Uso di Grid per il calcolo parallelo:

Il progetto TheoMPI per applicazioni di Fisica Teorica

Roberto Alfieri [1] , Silvia Arezzini [2] , Alberto Ciampa [2] , Roberto DePietri [1], Enrico Mazzoni [2]

[1] *INFN Parma e Università di Parma*

[2] *INFN Pisa*

Sommario

La **comunità dei Fisici Teorici** dell'INFN (gruppo IV) e le **necessità computazionali**

La Grid come soluzione per agevolare l'accesso distribuito a facility di calcolo massivamente parallele distrubuite su rete geografica

Lo stato del **supporto del calcolo parallelo in Grid** (Egee/EGI)

Le problematiche affrontate e le scelte effettuate nella configurazione del **CSN4cluster** in Grid

Esempi di **applicazioni di fisica teorica** eseguiti su CSN4cluster

Risultati e sviluppi futuri

Gruppo IV INFN

Aree di ricerca

Il Gruppo 4 dell'INFN svolge attività di ricerca in **Fisica Teorica.**

54 progetti di ricerca, denominati **Iniziative Specifiche** (IS), nelle seguenti aree:

- String & Field Theory:
QFT, Strings & M-Theory, Gravity, Lattice Gauge Th. and confinement, AdS/CFT;

- Particle Phenomenology:
SM and BSM (Susy, Extra Dim., Composite Higgs), QCD at colliders (MC simulations, finite T and μ), Flavor Physics (and lattice) and EFT for heavy flavors, AdS/CFT and QCD;

- Hadronic & Nuclear Physics:
nuclear structure and reactions (radioactive beams, stability valley and beyond), heavy ion collisions (quark-gluon plasma, saturation, jet quenching, T and μ phase transitions), confined hadronic matter (spin structure of hadrons, exotic spectroscopy, GPD);

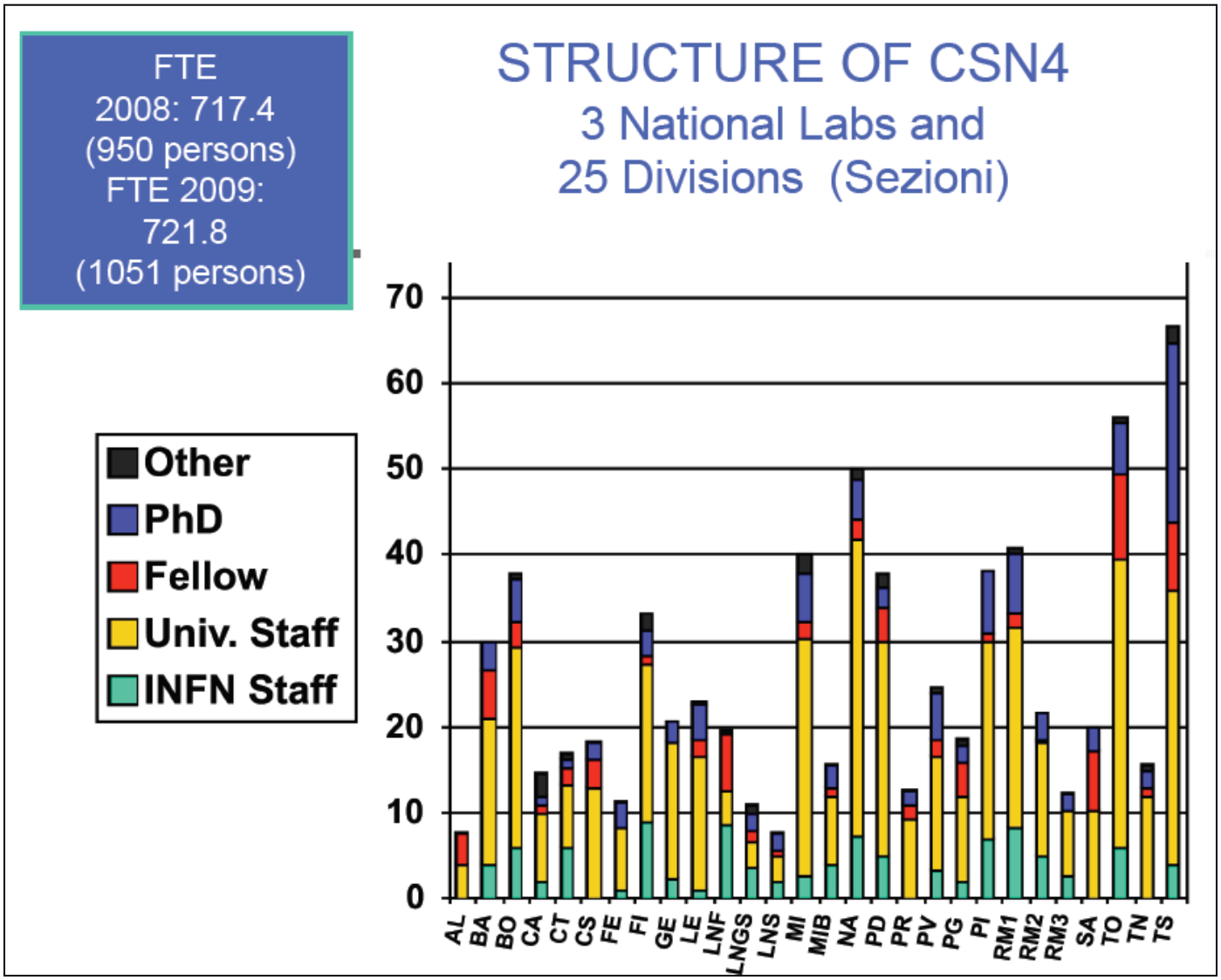
- Mathematical Methods:
general relativity (gravitational waves,..), quantum theory (foundations, chaos,..), conformal field theories;

- Astroparticle & Cosmology:
neutrino physics, "dark things" (matter, energy,..), astrophysical radiation sources, astronuclear physics, gravitational waves

- Statistical Field Theory & Applications:
complex & non-eq. systems, spin glasses, applications (quant. biology, turbulence,..).

Struttura della CSN4

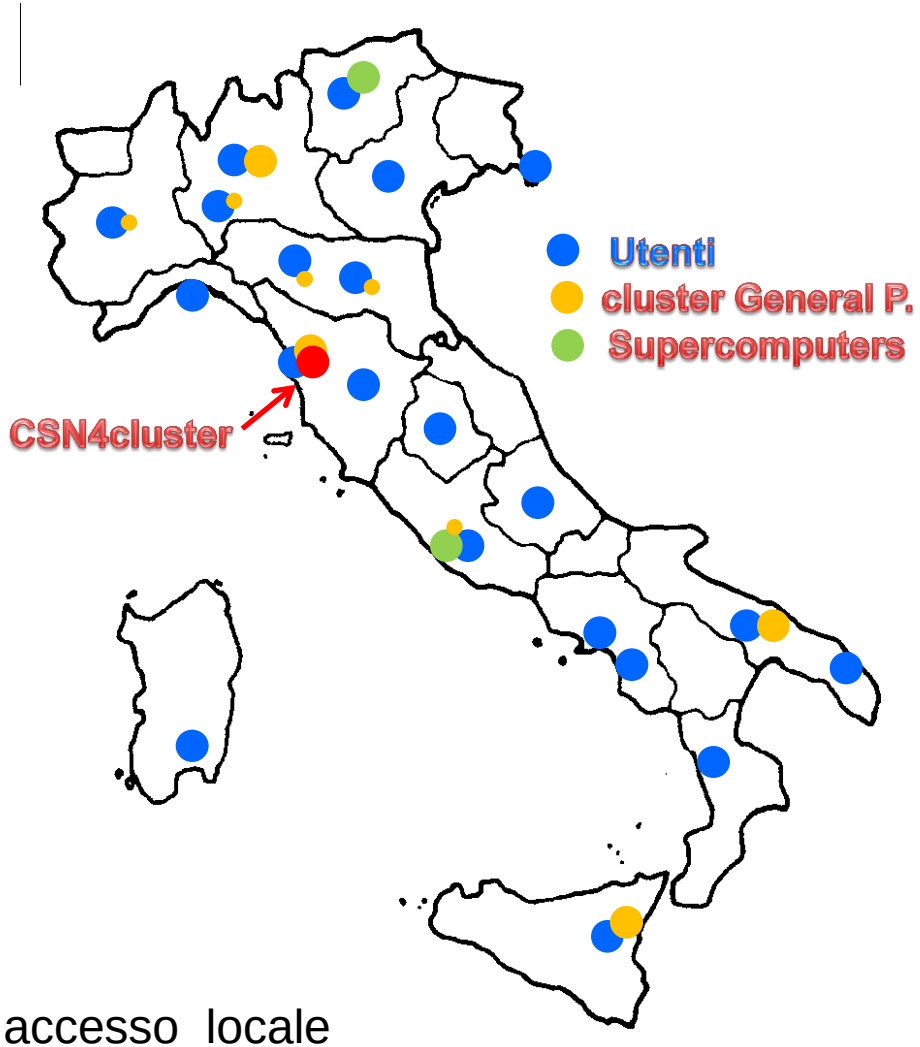
Le attività sono coordinate da Commissione Scientifica Nazionale 4 (CSN4) dell'INFN:



Necessità computazionali

SuperCalcolo:

- Simulazioni su reticolo
  (LGT, fisica gravitazionale, turbolenze, ..)
- Tipico Job: 2011 *O*(10TFlops-year)
- Risorse: Supercomputers famiglia **APE**



General Purpose:

- prevalentemente calcolo numerico
- Molti job seriali e paralleli
- circa 400 ricercatori coinvolti
  distribuiti su tutto il territorio nazionale
- Risorse di calcolo:
  -- In Passato: PC cluster medio-piccoli ad accesso locale
  -- Dal 2010 **progetto TheoMPI (CSN4cluster)**

Il progetto TheoMPI

Obiettivi

- Realizzare una infrastuttura nazionale distribuita per il calcolo in grado di soddisfare le necessita' di calcolo di gr. IV seriale **e parallelo** (per il momento solo Gen. Purpose)
- Condividere le risorse per ottimizzare e razionalizzare l'utilizzo
- Uniformare policy di Autenticazione e Autorizzazione
- Rendere trasparente accounting e monitoring

La Grid Italiana (IGI)

- 58 siti, 26Kcore, 25PB disk, 1100 utenti , middleware gLite
- Fornisce Autenticazione (certificati X509), Autorizzazione (VOMS) Accounting (HLRmon), Gestione/Supporto (Operation Center) ,…
- **Supporto per il parallelismo carente**; utilizzo quasi nullo.

  Diversi Working Group MPI attivati da EGEE per trovare soluzioni.

Il progetto

- Inizialmente unico cluster centrale (**CSN4cluster**)
- Utilizzo di Grid per job seriali e, **soprattutto,** paralleli (VO **theophys**)
- Progressiva integrazione/estensione delle altre risorse/comunità MPI

CSN4cluster
timeline

Dic 2009: definiti i requisiti del cluster

Feb 2010: approvata proposta INFN-Pisa

Giu 2010: cluster operativo per i job seriali

Lug 2010: call per proposal scientifici 2010-2011

Richieste: 130K day*core seriale + 250K day*core parallelo = 380K day*core

Set 2010: approvati 16 progetti e assegnate le quote di fair-share

Dic 2010: cluster operativo (sperimentalmente in grid) per job paralleli

Ott 2011: consuntivi utilizzo 10-11, call per proposal scientifici 11-12

Nov 2011: approvazione progetti e definizione nuove quote di fair-share

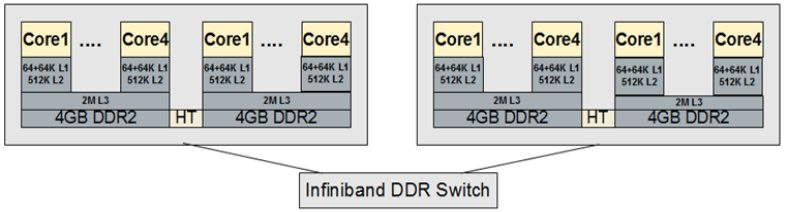
CSN4cluster
progetto tecnico

Il cluster è installato, configurato e manutenuto dallo staff del servizio di calcolo scientifico di **INFN-Pisa**

Computing:

128 WNs Opteron 2x4 cores, SL5/x86_64, openMPI
1024 total cores, 10TFlops peak perf

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



High Speed Network:

Infiniband DDR

Storage (30TB GPFS/InfiniBand):

home-dir condivisa tra i nodi di calcolo

1 SE gridsrm.pi.infn.it : Storm

Dettagli: ***http://wiki.infn.it/strutture/pi/datacenter/cluster_gruppo_iv/csn4cluster***

Stato attuale del supporto (gLite middleware)

L'utente Grid sottomette il proprio job accompagnato dal file JDL (Job Descriptor Language) in cui specifica le risorse richieste.

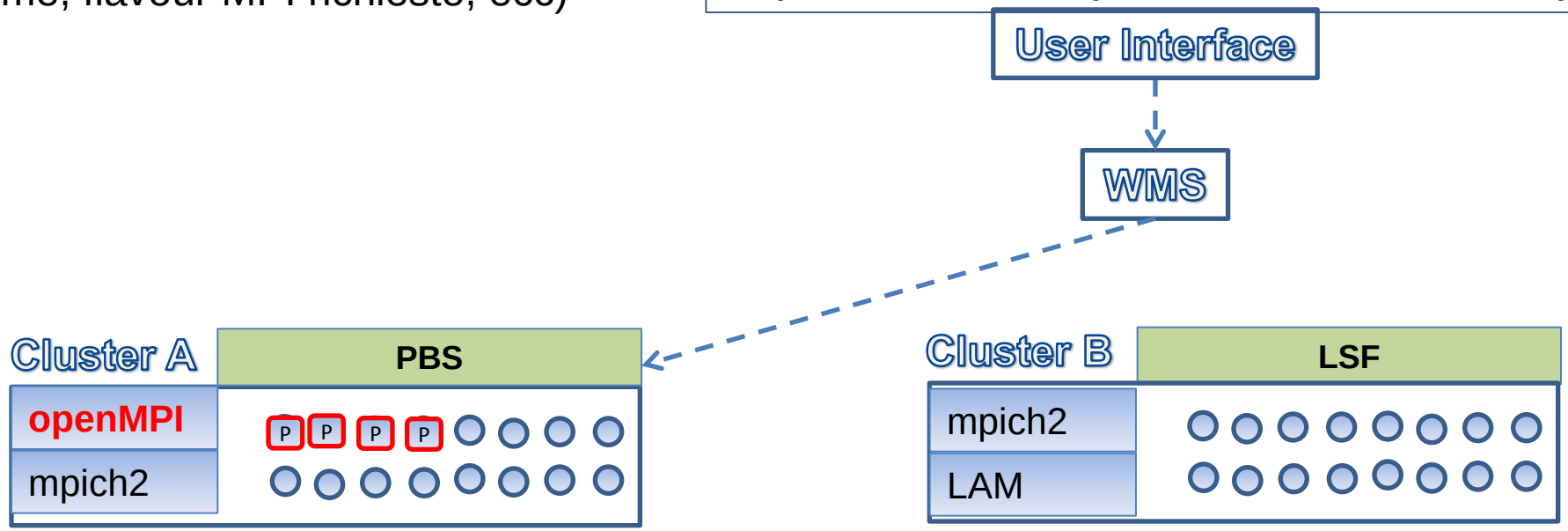
Il JDL viene elaborato dal WMS il quale individua la risorsa piu' adatta.

Se il job e' MPI occorre specificare il **numero di CPU** richieste.

Il wrapper **MPI-start** ha il compito di rendere traparente all'utente il contesto di esecuzione (shared home, flavour MPI richiesto, ecc)

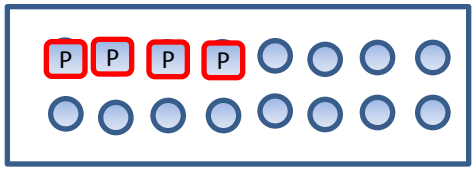
JDL file

```
CPUnumber = 4;
Executable = "mpi-start-wrapper.sh";
Arguments = "my-mpi-prog";
InputSandbox = "mpi-start-wrapper.sh  my-mpi-prog";
….
Requirements=Member("OPENMPI", other.GlueHostAp
```

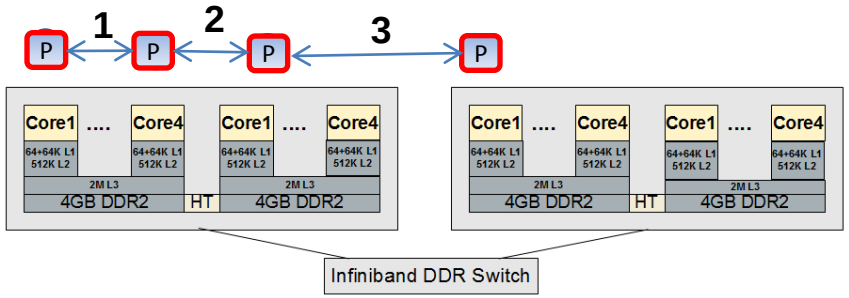


Tipi di comunicazioni nei moderni nodi di calcolo

Modello di cluster nell'attuale middleware Grid.



Nei moderni cluster i processori sono multicore e l'interazione tra processi avviene con almeno 3 livelli di comunicazione:



Measured network performance (on CSN4cluster, using NetPIPE):

	Comm Type	Comm. device	Latency	MAX Bandw.
1	Intra-socket	SHared Mem - L3 Cache or DDR	640 ns (DDR)	14 Gb/s
2	Intra-node	SHared Mem - NUMA (HT or QPI)	820 ns	12 Gb/s
3	Extra-node	Infiniband	3300 ns	11 Gb/s

Il middleware Grid deve supportare le novita' introdotte dall'architettura multicore:

-- **parallelismo multi-thread** (per sfruttare le comunicazioni Shared Memory)

-- **CPU/memory affinity** (è la capacita' di legare un processo ad uno specifico core o banco di memoria, per sfruttare l'effetto di Cache e limitare le comunicazioni NUMA)

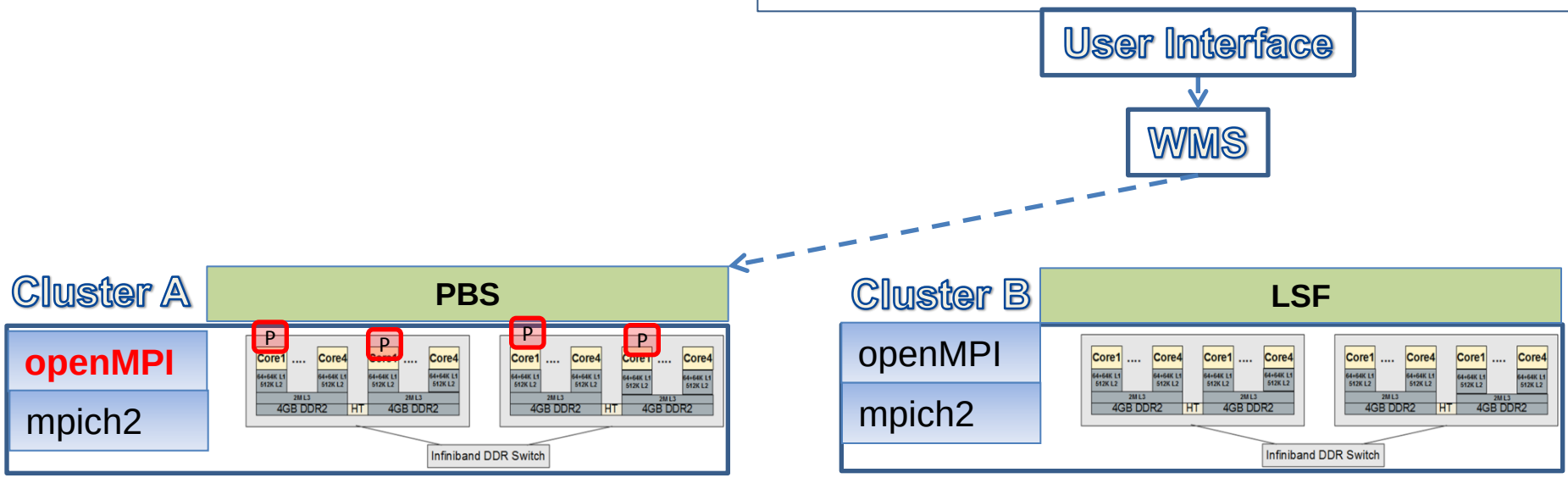
Nuovo supporto (EMI middleware)

Il nuovo middleware EMI supporterà la granularità, la programmazione multithread e la CPU/memory Affinity con l'introduzione **di nuovi attributi specifici nel JDL** (Hostnumber, WholeNodes e SMPGranularity) e una **nuova versione di Mpi-start.**

Queste nuove funzionalità non sono disponibili nell'attuale gLite middleware, ma un **versione sperimentale (development) è installata e operativa sul CSN4cluster.**

JDL file

```
#CPUnumber=4;
Wholenodes=true;
HostNumber=2;
SMPGranularity=8;
Executable= "mpi-start";
Arguments="-t openmpi -psoket -- my-prog";
InputSandbox="my-prog";
….
Requirements=Member("OPENMPI", other.GlueHostAp
```

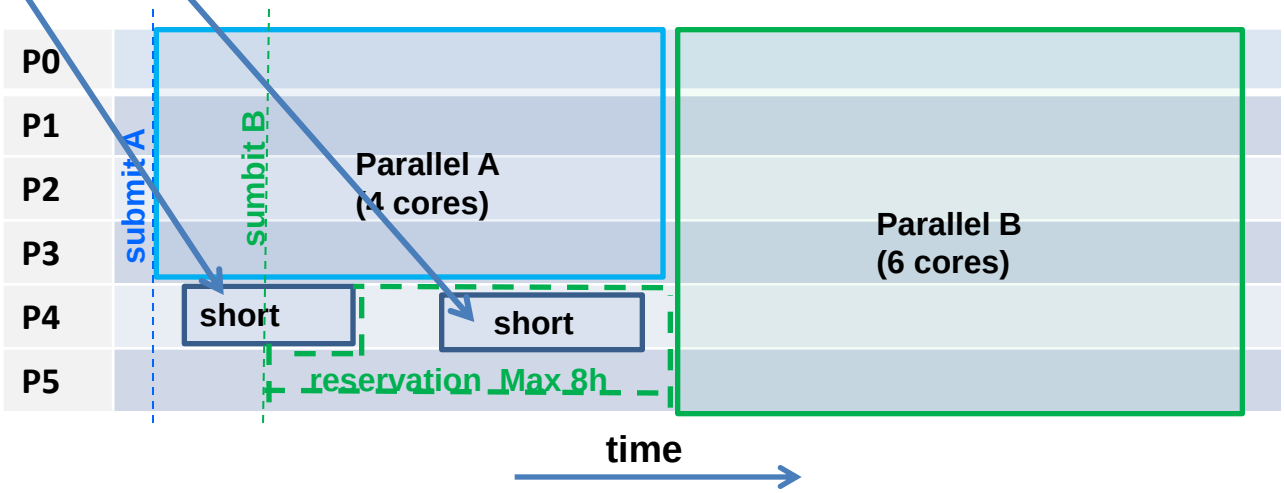


organizzazione delle code

Caratteristiche delle code sulla partizione parallela:

- **theompi** : solo job paralleli, runtime 72h, max 8h di **Reservation time** sui core liberi
- **theoshort** : job corti, runtime 4h (piu' corto del reservation time)

- La coda "theoshort" consente di sfruttare i core quando non sono utilizzati dalla coda parallela
- La tecnica **di Backfill** consente inoltre di sfruttare anche i core riservati dalla coda parallela



Autenticazione e autorizzazione

Il metodo di **autenticazione** si basa sul certificato X.509 rilasciato dalla CA INFN, che l'utente utilizza al momento dell'apertura della sessione con il comando

voms-proxy-init --voms theophys

L'autorizzazione è realizzata utilizzando **gruppi e ruoli:**

un gruppo VOMS è stato creato per ogni Iniziativa Specifica e assegnato ai membri della IS: **/theophys/IS_<nomeIS>** *esempio /theophys/IS_OG51*

La definizione di questi gruppi all'interno del JobManager (LSF) consente di assegnare le policy di FairShare a livello di Gruppo.

Per evitare l'accesso dei job seriali sulle code parallele abbiamo introdotto un ruolo specifico**:** **role=parallel** che l'utente autorizzato deve presentare alla risorsa

L'utente dovrà richiedere al VOMS il gruppo e il ruolo appropriato al momento dell'apertura della sessione grid. Esempi:

voms-proxy-init --voms theophys:/theophys/IS_OG51 #seriale

voms-proxy-init --voms theophys:/theophys/IS_OG51/Role=parallel #parallelo

Storage

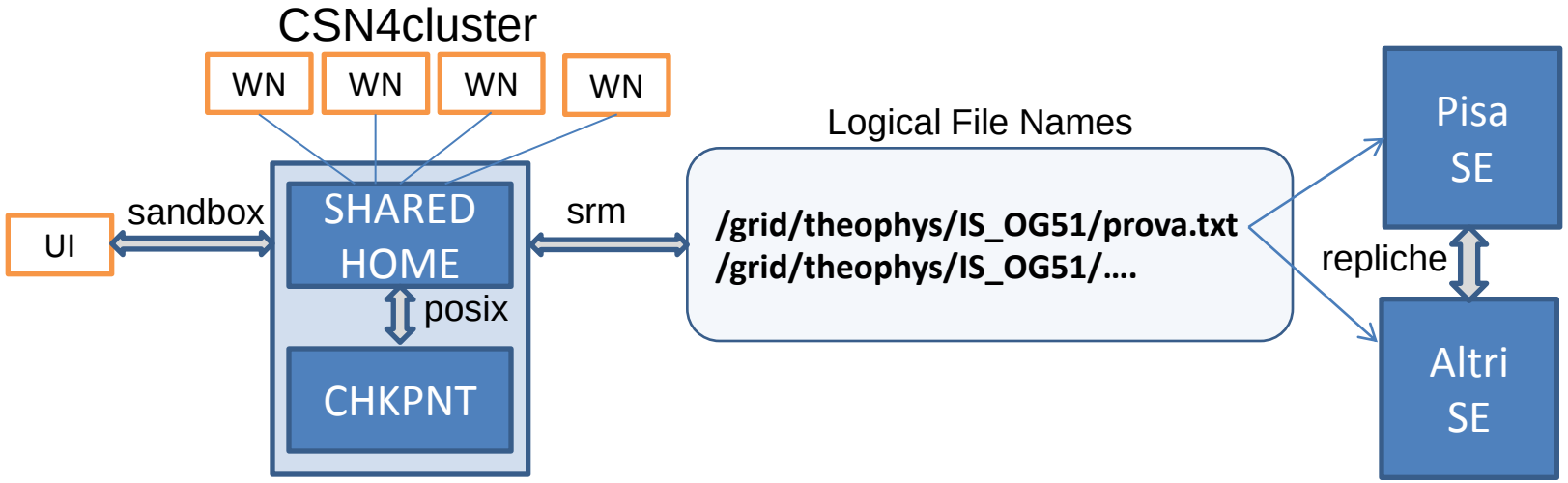
L'architettura dei dati in grid si basa sugli **Storage Element (SE)** in cui i dati vengono referenziati attraverso un nome logico e possono essere replicati per ridondanza o per avvicinarli al sito di elaborazione.

L'Input/output può avvenire anche direttamente dalla **User Interface (UI),** utilizzando lo strumento **Sandbox**, se il volume dei dati è limitato.

Organizzazione dell'area di Storage disponibile a **INFN-Pisa**:

- **HOME :** spazio di lavoro condiviso tra i nodi
- **SE :** area dati Grid
- **CHKPNT :** area di salvataggio per i CheckPoint e per i dati tra un run e il seccessivo

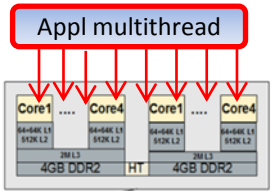
(i job paralleli possono avere una lunga durata (>10 giorni) e produrre grandi data-set (>100GB))



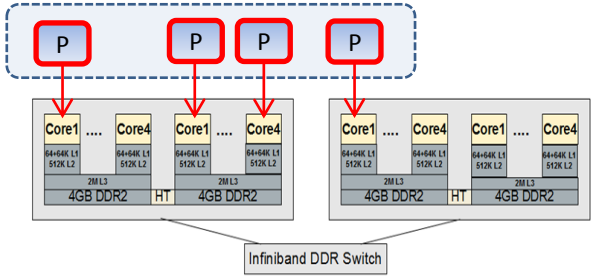
Applicazioni parallele

Le applicazioni parallele che possiamo eseguire sul cluster sono quindi di diverso tipo:

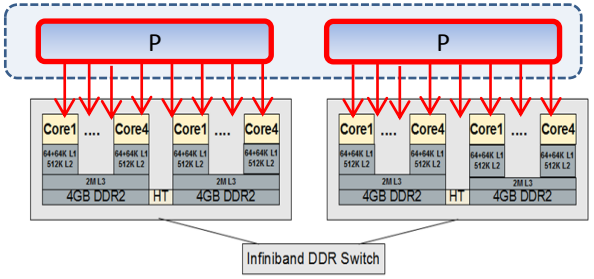
-- Multithread: sfruttano tutti i core di un solo nodo (un thread per core)



-- MPI puro : distribuite su più nodi (un processo per core)



-- Ibrido: combinazione dei due tipi precedenti

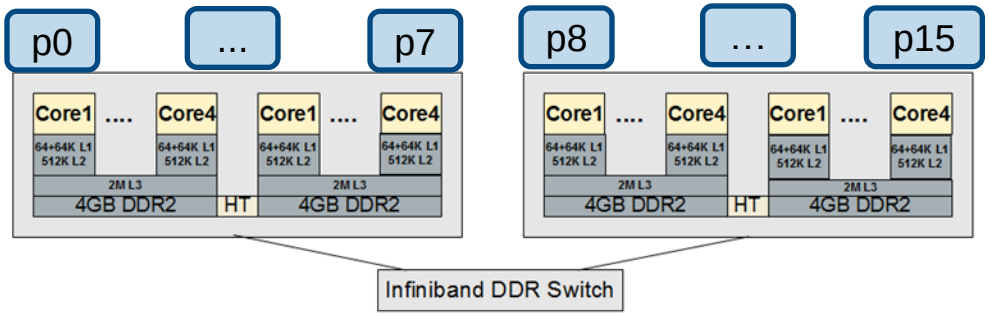


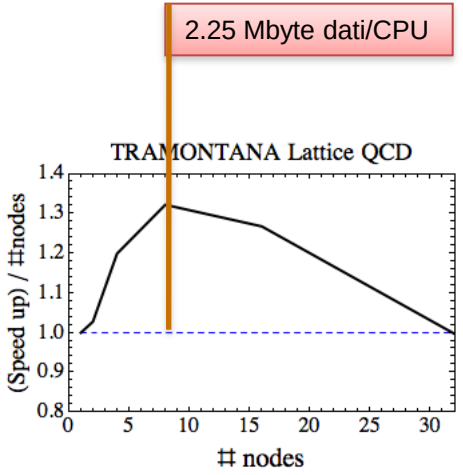
Applicazioni parallele
Chroma Library

Simulazione Hybrid-Montecarlo di Pure Gauge SU(3) su un reticolo 32x32x32x8 (2000 sweep) usando la libreria pubblica “Chroma” della **USQCD collaboration** (http://usqcd.jlab.org/usqcd-docs/chroma/).

- Codice puro MPI
- Occupazione totale di memoria del reticolo **~36MBytes**
- Importanza della **memory affinity** (quando tutti i dati non sono in cache)
- Effetto della Cache (efficienza >1)

Np	**8** (1x8)	**16** (2x8)	**32** (4x8)	**64** (8x8)	**128** (16x8)
Non-ranked	295 min	146 min	62 min	27 min	14 min
Ranked	287 min	139 min	59 min	27 min	14 min

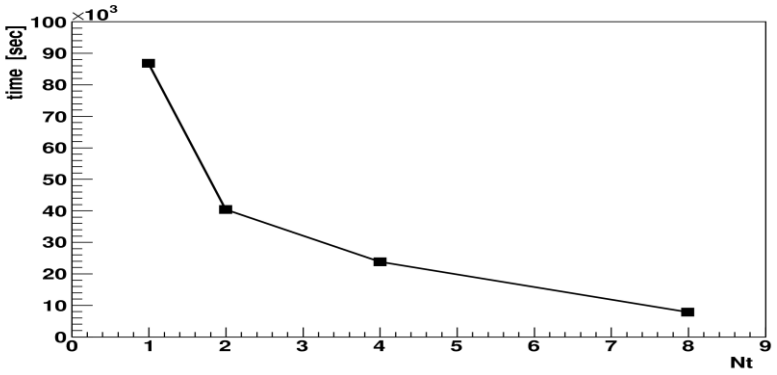




Grazie a A. Feo (Parma Univ.)

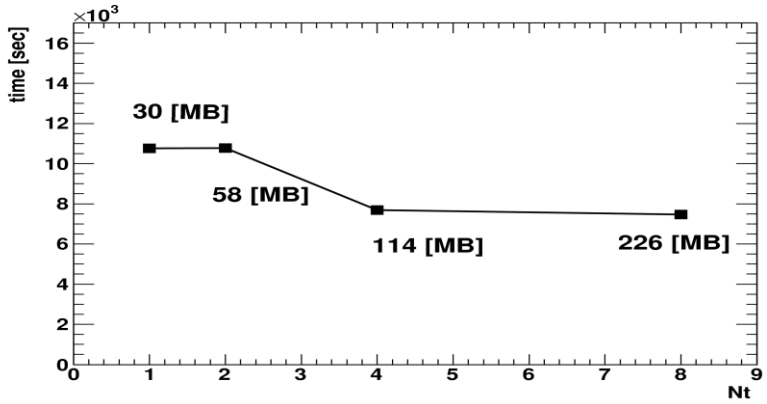
Teoria delle perturbazioni stocastiche numeriche

La Numerical Stochastic Perturbation Theory viene introdotta per eseguire calcoli perturbativi di ordine elevato nella teoria di Gauge su reticolo e integra numericamente le equazioni differenziali della Teoria di Perturbazione stocastica.
Il codice è stato scritto a Parma e supporta la parallelizzazione **openMP.**



Dimensione fissa del volume (16x16^3 ~ 226 MB) e numero crescente di threads

- Il tempo di esecuzione atteso e' ~ 1/Nt



Il volume V cresce con il numero di thread Nt
V = (Nt x 2) x 16^3.

- Il tempo di esecuzione atteso è costante

Grazie a M. Brambilla and M. Hasegawa (Parma Univ.)

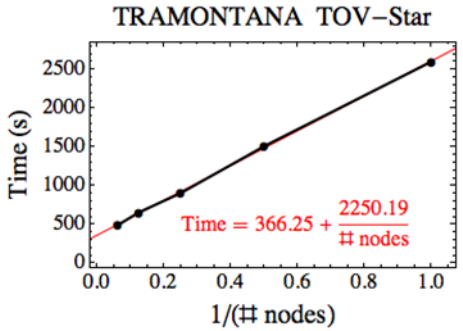
Applicazioni parallele
Einstein Toolkit

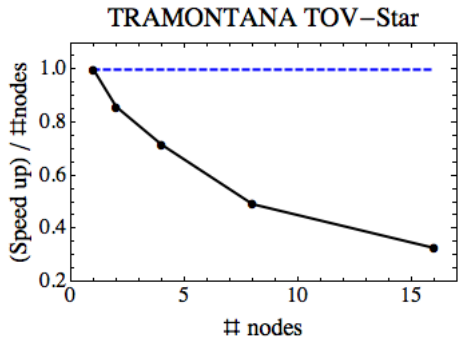
Evoluzione di una stella di Neutroni isolata usando il codice dell'**Einstein Toolkit consortium** (http://einsteintoolkit.org/).

Hydro-dynamical Simulation of a perfect fluid coupled to the full Einstein Equations (dynamical space-time) on a 3-dimensional grid with 5-level of refinement spanning an octant of radius 177 km with a maximum resolution within the star of 370 m.

- **Codice molto complesso** (F90, F77, C++, C sviluppato da piu' di 100 programmatori)
- **Hybrid MPI-openMP** (parallelizzazione parziale)
- Occupazione complessiva di memoria del reicolo **~8GByte**.

#node	Np=8x#	Np=4x#	Np=2x#	Np=#	Np=2x#
	Nt=1	Nt=2	Nt=4	Nt=8	Nt=4 (rank)
1	2291.90	2934.21	3126.73	3360.96	**2608.08**
2	1438.72	1619.83	1797.30	2061.55	**1516.04**
4	1007.71	993.79	1007.71	1268.79	**909.36**
6	767.45	783.07	694.31	927.35	**745.63**
8	663.03	638.81	694.31	753.79	**661.37**
16	461.85	448.77	484.20	552.89	**497.78**





Grazie a R. De Pietri (Parma Univ.)

Risultati

Dall'inizio dell'anno CSN4cluster e' in produzione per **job seriali, multithread, Mpi e ibridi.**

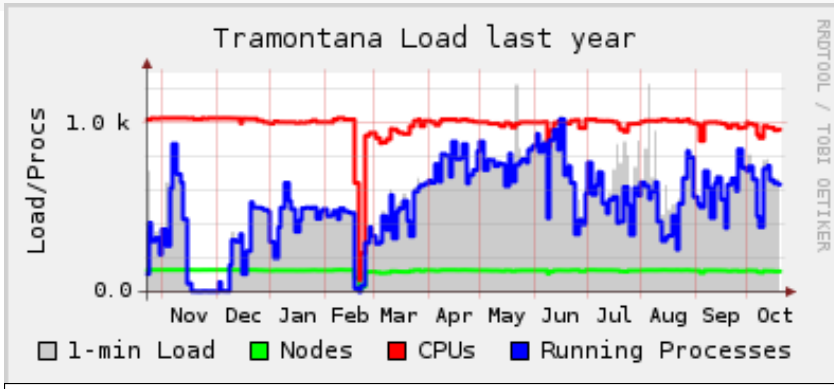
L'accesso avviene **unicamente via Grid,** anche se il middleware ancora non supporta ufficialmente tutti i tipi di job paralleli.

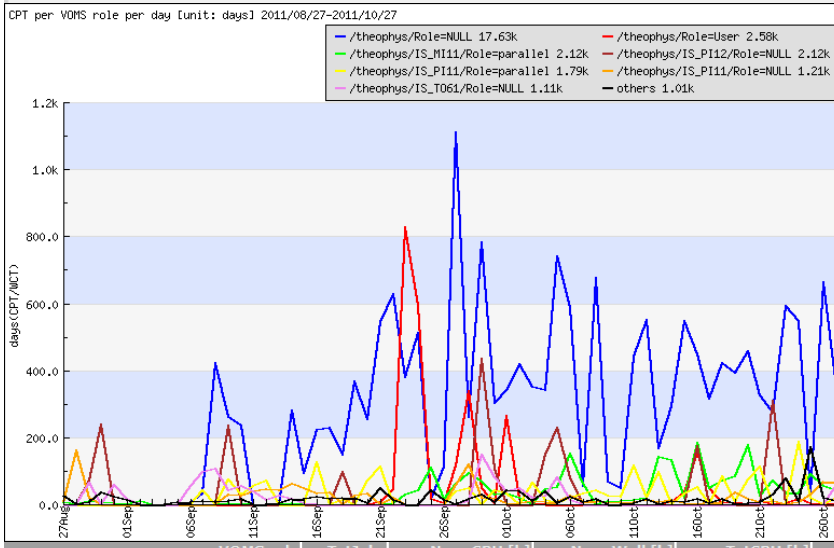
Le **policy di autorizzazione** per la gestione del parallelismo e dei gruppi sono impostate centralmente tramite il servizio VOMS.

Il cluster ha un **utilizzo attorno all'80%**

- La maggior parte dell'idle e' dovuta alla Reservation dei core dal parte di LSF.

Tramite il portale Web HLRmon
https://dgas.cnaf.infn.it/hlrmon/report/
è possibile consultare **l'accounting** di tutte le attività filtrate per sito, ruolo (seriale o parallelo), gruppo (Iniziativa Specifica), periodo di tempo e tipo di dato (numero job, CPU-time, ecc).





VOMS role	TotJobs	NormCPU [h]	NormWall [h]	TotCPU [h]
/theophys/IS_PI12/Role=NULL	39226	975787.94	1033691.51	487893.97
/theophys/Role=NULL	33402	2374458.90	2498682.60	1187229.45
/theophys/IS_PI11/Role=NULL	20816	426159.84	357943.60	213079.92
/theophys/IS_PR21/Role=NULL	10835	73616.76	93312.71	36808.38
/theophys/Role=User	6752	138151.81	185992.31	69075.90
/theophys/IS_TO61/Role=NULL	6154	184108.60	186744.16	92054.30
/theophys/IS_MI11/Role=parallel	1927	423857.29	93278.66	194215.19
/theophys/IS_PI11/Role=parallel	788	224309.71	47700.48	102611.33
/theophys/IS_OG51/Role=parallel	710	105417.21	17326.78	48201.74
/theophys/Role=parallel	240	13099.37	3699.51	5989.65
/theophys/IS_AD31/Role=NULL	63	1361.99	1369.96	680.99
/theophys/IS_MI41/Role=parallel	54	17768.56	2641.05	8124.62
/theophys/IS_TV62/Role=parallel	48	2247.63	343.19	1027.72

Sviluppi futuri

Installazione del **Middleware EMI** per avere il supporto completo ai nuovi attributi per la selezione della Granularity.

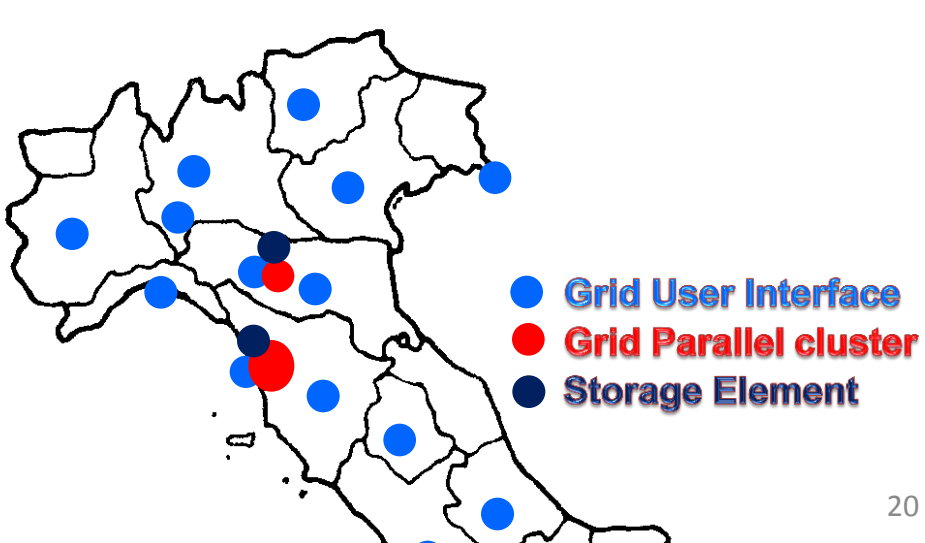
Introduzione di un **portale** per semplificare all'utente la gestione del ciclo di vita del job.

Ridefinizione **delle politiche di convivenza tra job seriali e paralleli** per massimizzare i tempi di utilizzo delle CPU

(ad esempio incrementando MAXreservation time e runtime della coda short)

La **scalabilità del modello** consente l'utilizzo della stessa interfaccia utente per l'accesso ad altri cluster paralleli dislocati in siti diversi della rete.

Per la verifica di questa architettura distribuita un secondo (piccolo) cluster parallelo e' stato installato a Parma con la stessa configurazione Grid del CSN4cluster.



Grazie per l’attenzione!