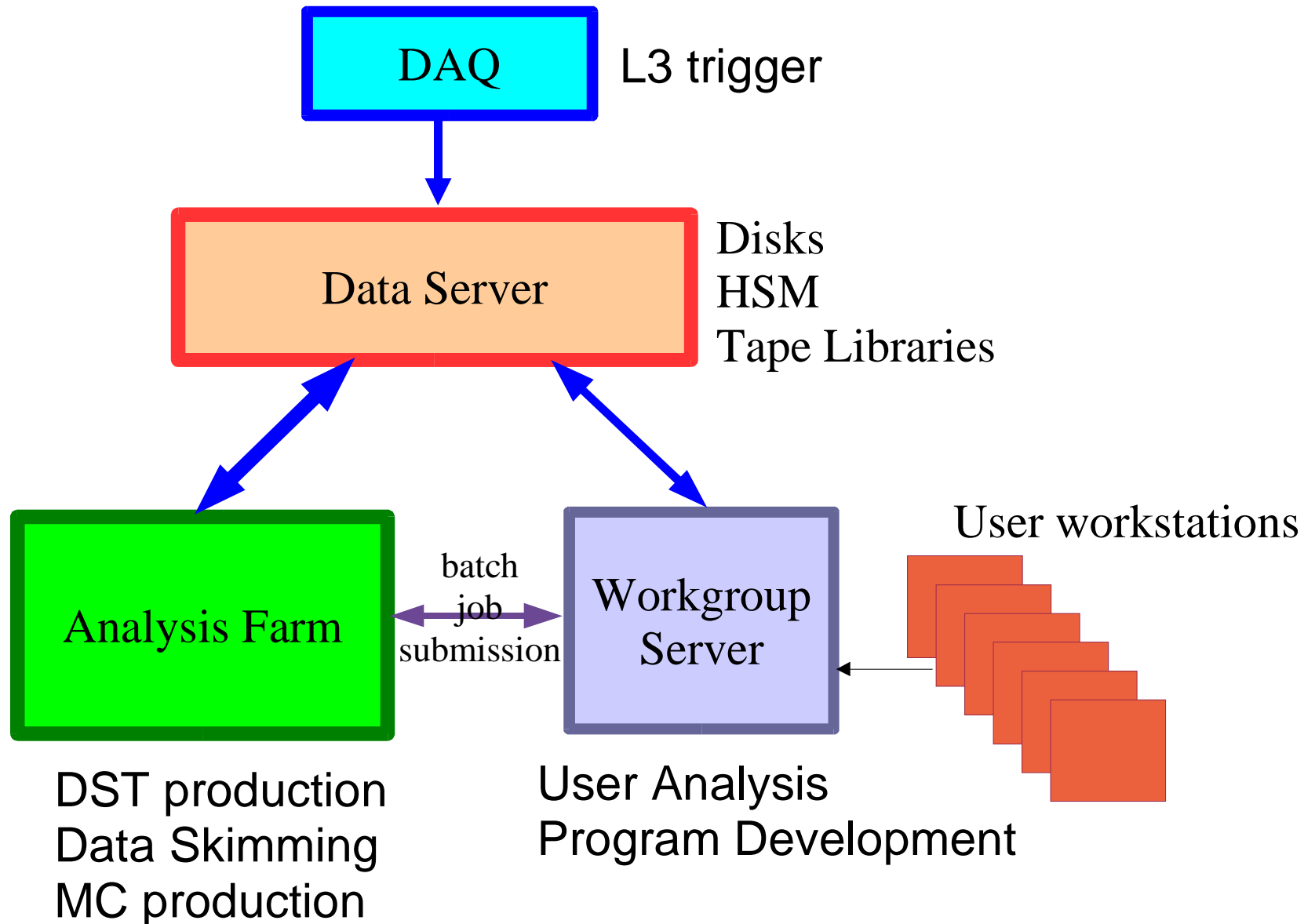
DAQ:

- Trigger rate : 200Hz (typical), 500Hz(max)
- Data size : 30KB/event
- Recording speed : 5MB/sec (typical) 15MB/sec(max)

Analysis:

- Data storage : 30TB/year
- CPU power : ~ 1000 Pentium@1GHz

Computing Model



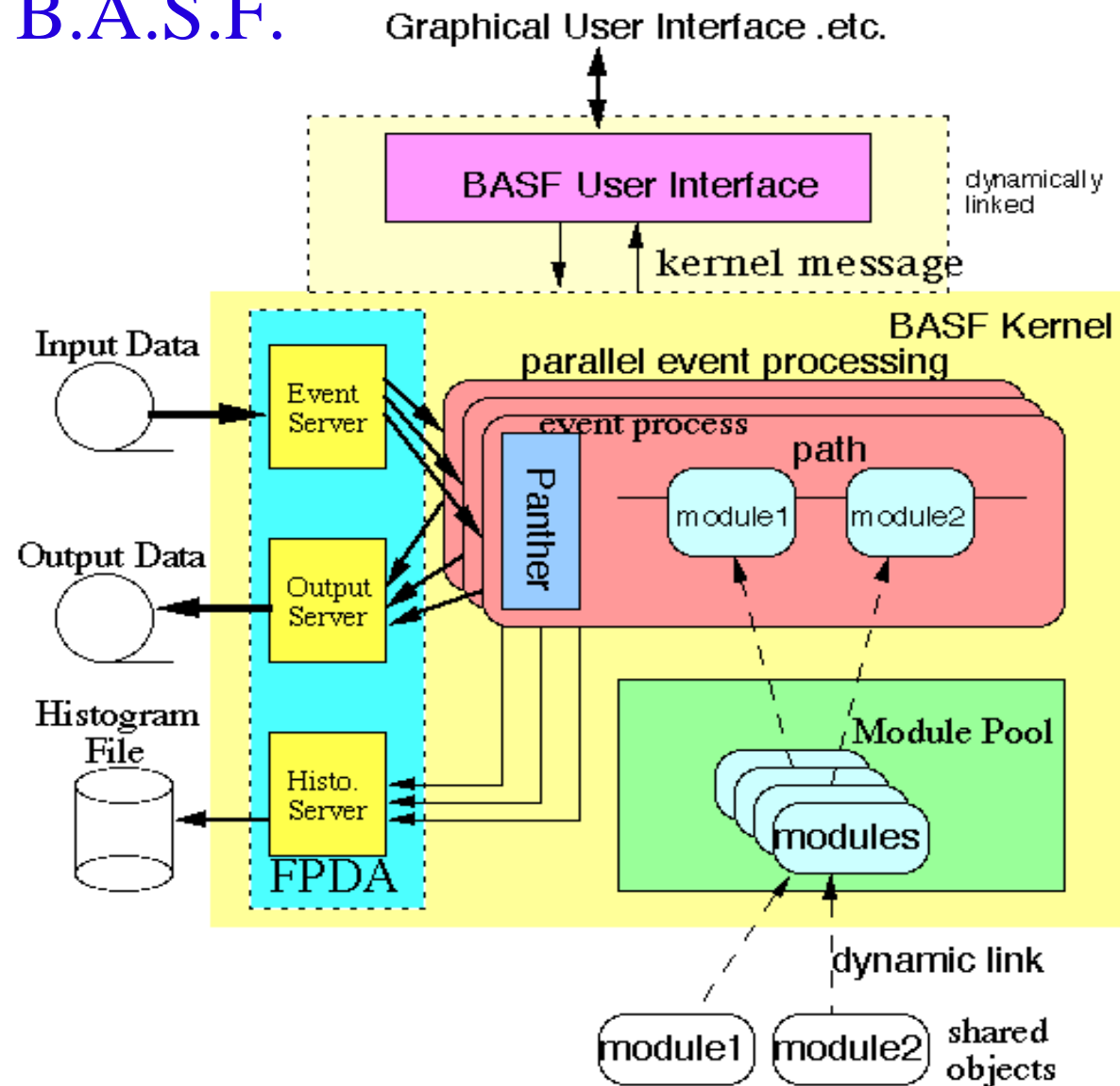
Common Software Tools

Programming Language : C++

Home grown tools:

- Data Management : "Panther"
 - non-OO Data Management System
 - ADAMO-like structure
 - C++ interface
- Analysis Framework : "B.A.S.F."
 - module and path structure
 - dynamic link of modules and I/O drivers
 - event-by-event parallel processing capability on SMP
- Communication tool over network : "NSM"
 - shared memory/message passing capability over network
 - to be used for slow control

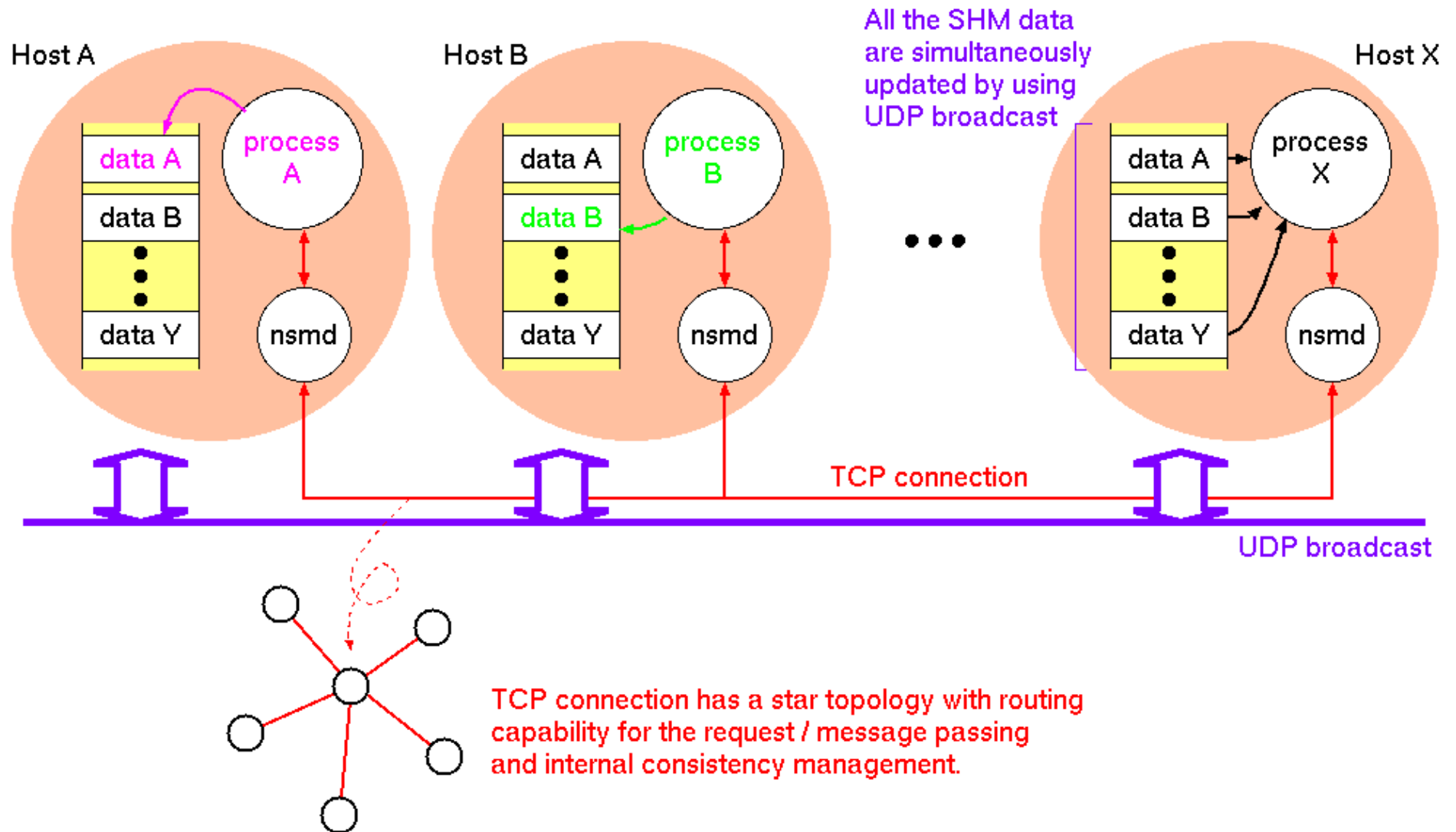
B.A.S.F.



* event process:
separate UNIX process
→ parallel event-by-event
processing on SMP

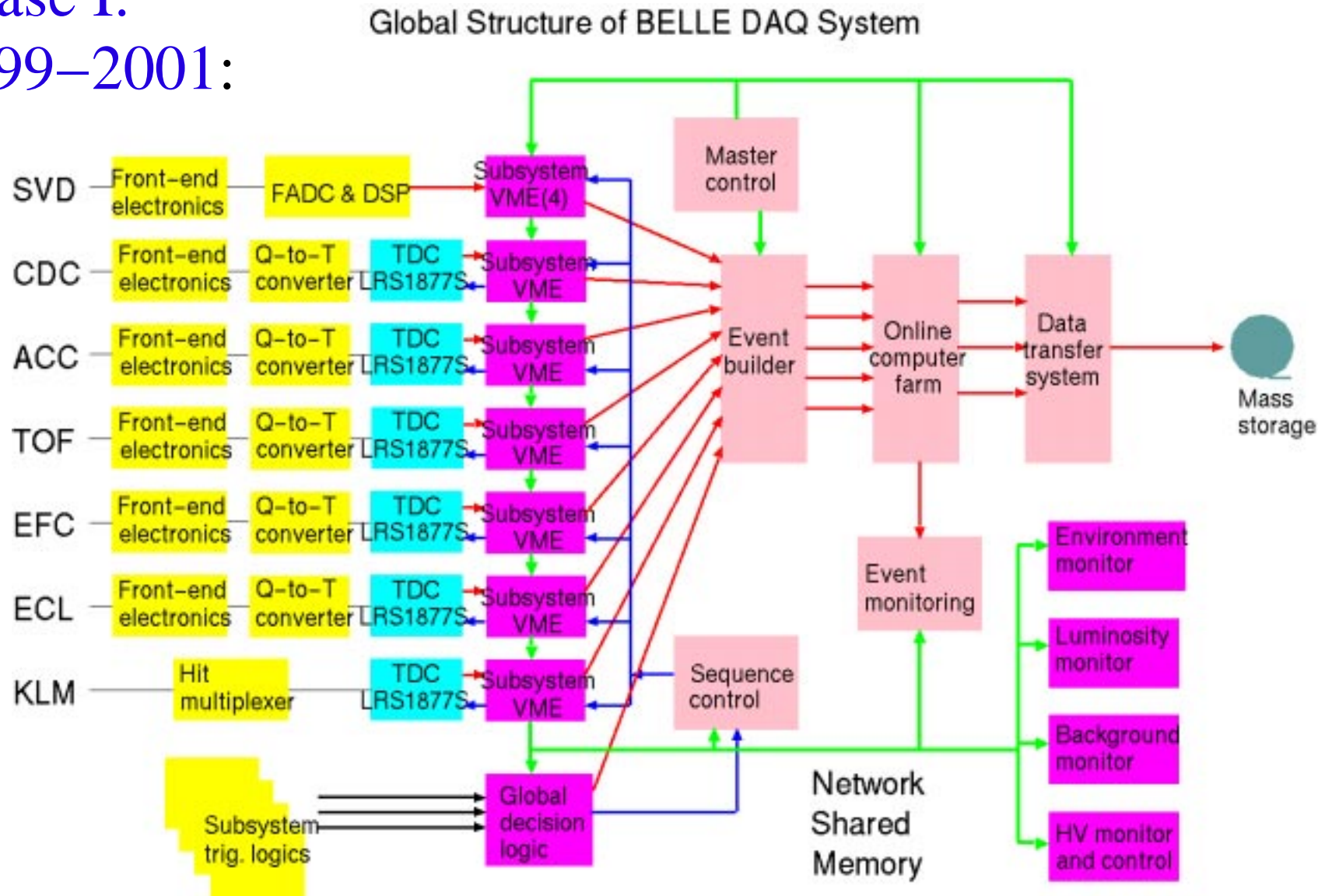
* analysis modules and
I/O packages are
dynamically linked

NSM



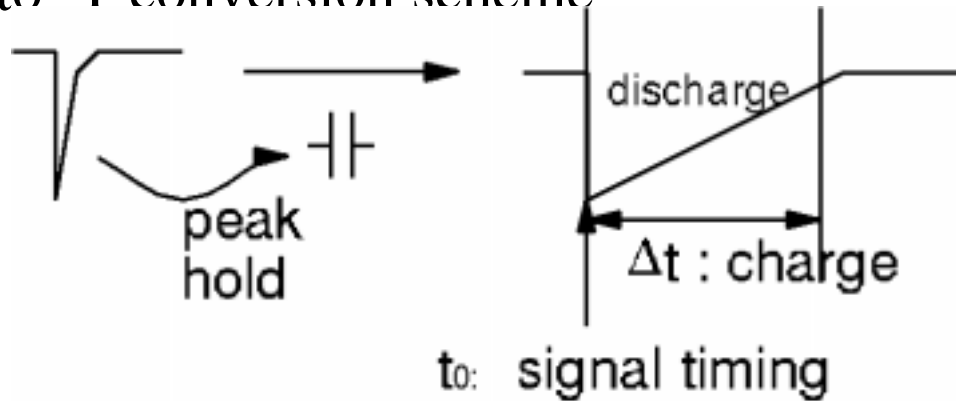
3. Data Acquisition

Phase I:
1999–2001:

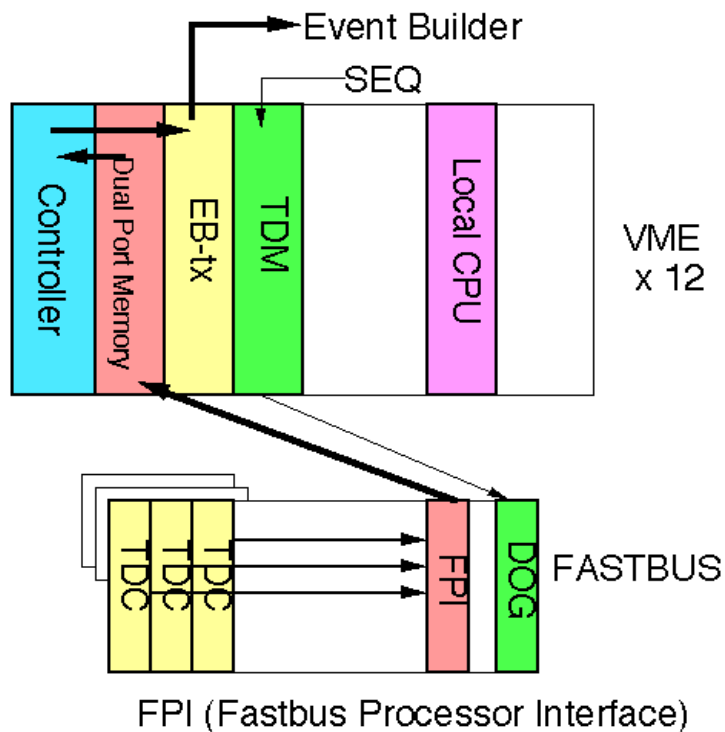


Frontend Readout

- Q-to-T conversion scheme

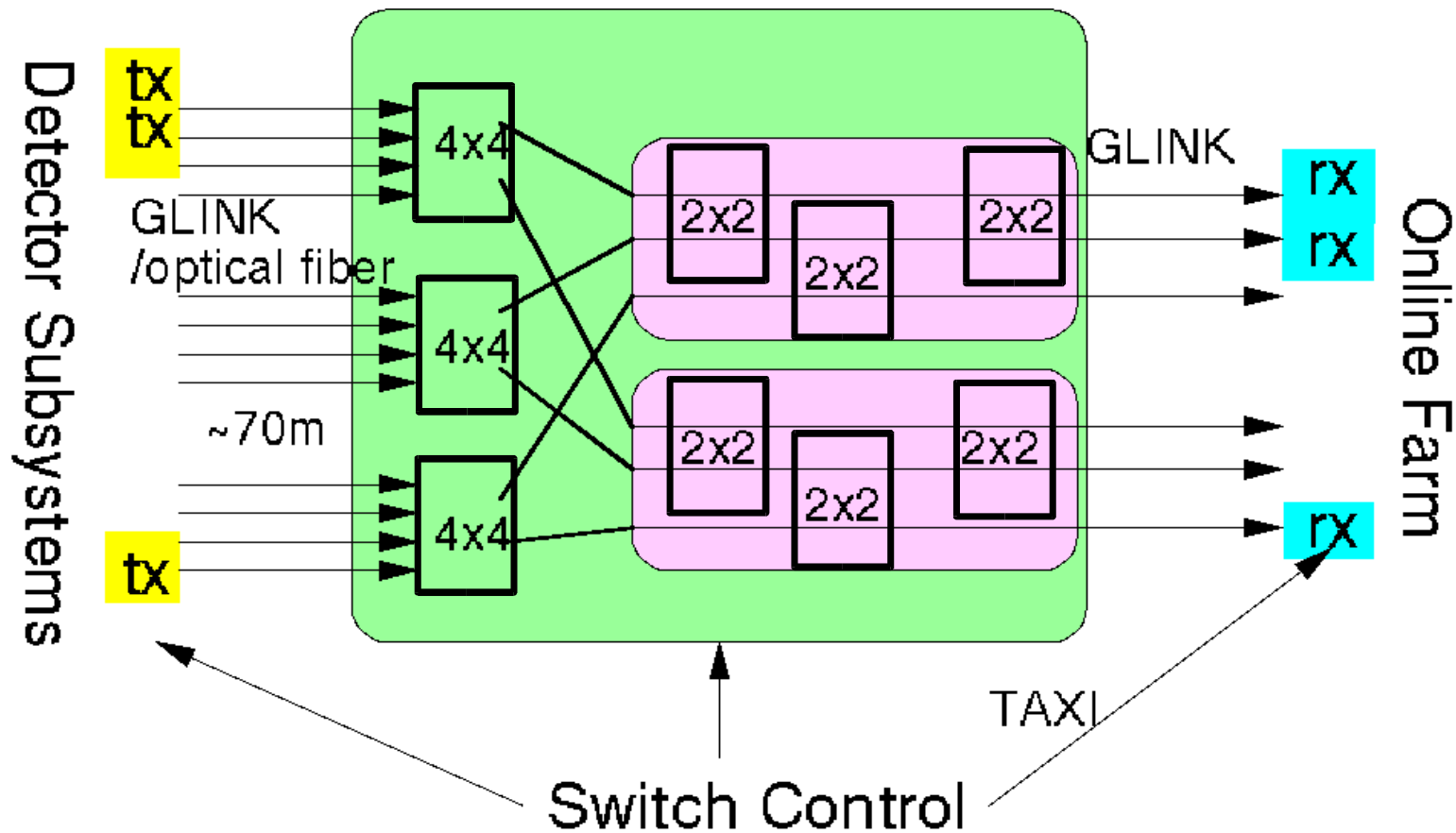


Q-to-T : LeCroy MQT300
multi-hit TDC : LRS1877S



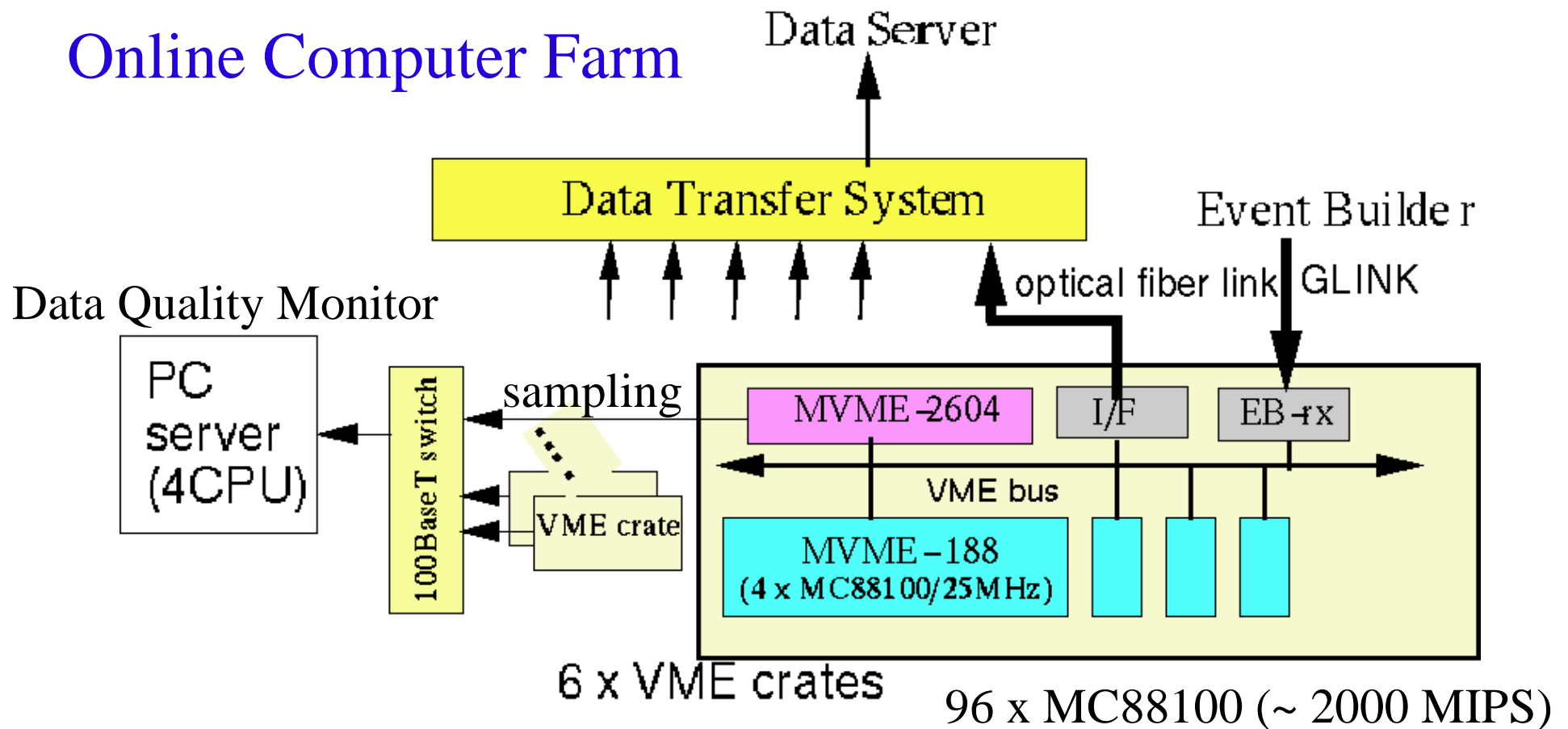
FASTBUS readout through
VME subsystem

Event Builder



- * External traffic control
- * 2x2 and 4x4 barrel shifting switches → 12 x 6 switch
- * GLINK connection

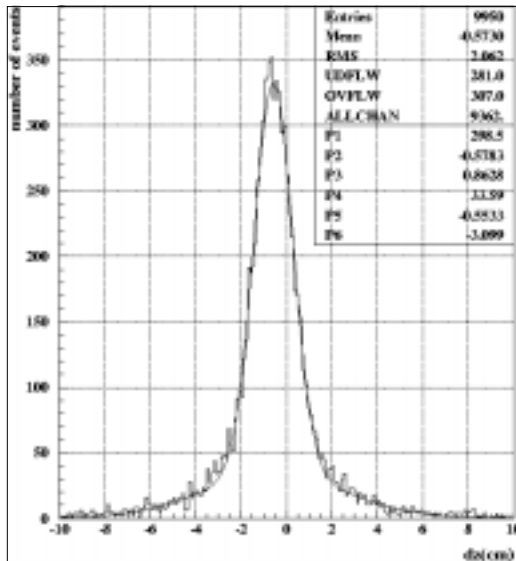
Online Computer Farm



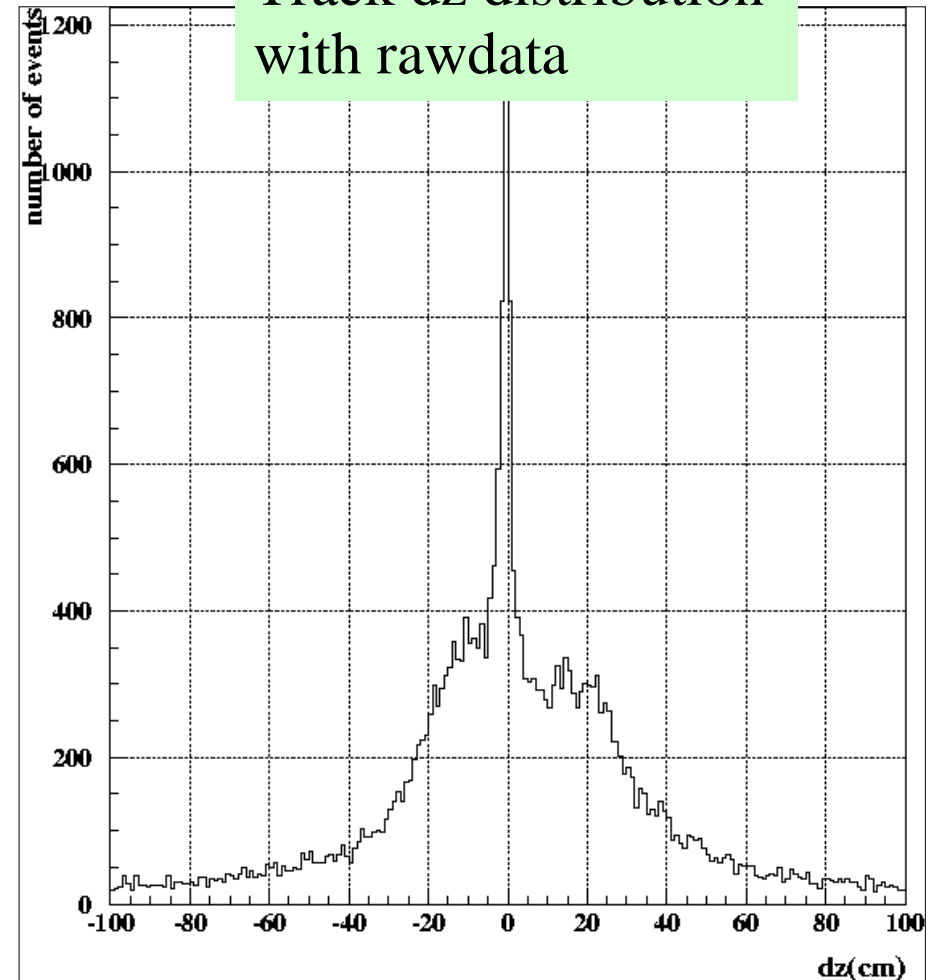
- * Raw data are converted into Panther format
- * B.A.S.F is used as the framework on MVME-188.
- * Level 3 trigger is implemented in the framework

Level 3 Ultra Fast Tracker

- Memory–lookup technique is used
 - Effective for $P_t > 350\text{MeV}/c$
- Drift–time info. is used
 - dz resolution $\sim 9\text{mm}$
 - obtained with $ee \rightarrow \mu\mu$



Track dz distribution with rawdata



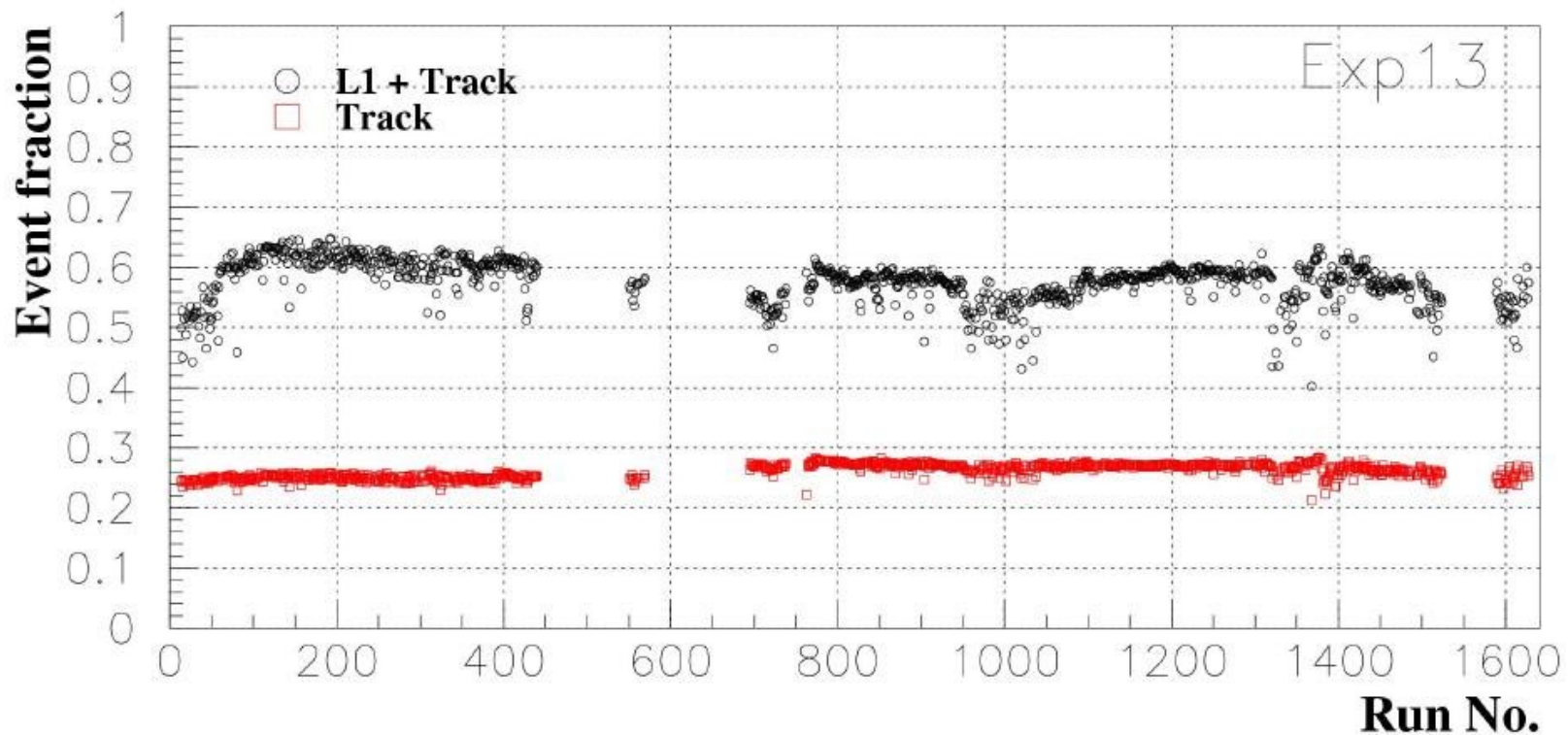
Performance of Level 3 Trigger

Efficiency $> 99\%$ for B events

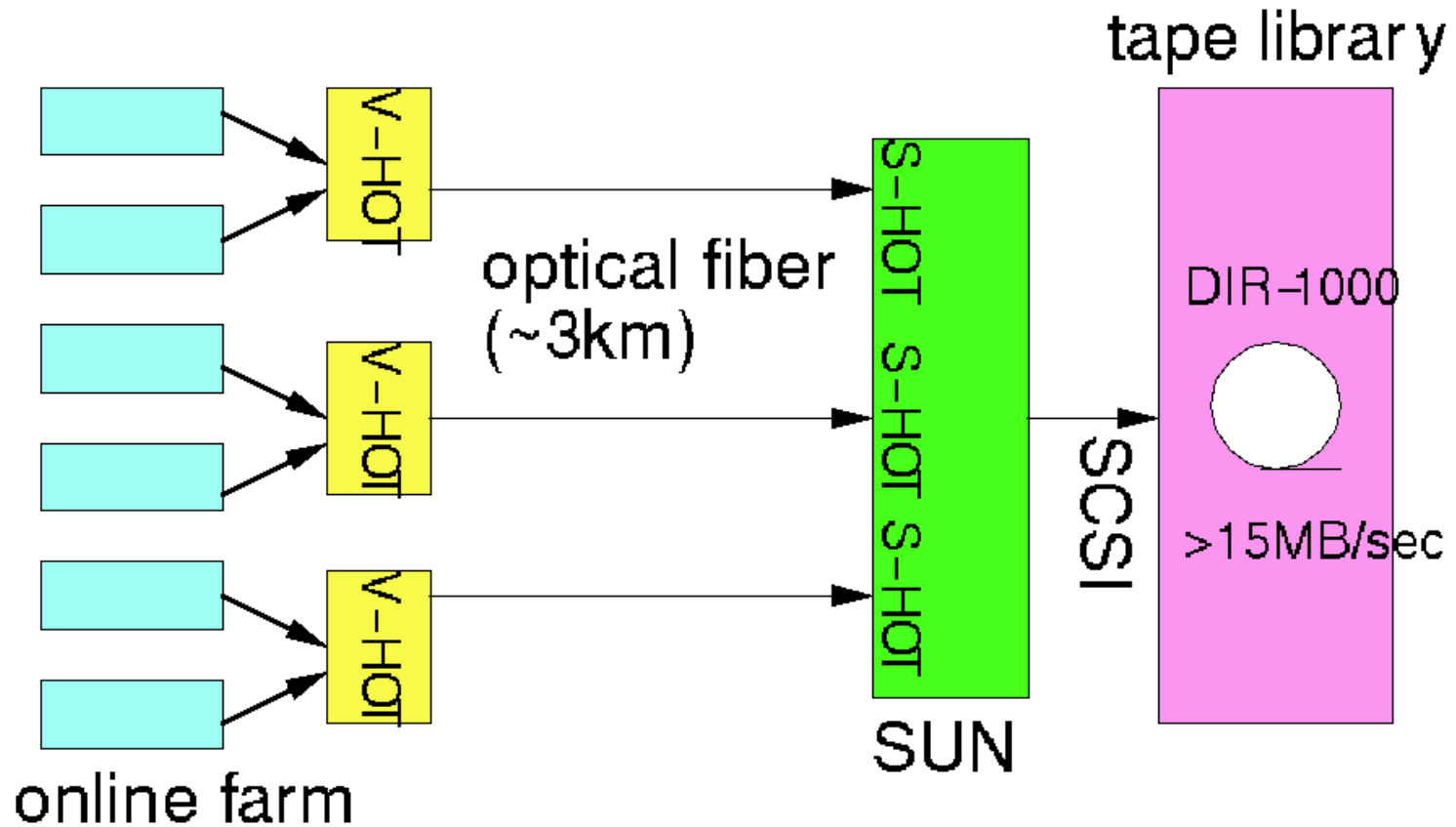
$> 99\%$ for $\tau+\tau^-$ candidates

No deadtime (OK up to $\sim 1\text{kHz}$)

Background rejection (see the figure below)



Data Recording System



The system worked very well until now
→ deadtime < 5% with ~250Hz L1 trigger rate.

Problems

1. The technology used to build the system is outdated.
→ difficult to maintain
2. CPU power of online farm is not enough (only 1/3 of design value)
→ difficult to add more CPUs (CPU modules are outdated....)



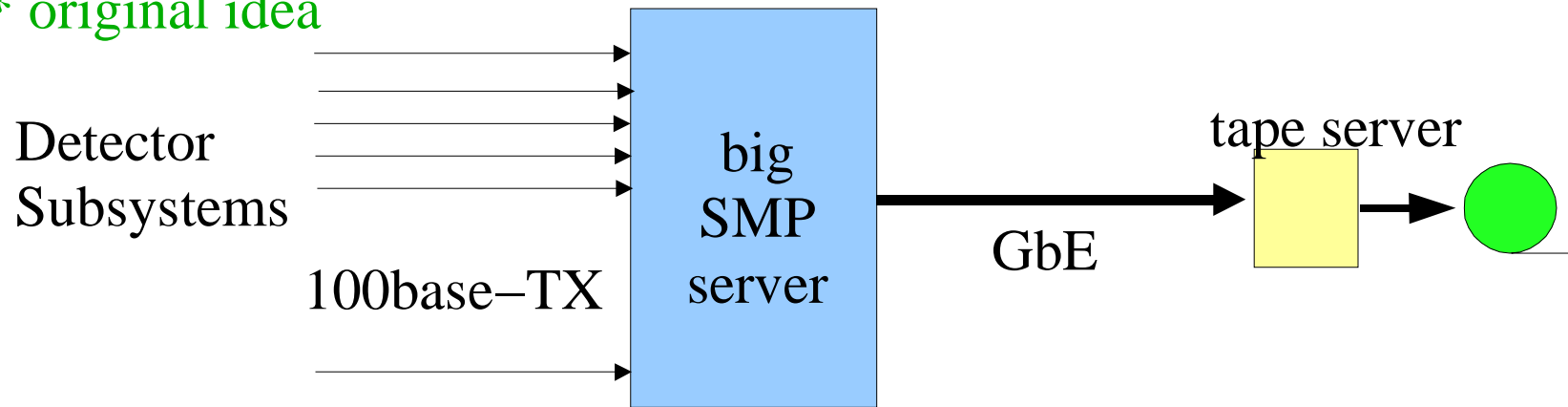
Replace

**Event builder, Online Farm and Recording system
with a set of PC's connected via network.**

- * New tape drive (SONY DTF2) can be used for data recording.
→ 24MB/sec recording speed.

Switchless Event Building

* original idea

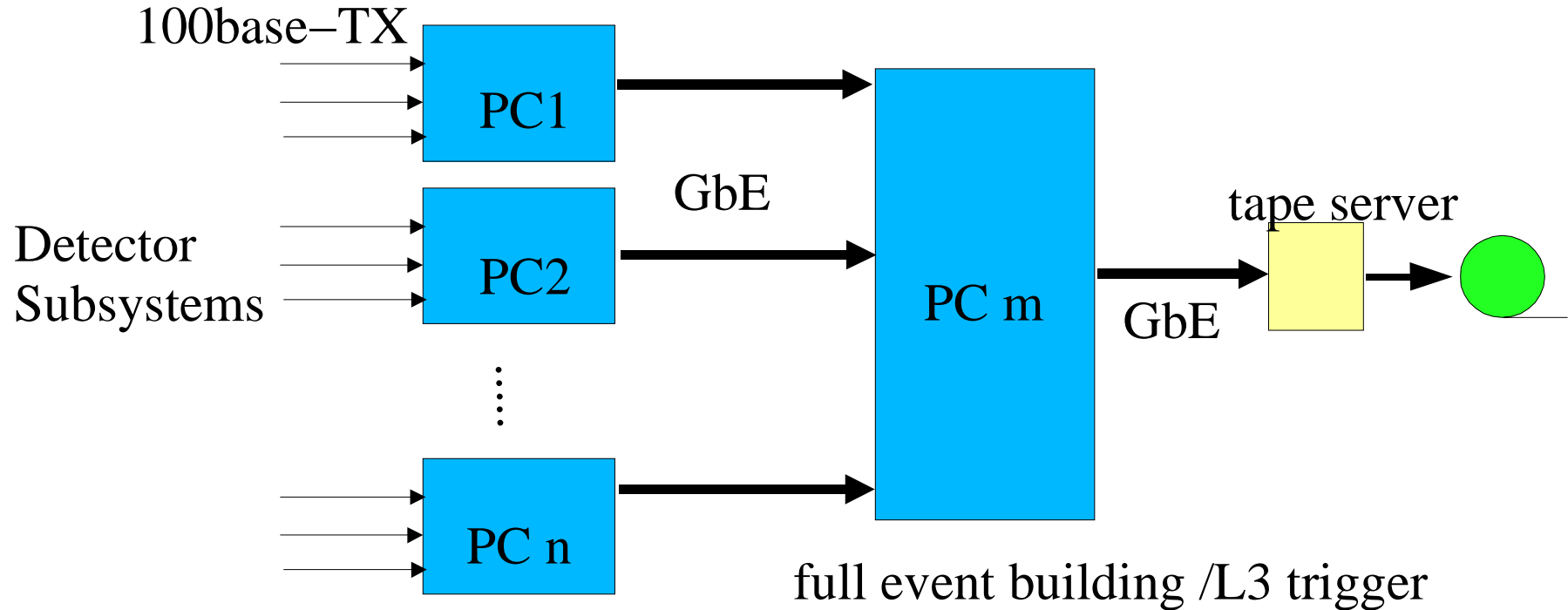


Event Building
L3 trigger processing

- Exactly the same offline environment → easy to write L3 code
 <– B.A.S.F. supports parallel processing on SMP
- Event building → trivial using "Panther" on shared memory
 (just a matter of handling pointers on shared mem.)

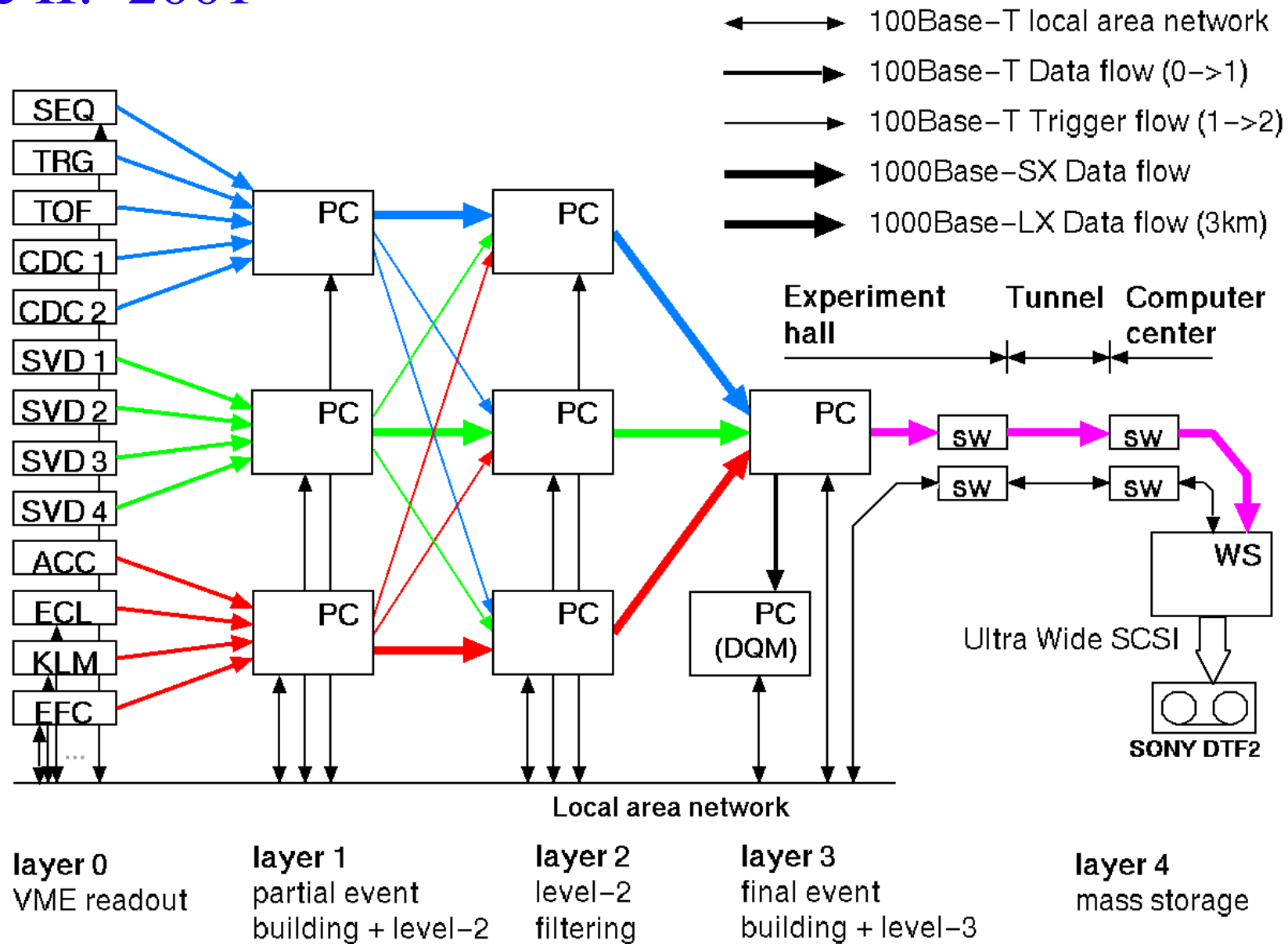
- * a large scale SMP server → expensive
- * bandwidth of CPU bus for SMP can be bottleneck
- * need to drive too many network cards by single OS → reliable?

Multi-step event building



- * Step-by-step event building
- * Enable to use small scale PCs → low cost
- * Level 2 trigger: efficient data reduction in upper stream
- * distribution of processing power to multiple PCs → scalability
- * point-to-point network connection
→ can avoid traffic control problem

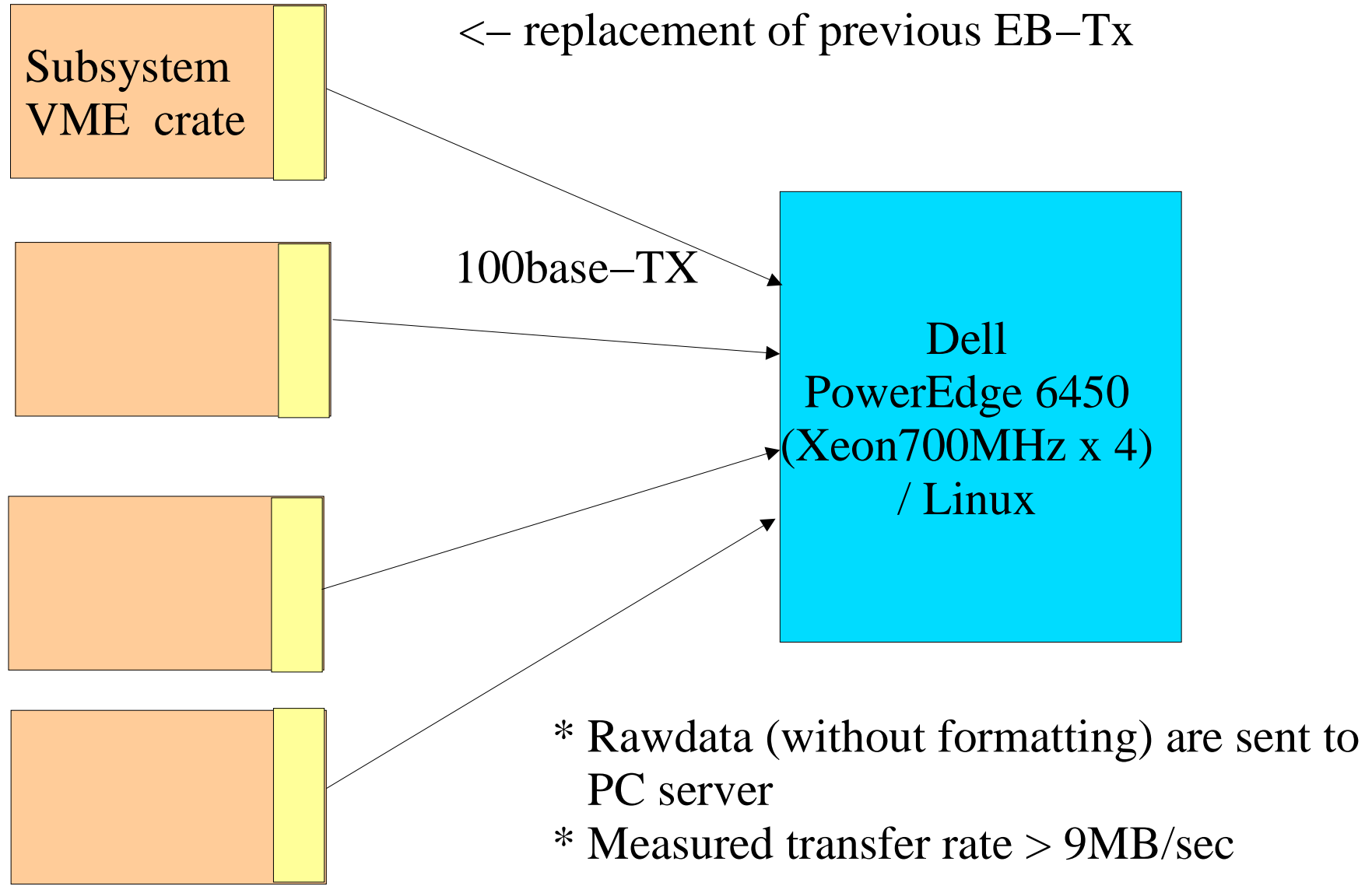
Phase II: 2001–



PC: SMP server with 4 CPUs

Layer0 → Layer1

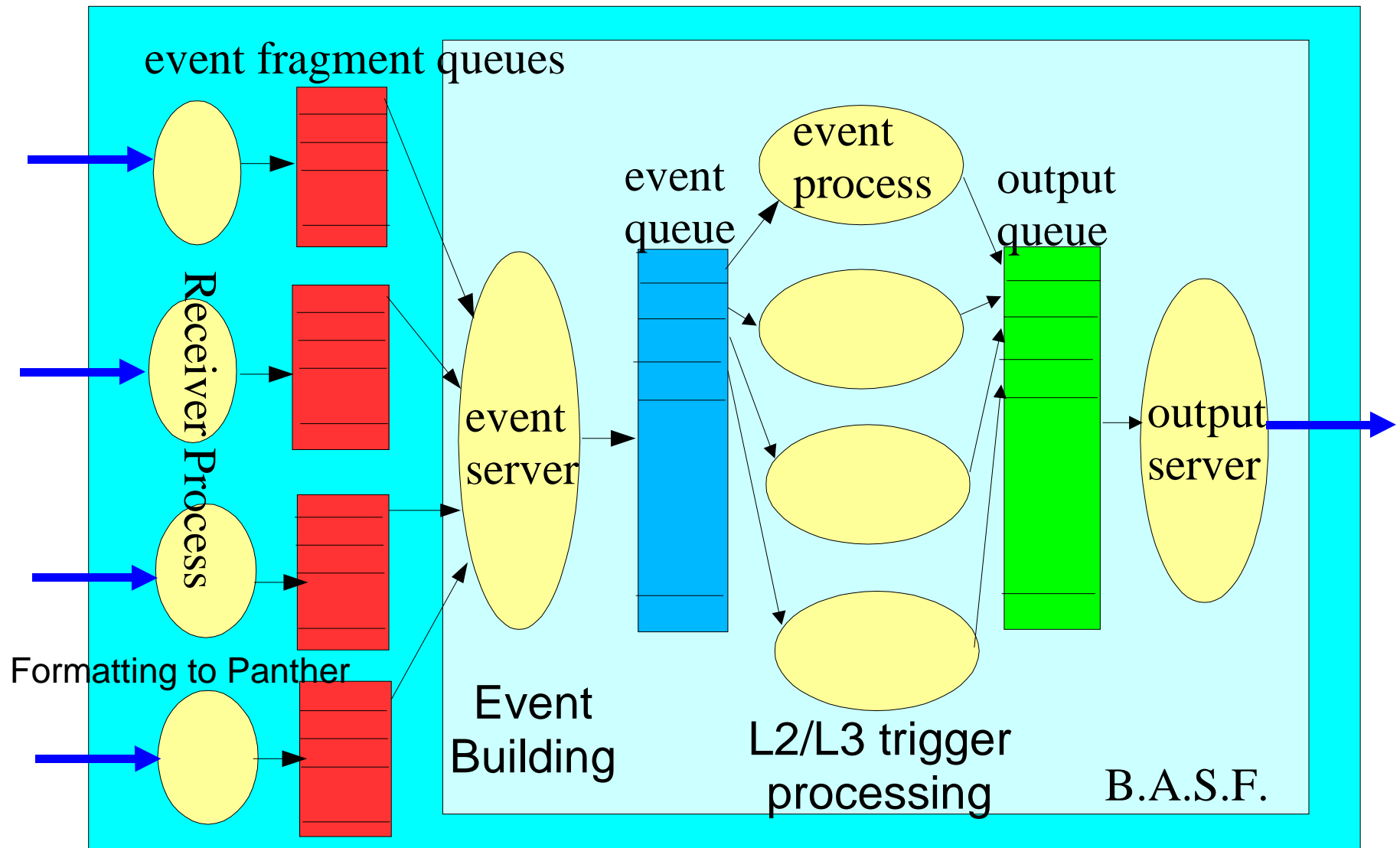
MVME5100/VxWorks



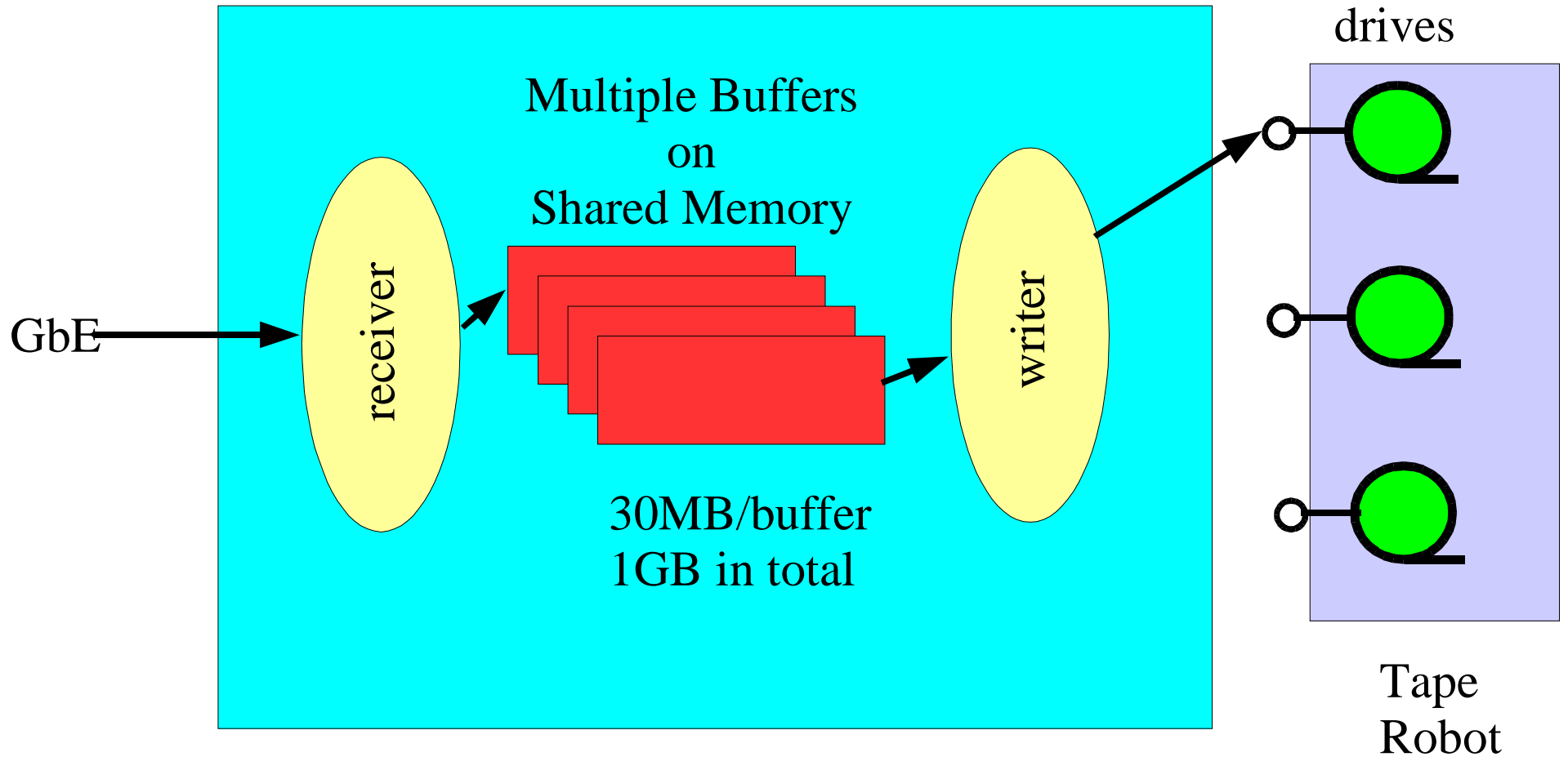
Layer 1/2/3

process

4 Xeon CPUs on Linux-SMP



Data Recording



The construction of system was completed
and now the system is being tested.

Measured performance:

- Point-to-point transfer speed on GbE : > 70MB/sec
- Overall speed:
 - dummy events on layer 1 node → tape
 - event building on layer 1 and layer 3
 - no event filtering on layer2 and 3

~ 20 MB/sec

→ satisfies our requirements

This system will be in charge from this autumn run
(from Oct. 1)

3. Data Analysis at Belle

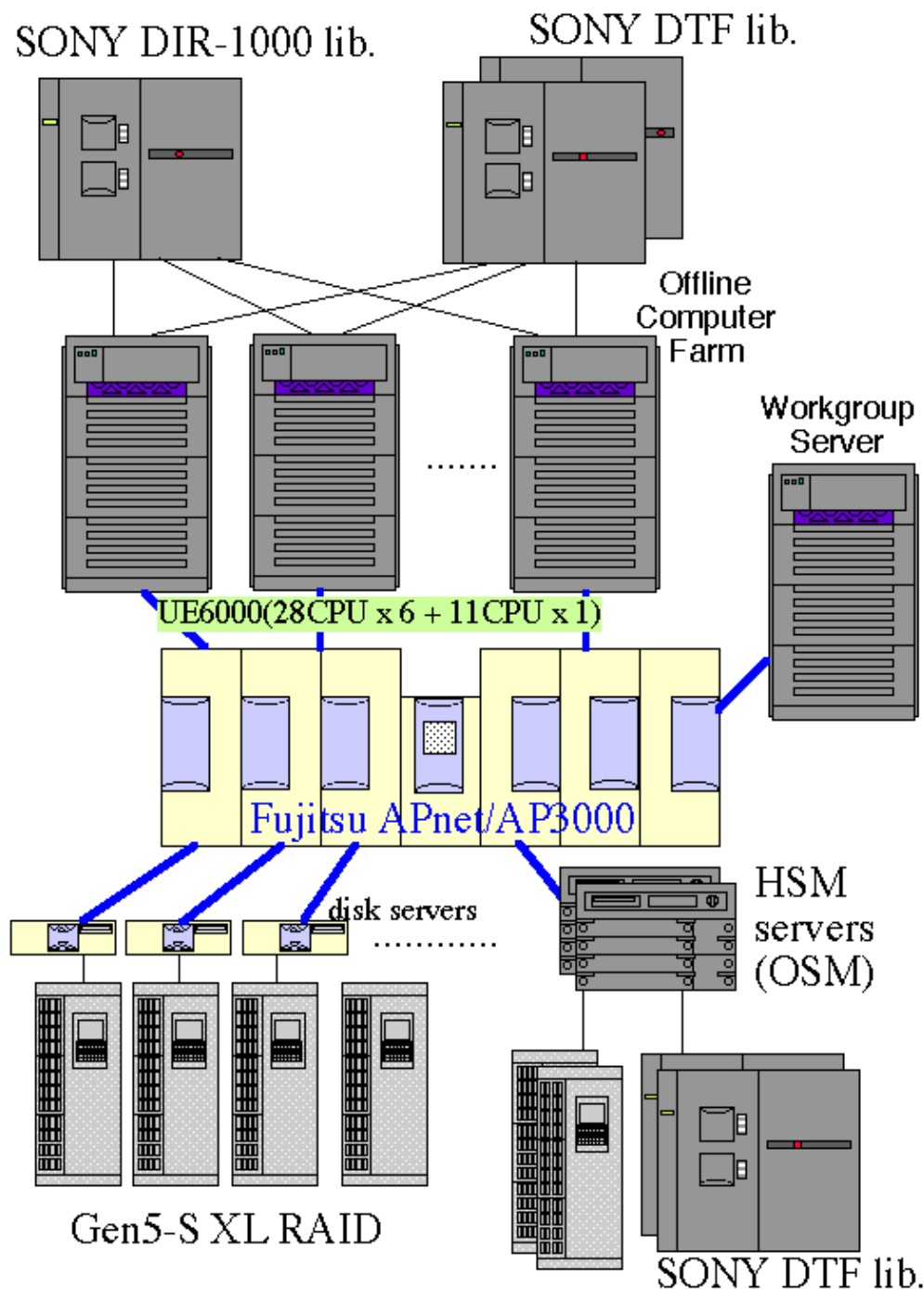
Computing Platform

Phase I (1997–2000)

- designed to obtain maximum performance of "B.A.S.F. + Panther"
 - based on a large SMP–server cluster + HSM
- ← design fixed in 1996

- All the system had to be purchased at the same time in 1997 from one company by bidding because of Lab's regulation.
- 4 year rental contract → can be upgraded in 2001

CPU power	: 1500 SPECint95
Disk size	: ~ 5Tbytes
Tape library capacity	: ~ 200Tbytes

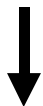


Cluster of 28 CPU
SMP servers
(SUN UltraEnterprise)

* SMP performance was tested by the actual software and proven to increase almost linear to the number of CPUs.

However,

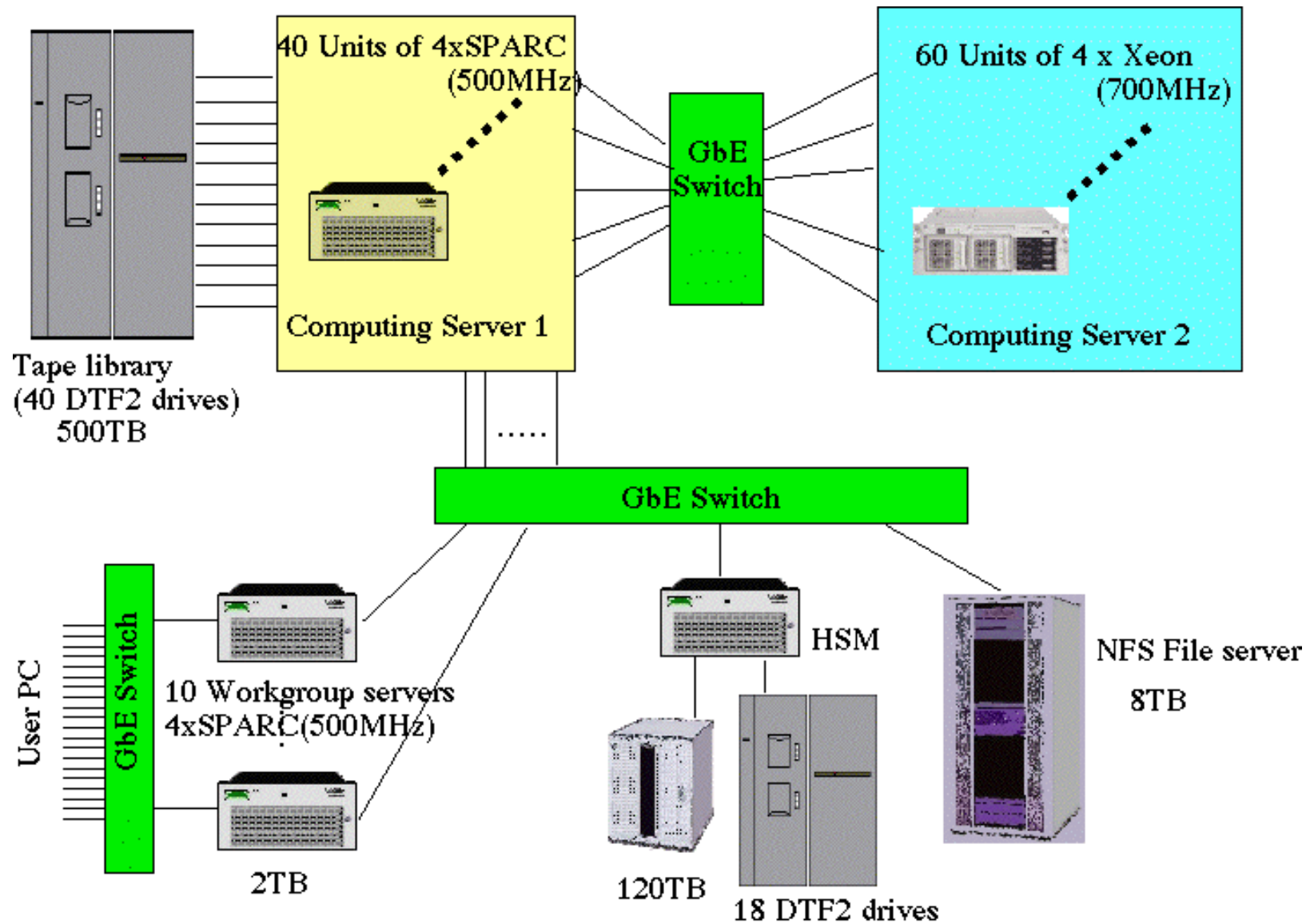
- CPU power became running short soon (as well as everywhere....)
 - added PC farms for MC production (2000)
 - ➔ PC server with 2 or 4 Pentium Xeon (500–700MHz) x ~80 units
 - ➔ Linux–SMP is used as the operating system
 - > proven that PC farm can be used for Belle analysis



Decided to replace main SMP cluster with PC farms
with 10 times of previous CPU power
in 2001 upgrade of computing platform

CPU power : ~ 15000 SPECint95
Disk size : ~ 10TB
Tape library capacity : > 500TB

Phase II (2001–)

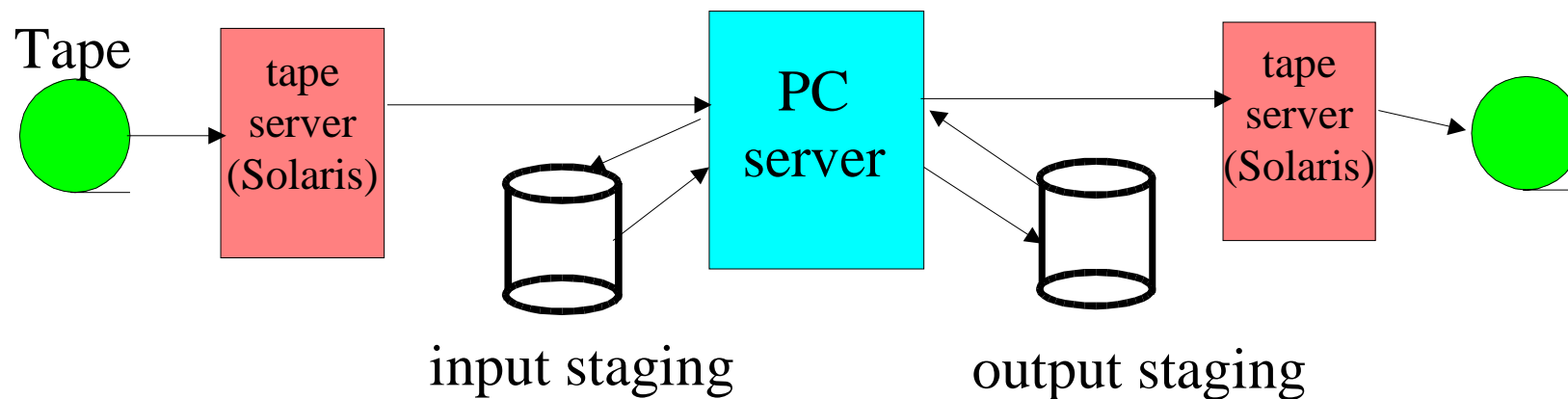


Extension of analysis framework:

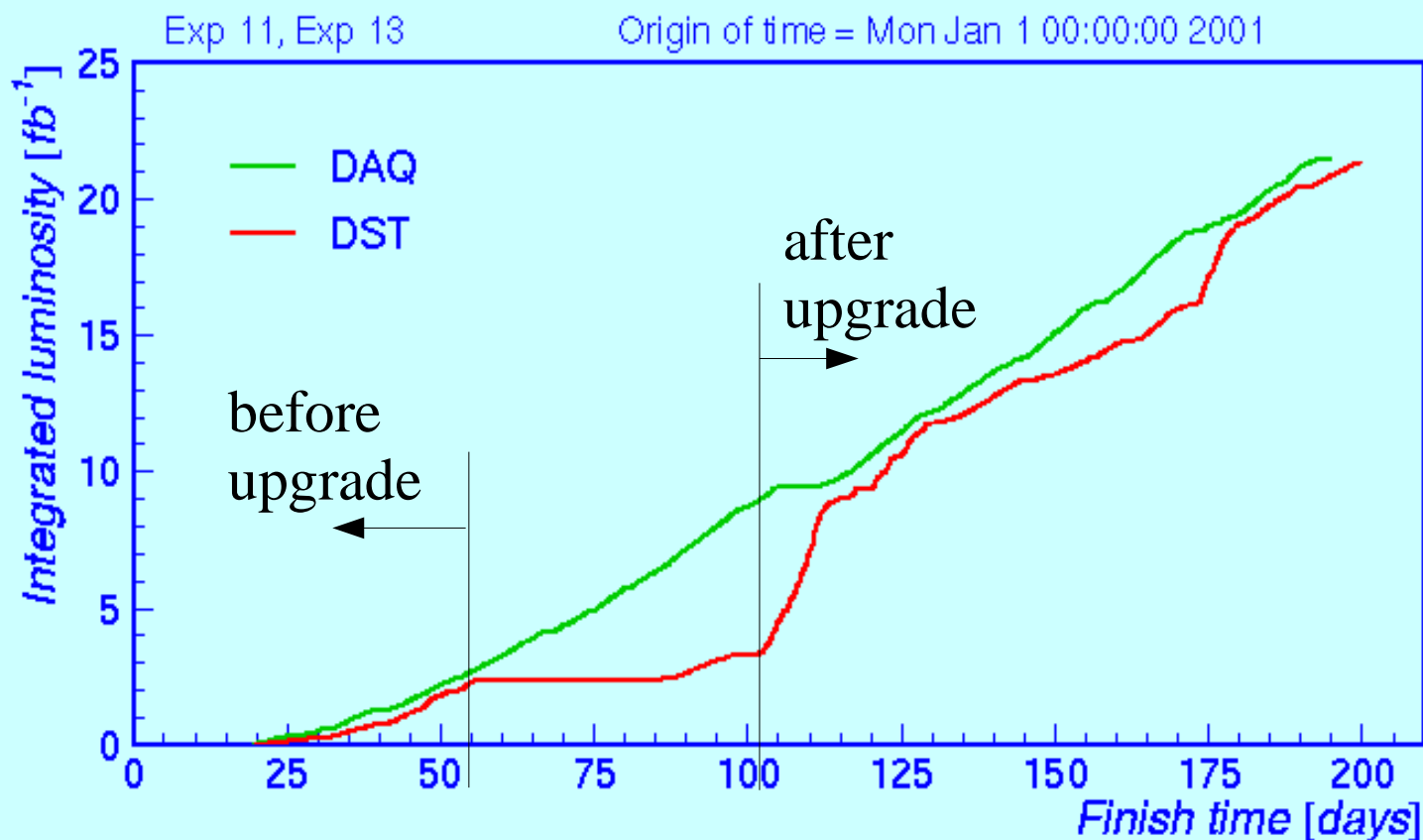
- * need to extend B.A.S.F. to utilize PC servers connected via network
 - parallel processing on network–distributed SMP machines

Intermediate approach (Mar 2001 – Jul 2001)

- disk staging
- run–by–run basis parallel processing



DST production event rate



Online Integrated Luminosity vs. DAQ/DST Job Finish Time

can process $1 \text{ fb}^{-1}/\text{day}!!$
($\sim 1\text{M}$ BB events)

– Drawbacks

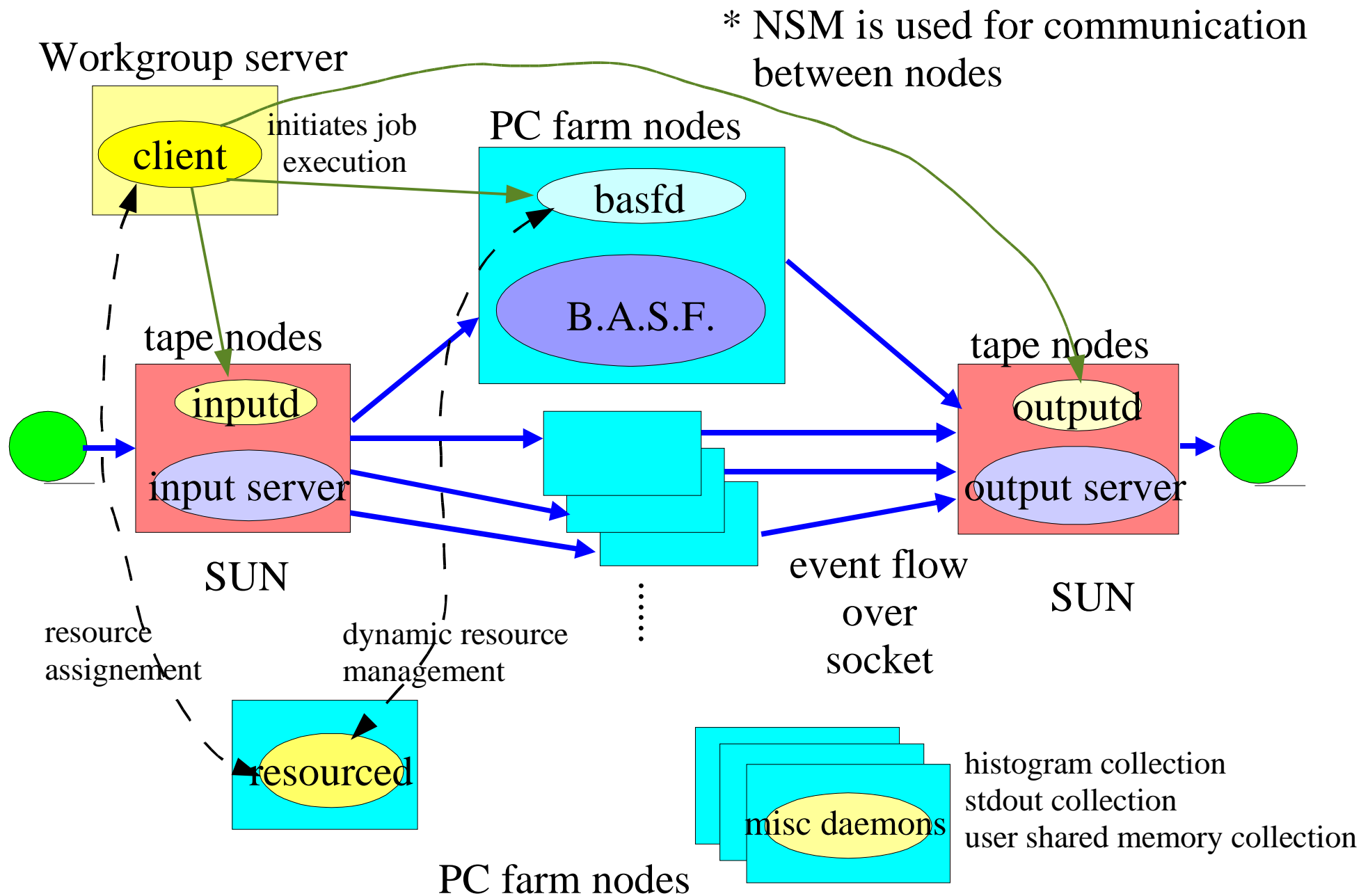
- * The management of staging disk size is difficult.
- * I/O network traffic becomes "bunched".
- * Real time processing is not possible

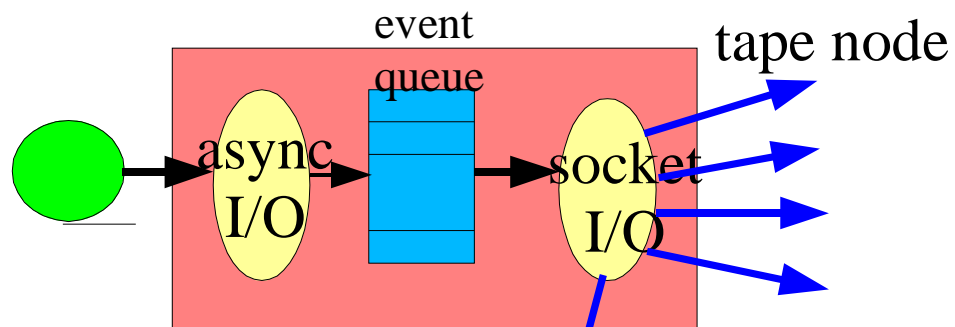


Distributed B.A.S.F. (dBASF)

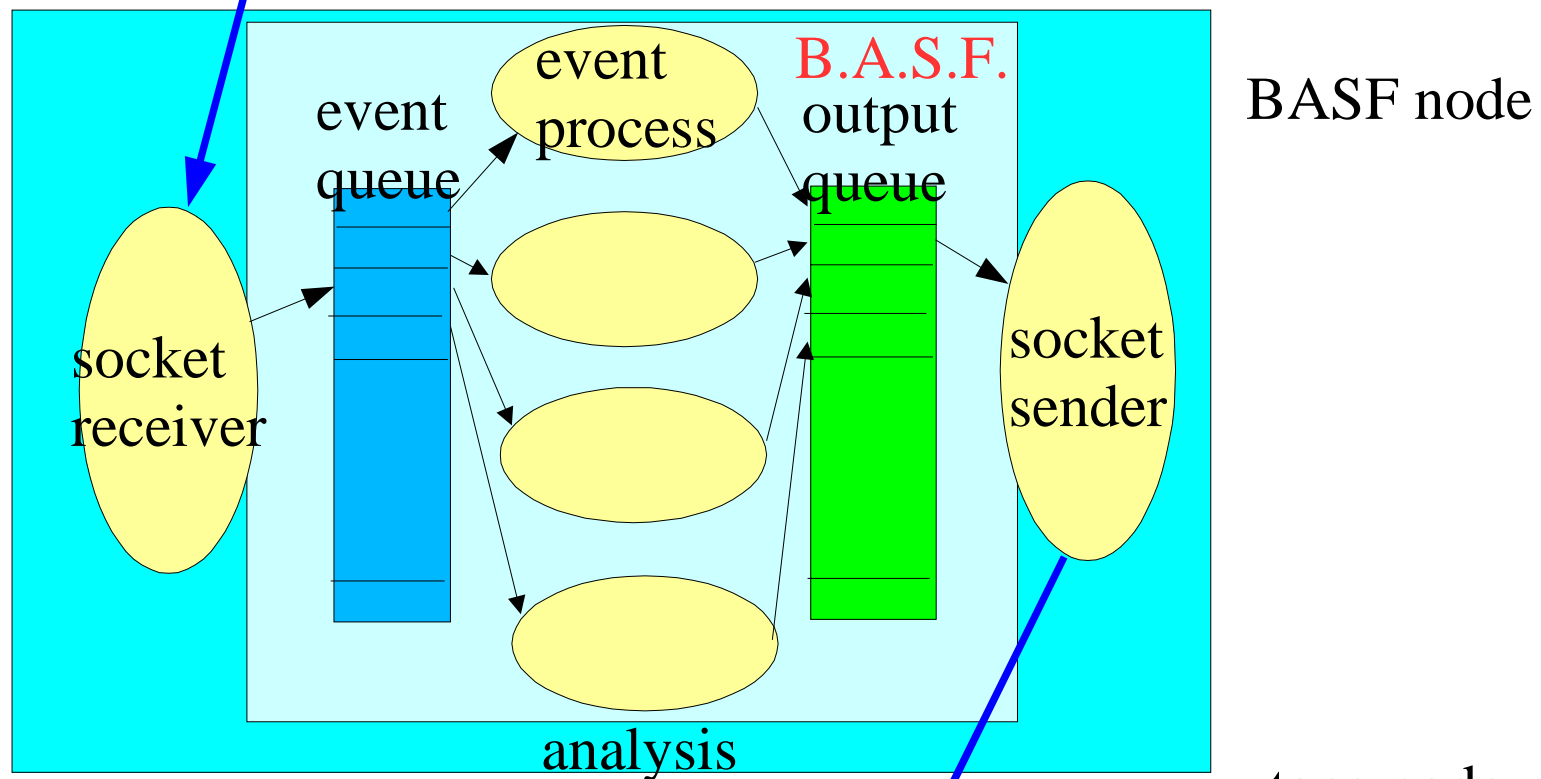
- extended event by event parallel processing capability over network
- dynamic optimization of resource usage

dBASF

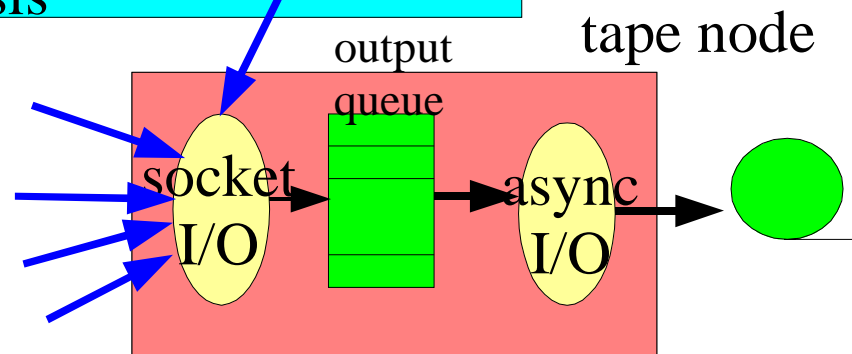




dBASf data flow



event data transmission over socket



dBASF was partially used for DST production in the previous runs.

→ quasi real time DST production with 15 PC server nodes

→ used in energy scan runs to obtain current beam energy from data



proven to work in situ

DST/MC production will be managed by dBASF
from coming autumn run (Oct.1).

* Dynamic resource management is not yet implemented.

→ work is in progress

5. Future of Belle

- Luminosity upgrade of KEKB accelerator is now being discussed.
→ Goal : New physics search in rare B meson decays

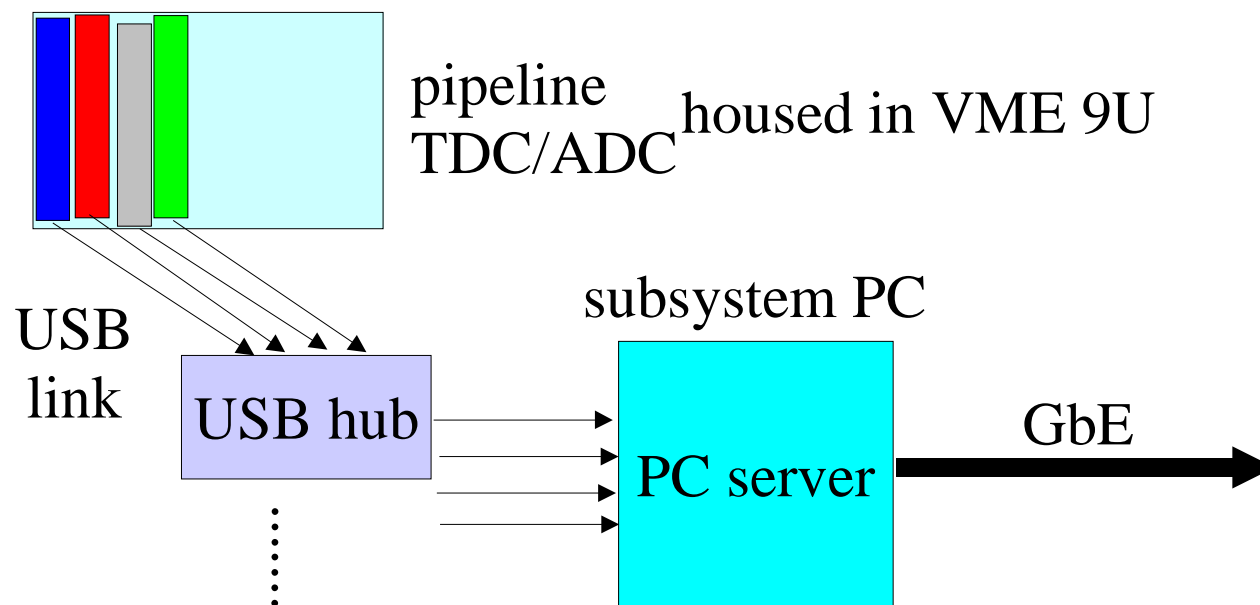
Target: Luminosity = $10^{35} \text{ cm}^{-2} \text{ sec}^{-1}$ in 2006

Change in the requirements

	Current	After upgrade
L1 trigger	100Hz(phys), 200Hz(BG)	1KHz(phys), 1–10KHz(BG)
Event Size	30KB/ev	100KB/ev (use of pixel detector, wave form sampling)
Storage speed	5–10MB/sec	300–500MB/sec
Storage size	~ 50TB/year	~1PB/year
CPU power	~1000 Pentium@1GHz	~10000 Pentium@4GHz

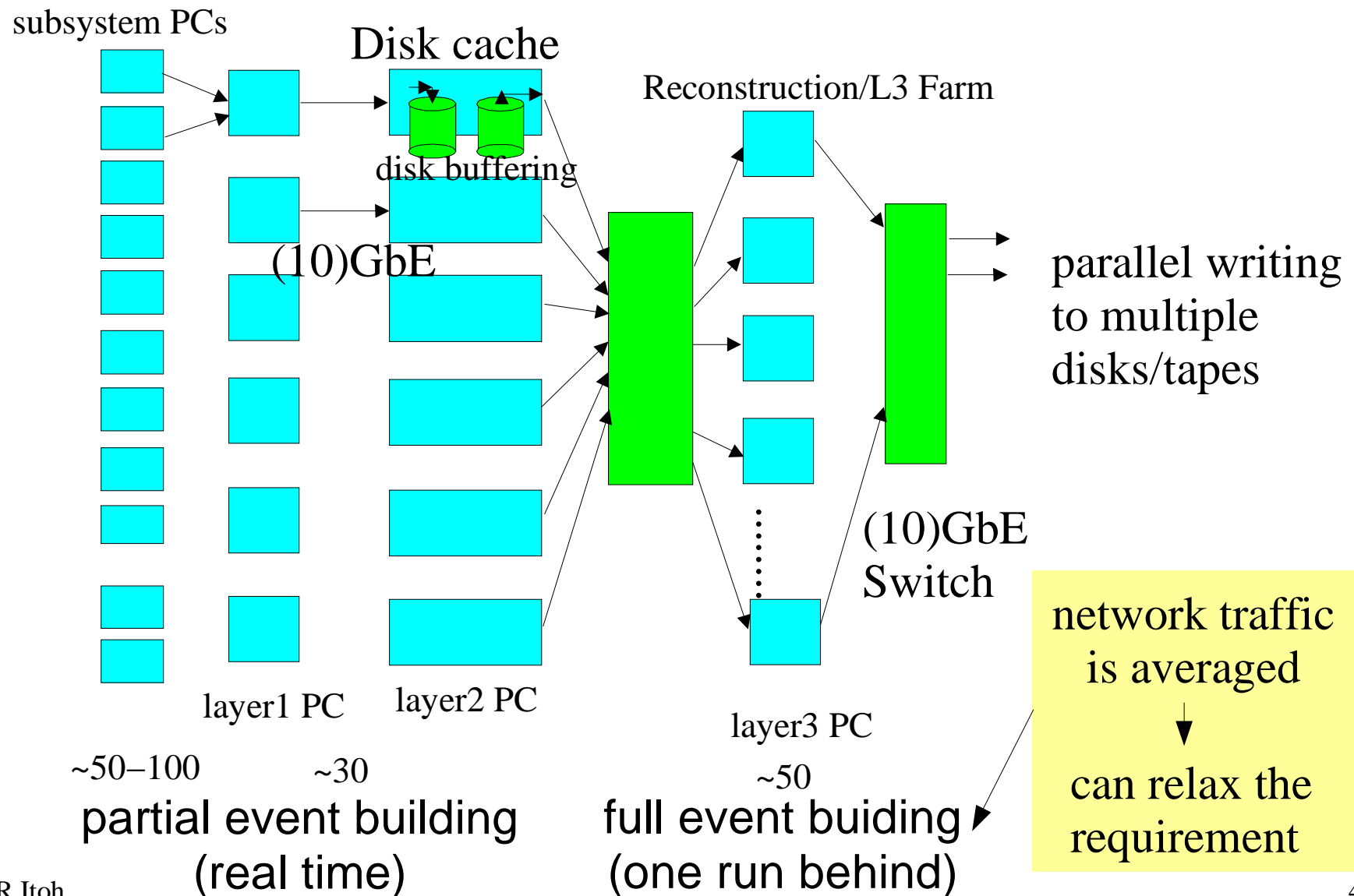
Possible design

- Frontend : * pipelined readout (TDC/ADC) is required.
 - * readout through serial link (USB2, IEEE1394...)
 - * VME still survives for packages



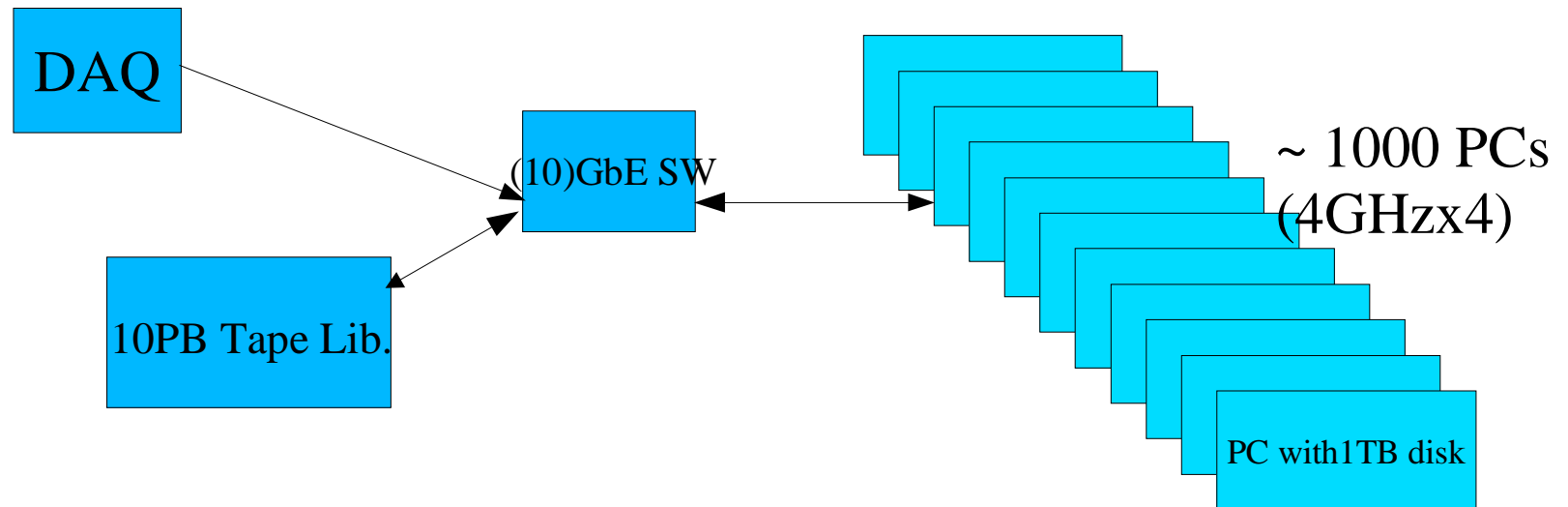
Event Building : * many ideas based on current system

example: Semi-offline event building



Data Analysis

- * Do we continue to use non-OO data management system?
 - Full OODB (Objectivity-like package) might not be used.
 - <– difficult considering man power
 - Event Stream I/O + Object handling in one event
 - "root" I/O like package?
- * Framework : we can still use dBASF with new event stream I/O
- * Computing system: "PC with data" (or "Data with PC")



6. Summary

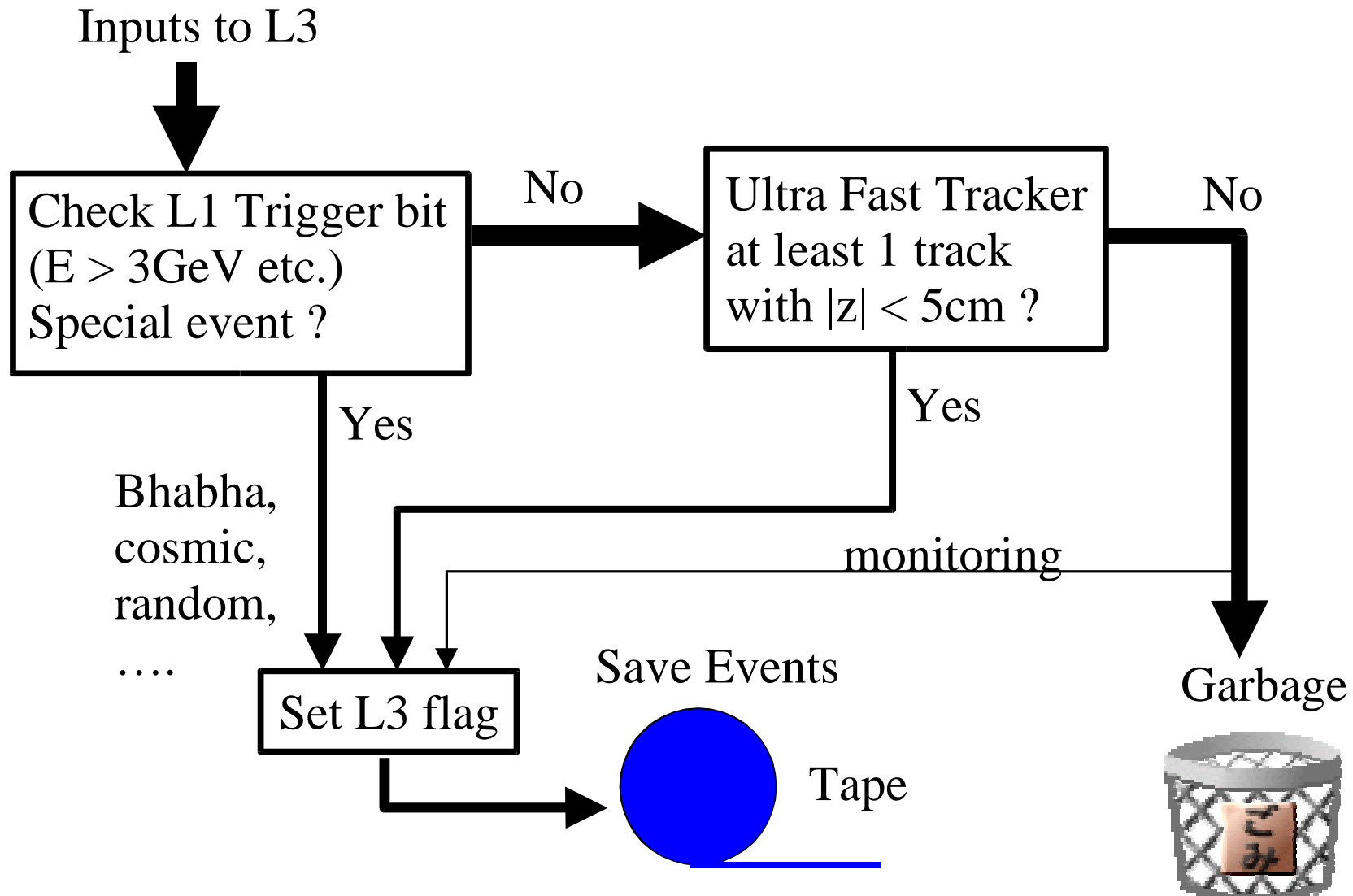


- The Belle Data Acquisition and Analysis Systems worked well in the previous runs achieving the expected performance.

Observation of CP violation in B meson system!!

- Upgrades of both systems have been made to prepare for the coming runs where higher luminosity is expected.
- The design of DAQ and Analysis systems for new high luminosity project has just been started.

Level 3 (L3) Trigger Data Flow



dBASf: consists of **Framework** and **Dataflow** and **Misc programs**

● **Framework** : client and a set of daemons running on different hosts

- * client : main program to start a job
- * daemons : invoke corresponding subprograms by receiving requests from client.
 - inputd : start "input_distributor"
 - outputd : start "ouput_collector"
 - basfd : start "B.A.S.F."
- * resource management : "resourced"
 - monitors the CPU loads and data transfer on each basfd nodes and dynamically changes the number of nodes
- * communication among client/daemons is done by NSM

● **Data Flow** : event-by-event data transfer over raw socket

- input_collector : read data file and distribute events to BASF running on multiple hosts.
- ouput_collector : receive events from BASF hosts and write them

● **Miscellaneous**: historam, stdout, user shared memory....