Sky Computing on FutureGrid and Grid'5000 with Nimbus

Pierre Riteau
Université de Rennes 1, IRISA
INRIA Rennes – Bretagne Atlantique
Rennes, France

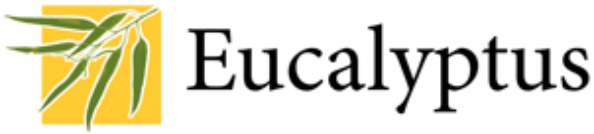
Outline

- Introduction to Sky Computing
- The Nimbus Project
- Large-Scale Sky Computing Experiments
- Conclusion

INTRODUCTION TO SKY COMPUTING

Cloud Computing

- Access to remote resources
  - On demand/elastic model
  - Pay as you go
- Multiple abstractions
  - IaaS, PaaS, SaaS
- Infrastructure as a Service (IaaS)
  - Access to virtual machines with administrator privileges
- Commercial providers
  - Amazon EC2, Rackspace, etc.
- Open-source implementations
  - Allow to create private clouds

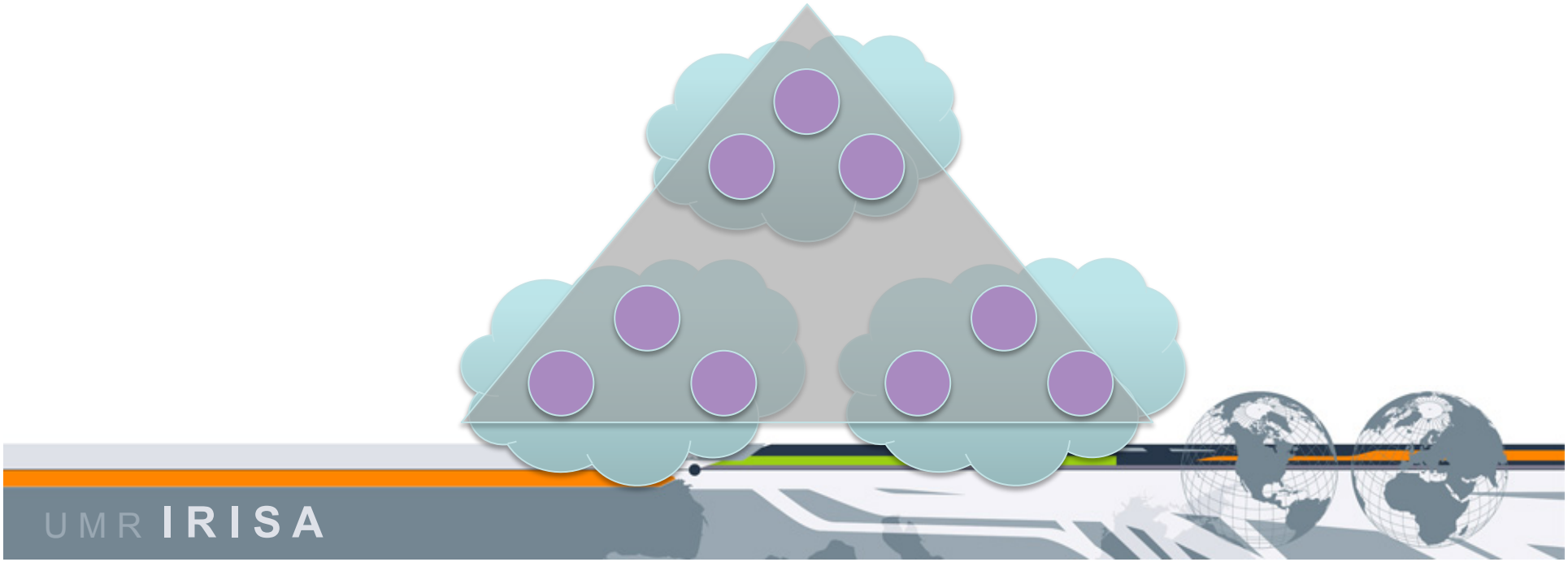


Infrastructure-as-a-Service

- Basic features
  - Run VM from VM image
  - Modify + save VM image
  - Terminate VM
- All operations accessible through an API
  - Autonomic infrastructure management
- Business model
  - Pay for CPU time + network traffic in/out + storage
- Initially targeting web service hosting
  - Also HPC now (Amazon Cluster Compute instances)

Sky Computing

- Federation of multiple clouds
- Creates large scale infrastructures
- Allows to run software requiring large computational power



Sky Computing Benefits

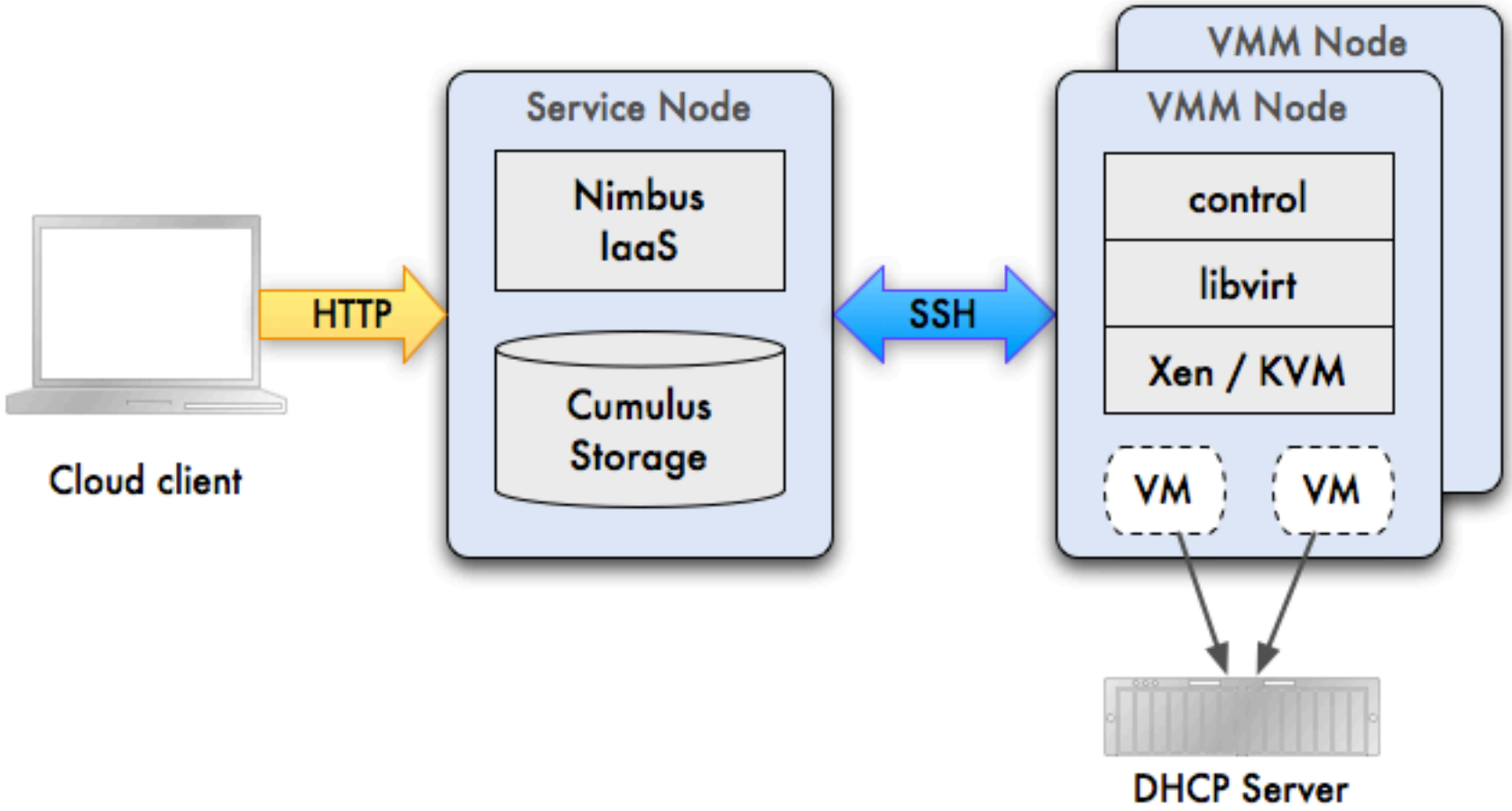
- Single networking context
  - All-to-all connectivity
- Single security context
  - Trust between all entities
- Equivalent to local cluster
  - Compatible with legacy code

THE NIMBUS PROJECT

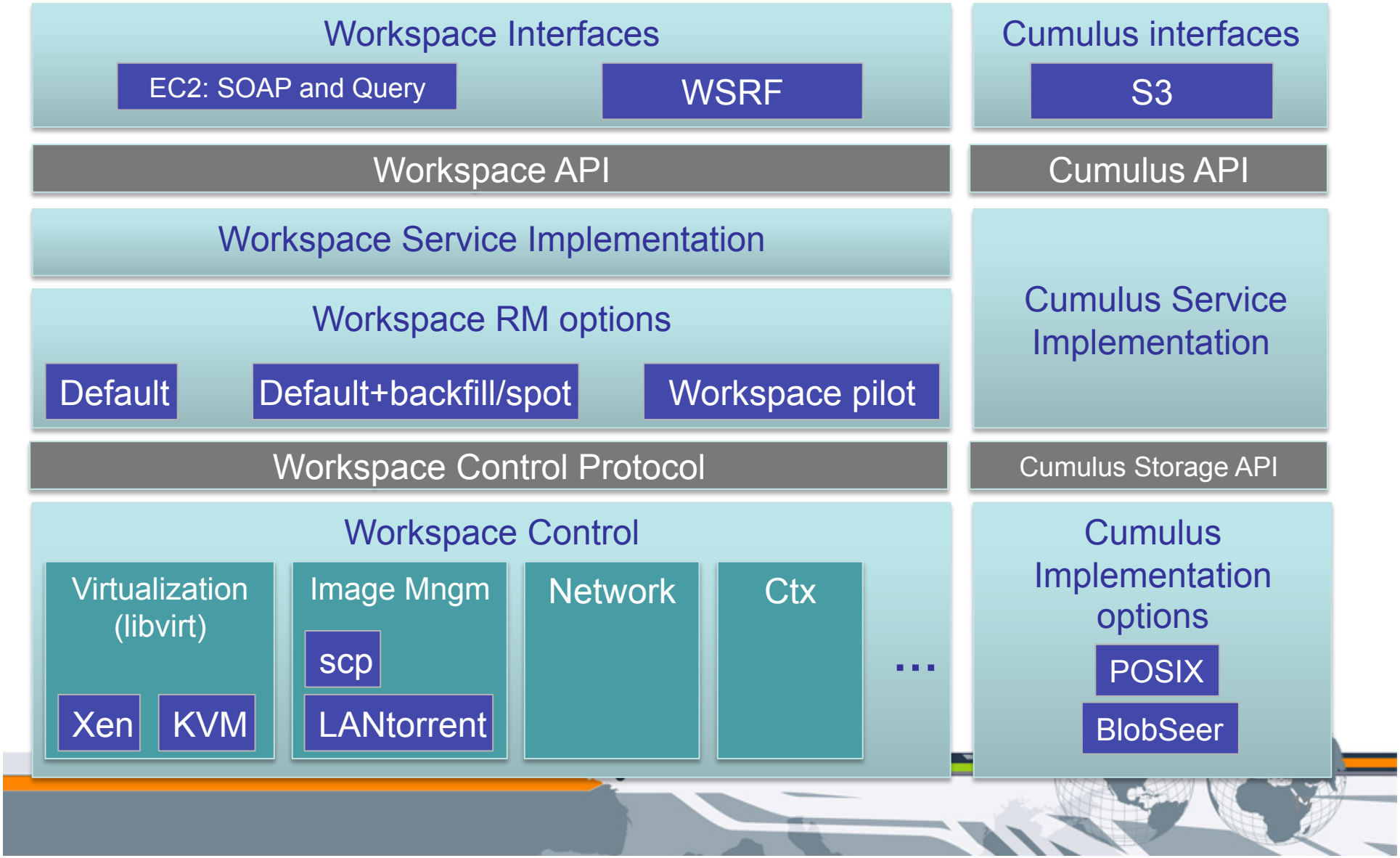
The Nimbus Project

- Started in 2005 by Kate Keahey (Argonne/UC)
- Cloud computing Toolkit
  - Open source IaaS implementation
    - Amazon EC2-compatible (WSDL and Query APIs)
  - Cumulus (Amazon S3-compatible storage cloud)
- Targets Clouds for Science
- Unique features
  - Context Broker
  - Workspace Pilot
  - LANTorrent
  - Spot instances

Nimbus Architecture



Nimbus: A Highly-Configurable IaaS Architecture



Context Broker

- Service to configure a complete cluster with different roles
- Works with a cluster distributed on multiple clouds (e.g. Nimbus and Amazon EC2)
- VMs contact the context broker to
  - Learn their role
  - Learn about other VMs in the cluster
- Ex. : Hadoop master + Hadoop slaves
    - Hadoop slaves configured to contact the master
    - Hadoop master configured to know the slaves

Cluster description

```
<?xml version="1.0" encoding="UTF-8"?>
<cluster xmlns="http://www.globus.org/2008/06/workspace/
      metadata/logistics">
 <workspace>
   <name>hadoop-master</name>
   <image>fc8-i386-nimbus-blast-cluster-004</
     image>
   <quantity>1</quantity>
  <nic wantlogin="true">public</nic>
  <ctx>
   <provides>
     <identity />
       <role>hadoop_master</role>
       <role>hadoop_slave</role>
   </provides>
   <requires>
     <identity />
     <role name="hadoop_slave" hostname="true"
    pubkey="true" />
     <role name="hadoop_master" hostname="true"
    pubkey="true" />
   </requires>
  </ctx>
 </workspace>
```

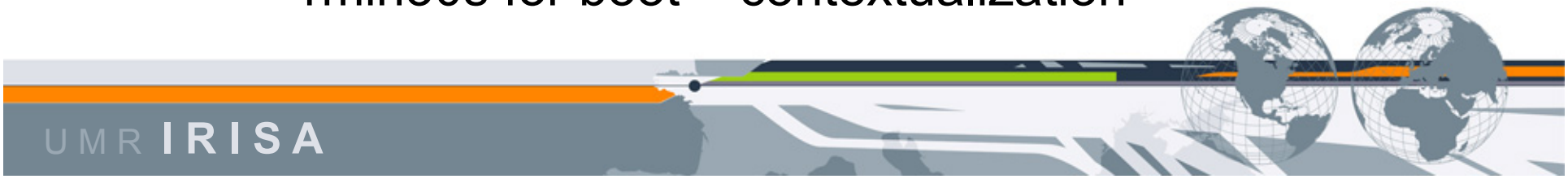
```
<workspace>
  <name>hadoop-slaves</name>
  <image>fc8-i386-nimbus-blast-
    cluster-004</image>
  <quantity>16</quantity>
 <nic wantlogin="true">public</nic>
 <ctx>
  <provides>
    <identity />
      <role>hadoop_slave</role>
  </provides>
  <requires>
    <identity />
    <role name="hadoop_master" hostname="true"
   pubkey="true" />
  </requires>
 </ctx>
</workspace>
</cluster>
```

Worskpace Pilot

- Integrates Nimbus in existing clusters (Torque-based)
- User request VM -> a “pilot” job reserves a node
- Seamlessly share a cluster between traditional submission and VM leases

LANTorrent

- Efficient VM image propagation
  - Based on the BitTorrent concept
  - But optimized for a LAN environment
  - Duplicate data detection
    - when sending multiple images to the same node
  - A thousand VMs deployed in 10 minutes on Magellan (Argonne)
  - 168 VMs on G5K (4 VMs/node)
    - 1min30s for propagation
    - 1min30s for boot + contextualization



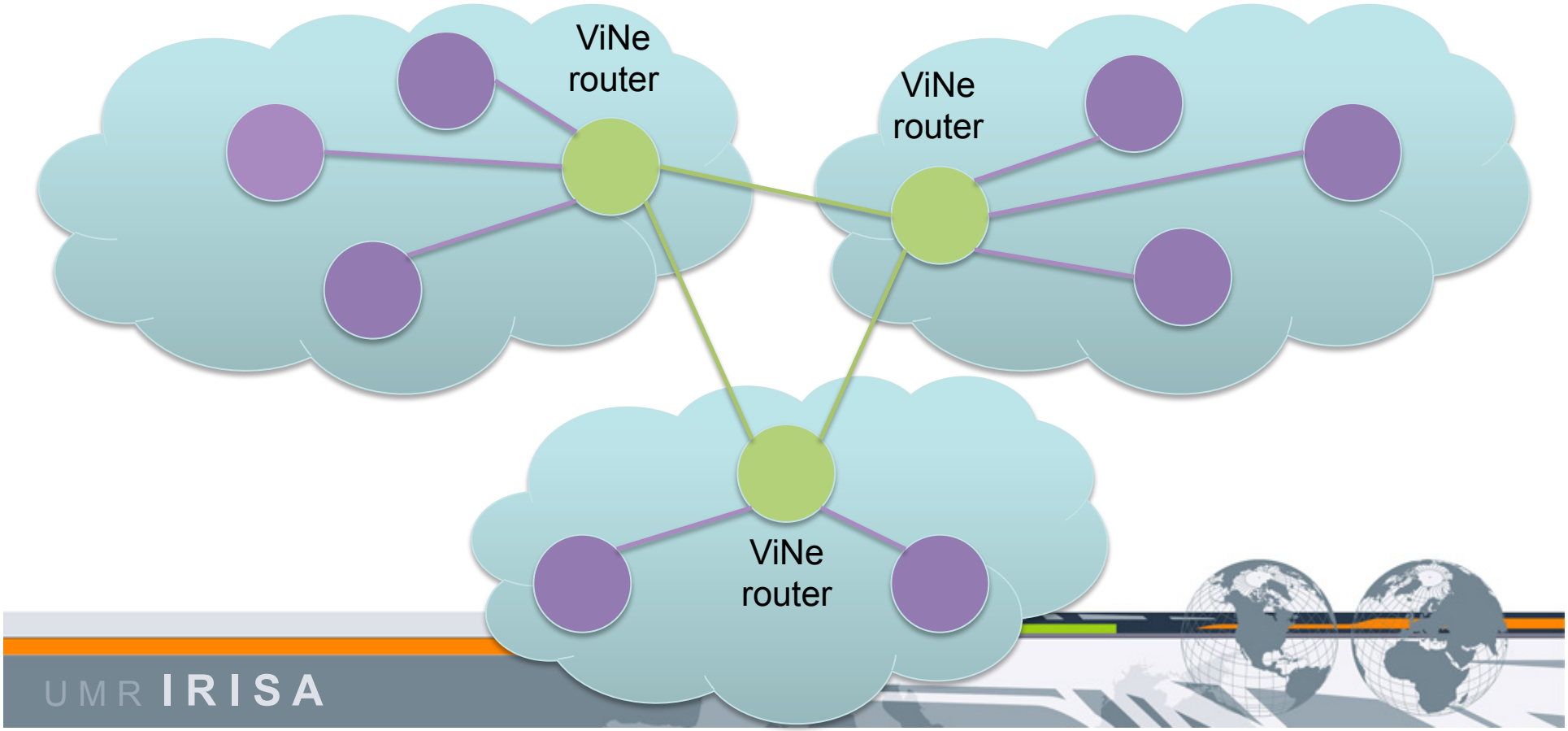
LARGE-SCALE SKY COMPUTING EXPERIMENTS

Sky Computing Toolkit

- Nimbus
  - Resource management
  - Contextualization
- ViNe
  - All-to-all connectivity
- Hadoop
  - Task distribution
  - Fault tolerance
  - Resource dynamicity

ViNe

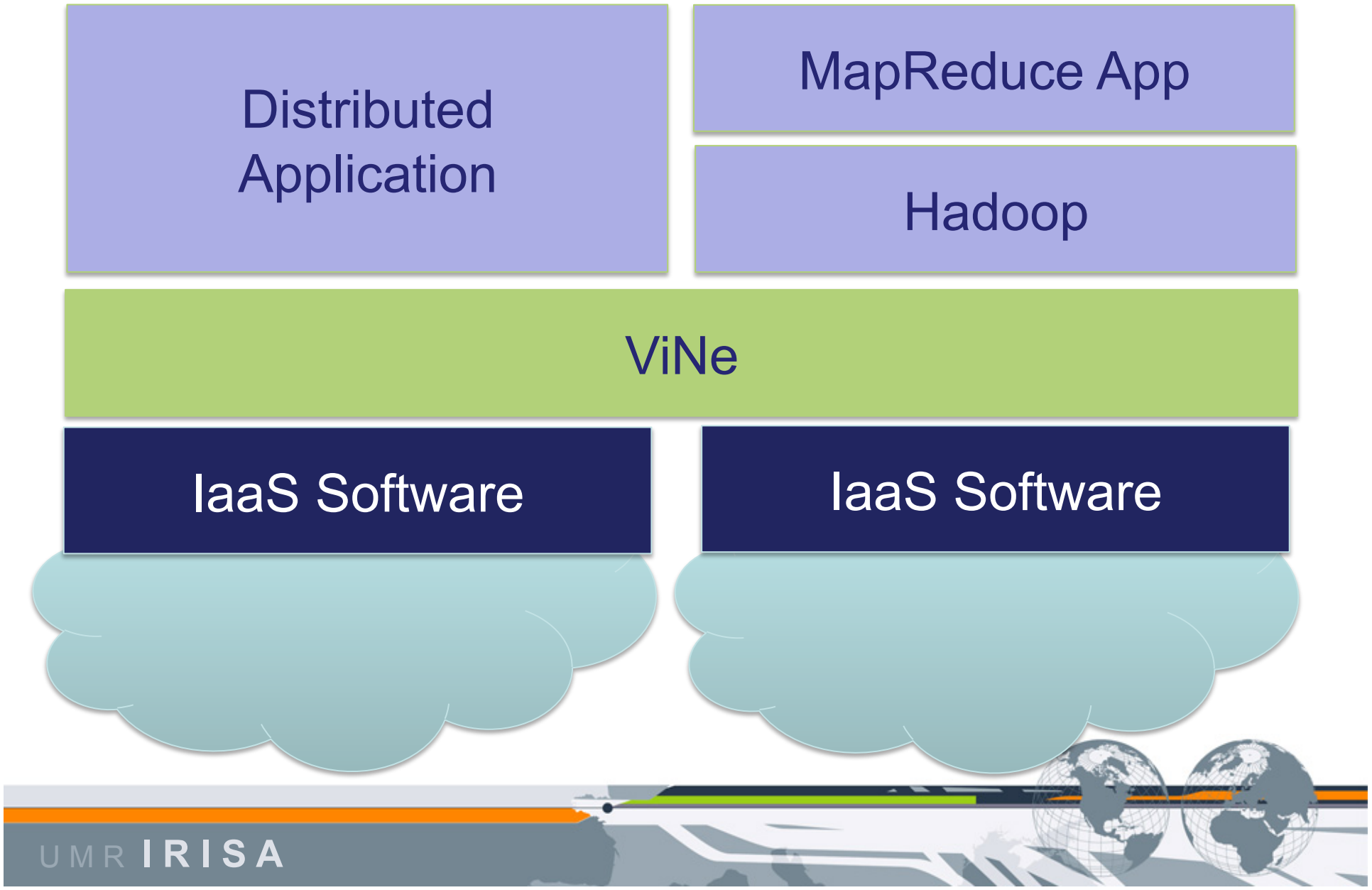
- Project of the University of Florida (M. Tsugawa et al.)
- High performance virtual network
- All-to-all connectivity



Hadoop

- Open-source MapReduce implementation
- Heavy industrial use (Yahoo, Facebook…)
- Efficient framework for distribution of tasks
- Built-in fault-tolerance
- Distributed file system (HDFS)

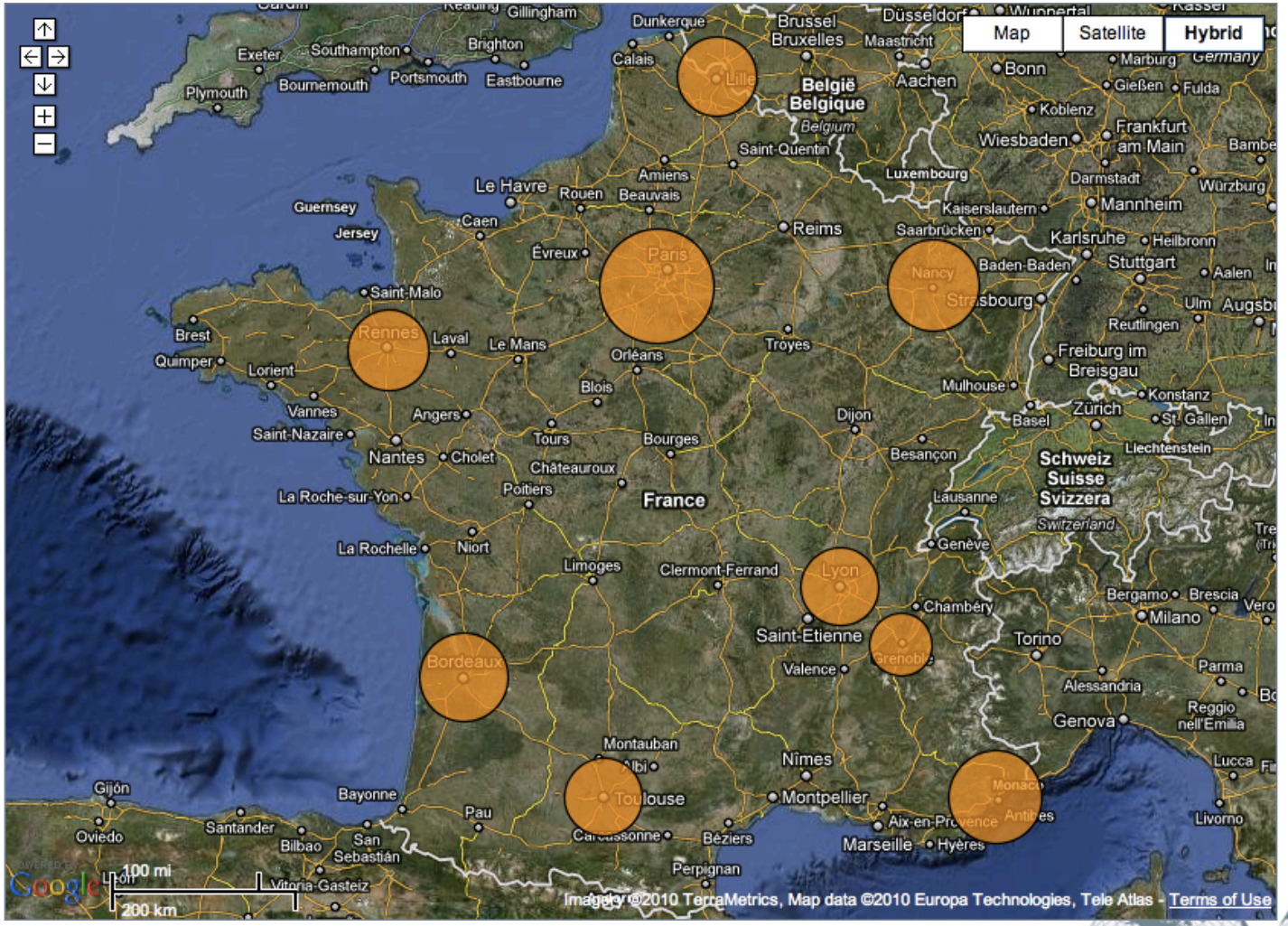
Sky Computing Architecture



Grid'5000 Overview

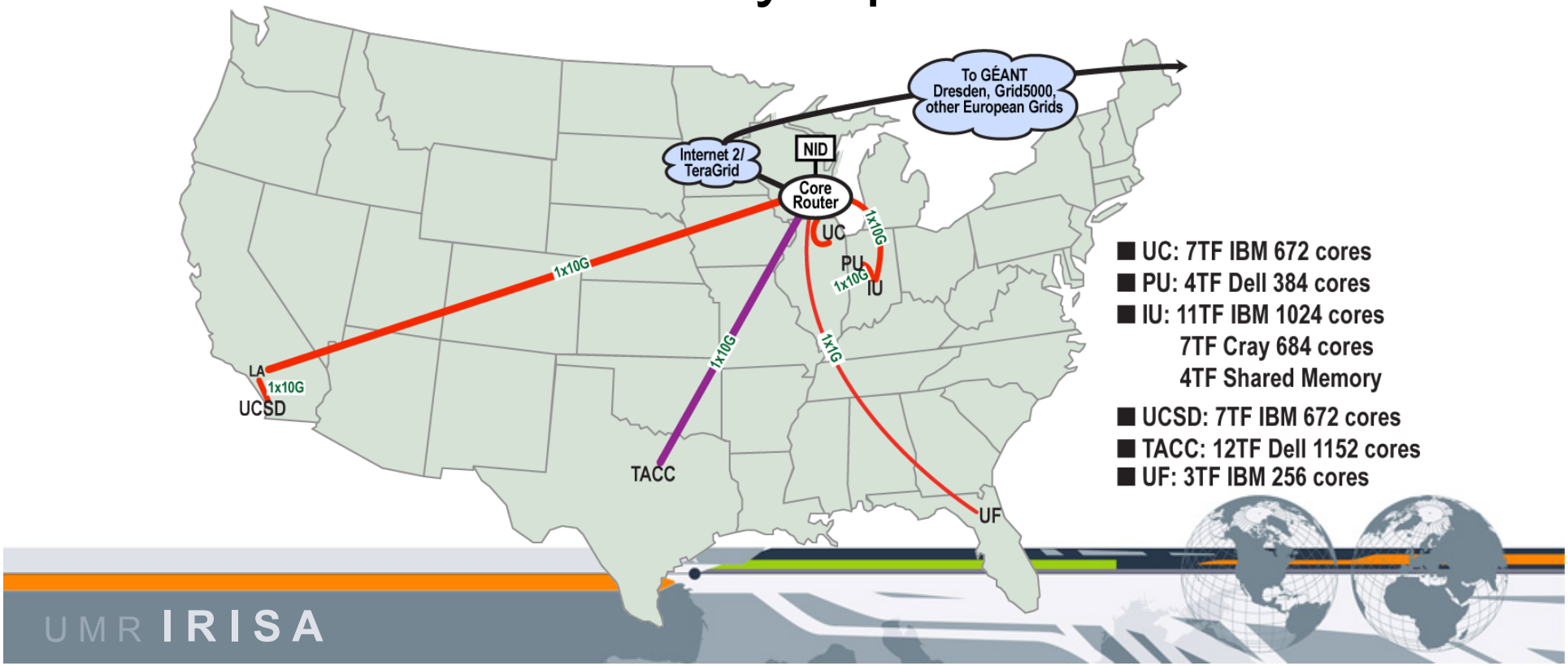
- Distributed over 9 sites in France
- ~1500 nodes, ~5500 CPUs
- Study of large scale parallel/distributed systems
- Features
  - Highly reconfigurable
    - Environment deployment over bare hardware
    - Can deploy many different Linux distributions
    - Even other OS such as FreeBSD
  - Controlable
  - Monitorable (metrics access)
- Experiments on all layers
  - network, OS, middleware, applications

Grid'5000 Node Distribution



FutureGrid: a Grid Testbed

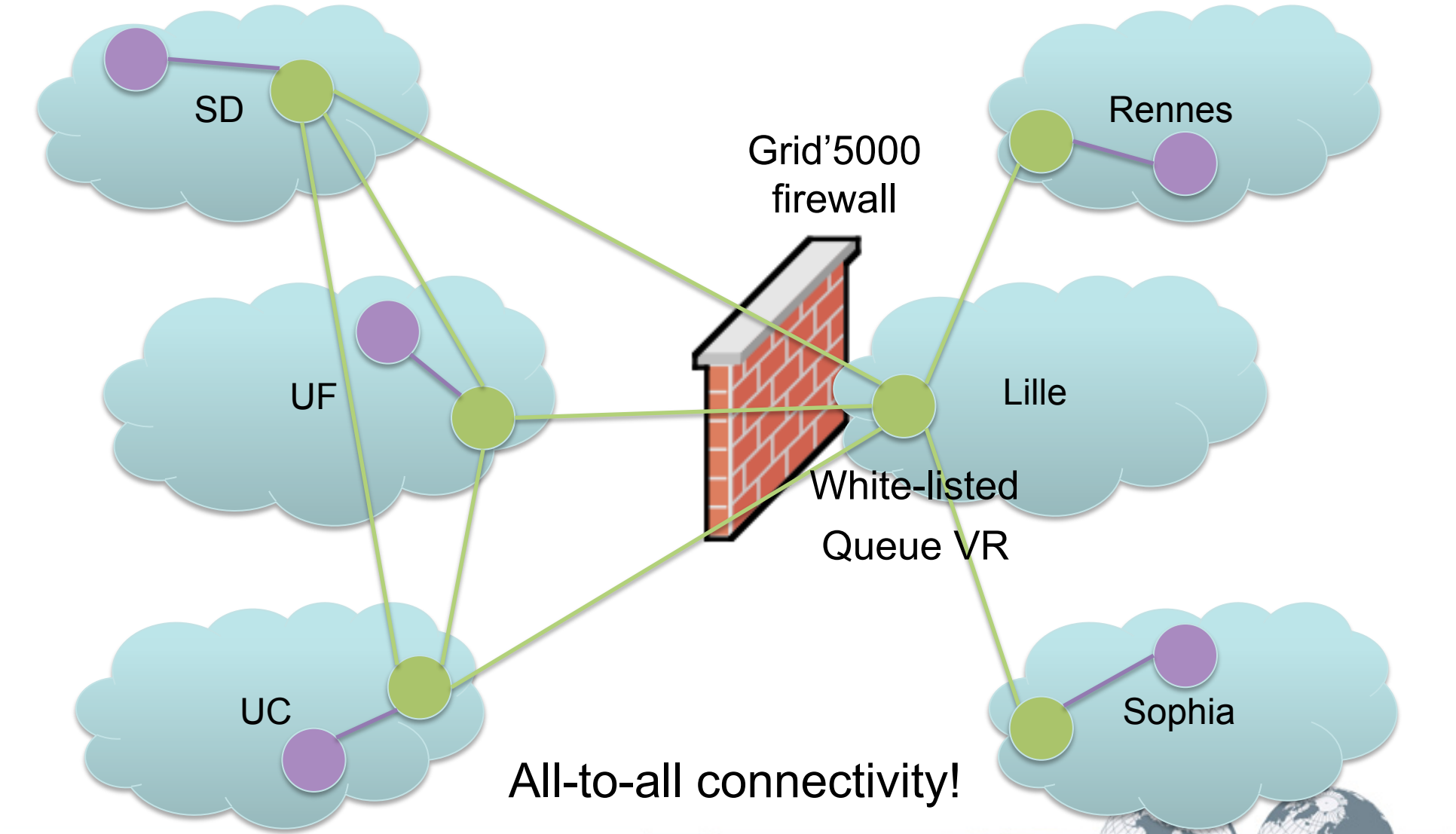
- NSF-funded experimental testbed
- ~5000 cores
- 6 sites connected by a private network



Resources used in Sky Computing Experiments

- 3 FutureGrid sites (US) with Nimbus installations
  - UCSD (San Diego)
  - UF (Florida)
  - UC (Chicago)
- Grid'5000 sites (France)
  - Lille (contains a white-listed gateway to FutureGrid)
  - Rennes, Sophia, Nancy, etc.
- Grid'5000 is fully isolated from the Internet
  - One machine white-listed to access FutureGrid
  - ViNe queue VR (Virtual Router) for other sites

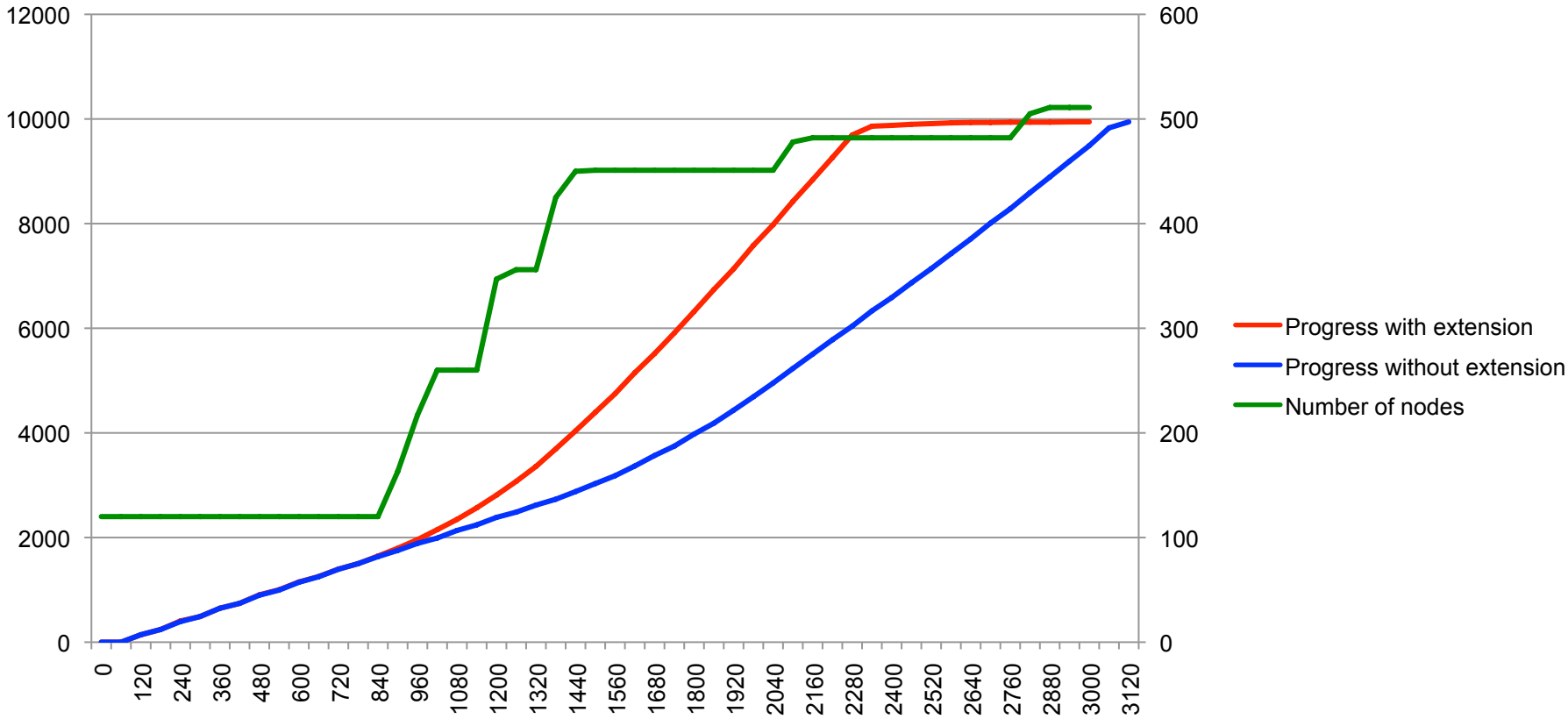
ViNe Deployment Topology



Experiment scenario

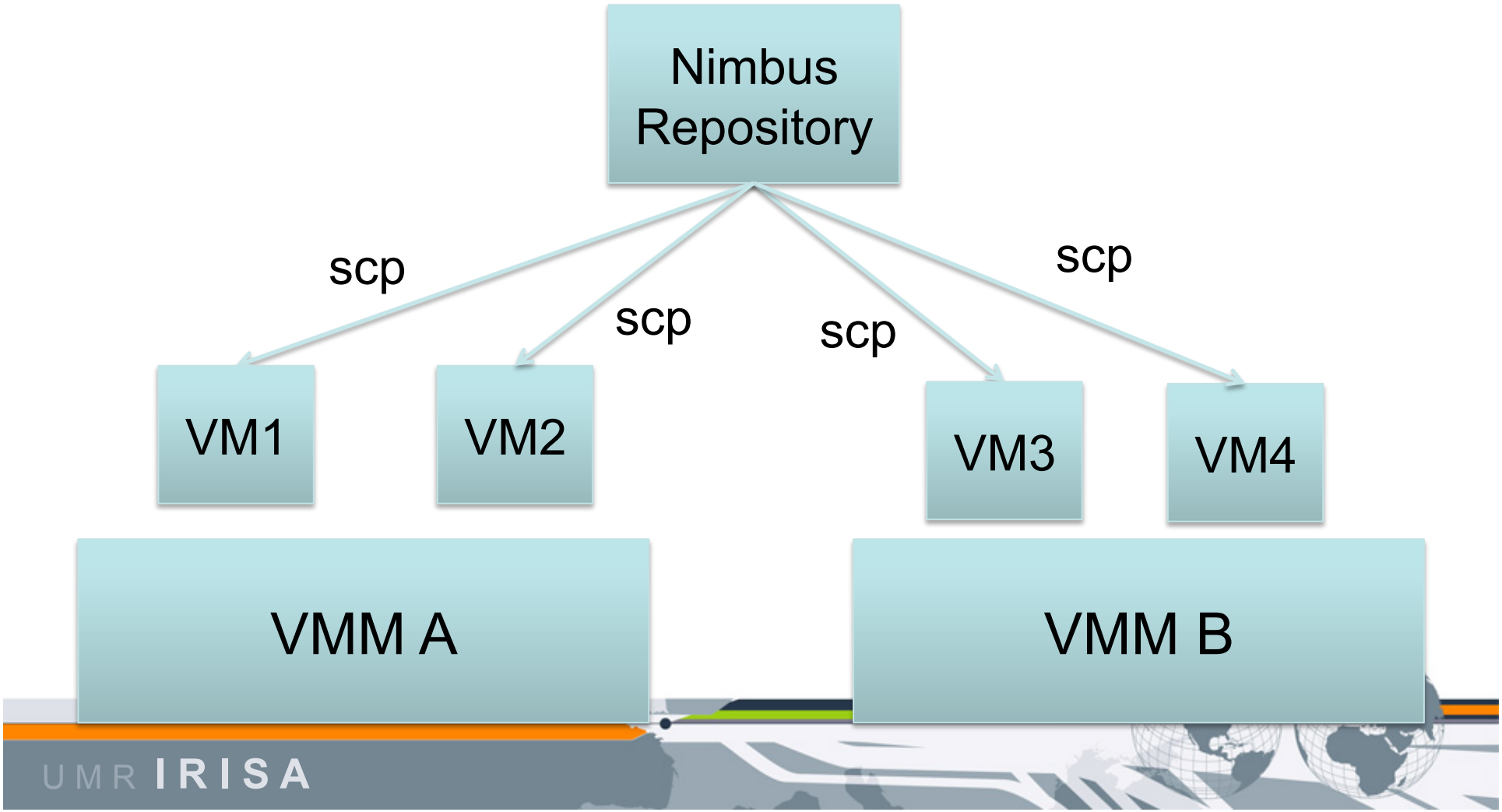
- Hadoop sky computing virtual cluster already running in FutureGrid (SD, UF, UC)
- Launch BLAST MapReduce job
- Start VMs on Grid'5000 resources
  - With contextualization to join the existing cluster
- Automatically extend the Hadoop cluster
  - Number of nodes increases
    - TaskTracker nodes (Map/Reduce tasks execution)
    - DataNode nodes (HDFS storage)
  - Hadoop starts distributing tasks in Grid'5000
  - Job completes faster!

Job progression with extension



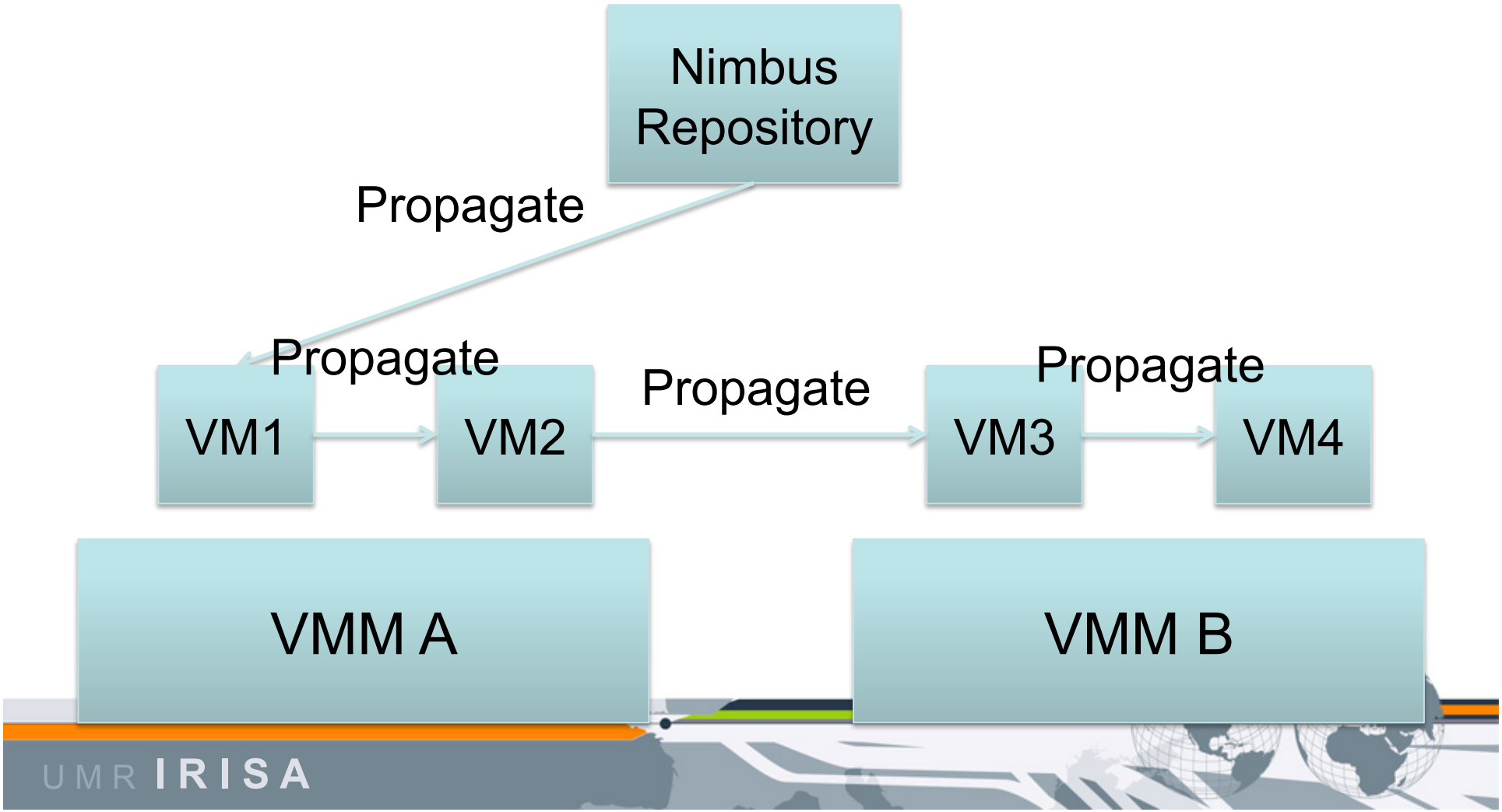
Fast Virtual Cluster Creation (1/3)

- Standard Nimbus propagation: scp



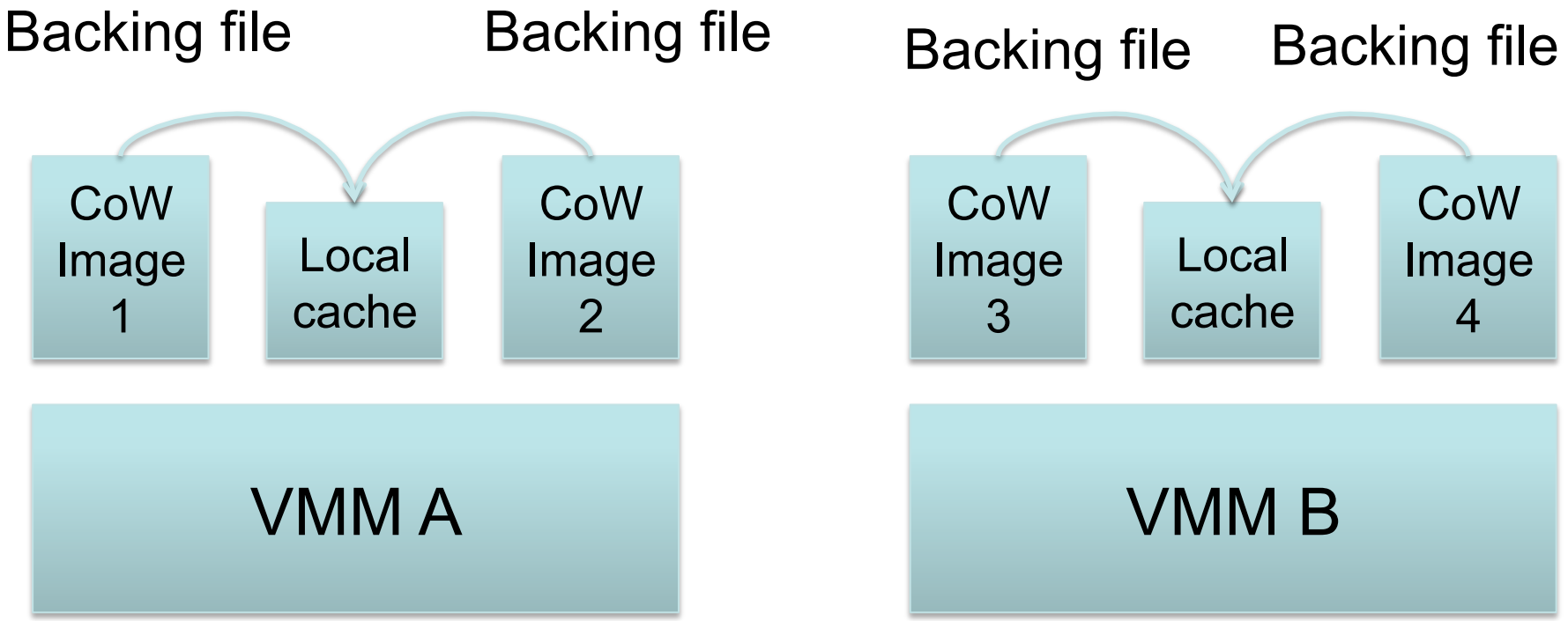
Fast Virtual Cluster Creation (2/3)

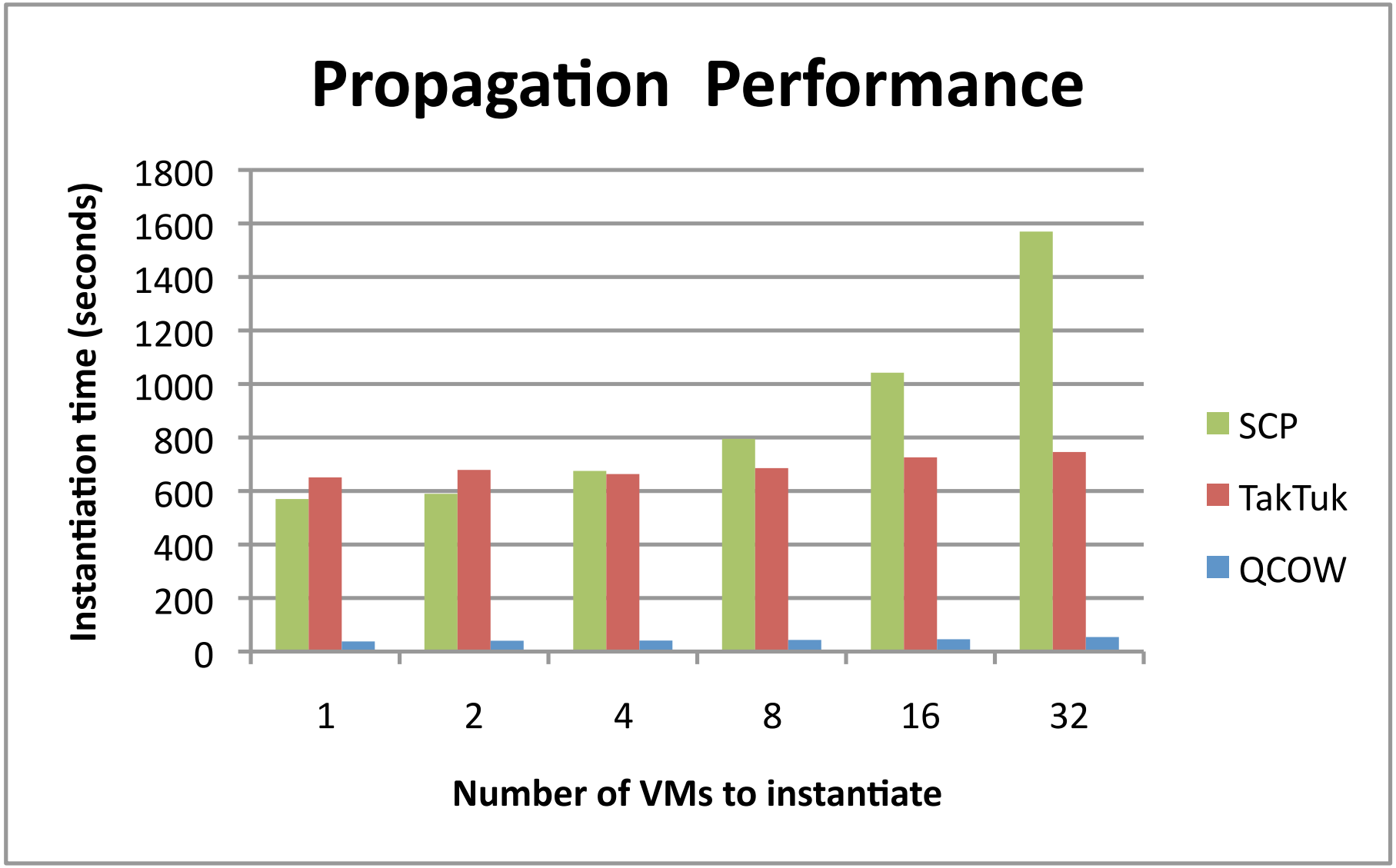
- Pipelined Nimbus propagation: Kastafior/TakTuk



Fast Virtual Cluster Creation (3/3)

- Leverage Xen Copy-on-Write (CoW) capabilities




Propagation Performance
Instantiation time (seconds)
1800
1600
1400
1200
1000
800
600
400
200
0
1
2
4
8
16
32
Number of VMs to instantiate
SCP
TakTuk
QCOW

CONCLUSION

Conclusion

- Sky Computing to create large scale distributed infrastructures
- Our approach relies on
  - Nimbus for resource management, contextualization and fast cluster instantiation
  - ViNe for all-to-all connectivity
  - Hadoop for dynamic cluster extension
- Provides both infrastructure and application elasticity

Ongoing & Future Works

- Migration support in ViNe (WAN migration)
- Elastic MapReduce implementation leveraging Sky Computing infrastructures
- Migration support in Nimbus
  - Leverage spot instances

Acknowledgments

- Tim Freeman, John Bresnahan, Kate Keahey (Argonne)
- David LaBissoniere (University of Chicago)
- Maurício Tsugawa, Andréa Matsunaga, José Fortes (University of Florida)
- Thierry Priol, Christine Morin (INRIA)

THANK YOU!

QUESTIONS?