

GNU/Linux (e il software
libero) nella fisica delle
particelle elementari

Roberto Ferrari

Parma GLUG

GNU/Linux Day 2011

22 ottobre 2011

ACRONIMI:

INFN:

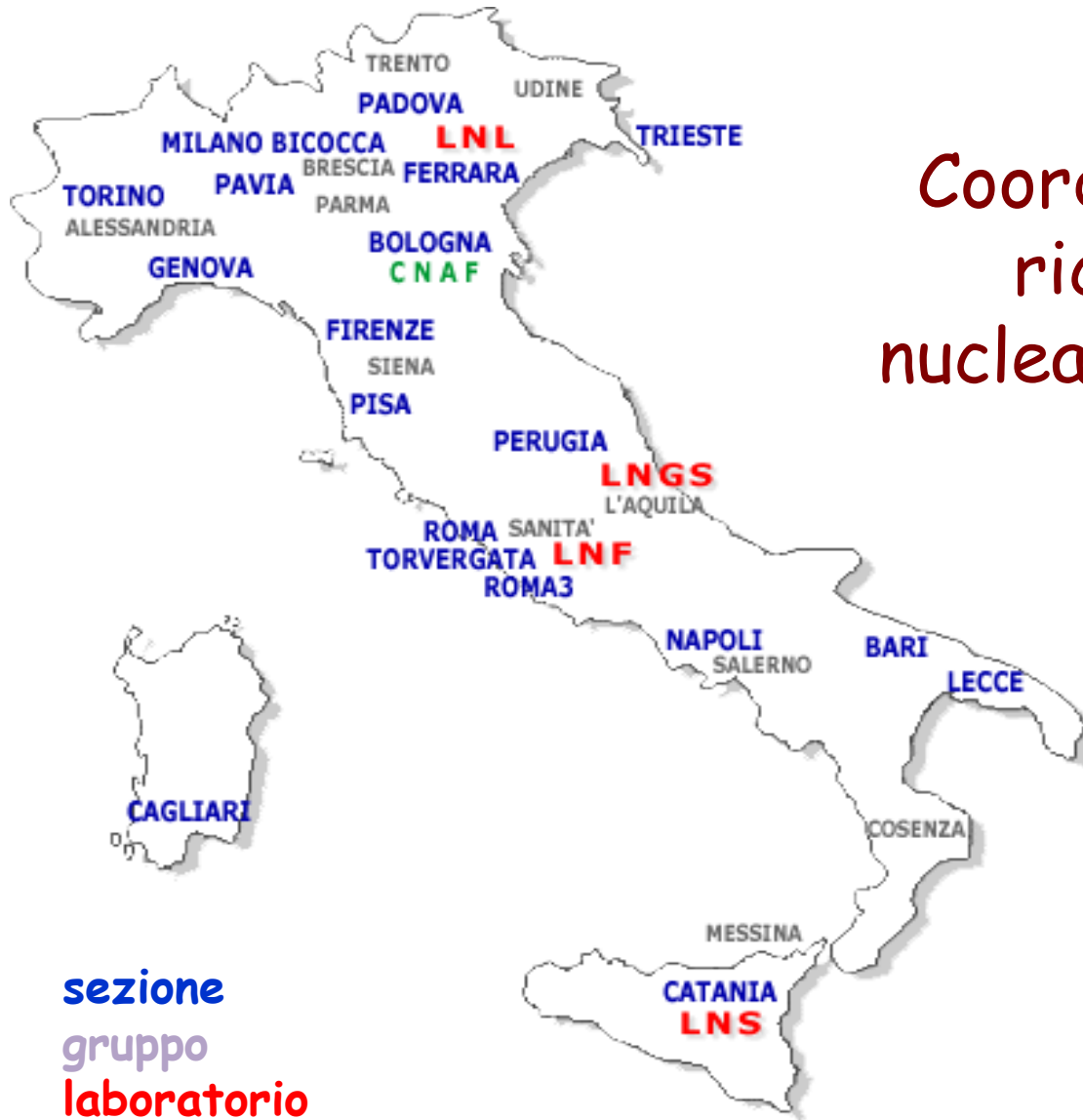
Istituto Nazionale di Fisica Nucleare

CERN:

European Organization for Nuclear Research

I'INFN

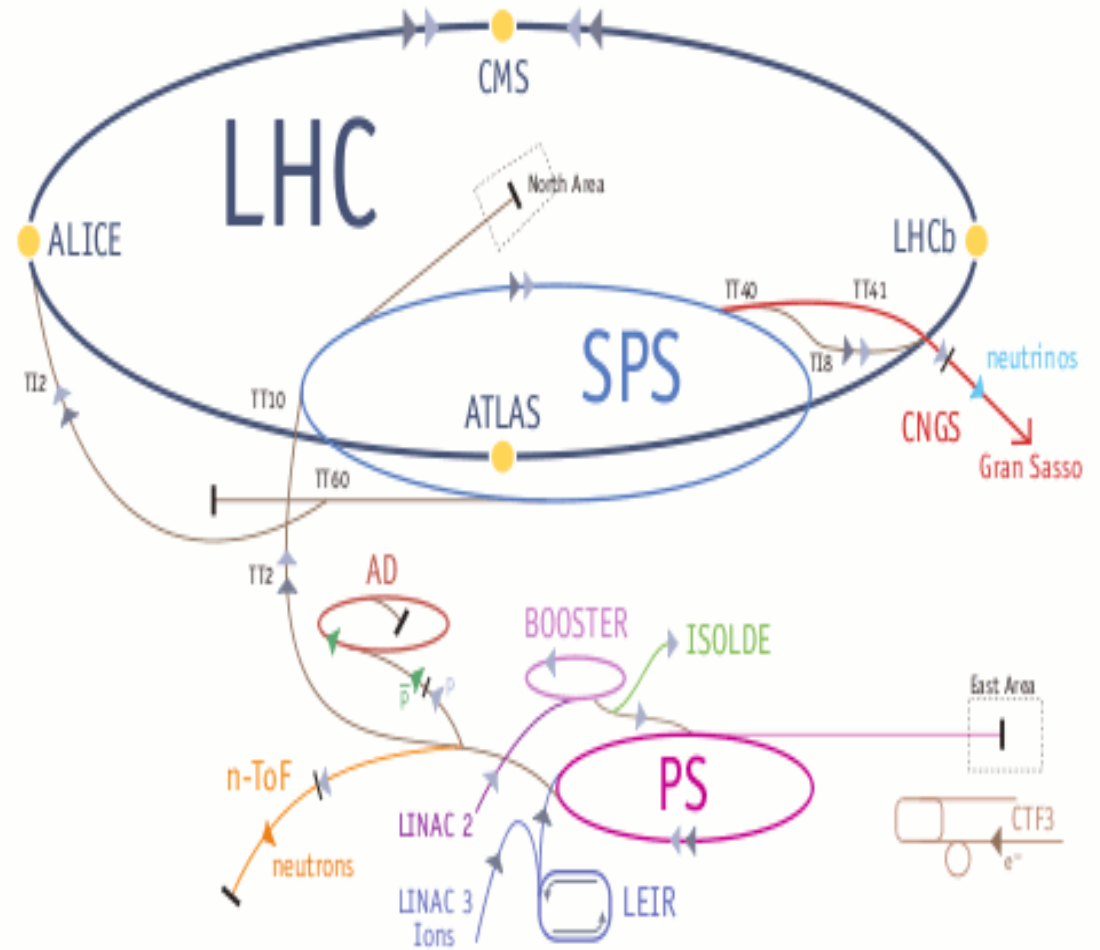
Coordina e finanzia la
ricerca in fisica
nucleare e sub-nucleare
in Italia



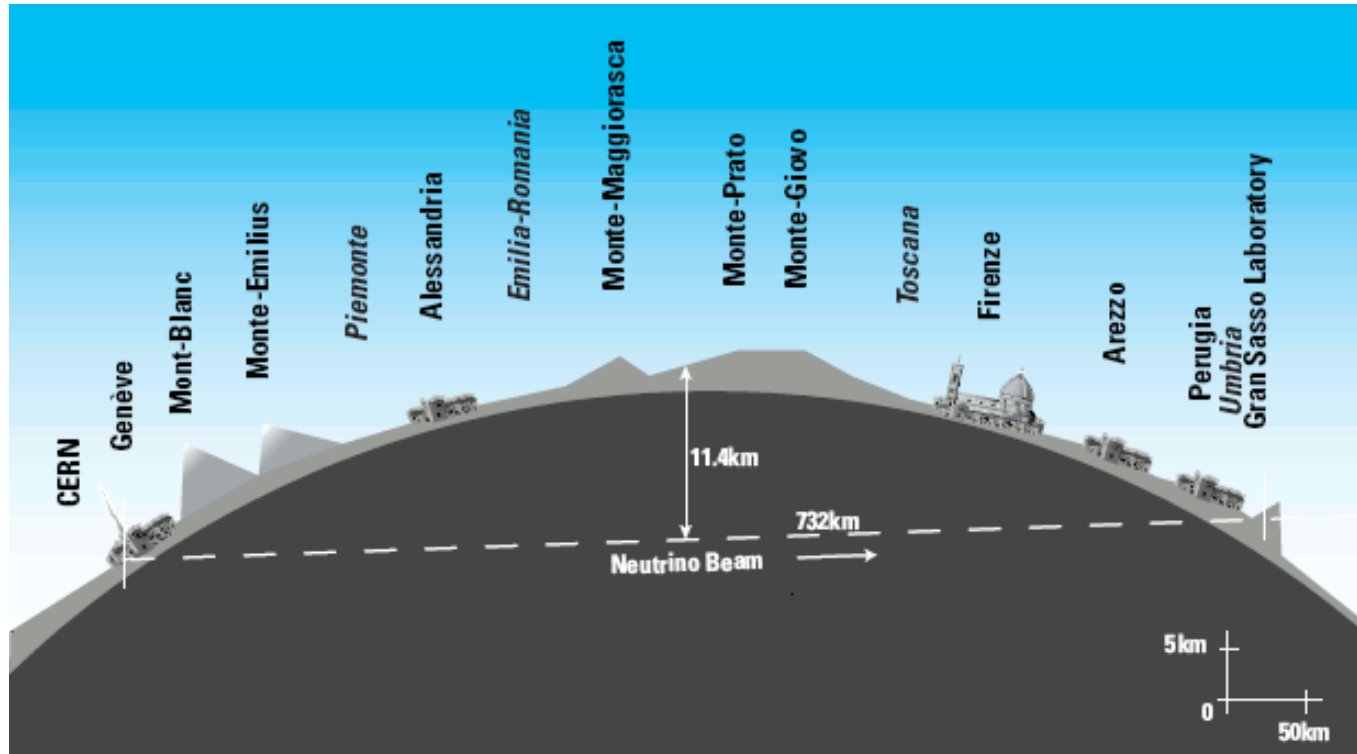
sezione
gruppo
laboratorio

il CERN

Laboratorio europeo
per la fisica delle
particelle

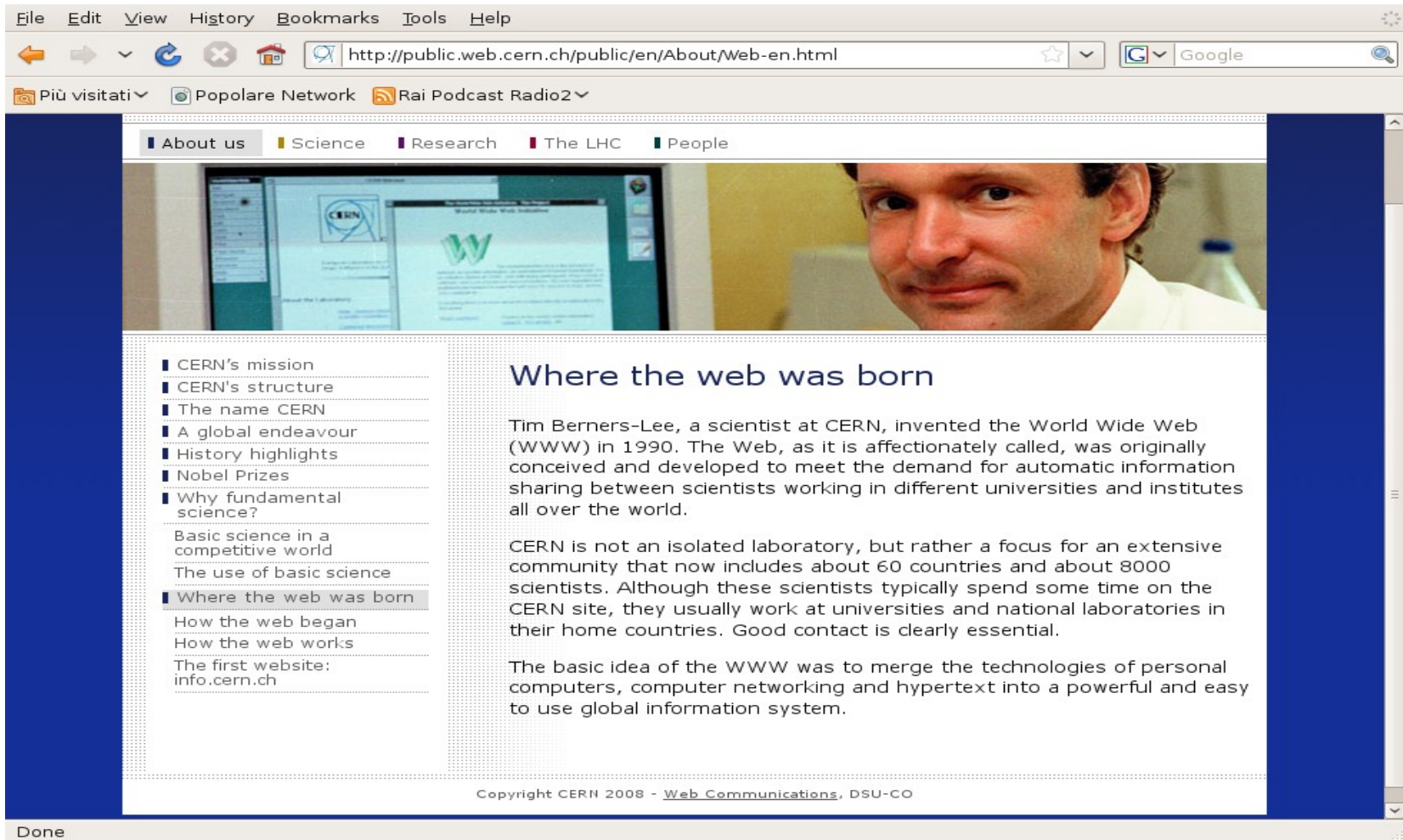


dal CERN al Gran Sasso (INFN)



punto medio ~ monte Maggiorasca (confine
Parma-Piacenza-Genova) vicino a Bardi

dove è nato il Web ?



The screenshot shows a web browser window with the address bar displaying <http://public.web.cern.ch/public/en/About/Web-en.html>. The browser's menu bar includes File, Edit, View, History, Bookmarks, Tools, and Help. The address bar also shows a search engine dropdown set to Google. Below the address bar, there are navigation icons and a search bar. The main content area features a navigation menu with links for About us, Science, Research, The LHC, and People. A large image shows a man in a white shirt looking at a computer monitor displaying the CERN website. Below the image, there is a sidebar with a list of links: CERN's mission, CERN's structure, The name CERN, A global endeavour, History highlights, Nobel Prizes, Why fundamental science?, Basic science in a competitive world, The use of basic science, Where the web was born (highlighted), How the web began, How the web works, and The first website: info.cern.ch. The main content area has the heading 'Where the web was born' and two paragraphs of text. The first paragraph describes Tim Berners-Lee's invention of the World Wide Web in 1990. The second paragraph describes CERN as a focus for an extensive community of scientists. The third paragraph explains the basic idea of the WWW as a merger of technologies. At the bottom of the page, there is a copyright notice: Copyright CERN 2008 - Web Communications, DSU-CO.

File Edit View History Bookmarks Tools Help

http://public.web.cern.ch/public/en/About/Web-en.html

Più visitati Popolare Network Rai Podcast Radio2

About us Science Research The LHC People

CERN's mission
CERN's structure
The name CERN
A global endeavour
History highlights
Nobel Prizes
Why fundamental science?
Basic science in a competitive world
The use of basic science
Where the web was born
How the web began
How the web works
The first website:
info.cern.ch

Where the web was born

Tim Berners-Lee, a scientist at CERN, invented the World Wide Web (WWW) in 1990. The Web, as it is affectionately called, was originally conceived and developed to meet the demand for automatic information sharing between scientists working in different universities and institutes all over the world.

CERN is not an isolated laboratory, but rather a focus for an extensive community that now includes about 60 countries and about 8000 scientists. Although these scientists typically spend some time on the CERN site, they usually work at universities and national laboratories in their home countries. Good contact is clearly essential.

The basic idea of the WWW was to merge the technologies of personal computers, computer networking and hypertext into a powerful and easy to use global information system.

Copyright CERN 2008 - [Web Communications](#), DSU-CO

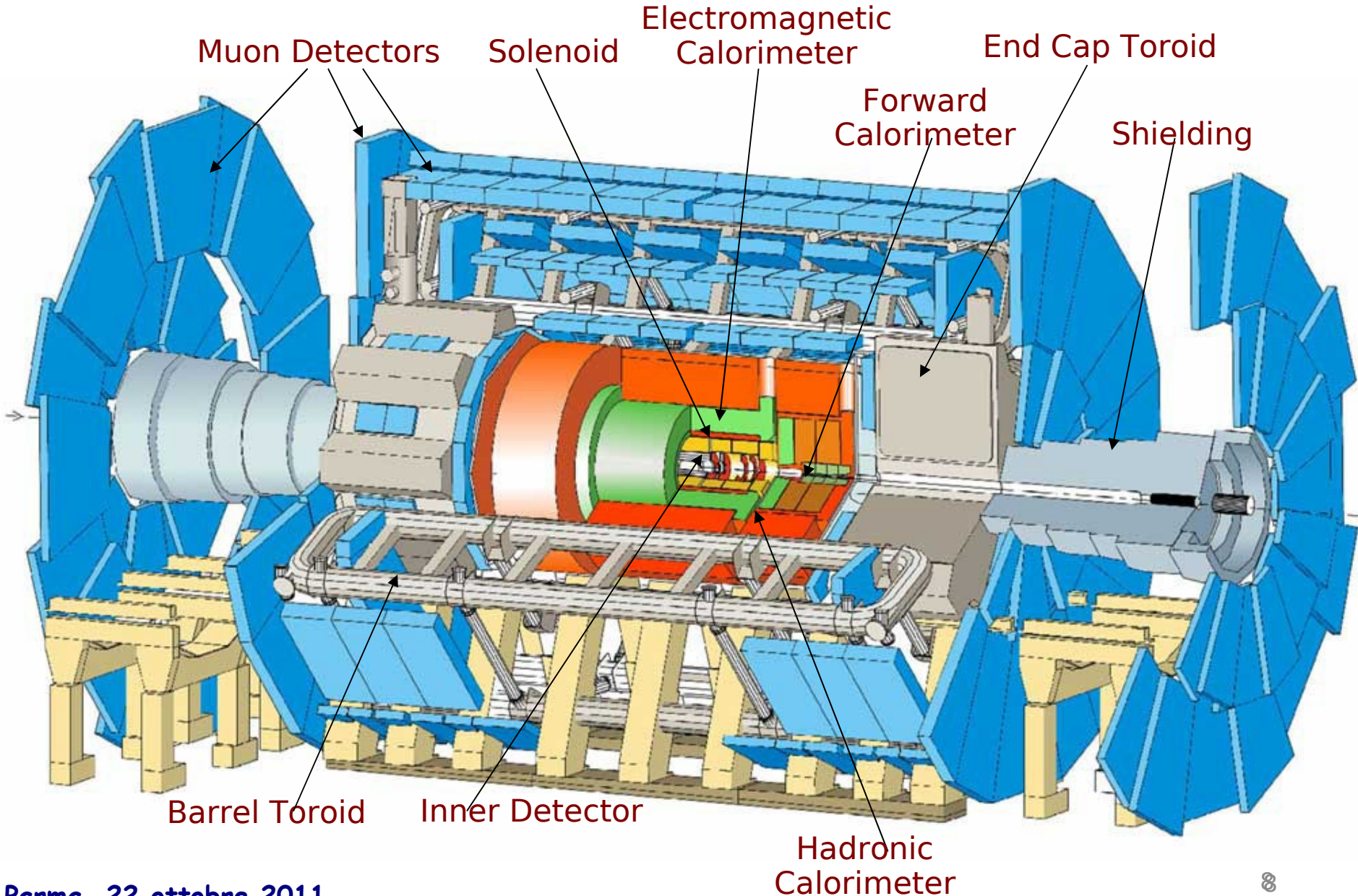
Done

nel 2009 ha festeggiato i 20 anni
<http://info.cern.ch/www20>

LHC



ATLAS: un microscopio alto 22 e lungo 46 m



~ 3000 scienziati di 174 istituti da 38 paesi diversi
più di 1000 studenti di dottorato!

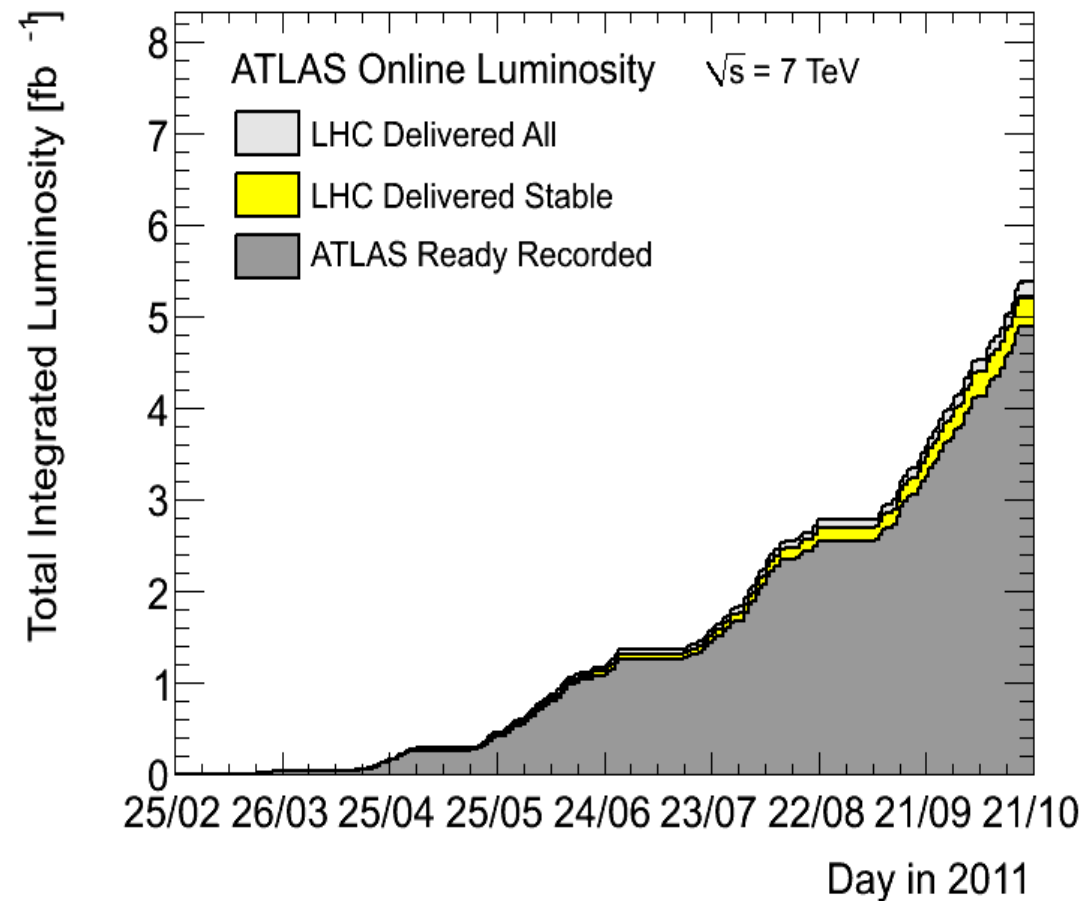


- | | |
|----------------|--------------|
| Argentina | Morocco |
| Armenia | Netherlands |
| Australia | Norway |
| Austria | Poland |
| Azerbaijan | Portugal |
| Belarus | Romania |
| Brazil | Russia |
| Canada | Serbia |
| Chile | Slovakia |
| China | Slovenia |
| Colombia | South Africa |
| Czech Republic | Spain |
| Denmark | Sweden |
| France | Switzerland |
| Georgia | Taiwan |
| Germany | Turkey |
| Greece | UK |
| Israel | USA |
| Italy | CERN |
| Japan | JINR |

ATLAS
Collaboration



Dati 2011



~ 400 eventi/s

1 evento ~ 1.5 MB

-50% "zip" in volo

~ 1 TB/ora

live time $\sim 33\%$

→

~ 3 PB/anno

[4 miliardi di eventi]

il Calcolo in ATLAS/LHC

- ONLINE: interattivo / real time (?)
 1. selezione e acquisizione dati (DAQ)
 2. trasferimento e storsaggio
- OFFLINE: ~ non interattivo (code batch)
 3. ricostruzione eventi
 4. analisi
 5. simulazione

Problematiche

- ONLINE

→ efficienza, velocità, robustezza, stabilità,
enormi flussi di dati, controllo strumentazione

- OFFLINE

→ precisione, ripetibilità

rete, storage, database, fogli elettronici, ...

versioning, documentazione ... "event display"

Real Time ?

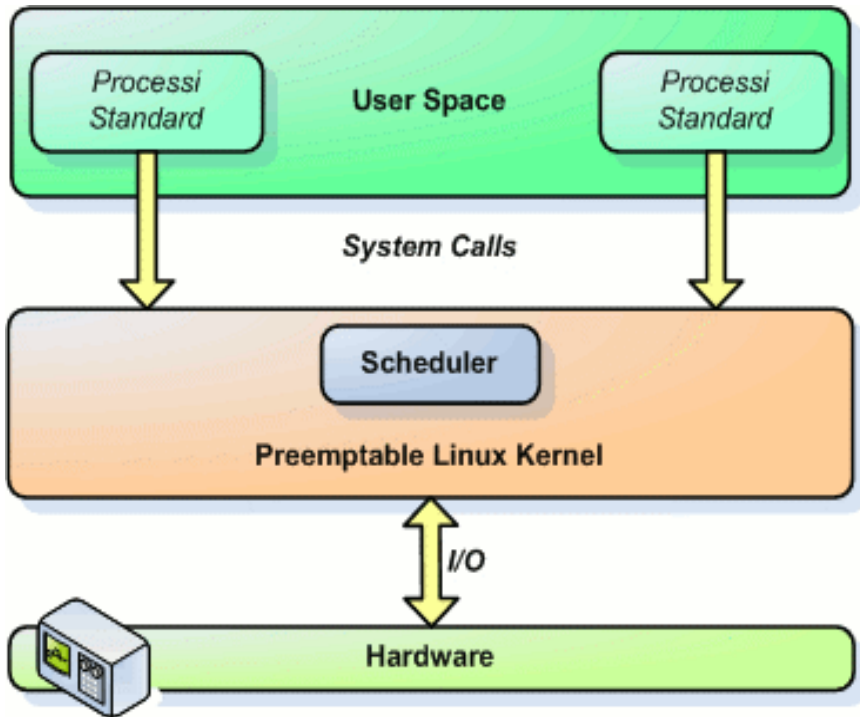
REAL TIME O.S. :

massimo ritardo di risposta definito

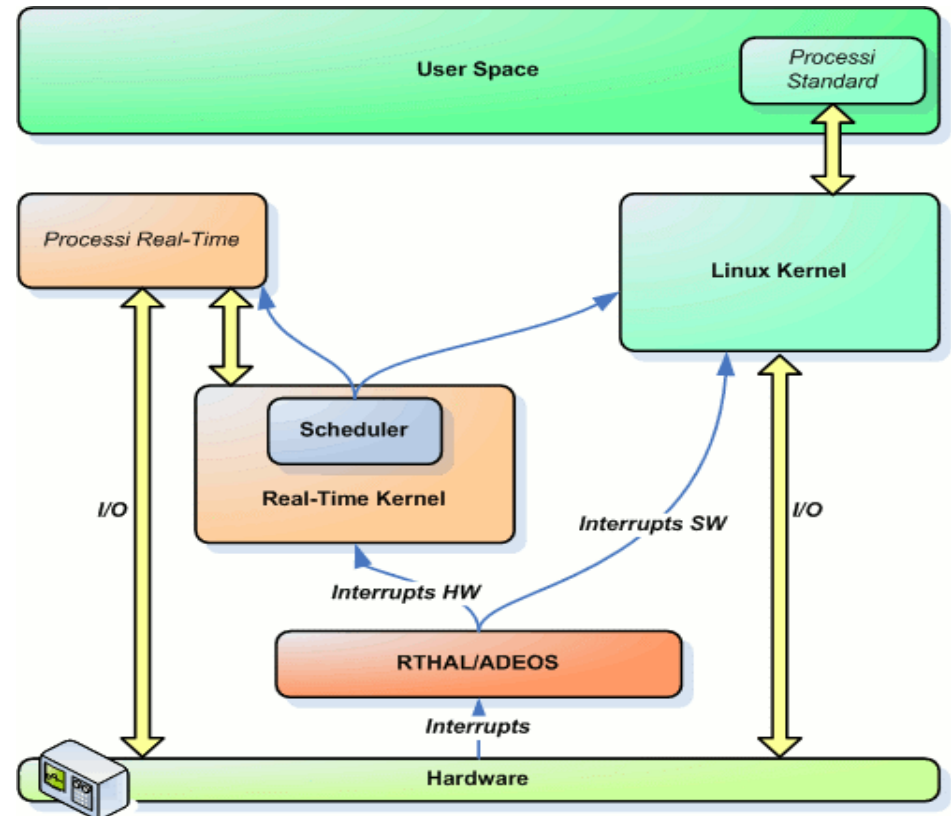
Il kernel "standard" UNIX non è real time:

una chiamata di sistema può richiedere
un tempo lungo a piacere ...

UNIX "Real-Time"



Low-latency patch
(Ubuntu Studio):
linux kernel interrompibile



RTAI: il kernel linux gira
come una applicazione a
priorità maggiore

S.L.C. (x86)

Scientific Linux: release creata e mantenuta da FermiLab e Cern (più altre università e laboratori nel mondo)

Nata nel 2004 a Fermilab

“Red Hat Enterprise Linux” ricompilata e integrata con pacchetti specifici:

<https://www.scientificlinux.org/>

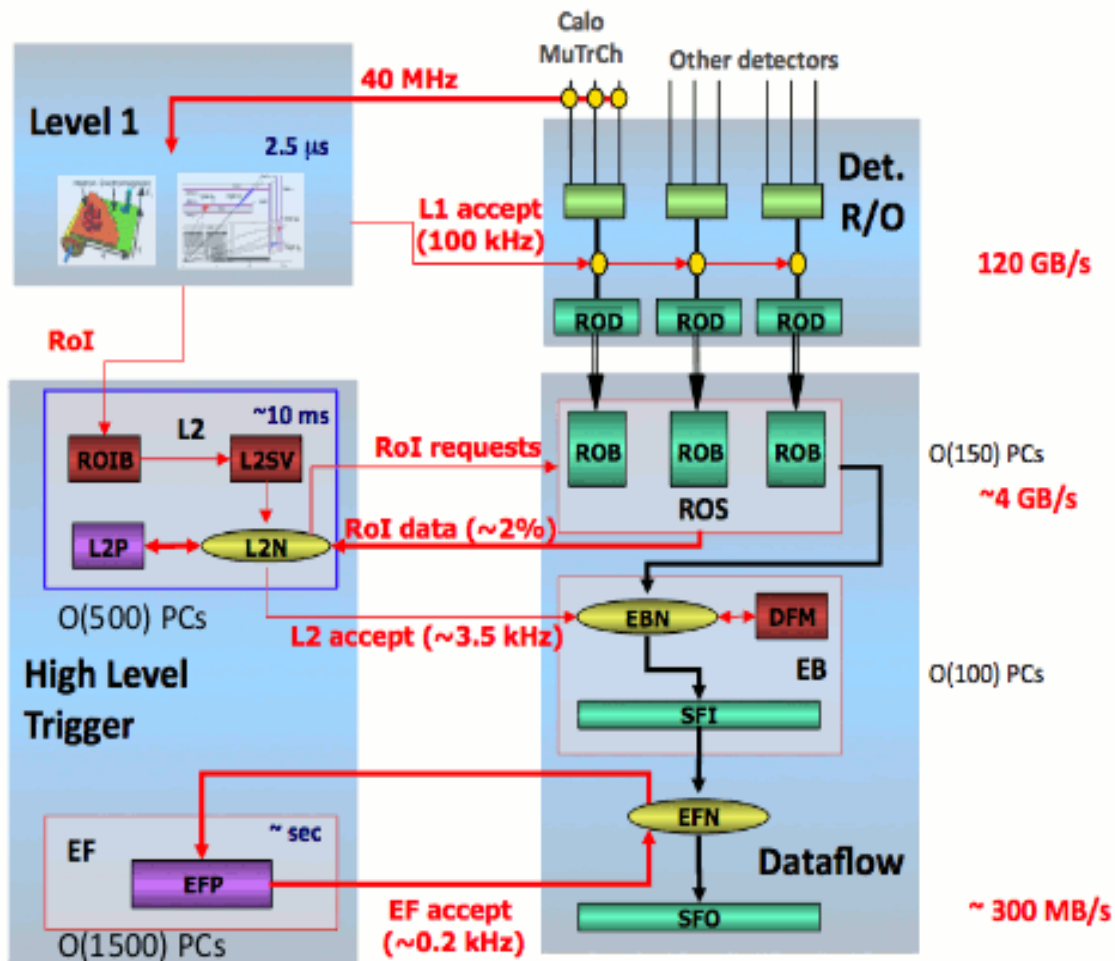
Scientific Linux Cern: sottovariante CERN

<http://linux.web.cern.ch/linux/scientific.shtml>

DAQ @ ATLAS

Selezione eventi "on-line"

- Elettronica e computer dedicati
- migliaia di processori in parallelo (hardware)
- decine di migliaia di processi da controllare (software)

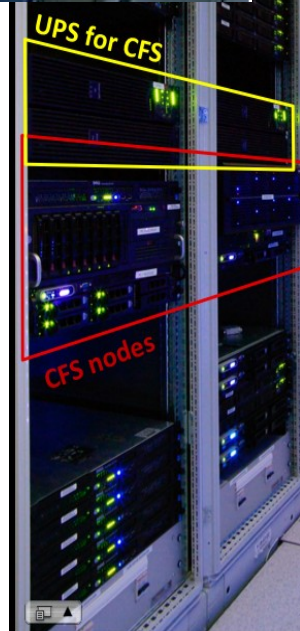




la Sala di Controllo



i Rack



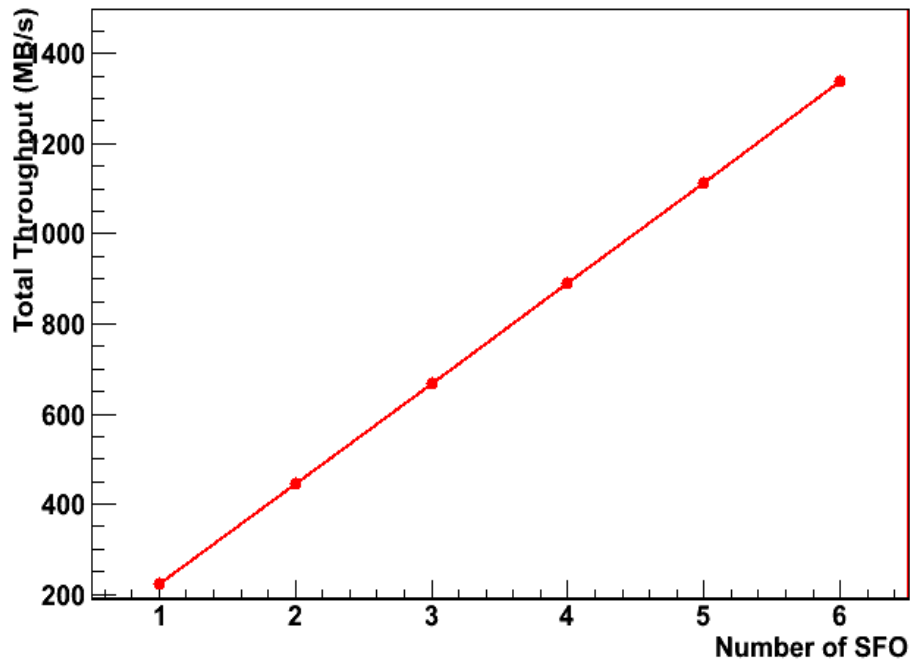
SFO = Online Storage (cache)

6 macchine:

24 dischi 1 TB = 144 TB

2x4 core (16 "processori"
ind.)

24 GB RAM



Parma, 22 ottobre 2011



fra la via Emilia e il West

E4 Computer Engineering - Scandiano

<http://www.e4company.com/>

500 sistemi di calcolo HPC all'INFN

Aggiudicandosi questa importante fornitura, E4 si riconferma ancora una volta come realtà di riferimento nell'ambito dei supercomputer a livello nazionale, ma anche internazionale

E4 Computer Engineering, Azienda specializzata nella produzione di sistemi informatici d'alta fascia e a elevate performance per l'industria e i centri di calcolo e ricerca scientifica, ha vinto l'ultima gara di appalto di IT procurement dell'INFN - CNAF, il Centro Nazionale per la Ricerca e Sviluppo nelle Tecnologie Informatiche e Telematiche.

Aggiudicandosi questa nuova fornitura di quasi 500 High Performance PC (HPC), E4 ha dimostrato ancora una volta sul campo la bontà del proprio modello di business e l'elevata professionalità del proprio team, in gran parte ingegneri provenienti dalle università emiliane, che le hanno consentito di competere ad armi pari e superare anche in questo caso i più importanti player della techno-



Vincenzo Nuti.

logia a livello mondiale.

Aggiudicandosi quest'ultima gara, la dinamica azienda emiliana può oggi vantare all'interno dell'INFN circa 2.300 computer ad alte prestazioni, che contribuiscono quotidianamente al lavoro dell'ente di ricerca, ai quali si aggiungono sistemi storage in grado di immagazzinare sui loro 6.700 dischi rigidi informazioni per oltre 14 Petabyte, ossia 14.000 Terabyte.

Le esperienze e il know-how

accumulato sul campo, anche grazie al CNAF dell'INFN, hanno consentito all'azienda di diventare da qualche anno fornitore di sistemi ad alte prestazioni anche per il prestigioso CERN di Ginevra, il più importante e famoso centro di ricerca mondiale. In questo "tempio" della ricerca, i sistemi di E4 costituiscono dal 25 al 33% dei sistemi di calcolo e d'immagazzinamento dei dati del data center e, a oggi, l'azienda è l'unica realtà tutta italiana che riesce a competere con successo con i "big" internazionali.

"I sistemi di calcolo e supercalcolo sono spesso totalmente sconosciuti al grande pubblico, ma i supercomputer giocano, in realtà, un ruolo straordinariamente importante quando si ragiona in termini di competi-

tività e sviluppo di un paese, poiché permettono di effettuare, ad esempio, simulazioni e test su nuovi prodotti, prima ancora di averli realizzati, con evidenti vantaggi in termini di tempo e denaro.", ha affermato Vincenzo Nuti, Amministratore Delegato di E4 Computer Engineering, che ha poi proseguito: "Ritengo, in ogni caso, che i livelli di eccellenza che abbiamo raggiunto vadano in parte sicuramente ricercati nel tessuto dell'Emilia Romagna, regione in cui abbiamo sede e dove abbiamo avuto modo di trovare sia professionalità, che opportunità, due ingredienti che hanno certamente contribuito al nostro successo anche in campo internazionale."

Per informazioni:
www.e4company.com

- PC / WS INTEL ENTRY
- WORKSTATION INTEL
- SERVER INTEL 1U RACK
- SERVER INTEL 2U/3U RACK
- SERVER INTEL 4U/TOWER
- SISTEMI EMBEDDED
- HIGH DENSITY
- SERVER AMD 1U/2U RACK
- GPU SOLUTIONS
- SISTEMI DI STORAGE
- SWITCH ETH / FC / IB
- ARMADI
- CPU-HD-RAM-SW-VGA
- Componenti ed accessori
- Supermicro - chassis M/B
- Adapter HBA / HCA
- Networking - KVM - UPS

Software ...

Trasferimento, processamento, monitoraggio dati:

C/C++ (protocolli di rete: UDP, TCP)

GUI: Java / JS / Qt / Python (tk/tcl)

Sistema Esperto: Common Lisp

Inter Process Communication: CORBA

Configurazioni/Calibrazioni/Allineamenti/Geometrie:

file, OKS (xml), COOL, ORACLE, SQLITE, Python ...

largo uso di Proxy

Documentazione, gestione problemi: WWW, Twiki, Savannah

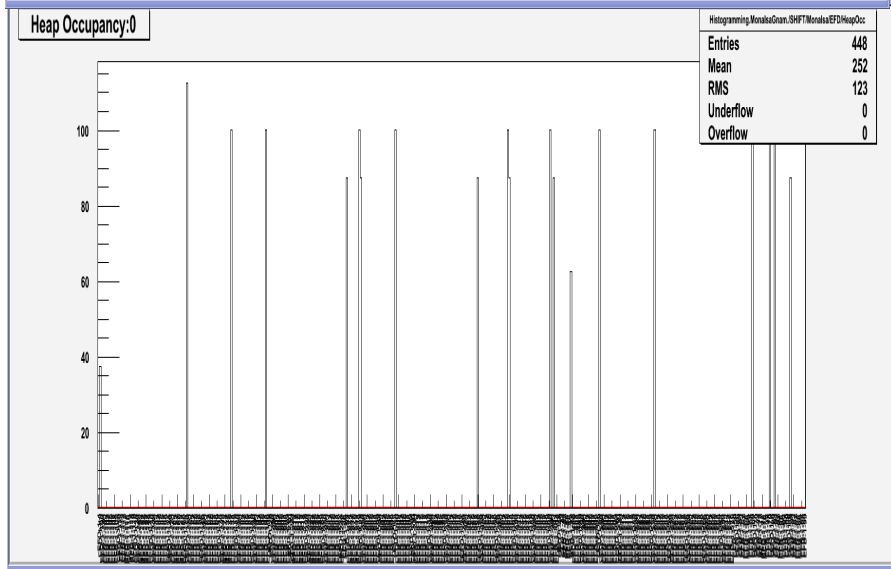
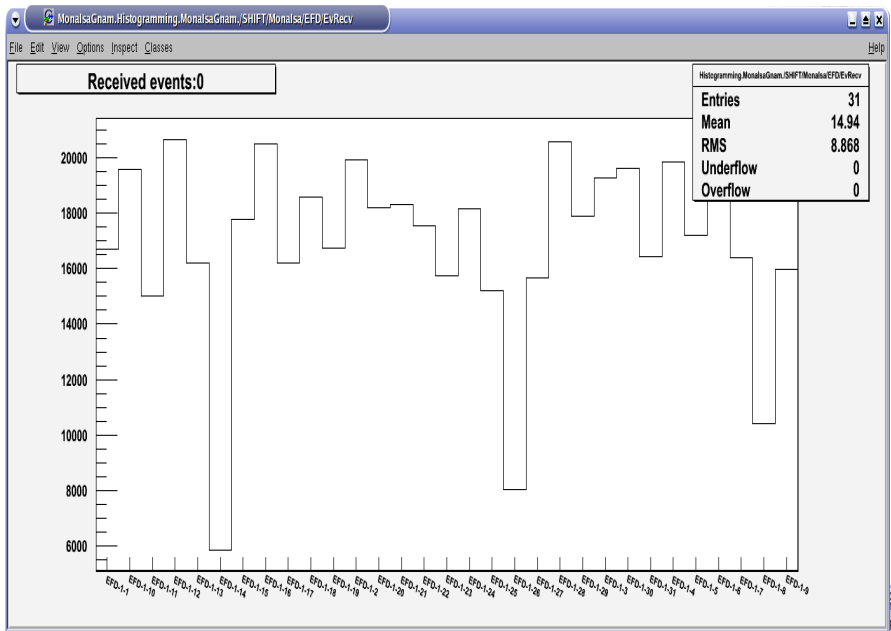
... Nagios (monitoraggio !), IPMI (controllo !) ...

Parole chiave: Macchine a Stati Finiti, Scalabilità,

Partizionabilità, Configurabilità, Sicurezza

Monitoraggio Online

Information Service



Partition 'ATLAS', server 'DF-EF-Segment-01-rack-Y03-06D2-iss'

| Name | Type | Modified | Description |
|----------|------|-------------------------|-------------|
| EFD-1-25 | EFD | 16/7/08 09:43:31,549965 | |
| EFD-1-26 | EFD | 16/7/08 09:43:34,503773 | |
| EFD-1-27 | EFD | 16/7/08 09:43:31,834124 | |
| EFD-1-28 | EFD | 16/7/08 09:43:31,946579 | |

| Value | Type | Name | Description |
|--|--------|-----------------------|--|
| pc-tdq-xpu-0245:/local_L/efHeap/sharedHeap.cmc.ATLAS | String | SharedHeap | SharedHeap file fullpath |
| 3 | UI16 | ConnNbrSFIs | Number of connected SFIs (sum over InputTasks) |
| 5 | UI16 | ConnNbrSFOs | Number of connected SFOs (sum over OutputTasks) |
| 4 | UI16 | ConnNbrPTs | Number of connected PTs (sum over ExtPTsTasks) |
| 87.54 | Double | HeapOcc | SharedHeap occupancy (%) |
| 1521 | S32 | EventsRcv | Number of received events |
| 1514 | S32 | EventsSent | Number of events sent to SFO (ie: Dismissed-Deleted) |
| 7 | S32 | EventsInside | Number of events inside |
| 0 | U32 | EventsWaitingForProc | Number of events waiting for processing |
| 2 | U32 | EventsWaitingForDeli | Number of events waiting to be sent to SFO |
| 0 | Double | RateIn | Current rate of incoming events (Hz) |
| 0 | Double | RateOut | Current rate of events sent to SFO (Hz) |
| 0 | Double | FluxIn | Current rate of space allocation in SH; >M data flux |
| 0 | Double | FluxOut | Current data flux to SFO (MB/s) |
| -1 | Double | FlowCtrlStopTime | Guess of the stop transition time (s) |
| 460 | U32 | FlowCtrlISleepTime | Current flow control sleep time (ms) |
| 538 | U32 | FlowCtrlBarrierLocks | Number of times the input barrier has been locked |
| 0 | S32 | ptionNbrProcTimeouts | Number of processing timeouts |
| 0 | S32 | ptionNbrSocketHungUps | Number of PT socket hung-ups |
| 0 | S32 | ptionNbrForceAccept | Number of force accepted events |
| 0 | S32 | efionNbrSFiBrokenConn | Number of broken connections to SFI |
| 0 | S32 | efionNbrSfoBrokenConn | Number of broken connections to SFO |
| 1521, 0, 0, 0, 0, 0 | S32[6] | EventTagTypesIn | Type counters: phys, calib, reserved, debug, unknow |
| 1519, 0, 0, 2, 0, 0 | S32[6] | EventTagTypesOut | Type counters: phys, calib, reserved, debug, unknow |

403 objects | 24 attributes

JavaScript + Web

Web Interface to Atlas Online Information Service

The WebIS service complements the Web Monitoring Interface by providing generic access to any object and histogram in the Atlas online Information Service. This allows to build simple HTML and/or Javascript based web pages that show up-to-date online information from Point 1.

The following list shows some general applications that will be useful for experts who are outside of P1 as well as some examples on how the information can be processed and presented with some simple Javascript code.

Simply look at the HTML source to see how to include e.g. the status display or a histogram into your own page.

Generic Applications

Based on the [ExtJS](#) framework

These are best viewed with a modern browser with a fast Javascript implementation (Firefox > 3.0, Opera > 10.0, Google Chrome, Internet Explorer 8.x). Older browsers will be either very slow or not work at all (e.g. Konqueror). In fact, in many cases IE will not work properly either, I suggest to use any other browser instead...

- [Histogram Browser](#)
- [Information Service](#)
- [Process Manager](#)
- [OKS Configuration Browser](#)
- [Combined Browser](#)

Simple HTML plus some Javascript

- [A simple example on how to integrate histograms into a web page](#)
- [Simple Browser](#)

DAQ Examples

- [Status Display for other partition Status message only](#)

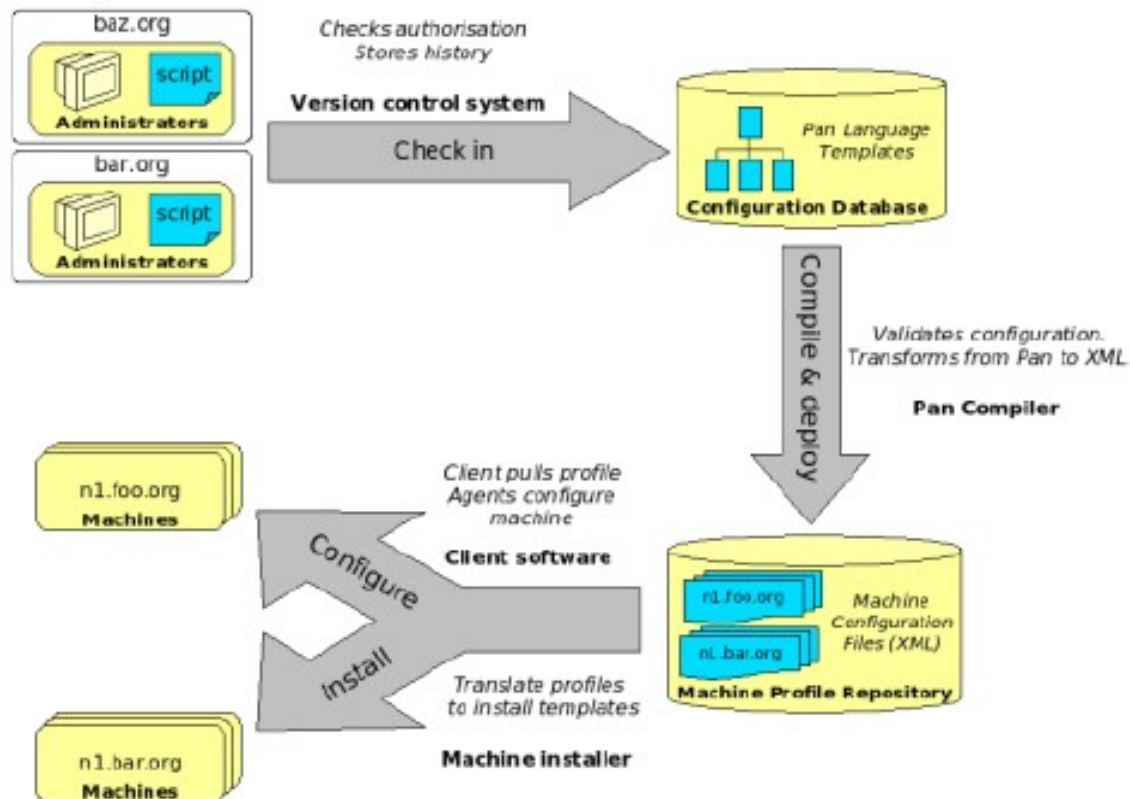
ATLAS: **RUNNING** Run Number: 167521 Run Type: Physics Start: 22/10/10 23:08:14 End: 1/1/70 01:00:00

- [Run Status](#)

System Management

Quattor Workflow

The quattor toolsuite caters for every stage of a typical system management workflow, from specification and management of configurations, to installation of new machines from scratch, to ongoing maintenance of software and services.



Offline

$O(1 \text{ miliardo})$ di eventi all'anno da ricostruire e analizzare
~ Altrettanti da simulare

STORAGE

~3 PB/anno

CPU

~ 7000 kSi2k*anno

Analisi Eventi

Ambiente complesso ... ogni livello richiede competenze specifiche:

Dall'online arrivano informazioni "grezze" (numeri):

→ misure di tempi, cariche elettriche, tensioni

Ricostruzione a più stadi (attività centralizzata):

→ informazioni fisiche (posizioni, velocità)

→ identificazione particelle, energia, quantità di moto

Analisi fisica (attività caotica):

→ criteri di separazione fondo / segnale (selezione eventi)

→ analisi statistica

Simulazione, Ricostruzione e Analisi Dati

Attività distribuita verticalmente e orizzontalmente::

Tier-0 (CERN) → Tier-1 (grossi centri nazionali)

→ Tier-2 (centri regionali) → Tier-3 (istituti)

Ampio uso della virtualizzazione

Dati distribuiti con ridondanza (almeno due copie di ogni dataset)

Cataloghi (database) per tenerne traccia

Esecuzione delocalizzata: nuovo strato software (middleware) che indirizza gli eseguibili dove si trovano i dati, raccoglie e assembla i risultati

la GRID

la Griglia (GRID)

Dati LHC equivalenti a ~20 milioni di CD (una pila alta 20 km) all'anno

Per l'analisi necessari ~100mila dei più veloci processori odierni



WWW: accesso a informazione archiviata in diverse località geografiche

GRID: accesso a risorse di calcolo e di archiviazione dati distribuite su tutto il pianeta



il Middleware

un ulteriore livello di astrazione:

connette applicazioni, componenti, sistemi, su
scale regionali, nazionali, internazionali

hardware → software → middleware

permette di analizzare dati o eseguire
applicazioni su macchine distribuite in tutta
la rete

Il Calcolo LHC in Italia

Tier-1: CNAF (Bologna) unico per tutti gli esperimenti LHC (e non solo)

Tier-2: ~10 (Roma, Legnaro, Torino, Napoli, Catania, CNAF, Pisa, Milano)

Investimento (ad oggi) ~ 30 M Euro (incluse infrastrutture CNAF)

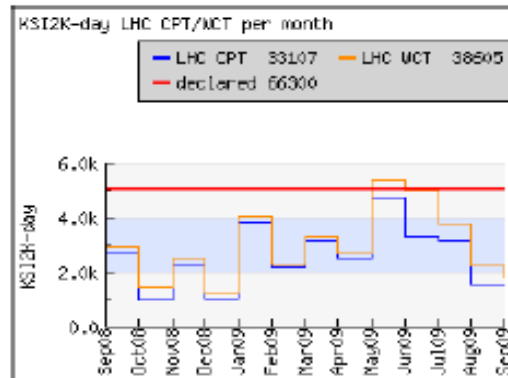
+ molti anni uomo di sviluppo sw (anche grazie a finanziamenti europei)

Il Portale di Monitoring

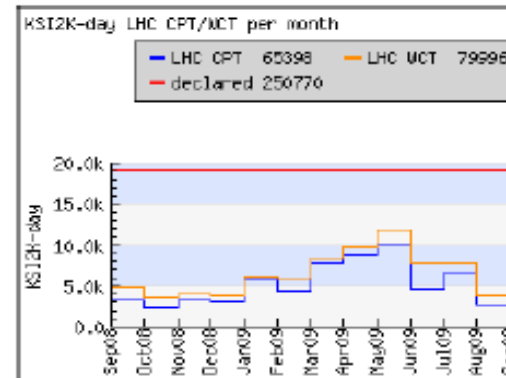


Uso CPU T2 ATLAS

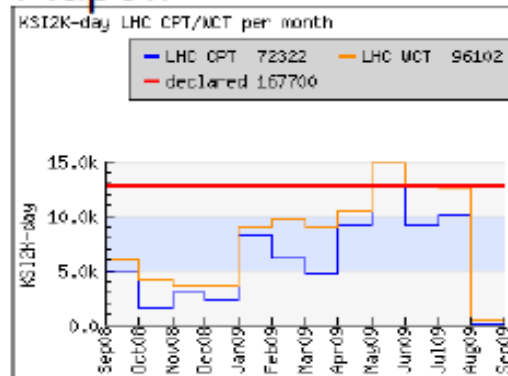
"Frascati"



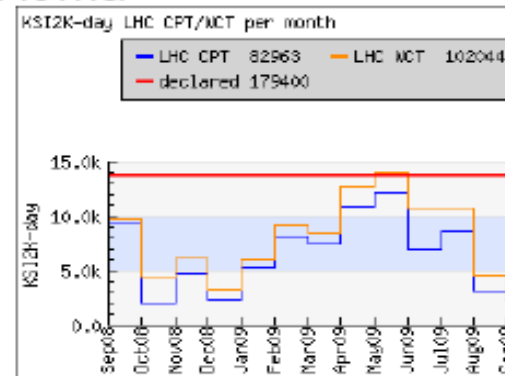
Milano



Napoli



Roma



Italian Grid Infrastructure



<http://www.italiangrid.org>

La GRID in casa

BOINC : <http://boinc.berkeley.edu/>

Open-source software for volunteer computing and grid computing.

Note: if your computer is equipped with a Graphics Processing Unit (GPU), you may be able to use it to compute faster.

| Projects | Windows | | Linux | | Mac OS X | | Solaris | |
|---------------------------------------|----------|----------------------|-------|--------|----------|-----|---------|-----|
| | 95/98/ME | XP/2000/2003/Vista/7 | x86 | x86-64 | PowerPC | x86 | SPARC | x86 |
| climateprediction.net | No | Yes | Yes | | | Yes | | |
| Einstein@home | Yes | Yes | Yes | Yes | Yes | Yes | Yes | |
| Leiden Classical | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| Rosetta@home | Yes | Yes | Yes | Yes | limited | Yes | No | No |
| SETI@home | 98 only | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SIMAP | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| SpinHenge@home | Yes | Yes | Yes | No | No | No | No | No |
| The Lattice Project | Yes | Yes | Yes | Yes | Yes | Yes | No | No |
| World Community Grid | No | Yes | Yes | | Yes | Yes | No | No |

Il Valore Aggiunto della Collettività ...

Risultati altrimenti impensabili possono essere raggiunti grazie al contributo di tutti ...

BOINC Berkeley Open Infrastructure for Network Computing

es:

<http://milkyway.cs.rpi.edu/milkyway/>

<http://einstein.phys.uwm.edu/>

... ricerche in campo medico, farmacologico, ...

TheoPhys/TheoMPI



TheoMpi
resources

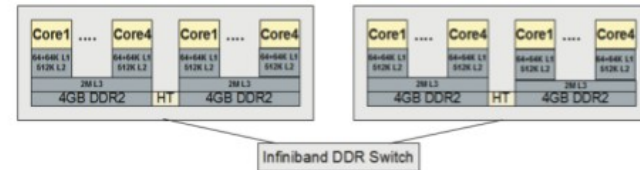
The cluster is installed, configured and maintained by the **INFN-Pisa computing center** ([web site](#))

Computing:

128 WNs dual Opteron 8356, 2x4 cores per node, 8GB ram, 1024 cores, 10TFlops peak perf.

- SW: Linux x86_64, openMPI

1 CE gridce3.pi.infn.it : Cream-CE, LSF



High Speed Network:

Infiniband DDR

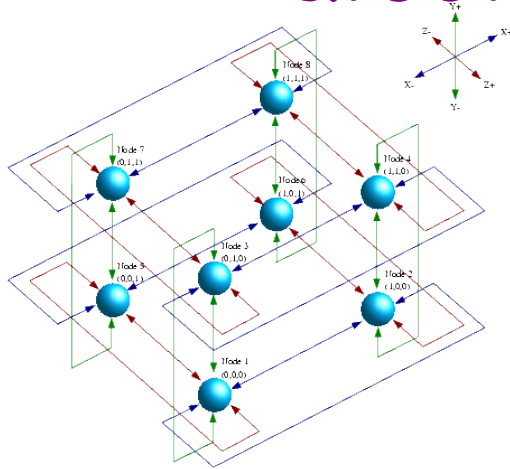
Storage:

1 SE gridsrm.pi.infn.it : STORM

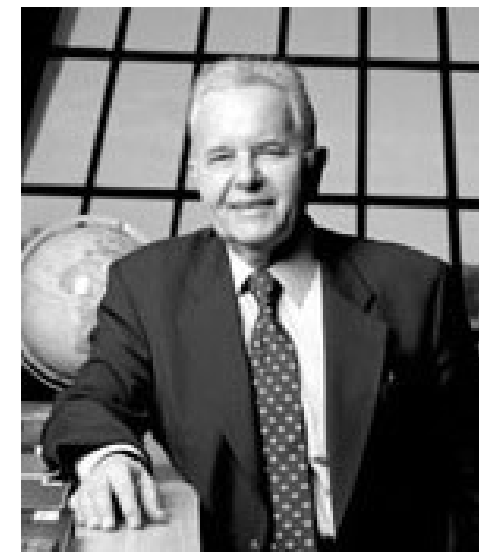
10 TBytes GPFS over Infiniband



Calcolo Parallelo ...



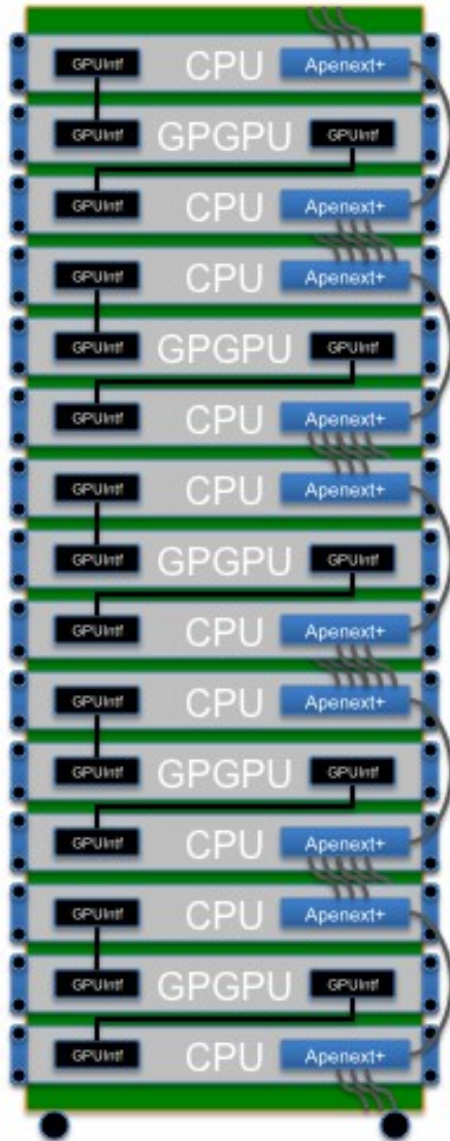
dalla Fisica Teorica
al Super-Computing
ovvero il progetto APE



N. Cabibbo

| | APE | APE100 | APEmille | APEnext |
|-----------------------|-------------|-------------------|-------------|-------------|
| Year | 1984-1988 | 1989-1993 | 1994-1999 | 2000-2005 |
| Number of processors | 16 | 2048 | 2048 | 4096 |
| Topology | Flexible 1D | Next Neighbour 3D | Flexible 3D | Flexible 3D |
| Total Memory | 256 MB | 8 GB | 64 GB | 1 TB |
| Clock | 8 MHz | 25 MHz | 66 MHz | 200 Mhz |
| Peak Processing Power | 1 GFlops | 100 GFlops | 1 TFlops | 7 TFlops |

QUOnG: GPU-based HPC system



- QUantum chromodynamics ON Gpu

PC clusters + GPU + 3D network
APEnet+ boards

- 42U rack system:
- 60 TFlops/rack peak
- 25 kW/rack (i.e. 0.4 kW/TFlops)
- 300 k€/rack (i.e. 5 k€/TFlops)

High Performance Supercomputing



[Home](#) ▶ [Statistics](#) ▶ [List Statistics](#)

Operating system Family share for 06/2011

In addition to the table below, you can view the visual charts using the [TOP500 charts page](#). A direct link to the charts is also [available](#).

| Operating system Family | Count | Share % | Rmax Sum (GF) | Rpeak Sum (GF) | Processor Sum |
|-------------------------|------------|-------------|--------------------|--------------------|----------------|
| Linux | 456 | 91.20 % | 53513545 | 78503517 | 6443648 |
| Windows | 6 | 1.20 % | 459520 | 563535 | 63140 |
| Unix | 22 | 4.40 % | 1718426 | 2205312 | 124976 |
| BSD Based | 1 | 0.20 % | 122400 | 131072 | 1280 |
| Mixed | 15 | 3.00 % | 3116134 | 3776512 | 1146880 |
| Totals | 500 | 100% | 58930025.59 | 85179949.00 | 7779924 |

Conclusioni

- il mondo gnu/linux/free sw ha un legame doppio con la ricerca in fisica delle particelle elementare
- è una formidabile piattaforma di crescita sia individuale che collettiva
- la condivisione della conoscenza è un valore primario che si riflette anche nelle modifiche delle politiche di accesso alle pubblicazioni scientifiche → "Open Access"
- la condivisione delle risorse permette di ottenere risultati significativi sia dal punto di vista scientifico che da quello sociale