Progetto per la realizzazione di una Cloud per l'area Padovana

Versione 0.3.2
14 Ottobre 2013

Introduzione

Il modello di calcolo basato su paradigma GRID si e' rivelato di grande successo perche' ha permesso di creare dei centri di calcolo molto potenti ed aperti a molti grandi esperimenti INFN, in particolare gli esperimenti LHC.

Per i piccoli esperimenti invece il successo e' stato minore per via di diverse rigidita` presenti nel modello GRID, quali:

- La modalita` di autenticazione basata su certificati X.509
- Il numero limitato di sistemi operativi e in generale di ambienti di esecuzione supportati
- La curva di apprendimento molto ripida nelle fasi iniziali, che richiede all'inizio grandi investimenti di risorse umane per gestire problemi piu' facilmente risolvibili in modo tradizionale.

Il modello GRID, inoltre, se da un lato ben supporta una modalita` di calcolo di tipo batch, sostanzialmente non e` usabile per un accesso alle risorse di tipo interattivo, necessaria anche ai grandi esperimenti per le attivita` di analisi finale.

Quindi, oltre alla presenza di infrastrutture Grid-based, c'e` ora una proliferazione di cluster di calcolo dedicati ai vari gruppi/esperimenti, che sono spesso sottoutilizzati per grandi periodi, e sono al contrario non sufficienti a soddisfare i picchi di richieste che si concentrano in certi periodi limitati.
Questi cluster sono inoltre molto dispendiosi dal punto di vista della gestione sistemistica, e inefficienti anche dal punto di vista della gestione delle infrastrutture di condizionamento e protezione elettrica.

Il nuovo paradigma del Cloud computing si propone di gestire al meglio questi scenari: la Cloud permette di assegnare risorse in modo semplice e con grande dinamicita', permette all'utente di creare on-demand (e poi rilasciare quando non piu` usate) risorse di calcolo delle dimensioni desiderate (in termini di core, di memoria, di spazio disco) e con l'ambiente di esecuzione (sistema operativo, librerie, compilatori, ecc.) a cui l'utente e` abituato.

In questo modo si potrebbero unire in un'unica facility di calcolo le risorse dei diversi esperimenti e gruppi, evitandone il partizionamento statico al momento dell'installazione, e controllandone in modo elastico ed efficiente il loro impiego.
Con l'implementazione di una infrastruttura Cloud ci si pone quindi l'obiettivo di migliorare sia l’efficienza dell'uso delle risorse hardware (che a regime dovrebbe ricadere anche sugli acquisti dello stesso), sia del necessario manpower di gestione.

Realizzazione di una Cloud INFN tra la Sezione di Padova e i LNL

Con il presente progetto si intende dotare la sezione INFN di Padova e i Laboratori Nazionali di Legnaro di un servizio di Cloud computing e storage.
Obiettivo e` quello di implementare una stessa infrastruttura che possa essere utile per diversi aspetti del calcolo nelle due strutture INFN: supporto ad esperimenti, servizi e attivita' di progettazione, sviluppo e conduzione degli attuali e futuri acceleratori (ALPI, SPES).

Particolare beneficio di questa infrastruttura comune tra le due unita` INFN sara` per quegli utenti che hanno gia` necessita` di operare in entrambe le strutture e che auspicano una maggiore integrazione tra

i due diversi siti: un unico servizio a disposizione degli esperimenti e degli utenti permettera`, in ambo le sedi, un unico ed uniformato modo di accesso e di lavoro.

Una infrastruttura Cloud comune INFN LNL-Sezione di Padova proseguira` la storia di collaborazione tra le due strutture (il Tier-2 Legnaro-Padova degli esperimenti ALICE e CMS ne e` un esempio) e potra` beneficiare dell'esperienza gia` acquisita, nonche' di parte dell'infrastruttura già esistente.

In questa Cloud si potrebbero anche mettere a disposizione degli specifici applicativi utilizzabili eventualmente anche da utenti INFN non Padovani. Questo permetterebbe un risparmio sulle spese di site license per quelli applicativi usati nelle varie Sezioni da un ristretto numero di persone (es. le NAG LIBRARY).

Implementazione

L'obiettivo e` quello di implementare una "Cloud di produzione" in tempi brevi, sfruttando al massimo software gia` esistenti (versioni stabili).
A tal fine si utilizzera` in particolare il software OpenStack, uno dei middleware Cloud opensource piu` usati (anche il Cern e molti altre sezioni o laboratori hanno deciso di adottare questo middleware Cloud per implementare servizi di Cloud).
Componenti innovative o particolari customizzazioni, sviluppate in particolare all'interno del Cloud Working Group della CCR dell'INFN (con cui si intende collaborare attivamente) verranno messe in produzione quando ritenute sufficientemente mature.

Fisicamente le risorse di questa Cloud saranno distribuite su entrambe le sedi, se pure separate dall'attuale istanza di produzione Tier-2. La Cloud così costituita potrebbe beneficiare delle fibre che oggi uniscono le sale computing del Tier2 nelle due sedi, di parte dell’infrastruttura di rete (router e switch almeno in una primissima fase), dell’esperienza acquisita nella gestione di risorse distribuite e non ultimo l’utilizzo di alcune risorse hardware sia di computing che di storage del Tier2, che pur essendo fuori manutenzione sono ancora funzionanti e quindi usabili.
Una stima di massima delle dimensioni che questa infrastruttura Cloud che si vuole implementare dovrebbe raggiungere nel breve-medio periodo sono:

- 900-1200 CPU-cores da usare sia per i servizi Cloud openstack (“controller node” e “network node”), sia per i compute node
- O(100) TB di storage, da utilizzarsi sia per i servizi Cloud (storage per le immagini e per le istanze), sia per storage per gli utenti. Sara` valutato nelle fasi iniziali del progetto i servizi di storage piu` adatti a soddisfare gli use case degli esperimenti (block storage attraverso il servizio Cinder e/o object storage via swift e/o un qualche file system distribuito, etc.)
- Connessioni di rete a 10Gbps tra compute server e storage

Parte delle risorse di questa infrastruttura distribuita saranno destinate ad attivita` di R&D, in numero allocabile in modo assolutamente dinamico a seconda delle necessita`.

Come sopra descritto si intende collaborare attivamente con il working group CLOUD della CCR, in modo anche che la Cloud dell'area Padovana possa poi essere integrata in una futura Cloud INFN Nazionale.

Esperienze

Esperienze su OpenStack sono gia` presenti tra il personale della Sezione di Padova.
In particolare il gruppo che fino ad ora si e` occupato delle attivita` Grid in Sezione ha fatto il deployment di una piccola infrastruttura OpenStack based, che e` ora utilizzata per ospitare i testbed per lo sviluppo e la certificazione del software Grid (in particolare CREAM).

Ci sono esperienze in Sezione anche per quel che riguarda l'interfacciamento tra Cloud e framework di job submission dell'esperimento CMS, dove la parte di resource provisioning e` ben distinta dal resource scheduling. Il tutto e` implementato attraverso i cosiddetti “pilot job”, che sono sottomessi alle risorse Grid in modo da creare un “overlay batch system” su cui poi vengono eseguiti i job degli utenti. CMS sta attualmente esplorando la possibilita` di usare risorse Cloud oltre che Grid: anziche` sottomettere pilot job su risorse Grid, vengono istanziate on-demand macchine virtuali che eseguono i pilot job. Un testbed OpenStack e` stato implementato, e questo e` stato utilizzato per dimostrare il funzionamento di questo modello, sottomettendo job di analisi via CRAB2.

Il gruppo Grid di Sezione e` anche attivo in diverse attivita` coordinate dal working group Cloud della CCR, in particolare su tematiche che coinvolgono la gestione delle immagini (inclusi test di un servizio OpenStack Glance replicato), di test di GlusterFS in wide area, dell'integrazione con il sistema di autenticazione e autorizzazione INFN AAI.
Da registrare anche attivita`, gia` presentate al working group Cloud, relative all'implementazione di un overlay SLURM batch system (la cui dimensione varia dinamicamente in base al workload da eseguire), usando risorse che espongono un'interfaccia Grid o Cloud.

Il know-how maturato ai Laboratori Nazionali di Legnaro si concentra invece in particolare sulle problematiche relative allo storage, per la gestione di quantita` di dati sempre piu` elevate a cui e` richiesto un accesso efficiente. Per questo il servizio calcolo dei Laboratori ha da tempo maturato esperienza nell’ambito dei file system distribuiti. E` attualmente in fase di finalizzazione l'implementazione di un servizio di produzione di circa 55TB, basato su GlusterFS, inizialmente accessibile all’esperimento Galileo e ad un paio di altri piccoli esperimenti.

Attivita` di R&D

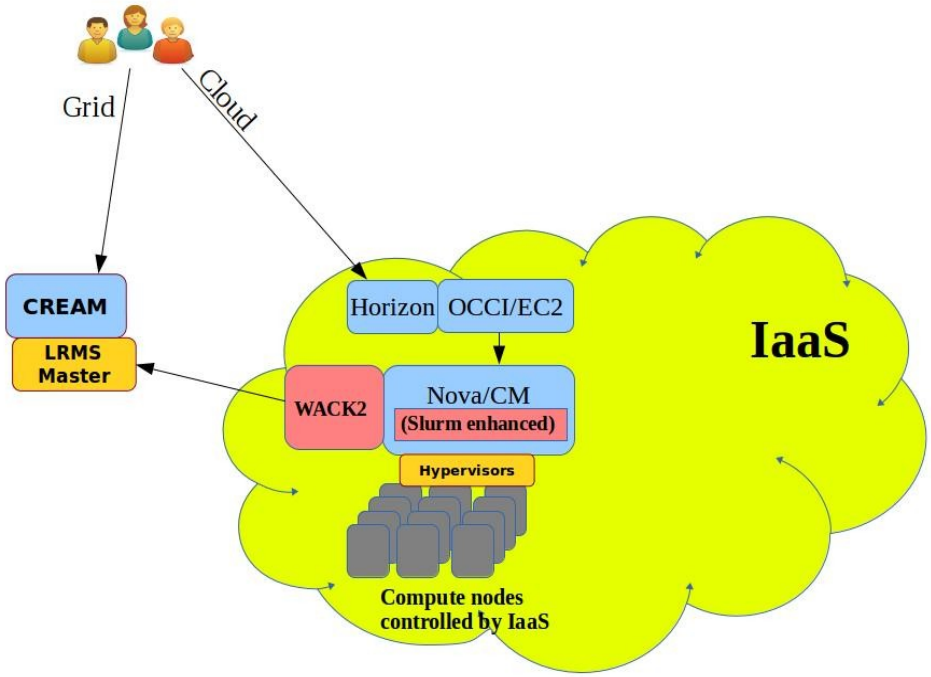
Nell'implementazione di una infrastruttura di Cloud si vuole sfruttare al massimo software esistente, evitando quindi pesanti (re)implementazioni.
Tuttavia specifiche customizzazioni e nuove funzionalita` dovranno essere implementate e integrate, in quanto non disponibili.
Si intende lavorare, compatibilmente con la disponibilita` di personale, su questi aspetti di ricerca e sviluppo in stretta collaborazione con il working group Cloud della CCR.

L'integrazione con il sistema di autenticazione e autorizzazione INFN AAI e con altri identity provider e` una delle aree in cui si intende operare.
L'obiettivo e` quello di semplificare le modalita' di accesso alle risorse di calcolo da parte delle comunita' scientifiche, permettendo l'autenticazione ai servizi di OpenStack tramite INFN-AAI (o altro identity provider, es. IDEM), a fianco del tradizionale accesso via username/password.
C'e` gia` molta attivita` a riguardo da parte di diversi soggetti (con cui si intende cooperare), ma non

c'e` ancora una soluzione matura pronta per essere utilizzata.

Un'altra area in cui sara` necessario qualche sviluppo riguarda la gestione delle risorse, fornite da diversi esperimenti e che non sono infinite: se da un lato si vuole assicurare massima priorita` nell'utilizzo delle “proprie” risorse da parte dei diversi utenti, si vuole anche garantire un uso ottimale delle risorse, che non sarebbe garantito con un partizionamento statico (implementato esempio attraverso quote) delle stesse.
Un possibile approccio che verra` esplorato per questo problema, che permetta di gestire in maniera efficiente il problema del fair-share dell'utilizzo delle risorse e la gestione delle richieste di risorse che non possono essere immediatamente soddisfatte, e` l'integrazione di un algoritmo di scheduling tipico di un resource management system (quale SLURM) nel componente Nova di OpenStack.
Si intende lavorare su questo aspetto in collaborazione in particolare con gli sviluppatori del sistema WACK (l'integrazione tra WNODES e OpenStack).

Altro aspetto su cui si intende essere attivi (anche se con priorita` piu` bassa rispetto alle altre problematiche sopra esposte) riguarda l'integrazione tra Grid e Cloud, per permettere l'accesso allo stesso set di risorse sia attraverso un'interfaccia Grid (implementata attraverso un CREAM Computing Element) che un'interfaccia Cloud.
Per risolvere questo problema, una possibile architettura che sara` esplorata e` quella mostrata nella seguente figura:



Quindi per quel che riguarda l'interfacciamento con Grid, questa sarebbe implementata dal componente in figura chiamato “WACK2”, responsabile di “prelevare” i job dalla coda del batch system (qui sottomessi da CREAM), istanziando macchine virtuali (o bare metal) via Openstack nova, che verrebbero configurati come worker node del batch system e dove quindi i job verrebbero eseguiti.
Anche in questo caso si intende collaborare in particolare con il team di sviluppo di WNODES/WACK.

Integrazione con il Tier-2

Non e` prevista, almeno in una fase iniziale, l'integrazione tra la le risorse della Cloud di produzione qui descritta e quelle del Tier-2 di Legnaro-Padova.
Questa potra` essere implementata quando una soluzione production-ready per l'integrazione tra Grid e Cloud sara` disponibile, se ci sara` ancora l'esigenza per gli esperimenti ALICE e CMS di una interfaccia GRID per poter utilizzare le risorse del Tier-2.

Sara` invece esplorata la possibilita`, per gli utenti ALICE e CMS della Cloud di produzione, di poter accedere in maniera diretta ed efficiente allo storage del Tier-2 (es. fornendo accesso attraverso il protocollo dcap allo storage dcache di CMS).

Collaborazione con l'Universita` di Padova

Anche il dipartimento di Fisica e Astronomia dell'Universita` di Padova sta pensando di dotarsi di un sistema di calcolo basato su Cloud, in collaborazione con altri dipartimenti dell'Universita' e con il centro di calcolo di ateneo (CCA). Sara` valutata una possibile integrazione tra questa infrastruttura e la Cloud INFN dell'area padovana.