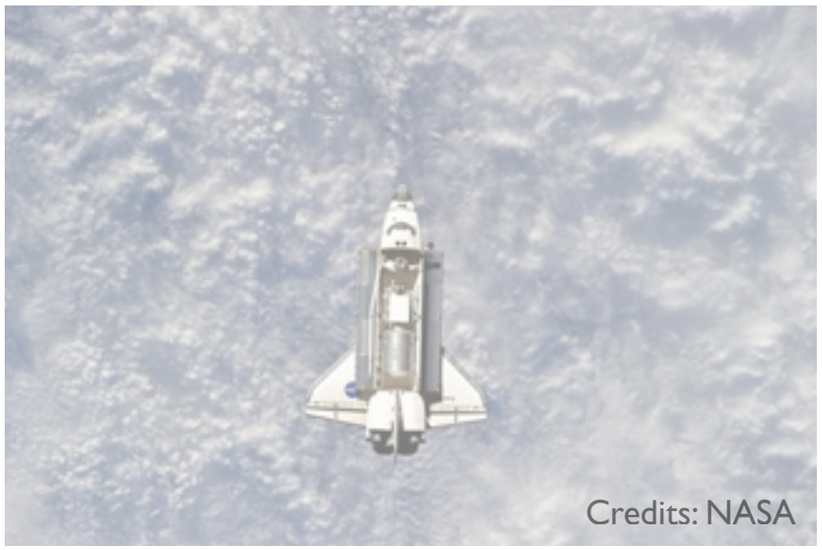
The Discovery Initiative
Beyond the Clouds



Revising OpenStack Internals to Operate Massively Distributed Clouds

Thierry Carrez

OpenStack ‘’Shepherd’’

Adrien Lebre

Discovery PI

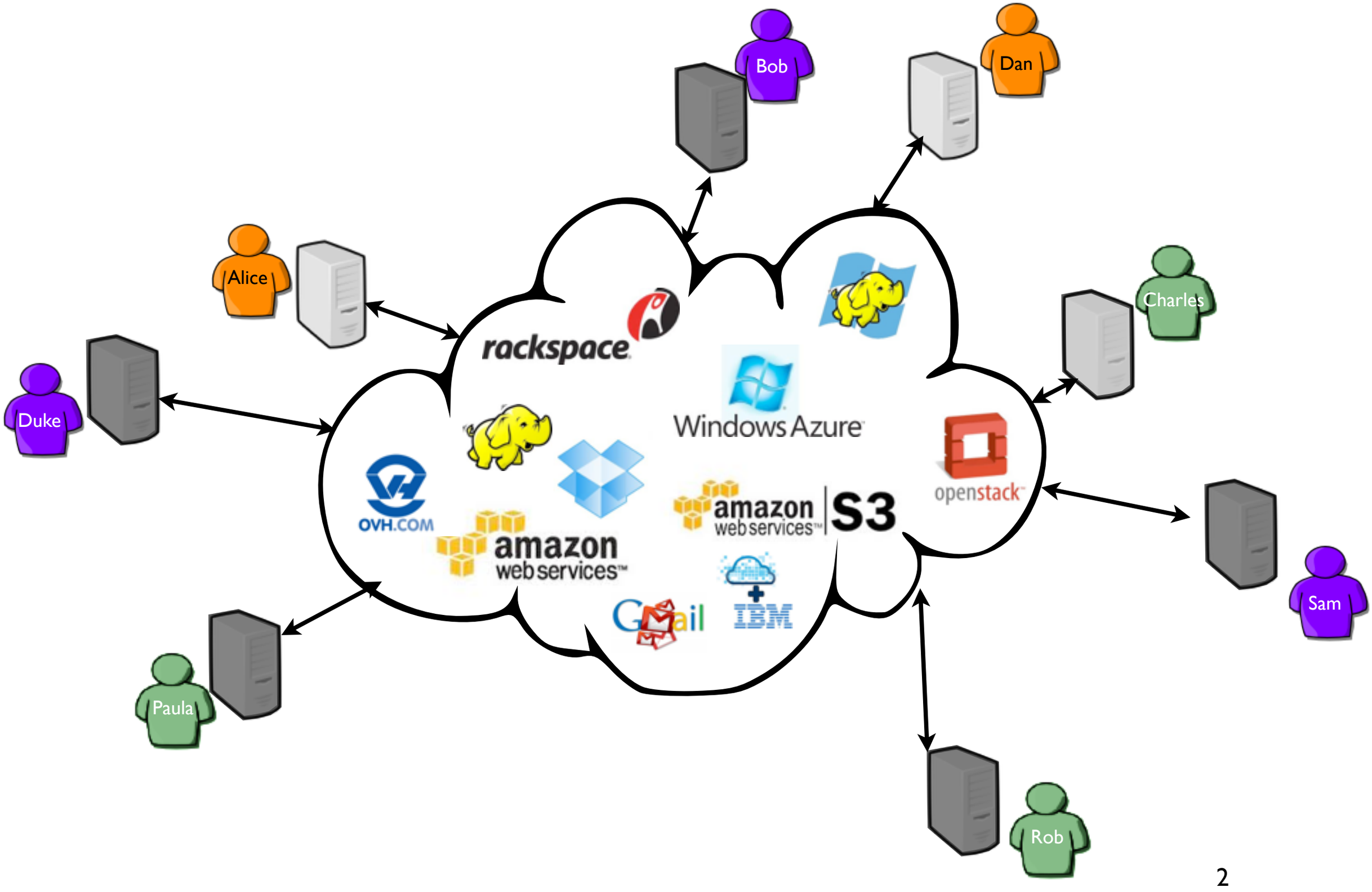
Jonathan Pastor

ROME Project Founder

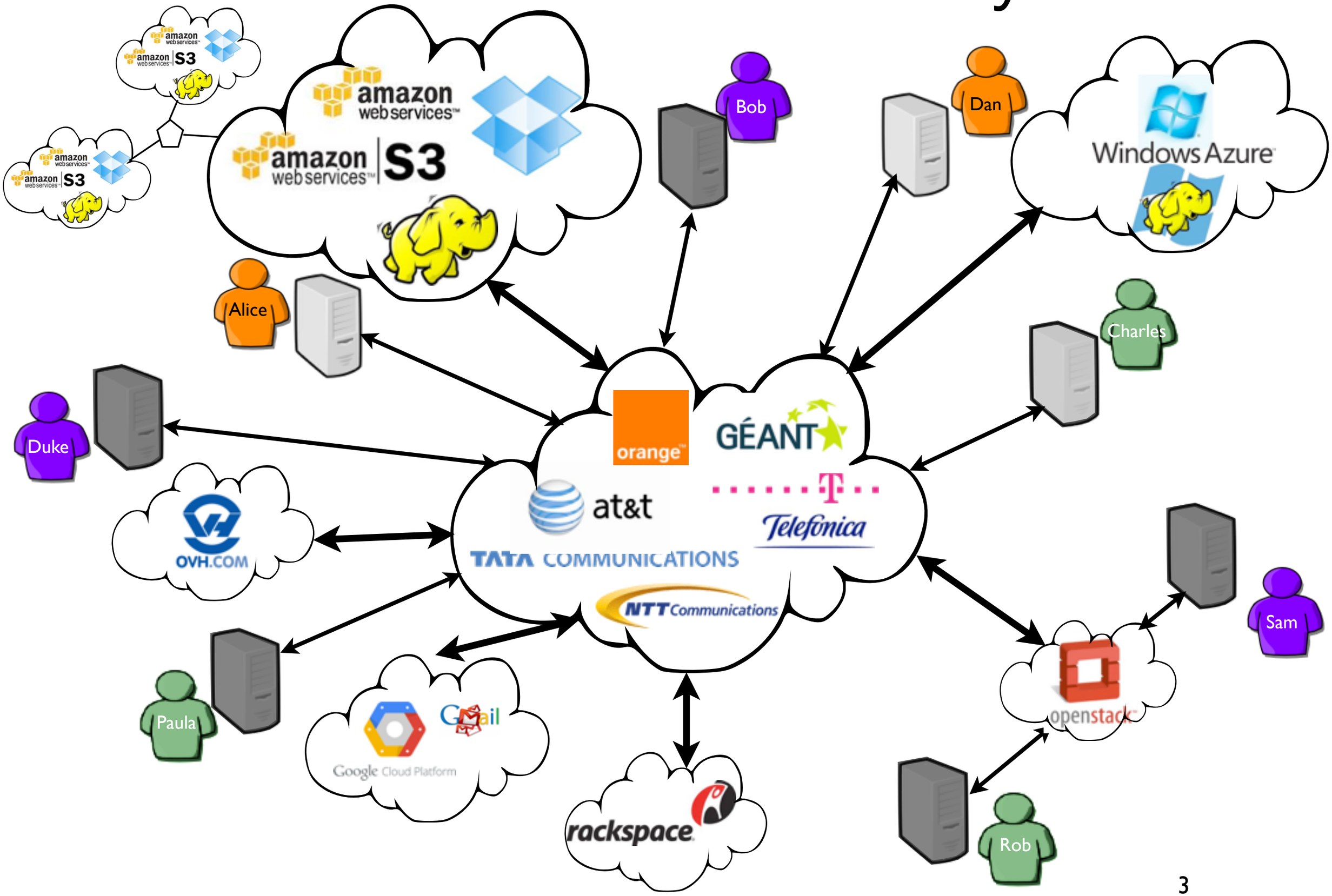
Matthieu Simonin

Discovery TechLead

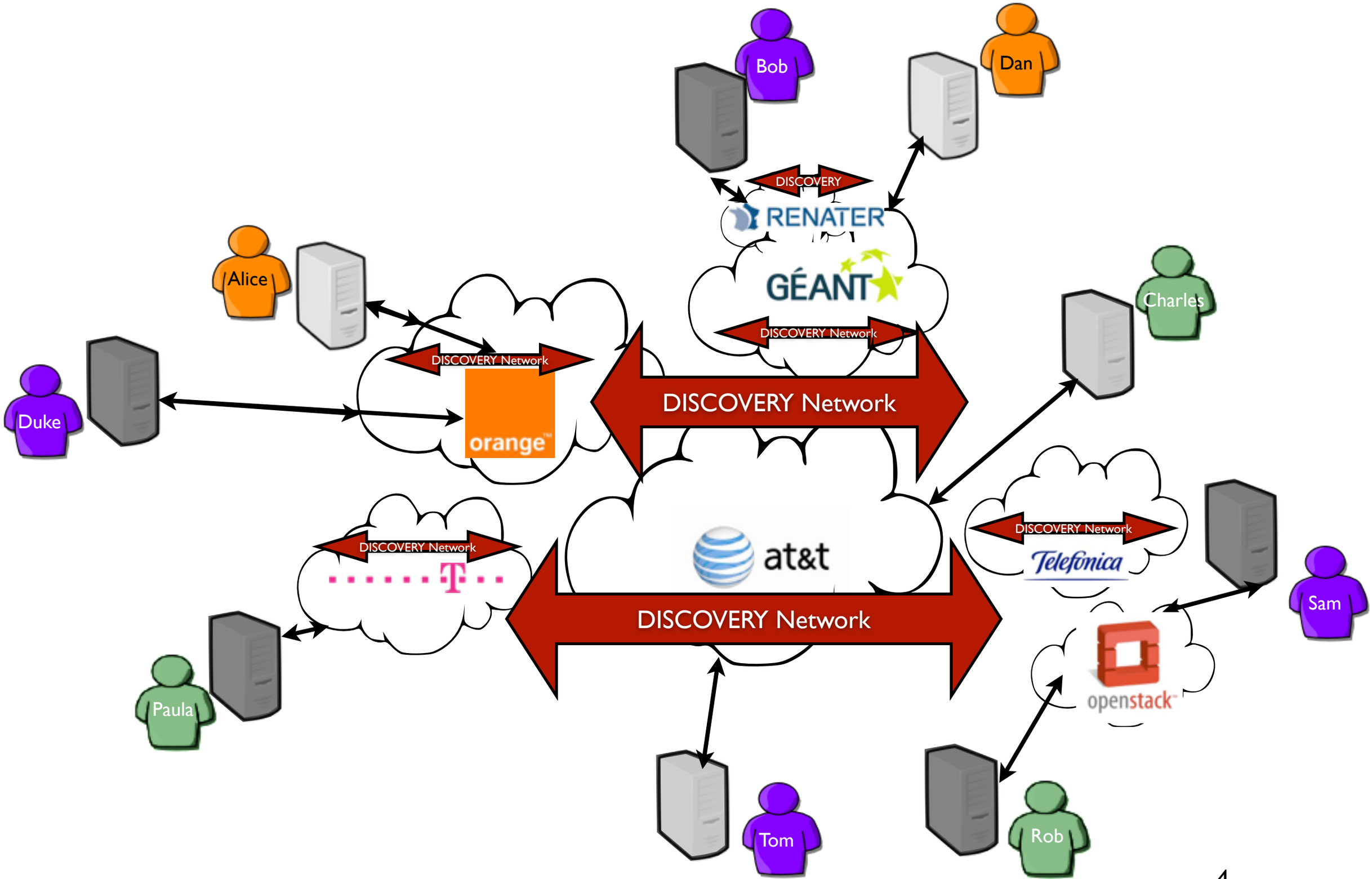
The cloud from end-users



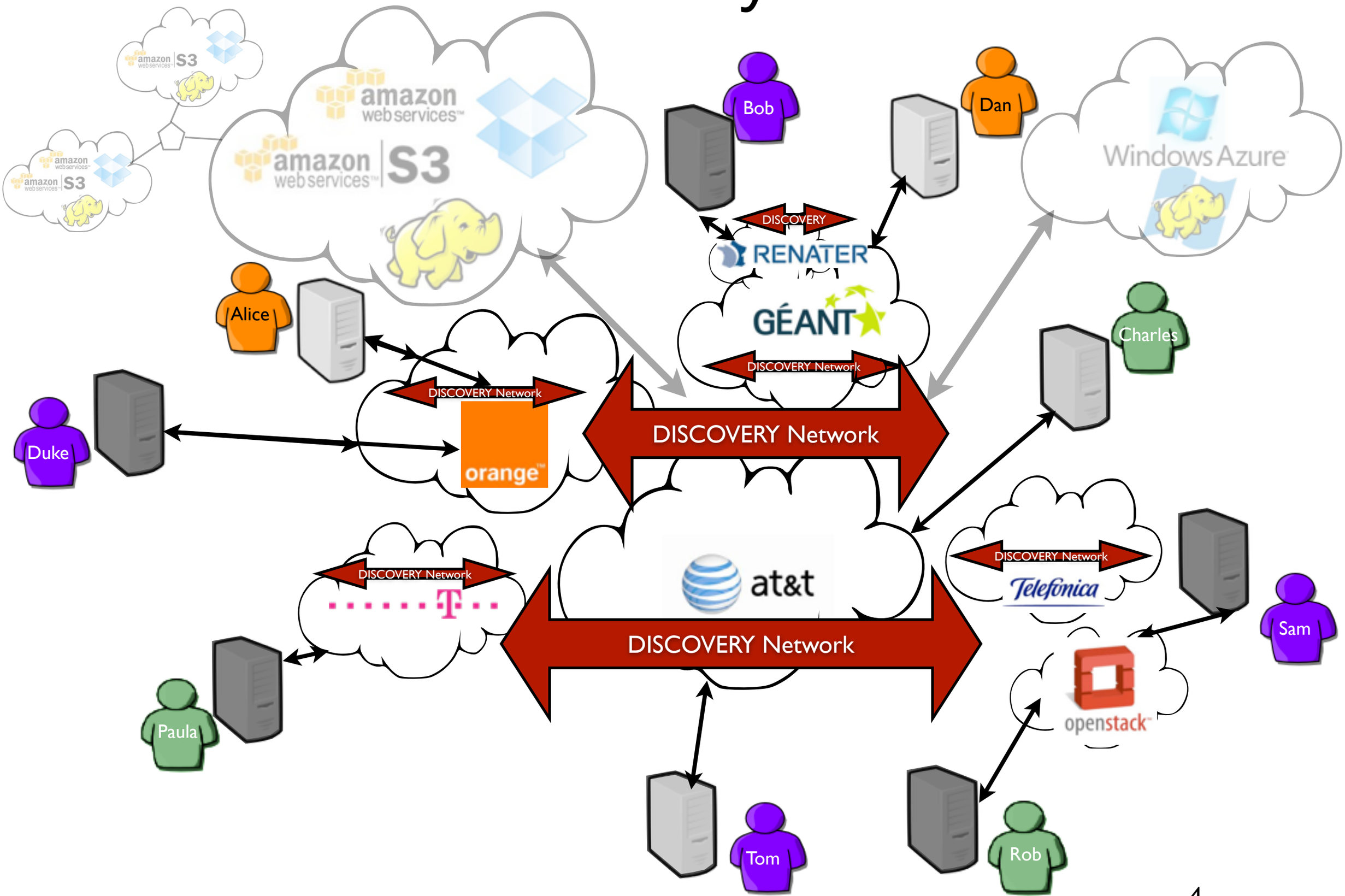
The cloud in reality



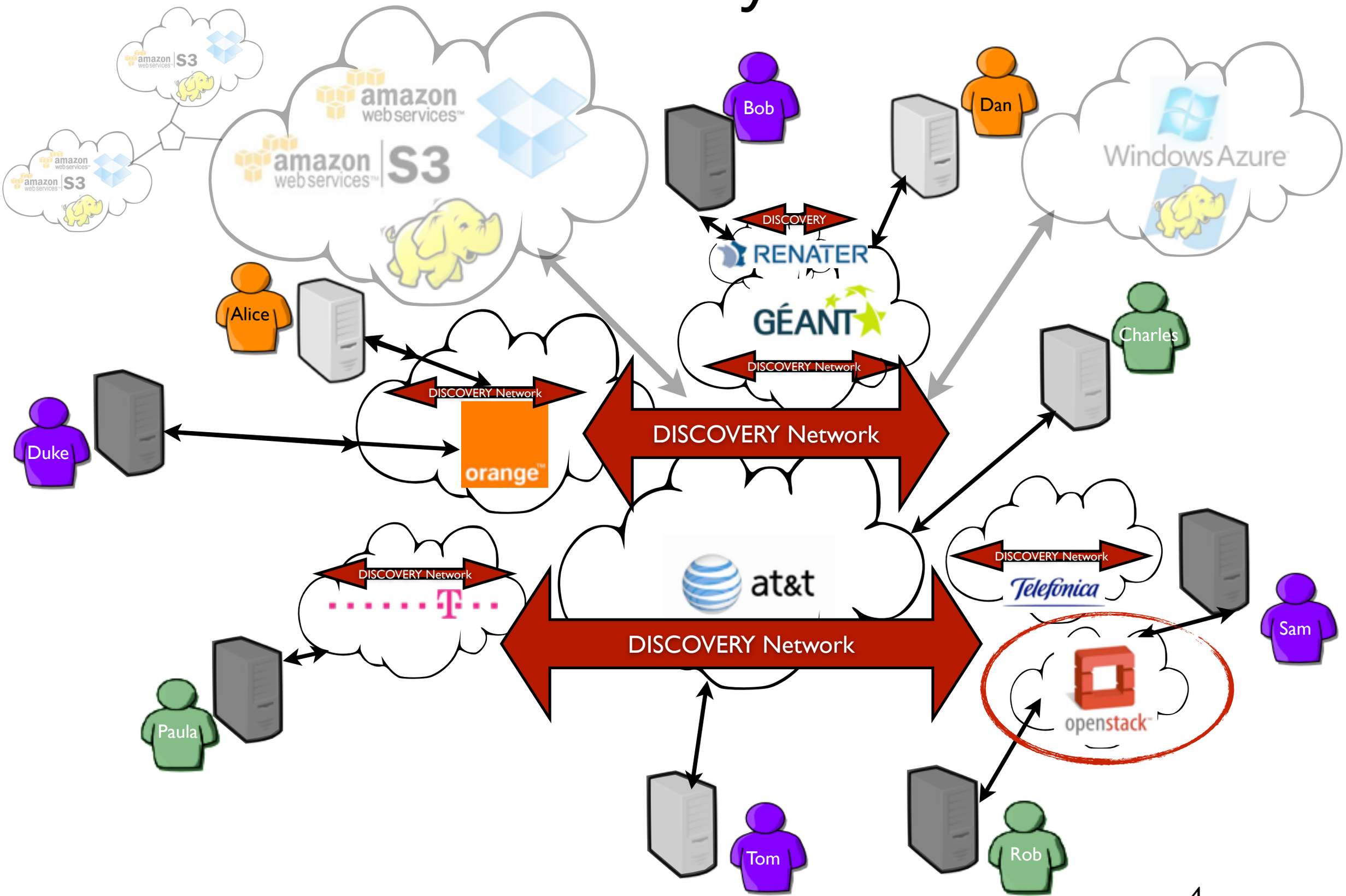
The Discovery Initiative



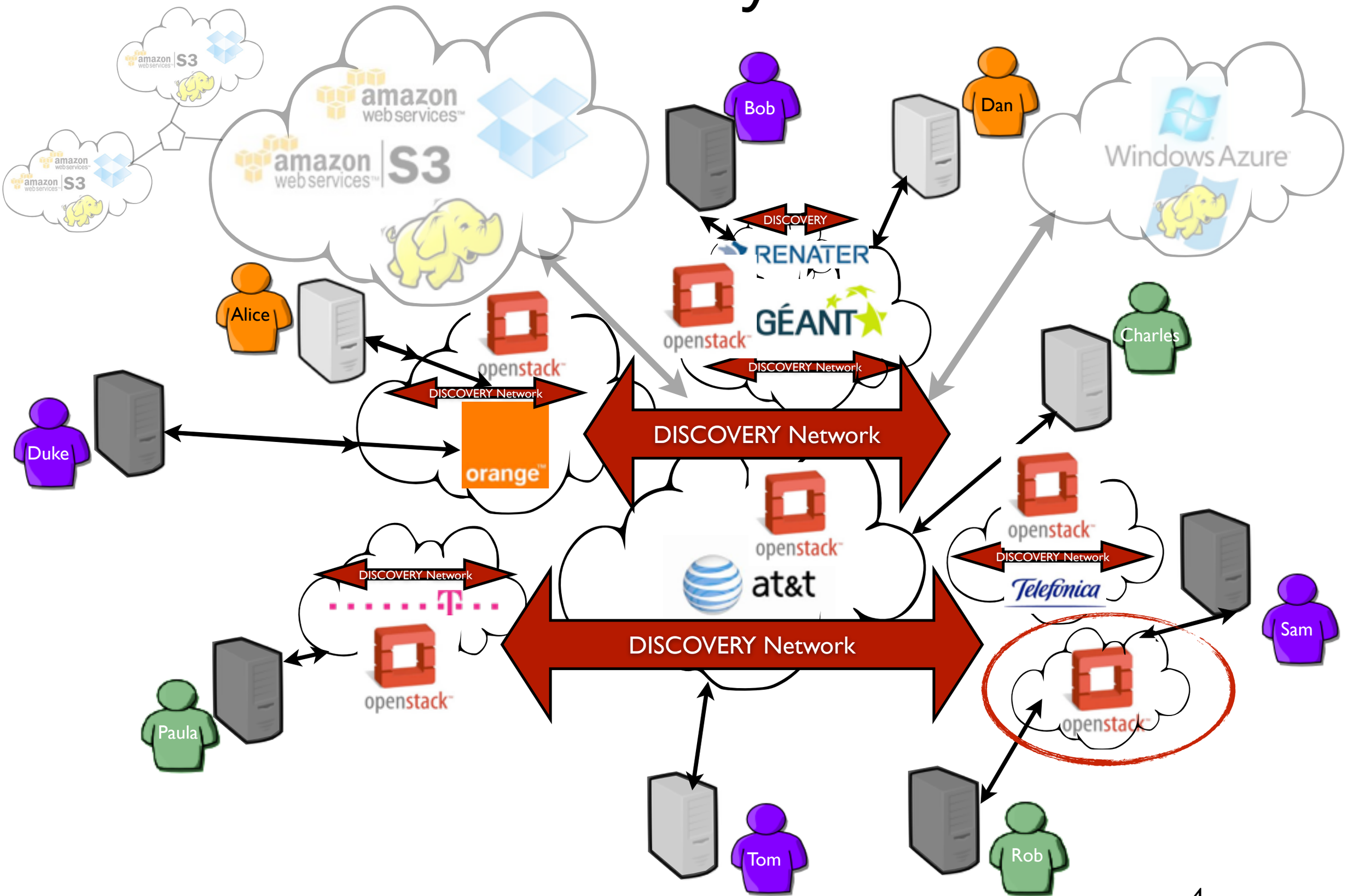
The Discovery Initiative



The Discovery Initiative



The Discovery Initiative



Why ?

Let’s give a look to the current situation

The Current Trend: Large off shore

- To cope with the increasing UC demand while handling energy concerns but…

credits: datacentertalk.com - Microsoft DC, Quincy, WA state

The Current Trend: Large off shore

- To cope with the increasing UC demand while handling energy concerns but…

credits: google map - Quincy

The Current Trend: Large off shore

- To cope with the increasing UC demand while handling energy concerns but…

credits: coloandcloud.com

The Current Trend: Large off shore

credits: coloandcloud.com

Jurisdiction concerns

Reliability

DC distance (network overheads)

Major brakes for the adoption of the CC model

2012 - 2013

Jurisdiction concerns

Reliability

DC distance (network overheads)

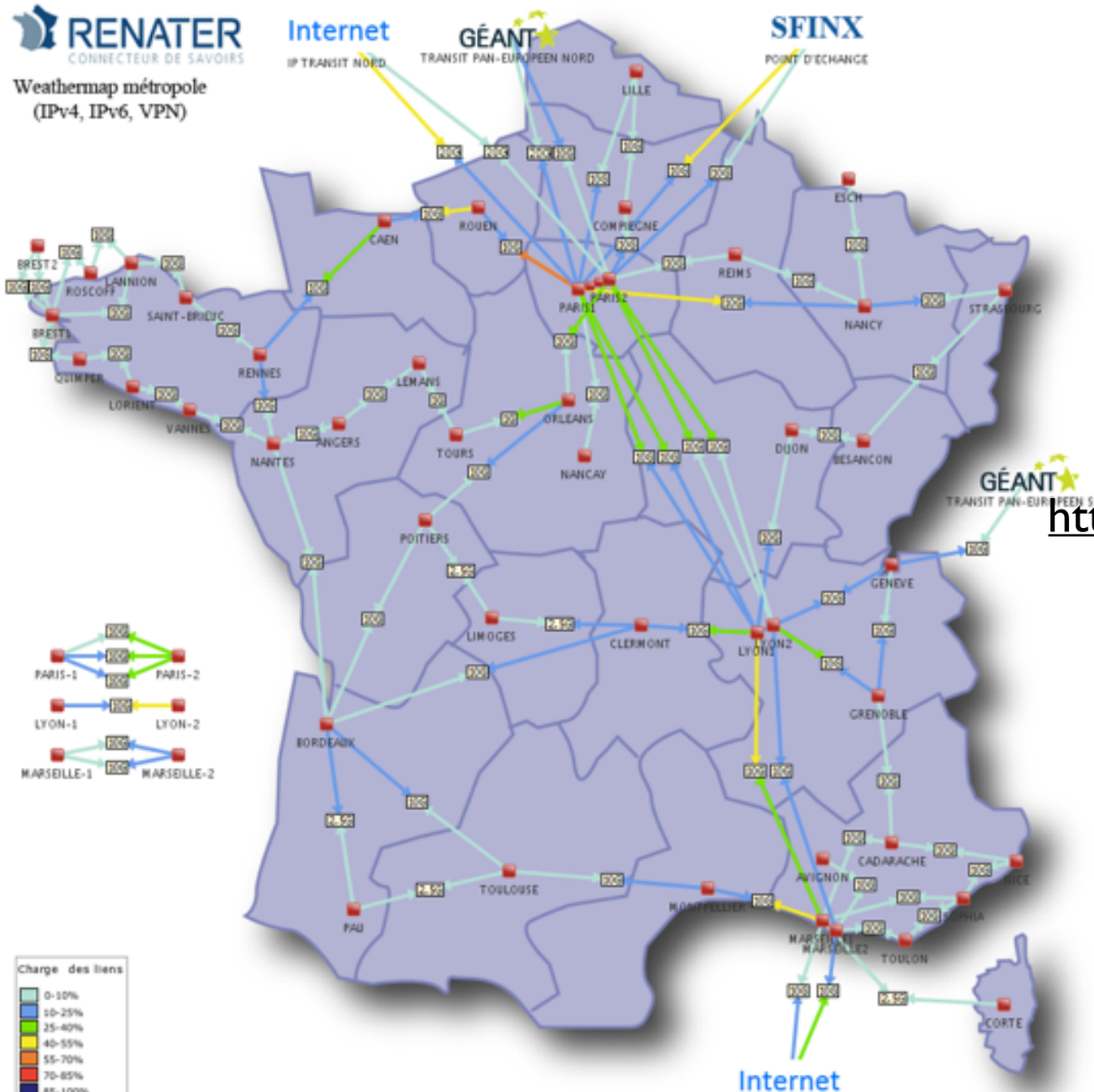
2016 - xxxx

Big Data - Internet of Things/Everything

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based Cloud infrastructures / Fog / Edge

  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.

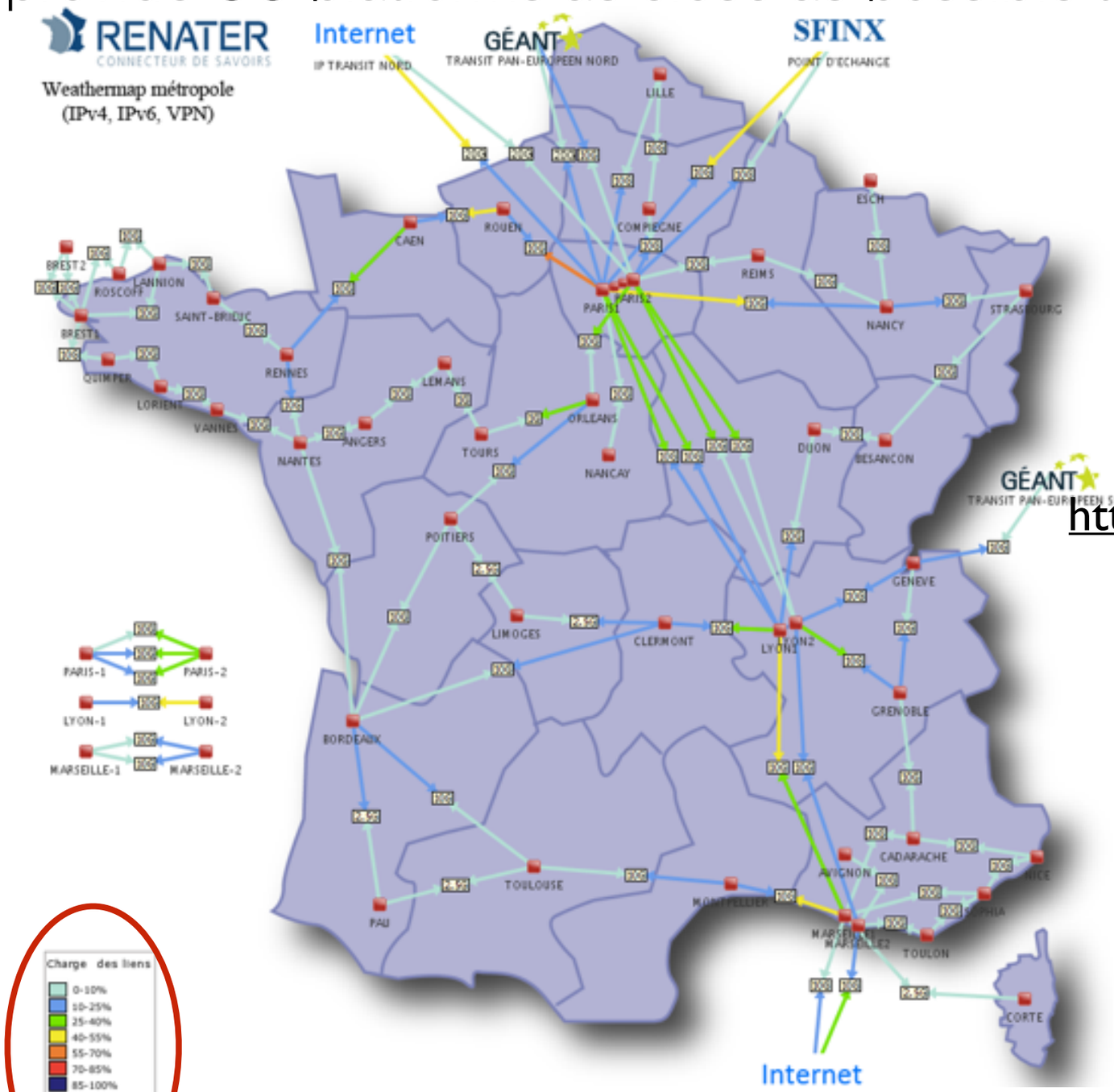


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based Cloud infrastructures / Fog / Edge

  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.

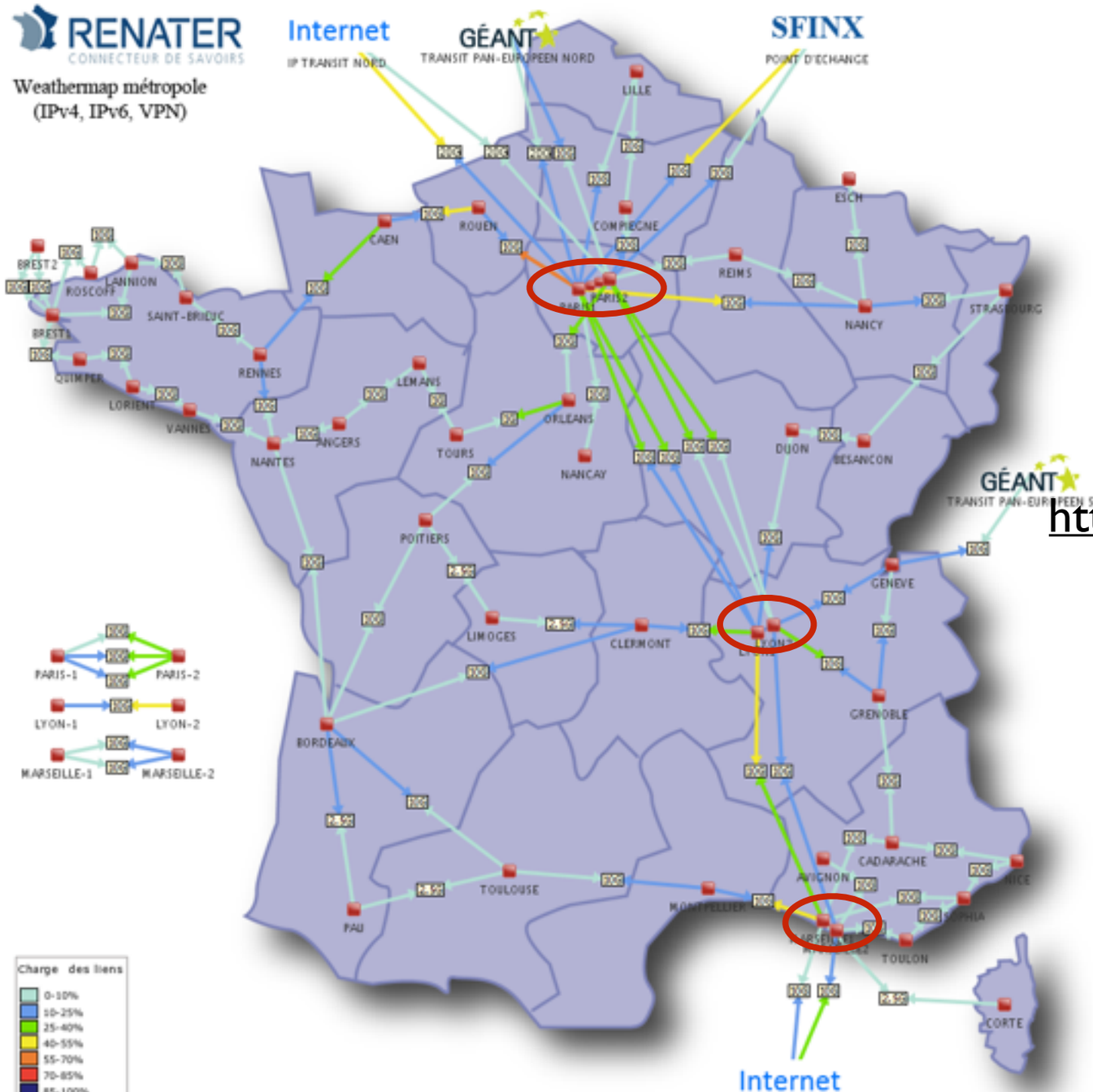


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based Cloud infrastructures / Fog / Edge

  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.

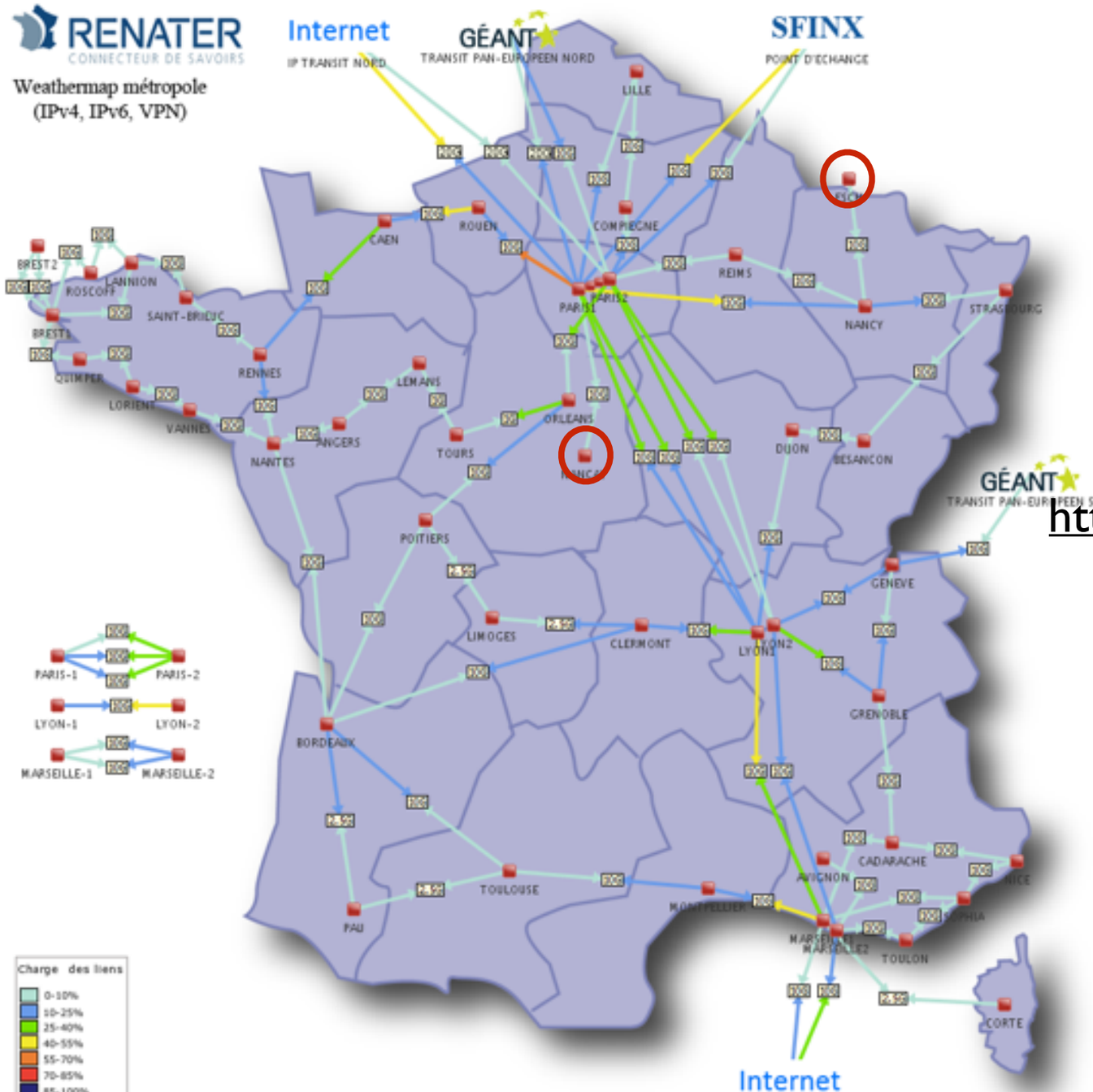


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based Cloud infrastructures / Fog / Edge

  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.

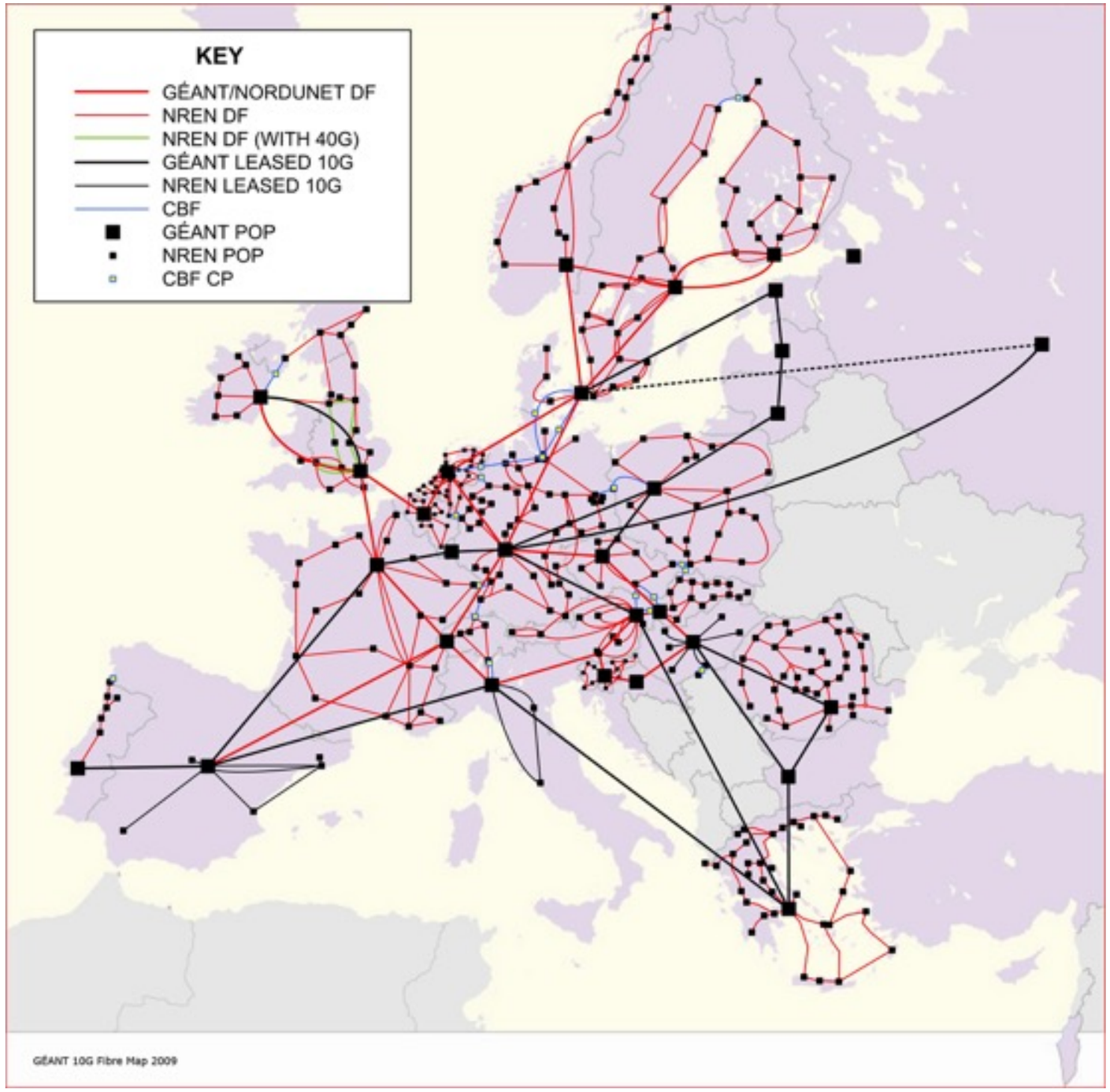


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based Cloud infrastructures / Fog / Edge

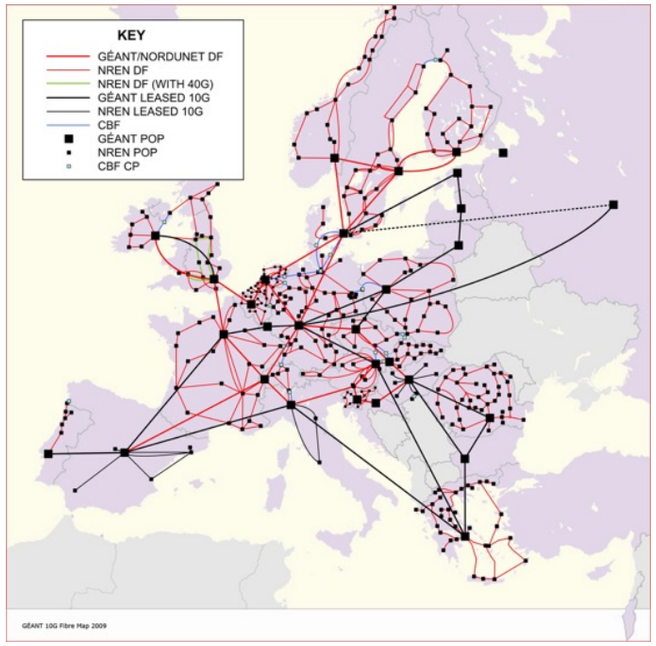
  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.



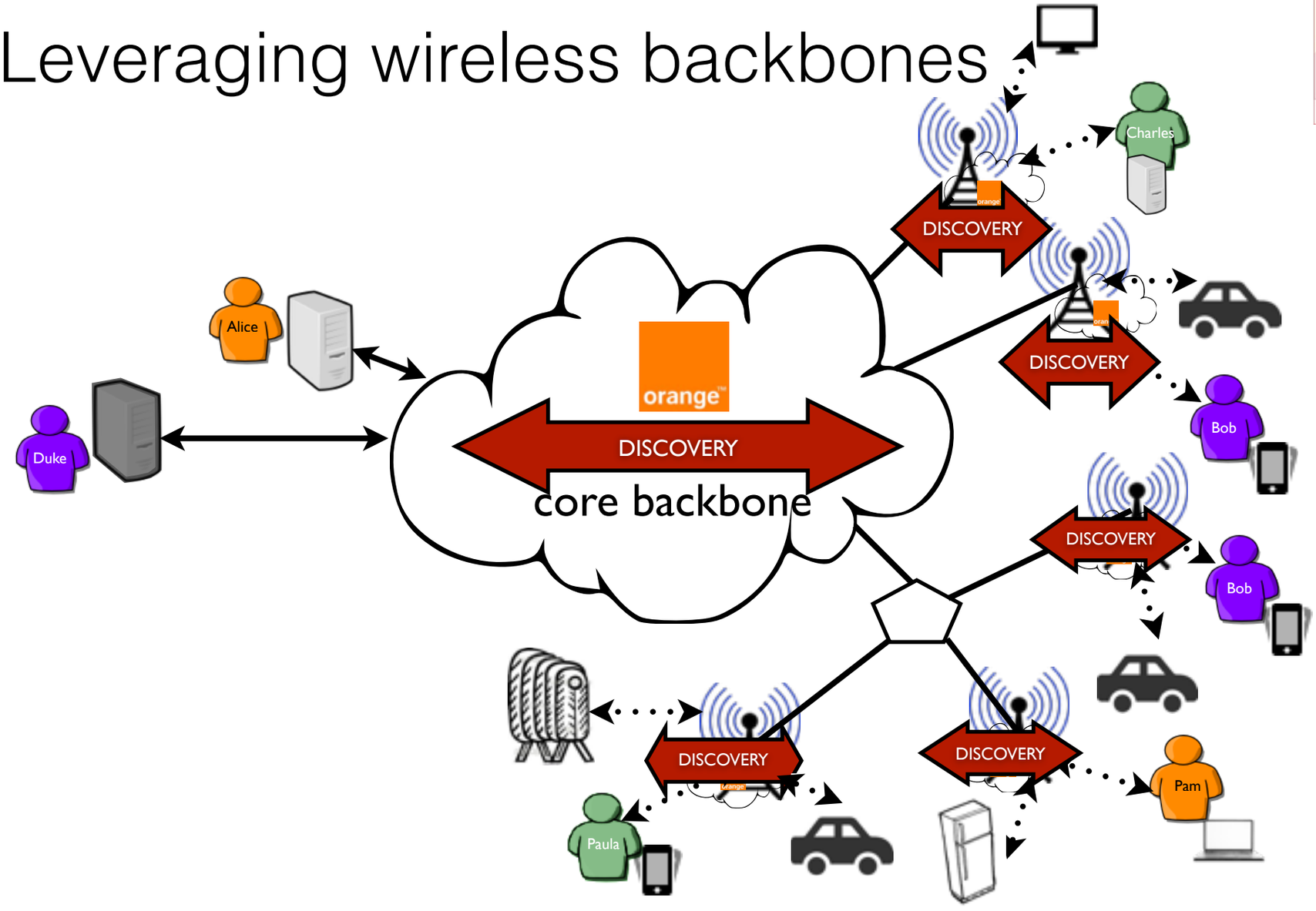
Beyond the Cloud, the DISCOVERY Initiative

- Leveraging network backbones

  Extend any point of presence of network backbones with UC servers (from network hubs up to major DSLAMs that are operated by telecom companies and network institutions).



- Leveraging wireless backbones

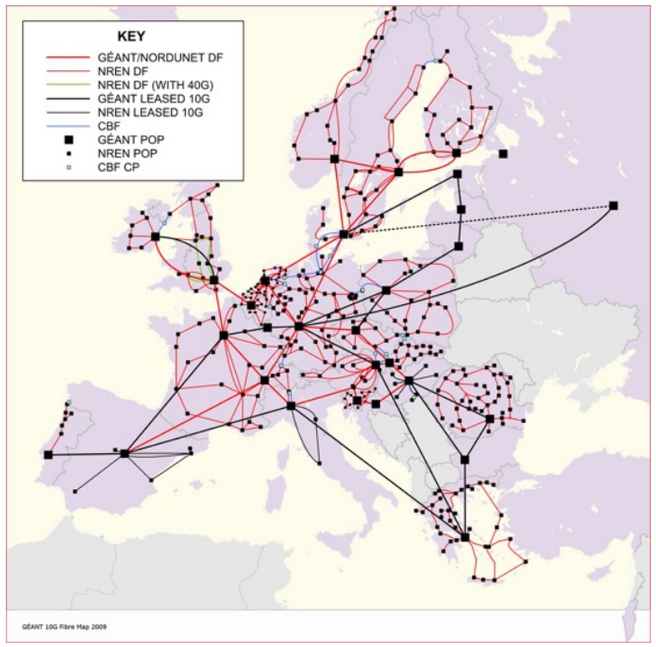


Beyond the Cloud, the DISCOVERY Initiative

- Leveraging network backbones

  Extend any point of presence of network backbones with UC servers (from network hubs up to major DSLAMs that are operated by telecom companies and network institutions).

- Leveraging wireless backbones



Beyond the Cloud, the DISCOVERY Initiative

*Localized or **micro data centers are a fact of life, but by applying a self-contained, scalable and remotely managed solution and process, CIOs can reduce costs, improve agility, and introduce new levels of compliance and service continuity**. Creating micro data centers is something companies have done for years, but often in an ad hoc manner.*

Gartner 2015

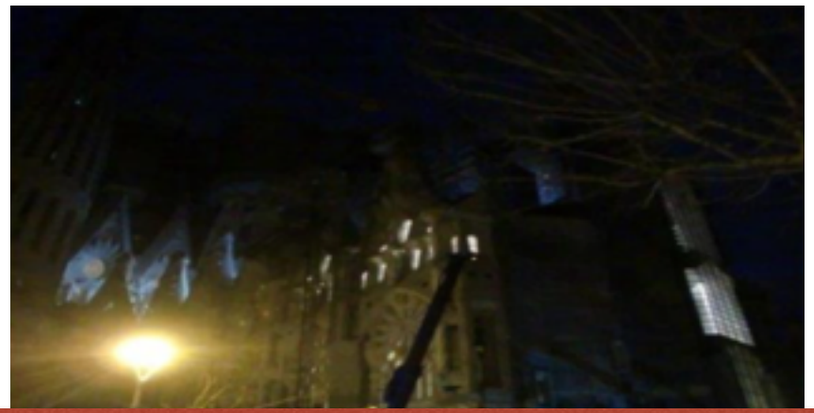
Sagrada Familia microDC (Barcelona, Spain)

Deployment of a new PoP of the Orange French backbone

Beyond the Cloud, the DISCOVERY Initiative

*Localized or **micro data centers are a fact of life, but by applying a self-contained, scalable and remotely managed solution and process, CIOs can reduce costs, improve agility, and introduce new levels of compliance and service continuity**. Creating micro data centers is something companies have done for years, but often in an ad hoc manner.*

Gartner 2015



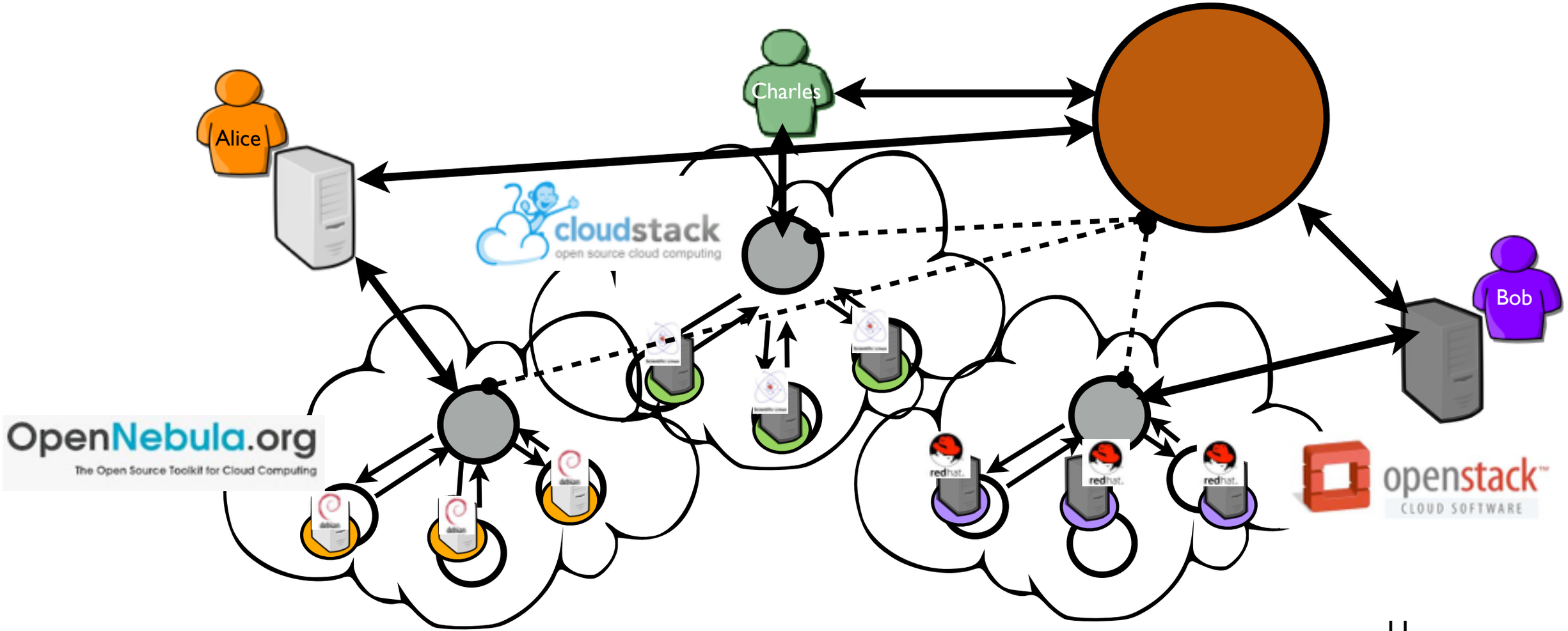
Development of a fully distributed system in charge of operating such a massively distributed infrastructure

Sagrada Familia microDC (Barcelona, Spain)

Deployment of a new PoP of the Orange French backbone

Why not a broker ?

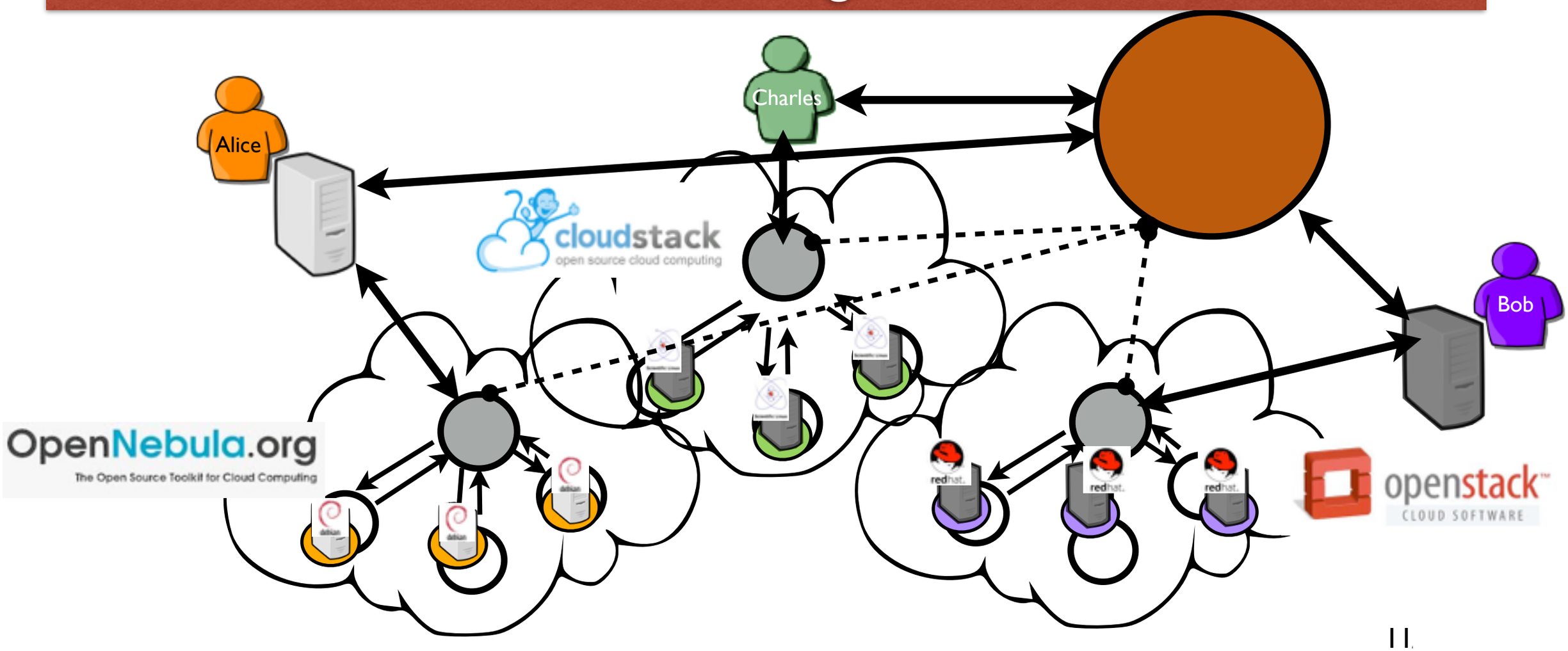
- Sporadic (hybrid computing/cloud bursting) almost ready for production
- While standards are coming (OCCI, OVF, ….), current brokers are rather limited



Why not a broker ?

- Sporadic (hybrid computing/cloud bursting) almost ready for production
- While standards are coming (OCCI, OVF, ….), current brokers are rather limited

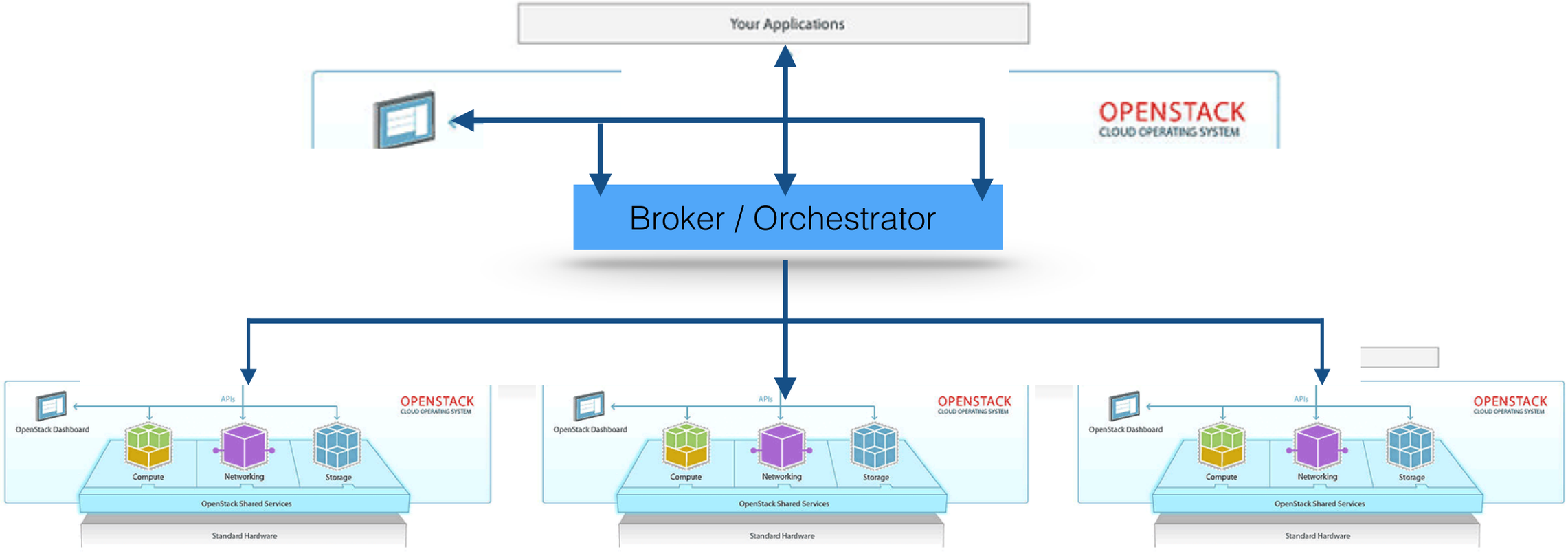
Advanced brokers must reimplement standard IaaS mechanism while facing the API limitation



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

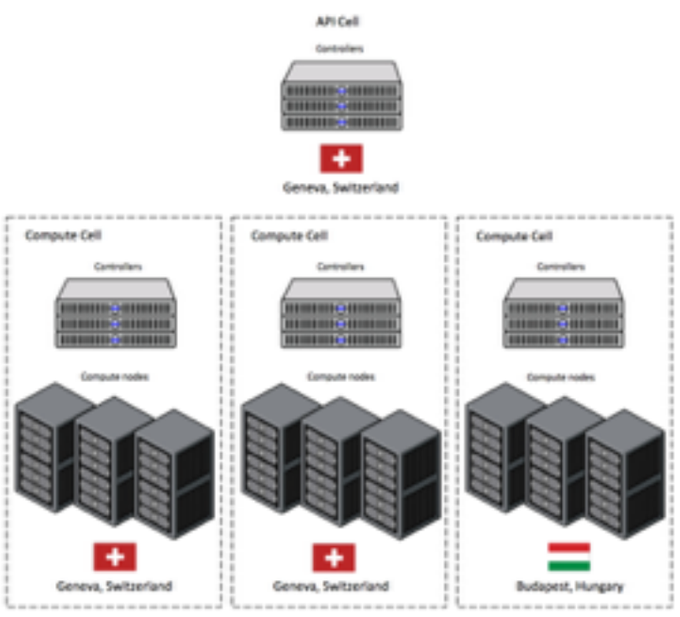
Top/Down: add a substrate to pilot independent OpenStack instances



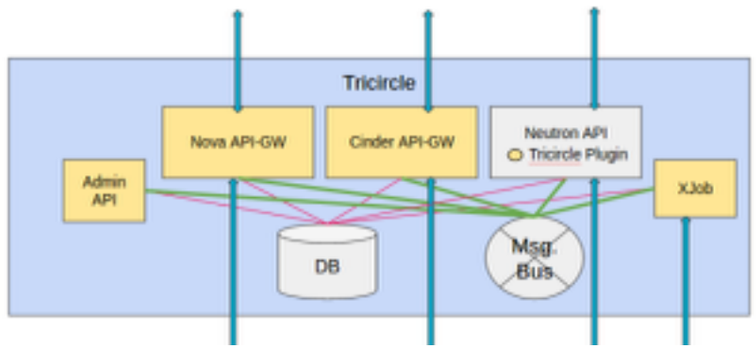
Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

Top/Down: add a substrate to pilot independent OpenStack instances



Cell v2 (Nova)

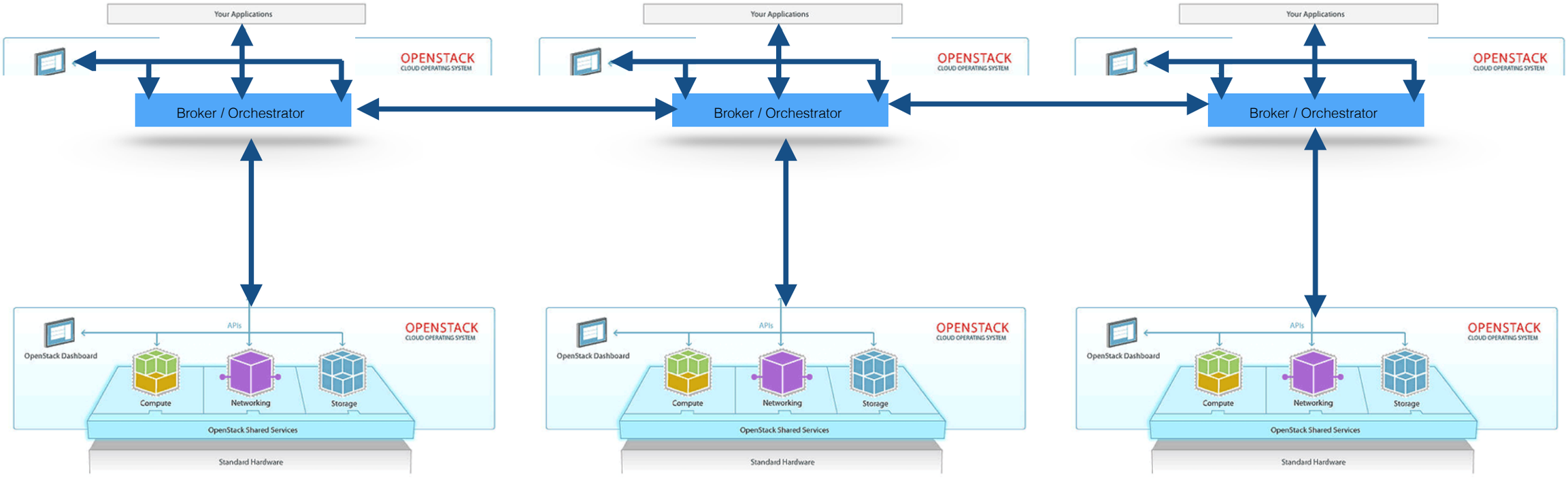


Tricircle (previously cascading OpenStack)

Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

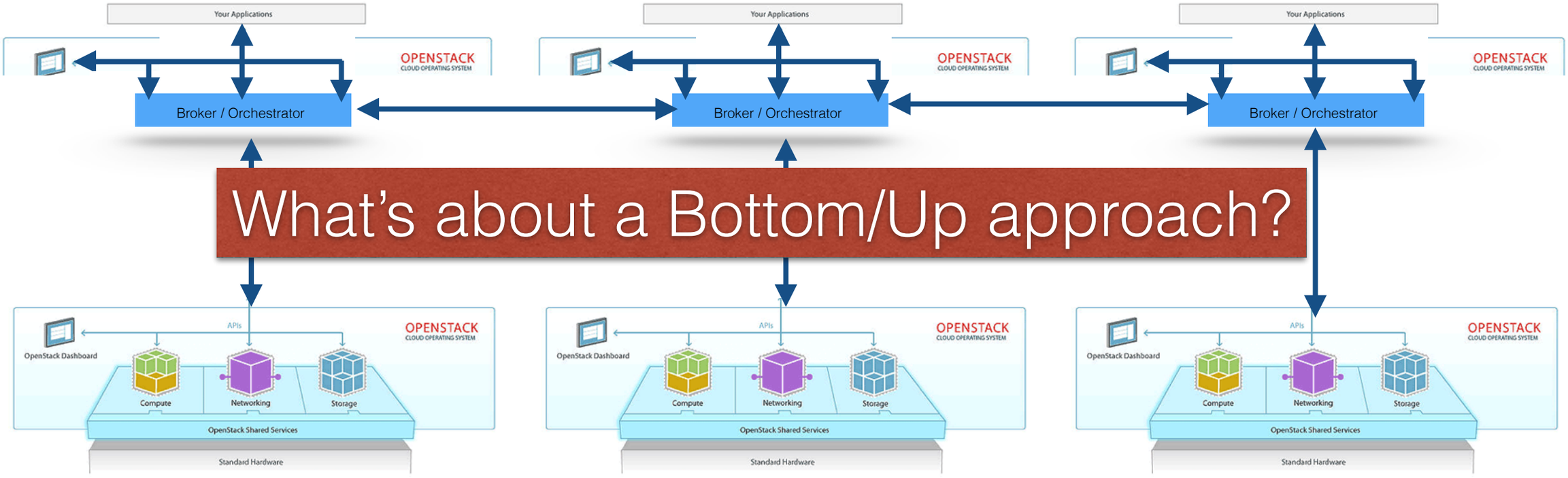
Top/Down: add a substrate to pilot independent OpenStack instances



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

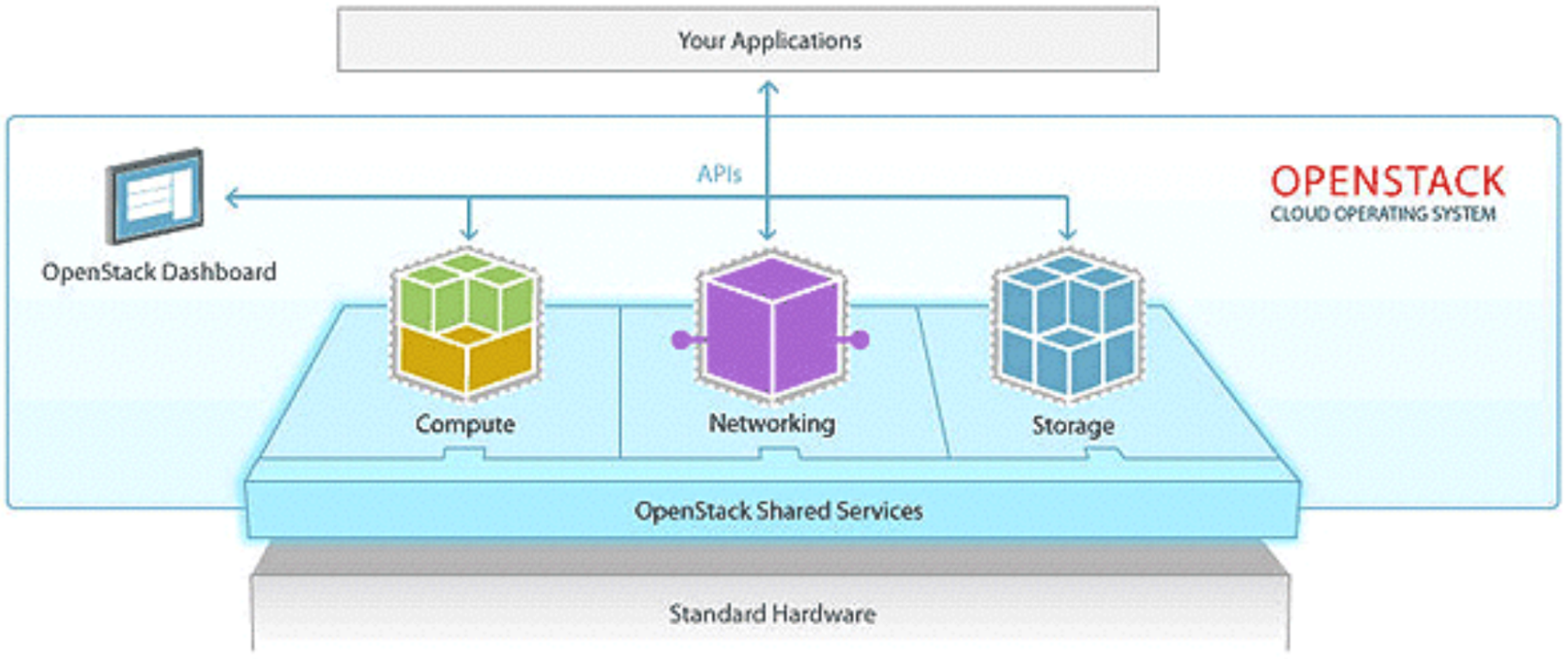
Top/Down: add a substrate to pilot independent OpenStack instances



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

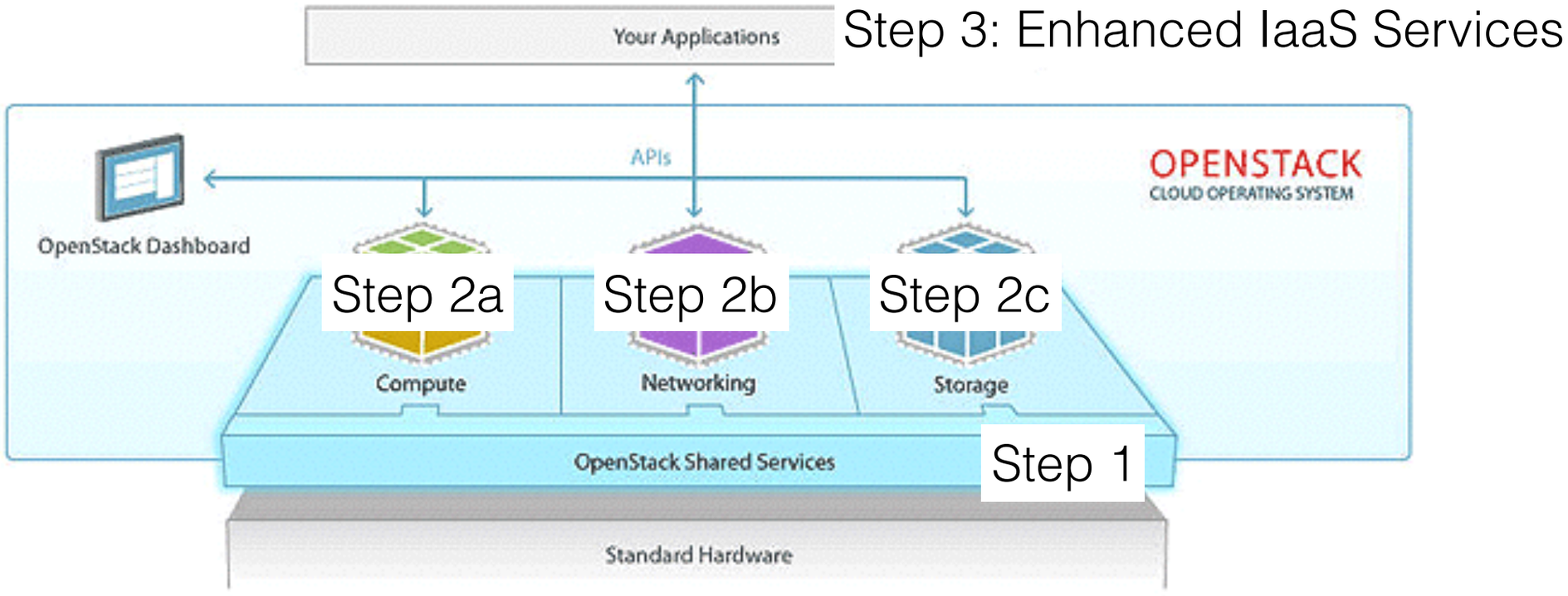
  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default

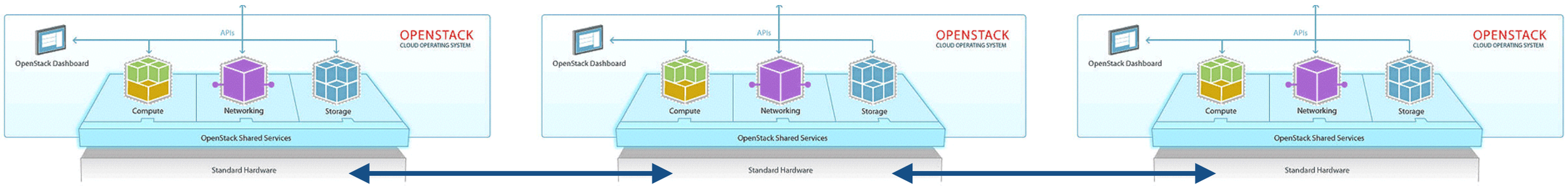


Step 3: Enhanced IaaS Services

Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default

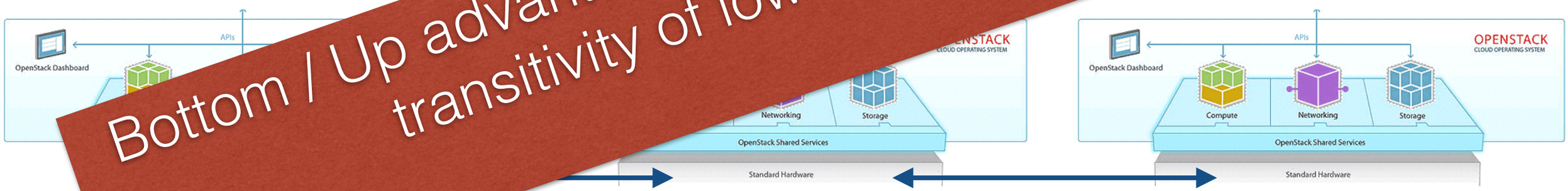


Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default

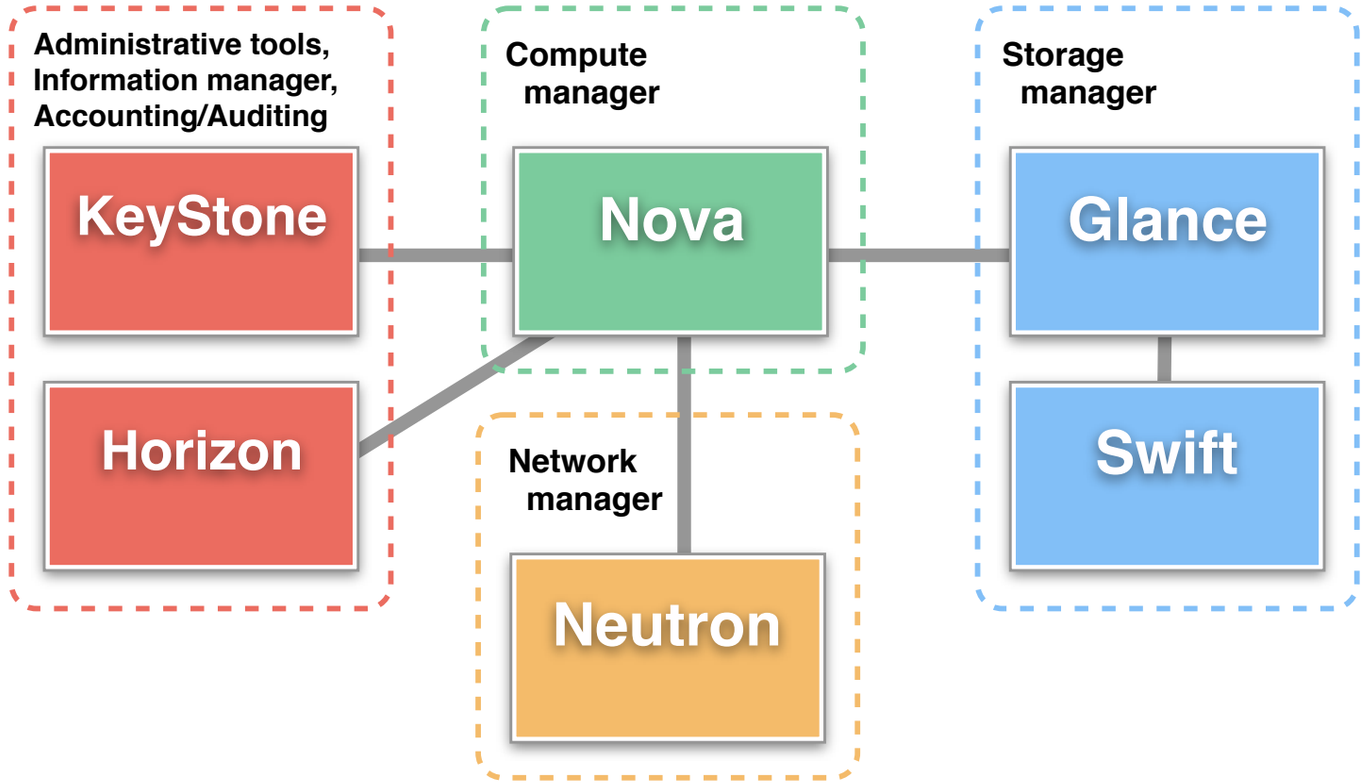
Bottom / Up advantage: high services can benefit by transitivity of low level changes.



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default



Would OpenStack be the solution?

- Do not reinvent the wheel… it is too late
- Few proposals to federate/operate distinct OpenStack DCs

  Bottom/Up - investigate whether/how OpenStack core services can be cooperative by default

openstack CLOUD SOFTWARE Technical considerations

SIDEBAR ◂ PREV | UP | NEXT ▸

CONTENTS SEARCH

OPENSTACK MANUALS ▸ OPENSTACK ARCHITECTURE DESIGN GUIDE - CURRENT

Technical considerations

Repurposing an existing OpenStack environment to be massively scalable is a formidable task. When building a massively scalable environment from the ground up, ensure you build the initial deployment with the same principles and choices that apply as the environment grows. For example, a good approach is to deploy the first site as a multi-site environment. This enables you to use the same deployment and segregation methods as the environment grows to separate locations across dedicated links or wide area networks. In a hyperscale cloud, scale trumps redundancy. Modify applications with this in mind, relying on the scale and homogeneity of the environment to provide reliability rather than redundant infrastructure provided by non-commodity hardware solutions.

Infrastructure segregation

OpenStack services support massive horizontal scale. Be aware that this is not the case for the entire supporting infrastructure. This is particularly a problem for the database management systems and message queues that OpenStack services use for data storage and remote procedure call communications.

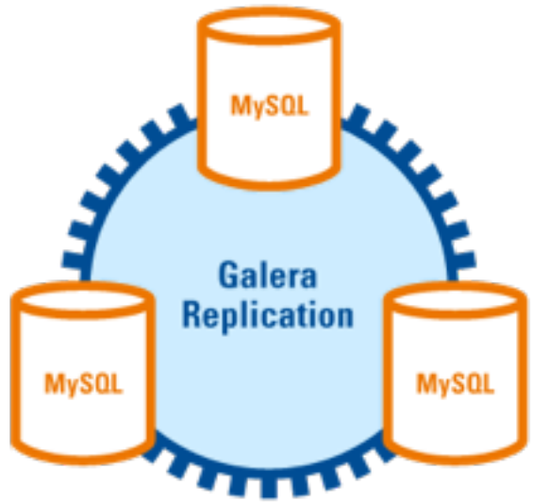
Traditional clustering techniques typically provide high availability and some additional scale for these environments. In the quest for massive scale, however, you must take additional steps to relieve the performance pressure on these components in order to prevent them from negatively impacting the overall performance of the environment. Ensure that all the components are in balance so that if the massively scalable environment fails, all the components are near maximum capacity and a single component is not causing the failure.

Distributing OpenStack Through a Bottom/Up Approach

- Step 1: OpenStack shared services

  A SQL database 

  A messaging queue 



- Active/Active replication
  Well-tested but does not scale to our target.

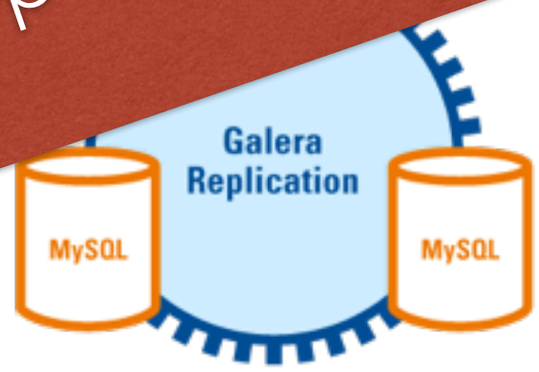
- Key/Value Store systems
  Alternate solutions for storing states over a highly distributed infrastructure

Distributing OpenStack Through a Bottom/Up Approach

- Step 1: OpenStack shared services

  A SQL database 

  A messaging queue

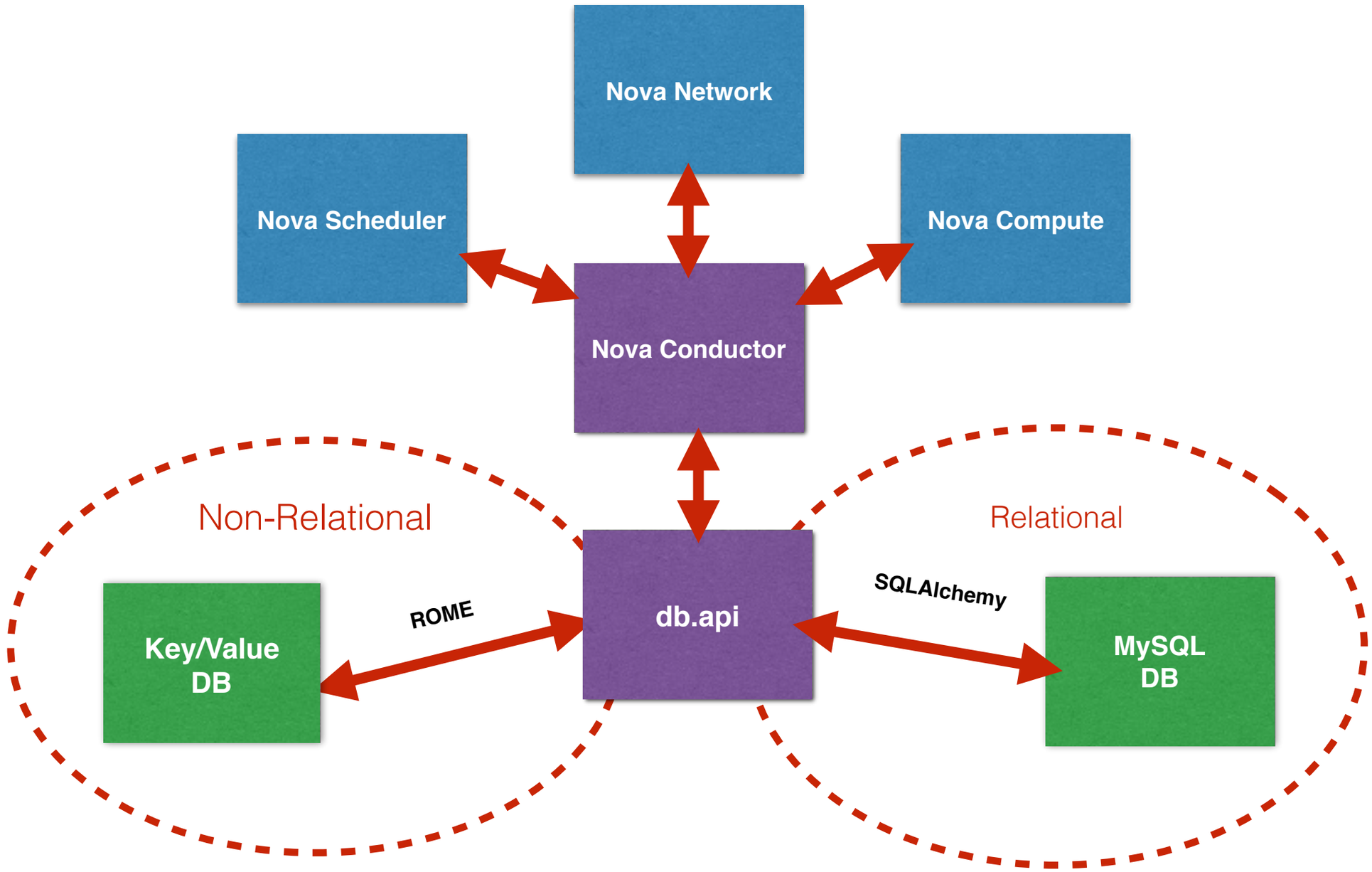


How can we switch between a SQL solution and a NoSQL system for storing inner states of OpenStack ?

- Acti… not scale to our target.
- …ide Store systems

  Alternate solutions for storing states over a highly distributed infrastructure

Leveraging a key/value store DB

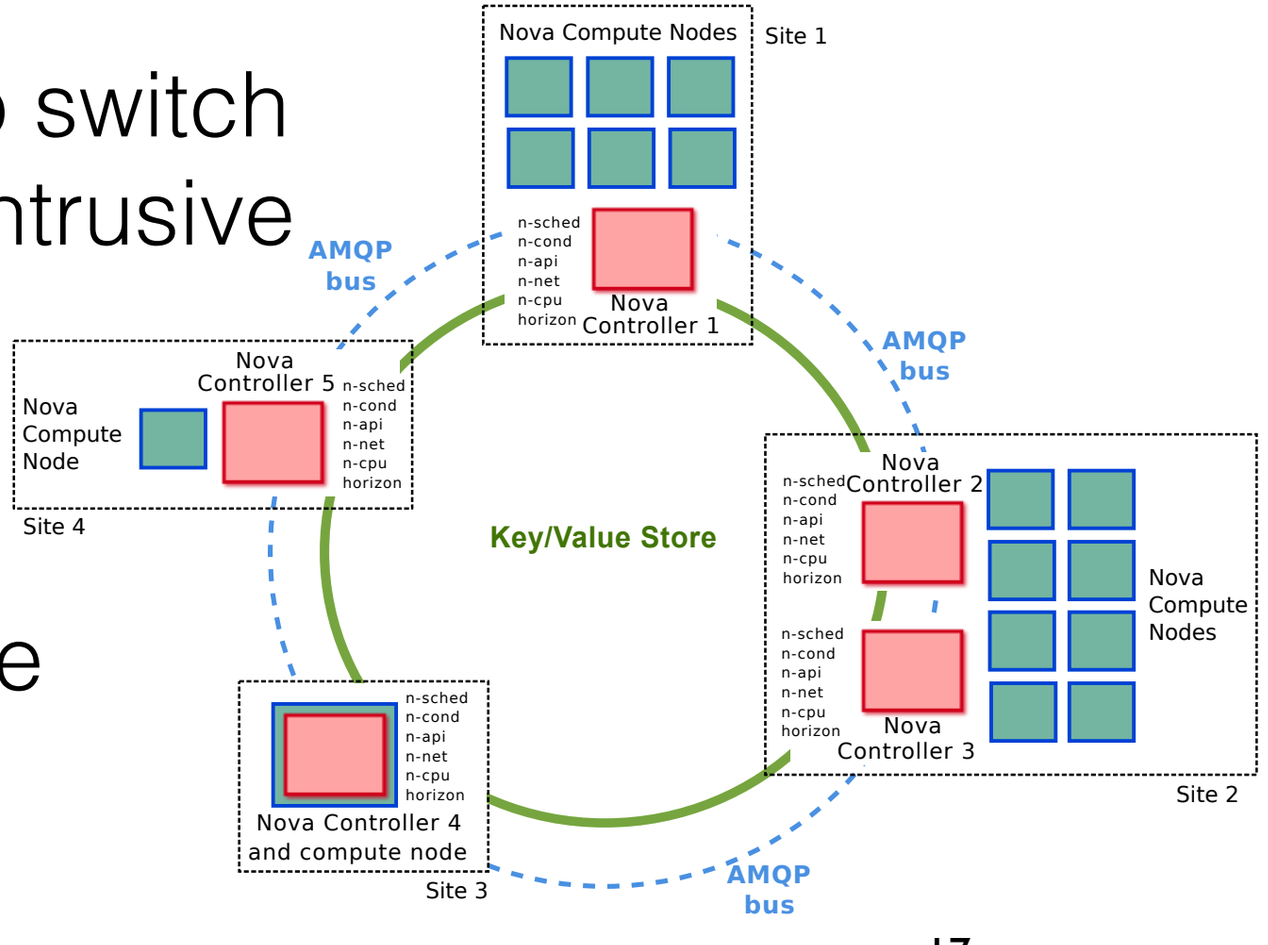


Nova (compute service) - software architecture

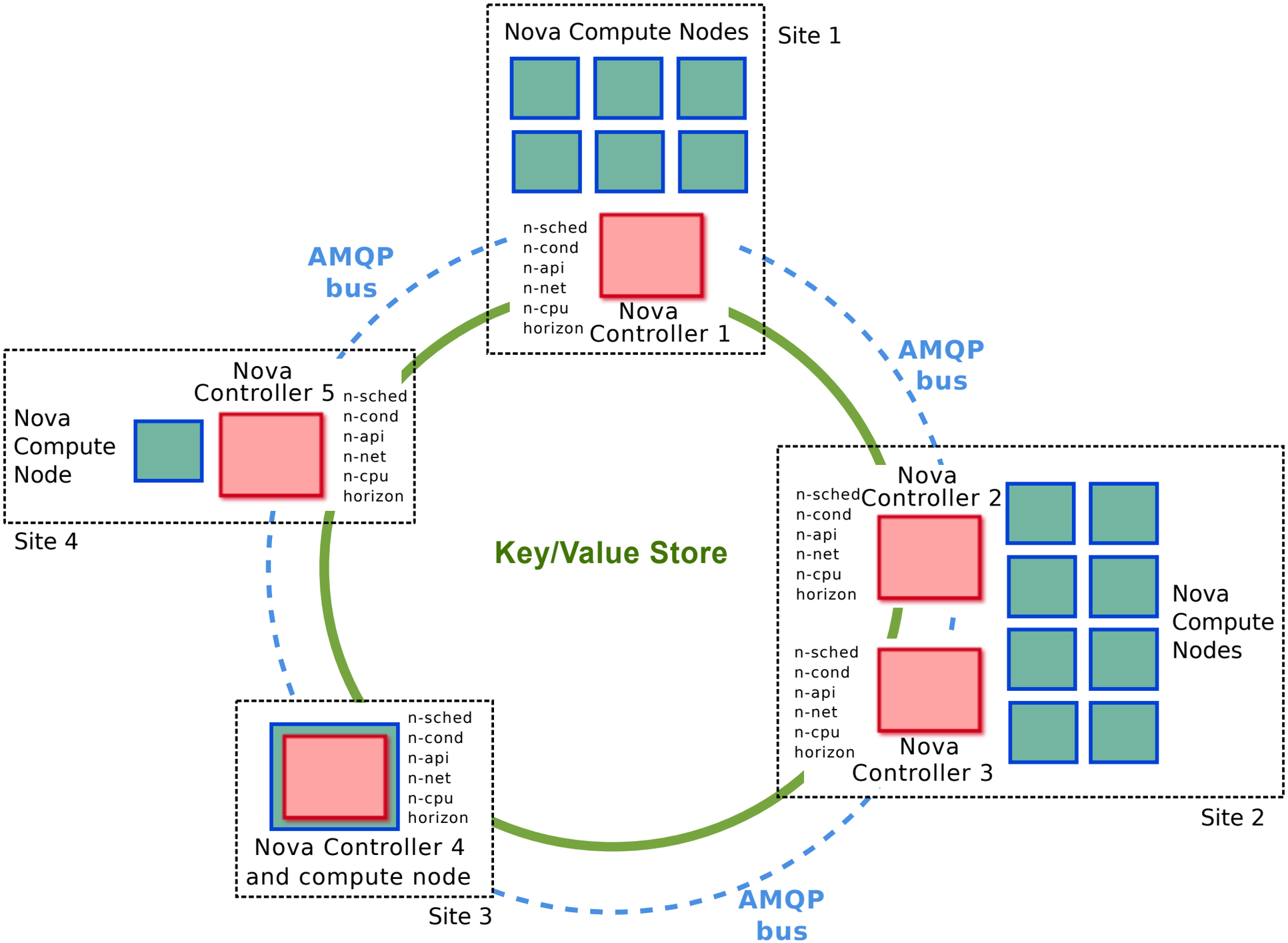
ROME

- Relational Object Mapping Extension for key/value stores
  Jonathan Pastor's Phd
  https://github.com/BeyondTheClouds/rome
- Enables the query of key/value store DB with the same interface as SQLAlchemy

- Enables Nova OpenStack to switch to a KVS without being too intrusive
- The KVS is distributed over (dedicated) nodes
- Nova services connect to the Key/value store cluster



Nova Proof-Of-Concept

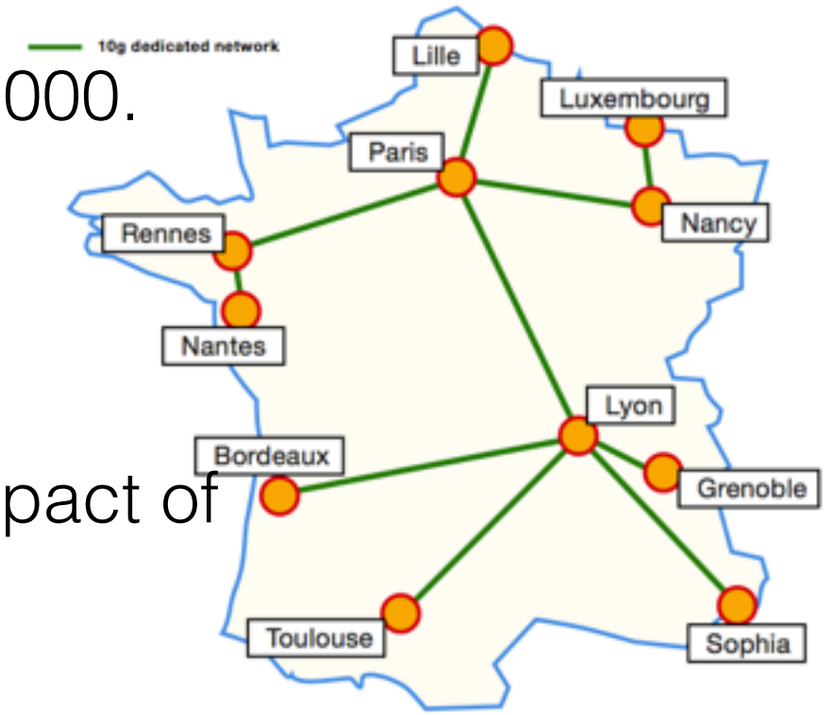


ROME Internals

- TODO Jonathan
- Compatibility with SQLAlchemy
- Almost no changes of the vanilla code (more than 97% of code)
- Model query

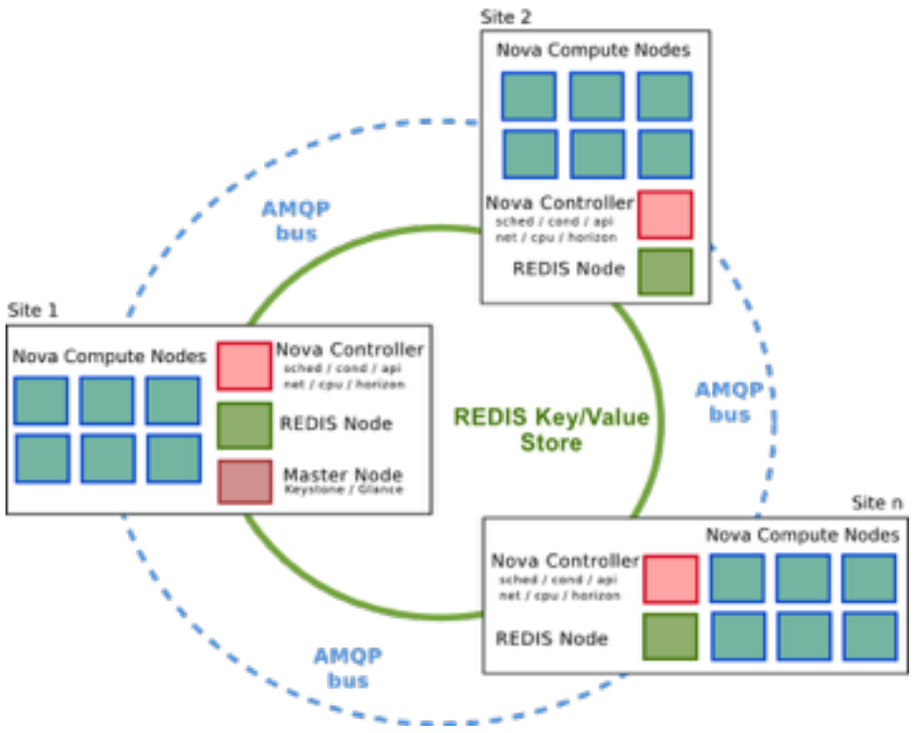
Experiments

- Experiments have been conducted on Grid'5000.
- ***mono-site experiments:*** *to evaluate the overhead of using REDIS and the network impact.*
- ***multi-site experiments*****:** To determine the impact of latency.
- High level mechanisms validation.

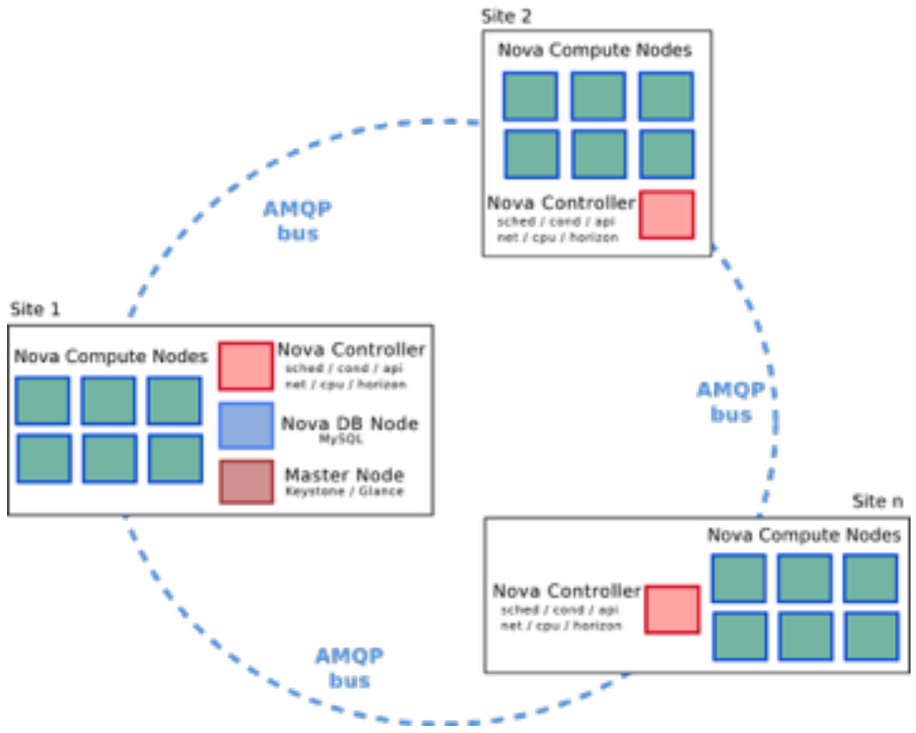


Experimental protocol

- Ask for the creation of 500 VMs, fairly distributed on each controller.



Rome+REDIS



SQLAlchemy+MySQL

Measuring the overhead

- Rome stores objects in a JSON format → *serialization/ deserialization cost*
- Rome reimplement some mechanisms: *join, transaction/session, …*

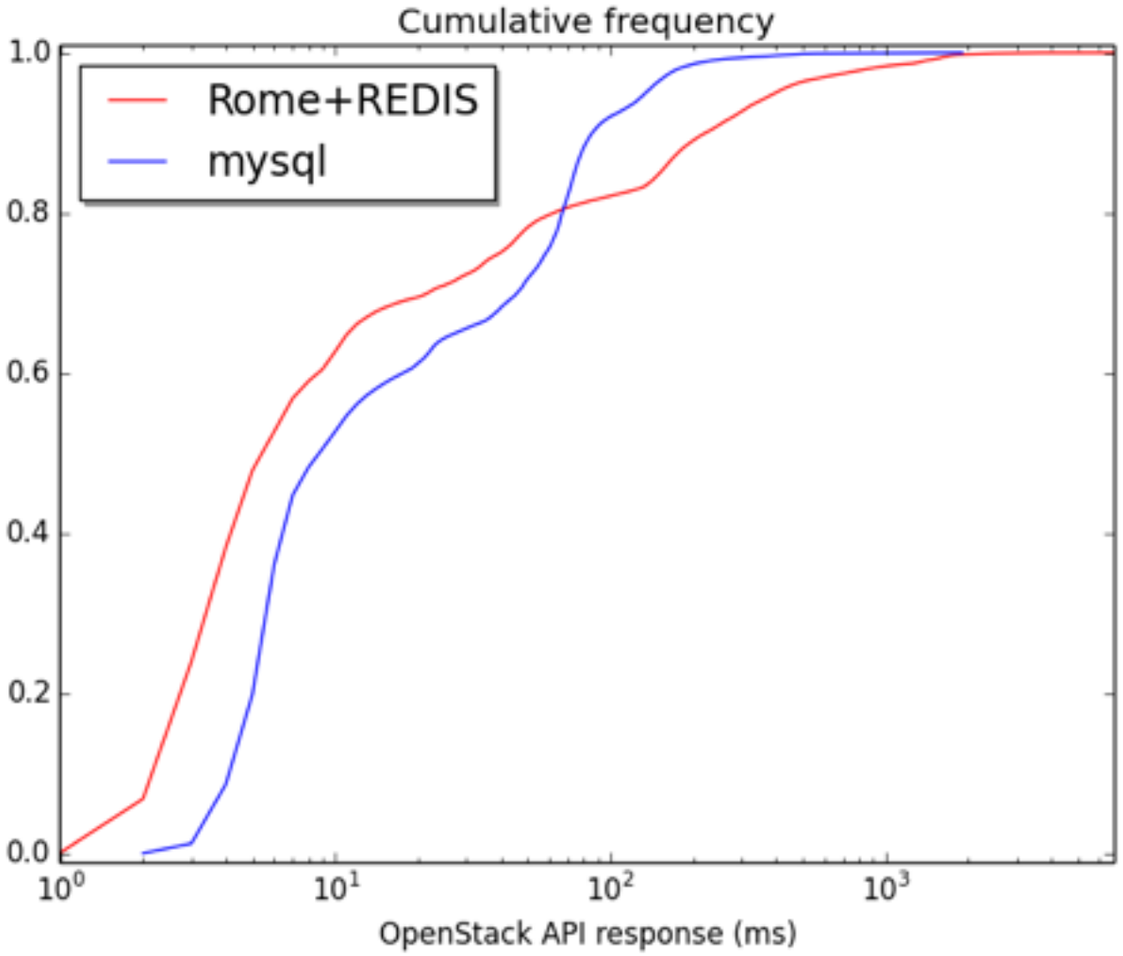


Table 1: Average response time to API requests for a mono-site deployment (in ms).

Backend configuration	REDIS	MySQL
1 node	83	37
4 nodes	82	-
4 nodes + repl	91	-

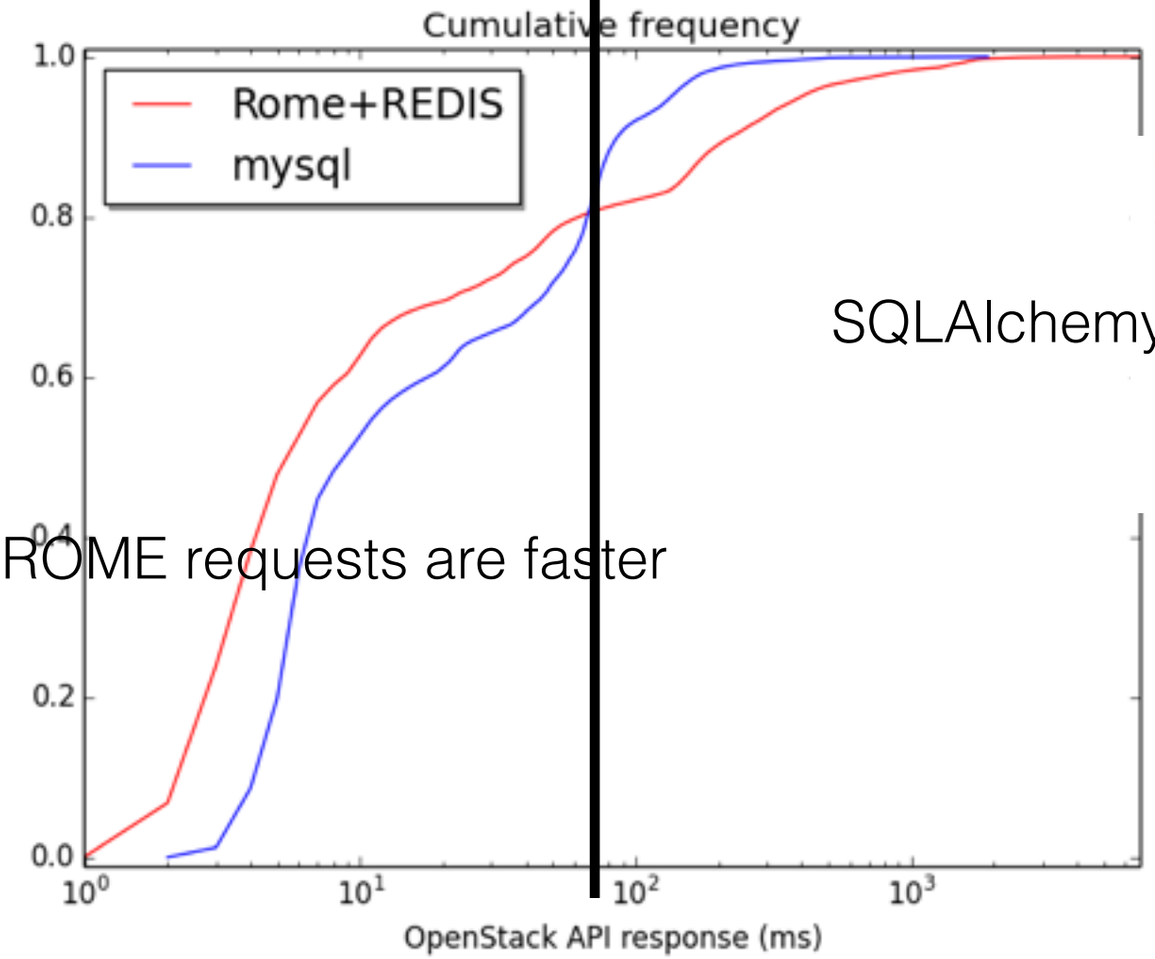
Table 2: Time used to create 500 VMs on a single cluster configuration (in sec.)

Backend configuration	REDIS	MySQL
1 node	322	298
4 nodes	327	-
4 nodes + repl	413	-

- With Rome+REDIS, 75% of requests are less than 40ms; 59ms with MySQL

Measuring the overhead

- Rome stores objects in a JSON format → *serialization/ deserialization cost*
- Rome reimplement some mechanisms: *join, transaction/session, …*



SQLAlchemy

ROME requests are faster

Table 5: Amount of data exchanged over the network (in MegaBytes.)

Backend configuration	REDIS	MySQL
1 node	2190	1794
4 nodes	2382	-
4 nodes + repl (1 replica)	4186	-

- Enabling replication causes a significant network overhead.

Multisite experiments

- Experiments with nodes from Rennes:
  - *virtual cluster by adding latency- thanks to the TC tools;*
  - *each cluster was containing 1 controller, 6 compute nodes (and 1 dedicated node in the case of REDIS).*
  - *Redis and SQL used*
- To fairly compare with MySQL, data replication was not activated.

Table 3: Time used to create 500 VMs with a 10ms inter-site latency (in sec.).

Nb of locations	REDIS	MySQL
2 clusters	271	209
4 clusters	263	139
6 clusters	229	123
8 clusters	223	422

one SQL server for all sites

Table 4: Time used to create 500 VMs with a 50ms inter-site latency (in sec.).

Nb of locations	REDIS	MySQL
2 clusters	723	268
4 clusters	427	203
6 clusters	341	184
8 clusters	302	759

SQL scalability limitations

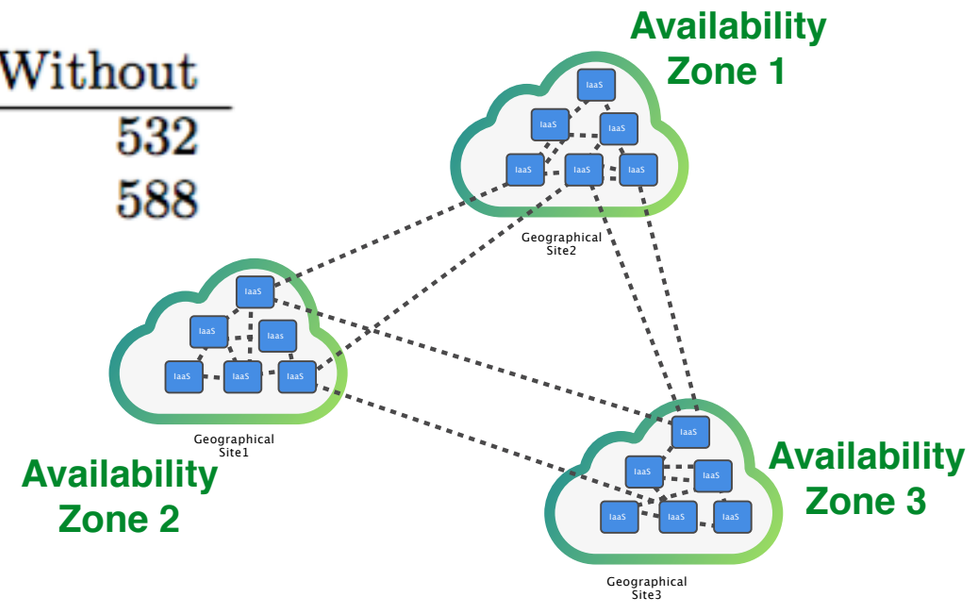
- Increasing the number of nodes leads to better reactivity.
- With 8 clusters, we saturate MySQL.

Compatibility with existing mechanisms

- Tested the usage of advanced OpenStack feature:
  host-aggregates / availability zones

Table 6: Time used to create 500 VMs with a 10ms inter-site latency (in sec.) and using Redis with and without host-aggregates.

Nb of locations	With	Without
4 clusters	582	532
8 clusters	492	588

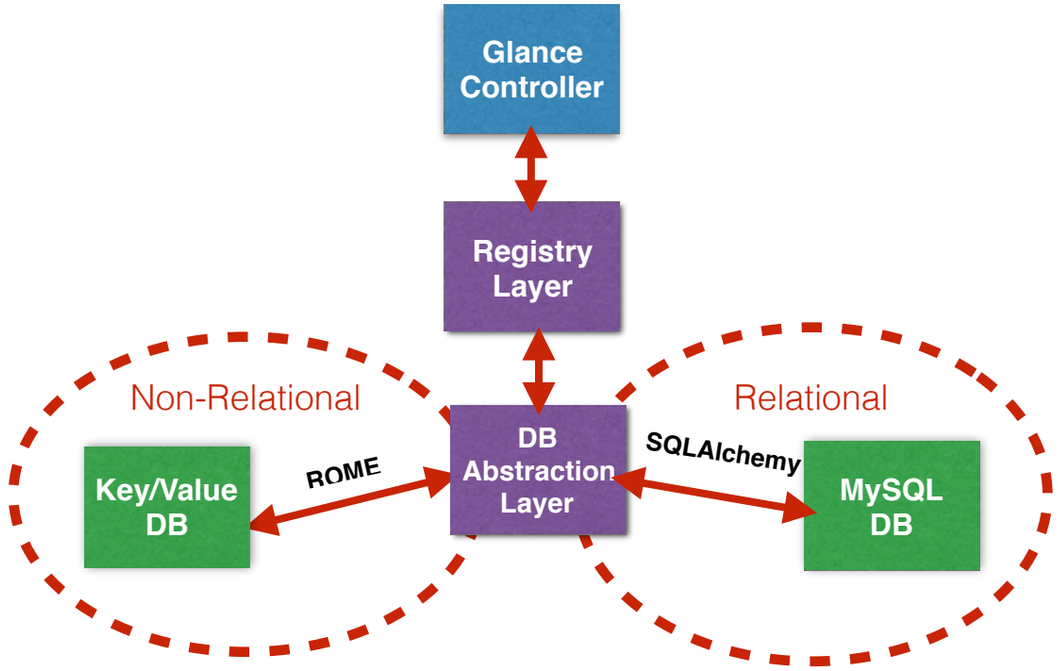


- As we targeted a low-level component, ROME is compatible with most of the existing features.
- Performances seem to be overall acceptable.

DISCOVERY - Where we are ?

- Code available at : https://github.com/BeyondTheClouds
- ROME
- Nova fork
- Glance fork [WiP]
- Development environment

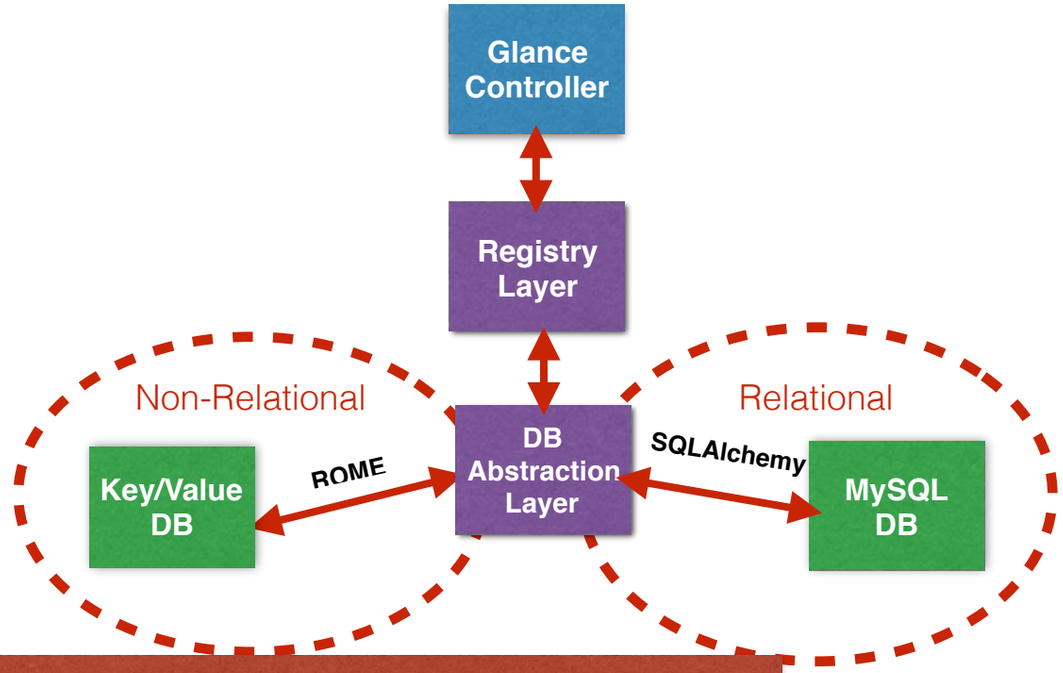
  Based on vagrant / Devstack
  - standalone
  - collaborative



DISCOVERY - Where we are ?

- Code available at : https://github.com/BeyondTheClouds
- ROME
- Nova fork
- Glance fork [WiP]
- Development environment

  Based on vagrant / Devstack
  - standalone
  - collaborative

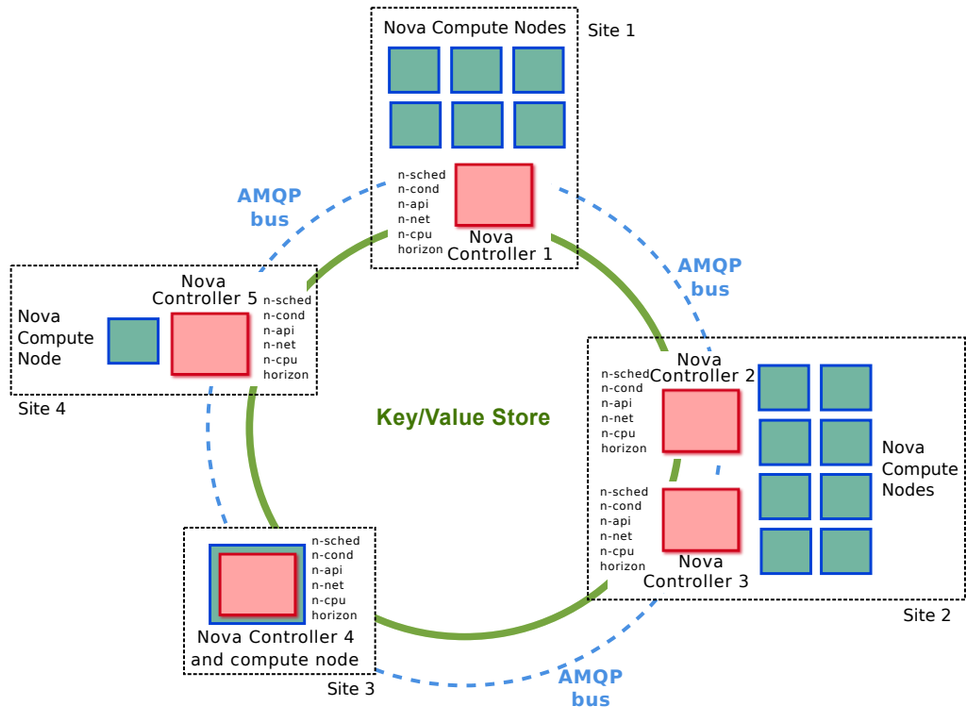


Comments/feedback Welcome
discovery-contact@inria.fr
#discovery on irc.oftc.net

KVS Integration - Where we go ?

Towards CONFORMITY

- Short-term : October. 2016 (first official release)
- Compatible with OpenStack Newton
  Redis Support for Nova and Glance (Cinder ?)
- **Feature Conformity :**
  - Testing (focused on nova.tests.unit.db tests)
  - Tempest
- **Performance Conformity**
  - Driver performance Improved
    (speed up the ~20% slowest requests)
  - Release of our continuous benchmarking tool based on Rally
  - Scalability as well as correctness on-going work: 100 sites, 10 servers each



KVS Integration - Where we go ?

Towards Locality

- Mid term: April. 2017 (second official release)
- Extension for other NoSQL backends (e.g cassandra)
- Locality improved (r/w locally and mitigate inter-site traffic)

  - reshape the data model

  - better control over key partionning

- Neutron support (Orange Labs)
  Postdoc ongoing and a PhD position open.
  DragonFlow on-going development

DISCOVERY - Long term roadmap

- A lot of scientific/technical challenges

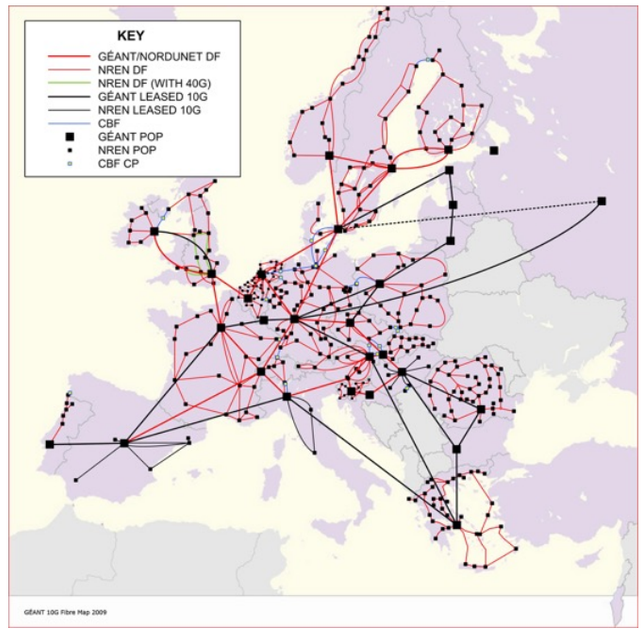
- Shared services
  - storage backend for Fog/Edge computing (KVS like)
  - communication layer (scalability, inter-site control,…)
- Compute: Locality at every level (API, scheduler, …)
- Network: Revision of Neutron internals (REDIS but also SDN functions).

- Storage
  S3-like service for Fog/Edge Computing (SWIFT / RADOS under high latency ?)
  Multi sites VM Image Management (replication/prefetching mechanisms)

- Distribution e.g user authentication, tenant network, quota management

- On the OpenStack software platform itself !
  Deployment / reconfiguration at each new release/upgrade throughout the whole infrastructure.

DISCOVERY - Long term roadmap



- A lot of scientific/technical challenges

- Shared services
  - storage backend for Fog/Edge computing (KVS like)
  - communication layer (scalability, inter-site control,…)
- Compute: Locality at every level (API, scheduler, …)
- Network: Revision of Neutron internals (REDIS but also SDN functions).

- Storage
  S3-like service for Fog/Edge Computing (SWIFT / RADOS under high latency ?)
  Multi sites VM Image Management (replication/prefetching mechanisms)

- Distribution e.g user authentication, tenant network, quota management

- On the OpenStack software platform itself !
  Deployment / reconfiguration at each new release/upgrade throughout the whole infrastructure.

Discovery Task forces

- Today: provided by Orange and Inria

6 Phds

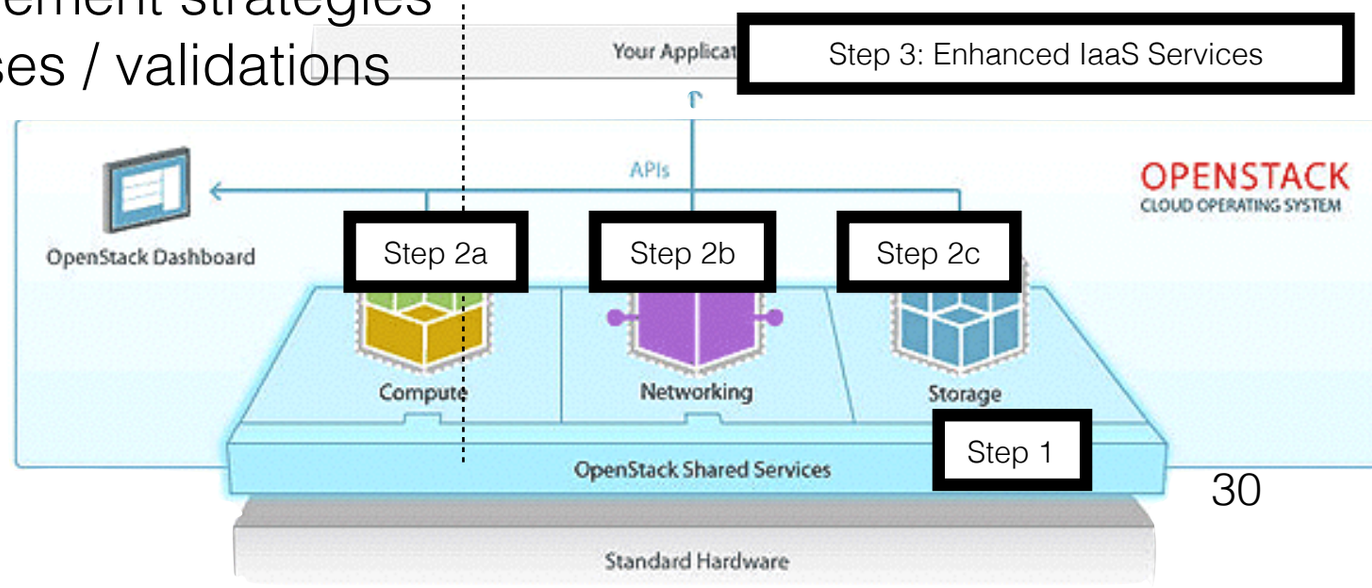
- Locality based Overlay networks - step 1
- Monitoring - step 1
- Security enforcement - step 1
- Distributed SDN capabilities - step 2b
- Image management - step 2c
- Locality from the application elasticity view point - step 3

6 post docs

- Cost benefit analysis and energy opportunity
- Identification of Neutron challenges
- Deployment and reconfiguration of OpenStack
- Data scheduling policies
- VM placement strategies
- Use-cases / validations

3 engineers

- Core developper (soon !)
- Sys Admin
- GUI/command line developper



Discovery Task forces

- Today: provided by Orange and Inria

6 Phds

- Locality based Overlay networks - step 1
- Monitoring - step 1
- Security enforcement - step 1
- Distribut... co...
- ...
- L...ality from the application elasticity view point - step 3

6 post docs

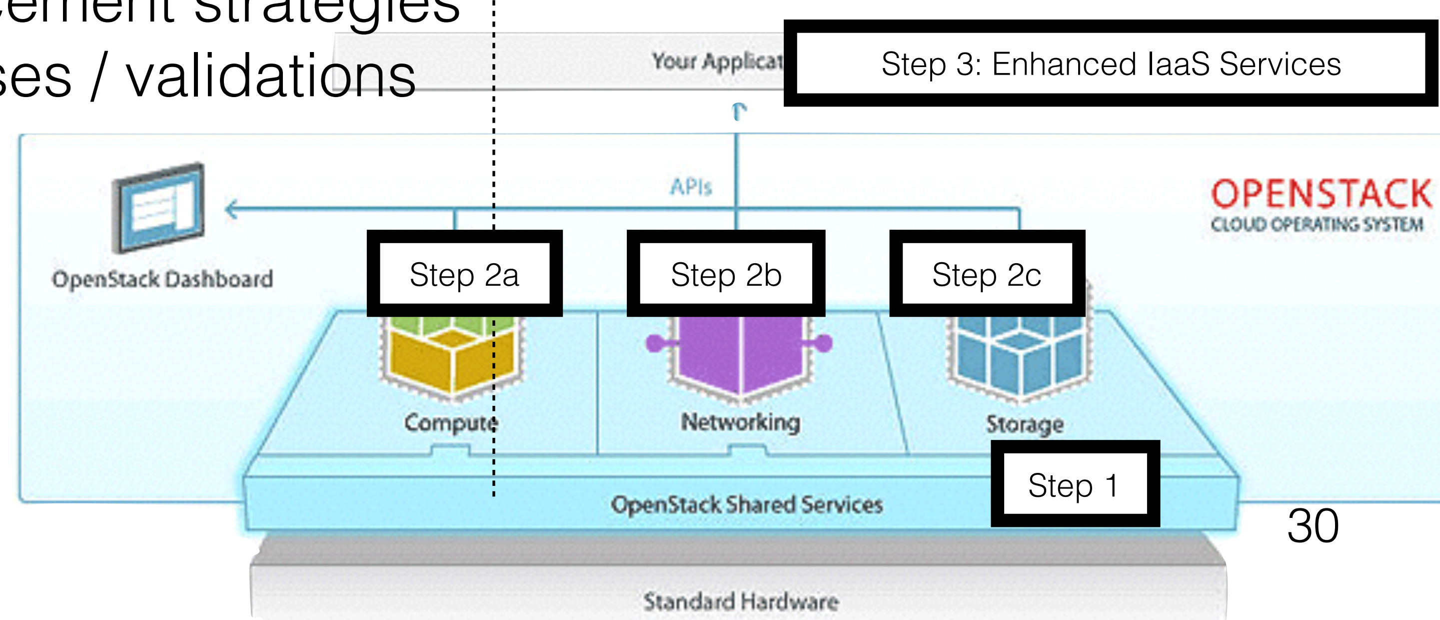
- Cost benefit analysis a... energy opport...
- Identifi...
- ...on of ...Stack
- Data scheduling policies
- VM placement strategies
- Use-cases / validations

3 e...

...n !)

...mmand line developper

See Open Positions on the Discovery website.
or just come and contribute ;)



Take away message

- Academics and Industrials agree :
  Fog/Edge Computing will become

- Do not reinvent the wheel !

- We want to help OpenStack to support the Massively Distributed Use Case

- Several companies/instituts expressed their interests w-r-t the Discovery objectives (Orange, Thales, EU NRENs, …)

- Creation of a massively distributed clouds WG
  (federate and favour synergy between on-going actions).

Massively distributed (MD) working group

4:40pm-6:10pm Hilton Austin - Level 6 - Salon J My calendar

Massively Distributed Clouds Working Group Inaugural Meeting

Working Groups PRESENTATION

The objectives of this inaugural meeting are:
- To present in details the massively distributed cloud use-case (aka, the Fog/Edge Computing paradigm)
- To identify persons/institutions that might be interested to take part to such a WG;
- To identify on-going related activities addressed inside but also outside the OpenStack community (short presentations will be given);
- To define a roadmap and identify key persons that can contribute to make OpenStack cooperative by default.

The major difference with respect to existing WGs such as the Large-Scale deployment one, is that we would like to address challenges related to the geo-distribution of resources and WANwide exchanges (one single cloud/cell but deployed across multiple small sites). Concretely, we want to study how the vanilla OpenStack code can support the Fog/Edge Computing use-case while minimising the changes on the OpenStack internals.

Further information available at: https://etherpad.openstack.org/p/massively-distributed-clouds

Adrien Lebre
INRIA

Matthieu Simonin
Research Engineer - Discovery Technical Manager

Craig Lee
Senior Scientist

Chaoyi Huang
Huawei

Level: Intermediate

Tags: Architect, Telecom, Research

EVENT DETAILS

https://etherpad.openstack.org/p/massively-distributed-clouds

Massively distributed (MD) working group

4:40pm-6:10pm Hilton Austin - Level 6 - Salon J My calendar

Massively Distributed Clouds Working Group Inaugural Meeting

Working Groups PRESENTATION

The objectives of this inaugural meeting are:
- To present in details the massively distributed cloud use-case (aka, the Fog/Edge Computing paradigm)
- To identify persons/institutions that might be interested to take part to such a WG;
- To identify on-going related activities addressed inside but also outside the OpenStack community (short presentations will be given);
- To define a roadmap and identify key persons that can contribute to make OpenStack cooperative by default.

The major difference with respect to existing WGs such as the Large-Scale deployment one, is that we would like to address challenges related to the geo-distribution of resources and WANwide exchanges (one single cloud/cell but deployed across multiple small sites). Concretely, we want to study how the vanilla OpenStack code can support the Fog/Edge Computing use-case while minimising the changes on the OpenStack internals.

Further information available at: https://etherpad.openstack.org/p/massively-distributed-clouds

Join us

EVENT DETAILS

Adrien Lebre
INRIA

Matthieu Simonin
Research Engineer - Discovery Technical Manager

Craig Lee
Senior Scientist

Chaoyi Huang
Huawei

Level: Intermediate

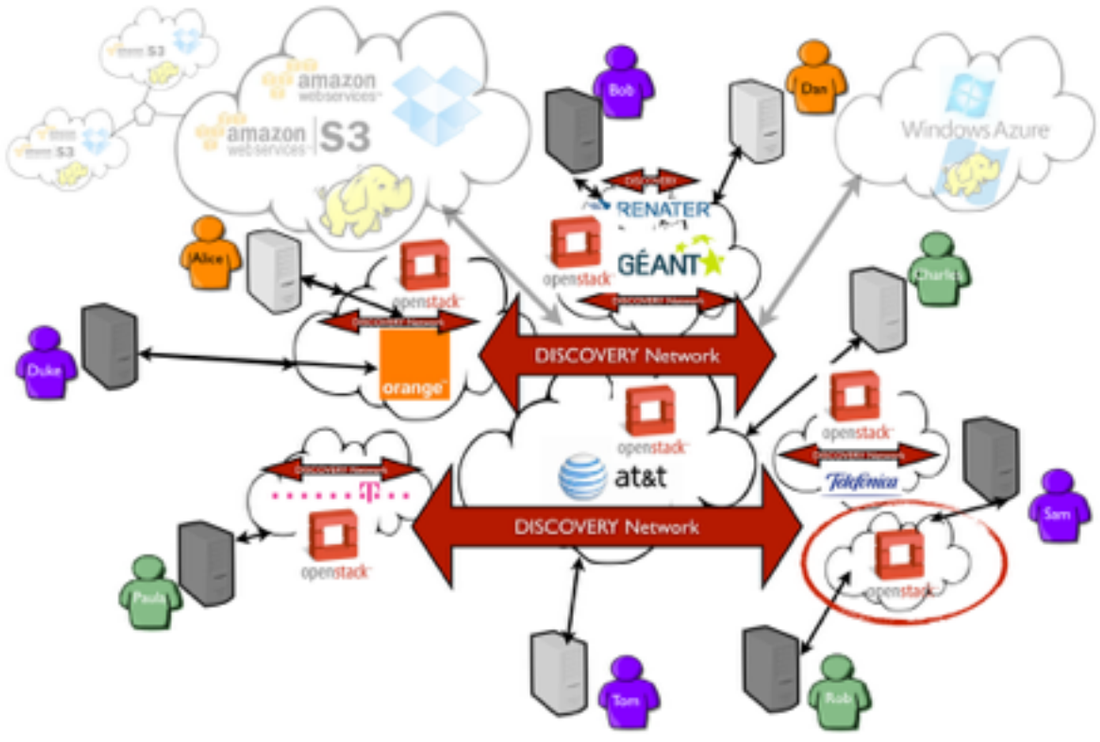
Tags: Architect, Telecom, Research

https://etherpad.openstack.org/p/massively-distributed-clouds

The DISCOVERY Initiative

- Several researchers, engineers, stakeholders of important EU institutions and SMEs have been taking part to numerous brainstorming sessions (BSC, CRS4, Unine, EPFL, PSNC, Interoute, Orange Labs, Peerialism, TBS Group, XLAB, …)

http://beyondtheclouds.github.io/



RENATER
CONNECTEUR DE SAVOIRS

Inria

orange™

discovery-contact@inria.fr