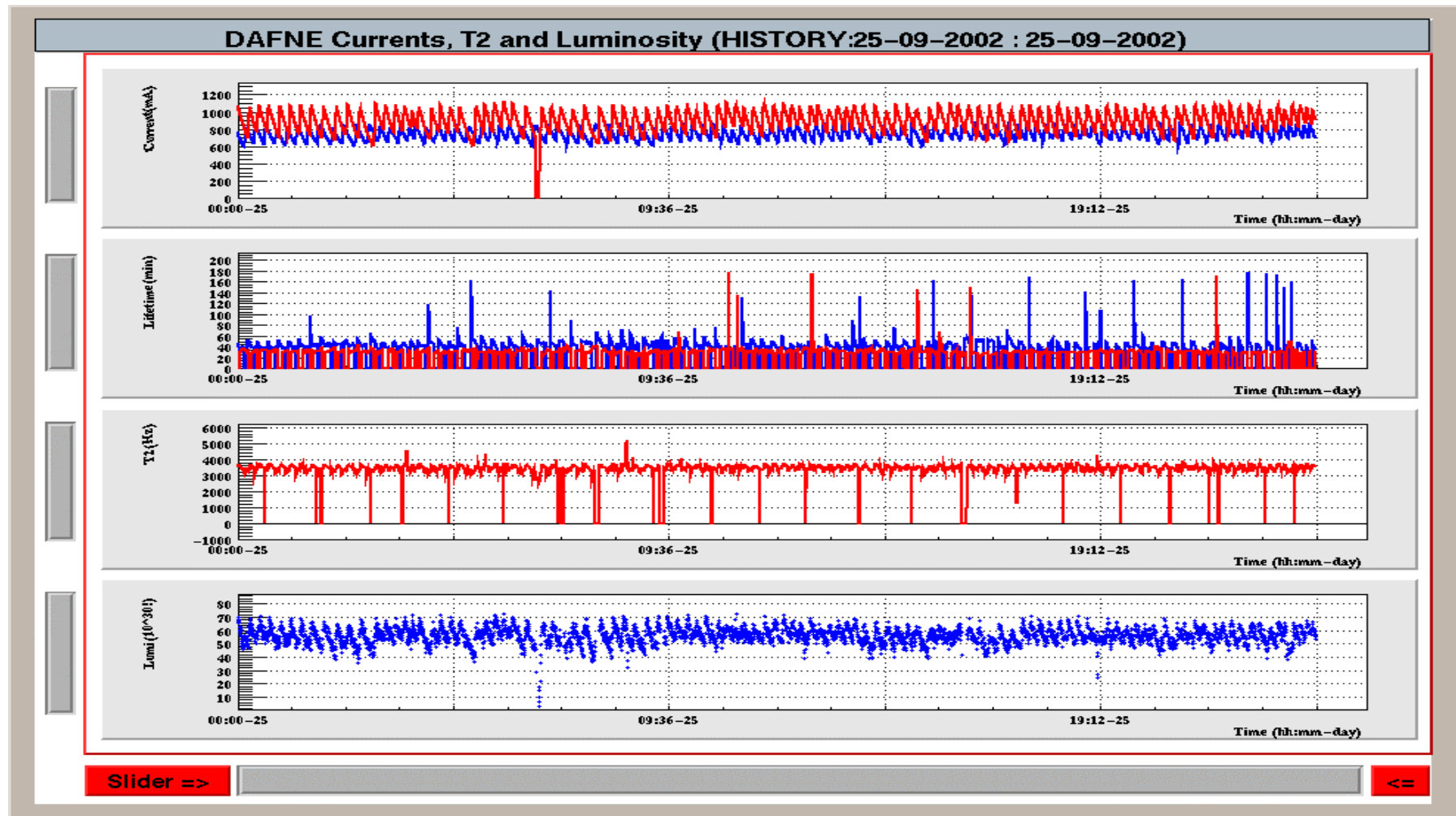


KLOE Computing

Paolo Santangelo
INFN LNF

Commissione Scientifica Nazionale 1
Perugia, 11-12 Novembre 2002

2002 – 3.6 kHz DAQ – 1.6 kHz T3



$$L_{\text{peak}} \sim 7 \cdot 10^{31} \text{ cm}^{-2} \text{ s}^{-1}$$

$$\langle L \rangle \sim 5.4 \cdot 10^{31} \text{ cm}^{-2} \text{ s}^{-1}$$

$$L_{\text{int max}} = 4.8 \text{ pb}^{-1} / \text{day}$$

on-line farm computers

1 run control

3 data acquisition

1 online calibration

1 data quality control

2 tape servers

1 database server (DB2)

500 SpecInt95

caption: IBM F50 (4 way 166 MHz PowerPC)
IBM H50 (4 way 332 MHz PowerPC)

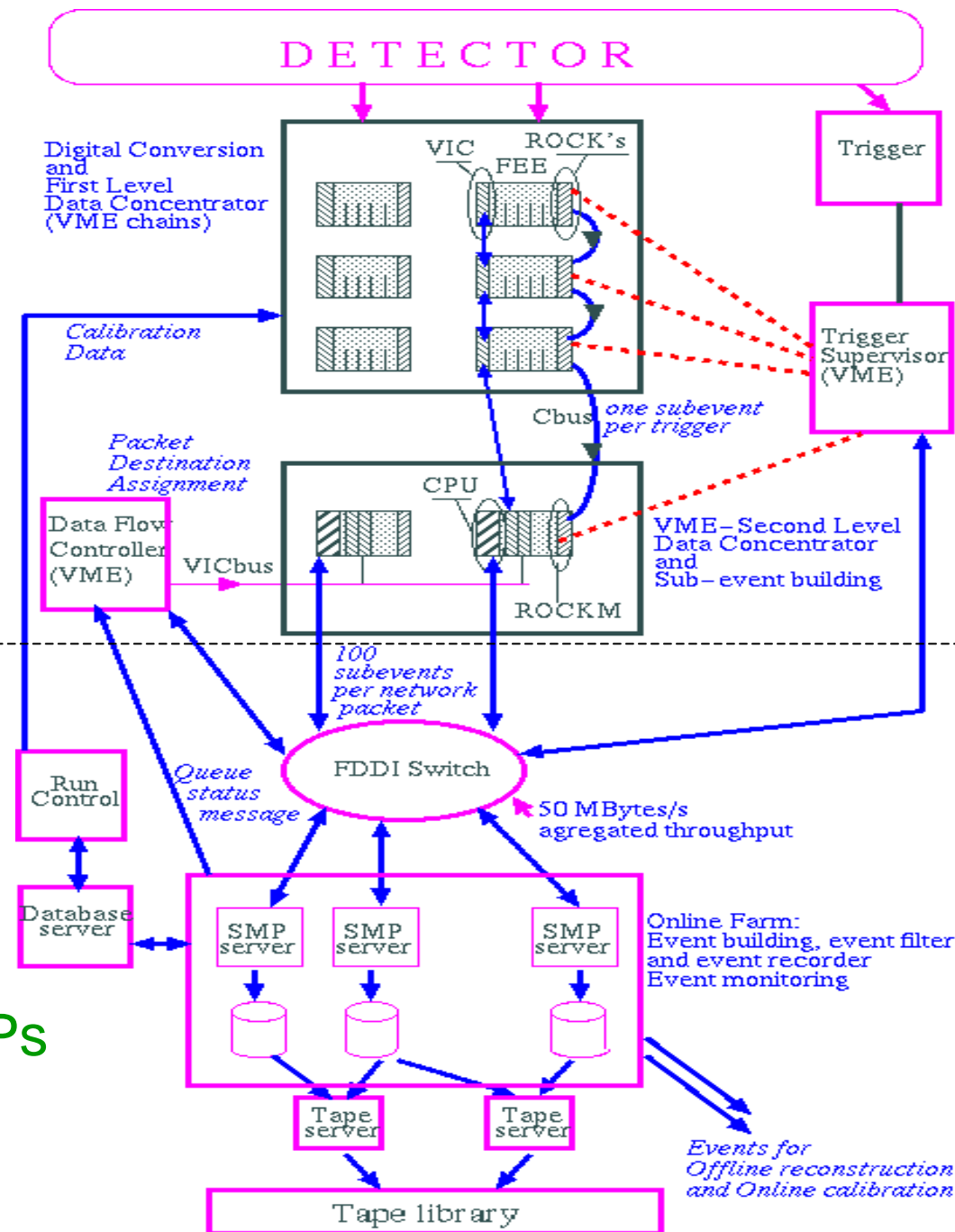
DAQ layout

10 L2 CPUs

FDDI

3-7 4 way SMPs

Fast Ethernet and
Gigabit Ethernet



FEE and L2 Processors

DAQ Computing

DAQ dataflow

- L2 processors
 1. collect detector data from VME
 2. send data to on-line farm computers
- on-line farm computers
 1. receive data from L2 processors
 2. build events
 3. filter events (L3, fast tracking rejects cosmics)
 4. write events to storage

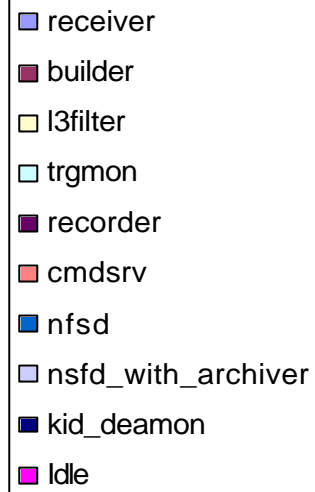
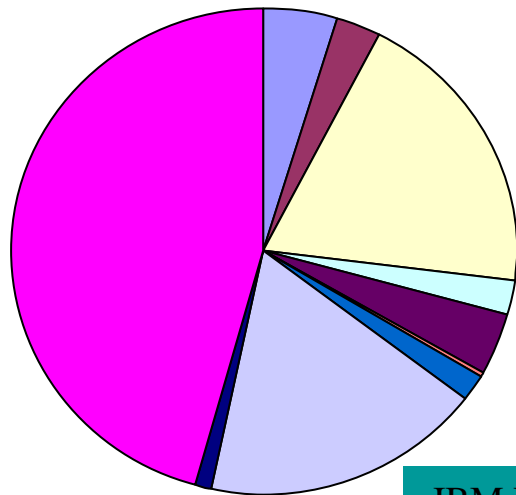
also

DAQ dataflow is sampled for data quality controls, calibrations, monitoring, event display

on-line farm

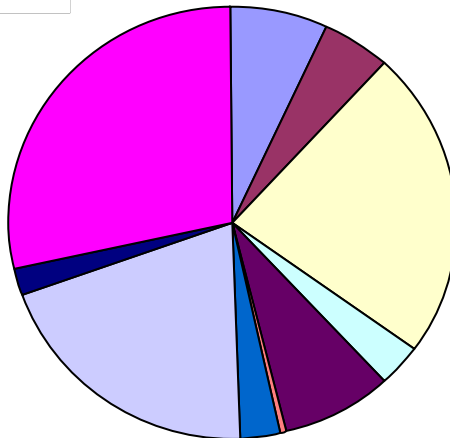
- processes not limited by processor speed
- unix *fixed priorities* for *DAQ processes*
quasi real-time OS
- DAQ rate scales with number of machines used
- with 3 (4 way) machines the rates are
up to 5 kHz of DAQ
now L3 filter limits DAQ output to 1.6 kHz
- 2-way Fast EtherChannel to processing/storage
tape drive speed is 14 MB/s

0.8 kHz / machine



IBM H50 4-way
58 Specint95

1.6 kHz / machine



event size 2.5 KBytes

2.4 kHz DAQ input
3 computers

each computer
4 way SMP

data moving
simultaneous
with smooth DAQ

processes are
compatible
with processors

4.8 kHz DAQ input
3 computers

data server and data processing nodes

2 disk and tape servers

2 *AFS servers*

2 *AFS clients (analysis)*

8 *montecarlo*

700 SpecInt95

40 processors

0.8 kHz nominal reconstruction rate

4 *AFS clients (analysis)*

28 *data processing*

4900 SpecFp95

96 processors

4.5 kHz nominal reconstruction rate

caption: IBM F80 (6 way 500 MHz RS64 III)

IBM H70 (4 way 340 MHz RS64 III)

Sun Enterprise 450 (4 way 400 MHz Ultra Sparc 2)

IBM B80 (4 way 375 MHz Power3 II)

long-term storage – tapes - hw

- tape library
 - 15 (+2) box long IBM 3494 tape library
 - 5,500 cartridge slots
 - dual active accessors
 - dual high-availability library control (standby takeover)
- 12 tape drives
 - 14 MB/s IBM Magstar (linear, high reliability)
 - presently 40 GB per cartridge (uncompressed)
 - upgrade to 60 GB per cartridge (ordered)
- safe operations
 - some cartridges mounted up to 10,000 times

long-term storage – tapes - hw

- full usage of investment protection
KLOE used a full generation of drive/media
from 10 -> 60 GB per cartridge
- what next ?
a new generation of drives and media
in the same library (year 2003)
higher track density (300 GB to 1 TB per cartridge)
tape length per cartridge, roughly expected constant
- expected costs for the new generation ?
cheaper tape drives
more expensive cartridges
total cost similar (in numbers of automated cartridges)

long-term storage – tapes - sw

- software
HPSS vs. ADSM and similar
- adopted: ADSM (now TSM)
low cost (no annual fee)
good performance
robust database
easy to install, easy to use
important developments (SAN, server free)
- transparent integration in KLOE sw environment
using TSM API

KLOE archived Data - October 2002

1999		raw	6 TB	GONE
2000	~20 pb ⁻¹	raw	22 TB	
		reconstructed	12 TB	
2001	~180 pb ⁻¹	raw	48 TB	
		reconstructed	37 TB	
2002	~288 pb ⁻¹	raw	35 TB	
		reconstructed	29 TB	
total			183 TB	

tape library capacity is presently 200 TB + compression
also used for MC, AFS analysis archives, user backups
upgrade to 300 TB (ordered)

disk space usage

- **DAQ (1.5 TB)**
5 strings - 300 GB each - RAID 0
can buffer 8 hours of DAQ data at 50 MB/s
- **disk and tape servers (3.5 TB)**
12 strings - 300 TB each - RAID 0
1+1 for reconstruction output
5+5 for data staging for reprocessing or analysis
- **AFS (2.0 TB)**
several RAID 5 strings
user volumes
analysis group volumes
- **all disks are**
directly attached storage
IBM SSA 160 MB/s technology

disk and tape servers

- two large servers are the core of the KLOE offline farm
 - several directly attached storage devices (plus GEth and others)
 - 12 Magstar E1A drives
 - 12 SSA loops, 96 x 36.4 GB SSA disks
- data moving speeds
 - aggregate server I/O rate scales with these numbers
 - 40 MB/s per filesystem
 - 40 MB/s per remote NFS v3 filesystem
 - 14 MB/s per tape drive
- client production is not constrained by server resources
- scaling with number of production clients
 - presently, up to 100 client processes use server data
 - more reconstruction power can be added safely

offline farm – software

- raw data production
 - output on a *per-stream* basis
 - makes reprocessing faster
- production and analysis *control software*
 - AC (FNAL's Analysis Control)
 - KID (KLOE Integrated Dataflow)
 - a distributed daemon designed to manage data
 - with data location fully transparent to users
 - tracks data by means of database information and the TSM API
 - example:
 - input ybos:rad01010%N_ALL_f06_1_1_1.000
 - input dbraw:(run_nr between 10100 and 10200) AND (stream_code='L3BHA')

reconstruction farm

- 24 IBM B80 servers
 - 96 processors
 - 4900 SpecFp95
 - 4-way 375 MHz Power3 II (4 x 51 Specfp95)
- delivers a maximum 5 kHz reconstruction rate
- 10 SUN E450 servers
 - 40 processors
 - 4 way 400 MHz UltraSparc II (4 x 25 Specfp95)
- processor performance
 - evaluated on the basis KLOE specific benchmarks
 - SPEC metrics, almost meaningless

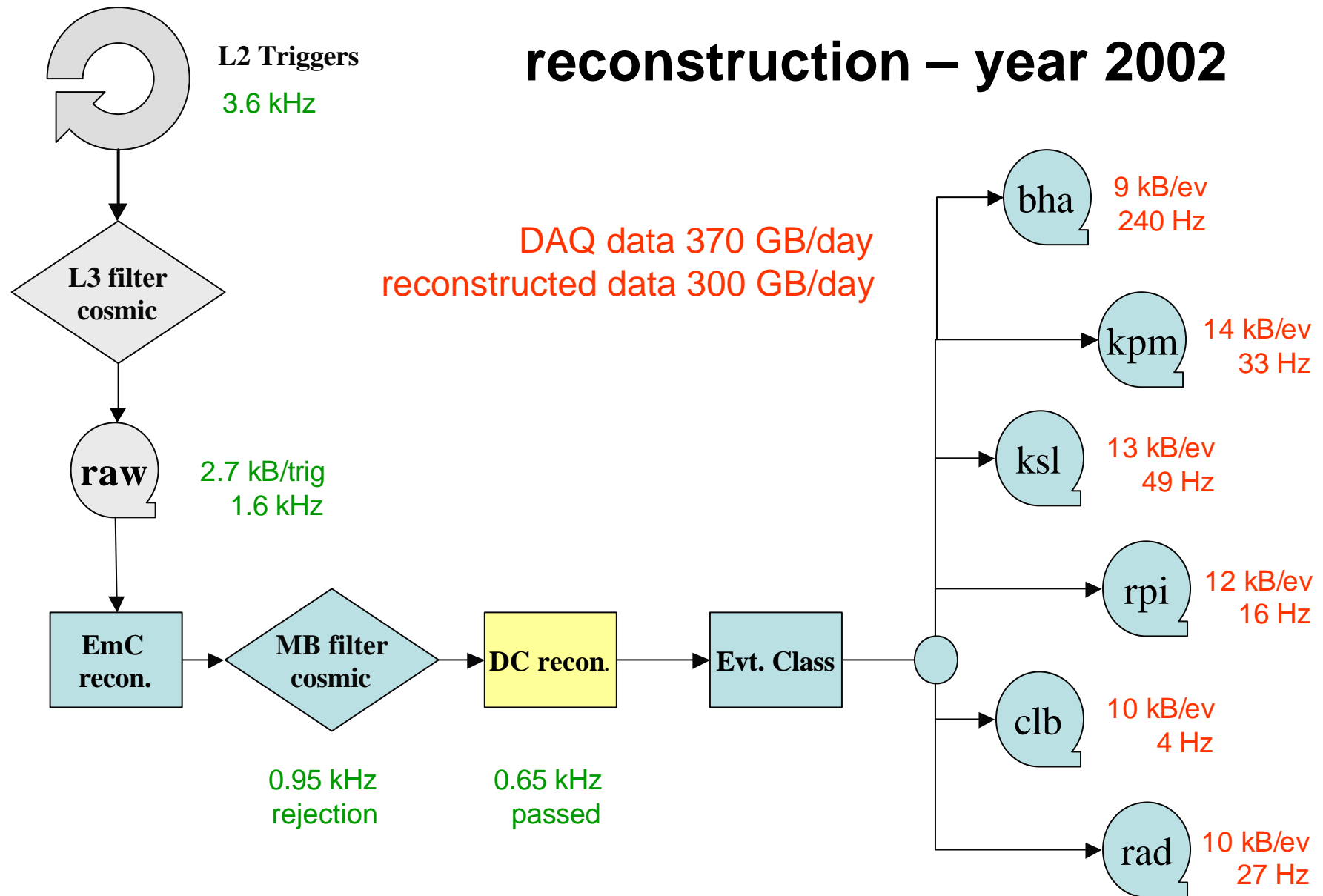
Processor Comparison for KLOE Tasks

	IBM Power 3 375 MHz	IBM Power 4 1 GHz	Sun ES 450 450 MHz	PentiumIII 1 GHz	Athlon XP2000+
--	---------------------------	-------------------------	--------------------------	---------------------	-------------------

ms/trigger					
data, full reconstruction	17	8	40	29	18

ms/event					
data, tracking only	24	12	66	53	32
MC-1 $\pi^+\pi^-$ production	210	110	420	270	150
MC-2 $K_S^0 \rightarrow \pi^+\pi^-$	120	60	240	160	90
MC-1 reconstruction	70	35	170	130	76
MC-2 reconstruction	120	60	280	210	123

reconstruction – year 2002



trigger composition and reconstruction timings

??+ Bha	background		
	filtered	tracked	total

triggers	4%	74%	26%	96%
reconstruction time	63 ms	1 ms	51 ms	14 ms
	16%	4%	80%	84%

year 2000

physics is a tiny fraction

computing is used for tracking of background events

triggers	11%	67%	33%	89%
reconstruction time	63 ms	1 ms	50 ms	17 ms
	31%	3%	66%	69%

year 2001

DA? NE gives more physics

triggers	23%	78%	22%	77%
reconstruction time	63 ms	1 ms	33 ms	8 ms
	70%	3%	27%	30%

year 2002

physics is now 23 %

computing is now used for useful physics

KLOE data taking conditions and CPUs for data processing

year	trigger rate, Hz	luminosity $10^{31} \text{ cm}^{-2} \text{ s}^{-1}$	$\mu\mu$ Bhabha Rate, Hz	data taking DAQ hours/pb ⁻¹	data recon. hours*CPU/pb ⁻¹	total Gb/pb ⁻¹
2000	2100	0.9	77	33	970	1500
2001	2000	2.4	220	11	520	470
2002	1600	4.1	375	6.8	230	210
2003	2150	10.0	920	2.7	190	145
200x	5800	50.0	4600	0.6	167	115

extrapolated assuming 2002 background and trigger conditions

nominal processing power for concurrent reconstruction (in units of B80 CPUs)
is 34, 70 and 300 CPU units for years 2002, 2003 and 200x respectively

these numbers do not include the sources of inefficiencies, MC production and concurrent reprocessing

CPU power for data processing and MC generation

1 fb ⁻¹	reprocessing from raw data	reprocessing from streamed data	MC $\gamma\gamma$ decay	
day CPU	9600	kpm 1440	simulation	6650
		ksl 1142	reconstruction	5100
		rad 198		
		bha 1440		
		4220		11750

these numbers do not include the sources of inefficiencies

data volume for data and MC samples

1 fb ⁻¹	raw data	reconstructed	DSTs	MC files	MC DSTs
	115 TB	90 TB	10 TB	83 TB	20 TB

- using 2002 background and trigger conditions
- all numbers refer to a sample of 1 fb⁻¹
- day CPU number are in units of B80 CPUs

KLOE database (DB2)

- present database size larger than 2 GB
 - runs and run conditions (20 kfiles)
 - raw data file indexing (160 kfiles)
 - reconstructed data file indexing (640 kfiles)
 - 100 kB per run
 - 2.5 kB per file
- almost no manpower needed to operate DB2
- reliability
 - augmented by a semi-standby and takeover machine
 - on-line backups at full DB level
 - on-line fine time-scale backup by means of archival of DB logs
- also
 - a minimal hardware
 - no cost DB for academia (IBM Scholars Program)

networking

- Networking and optimizations
 - FDDI
 - GigaSwitch (L2 to on-line Farm)
 - CISCO Catalyst 6000
 - Ethernet (on-line and production farm)
- Gigabit Ethernet at KLOE
 - server bandwidth
 - 100 MB/s with Jumbo Frames (9000 byte MTU)
 - FEth client bandwidth usage from a single GEth server
 - flattens at 70 MB/s for more than 6 clients at 10 MB/s each
 - all numbers double in full duplex mode
- networking and related optimizations
 - simple IP and TCP tuning
 - other TCP tuning for complex bandwidth allocations (in progress)

remote access

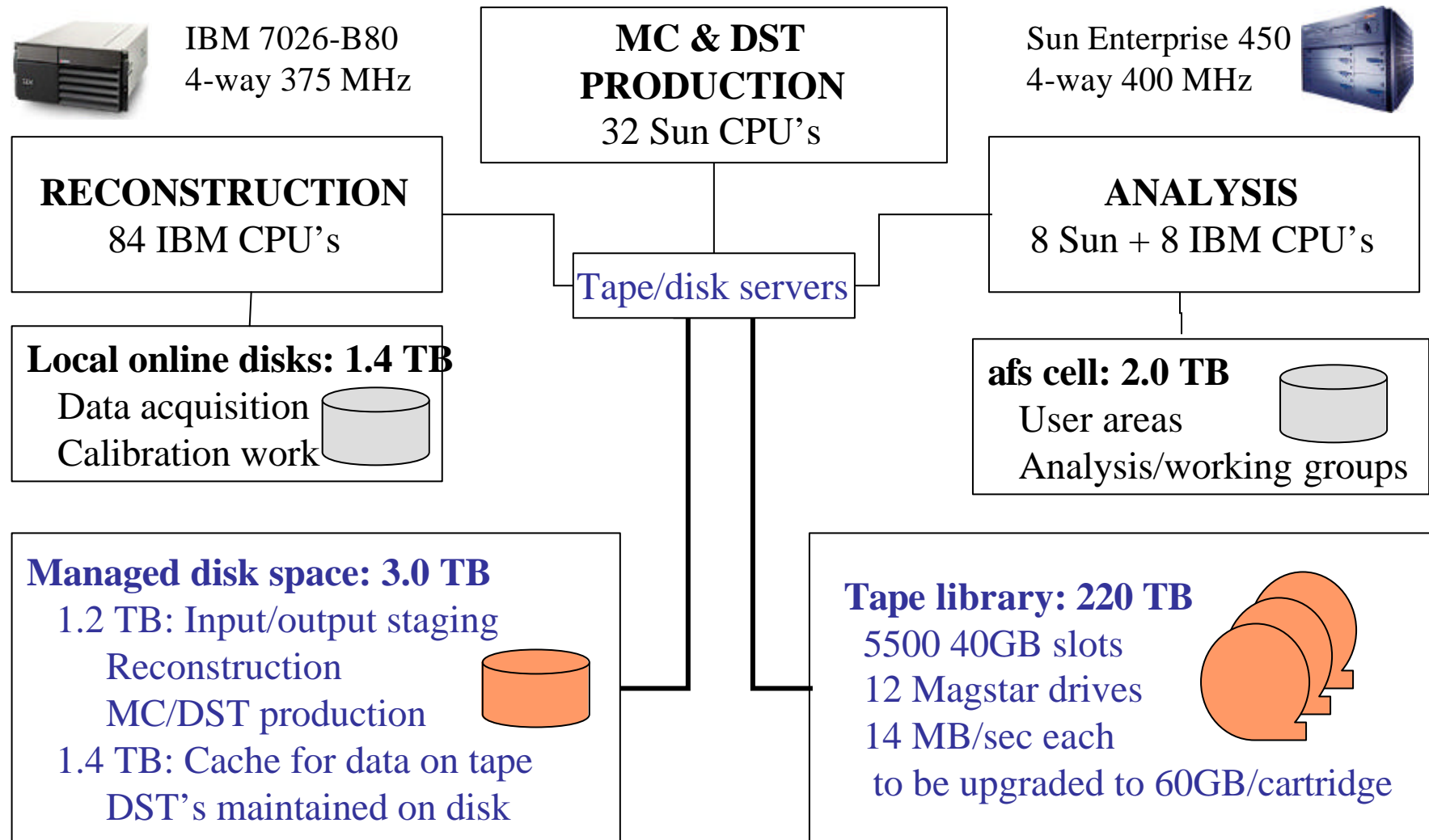
- remote computers
 - can access KLOE data
 - AFS data serving at the core of KLOE analysis
 - raw & reconstructed data managed and served by KID
 - metadata managed by the KLOE DB2 database
- AFS demonstrated and operated with
 - large server volumes (up to 100 GB)
 - high server throughput (20 MB/s per disk string)
 - high client performance (8 MB/s with FastEthernet)
- but end-of-life announced for AFS ...

conclusions

- KLOE computing runs smoothly
- uptime only constrained by external events
- hardware will be upgraded for 2003 data taking
 - +1 tape library (+1 PB)
 - +10 TB disk space
 - +80 CPU power

Backup Slides

offline computing resources



data reconstruction for 2002 data taking

