Realizzazione di alta affidabilità per componenti di OpenStack

Marco Caberletti
INFN-CNAF

Overview

- Descrizione infrastruttura OpenStack al CNAF
- Soluzioni per servizi in Alta Affidabilita’
  - Corosync & Pacemaker
  - MySQL
  - RabbitMQ/QPid
  - Keystone
  - Glance
  - Servizi Nova
  - Quantum
- Sviluppi futuri

OpenStack@CNAF - 1

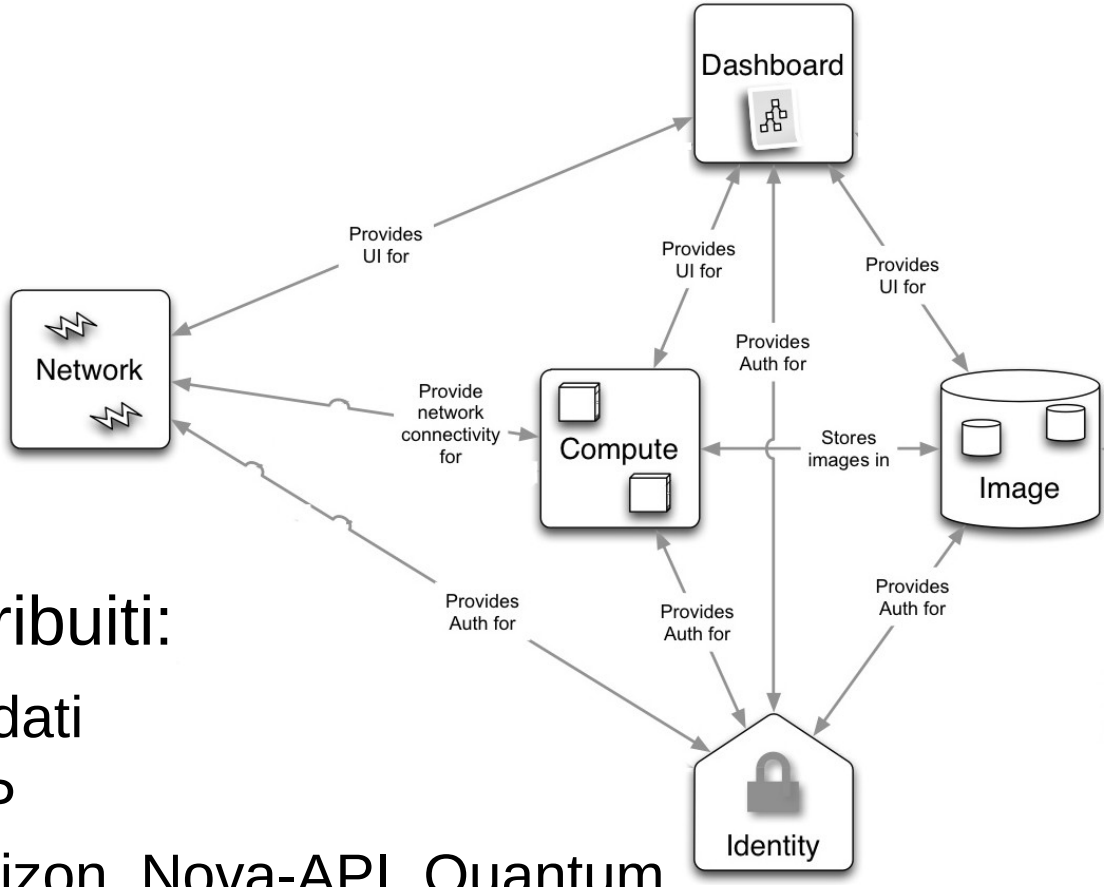
- Nel CdC di aprile si è deciso di installare un Cloud IaaS basato su OpenStack per il provisioning di VM agli utenti interni al CNAF (VM per sviluppo, test, ...)
  - Servizio di **PRE-PRODUZIONE**
  - Attività trasversale a diversi reparti del CNAF
- L'idea è partire dall'esperienza maturata in MarcheCloud:
  - Setup base per acquisire competetenze nella gestione dell'infrastruttura
  - Fornire un primo set di servizi agli utenti CNAF

OpenStack@CNAF - 2

- L'infrastruttura, sebbene di pre-produzione, deve essere disegnata in **alta affidabilità**:
  - Ogni componente deve essere almeno ridondato
- Verrà usata l'ultima release di OpenStack, denominata **Grizzly**
- Seguirà la messa in produzione ed estensione dell'infrastruttura:
  - Provisioning di servizi Nazionali ed Internazionali
  - Cloud IaaS per utenti INFN in generale

Schema generale

- 5 macro componenti:
  - Dashboard (Horizon)
  - Image (Glance)
  - Identity (Keystone)
  - Compute (Nova)
  - Network (Quantum)
- Iniziale saranno così distribuiti:
  - Due cloud controller ridondati
    - MySQL e server AMQP
    - Keystone, Glance, Horizon, Nova-API, Quantum
  - Due Compute Nodes
    - Hypervisor KVM, nova-compute e quantum-agent



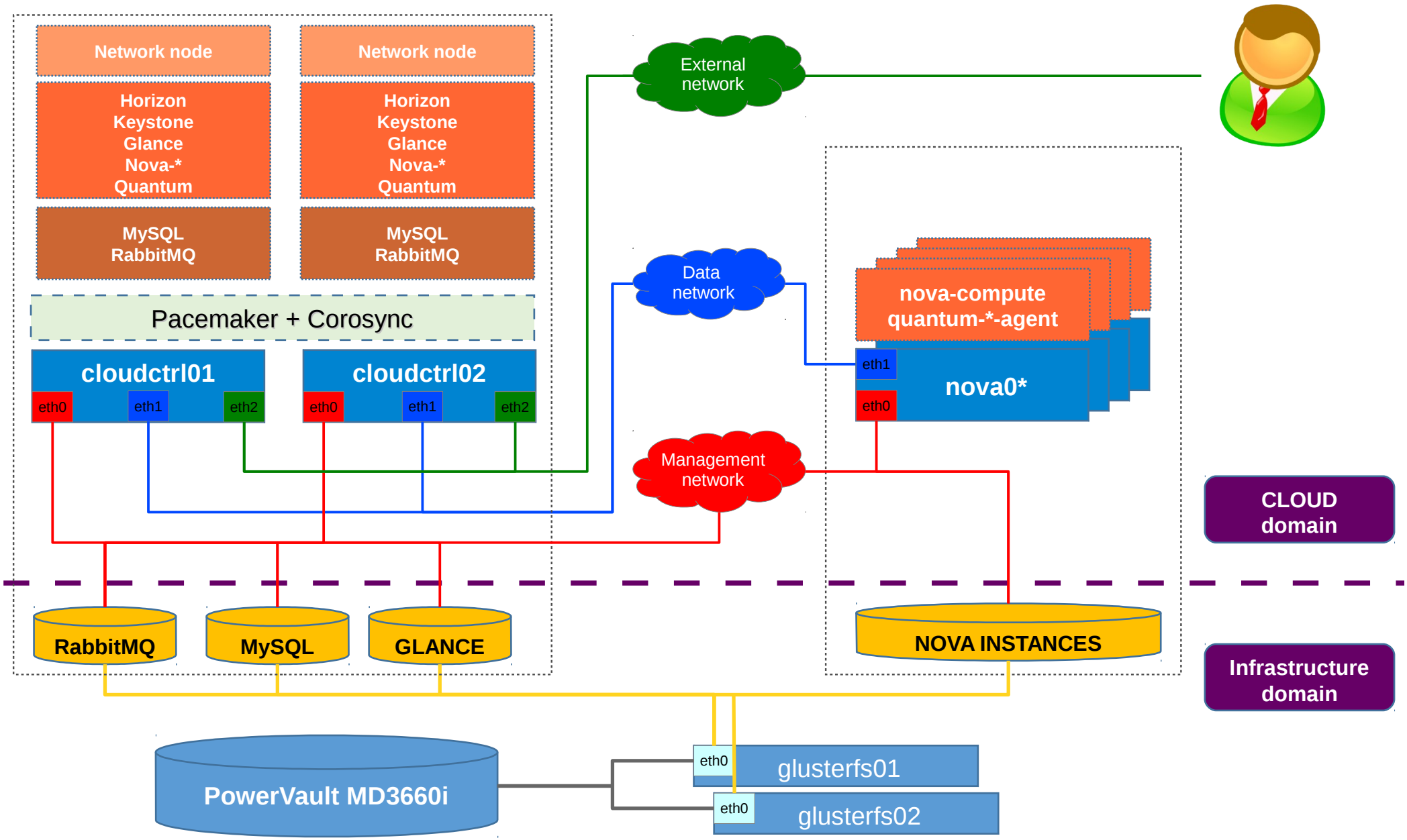
Quantum

- Nelle prime release di OpenStack, la rete veniva gestita con il modulo *nova-network*
- Dalla release Folsom è disponibile Quantum, componente che implenta il concetto di **“Network as a Service”**
- Quantum consente setup di rete complessi e di interfacciarsi direttamente a device di rete fisici
- Ha un'architettura **modulare a plugins**:
  - Grande flessibilità e facilità di espansione
- Tuttavia, intruduce **complessità** ed è un servizio relativamente “giovane”

Setup di rete

- Per l'infrastruttura CNAF, si è scelto di usare una conifigurazione del tipo “Provider Router con reti private”
- **Ogni tenant può disporre di più reti private** che connette al mondo esterno mediante un *quantum router*
- C'è un solo router virtuale, gestito dall'amministratore, che può mettere in comunicazione VM di tenant diversi
- Le **subnet sono univoche**, non è permesso overlap degli indirizzi IP

FASE I – Infrastruttura di base



Elementi caratterizzanti l'infrastruttura

- **GlusterFS**
  - Filesystem distribuito e parallelo, usato per condividere dati fra i due controller ridondati
  - Esporta l'archivio delle immagini sui controller e mette in condivisione fra i compute note la directory delle istanze running (per *live migration*)
- **Corosync**
  - Messaging layer per creare cluster affidabili
  - Usa protocolli di membership e quorum message-based
- **Pacemaker**
  - Interagisce con le applicazioni mediante *resource agents (RA)*
  - OpenStack fornisce i RA per i propri servizi

Red Hat RDO

- RDO è una distribuzione di OpenStack per sistemi RedHat-based
- Fornisce un **proprio repository** con l'ultima release di OpenStack
- Il tool ***packstack*** automatizza e semplifica notevolmente l'installazione e configurazione dei vari componenti di OpenStack
  - Installazione interattiva o con *answer file*

HA per MySQL

- Analizzate e testate diverse soluzioni
  - Percona XtraDB con HAProxy
  - Galera
  - MySQL Cluster community edition
- Condiderato che OpenStack non carica in modo eccessivo il database, si è optato per:
  - Uso di **MySQL standard**
  - Setup **Attivo/Passivo** con la directory dei DB frame condivisa fra le due istanze
- Questo è il setup ad oggi consigliato da OpenStack

HA per RabbitMQ

- Analizzate e testate le due possibili soluzioni:

1) Code persistenti
   - Le code e il loro contenuto viene sincronizzato su file
   - Le due istanze Attivo/Passivo condividono la directory con questi file
2) Cluster con queue mirroring
   - Più istanze, tutte attive, di RabbitMQ replicano fra loro le code e i messaggi contenuti

- Abbiamo optato per l'uso delle **code persistenti**, soluzione ad oggi consigliata da OpenStack

Cloud Controller in HA

- Tutti i servizi del Cloud Controller (Keystone, Glance e i servizi di Nova) sono configurati nel medesimo modo
- **Ogni servizio è in modalità Attivo/Passivo** ed ognuno viene associato ad un **Virtual IP** (VIP)
- Ogni servizio viene indirizzato esclusivamente mediante il VIP
- Se l'Attivo fallisce, Pacemaker sposta il VIP e avvia il servizio sull'altro nodo

Quantum in HA

- Quantum ad oggi richiede **necessariamente** la configurazione **Attivo/Passivo** per il componente che esegue il NAT/routing (*quantum-l3-agent*) e per i plugins
- Come per il Cloud Controller, Pacemaker mediante i RA si occupa di spostare l'esecuzione dei componenti assieme al rispettivo VIP

Workflow

- Dalla nostra esperienza è emerso che è preferibile il seguente workflow:
  1. Creare il cluster di controller con Corosync/Pacemaker
  2. Inserire i VIP di tutti i servizi
  3. Mettere in standby *n*-1 nodi controller
  4. Sull'unico controller attivo installare tutto via *packstack*
  5. Ripetere l'installazione sugli altri nodi
  6. Inserire i servizi in Pacemaker facendo attenzione a dipendenze e priorità
  7. Riattivare i nodi

Questioni aperte

- Al momento della creazione di una VM, OpenStack assegna un **MAC address pseudo casuale**
  - Questo aspetto complica la gestione delle VM, per le modalità usate al CNAF nella gestione delle macchine
  - Al CERN hanno risolto con una modifica del driver di *nova-network*, ma la loro soluzione non può essere riutilizzata al CNAF
  - Si tratta di modificare alcuni moduli affinchè il MAC non venga generato random, ma prelevato da una lista (es. file o db) specificata dall'amministratore
- Richiesta di **autenticazione via AAI INFN**
  - Problema mappatura dinamica utenti-tenant e rispettivi privilegi

Sviluppi futuri

- Integrazione nell'infrastuttura di Swift e Cinder per fornire un servizio di Cloud Storage
- Utilizzo di *cells* e *availability zone* di OpenStack per includere datacenter esterni al CNAF
  → unico cloud IaaS INFN
- Valutazione ed eventuale introduzione dei nuovi componenti OpenStack:
  - Es. Heat e Ceilometer
- Introduzione nuovi servizi nell'infrastruttura a seconda dei requisiti evendenziati dagli utenti del CNAF