

Distributed Computing A Historical Perspective: The SHIFT Project

Frédéric Hemmer
CERN – EP Department
LHCb Experiment

13 May 2022

Microsoft PowerPoint



PowerPoint can't read C:\Cernbox\Documents\2022\1997-06-03 - COMPASS - PCSF (X!).ppt because the PowerPoint 95 file format is no longer supported.

OK

Help

Aknowledgements

This represents the work of many people, including A. Baran, J-P. Baud, C. Boissat, F. Cane, F. Hemmer, E. Jagel, S. Jarp, A. Kumar, A. Lee, G. Lee, T. Mouthuy, M. Schroeder, A. Simmins, B. Panzer-Steindel, L. Robertson, B. Segal, A. Trannoy, I. Zacharov and all the ones I forgot..

I have tried to reuse old slides with vintage PowerPoint templates as far as possible. Sometimes with surprises... Clarifications added in grey.

Outline

- Looking backwards
- HOPE, CSF and SHIFT
- Birth of PCs as a Physics Processing platform



Back in History

13 May 2022

Frédéric Hemmer







We were using and ... mounting tapes 24x7

CERN Central Computing Services

- 1970s:
 - CDC 7600 & IBM MVS mainframes
 - general purpose services
 - batch, tapes
 - interactive job submission, editing
- early 1980s:
 - VAX/VMS interactive services
 - special purpose services (DB, CAD/CAM, engineering)
 - rapid growth of external and internal networking
 - first Unix (VAX) service and Apollo workstations - 1982
- second half of the 1980s:
 - IBM/VM+VAX/VMS: interactive services
 - IBM/VM+Cray batch,tapes & growing disk configuration
 - explosion of networking
 - support for distributed computing
 - Workstations : VAXstation 2000 & Apollo
 - PC's



Workstation clusters for Physics Production

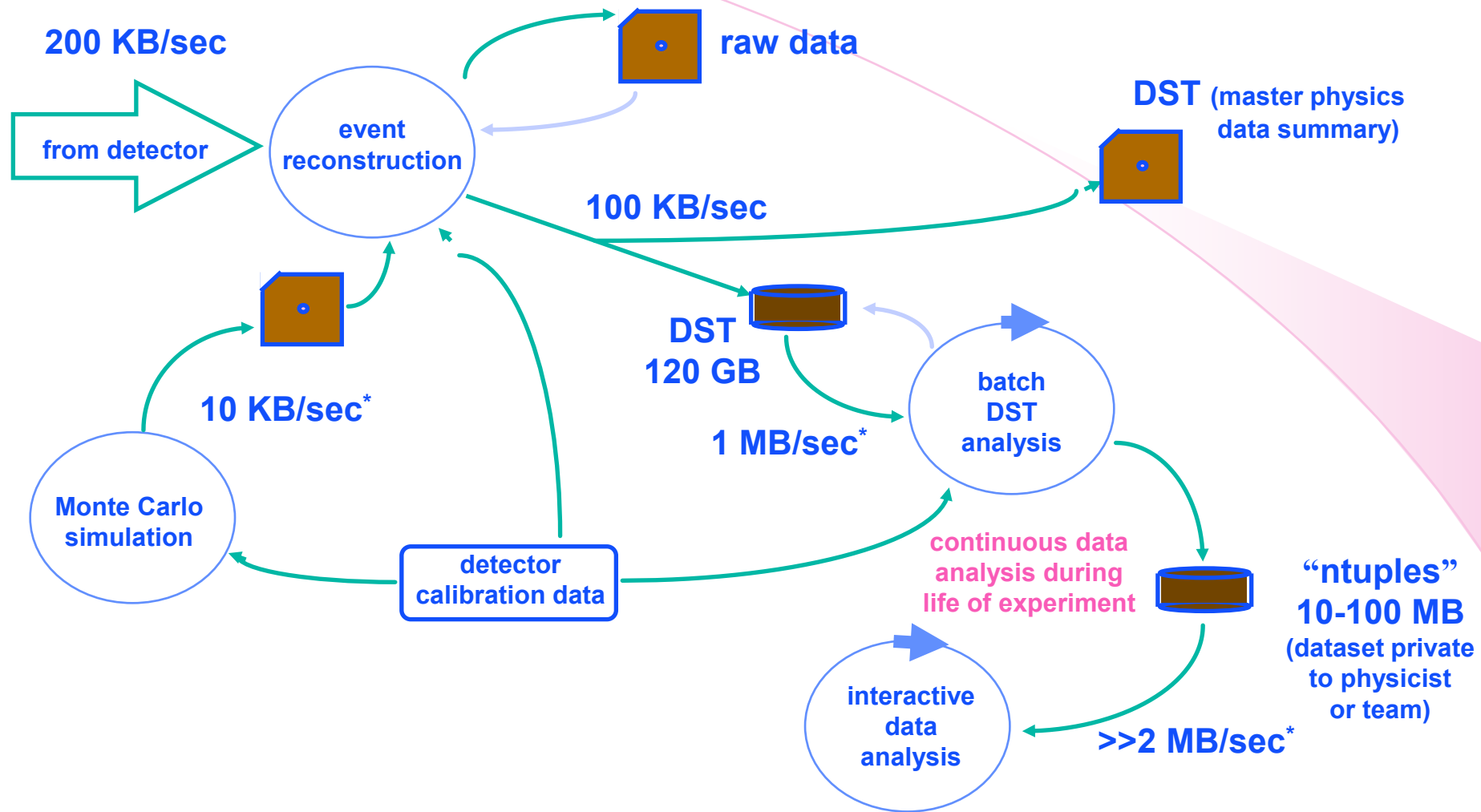
- 1988 LEP accelerator begins operation
 - Workstation clusters used for data reconstruction at experimental areas, replacing custom-built processors (“IBM emulators”)
 - growing number of private clusters (mostly VAX/VMS) for data analysis
 - BUT: IBM/VM remains the principal provider of
 - Batch+tape+disk services
 - public interactive service

RISC & SCSI

- 1989: Arrival of RISC processors
 - power comparable with mainframes (for HEP codes)
 - astounding price/performance
- ... and inexpensive SCSI disks

Workstations look attractive as a platform for central services

LEP Experiment Data Flow - 1992



data rates denoted * are per fast RISC CPU



Central Computing Services **History**

- **1989** ***HOPE***
- **1990** ***CSF***
- **1991** ***SHIFT***
- **1992** **CRAY X/MP disappears**
- **1994** **ES/9000 -> 3090/J@6 procs**
- **1995** **3090J@3 procs**
- **1996** **IBM mainframe disappears**

HOPE 1990-1991

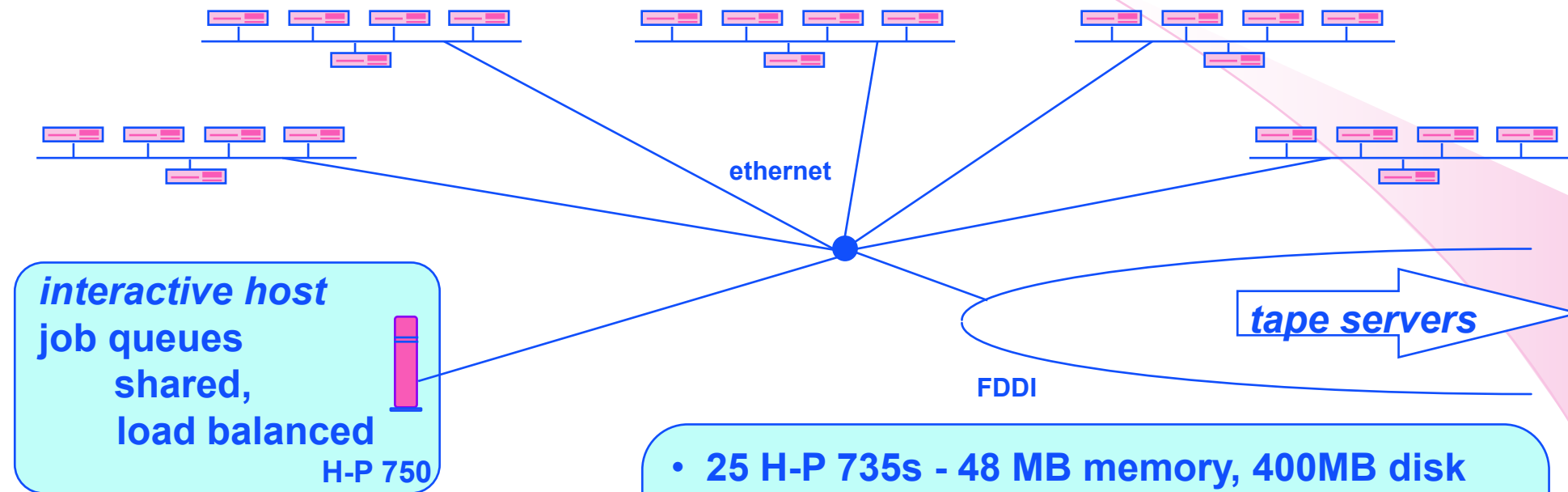
- *HP provided funding for an experimental batch service using Apollo DN10000 processors*
 - 3 Apollo DN10040 systems (12 CPU's)
 - NQS batch system (as on CRAY)
 - STK tape drives
 - Added FDDI later
- *usage*
 - data reconstruction (moderate I/O): tape -disk -tape
 - but mainly used for simulation (low I/O)
- *limitations*
 - network performance
 - tape support

CSF

- Started in 1990 - successor to HOPE
- Low cost HP 9000/725
- 16->25->45 machines
- CSF/2 HP 715 lower cost/CERN Unit, but speed actually lower than CSF/1
- CSF/3 will use PC's (1996)

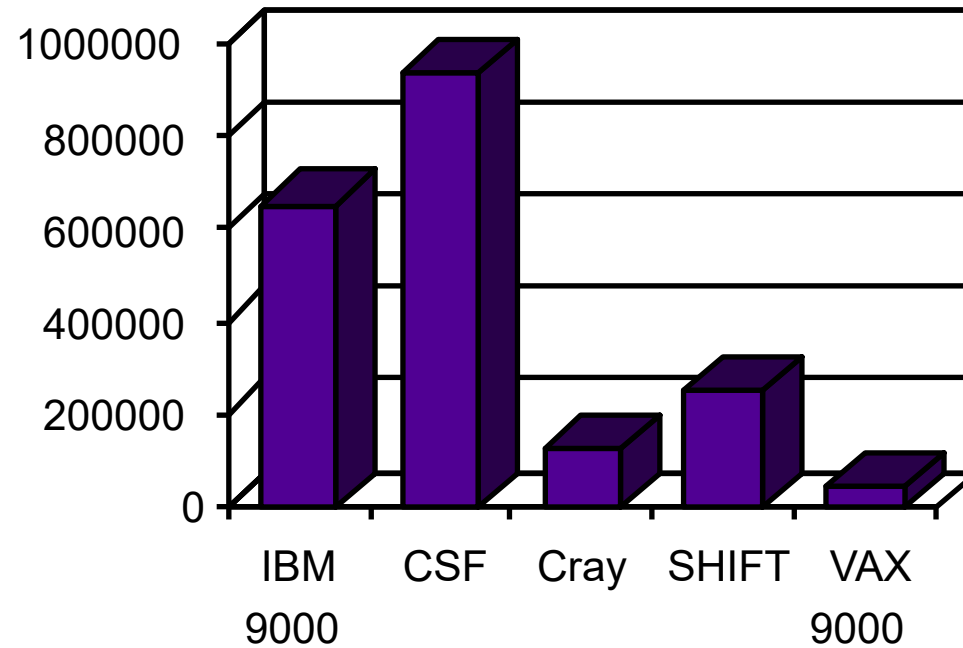
CSF - Central Simulation Facility

□ second generation, joint project with H-P



- 25 H-P 735s - 48 MB memory, 400MB disk
- one job per processor
- generates data on local disk
- staged out to tape at end of job
- long jobs (4 to 48 hours)
- very high cpu utilisation : >97%
- very reliable : > 1 month MTBI

Delivered CPU 1992



SHIFT

- Designed in 1990
 - fast access to large amounts of data
 - good tape support
 - cheap & easy to expand
 - vendor independent

- First implementation in operation in Jan 1991

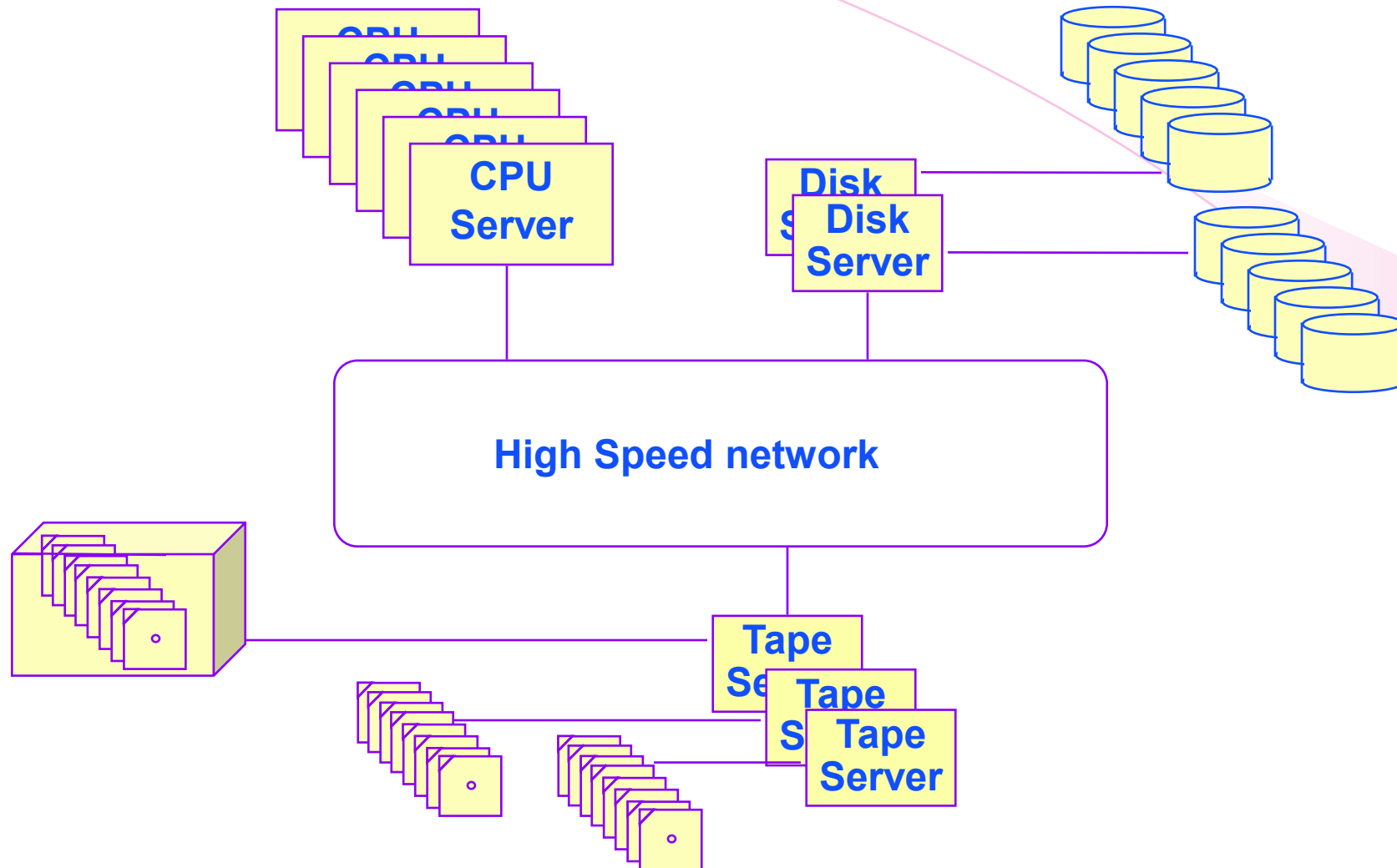
Many performance studies (network – and disk):

An example from sink.c – January 1990

```
/* This program consumes a continuous stream of records of length
   "record_length" bytes from the TCP socket specified by "host_name"
   and "port_number". */
/* Les Robertson, CERN/DD, January 1990 (adapted from an example in
   "C Programming in the Berkeley Unix Environment", R.N.Horspool,
   Prentice-Hall Canada - 1986. */
/*
 * The code was also modified by Ben Segal to include SORECVSIZE socket option
 *
 */
```



SHIFT Model



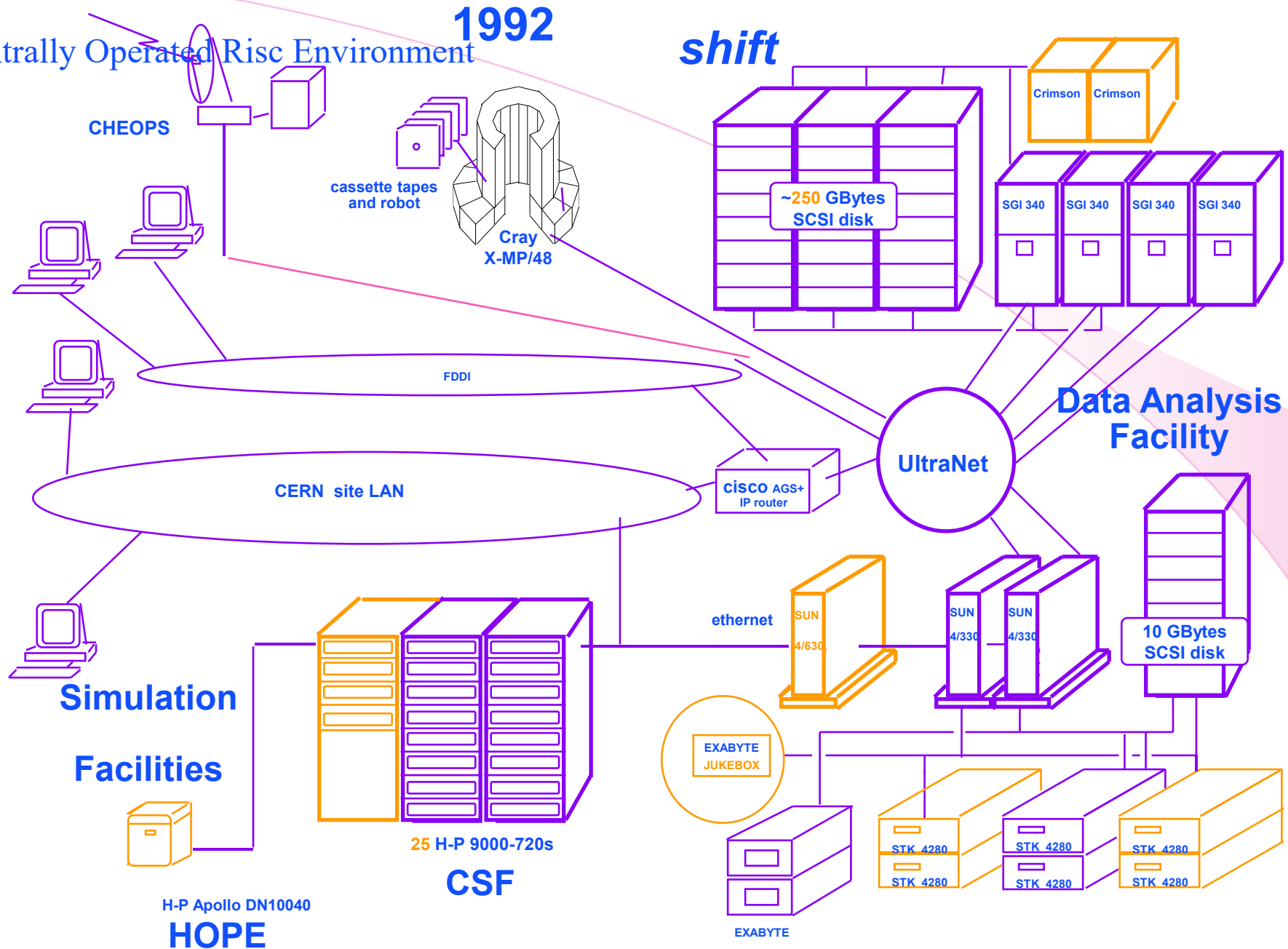
Design choices

- Unix + TCP/IP
- system-wide batch job queues

*“single system image”
target Cray style & service quality*

- pseudo distributed file system
assumes no read/write file sharing
- distributed tape *staging* model
(disk cache of tape files)
 - the tape access primitives are
copy disk file to tape
copy tape file to disk

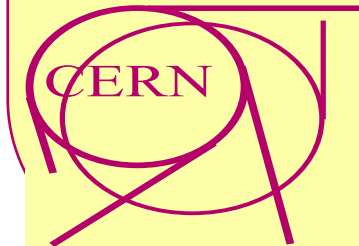
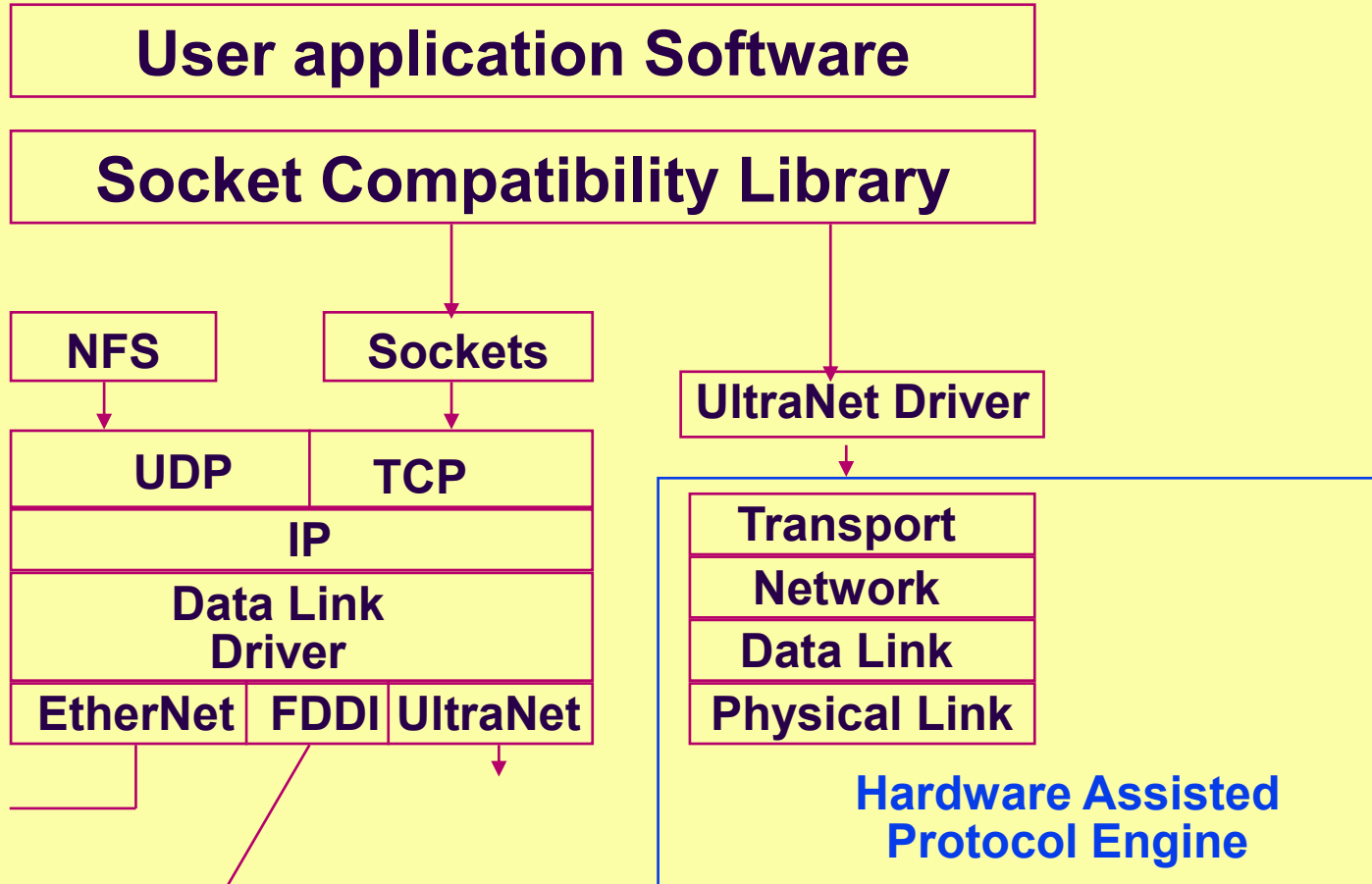
CORE – Centrally Operated Risc Environment





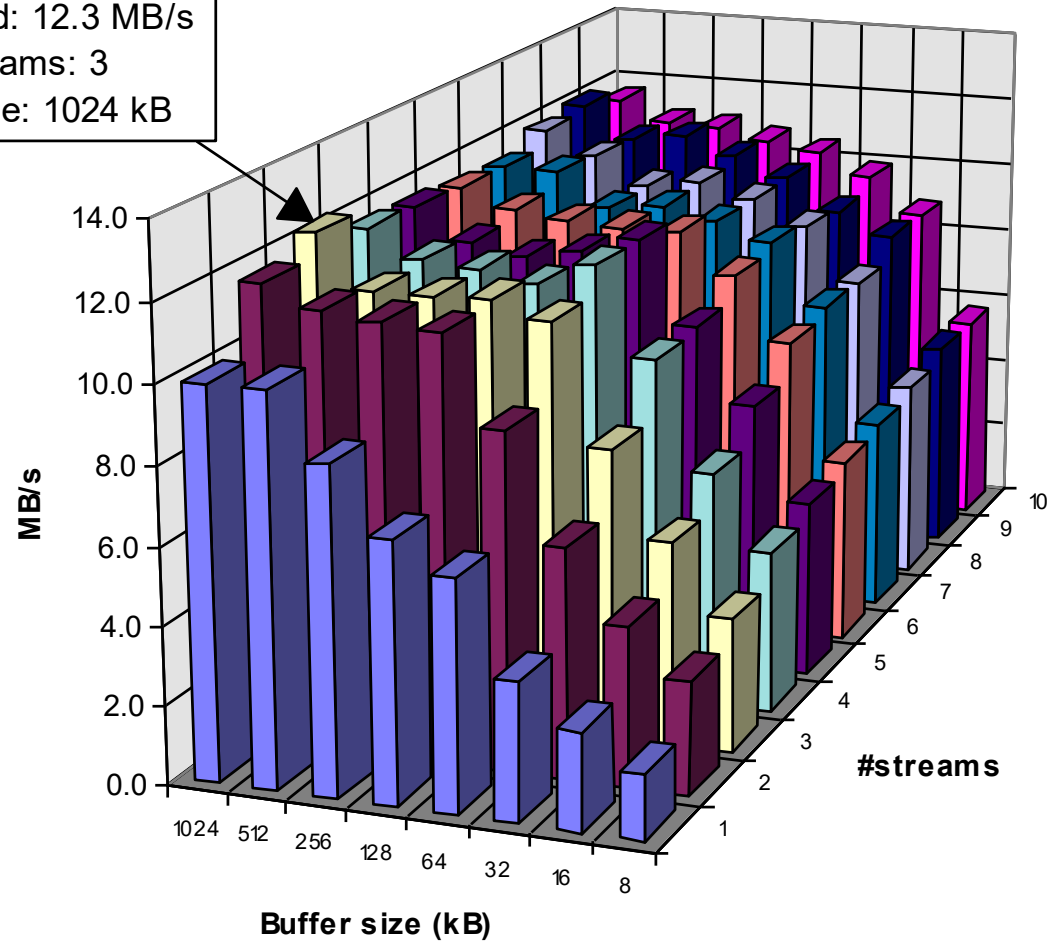
SHIFT in 1992

UltraNet Software



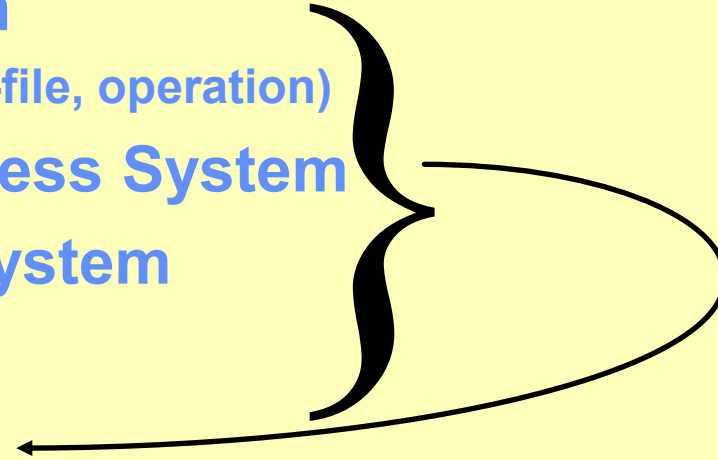
UltraNet transfer speeds pure TCP (SGI to SGI)

512kB socket buffering
Max speed: 12.3 MB/s
#streams: 3
Buffer size: 1024 kB



SHIFT Basic Software

- **Unix Tape Subsystem**
 - (multi-user, labels, multi-file, operation)
- **Fast Remote File Access System**
- **Remote Tape Copy System**
- **Disk Pool Manager**
- **Tape Stager**
- **Clustered NQS batch system**
- **Integration with standard I/O packages**
 - FATMEN, RZ, FZ, EPIO, ..
- **Network Operation**
- **Monitoring**

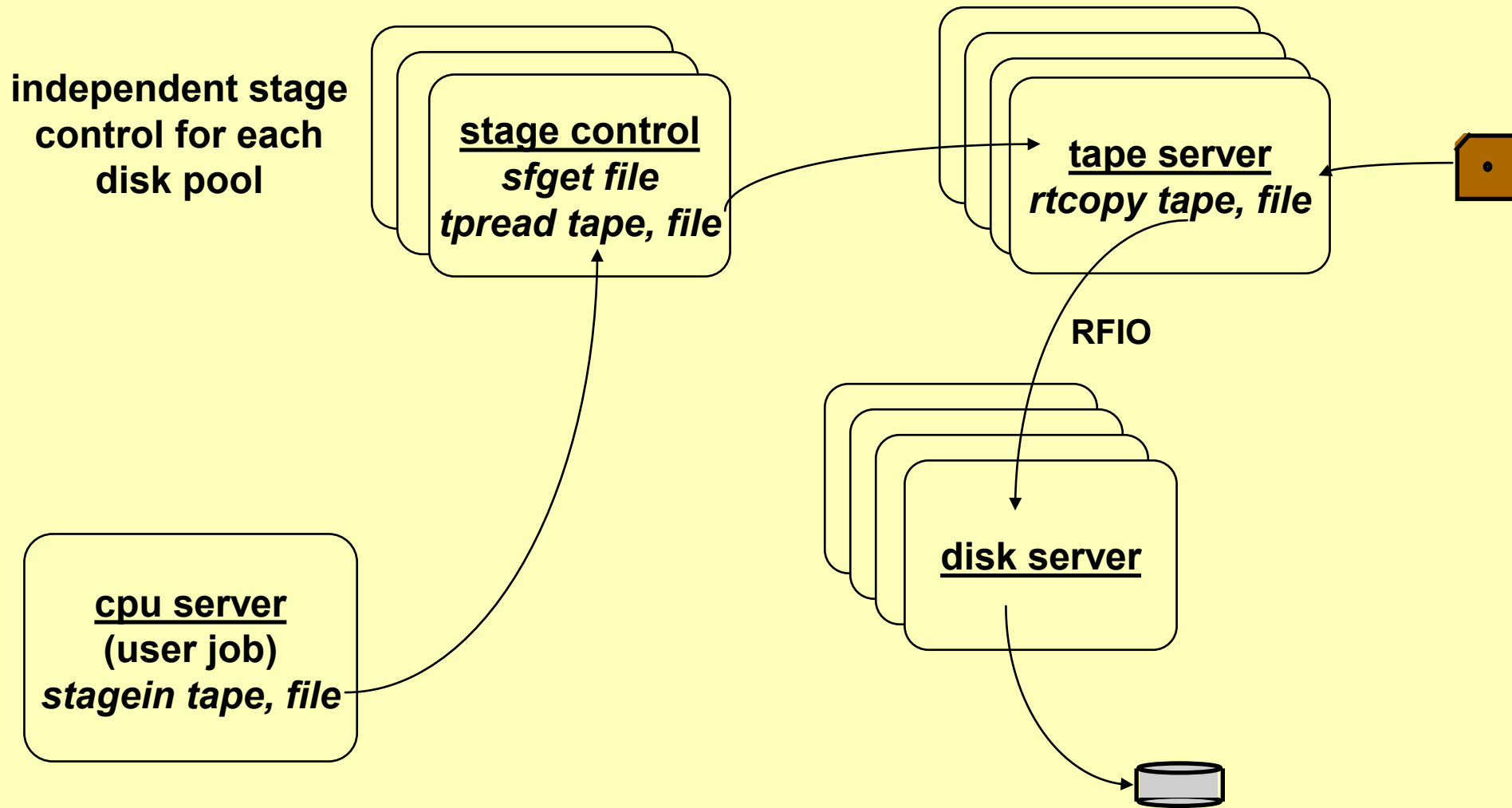


Remote File Access System - RFIO

high performance, reliability (improve on NFS)

- C I/O compatibility library
 - Fortran subroutine interface
- *rfio daemon* started by *open* on remote machine
- optimised for specific networks
- asynchronous operation (read ahead)
- optional vector pre-seek
 - ordered list of the records which will probably be read next


Tape Stager





CORE Physics Services

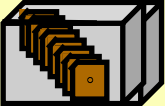
CSF
Simulation Facility



25 H-P 9000-735
H-P 9000-750

Central Data Services


Shared Tape Servers



7 IBM, SUN servers

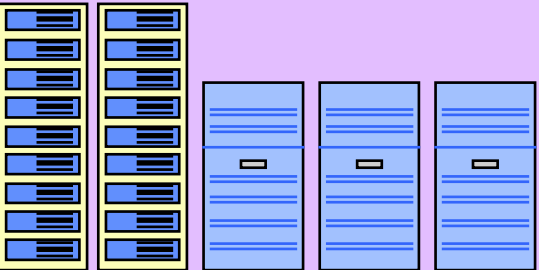
3 tape robots
21 tape drives
6 EXABYTES

Shared Disk Servers




260 GBytes
6 SGI, DEC, IBM servers

SHIFT
Data intensive services




Processors: 24 SGI; 11 DEC Alpha;
9 H-P; 2 SUN; 1 IBM
Embedded disk: 1.1 TeraBytes

PIAF - Interactive Analysis Facility




5 H-P 9000-755
100 GB RAID disk

Home directories & registry




SPARCservers
Baydel RAID disks
tape juke box

consoles & monitors



Scalable Parallel Processors



8 node SPARCcenter
32 node Meiko CS-2
(Early 1994)

CERN Network

SHIFT Status

equipment installed or on order September 1994

configuration

-- capacity --

cpu(CU*) disk(GB)

ALEPH	8-cpu SGI Challenge XL (R4400 - 150MHz) Seven DEC 9000-400	300	210
DELPHI	Three H-P 9000-735/99, two 735/125	149	200
L3	20-cpu SGI Challenge/XL (R4400 - 150MHz)	440	300
OPAL	Two 8-cpu SGI Challenge/XL (R4400 - 150 MHz) Two SGI 340S 4-cpu (R3000 - 33MHz)	378	674
ATLAS	H-P 9000-755/99, 735/125	61	56
CMS	H-P 9000-735/99, two 735/125	95	15
CHORUS	IBM RS/6000-370	17	38
CPLEAR	Two DEC 3000-300AXP, 400AXP	48	10
NA49	Two H-P 9000-735/125	68	30
NOMAD	DEC 3000-500 AXP	19	10
SMC	SUN SPARCserver10, 4/630	22	4
Totals		1597	1547

* CERN-Units:one CU equals approx. 4 SPECints



Central Data Recording

F. Hemmer

CERN - CN/PDP

11.1995

Planning

- The most important
- Plan disk space 6 months in advance
 - » At the pit
 - » In the disk pool
- Plan network bandwidth over years !

Writing tapes

- tpwrite vs. stageout/put
- stageout reserves space
- stagexxx commands do retries
- tpwrite now seen as an internal command, will disappear
- stagewrt can be used to recover or as a replacement to tpwrite
- Use large block sizes (> 8KB or even 32 KB, multiple of 8KB)!
- Don't use small files (> 200 MB)
- Write several files at a time (~10 * 200 MB on a DLT)
- Avoid file positioning by name

□ Optimize transfers

Copying files to disk

- All methods are fine, rfcop preferred
 - Check transfer success
 - Do not delete file at the pit before successfully written onto tape
 - Keep buffer large enough to cover ~ 4 days data taking
- reliable system

Summary

- Plan well in advance (disk space)
- use stageout/put, rfcop
- Optimize transfers (block & file sizes)
- Inform us
- VMS support weak

Disk types & capacity (96)

Vendor	Model	Formatted capacity	Number	Total size (MB)
IBM OEM	0664CSH	3840	1	3840
IBM OEM	0664M1H	1920	18	34560
IBM OEM	0664N1D	1920	230	441600
SGI	0664N1D	1920	2	3840
MICROP	2112	1001	1	1001
HP	C2247	1001	10	10010
DEC	DSP5350S	3406	31	105586
SGI	IBM DFHSS4E	4303	96	413088
DEC	RZ26-VA	1075	14	15050
DEC	RZ26L-VA	1075	3	3225
DEC	RZ28-VA	2150	189	406350
DEC	RZ29-VA	4091	2	8182
DEC	RZ74	3406	88	299728
SEAGATE	ST12400N	2048	52	106496
SEAGATE	ST2383N	317	2	634
SEAGATE	ST410800W	8668	310	2687080
SEAGATE	ST41200N	990	44	43560
SEAGATE	ST41600N	1331	1	1331
SEAGATE	ST41650	1351	1	1351
SEAGATE	ST41650N	1351	162	218862
SEAGATE	ST41651N	1350	6	8100
SEAGATE	ST42100N	1812	12	21744
SEAGATE	ST43400N	2778	61	169458
SEAGATE	ST43401N	2778	36	100008
SEAGATE	ST4766N	669	1	669
DEC	TLZ06	0	1	0
QUANTUM	XP34300W589C	4303	96	413088
Total			1470	5518441



SHIFT in 1997

SHIFT in 1998



1997: start of the PC as a Physics
Data Processing platform

PCSF Status

- ❑ RD47 - W/NT, report available in 2 weeks
- ❑ 10 Dual Ppro @ 200 Mhz (800 CU)
- ❑ Almost ready, LSF 3.0b, SHIFT client
- ❑ ATLAS/CMS code have been ported
- ❑ Physics Production targeted to 12/97
- ❑ Proposal to increase capacity by 3000 CU

First PC Servers

1997



10 Dual Ppro 200 Mhz clients, 96 MB

1998



25 dual PII's @ 300 MHz, 128 MB

First PC Servers

*First PC Servers for Physics
in the 513 Computer Center!
Running Windows NT...*

1997



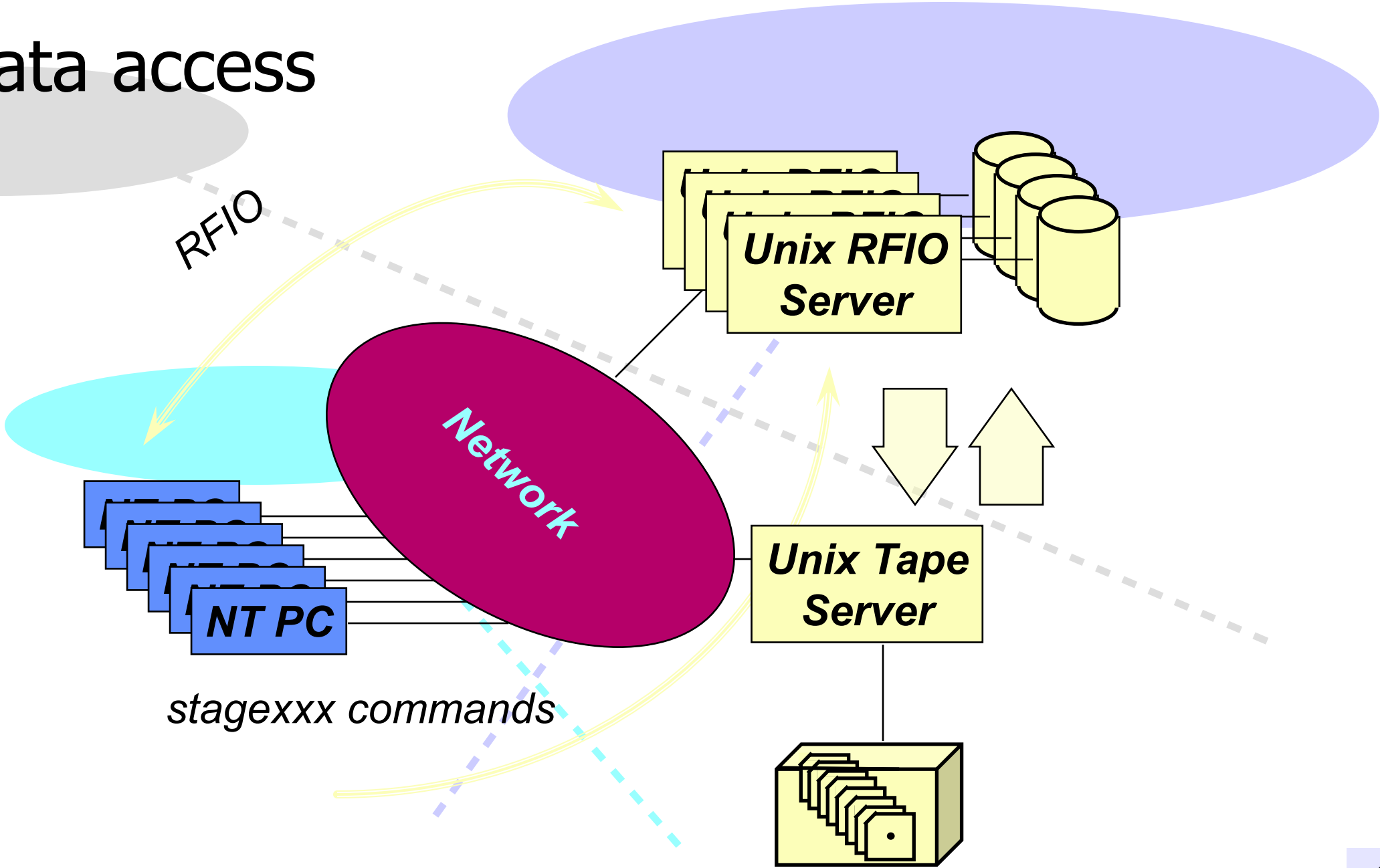
10 Dual Ppro 200 Mhz clients, 96 MB

1998



25 dual PII's @ 300 MHz, 128 MB

Data access



My Computer

Network Neighborhood

Recycle Bin

Connect to

LSF Batch Job Submission

File Edit Help

Batch

Command Line: From File

Queue:

Hosts:

Resources:

Exit

15:42

LSF Batch System

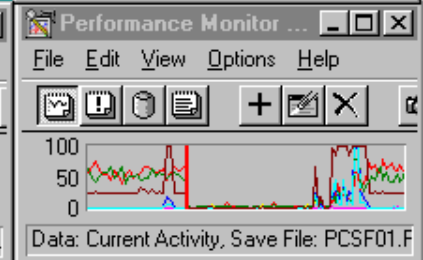
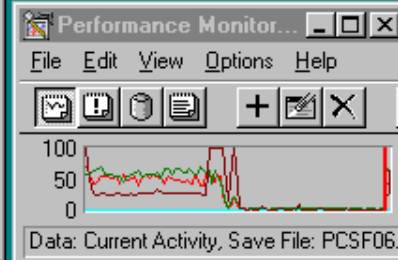
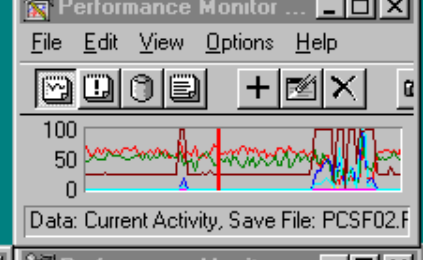
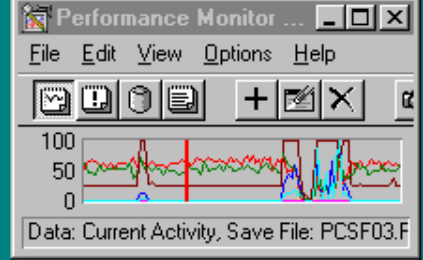
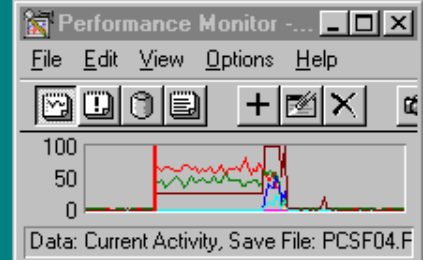
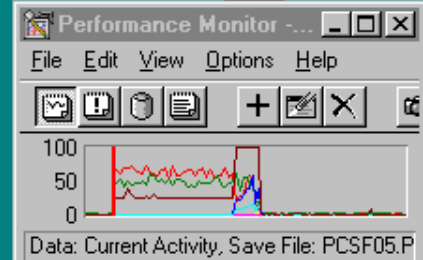
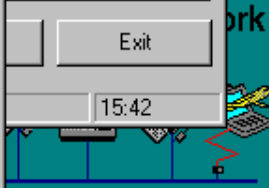
File Job Host Queue User Parameter Options Help

JobID	User	Status	Queue	From_Host	Exec_Host	Submit_Time
14123	hemmer	RUN	priority	pcfam00.ce...	pcsf06.cern...	Fri Nov 28 16:4
14128	hemmer	RUN	priority	pcfam00.ce...	pcsf06.cern...	Fri Nov 28 16:4
14124	hemmer	RUN	priority	pcfam00.ce...	pcsf05.cern...	Fri Nov 28 16:4
14125	hemmer	RUN	priority	pcfam00.ce...	pcsf04.cern...	Fri Nov 28 16:4
14126	hemmer	RUN	priority	pcfam00.ce...	pcsf03.cern...	Fri Nov 28 16:4
14127	hemmer	RUN	priority	pcfam00.ce...	pcsf02.cern...	Fri Nov 28 16:4
14129	hemmer	PEND	priority	pcfam00.ce...		Fri Nov 28 16:4
14130	hemmer	PEND	priority	pcfam00.ce...		Fri Nov 28 16:4

Host_Name	Status	JL/U	MAX	NJobs	RUN	SSUSP
pcsf01.cern.ch	ok	2	3	0	0	0
pcsf02.cern.ch	ok	2	3	1	1	0
pcsf03.cern.ch	ok	2	3	1	1	0
pcsf04.cern.ch	ok	2	3	1	1	0
pcsf05.cern.ch	ok	2	3	1	1	0
pcsf06.cern.ch	ok	2	3	2	2	0
pcsf901.cern...	closed	1	1	0	0	0

Queue_Name	Priority	Status	MAX	JL/U	J
1nd	30	Open:Active	1	1	
1nh	50	Open:Active	6	6	
1nw	20	Open:Active	1	1	
8nh	40	Open:Active	1	1	
8nm	60	Open:Active	12	12	
priority	90	Open:Active	6	6	
...	10	Open:Active	11	11	

Updating done. Updated 16:42:29



Taskbar with Start button, application icons (Excel, Telnet, Re: p..., PC K..., Perfo...), and system tray (LSF, WPc..., LSF..., Micro..., 16:43).

Nomad PC Status

- 1 Dual PII @ 266 Mhz - Linux
- All running fine
- 3 Other Dual PII ordered
- Proposal to add 700 CU in 98

Nomad PC Status

- ❑ 1 Dual PII @ 266 Mhz - **Linux**
- ❑ All running fine
- ❑ 3 Other Dual PII ordered
- ❑ Proposal to add 700 CU in 98

*First Linux machines in the 513
Computer Center!*

Lessons (1) From the 1996 talk

- Things change fast
 - CPU increase in speed
 - disk explosion
 - tape technology change
- Network is the core
 - But complicated
 - even more with high speed networking

Lessons (2) From the 1996 talk

- More equipment need more effort
 - but manpower shrinks
 - we cannot reduce diversity
 - outsourcing may be the answer
- We are probably too distributed
 - Subgroups that do well only have a few machines closely interconnected

Lessons (3) From the 1996 talk

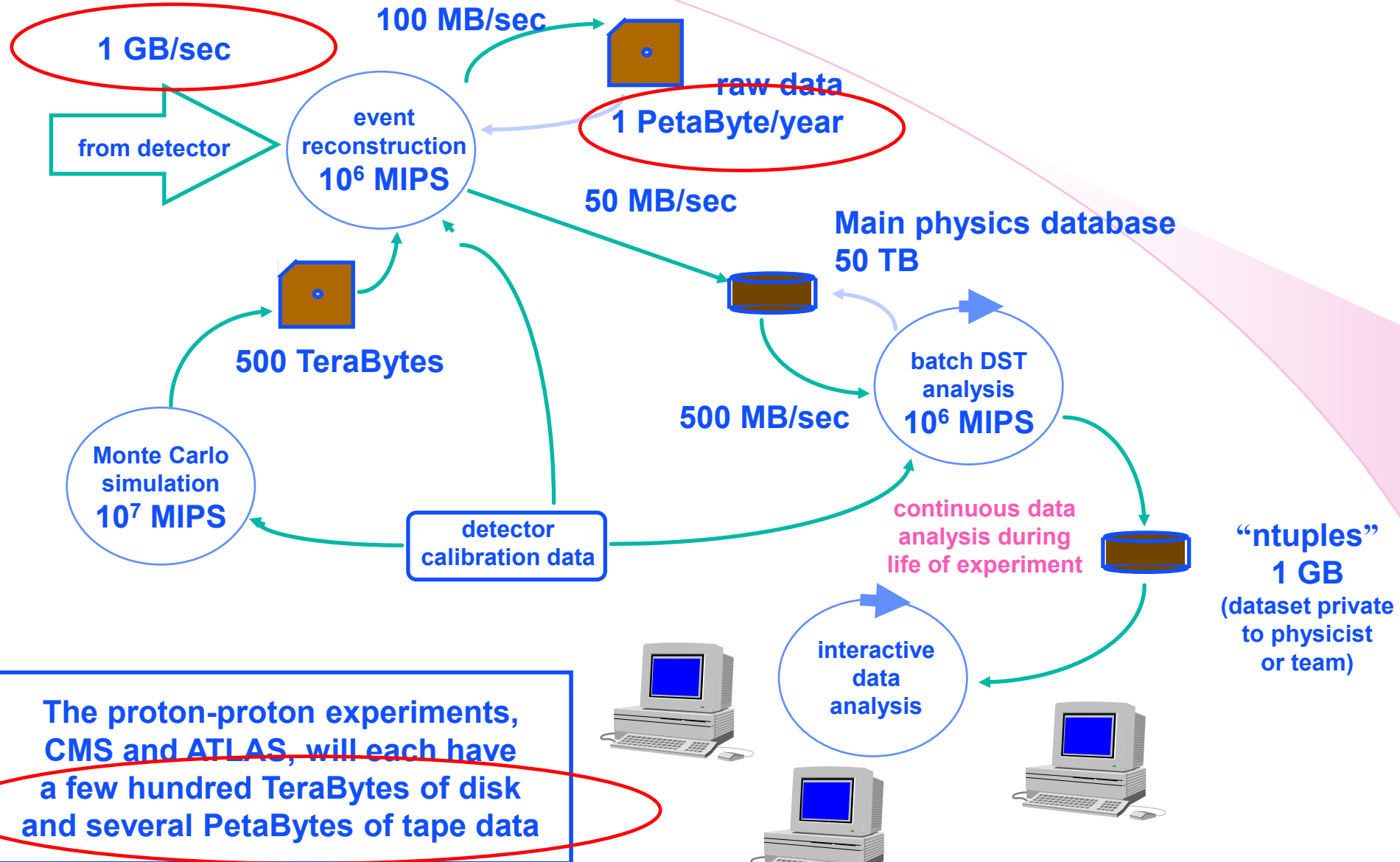
- Although the economics is paying, we start to pay the price ...
 - system management is the real issue now !
 - ... try to centralized it, automate, clone ...
 - there is a lack of multivendor solution (or too expensive)

Future ...

From the 1996 talk

- We hope outsourcing is an answer
- Again a technology change ?
 - PC's
- Major challenges arrive ... LHC

LHC pp Experiment Data Flow - 200x



Final personal considerations

- We have not created new technologies, but rather acted as integrators of
 - New technologies coming (and going) from the computing market
 - De facto standards such as Unix and TCP
 - Improvements and exploitation of new networking solutions
 - Driven by the needs of Physics
- Inspired by guidance of people such as Les Robertson and Ben Segal
- Benefitting from budgets from the IT department (at that time CN for Computing and Networks) and supervised by people who knew what delivering computing services implied!
- Constantly pushed by the insatiable appetite of the experimental physicist for data (and hence for cycles to process the data). It is essential to work *with* them!
- Many lessons and conclusions from that period are still valid today, the numbers have just increased by at least 3 orders of magnitude (but not the people nor the budgets have increased that much...)

Some References

- [Ian Willers, Parallel Computing: Some activities in High Energy Physics, CERN/ECP 90-4, August 1990](#)
- [J-P. Baud et al., «Mainframe Services from Gigabit-Networked Workstation”, Summer ‘92 Usenix, June 1992](#)
- [S. Jarp et al., “PC as Physics Computer for LHC ?, CN/95/14, September 1995](#)
- [Ben Segal, «A major SHIFT in outlook», CERN Courier, July 2001](#)
- [F. Hemmer & P.G. Innocenti \(ed.\), Technology meets research : 60 years of CERN technology : selected highlights: Data Handling and Communications, 2017](#)