Beyond the Clouds, The Discovery Initiative

How Should Next Generation Utility Computing Infrastructures Be Designed to Solve Sustainability & Efficiency Challenges ?

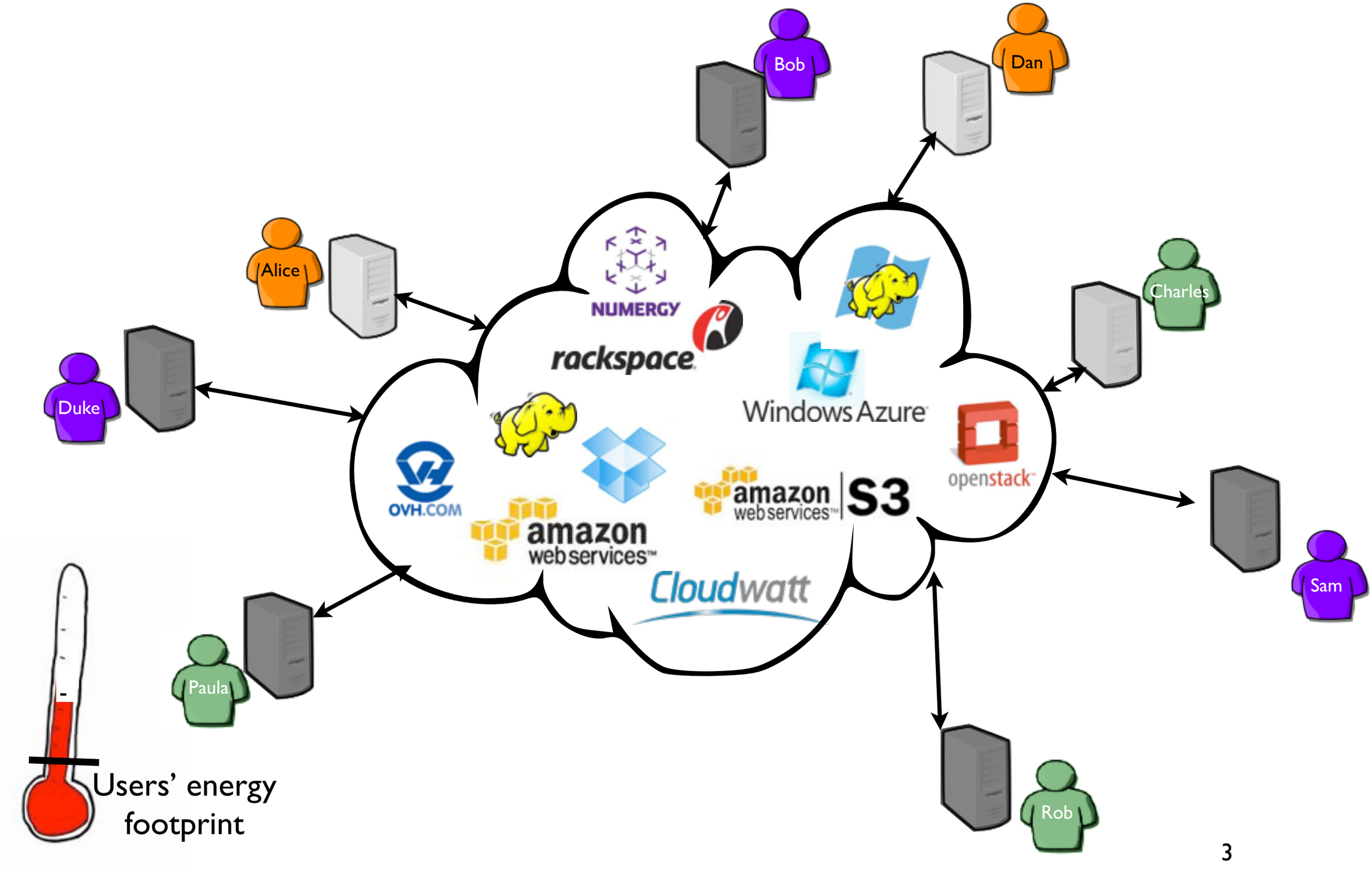
Adrien Lebre
Jan, 2015 - Inria /Alcatel Lucent Bell Labs

Localization is a key element to deliver *efficient* as well as *sustainable* **Utility Computing** solutions

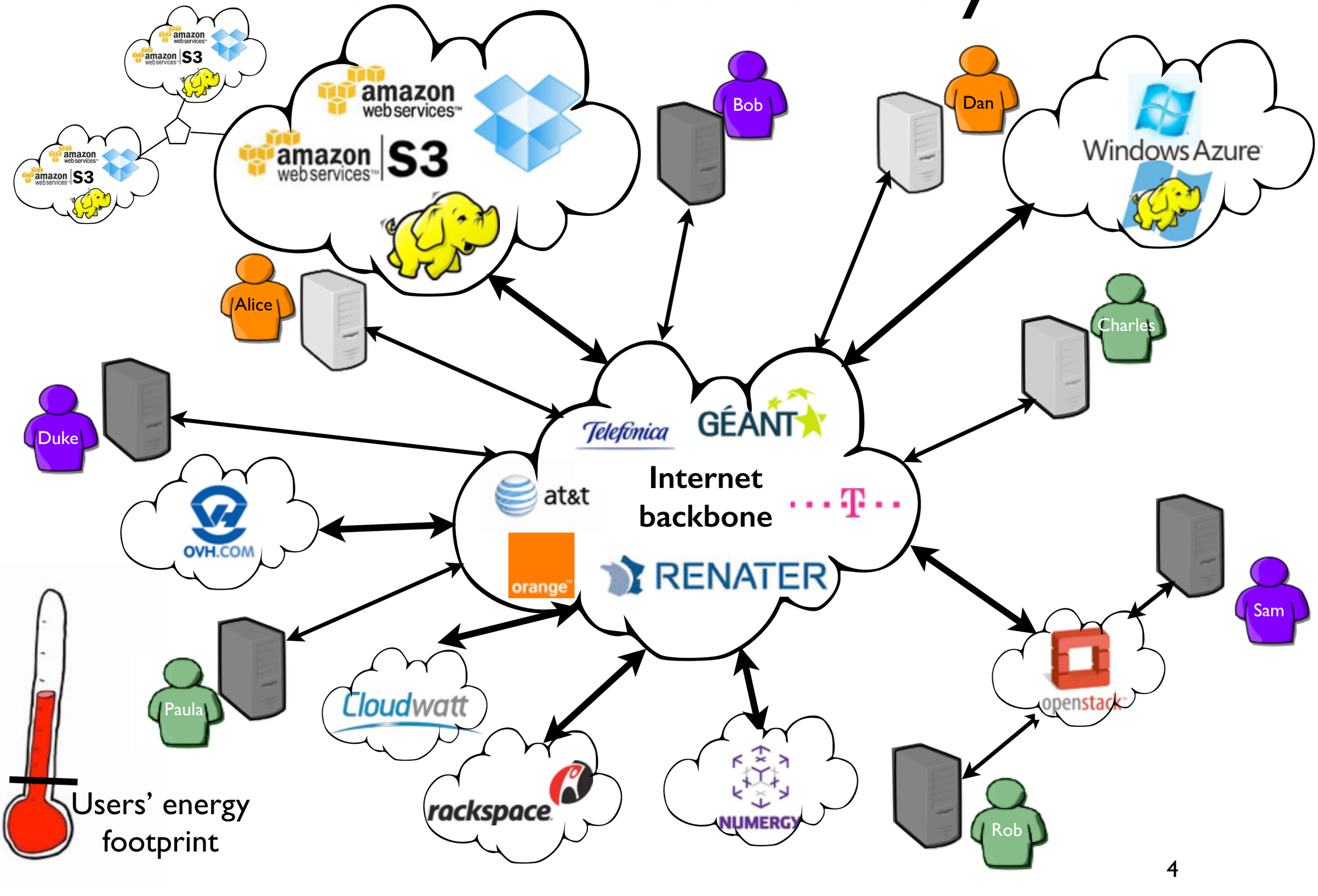
A simple Idea

Bring Clouds back to the cloud

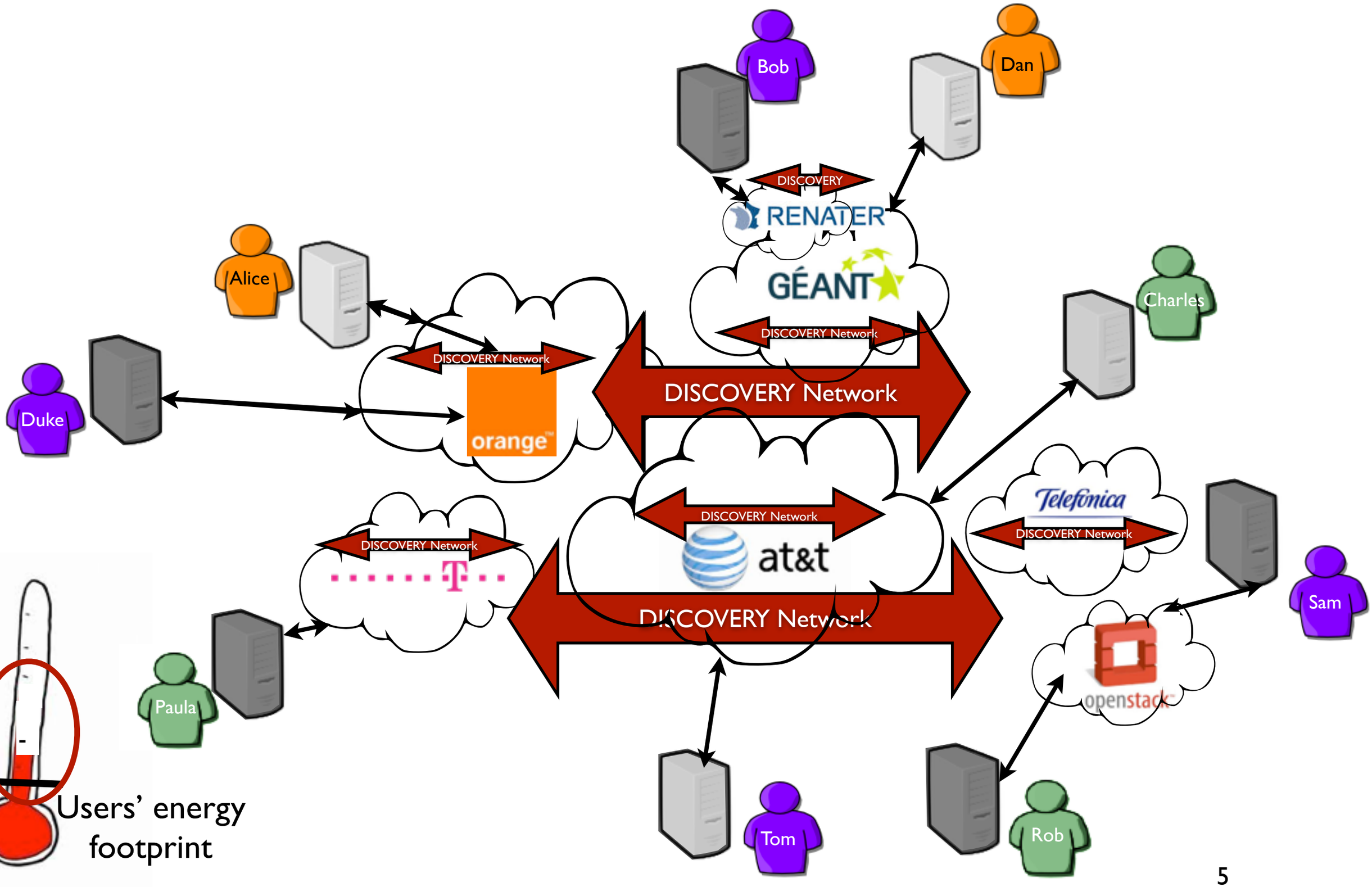
The cloud from end-users



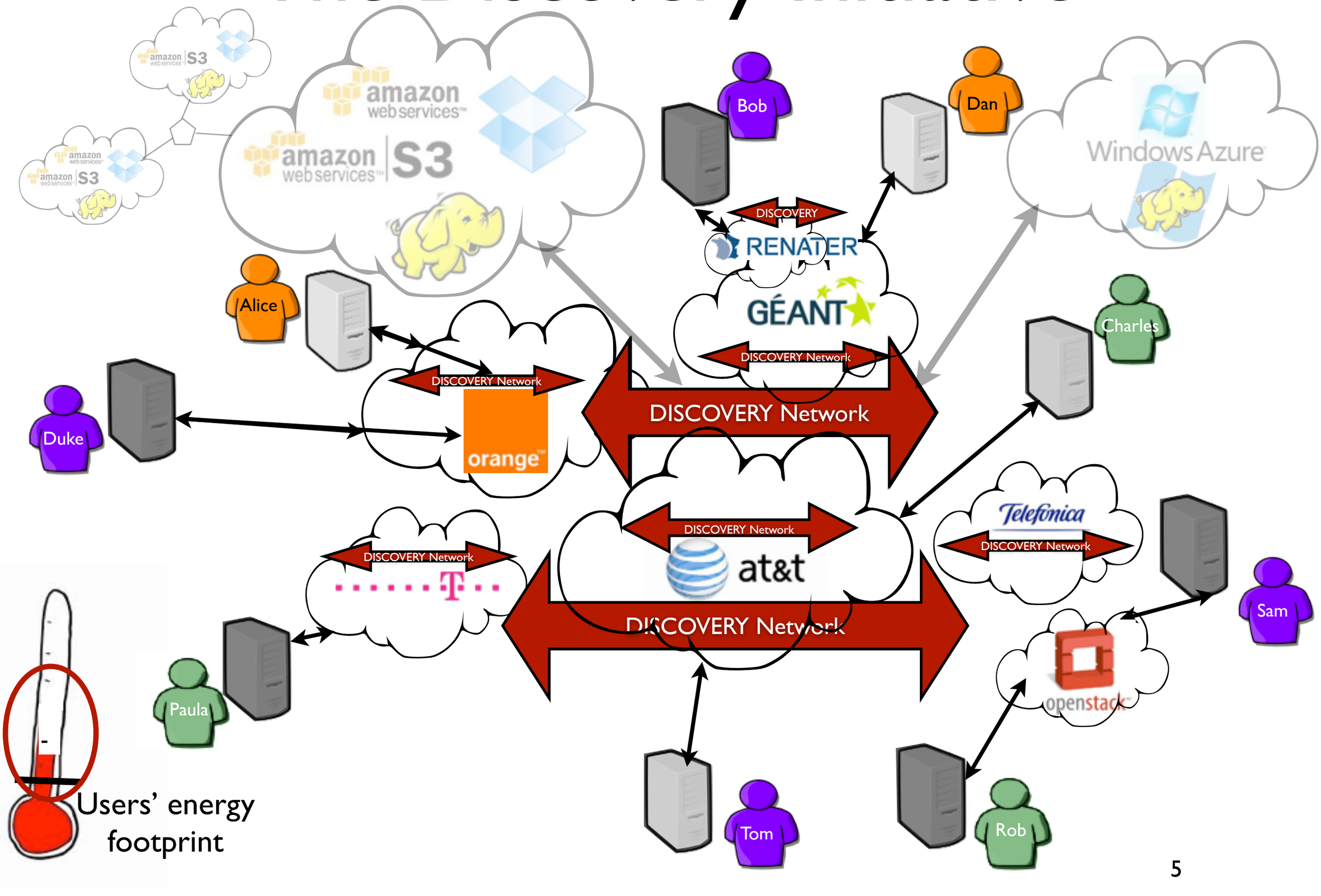
The cloud in reality



The Discovery Initiative



The Discovery Initiative



Users' energy footprint

Why ?

Let's give a look to the current situation

The Current Trend: Large off shore DCs

- 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 DCs

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

credits: google map - Quincy

The Current Trend: Large off shore DCs

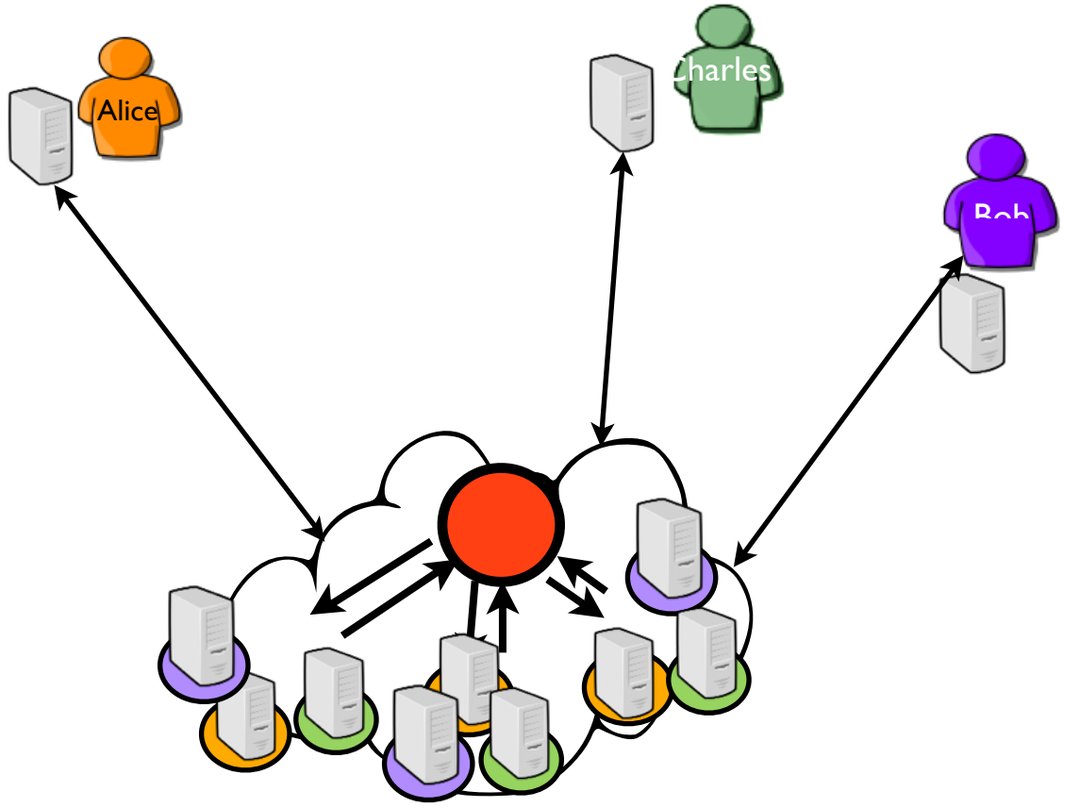
- To cope with the increasing UC demand while handling energy concerns but…

credits: coloandcloud.com

Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

  1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

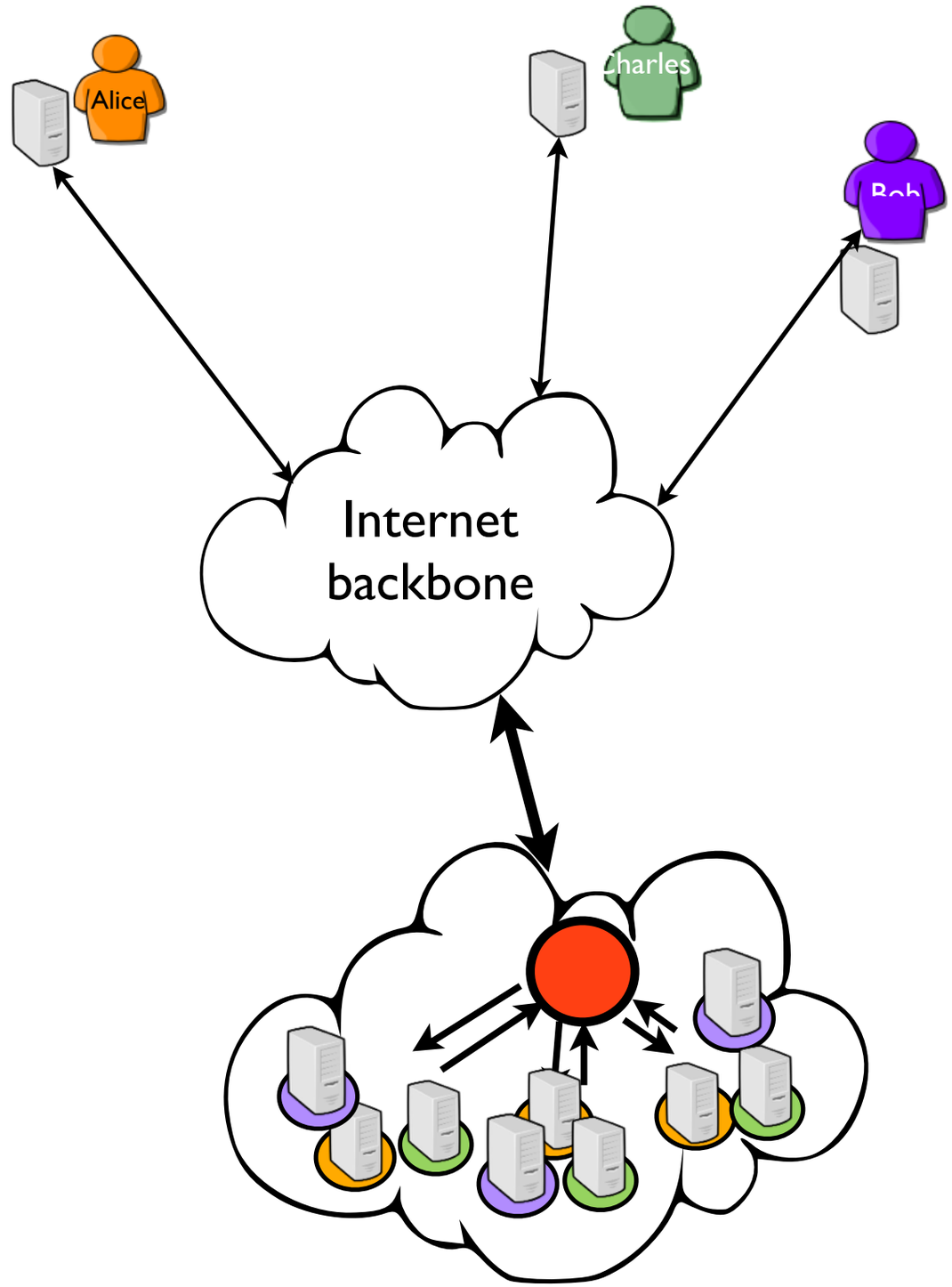


Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

  1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

  2. Overhead implied by the unavoidable use of the Internet to reach distant platforms



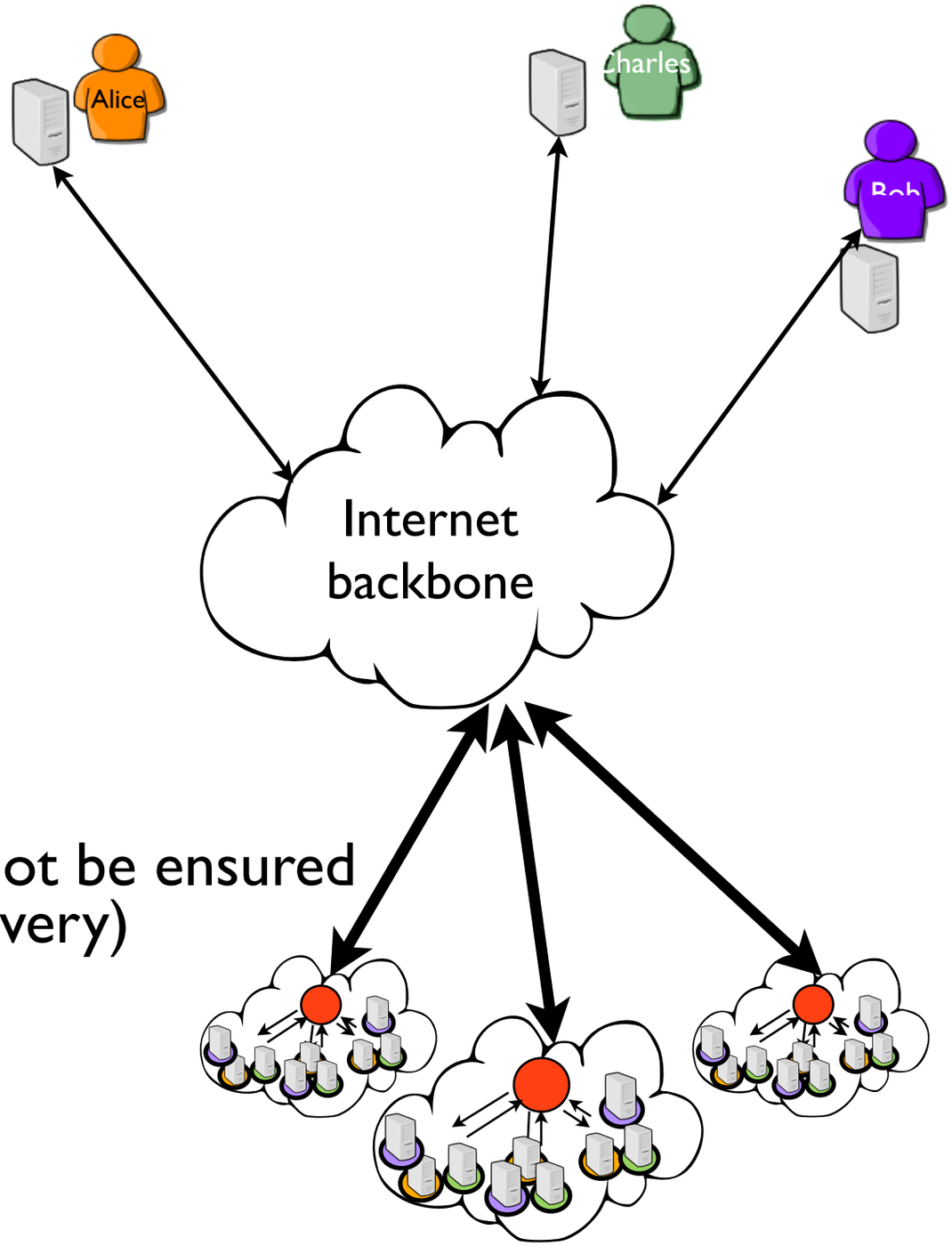
Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

2. Overhead implied by the unavoidable use of the Internet to reach distant platforms

3. The connectivity to the application/data cannot be ensured by centralized dedicated centers (disaster recovery)



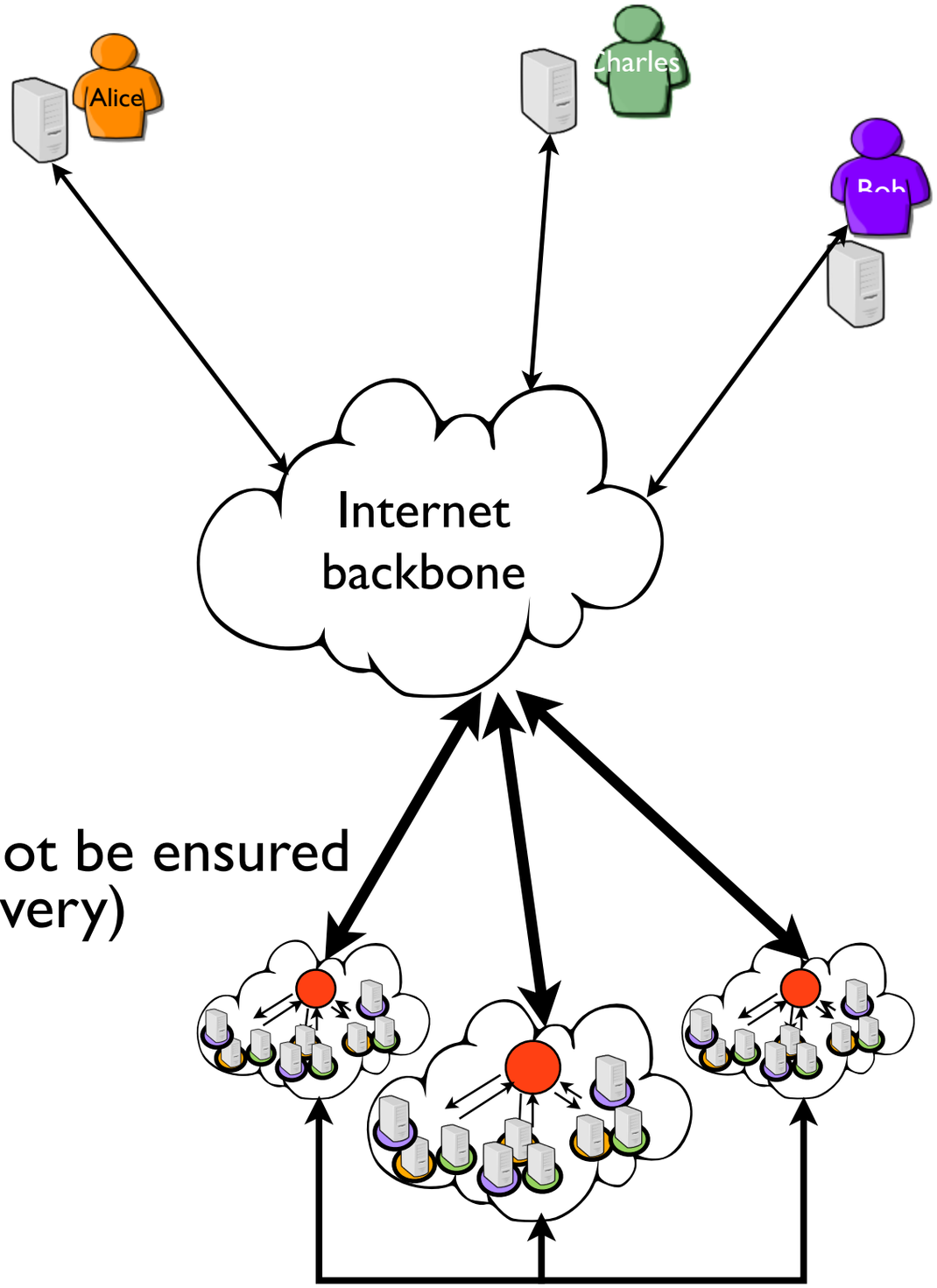
Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

2. Overhead implied by the unavoidable use of the Internet to reach distant platforms

3. The connectivity to the application/data cannot be ensured by centralized dedicated centers (disaster recovery)



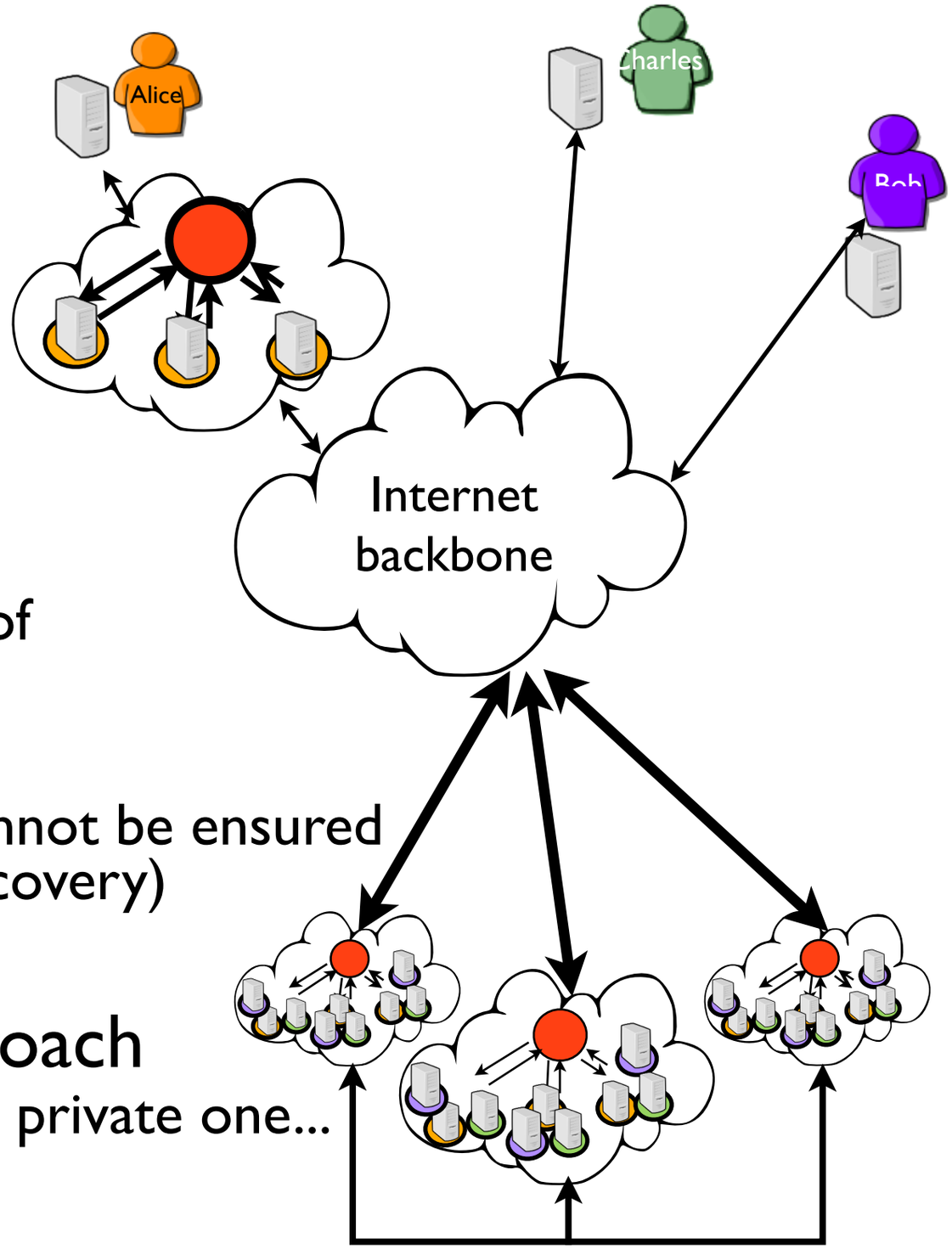
Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

  1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

  2. Overhead implied by the unavoidable use of the Internet to reach distant platforms

  3. The connectivity to the application/data cannot be ensured by centralized dedicated centers (disaster recovery)



- Hybrid platforms: a promising approach

  It depends how you are going to extend the private one...

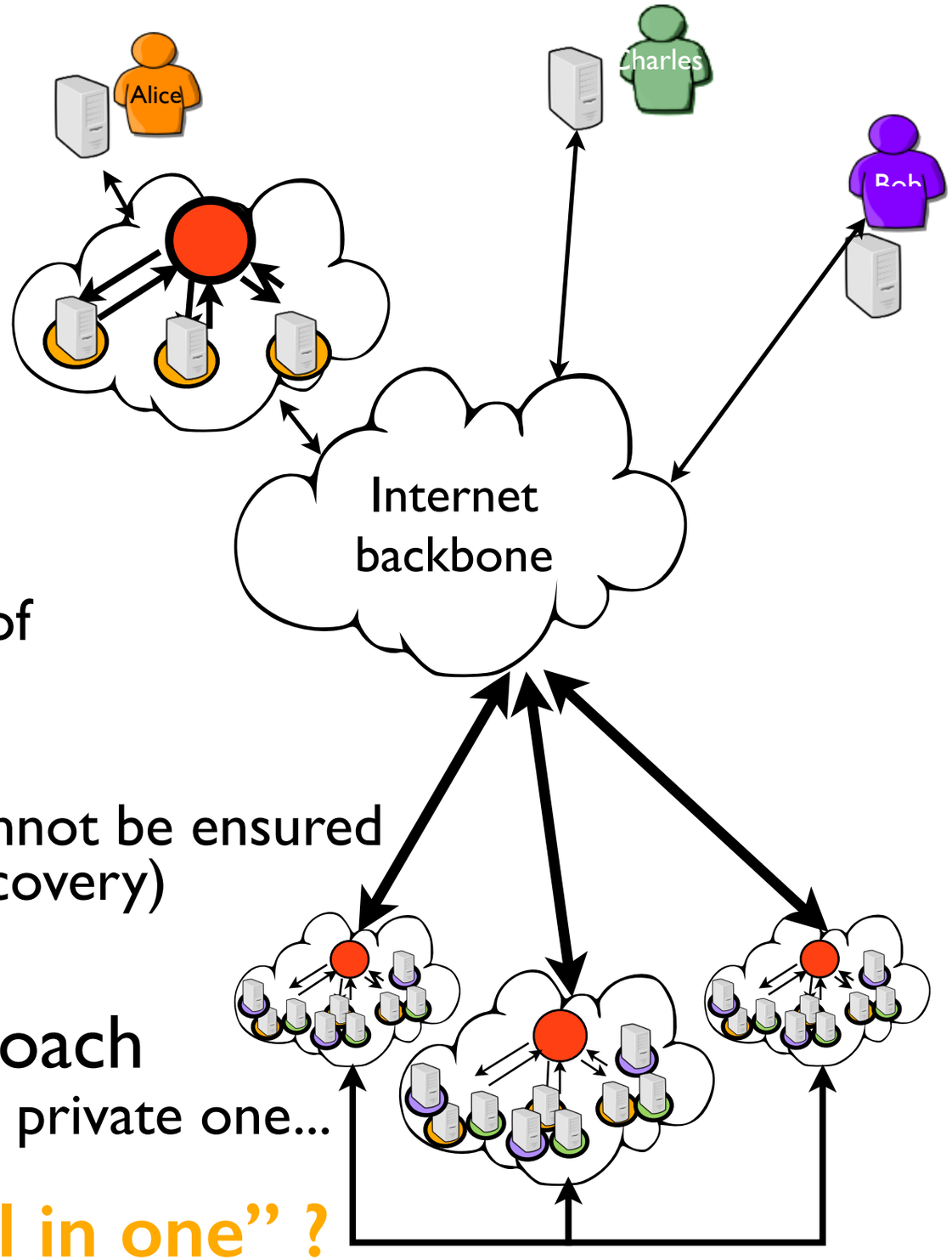
Inherent limitations of current solutions

- Large off shore DCs to cope with the increasing UC demand while handling energy concerns but…

  1. Externalization of private applications/data (jurisdiction concerns, PRISM NSA scandal, Patriot Act)

  2. Overhead implied by the unavoidable use of the Internet to reach distant platforms

  3. The connectivity to the application/data cannot be ensured by centralized dedicated centers (disaster recovery)



- Hybrid platforms: a promising approach
  It depends how you are going to extend the private one...

Can we address these concerns "all in one" ?

µ/nDC concept

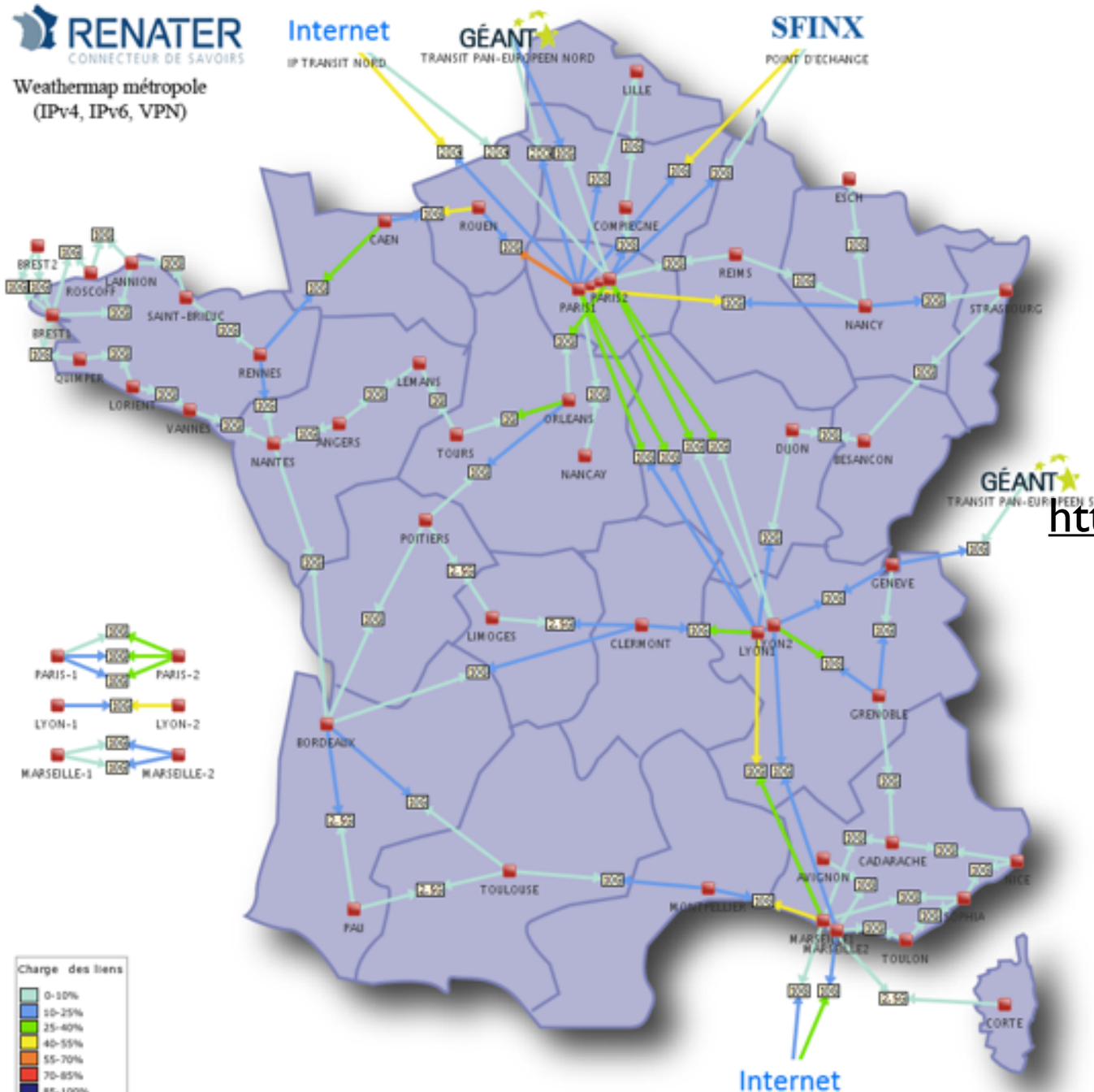
How and where the μDC concept can be deployed ?

Locality Based Utility Computing Toward LUC Infrastructures

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

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

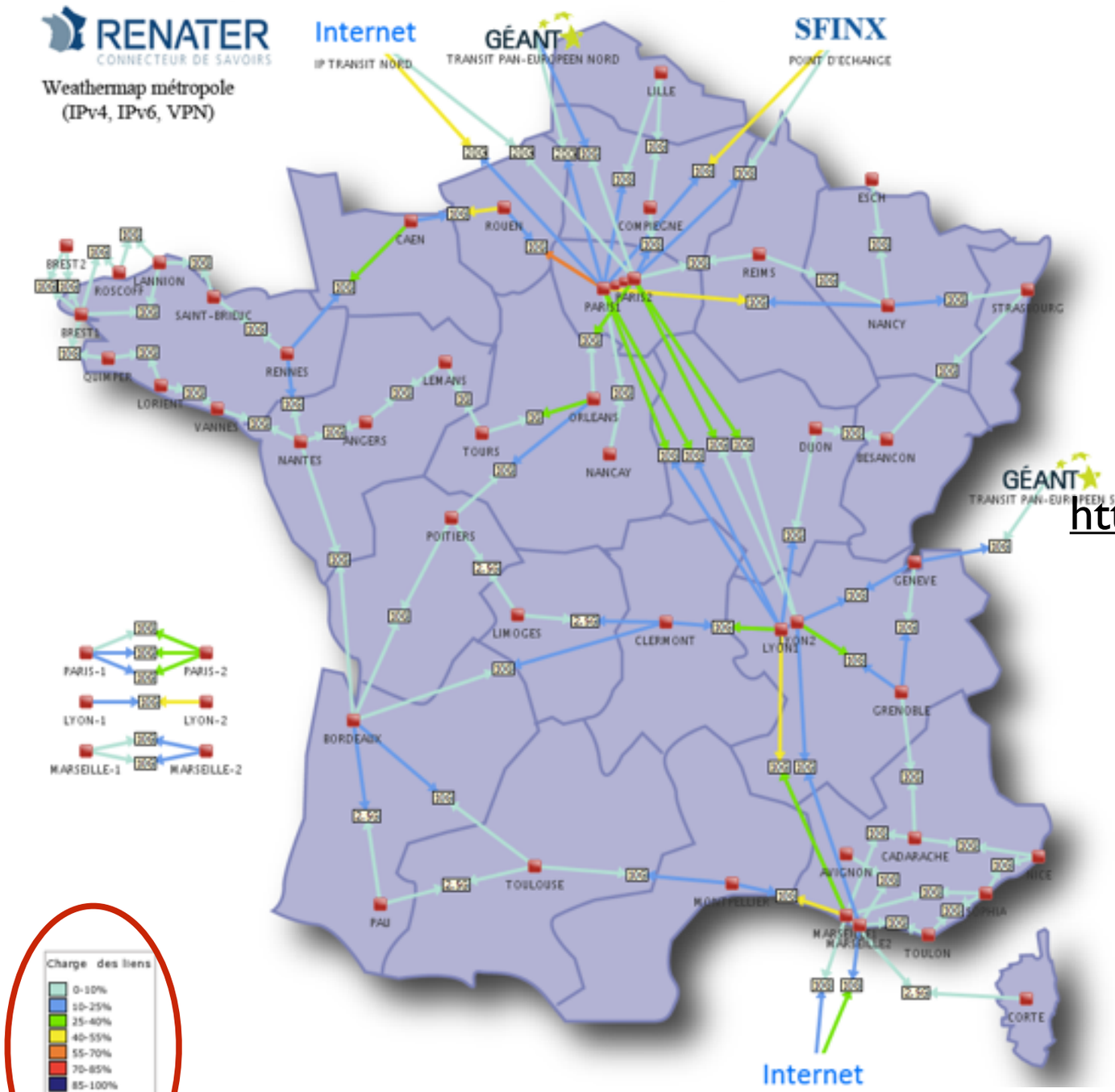


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

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

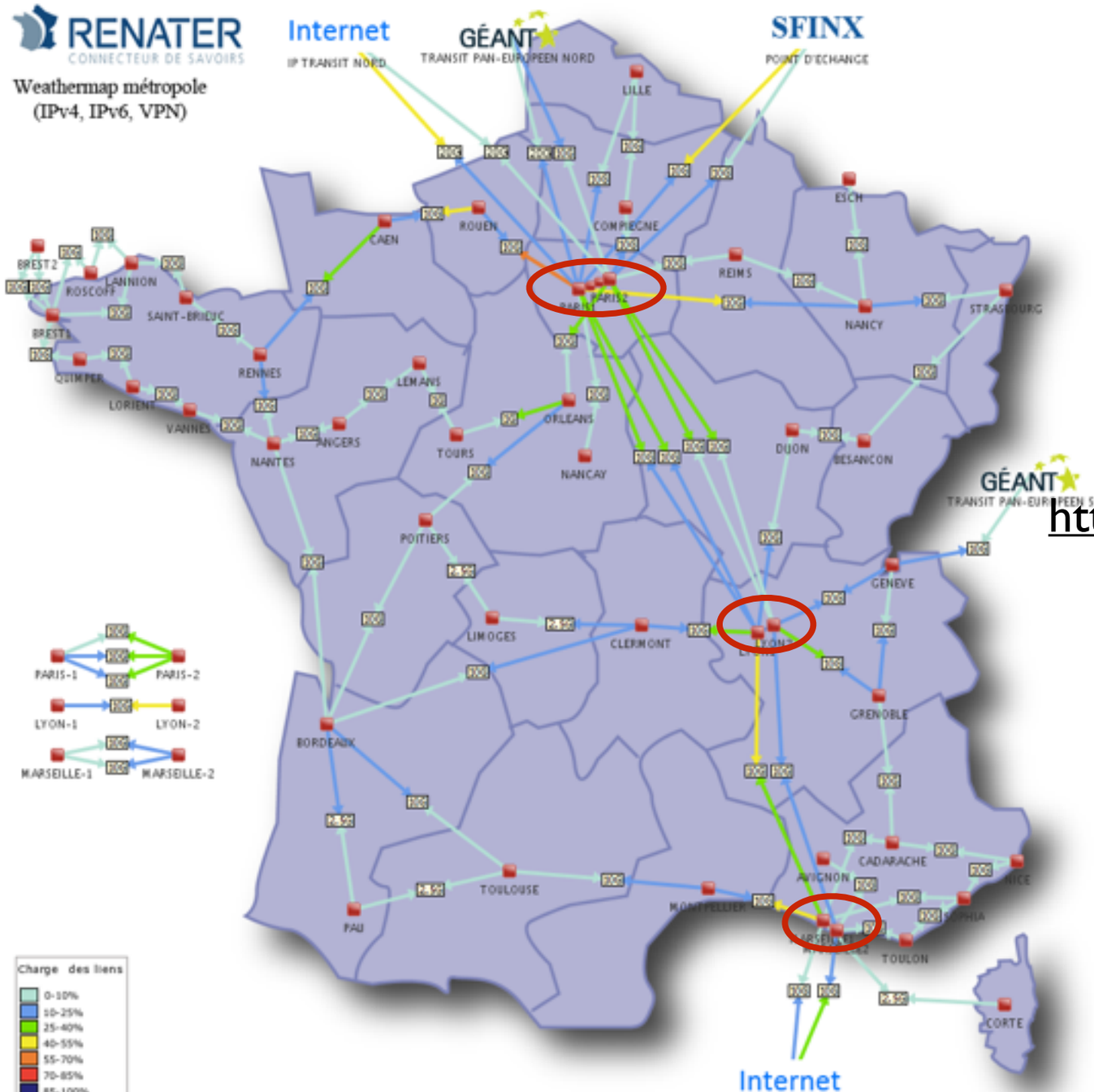


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

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

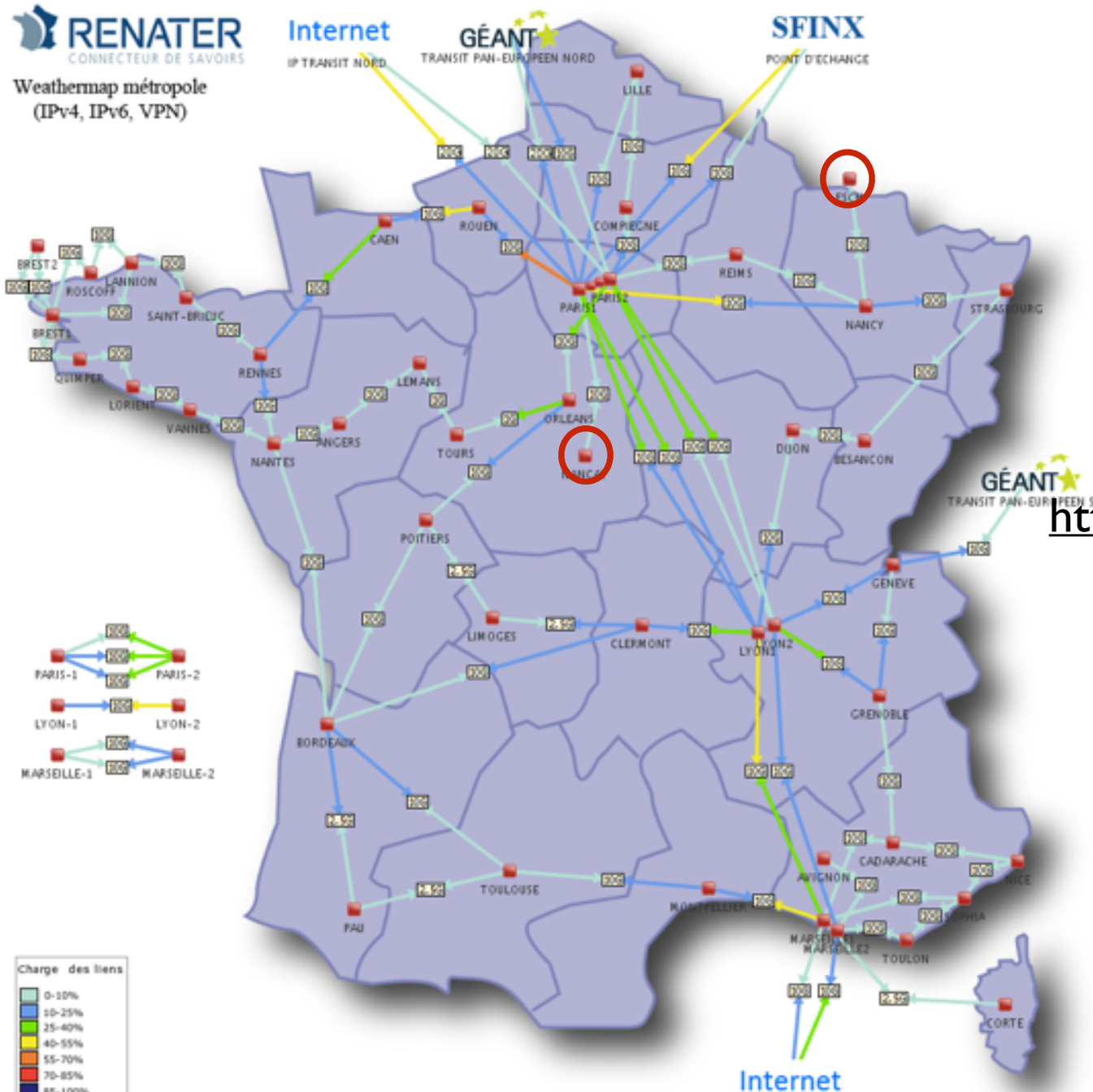


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

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

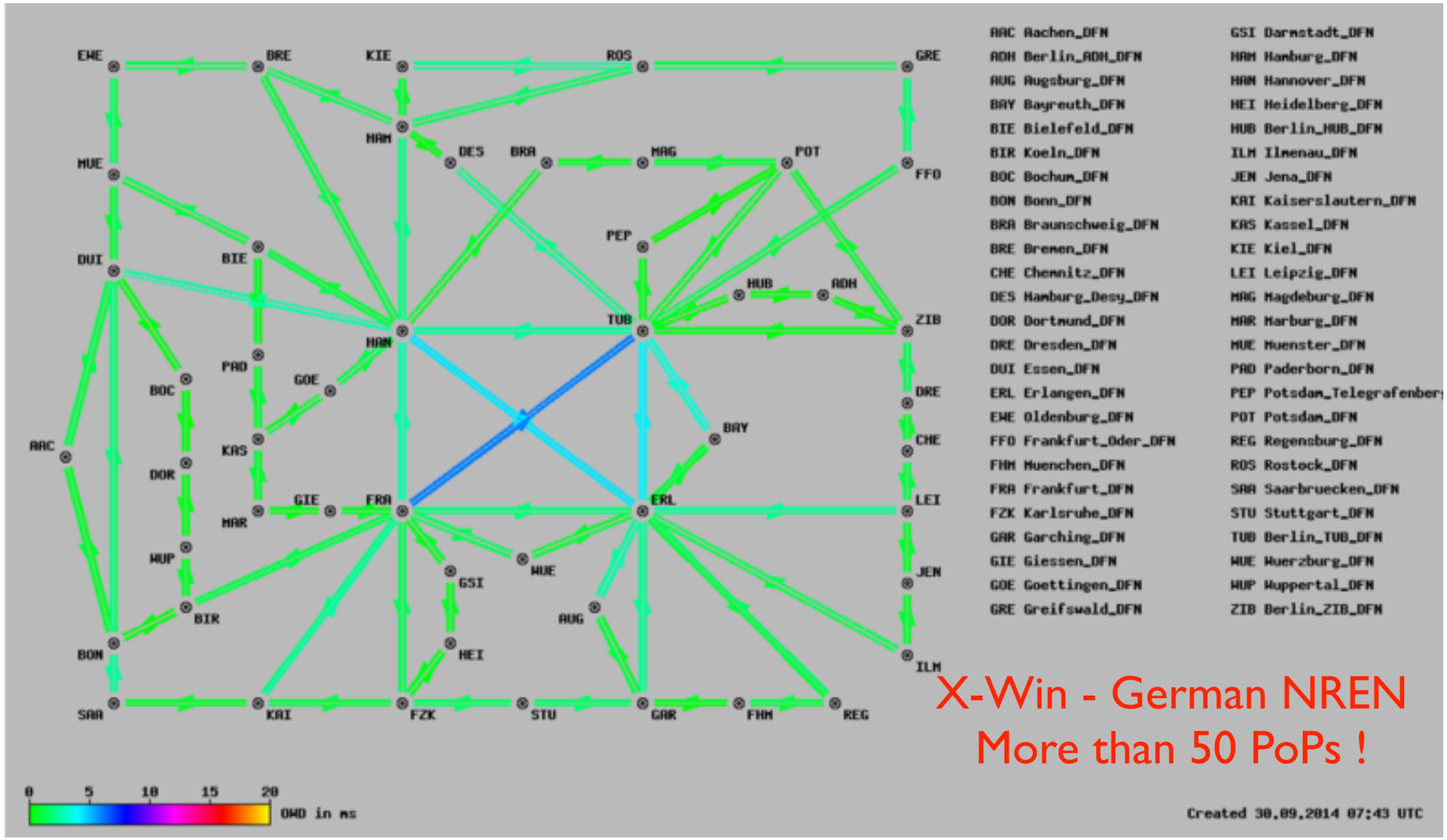


http://www.renater.fr/raccourci?lang=fr

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

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

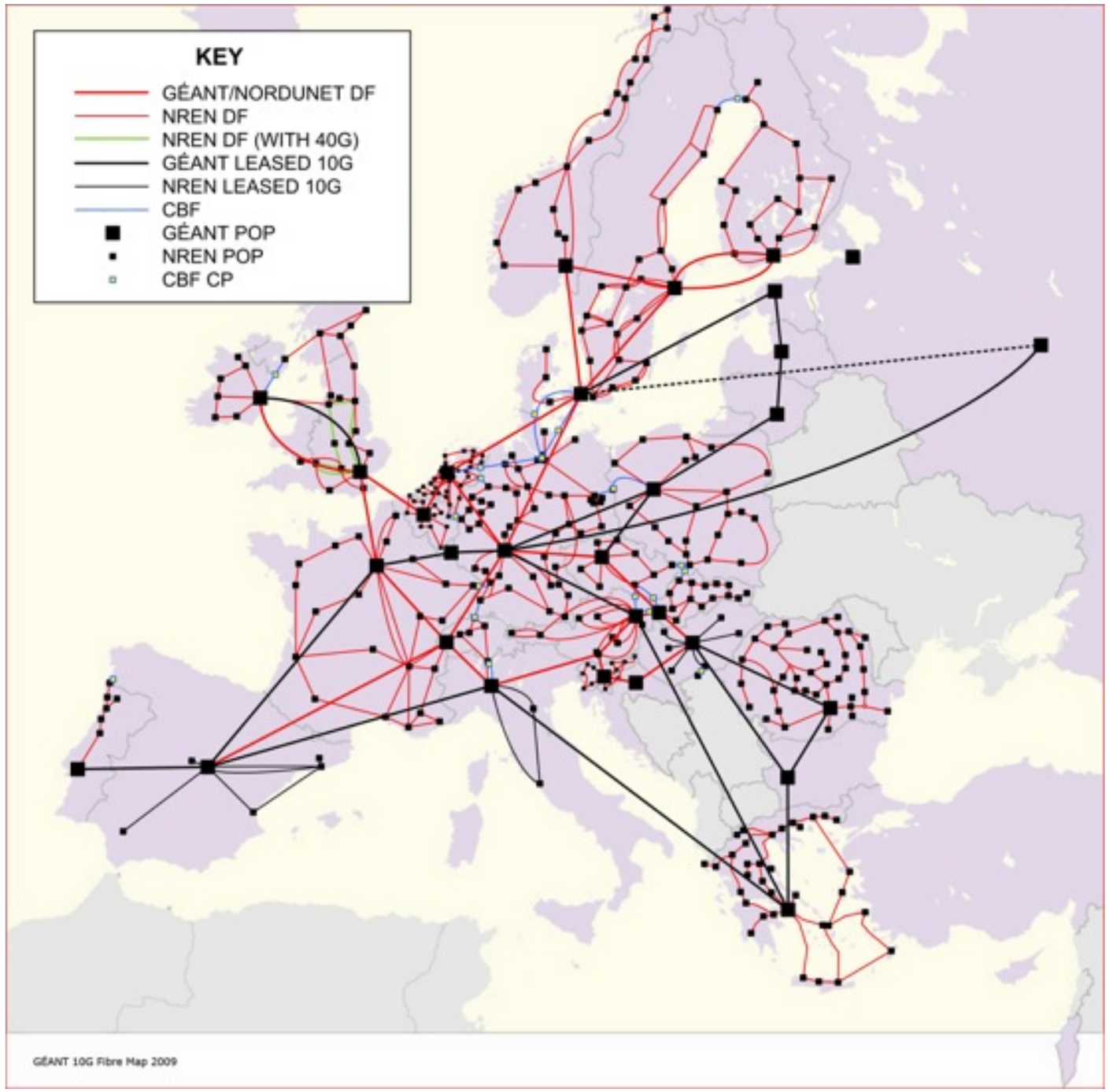


http://www.win-labor.dfn.de/cgi-bin/hades/map.pl?config=win

Beyond the Clouds, the DISCOVERY Initiative

- Locality-based UC infrastructures

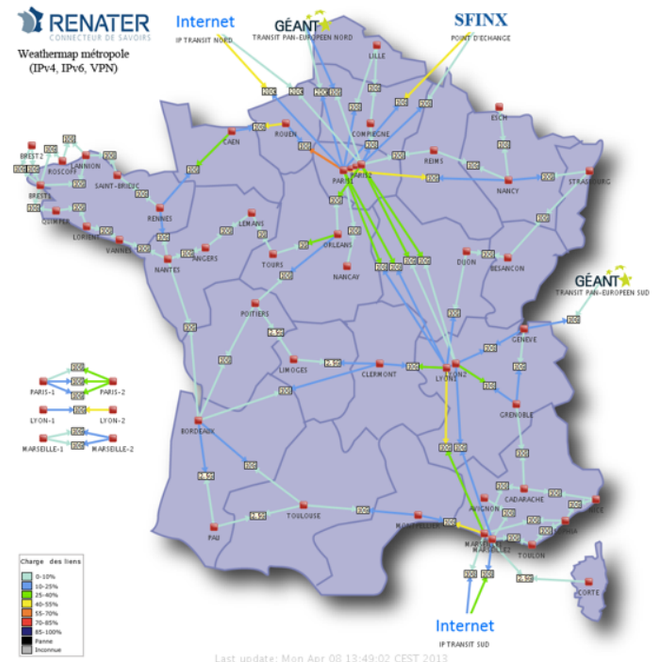
  A promising way to deliver highly efficient and sustainable UC services is to provide UC platforms as close as possible to the end-users.



Beyond the Cloud, the DISCOVERY Initiative

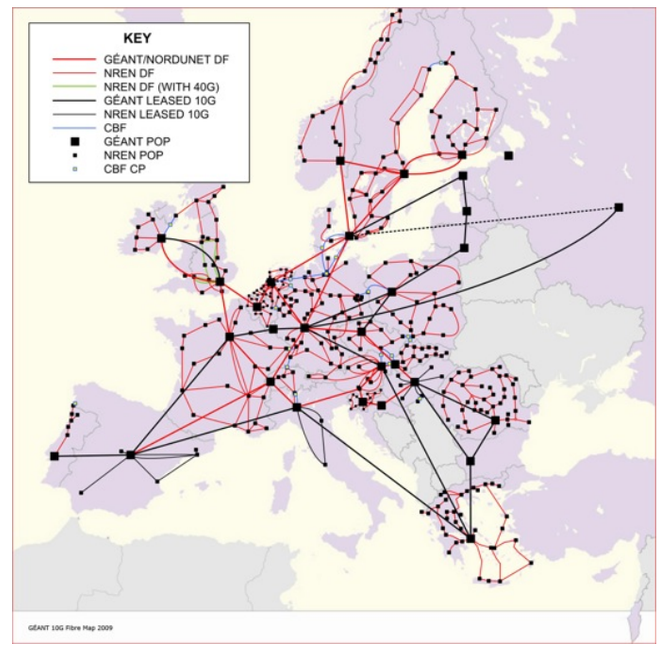
- Locality-based UC infrastructures

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



- Leveraging network backbones

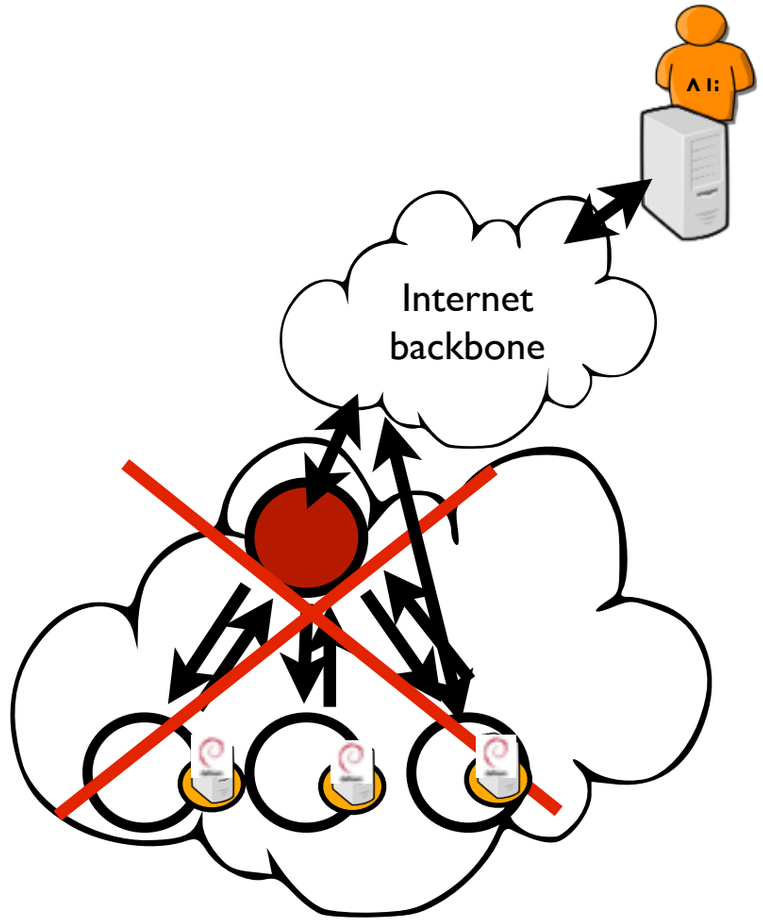
  Extend any point of presence of network backbones with UC servers (from network hubs up to major DSLAMs that are operated by telecom companies and network institutions).



⇒ Operating such widely distributed resources requires the definition of a fully distributed system

The DISCOVERY Proposal

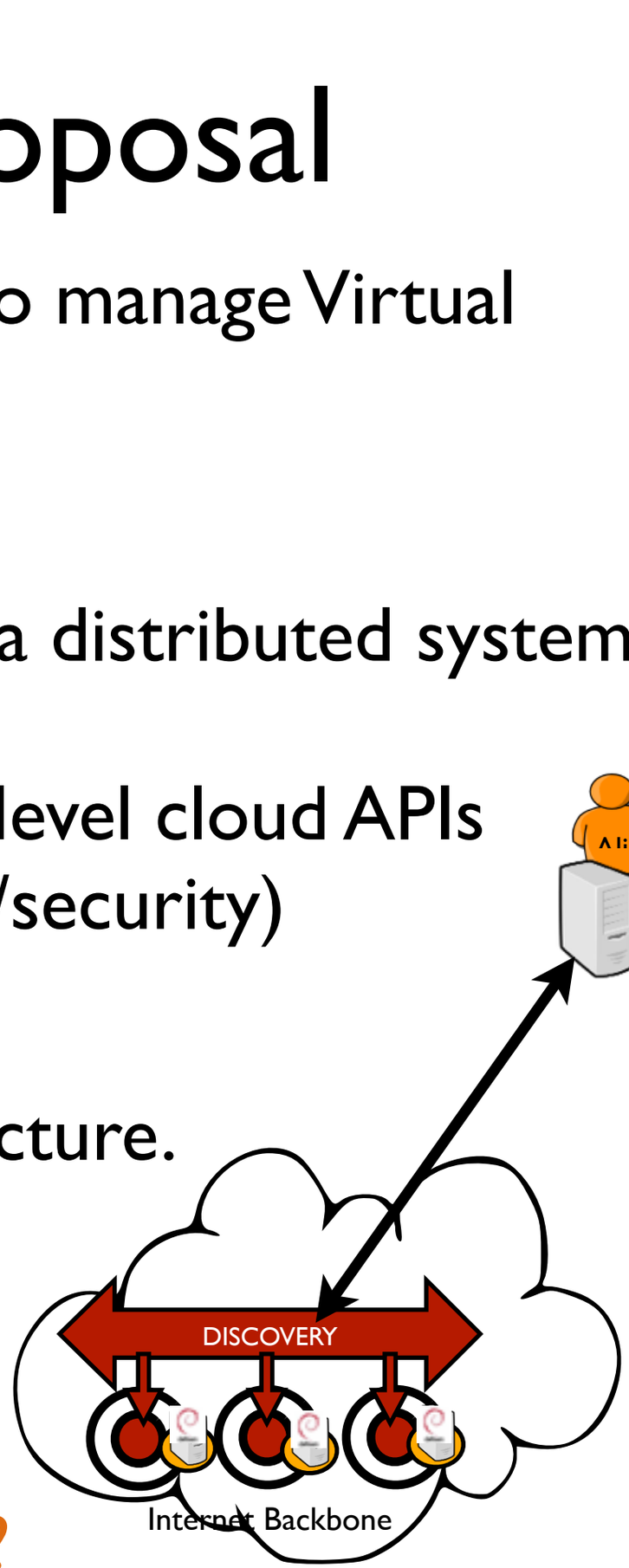
- DIStributed and COoperative framework to manage Virtual EnviRonments autonomously



The DISCOVERY Proposal

- DIStributed and COoperative framework to manage Virtual EnviRonments autonomously
- The LUC OS
  - A fully distributed IaaS system and not a distributed system of IaaS systemS
    We want to/must go further than high level cloud APIs (cross-cutting concerns such as energy/security)
  - Leverage P2P algorithms and self-* approaches to operate a LUC infrastructure.

?? A distributed version of the EGI Core that directly manipulates resources http://www.egi.eu/infrastructure/cloud/ ??



Where We Try To Go (few details)

- The LUC OS

  Based on VMs and VEs (group of VMs) as the fundamental granularity

  Scalability, targeting the management of hundred thousands of VMs upon thousands of physical machines spread throughout hundreds of sites
  Reliability, considering "hardware failures as the norm rather the exception" (but this is not a BitTorent system !)
  Reactivity, handling each reconfiguration event as swiftly as possible to maintain VEs' QoS.

- **lots of scientific/technical challenges**

  Cost of the DISCOVERY network !? partial view of the system !?
  Impact on the others VMs !?, management of VM images !?
  Which software abstractions to make the development easier and more reliable (distributed event programming)?
  How to take into account locality aspects ?

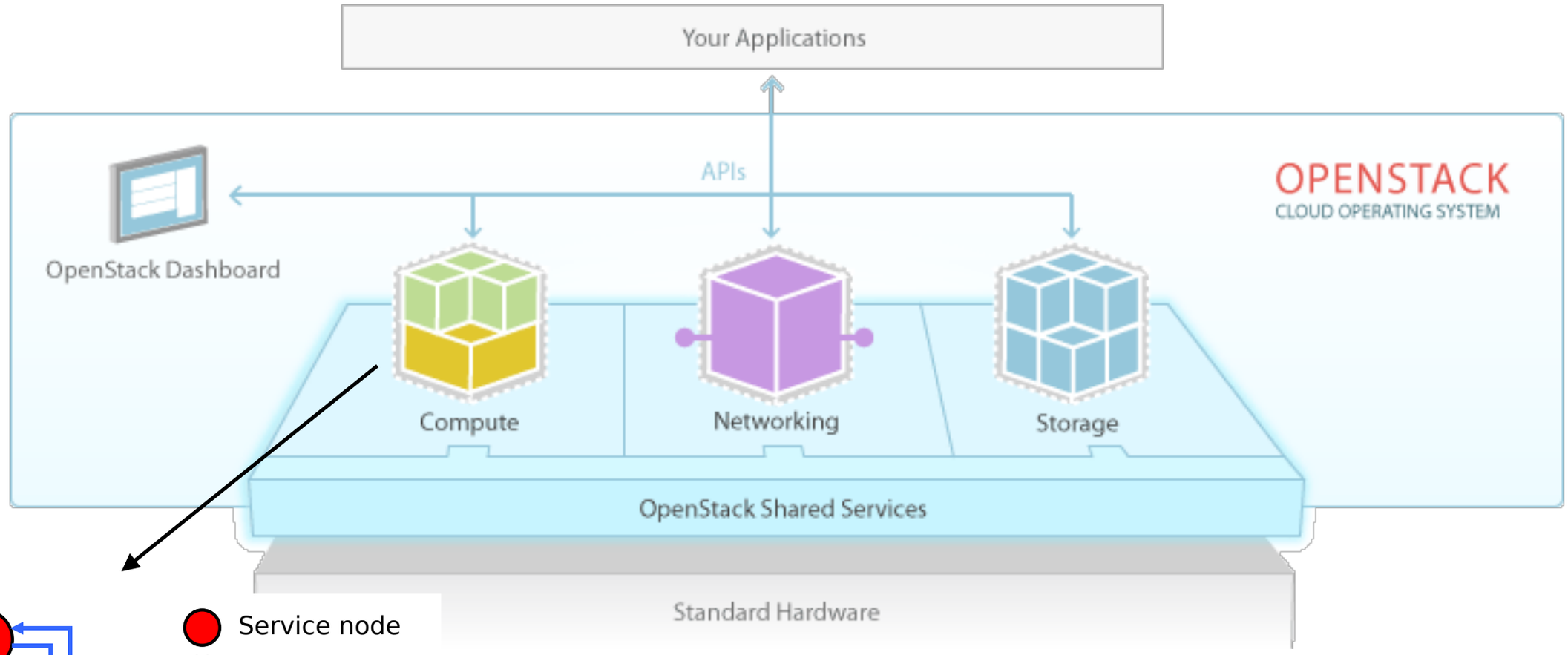
Where We Are

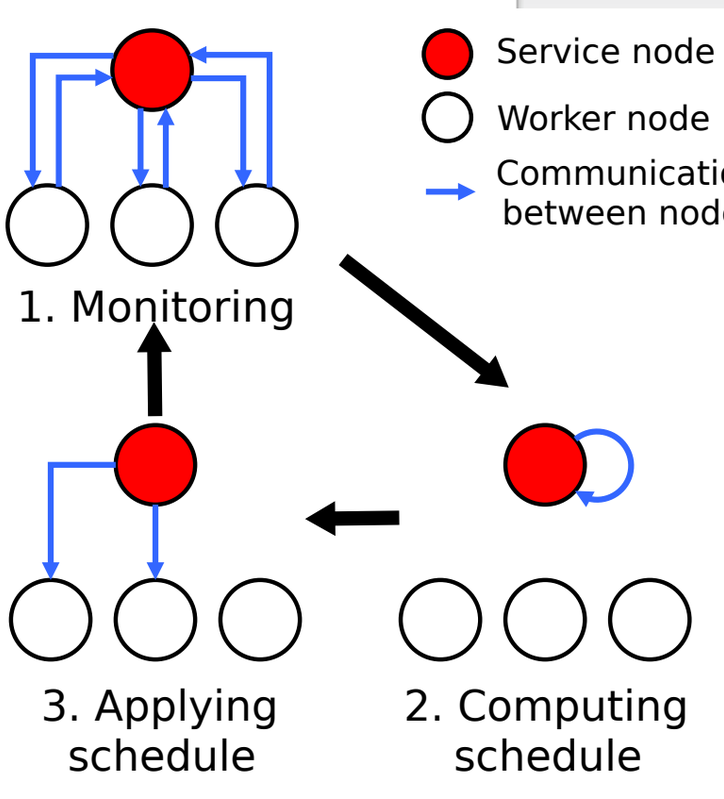
- Validation of the LUC model
  (on-going work with RENATER, the French NREN)
  - From energy/efficiency/economical point of views
  - On a brick basis (100 VMs) and by considering the cost of the network.
- An academic POC for validating the feasibility of major blocks (scheduling of VMs, migration between distinct sites…)
  - Two PhDs, Two PostDocs
  - Managing 10K VMs on top of Grid'5000 like normal processes on a laptop.
- A POC is nice but can we push this idea further ?
  - Making a complete system is a huge/non sense effort for researchers
    ⇒ Revisit OpenStack (on-going work, started 9 months ago),

Where We Are

- Validation of the LUC model
  (on-going work with RENATER, the French NREN)
  - From energy/efficiency/economical point of views
  - On a brick basis (100 VMs) and by considering the cost of the network.
- An academic POC for validating the feasibility of major blocks (scheduling of VMs, migration between distinct sites…)
  - Two PhDs, Two PostDocs
  - Managing 10K VMs on top of Grid'5000 like normal processes on a laptop.
- A POC is nice but can we push this idea further ?
  - Making a complete system is a huge/non sense effort for researchers
    ⇒ Revisit OpenStack (on-going work, started 9 months ago),

Focus on Compute and the scheduling challenge

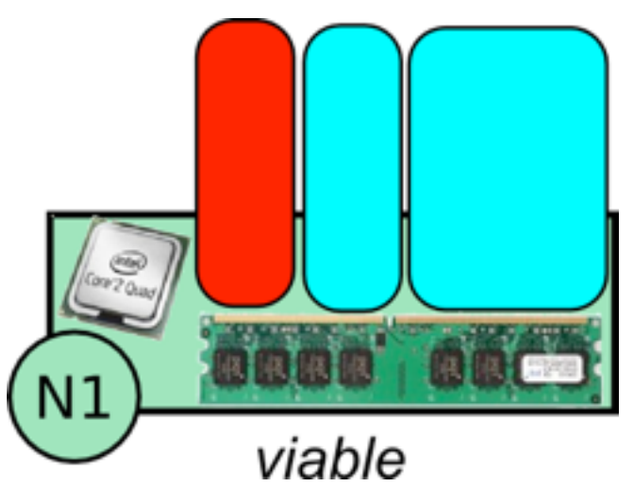




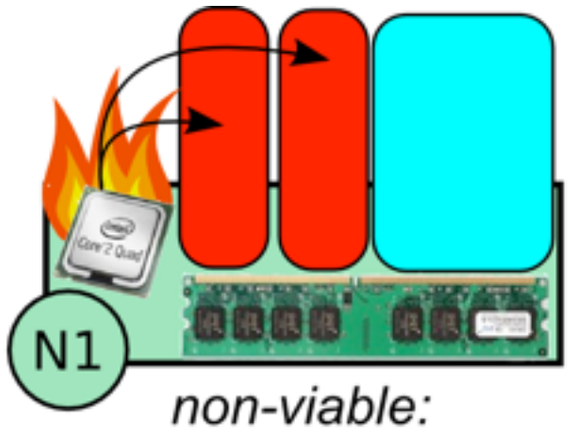
- The main objective of Compute (aka Nova) is to create/configure and schedule VMs in order to satisfy CPU, memory, … requirements while maximising the resource usage of the infrastructure.

Viable and Non Viable Configurations

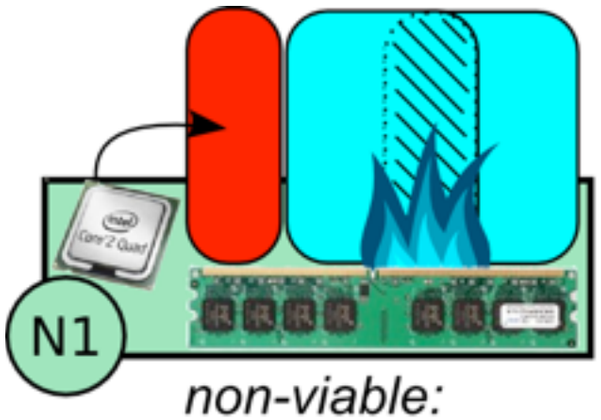
- Fine management of resources (efficiency and energy constraints)
- Find the "right" mapping between needs of VMs and resources provided by PMs



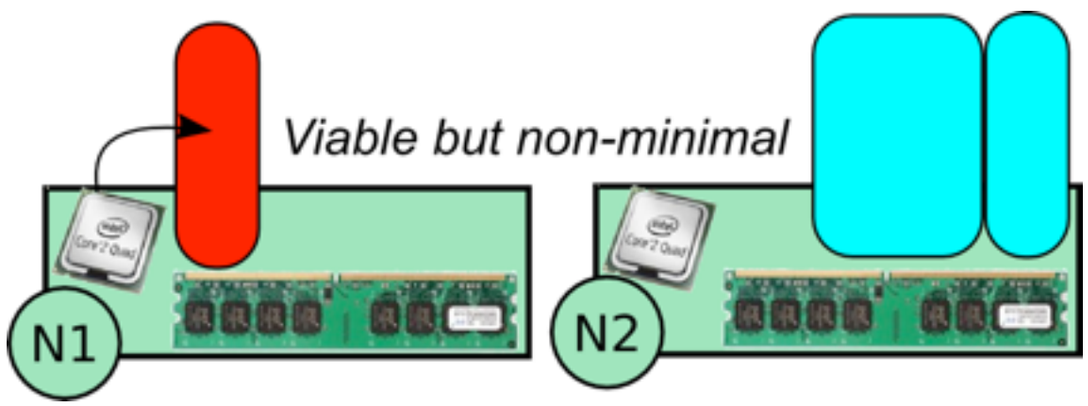
viable

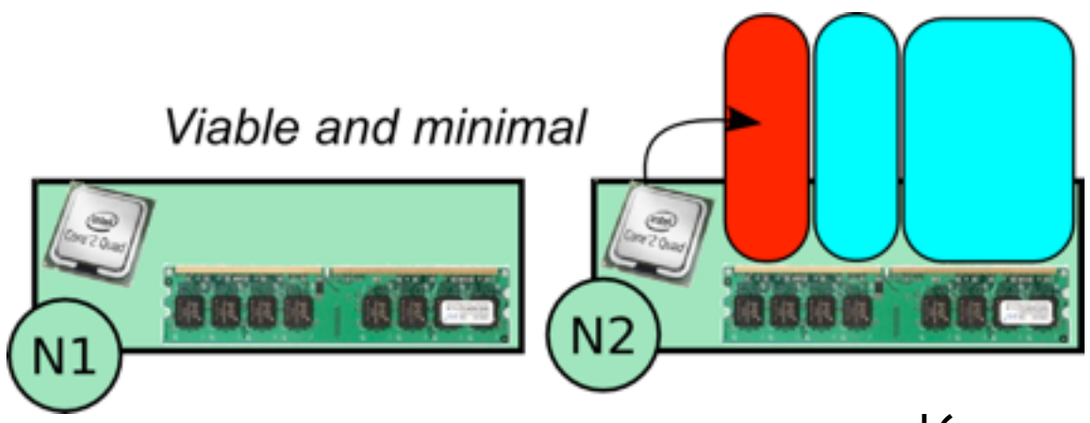


non-viable:
2 active VMs for one CPU



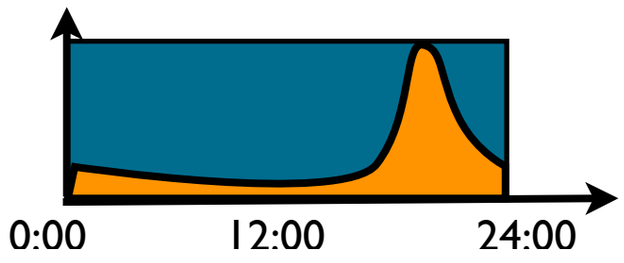
non-viable:
memory overcommitment



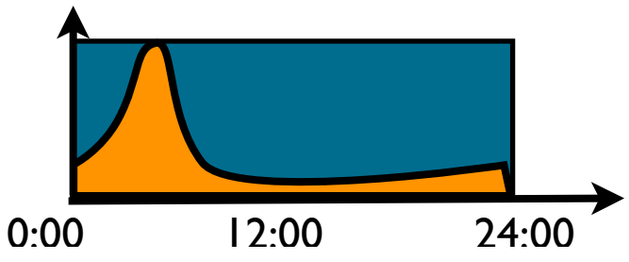


credits: F. Hermenier, OSDI poster session 2008

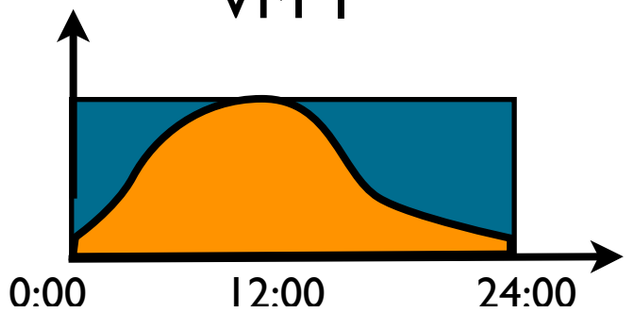
Fluctuations of VM Requirements



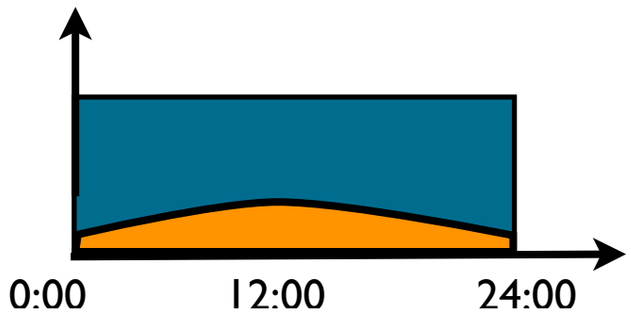
VM 1



VM 2



VM 3



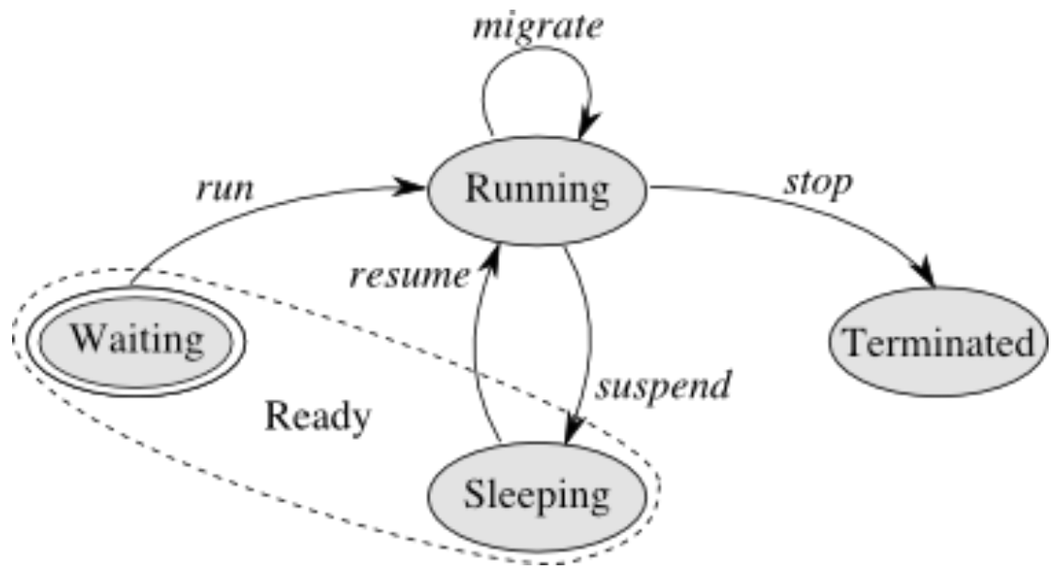
VM 4

- Static placement policies (as delivered by most of the popular Cloud Computing management systems)
  ⇒ Simple but prevent CC providers to maximize the usage of CC resources while guaranteeing VM resource requirements

- Used advanced dynamic placement strategies to relocate VMs according to the scheduler objectives / available resources / waiting queue / ...

Dynamic VM Placement Policies

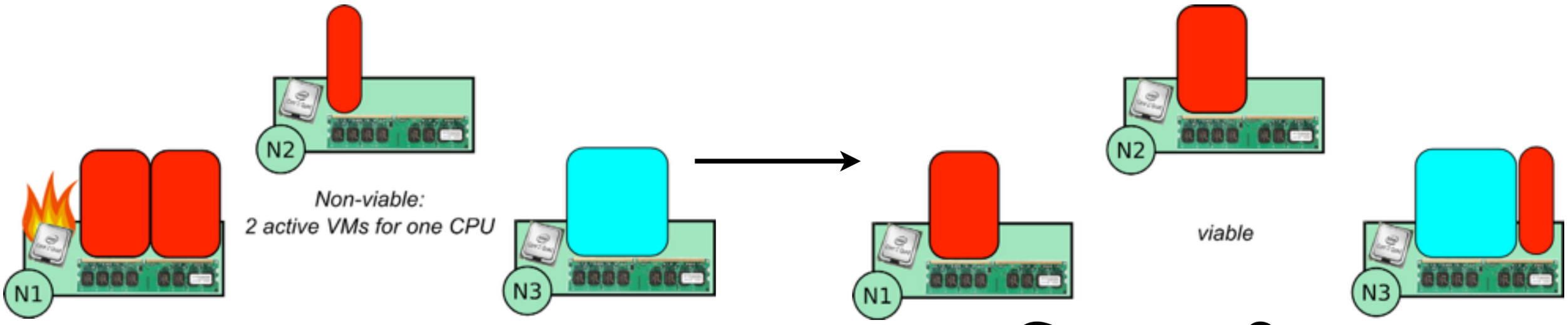
- Generale idea: leverage VM capabilities to manipulate VEs in a similar way of usual processes on a laptop (a VE is a users' working environment, possibly composed of several interconnected VMs)



- Each VE is in a particular state

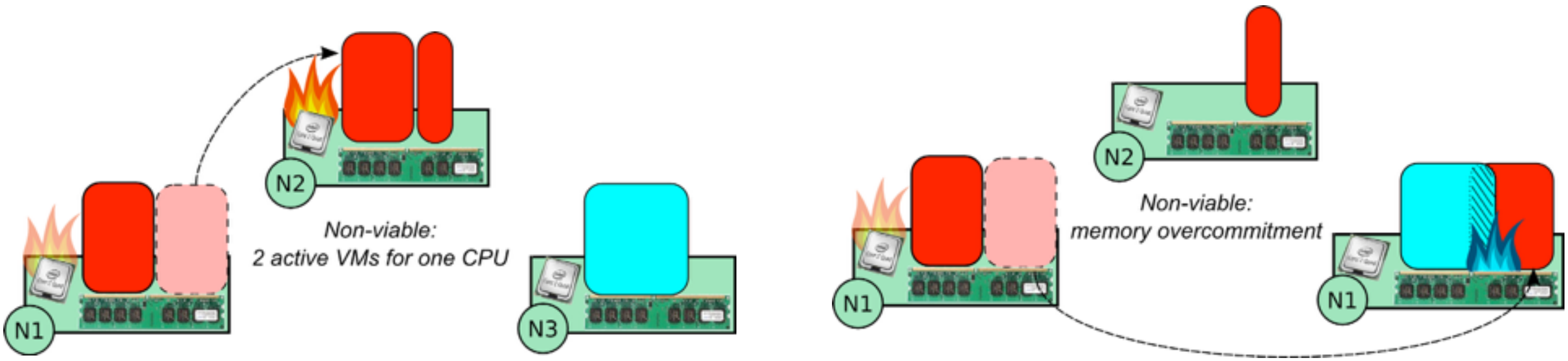
- Perform VE context switches (a set of VM context switches) to reschedule/rebalance the LUC infrastructure

Viable and Non Viable Manipulations



Current Status

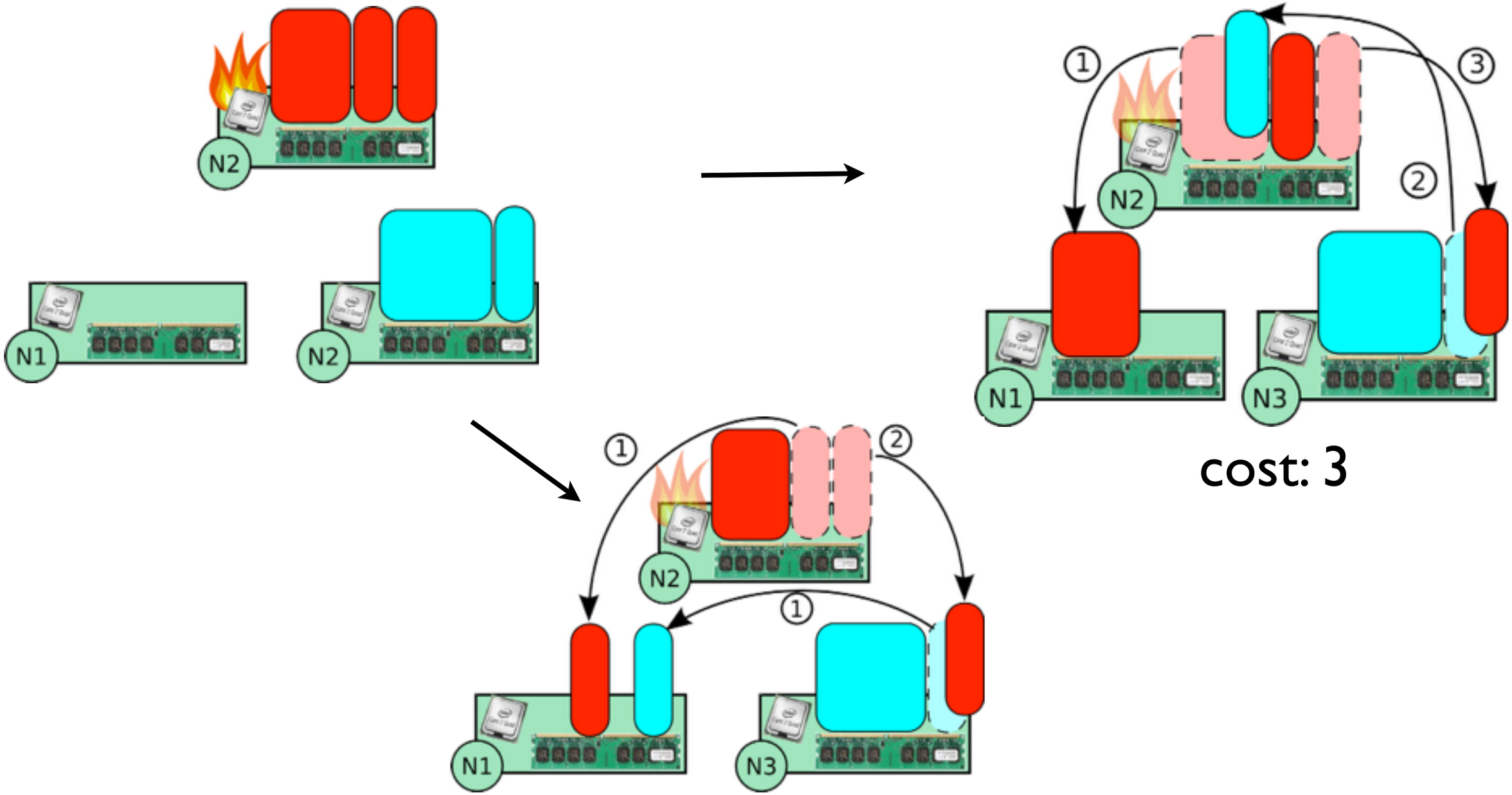
Correct Status



Non-viable manipulations

credits: F. Hermenier, OSDI poster session 2008

Optimizing The VE Context Switch



credits: F. Hermenier, OSDI poster session 2008

Entropy / btrPlace

- An autonomic framework to maintain viable VE placements

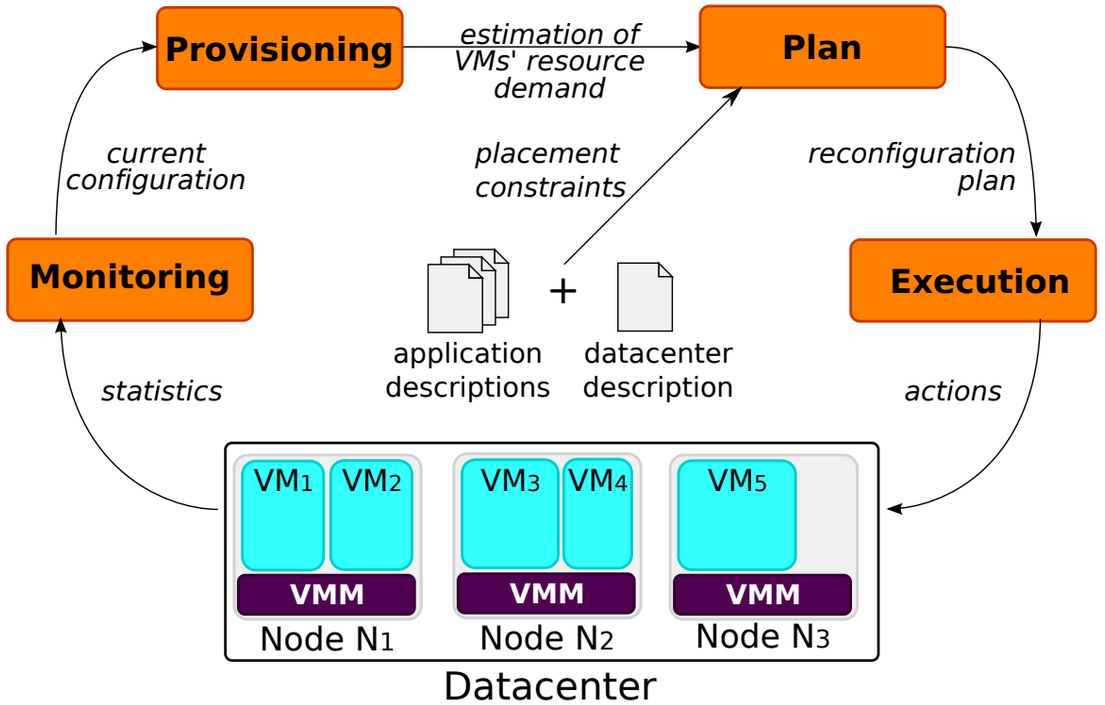
ASCOLA Research Group (ANR SelfXL/Emergence, EasyVirt)
http://www.btrcloud.org/

Oasis/Scale Research Group (University Of Nice Sophia Antipolis)
http://btrp.inria.fr

Additional placement constraints

Autonomic but…

….Centralized



credits: F. Hermenier, Plasma Control Loop, Feb 2011

Revisiting Entropy/BtrPlace for Discovery

- Cooperation between direct neighbours to solve events

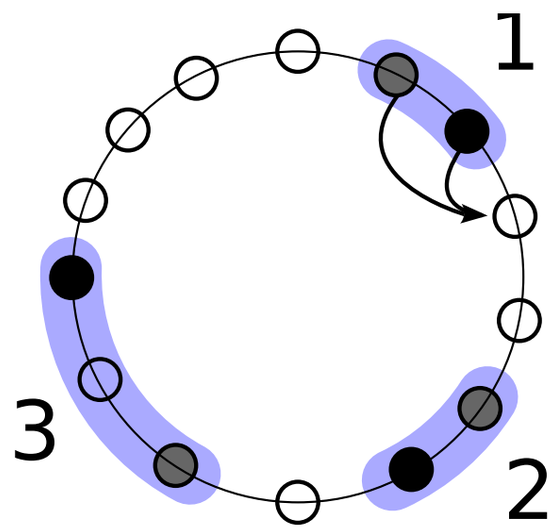
  Event driven

  Peer to peer, no service node

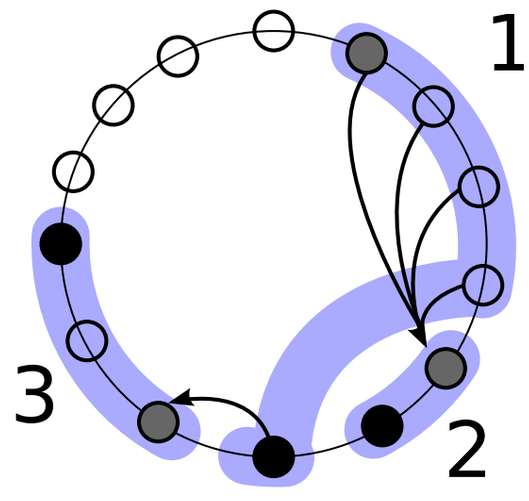
  Local interactions between nodes

  Nodes have a partial view of the system

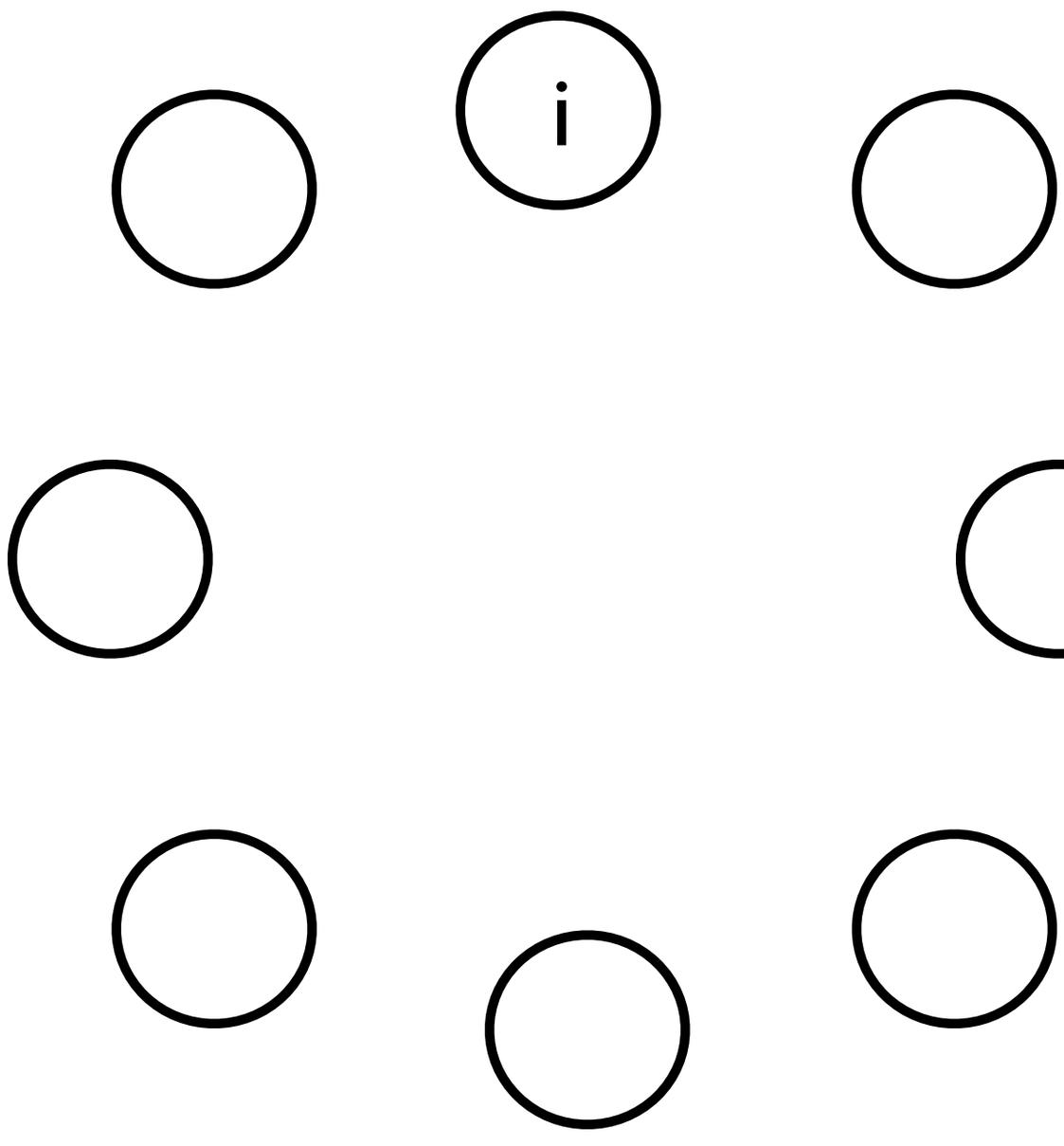
  Local invocation of the resolution algorithm



credits: F. Quesnel et al., DVMS April 2012

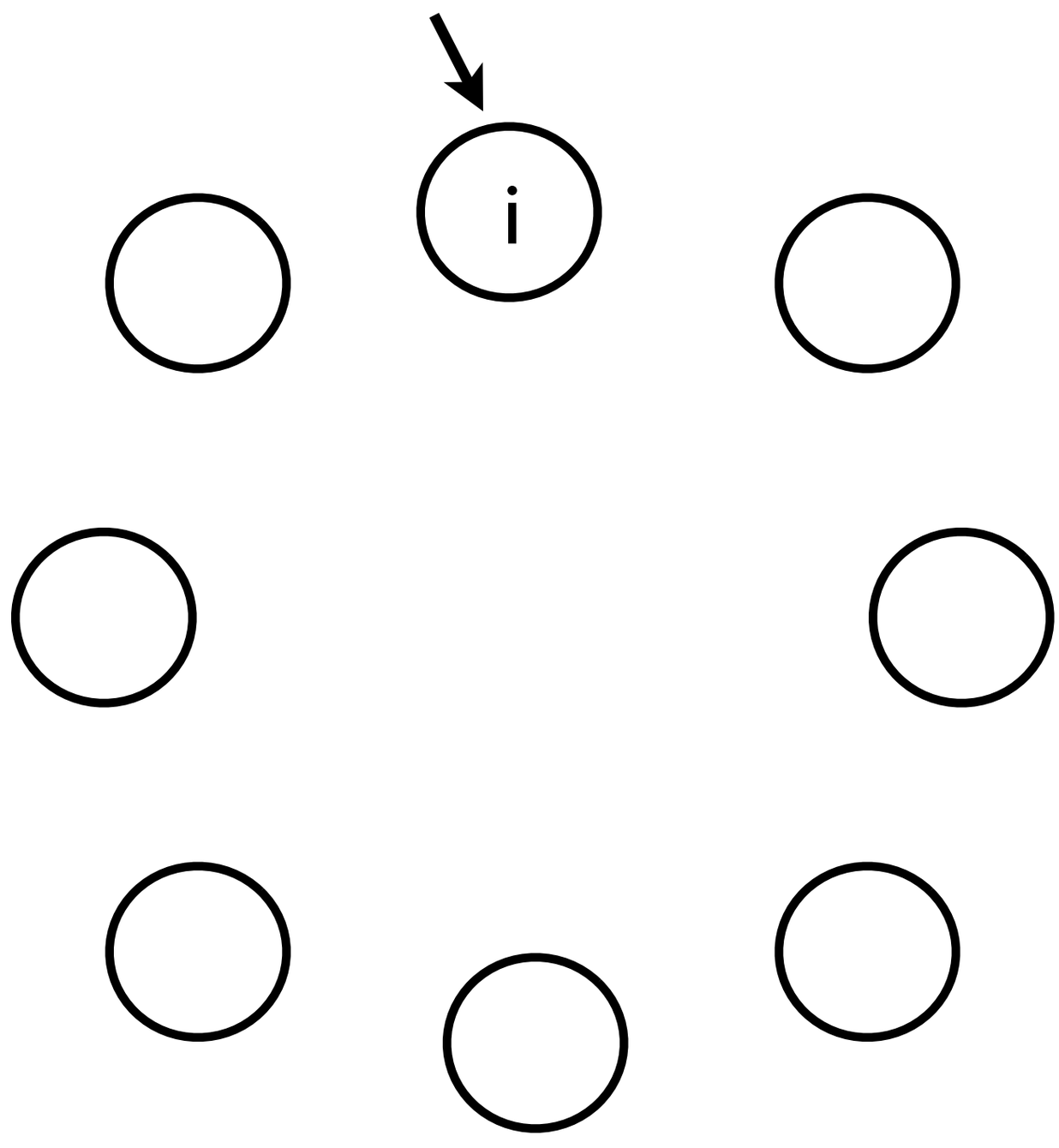


Revisiting Entropy/BtrPlace for Discovery



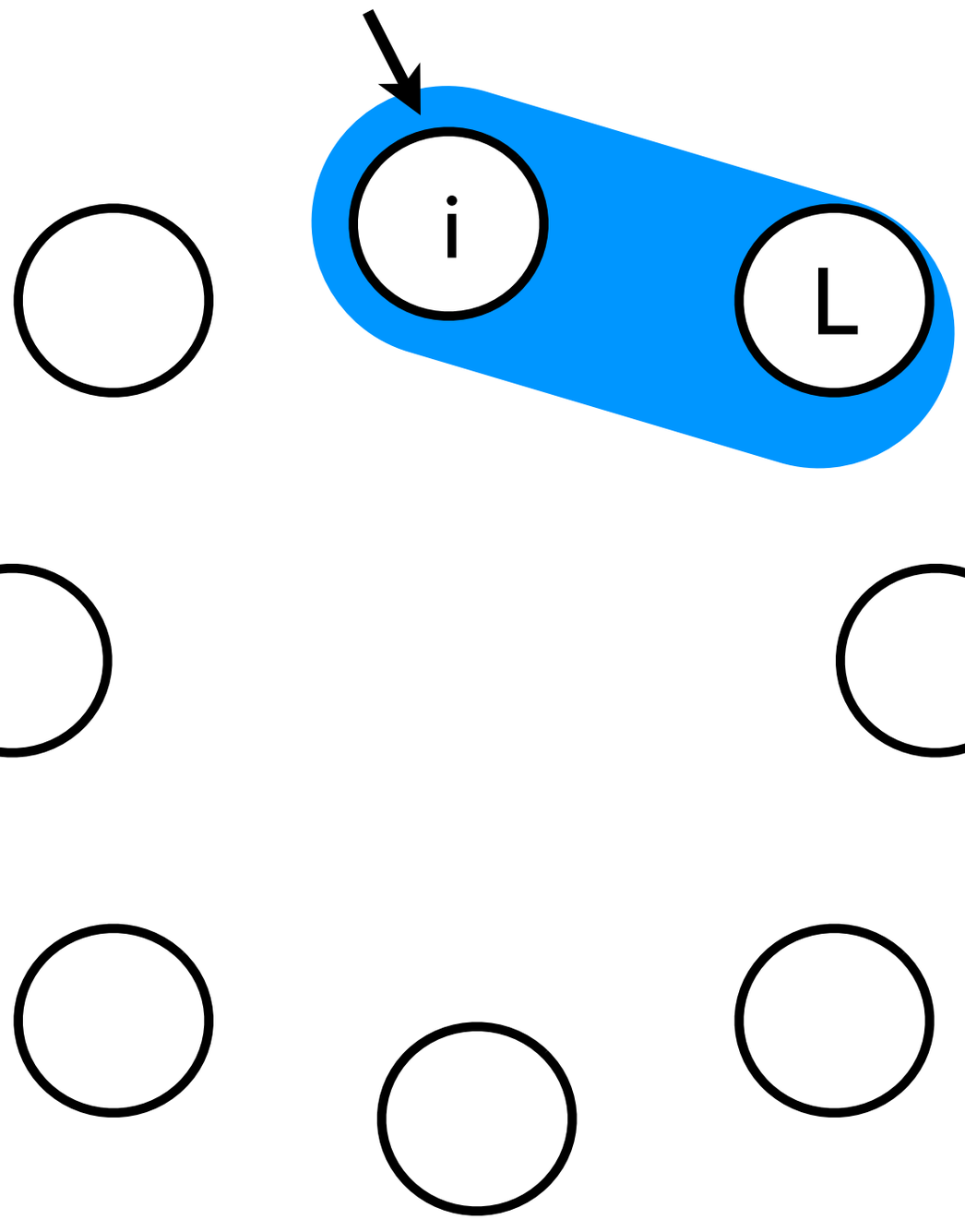
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



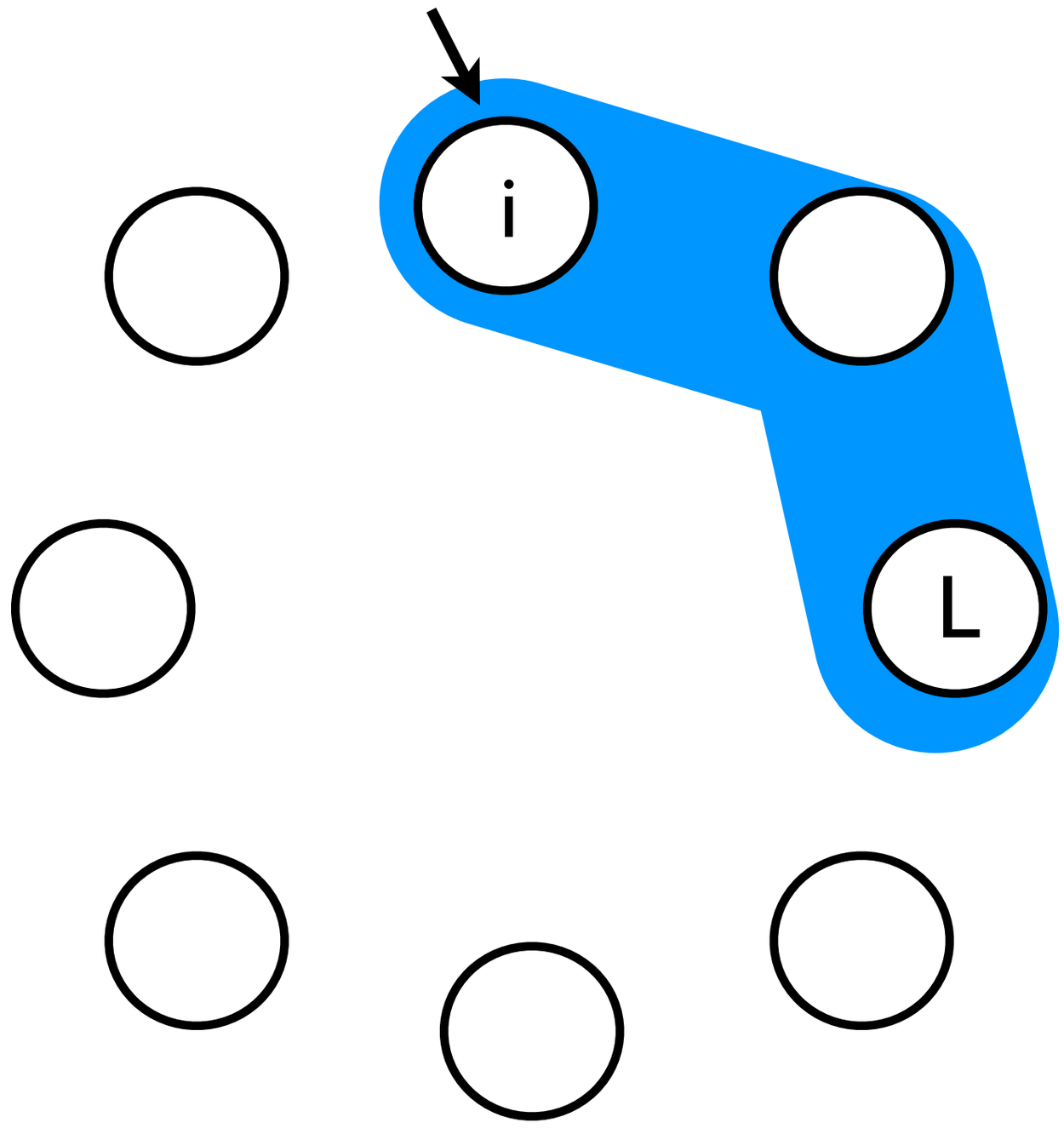
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



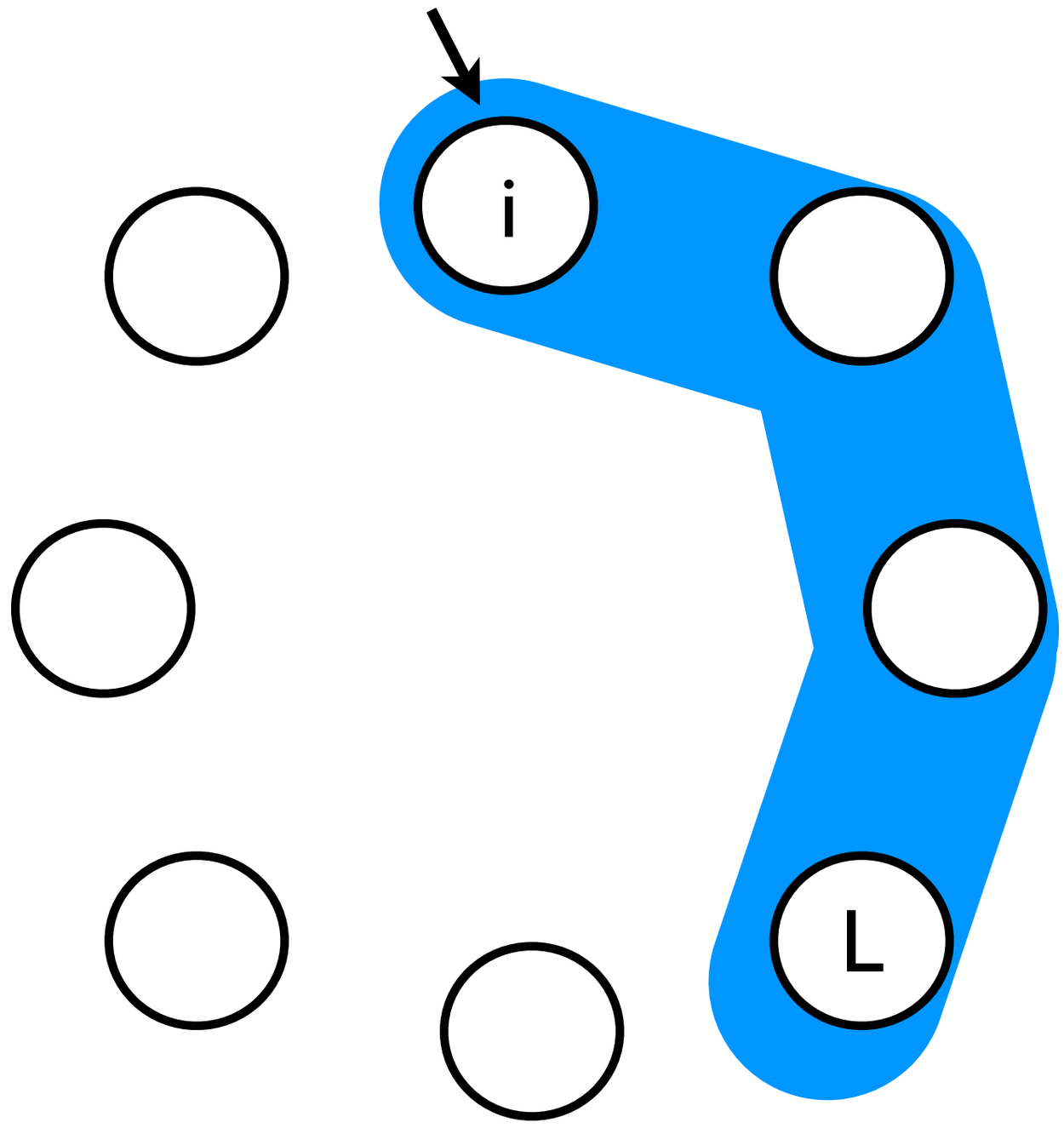
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



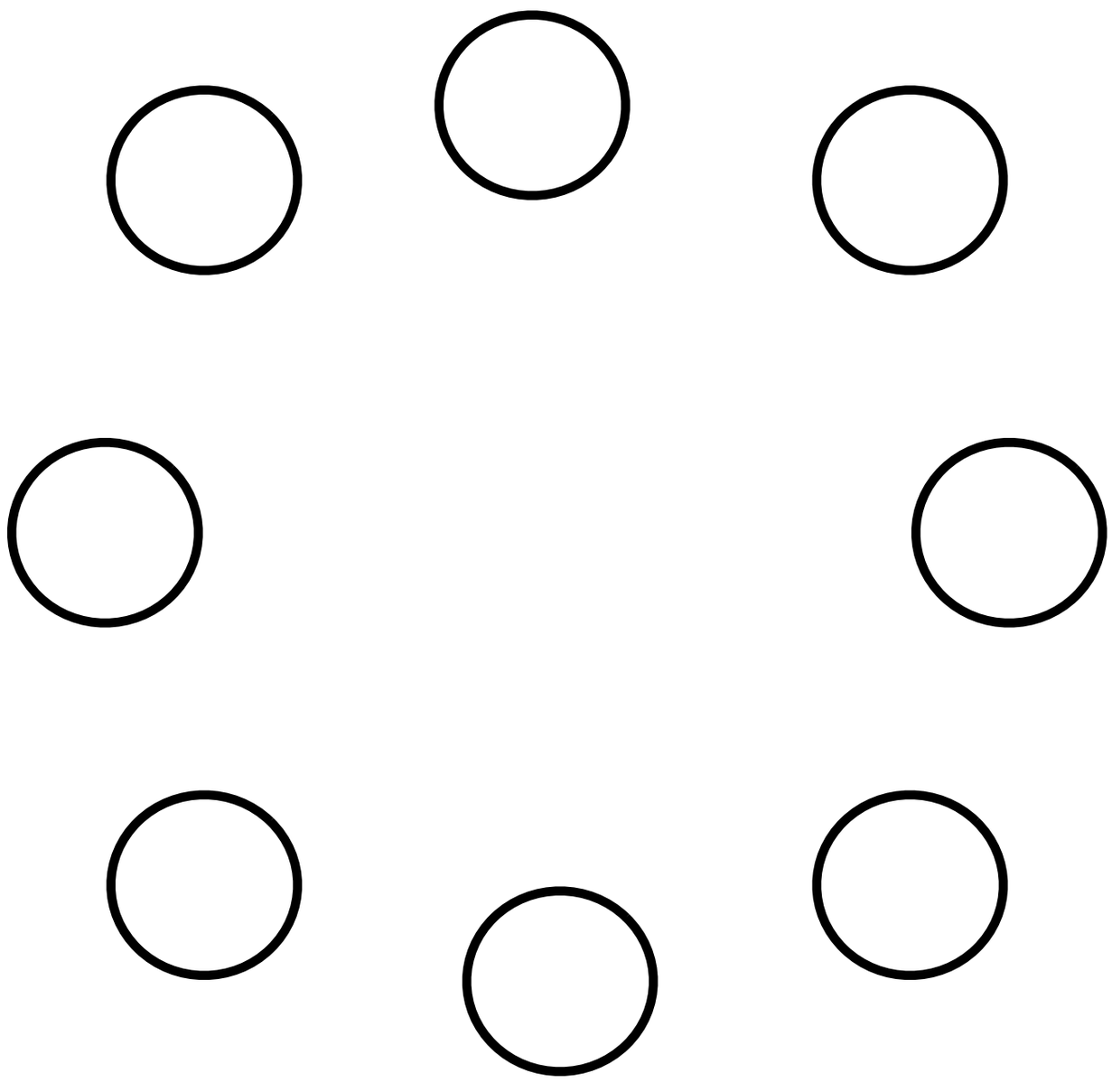
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



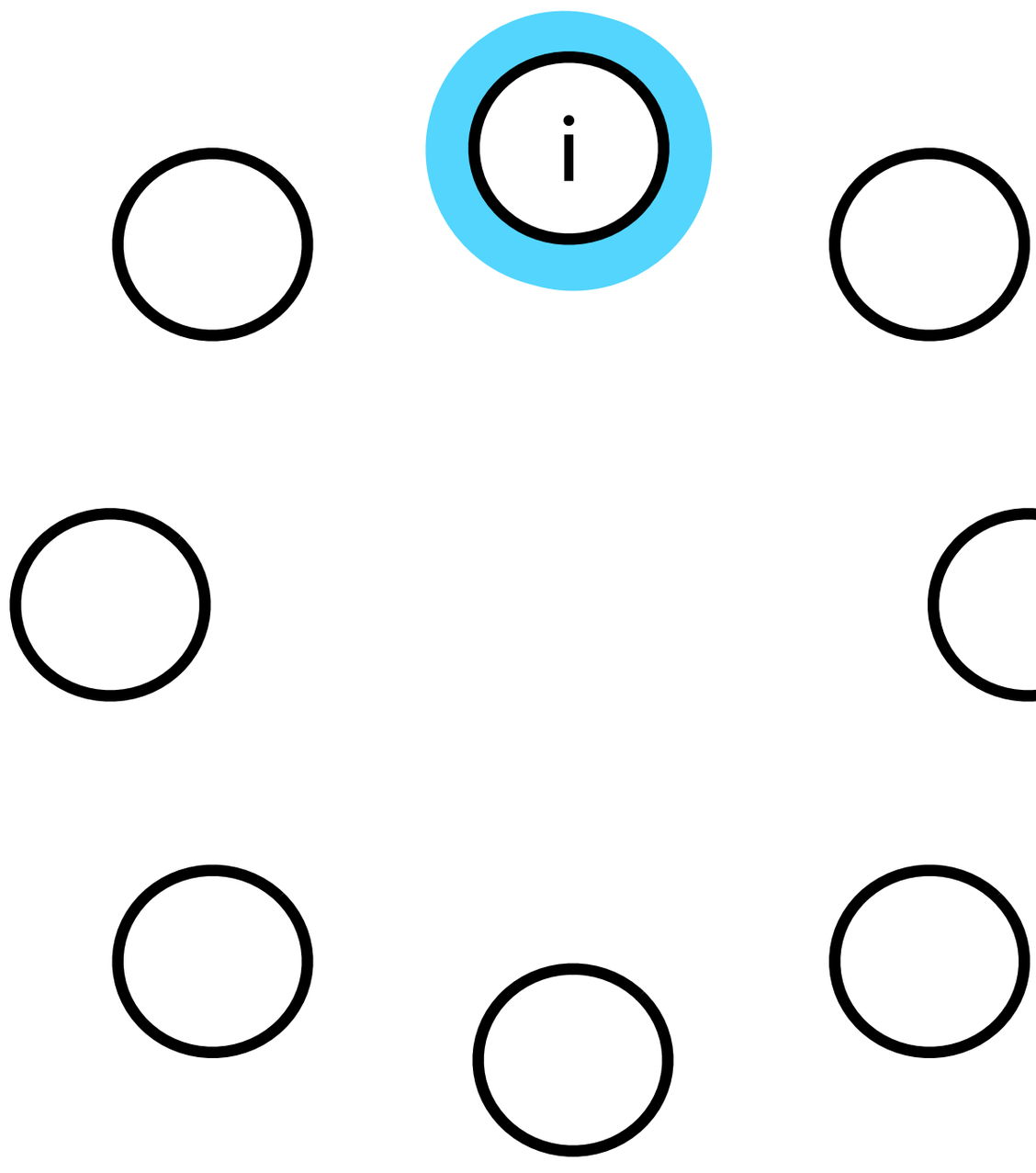
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



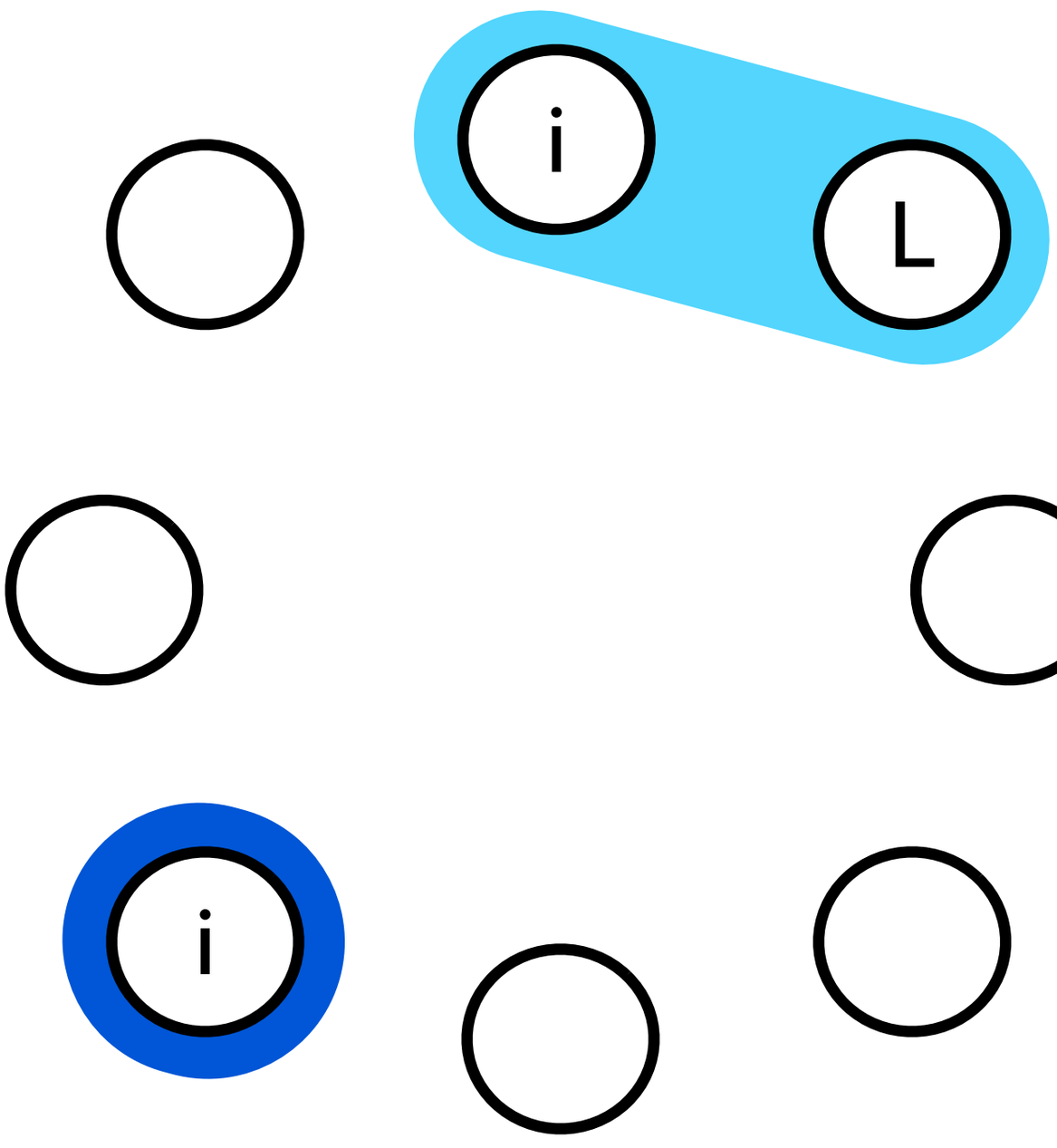
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



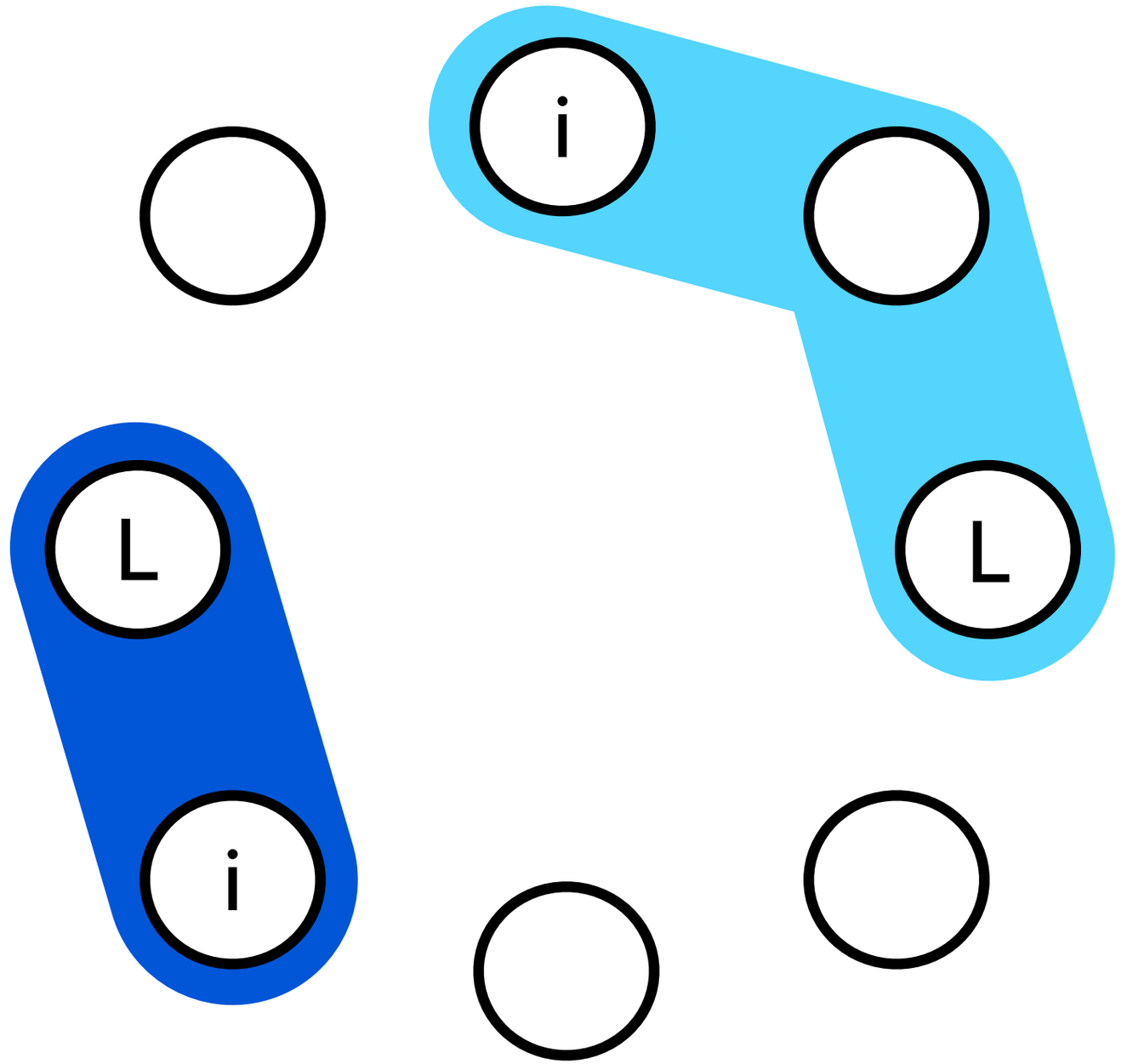
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



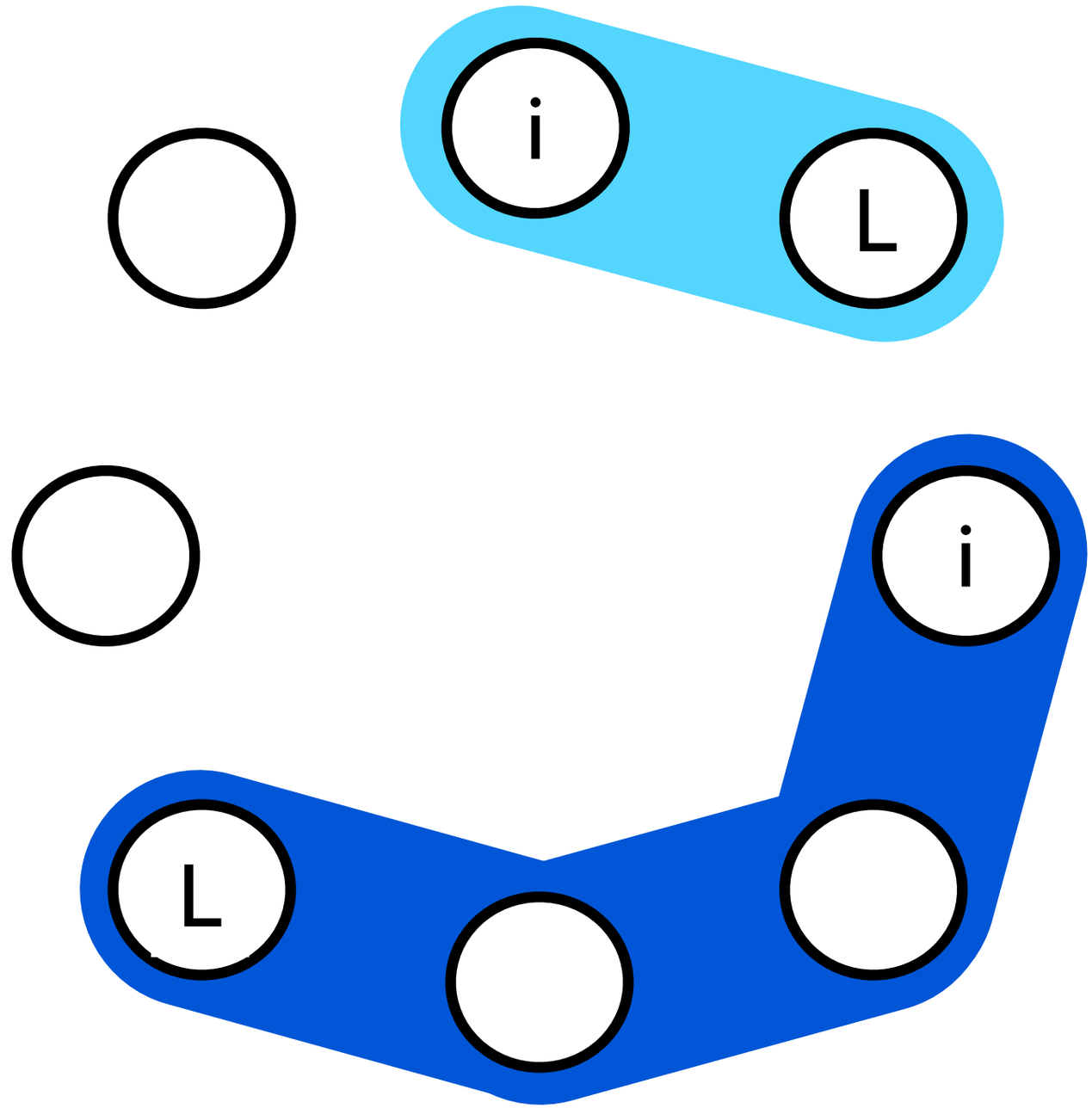
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



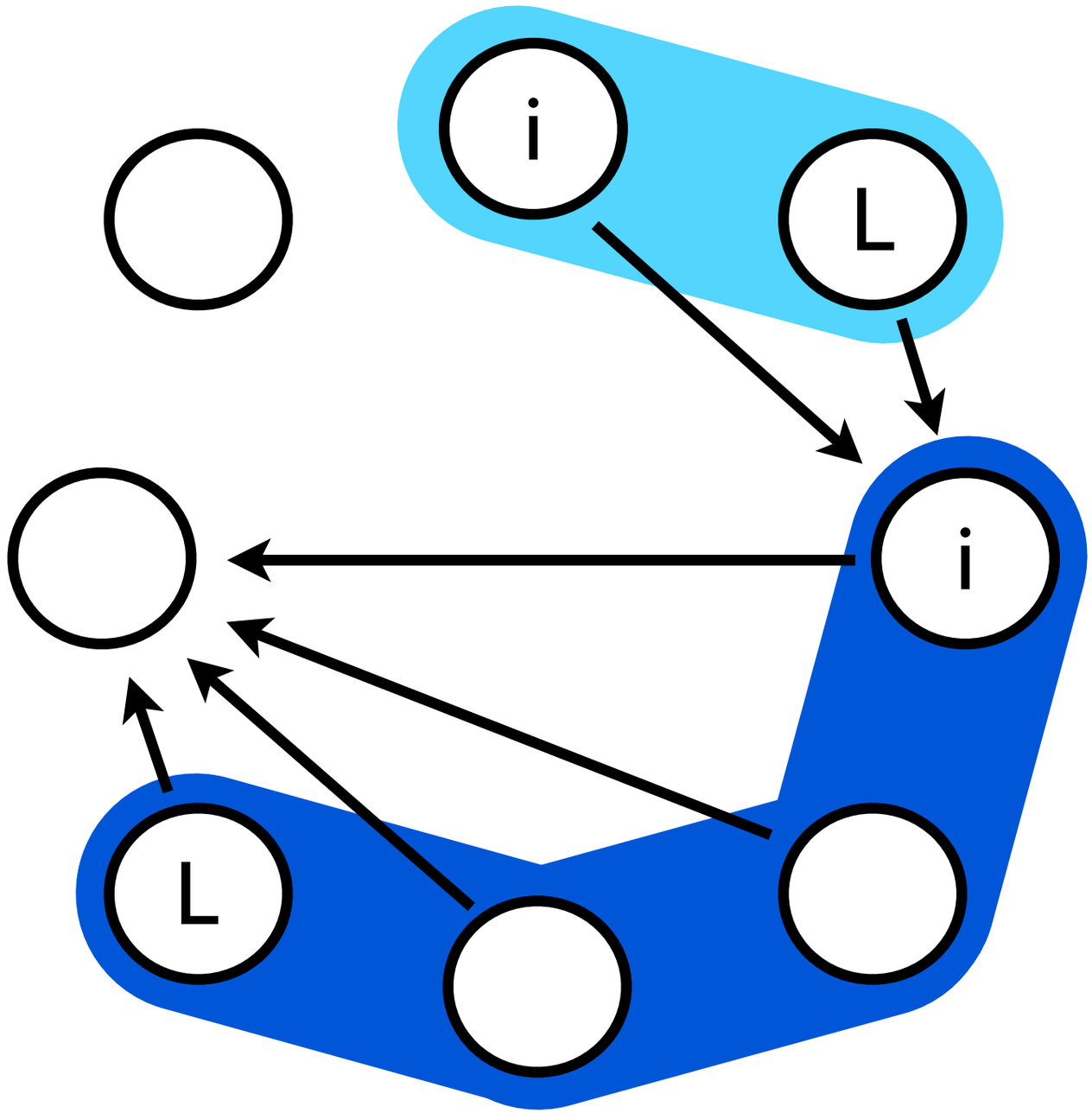
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



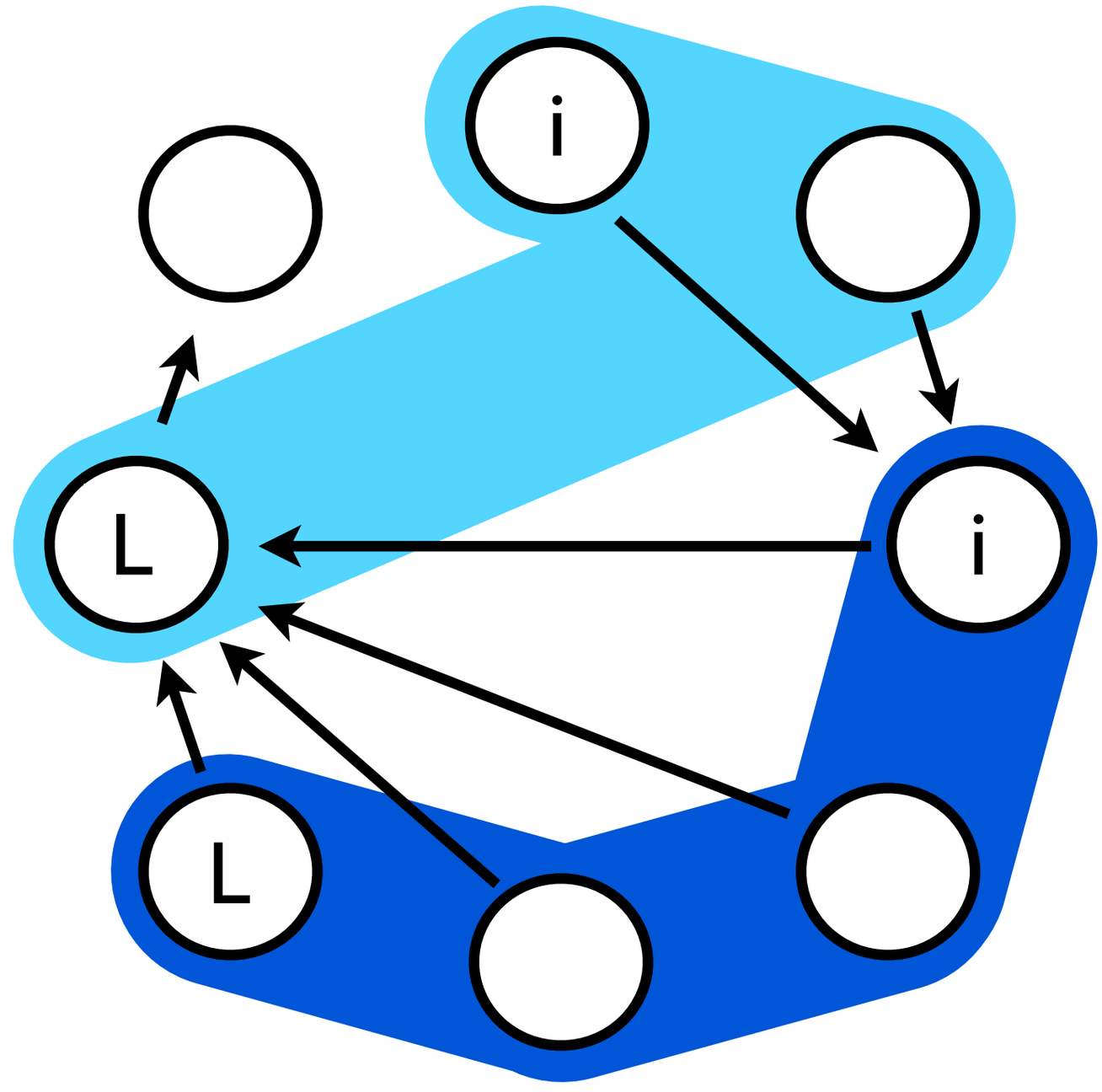
The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



The DVMS proposal

Revisiting Entropy/BtrPlace for Discovery



The DVMS proposal

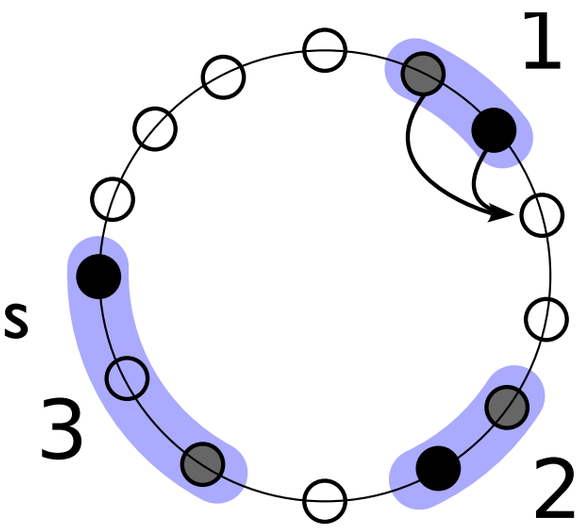
Revisiting Entropy/BtrPlace for Discovery

- Cooperation between direct neighbours to solve events

Nodes have a local view of the system
Local invocation of the resolution algorithm

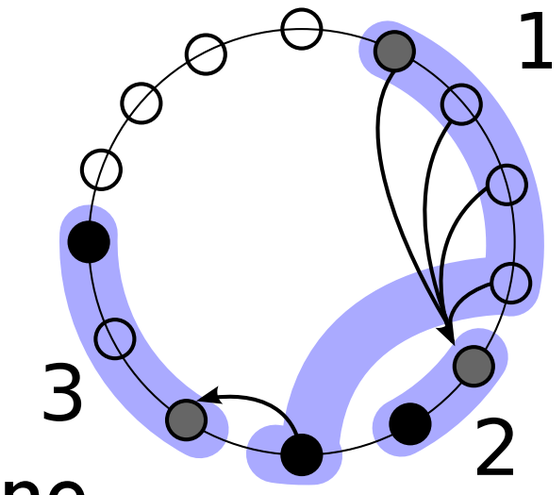
In vivo (on Grid5000) 500 physical machines, 4500 VMs

Simulation (using Simgrid) 10K PMs, 80K VMs



http://beyondtheclouds.github.io/DVMS/

credits: F. Quesnel et al., DVMS April 2012

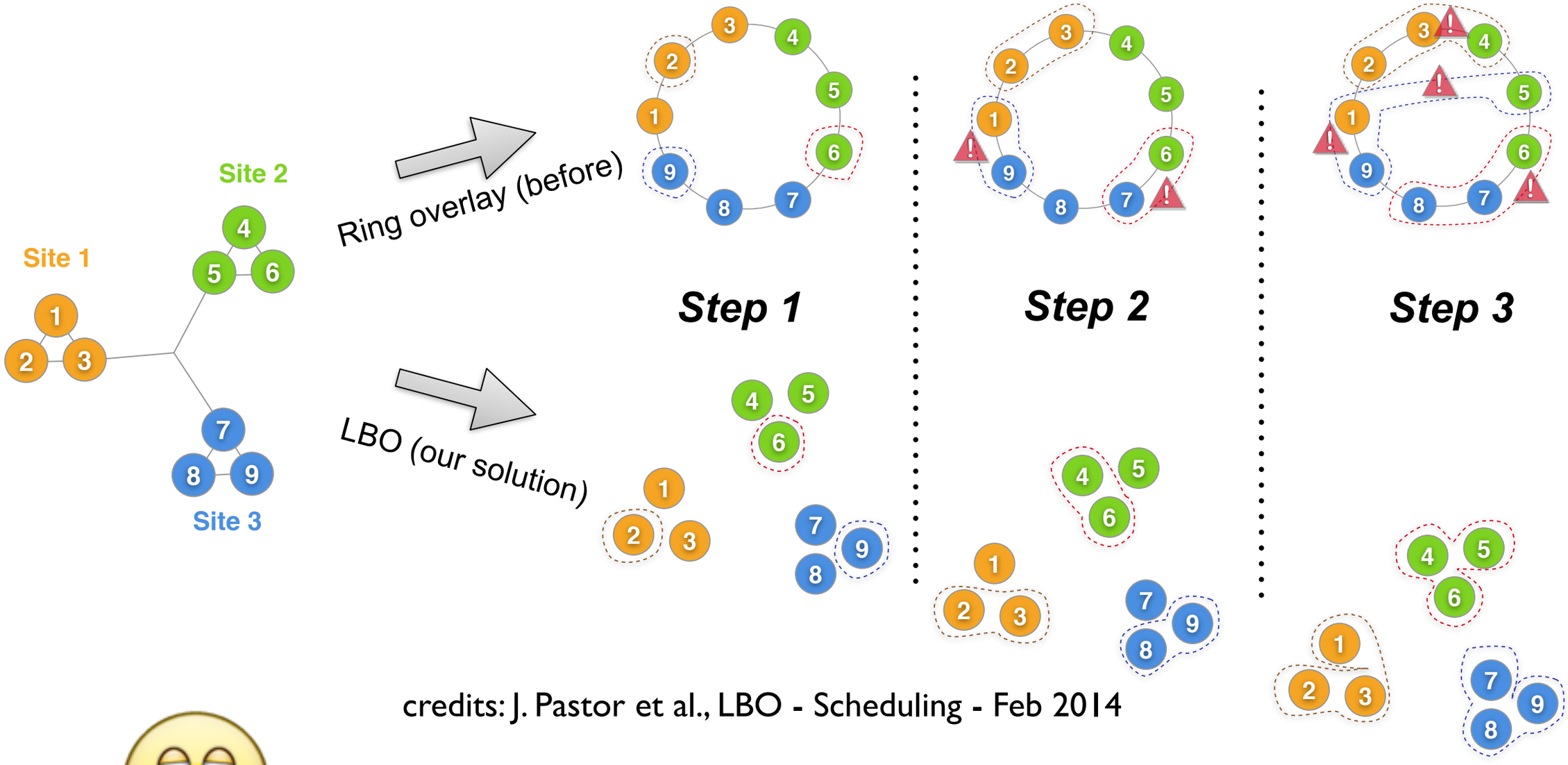


Scalability/reactivity but….

…matching a ring on a real network backbone

Distributed and Locality-aware

- Leverage a locality based overlay (vivaldi) + a shortest path algorithm to favour cooperations between close nodes

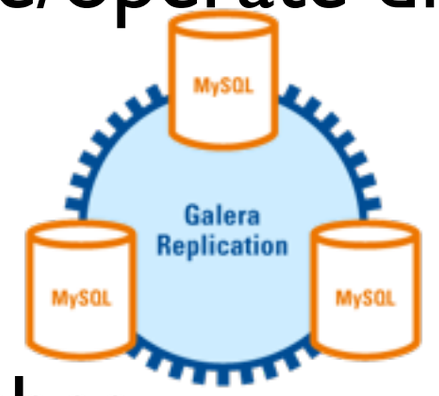


credits: J. Pastor et al., LBO - Scheduling - Feb 2014

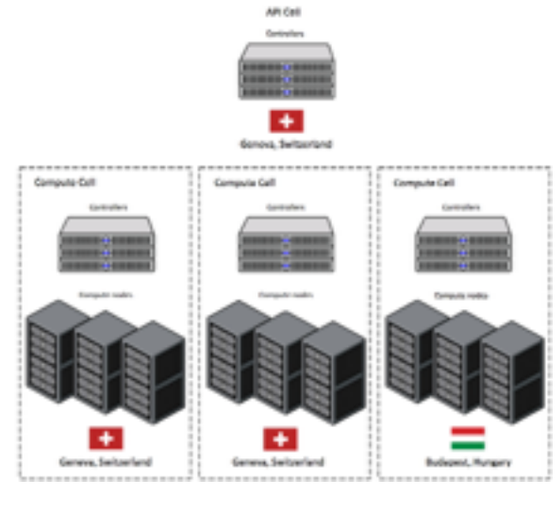
Next: network and storage dimensions

Revisiting OpenStack (on-going)

- Few proposals to federate/operate distinct OpenStack DCs
  - Leveraging Galera
    ⇒ Scalability issues



  - Hierarchical approaches
    - Cells based (CERN: 3 Sites / 50K cores)
      ⇒ the top cell is a central point

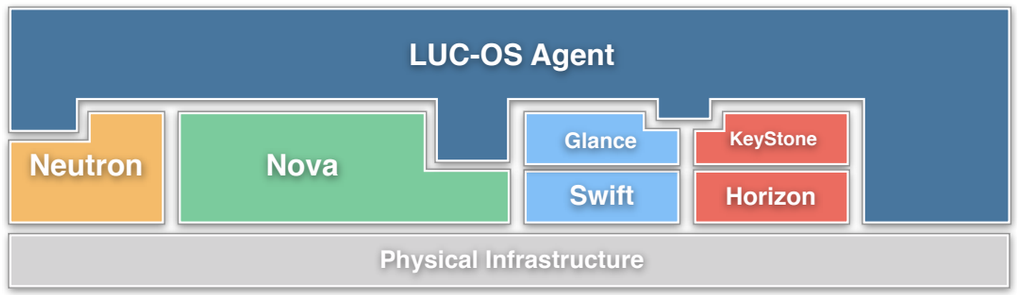


    - Cascading OpenStack (Huawei, Oct 2014 summit)
      A more advanced CELLS like approach (nova, neutron,…)
      ⇒ a unique cascading OpenStack,
      build on top of the OpenStack API (a system of systems)

- You know others ! ? please mail us.
  We try to maintain a dedicated webpage

Revisiting OpenStack (on-going)

- Our target is to deliver a multi-agent architecture composed of several OpenStack that can natively cooperate (i.e. without specific or priority elements)

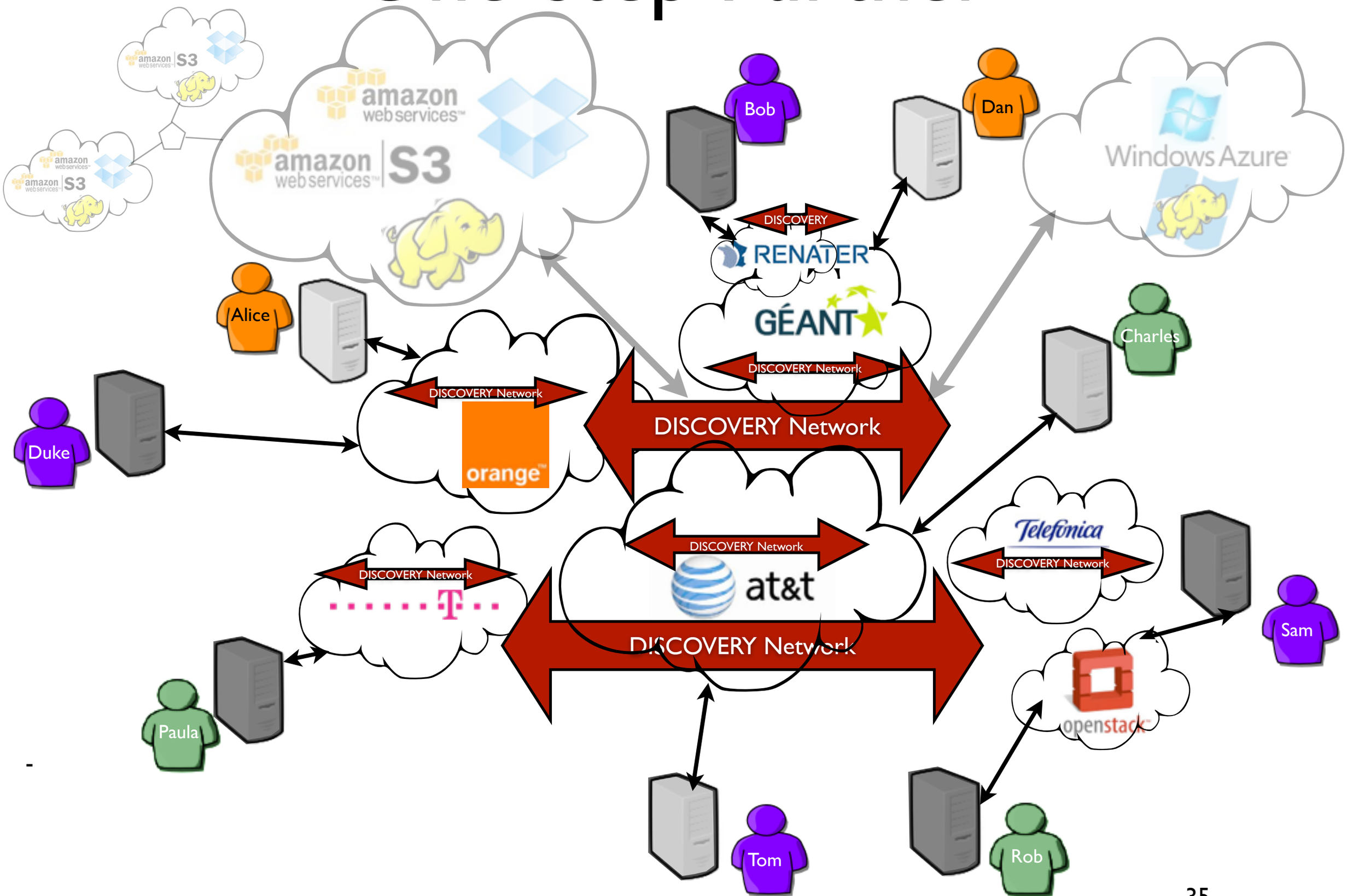


- Some components should require only minor changes / extensions to fit the Discovery's requirements (Swift, CephFS).
- Others, which have been built on top of centralised components (such as SQL DBs) must be revisited
- Identify centralized architecture issues (almost done for Nova, Neutron, Cinder/Glance and Keystone) and propose appropriate mechanisms to distribute them
- A Nova POC soon (Nova + RIAK, validation is on-going)

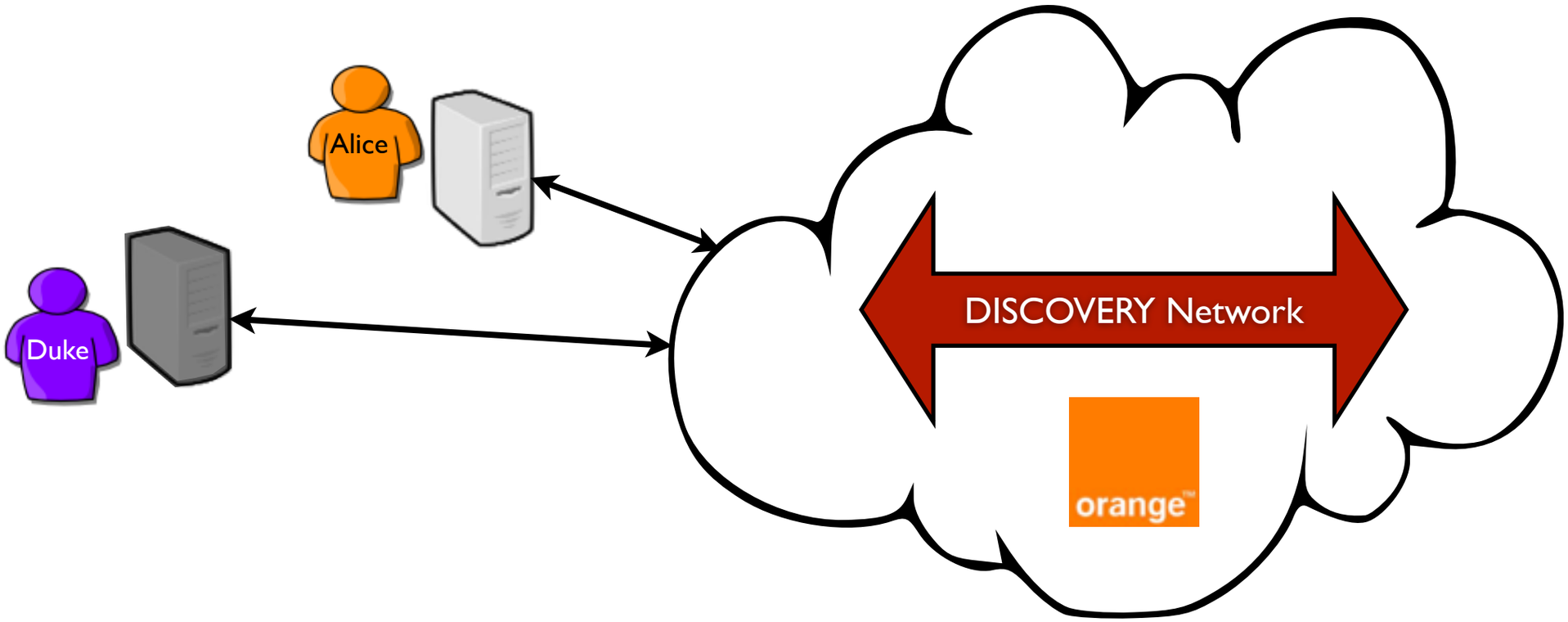
The Discovery Initiative

- Lots of challenges (network, VM images, security, …) that require to be addressed in a distributed and autonomous way.
- Leveraging former projects but still on the starting blocks!
- Preliminary works with promising results
- Long term objective: impact on the design of distributed applications in order to take advantage of the locality (building S3 like system)
- Important actors to follow:
  Akamai (micro DCs, Akamai/Aspera)
  Amazon (micro DCs at the Edge, cloudFront)
  Huawei (cascading openstack, several use-casesma)

One Step Further

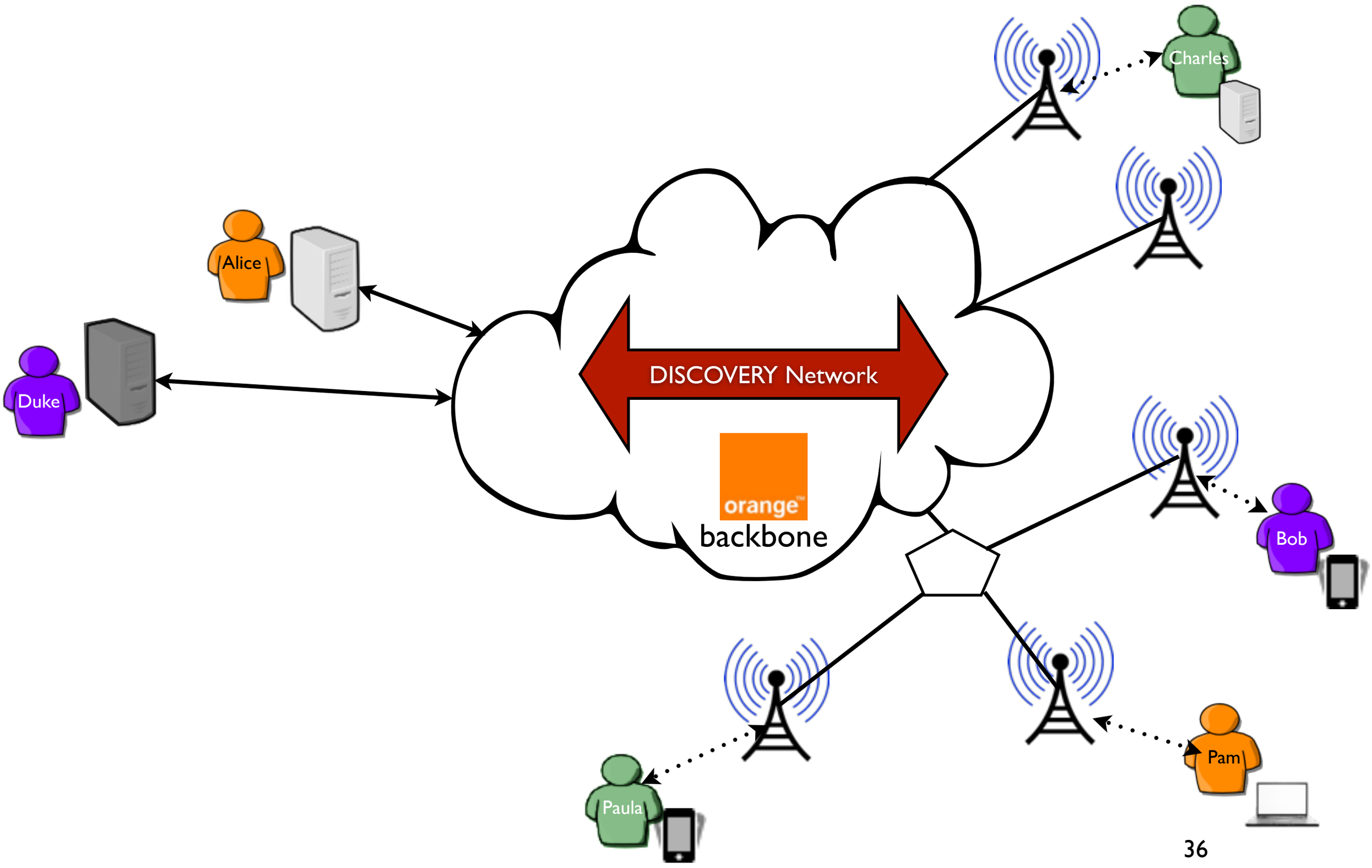


One Step Further

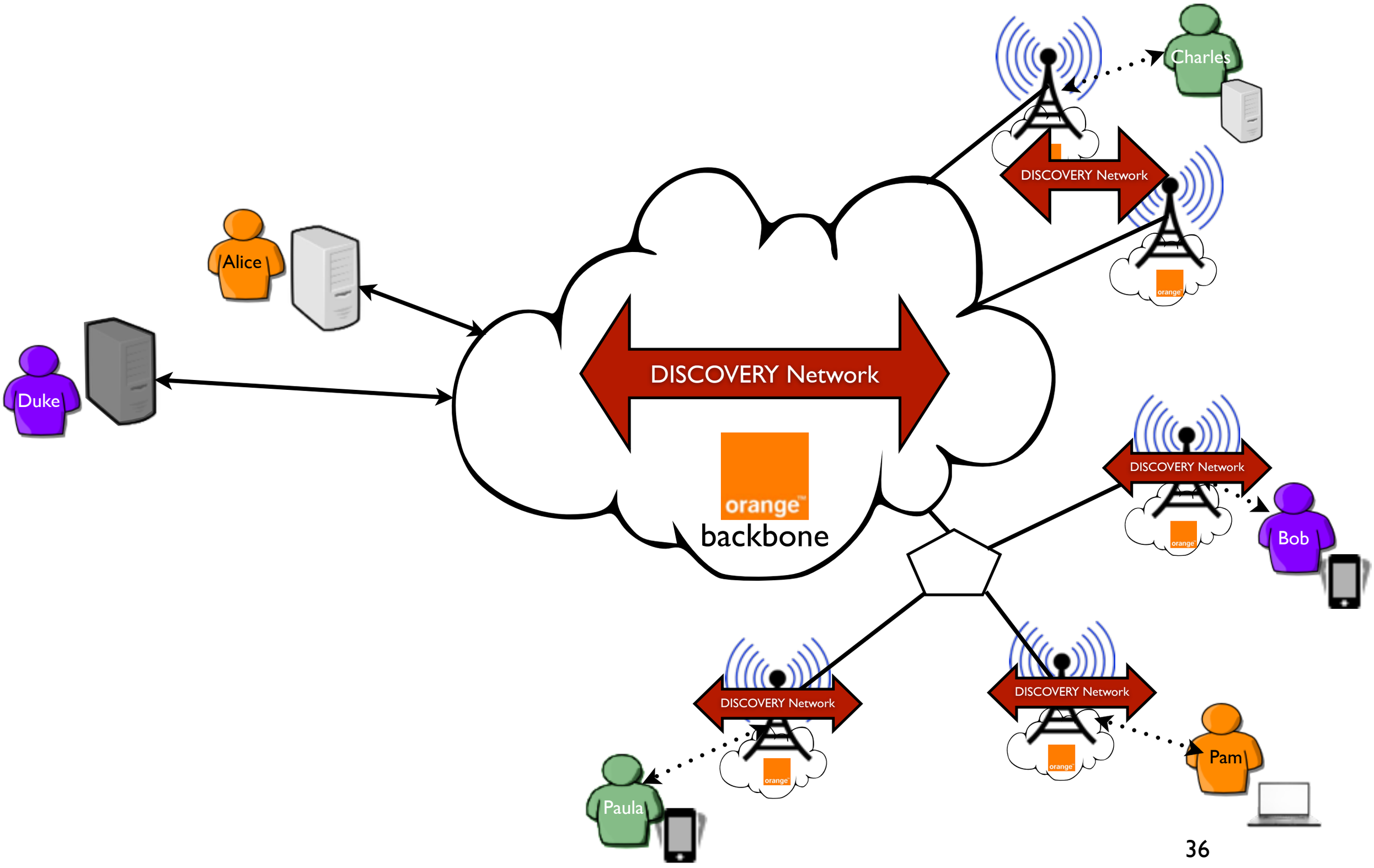


-

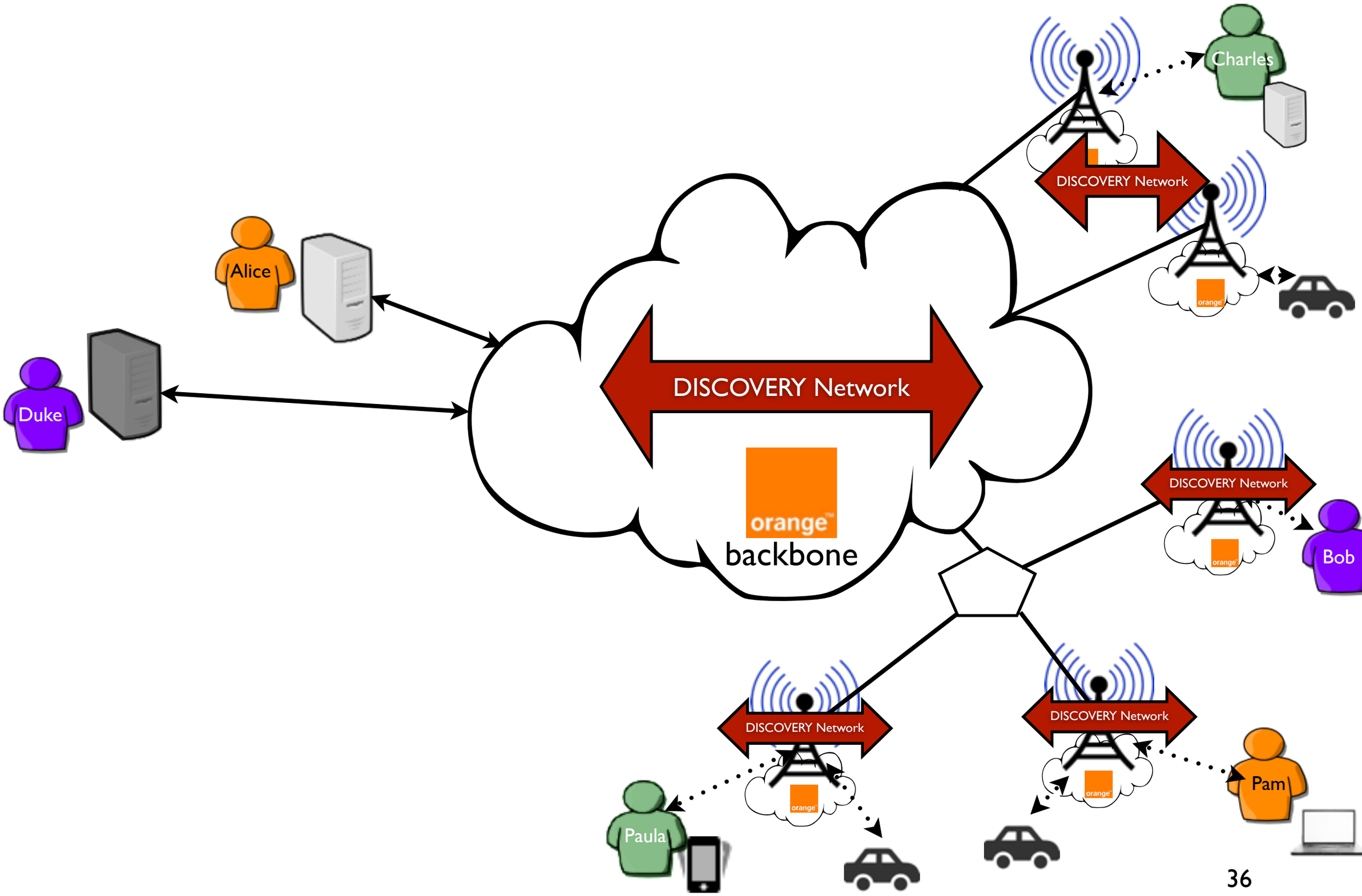
Radio Base Stations



Radio Base Stations



Radio Base Stations



The Discovery Initiative Pros/Cons

- Pros

  Locality (jurisdiction concerns, latency-aware apps, minimize network overhead)

  Reliability/redundancy (no critical point/location/center)
  The infrastructure is naturally distributed throughout multiple areas

  Lead time to delivery
  Leverage current PoPs and extend them according to UC demands

  Energy footprint (to be confirmed)

  Bring back part of the revenue to NRENs/Telcos

- Cons

  Security concerns (in terms of who can access to the PoPs)

  Operate a fully IaaS in a unified but distributed manner at WAN level

  Not suited for all kinds of applications : Large tightly coupled HPC workloads
  50 nodes/1000 cores, 200 nodes / 4000 cores (5 racks),
  so1000 nodes in one PoP does not look realistic …

  Peering agreement / economic model between network operators

Conclusion

- Cloud Computing technology is changing every day

  New features, new requirements (IaaS ++ services)

  One more challenge will be to ensure that such new features/ mechanisms can run in a distributed manner.

- Distributed Cloud Computing is happening !

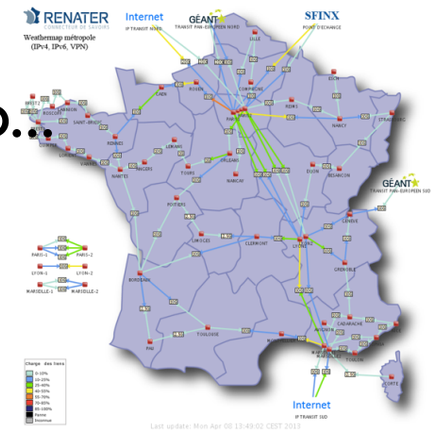
  Dist. CC workshop (2 editions UCC 2013, SIGCOMM 2014)
  FOG Computing workshop (collocated with IEEE ICC 2013)

  More and more academic papers
  Decentralizing the Cloud: How Can Small Data Centers Cooperate IEEE P2P 2014…

Beyond Discovery !

- From sustainable data centers to a new source of energy

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



- Leverage "green" energy (solar, wind turbines...)

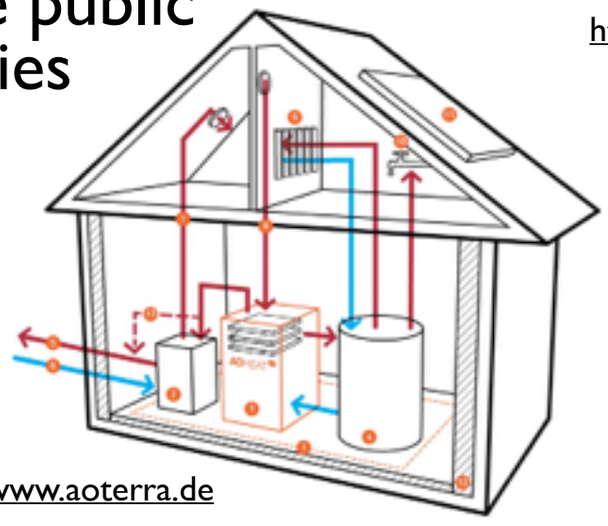
  Transfer the green micro/nano DCs concept to the network PoP

  Take the advantage of the geographical distribution

- Leveraging the data furnaces concept

  Deploy UC servers in medium and large institutions and use them as sources of heat inside public buildings such as hospitals or universities

http://parasol.cs.rutgers.edu



https://www.aoterra.de

The DISCOVERY Initiative

- Thank you / Questions ?
- 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/

adrien.lebre@inria.fr / jonathan.pastor@inria.fr

Additional slides

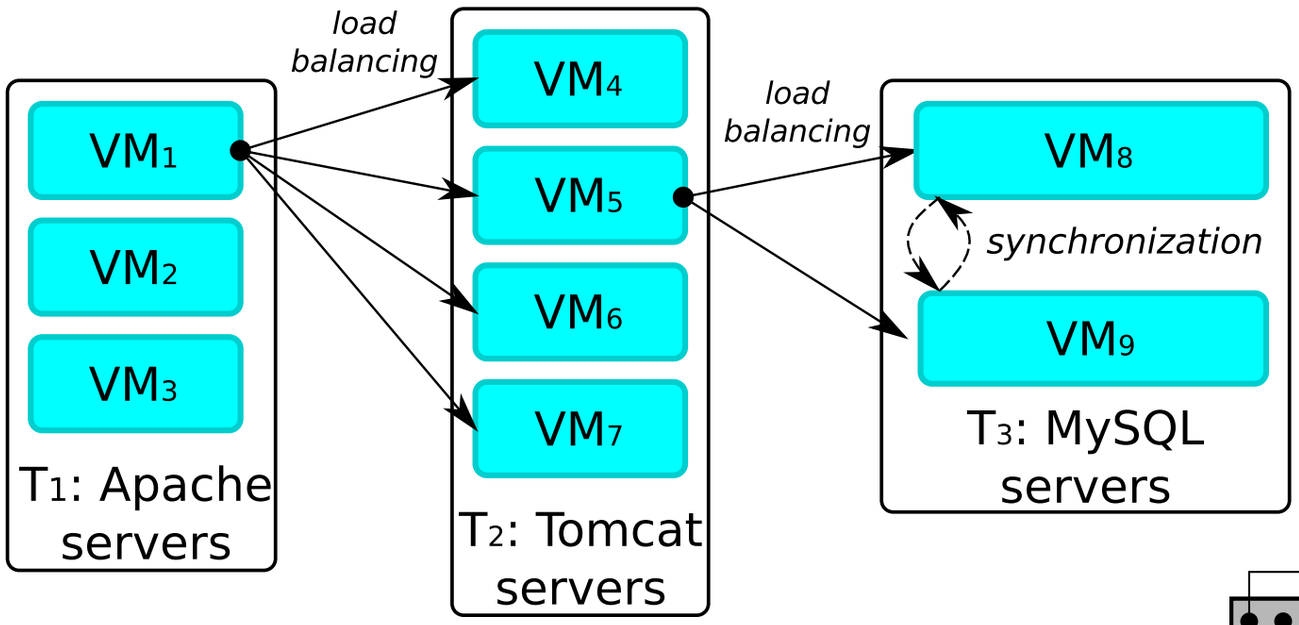
BtrPlace/ Plasma
More Constraints

- Manipulate VEs dynamically can lead to non desired configurations
- Additional constraints should be considered

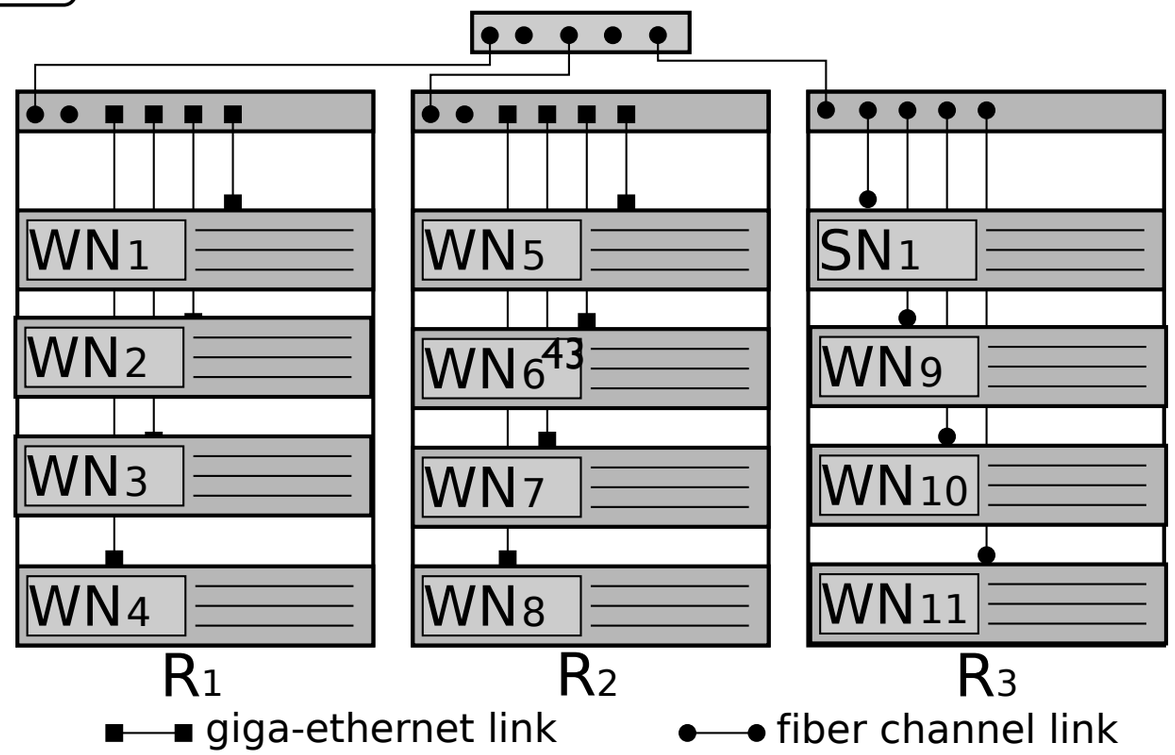
To take into account particular requirements according to the infrastructure (performance, HA, maintenance operations....)

To maintain VE "consistency" during reconfigurations

Background - Plasma and BtrPlace

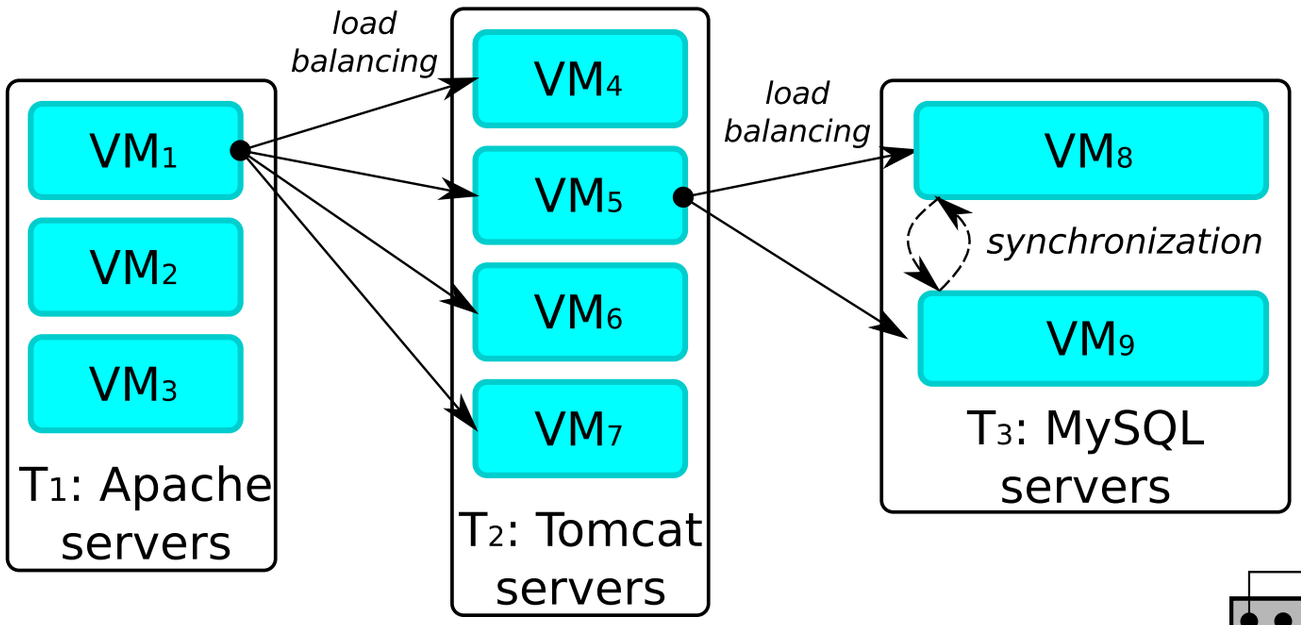


Virtualized HA application



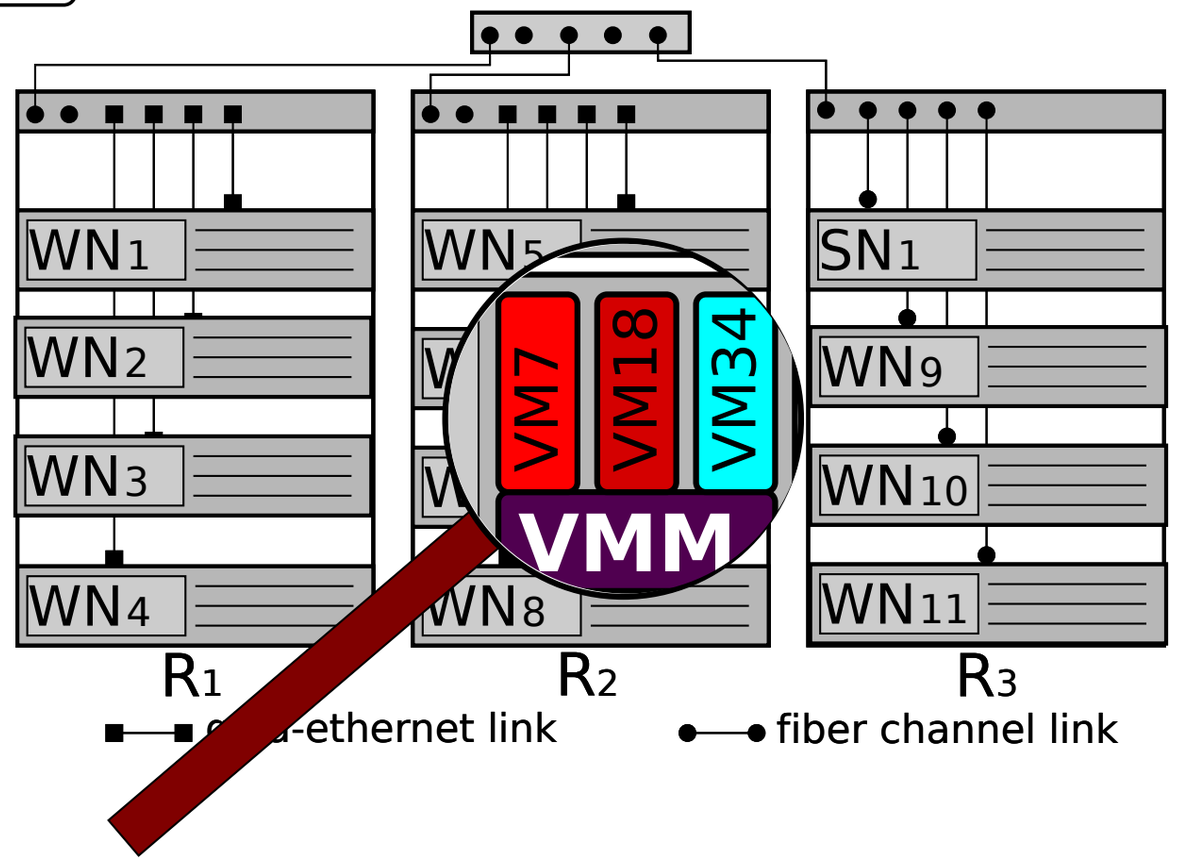
Credits: Hermenier et al., RR-7545 INRIA

Background - Plasma and BtrPlace



Virtualized HA application

Plasma, a DSL to describe:
- the infrastructure
- the VEs and their placement constraints



Credits: Hermenier et al., RR-7545 INRIA

Background - Plasma and Entropy

- ban({VM1, VM2}, {N1, N2})

  Prevents a set of VMs from being hosted on a given set of nodes

- fence({VM1, VM2}, {N1, N2})

  Forces a set of VMs to be hosted on a set of nodes

- spread({VM1, VM2})

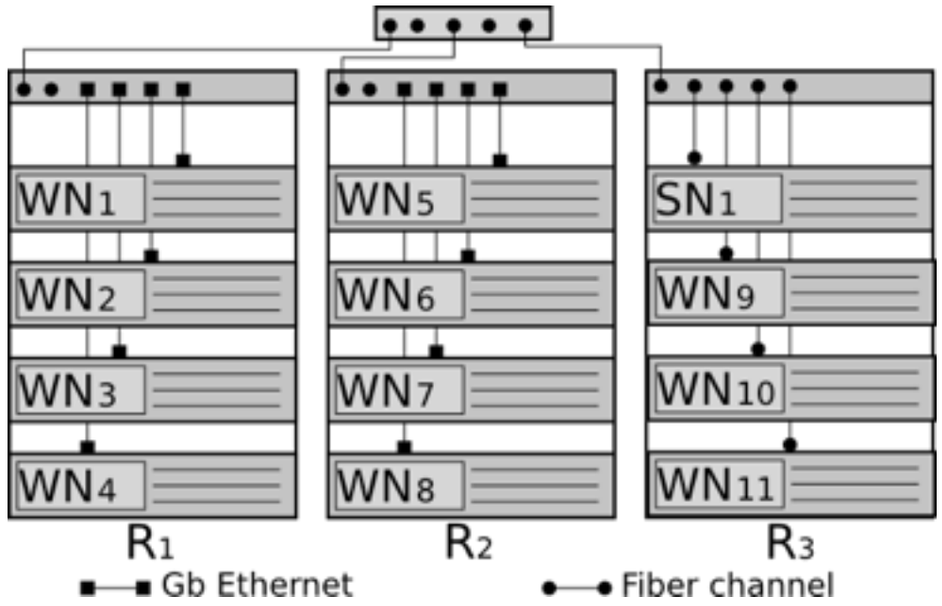
  Ensures that the specified VMs are never hosted on the same node at the same time

- latency({VM1, VM2}, {{N1, N2}, {N3, N4}})

  Forces a set of VMs to be hosted on a single group of nodes

- See more on **http://btrp.inria.fr/**

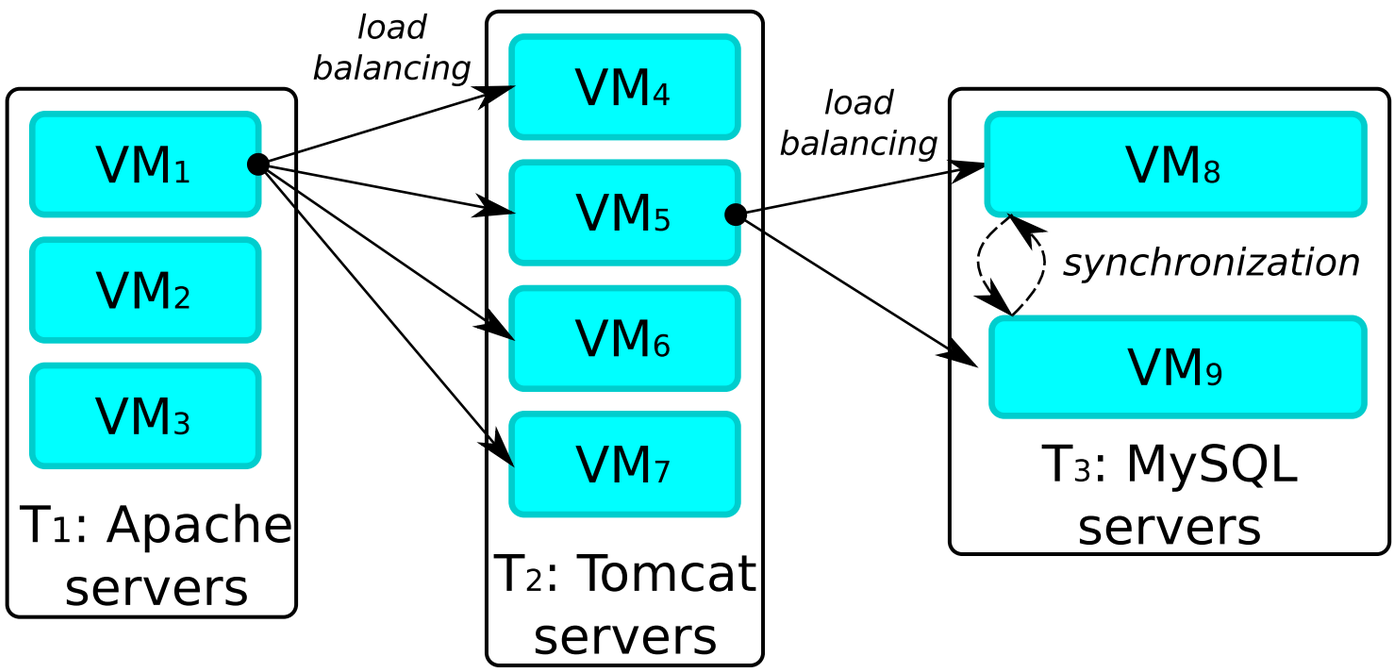
Infrastructure/Application Description




```
// Infrastructure
$R1 = {WN1 ,WN2 ,WN3 ,WN4 };
$R2 = WN [5..8];
$R3 = WN [9..11] + {SN1 };

// Classes of latency
$small = {$R3 };
$medium = $R [1..3];

// Constraints
ban ( $ALL_VMS , {SN1 });
ban ( $ALL_VMS , {WN5 });
fence ($A1 , $R2 + $R3 );
```




```
// The 3 tiers
$T1 = {VM1 ,VM2 ,VM3 };
$T2 = VM [4..7];
$T3 = VM [8..9];

// Fault tolerance to hw. failures
spread($T1);
spread($T2);
spread($T3);

// Efficient synchronization
latency ($T3 , $small );
```