Il nuovo cluster INFN per la fisica teorica

Incontro di lavoro della CCR – Napoli, 26 Gennaio 2010

Roberto Alfieri

Necessità di calcolo del Gr. IV

Attività nelle aree scientifiche:

Fenomenologia delle particelle elementari, Fisica teorica nucleare, Meccanica statistica, Turbolenza , Sistemi complessi, Biologia quantitativa, teoria di campo, teoria di Gauge su reticolo, ..

20% delle attività : simulazioni su reticolo (elevato parallelismo locale)

Necessita' tipico Job (Esempio SuberB):

2010: *O*(TFlops-year) - 2011: *O*(10 TFlops-year) - 2012: *O*(PFlops-year)

Soluzione CSN4: PROGETTI APE (attualmente ApeNext 12 TFlops)

80% delle attività : calcolo "generale"

Numerosi Job sequenziali o paralleli (multicomputer e/o multicore)

Necessita' tipo job < *O*(100 Gflops-year)

Strumenti Software: tools e librerie per il calcolo scientifico e simbolico (Mathematica, Maple, Matlab, NAG , GSL, GMP, ..)

Soluzione CSN4 : CLUSTER di PC

CLUSTER DI PC FINANZIATI DA CSN4

→ 2005 : Numerose farm o cluster MPI medio piccoli distribuiti diverse sedi
- spesso al di fuori dei centri di calcolo

2005-2006 : Centralizzazione al CNAF di un cluster MPI (CusterQuarto)
- 24 dual Xeon, Infiniband

2007-2008 : 4 PC cluster (BA, CT, MIB e PI) su proposte consorziate
- apertura a InfnGrid (VO theophys)

2010 → : Richiesta alla Presidenza un supporto per il finanziamento di un nuovo cluster centrale (CSN4cluster)

CSN4CLUSTER: LA SEDE

Ottobre 2009: Chiesta la collaborazione alla CCR per la definizione dei requisiti da inviare alle sedi candidate

Novembre 2009: Inviato "Call for proposal" alle Sezioni INFN. Requisiti:

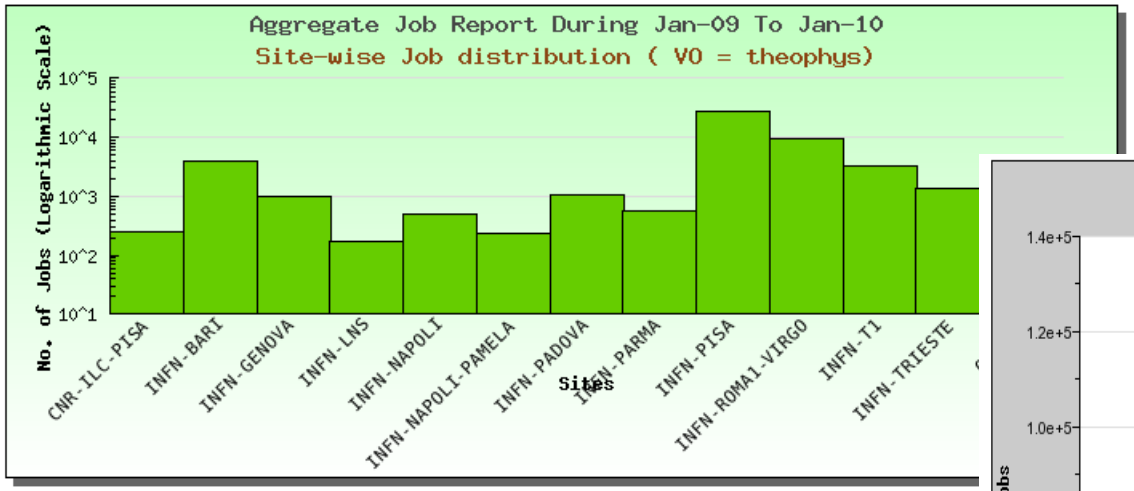
- Infrastruttura: procedura già utilizzata per i Tier2 (descrizione dei locali, impianto elettrico e condizionamento, personale tecnico e scientifico, costi di installazione e di esercizio)
- Hardware: 5Tflops di picco, nodi multicore x86_64, almeno 1 GB per core
- Network: Rete a bassa latenza (Infiniband o equivalenti)
- Storage: almeno 10TB condivisi tra i nodi di calcolo.
- Software: Linux x84_64 supportato da InfnGrid, librerie MPI-2
- Accessibilità: via InfnGrid (job sia sequenziali che paralleli)

Dicembre 2009: Tre delle proposte ricevute sono state prese in considerazione

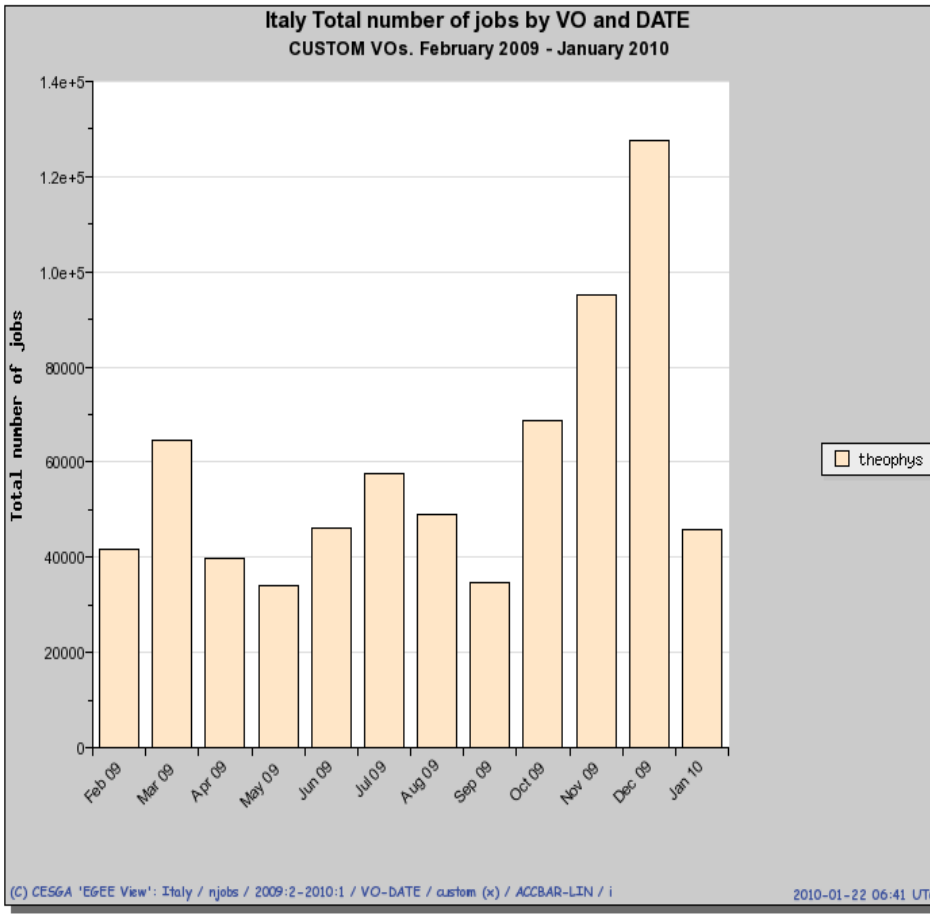
Gennaio 2010 : Richiesti alle sedi (BA, NA, PI) chiarimenti e ulteriori dettagli

Febbraio 2010: Scelta finale

Fisica teorica in Infn-Grid: THEOPHYS



http://gridview.cern.ch/GRIDVIEW/dt_index.php



https://www3.egee.cesga.es/gridsite/accounting/CESGA/egee_view.php

MPI/GRID NELLA FISICA TEORICA

I teorici utilizzano i cluster di PC sia con job sequenziale che paralleli, ma attualmente i job paralleli vanno **quasi esclusivamente su risorse locali**.

Possibili motivazioni nella preferenza di risorse locali per MPI:

Migliore conoscenza dell'ambiente hardware, software e della gestione:
possiamo accordarci con gli amministratori per adattarlo alle nostre esigenze
configurazioni specifiche, supporto per accesso e programmazione

Accesso : disponibilità di facility di scheduling (Reservation, Check-Pointing)
maggiori sicurezze sul workflow del Job (tempi di attesa, job aborted, ..)

Carenze di MPI/GRID :
limitazioni nell'ambiente software e nel middleware
documentazione, supporto tecnico
supporto 64 bit
affidabilità

MPI/GRID RELIABILITY

VO THEOPHYS con supporto MPI-START

Tabella presentata a Palau – maggio 2009

04/12/08	10/16	mpich1.2.7: 10 openmpi-1.2.6: 1	LSF: 4 PBS: 6	Inifinband: 1	OK: 3 SL64/mpi32: 4 no-gcc: 1 no-hostbasedauth: 1 aborted:1
19/02/09	9/16	mpich-1.2.7: 9 openmpi-1.2.6: 1	LSF: 4 PBS: 5	Infiniband: 2	OK: 3 SL64/mpi32: 2 no-gcc: 1 no-hostbasedauth: 1 aborted: 2
21/04/09	10/16	mpich-1.2.7: 10 openmpi-1.3: 1	LSF: 6 PBS: 4	Infiniband: 3	OK: 4 SL64/mpi32: 3 no-hostbasedauth: 1 aborted: 2

Test SAM (https://lcg-sam.gridops.org:8443/sam/sam.py)

Al 21/01/2010 su 46 risorse nazionali che supportano MPI solo 16 sono OK

PERCHE' MPI/GRID SU CSN4CLUSTER?

Grid e' una scelta INFN

- molte risorse disponibili, punti di accesso, prospettive future

Ottimizzazione delle risorse HPC

- Contenimento dei costi

Standardizzazione

- procedure di accesso, programmazione, condivisione esperienze

Occasione concreta per far funzionare MPI in INFN-Grid

Ci sono rischi da evitare. Ad esempio:

Utilizzo esclusivo di job sequenziali.

- Spreco acquisto delle rete a bassa latenza e manpower di supporto specializzato (vedi cluster Quarto al Cnaf).

Starvation dei job (sequenziali e paralleli)

- gli slot riservati per i job paralleli rimangono inutilizzati

...

MPI IN INFNGRID (THEOPHYS): SPUNTI PER LA DISCUSSIONE

Discriminazione job sequenziali e paralleli:

- gruppi/ruoli VOMS, ApplicationSoftwareRunTimeEnvironment=”PARALLEL” (altro?)
- **Griglia separata per le code parallele** (Bencivenni - Aiftimiei) :

Code parallele separate mediante l'attributo GlueCESiteStatus =“PARALLEL”

Scheduling: Pool di WN condiviso (completamente o parzialmente) con preemption dei job sequenziali (o trasferiti su altri core in overbooking?)

- Disponibilità effettiva di slot per coda?

Numero di siti limitato (almeno inizialmente) e con configurazioni concordate e omogenee (es: csn4cluster + altri cluster Consorziati)

Altri problemi:

- **Quale Software:** x86_64, flavour MPI, compilatori, librerie
- **LowLatency Network:** Ethernet, Infiniband, 10Gb Eth?, ..

Formazione: Organizzato nel 2010 a Catania corso per Fisica Teorica

Commenti / domande?