ReCaS Datacenter cluster HPC, guida al suo utilizzo

Febbraio, 2016

Alessandro Italiano

Indice:

1. Caratteristiche del cluster HPC del ReCaS Datacenter

Il cluster e' composto da 20 macchine, naturalmente identiche,che hanno le seguenti caratteristiche:

- 2 socket da 10 core Intel(R) Xeon(R) CPU E5-2650L v2 @ 1.70GHz
- Una interfaccia InfiniBand: QLogic Corp. IBA7322 QDR InfiniBand HCA (rev 02)
- Una scheda grafica NVIDIA Corporation GK110BGL [Tesla K40m] (rev a1)
- Una scheda Ethernet a 10Gb/s in fibra ottica

In nodi sono tra loro interconnessi tramite una rete a bassa latenza con banda passata fino a 40Gb/s tramite uno switch Intel 12300 True Scale Fabric.

2. Accesso alle risorse.

L'accesso alle risorse del cluster HPC del ReCaS Datacenter e garantito solo ed esclusivamente in modalità batch. Ne consegue che la sottomissione di job batch possa essere effettuata solo dopo avere fatto il login via ssh sulla seguente macchine:

- frontend.recas.ba.infn.it

DNS alias dei seguenti due server Unix
- ui02.recas.ba.infn.it
- ui03.recas.ba.infn.it

Su tali macchine l'utente avrà a disposizione una HOME directory per creare ed editare file necessari all'esecuzione delle propria applicazione HPC. Il codice applicativo **non può essere** compilato sui frontend nel caso il codice richieda il supporto alla versione locale di openMPI poiché infiniband non è presente sui frontend.

3. Compilazione codice

Nel caso il codice applicativo sia parallelo e si voglia usare la versione locale disponibile via module di openMPI si dove procedere alla compilazione attraverso un job interattivo sottomesso al batch System. In questo modo ci sarà a disposizione una sessione interattiva sui uno dei nodi del cluster sul quale si potrà procedere alla compilazione del codice poiché sui nodi è presente infiniband con relativo codice sorgente.

Comando per sottomettere un job interattivo
- qsub -I -q *bigmpi2@sauron.recas.ba.infn.it*
- > module avail
- > module load gcc-447/openmpi-1.10.2

4. Filesystem condiviso

Sul cluster HPC oltre che sulle due macchine di frontend saranno disponibili i seguenti filesystem condivisi

- /lustre:
  - IBM GPFS filesystem usato per memorizzare i dati degli utenti che vengono scritti in un'unica copia cio' vuol dire che non e' garantita la consistenza del contenuto della propria aria. Su questo filesystem viene applicata una quota a livello di gruppo locale/esperimento di appartenenza superata la quale l'intero gruppo non potra' piu' usare il filesystem per scrivere dati.

- /lustrehome:
  - IBM GPFS filesystem usato per fornire spazio disco per le home degli utenti. I file in questo filesystem vengono gestiti in doppia copia di modo da garantire la consistenza del contenuto della propria area.

- /opt/exp_soft:
  - NFS filesystem usato per condividere il codice disponibile. Anche questo filesystem gestisce i file in doppia copia.

5. Compilatori

Allo stato attuale il compilatore disponibile di default e' il seguente:

- gcc

disponibile in più versioni ed accessibile utilizzando il comando module

6. Module

L'inizializzazione dell'ambiante di lavoro sia per l'utente che per i job mpi va fatta utilizzando il comando module, il quale in base ai comandi immessi caricherà l'ambiente richiesto. Di seguito si riportano alcuni esempi esplicativi:

- Vedere i moduli disponibili sul cluster HPC
  - *module avail*
- Caricare un modulo
  - *module load compilers/gcc-447*
- Vedere i moduli caricati
  - *module list*
- Rimuovere un modulo precedentemente caricato
  - *module unload compilers/gcc-447*

7. Batch system

Le risorse di calcolo presenti nel cluster HPC siano queste le GPU che le CPU vengono gestite da un batch system che nel nostro caso specifico e'

- Torque versione 6.0.0.1 - Resource Manager
- MAUI versione 3.3.1 - Scheduling

In un sistema multi utente, come quello descritto in questo documento, il Batch system svolge le seguenti funzioni:

- Fornisce un sistema di accodamento per le richieste di sottomissione dei batch job. I job rimango in coda fino a quando non ci sono 0e risorse necessarie per la loro esecuzione
- Gestisce le priorità di accesso alle risorse garantendo la quota di risorse per le quale il gruppo di appartenenza dell'utente ha pagato.
- Alloca correttamente le risorse richieste dall'utente al momento della sottomissione in modo da rispettare i requisiti del batch job. Le risorse allocate saranno ad uso esclusivo per l'esecuzione del batch job.
- Si occupa in maniera completamente trasparente all'utente di eseguire remotamente rispetto alla macchina di sottomissione, cioè il frontend, l'applicazione HPC.
- L'utente può interagire col batch system attraverso comandi disponibili sul frontend nel default path della propria bash. Per esempio può recuperare lo stato dei job sottomessi o lo stato del batch system stesso.

8. Sottomettere un batch job

Allo stato attuale sul sistema batch e' presente una sola coda per cui i job andranno sottomessi usando solo questa.

- batch queue: bigmpi2, con limite di esecuzione di 1000 ore

Un esempio di comando per la sottomissione e' il seguente:

- *qsub -q bigmpi2@sauron.recas.ba.infn.it -l nodes=2:ppn=40 my_mpi_script*
  - questo job richiede l'esecuzione di "my_mpi_script" con 80 core allocati su due macchine
- *qsub -q bigmpi2@sauron.recas.ba.infn.it -l ppn=40 my_mpi_script*
  - questo job richiede l'esecuzione di "my_mpi_script" con 40 core

Un esempio di script per l'esecuzione di un job mpi può essere il seguente

```
#/bin/bash

module load gcc-447/openmpi-1.10.2


/opt/exp_soft/misc/gcc-447/openmpi-1.10.2/bin/mpirun -n $PBS_NP -machinefile $PBS_NODEFILE --mca mtl psm /lustre/home/italiano/OSU/libexec/osu-micro-benchmarks/mpi/pt2pt/osu_latency D D
```