Realizzazione di una infrastruttura Cloud pilota basata su OpenStack

Paolo Veronesi (INFN-CNAF/IGI)

Livio Fano' Illic (INFN-PERUGIA)
Enrico Fattibene (INFN-CNAF/IGI)
Matteo Manzali (INFN-CNAF)
Hassen Riahi (INFN-PERUGIA)
Davide Salomoni (INFN-CNAF)
Andrea Valentini (INFN-PERUGIA)
Valerio Venturi(INFN-CNAF/IGI)

Overview

- Il progetto MarcheCloud
- WP1 - L'infrastruttura pilota
  - Architettura e servizi OpenStack utilizzati
  - Features di interesse generale testate
- WP2 – Monitoring e Allarmistica
  - Categorie di dati
  - Analisi e soluzione
- Sviluppi futuri

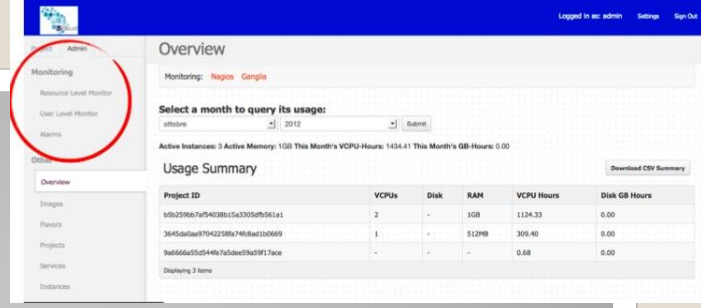
Marche Cloud

- Una soluzione pilota per la fornitura flessibile ed innovativa di servizi al cittadino:
  - **Con strumenti Open Source**
    - Cf. Decreto Sviluppo 2012, art. 22 comma 10: "Solo quando la valutazione comparativa di tipo tecnico ed economico dimostri l'impossibilità di accedere a soluzioni open source o già sviluppate all'interno della pubblica amministrazione ad un prezzo inferiore, è consentita l'acquisizione di programmi informatici di tipo proprietario mediante ricorso a licenza d'uso."
  - **Adattabile** alla fornitura di nuove applicazioni o prodotti
  - **Interoperabile** attraverso interfacce standard con fornitori pubblici o privati
  - **Fruibile** con moderne interfacce utente
  - **Espandibile** verso una infrastruttura per l'offerta di risorse di calcolo, storage, software as a service da parte di soggetti pubblici o privati

Caratteristiche di Marche Cloud

- Cloud computing attraverso OpenStack per virtualizzazione calcolo, network e storage. L'infrastruttura pilota deve essere ridondante e auto-consistente, integrabile in installazioni e infrastrutture pre-esistenti (**WP1 – INFN-CNAF/IGI**)
- Monitoraggio e allarmistica integrati ed espandibili (**WP2 – INFN-PERUGIA**)
- Autenticazione alle risorse attraverso sistemi locali o federati (WP3 – Università di Camerino)
- Interfacce utente per software pilota di accesso a referti on-line Regione Marche con sistema operativo Android o attraverso Web TV (WP4 - Università Politecnica delle Marche)

Architettura



Applicazioni e utenti

Infrastruttura virtuale pilota
(sviluppabile su **MCloud.Gov [pubblico]** e **MCloud.B&R [privato]**)

OpenStack: servizi di identità, gestione calcolo, storage, monitoring; virtualizzazione di rete, sistemi e applicazioni

Calcolo | Dati | Rete

Infrastruttura fisica

Interoperabilità con altri software o infrastrutture Cloud

WP1

- Definizione dell'infrastruttura pilota basata su OpenStack
  - (**WP1 – INFN-CNAF/IGI**)

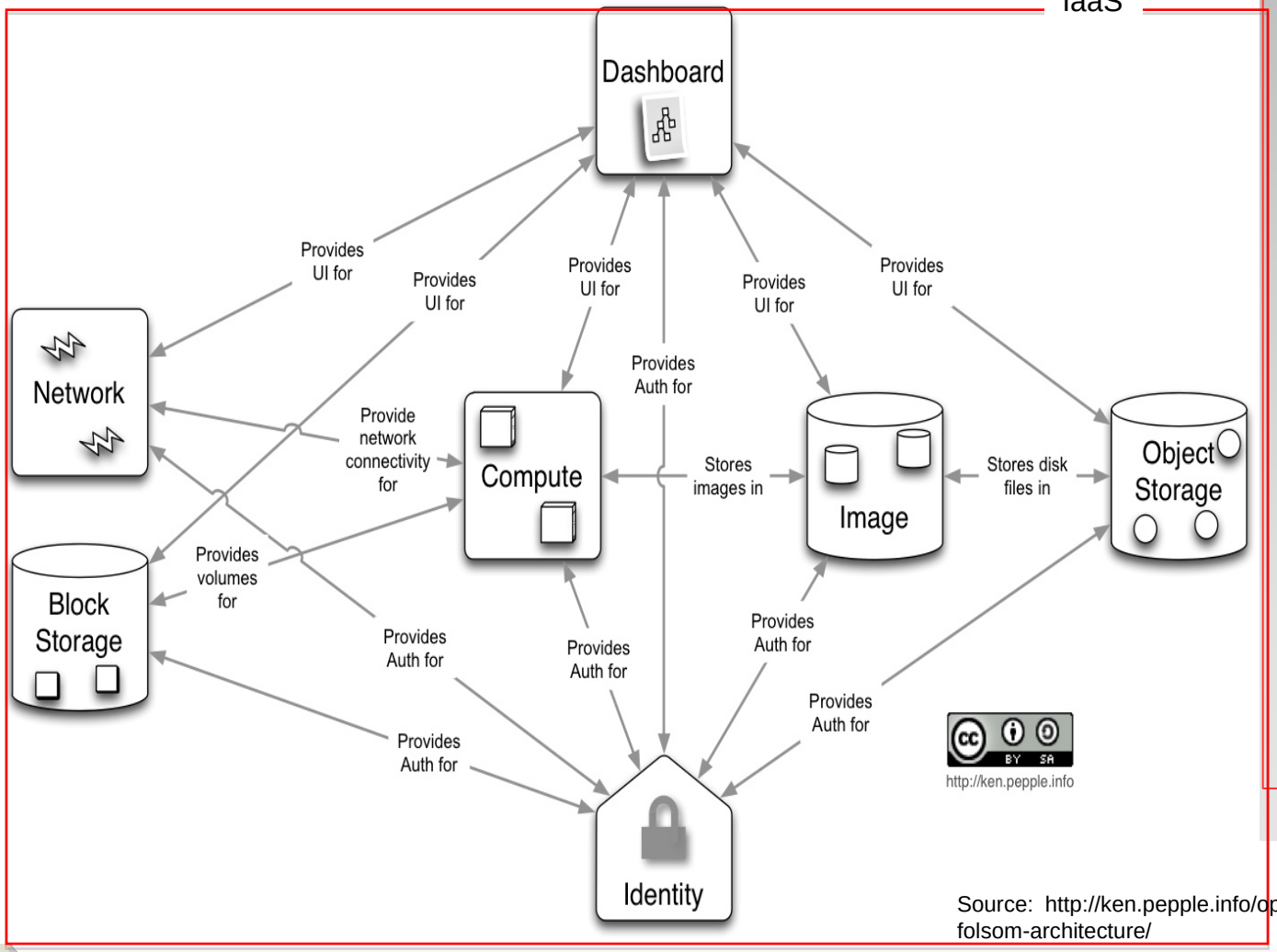
OpenStack

- OpenStack e' una suite di servizi open source per la costruzione di infrastrutture Cloud private e pubbliche di tipo IaaS
- Il progetto nasce nel 2010 come unione di due progetti separati con funzioni specifiche: Nova, cioè il gestore della parte di computing, e Swift, un gestore di tipo object storage.
  - ora sono ben 150 le compagnie che hanno aderito al progetto
    - Tra i membri platino (oltre alle due fondatrici) sono SUSE linux, Red Hat, IBM, HP, Nebula e AT&T
    - Tra imembri oro spiccano Intel, Cisco, Dell e Yahoo!.
- Il software è completamente libero e viene rilasciato sotto la licenza Apache.

OpenStack Leadership's vision statement
"essential Infrastructure, support platform"

I componenti di OpenStack per una infrastruttura IaaS

IaaS



IaaS

- **Identity (Keystone)**
- **Image Catalog (Glance)**
- **Compute (Nova)**
- **Network (Nova-Network/Quantum)**
- **Dashboard (Horizon)**
- **Storage (Nova-Volume/Cinder)**
- Object Storage (Swift)

Source: http://ken.pepple.info/openstack/2012/09/25/openstack-folsom-architecture/

Keystone (Identity Service) e Glance (Image Service)

Keystone unifica diversi progetti core in un unico sistema di autenticazione. Vengono fornite autorizzazioni per credenziali di log-in multiple, garantendo login token-based e in stile AWS.

- Fornisce inoltre funzionalita' di catalogo dei servizi e policy

- Come identity provider supporta diversi backend (non ancora saml, shibboleth)

- **Glance,** progettato per essere un servizio standalone, fornisce un servizio di catalogo per la memorizzazione e l'interrogazione di immagini.
- L'architettura di Glance è composta da tre parti: le API, il registro e l'image store.
- Supporta immagini di vari formati: Raw, Machine (kernel/ramdisk outside of image, a.k.a. AMI), VHD (Hyper-V), VDI (VirtualBox), qcow2 (Qemu/KVM), VMDK (VMWare), OVF (VMWare, others)

Nova-*

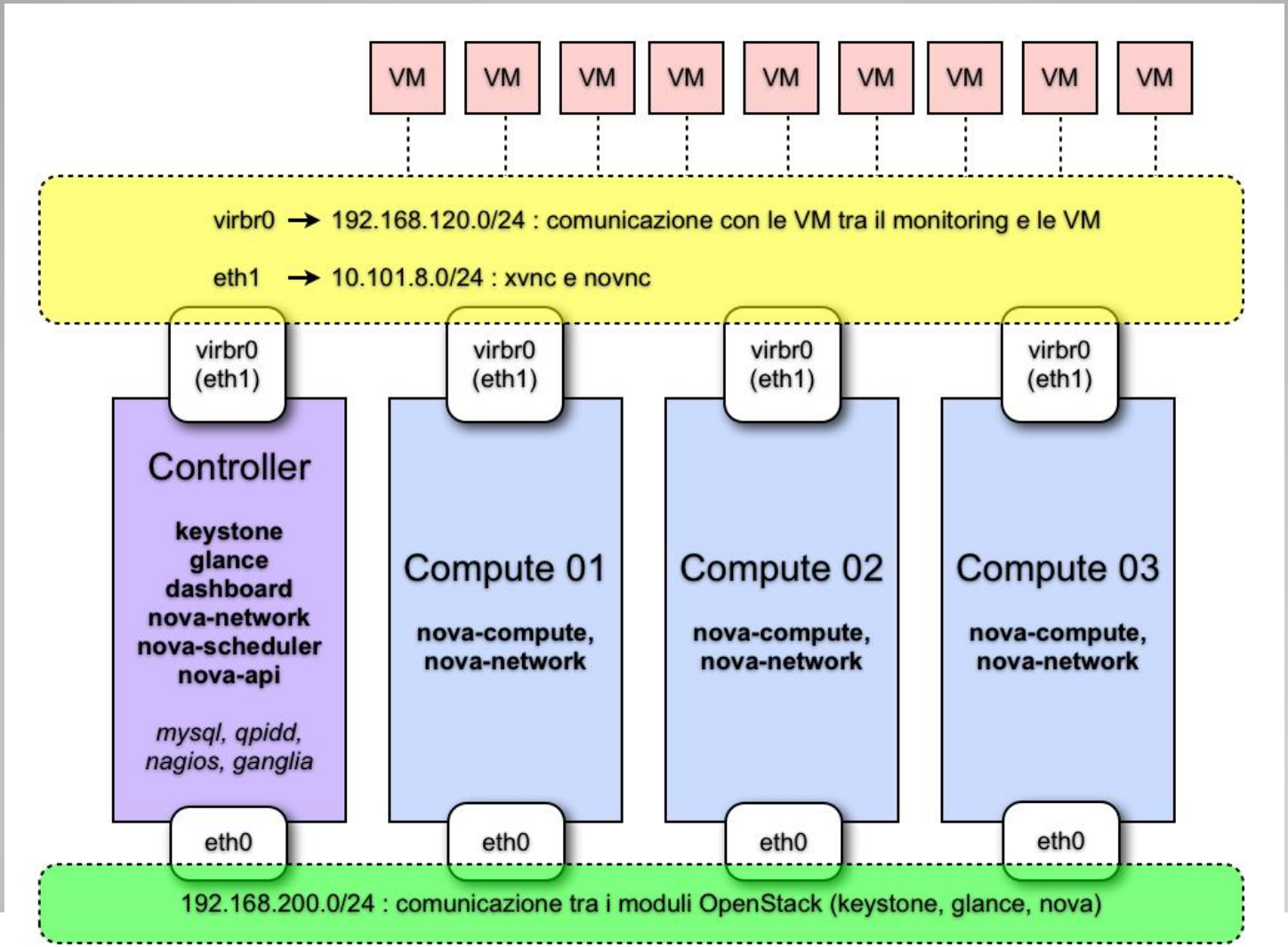
- Nova fornisce un framework per il provisioning e la gestione su larga scala di istanze di calcolo virtuali. Simile nelle sue funzionalità e scopo al servizio Amazon EC2, permette di creare, gestire, ed eliminare server virtuali basandosi sulle proprie immagini di sistema attraverso un'API programmabile.
- Supporta diversi tipi di hypervisor (XenServer/XCP, KVM, QEMU, LXC, ESXi, Hyper-V), ma alcuni con funzionalita' limitate (dettagli in http://wiki.openstack.org/HypervisorSupportMatrix)
- Il componente **nova-network/Quantum** fornisce funzionalita' di rete avanzate (Flat, Flat DHCP, VLAN DHCP, IPv6, floating IPs)
- Il **cloud controller** si occupa dell'orchestrazione della comunicazione dei vari componenti (basato su AMQP)
- Il database archivia gli stati di configurazione e run-time per l'infrastruttura cloud (i tipi di istanza che sono disponibili per l'uso, le istanze in uso, le reti disponibili, i progetti, utenti, etc).

Swift vs nova-volume/Cinder

- **Object Storage**
- Può interfacciarsi con keystone
- Garantisce alta affidabilità implementando meccanismi di replicazione e distribuzione
- Da considerare come un distributed storage system, non come un file system, supporta S3 API

- **Block Storage**
- Può interfacciarsi con keystone
- Non garantisce alta affidabilità (demandata al filesystem sottostante)
- Permette di collegare i volumi alle macchine virtuali

L'infrastruttura Pilota



Funzionalità disponibili nell'infrastruttura

- Volumi persistenti
- Live migration
- Floating Ips (asssegnazione dinamica di ip pubblici)
- Snapshot

Sulla contestualizzazioni delle immagini

- Oz (tool per la creazione automatizzata di immagini)
  - http://www.aeolusproject.org/oz.html

VOLUME PERSISTENTE

- Attraverso **nova-volume** è possibile creare un volume persistente e renderlo visibile e modificabile ad un'istanza
- Il volume inizialmente non è formattato e non viene montato automaticamente
- Il volume è indipendente dall'istanza e può essere associato ad una sola istanza per volta
- Le modifiche fatte al volume vengono mantenute anche dopo la terminazione dell'istanza e sono visibili ad una successiva istanza a cui viene associato il volume stesso
- E' possibile lanciare un'immagine da un volume, invece che dal repository glance

LIVE MIGRATION

- Con live migration si intende la possibilità di migrare una VM da un compute node ad un altro senza discontinuità di servizio
- Questo meccanismo è fornito da nova attraverso un semplice comando
- Per poter effettuare questa operazione è necessario che la directory contenente le istanze sia condivisa tra tutti i compute node (o almeno tra i compute node tra cui si vuole effettuare la migrazione)
- Questa operazione è utile per fare ad esempio manutenzione su un compute node

FLOATING IP

- In OpenStack viene fornita la possibilità di associare in maniera dinamica un determinato IP ad una macchina virtuale, questo tipo di indirizzo viene chiamato “floating”
- E’ possibile definire più gruppi (pools) e specificare da quale gruppo deve essere scelto l’indirizzo IP
- Si può scegliere se associare in automatico l’indirizzo alla VM durante la sua creazione oppure aggiungerlo/rimuoverlo manualmente mentre la VM è in esecuzione
- In questo modo risulta immediato sostituire la macchina virtuale che risponde ad un determinato indirizzo pubblico con un’altra macchina virtuale, ad esempio a causa di un malfunzionamento

SNAPSHOTS

- In OpenStack le istanze sono per definizione volatili, una volta terminata un'instanza vengono perse tutte le modifiche apportate
- Per ovviare a questo problema è stato implementato il meccanismo di snapshots, una feature fornita da nova
- E' possibile in qualsiasi momento creare una fotografia (snapshot) della macchina virtuale e inserire questa nuova immagine nel repository di glance
- Infine si possono istanziare nuove macchine virtuali a partire dallo snapshot creato

WP2

- Monitoraggio e allarmistica integrati ed espandibili (**WP2 – INFN-PERUGIA**)

Categorie dei Dati

infrastrutturali (da dashboard di Openstack)

- quanti e quale istanze vengono consumate dagli utenti ?
- quali utenti consumano cosa ?
- a quali progetti/gruppi le risorse/utenti appartengono ?
- prestazioni dell'istanza
- disponibilità delle risorse

virtualizzatore (da KVM)

- efficienze interne delle singole istanze
- consumi specifici (cpu, memoria, storage, banda...)
- allarmistica
- sub-networking

Categorie dei Dati

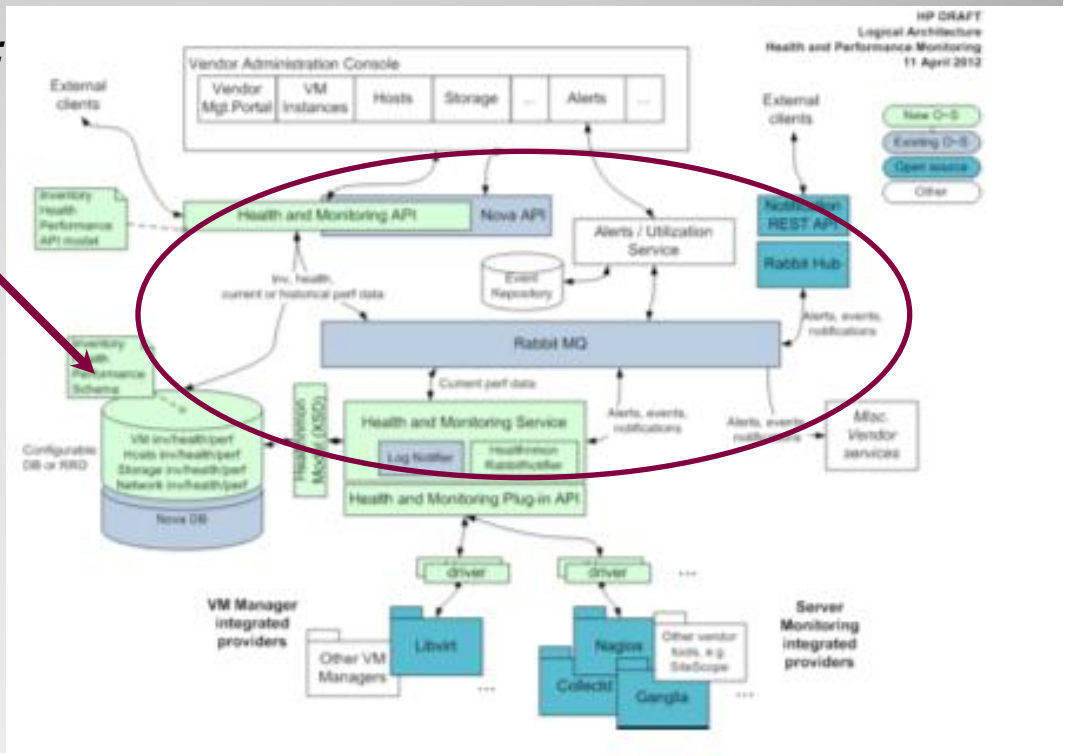
infrastrutturali

AMQP/Nova

Principalmente legato alle istanze:

+ quante/quali
+ metadati dell'istanza
+ disponibilità
+ storico su DB-Nova interno

virtualizzatore



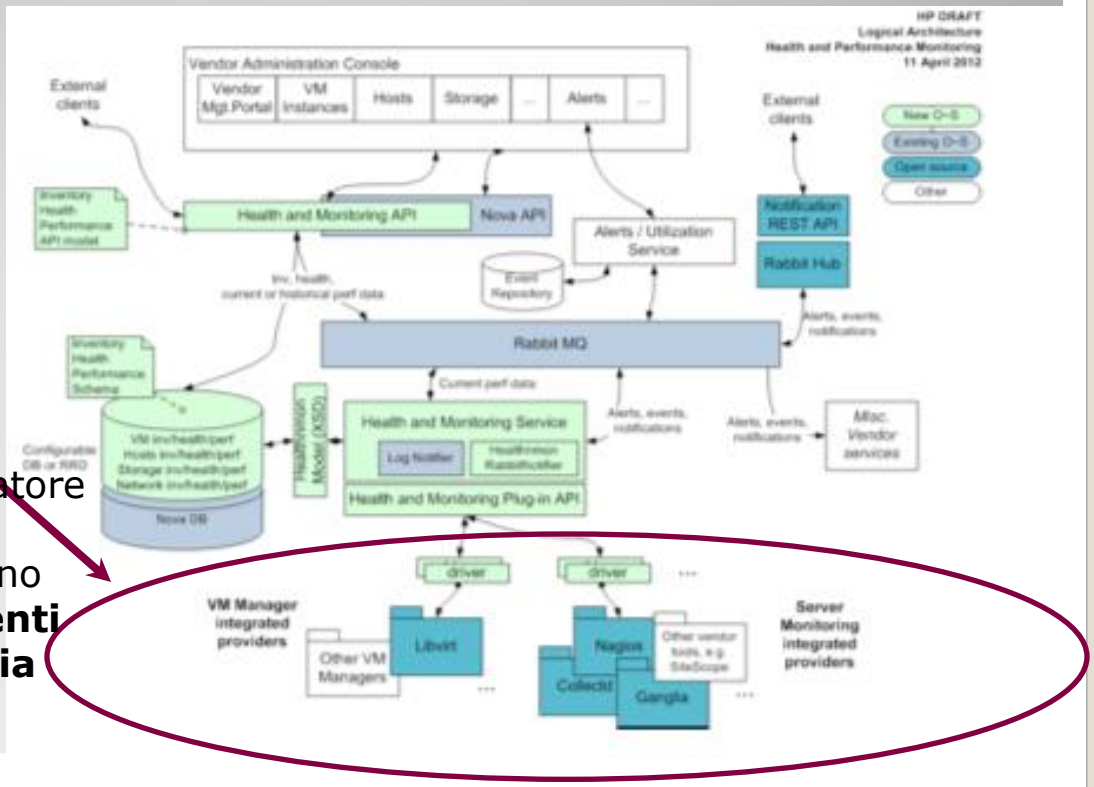
Categorie dei Dati

infrastrutturali

virtualizzatore

Il monitor di Openstack non interfaccia direttamente il virtualizzatore

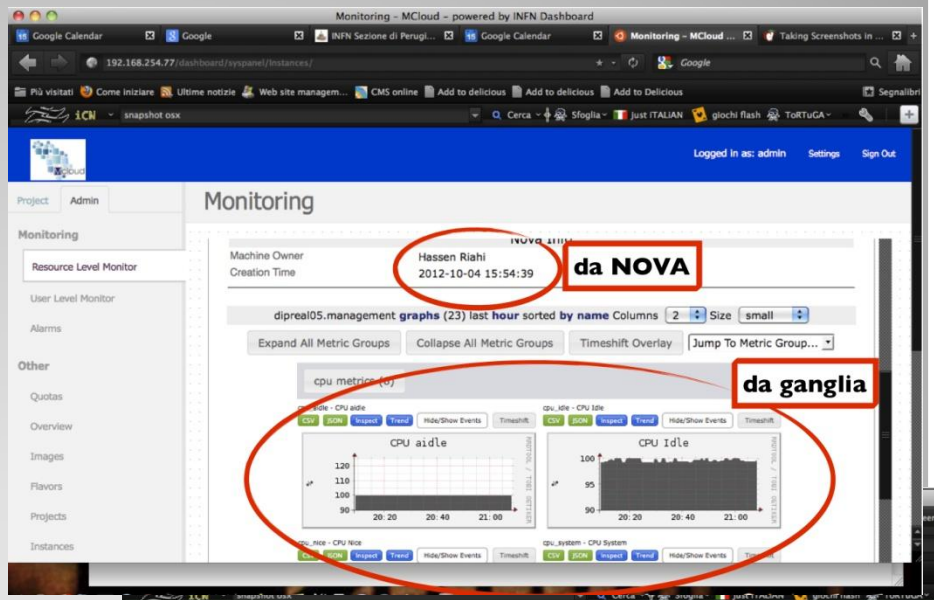
espone però delle API che permettono l'interfacciamento a **tool indipendenti sia tramite plugin specifici che via libVirt**

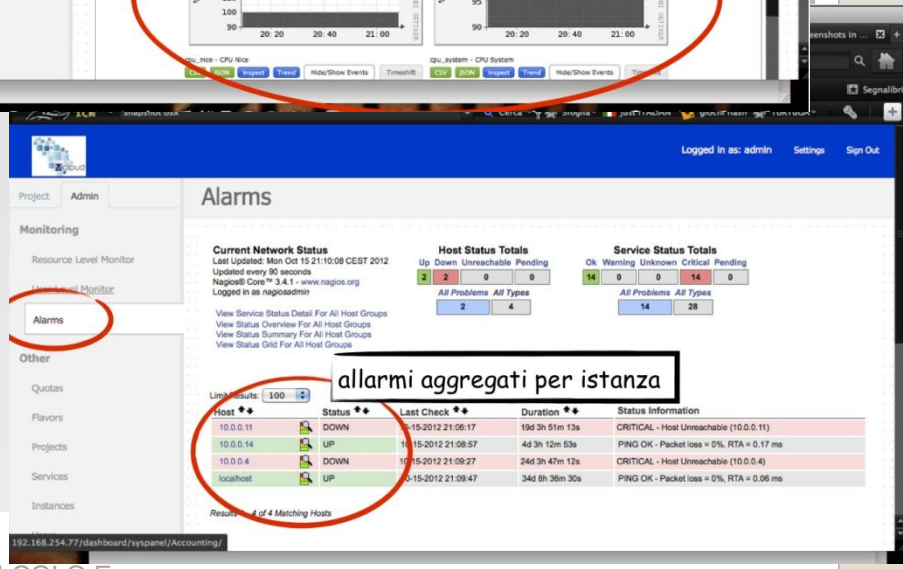


Analisi soluzioni

Ganglia + Nagios è la soluzione scelta per un monitor completo della Cloud

- ➡ Molto diffusi in combinazione con Openstack
- ➡ Minimizza lo sforzo di integrazione grazie ai numerosi plugin
- ➡ Espandibili
- ➡ Richiede una *customizzazione*, anche se minima, delle immagini virtuali (sono soluzioni *demon-based*)





Sviluppi futuri

- Collaborazione con progetti PON come Prisma, partecipazione ad altri bandi Smart Cities
- Ri-utilizzo delle competenze architetturali e tecnologiche acquisite con Marche Cloud
- Interoperabilità con altri Cloud provider e con altri stack software
- Definizione di una architettura pienamente ridondata
- Definizione di reti virtuali dinamiche
- Alta affidabilità e tecniche di disaster recovery geograficamente distribuito
- Setup di una prima infrastruttura per offrire servizi agli utenti locali del CNAF (sviluppo, testing, deployment di nuovi servizi) e successivamente da aprire all'utenza scientifica per necessità non Grid

Riferimenti e Ringraziamenti

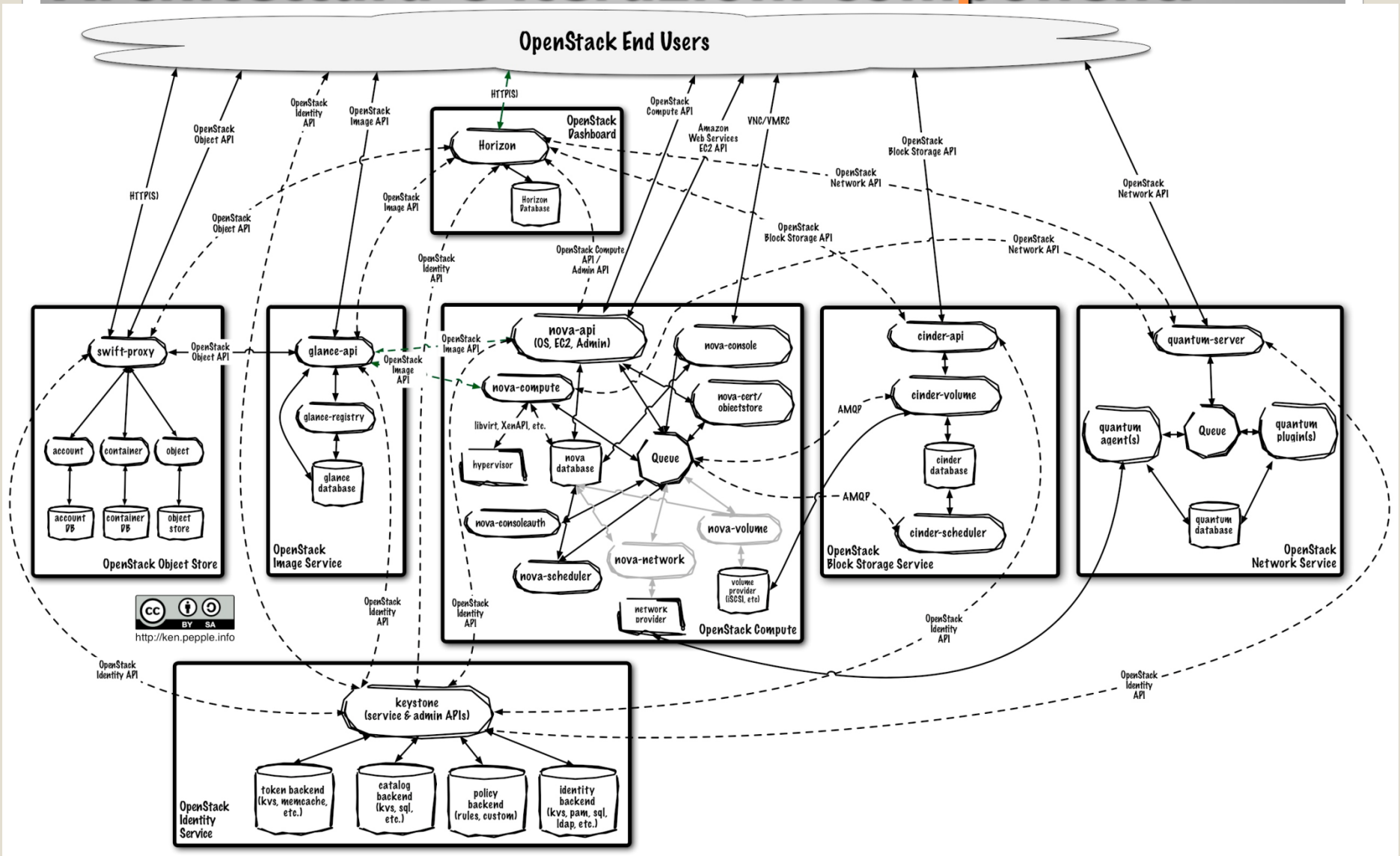
- OpenStack
  - Wiki
    - http://wiki.openstack.org/
  - Documentation
    - http://docs.openstack.org/

Hanno collaborato e contribuito al progetto MarcheCloud e alla preparazione di queste slide:

- Livio Fano' Illic (INFN-PERUGIA)
- Enrico Fattibene (INFN-CNAF/IGI)
- Matteo Manzali (INFN-CNAF)
- Hassen Riahi (INFN-PERUGIA)
- Davide Salomoni (INFN-CNAF)
- Andrea Valentini (INFN-PERUGIA)
- Valerio Venturi (INFN-CNAF/IGI)
- Paolo Veronesi (INFN-CNAF/IGI)

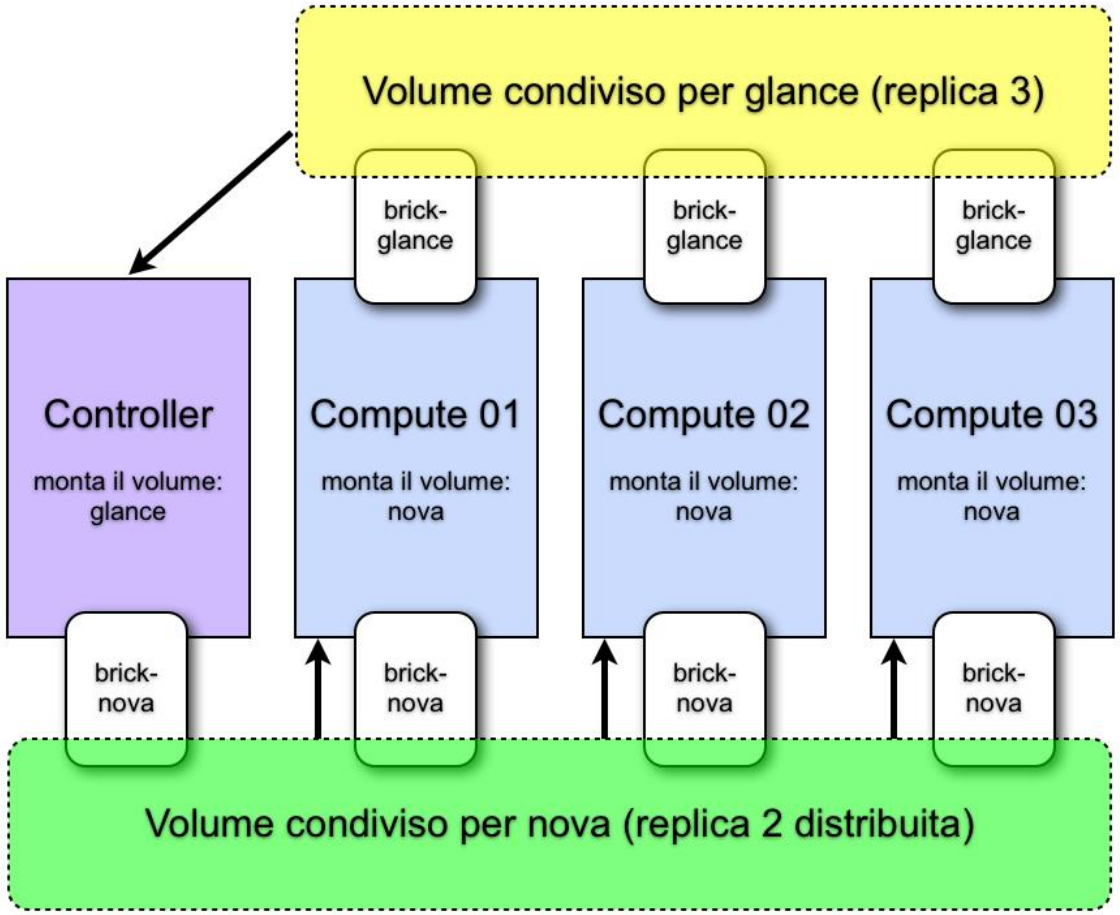
BACKUP

OpenStack
Architettura e iterazioni componenti



Source: http://ken.pepple.info/openstack/2012/09/25/openstack-folsom-architecture/

L’infrastruttura Pilota
Alta affidabilità a livello storage nel pilota



GlusterFS utilizzato per simulare una SAN che offra spazio per l’Image Repository (Glance) e per gli hypervisor (Nova-Compute)

oz

- E’ un tool command line per creare immagini di comuni distribuzioni Linux e Windows
- Usa KVM e libvirt per generare le immagini virtuali partendo dai sorgenti del SO o da una ISO
- Usa un file XML (il Template Description Language) per definire i principali parametri dell’immagine virtuale (sistema operativo, repository e pacchetti da installare) e le customizzazioni da apportare (files da creare e/o operazioni da compiere) post installazione
- Fa parte del progetto AEOLUS http://www.aeolusproject.org

Analisi Alta Affidabilita’ (1/2)

- DB (MySQL, PostgreSQL, …)
  - cluster mysql;
  - cluster con pacemaker and corosync che si occupano dell'HA di MySQL
- APMQ (Qpid, RabbitMQ, …)
  - configurazione di broker di messaggistica oppure tramite pacemaker e corosync + modifiche a codice openstack (in attesa di integrazione in una release di openstack ufficiale).
- Keystone
  - backend: usa come backend un db;
  - frontend: configurazione basata su pacemaker and corosync che si occupano dell'HA del servizio

Analisi Alta Affidabilita' (2/2)

- Glance
  - backend: usa come backend un db
  - storage: HA demandata a SAN (glusterFS, SWIFT, etc)
  - frontend: configurazione basata su pacemaker and corosync che si occupano dell'HA del servizio
- Nova
  - backend: usa come backend un db
  - nova-compute (scalabilità orrizontale)
  - nova-network/Quantum (dipende molto dalla tipologia di networking usata). Per FlatDHCP è stato replicato il servizio su tutti i nova-compute
- Horizon (Dashboard)
  - backend: usa come backend un db
  - frontend: configurazione basata su pacemaker and corosync che si occupano dell'HA del servizio

WP2 – Analisi soluzioni

	Performance monitoring	User-friendly Web App	Notifications	Log monitoring	Libvirt Plugin	Support of Windows	Plugin for OpenStack	Plugin based metrics	Richness in existing metrics	Popularity
Collectd	X		X	X	X	X		X		
Ganglia	X	X				X	X	X	X	X
Nagios		X	X	X	X	X	X	X		X
Zenoss	X	X	X	X	X	X	X	X		
Own libvirt-based script					X			X		