



Data Acquisition Systems and Mass Storage for Experiments at SuperCollider

P. Vande Vyvre - CERN/EP

Workshop on Innovative Detectors for Supercolliders

Erice September 2003



DAQ for Super Collider Experiments

- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ Technology trends
- ◆ DAQ and HLT for SLHC experiments
- ◆ R&D
- ◆ Conclusions

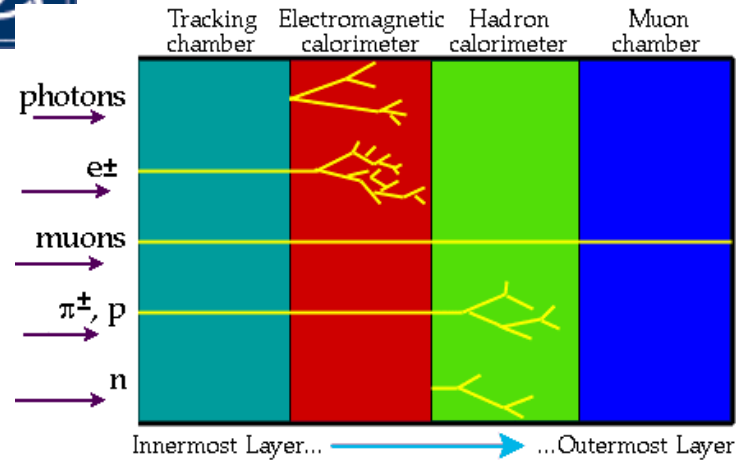


DAQ for Super Collider Experiments

- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ Technology trends
- ◆ DAQ and HLT for SLHC experiments
- ◆ R&D
- ◆ Conclusions



Trigger

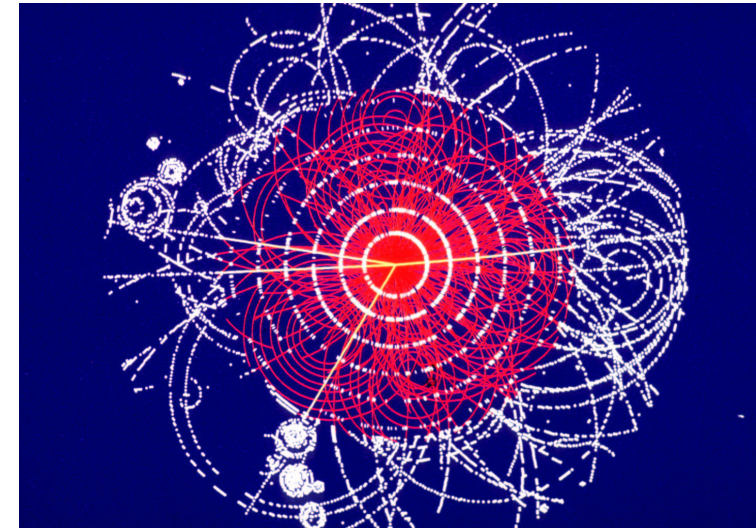
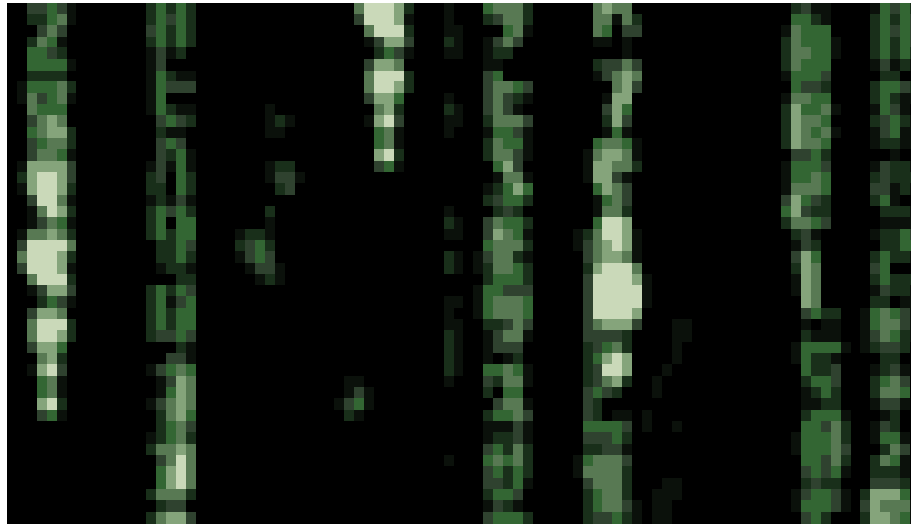
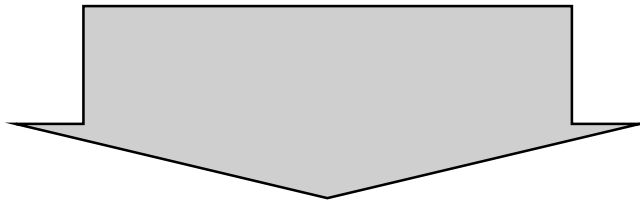


Multi-level trigger system

Reject background

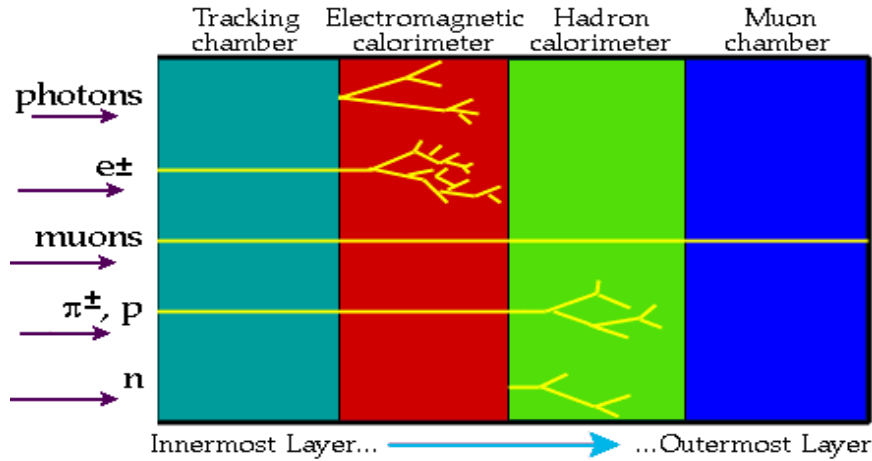
Select most interesting collisions

Reduce total data volume

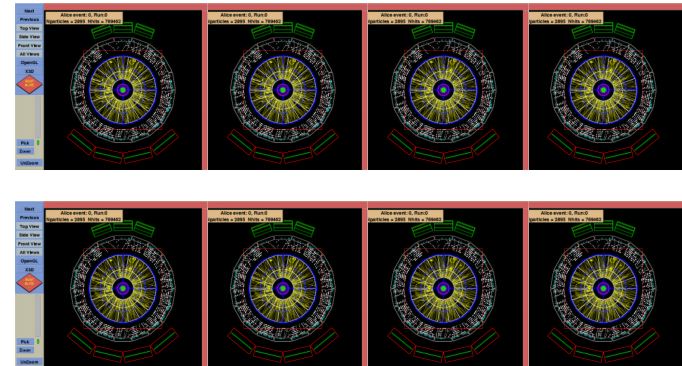
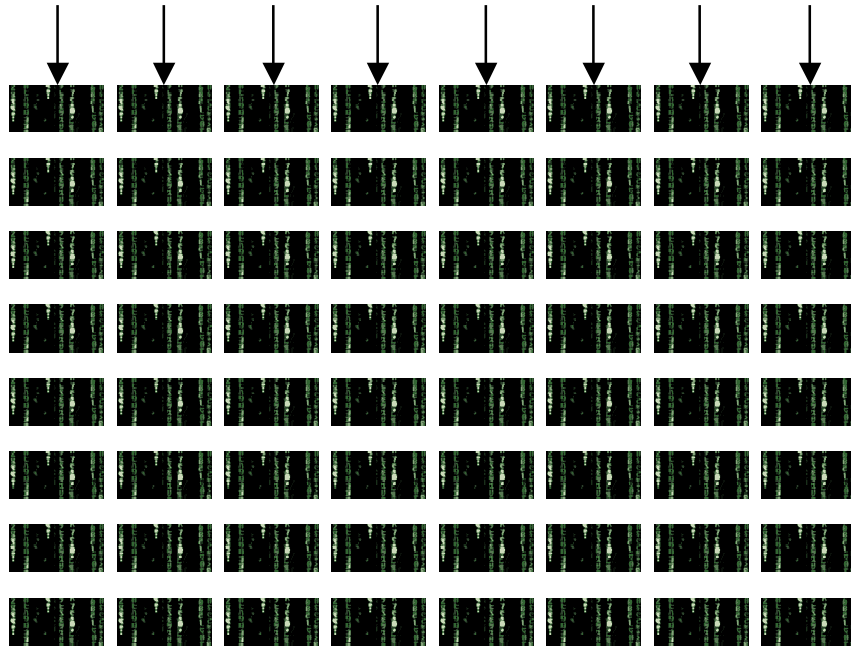




Data acquisition



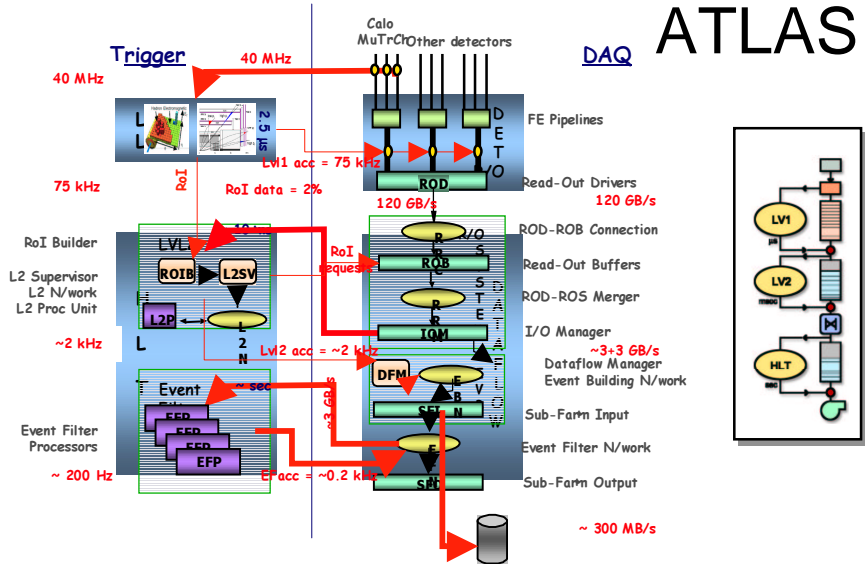
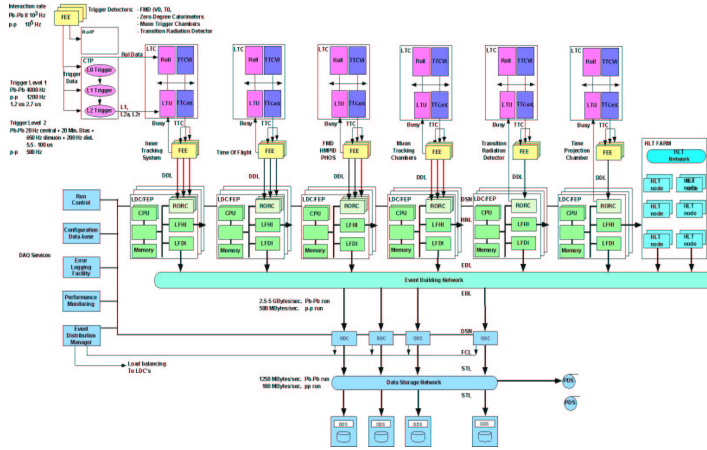
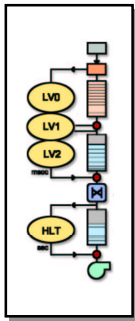
Acquire data from 1000's of sources
Reassemble all the data of same event





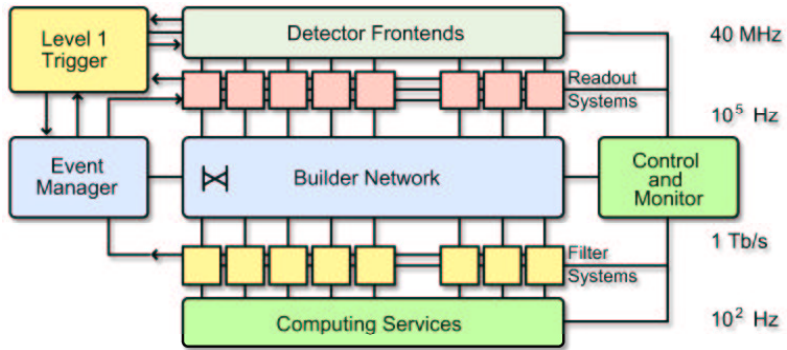
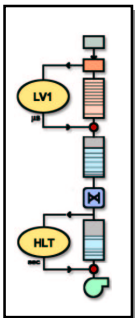
TRG/DAQ/HLT @ LHC

ALICE

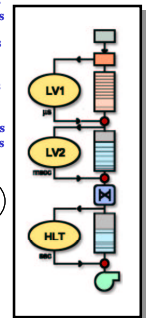
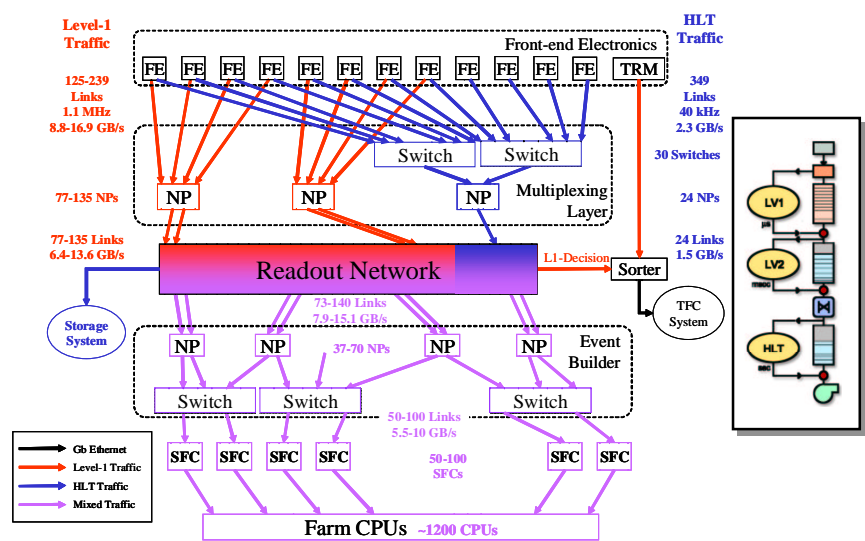


ATLAS

CMS

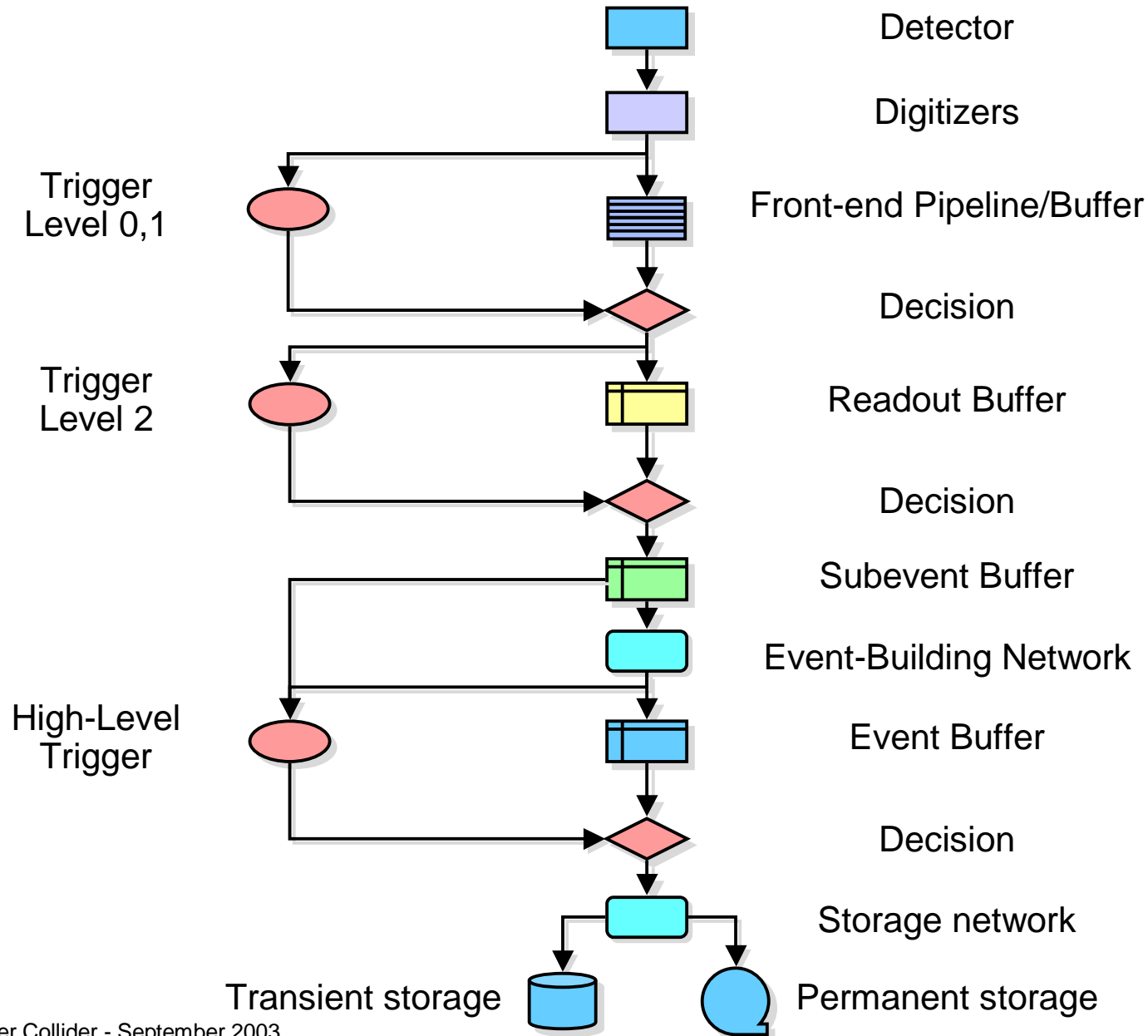


LHCb





Reference TRG/DAQ/HLT



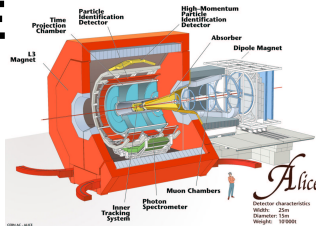


TRG @ LHC (1)

Trigger Levels

Rate First Level Trigger (Hz)

ALICE

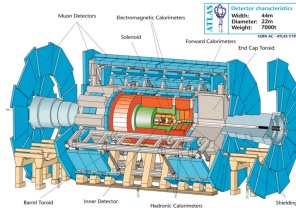


4

Pb-Pb
p-p

6×10^3
 10^3

ATLAS

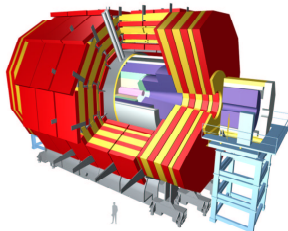


3

L 1
L 2

10^5
 2×10^3

CMS

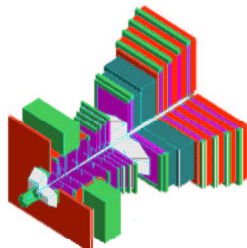


2

L 1

10^5

LHCb



3

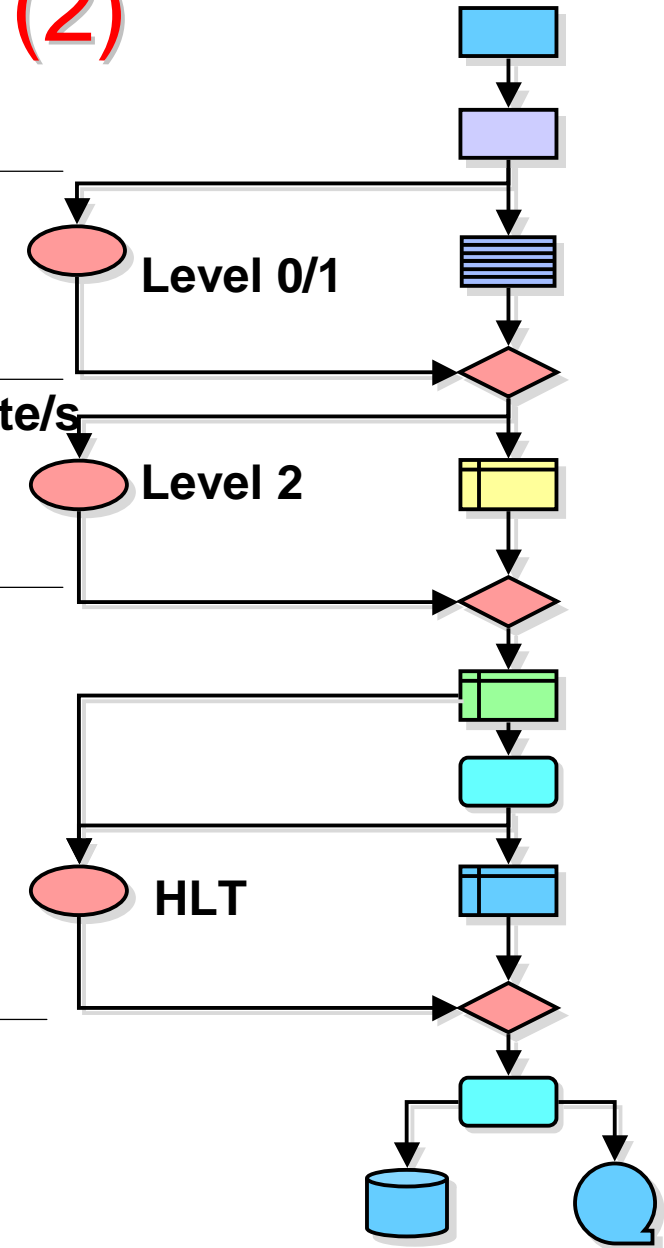
L 0
L 1

10^6
 4×10^4



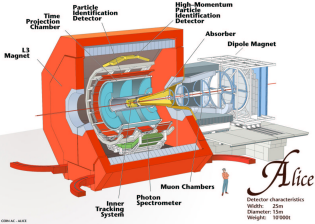
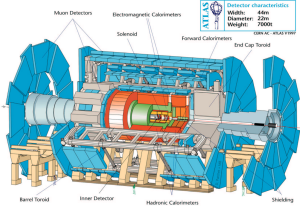
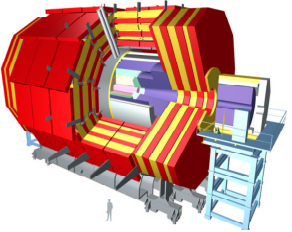
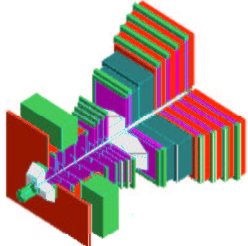
TRG @ LHC (2)

ALICE	ATLAS	CMS	LHCb	
40	40	40	40	MHz
0.9/5.2	2.5	2.5	4/<2000	μ s
6	75	100	1100/40	kHz
	120			GByte/s
0.08	10			ms
2	2			kHz
~200	~100	~100	~200	Hz





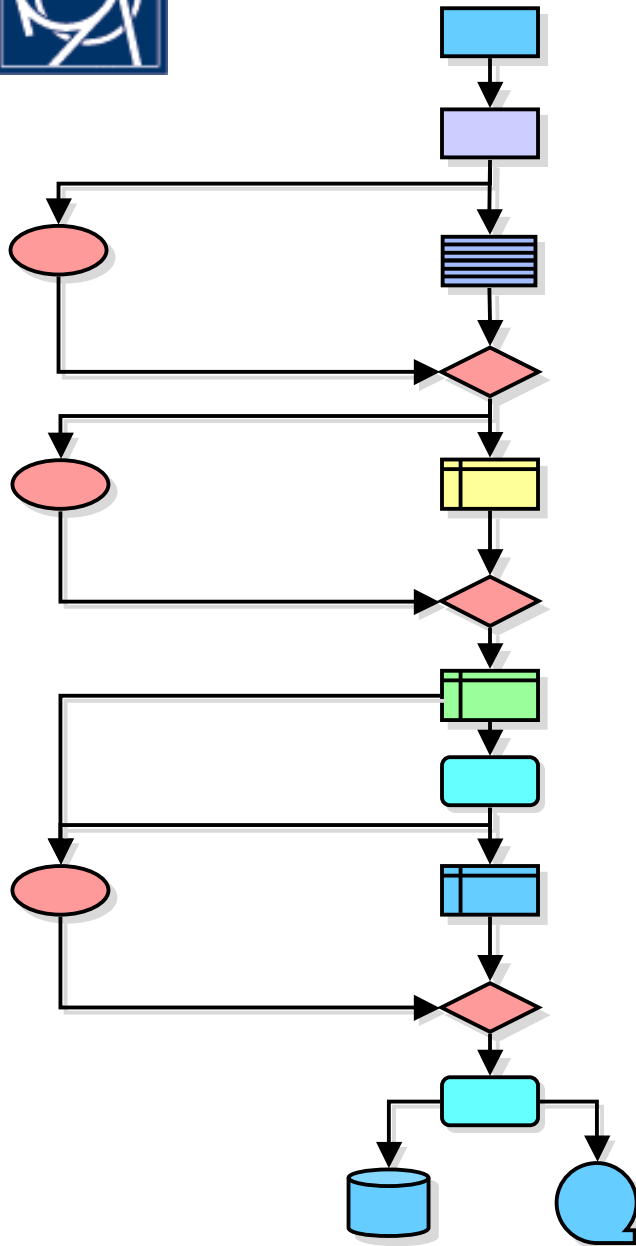
DAQ @ LHC (1)

		Event Size (Byte)	Readout (HLT input) (Events/s.) (GByte/s)
ALICE 	Pb-Pb pp	5×10^7 2×10^6	2×10^3 10^2
ATLAS 		10^6	2×10^3 10
CMS 		10^6	10^5 100
LHCb 		2×10^5	40×10^4 4



DAQ @ LHC (2)

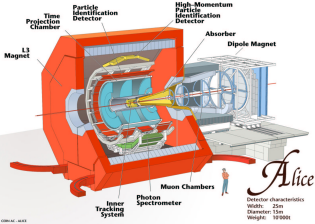
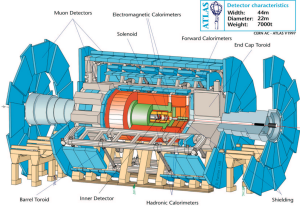
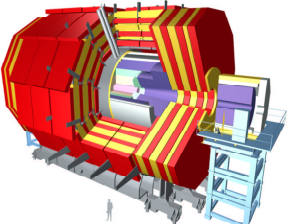
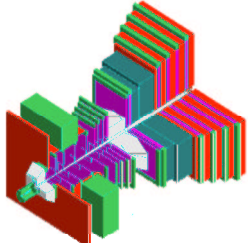
ALICE ATLAS CMS LHCb



	ALICE	ATLAS	CMS	LHCb	
	25	10	100	4	GByte/s
	2.5	6			GByte/s
	200	100	100	40	MByte/s
	1250	300	100		MByte/s

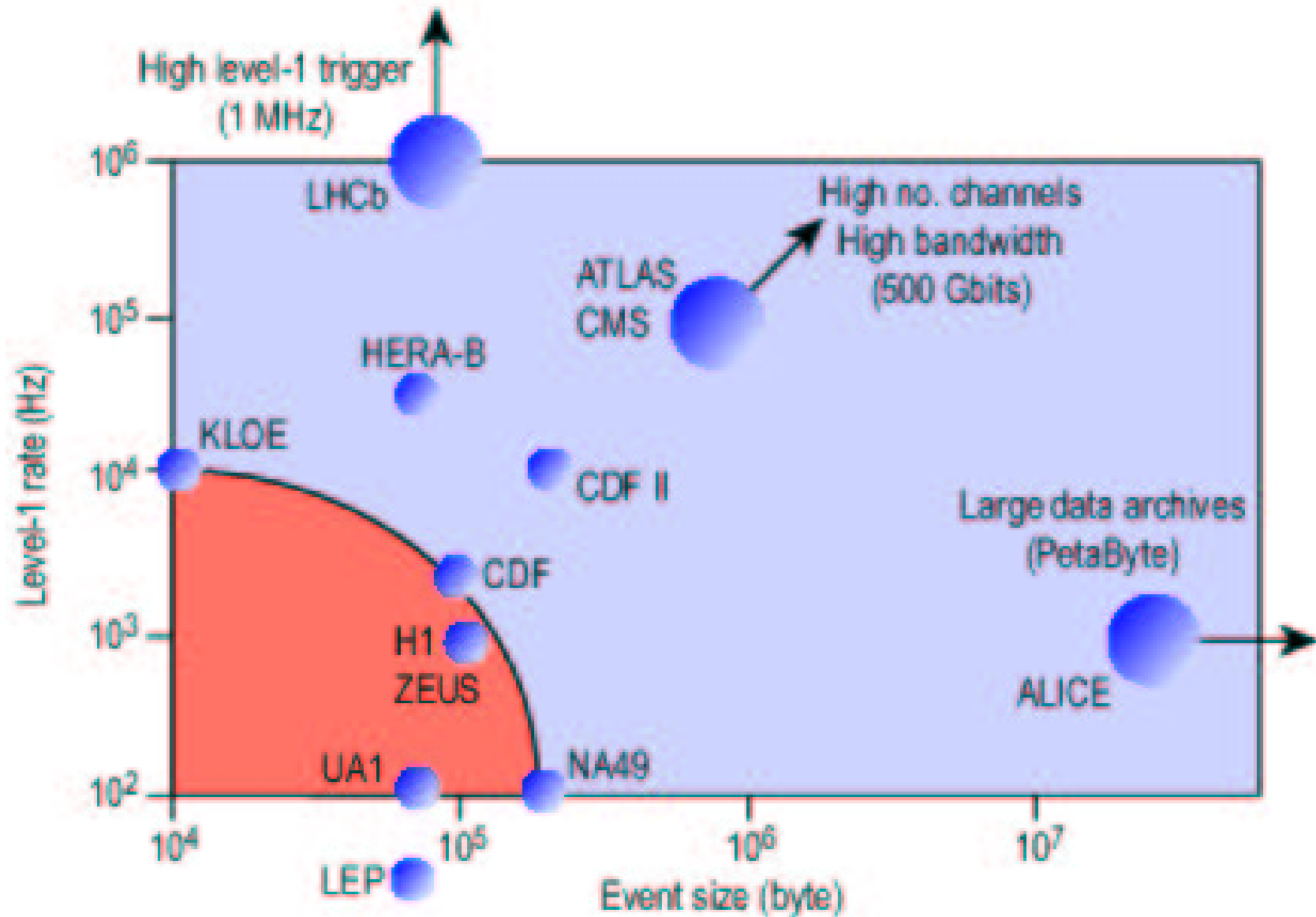


Mass Storage @ LHC

		Readout (HLT output) (Events/s.) (MByte/s)	Data archived Total/year (PBytes)	
ALICE 	Pb-Pb pp	2×10^2 10^2	1250 200	2.3
ATLAS 	Pb-Pb pp	10^2	300 100	6.0
CMS 	Pb-Pb pp	10^2	100 100	3.0
LHCb 		2×10^2	40	1.0



Rates & Bandwidths @ LHC





DAQ for Super Collider Experiments

- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ Technology trends
- ◆ DAQ and HLT for SLHC experiments
- ◆ R&D
- ◆ Conclusions



Super collider reference

References:

- hep-ph/0204087 “Physics potential and experimental challenges of the LHC luminosity upgrade”
- ICFA workshop October 2002 on advanced hadron colliders

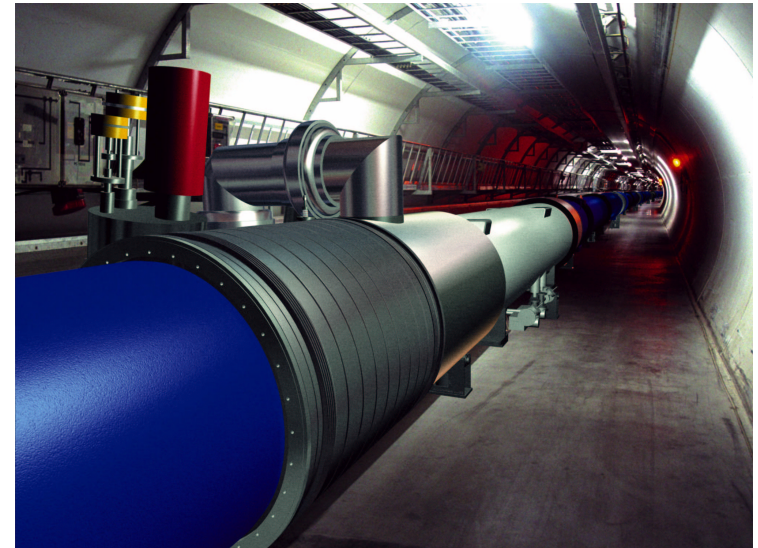
Complement to LHC in TeV region

- e^+e^- colliders
- $\mu^+\mu^-$ colliders

After LHC

- Multi-10-100 TeV
- LHC energy upgrade
 - New magnets, new machine
 - Technical feasibility being studied
- LHC luminosity upgrade $L=10^{35} \text{c m}^{-2}\text{s}^{-1}$, bunch crossing 12.5 ns
 - “Modest” change to machine
 - Major upgrade for experiments
 - Tracker occupancy increased by 10
 - Used here as reference collider

■ VLHC, CLIC





Consequences for DAQ

- ◆ Rate increase
- ◆ Data volume increase
- ◆ Massive need for data transfer, processing and storage
 - 1000's of links to transfer 10's TByte/s off-detector
 - Event building at TByte/s
 - Data storage at GByte/s
- ◆ Impact of duration and complexity
- ◆ DAQ and HLT based on commodity components
- ◆ Need for R&D and prototyping



Trigger, DAQ, HLT

◆ Trigger Level 1

- Custom logic
- Special architectures
- Computing farm

◆ Trigger Level 2

- Special architectures
- Computing farm

◆ DAQ

- Ad-hoc solution (readout)
- Computing farm

◆ High Level Trigger (HLT)

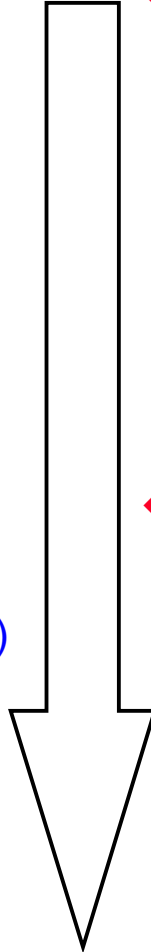
- Computing farm

◆ HEP specific

- Home-made development
- Custom building blocks
- Fast but rigid
- Obsolescence of dev. tools
- Programmable by “a few experts”

◆ General-purpose

- Home-made software
- Commodity building blocks
- Slow but flexible
- Long-term availability tools
- Programmable by “all”



- ◆ For DAQ and HLT: custom if no alternative
- ◆ Evolution of industry will be the driving force

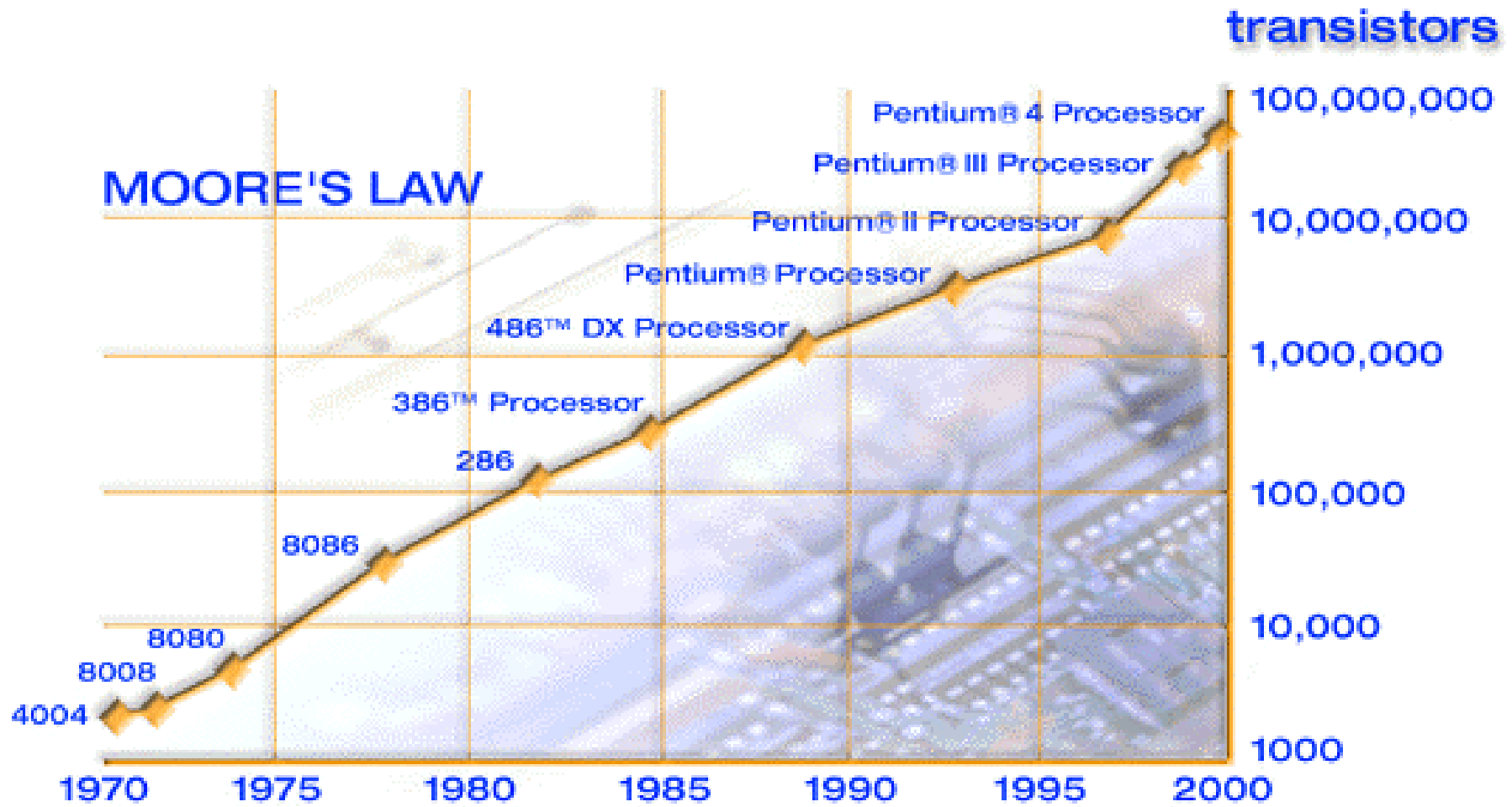


DAQ for Super Collider Experiments

- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ **Technology trends**
- ◆ DAQ and HLT for SLHC experiments
- ◆ R&D
- ◆ Conclusions



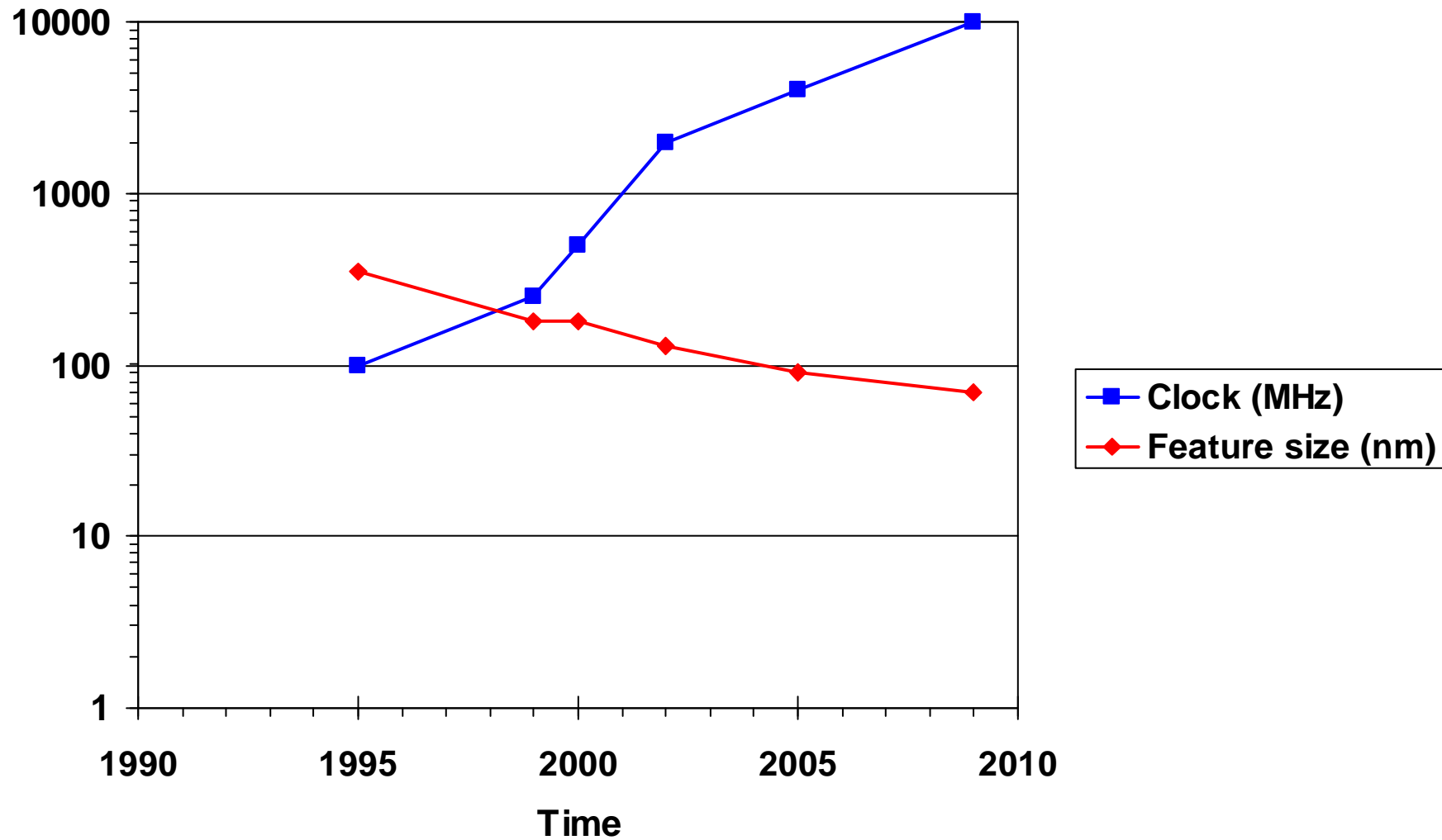
Moore's Law



© Intel corp.

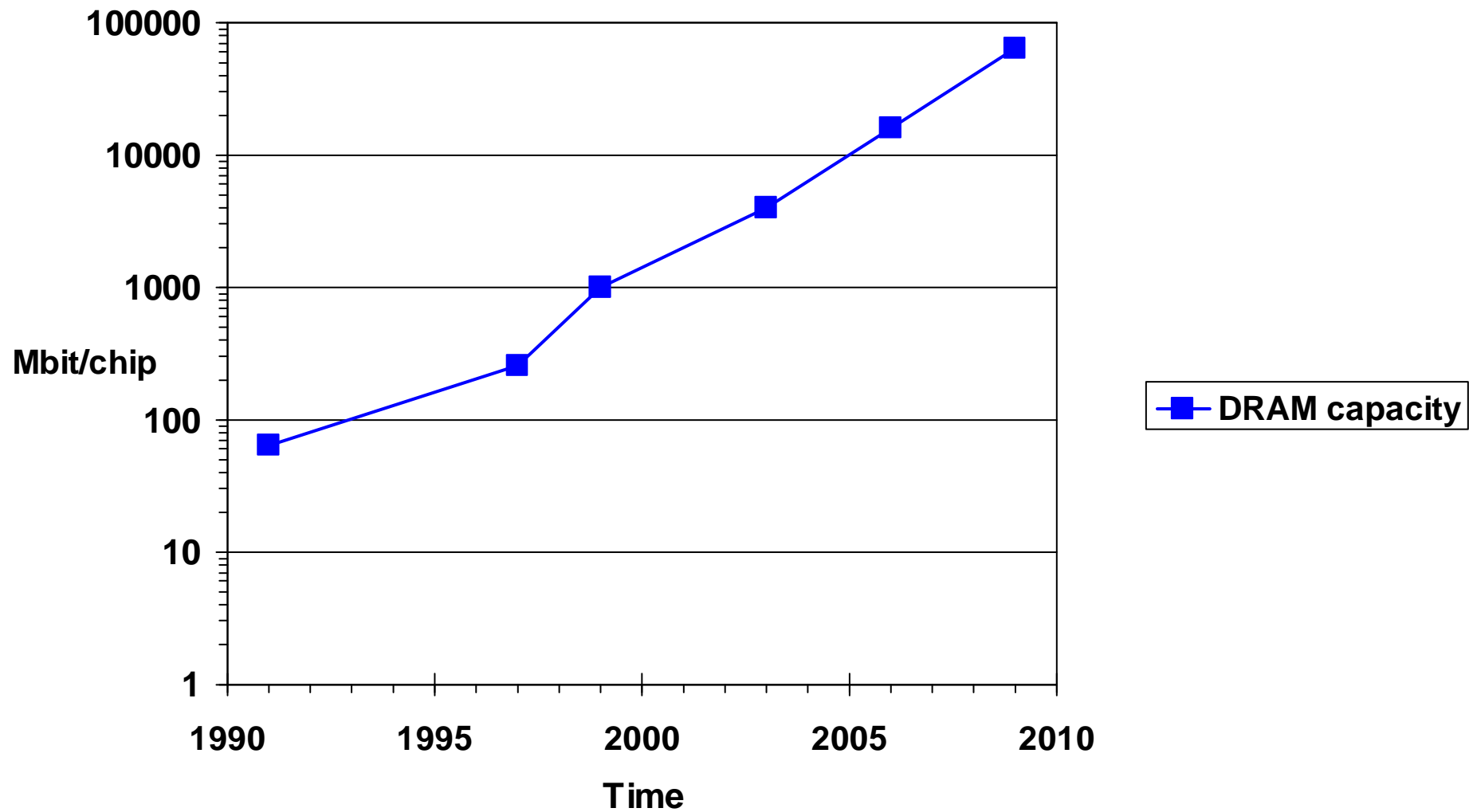


Chip key parameters



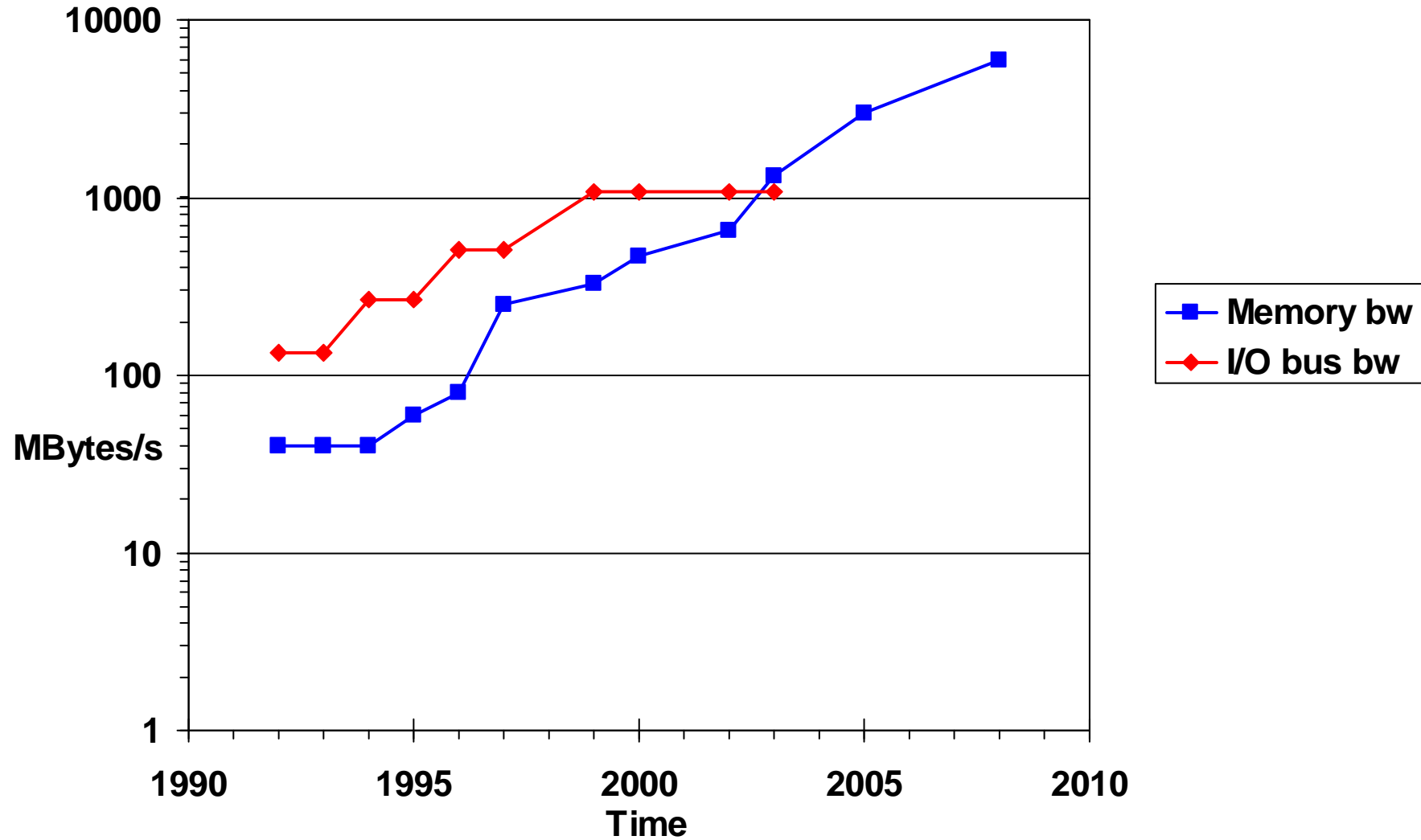


Memory capacity





Memory and I/O bus Bandwidth





On and off board data communication

- ◆ Standardize in the box (Rapid I/O, Hyper-Transport, etc)
- ◆ The RapidIO Interconnect Architecture:
 - Chip-to-chip and board-to-board communications
 - Gbit/s and beyond.
 - High-performance, packet-switched, interconnect technology
 - Switches on the board
- ◆ The RapidIO Trade Association:
 - Non-profit corporation controlled by its members
 - Direct the future development
 - For networking products: increased bandwidth, lower costs, and a faster time-to-market than other more computer-centric bus standards.
 - Steering Committee: Alcatel, Cisco Systems, EMC Corporation, Ericsson, Lucent Technologies, Mercury Computer Systems, Motorola, and Nortel Networks



I/O bus evolution

- ◆ PCI is today's de-facto standard
- ◆ Initiative of Intel
- ◆ Public from the start, "imposed" to industry
 - Exceptional period of stability and compatibility
- ◆ Industry de-facto standard for local I/O: PCI (PCI SIG)
 - 1992: origin 32 bits 33 MHz 133 MBytes/s
 - 1993: V2.0 32 bits
 - 1994: V2.1
 - 1996: V2.2 64 bits 66 MHz 512 MBytes/s
 - ■ 1999: PCI-X 1.0 64 bits 133 MHz 1 GBytes/s
 - 2002: PCI-X 2.0 64 bits 512 MHz 4 Gbytes/s
- ◆ Future: PCI-X 2.0, 3GIO, PCI-Express



I/O and system busses

	Bus	Industrial Support	Bus width (bits)	Bus clock (MHz)	Max. bw on single channel	Type
I/O	PCI 32 bits/33 MHz	1990, Intel	32	33	132	Bus
	PCI 64 bits/33 MHz		64	66	264	Bus
	PCI 64 bits/66 MHz	1995, PCI SIG	64	66	533	Bus
	PCI-X	2000, IBM, Compaq, HP	64	133	1056	Bus
System	Future I/O	IBM, Compaq, HP Adaptec, 3COM				Channel
	NGIO 2.5 Gb	Intel, Sun, Dell, Hitachi, NEC, Siemens	serial	2500	500	Channel
	Infiniband	Intel, Sun, Dell, IBM Compaq, HP, Microsoft	serial	2500		

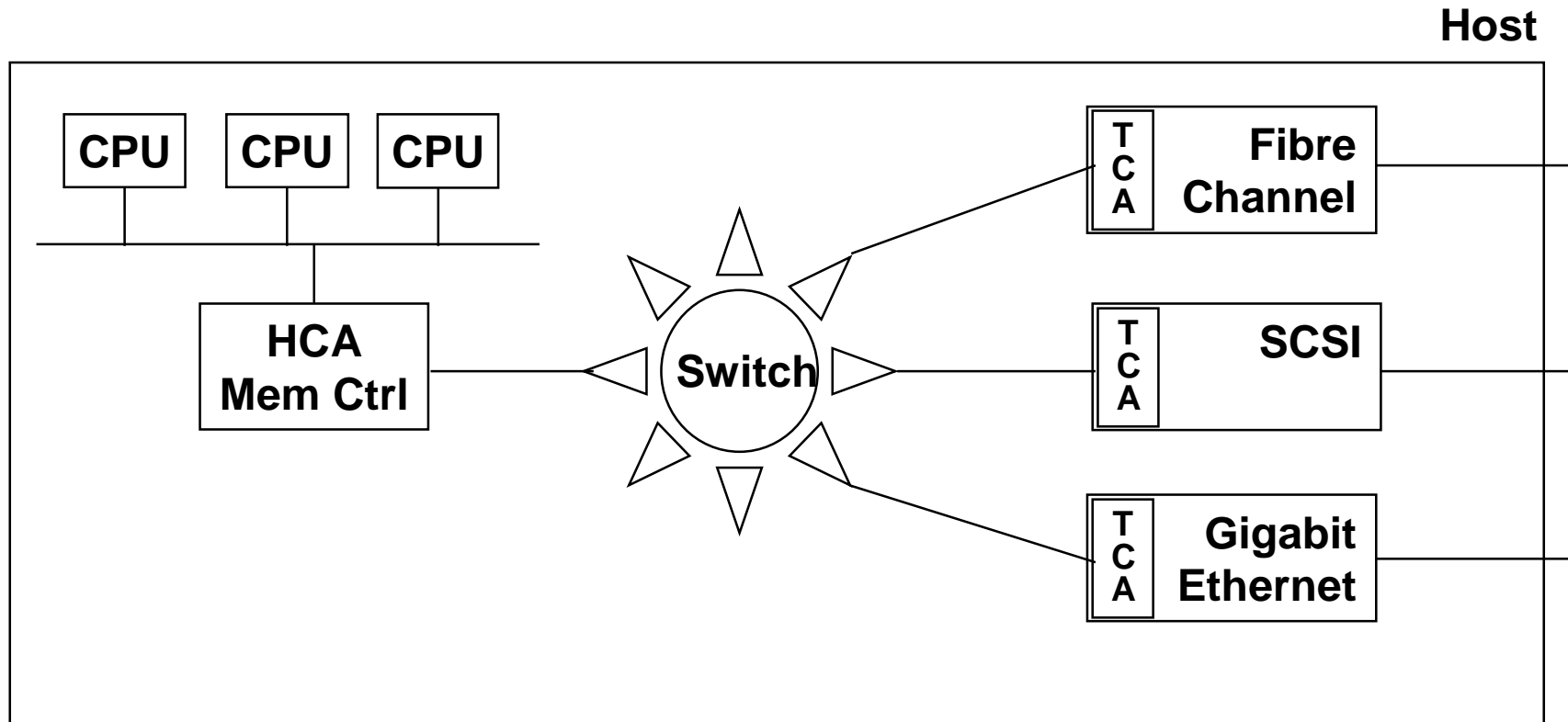


Infiniband

- ◆ Techno
 - 2.5 Gbit/s line rate
 - 1, 4 or 12 lines giving 0.5, 2, 6 GB/S
 - Switch-based system
 - Transport: reliable connection and datagram, unreliable connection and datagram, IPV6, ethertype
- ◆ Common link architecture and components with Fibre Channel and Ethernet
- ◆ Chips: Cypress, IBM, Intel, LSI logic, Lucent, Mellanox, Redswitch
- ◆ Products: Adaptec, Agilent



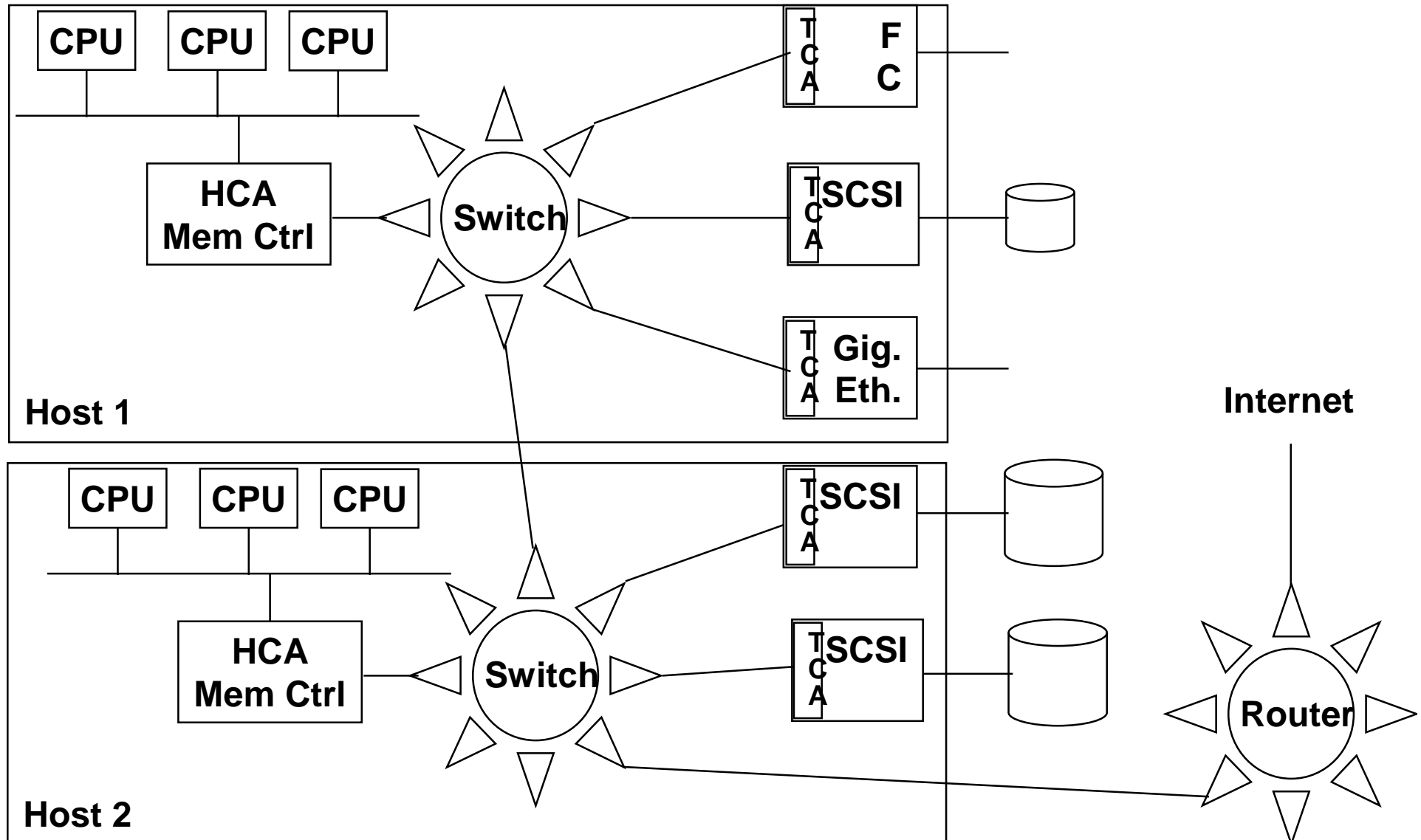
Infiniband



**Host Channel Adapter (HCA)
Target Channel Adapter (TCA)**

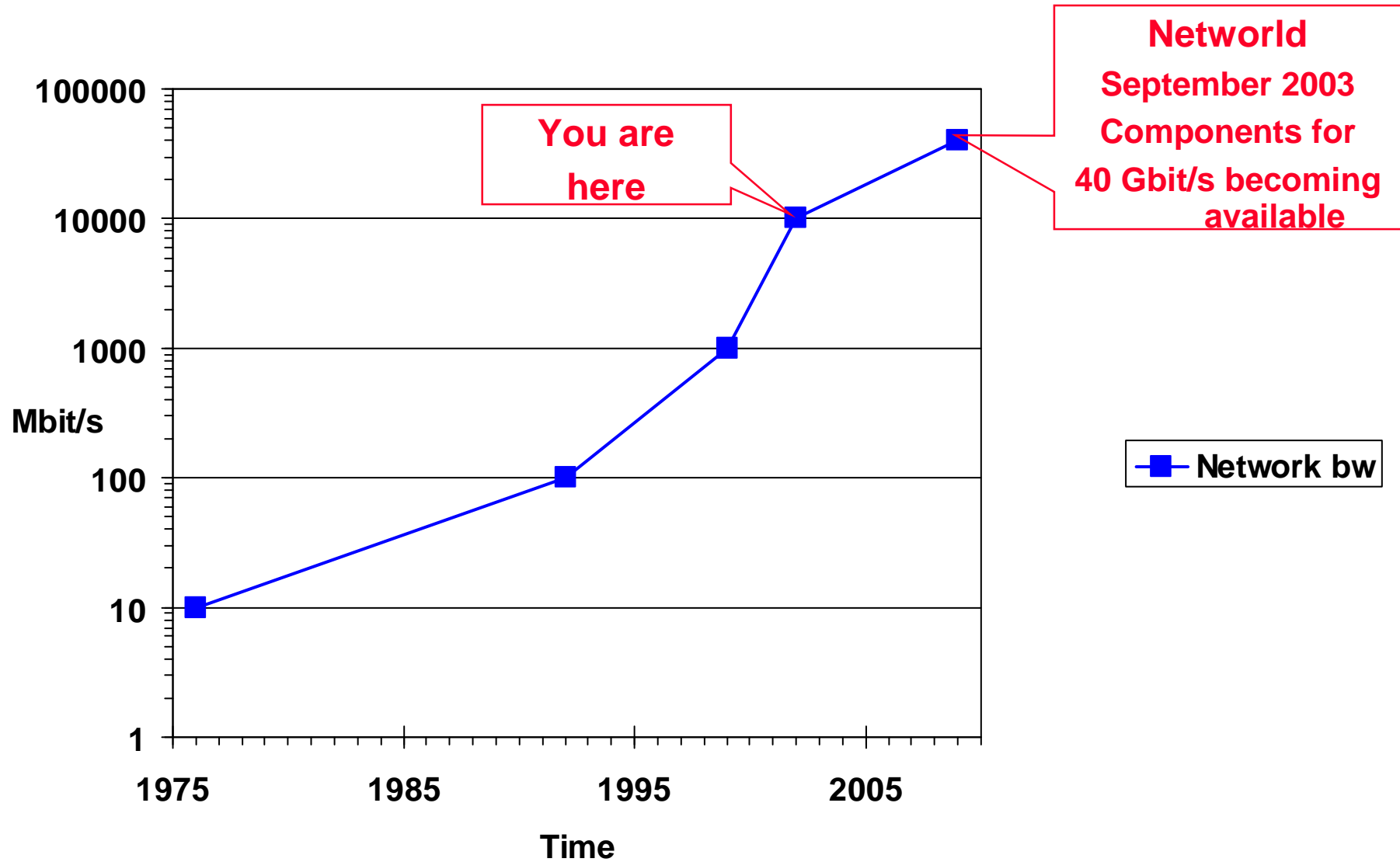


Infiniband: multiple hosts





Networking technology





DAQ for Super Collider Experiments

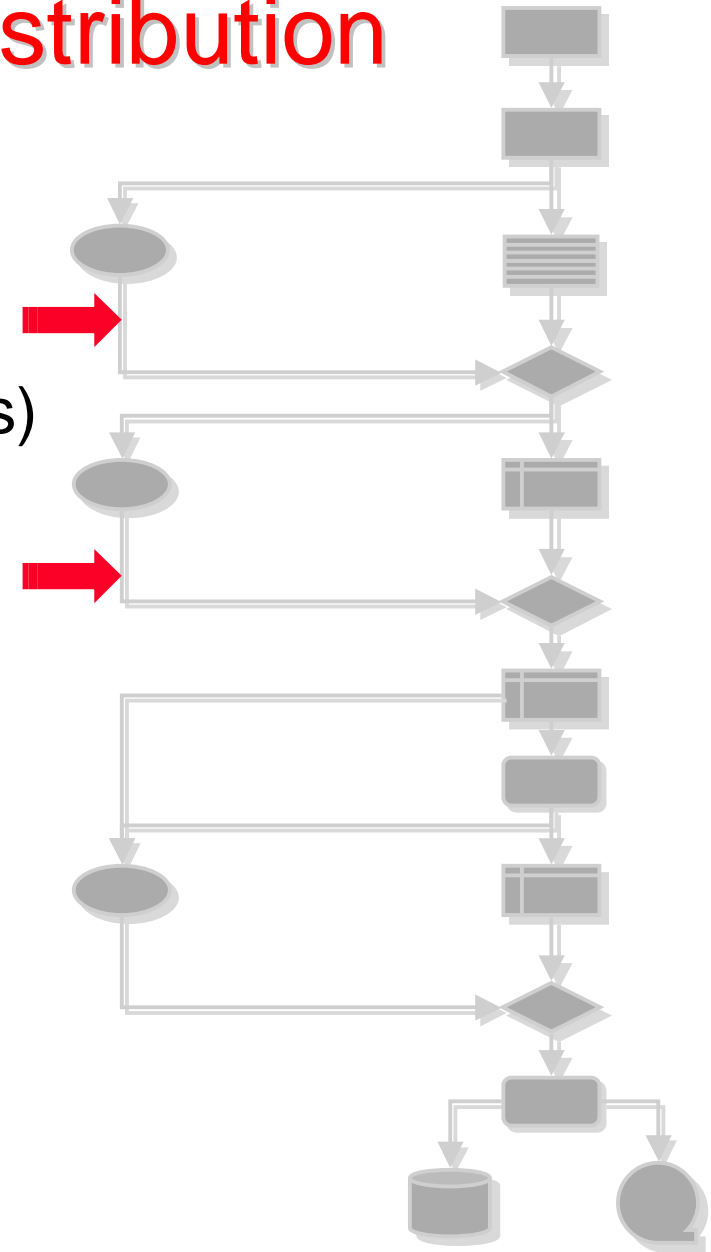
- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ Technology trends
- ◆ **DAQ and HLT for SLHC experiments**
- ◆ R&D
- ◆ Conclusions



Trigger & Timing distribution

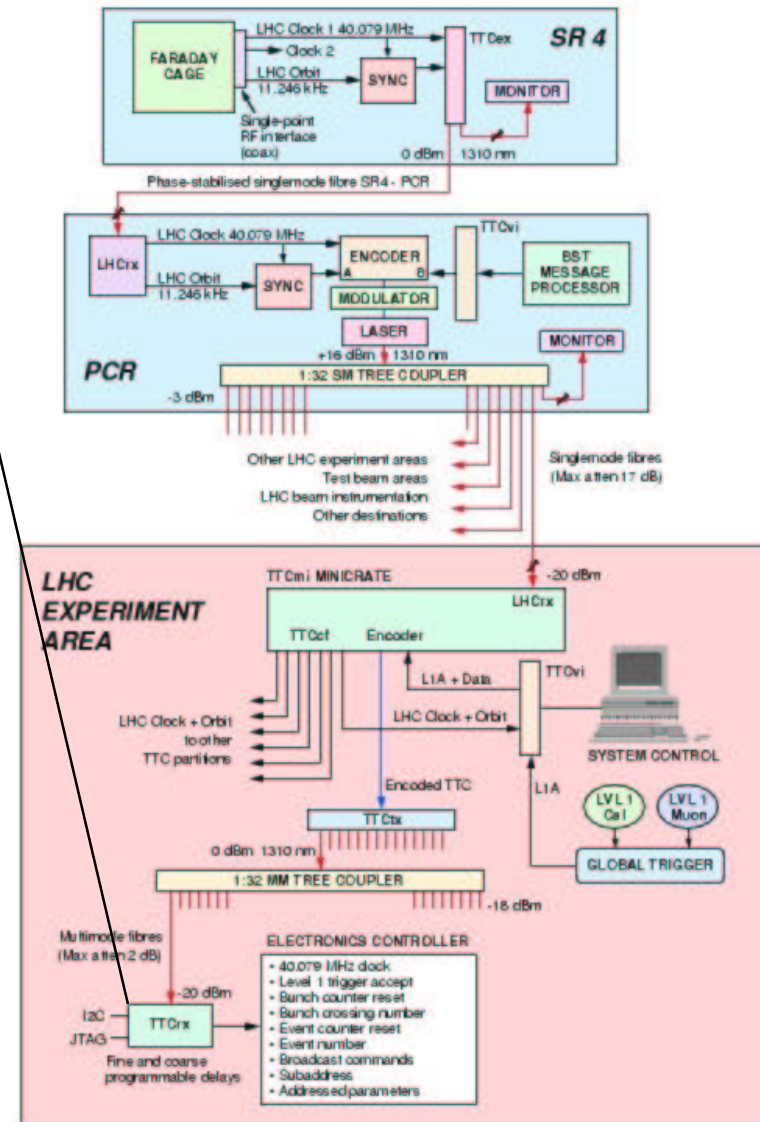
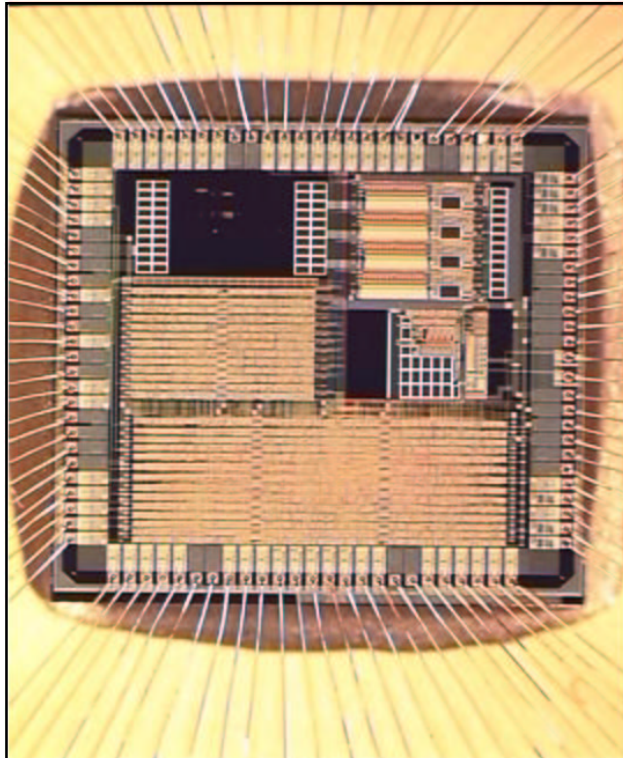
- ◆ Transfer from TRG to electronics
- ◆ One to many
- ◆ Massive broadcast (100's to 1000's)

- ◆ Optical, Digital
 - HEP-specific components
 - HEP developments





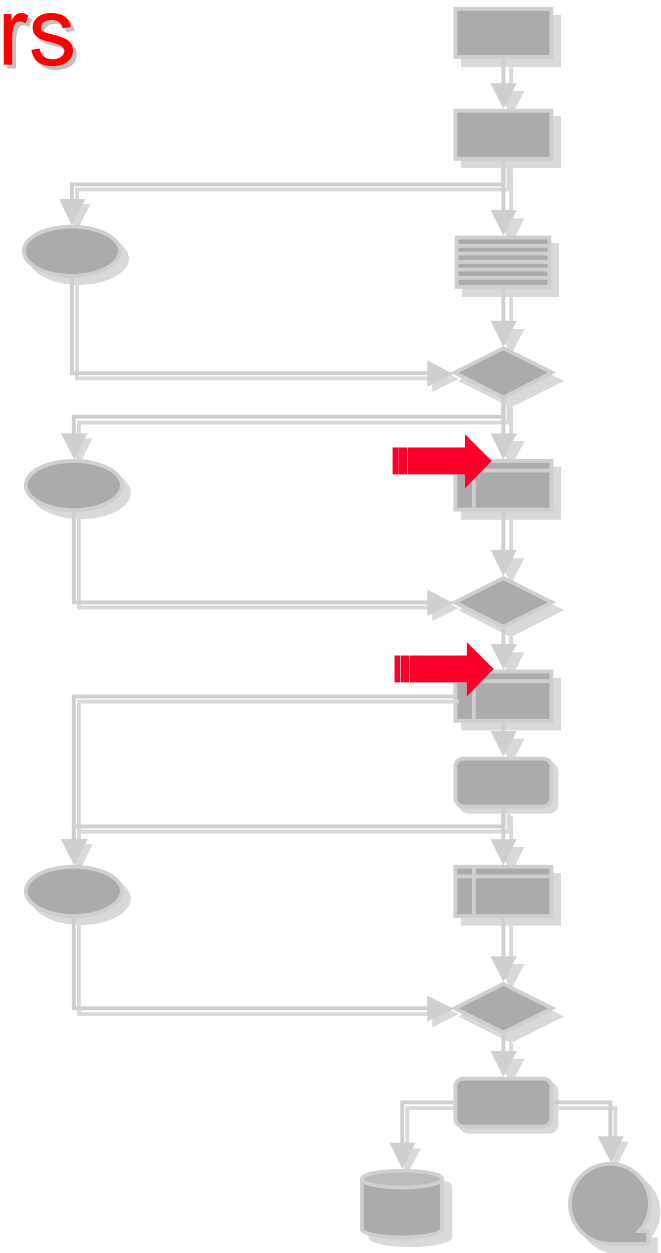
LHC Trigger & Timing distribution





Links Adapters

- ◆ Adapter for 1 or a few links to I/O bus of the memory or the computer
- ◆ Many-to-one
- ◆ Massive parallelism (100's to 1000's)
- ◆ Physical interface realized by
 - Custom chip
 - IP core (VHDL code synthesized in FPGA)
- ◆ Implementation depend upon I/O bus evolution





Link and adapter performance

- PCI 32 bits 66 MHz with commercial IP core
- No large local memory. Fast transfer to PC memory

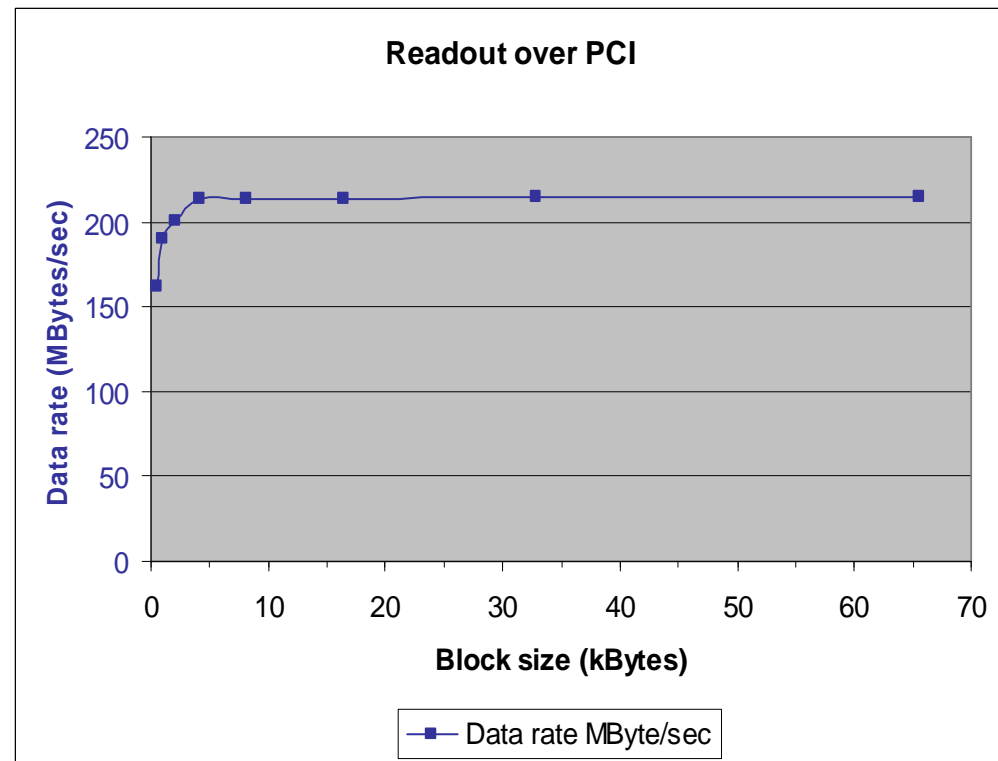
Reach 200 MB/s for block size above 2 kBytes.

Total PCI load: 92 %

Data transfer PCI load: 83 %

Lots of bw available.

Major fraction available to end application.





Subevent & event buffer

◆ Baseline:

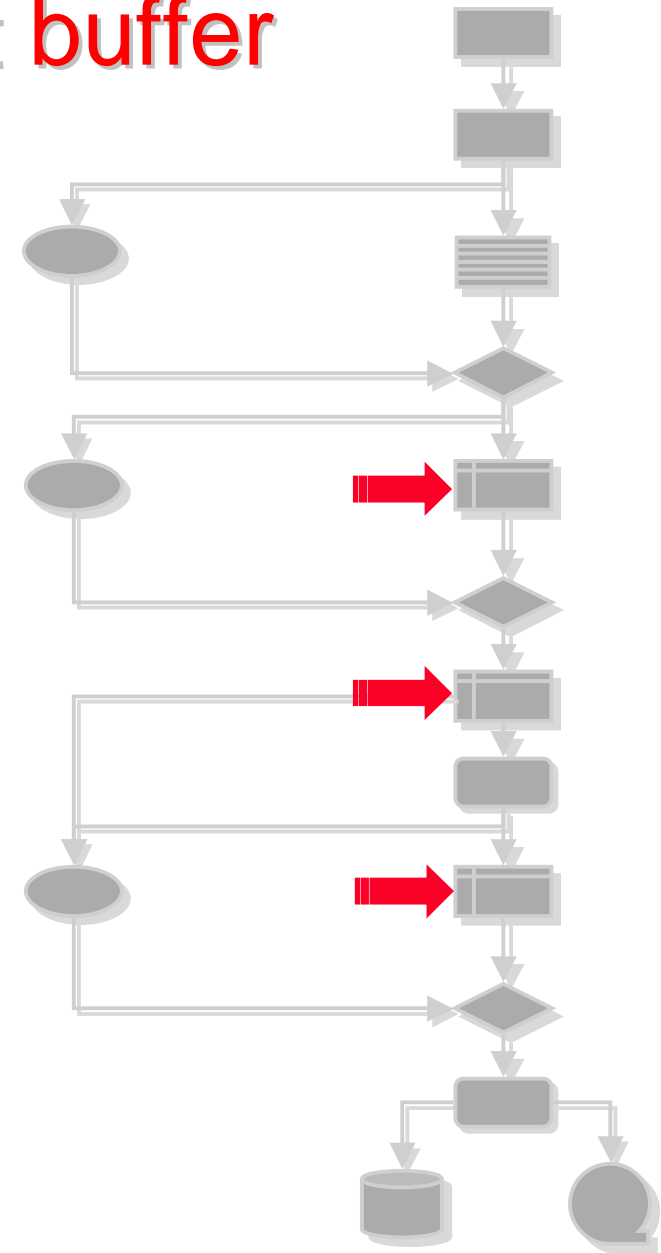
- Function: fast dual-port memories
- Adopt commodity component (PC)

◆ Key parameters:

- Cost/performance
- Performance: memory bandwidth

◆ Future

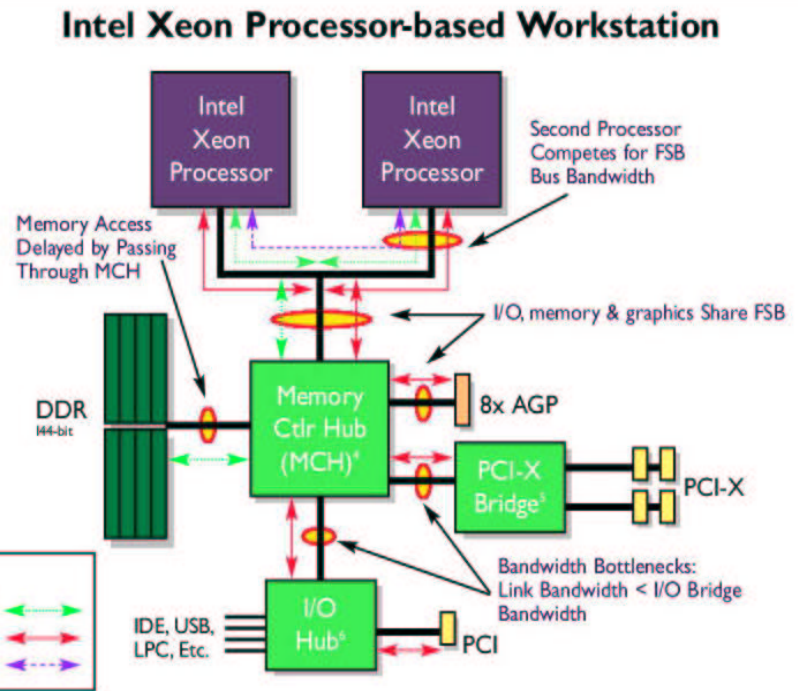
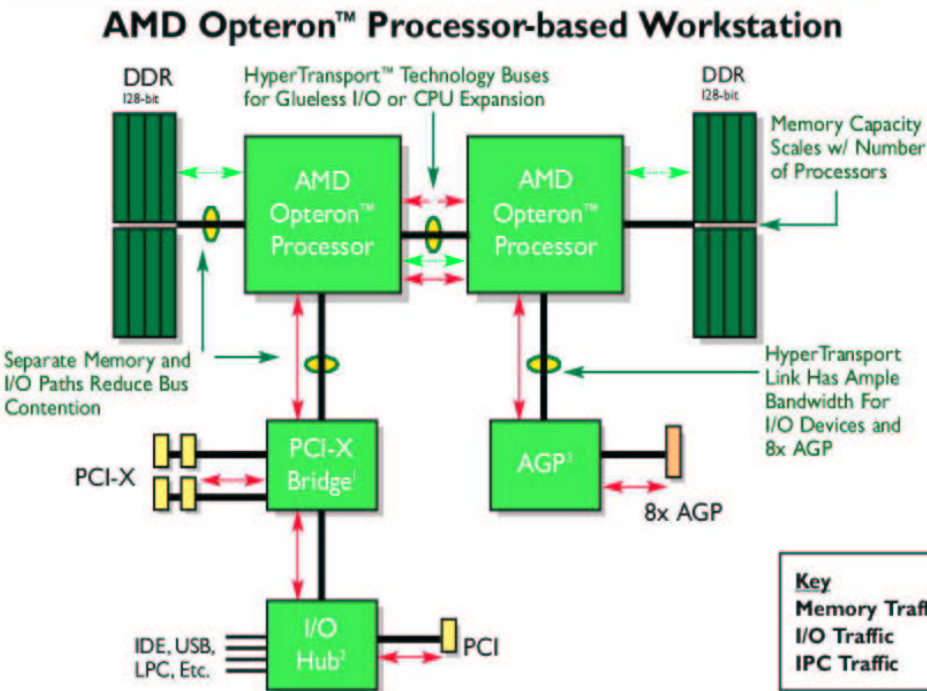
- Faster memory clock
- Wider data bus





Dual CPU Architectures

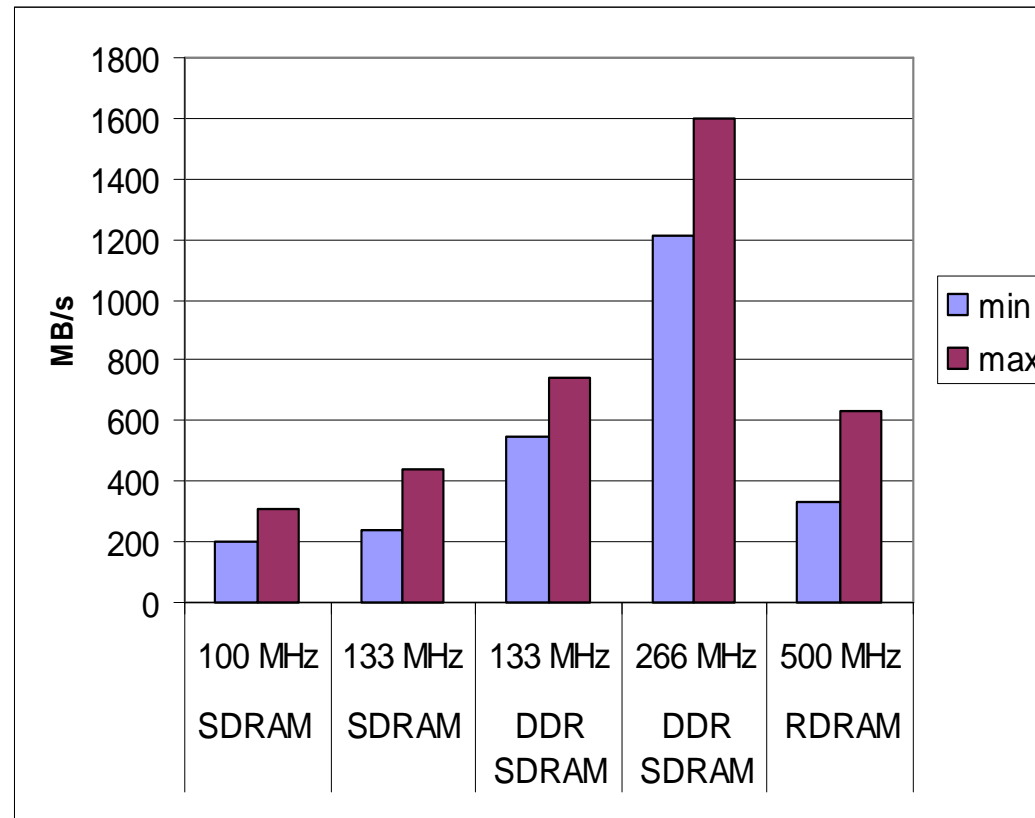
2 players in commodity market: AMD, Intel





Memory Benchmarks

1x Stream:

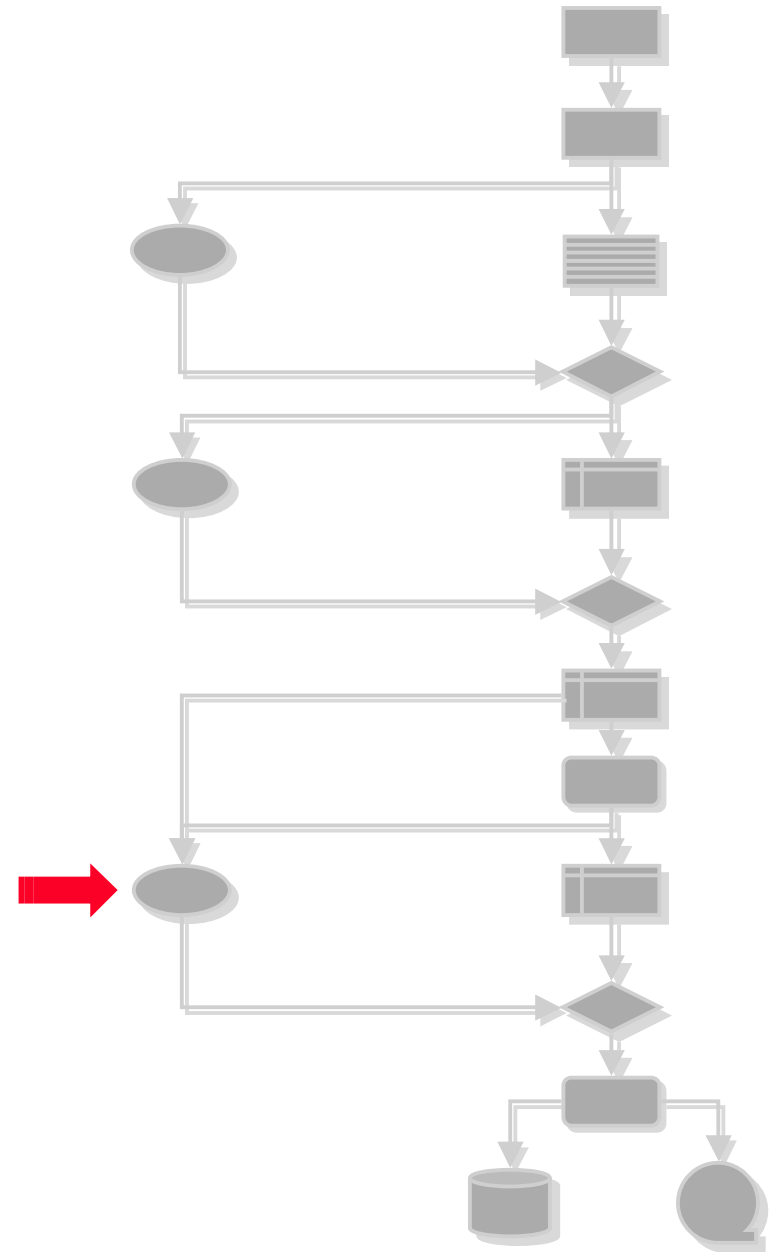


DDR266	1x Stream:	2x Stream:	4x Stream:
2x Opteron, 1.8 GHz, HyperTransport:	1006 – 1671 MB/s	975 – 1178 MB/s	924 – 1133 MB/s
2x Xeon, 2.4 GHz, 400 MHz FSB:	1202 – 1404 MB/s	561 – 785 MB/s	365 – 753 MB/s



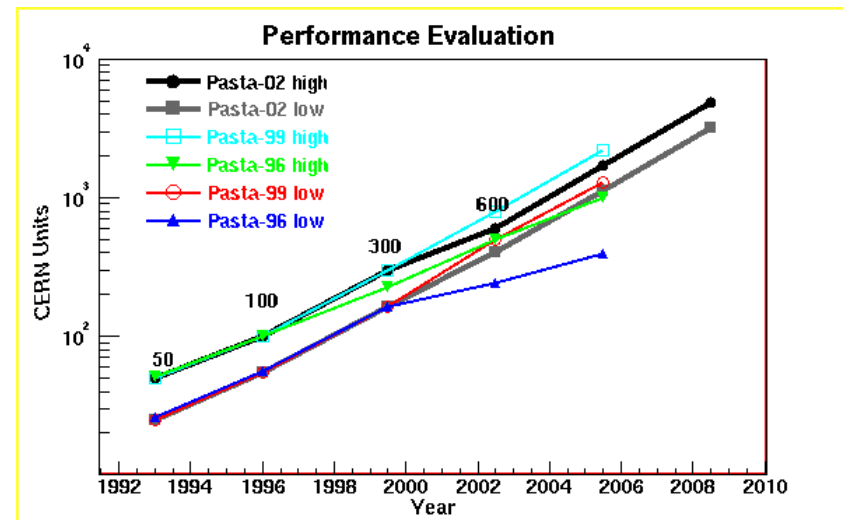
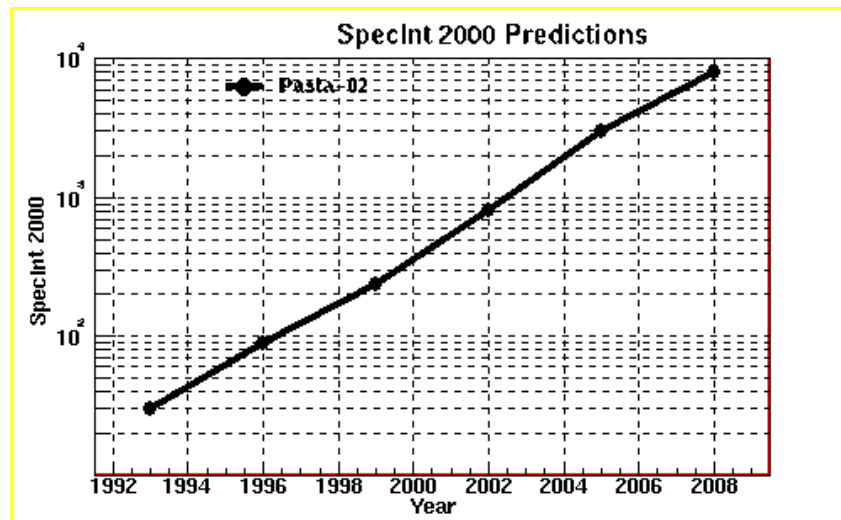
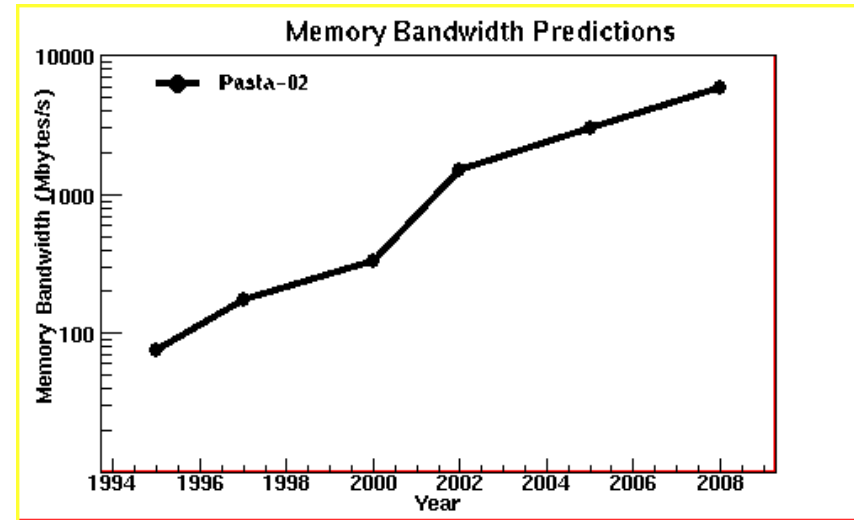
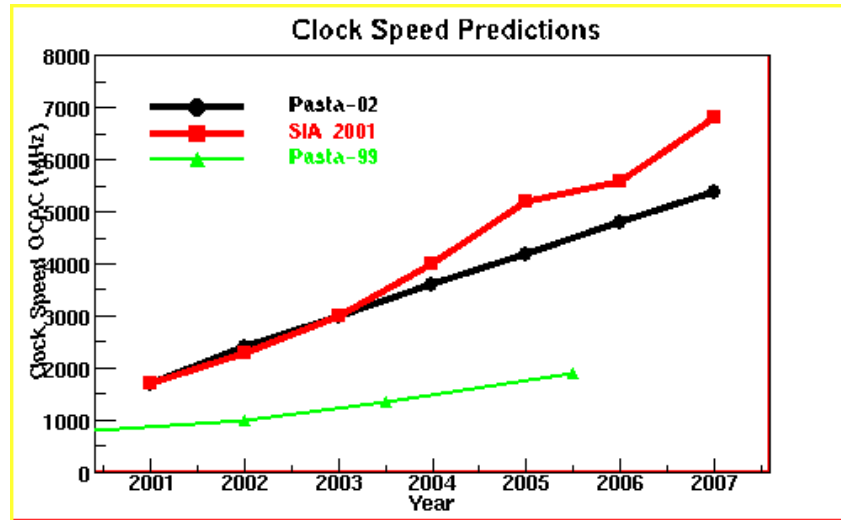
HLT

- ◆ **Baseline:**
 - Function: fast dual-port memories and data processing
 - Adopt commodity component (PC)
- ◆ **Key parameters:**
 - Cost/performance
 - Performance: memory bandwidth & CPU performance
- ◆ **Future**
 - Faster CPU clock
 - Multi CPUs chips (3G, human I/O)
 - Wider data bus





Performance predictions



Raw performance usable by HEP !

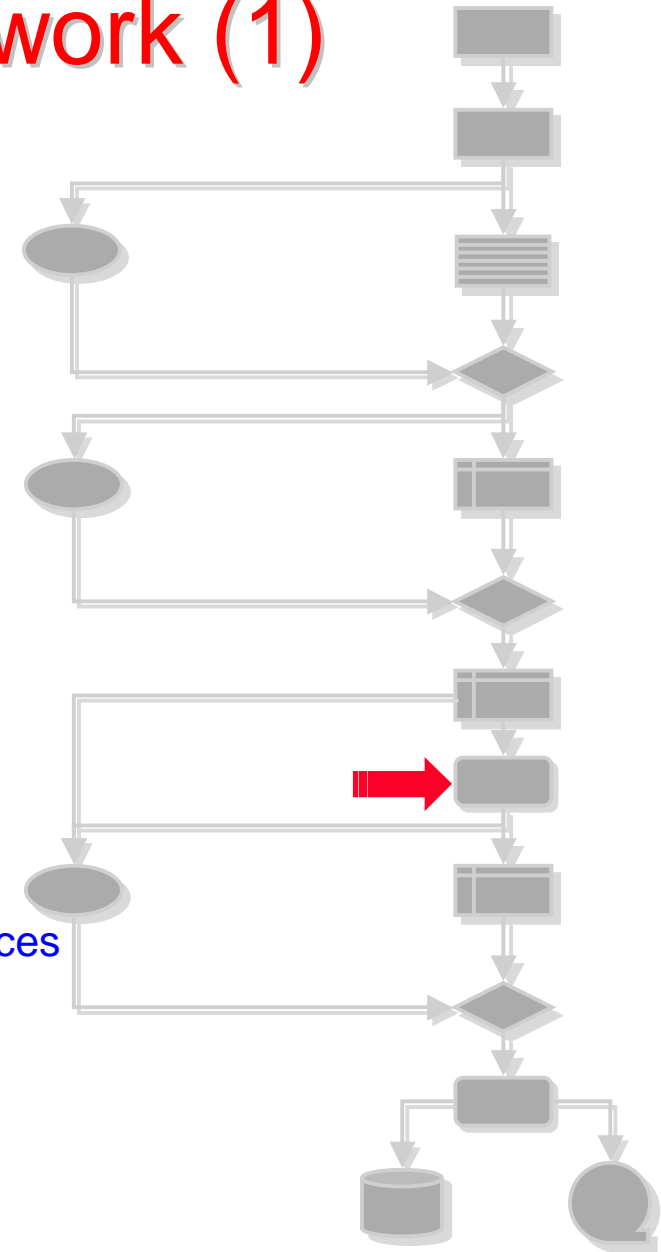


Event Building Network (1)

- ◆ **Baseline:**
 - Adopt broadly exploited standards
Switched Ethernet (ALICE, ATLAS, LHCb)
 - Adopt a performing commercial product
CMS: Myrinet baseline, Gbit Eth. as backup

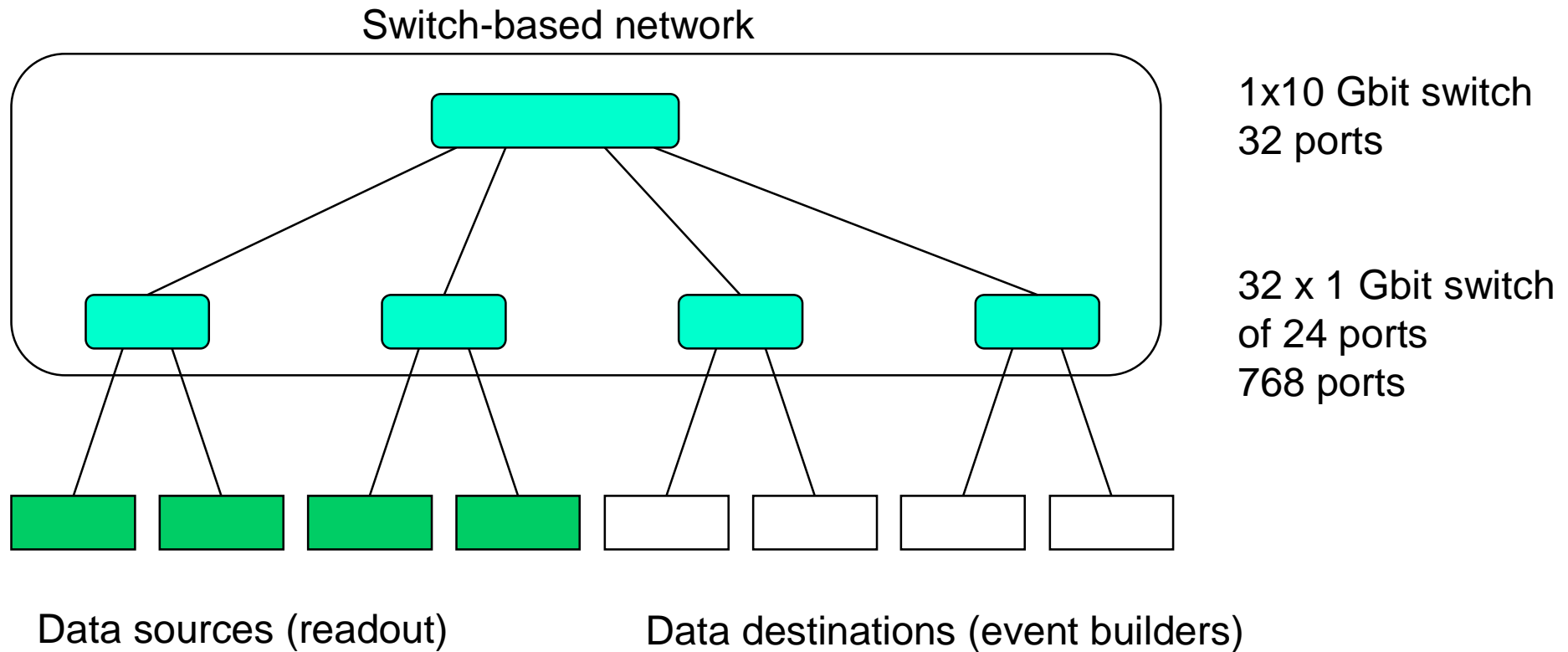
- ◆ **Motivations for switched Ethernet:**
 - Performance of Gigabit Ethernet switches already adequate for DAQ @ LHC
256 Gbit/s of aggregate bandwidth
 - Use of commodity items: network switches and interfaces
 - Easy (re)configuration and reallocation of resources

- ◆ **Future: 40 or 100 Gbit/s Eth.**





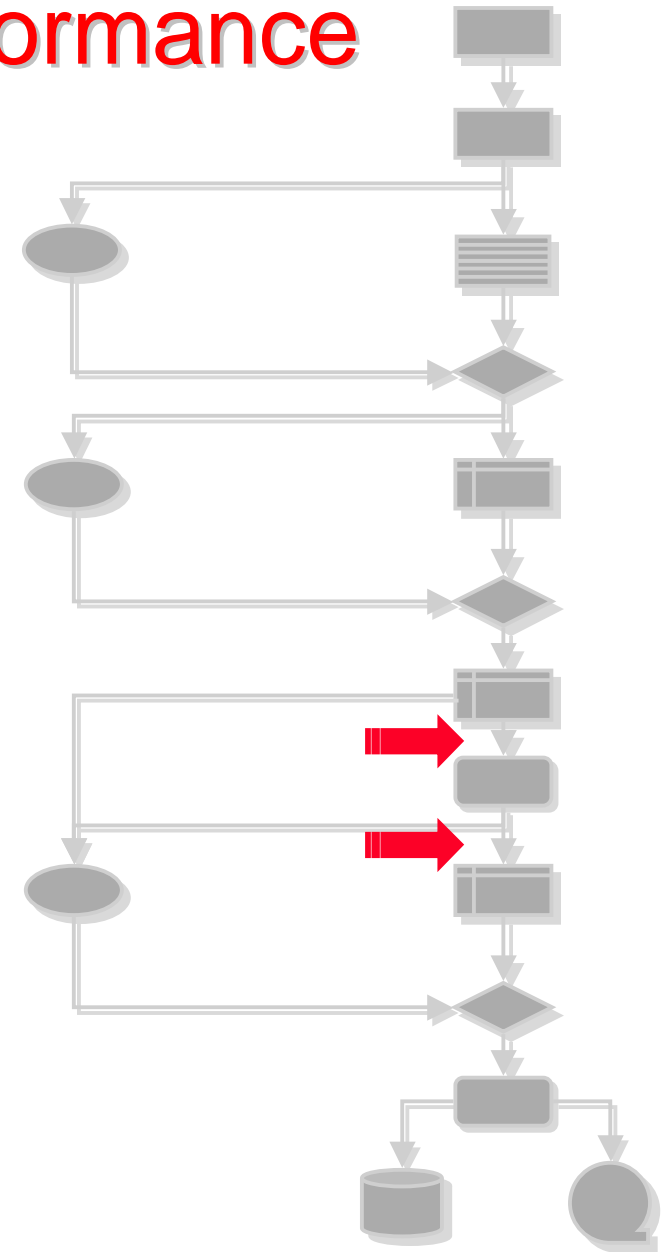
Event Building Network (2)





Ethernet NIC's Performance

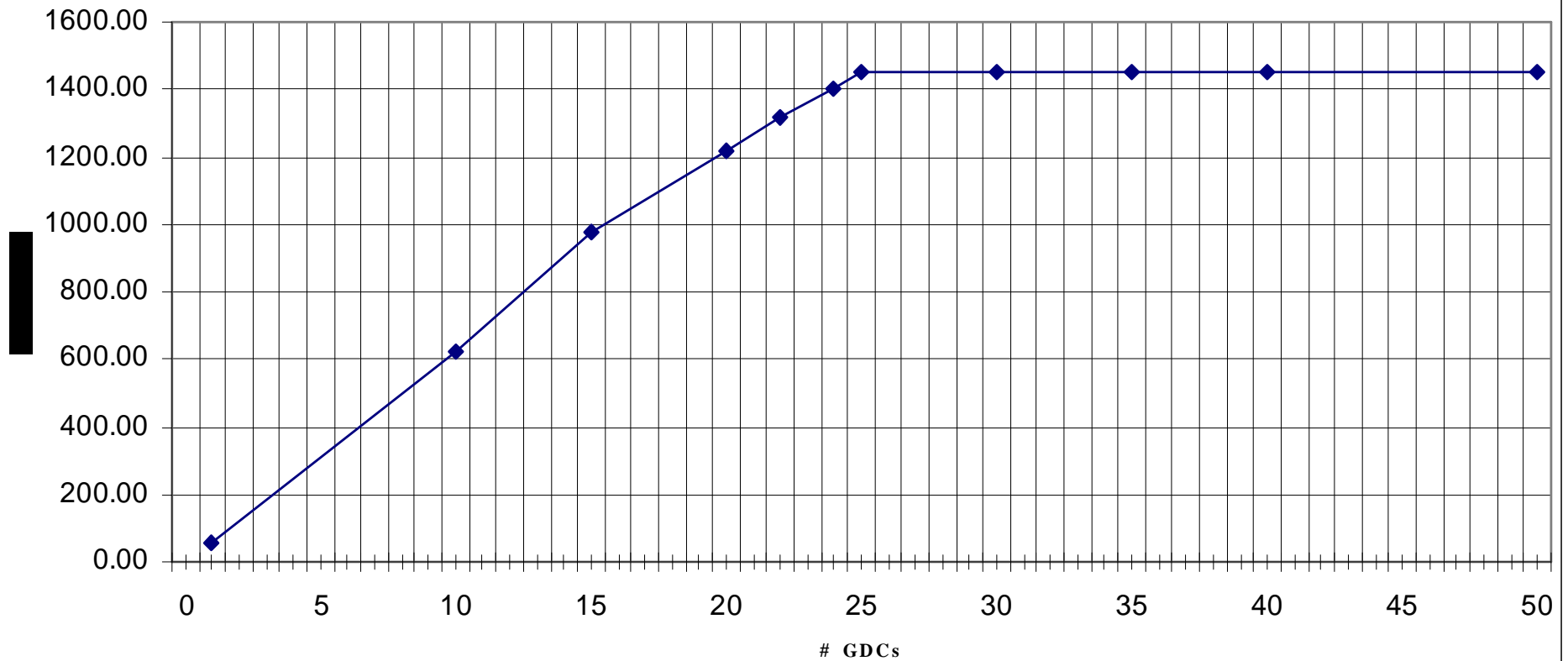
- ◆ Gigabit Ethernet
 - New generation of PC motherboard includes 2 Gbit Eth ports
 - ◆ Active market with several players
 - ◆ 3Com, Broadcom, Intel, NetGear
 - ◆ Fast evolution since 3 years
 - ◆ BW: from 50 to 110 MB/s
 - ◆ CPU usage: 150 to 60 %
- ◆ TCP/IP Offload Engine (TOE)
 - ◆ Dedicated processor to execute IP stack
- ◆ 10 Gigabit Ethernet
 - ◆ Up to 700 MB/s





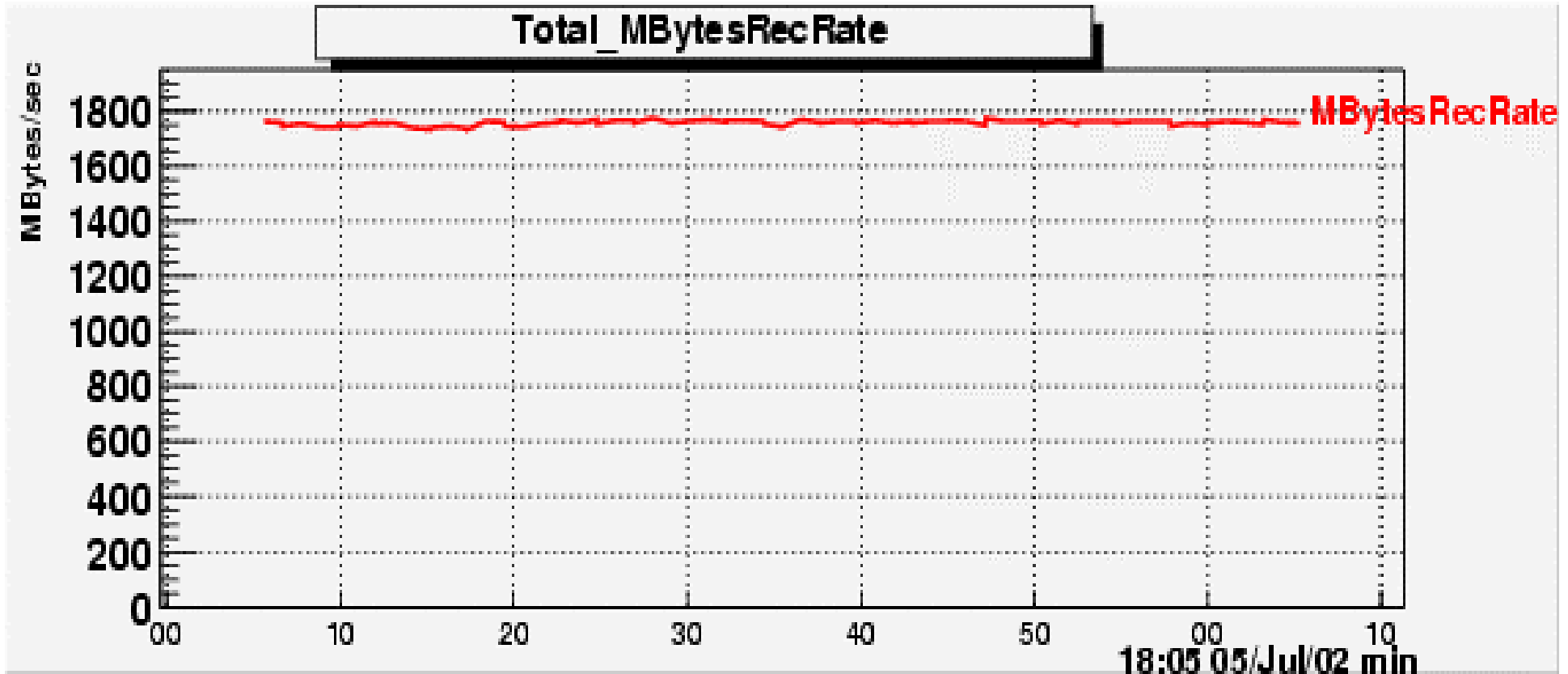
Scalability of network-based event building

DATE COLE EQUIPMENT FLAT, 1 MB events, 21 LDCs





Performance of network-based event building

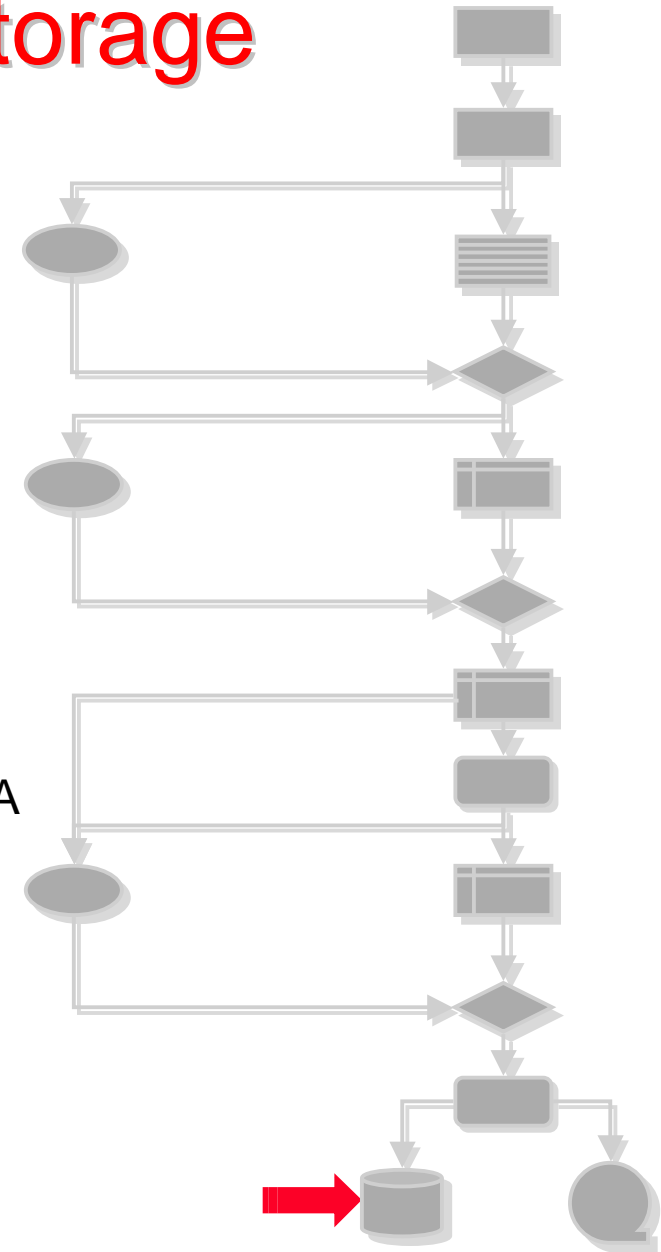


- Event building
- No recording
- 5 days non-stop
- 1750 MBytes/s sustained (goal was 1000)



Transient Data Storage

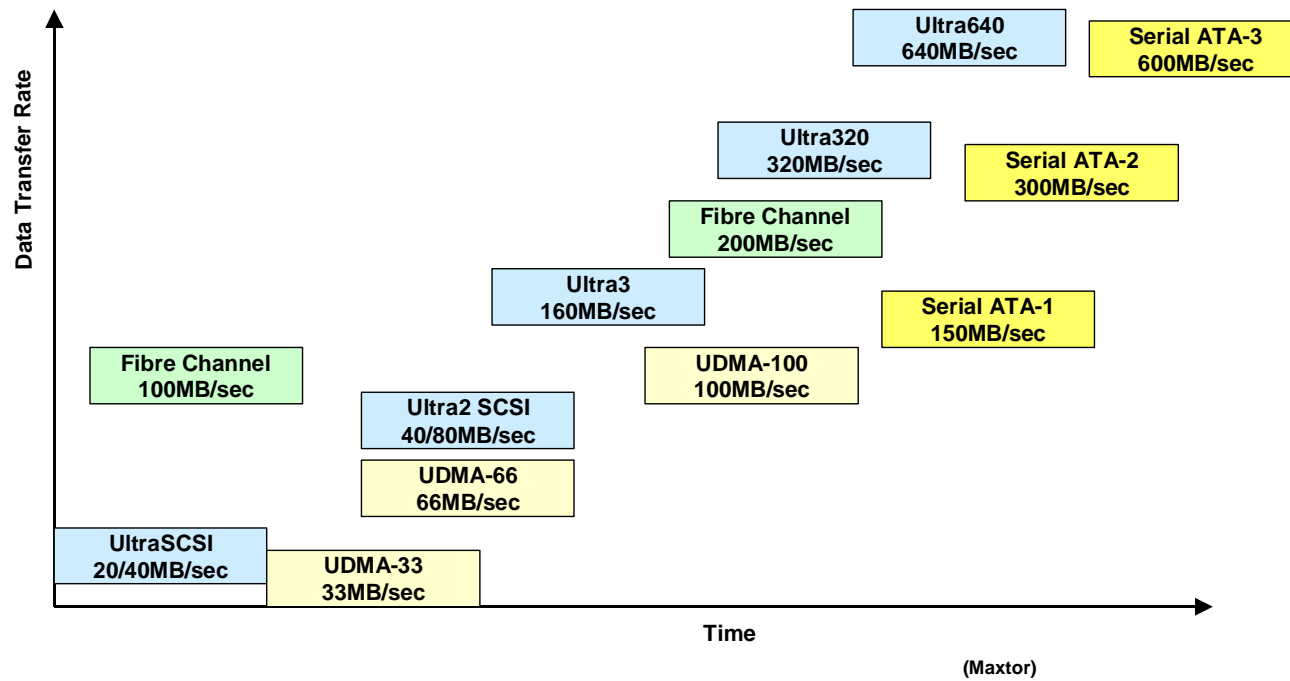
- ◆ Transient Data Storage
- ◆ Before archiving to tape, if any
- ◆ Several options
 - **Disk Technology**
 - IDE: 2 SFr/GB naked, 8 SFr/GB with infra.
 - Density: 2 Gbit/in²
 - **Disk attachment:**
 - DAS: IDE, SCSI, Fiber Channel, serial-ATA
 - NAS: disk server
 - SAN: Fiber Channel
 - **RAID-level**
- ◆ Key selection criteria:
cost/performance & bandwidth/box





Disk attachment

Disk Connection Technology Evolution

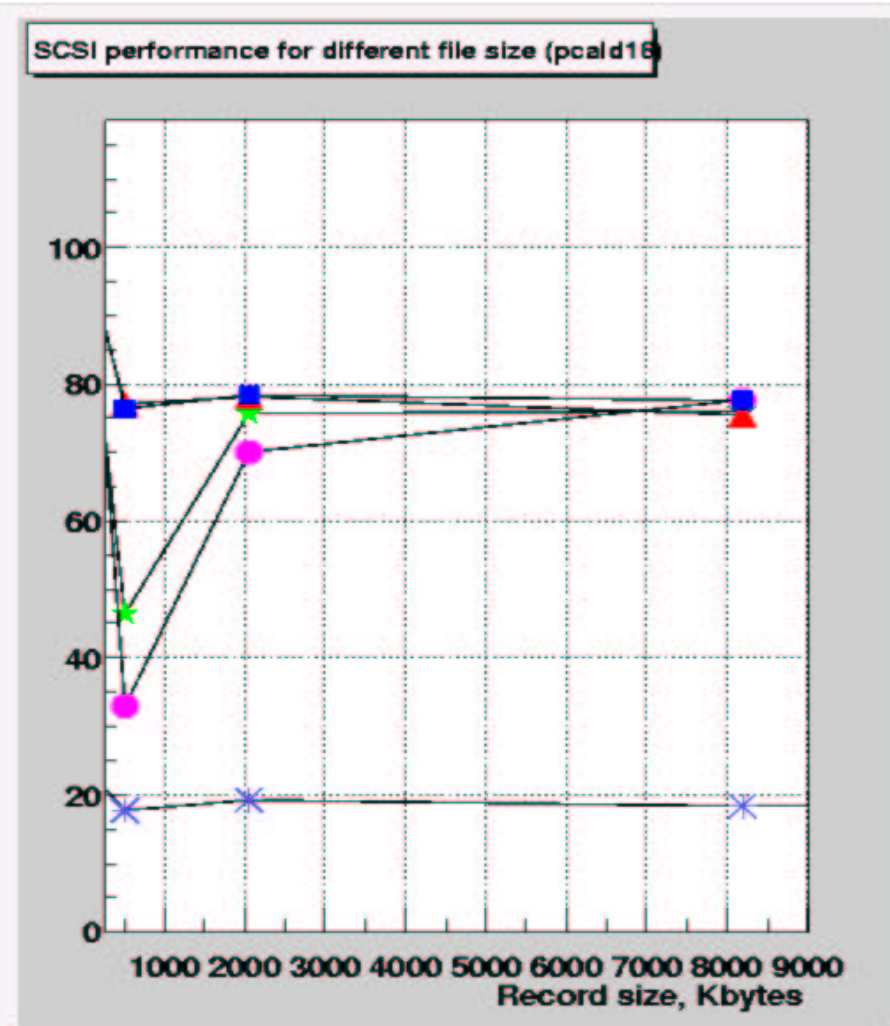
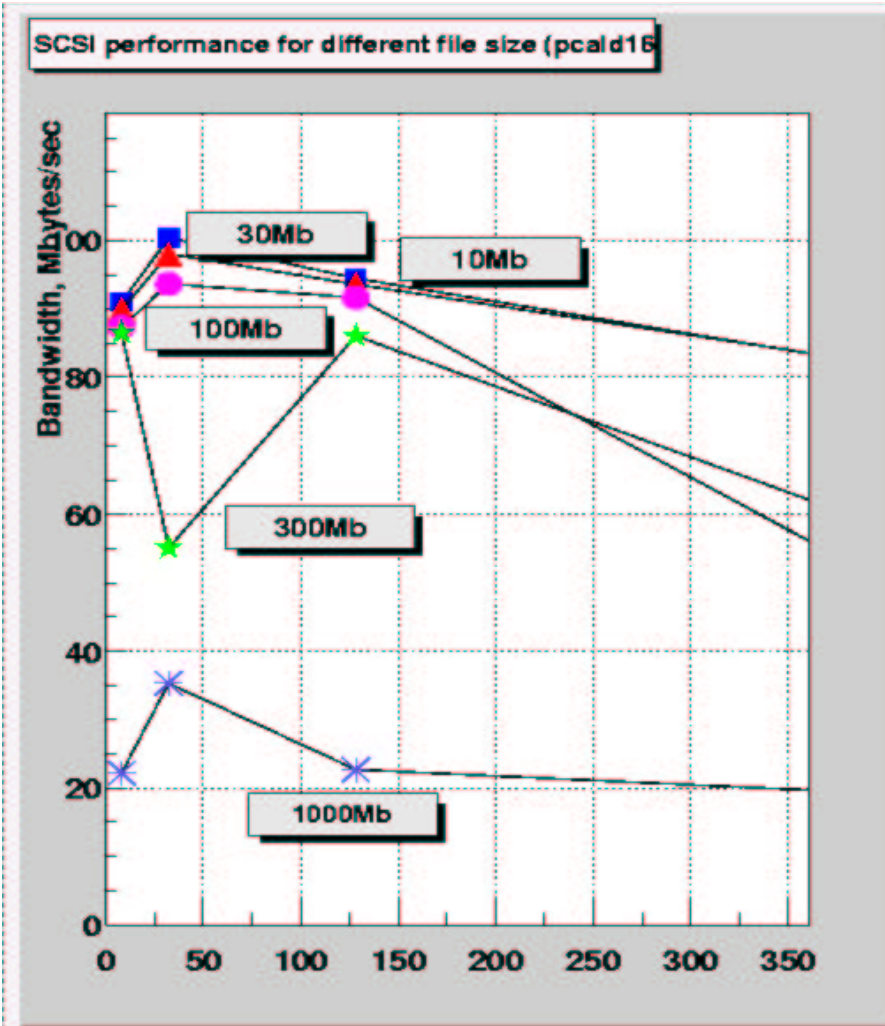


8 May 2002

1



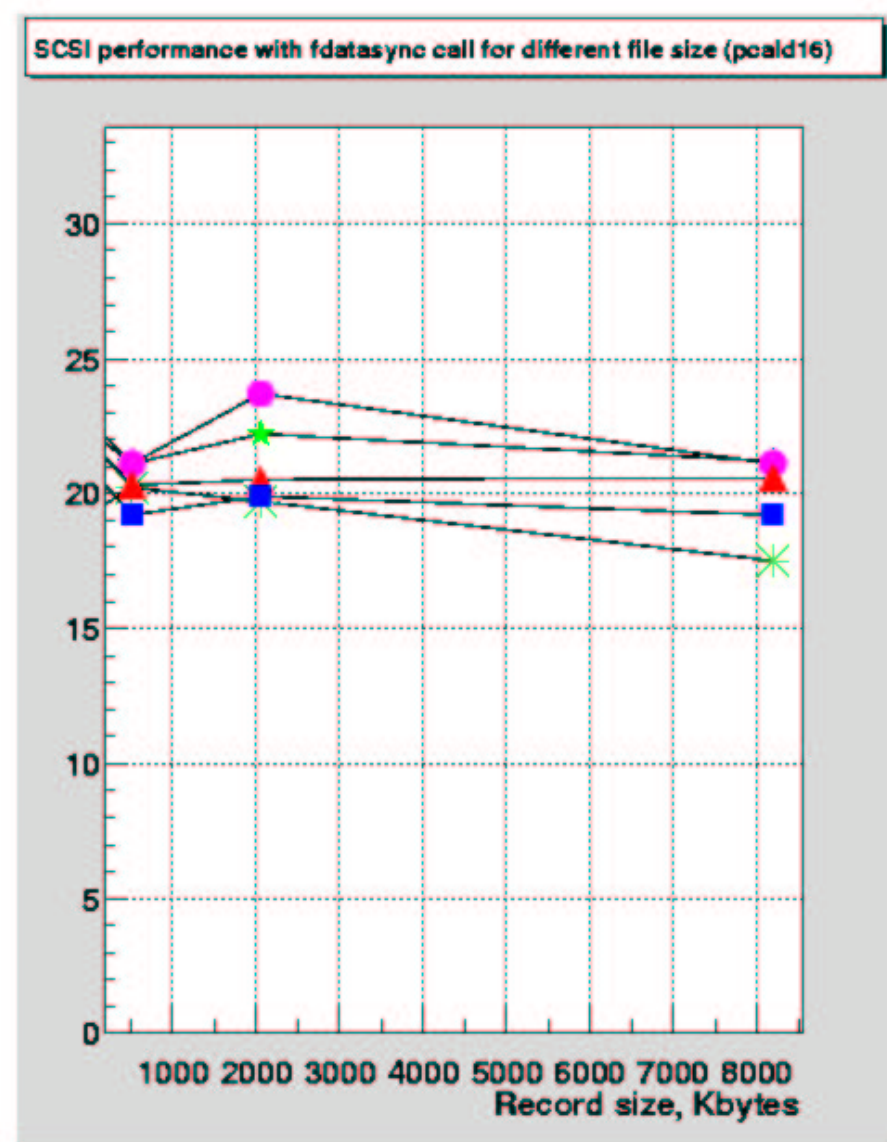
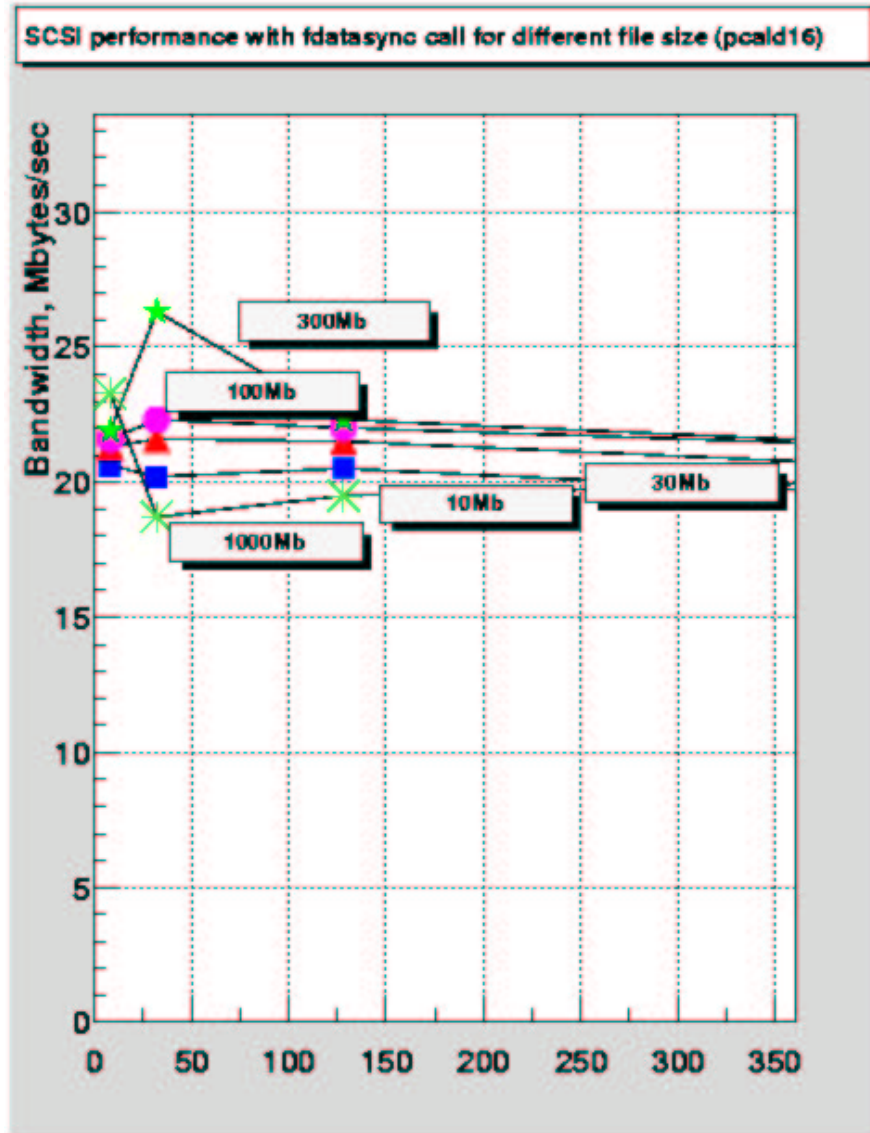
Storage: file & record size (file cache active)



Burst performance ! Irrelevant for HEP !

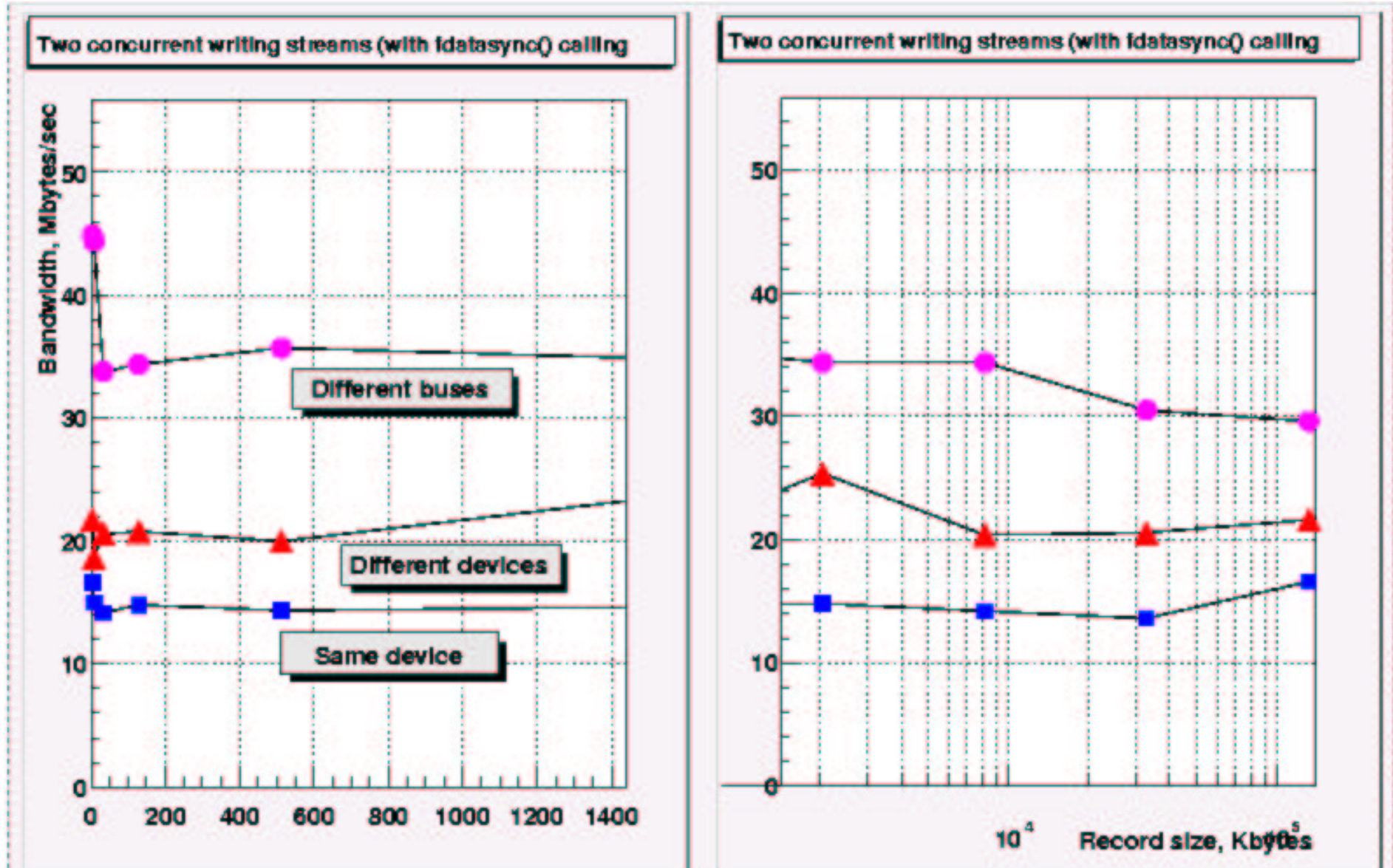


Storage: file & record size (file cache inactive)





Storage: effect of connectivity





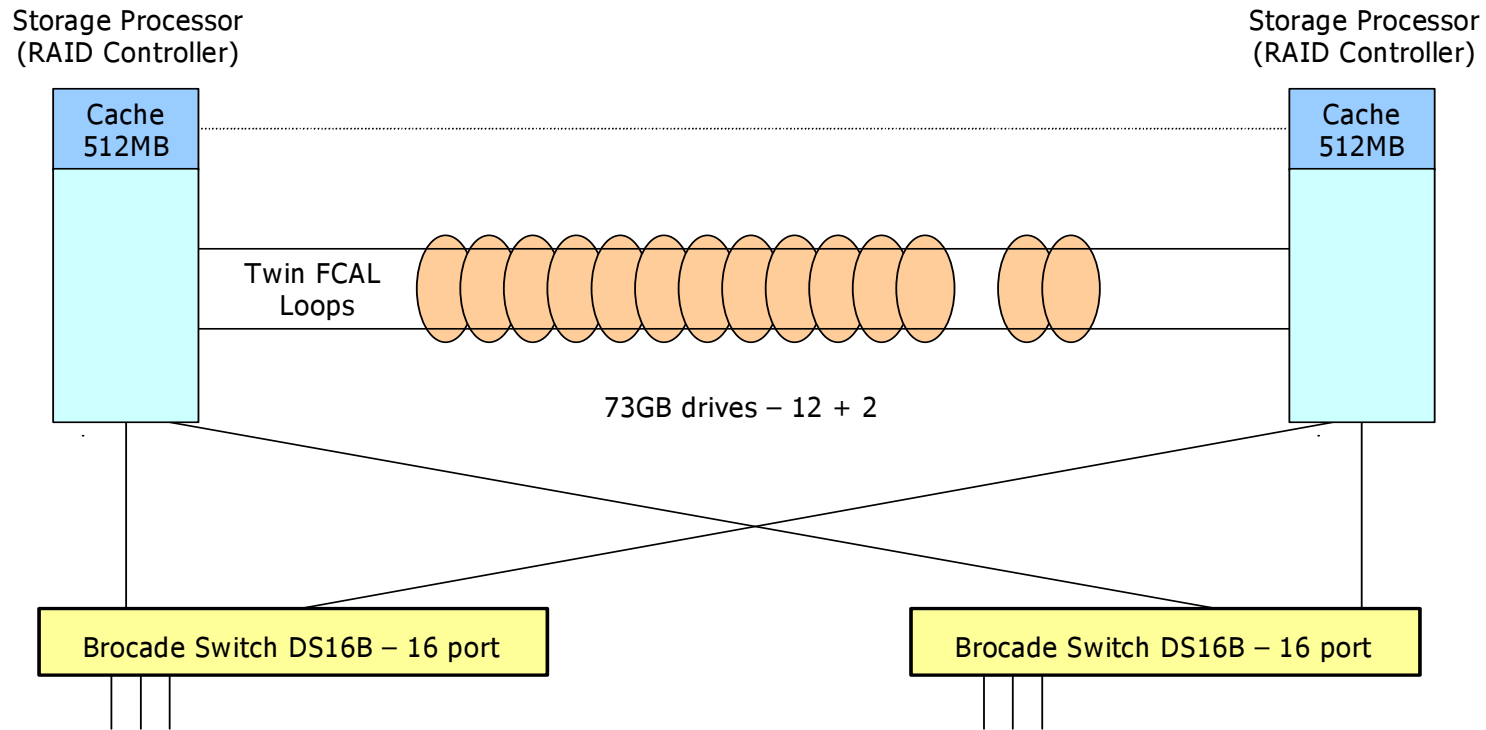
Transient Data Storage

- ◆ Disk storage highly non scalable
- ◆ To achieve high bandwidth performance
 - 1 stream, 1 device, 1 controller, 1 bus
 - With these conditions, sustained transfer bw to media:
 - 15-20 MB/s with 7.5 kRPM IDE disks
 - 18-20 MB/s with 15 kRPM SCSI disks
- ◆ To obtain high bandwidth with commodity solutions
 - Footprint too big
 - Infrastructure cost too high
- ◆ More compact and stable performance
 - RAID (Redundant Array of Inexpensive Disks)
 - RAID 5, large caches, intelligent controllers
 - Lots of provider (Dot Hill, EMC, IBM, HP)
 - Bw: 30-90 Mbytes/s sustained



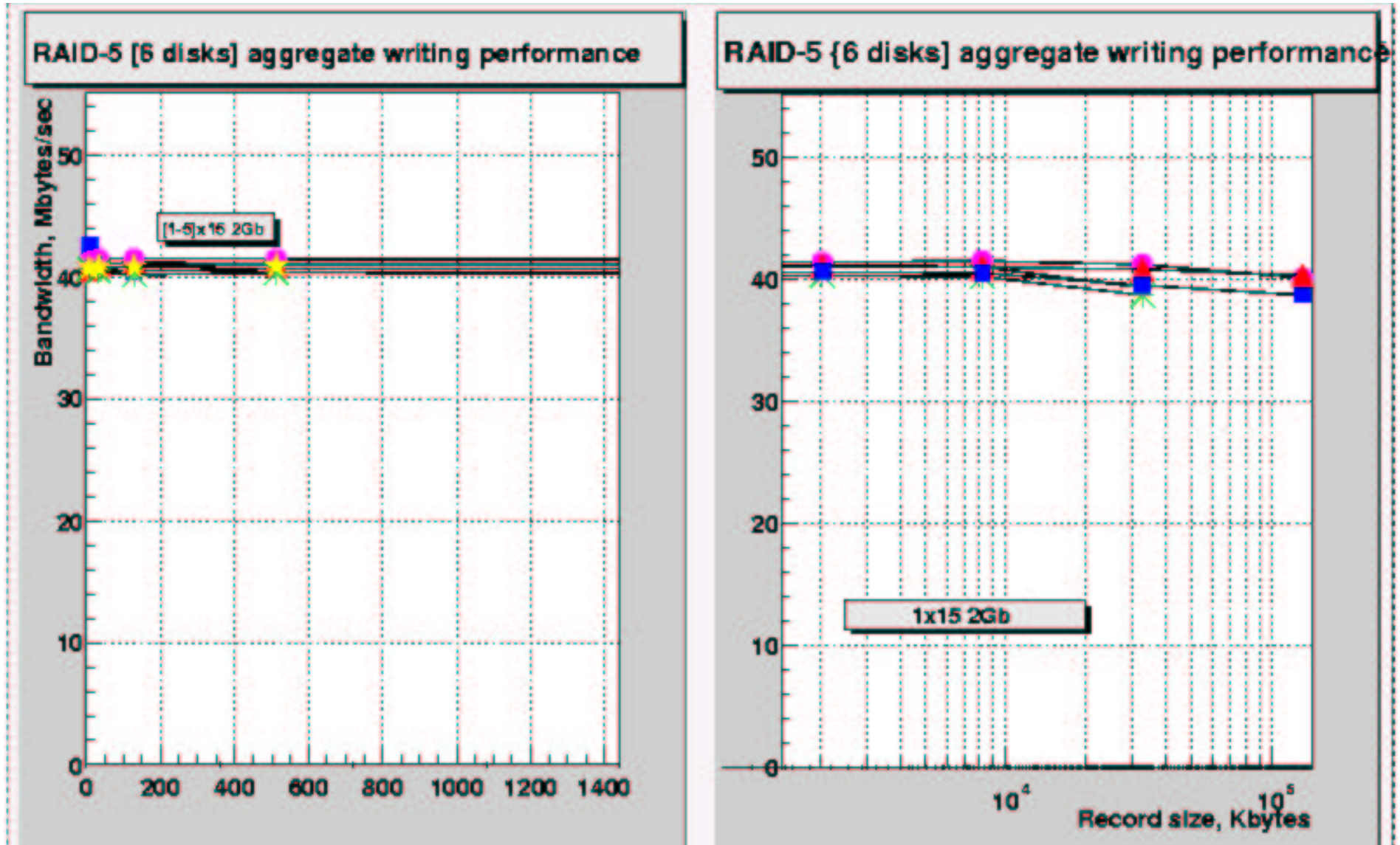
Storage Array

EMC CLARiiON FC4500 RAID Hardware





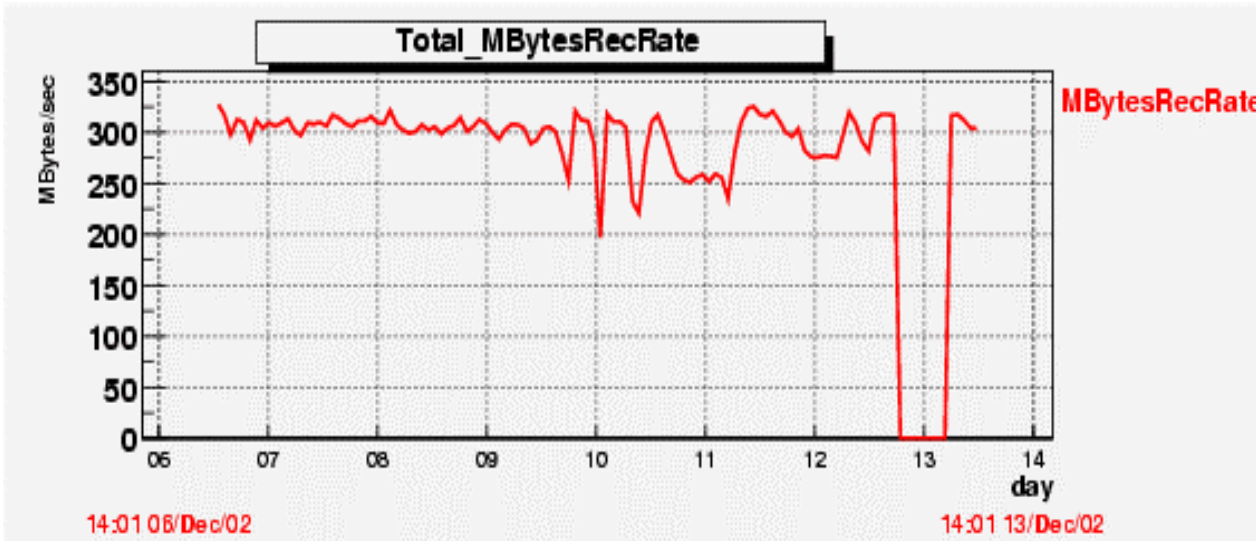
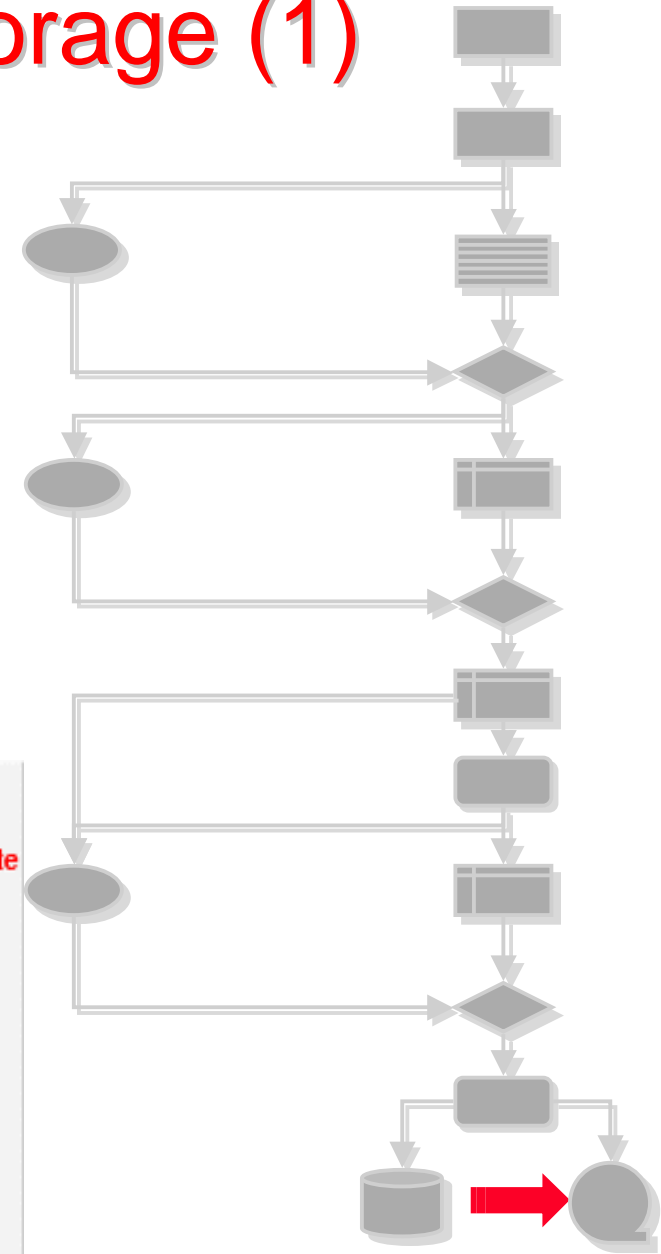
Storage: effect of SCSI RAID





Permanent Data Storage (1)

- ◆ Infinite storage at very low cost
- ◆ 1 realistic solution: magnetic tape
 - Media: 0.3 \$/GByte
 - Density: 0.1 Gbit/in²
- ◆ Critical areas
 - Must be hidden by a MSS
 - Limited market, different application
 - Limited competition, no real alternative
- ◆ Demonstrated solution for LHC
 - 15 parallel streams





Permanent Data Storage (2)



Tape Drive
STK 9940A 10 MB/s
 60 GB/Volume
 SCSI
STK 9940B 30 MB/s
 200 GB/Volume
 Fibre Channel



Tape Library
Several tape drives of both generations

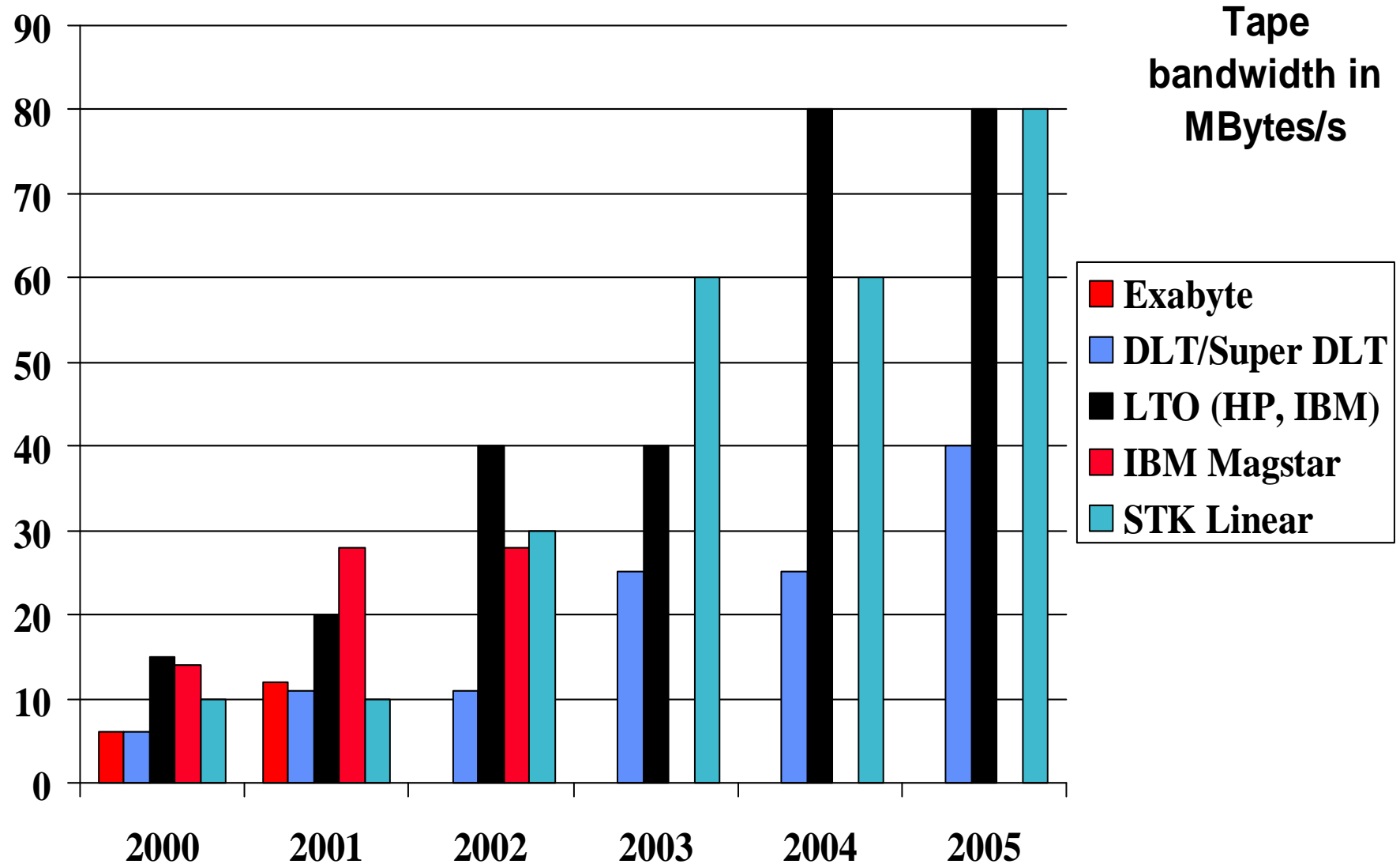


Permanent Data Storage (3)



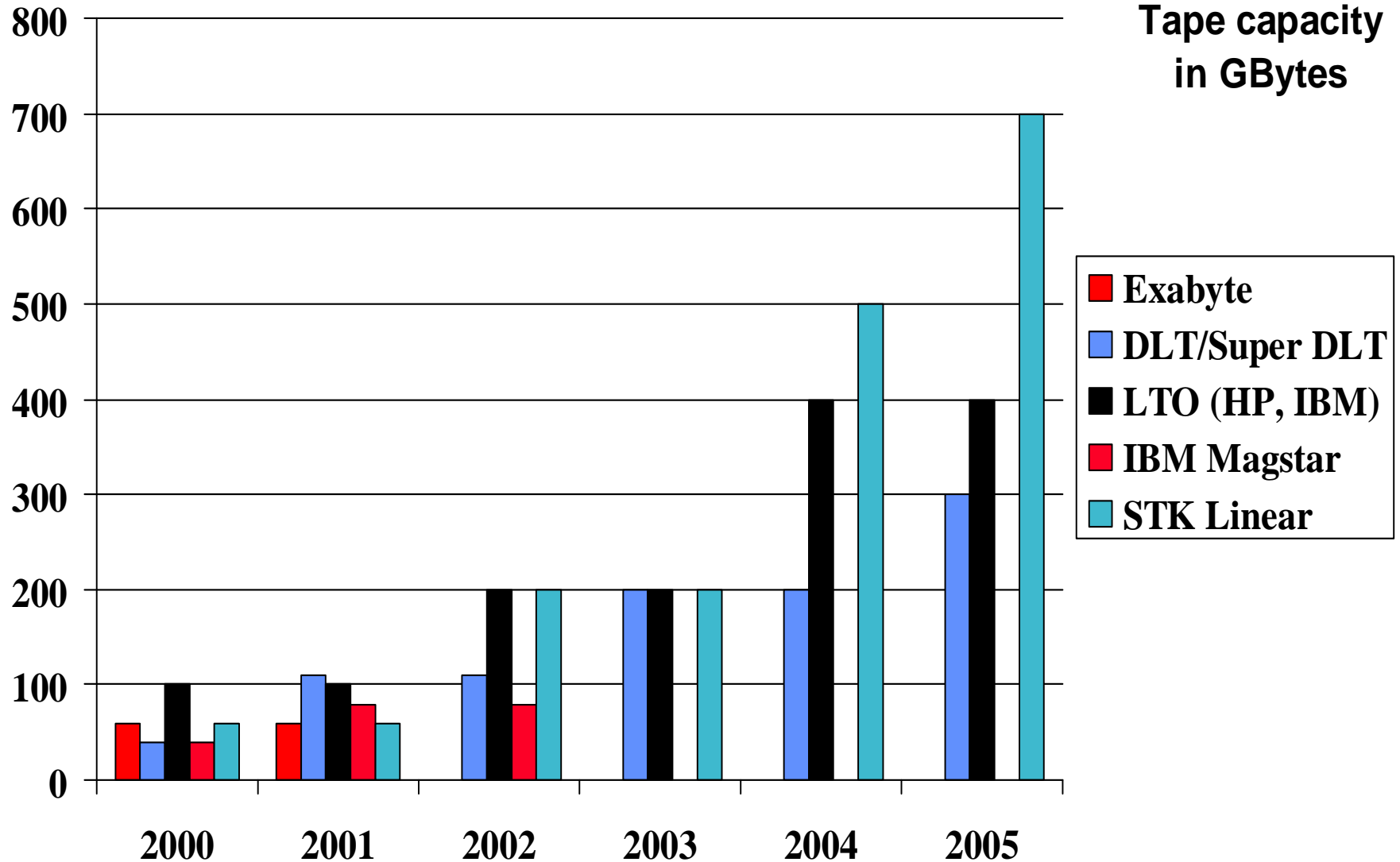


Storage: Tape Bandwidth (prevision)





Storage: Tape Capacity (prevision)





DAQ Software Framework

- ◆ DAQ Software Framework
 - Common interfaces for detector-dependant applications
 - Address all configurations and all phases from the start
 - For SLHC: handle more and more complexity

- ◆ DAQ Software
 - Complete ALICE DAQ software framework in 3 packages:
 - **DATE:**
 - ◆ Data-flow: detector readout, event building
 - ◆ System configuration, control (1000's of programs to start, stop, synchronize)
 - **AFFAIR: Performance monitoring**
 - **MOOD: Data quality monitoring**
 - Production-quality releases
 - Evolving with requirements and technology

- ◆ Key issues
 - Scalability (1 to 1000, demonstrate it)
 - Support and documentation

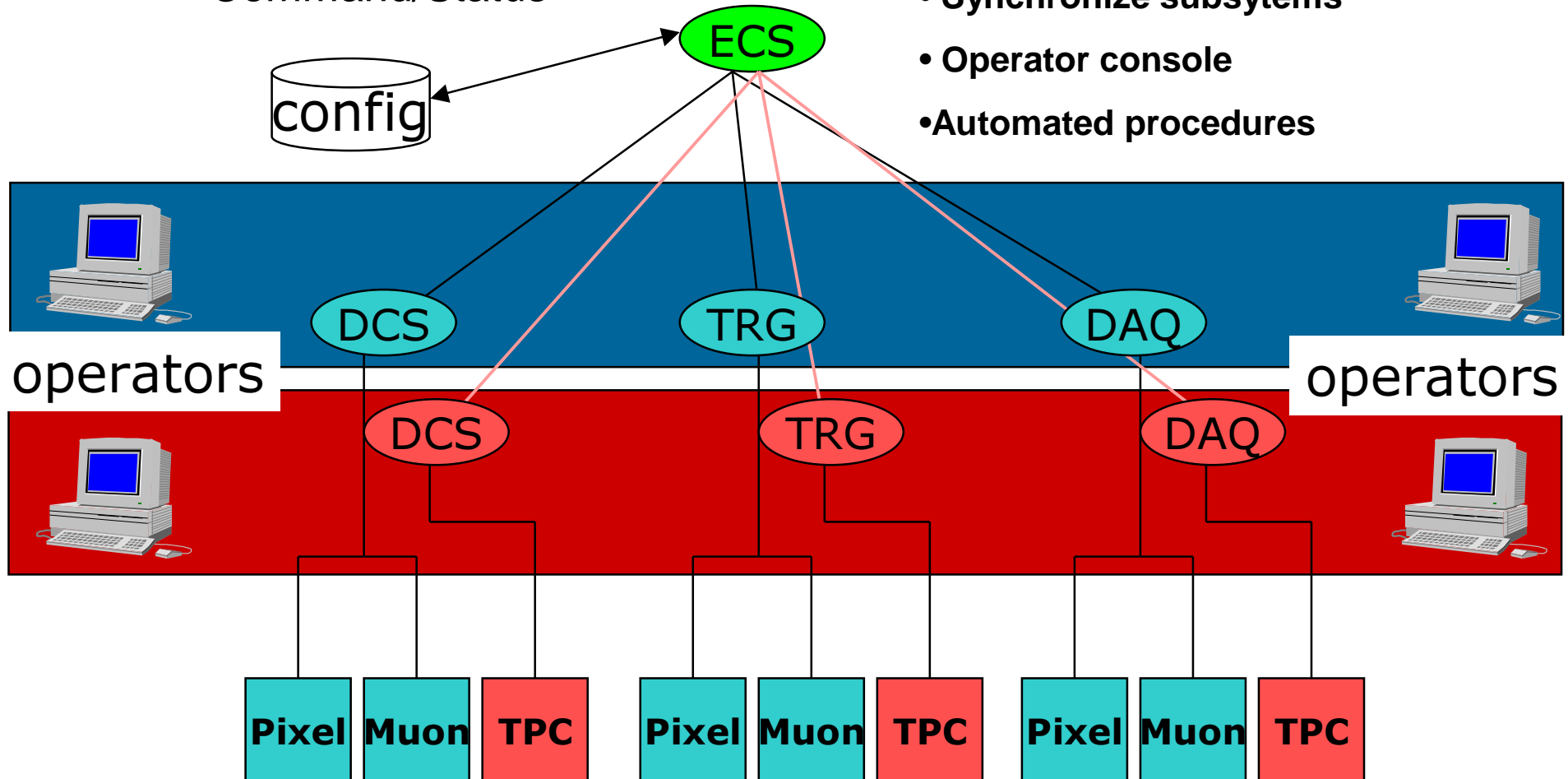


Experiment Control System

- State Machines
- Command/Status

ECS functions

- Configuration and booking
- Synchronize subsystems
- Operator console
- Automated procedures





Data Flow - DATE

DAQ - Run Control

DOMAIN: divia23073

Configuration
Run Parameters
Ready to start
Data Taking

Define

Show

Define

Show

Start run

Stop run

Run AutoStart

Autoset GDC

Recording Enabled

AFFAIR

EDM

ALIMDC

HLT

Pause trigger

Abort run

RUN NUMBER : 1785 DAQ Logic Engine Status : RUNNING

Info: Run 1785 running

Trace Fri 13 11:07 Run 1785 running

Clear Fri 13 11:07 Run number saved on /dateSiteAdc/configurationFiles/runNumber.config

 Fri 13 11:07 Starting run 1785

Debug Fri 13 11:07 * Message from tbed0029gdc: TRACE STOP_PROCESS: EVB 3223 has been killed as r

 Fri 13 11:07 * Message from tbed0029gdc: ACTION End of run requested with error

Pause Fri 13 09:10 * Message from tbed0049ldc: ERROR file /date/runControl/Linux/checkProc.sh problem

Bigger Fri 13 08:05 Run 1784 running

Smaller Fri 13 08:05 Run number saved on /dateSiteAdc/configurationFiles/runNumber.config

```

11:21am up 78 days, 22:29, 1 user, load average: 1.73, 1.69, 1.62
90 processes: 87 sleeping, 3 running, 0 zombie, 0 stopped
CPU0 states: 2.0% user, 50.5% system, 1.2% nice, 46.1% idle
CPU1 states: 3.0% user, 75.3% system, 2.0% nice, 21.0% idle
Mem: 384356K av, 374564K used, 9792K free, 3020K shrd, 147540K buff
Swap: 1044184K av, 26364K used, 1017820K free, 152456K cached
          
```

PID	USER	PRI	NI	SIZE	RSS	SHARE	STAT	%CPU	%MEM	TIME	COMMAND
15208	nobody	14	5	4080	4080	3644	R N	99.9	1.0	13:09	recorder
1334	root	9	0	2332	2284	1592	S	0.5	0.5	21:39	sshd
1574	root	9	0	1060	1060	820	R	0.3	0.2	30:05	top
3	root	19	19	0	0	0	SWN	0.1	0.0	13:47	ksoftirqd_CPU0
1337	root	9	0	2368	2364	1856	R	0.1	0.6	3:17	xterm
5070	nobody	8	0	4004	3976	1468	S	0.1	1.0	10:00	rcServer
1	root	9	0	496	448	448	S	0.0	0.1	0:12	init
2	root	8	0	0	0	0	SW	0.0	0.0	0:00	keventd
4	root	19	19	0	0	0	SWN	0.0	0.0	12:06	ksoftirqd_CPU1
5	root	9	0	0	0	0	SW	0.0	0.0	1:58	kswapd
6	root	9	0	0	0	0	SW	0.0	0.0	0:00	kreclaimd
7	root	9	0	0	0	0	SW	0.0	0.0	0:00	bdflush
8	root	9	0	0	0	0	SW	0.0	0.0	0:01	kupdated
9	root	-1	-20	0	0	0	SW<	0.0	0.0	0:00	mdrecoveryd
15	root	9	0	0	0	0	SW	0.0	0.0	0:00	scsi_ah_0
16	root	9	0	0	0	0	SW	0.0	0.0	0:00	scsi_ah_1

SD

LDC status display

LDC name	tbed0001ldc	tbed0013ldc	tbed0030ldc	tbed0037ldc
Event rate	13	13	14	13
Bytes recorded rate	40.182 M	41.203 M	41.938 M	40.163 M
Bytes in buffer	C 1192% M 1195%	C 1188% M 1193%	C 1192% M 1194%	C 1187% M 1191%
Number of events	10453	10462	10457	10450
Events recorded	9816	9825	9820	9813
Bytes injected	31'031'205'136	31'057'922'896	31'043'079'696	31'022'299'216
Bytes recorded	29'141'863'284	29'175'752'364	29'154'396'136	29'140'480'912
Readout SOR/EOR phases	0	0	0	0
Recorder SOR/EOR phases	0	0	0	0

GDC status display

GDC name	tbed0003gdc	tbed0004gdc	tbed0014gdc	tbed0015gdc
Events received	4924	5170	5438	3505
Events recorded	622	639	673	432
Bytes received	14'588'026'944	15'347'910'144	16'167'256'896	10'428'860'800
Bytes recorded	14'392'096'256	15'175'728'576	15'983'200'832	10'259'648'000
Event builder SOR/EOR phases	0	0	0	0
Status	FULL	FULL	FULL	

EDM status display

EDM name	tbed0015edm
wakeUpId received	(nblnRun:10442)
maxWakeUpId	(nblnRun:10442)
lastThresholdSent	(nblnRun:10454)
lastUpperBoundSent	(nblnRun:10464)
edmMask	[0]:00040000 [1]:00000100
Excluded	3 4 14 26 29 41 50 51 64 65 74 75 96 97

```

11:21am up 78 days, 22:29, 1 user, load average: 1.73, 1.69, 1.62
90 processes: 87 sleeping, 3 running, 0 zombie, 0 stopped
CPU0 states: 2.0% user, 50.5% system, 1.2% nice, 46.1% idle
CPU1 states: 3.0% user, 75.3% system, 2.0% nice, 21.0% idle
Mem: 384356K av, 374564K used, 9792K free, 3020K shrd, 147540K buff
Swap: 1044184K av, 26364K used, 1017820K free, 152456K cached
          
```

PID	USER	PRI	NI	SIZE	RSS	SHARE	STAT	%CPU	%MEM	TIME	COMMAND
31330	nobody	14	5	187M	187M	187M	R N	44.8	50.0	5:23	eventBuilder
31363	alicemdc	13	5	188M	188M	187M	S N	23.5	50.1	3:47	writeCastor_v3
28903	pvv	9	0	992	948	748	S	1.9	0.2	40:30	top
15701	root	14	0	1052	1052	820	R	1.7	0.2	1:56	top
3	root	19	19	0	0	0	RWN	0.7	0.0	21:16	ksoftirqd_CPU0
4131	root	9	0	2176	1724	1496	S	0.3	0.4	23:19	sshd
496	root	9	0	532	532	448	S	0.3	0.1	0:00	sleep
838	ntp	9	0	1924	1924	1732	S	0.1	0.5	0:14	ntpd
18454	root	9	0	1236	1080	964	S	0.1	0.2	0:05	xload
1	root	8	0	488	440	424	S	0.0	0.1	0:20	init
2	root	8	0	0	0	0	SW	0.0	0.0	0:00	keventd
4	root	19	19	0	0	0	RWN	0.0	0.0	21:13	ksoftirqd_CPU1
5	root	9	0	0	0	0	SW	0.0	0.0	1:38	kswapd
6	root	9	0	0	0	0	SW	0.0	0.0	0:00	kreclaimd
7	root	9	0	0	0	0	SW	0.0	0.0	0:00	bdflush
8	root	9	0	0	0	0	SW	0.0	0.0	0:03	kupdated



Run Control - DATE

State of one node

RUN Control

File View Options Windows

DAQ - Run Control

SMI Status

GDC (72)

- NOT_RUNNING
- STARTING
- STARTING_ALIMDC
- STARTING_EVB
- RUNNING
- RUNNING_ERR
- STOPPING_SHM_EVB
- STOPPING_EVB
- STOPPING_ALIMDC
- WAIT_STOPPED
- STOPPED
- RESETTING

LDC (72)

- NOT_RUNNING
- STARTING
- STARTING_EDMC
- STARTING_RECORDER
- STARTING_READOUT
- RUNNING
- STOPPING_SHM_READOUT
- STOPPING_READOUT
- STOPPING_SHM_RECORDER
- STOPPING_RECORDER
- STOPPING_SHM_EDMC
- WAIT_STOPPED
- STOPPED
- RESETTING

Ready to start

Start run

Stop run

Run AutoStart

Autoset GDC

Recording Enabled

Affair on

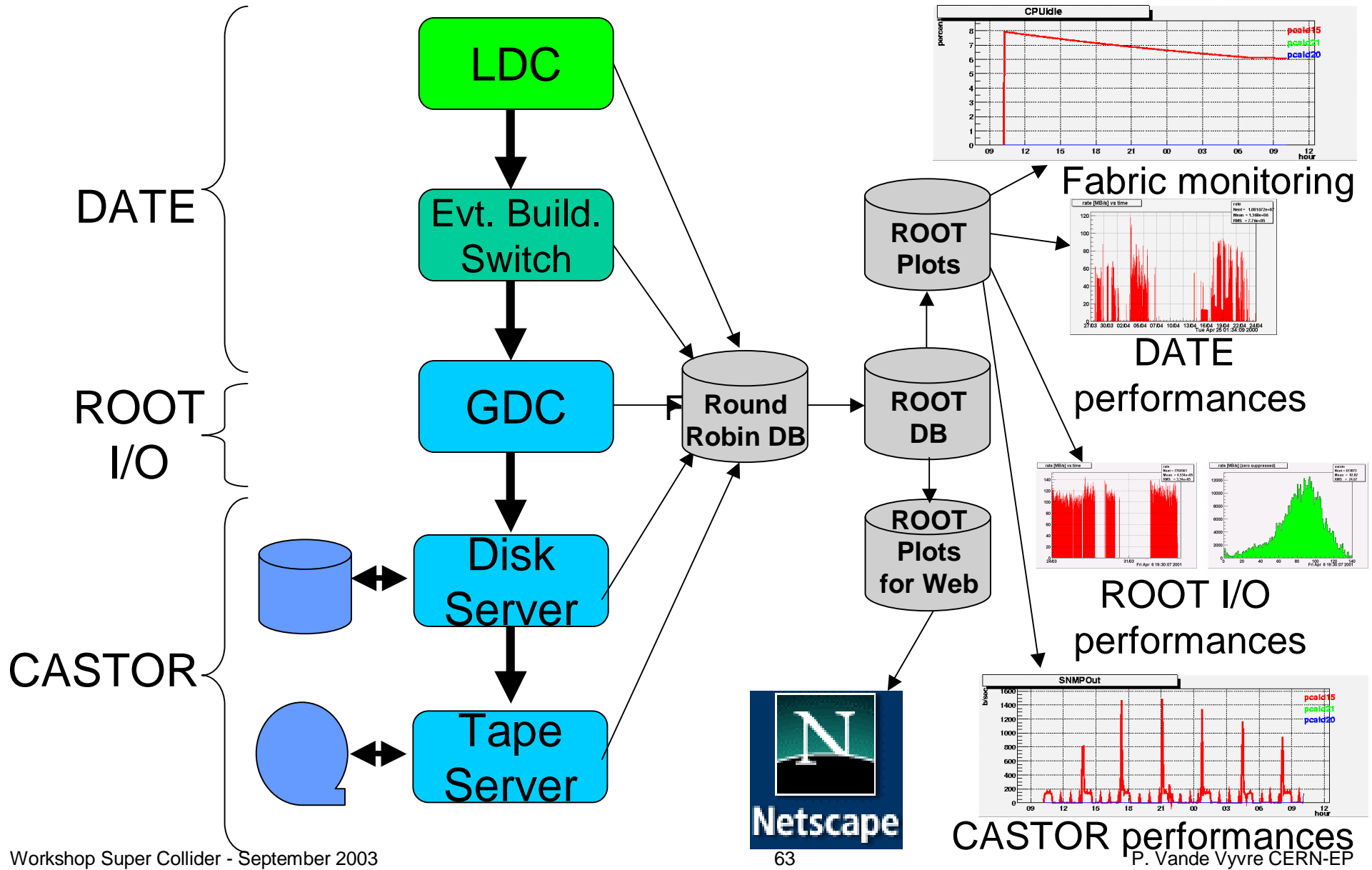
ALIMDC on

STOPPING_LDCS

...eAdc/configurationFiles/runNumber.
START_PROCESS TBED0027LDC_F
START_PROCESS TBED0020LDC_F
ic: START_PROCESS LXSHARE005I
START_PROCESS TBED0005LDC_F
START_PROCESS TBED0052LDC_F

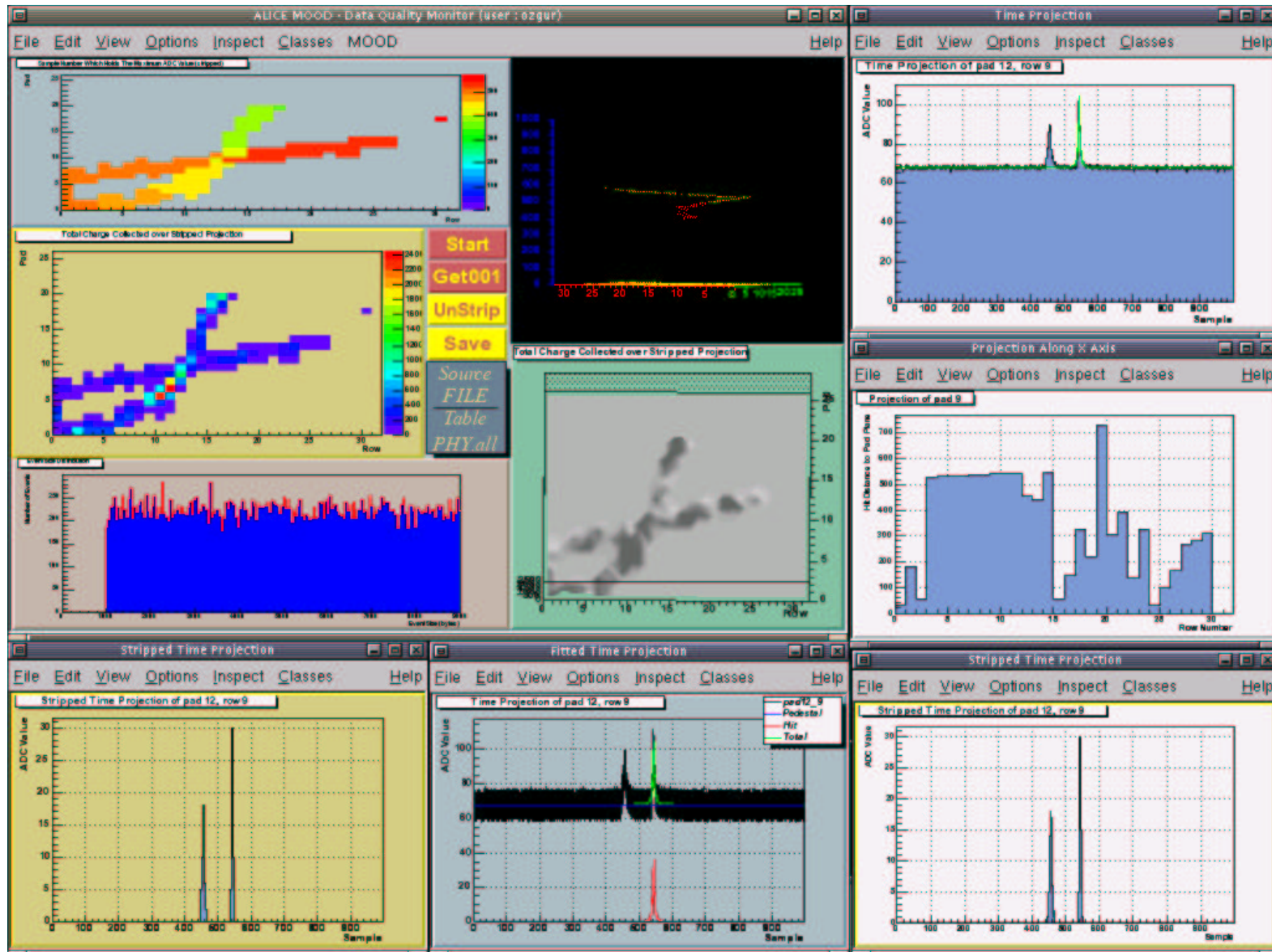


Performance monitoring - AFFAIR





Data Quality Monitoring - MOOD





DAQ for Super Collider Experiments

- ◆ DAQ and HLT of LHC experiments
- ◆ Supercollider reference
- ◆ Technology trends
- ◆ DAQ and HLT for SLHC experiments
- ◆ **R&D**
- ◆ Conclusions



R&D for the SLHC

- ◆ Semiconductor industry is the driving force:
 - Industry has learned to do switches for Telco:
 - Silicon has been developed
 - Exponential development of Internet: commodity networking
 - Switches at all levels in Trigger/DAQ architecture
 - Chips
 - Boards (Rapid I/O, HyperTransport)
 - Systems (switched LAN)
 - Collaboration (WAN at OC192-10 Gbit/s and OC768-40 Gbit/s)

- ◆ Questions to be considered
 - Permanent technological progress: hype or reality ?
 - Industry evolution: taking a “good” direction ?
 - Will HEP afford cost of R&D ?
 - How should the R&D be performed ?

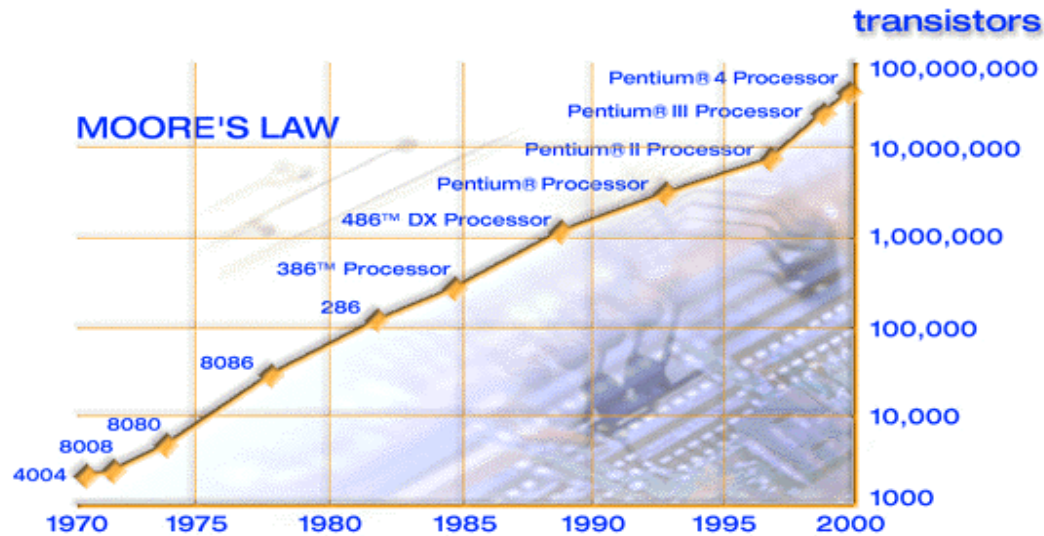


Moore's law: myth and reality (1)

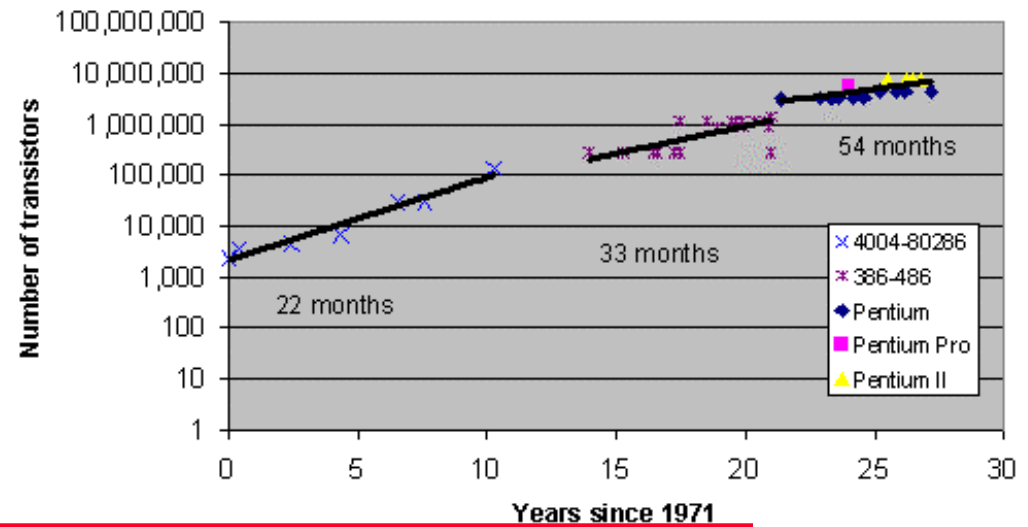
- ◆ Observation by G. Moore in 1965 when working at Fairchild
 - “Cramming more components onto integrated circuits”,
in Electronics Vol. 38 Nb 8, April 19, 1965
 - “Complexity of minimum cost semiconductor component had doubled every year”.
 - Cost per integrated component $\approx 1/\text{number of components integrated}$
But yield decreases when components added
∃ Minimum cost at any point in time
- ◆ In 1975, prediction that doubling every 2 years
 - G. Moore co-founded Intel
 - His law became the Intel business model
 - Initially applied to memory chips, then to processors
- ◆ Interpretation and evolution of Moore's law
 - In the 1980's: \Rightarrow doubling of transistors on a chip every 18 months
 - In the 1990's: \Rightarrow doubling of microprocessor power every 18 months
- ◆ Subject of debate in the semiconductor industry. However...
 - Intel: in 1971 the 4004 had 2250 transistors, in 2000 the PIV had 42 Millions
 - Exponential evolution over 30 years



Moore's law: myth and reality (2)



© Intel corp.



Mr. Ilkka Tuomi

Verify real performance with HEP application



Evolution could go in a bad direction...

- ◆ Vulnerability
 - HEP depends upon evolution of commodity markets
- ◆ A typical example
 - PC form factor not well adapted to the vast majority of end-users
 - **Who wants to change graphics card ?**
 - The present format (desktop with a PCI bus) handy for HEP
 - Mass market could go for a closed box (such as video games)
 - Video games platform:
 - **Hw and system Sw fixed; only application sw change**
 - **Price does not cover the cost. Benefits done with the appl. sw**
 - **Unusable for HEP.**
- ◆ Situation not so bad
 - HEP using 2 CPUs machines
 - HEP is not alone. Lots of applications: computing centres, ISP etc.



...or in the good direction

- ◆ Need to move data continues to increase
 - The cost of moving data continues to decrease
- ◆ Largest Gbit Eth. switches: Multi Tbits/s
- ◆ 10 Gbit/s networking
 - Components exist but the price is high or even outrageous
 - LAN (10 Gbit/s Eth port): 25-75 k\$, 5k\$ in 2006
 - WAN (10 Gbit/s SONET/SDH): 150-325 K\$
 - Present period of economic restriction not favorable but the deployment has started
- ◆ Optical switching is the next big evolution
 - Components exist
 - Application exist
 - Commercialization requires huge investments and will take time



Can HEP afford R&D ?

- ◆ Resources needed
- ◆ Collaboration extremely collaborative with networking and computing industry
 - Early access to new products
 - HEP has demanding needs and contributes efficiently to field-testing
 - Substantial contribution to R&D
- ◆ Might be more difficult for chip development
 - New semi-conductor fab for 90 nm: \cong 1 B\$
 - Small number of players
 - Investment can only be absorbed by very large volumes
 - Commodity products: mobile phones, PDA, PC (CPU and DRAM)
 - Little room for tests of new ideas or for small productions



“R&D humanum est” (1)

◆ RD-27

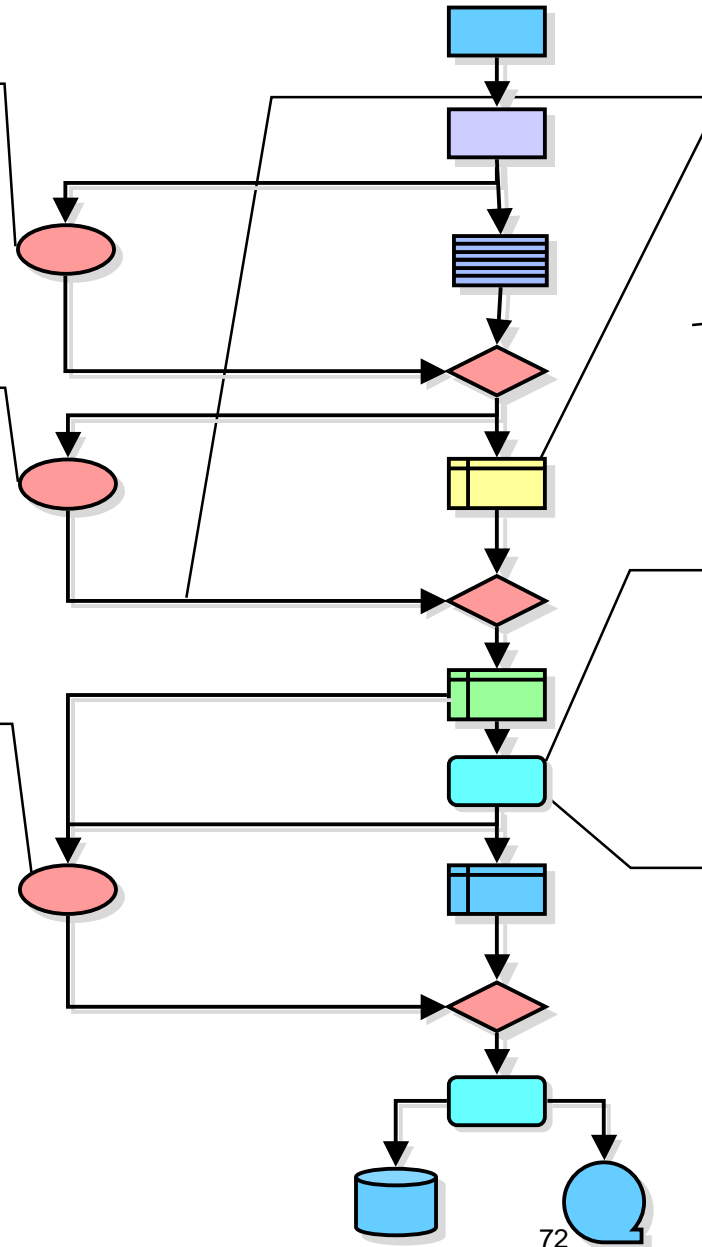
First-level trigger systems for LHC experiments.

◆ RD-11

EAST Embedded architectures for second-level triggering in LHC experiments

◆ LCB_005

Event Filter Farm



◆ RD-12

Readout system test benches.

◆ RD-13

A scalable data taking system at a test beam for LHC.

◆ RD-24

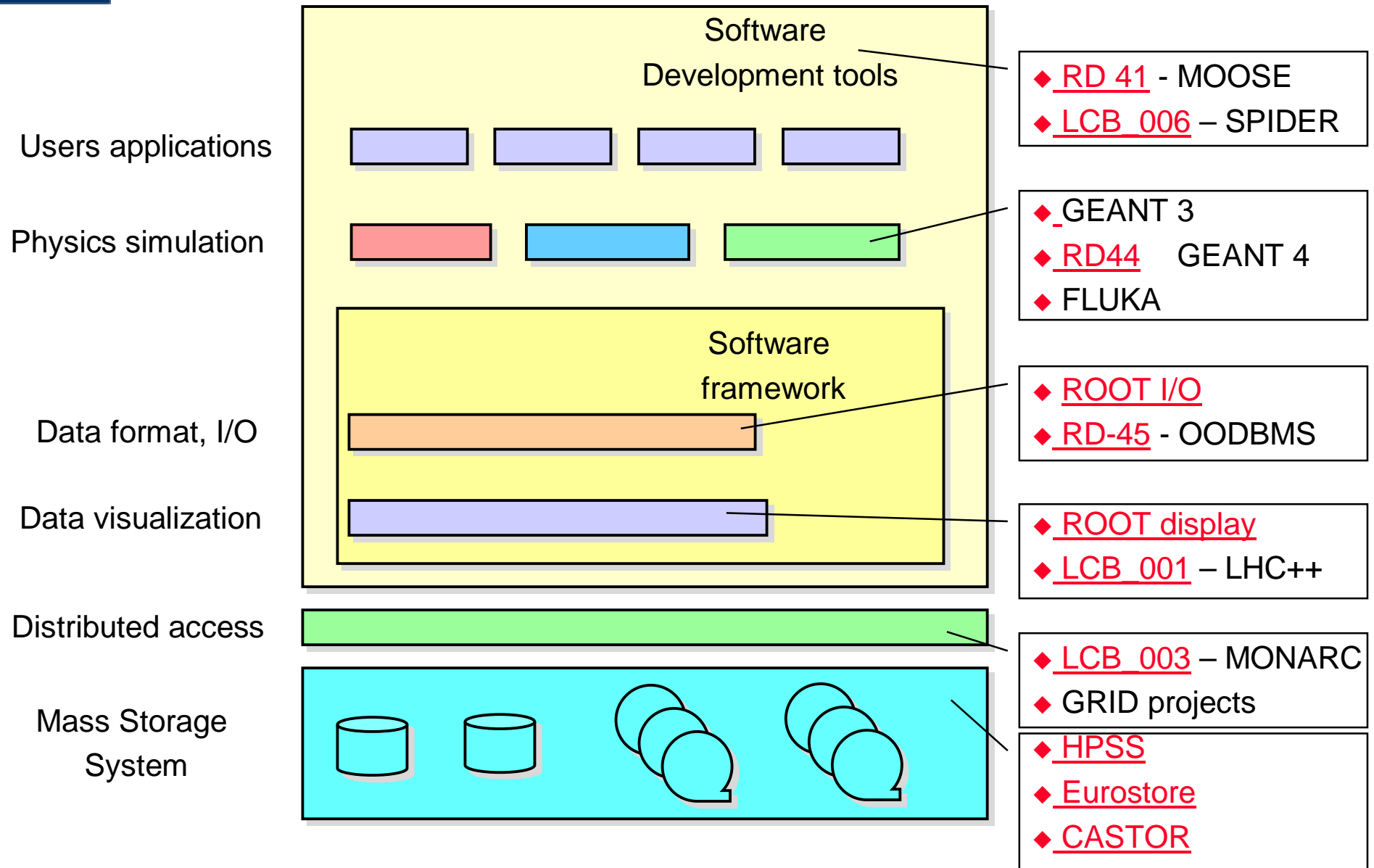
Applications of the scalable coherent interface to data acquisition at LHC (SCI).

◆ RD-31

NEBULAS: An asynchronous self-routing packet-switching network architecture for event building in high rate experiments (ATM).



“R&D humanum est” (2)





Outcome of LHC R&D

- ◆ Design and implementation of hardware components
 - TTC system for the trigger distribution
- ◆ Design and implementation of software packages
 - ROOT package e.g.
- ◆ Proof of concept of major concepts
 - Positive recommendation of using a communication switch for the event building based on tests with ATM. Different technologies considered today (Gigabit Ethernet, Myrinet).
- ◆ Positive recommendation of technologies
 - Object Oriented (OO) programming for the LHC software.
- ◆ None or few negative recommendations but some recommended technologies have not been adopted by experiments
 - Commercial software for offline framework
 - OO database for the storage of raw data
 - Usage of Microsoft Windows for physics data processing



Lessons from LHC R&D for DAQ and HLT

- ◆ HEP specific but ample usage of commercial elements
- ◆ R&D ? Not really...
 - Influence of industrial developments: track technology
 - Maintain and develop competence
- ◆ Best result for problem-oriented not technology oriented
 - Risks associated with cutting-edge technology
 - Technology development failure
 - Not adopted by industry
 - Taken-over by the next technological wave
 - Push 1 technology at all costs (e.g. OODB for raw data)
- ◆ Different approaches
 - Event building: network-based (ATM, FCS) or memory-based (SCI)
 - Network-based was the undisputed winner but with different technologies (switched Ethernet and Myrinet)
- ◆ Progress monitoring
 - Factual deliverables (“paperware” is not enough)
 - Open development
 - Early exposure to end application
- ◆ Long and repeated delays for computer-technology based R&D project indicate a lack or diminishing interest from industry



Conclusions

- ◆ DAQ and HLT of LHC experiments (reference architecture)
 - Similar architecture, comparable concepts
 - Large and complex systems made of 1000s of commodity components
- ◆ Super Collider reference model: LHC luminosity upgrade
 - Higher tracker occupancy
 - DAQ and HLT: increased needs for data transfer and processing
- ◆ Technology evolution
 - Data processing: current evolution will carry at least up for the next few years
 - Data transmission
 - 10 Gbit/s point-to-point, optical switching
 - Fractal explosion of switched architecture (boards, subsystems, DAQ, HLT)
- ◆ DAQ for SLHC:
 - Ingredients: 2 CPUs PCs, Linux, switched Eth., IDE disks with RAID and SAN, mag. tape
- ◆ R&D
 - DAQ and HLT: more technology tracking than pure R&D. Application driven.
 - Strong links with industry
 - Critical areas: access to micro-electronics fabs, R&D process