Stato attività SLURM

Giacinto DONVITO
Alessandro ITALIANO

Test di funzionalità

- MEMLIMIT:
  - **Funziona come in LSF: se il job supera la soglia di memoria imposta viene killato**
- Forward dell'X nel job interattivi:
  - **Ho trovato e patchato un semplice script shell che consente di fare un job interattivo completo di X11**
    - srun.X11
      - In pratica sottomette un job e poi fa login ssh sulla macchina dove il job è finito usando gli stream di input/output di quel job
        - » Richiede il login sui WN

Test di funzionalità

- Add/Remove NODE:
  - **Cambiare il file di conf…**
  - **scontrol reconfigure**
- Copia automatica dello standard output/ error:
  - **Non sembra possibile neanche usando epilogue script**
    - **Mancano diverse informazioni necessarie**
  - Forse però l'uso di un file-system distribuito e condiviso per le home degli utenti di un cluster non è un problema troppo grave.

Test di funzionalità

- Copia automatica dello standard output/error (cont):
  - Per i job grid il CREAM-CE dovrebbe aver risolto la cosa con il wrapper di sottomissione
  - Per i job locali:
    - A bari abbiamo già un file system distribuito condiviso fra i front-end e il cluster di WN per permette l'esecuzione di jobs che altrimenti sarebbe piuttosto difficile per la media degli utenti
  - Farm ben più grandi dei nostri centri già usano questo sistema per eseguire job degli utenti
  - Inoltre, i job MPI più complessi richiedono obbligatoriamente un file system condiviso fra i nodi
  - Di seguito alcuni esempi di datacenter di grandi dimensioni che usano file-system Lustre per le home (o la Scratch area)
    - Sicuramente ci saranno esempi simili con GPFS…

Esperienza di datacenter con home condivisa

- NASA Pleiades:
  - Total processors: 23,552
  - Total cores: 126,720
  - SGI® InfiniteStorage NEXIS 9000 home filesystem
  - 12 DDN RAIDs, 9.3 PB total
  - 6 Oracle Lustre cluster-wide filesystems

Esperienza di datacenter con home condivisa

Australian NCI National Facility

- Tutti i nodi sono diskless e il sistema operativo è su Lustre
  - anche le home sono su Lustre
- Current machine "vayu"
  - ~1500 nodes, ~12k Nehalem cores
  - 26 OSS's, 4 MDS's
- Root on Lustre – Why?
  - Simplicity
    - Fewer things to fail
    - No NFS or local disks involved Reliability and Scalability
  - Use centralised scalable and reliable hardware
    - If Lustre is down then jobs are hung anyway. May as well put the OS there too
  - Maintainability
    - One rsync from the master OS image to the OS image on Lustre updates every node immediately
    - Unlimited space for OS packages, OS variations, ...

Esperienza di datacenter con home condivisa

National Climate Computing Center

- Capacity: fit the use cases that need performance
  - Scratch
  - Hot dataset cache
  - Semi-persistent library
  - Staging and buffering for WAN transfer
- Consistency: use cases increase variability
  - Some demand capability (scratch, hot cache)
    - Significantly more random access
  - Some are more about capacity (library, staging)
    - More sequential access
- Cost: Always an issue
  - On a fixed budget, I/O robs compute
  - Capability costs compute resources (more I/O nodes)

- Phase 1: Cray XT
  62,576 AMD Opteron 6174
- Phase 2: Cray XE
  65,200 AMD Opteron 16-core

Esperienza di datacenter con home condivisa

National Climate Computing Center

- Fast Scratch
  - 18x DDN SFA10000
  - 2,160 active 600GB SAS 15000 RPM disks
  - 36 OSS
  - InfiniBand QDR

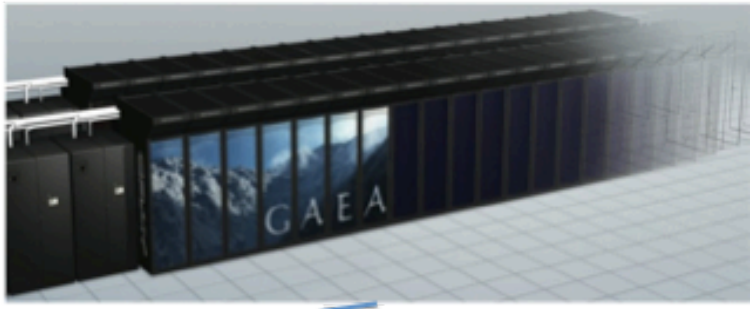
120 Disk per DDN system

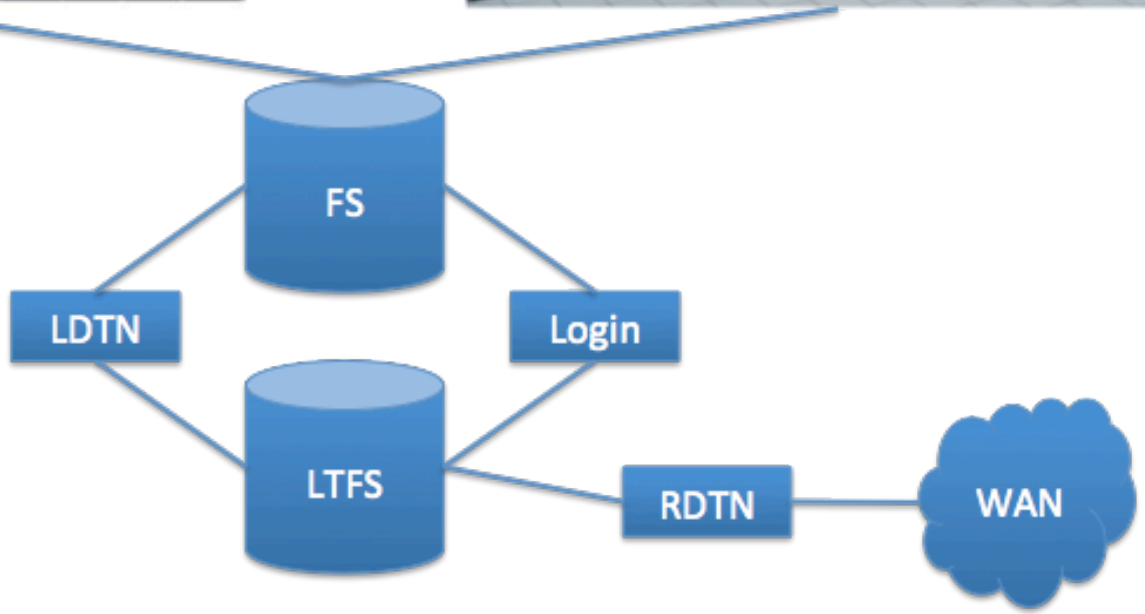
- Long Term Fast Scratch8x DDN SFA100002,240 active 2TB SATA 7200 RPM disks16 OSSInfiniBand QDR

280 Disk per DDN system

Esperienza di datacenter con home condivisa

National Climate Computing Center





Gaea filesystem architecture

ToDo & Future Work

- Alessandro ha finito lo sviluppo del plugin per WNoDeS
  - Nelle prossime settimane si proverà in una istanza di test
- Testing CREAM-CE
- Scalability test:
  - Potremmo usare le macchine nuove che stanno arrivando a Bari:
    - 40 host con 24 core e 80GB di RAM ciascuno
      - Potremmo far partire circa un migliaio di nodi virtuali e configurarli come WN di SLURM