Linux e la ricerca scientifica

Roberto Ferrari

Parma LUG

Linux Day 2010

23 ottobre 2010

DISCLAIMER

a) non ho fatto progressi rispetto allo scorso anno … :-(

b) mi limiterò all'ambiente della fisica delle particelle (INFN, CERN)

c) non prendetemi troppo sul serio !

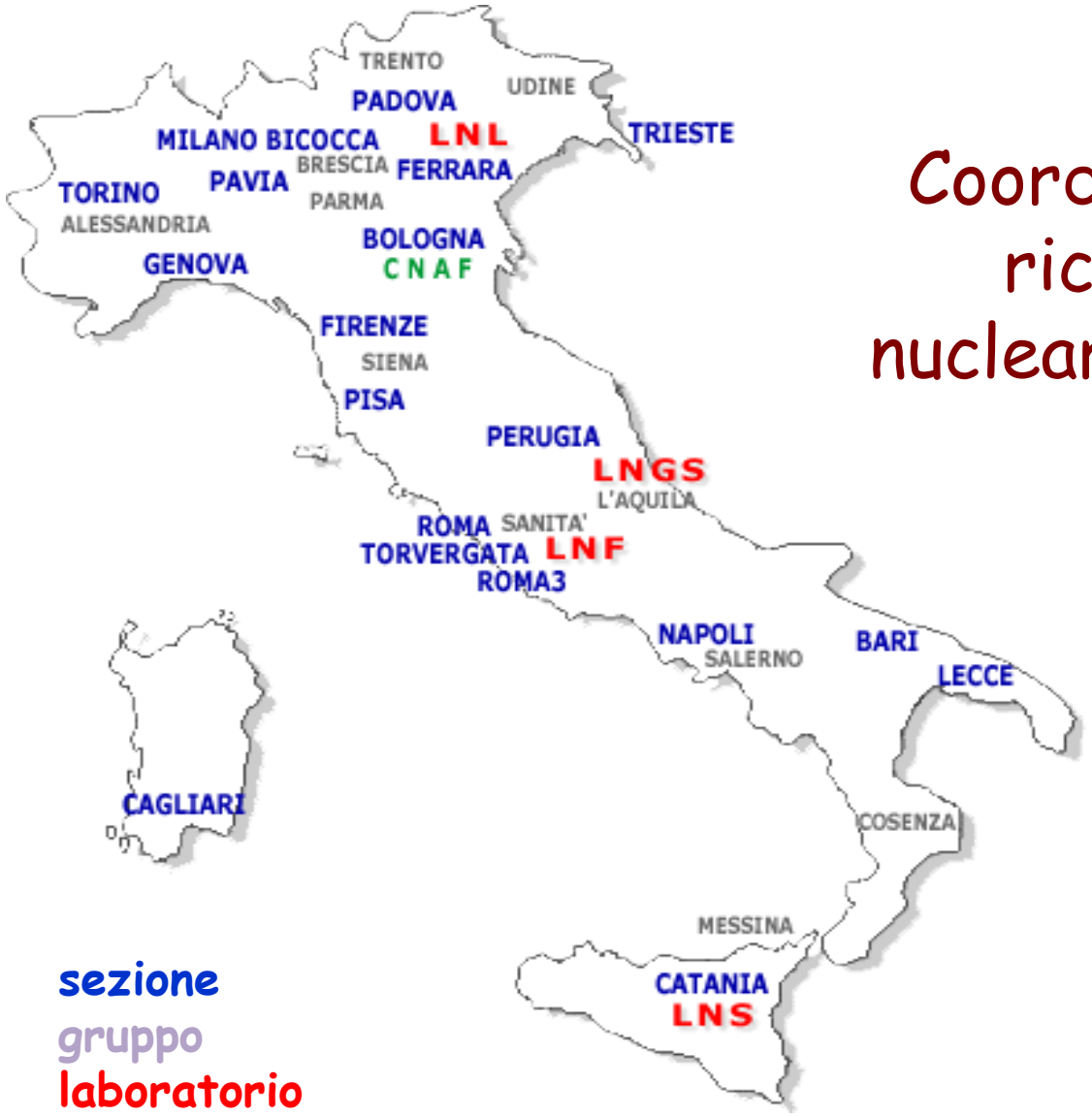
I.N.F.N.:

Istituto Nazionale di Fisica Nucleare

CERN:

European Organization for Nuclear Research

L'INFN



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

sezione
gruppo
laboratorio

CERN

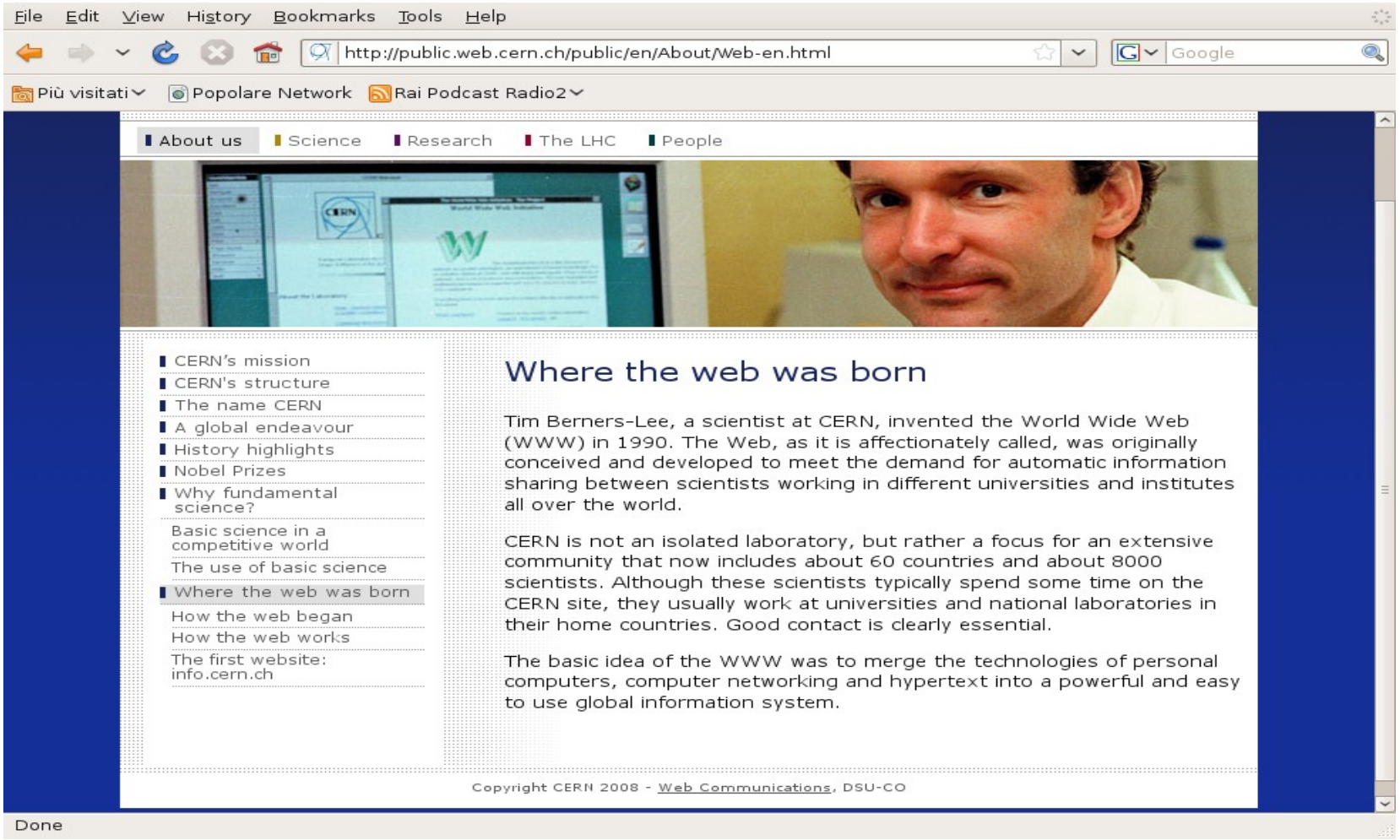
Nasce come laboratorio di ricerca europeo → ora è il più grande del mondo nel settore: luogo di confronto e collaborazione per ricercatori di ogni dove ...

Italia: stato membro fin dalla fondazione

Esperimenti: collaborazioni internazionali (Italia -> gruppi INFN)


LHC
SPS
CERN Site (Meyrin)

Dove è nato il Web ?



ha compiuto 20 anni - http://info.cern.ch/www20

Il Calcolo

Problematiche diverse:

A) ONLINE (acquisizione dati: DAQ)

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

B) OFFLINE (simulazione, ricostruzione e analisi dati)

→ precisione, ripetibilità, enormi quantità di dati

~ NON INTERATTIVO (code batch)

rete, storage, database, fogli elettronici, …

versioning, documentazione … "event display"

… collaborazioni internazionali

Tanti istituti e gruppi di ricerca diversi

Scelte interne poco coordinate con altri:

→ guidate dalla competenze e dalle esperienze disponibili nelle rispettive sedi

→ a volte poco esportabili (es. sviluppate in casa)

C'è chi ha scritto linguaggi, chi sistemi operativi, chi ha violato i primi MAC (1984), segando il case per interfacciarli a schede esterne …

Negli ultimi 30 anni ...

Si passa da scelte abbastanza (o molto) impegnative (acquisto → supporto nel tempo) a soluzioni scalabili:

'80: minicomputer, mainframe, supercomputer VAX, IBM, CDC, CRAY (s.o. proprietari)

'90: workstation e single-board-computer specializzati (unix proprietari): scalabilità ...

2k: pc desktop, pc rack mounted, single board computer "off the shelf" (linux)

Programmazione:

fortran 77 (+ vari assembler) → C → C++

CRAY X-MP 48 con UNICOS (1988)

- calcolo vettoriale
- UNIX con code batch (sviluppate in casa)
- clock ~ 118 MHz, RAM 128 MB
- potenza di calcolo
  <~ ½ Xbox
- costo ~10 M$

Architettura CPU

Vivace polemica durante tutti gli anni '90:

CISC .vs. RISC

CISC: chiusura del "semantic gap" fra istruzioni di alto livello e microcodice (molti registri, molte istruzioni complesse, molti modi di accedere la memoria)

RISC: il contrario → complessità nel software (unix & C), semplicità nell'hardware

(migliori compilatori, memorie meno costose)

Acquisizione Dati (DAQ)

* REAL TIME O.S. : ritardo massimo di risposta definito *

Il kernel UNIX "standard" non è real time:

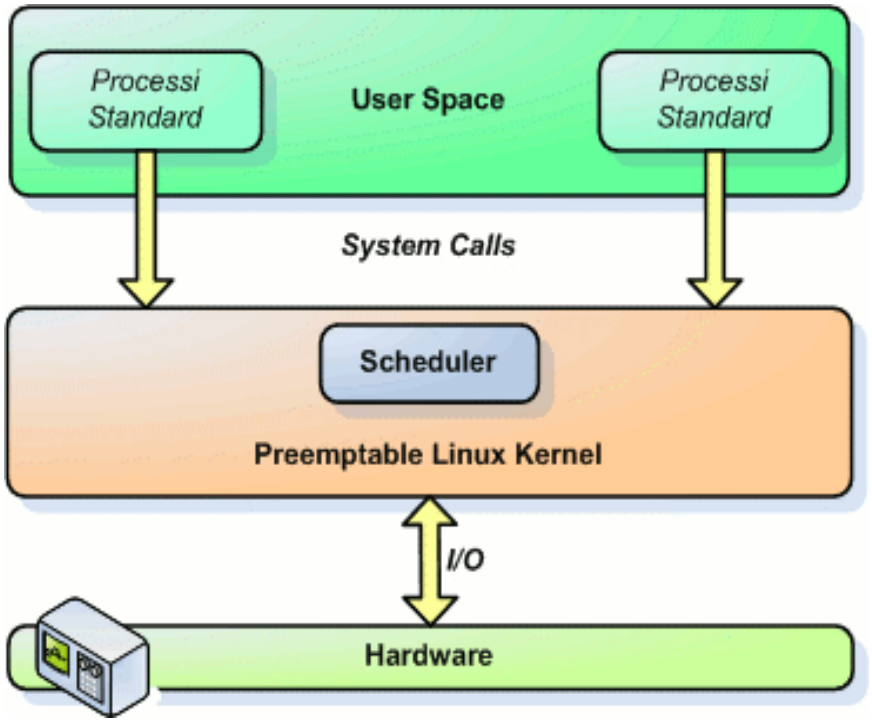
una chiamata di sistema può richiedere un tempo indefinito

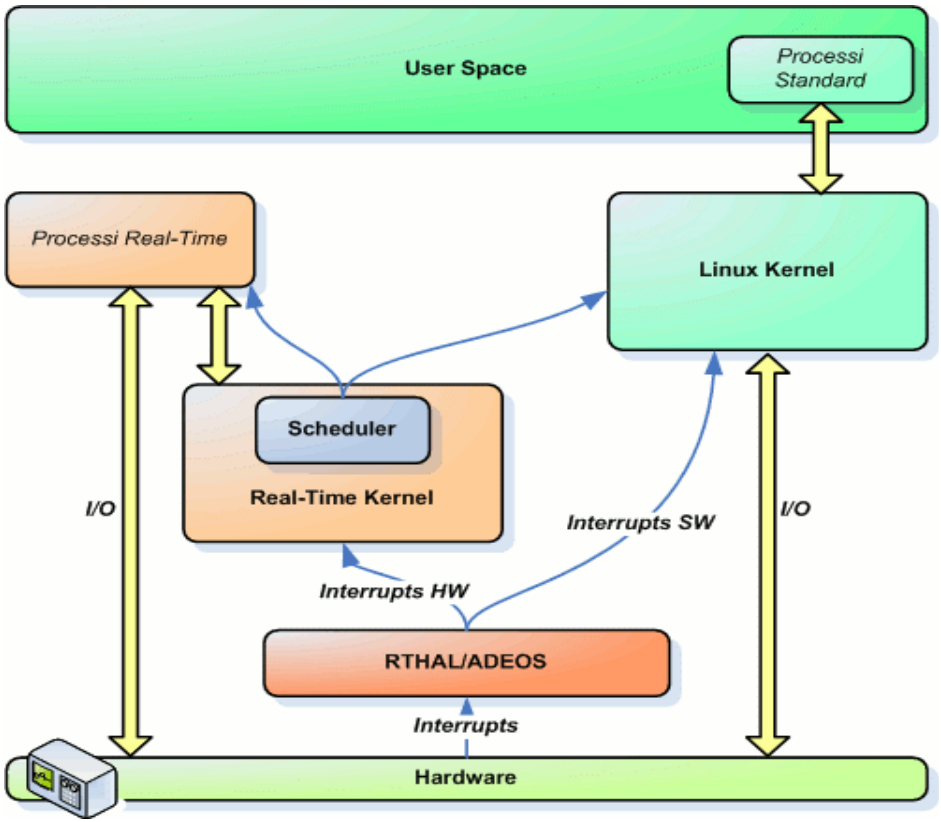
Dalla ~ metà degli anni 80 (LEP):

da VAX/VMS → Single Board Computer (SBC) VME

- processori 680x0
- sistema operativo real time OS-9

UNIX Real-Time





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

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

DAQ @ LHC (ATLAS) ...

Anni 90:

parole chiave:

VME + RISC + **UNIX** real time + **SCALABILITA'**

Es: MIPS R3000/R4000, PowerPC con LynxOs

la richiesta di "real time" cala rapidamente ...

rimangono dubbi rispetto a soluzioni "open source"

Ultimi 10 anni, graduale convergenza verso:

Red Hat Linux → S.L.C.

front-end (SBC): 80x86 + linux

back-end: rack di macchine linux

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

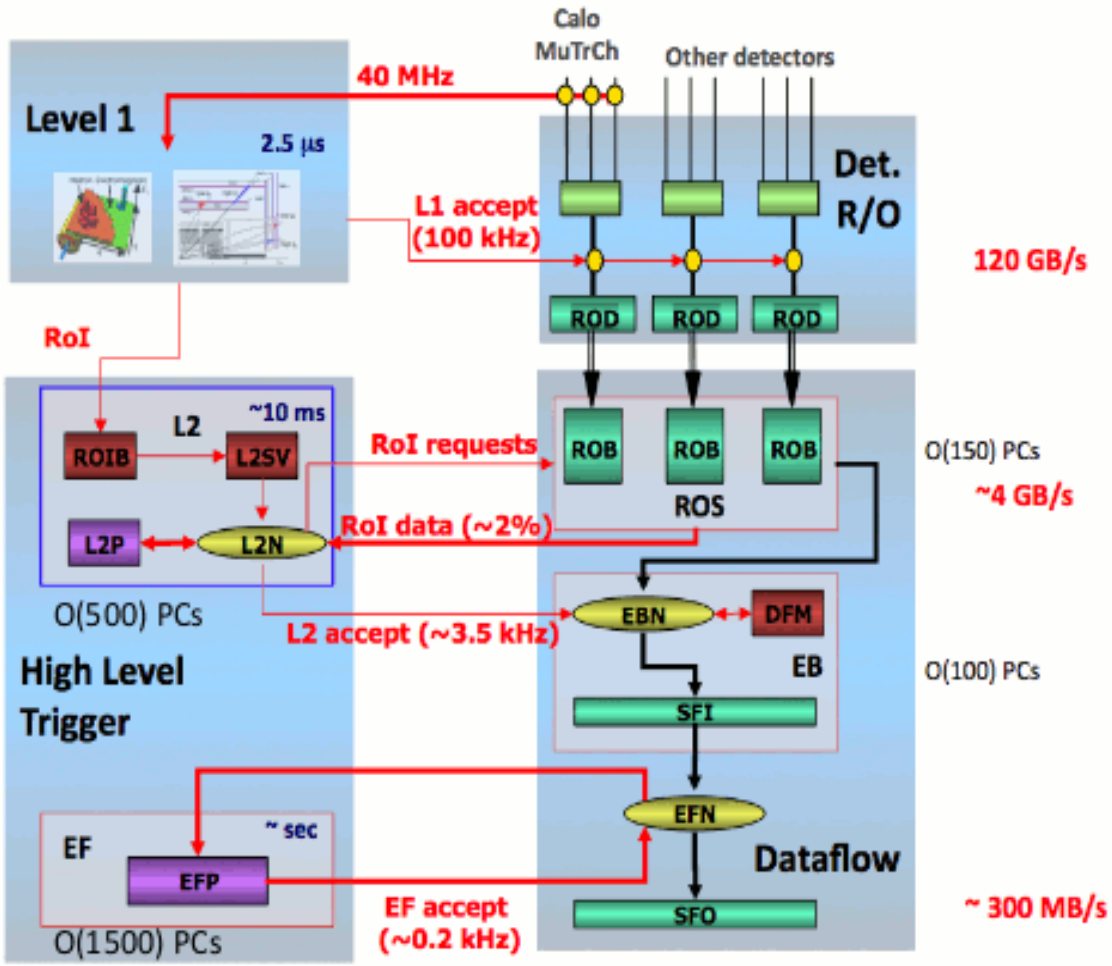
~ 40 M eventi / sec

~ 1 evento / 10^9
INTERESSANTE

~ 100 M di segnali → 1.5 MB/evento

<u>Selezione eventi "on-line"</u>

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



Inventario

~ 100 rack x 30 macchine = ~ 3000 macchine "rack mounted"

~ dual cpu / quad core / 16-24 GB

inoltre ~ 160 SBC (VME) + 150 ROS (readout system)

Performance:

1 macchina (8 core) : 200 eventi/s al livello 2 (40 ms/core)

2 eventi/s all'event filter (4 s/core)

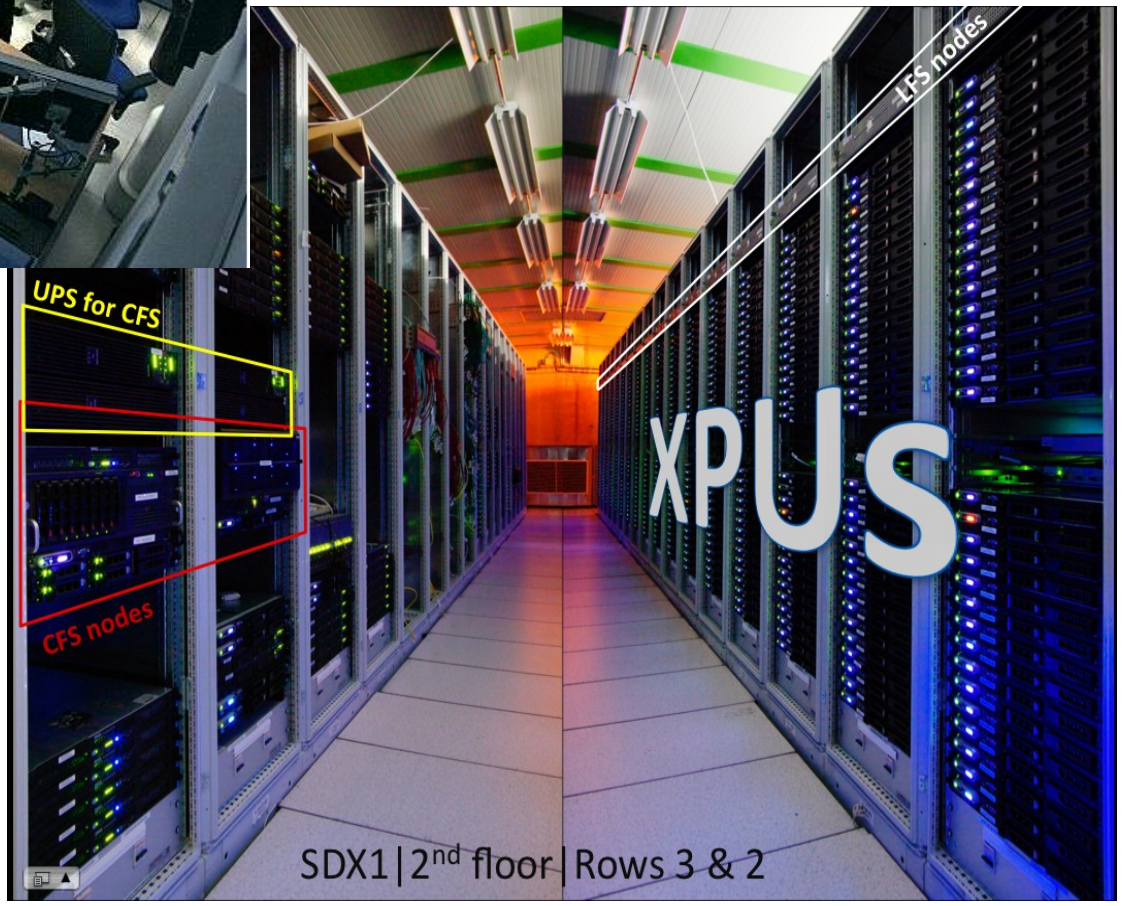
Storage (cache): 6 macchine x 24 dischi cad. (raid5) = 72 TB

E4 Computer Engineering (Scandiano):

24 TB (24 disch x 1 TB), 2x4 core (16 "processori" ind.), 24 GB

Link verso il centro di calcolo: 2 x 10 Gb/s

La Sala di Controllo



I Rack

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

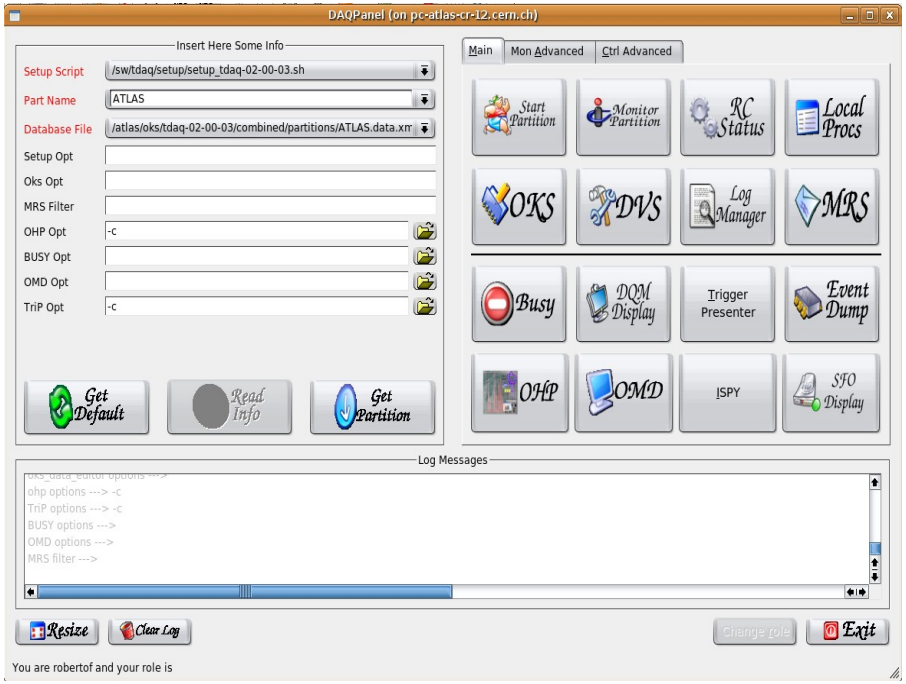
Documentazione, gestione problemi: WWW, Twiki, Savannah

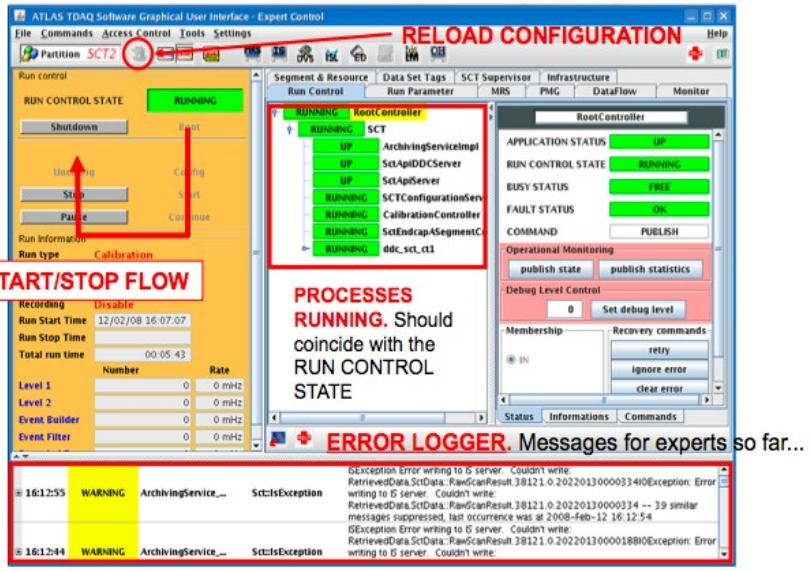
... Nagios (monitoraggio !), IPMI (controllo !) ...

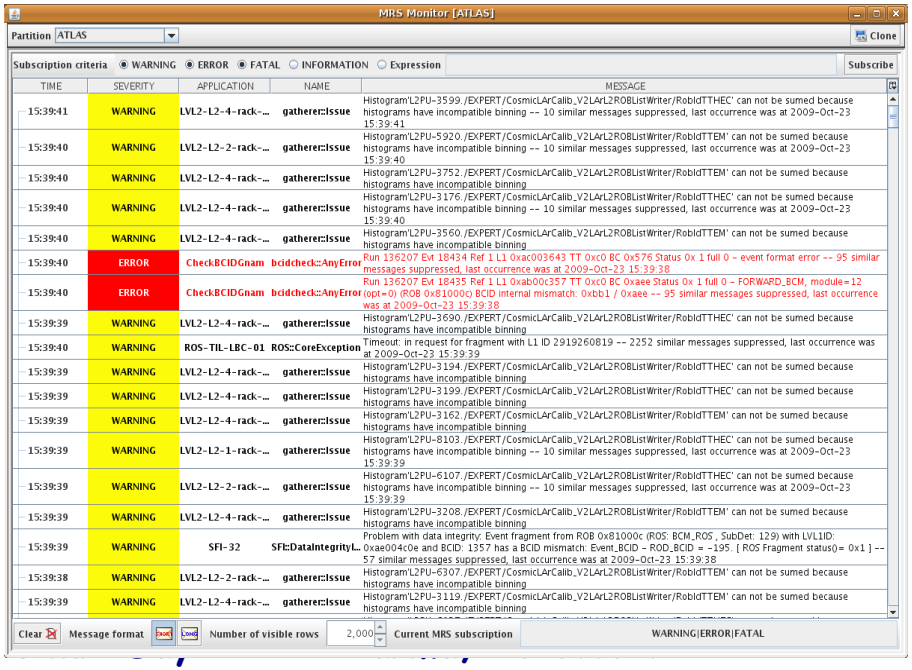
Parole chiave: Macchine a Stati Finiti, Scalabilità,

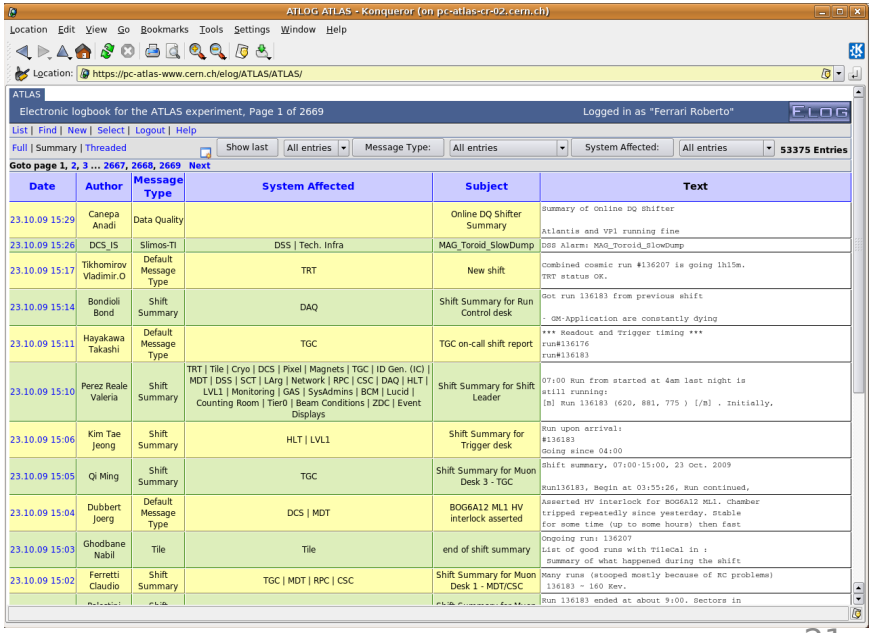
Partizionabilità, Configurabilità, Sicurezza

Run Control
Macchina a Stati Finiti



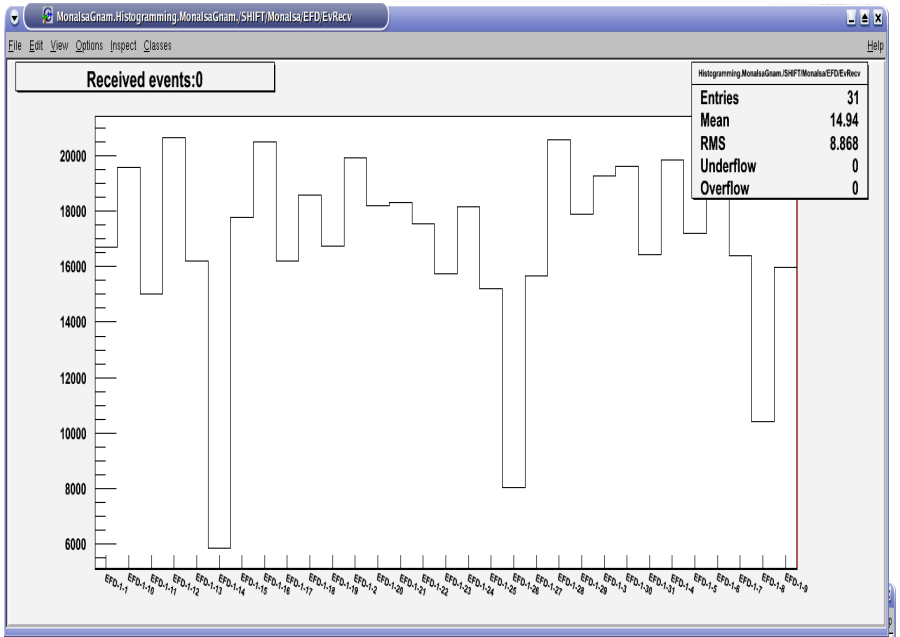


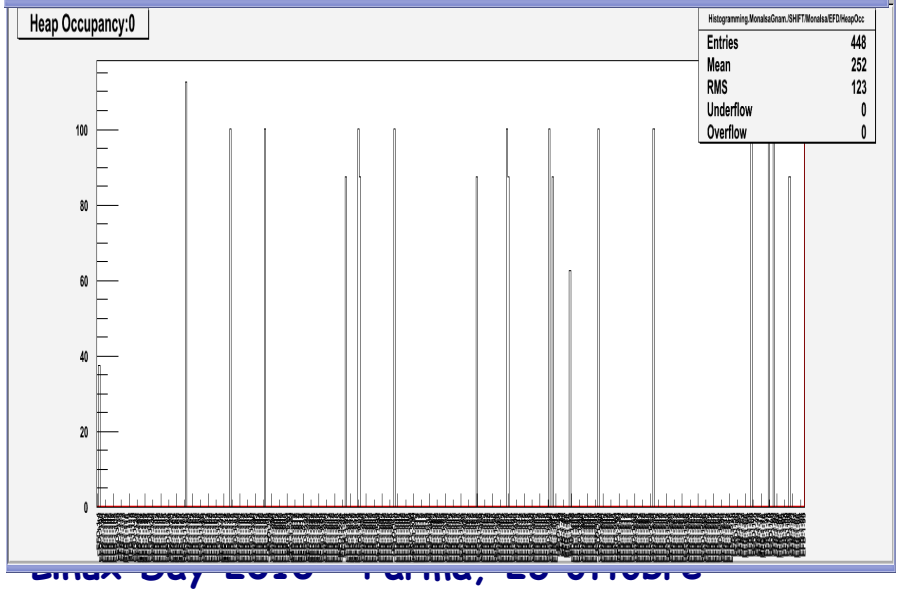


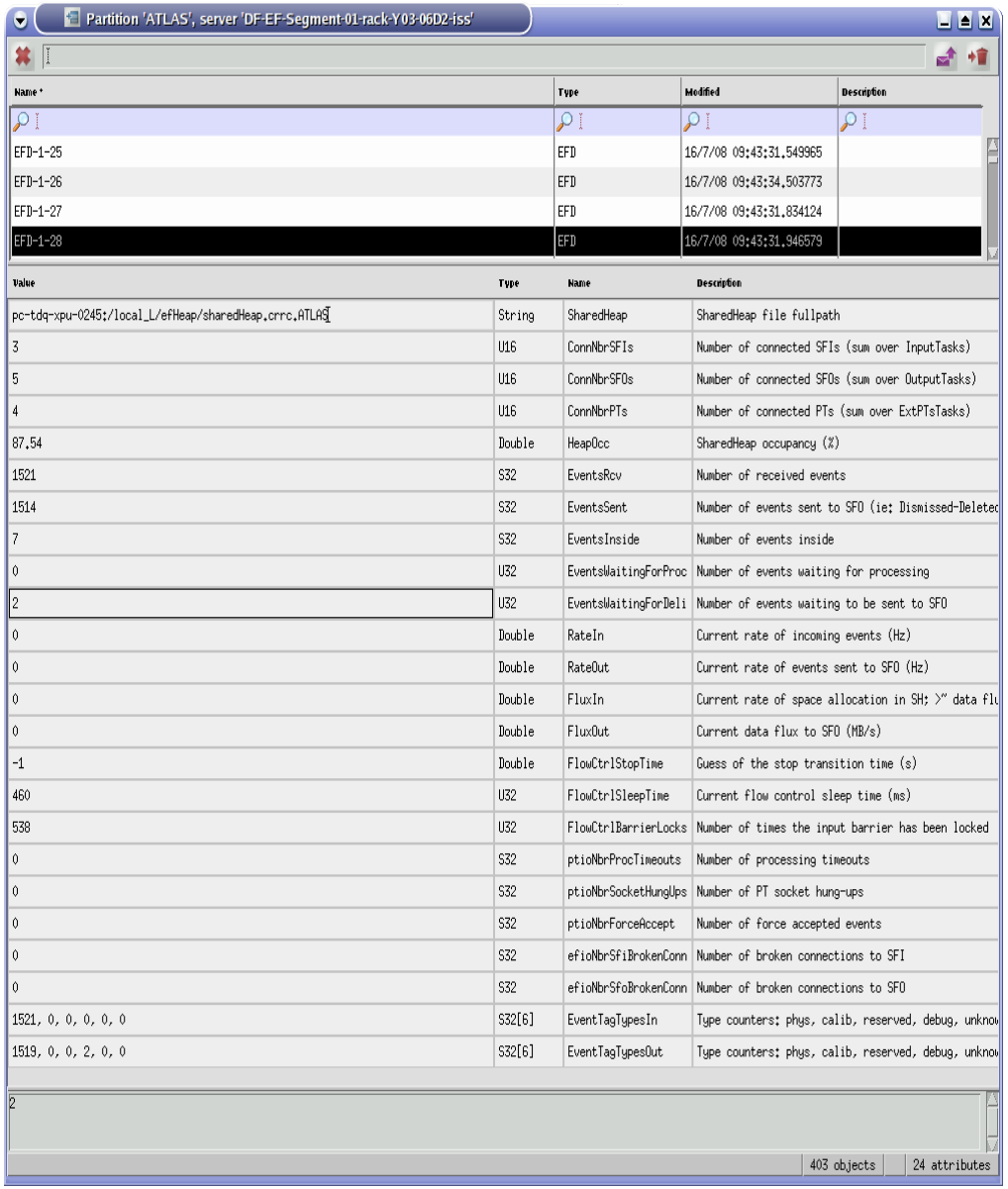


Monitoraggio Online

Information Service







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 a 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

Offline

O(1 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 !

(N.B.: il paradigma della Grid è ancora più forte)

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

Organizzazione Virtuale (ATLAS)

user → GANGA (frontend) → GLITE-WMS (backend)

→ Risorse fisiche (GRID)

** L'implementazione può variare da sito a sito **

"Ganga allows for the specification, submission, bookkeeping and post-processing of computational tasks on a wide set of distributed resource"

Workload Management System (WMS):

responsabili della distribuzione e gestione di processi sulle risorse fisiche Grid

Al livello hardware:

Allocazione ~ dinamica delle risorse di calcolo (CPU)

Allocazione ~ statica dello storage

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)

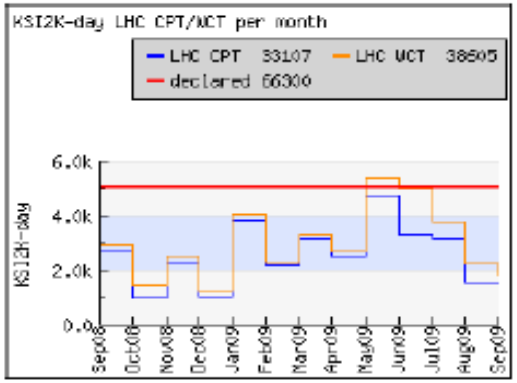
+ molti anni uomo di sviluppo sw (anche grazie a finanziamenti europei)

Il Portale di Monitoring

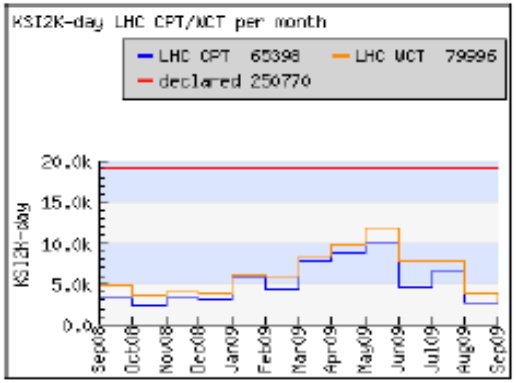
HLRmon Distributed Grid Accounting System

Uso CPU T2 ATLAS

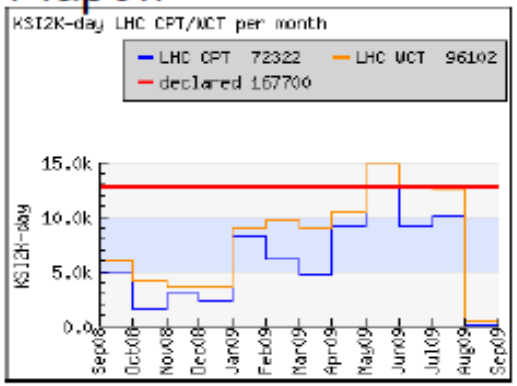
"Frascati"



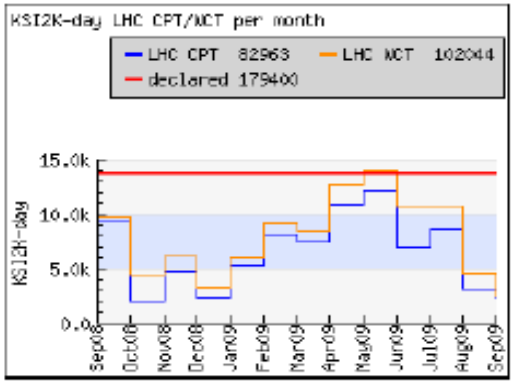
Milano



Napoli



Roma

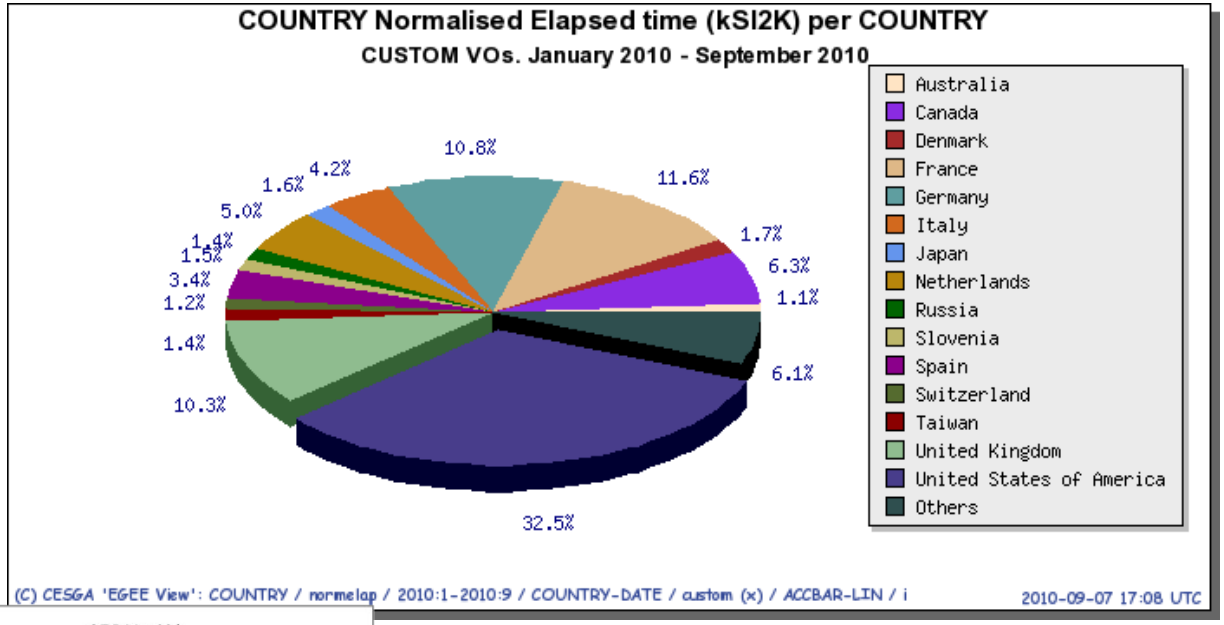


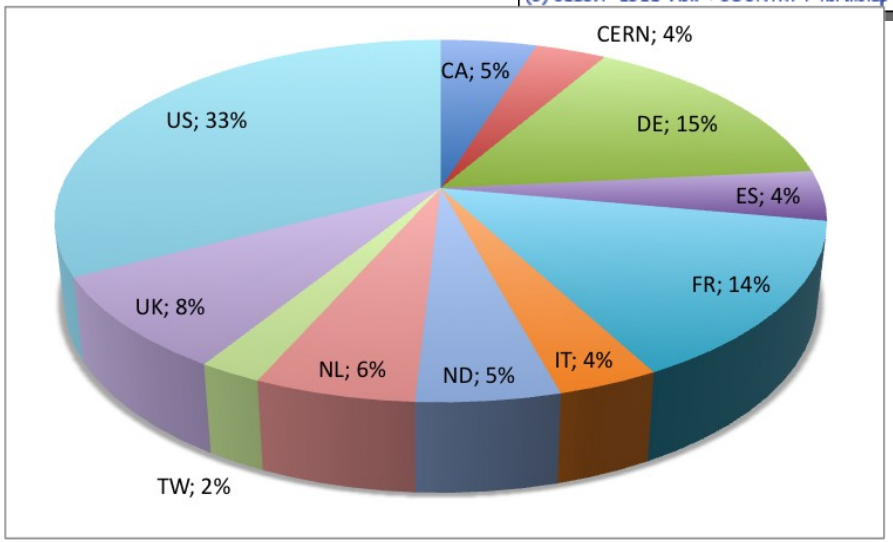
Uso risorse

VO ATLAS only

Uso delle CPU nella Grid per "Country" nei Tier1 e Tier2

(EGEE portal)

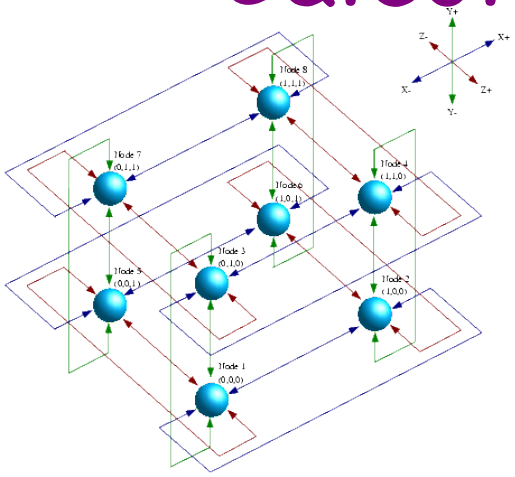




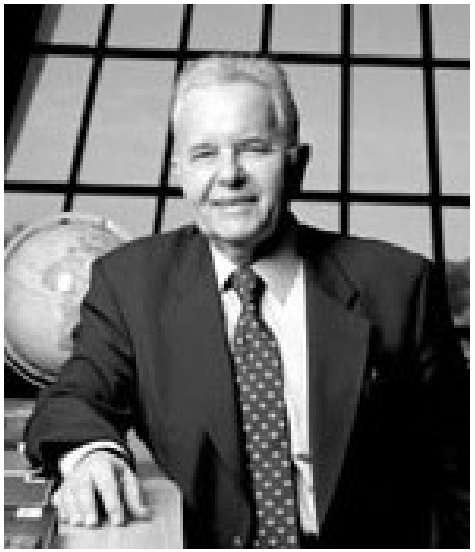
Numero totale di successful job (Analisi + Produzione):
Aprile – Settembre 2010

(ATLAS dashboard)

Calcolo Parallelo …



dalla Fisica Teorica
al Super-Computing
ovvero il progetto APE



N. Cabibbo

Progetto INFN, ora collaborazione con DESY Zeuthen e Université Paris-Sud 11

	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

Conclusioni

- tutta la nostra attività di ricerca si fermerebbe in questo momento senza Linux + mondo open source (GNU !)

- non mette limiti alla possibilità di sviluppo di soluzioni ad hoc (… di cui ogni tendiamo ad abusare)

- formidabile piattaforma "educativa"

- è un problema diverso, ma anche nella pubblicazione dei risultati (articoli) la politica "proprietaria" sta per essere abbandonata (circolazione riviste a pagamento limitata al primo mondo!) → "Open Access"

Presa Dati 2010 (1 evt = 1.5 MB)