PHYSICS overview and requirements

The KLOE experiment is programmed to run in fall 1997, at the Frascati Φ-factory DAΦNE

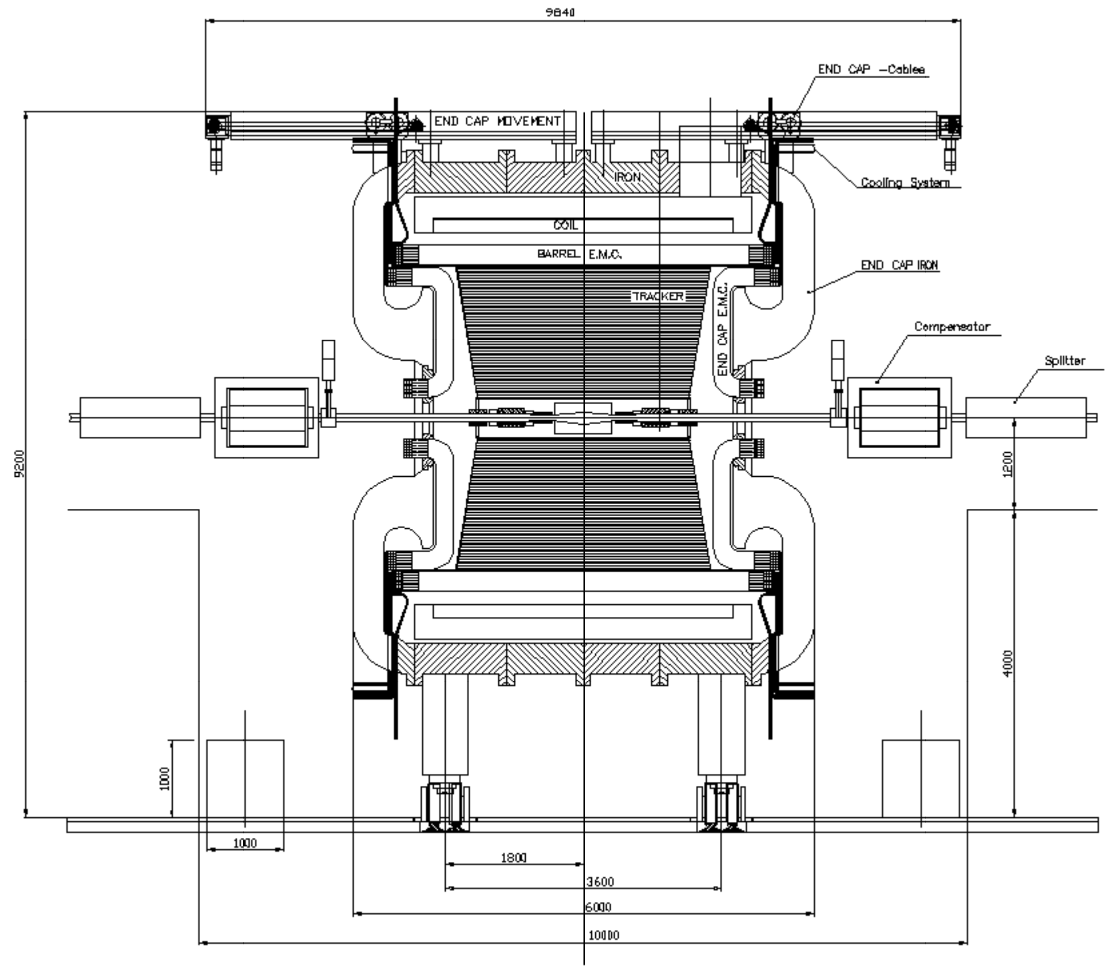
$$e^+e^- \to \Phi \to KK$$

- To study direct CP violation looking at final states involving 4 pions

- requires:
  - high purity $K_S - K_L$ event samples
  - accuracy in the measurement of $\mathcal{R}(\epsilon'/\epsilon)$ of $\sim \mathcal{O}(10^{-4})$

- imposes stringent requirements on:
  - Detector performance
  - Trigger system (inefficiencies $<$ few x10^{-3} and redundancy to measure inefficiencies)
  - DAQ (democracy of trigger in all channels)

The KLOE Detector



- 4 m diameter and 4 m long, B=6 kG
- maximum drift time of 2μs
- 25.000 electronic channels
- FEE, READOUT and TRIGGER under test
- Detectors under construction, end 1996 DAQ

The Trigger

- DAΦNE bunch crossings time $\sim$ 2.7 ns
- DAΦNE target luminosity $10^{33} cm^{-2} s^{-1}$
- 5 kHz of ϕ's produced
- expected background: Bhabha events, cosmic rays and machine background
- Trigger must operate in a continous mode
- Two level scheme based on
  - local energy deposition in Electromagnetic Calorimeter (E.M.) sectors
  - multiplicity information from the Drift Chamber (D.C.)
- First level:
  - is produced within 150 ns
  - uses E.M. and early D.C. multiplicity
  - good timing information to start the FEE operation (E.M. TDC's).
  - rate of about 25 kHz (simulations)

– dead-time of 2 μs $\Rightarrow$ 5% total dead-time

– can be vetoed if the event was identified as a Bhabha or a cosmic ray event

- Second level:
  – uses additional information collected from the D.C. within $\sim$ 850 ns and E.M.
  – no reset trigger signal but validation
  – 2 μs after physics event, no dead time
  – originates local trigger numbering on each ADC, TDC, ROC and copy on local buffers

Events	L1 eff. (%)	L2 eff. (%)	Events	L1 (kHz)	L2 (kHz)
$\pi^{\pm,\pm}$	99.8	99.6	Total ϕ	5	5
$\pi^{\pm,00}$	99.9	99.4	Bhabha	10	3
$\pi^{\pm,\pm}$	99.9	99.8	Machine	9	5
$\pi^{\pm,\pi\ell}$	99.9	99.7	Cosmics	1	1
			Total	25	12

Trigger efficiencies and expected rates.

- Simulation program: GEANFI (a modified version of GEANT including new/better treatment of the KLOE physics)
- The expected physics event size is 5 kByte.
- The requirement of sustained throughput in KLOE is about 50 MByte/s

- Trigger supervisor and FEE interface are defined
  - every TDC, ADC, ROC generates local (12 bits) trigger number
  - synch cycles (4 μs) are periodically generated to check trigger number consistency
  - EVENT integrity is assured by generating early busy signal in the FEE boards when only space for a new event is available

The Readout system
(presented by A. Aloisio)

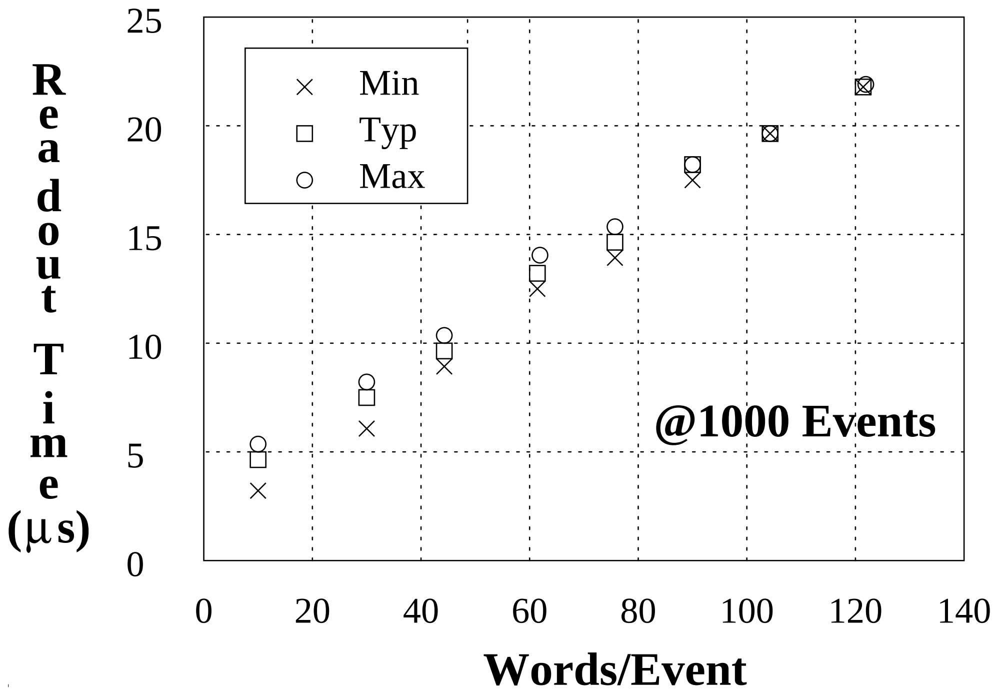
- custom system, VME based
- 1st level data concentration (L1-VME)
  - auxiliary bus AUX-bus interconnection
  - the Readout controller for KLOE (ROCK)
  - up to 16 ADC/TDC modules (30 or 96 channels/module)
  - sparse data scan operations
  - hardware zero suppresion
- 2nd level of data concentration (L2-VME). No CPU
  - a ROCK Manager (ROCKM)
  - a custom C-bus to interconnect up to 8 ROCK's in a chain
    * up to 10 MHz Data Transfer rate
    * 10 m total cable length
    * 27 bits data bus $\rightarrow$ $\sim$40 MB/s

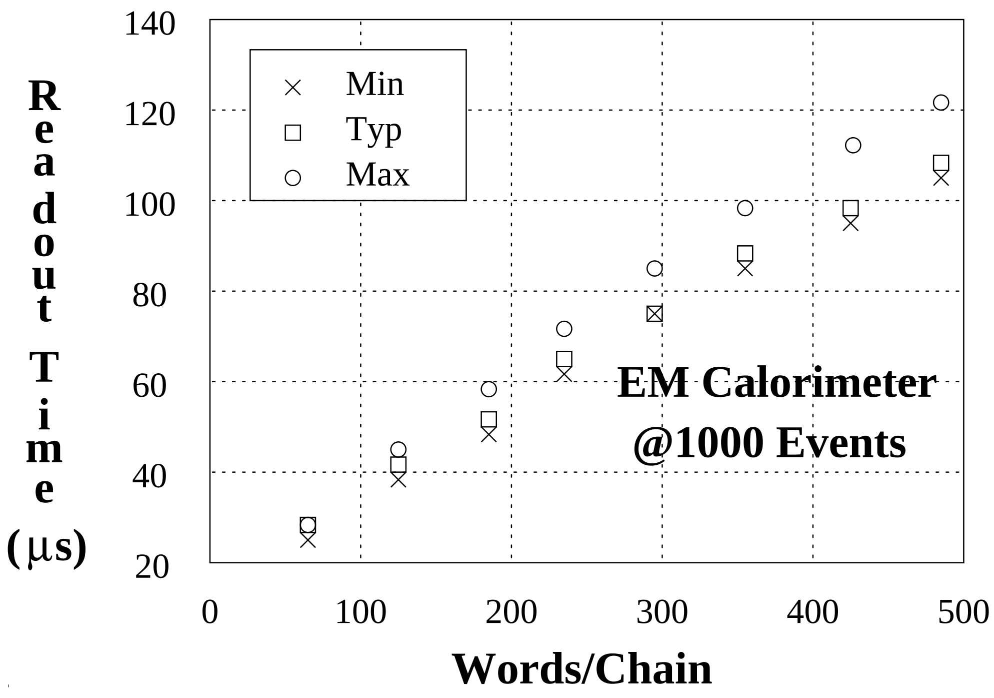
- ROCK and ROCKM have local buffering of 16 and 32 kBytes respectively
- work asynchronously to the trigger signal. Trigger numbers are memorized in local FIFO's
- trigger misalignment detection
- produce a first derandomization of data
- data integrity check
- minimum software overhead & complexity

C-bus is unidirectional (simplex) link

- L1-VME crates in a chain are interconnected via VIC-bus (from CES) for initial settings and monitoring
- L2-VME crates are interconnected via VIC-bus for Data Flow control

AUXbus and C-bus simulated performance, E.M.





- 10 kHz trigger rate (100 μs)
- maximum occupancy of 400 words/chain randomly distributed
- a chain can handle 15 MB/s bandwidth
- maximum occupancy at ROCK FIFO (simulation) is 6 kB

- Every chain in KLOE is configured (variable chain length) to sustain a 5 MB/s throughput $\Rightarrow$ 10 chains (50 L1-VME crates)
- $\frac{1}{2}$ of the FDDI nominal throughput

Using 10 chains connected to the switch

- The KLOE Readout can support 150 MB/s
- The KLOE Event Building can support a throughput of ~75 MB/s using communication interfaces in the market now.

KLOE uses the FDDI GIGAswitch from DEC (a switched bridge)

Data Flow Control

- Data of the Sub-event are memorized in the ROCKMs for each trigger
  - 0.5 kB per event in average
  - empty sub-events have only a two words record
- CPUs in the L2-VME crates collect sub-events in a local circular buffer
- pack N ($\sim$ 100) of them (50 kB) (N defined by DFC protocol)
- send every packet using TCP/IP to a defined processor in a Subfarm

- The Data Flow Control is in charge of the source - destination assignment
  - assuring EVENT integrity
  - optimizing traffic in the switch (consecutive packets sent to processors connected to different switch ports)

– reducing packet losses because buffer full at receiver side

– TCP/IP overcome this problem by negotiating window size

– Implementation:

  * Network map (processor and switch port list) known at the start of run

  * Ordered list of processors for distribution of traffic in a static table (10 kHz trigger, 100 sub-event/packet, 10 chains $\Rightarrow$ a new event per port every 0.1 sec)

  * Tables containing ordered list of processor flags (enable/disable) according to the queue status (maximum trigger number for table validity)

  * Shared memory (VIC-bus) to distribute tables to CPUs

- Study of the DAQ network as a Telecommunication problem
- The Simple Network Management Protocol (SNMP) is largely used to manage network devices
- managed devices are defined in standard MIB
- proprietary MIBs can be developed for new hardware/software objects
- Implementing DFC, SNMP is used for
  - Internet to MAC addresses correspondence (ARP-server)
  - Network mapping using GIGAswitch Digital MIBs
  - Identify the role of the processor in the DAQ (private KLOE MIBs)

- SNMP version 1 is used
- Software from Public domain (CMU) implementing
  - agents
  - utilities for set/get variables and walking in the MIB tree

- For Run Control commands distributions, specialized agents are under developing for
  - distribute commands between processors (setting MIB variables)
  - obtain acknowledgements (getting MIB variables)

The Event Builder

- based on the FDDI GIGAswitch from DEC
- the switch (FDDI bridge) was ready more than 2 years ago
- fullfil the KLOE requirement of total throughput
- FDDI is a stable network solution allowing studies of the general philosophy of the Event builder system and full test
- migration to ATM is possible (easy)
- workstations from DEC (AXP 3000/600) implement TCP/IP obtaining more than 90% throughput of the physical channel speed
- FDDI is a complete standard where everything can be tested and optimized NOW
- full defined network channels are late

Comparison between FDDI and ATM GIGAswitches from DEC

	FDDI	ATM
Throughput	100 Mbps	155 Mb/s 622 Mb/s
Ports	$\leq$22, 2 prts/brd $\leq$34, 4 prts/brd	$\leq$52, 4 prts/brd
Slots	11	13
Total thr-put	$\leq$3.6 Gbps	$\leq$10.4 Gbps
Switch	by packet	by cell 48/53 bits
Queues	input/output	input
Que. services	FIFO	first free path
Prt reservat.	none	Credit mechanism
TCP/IP	11 MB/s	16 MB/s

Credits mechanism from DEC eliminates cell retransmission $\Rightarrow$ increasing performances

TCP/IP over FDDI and ATM: throughput and CPU cost on DEC 3000/600, 114 SPECint92

	FDDI	ATM
Channel speed (Mb/s)	100	155
Phys. troughput (MB/s)	12.5 MB/s	17.5 MB/s
TCP/IP (MB/s)	11.5	16.5
UDP (MB/s)	12.5	16.5
CPU %	50-60	70-80
SPECint92	~60	~85

- TCP and UDP require similar resources on CPU
- Estimation of CPU cost at 622 Mb/s ATM (70 MB/s) is 350 SPECint92
- HP stations offer now equivalent performances on FDDI

- FDDI is a complete standard where everything can be tested and optimized NOW
- full defined network channels are late
- optimization of the TCP/IP protocol is possible
  - increasing the window size (now ∼43 kB LynxOS, ∼57 kB HP-UX, 64 kB OSF/1)
    * already proposed as standard (RFC 1323: TCP Extensions for High Performance), already implemented on OSF/1
    * will also be implemented in HP-RT 2.1, allowing TCP windows up to 4GB.
  - reducing the calculation of CRC in the CPU
  - reducing the number of copies in memory
  - increasing the memory bandwidth
- waiting for standard reliable UDP (=TCP??)

Optimize TCP is a good choice and of general interest independently of the physical channel

TCP/IP based Event builder requires multiple connections active in any processor:

- L2-VME CPU $\rightarrow$ A network processor in any subfarm ($\sim$10)
- Network processor in a subfarm $\rightarrow$ any L2-VME CPU, Trigger and DFC system ($\sim$12)
- access to TCP/IP through the Berkeley socket interface (like file descriptor)
- Software tools to maintain multiply connections (HP-UX, HP-RT, DEC-VMS, DEC-OSF/1, LynxOS): two methods tested
  - SELECT CALL
  - POSIX-Threads/Process
- Creating portable code mechanism
  - maintain multi-platform code on NFS mounted disks
  - only ONE set of source files
  - platform (or flavor) specific directories

– SELECT CALL:
  * a single process/thread doing the job
  * block until there is something to read or write
  * socket can be set in non-blocking mode
  * timeout available
  * loop about file-descriptors
  * INCOMPLETE string of events will originate timeout and Trigger busy
– POSIX-Threads:
  * a thread for each connection running independently using the first free socket
  * mutex lock/unlock for file descriptor list

– Results increasing connections (up to 60)
  * total throughput independent of number of sockets / connections
  * maximum throughput obtained with small window size (8 kByte)
  * no differences using select o thread

CPU and VME

- HP-RT, LynxOS, VxWorks, OSF/1 with real time extensions
- UNIX platform choosen
- Dificulties in using different vendors
  - different methods mapping IO bus into virtual user address
  - no standardized DMA for IO bus VME callable from user level

- No advantages in KLOE using Real-time operating systems
  - real-time implemented in hardware
  - better network performances using HP-UX on HP743rt
  - more difficult to port applications
  - communication drivers from different vendors

Buffer manager

- Circular Buffers (DAQ test stand)
  - FIFO for multi-producer and single-consumer
  - only one memory copy (shared memory)
  - no handshake sender - receiver
  - semaphore locking time $< 1\ \mu$s
  - $>$ 120.000 events/messages)/s
  - used between ROCKM and network, distributing events to analysis processors, collecting analysed events for storage
  - usable for message system
- Monitoring Buffer
  - maintain most recent events
  - single task never gets same event twice
  - demand driven, transfer only prerequested events
  - no blocking semaphores
  - minimized task switching
  - $\sim$ 100.000 events/s