Survey MPI

Roberto Alfieri - Università di Parma & INFN, Gr.Coll. di Parma

Catania, 18 Maggio 2010

Sommario

- Survey sull’uso di MPI

- Stato dei cluster che supportano MPI e del nuovo cluster di gr. IV

- Report sulle attività dell’ MPI WG di Egee-III

- Sottomissione MPI in gLite: stato e piani futuri (Salvo Monforte)

Utilizzo di MPI: survey del 2009

Survey dell'MPI-WG di Egee del Maggio 09, rivolto a utenti del calcolo parallelo

- Dati presentati al WorkShop di Palau - 05/2009

- Feed back da molte discipline
  -- nell'ordine: fisici, chimici, biologi, ricerca medica, astronomi, geologi, ..
  -- Nell'INFN utilizzo prevalente nella Fisica Teorica

- Interesse principale per Mpi2, openMP (vs Mpich, unico flavour MPI in Egee)

- Risorse prevalentemente locali (60%) , poi provider commerciali e Grid

Utilizzo di MPI : aggiornamento

superB (e in generale calcolo sperimentale)

- Interesse per applicazioni parallele multi-thread su singolo nodo

EUIndia-Grid

- diverse applicazioni MPI (Moose, Quantum Montecarlo, Quantum Espresso)
- Interesse nel poter riservare interi nodi SMP. Sviluppo della patch "reserve_SMP_nodes": http://euindia.ictp.it/grid-tools-and-utilities/reserve-smp-nodes-1/

Fisica teorica (gruppo IV)

- Numerose applicazioni parallele (LQCD, turbolenza, gravitazione, ..)
- Sviluppo di sistemi paralleli dedicati (APE) per le simulazioni su reticolo
- Finanziamento di cluster tradizionali (eventualmente con MPI e Infiniband) per le altre applicazioni. I cluster piu' recenti aperti anche a Grid.

Risorse MPI: SAM tests

Attualmente SAM verifica la corretta configurazione di MPI-start ed esegue un test di compilazione ed esecuzione del programma testjob-mpi. SAM test del 16/5/2010:

No	RegionName	SiteName	NodeName	Status	ops	
					mpi-js	mpi
1	Italy	GILDA-INFN-CATANIA	grid010.ct.infn.it	OK	ok	ok
2	Italy	GILDA-PADOVA	gilda-01.pd.infn.it	OK	ok	ok
3	Italy	GRISU-COMETA-INAF-CT	inaf-ce-01.ct.pi2s2.it	OK	ok	ok
4	Italy	GRISU-COMETA-INFN-CT	infn-ce-01.ct.pi2s2.it	OK	ok	ok
5	Italy	GRISU-COMETA-INFN-LNS	infnlns-ce-01.ct.pi2s2.it	OK	ok	ok
6	Italy	GRISU-COMETA-ING-MESSINA	unime-ce-01.me.pi2s2.it	OK	ok	ok
7	Italy	GRISU-COMETA-UNICT-DIIT	unict-diit-ce-01.ct.pi2s2.it	ERROR	error	ok
8	Italy	GRISU-COMETA-UNICT-DMI	unict-dmi-ce-01.ct.pi2s2.it	ERROR	error	ok
9	Italy	GRISU-COMETA-UNIPA	unipa-ce-01.pa.pi2s2.it	OK	ok	ok
10	Italy	GRISU-ENEA-GRID	egce1-cresco.portici.enea.it	MAINT	error	ok
11	Italy	ICEAGE-CATANIA	iceage-ce-01.ct.infn.it	OK	ok	ok
12	Italy	INFN-LNS	grid-ce.lns.infn.it	OK	ok	ok
13	Italy	INFN-NAPOLI	griditce01.na.infn.it	OK	ok	ok
14	Italy	INFN-NAPOLI-ARGO	argoce01.na.infn.it	OK	ok	ok
15	Italy	INFN-PADOVA	prod-ce-01.pd.infn.it	OK	ok	ok
16	Italy	INFN-PADOVA	prod-ce-02.pd.infn.it	OK	ok	ok
17	Italy	INFN-TORINO	t2-ce-02.to.infn.it	OK	ok	ok
18	Italy	UNI-PERUGIA	cex.grid.unipg.it	OK	ok	ok

Risorse MPI in INFNGRID

SITO	MPI	Shared Home	Lrms	CPU
INFN-LNS	mpich	NO	PBS	15x2
INFN-NA	mpich 1.2.7	NO	PBS	34x4
INFN-NA-ARGO	mpich 1.2.7	NO	PBS	32x8
INFN-PD	openMPI 1.2.8	NO	LSF	54x2
INFN-PG	mpich (no mpistart)	NO	PBS	178x2
INFN-TO	mpich 1.2.7	NO	PBS	48x2

Altre risorse MPI regionali

SITO	MPI	Shared Home	Lrms	CPU	Infini band
Grisu-cometa-inaf-ct	mpich mpich2	NO	LSF	140x2	SI
Grisu-cometa-infn-ct	mpich mpich2	NO	LSF	134x2	SI
Grisu-cometa-infn-lns	mpich mpich2	NO	LSF	54x2	SI
Gridu-cometa-unict-diit	mpich mpich2	NO	LSF	84x2	SI
Grisu-cometa-unict-dmi	mpich mpich2	NO	LSF	62x2	SI
Grisu-cometa-unipa	mpich mpich2	NO	LSF	289x2	SI
Grisu-enea-grid	openMPI	SI	LSF	0x4	
Iceage-catania	mpich	SI	PBS	24x2	
uni-perugia	mpich	SI	PBS	16x1	

Risorse MPI in Theophys 2009-2010

RUN	siti mpi in theophys	MPI	Lrms	Altro	Stato/Errori
aprile 2009	10/16	mpich-1.2.7 9 openMPI-1.2.6 1	LSF 6 PBS 4	Infiniband 3	OK 5 SL64/mpi32 2 no-hostbased 1 Aborted 2
maggio 2010	4/10	mpich -1.2.7 3 openMPI-1.2.8 1	LSF 1 PBS 3	Infiniband 1 no mpi-start 1	OK 3

Dopo attivazione dei ticket SAM per MPI

CSN4cluster : Caratteristiche

Installazione: INFN-PISA, Maggio 2010

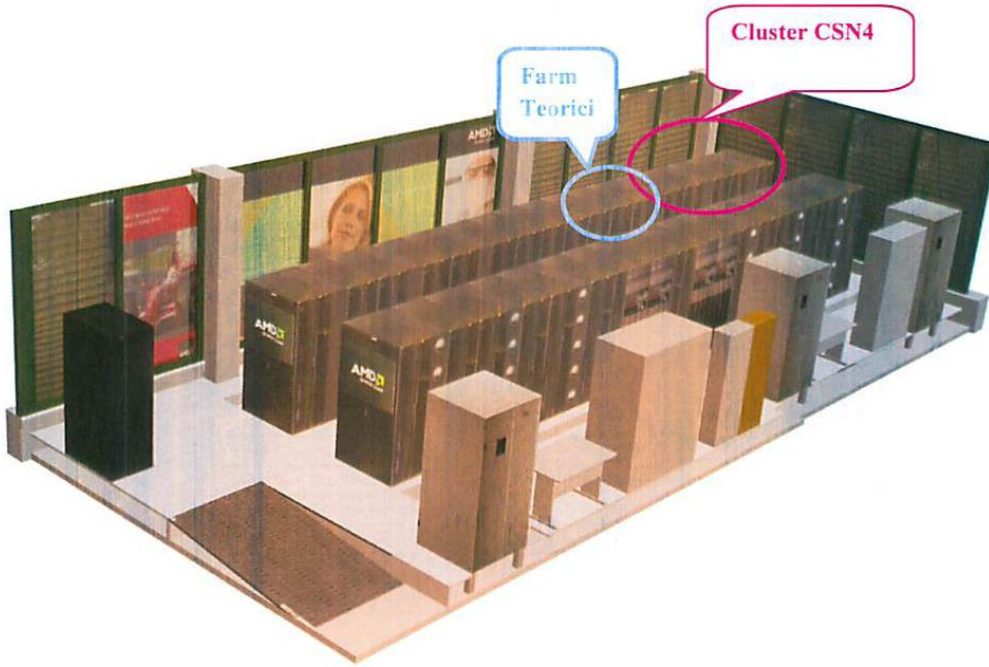
Nodi Calcolo: Biprocessore Acer 2x4 cores Opteron2356 2.3GHz, 8GB ram

Cluster: 128 server (1024 cores) , 10 Tflops di picco

SL5 x86_64, IG gLite 3.2, openMPI, mpich2, openMP

Rete Veloce: Switch Infiniband DDR Cisco 144 porte

Storage: 10 TB shared tra i WNs (GPFS via IP over IB?).



Rendering tridimensionale della Sala Calcolo

CSN4cluster : politiche e metodi di accesso

Ripartizione delle risorse con Fair-Share (LSF) gestito dalla CSN4

- periodicamente (6 mesi?) valutazione dei rendiconti del periodo concluso e delle richieste per il nuovo periodo

A regime unico metodo di **accesso via Grid**.

Temporaneamente la sottomissione di job MPI potrà avvenire anche via SSH utilizzando **la nuova infrastruttura AAI dell'INFN**, in attesa di una piena integrazione di MPI in Grid**.**

CSN4cluster : Job paralleli e sequenziali

Il sistema di code e' in fase di definizione. Possibile organizzazione:

Coda parallela MPI

- Job sequenziali bloccati utilizzando il ruolo "parallel"
- MAX_RESERVE_TIME (tempo massimo di reservation dei cores in attesa che venga accumulato il numero di cores richiesto) da determinare (24 ore?)

Coda sequenziale SHORT

- per job sequenziali con WallClockTime inferiore al MAX_RESERVE_TIME.
- Consente di sfruttare i cores inutilizzati, con ritardi contenuti dei job MPI

Coda sequenziale LONG

- per job sequenziali con WallClockTime > 24 ore
- Viene utilizzata una partizione del cluster per evitare starvation dei job paralleli

Coda parallela NODE

- Numero massimo di cores per job pari al numero di cores per nodo (8).
- Coda per job che richiedono un nodo in modo esclusivo (es: multi-thread)

Report sulle attività dell'MPI-WG di EGEE III

Inizio attività: 02/2009

Fine attività: 04/2010 (conclusione di Egee III)

Mandato:

- completare il lavoro del precedente WG (2007-08 http://www.grid.ie/mpi/wiki): raccomandare un metodo per il supporto di MPI per amministratori e utenti
- Proporre una soluzione per la granularita': come allocare i cores sui nodi

Documento finale: http://www.grid.ie/mpi/wiki/WorkingGroup

MPI-WG : principali raccomandazioni

Pacchetti

mpi-start, MPICH-2 e openMPI.

Shared file-system e SSH password-less tra i WNs.

Se il file-system non e' condiviso i file possono essere distribuiti automaticamente da mpi-start.

SAM test piu' dettagliati

CPU Time limits

se impostato, deve tener conto del parallelismo

Accounting

deve tenere conto del parallelismo (numero di CPU utilizzate, efficienza di MPI, ecc).

https://dgas.cnaf.infn.it/hlrmon/report/charts.php Vedi presentazione di M. Bencivenni.

MPI-WG : nuovi attributi nel JDL

Tre nuovi attributi:

- **SMPGranularity:** numero minimo di cores che devono essere allocati per nodo (default: 1)

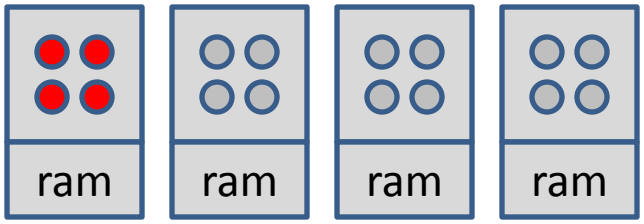
- **WholeNodes**: booleano per allocare un nodo intero (default: false)

- **NodeNumber:** numero di nodi da utilizzare (default: 1)

MPI-WG : esempi JDL

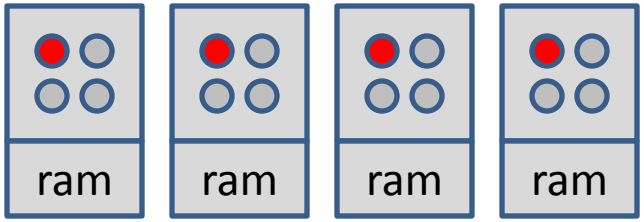
Applicazione Multi-thread con almeno 4 cores:

```
JobType  = “Normal”;
WholeNodes = “true”;
SMPGranularity= “4”;
```



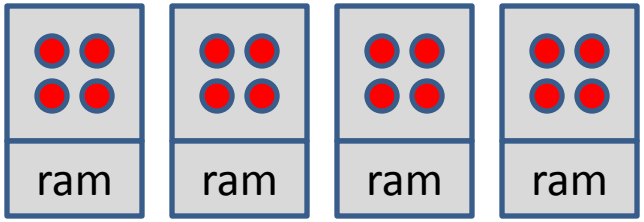
Applicazione MPI che chiede 4 cores su 4 nodi diversi:

```
JobType = “Normal”;
NodeNumber = “4”;
CPUNumber  = “4”;
```



Applicazione ibrida con 4 nodi interi e almeno 4 cores per nodo:

```
JobType  = “Normal”;
WholeNodes = “true”;
NodeNumber  = “4”;
SMPGranularity =“4”;
```



Grazie per l’attenzione!