Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

"Proposta di progettazione per un cluster HPC per il gruppo DEMACS"

Antonio Francesco Gentile ,Davide Macrì

RT-ICAR-CS-23-02 **Gennaio 2023**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 8-9C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.icar.cnr.it*
– Sezione di Palermo, Via Ugo La Malfa, 153, 90146 Palermo, URL: *www.icar.cnr.it*

Proposta di progettazione per un cluster HPC per il gruppo DEMACS

Progettazione di un cluster HPC per la facoltà di Matematica dell'Unical

Antonio Francesco Gentile, Davide Macrì

Introduzione

• Desiderata:

- Conversione di 3/4 nodi singoli già disponibili presso il centro di calcolo della facoltà in un cluster di calcolo
- Gestione dello stesso tramite il software SLURM
- Utilizzo delle risorse di storage disponibili (singoli HD DISK locali in logica condivisa)
- Sfruttamento di tutte le risorse di rete disponibili

Soluzioni proposte

- Gestione dei singoli HD dei singoli nodi in logica LVM, in modo da permettere la creazione di uno o più volumi logici di dimensione pari alla somma degli spazi fisici degli Hard Disk (meno la quota di gestione dell'LVM)
- Gestione del cluster tramite SLURM via chiavi gestite dal software MUNGE
- Pubblicazione degli LVM come cartelle pubbliche gestite via NFS
- Creazione di politiche di accounting per accesso al cluster mediante tecnologia NIS
- Creazione di servizi DHCP DNS per i nodi del cluster di calcolo
- Creazione di servizi NTP per sincronia temporale intra/extra cluster
- Condivisione dell HOME directories e di un folser DATA per tutte le utenze NIS
- Creazione di un firewall perimetrale per protezione del sistema di calcolo

Panoramica di Slurm 1/2

- sistema open source
- tolleranza ai guasti
- scalabilità per gestione e pianificazione dei lavori
- Non richiede modifiche al kernel per funzionare
- relativamente autonomo

Panoramica di Slurm 2/2

Slurm fornisce tre funzioni chiave per la gestione del carico di lavoro del cluster

- Assegnazione agli utenti dell'accesso esclusivo/non esclusivo ai nodi di calcolo per un certo periodo di tempo in modo che possano eseguire il lavoro.
- Pubblicazione di un framework per l'avvio, l'esecuzione e il monitoraggio del lavoro (solitamente in parallelo) sull'insieme di nodi allocati.
- Arbitraggio della contesa per le risorse e gestione delle code di lavori in sospeso.

Architettura di Slurm

Le entità gestite dall'ecosistema Slurm includono

- nodi - la risorsa di calcolo in Slurm
- partizioni - raggruppamenti dei nodi in insiemi logici possibilmente sovrapposti
- lavori - allocazioni di risorse assegnate a un utente per una quantità specificata di tempo e fasi di lavoro, cioè insiemi di attività all'interno di un lavoro.

Architettura di Slurm

Slurm è un sistema di calcolo modulare composto da:

- un daemon slurmd in esecuzione su ciascun nodo di calcolo
- un daemon slurmctld centrale in esecuzione su un nodo di gestione (con failover twin opzionale)
- un daemon slurmdbd centrale in esecuzione su un nodo di gestione (per gestione dati via SQL opzionale).

In figura si evince l'architettura virtualizzata necessaria a realizzare un cluster Slurm con un frontend e 3 nodi di calcolo:

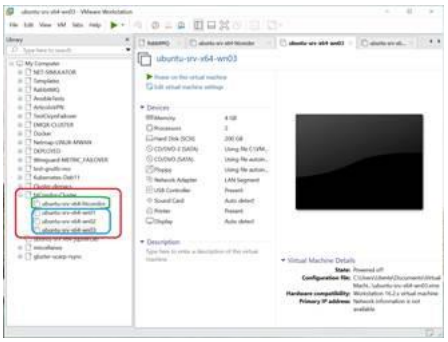
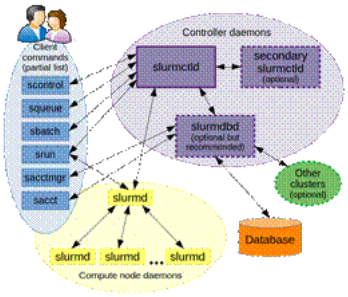
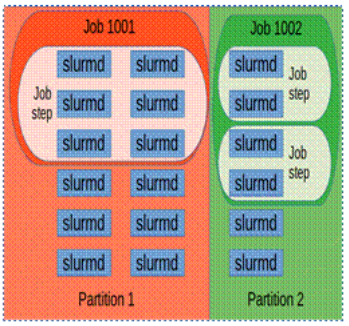


Figure 1: Vmware cluster demacs.



General Slurm Architecture General Slurm Architecture



General Slurm Partition Architecture

Architettura di NIS

Il Network Information Service (NIS)

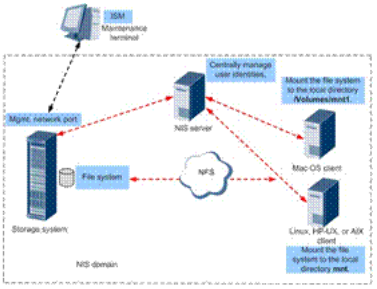
- protocollo di servizio di directory client-server per la distribuzione dei dati di configurazione del sistema come nomi utente e host tra computer su una rete di computer
- Un sistema NIS/YP mantiene e distribuisce una directory centrale di informazioni su utenti e gruppi, nomi host, alias di posta elettronica e altre tabelle di informazioni basate su testo in una rete di computer.
- Gli amministratori hanno la possibilità di configurare NIS per fornire i dati delle password a processi esterni per autenticare gli utenti utilizzando varie versioni degli algoritmi hash Unix crypt(3). Tuttavia, in tali casi, qualsiasi client NIS può recuperare l'intero database delle password per l'ispezione offline.
- Segmentazione a livello firewall relativa solo alla rete locale del cluster.
- Su LAN di grandi dimensioni , i server DNS possono fornire una migliore funzionalità del server dei nomi rispetto a quella fornita da NIS o LDAP, lasciando solo le informazioni di identificazione a livello di sito per i sistemi master e slave NIS da servire.

Panoramica di NIS 1/2

- Fornisce un semplice servizio di ricerca in rete costituito da database e processi
- Precedentemente noto come Sun Yellow Pages (YP).
- Fornisce informazioni, che devono essere note in tutta la rete:
  - nomi di accesso/password/home directory (/etc/passwd)
  - informazioni sul gruppo (/etc/group)
  - nomi host e numeri IP (/etc/hosts)
- La versione NIS+ fornisce il supporto per la crittografia dei dati e l'autenticazione su RPC sicuro.
- Il modello di assegnazione dei nomi di NIS+ è basato su una struttura ad albero. Ogni nodo dell'albero corrisponde ad un oggetto di NIS+, del quale si hanno sei tipi: directory, entry, group, link, table e private.

Panoramica di NIS 2/2

Ad esempio, se la propria password è registrata nel database NIS passwd, si sarà in grado di accedere a tutte le macchine in rete che hanno i programmi client NIS in esecuzione.



[0]General NIS + NFS Architecture

Panoramica del filesystem distribuito NFS (Network File System) 1/2

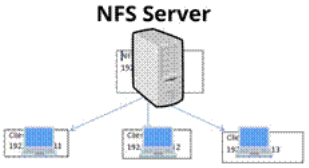
In ambiente Linux uno dei sistemi maggiormente utilizzati per la condivisione di dati in una rete di computer / nodi di calcolo di un cluster

- Il folder /home del nodo frontend, configurato via LVM, viene condiviso con i computer client
- Il folder /data del nodo frontend, configurato via LVM, viene condiviso con i computer client

In questo modo i contenuti delle cartelle sul server saranno accessibili direttamente dai client.

Panoramica del filesystem distribuito NFS (Network File System) 2/2

- Il servizio NFS sarà disponibile solo all'interno del network dei nodi di calcolo
- Il servizio NFS non sarà accessibile fuori dalla rete d'ateneo / da quella di dipartimento

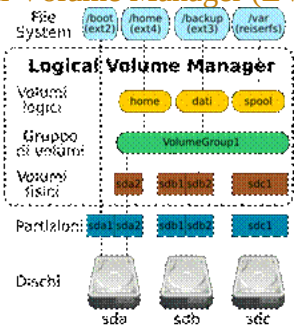


[0]General NFS Architecture

Panoramica del Logical Volume Manager (LVM) 1/2

- Ciascun disco fisico contiene al suo interno un certo numero di partizioni che possono o meno essere usate in LVM.
- Ogni partizione che viene inserita in LVM prende il nome di volume fisico (PV = Physical Volume)
- Tutti o solo alcuni volumi fisici possono essere assegnati a un gruppo di volumi (VG = Volume group). Questo consente di sfruttare più dischi/partizioni unite in un'unica struttura dati.
- Tutto lo spazio dedicato a un gruppo di volumi può essere suddiviso in uno o più volumi logici (LV = Logical Volume) che verranno usati dal sistema per costituire il file system, semplicemente formattandoli nella maniera preferita, come se fossero delle partizioni.

Panoramica del Logical Volume Manager (LVM) 2/2 L'LVM NON SOSTITUISCE I BACKUP.

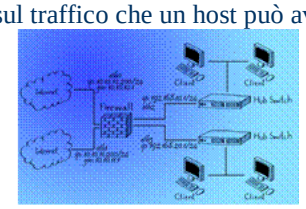


[0]General LVM Architecture

- Un firewall di rete classico permette di filtrare il traffico esterno prima che raggiuna dei server in una DMZ. Si ha a che fare con 2 interfacce
  - una esterna, esposta ad Internet
  - una sulla DMZ, che costituisce il default gateway dei server pubblici Può agire in 2 modi principali:
  - Routing fra rete esterna e DMZ con IP pubblic
  - Natting fra rete esterna e DMZ con IP privati nattati dal firewall

Il controllo su quali host della rete interna possono accedere a Internet può essere molto più fine, agendo a livello di:

- porte/protocolli concesse/i
- a livello di MAC address degli host abilitati
- introducendo un limite sul traffico che un host può avere



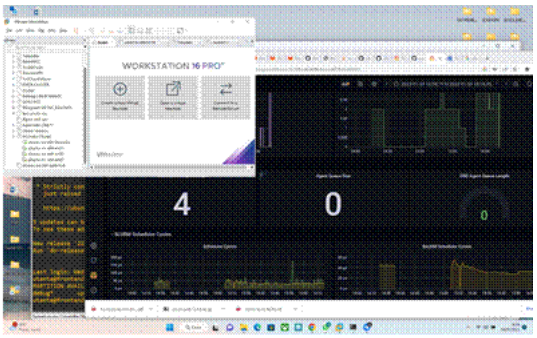
[0]Basic Firewall Architecture

Il servizio DHCP/DNS fornisce internamente al Cluster sia la configurazione di rete automatica dei nodi, sia il servizio di risoluzione dei nomi, interno ed esterno al dominio NIS. Il servizio NTP serve a garantire la sincronia interna del clock dei client rispetto al tempo di sistema del server.

La Slurm dashboard può essere utilizzata per visualizzare lo stato di un cluster Linux gestito tramite SLURM .

- Integrazione con slurm mediante SLURM exporter per Prometheus.
- Visualizza le seguenti metriche:
  - Stato dei nodi
  - Stato delle CPU/GPU
  - Stato dei lavori: include anche informazioni sui lavori in esecuzione/in sospeso/sospesi per account/utente
  - Informazioni sull'agenda
  - Condividere informazioni

Grafana Server che permette di visualizzare infografiche e allarmi per il web, unificando varie sorgenti di dati Visualizzazione su appositi pannelli (dashboard). Creazione di questi pannelli è effettuabile tramite query builder interattivi
Il supporto ad alcune sorgenti è incluso in maniera predefinita come ad esempio Elasticsearch, MySQL e Prometheus. Altre tecnologie sono integrabili per mezzo di componenti aggiuntivi fra
cui GitLab, Jira, PostgreSQL, Solr e Zabbix Mediante i componenti aggiuntivi si raggiunge la copertura di circa un centinaio di altre sorgenti di dati
Prometheus Tool di monitoraggio e alerting che archivia metriche in un database proprietario di timeseries.
Una timeseries, o serie temporale, è una serie di data point indicizzati (o elencati o rappresentati in un grafico) in ordine temporale.



[0]Prometheus Slurm Grafana Architecture