THE ITALIAN RESEARCH & EDUCATION NETWORK | Consortium GARR

Implementazione multisite di un datacenter OpenStack altamente affidabile e scalabile basato su LXC

Lazio Pulse

INFN - LNF - 7 Giugno 2016

Alex Barchiesi, Alberto Colla, **Fulvio Galeazzi**, Mario Reale, Giancarlo Viola

Indice

Obiettivi

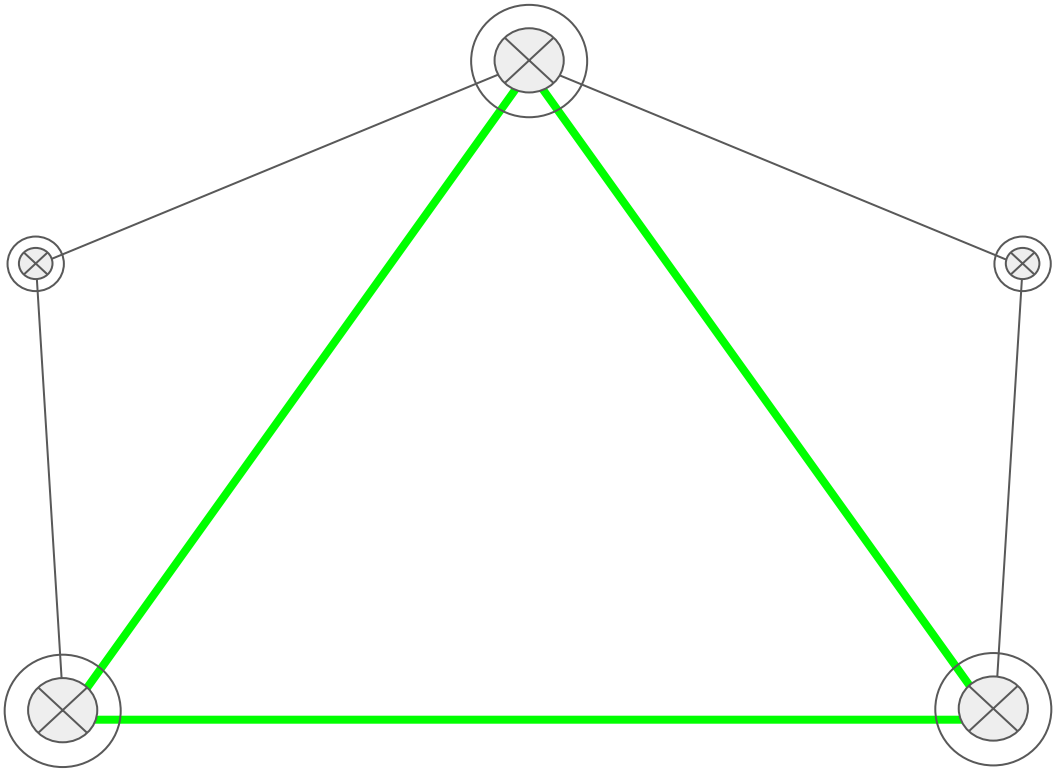
- Massimizzare SLA per servizi
  - Resilienza all'indisponibilità di un sito
  - Ridondanza:
    - Hardware
    - Rete
    - Software
- *Gestione* (utente) unificata datacenter distribuito geograficamente
- Gestione del carico locale e geografico
  - Modularità e scalabilità
- Sicurezza dei livelli dell'infrastruttura e dei servizi agli utenti
  - Separazione reti (VLAN, VXLAN, fisica)
  - ACL a diversi livelli (Router, OS, LXC, OpenStack Tenant)

8500 core
10 PB

...In 11 rack/moduli-CSD

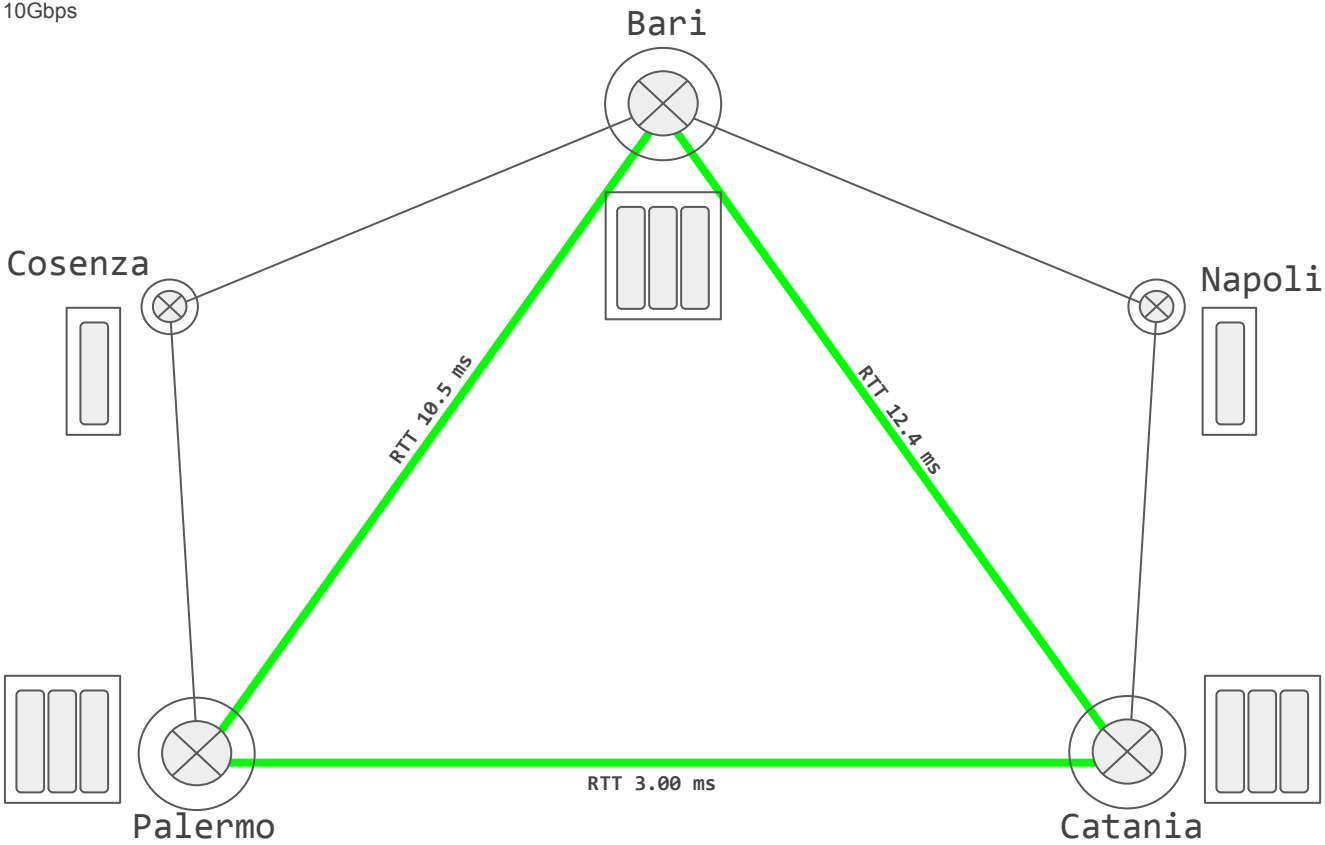
Network

40Gbps
10Gbps



Network

40Gbps
10Gbps



Chassis Blade **Dell M1000e:**

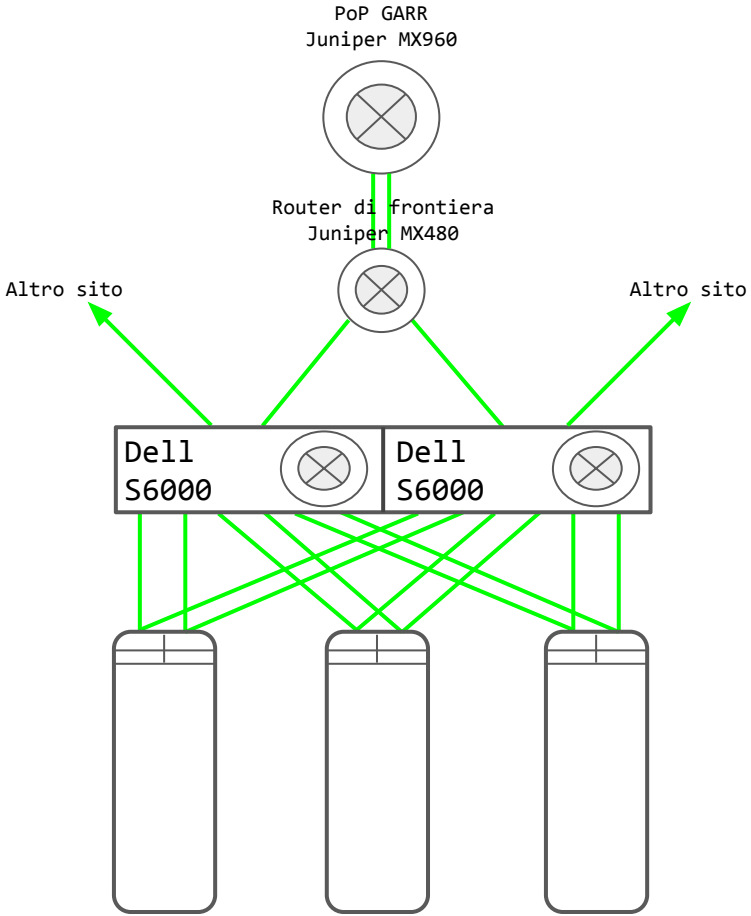
- 16 server (lame) Dell Poweredge M620
- 2 switch integrati Ethernet (Dell MXL)
  - 2x16 porte 10 Gbps -> server
  - 4 uplink 40 Gbps -> centro stella;
- 2 switch Fibre Channel (BCM57800S)
  - 2x16 porte a 16 Gbps verso i server
  - 8 uplink a 16 Gbps verso gli storage controller;

2 Storage Array **MD3860f FC:**

- Dischi SAS 116x4TB + 4xSSD 1.6TB
- FiberChannel brocade controller 2x16 Gbps (2x4 porte)

Modulo CSD

Sito



Rete Dati 40 Gbps

- 2 Switch ToR Dell MXL in ciascun modulo-CSD
- 2 Router/Switch centro stella Dell S6000
  - 32 porte x 40 Gbps

Rete di gestione (“ILO”) separata da rete Dati

- 2 Switch management ToR Dell S55 in ciascun modulo-CSD
- 2 Switch management centro stella Dell S4810
  - 48 porte x 1 Gbps

-Sistema operativo: **ubuntu** 14.04

-4 schede di rete - link aggregation **40Gbps**

-vlan + **Bridge** linux

-lama fisica fa da **GW per LXC**

-iptables su lama per:

-fwd, nat + sicurezza LXC

-indistinguibilità lame x LXC

LinuXContainer

-opensource

-supportati da kernel moderni

-Uso efficiente HW (rispetto VM)

-Near Bare Metal runtime performance

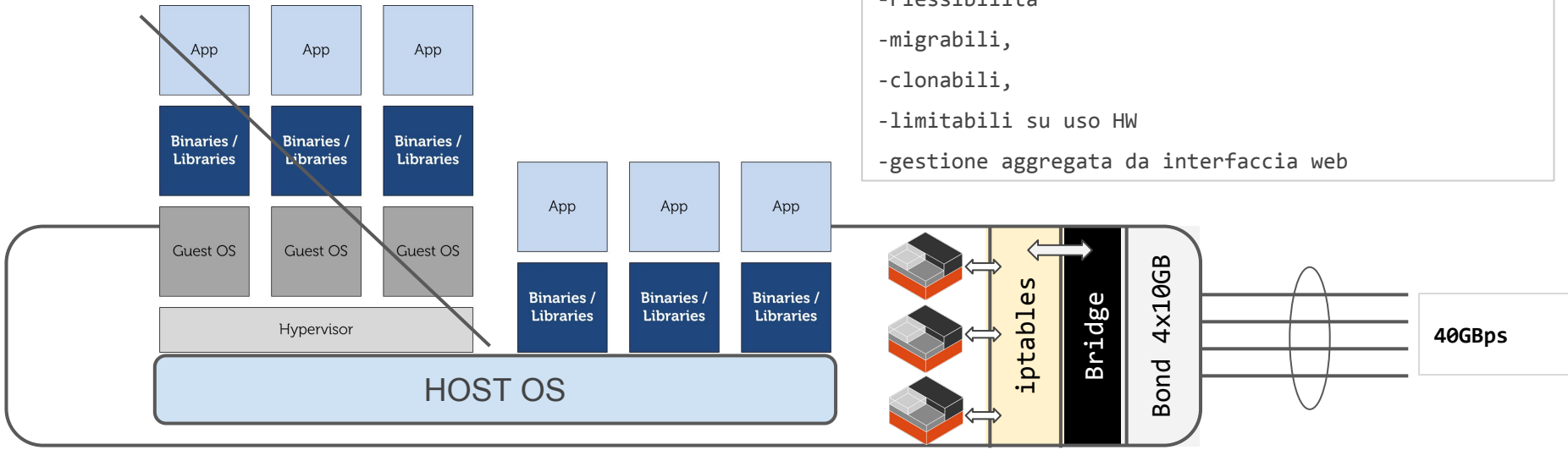
-Flessibilità

-migrabili,

-clonabili,

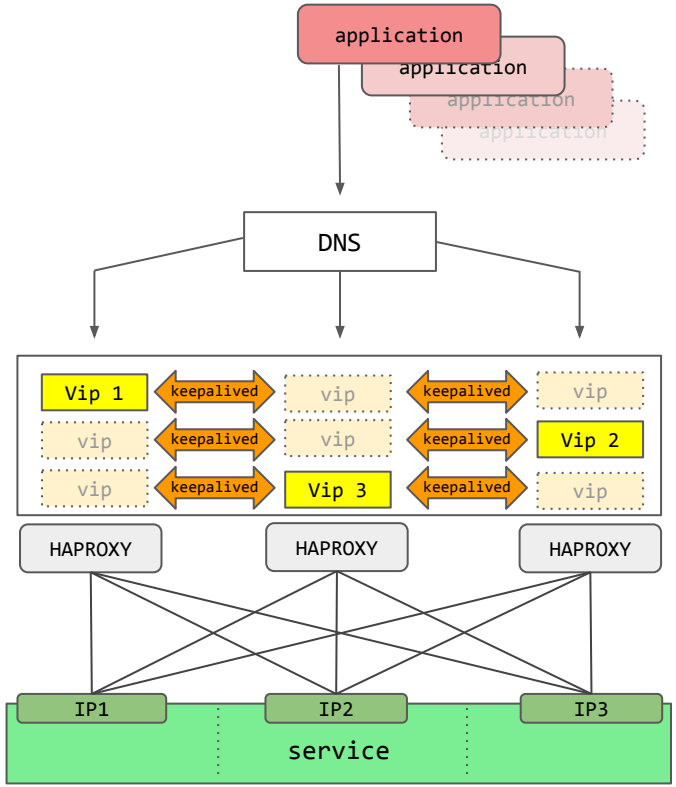
-limitabili su uso HW

-gestione aggregata da interfaccia web



Lama

Schema flusso richieste servizi



Criteri implementativi

No vendor lock in

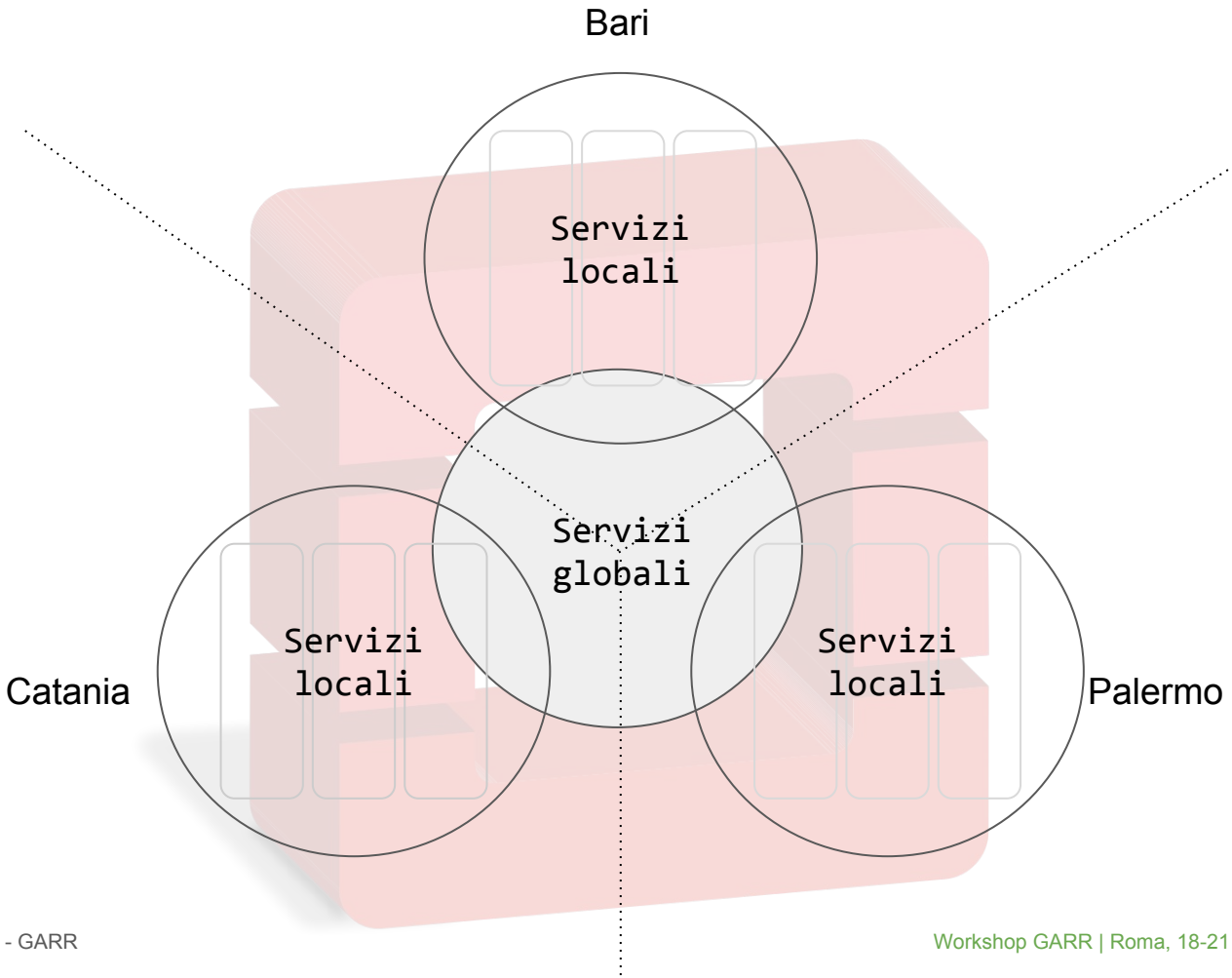
- Openstack per la piattaforma virtuale
  - Release Liberty
- Ceph (block) e Swift (object) per la fornitura di storage

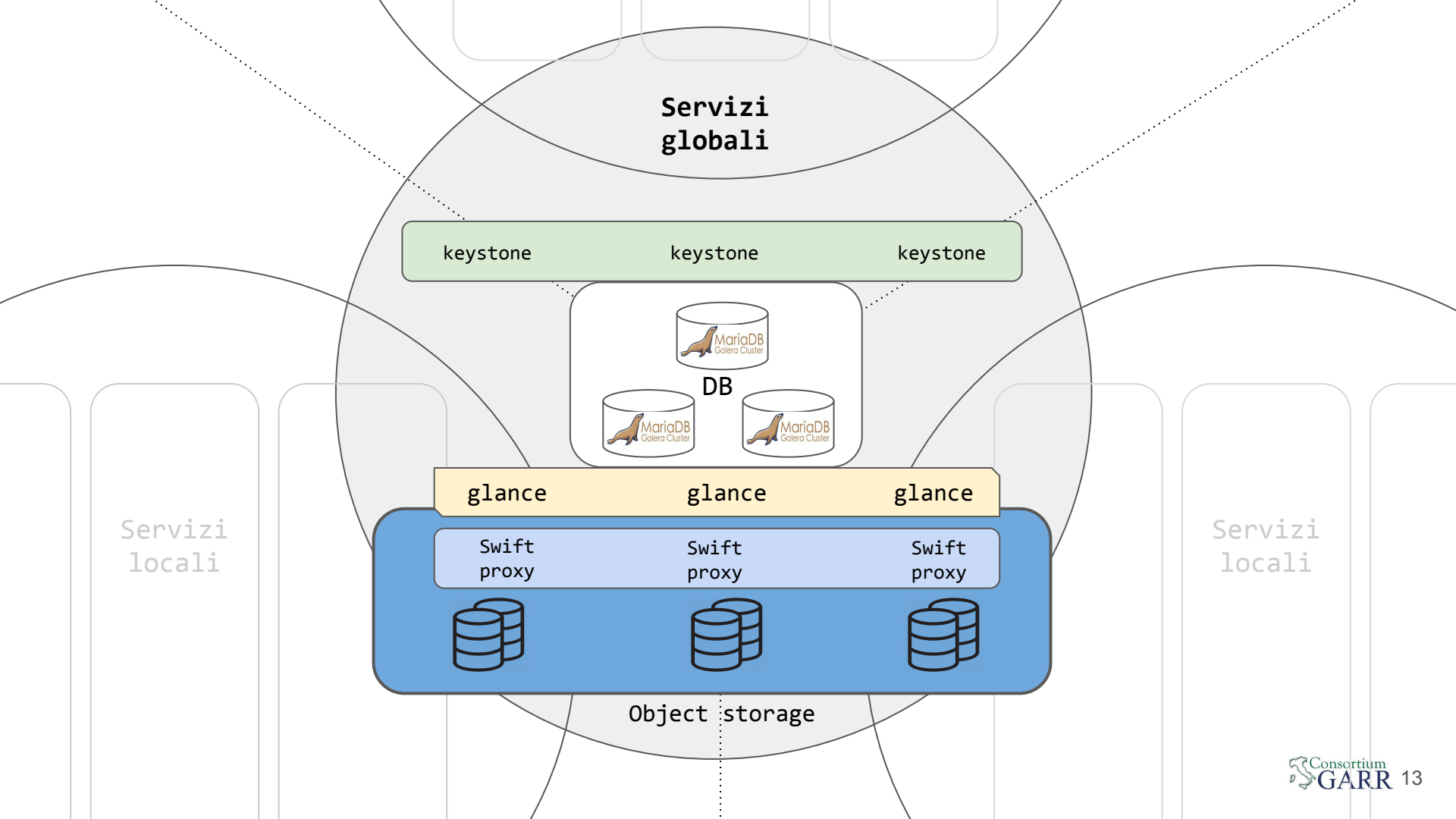
Suddivisione dei servizi di base:

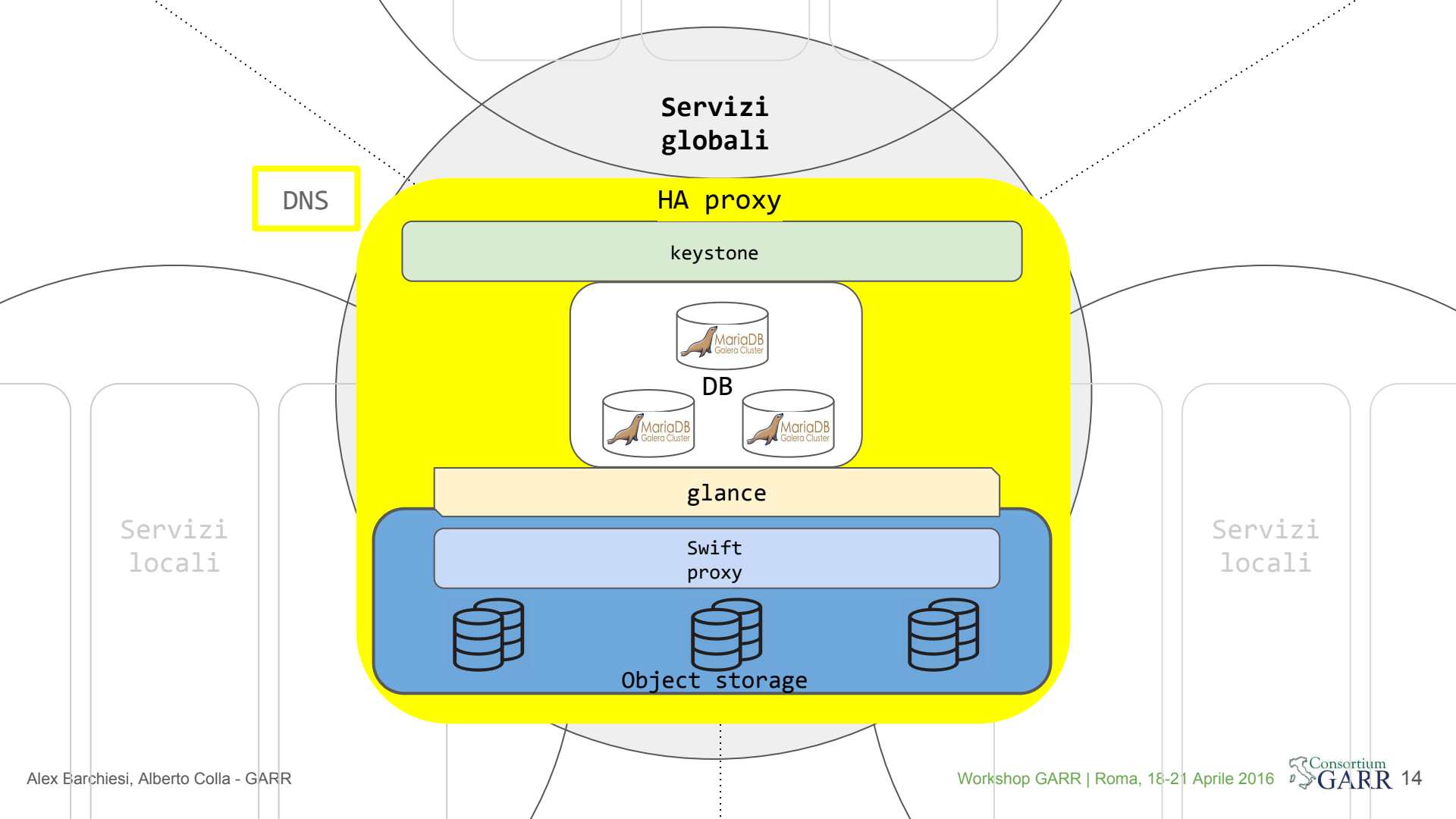
- **Globali** (unici sull'intero cluster - ridondati su **3 siti**)
  - Identity service / Keystone
  - Image service / Glance
  - Object Storage / Swift
- **Locali** (individuali su ciascun sito - ridondati su **3 rack**)
  - Controller service / Nova
  - Network service / Neutron
  - Block Storage / Ceph

Ciascun sito individua una Openstack **Region**

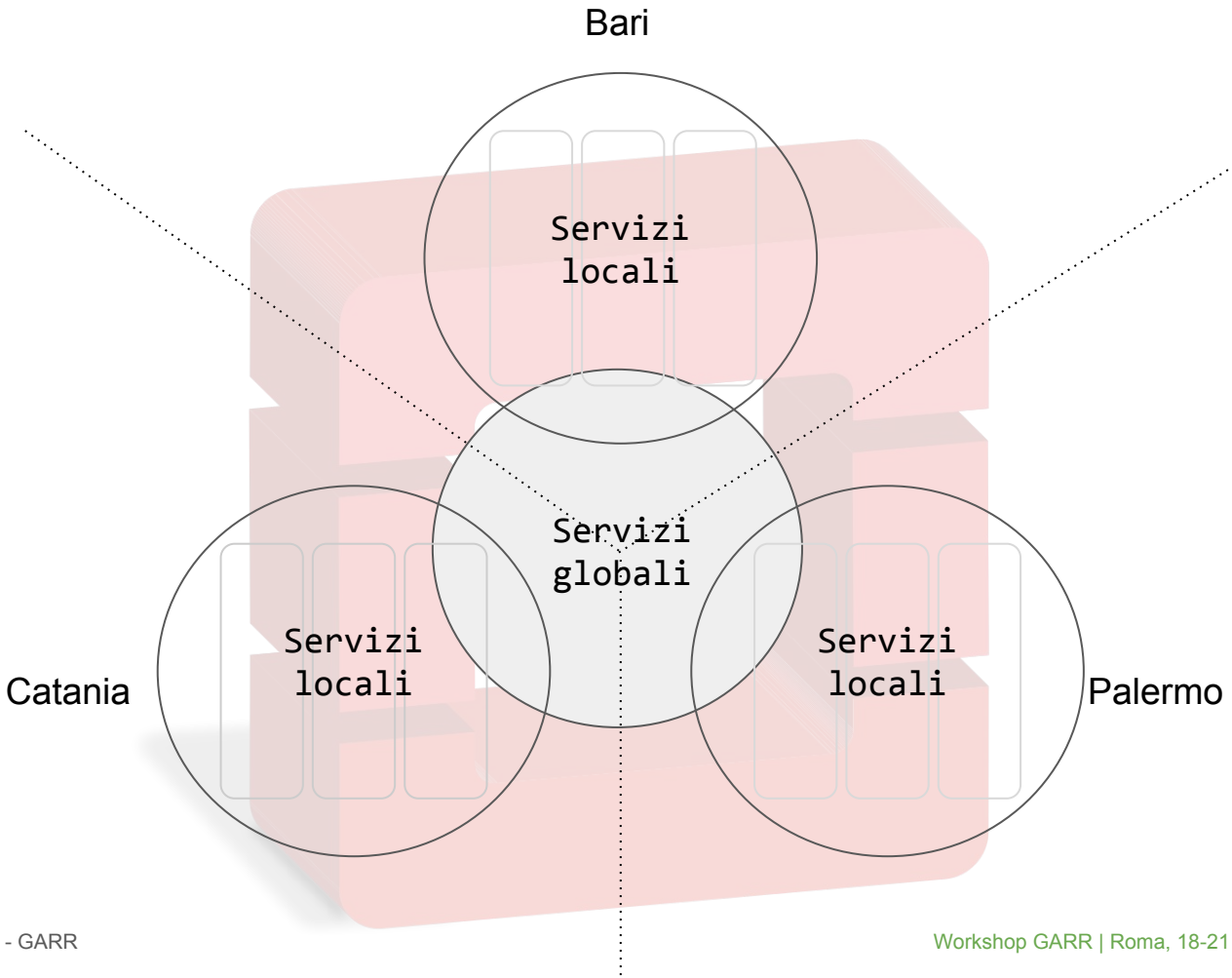
Architettura



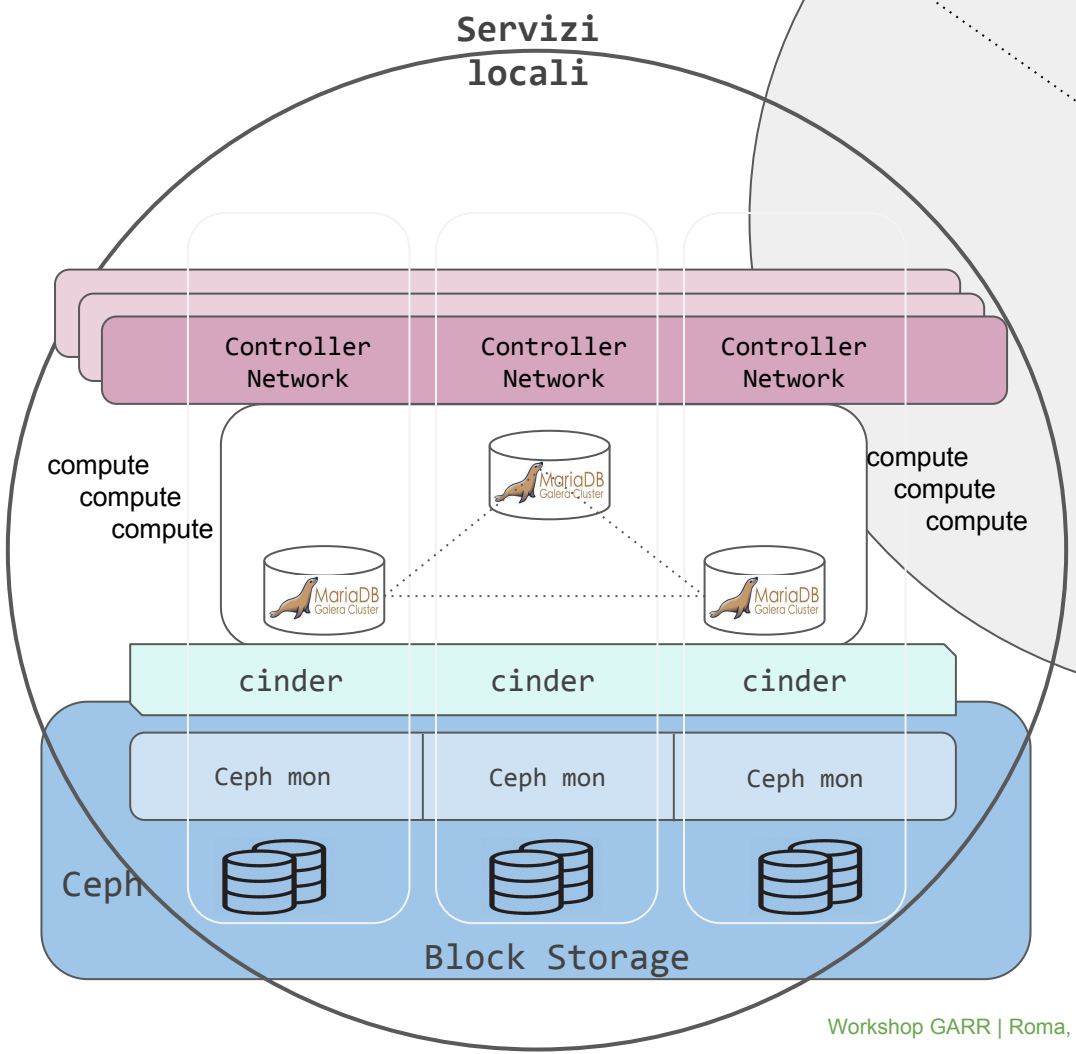

Servizi
globali
keystone
keystone
keystone
MariaDB
Galera Cluster
DB
MariaDB
Galera Cluster
MariaDB
Galera Cluster
glance
glance
glance
Swift
proxy
Swift
proxy
Swift
proxy
Object storage
Servizi
locali
Servizi
locali


Servizi
globali
DNS
HA proxy
keystone
MariaDB
Galera Cluster
DB
MariaDB
Galera Cluster
MariaDB
Galera Cluster
glance
Swift
proxy
Object storage
Servizi
locali
Servizi
locali

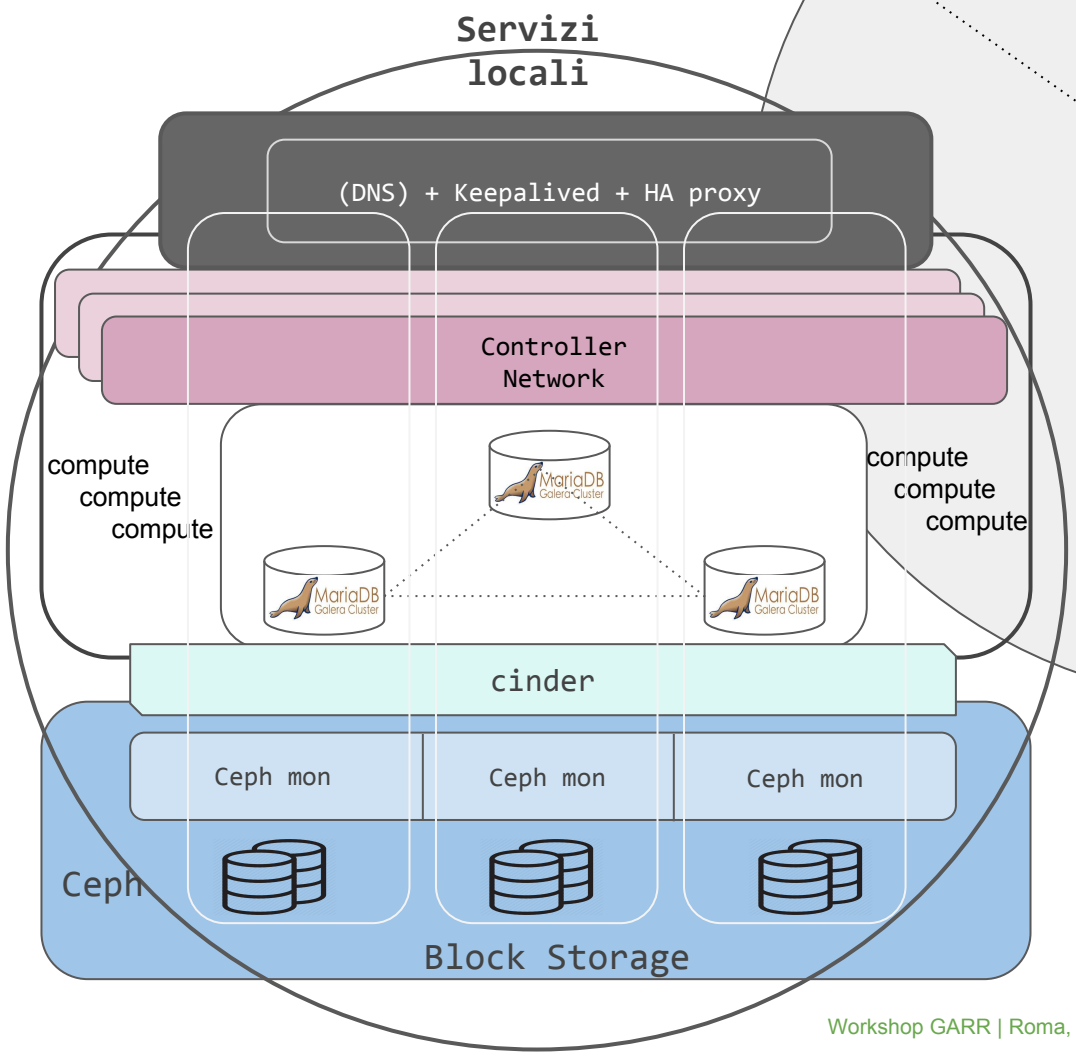
Architettura



3x


Servizi locali
Servizi globali
Controller Network
Controller Network
Controller Network
compute
compute
compute
compute
compute
compute
MariaDB Galera Cluster
MariaDB Galera Cluster
MariaDB Galera Cluster
cinder
cinder
cinder
Ceph mon
Ceph mon
Ceph mon
Ceph
Block Storage

3x


Servizi locali
Servizi globali
(DNS) + Keepalived + HA proxy
Controller
Network
compute
compute
compute
compute
compute
compute
MariaDB Galera Cluster
MariaDB Galera Cluster
MariaDB Galera Cluster
cinder
Ceph mon
Ceph mon
Ceph mon
Ceph
Block Storage

Highlight scelte implementative (1/4)

Compattezza

- Core services Openstack ospitati su Linux Containers
  - Componenti locali replicati su **3 lame su 3 rack diversi**
  - Componenti globali replicati su **3 siti**

Throughput

- Schede di rete delle lame configurate in bonding active/active
  - Bandwidth **40 Gbps** verificata con test iperf

Highlight scelte implementative (2/4)

Resilienza/Load balancing

- servizi globali: ridondanza via **DNS**
  - 1 hostname globale risolto da più record-A
  - Resilienza servizi globali verificata contro:
    - Shutdown processo sul container
    - Shutdown container
    - Breakdown networking intero sito
- servizi locali: ridondanza e load balancing via **DNS + Keepalived + HAproxy**
  - tempi di risposta uguali anche in caso di perdita di un membro del cluster
- **Galera multi-master** per i database
- HA Keystone via **Fernet tokens**
- **Rabbit cluster** (3 membri) locale

Highlight scelte implementative (3/4)

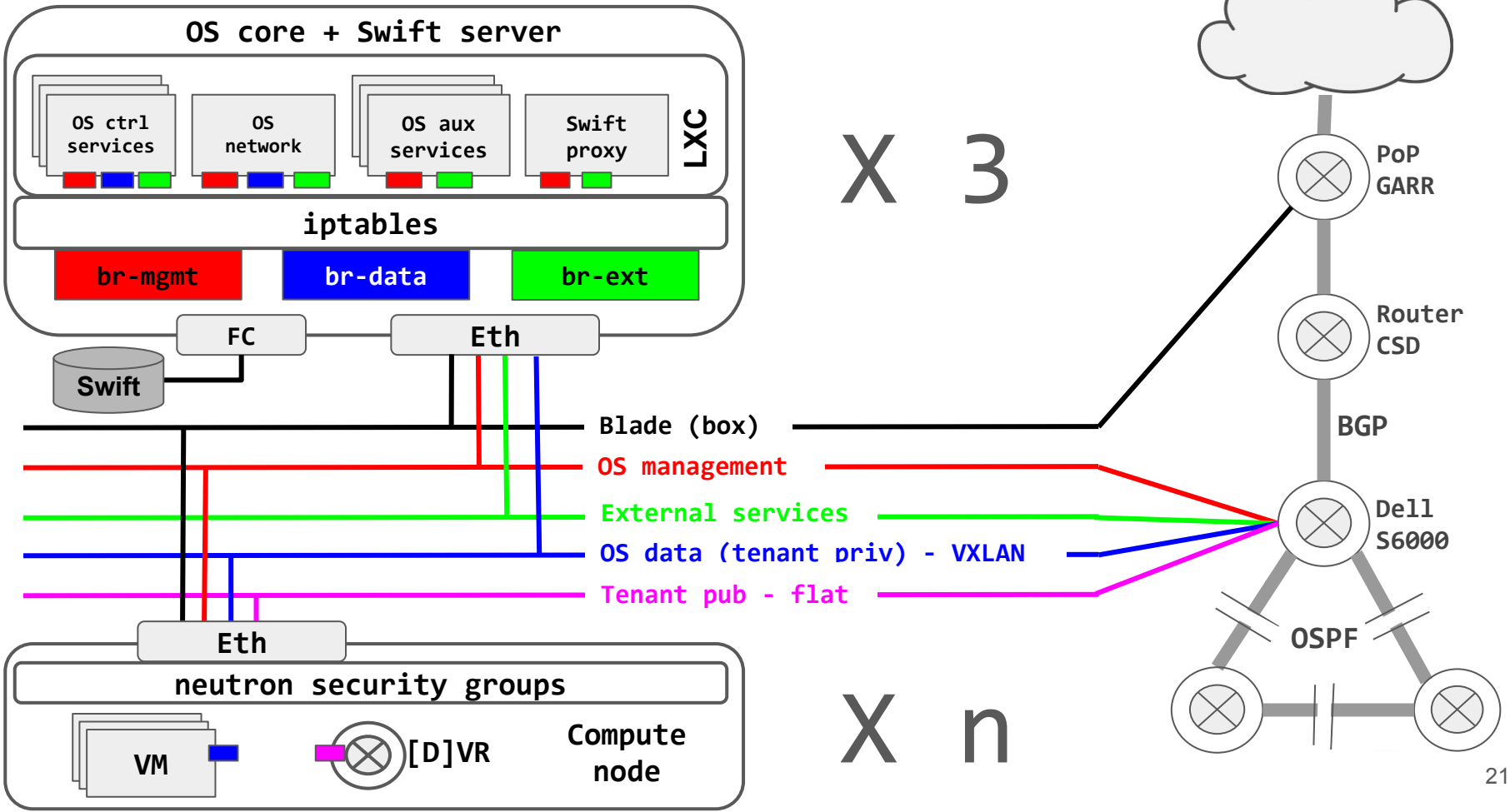
Networking

- Separazione L2 (VLAN) delle reti server e delle reti di Openstack
- Separazione reti tenant via VXLAN
- Networking inter-sito servizi openstack e tenant via IP su link dedicato

Sicurezza

- Servizi Openstack su LXC: iptables sui server fisici ospitanti
- VM Openstack sui compute: Neutron Security groups (iptables)
- ACL sul router di frontiera

Networking Openstack



Highlight scelte implementative (4/4)

Storage

- Object storage (**Swift**) globale
  - replica 3 su 2 siti
  - Gestione utenti via Keystone
  - default backend per immagini Glance
- Block storage (**Ceph**) locale
  - replica 3
  - distribuito su 3 rack
  - default backend per volumi VM Openstack

Gestione failover: scenari

Indisponibilità container / servizio su container

- Resilienza garantita da DNS + Keepalived + HAproxy

Perdita di un modulo-CSD

- Servizi core openstack in HA
- VM istanziate su volumi Ceph evacuate (*nova evacuate*) e riattivate sugli altri moduli senza perdita dati
  - Possibilità di gestione automatica failure

Perdita di un sito

- Servizi globali openstack resilienti (verificato)
- VM con ephemeral disk:
  - Snapshot periodica su Glance / Swift -> immagini disponibili sull'intero cluster
  - Respawn da snapshot su altro sito (implica replica reti)
- VM su Ceph: in studio

Stato attuale implementazione

- Implementazione componenti globali (Keystone, Glance, Swift) completata
  - Installati al momento 2 server x 3 siti
- Installazione componenti locali in corso
- Swift+Keystone utilizzati per sperimentazione su Object Storage federato INFN+GARR

Roadmap

- Completamento implementazione cluster multisito
- Full stress test
- Monitoring infrastruttura
  - Gestione failover via update automatico record DNS
- Migrazione servizi GARR sul cluster multisito

Sommario

- GARR puo’ offrire un ambiente cloud distribuito, sicuro, performante, resiliente, scalabile
  - Il modello potrebbe essere esteso ad altri siti
- IaaS e’ utile, ma la vera ”intelligenza” deve stare al livello superiore
  - Meccanismi di adattamento dinamico, nascondere i dettagli del “ferro”
- Federazione di risorse, federazione di utenti
  - Centralizzare vs. aggregare

Grazie