Implementazione della Cloud dell'area Padovana

Status Report

Versione 1.3 (31 Marzo 2014)

Introduzione

Come indicato in [1], obiettivo del progetto “Cloud dell’area Padovana” e` l’implementazione di un servizio di Cloud computing e storage rivolto in particolare agli utenti della sezione INFN di Padova e dei Laboratori Nazionali di Legnaro, ma che sia integrato in una futura Cloud INFN nazionale.

Viene qui descritto lo stato di realizzazione di questo progetto

Definizione del layout della Cloud dell’area Padovana

Considerando l'obiettivo di implementare una Cloud di produzione in tempi brevi, sfruttando al massimo software gia` esistenti, si e` deciso di fare riferimento al software Openstack come middleware Cloud da usare. Si e` scelto di considerare Havana come prima versione di cui fare il deployment.
Si e` inoltre deciso di implementare un’unica Cloud, con risorse fisicamente distribuite tra i due siti.

Il layout che si e` deciso di considerare e` simmetrico rispetto a quello adottato dal Tier-2 Legnaro-Padova (in cui i servizi sono a Legnaro mentre i worker node sono distribuiti in entrambi i siti) ovvero:

- 2 Controller node (in HA) a Padova
- 2 Network node (in HA) a Padova
- Storage di servizio (per le istanze e per le immagini) a Padova
- Compute node a Padova e Legnaro

Per quel che riguarda lo storage delle istanze, si e` deciso di considerare una configurazione “Off Compute Node Storage—Shared File System”, con uno shared file system per supportare la live migration, e implementato attraverso storage “esterno” ai compute nodes.

Si e` scelto di usare GlusterFS per l’implementazione di questo file system condiviso.

Per quel che riguarda i servizi di storage rivolti all’end-user, verra` fatto il deployment del servizio block storage Openstack Cinder.
Oltre a Cinder sara` inoltre possibile avere accesso ad altro storage Posix, implementato attraverso GlusterFS.

Non verra` invece installato il servizio object storage Swift, almeno in una prima fase, visto che non sono stati individuati use case che ne giustifichino la necessita`.

Hardware per la Cloud di produzione

Per l'implementazione della Cloud dell'area Padovana, considerando il layout sopra descritto, a fine 2013 e` stato acquistato a Padova il seguente hardware:

- 1 Blade enclosure DELL M1000e
- 4 lame DELL M620 ciascuna con 1 processore E52609, 32 GB RAM, da usarsi come controller node (in HA) e network node (in HA)
- 5 lame DELL M620 ciascuna con 2 processori E5-2670v2, 80 GB RAM, da usarsi come compute node
- 2 switch DELL Force 10 MXL
- 1 server iSCSI DELL MD3620i, con 23 dischi SAS da 900 GB, da usarsi per lo storage delle istanze, per lo storage delle immagini (Glance) e per il servizio di block storage Cinder

Il costo totale e` stato di circa 70Keuro, di cui 10Keuro finanziati dalla CCR.

Il materiale e` arrivato a fine Febbraio e, dopo averne fatto il testing e profiling, lo si sta configurando per il suo uso previsto.

A Legnaro invece sono stati acquistati:

- 4 Fujitsu Primergy RX300S7 (2 processori XEON E5-2665, 96 GB RAM) da usarsi come compute node
- Server storage DELL PowerVault MD3600f (FC) con 12 dischi da 4 TB, da usarsi per storage utente
- Ottica 10 Gbps per update link con Padova

Il costo totale e` stato di circa 27Keuro, di cui 10Keuro finanziati dalla CCR (il resto dal Laboratorio).

Altro hardware

Oltre all'harware che e` stato acquistato per il deployment della Cloud di produzione, sono usate molte altre risorse hardware per le attivita` sotto descritte, In particolare per:

- Servizi per l'installazione e configurazione del sistema operativo e dei servizi
- Servizi di monitoring
- Cluster mysql in HA

- Cluster HAProxy
- Testing delle diverse configurazioni della Cloud
- Sviluppo e testing del software
- Servizio Keystone nazionale
- Testing di soluzioni di storage distribuito
- EGI Federated Cloud Task Force

A tal fine si sta utilizzando in particolare hardware fornito dagli esperimenti: non molto prestante, ma adeguato per questi utilizzi.

Infrastruttura di rete

Gli switch a cui sono collegate le blade Cloud a Padova hanno un uplink a 10 Gbps verso lo switch/router centrale.
Il server i-SCSI e` collegato agli switch delle blade attraverso 4 link a 10 Gbps.

Il link che collega la Sezione di Padova con i Laboratori di Legnaro su cui passa tutto il traffico tra i due siti (escluso quello del Tier-2, per il quale esiste un'altra fibra dedicata) e che verra` utilizzato anche per questa Cloud, e` stato recentemente (20 Marzo 2014) upgradato da 1 Gbps a 10 Gbps.

A Legnaro e` stato recentemente fatto l'update (a 10 Gbps) del link tra il centro stella e lo switch su cui sono collegati i nodi Cloud.

Implementazione della Cloud dell'area Padovana

Servizi per l'High Availability

Per il deployment in modalita` High Availability dei servizi Openstack, e` stato implementato un cluster Mysql HA usando Percona XtraDB Multi-Master.
Tale cluster e` stato implementato su 3 macchine virtuali.

Oltre a questo e` stato implementato un HAProxy/Keepalived cluster per i servizi di load balancing e Virtual IP, per consentire l'accesso al database ridondato mediante endpoint univoco. Anche questo servizio e` stato implementato su 3 macchine virtuali.

Questo cluster Mysql HA, oltre che per i servizi Openstack, viene utilizzato anche per il database del servizio Foreman.

Configurazione dei servizi Openstack in modalita` HA

Per Keystone e Glance si e` scelto di considerare una modalita` di tipo Active/Active.

Per Neutron si sta provando la modalita`attivo/attivo con un processo esterno per la migrazione dei router virtuali da un l l3-agent ad un altro.

In mancanza dell'hardware definitivo (arrivato da poco) le prove di configurazione dei servizi Openstack in modalita` HA viene effettuata su hardware dismesso dal GR1.

Networking

Sono state definite e assegnate le varie reti per Openstack (management, data, public) [2].

Floating IP verranno assegnati solo per servizi che avranno davvero la necessita` di un indirizzo pubblico.

Le VM che verranno istanziate sulla Cloud dell'area Padovana senza floating IP pubblico saranno comunque accedibili dalle LAN di Padova e di Legnaro. La configurazione del servizio Neutron di Openstack per l'implementazione di questa funzionalita` e` stata definita e testata con successo.

E` stato anche analizzato come le VM possano accedere in maniera efficiente a storage esterno alla Cloud (es. un NFS server), senza necessita` di passare per il Network Node. E` stato individuato il setup necessario [3] che e` stato testato con successo.

Autenticazione, autorizzazione

E` stato definito il modello di autenticazione e autorizzazione ai Servizi della Cloud dell'area Padovana.
Per quel che riguarda l'autenticazione, oltre all'accesso via username/password sara` supportata anche l'autenticazione via IDP (INFN-AAI e piu` in generale IDEM).
L'autorizzazione sara` implementata attraverso i meccanismi nativi di Openstack.

Gli utenti saranno organizzati in Progetti (Tenant), ognuno dei quali avra` un responsabile che si fara` carico di accettare/rifiutare le richieste di membership per quel particolare tenant.

Sara` predisposto anche un tenant “Guest” (con poche risorse disponibili) disponibile agli utenti che vorranno provare le funzionalita` della Cloud e che non sono associabili a nessun tenant gia` esistente.

I dettagli sulle modalita` di registrazione, di richiesta membership a un tenant esistente, di creazione di un nuovo tenant, sono definite in [4].
L’implementazione e` in fase di finalizzazione.
Le modalita` di integrazione di questo software con Openstack sono descritte in [5].

Tool per l’installazione e configurazione

Per l’installazione e configurazione dei nodi dell’infrastruttura Cloud, si e` deciso di fare riferimento a un tool di gestione centralizzata e automatizzata.
Dopo un’analisi degli strumenti disponibili, si e` scelto di adottare Foreman e Puppet.
Foreman e Puppet sono stati scelti, nonostante non vi fossero gia` competenze su questi due tool, perche` da un’analisi fatta sembrano soddisfare i requirement per una gestione automatizzata, e perche` sono tool che ultimamente vengono adottati in molti siti (es. il CERN).

Foreman/Puppet master sono stati installati su un server dismesso dal GR1 (cld-foreman.cloud.pd.infn.it).
Attualmente questo sistema viene utilizzato sia per l’installazione, che per la configurazione di diversi servizi delle macchine di produzione o test Cloud.

Monitoring

Oltre ai servizi di monitoring propri di Foreman, per il monitoring della Cloud dell’area Padovana e` stata predisposta un’infrastruttura di monitoring basata su Nagios e Ganglia.
Oltre a monitorare lo stato delle diverse macchine, tale infrastruttura viene utilizzata anche per controllare la funzionalita` e l’efficienza dei singoli servizi Openstack.
I servizi Nagios (cld-nagios.cloud.pd.infn.it) e Ganglia (cld-ganglia.cloud.pd.infn.it) sono stati installati su due macchine virtuali ospitate nel cluster di virtualizzazione di Sezione.

L’installazione e configurazione dei plugin Nagios e Ganglia nei diversi nodi da monitorare viene fatta via Puppet.

Documentazione

La documentazione inerente le varie attivita` in corso viene ospitata nella wiki dell’INFN ([6]).

Le informazioni relative alle risorse hardware disponibili e il logbook delle operazioni effettuate su queste vengono invece mantenute su un'istanza DOCET. DOCET (Data Oriented Centre Tool) e` un tool sviluppato a Padova, che permette la gestione di tutte le informazioni relative ad un data centre. Viene usato da anni anche per le attivita` del Tier-2 di Legnaro-Padova.

Partecipazione alle attivita` del Cloud Working Group della CCR

Oltre alle attivita` direttamente finalizzate al servizio di produzione Cloud dell'area Padovana, il personale della Sezione di Padova e` coinvolto anche in altre attivita`, coordinate nell'ambito del working group Cloud della Commissione Calcolo e Reti, e qui sotto riportate.

QUACK

Il team di Padova e` attivo nell'implementazione del sistema QUACK [7] che, si prefigge:

- di integrare Grid e Cloud
- di gestire l'allocazione delle risorse senza un partizionamento statico delle stesse tra i diversi gruppi
- di gestire situazioni in cui le risorse sono pienamente utilizzate (in tali scenari Openstack semplicemente rifiuta la richiesta di allocazione di nuove risorse).

In particolare il gruppo di Padova e` responsabile dell'integrazione in Openstack di uno scheduler efficiente (e` stato scelto di fare riferimento allo scheduler SLURM).
E` stato implementato un prototipo che:

- permette di gestire l'accodamento di richieste di istanziazione di VMs nel caso di pieno utilizzo della Cloud; tali richieste verranno poi soddisfatte quando ci saranno risorse disponibili
- permette di implementare il fair-share sull'utilizzo delle risorse tra diversi gruppi (tenant) e/o tra diversi utenti appartenenti allo stesso tenant.

Le funzionalita` di questo prototipo sono state dimostrate al meeting QUACK del 06/03/2014 al CNAF [8]. Una demo di questo prototipo verra` effettuata a un prossimo (pre)GDB WLCG (non e` stato possibile farlo al GDB di Marzo al CNAF, vista l'indisponibilita` del chairmain).

Servizio Keystone INFN Nazionale

Per permettere l'accesso e l'utilizzo di risorse cloud distribuite nelle varie sezioni INFN e' stato realizzato un cluster di tre server Keystone in tre sedi INFN: Padova, LNGS, e Bari. Il database

SQL utilizzato da Keystone è replicato sui tre siti utilizzando la soluzione Percona XtraDB Cluster.
HAProxy viene utilizzato per il load-balancing e come soluzione per la High Availability: l' “alta” disponibilità dei nodi di ha-proxy su base geografica è realizzata utilizzando un DNS multimaster dinamico (ha.infn.it) che può aggiornare automaticamente le sue mappe in base ai controlli di integrità effettuati sulle singole istanze.

Per questo servizio/cluster a Padova vengono usate 2 macchine: una per il servizio Keystone vero e proprio, e l'altra per HAProxy.

Nell’ambito del CLOUD-WG della CCR e' stata formulata e presentata un'ipotesi di prima architettura distribuita per la Cloud INFN che prevede un singolo dominio amministrativo, con possibilita di delega ai vari siti per la gestione di utenti, gruppi e progetti.
Tale architettura deve essere opportunamente validata, con test di funzionalità' dei domini, regioni e availability zone, test di performance, test per verificarne l’effettiva fault tolerance.
A tal fine a Padova sono state messe a disposizione alcune macchine per l'istallazione di un testbed Openstack che usera` il servizio Keystone nazionale distribuito.

Test sullo storage distribuito

La Sezione di Padova sta anche partecipando alle attivita` di testing del servizio object storage Swift, distribuito geograficamente in piu` siti (Padova, Bari, LNGS).
Per questi test vengono utilizzati 2 storage server piu` un proxy node.

E` stato inoltre predisposto l’hardware per test su CEPH che verranno effettuati sempre in ambiente distribuito [9].

EGI Federated Cloud Task Force

La sezione di Padova e` anche attiva nelle attivita` della EGI Federated Cloud Task Force.
Si sta predisponendo un’installazione Cloud (Openstack integrato con altri servizi, quali sensori di accounting, VOMS plugin, BDII, etc.) che sara` integrata nell’infrastruttura EGI Cloud [10].

Questa istanza Openstack e` distinta da quella di produzione della Cloud dell’area Padovana.

Personale coinvolto, organizzazione

Le persone coinvolte nelle attivita` sopra descritte sono impiegate nei servizi Calcolo e/o sono impiegati nelle attivita` del Tier-2 e/o sono coinvolte nelle attivita` Grid.
In dettaglio le persone coinvolte sono:

- Cristina Aiftimiei (Padova)
- Paolo Andreetto (Padova)
- Sara Bertocco (Padova)
- Massimo Biasotto (LNL)
- Alberto Crescente (Padova)
- Fulvia Costa (Padova)
- Alvise Dorigo (Padova)
- Sergio Fantinel (LNL)
- Eric Frizziero (Padova)
- Michele Gulmini (LNL)
- Michele Michelotto (Padova)
- Massimo Sgaravatto (Padova)
- Sergio Traldi (Padova)
- Marco Verlato (Padova)
- Lisa Zangrando (Padova)

Tutte le attivita` sopra descritte, sia quelle strettamente finalizzate all'implementazione della Cloud dell'area Padovana, sia quelle condotte nell'ambito del working group Cloud della CCR, vengono definite e coordinate in meeting settimanali.
Agenda e minute di questi meeting sono disponibili su Indico [11].

Riferimenti

[1] http://wiki.infn.it/_media/progetti/cloud-areapd/progetto-cloud-padovana.pdf

[2] http://wiki.infn.it/progetti/cloud-areapd/networking

[3] http://wiki.infn.it/progetti/cloud-areapd/best_practices/storage_external

[4] http://wiki.infn.it/progetti/cloud-areapd/registration

[5] http://wiki.infn.it/progetti/cloud-areapd/integration_of_the_infn-aai_authentication_in_the_openstack_keystone

[6] http://wiki.infn.it/progetti/cloud-areapd/home

[7] https://agenda.infn.it/getFile.py/access?contribId=32&sessionId=2&resId=0&materialId=slides&confId=6179

[8] https://agenda.cnaf.infn.it/conferenceDisplay.py?confId=620

[9] https://agenda.cnaf.infn.it/getFile.py/access?contribId=4&resId=0&materialId=slides&confId=620 (slide 20)

[10] https://wiki.egi.eu/wiki/Fedcloud-tf:ResourceProviders

[11] https://agenda.infn.it/categoryDisplay.py?categId=658