ITALY
OpenInfra Days

OPENINFRA DAYS ITALY

Organized by

IRIDEOS

Binario Etico

Under the patronage of

AGID | Agenzia per l'Italia Digitale

Sponsored by

Mellanox TECHNOLOGIES

MESOSPHERE

gci SERVICE FACTORY

Alex Barchiesi - GARR

Rome, October 3rd, 2019

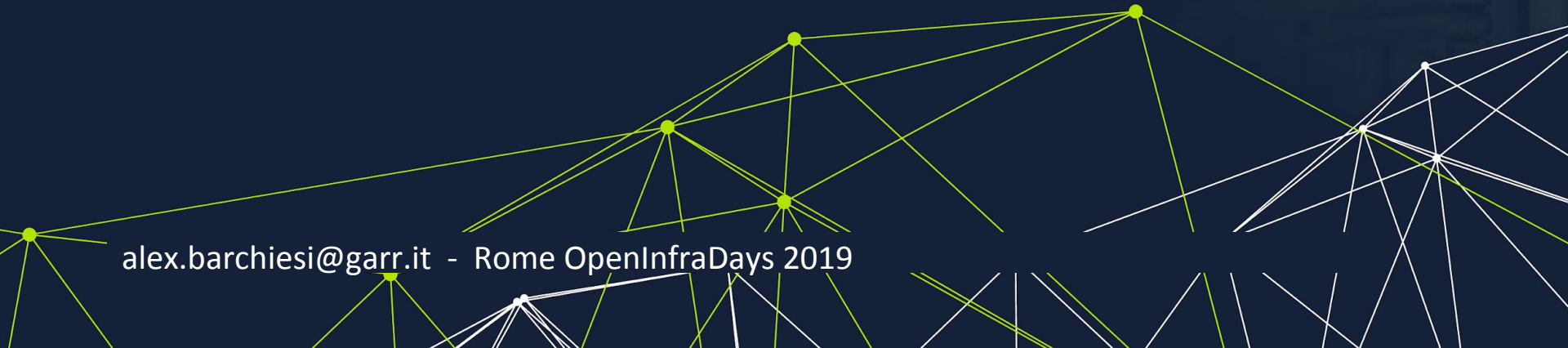
GARR*bage* collection:

aka how to build a federation

Consortium GARR
THE ITALIAN EDUCATION & RESEARCH NETWORK

GARR*bage* collection:

aka how to build a federation



alex.barchiesi@garr.it - Rome OpenInfraDays 2019

you cannot fly...until you *let go*

[Joe, volem volar trapeze company]

Re-cap of GARR mission and infrastructure

- […] fornire servizi per favorire **l’armonizzazione, l’implementazione e la gestione delle e-Infrastructure a vantaggio della comunità scientifica e accademica** nazionale; […]
- […] sostenere e **stimolare lo sviluppo di strumenti atti a facilitare l’accesso** alle risorse di calcolo, supercalcolo e storage a livello nazionale ed internazionale, **fornendo gli opportuni metodi, interventi e funzionalità** necessari a mantenere le e-Infrastructure ai livelli degli standard internazionali; […]
- […] svolgere le connesse attività di ricerca tecnologica, sperimentazione, trasferimento tecnologico e **formazione del personale**. […]

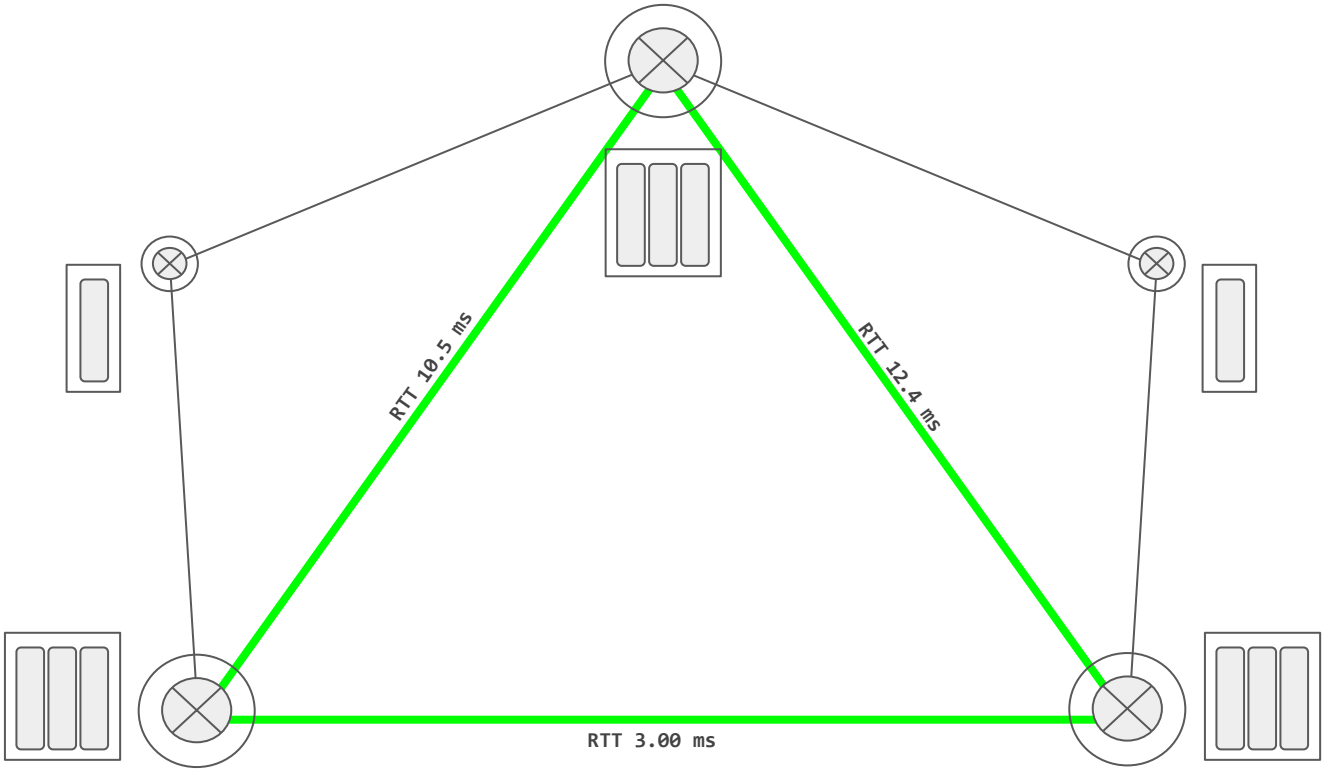
- **Who:**
  - no profit association: **CNR, ENEA, INFN,** Fondazione **CRUI** (universities)
- **Role**:
  - resource **aggregator** (federation)
- **Goals**:
  - **simplify** provisioning of storage and computing
  - serve **different** organizations
  - **Simplify** experience (empower administrators)

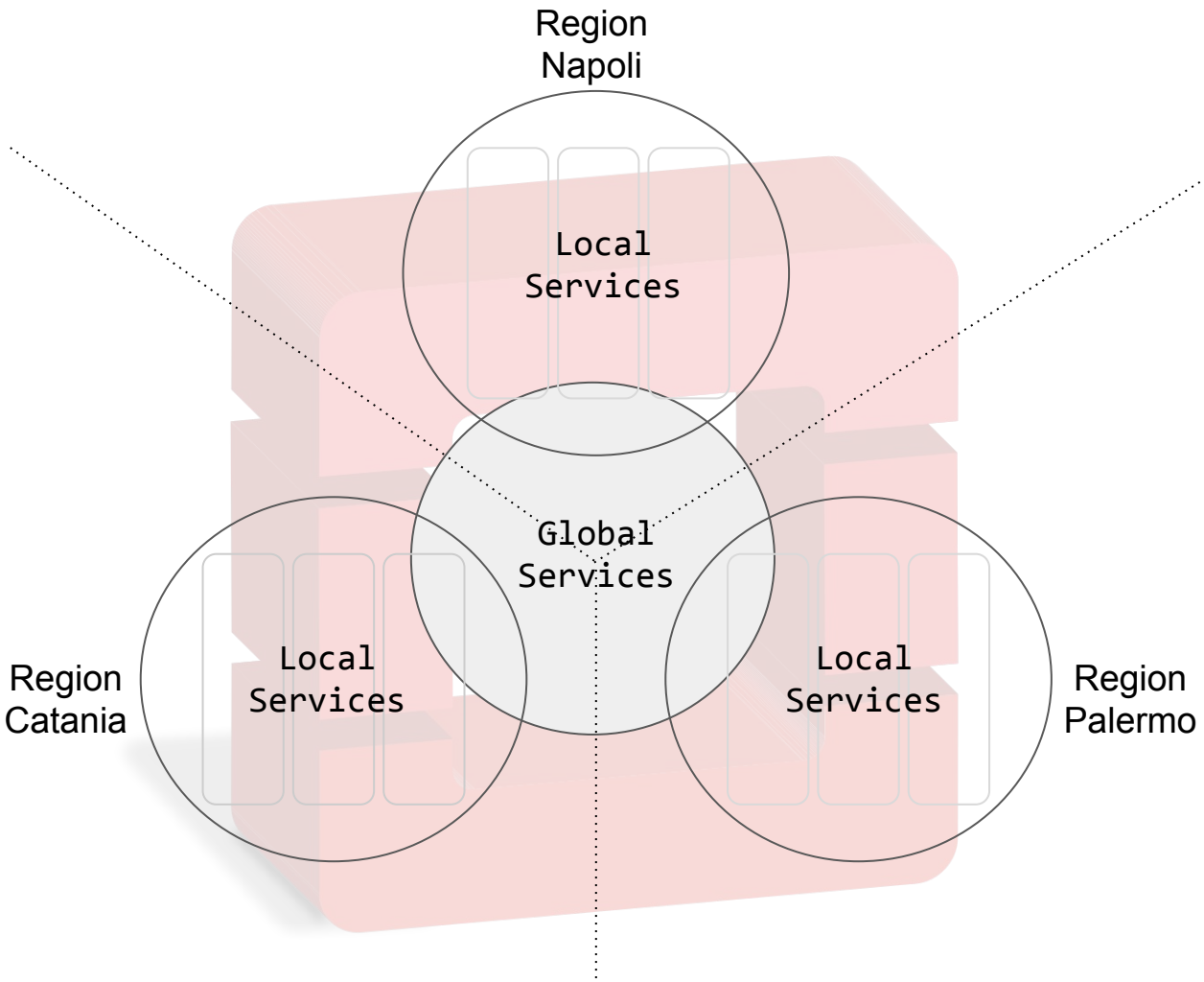
8500 core
10 PB

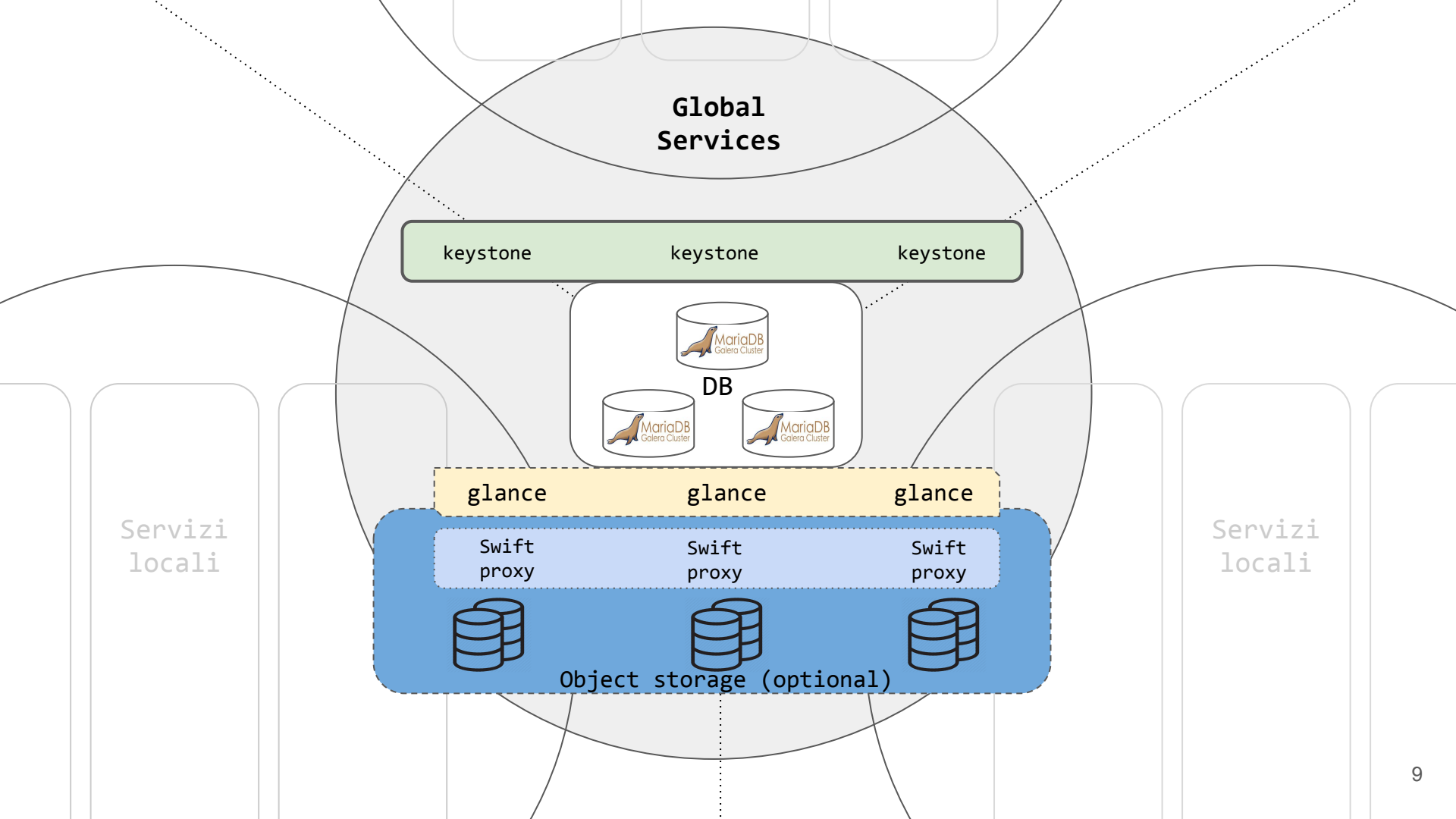
... 11 rack/CSD-modules

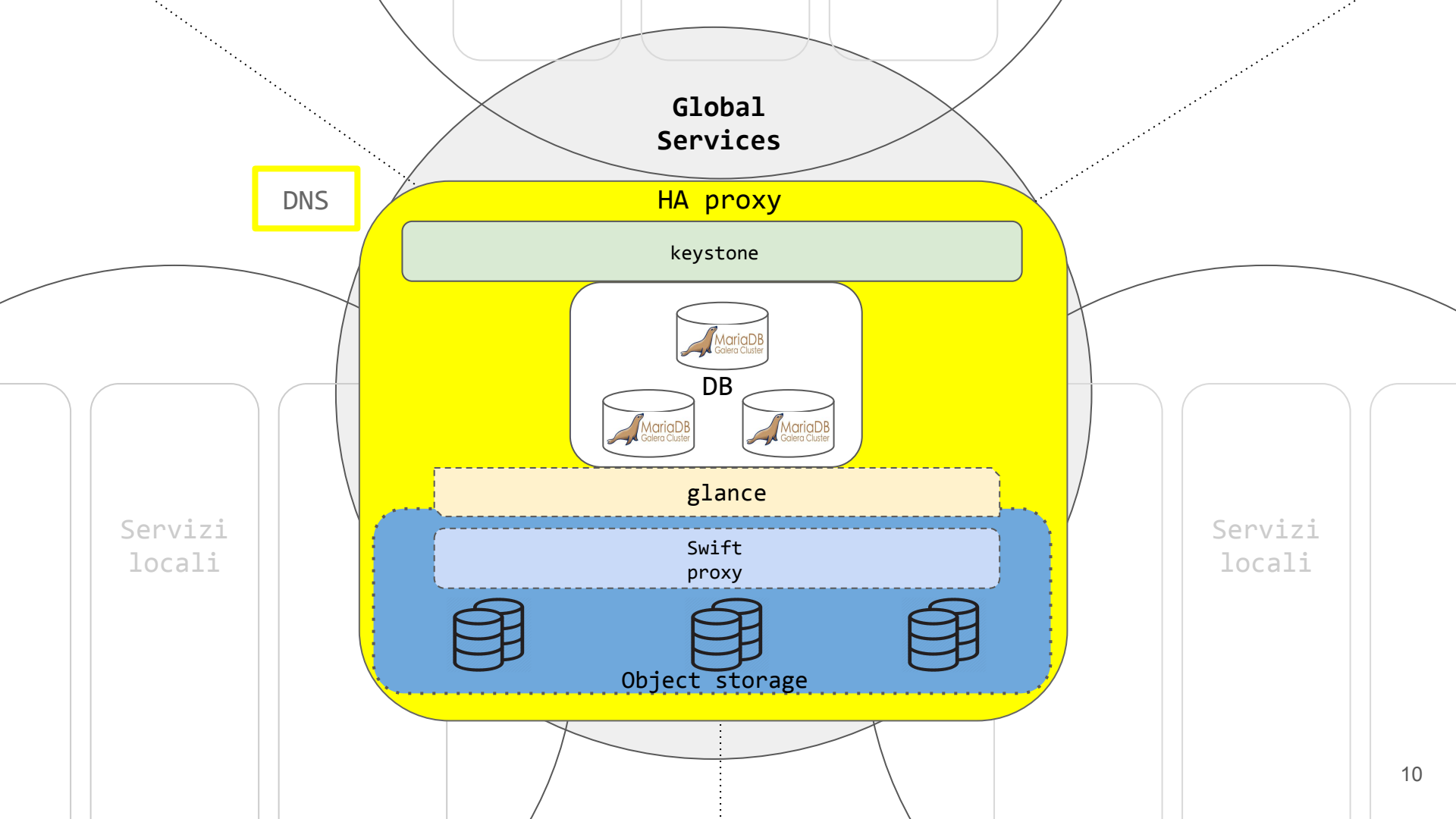
Network

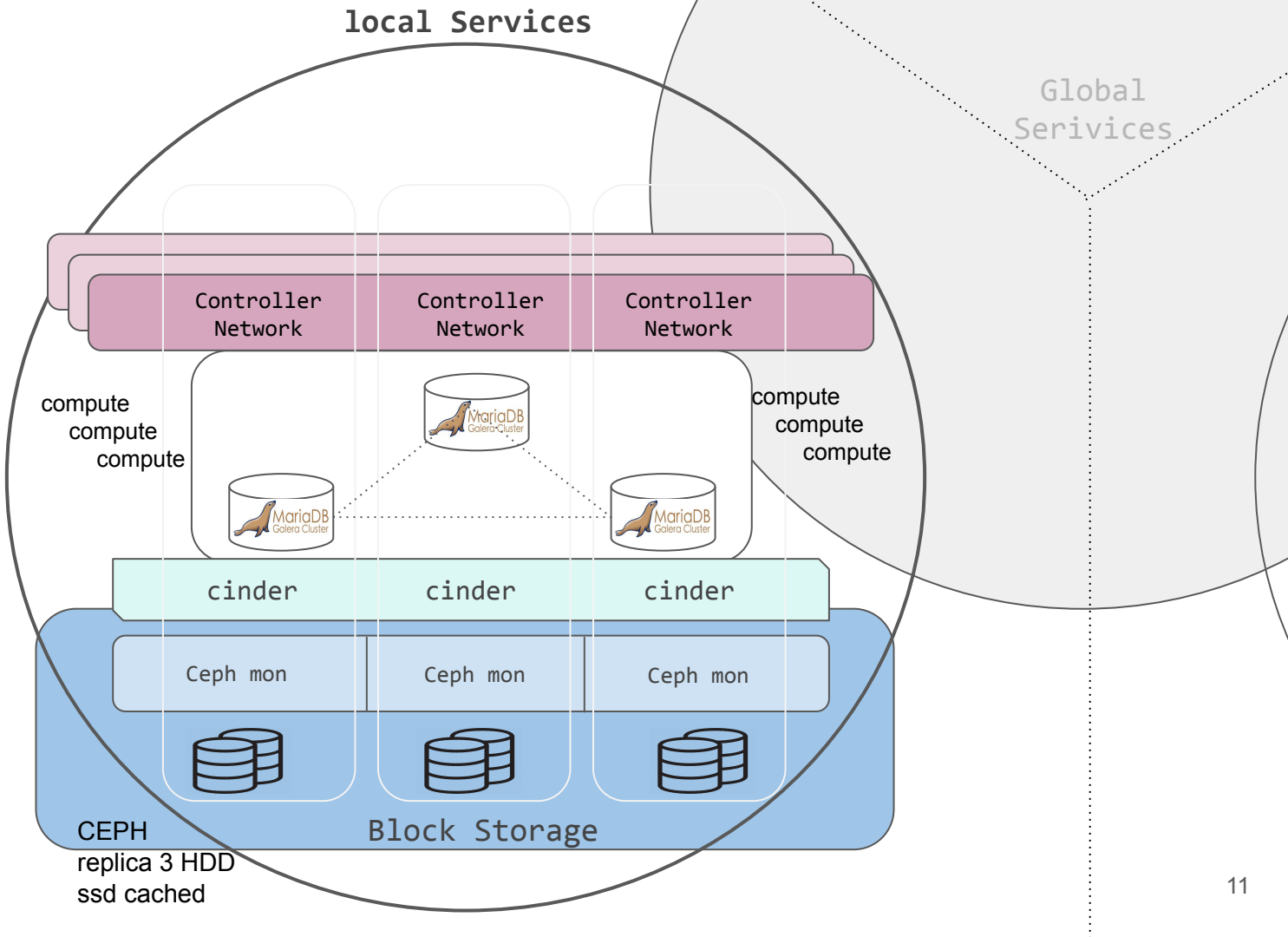
40Gbps
10Gbps



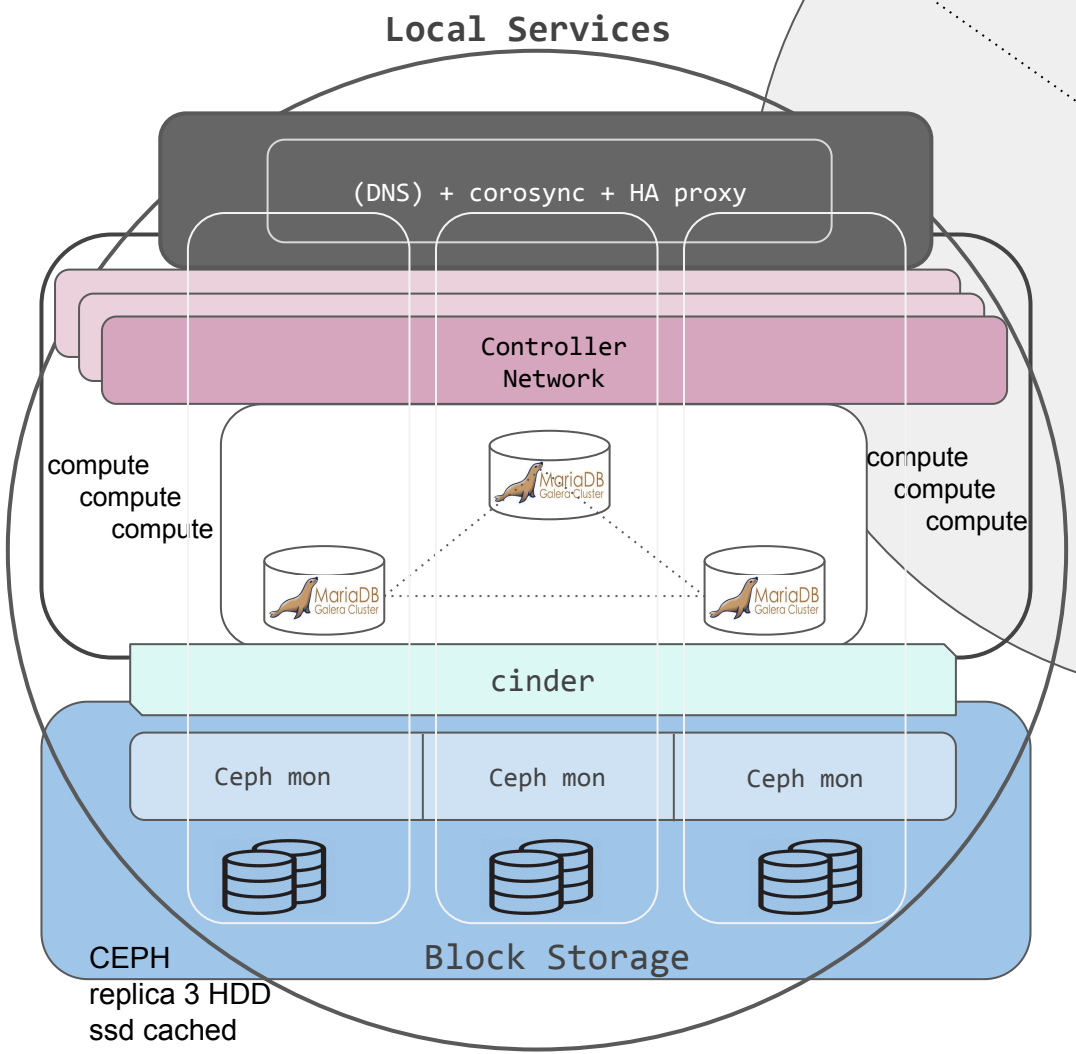

Region
Napoli
dns
Local
Services
Global
Services
Region
Catania
Local
Services
Local
Services
Region
Palermo


Global
Services
keystone
keystone
keystone
MariaDB
Galera Cluster
DB
MariaDB
Galera Cluster
MariaDB
Galera Cluster
glance
glance
glance
Swift
proxy
Swift
proxy
Swift
proxy
Object storage (optional)
Servizi
locali
Servizi
locali


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

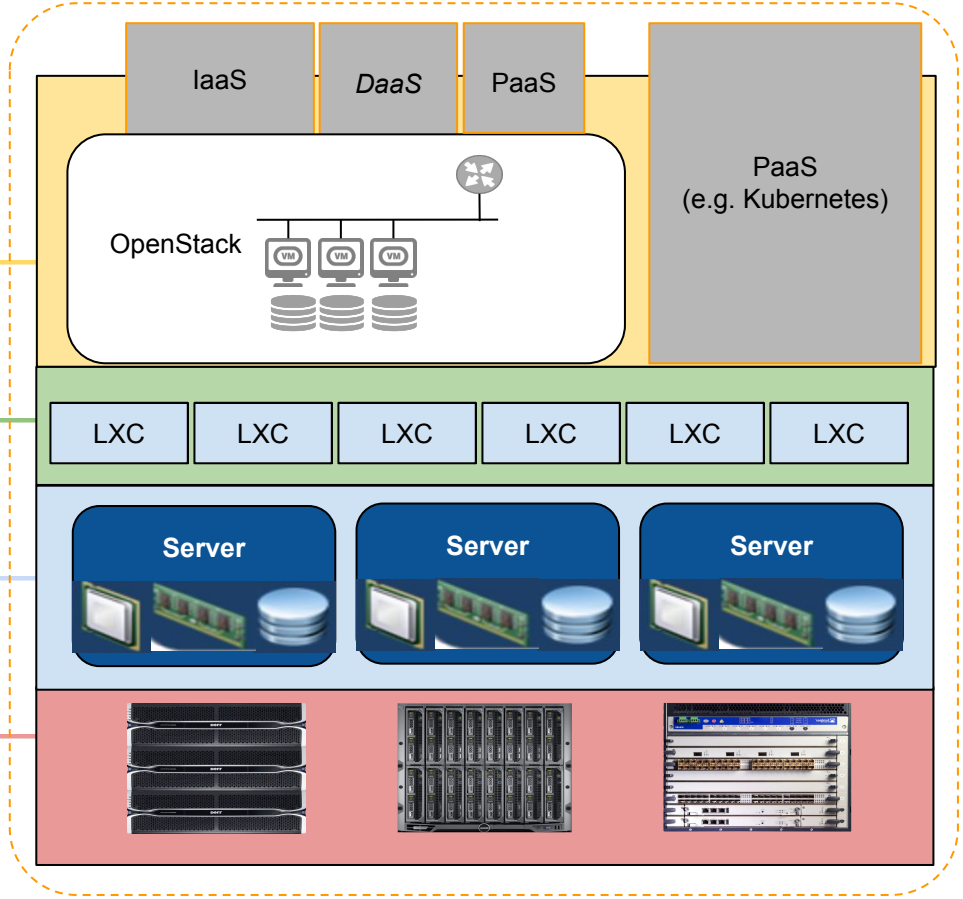

local Services
Global
Serivices
Controller
Network
Controller
Network
Controller
Network
compute
compute
compute
compute
compute
compute
MariaDB
Galera Cluster
MariaDB
Galera Cluster
MariaDB
Galera Cluster
cinder
cinder
cinder
Ceph mon
Ceph mon
Ceph mon
Block Storage
CEPH
replica 3 HDD
ssd cached

GARR = 3x


Local Services
Global Services
(DNS) + corosync + 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
Block Storage
CEPH
replica 3 HDD
ssd cached

4 Layers recipe:

1. Application Services
2. Infrastructure *Virtualization*
3. Operating System
4. Physical resources



4 Layers recipe:

1. Application Services
2. Infrastructure *Virtualization*
3. Operating System
4. Physical resources

IaaS
DaaS
PaaS
PaaS
(e.g. Kubernetes)
CANONICAL
OpenStack
juju
+
MAAS

-Operating system: **Ubuntu** Xenial

-4 NIC - link aggregation **40Gbps**

-vlan + Bridge linux

-blade is the **GW** for **LXC**

-iptables:

-fwd, nat + security LXC

-blade interchangeable from p.v. LXC

MAAS + juju *

LinuXContainer (LXC with LXD managment)

-opensource

-modern kernel support

-Efficient resource usage (compared to VM)

-Near Bare Metal runtime performance

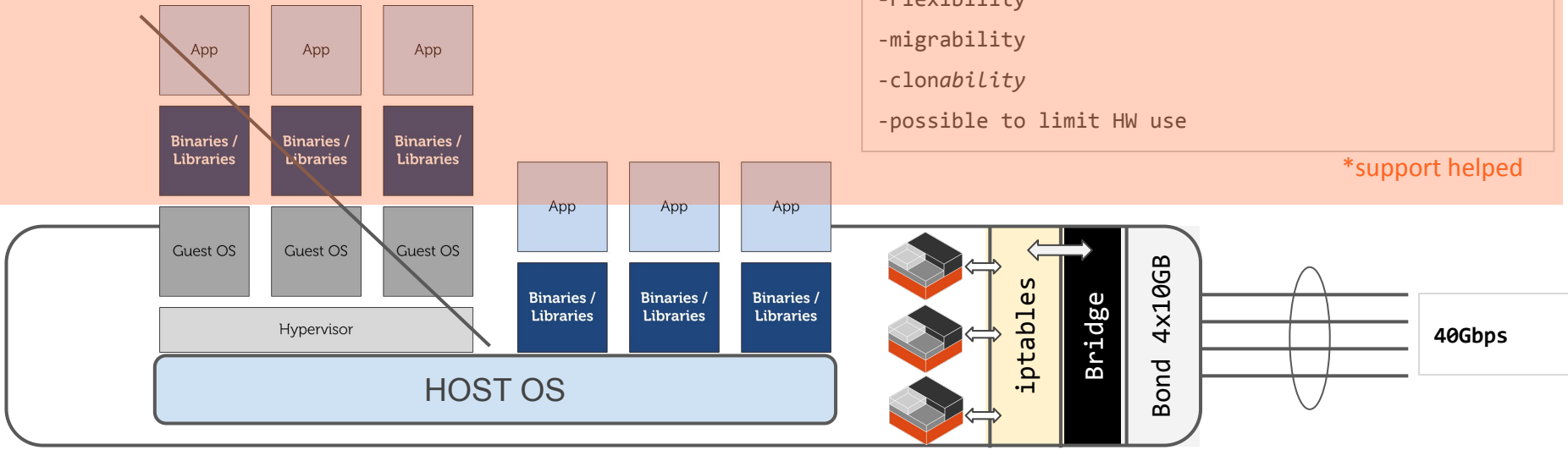
-Flexibility

-migrability

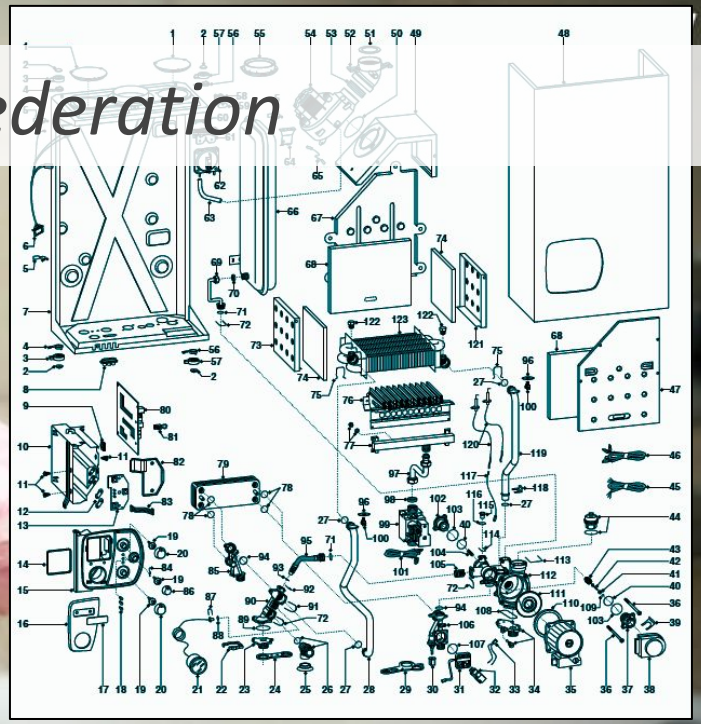
-clon*ability*

-possible to limit HW use

*support helped



our concept of *Federation*



- the simplest (for federated region admin) the better
  - the less requirements the more inclusive

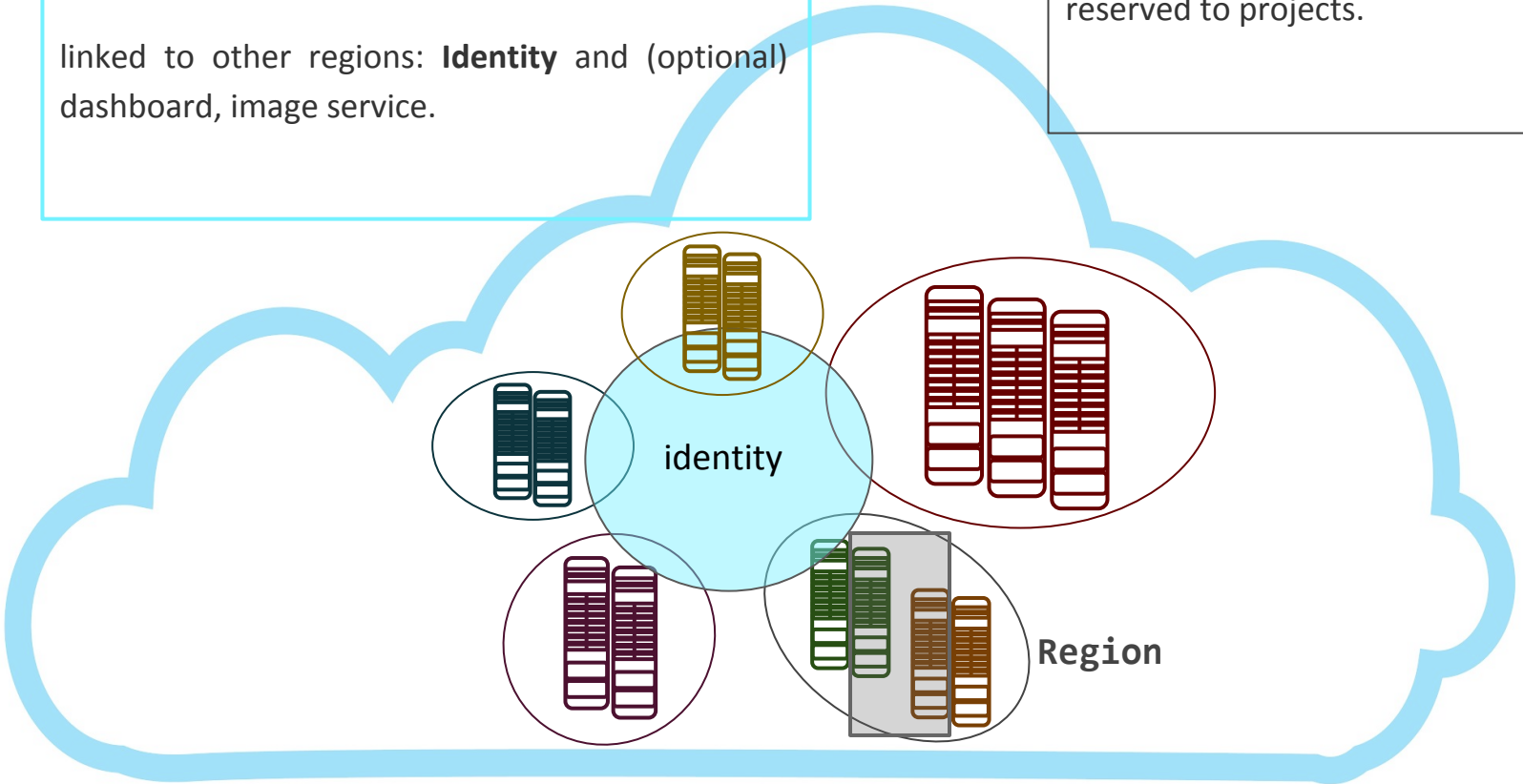
multi-region/multi-domain model

Region: own deployment of OpenStack

linked to other regions: **Identity** and (optional) dashboard, image service.

Inside a **Region**: advanced scheduling

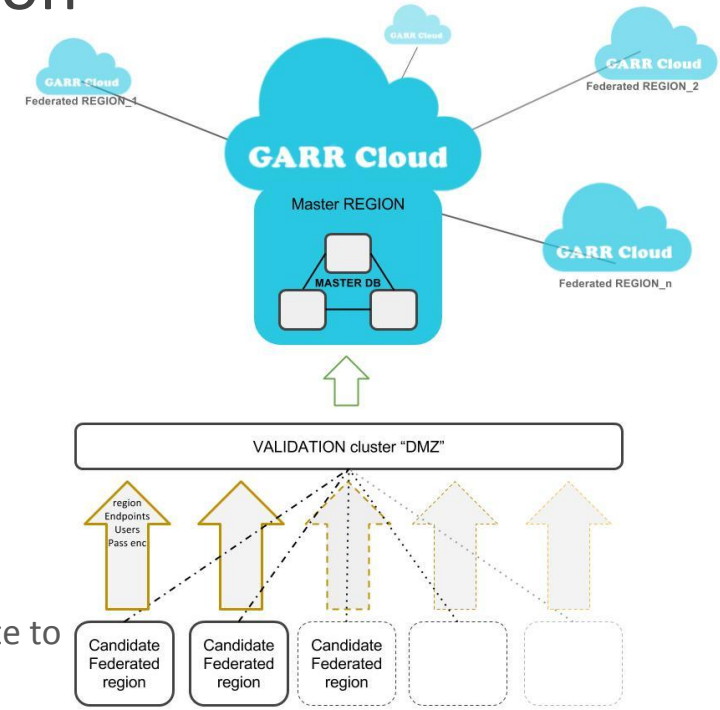
Availability Zone:
nodes can be logically grouped into AZ and reserved to projects.



joining the Federation

Procedure of inclusion

- Bundle OpenStack - attaches to validation cluster
- Validation in “DMZ” cluster
- **No cleartext credentials exchange**



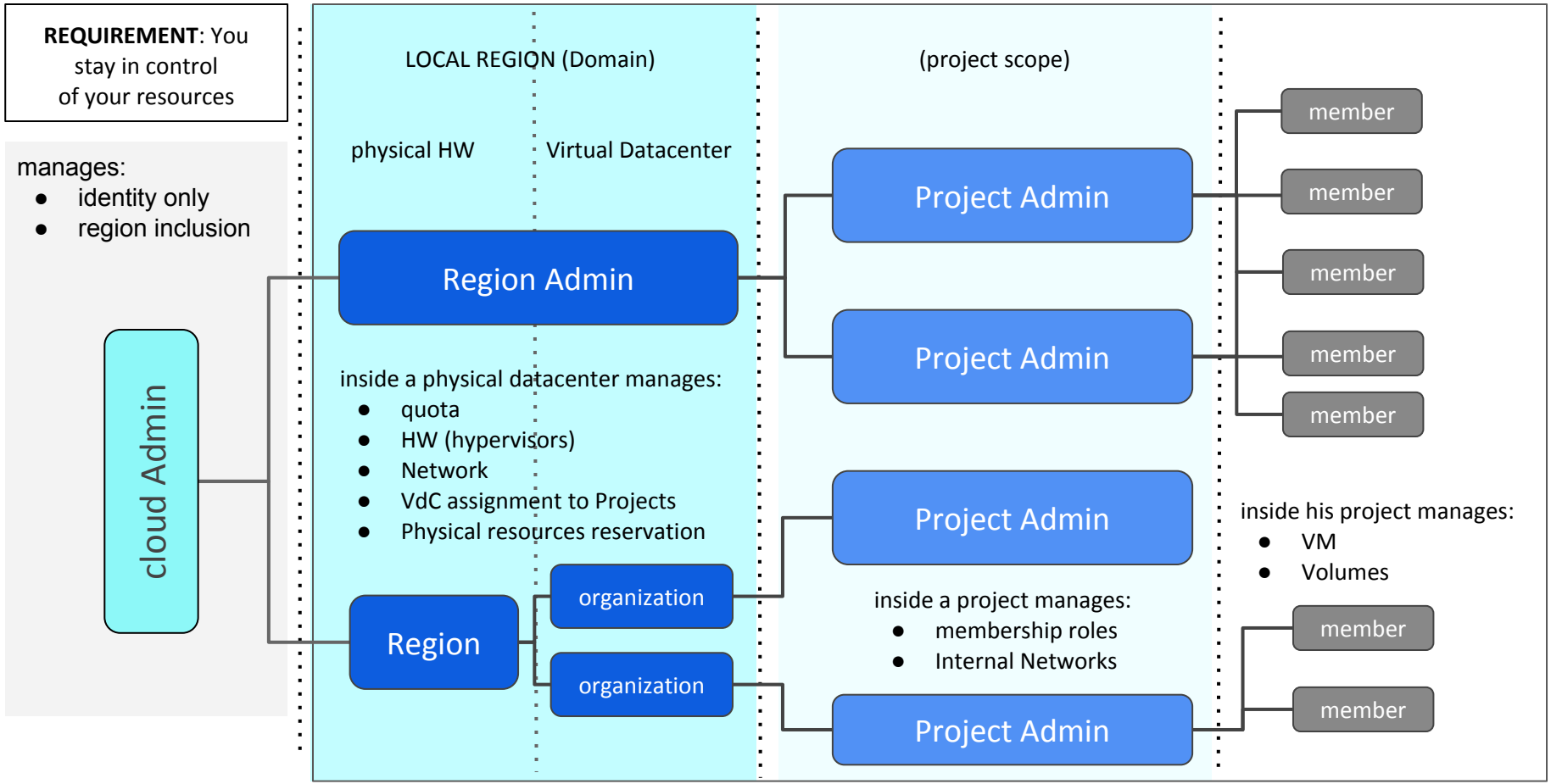
for research institutes:

1. You own HW, but have no manpower/knowledge (yet)
2. You already have an OpenStack deployment (recent one)
3. None of the previous, but you have men-power

but for everyone...**everything is publicly available** (see next) and we contribute to community projects:

- Horizon
- K8s-keystone-auth
- juju charms: ceph, keystone saml/sso, default gw, moodle...

Delegation of authority (via domains, policy and metadata filters)



(few) details about deployment *requirements*

networking prescriptions

1. private network
   - 1.1. ipmi (usually untagged)
   - 1.2. pxe/boot/management (best practice untagged - simplifies the setup of pxe)
   - 1.3. Storage ceph “priv” (to be used by OSDs only)
   - 1.4. Storage ceph “pub” (used by MON and Ceph clients, it normally is a private network)
   - 1.5. OpenStack data+mgmt

2. public network
   - 2.1. Public ip (for infrastructure and Cloud service frontend)
     Needed minimum 40 IP (suggested to have a 3 members HA of each service: /25 subnet)
   - 2.2. OpenStack floating ip - (this is basically the number of VM that you foresee to have publicly exposed)
     To be evaluated according to computing resources/use case

Best practice:

- link aggregate (bond) all interfaces and set a virtual interface (vlan) for each of the previously mentioned networks, except PXE.
- keep IPMI network separated.
- NB configure ILO/IDRAC with IPMI over LAN

Note: The networking must be configured on the switches/routers while MAAS takes into account the server configuration (ref https://docs.maas.io/2.5/en/installconfig-networking)

firewall: ports needed

egress: 80, 443, 5000, 35357, 8774, 8776, 8778, 9292, 9696, 6080

ingress (only needed on public network from subnets which need access to OpenStack and NB from GARR-keystone network): 80, 443, 8774, 8776, 8778, 9292, 9696, 6080, 5000

N.B. open port 5000 to test “local” keystone functionality, after including the region in the federation it can be closed again.

Automation and deployment: MAAS + juju

MAAS (responsible of physical machine deployment - ref: https://maas.io/)

Note MAAS is not a demanding service in terms of CPU, RAM, Disk, Bandwidth.

1 physical machine (2 if in HA):

- LXD host (MAAS + Juju + OpenStack clients)
- Hypervisor host (mainly for VM hosting Juju controller)

container LXC:

- Region controller (needs to reach ipmi network (ipmi network routed towards MAAS)
- Rack controller (may be co-located with Region)
- NAT (Note this is a Gateway for outgoing requests of services with only private networks, e.g. installation/upgrade of DB)

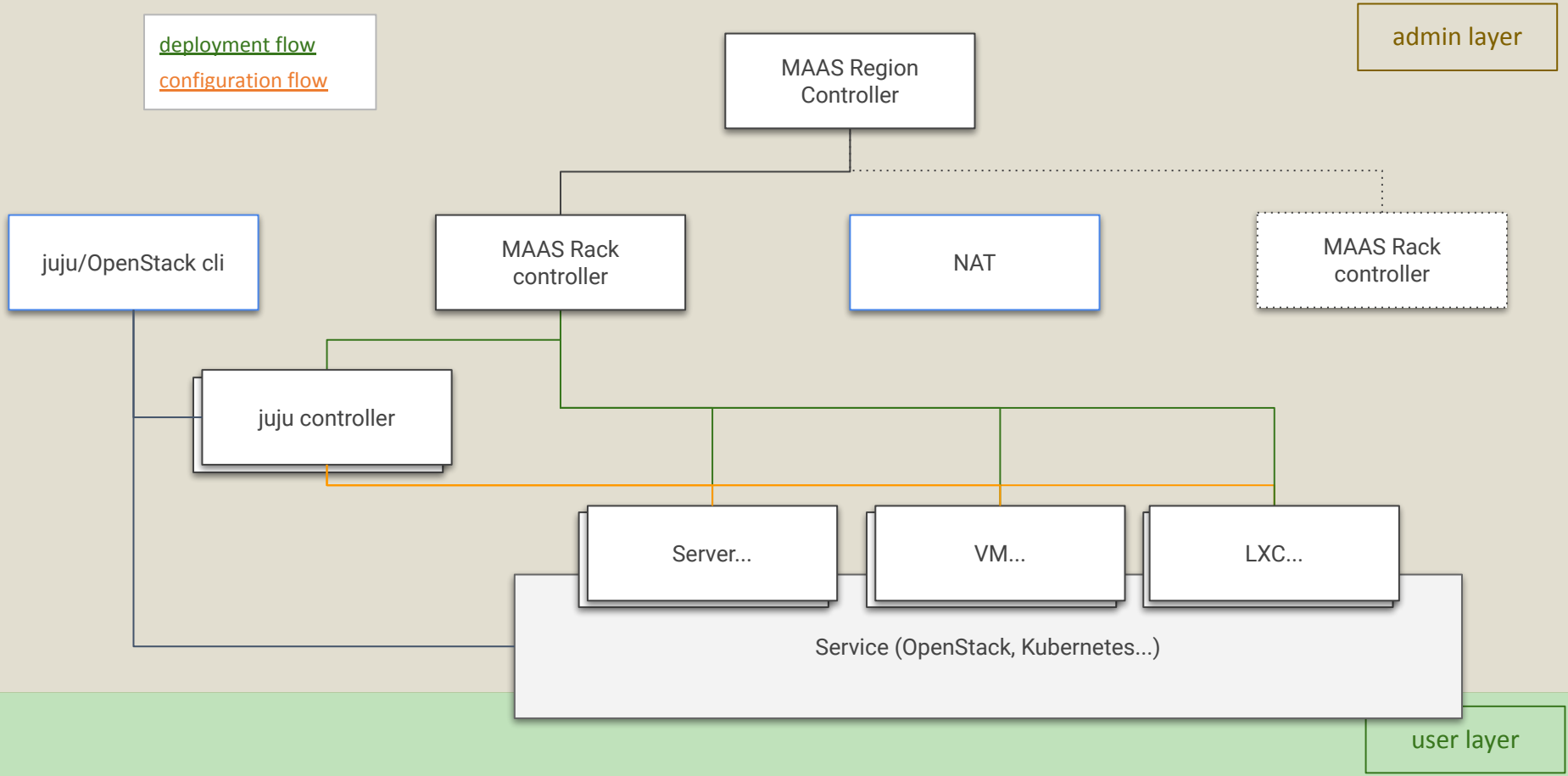
Best practice:

- configure HA on the hosted services
- Link aggregation of network interfaces (bond) + virtual interfaces (vlan)
- You can also install Region and Rack controller on the same LXC
- You can install juju and OpenStack client on the same LXC container (may be also deployed on separate containers)
- NAT configuration: (iptables -t nat -A POSTROUTING -o <NIC_NAME> -j MASQUERADE)

juju controller (ref: https://docs.jujucharms.com/2.5/en/)

- will be *bootstrapped* under MAAS supervision as a VM (Best-practice: several VMs on separate servers for HA)

Note juju controller is a key service to deploy and manage any service (i.e. OpenStack)


deployment flow
configuration flow
admin layer
MAAS Region Controller
juju/OpenStack cli
MAAS Rack controller
NAT
MAAS Rack controller
juju controller
Server...
VM...
LXC...
Service (OpenStack, Kubernetes...)
user layer

federation recipe:

https://cloud.garr.it

https://git.garr.it/cloud/federation

what is available till now: quite a lot...

- 3 + 2 regions up and running
- 4 deployments openstack for a total of about **20.000 vcpu** (and counting till 100k - 120k)
- more than **500** users (demo account for 6 months)
- **Virtual Data Centers** delivery in minutes (~500)
- **Physical Data Center** administrative **delegation** (you administer what you own and offload to other regions)
- ***DaaS*** a GARR hack for advanced PaaS or simplified IaaS (via juju with OpenStack cloud backend)
- **Kubernetes** container platform 4 NVIDIA GPUs on bare metal
- **Federated access** (SAML-idem and OIDC-google)

but the important thing is…

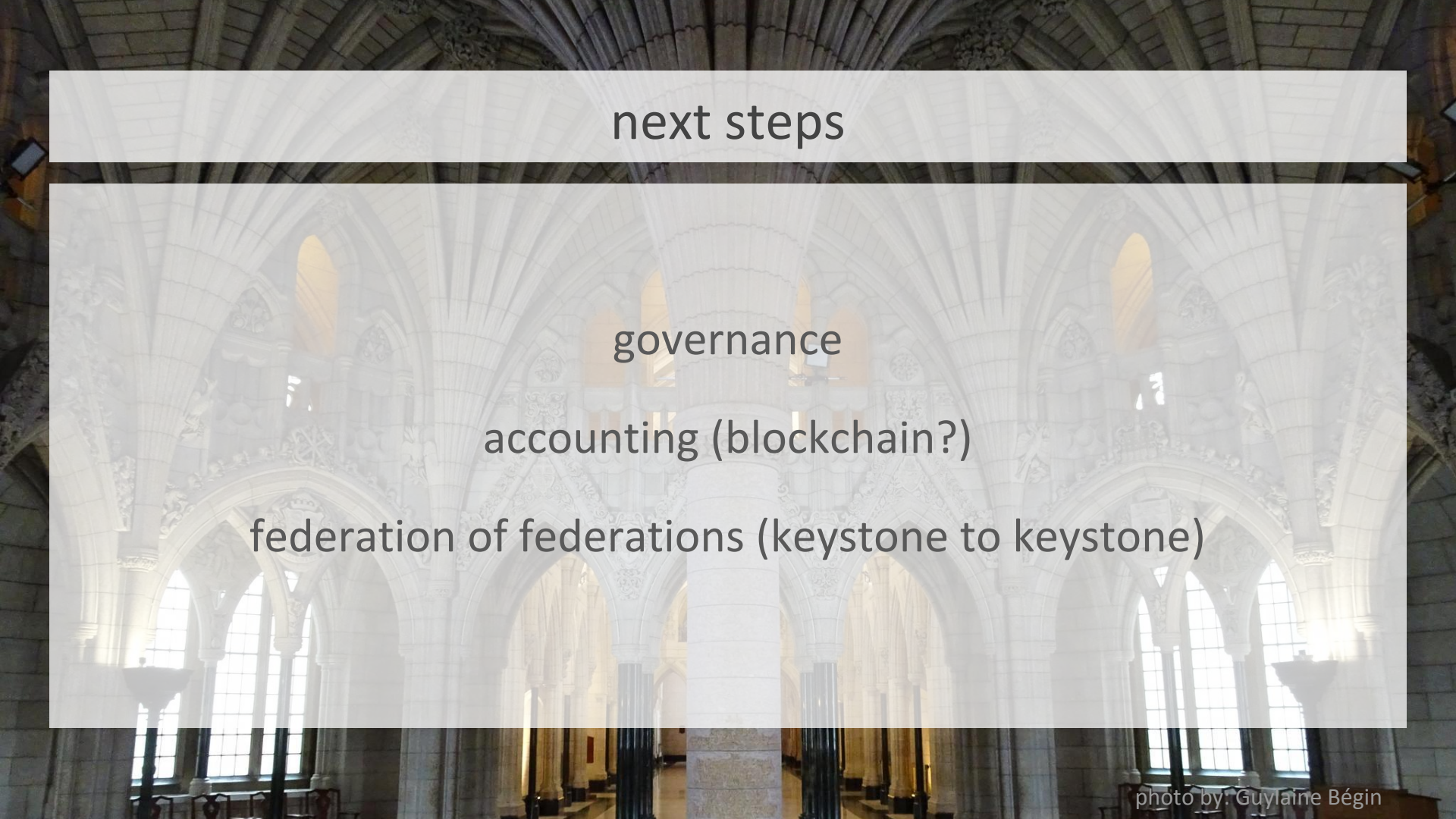
- simple **Federation *recipe*** (git and knowledge base available, references before)
- **Deployment** of OpenStack bare metal (+LXD) region up & running in a **couple of hours**
- **7** federated (**3** federations using the model) regions (ongoing):
  - HPC4AI project,
  - Politecnico To,
  - Uni Padova,
  - 3 INGV region (external federation, see confederation)
  - EAPConnect (East EU Nren, external federation)
  - Hungary
  - RECAS INFN Bari (maintenance)

next steps

governance

accounting (blockchain?)

federation of federations (keystone to keystone)



thanks

interested in a fed-federation working group?
drop us a line: alex.barchiesi@garr.it

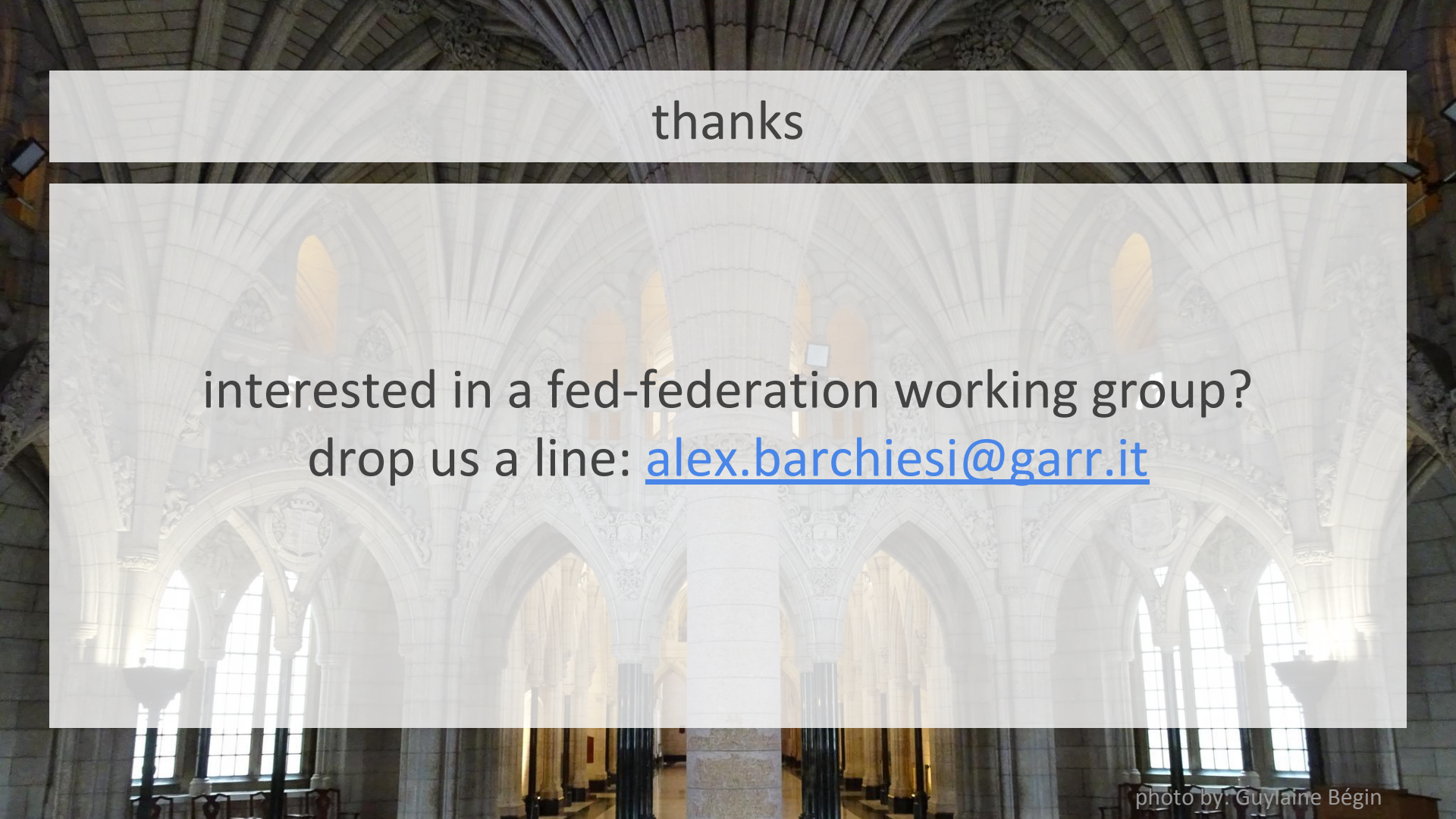


photo by: Guylaine Bégin