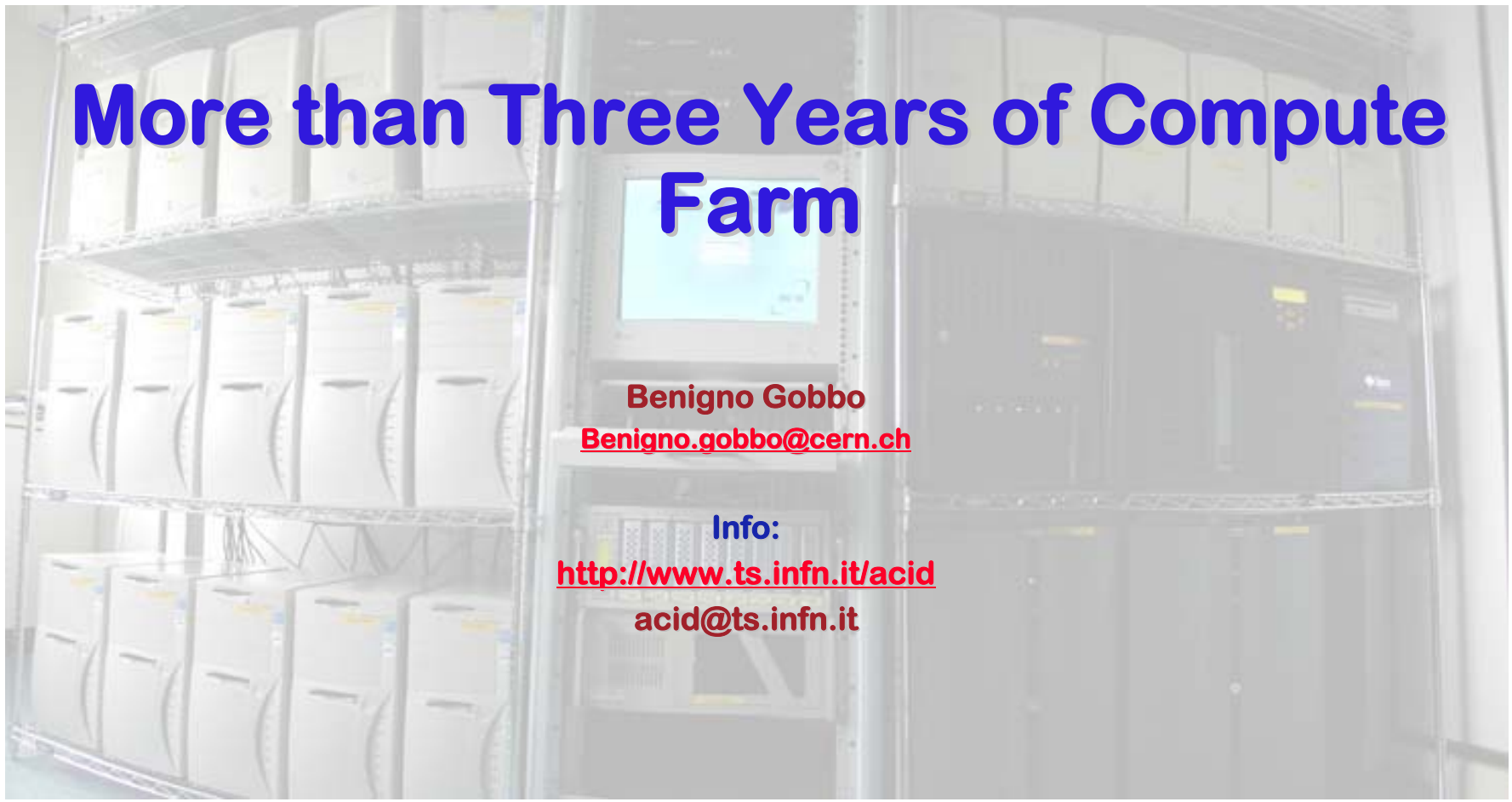


More than Three Years of Compute Farm

Benigno Gobbo

Benigno.gobbo@cern.ch

Info:

http://www.ts.infn.it/acid

acid@ts.infn.it

Requirements

COMPASS: High statistics - Medium event complexity

- ~ 10^{10} events/year
- ~ 10 "good" tracks/event
  - More than 200 tracking planes in non uniform magnetic field
  - Particle Identification: RICH, calorimeters, …
- Non trivial event reconstruction
  - Production time: ~0.5 s/ 1 GHz PIII CPU

DATA STORAGE, PRODUCTION and ANALYSIS model

- Raw data stored at CERN (~300 TB/year)
- Production at CERN: up to 400 reserved batch queues (↔CPUs)
- Monte Carlo Production and Data Analysis at Home-Labs

Need of Compute Farms at Home Laboratories

- Also due to usual CERN request of computing redistribution:
  - 33% at CERN, 67% outside

A different Computing Model

1998. Definition of a Computing Model for the post–LEP era

- January 1998. A Task Force was established at CERN (1)
  - To achieve: agreement with time scale and requirements of experiments, flexibility of environment, constraints from used commercial software, realistic assessment of costs, …
- April 1998. Conclusions (Recommendations): Hybrid Architecture
  - using PCs for computation (preferred: Windows NT, “tolerated”: Linux)
  - using at present RISC systems for I/O (legacy Unix)

1999. Evolution of the model

- Sensitive Linux improvements: now stable and better performing than Win NT
- Development of “low price + good enough quality” IDE disk based PC servers

COMPASS Definitive choice:

- PCs for both server and computation machines
- (RedHat) Linux OS

The Farm History

Sep. 2000. Approved (and above all "sponsored"!) by CSN I

- Financed in two years
  - 200M ITL in 2000
  - 124k € in 2001

Oct. 2000. Definition of a schema for the farm "initial setup"

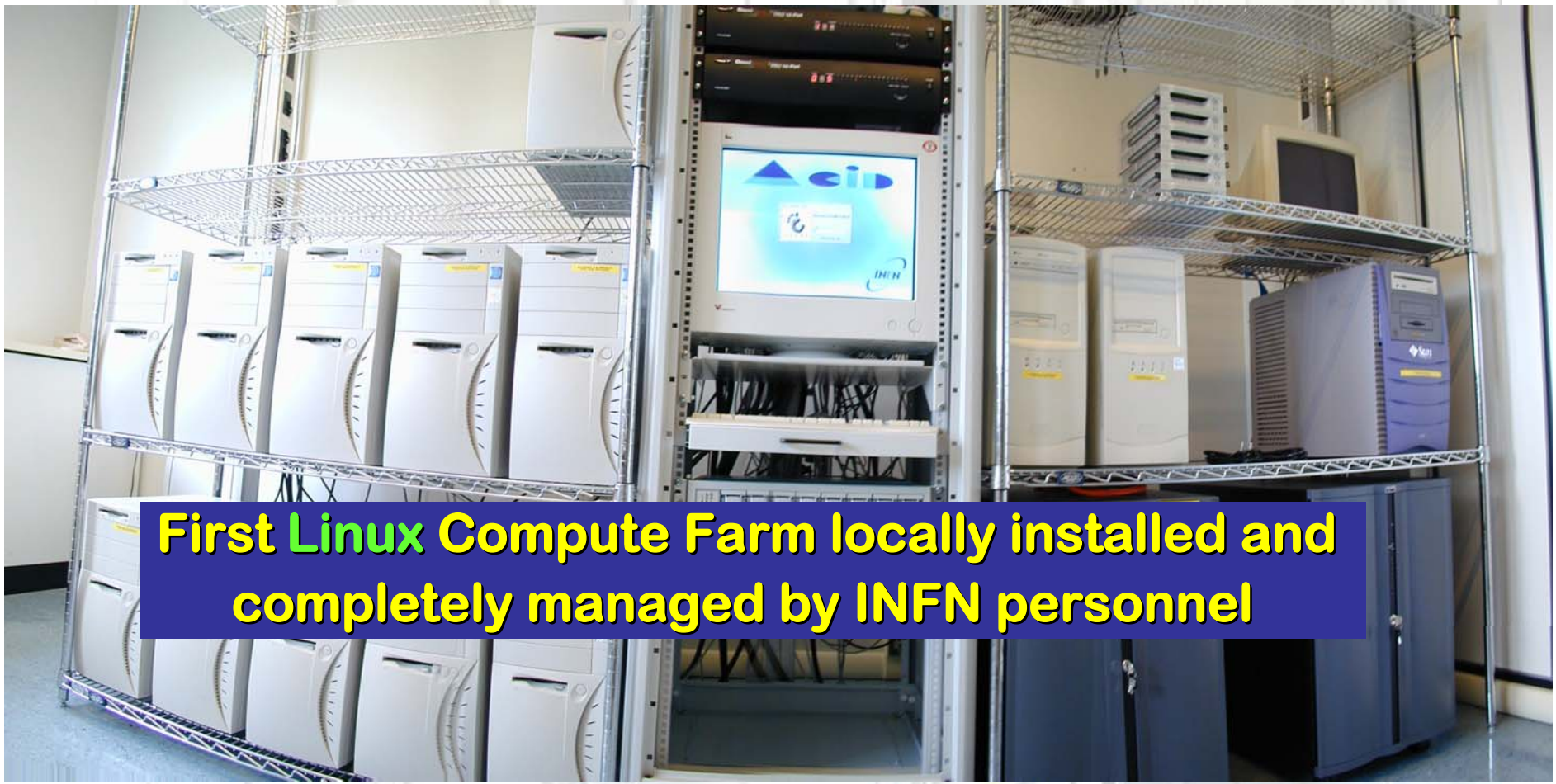
- The farm has to be as much as possible compatible with the CERN one
  - But not CERN-dependent
- The "initial setup" must guarantee a "production environment"
  - Enough disk space (for data storage and MC production)
  - Enough CPU power (i.e. PC clients)
- It must be scalable to the final configuration without (major) modifications
- It must fit with approved financing

History: first steps

Nov. 2000. “Initial setup” decided, orders submitted

- 1 PC Server with large EIDE disk space (with 14 x 75 GB EIDE disks)
  - RAID1 (mirroring) configured, it allowed 0.5 TB of (cheap) disk storage
  - The machine was assembled by ELONEX following a CERN R&D
- 1 Sun Server with external SCSI disks (8 x 73 GB)
  - Configured RAID5, gave a 0.47 TB of more reliable disk storage
  - Different OS (Solaris) and architecture (SPARC): allows better test and debugging of software
- 1 PC Supervision Server
  - Nothing special: just a white-box PC with better components. Used as a supervisor or master in monitoring or client-server software
- 12 PC Clients
  - Value white-box PC, to stay into available budget
- All machines are dual processor to improve performances/costs
  - Well… Sun was bought as single processor (it was so expansive…) and upgraded subsequently
- Network switch (36 100BaseT + 3 1000BaseSX ports)
- KVM switches, rack, shelves, monitor, keyboard, etc.
- UPS and cooling system (thanks to A. Mansutti & S. Rizzarelli)

History. Feb. 2001: “First setup” in production

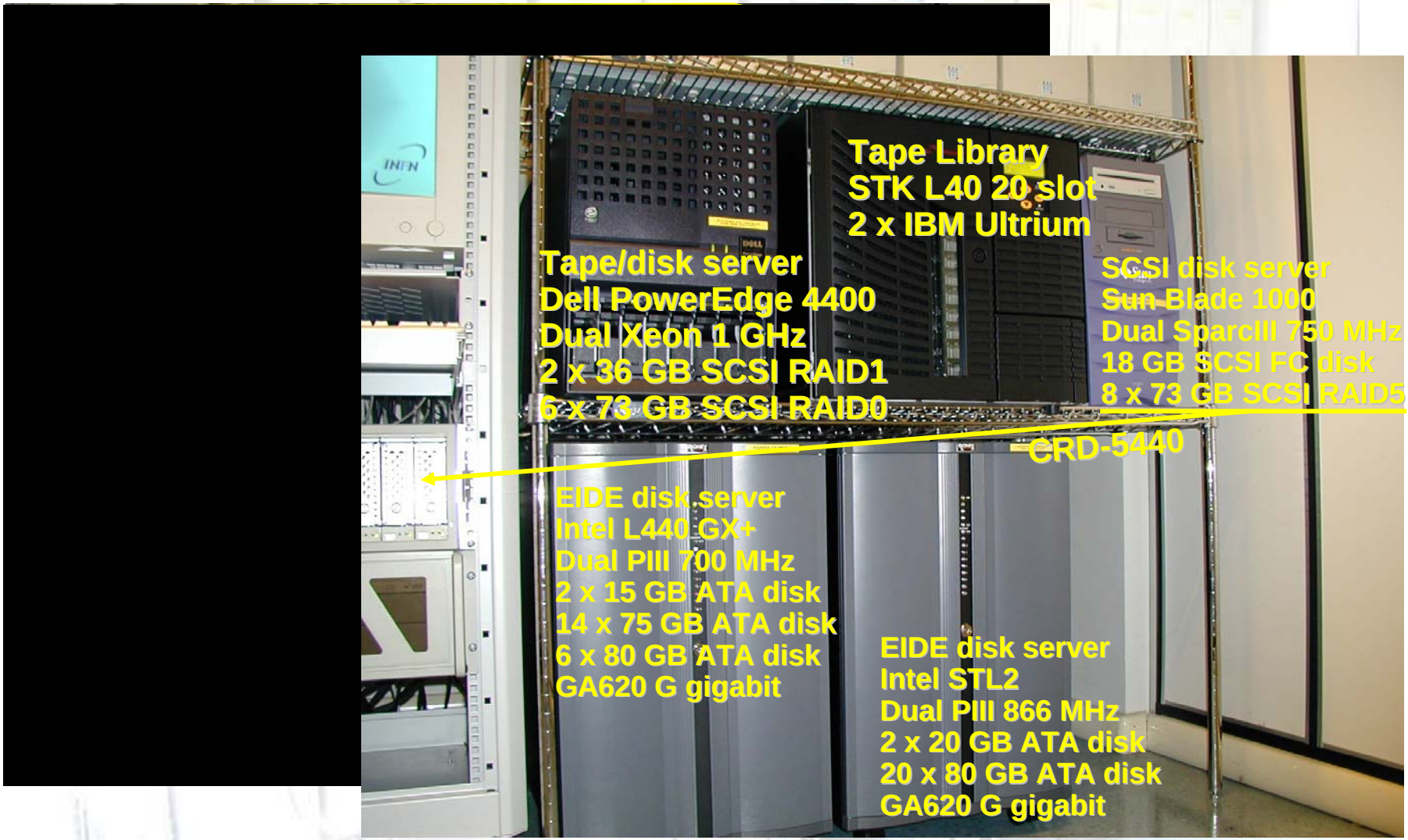


History: the final setup

Sep. 2001. Start Farm upgrade to Final Setup

- 1 more EIDE PC Server (with 20 x 80 GB EIDE disks)
  - Configured RAID1: 0.75 GB
- Upgrade of previous EIDE Server with 6 additional 80 GB disks
  - Now it provides 0.72 TB (RAID1)
- Upgrade of the Sun to dual processor
- STK Tape Library: 20 slots (can be upgraded to 40) , 2 IBM Ultrium drives (can have 4 drives)
  - It can store up to 4 TB of data. Drives transfer rate up to 30 MB/s
- 1 Dell PC Tape Server, with 6 x 73 GB SCSI disks configured RAID 0 (striping)
  - To be used with Tape Lib forming HSM system
- 19 PC clients
  - white-box machines, dual 1 GHz P III
- 12 ports 1000BaseSX switch
- KVM switches, etc.

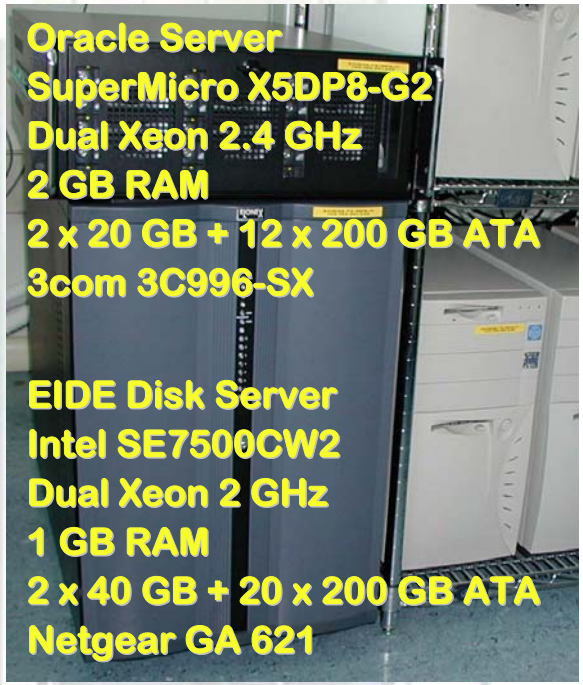
History: the 2002 “Final Setup”



History: up to now and in the near future

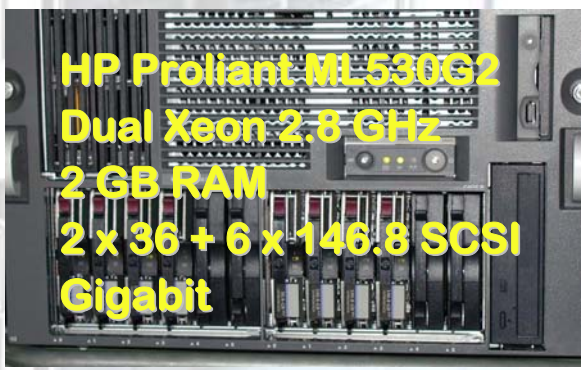
2002 - 2003. Upgrades

- Additional EIDE PC Server with 20 x 200 GB disks
  - Powerful machine (Dual Xeon). 4 RAID5 partitions allowing 3 TB of disk space
- PC server for Oracle/DB with 12 x 200 GB disks
  - To contain event database
- HP PC Server with 6 x 142 GB SCSI disks
- STK Tape Library upgrade from 20 to 40 slots
  - Now allows to store up to 8 TB of data



2004. Financed

- Ultrium2 Tape Drive for STK Tape Library
  - Up to 400 GB/cartridge, up to 70 MB/s transfer rate
- ~10 PC Clients
  - Rack mount Dual Xeon processor machines



ACID Farm w.r.t. CERN farm: Hardware

The choices (1) (2)

- **Clients. No alternatives due to cost difference: use PCs. But…**
  - At CERN there are short hardware upgrade periods → use "old", good quality (e.g. Intel chipsets), well Linux tested (certified) hardware
  - Here hardware lifetime is longer → use "recent" hardware (as it becomes "dated" really fastly), middle quality (e.g. VIA chipset, for cost reasons), may be not yet completely Linux certified
- **EIDE disk server shows a great performance/cost ratio**
  - Not completely tested at beginning, but looked nice and the difference in cost with SCSI based servers (a factor three) looked too attractive
- **The Sun**
  - Also at CERN the is a SUNDEV cluster made available for code quality checking. In addition, there are some services still run on Suns for stability or commercial software requirement reasons

ACID Farm w.r.t. CERN Farm: Software

Requirements and solutions (1) (2)

- **Compatible as much as possible**
  - Programs should run without recompilation → Use same kernel and compilers
  - Users should find similar environment → Use same Linux distribution
  - Use CERN patches if they help
- **Independent as much as possible**
  - Do not use too-CERN-specific tools like SUE (hard to port, not so useful)
  - Use official distributions (RedHat) and not CERN "adapted" ones
  - Do not use CERN patches if they do not help
  - Use INFN-Trieste (e.g. LinuxUpdate [L.Strizzolo, T.Macorini] , local CUPS implementation [L.Strizzolo]) or INFN solutions whenever available
- **Chose something else if nothing available or simply if there is something better around:**
  - CERN batch solution too expensive (LSF), nothing interesting at INFN level → use SGE: free, good, supported
  - Monitoring: BigBrother is fee and looks nice (1) (2) (3) (4)
  - Software documenting too: found Doxygen, it is so good that it was subsequently adopted by CERN

ACID w.r.t. CERN Farm: Commercial Software

We try to avoid it, if possible (it costs and it is source of troubles)

- **CERN attempt to go for "commercial-only software" dramatically failed!**
  - In general: too difficult to interface to HEP environment
  - In general: it never completely fits with HEP requirements
  - In general: not able to follow the fast Linux and GNU software evolution (e.g. compiler: we are forced to use quite outdated and now unsupported gcc compilers. Objectivity/DB needed gcc 2.95.2, ORACLE needs gcc 2.95.3 or 2.96; current gcc version is 3.3)
  - Expansive or whit unsatisfactory support (and, in any case, no source code available: so no way to fix problems by ourselves)
- **So, the current idea is to use commercial software only where there are not alternatives**
  - Basically only DBMS (Objectivity/DB 6 before, ORACLE 9i after): too difficult to develop an HEP specific DBMS. Well, free DBMS are available too (e.g. MySQL), but it is too dangerous to follow a solution different with the CERN one on this subject…

ACID w.r.t. CERN Farm: HEP Linux, what is going on

Recent (~2003) RedHat change of philosophy

- Free distribution "Fedora Project"
  - Free distribution with a release period of 4-6 month (too fast for HEP needs) and just 3 months support/patching of previous release (too short for HEP needs)
- Commercial distribution "Enterprise"
  - Commercial distribution with 5 years support of previous release but too expensive!

HEP Reactions

- Mandate to the 3 HEP big labs to negotiate with RedHat, but at the end…
- FNAL
  - Rebuild RHEL from source (legal if done without violating RedHat copyrights!) LTS 3.0.1 (now available also cleared from FNAL specifics and renamed HEPL). FNAL would like to collaborate with other HEP labs
- SLAC
  - Negotiated with RedHat "via" DOE. For one year RHEL will be used. And after, who knows?
- CERN
  - As FNAL (CEL3 rebuild) as main line. But some RHEL3-WS (~200) is being bought. CEL3 is now under certification (to be finalized by 2Q2004 or so).

ACID w.r.t. CERN Farm: software, what will change

Keep CERN compatibility. Will it be easier? Expensive?

- **Good**
  - CERN port will be less specific (no more SUE, etc.)
  - No more “alternative gcc” compilers (if possible)
  - But with additional “wanted” packages (PINE, …) no more available from RedHat distribution to avoid license violations.
  - ACID could probably use CERN distribution without major problems (to be checked) instead of use RedHat distribution plus add-ons.
- **And Bad**
  - The port will be supported for 1-2 years. And after?
  - The RHEL option still present. That could mean extra costs for software (now we use RHEL (AS2.1) just on the ORACLE server machine). In that case an I.N.F.N. wide license solution would be a better solution. Or we could try to user FANL HEPL. We will see…

Farm management: man power costs (SW)

Distribution Upgrade

- It is a major task as a local certification is needed too
  - All applications need to be tested
  - All nodes need to be re-installed from scratch
  - In general it requires more than a month preparation time
  - Not too frequent: one every few years (~2)

Software Installation

- Complexity and test-debug period depend on package
  - Could be a strong work (e.g. CASTOR/HSM porting: many months of work)
  - Time-to-time, upgrades/updates are needed

Patching

- In general simple but quite frequent (security patches)
  - Could need a lot of time (e.g. as we use a locally patched kernel, we need a complete kernel recompilation after every official patch)
  - And the risk of troubles after a patch is not negligible: in particular after Kernel updates

Farm management: man power costs (HW)

New hardware

- **Purchase**
  - Product choice, offers requests, “CONSIP”, …Very time consuming and generally boring
- **Installation and/or integration**
  - In general non complex, but in some cases needs time

Maintenance

- **Many parts of the farm are no more covered by warranty nor under outsourced maintenance**
  - Broken parts (disks, boards, …) need to be replaced by hand. That takes a lot of time (1)
  - An Example:

MicroStar 694D Pro mainboards mount bad quality electrolytic capacitors (from TAYEH). Over 11 boards, on 7 there were failures due to that capacitors leakage. Intervention requires a complete PC dismount, board removal, capacitor replacement and re-mount. On two boards capacitor failure damaged following electronics: in those cases mainboard replacement where necessary.

- **Power loss (HW failures were many times due to overheating).**
  - Quite (better: too) frequent in AREA. No cooling for long periods with consequent machines overheating (In addition, as I always said, that T02 room is definitively too small compared to the hardware installed inside, this will fortunately change soon).

The good and the bad

- As said: the first Linux Compute Farm installed and managed in an INFN Lab
- First COMPASS home-lab farm in production
- One of the first CASTOR/HSM installation outside CERN
  - and probably the first one in production
- First "in production" ORACLE database replica of part of (COMPASS) events outside CERN
- Heavily used by COMPASS-Trieste group
  - Data analysis, Monte Carlo production, RICH software development and analysis, …
- "Borrowed" for other Trieste groups works (LEP, …)

- It is an "in production" apparatus
  - Interventions have to be immediate, quick (& NOT dirt)
  - It requires a continuous monitoring: i.e. someone always has to be present "nearby T02"
  - It always "evolve" (software updates, hardware upgrades) and that requires manpower
  - It is fragile: the probability of failures is high
  - Parts of software need to be updated and checked very frequently (even every day or so)
  - It is difficult to have a day without need of interventions somewhere inside the farm

What next

A new project: the *"Farm di Sezione"*

- To (try to) merge all local farms in a kind of unique entity.
  - It is again something relatively new inside INFN sites
- It involves *Gruppo Calcolo* and several experiments people from existing farms (ALICE and COMPASS) and new ones
- Discussion started: to find common requirements and evaluate incompatibilities
- Place was found: T02 →T01+T02
- Cooling is being powered
- Some hardware was already acquired
- R&D will start soon (compatibility tests between different present farms environments, etc.)
- Consequences on the ACIDs: too early to say anything, we will see…

Acknowledges and Conclusions

Thanks to

- R. Birsa
  - Sun Management
  - Help in software installation and debugging (e.g. CASTOR would never be installed without his accurate work on it)
- V. Duic
  - Data (DB) import, job parallelization tools
- All people of *Gruppo Calcolo*
  - Offer requests
  - Consultancy
  - "Linux Update"

To conclude

- This farm shows that at INFN-Trieste there is a not negligible IT knowledge (compared to other INFN sites)
- Computing is becoming more and more relevant in HEP experiments. It will probably be dominant (in good and bad) at LHC
- Unfortunately INFN looks NOT so pioneering on that field…