Il calcolo MPI nell'INFN

esigenze della comunità scientifica, stato attuale e rapporto dal WG MPI di EGEE-III

Roberto Alfieri - INFN Parma

Workshop CCR e INFN-GRID 2009 Palau, 11–15 Maggio 2009

Information Society and Media

Sommario

Il supporto di MPI in Egee

- Supporto “classico” di MPI in gLite e i suoi limiti
- La soluzione attualmente in produzione
- Una soluzione alternativa: il supporto di MPI in Cometa

Rapporto dal nuovo MPI-WG di EGEE-III

- persone, meeting, attività.

Esigenze e utilizzo attuale di MPI

- Risultati del Survey MPI per Utenti e Amministratori
- Risultati del Monitoraggio dei siti MPI utilizzando Gridstat

Conclusioni

- Quale modello per in calcolo MPI in INFN-GRID?

Calcolo HPC nell'INFN

Dati anticipati dal Survey Utenti:

Risorse di sito (Cluster di PC di gruppo, singoli PC multicore**):** **60%**

Provider commerciali (Cineca, ..) **7%**

Federazioni di comunità scientifiche : **33%**

- Progetti APE
- Grid
- Altri progetti

L'utilizzo prevalente è in Gr. IV e Virgo

Il supporto di MPI in Egee prima dell'update 14 di gLite (02/2008)

Sull'UI occorreva compilare con MPICH e trasferire l'eseguibile

```
JobType = "MPICH";
NodeNumber = 8;
Executable = "mpich-executable";
StdOutput = "mpi-test.out";
StdError = "mpi-test.err";
InputSandbox = {"mpich-executable"};
OutputSandbox = {"mpi-test.err","mpi-test.out"};
```

Sul WN veniva eseguito: **`mpirun -np 8 mpich-executable`**

Carenze dell'architettura

Cosa mancava

- poter scegliere il **flavour di MPI** (mpich, mpich2, openmpi, ..)
- poter scegliere **il compilatore** (gcc, icc, pgi,..)
- poter scegliere **l'infrastruttura di rete** (Infiniband, GbEthernet, ..)
- poter scegliere **la granularità** (quante CPU/core per nodo)
- pre-processing (compilare sul sito, ..)
- post-processing (gestire i risultati sul sito, ..)
- la distribuzione automatica dei file se la Home non e' shared
- ....

Il supporto attuale di MPI in Egee

Deriva dalle raccomandazioni dell'MPI WG di Egee (2007-2008)

Coordinato dal Trinity College di Dublino - http://www.grid.ie/mpi/wiki/

Febbraio 2008: Update 14 di gLite 3.1

- Installazione e configurazione di **mpi-start (HLRS Stoccarda)**
  - Supporto di più flavour MPI (openmpi, mpich, mpich2, lam)
  - Estrae automaticamente il machinefile dallo scheduler (pbs, lsf, sge)
  - Supporto alla pre e post esecuzione (mpi-hooks.sh)
  - Distribuzione dei file sui WNs se la Home non è shared

Febbraio 2009: Update 3.1.12 di gLiteWMS

- Per i Job paralleli è ammesso JobType=Normal.

Ad esempio: `JobType=Normal; NodeNumber=4;`

Documentazione: https://twiki.cern.ch/twiki/bin/view/EGEE/MpiTools

Installazione e configurazione

Gestita da Yaim con pochi TAG in site-info.def

Esempio di configurazione:

```
MPI_MPICH_ENABLE="yes"
MPI_MPICH_PATH="/opt/mpich-1.2.7p1/"
MPI_MPICH_VERSION="1.2.7"
MPI_MPICH_MPIEXEC="/opt/mpiexec-0.82/bin/mpiexec"
MPI_OPENMPI_ENABLE="yes"
MPI_OPENMPI_PATH="/opt/openmpi/1.2.6/"
MPI_OPENMPI_VERSION="1.2.6"
MPI_OPENMPI_MPIEXEC="/opt/openmpi/1.2.6/bin/mpiexec"
MPI_SHARED_HOME="yes"
MPI_SSH_HOST_BASED_AUTH="yes"
```

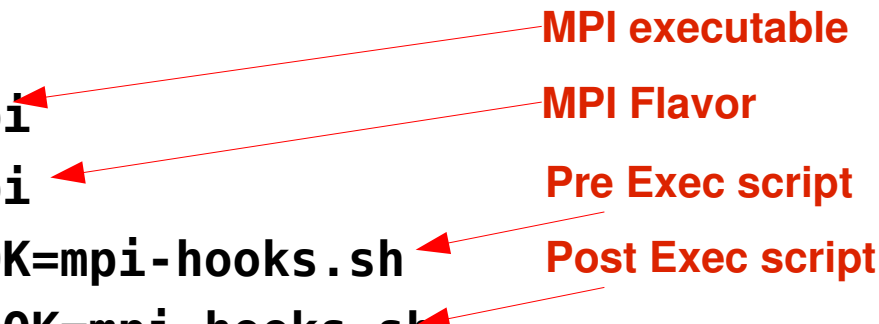
Viene installato mpi-start e solo mpich, altri eventuali flavour vanno installati manualmente.

mpi-start-wrapper.sh

L'utente mette in esecuzione un Wrapper che definisce alcune variabili d'ambiente e quindi esegue mpi-start.

Esempio semplificato:

```
export I2G_MPI_APP=test-mpi
export I2G_MPI_TYPE=openmpi
export I2G_MPI_PRE_RUN_HOOK=mpi-hooks.sh
export I2G_MPI_POST_RUN_HOOK=mpi-hooks.sh
# Invoke mpi-start:
$I2G_MPI_START
```



Mpi-start: mpi-hooks.sh

ESEMPIO DI CONFIGURAZIONE

```
#!/bin/sh
pre_run_hook () {
echo "Compiling ${I2G_MPI_APPLICATION}"
cmd="mpicc ${MPI_MPICC_OPTS} -o ${I2G_MPI_APPLICATION} $
{I2G_MPI_APPLICATION}.c"
echo $cmd
$cmd
if [ ! $? -eq 0 ]; then
echo "Error compiling program. Exiting..."
exit 1
fi
echo "Successfully compiled ${I2G_MPI_APPLICATION}"
return 0
}

post_run_hook () {
echo "Executing post hook."
return 0
}
```

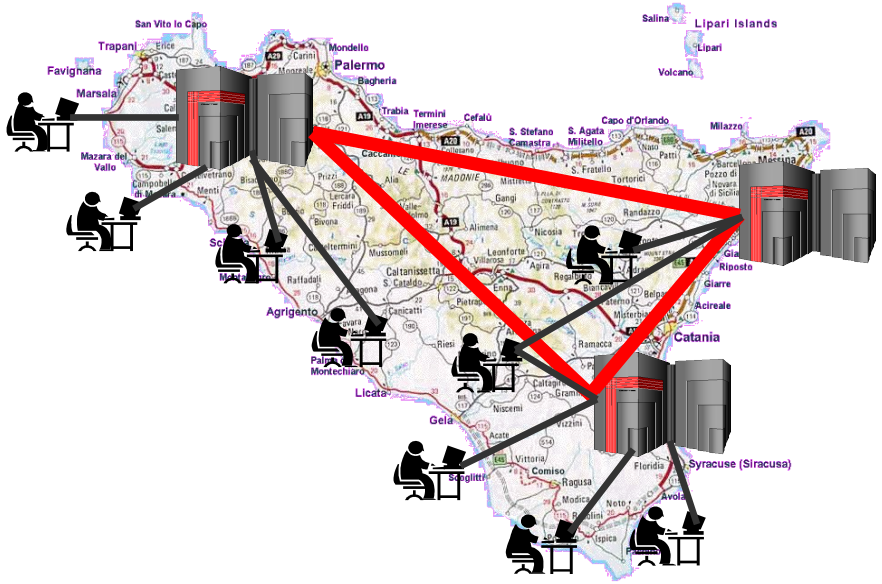
Compilazione nella fase di pre-esecuzione

Submit di un Job MPI: il file JDL

```
JobType = "Normal";
NodeNumber = 10;
Executable = "mpi-start-wrapper.sh";
Arguments = "cpi OPENMPI";
StdOutput = "cpi.out";
StdError = "cpi.err";
InputSandbox = {"cpi", "mpi-start-wrapper.sh"};
OutputSandbox = {"cpi.out", "cpi.err"};
Requirements = Member("MPI-START",
           other.GlueHostApplicationSoftwareRunTimeEnvironment)
&& Member("MPI-INFINIBAND",
          other.GlueHostApplicationSoftwareRunTimeEnvironment)
&& Member("OPENMPI-1.2.7",
          other.GlueHostApplicationSoftwareRunTimeEnvironment);
```

Supporto di MPI in Cometa

Cometa e' un progetto per l'implementazione e lo sviluppo di una infrastruttura Grid in Sicilia basata sul middleware gLite.



Cometa ha modificato il supporto di MPI del Middleware consentendo ulteriori flessibilità per l'utente, mediante l'introduzione di:

- **nuovi TAG JDL (MPIType e MPIGranularity)**
- **script di pre e post esecuzione (mpi.pre.sh e mpi.post.sh)**

MPI in Cometa: nuovi TAG JDL

MPIType : descrive il tipo di libreria e il compilatore da utilizzare per l'applicazione secondo il seguente formato:

<MPI_library>_<compiler_name><compiler_version>

Esempio: `MPIType= MPICC_GCC4;`

MPIGranularity: Consente agli utenti di fornire indicazioni sull'allocazione delle risorse in termini di concentrazione di cores

Esempio: `MPIGranularity = 4;`

MPI in Cometa: il JDL

```
Type = "Job" ;
JobType = "MPICH" ;
MpiType = "mpich_sh_gcc4" ;
NodeNumber = 8 ;
MPIGranularity = 4;
Executable = "cpi-mpich1-gcc4" ;
Arguments = "10" ;
StdOutput = "mpi.out" ;
StdError = "mpi.err" ;
InputSandbox = {"cpi-mpich1-gcc4","mpi.pre.sh", "mpi.post.sh" };
OutputSandbox = {"mpi.err", "mpi.out" };
```

MPI-WG di EGEE-III

Scopo:

Individuare i motivi dello scarso utilizzo di MPI in Egee.

Update delle raccomandazioni per amministratori e utenti, con particolare attenzione alle nuove architetture multicore.

Coordinatore: Jeroen Engelberts (Sara, NL)

Partecipanti: Steve Traylen, Karolis Eigelis, Steven Newhouse, John Ryan, Vangelis Koukis, Dennis van Dok, Jeroen Engelberts, Alvarez Lopez, Salvatore Montforte (INFN) , Fokke Dijkstra, Roberto Alfieri (INFN), Ugo Becciani.

Sito: www.grid.ie/mpi/wiki (lo stesso del precedente MPI-WG)

Prima fonoconferenza: 23/02/2009

Meeting: 18/03/2009 ad Amsterdam

MPI-WG di EGEE-III: Action list

Survey per Amministratori e utenti

Raccomandazioni

- New JDL JobType: Parallel (MPICH deprecated)
- MPI Packages: allargare l'insieme dei Flavour MPI distribuiti da gLite
- Shared file-system
- SSH password-less tra i WNs
- Esecuzione dei SAM test
- Documentazione e formazione
- **New JDL TAG : Nuovo Attributo SMPGranularity**
- **Problemi aperti: strarvation, advanced reservation, check-pointing,..**
- **Work in progress..**

MPI-WG Survey per utenti

Questionario realizzato e diffuso dal MPI-WG di Egee-III.

Abbiamo chiesto agli utenti di inviare i dati anche al Cnaf per poter analizzare i dati nazionali.

Grazie a Marco Bencivenni per l'elaborazione!

Dati rilevanti:

- **Solo 10 questionari ricevuti**
- **Tutti interessati al calcolo parallelo, ma solo 4 usano MPI in Grid**
- **Interessi in diversi ambiti scientifici: fisica 35%, chimica 23%, astronomia 18%,**
- **Strumenti software; mpi2 54%, openmp 20% , mpi1 13% ..**
- **Interesse in architetture multicore**
- **Uso prevalente su risorse locali: 60%, grid 33%, provider commerciali 7%**
- **Difficile trovare documentazione per l'uso di MPI in grid (100%)**

MPI-WG Survey per amministratori

Questionario sono stati realizzato e diffuso dal MPI-WG di Egee-III.

Abbiamo chiesto agli utenti di inviare i dati anche al Cnaf per poter analizzare i dati nazionali.

Grazie a Marco Bencivenni per l'elaborazione!

Dati rilevanti:

- I siti non hanno infrastrutture di rete adeguate: ethernet 70%, Infiniband 23%
- L'implementazione MPI maggiormente supportata è MPICH1 e ciò è in contrasto con le esigenze degli utenti che richiedono un maggiore supporto a MPICH2
- Il processo d'installazione e configurazione risulta abbastanza semplice ma sarebbe necessaria una documentazione più accurata
- Scarso utilizzo di Job MPI rispetto al numero complessivo: < 10%

gridstat

Script “probe” sviluppato a Parma ed eseguito su tutti i siti che supportano Mpi-start e la VO di Theophys.

Scopo:

Monitorare lo stato dei siti e delle risorse disponibili

Determinare direttamente sulle risorse i parametri utili per MPI:

- Tipo di Interconnessione (Infiniband, Gigabit Eth, ..)
- Latenze e Banda nell'interconnessione tra i nodi
- Prestazioni di calcolo Floating Point (benchmark custom)
- Numero di nodi effettivamente disponibili.

http://www.fis.unipr.it/dokuwiki/doku.php?id=grid:gridstat

Gridstat: risultati

VO THEOPHYS con supporto MPI-START:

RUN	Siti	Mpi	LRMS	Infiniband	Stato/errori
04/12/08	10/16	mpich1.2.7: 10 openmpi-1.2.6: 1	LSF: 4 PBS: 6	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	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	3	OK: 4 SL64/mpi32: 3 no-hostbasedauth: 1 aborted: 2

Conclusioni

- INFN-Grid e' gestita ed utilizzata prevalentemente per Job sequenziali
- C'e' un interesse crescente a MPI su Grid, ma il supporto in Egee è complesso ed ancora in evoluzione.
- Ad oggi l'uso di MPI in INFN-Grid e' → 0

Quale modello di supporto al calcolo MPI per Infn-Grid?

Possibile proposta:

Individuare in una prima fase un ristretto insieme di siti medio-grandi che concordano un minimo di funzionalità comuni (Es. Infiniband, MPI1 e MPI2 openmp, shared Home, LRMS?, ..), eventualmente in WMS separato, con test SAM specifici.

Successivamente estendere le funzionalità mancanti ed i siti coinvolti.

Domande?