Cloud@CNAF Next Generation

Diego Michelotto (diego.michelotto@cnaf.infn.it)

Cristina Duma (cristina.aiftimiei@cnaf.infn.it)

3/12/2019

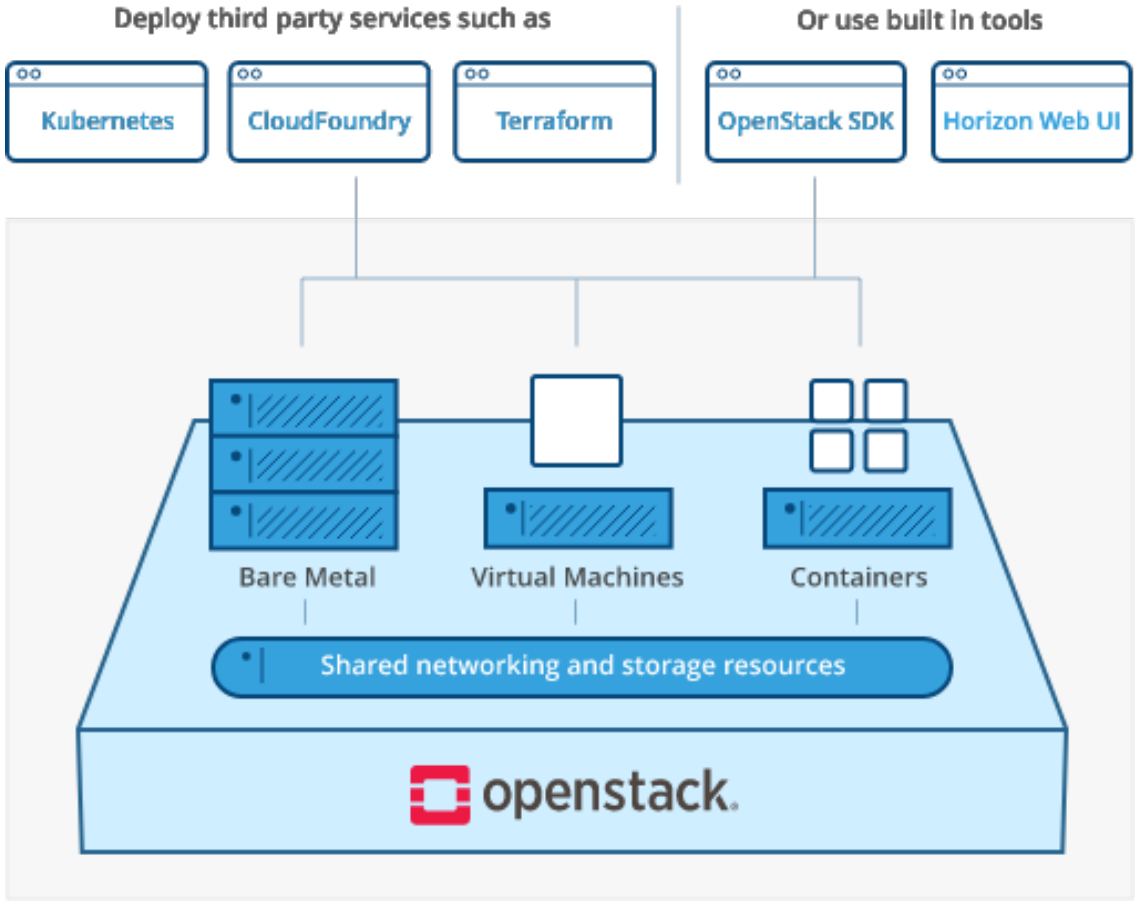
Indice

- OpenStack
- Infrastruttura
  - Network
  - Servizi ancillari
  - Storage
  - Servizi cloud comuni
  - Servizi cloud per regione
- Operations
  - Monitoraggio
  - Accounting
  - Installazione/Configurazione
  - Backup
- Integrazione Storage Esterni
- Migrazione
- Utilizzo Cloud
  - Autenticazione
  - Dashboard
  - API
- Nuovi utenti
- Use case
  - EEE
  - MW-DEVEL
  - DODAS
  - USER-SUPPORT
- Next steps
  - Aggiunta servizi cloud
  - Raccolta log
  - Dyn Part / Spot Instances
  - Integrazione Kubernetes
  - Security
  - Misc.
  - Cloud INFN
- Conclusioni

OpenStack (1/2)

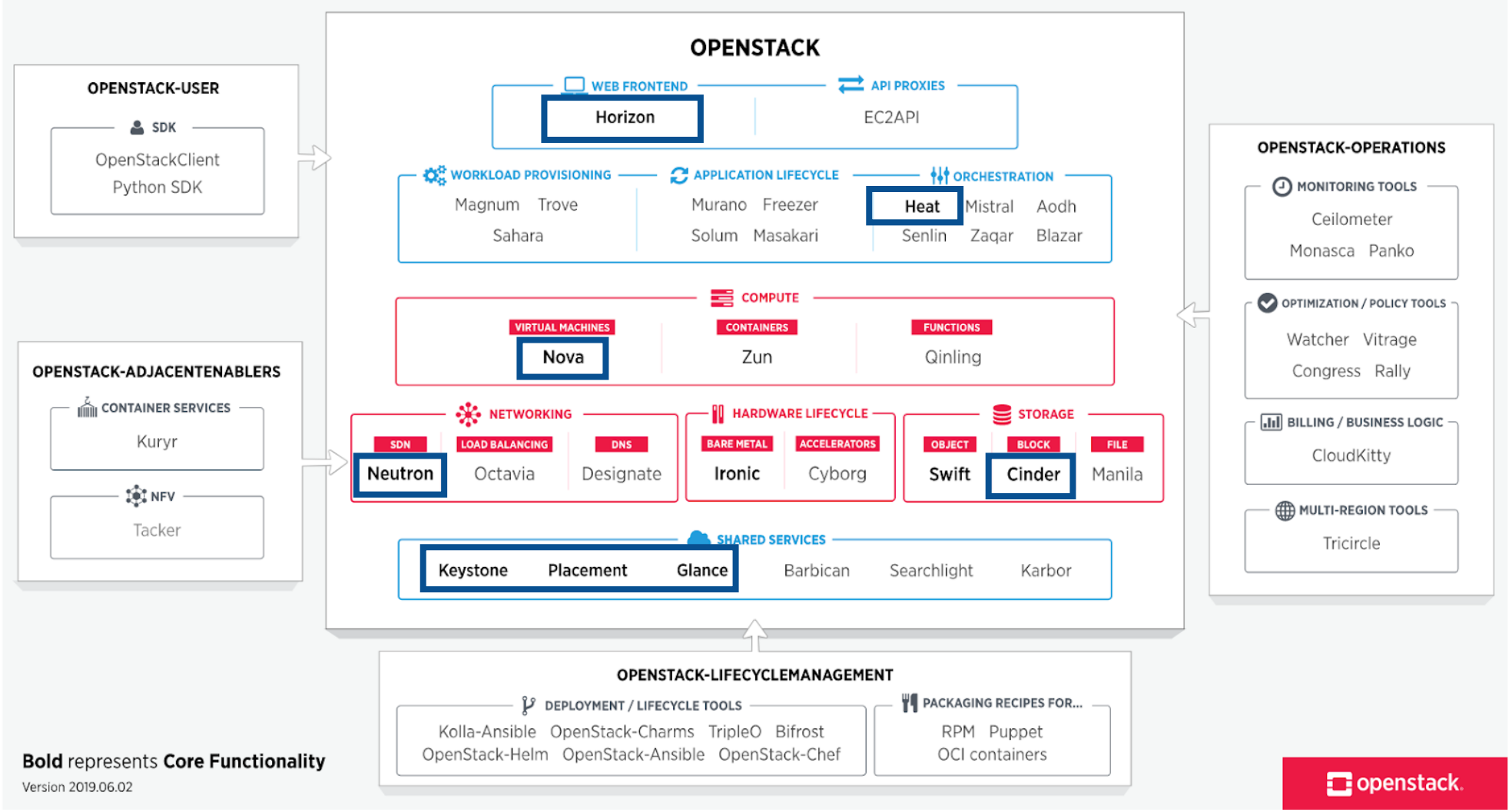
OpenStack is a **cloud operating system** that controls large pools of **compute**, **storage**, and **networking** resources throughout a datacenter, all managed and provisioned through APIs with common authentication mechanisms.

Beyond standard **infrastructure-as-a-service** functionality, additional components provide **orchestration**, **fault management** and **service management** amongst other services to ensure **high availability** of user applications.



ref: https://www.openstack.org/software/

OpenStack (2/2)



Infrastruttura (1/3)

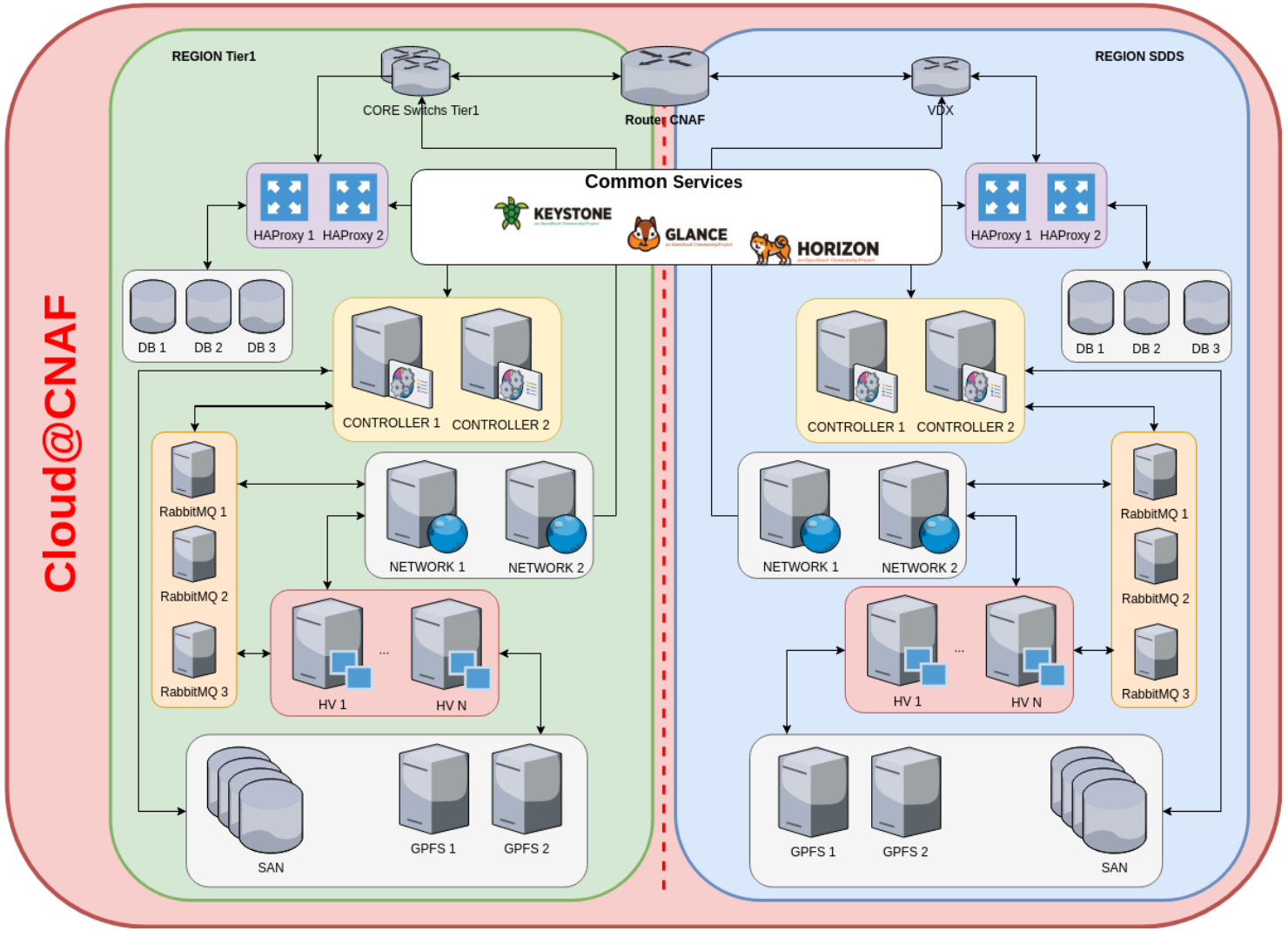
Obbiettivi:

- Unica infrastruttura condivisa tra Tier1, SDDS e Tier3
- Riutilizzare non replicare il lavoro già fatto.
- Migliorare l'alta affidabilità
- Dare risorse cloud a utenti/comunità del Tier1
  - Risorse vicino allo storage del Tier1

Perché una nuova infrastruttura?

- Infrastruttura vecchia di 4 versioni
  - complicato fare 4 upgrade
  - live upgrade introdotti solo da versioni più nuove

Infrastruttura (2/3)

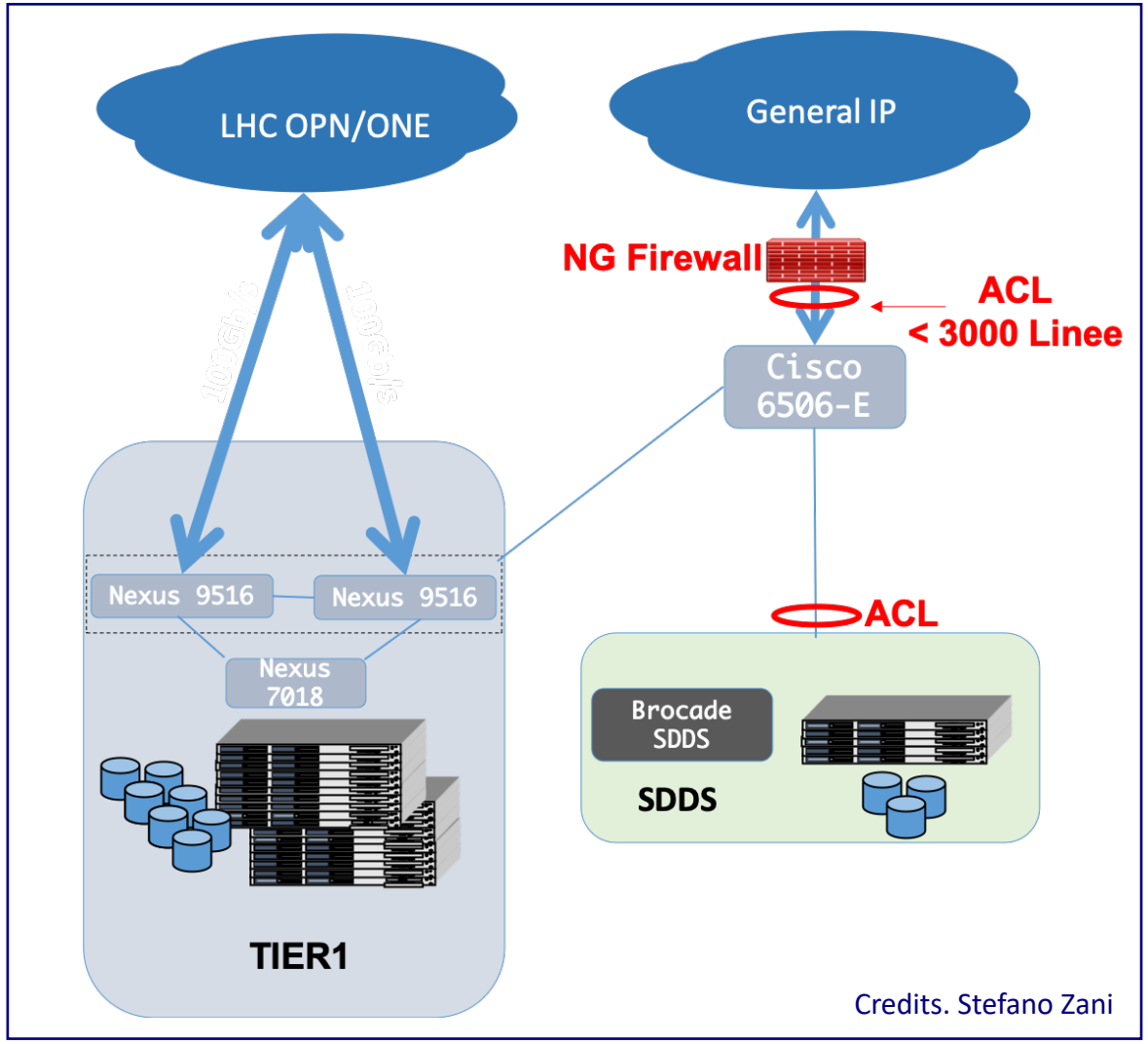


Infrastruttura (3/3)

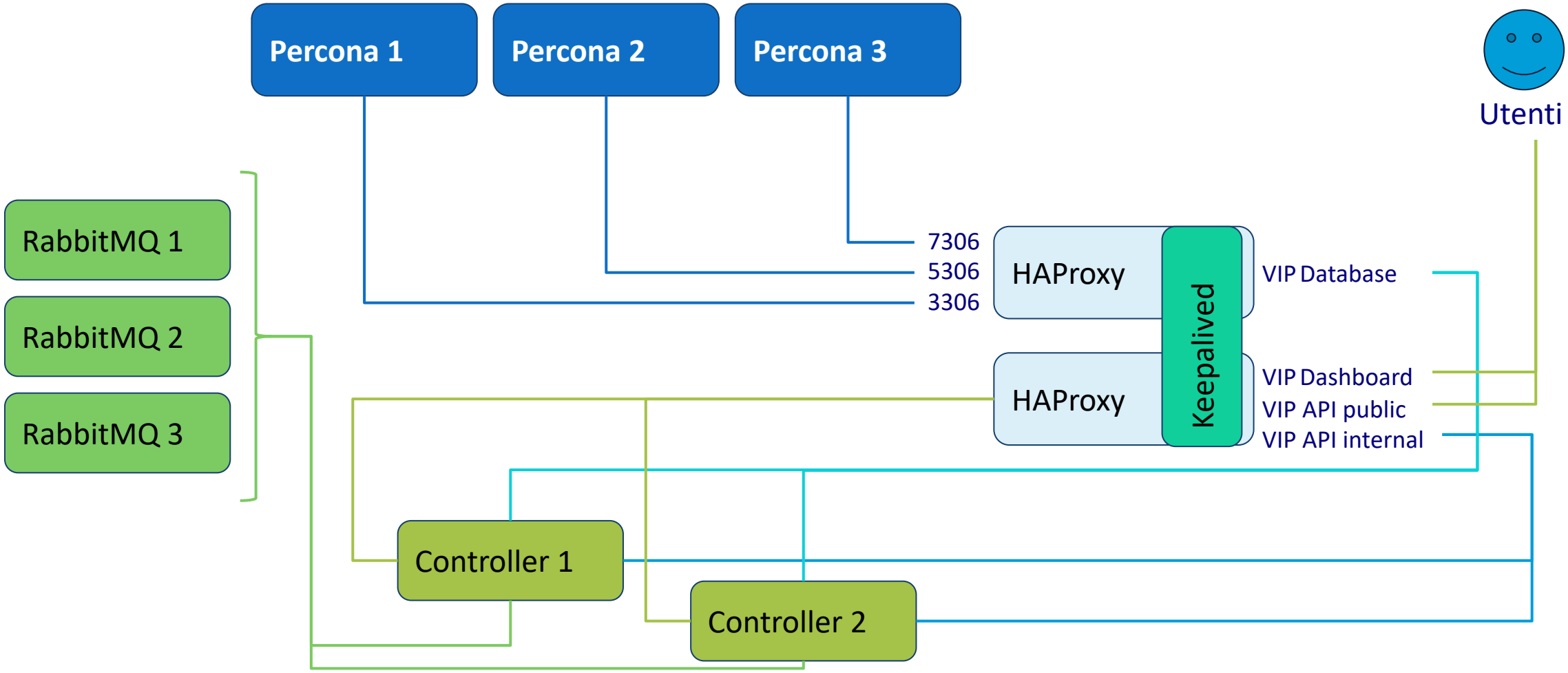
- 2 infrastrutture
  - **Testbed**
    - Uguale a Production ma con meno risorse
      - Servizi in alta disponibilità
    - Usata per test nuove funzionalità, test classi puppet, test upgrade
  - **Production**
    - Infrastruttura di produzione
      - SDDS: 368 Core, 1.4 TB RAM, 16 TB storage distribuito (Nova, Cinder), 510 FIP
      - Tier1: 496 Core, 1.8TB RAM, 500 TB storage distribuito (Nova, Cinder, Glance), 1022 FIP
    - Attenzione **all'alta affidabilità e disponibilità**
      - Hardware e switch ridondati
      - Servizi replicati distribuiti all'interno del datacenter, rack differenti per servizi replicati, sistemi di virtualizzazione differenti
      - Servizi ancillari ospitati su infrastrutture di virtualizzazione differenti (oVirt, VMWare)

Infrastruttura - Network

- Uso delle self service network
- Per ogni regione
  - 1 rete privata per management
  - 200 VLAN
  - 1 rete pubblica per external network
    - SDDS: /23
    - Tier1: /22
- Per il testbed vengono prese una piccola porzione delle reti e VLAN di sopra
- Infrastruttura di produzione
  - Utilizzo di LACP o bond AB su switch differenti
  - O servizi replicati attestati su switch differenti



Infrastruttura - Servizi ancillari



Infrastruttura - Storage

- Filesystem condiviso basato su **GPFS**
- Usato come backend per Immagini, Volumi e VM
- Uso **POSIX** non block storage
- Immagini, Volumi e VM sono file in formato **qcow2**
- Tier1 513TB
  - **Gestito da gruppo Storage**
- SDDS 16TB (estendibili)
- Testbed 2TB

Infrastruttura - Servizi cloud comuni

- Ospitati su risorse del Tier1
- **Keystone** gestione **autenticazione**, **autorizzazione**, **utenti**, **progetti**, **domini**
  - Ogni progetto è visibile su entrambe le regioni, mentre per entrambe le regioni gli utenti e i ruoli sono gli stessi, le **quote** possono essere **differenti per regione**
    - Alla creazione di un progetto **le quote di default sono 0 su Tier1** e **risorse limitate su SDDS**, così da incentivare i test su SDDS e produzione su richiesta (pledge pagati) su Tier1
- **Glance** gestione **immagini**
  - **Stesse immagini per entrambe le regioni**
  - Lo snapshot delle VM (diverso da snapshot di volumi) viene caricato in glance ed è possibile usarlo come sorgente per una nuova VM
  - **Immagini pubbliche personalizzate** dagli amministratori per motivi di **sicurezza**
    - iniettata chiave cloud-security per utente root
    - configurato rsyslog per invio log su logserver
- **Horizon**
  - **Dashboard unica** per l'accesso alla risorse cloud di entrambe le regioni

Infrastruttura - Servizi cloud per regione (1/3)

- **Nova**
  - Gestisce il ciclo di vita delle VM
  - Si interfaccia con **libvirt** e **GPFS** (vedi patch sottomessa a libvirt https://bugzilla.redhat.com/show_bug.cgi?id=1679528)
  - Gestisce le **migrazioni** delle VM tra host. **Live migration** possibile perché c'è **filesystem condiviso**.
  - Usa come **storage backend GPFS**, le immagini della VM sono in **formato compresso** qcow2.
  - Le risorse sono **raggruppate** per cpu affinity usando **«host aggregate»**
    - Le VM possono essere **migrate** in modalità **live** solo all'interno dello **stesso host aggregate**
    - Per **gruppi di host piccoli** ma con CPU leggermente differenti viene usato il **CPU model inferiore per compatibilità**
    - **Unico host aggregate** per le risorse del **Tier3**, usando metadati e filtri per lo scheduling
    - Host aggregate **LSD** (Local Storage Disk) per VM che richiedono **accesso diretto al disco** dell'hypervisor

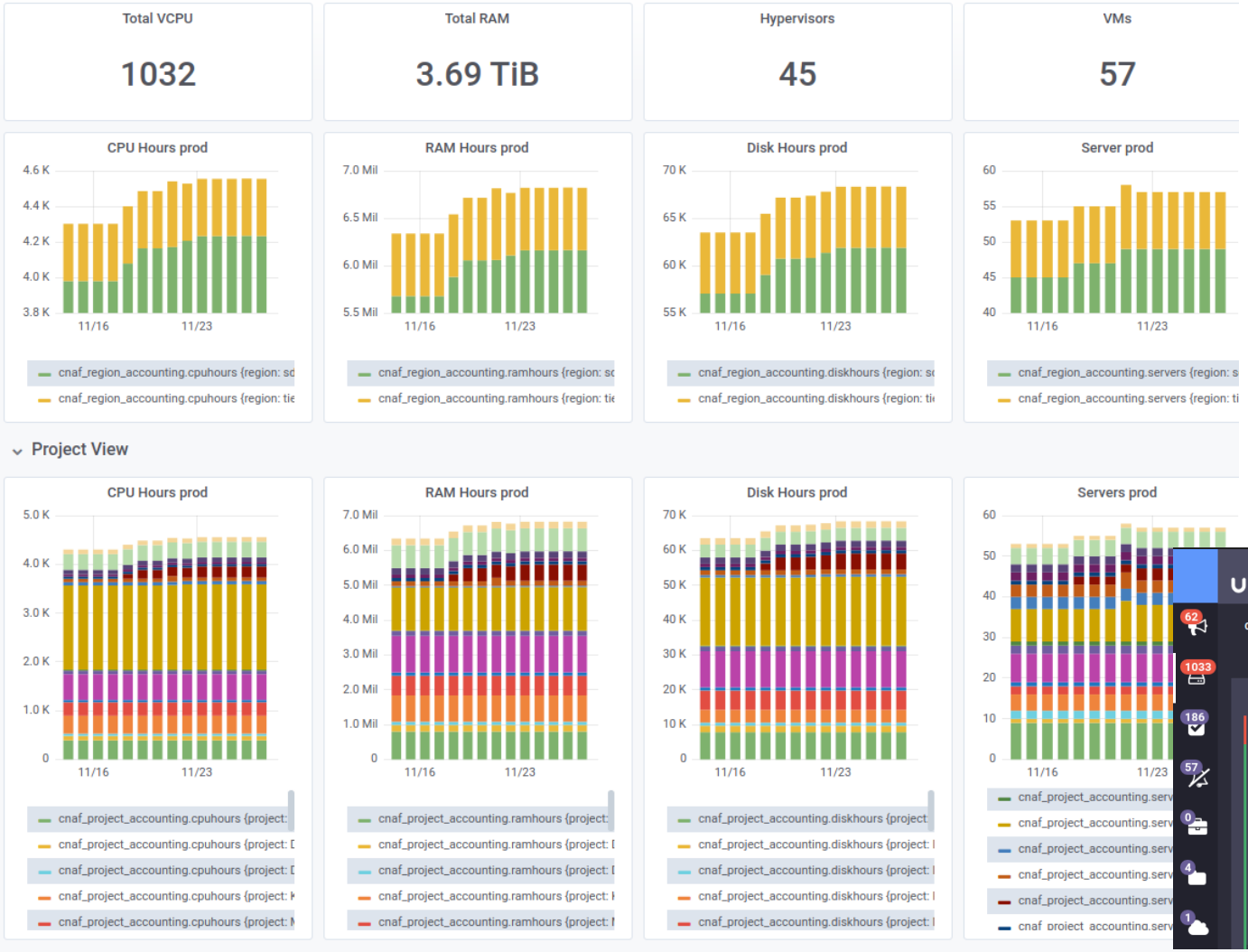
Infrastruttura - Servizi cloud per regione (2/3)

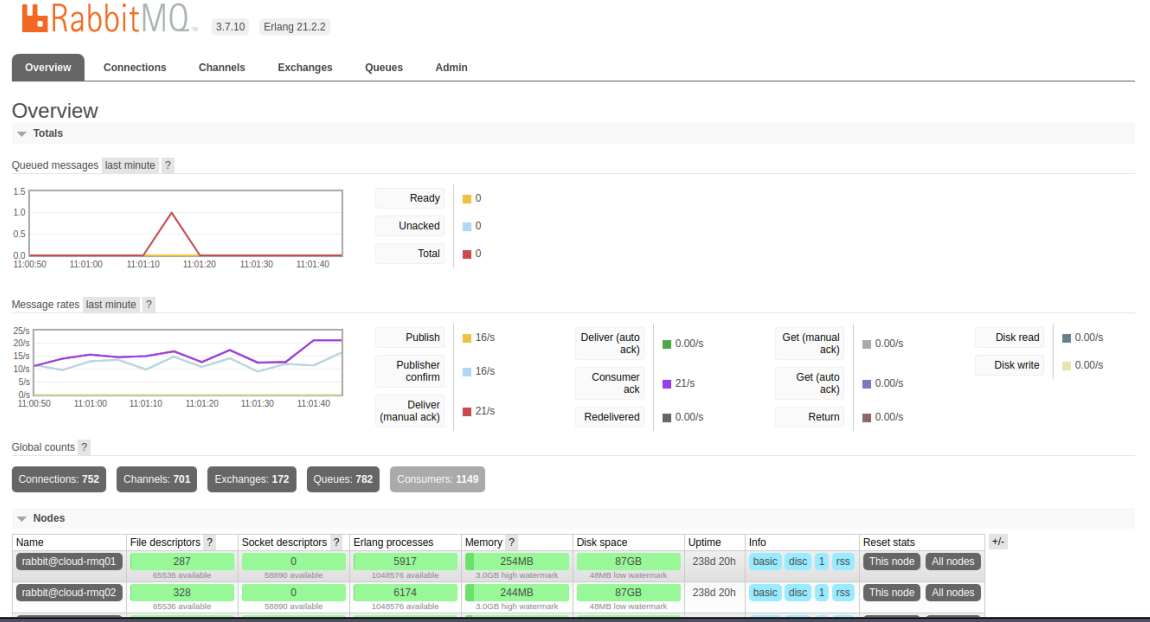
- **Neutron**
  - Gestisce il **networking** delle VM
    - Configurato per uso delle **self-service network** con uso di **linuxbridge**, **VLAN** e **Floating IP**
    - Ogni **rete privata** dell'utente è mappato su una **VLAN** fisica
    - Esiste una **rete external** che se collegata ad una rete privata tramite un **virtual router** da **outbound connectivity** alle reti private
    - Per avere **inbound connectivity** è necessario assegnare un **FIP** alle VM che la richiedono.
    - Gestisce le **regole di accesso** alle **VM** tramite **«security group»**
      - Di **default** le VM si parlano tra di loro se appartengono allo **stesso security group** e hanno **tutte** le **porte** e i **protocolli aperti in uscita**.
      - Per tutte le **connessioni in ingresso** vanno **espressamente specificate** come regole **in security group ad hoc**
        - ✓ Per l'accesso da **general internet** sono applicate le **policy del CNAF** per quanto riguarda le risorse di rete

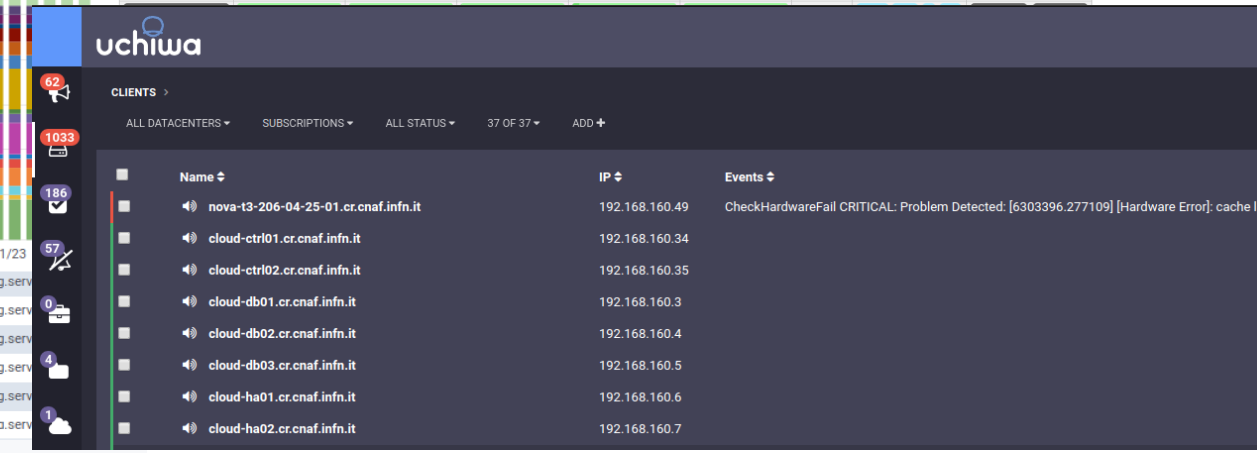
Infrastruttura - Servizi cloud per regione (3/3)

- **Cinder**
  - Gestisce **block storage** per le VM
    - Le **VM** possono essere **istanziate** da immagine direttamente su **Volume** (Default)
      - Il volume **persistere alla cancellazione** della VM se non diversamente specificato
      - Il volume così creato può non rispettare le dimensioni del disco specificate nel flavour
    - Alle VM può essere **collegato un ulteriore volume** che verrà visto come disco aggiuntivo
      - Usato per salvare **dati**
      - **Non viene eliminato se viene cancellata la VM**, può essere scollegato e riutilizzato
  - Gestisce lo **snapshot** dei volumi
  - Il Backup dei volumi è al momento disabilitato
- **Heat**
  - Orchestra risorse cloud tramite template

Operations – Monitoraggio (1/2)







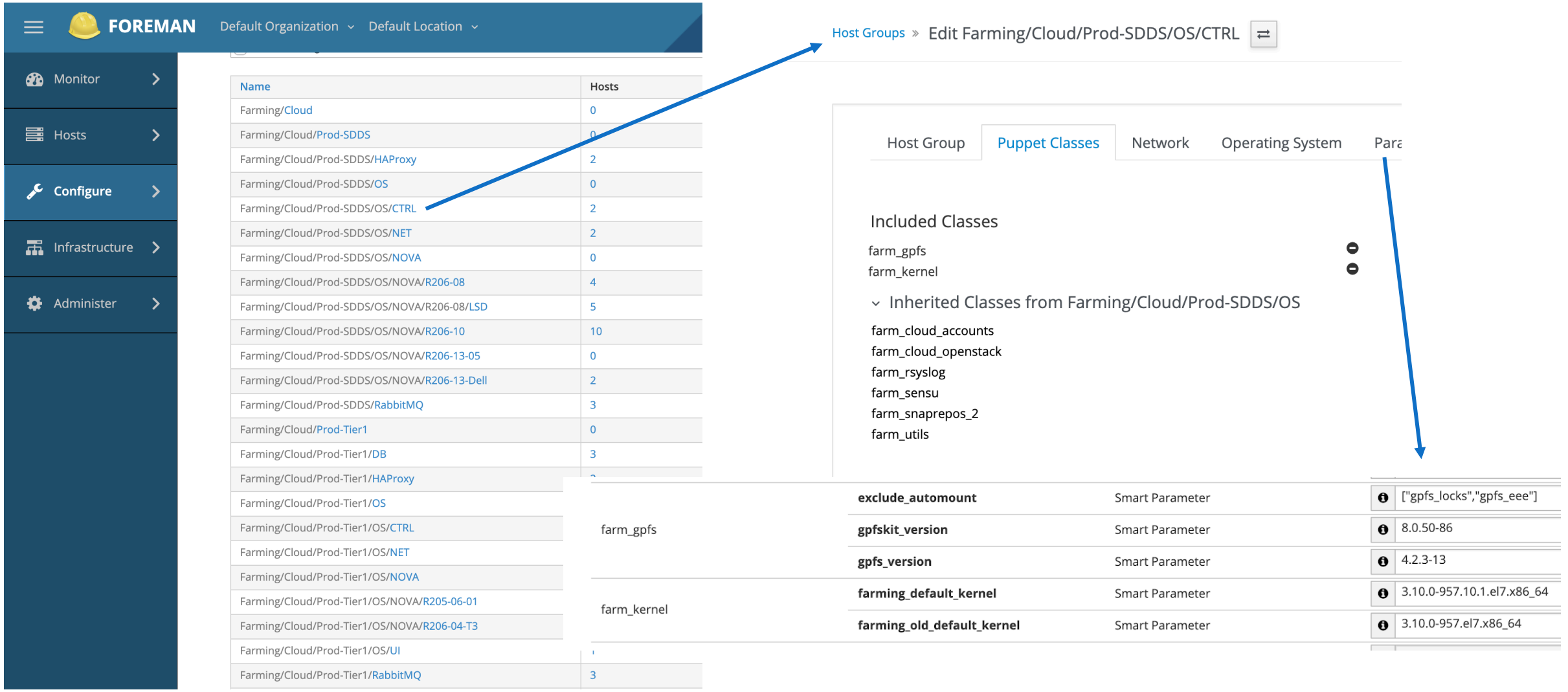
Operations – Monitoraggio (2/2)

- Metriche
  - Sensu + InfluxDB + Grafana
  - Cloud Overview: https://t1metria.cr.cnaf.infn.it/d/rrKsOx5Zz/cloudatcnaf-overview?orgId=2
  - HAProxy:
    - Metriche: https://t1metria.cr.cnaf.infn.it/d/000000035/haproxy?orgId=2
  - RabbitMQ:
    - Metriche: https://t1metria.cr.cnaf.infn.it/d/000000036/rabbitmq?orgId=2
  - Percona:
    - Metriche percone: https://t1metria.cr.cnaf.infn.it/d/000000033/percona_cluster?orgId=2
    - Metriche mysql: https://t1metria.cr.cnaf.infn.it/d/000000034/mysql_host?orgId=2
- Allarmistica
  - Sensu + Uchiwa
  - Notifiche su Slack/Mail

Operations - Accounting

- Accounting giornaliero: https://t1metria.cr.cnaf.infn.it/d/k0dz1eZZz/cloudatcnaf-accounting?orgId=2
- Raccolto con **openstack usage list -f csv --start 2019-12-02 --end 2019-12-04** a mezzanotte di tutti i giorni
- I dati vengono aggregati, manipolati e inviati a influxdb
- Cosa manca:
  - Accounting fine (**cpu hours vm * potenza dell'hypervisor**)
    - **cpu hours == wall clock time**
  - Integrazione con i pledge del T1

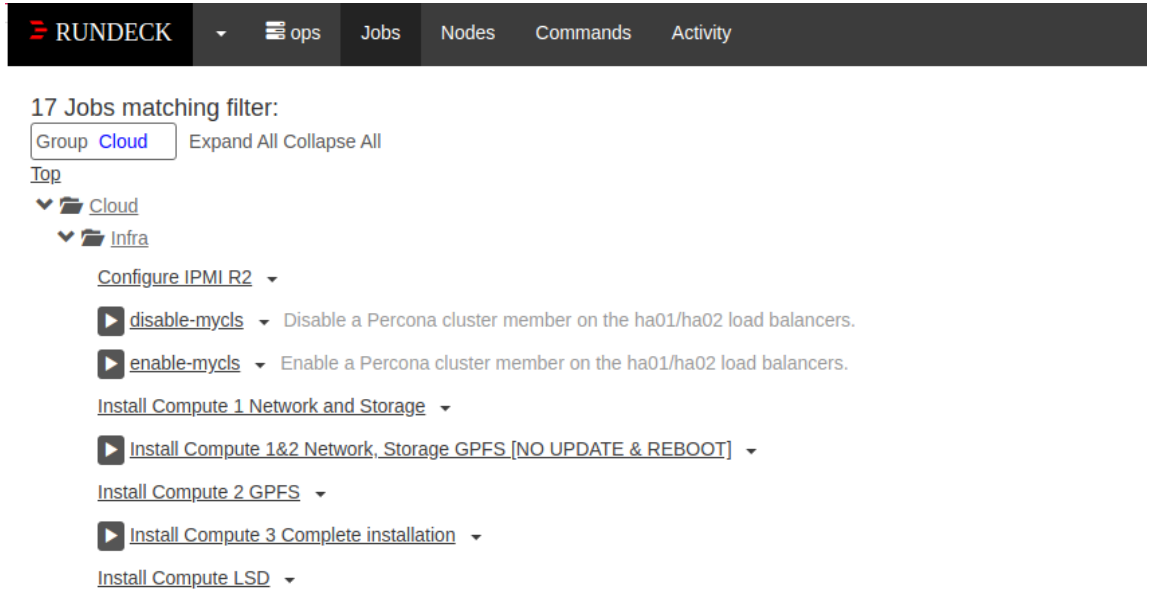
Operations - Installazione/Configurazione (1/2)



Operations - Installazione/Configurazione (2/2)

L'installazione del sistema operativo e del software avviene tramite l'utilizzo di **snapshot**, da un **repository locale**

- riproducibilità delle stesse installazioni
- versioni software sotto controllo



La parte di installazione, gestione e configurazione è gestita tramite Foreman + Rundeck

- **Foreman** definisce le **configurazioni software /profili** da installare sulle macchine tramite gli hostgroup che vengono applicati da puppet
- **Rundeck** automatizza **task ripetitivi**, si integra direttamente con le API di foreman per recuperare le informazioni degli host, esegue in batch script, workflow e fa reporting dei task.

Operations - Backup

Assunto che tutta la configurazione viene fatta da puppet e che le classi puppet sono su repository git, l'unica cosa di cui fare i backup sono i DB

Backup:

- **backup full 1° giorno del mese** (cancellazione del backup precedente)
- backup **incrementale** il resto dei giorni (max 30 incrementali)
- backup effettuato su **mount nfs** esportati da sistemi di **storage/backup dei reparti**

Restore:

- generazione del backup completo **applicando in ordine cronologico gli incrementali al full** prima di effettuare il restore

NB: no backup VM

Integrazione Storage Esterni

Dare accesso allo storage del Tier1 dalle risorse cloud

- **Accesso veloce** ai dati dei rispettivi esperimenti
- Per alcuni **casi «stabili»** dare **accesso** diretto al **cluster GPFS**
- Per altre risorse **accesso al cluster GPFS via CNFS** (caso d'uso «**UI** esperimenti»)
- Studiare e testare **nuove modalità** di accesso allo storage del Tier1
  - **webdav**
  - **Onedata**

Permettere accesso ad altre risorse **storage esterne** all'infrastruttura cloud

- Risorse storage Tier3 (INFN-BO)
- Risorse storage EEE (Vedi prossime slide)

Migrazione

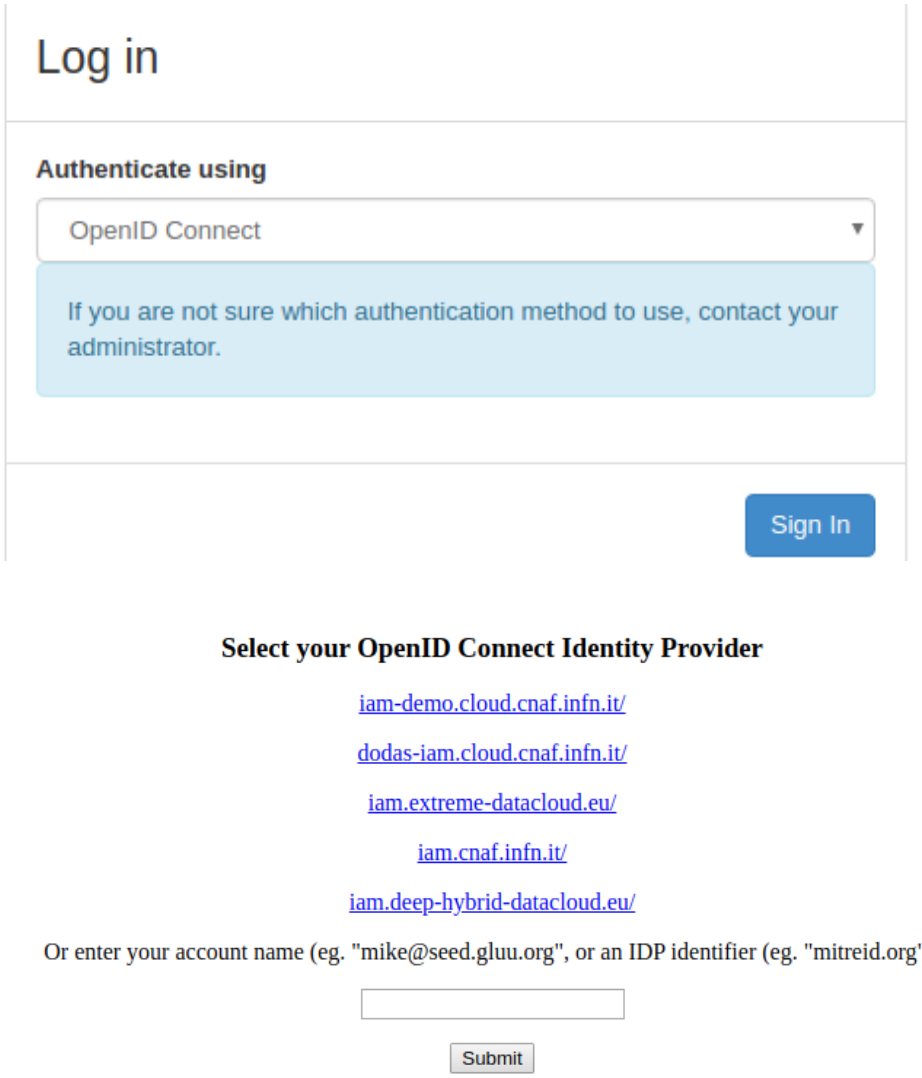
Stato: ~40% DONE, ~60 VM su 200 VM

- **Opzione 1**:
  - Ricreare le risorse sulla nuova infrastruttura
- **Opzione 2**:
  - Creare uno snapshot della VM sulla vecchia infrastruttura, scaricarla e caricarla sulla nuova infrastruttura e ricreare la VM
  - Per i volumi, creare un'immagine dal volume, scaricarla e caricarla sulla nuova infrastruttura e creare un volume partendo dall'immagine.

- Esperimenti migrati:
  - **AMS**: 1 VM, 4 VCPU, 8GB RAM, 250GB Disco, Alias ams-mongodb.cloud.cnaf.infn.it
  - **ICARUS**: 1 VM, 4 VCPU, 8GB RAM, 80GB Disco, Alias ui-icarus.cloud.cnaf.infn.it
  - **NTOF**: 4 VM, 64 VCPU, 32GB RAM, 1TB Disco, Alias ui-ntof.cloud.cnaf.infn.it
  - **FAZIA**: 1 snapshot bastion, 1 snaptshot wn, 2 TB Volume Dati

Uso Cloud - Autenticazione

- Metodo di autenticazione **INDIGO-IAM**
- iam.cnaf.infn.it con **utenti mappati in gruppi diversi a seconda delle richieste** di utilizzo:
  - utente **«non classificato»** → progetto **«CNAF»** condiviso con altri utenti simili, solo poche risorse nella regione SDDS
  - utente **«classificato»** → **progetto personale** dedicato con risorse minime su regione SDDS
  - utente appartenente ad un **gruppo specifico** → **progetto condiviso con il proprio gruppo**, quote e regioni stabilite con il gruppo
  - utente **admin** (amministratori dell'infrastruttura cloud) → mappato direttamente su un **utente keystone**
  - **un utente può appartenere a più progetti → mapping in keystone**
- iam-demo.cloud.cnaf.infn.it usato per **corsi formazione**
- gli altri IAM, tutti gli **utenti vengono mappati su un unico progetto** con quote e regioni stabiliti con il **progetto** proprietario dell'IAM
- **NB**: stesso utente su iam differenti equivale a utenti differenti per OpenStask



Uso Cloud - Dashboard

Dashboard URL https://cloud-dashboard.cnaf.infn.it

(Demo)

Uso Cloud - API

Possibile utilizzare la cloud tramite le sue API:

- **OpenStack**:
  - OpenStack CLI / OpenStack python SDK
    - Doc: https://docs.openstack.org/rocky/user/
  - API
    - Doc: https://docs.openstack.org/api-quick-start
- **PaaS services**:
  - INDIGO Orchestrator
  - INDIGO IM
  - DODAS
  - Rancher

Nuovi utenti

- In fase di definizione la procedura per l'aggiunta di nuovi utenti
  - Firma **AUP del CNAF**
  - Account su **bastion.cnaf.infn.it**
  - **Registrazione** su https://iam.cnaf.infn.it
    - Approvazione dagli amministratori, verifica dei punti precedenti e inserimento nel gruppo più adatto
  - A secondo dei casi potrebbe essere necessaria la **modifica delle quote** del progetto
- Supportare nuovi use-case, nuovi utenti
  - **Stretta interazione con User Support**
  - **Identificazioni nuovi servizi da rendere disponibili**
- Training e Consulenza
  - User hands on
  - Debugging sections
  - Helpdesk

Use case - EEE

Aggregazione e analisi dei dati raccolti da diversi telescopi sparsi tra le scuole italiane e il CERN https://eee.centrofermi.it/

- **Area storage** di progetto **ospitata** da SDDS su **GPFS** esportata via **CNFS (60TB)**
- 1 VM che fa da **frontend**
  - monta l'area storage e fornisce un server **syncthing** che sincronizza i dati dai vari telescopi
- 1 VM che fa **server web**
  - monta l'area storage e fornisce strumenti per la pubblicazione dei dati
- Diverse VM che montano l'area storage e mettono a disposizione strumenti di **analisi per gli utenti**.

Punti deboli:

- CNFS, non supporta grosso I/O nelle fasi di analisi o ricostruzione db di syncthing

Use case - MW-DEVEL

Risorse del gruppo Software Development di SDDS per supporto allo sviluppo software

- **Cluster Kubernets** installato su risorse cloud (17VM, 1 master , 1 ingress, 1 nfs server, 1 logger, 13 executor) ~80 VCPU e ~ 200GB RAM
- deployment di:
  - **Strumenti di sviluppo**: Jenkins, sonarqube, repository software, …
    - jenkins tramite client openstack crea vm on demand per l'esecuzione di alcuni test su progetto dedicato.
    - slave jenkins
  - **Servizi per vari progetti**:
    - IAM, TTS
    - Namespaces separati

Use case - DODAS

Deployment di infrastrutture su cloud

- **batch system as a service**
  - deploy di cluster mesos + marathon sul quale gira un batch system htcondor autoconsistente o che si collega a pool globali come quello per CMS
- **big data analysis as a service**
  - deploy di cluster spark e/o hadoop per analisi di big data

Utilizza **INDIGO PaaS** per la gestione di queste infrastrutture su **cloud pubbliche e/o private**

Use case - USER- SUPPORT

Progetto dedicato al gruppo USER SUPPORT che **istanzia vm/servizi su risorse cloud per conto degli esperimenti**

- **FAZIA**: 1 bastion + 16 worker per analisi dati, VM con **accesso al disco locale** dell'hypervisor per accesso I/O pesante
- **AMS**: 1 VM con **mongoDB** utilizzato da software di sottomissione job al CNAF e al CERN
- **ICARUS**: 1 user interface per gli utenti icarus
- **NTOF**: 4 VM con home condivisa per dati, **utilizzate per analisi**

Esperimenti isolati tra di loro tramite utilizzo di **reti private separate** e **security group dedicati**

Next steps - Aggiunta servizi cloud

- **LBaaS**: Load Balancer as a Service
- **Manila** : NFS as a Service
- **Trove**: DB as a Service
- **Magnum** : Cluster as a Service (Kubernets, Mesos+Marathon,)
- **KataContainer** : Container as a Service
- **Mistral** : Workflow as a Service
  - Implementazione di workflow sul **lifecycle delle VM**
- **Ironic** (con eventuale **GPU**) : Bare Metal as a Service

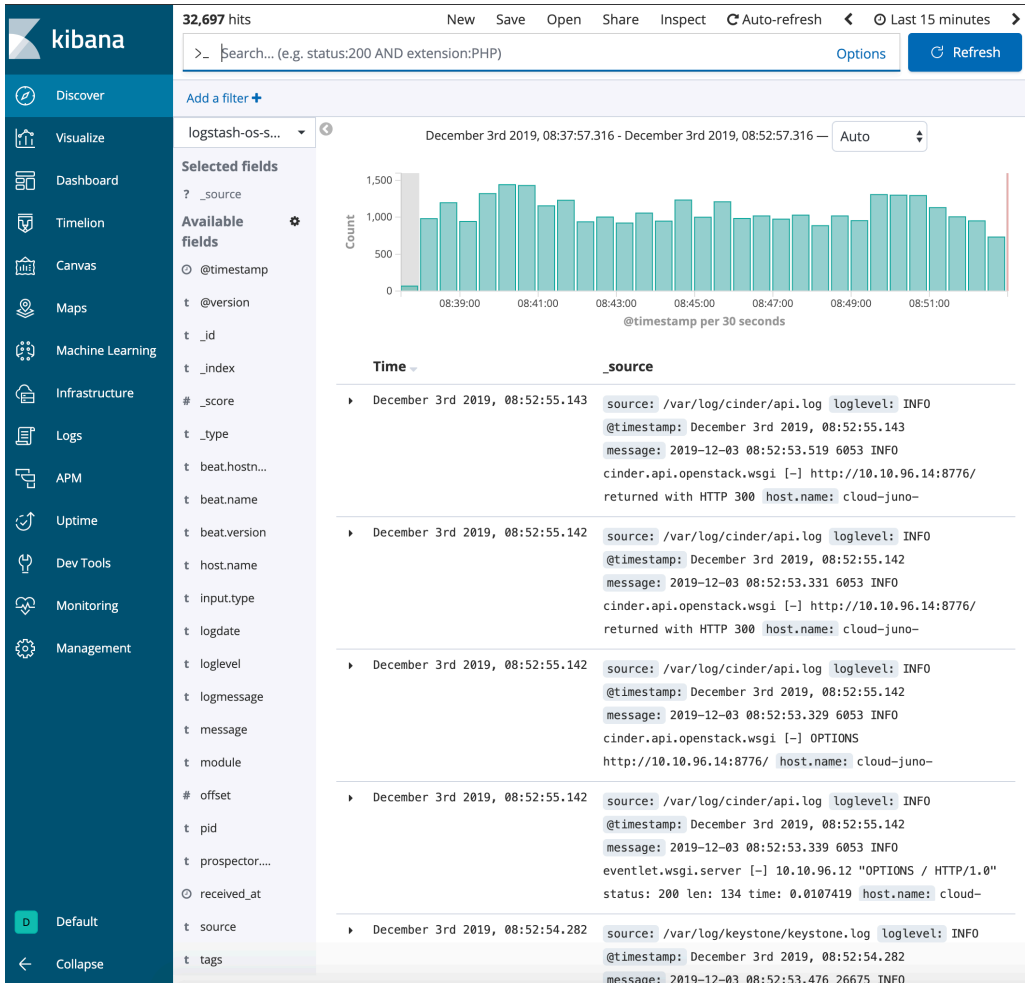
Next steps - Raccolta Log

Situazione attuale:

- **Immagini** “standard” modificate per inviare i log tramite **syslog** a **logserver**
- Testbed:
  - **Raccolta log** dei servizi openstack tramite **Filebeat**
  - Invio log a **logstash** per **filtering**
  - Invio log a cluster **ELK** di test

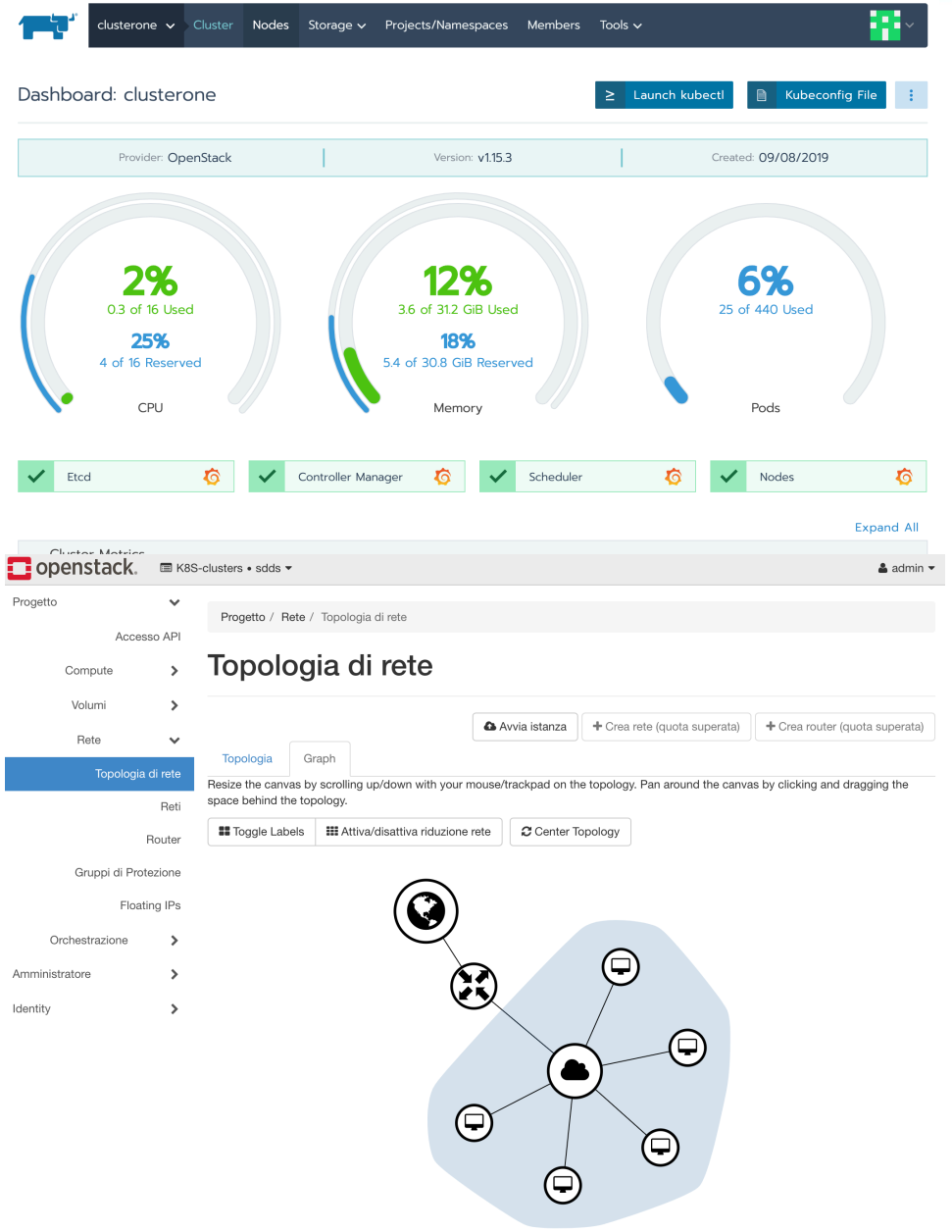
Manca:

- Terminare **definizione filtri Logstash** per ogni log
- Inserire in **Puppet e Foreman** tutte queste configurazioni



Next steps - Kubernetes

- **Rancher** https://rancher.com/
  - KaaS (**Kubernets as a Service**)
  - Gestore di cluster Kubernetes - **locali o remoti**
  - https://rancher.cloud.cnaf.infn.it
  - **Servizio online, in fase di test**
    - Potrà interagire con LBaaS quando sarà disponibile
- **Magnum**
  - Gestore di **cluster as a Service** (Kubernetes, Mesos, Swarm)
  - Servizio di **OpenStack**, integrato anche nella dashboard
  - Utilizza **Heat**



Next steps - Dyn Part / Spot Instances

Obbiettivo:

- **Utilizzare al massimo le risorse della farm** integrando l'accounting cloud con il calcolo dei pledge del Tier1

2 soluzioni in fase di studio:

- **Dyn Part**:
  - sviluppato in **INDIGO-DataCloud**
  - permette di **sottrarre** al batch system in maniera **automatica** e **controllata worker node** per poterli utilizzare come **compute node** e vice versa
- **Spot instances**:
  - **creazione di vm** (worker node) sulle **risorse libere** della cloud
  - nel caso in cui vengano **richieste risorse cloud** dagli utenti vengono **eliminate** tante **spot instances** quante le risorse richieste.

Next steps - Security

Implementazione di **meccanismi** per l'esecuzione **automatica** di **test di sicurezza** periodici sui **FIP**

- Scansione di tutti i FIP
- **Report di vulnerabilità**
  - Report di vulnerabilità ai **proprietari** delle VM
- Blocco degli utenti e/o delle VM in caso di mancata risposta alle segnalazioni

Misure Minime e GDPR

- Procedure **creazione** e **cancellazione account** e **risorse** associate
- Gestione **scadenza e complessità password**
- **Tracciabilità attività VM e utenti**
- … e molto altro

Next steps - Misc.

- **Upgrade infrastruttura**
  - La versione attuale è Rocky, ultima release Train (R+2)
- **Replica dei servizi core** (Keystone, Glance, Horizon)
  - anche su SDDS + studio soluzione migliore
- Completare **documentazione** per utilizzo risorse e richiesta supporto https://wiki.infn.it
- Miglioramento **status page** - http://status.cloud.cnaf.infn.it:25554/
- **Monitoring e accounting fine delle VM**
  - Per monitorare l'effettivo utilizzo di risorse per VM/Progetto/Utente
  - Calcolo pledge
  - Gestire eventualmente più o meno overcommit di risorse
- **Virtualizzazione di GPU**
  - Sfruttare soluzioni sviluppate nei vari progetti (Deep-HybridDataCloud)
- Studio e integrazione di **Ceph** come **backend dei servizi**
- **Consolidamento infrastruttura**
  - rete su SDDS
  - rinnovo hardware

Next steps - Cloud INFN (1/2)

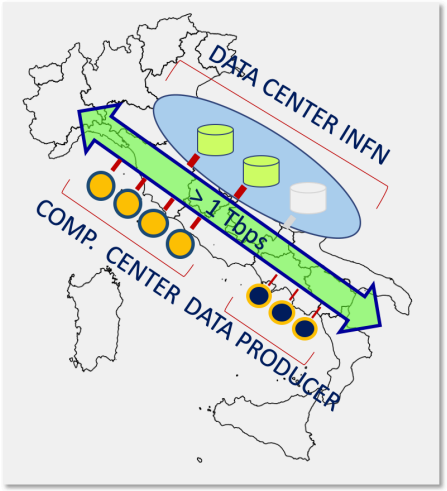
- Iniziativa che **integra ed espande** quanto fatto negli scorsi anni in diverse attività (INFN-CC, progetti nazionali ed europei, implementazioni ed expertise locali, etc.)
- È stato definito un **catalogo di servizi**, discusso – insieme ad una serie di priorità – in una riunione il 3/10/2019 a Bari.
- Il 18/11/2019 il Presidente ha dato mandato a Davide Salomoni di **«coordinare le attività per giungere nel più breve tempo possibile ad avere una Cloud INFN funzionante ed operativa da mettere a disposizione al nostro personale».**
  - Davide sta preparando un **documento sull'architettura** implementativa di questa INFN Cloud, che verrà discusso in primis con Gaetano e Claudio (come da richiesta esplicita del Presidente) e poi con altre realtà INFN.
  - In questa architettura di INFN Cloud, sarà prevista – in accordo con le proposte per un «INFN Datalake» presentate a luglio 2018 – la presenza di un **backbone per la gestione di casi d'uso specifici** e la possibilità di interconnettere («**federare**») data center INFN (come la cloud@CNAF) che espongano ad esempio **endpoint** noti per lo **storage** e per il **calcolo**, con policy di autenticazione e autorizzazione consistenti tra i vari siti. Maggiori dettagli si troveranno nel documento in preparazione menzionato sopra.

Next steps - Cloud INFN (2/2)

Una possibile infrastruttura: data lake Italia

1. Un grande data center (DC), con risorse sia HTC che HPC, da esporre nel "Data Lake" di WLCG composto da almeno 2 DC fisici (i.e. CNAF al tecnopolo ed un altro sito principale della infrastruttura. i.e. uno del PON-SUD)

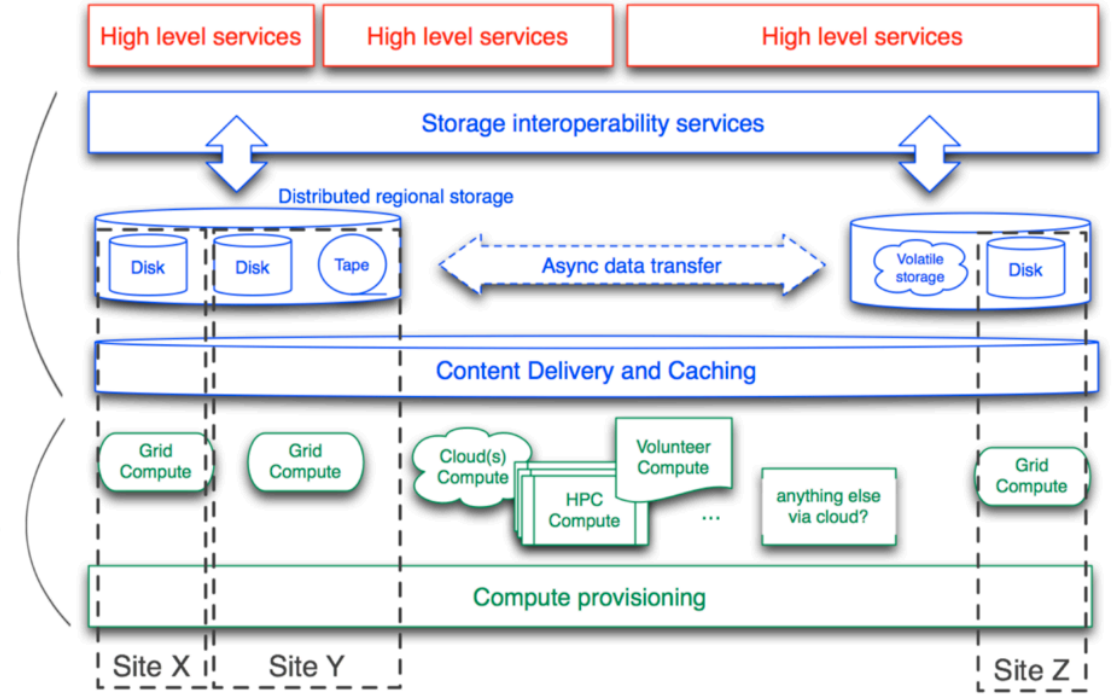
2. Un insieme di centri che abbiano la funzione di computing center (CC) con CPU e cache (i.e. Tier-2) ed eventualmente disco in funzione della capacità di supporto del sito



3. I DC e tutti i CC INFN dovranno essere connessi da una rete ad altissima velocità basata su link tipo DCI (Data Center Interconnect)

4. I laboratori/esperimenti I che producono dati (TierC dati raw dovranno essere conservati nel DC INFN (+ fornendo agli esperiment questione strumenti stan per la gestione dei dati.

5. CLOUD@INFN per accesso alla infrastruttura



Conclusioni

- Infrastruttura con servizi basilari consolidata
- Ancora molto lavoro da fare per integrare nuovi servizi
- Disponibilità per nuove interazioni

- Corso Cloud **«Amministratori di Infrastrutture di Cloud Computing, IaaS & PaaS»** coming soon.

- Contattateci
  - mailing list: **cloud-ops@lists.cnaf.infn.it**