How Virtualization Changed The Grid Perspective

Adrien Lèbre
Ecole des Mines de Nantes

"Des grilles aux Clouds, nouveaux problèmes et nouvelles solutions"
13 December 2010, Ecole Normale Supérieure de Lyon

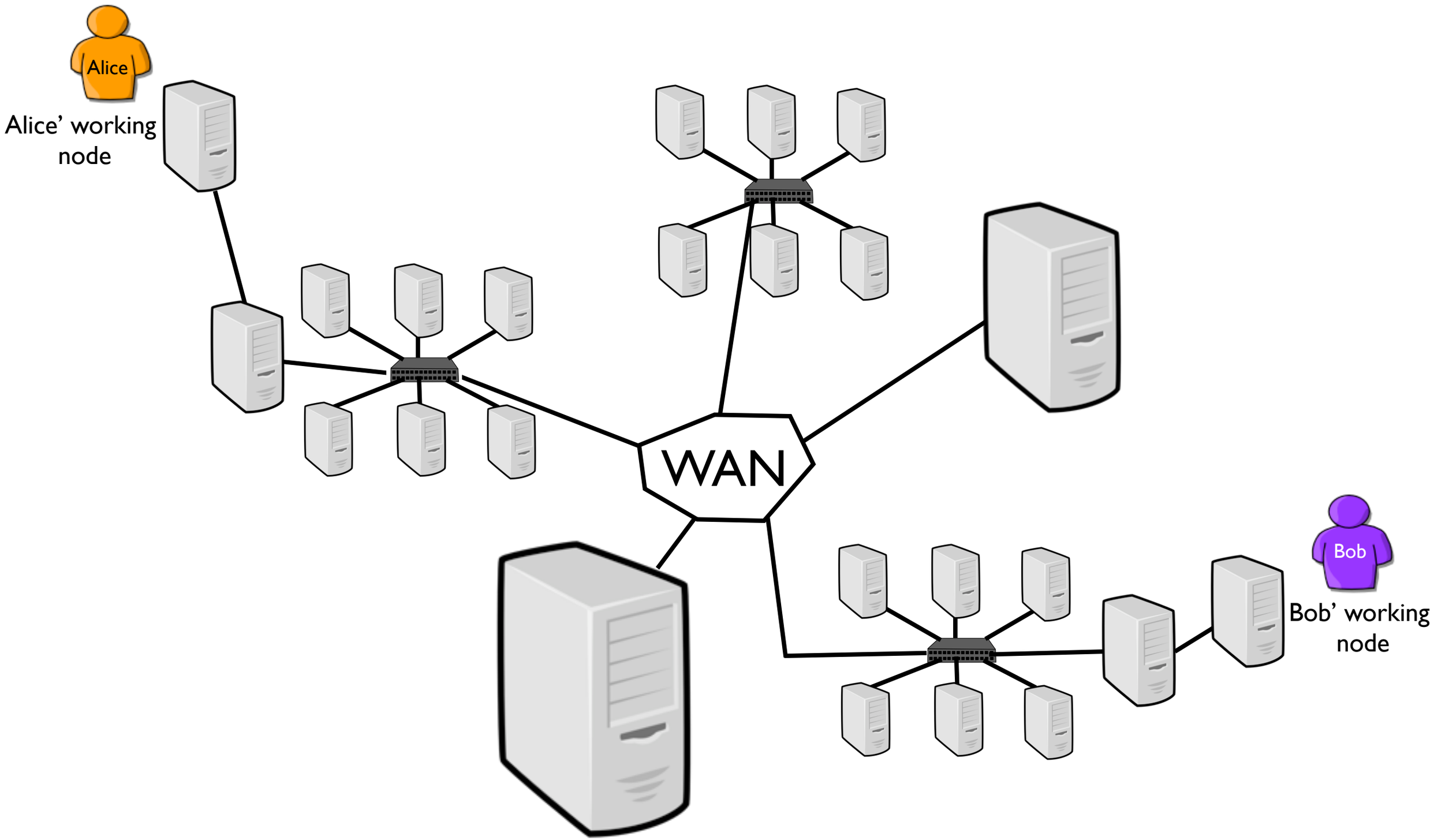
xxx Computing

- xxx as Distributed
  (Cluster / Grid / Desktop / "Hive" / Cloud / Sky / ...)

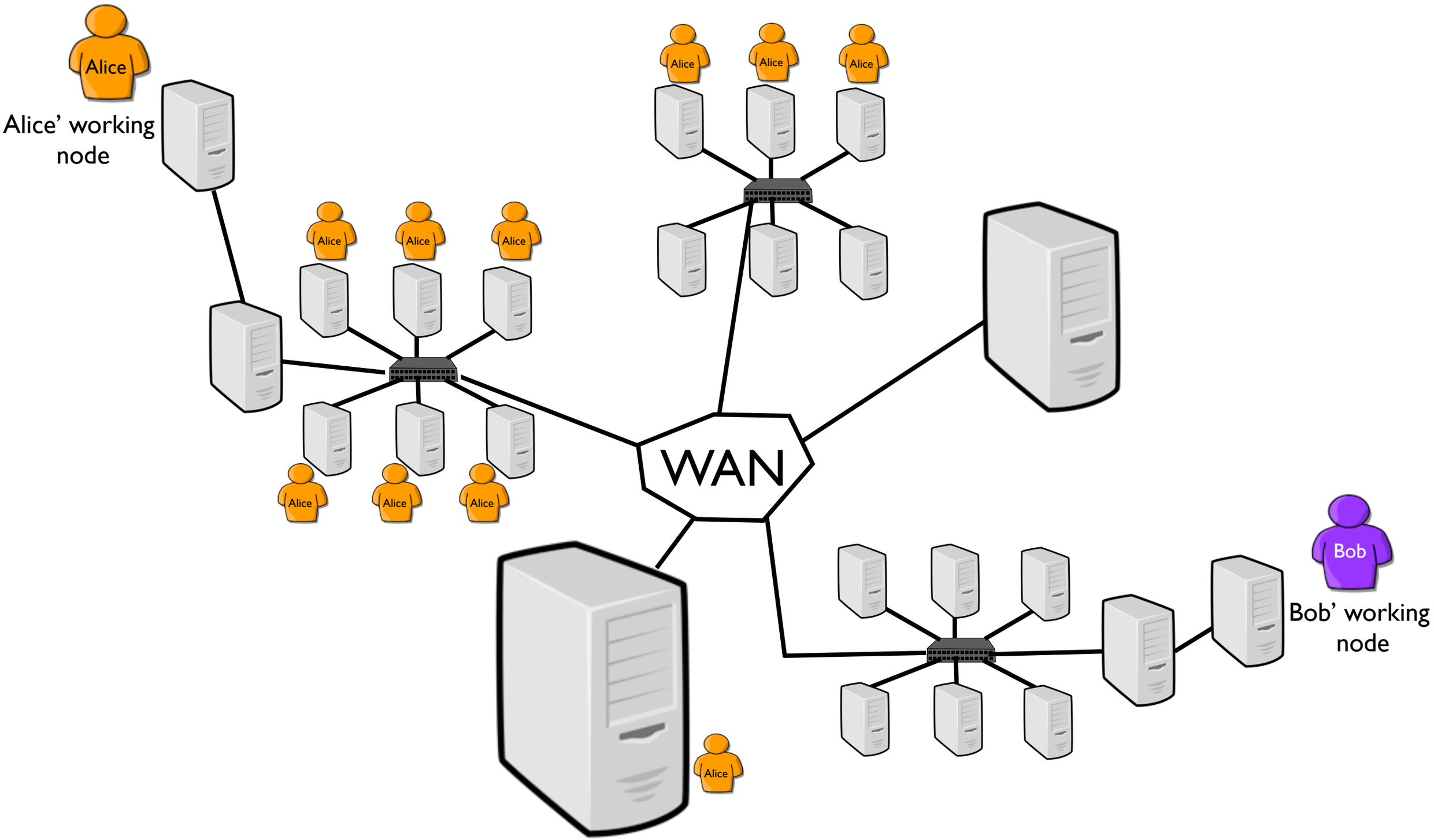
- A common objective

  provide computing resources (both hardware and software)
  in a flexible, transparent, secure, ... way

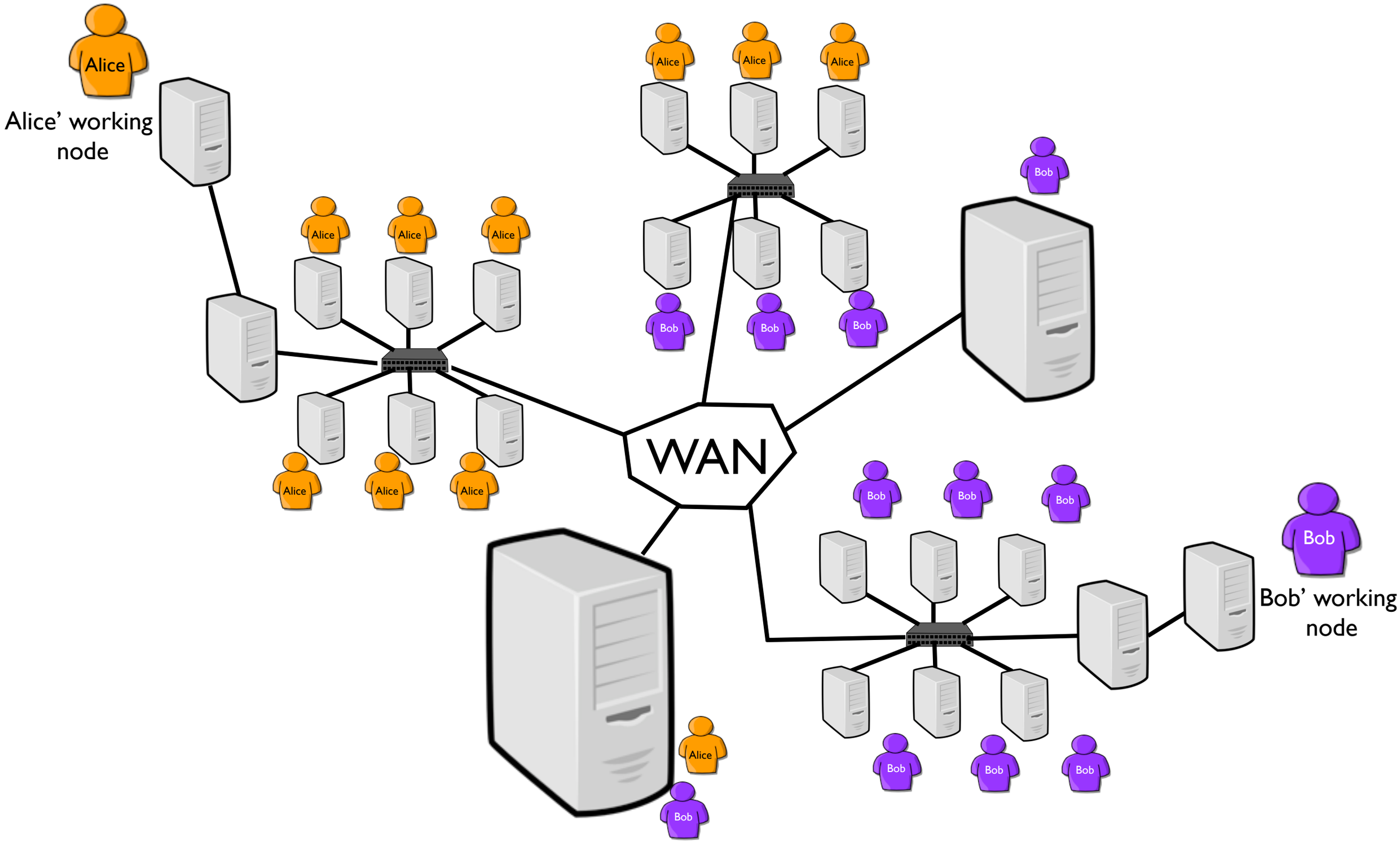
The Alice/Bob Example



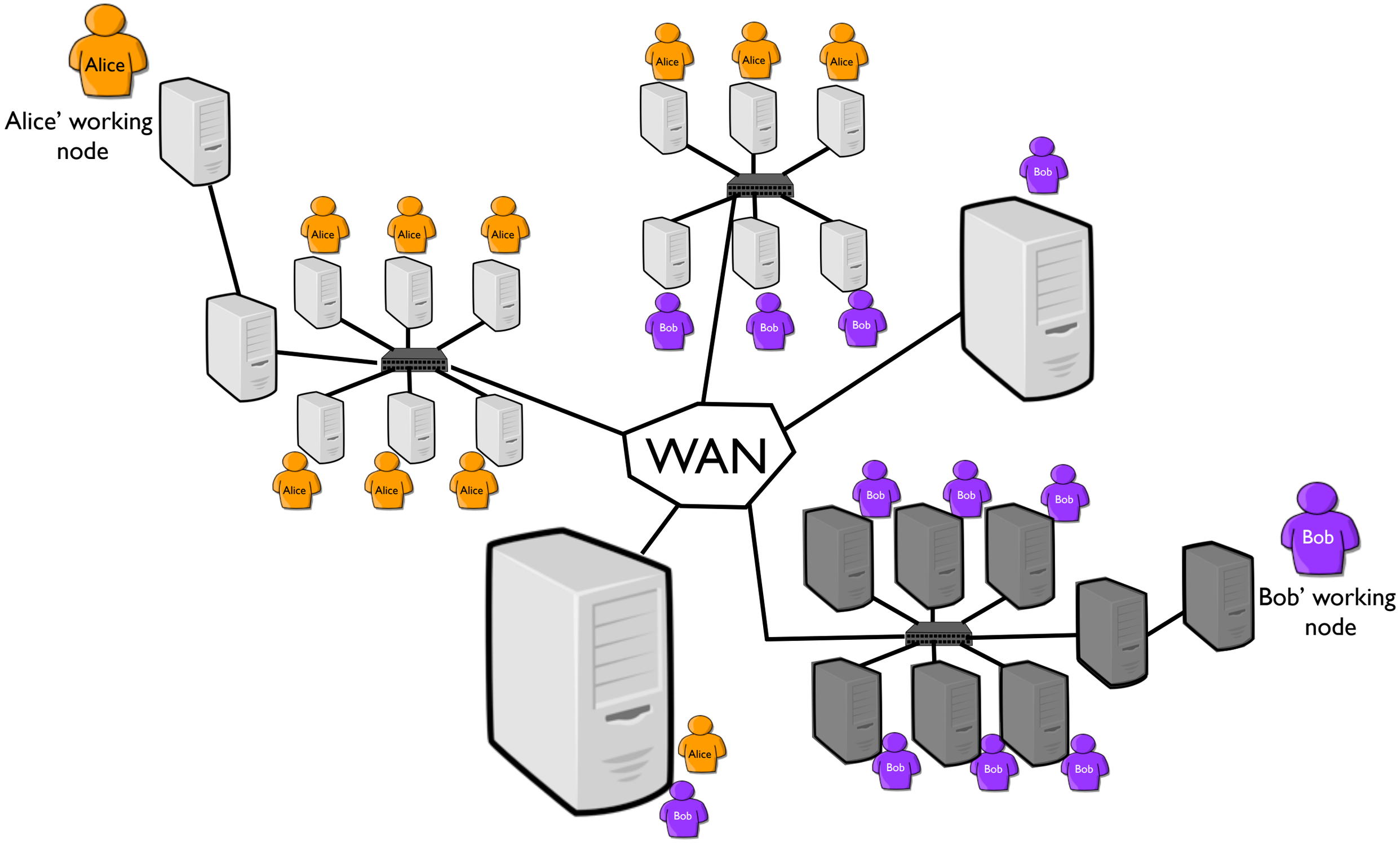
The Alice/Bob Example



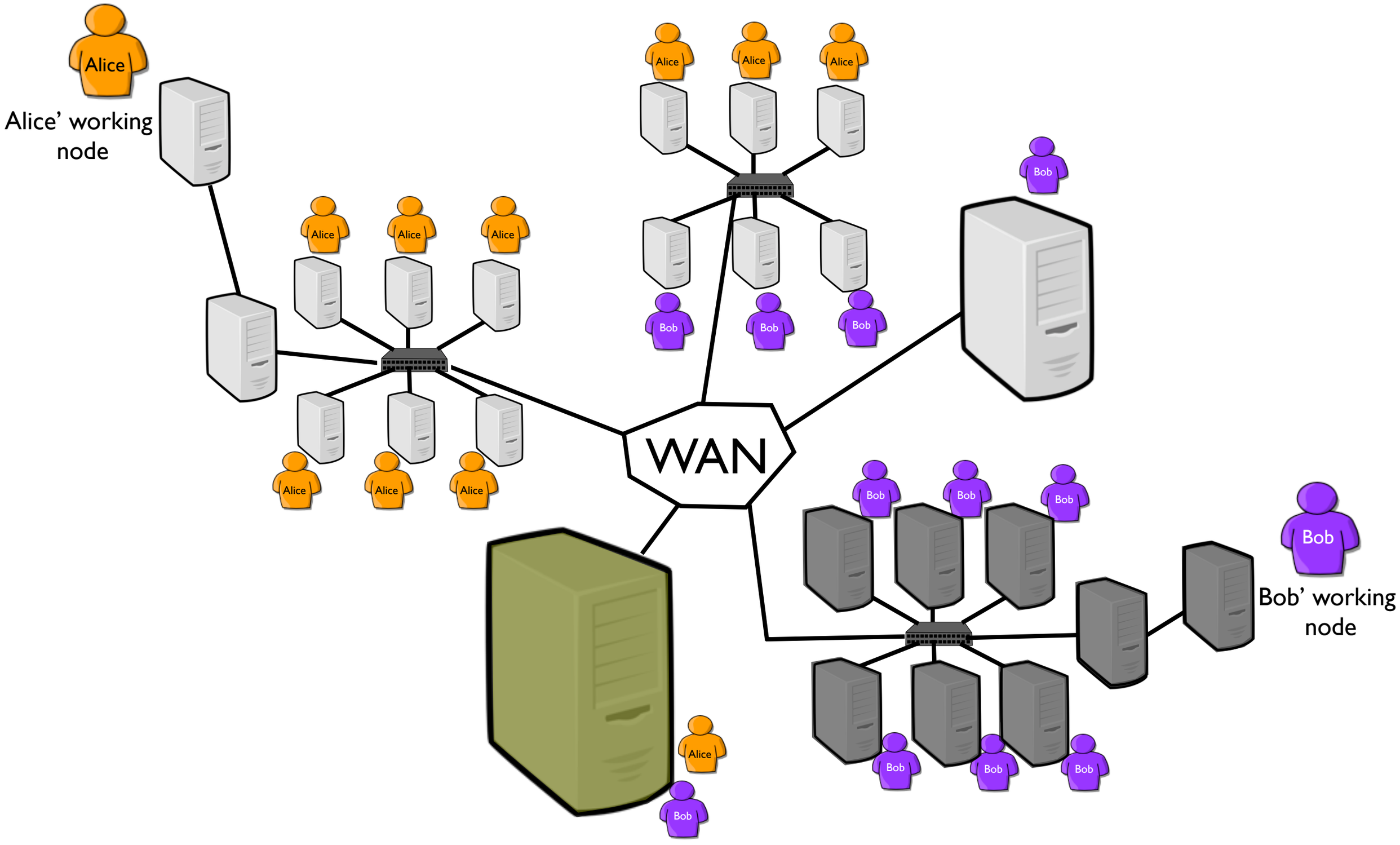
The Alice/Bob Example



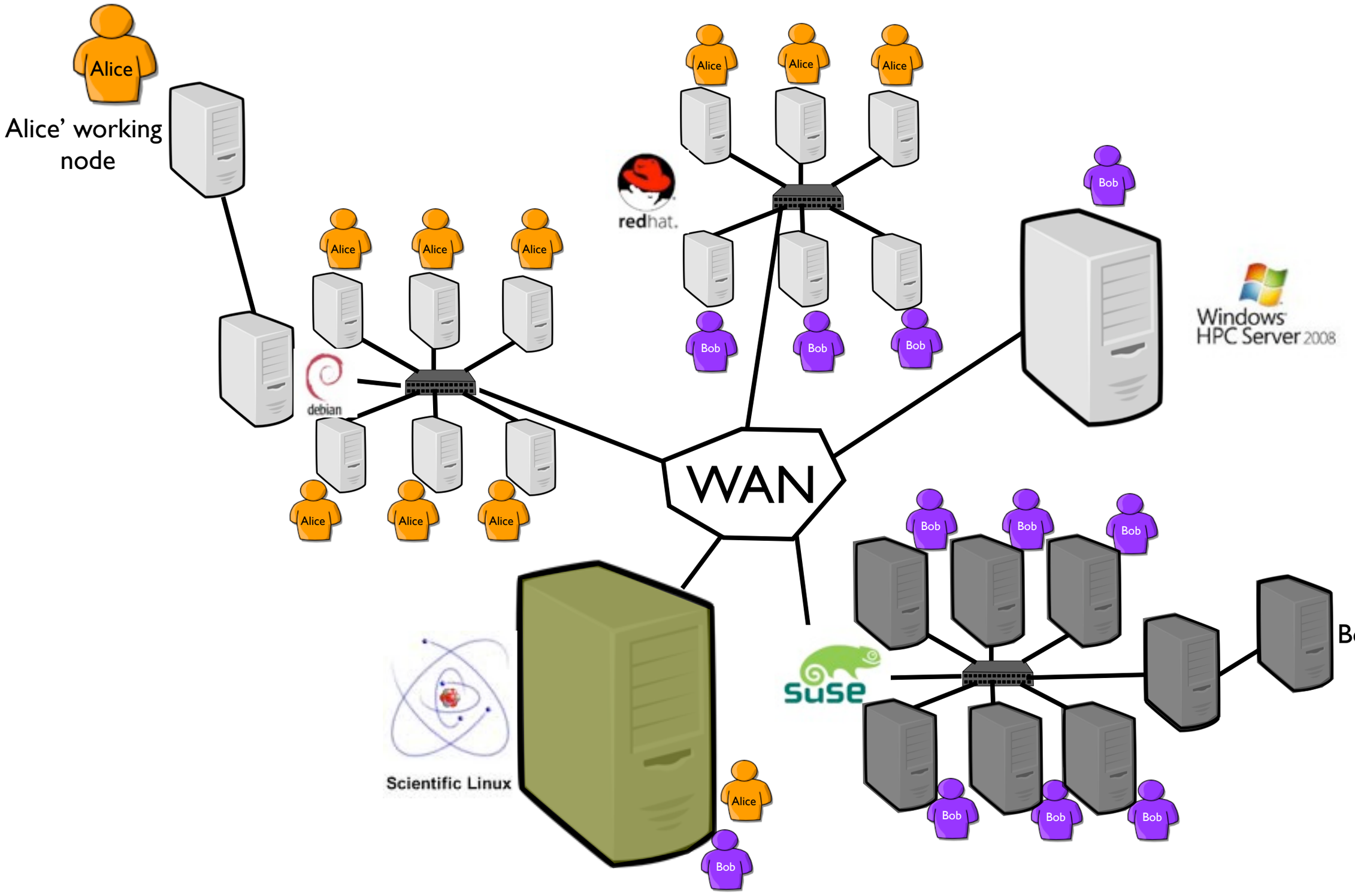
The Alice/Bob Example



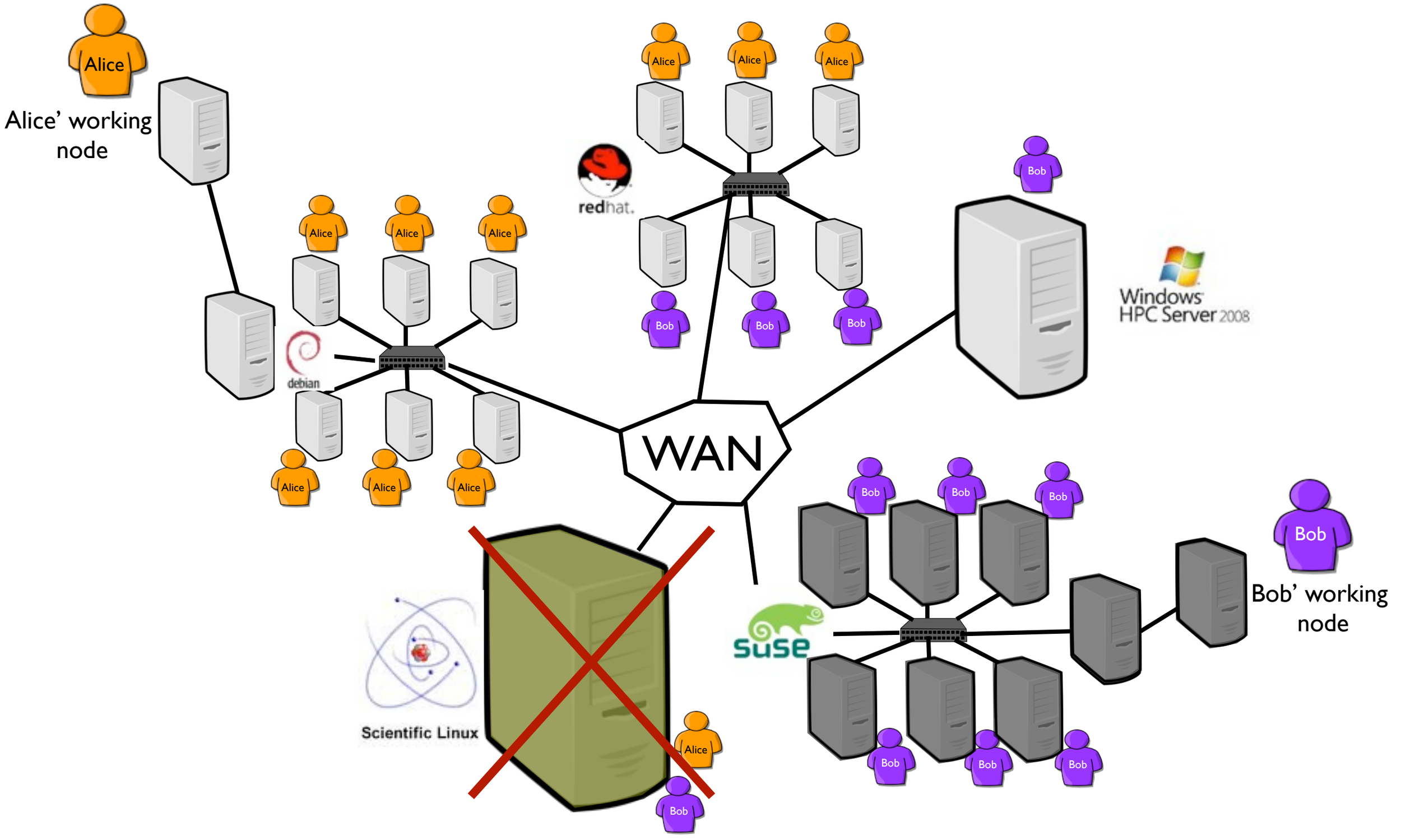
The Alice/Bob Example



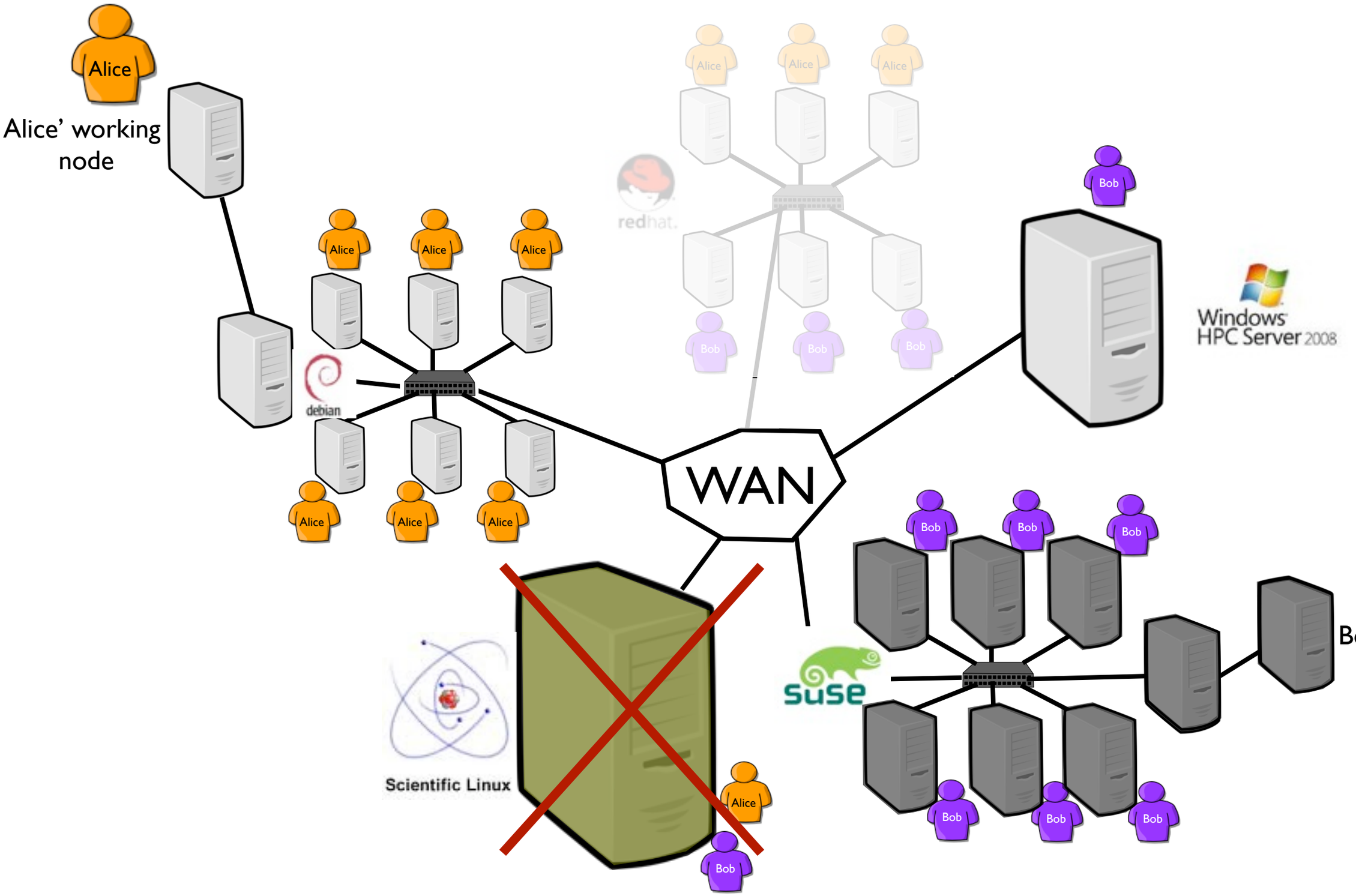
The Alice/Bob Example



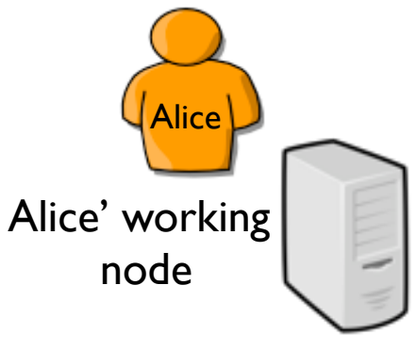
The Alice/Bob Example



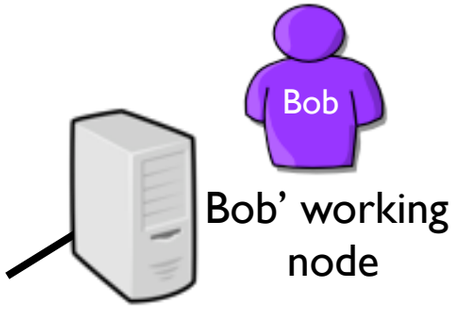
The Alice/Bob Example



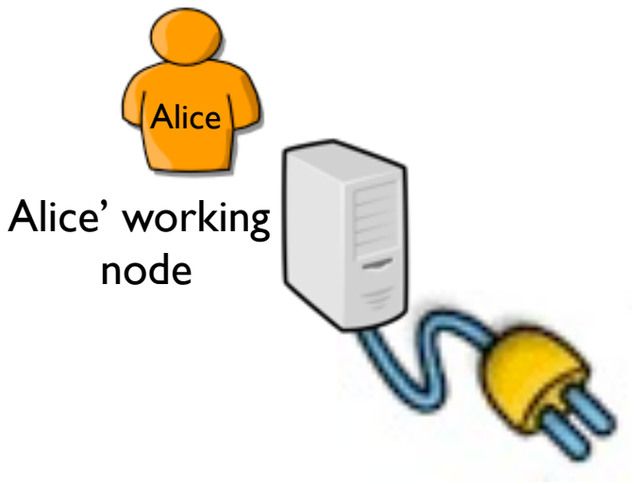
What a Grid ! ? !



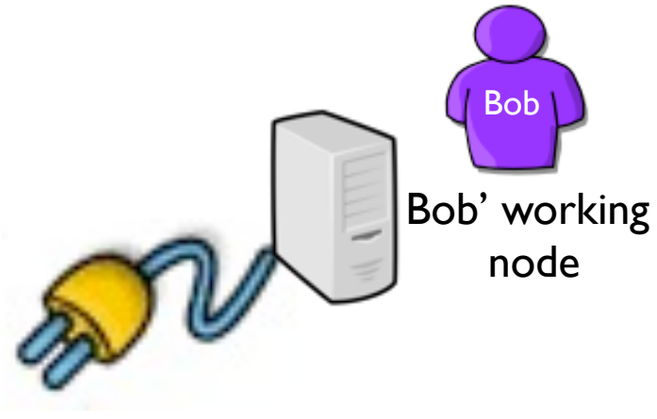
Resource booking (based on user's estimates)
Security concerns (job isolation)
Heterogeneity concerns (hardware and software)
Scheduling limitations (a job cannot be easily relocated)
Fault tolerance issues
...



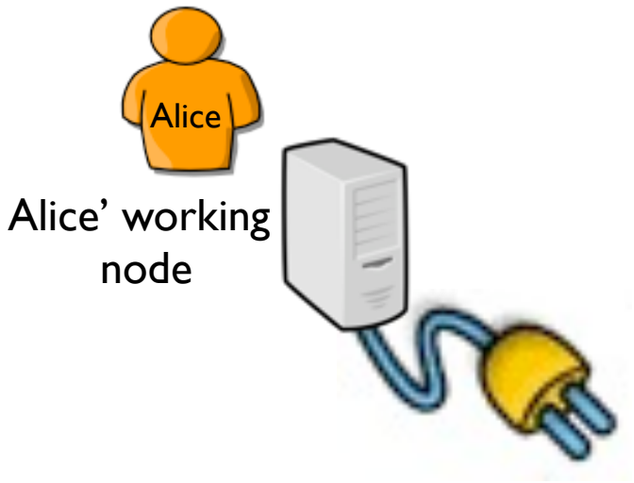
What a Grid ! ? !



Resource booking (based on user's estimates)
Security concerns (job isolation)
Heterogeneity concerns (hardware and software)
Scheduling limitations (a job cannot be easily relocated)
Fault tolerance issues
...



What a Grid ! ? !

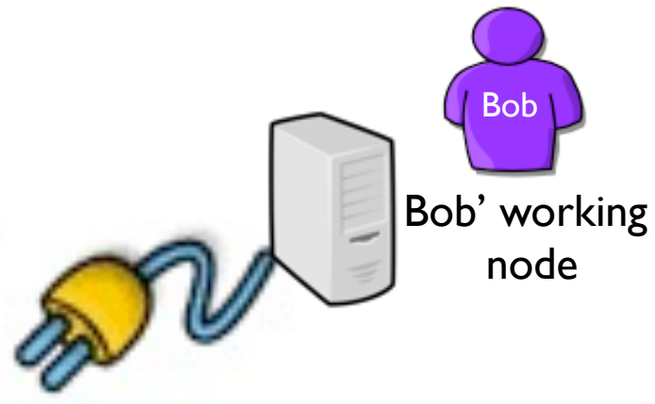


Resource b...

A lot of progress has been done since the 90's and several proposals partially addressed these concerns.

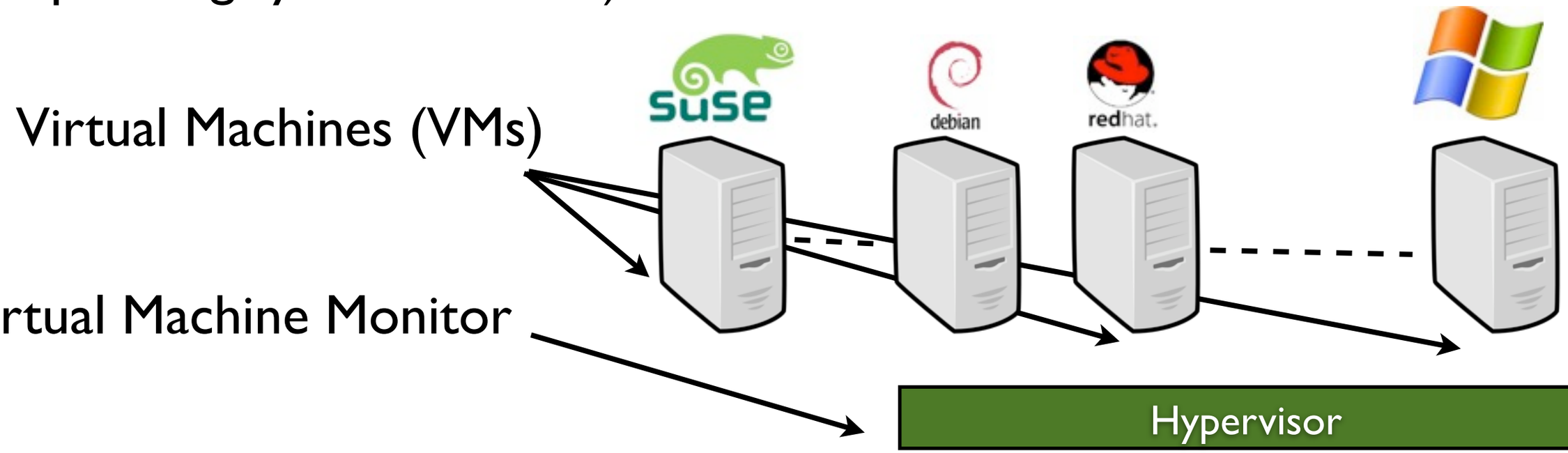
However none of them is mature enough and Strong limitations still persist !

...



Here Comes *System Virtualization*

- One to multiple OSes on a physical node thanks to a hypervisor (an operating system of OSes)

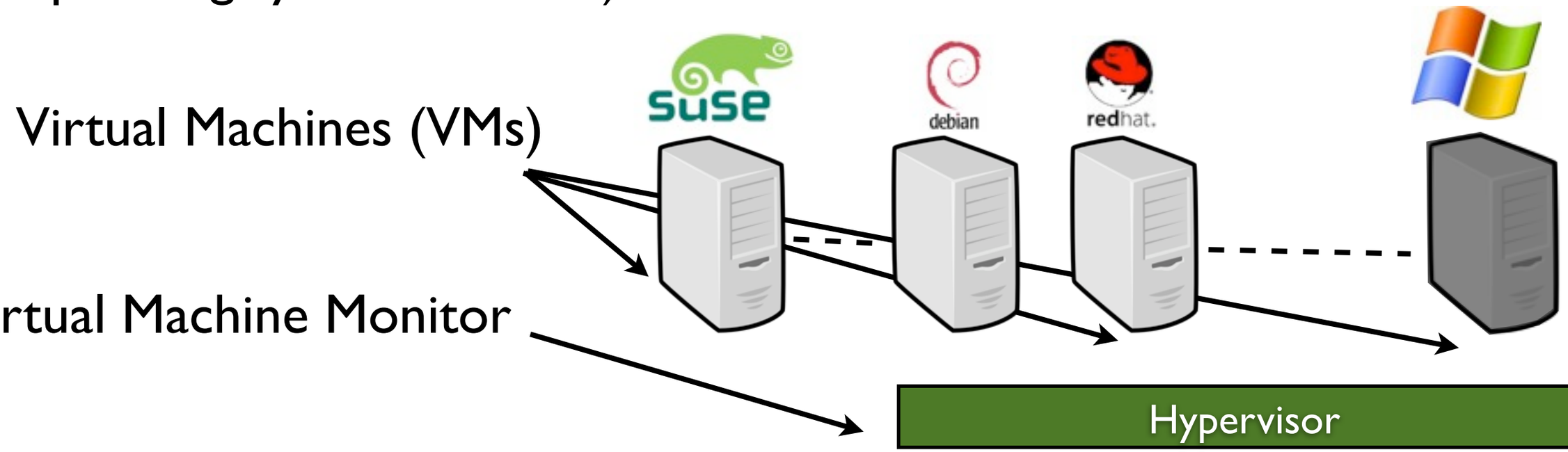


"A ***virtual machine*** (VM) provides a faithful implementation of a physical processor's hardware running in a protected and isolated environment.
Virtual machines are created by a software layer called the ***virtual machine monitor*** (VMM) that runs as a privileged task on a physical processor."

Physical Machine (PM)

Here Comes *System Virtualization*

- One to multiple OSes on a physical node thanks to a hypervisor (an operating system of OSes)



"A ***virtual machine*** (VM) provides a faithful implementation of a physical processor's hardware running in a protected and isolated environment.
Virtual machines are created by a software layer called the ***virtual machine monitor*** (VMM) that runs as a privileged task on a physical processor."

Virtualization History

- Proposed in the 60's by IBM

  More than 70 publications between 66 and 73

  *"Virtual Machines have finally arrived. Dismissed for a number of years as merely academic curiosities, **they are now seen as cost-effective techniques for organizing computer systems resources to provide extraordinary system flexibility** and support for certain unique applications" .*

  Goldberg, Survey of Virtual Machine Research, 1974

Virtualization History

- The 80's

  No real improvements
  Virtualization seems given up

- End of the 90's:

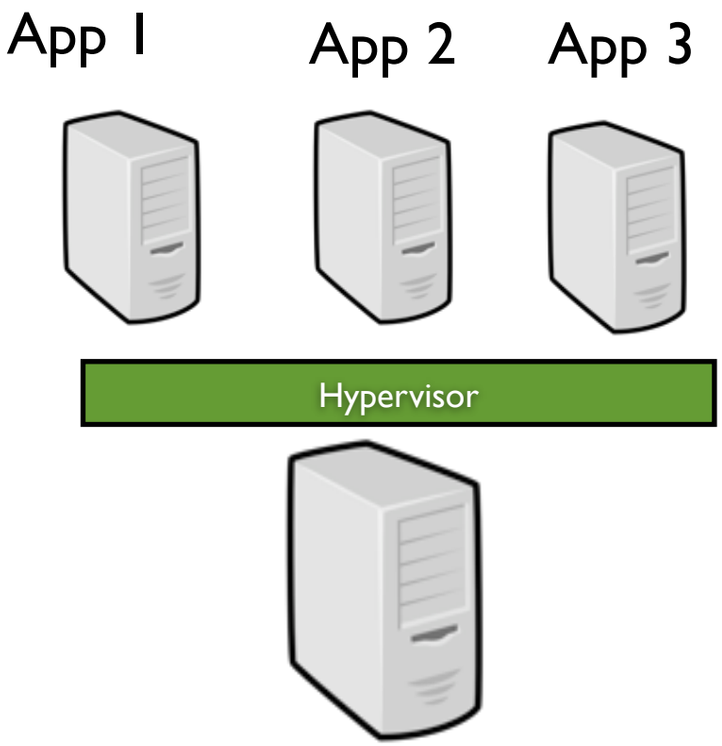
  HLL-VM : High-Level Language VM
  Java and its famous JVM !

  Virtual Server: Exploit for Web hosting
  (Linux `chroot` / containers)

  Revival of System Virtualization approach (VmWare/Xen)

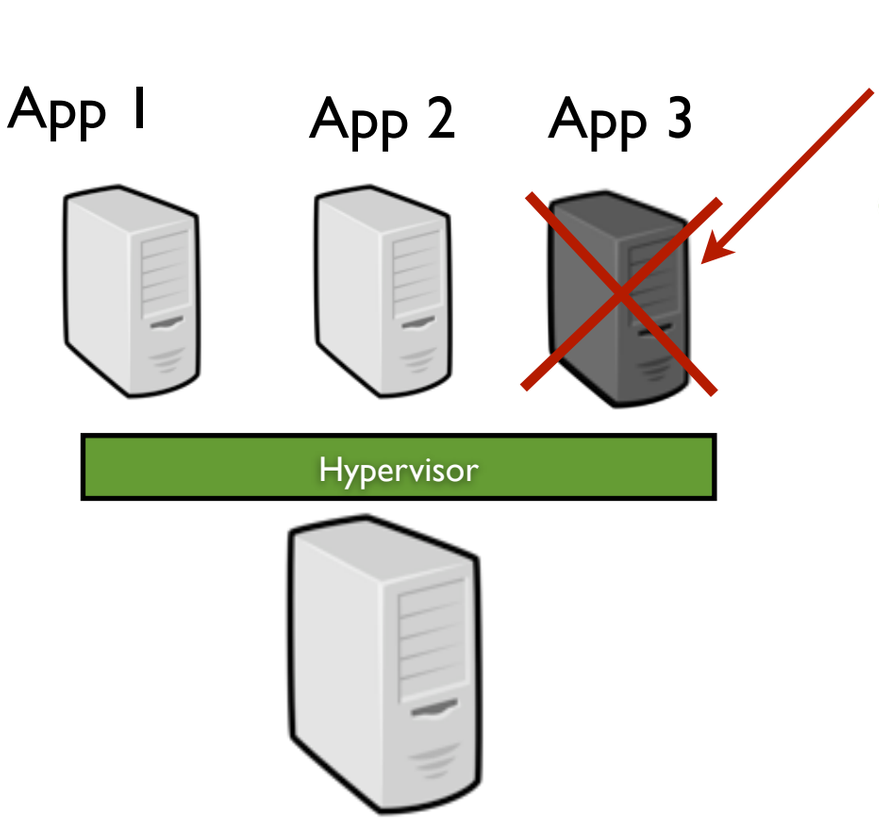
  Hard or soft partitioning of SMP/Numa Server

VM Capabilities



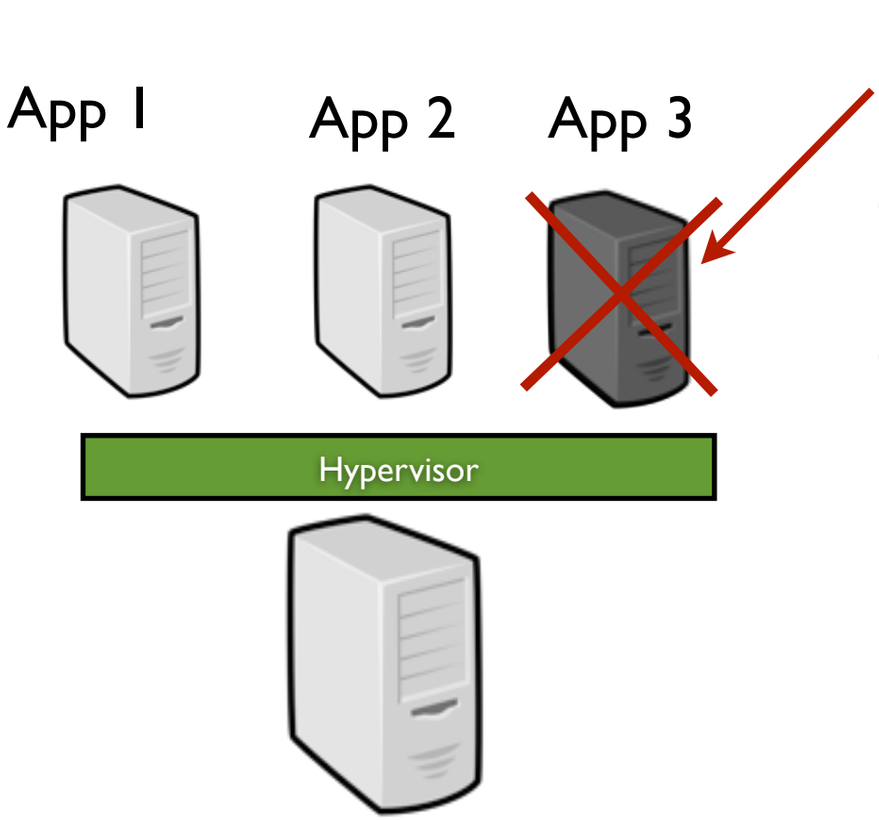
- Isolation ("security" between each VM)

VM Capabilities



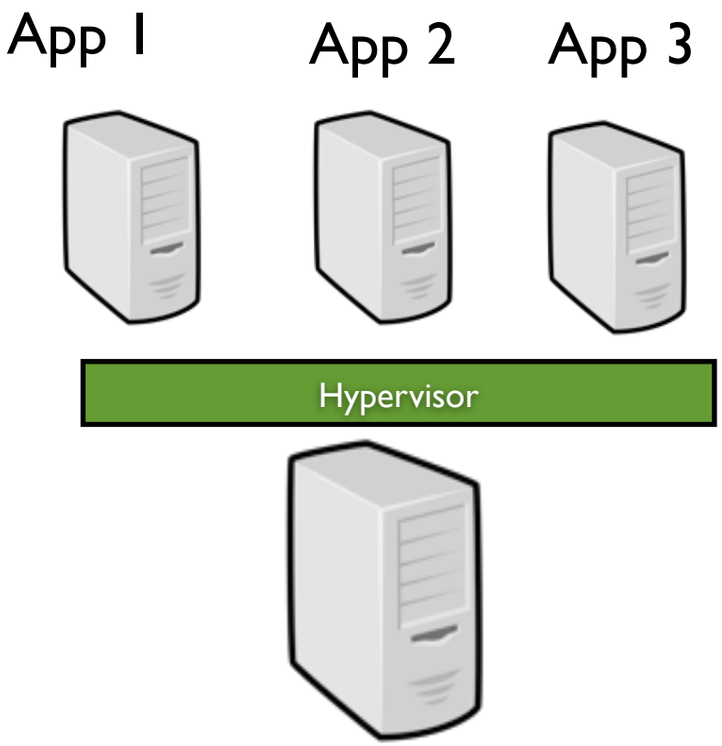
- Isolation ("security" between each VM)

VM Capabilities



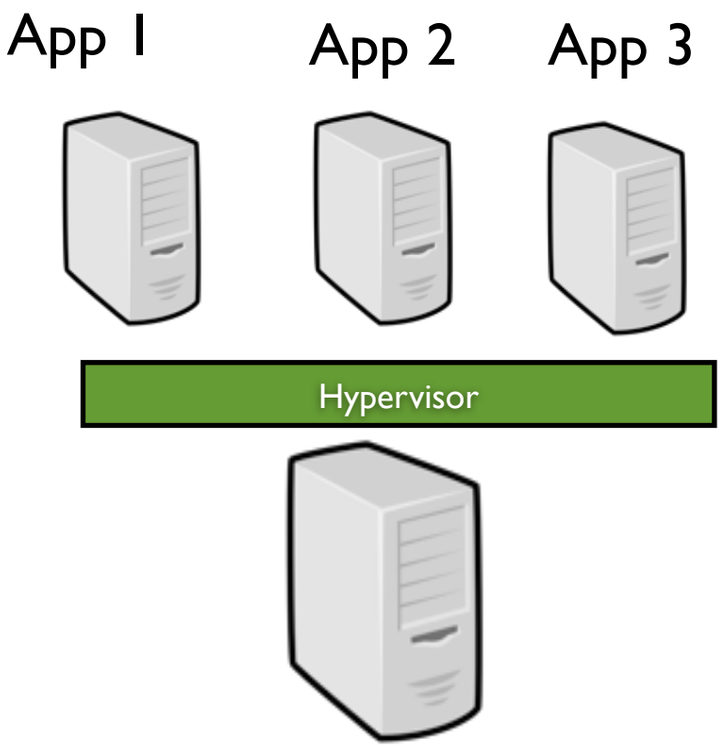
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

VM Capabilities



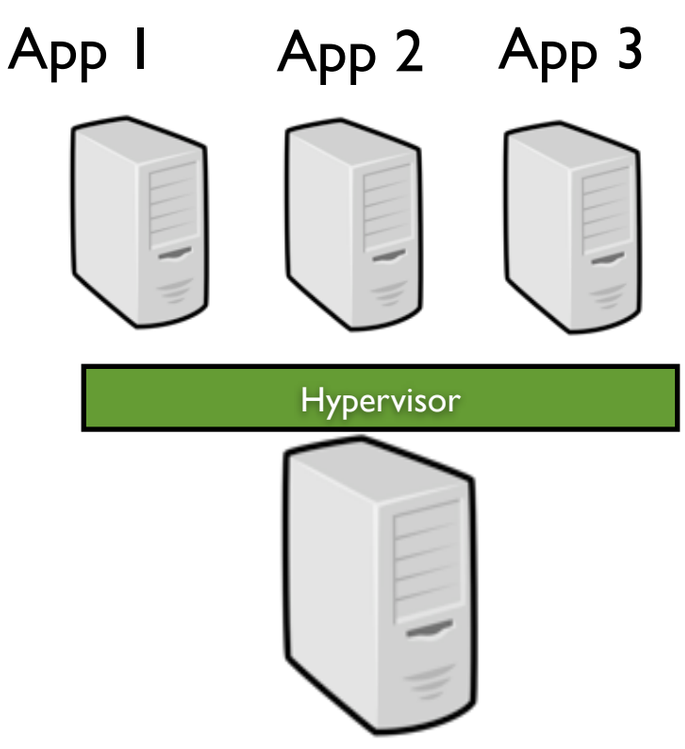
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

VM Capabilities

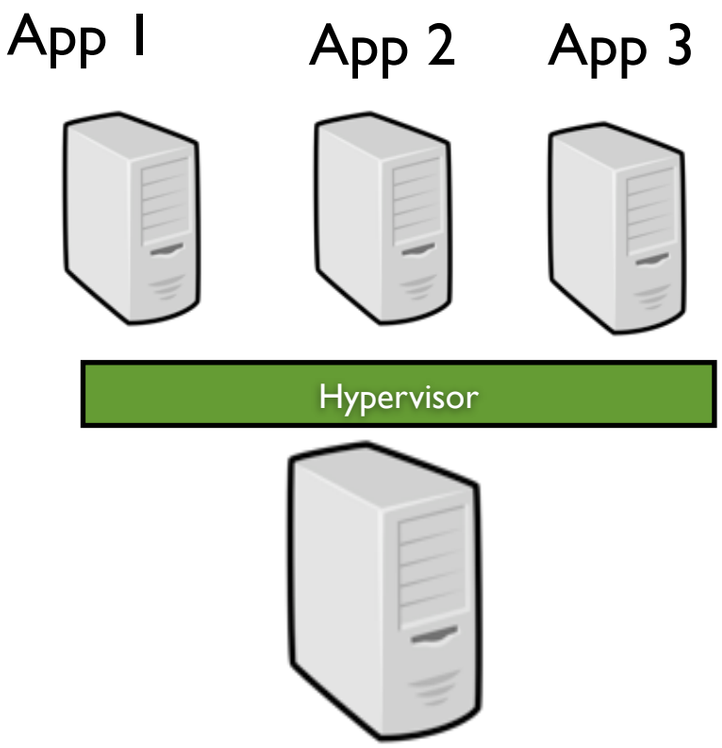


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

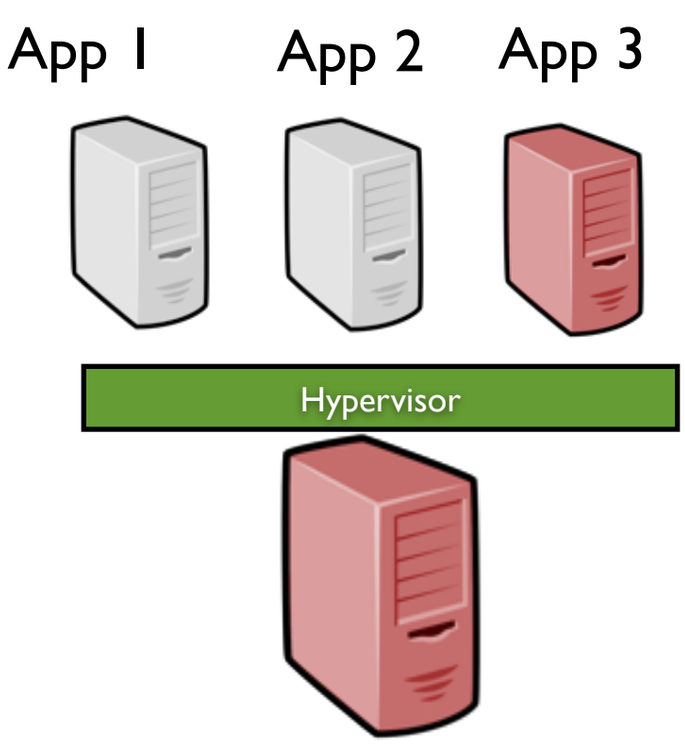
- Suspend/Resume



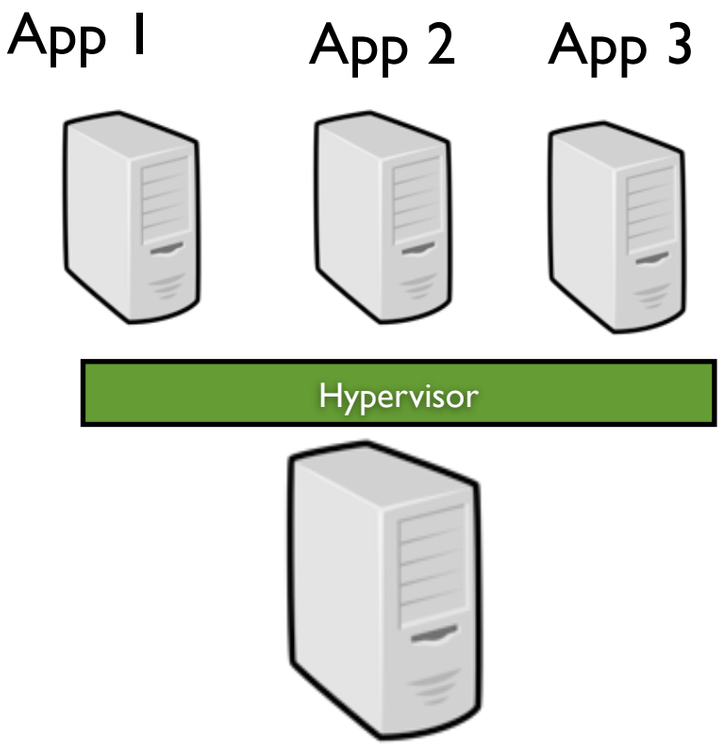
VM Capabilities



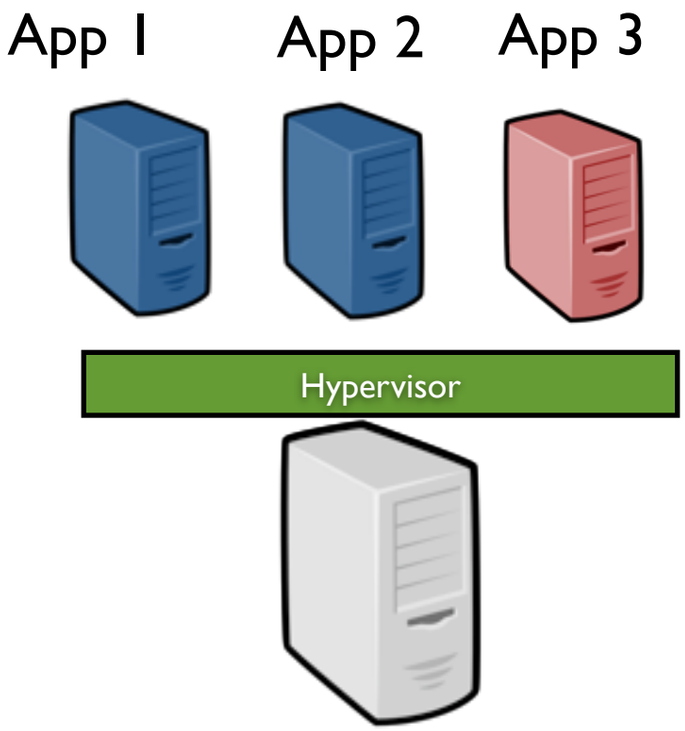
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)
- Suspend/Resume



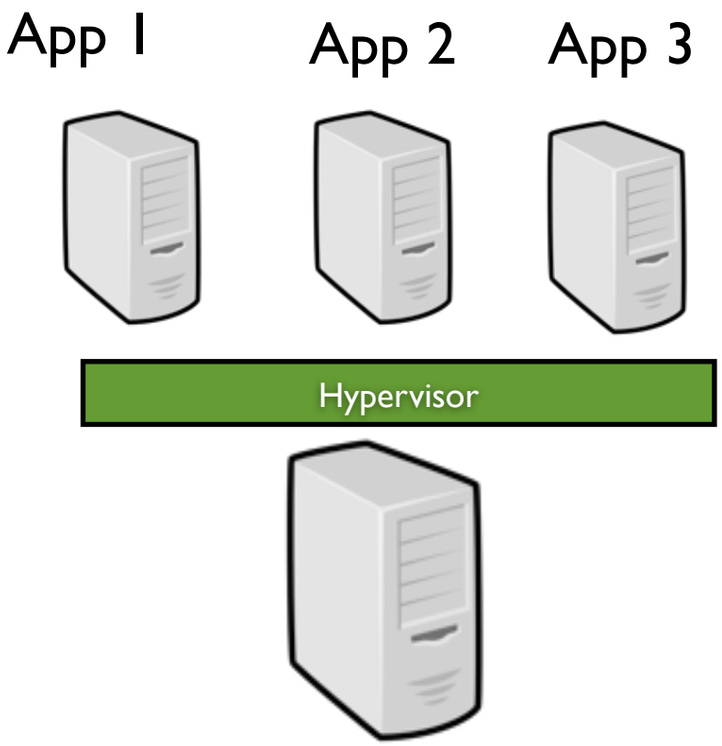
VM Capabilities



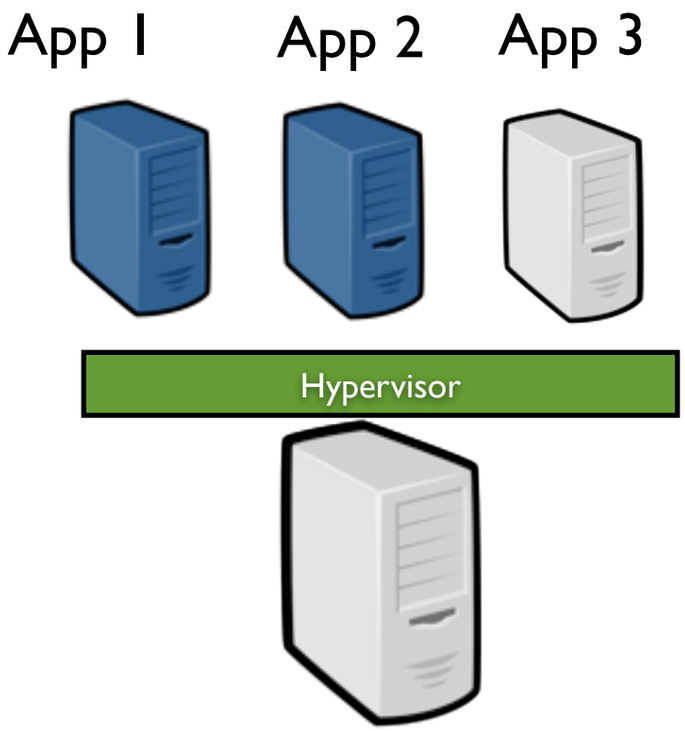
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)
- Suspend/Resume



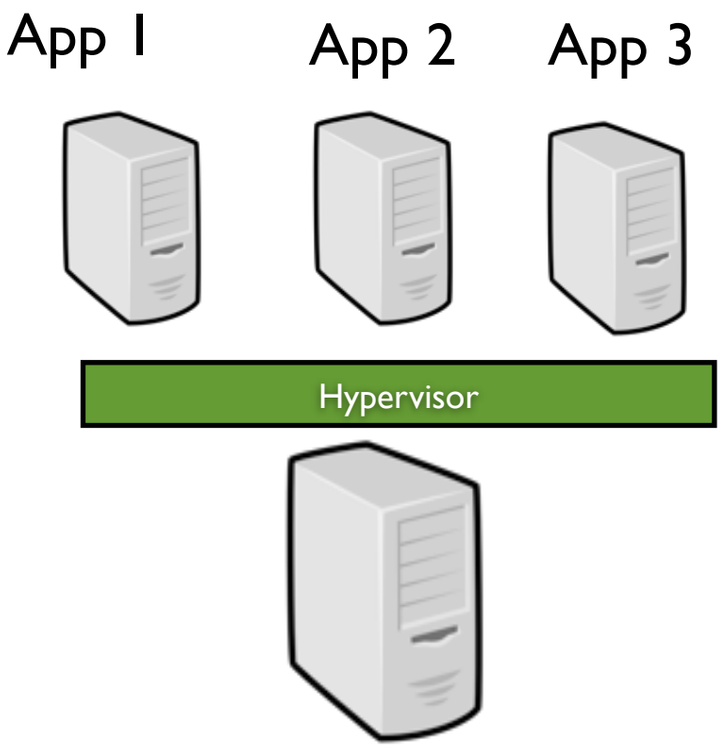
VM Capabilities



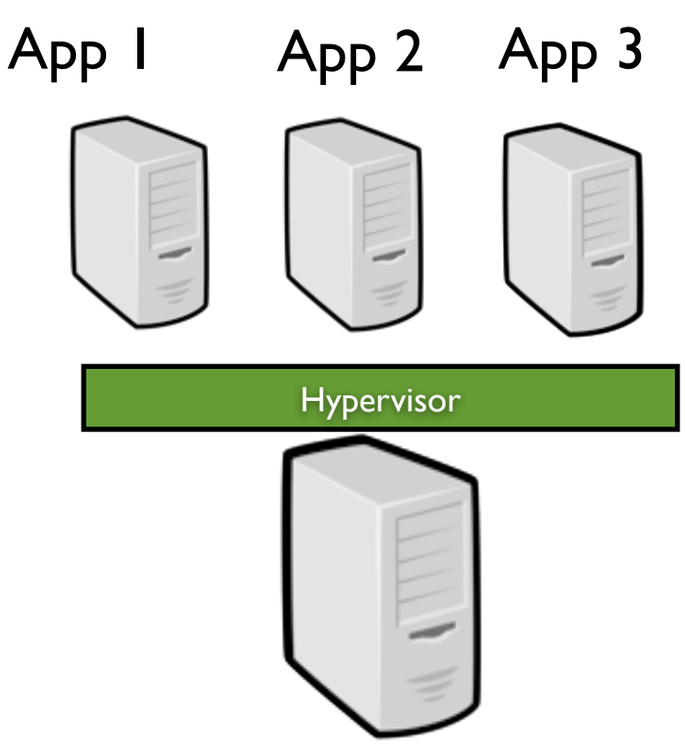
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)
- Suspend/Resume



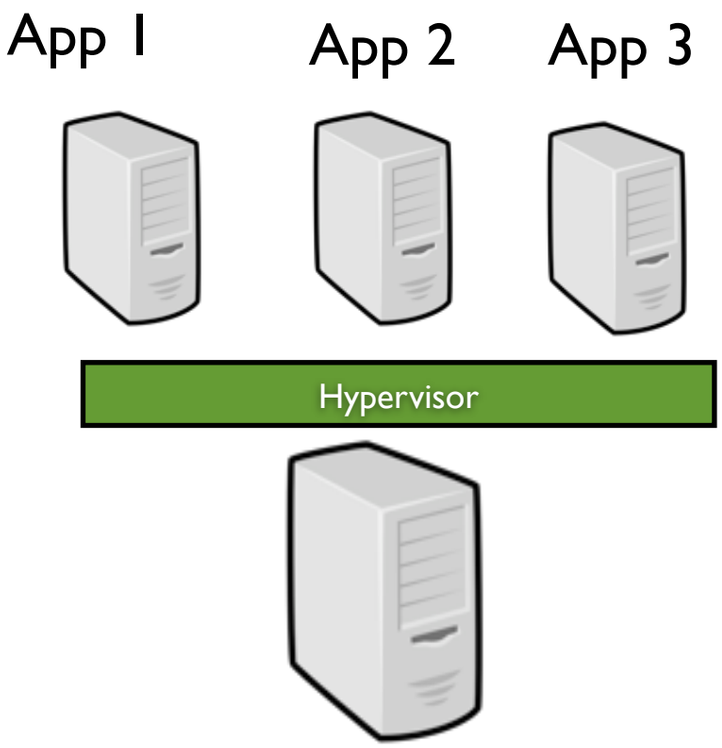
VM Capabilities



- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)
- Suspend/Resume



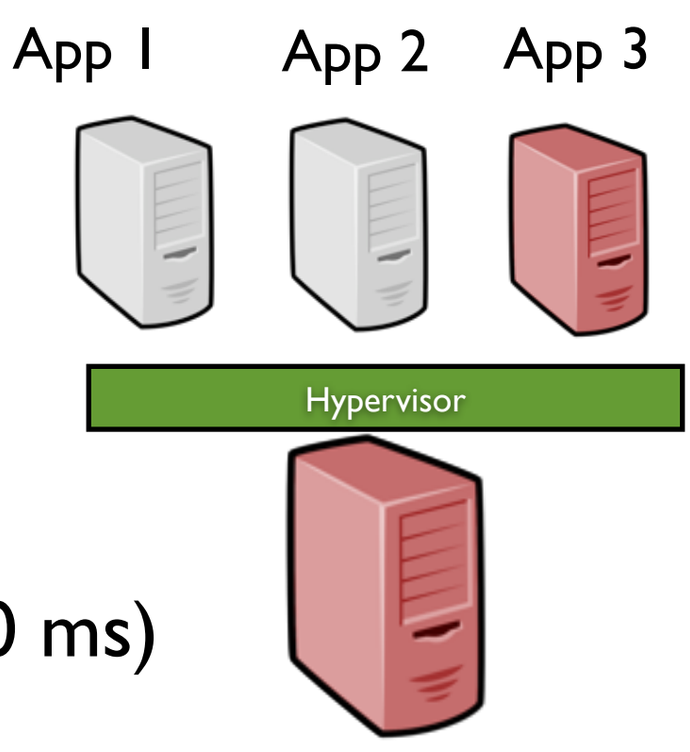
VM Capabilities



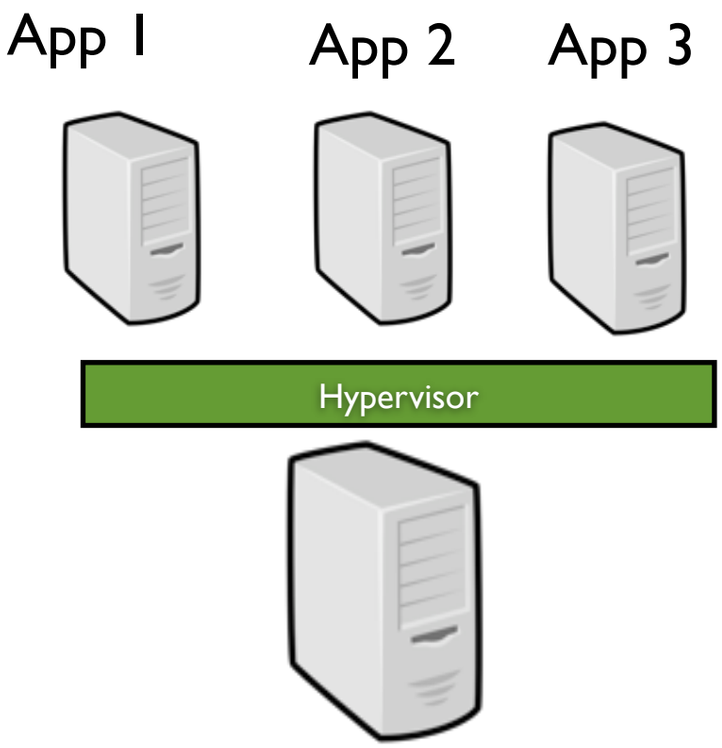
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

- Suspend/Resume

- Live migration
(negligible downtime ~ 60 ms)



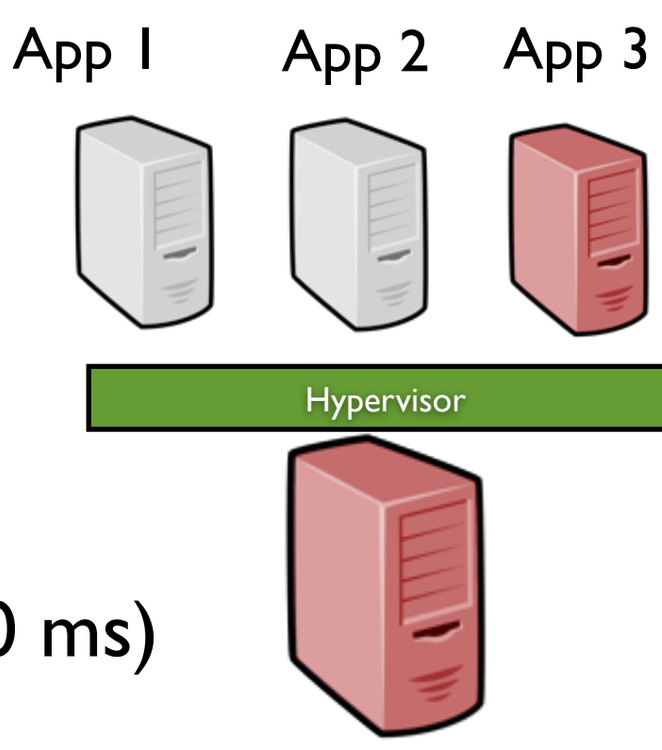
VM Capabilities

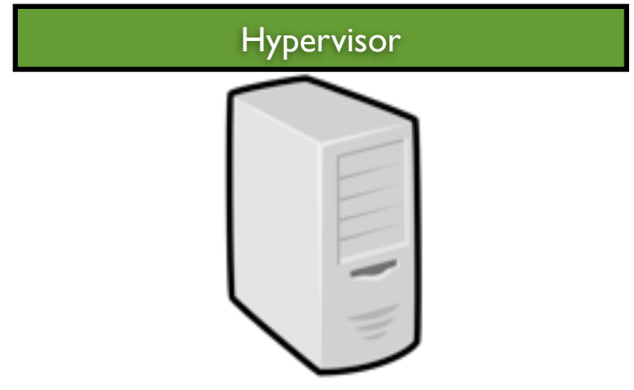


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

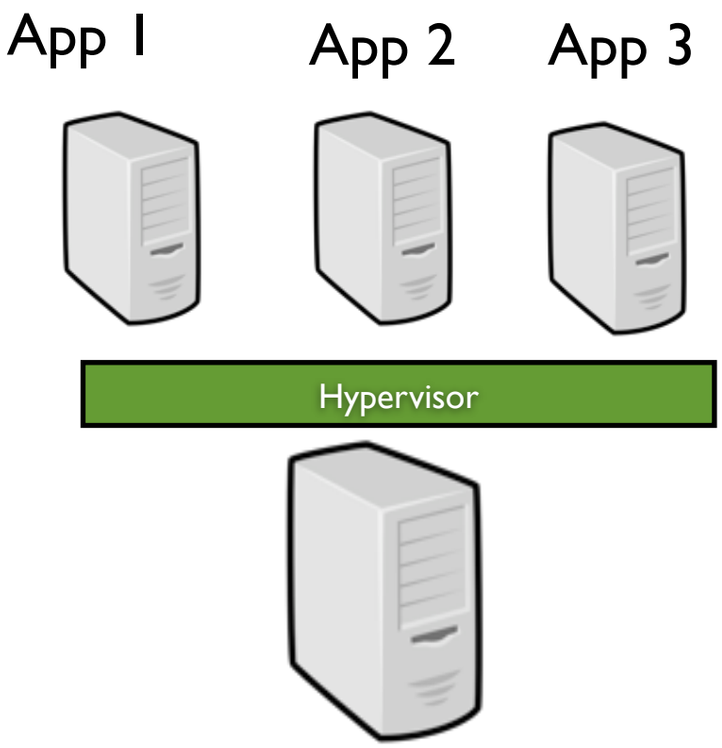
- Suspend/Resume

- Live migration
(negligible downtime ~ 60 ms)





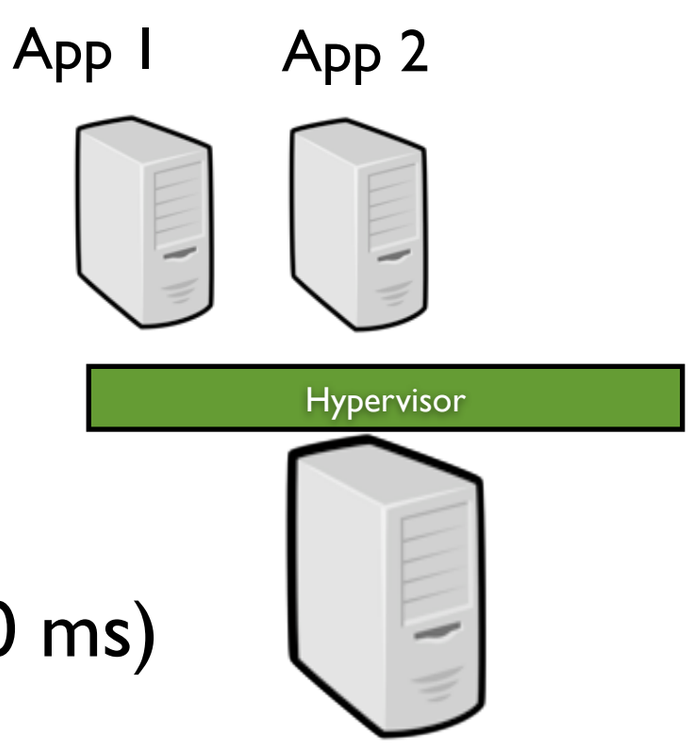
VM Capabilities

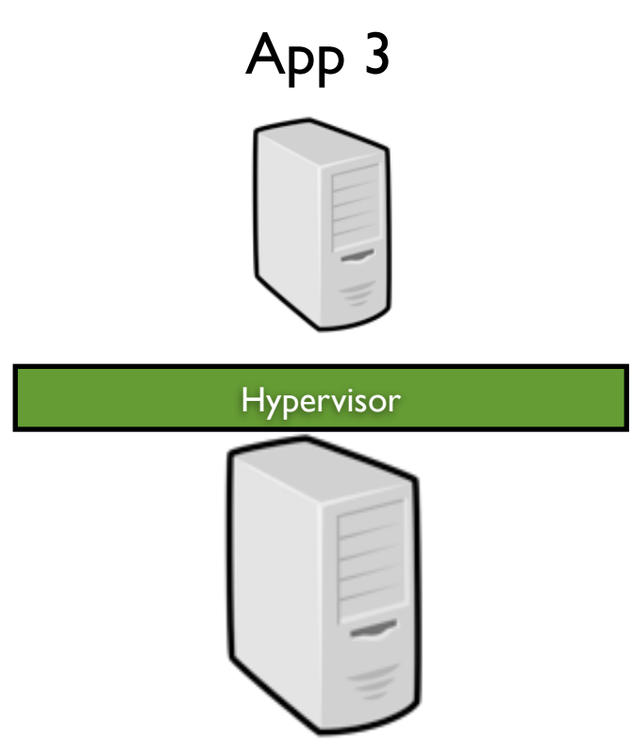


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

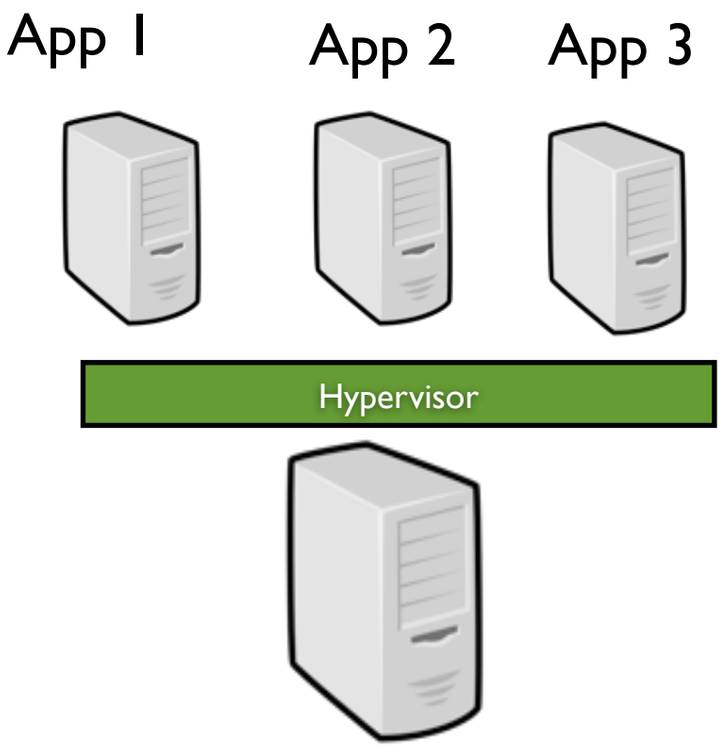
- Suspend/Resume

- Live migration
(negligible downtime ~ 60 ms)





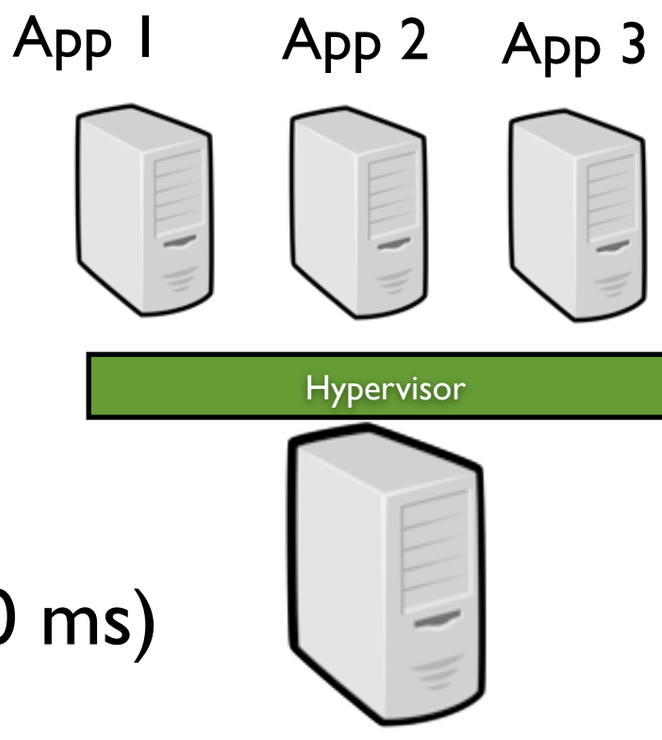
VM Capabilities

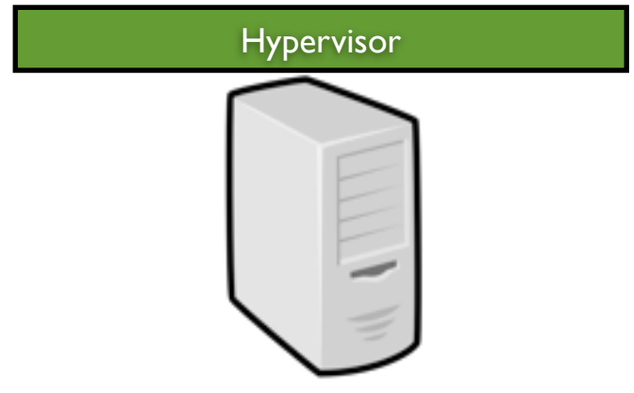


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

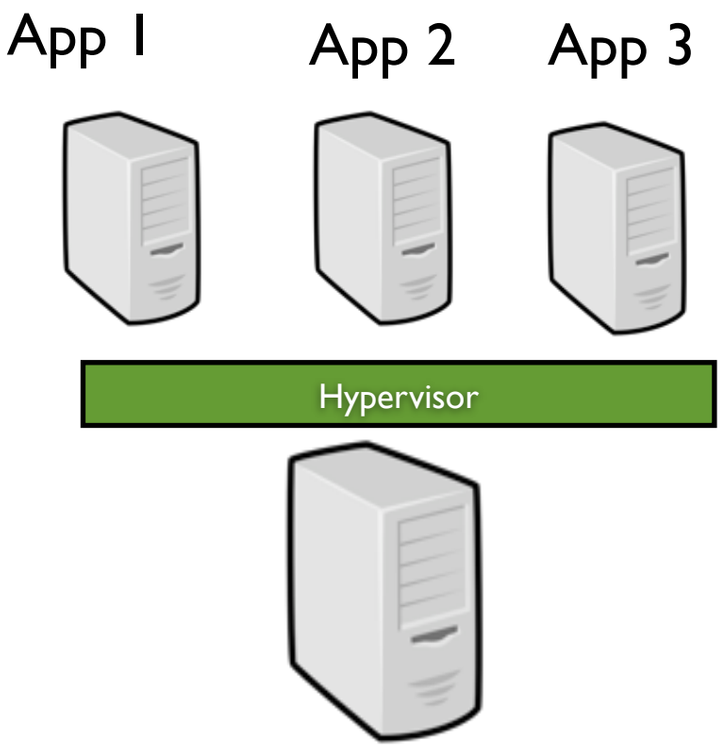
- Suspend/Resume

- Live migration
(negligible downtime ~ 60 ms)





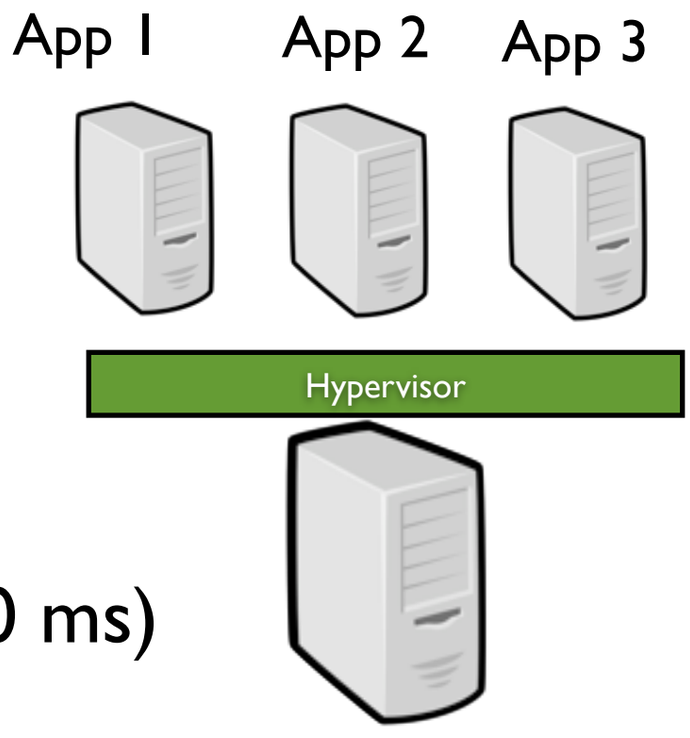
VM Capabilities



- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resume from its latest consistent state)

- Suspend/Resume

- Live migration
(negligible downtime ~ 60 ms)



Let's start

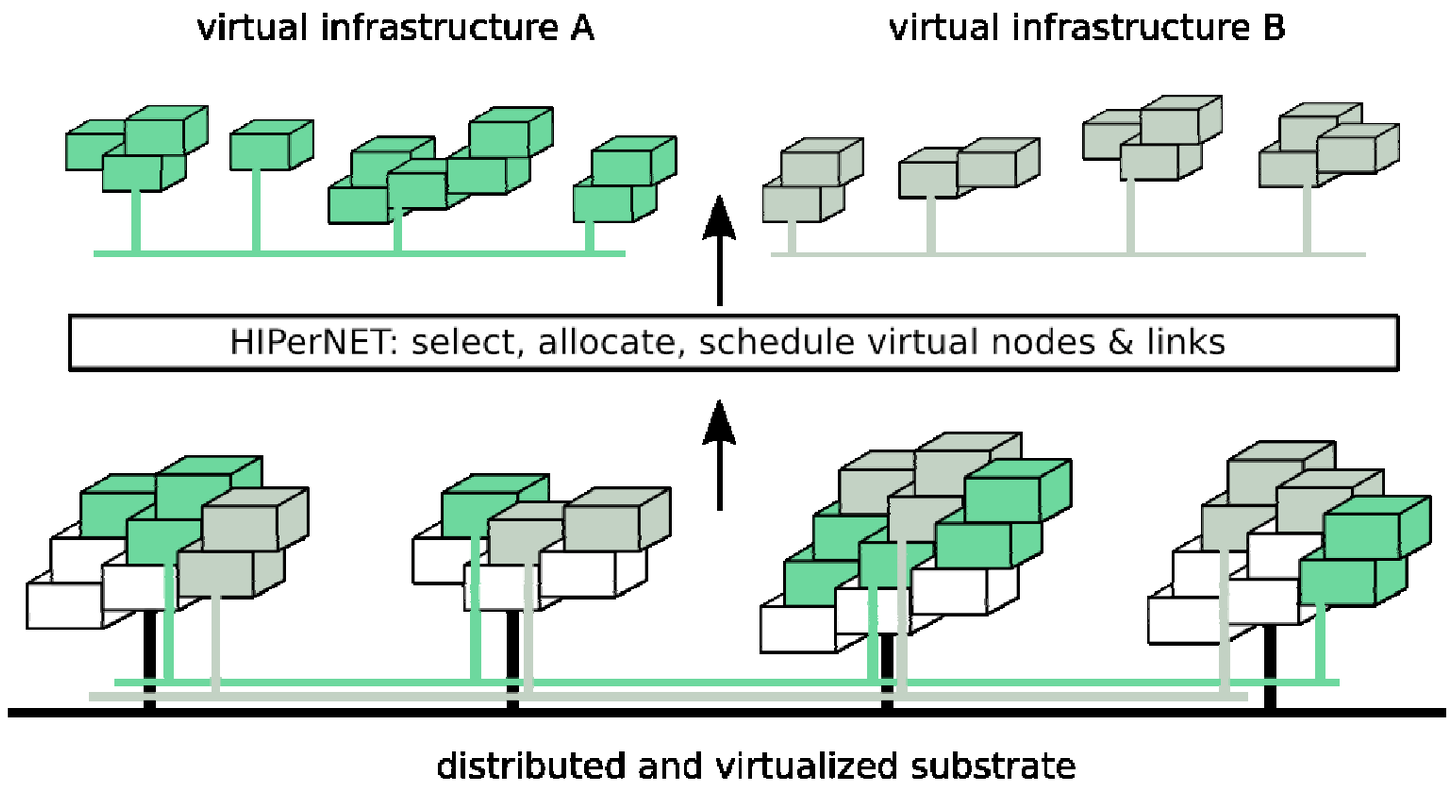
- Isolation to build distributed infrastructures according to user's expectations.
- Suspend/Resume and Live migration to implement advanced scheduling strategies
- Snapshotting to better address fault tolerance concerns

Isolation Capability

- The HiperNet proposal

  ANR HipCal (2007-2010, http://hipcal.lri.fr), INRIA Reso team,

  Combine resource virtualization and network virtualization to give the user the illusion he is using a private distributed system, while in reality he is using multiple systems parts of a virtualized physical substrate



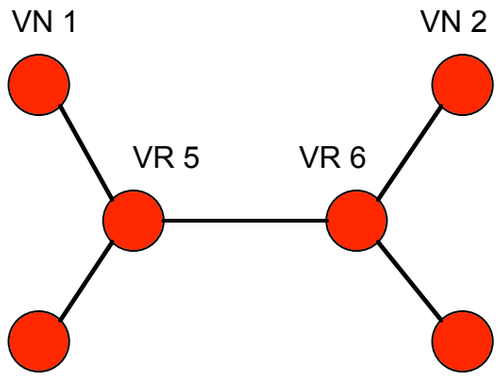
The HIPerNet Proposal

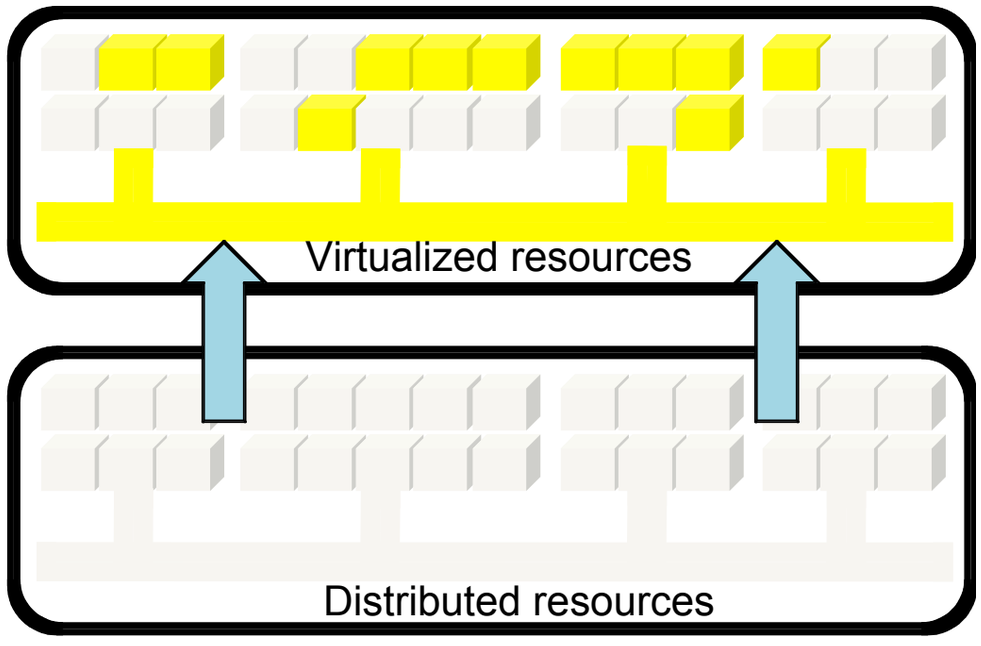
- HIPerNet is a software solution virtualizing a physical infrastructure and orchestrating the virtual infrastructures composed and provisioned over it.
- At the lowest level, the HIPerSpace manager plays the role of an ”infrastructure hypervisor”.

  It pilots a set of virtualized and exposed computing network resources (resources that can be managed by the HIPerNet framework) .

  At the highest level, HIPerNet manages ViPXis (VIrtual Private eXecution Infrastructures)

- A VipXi defines a federation of virtual capacities interconnected by a virtual network.





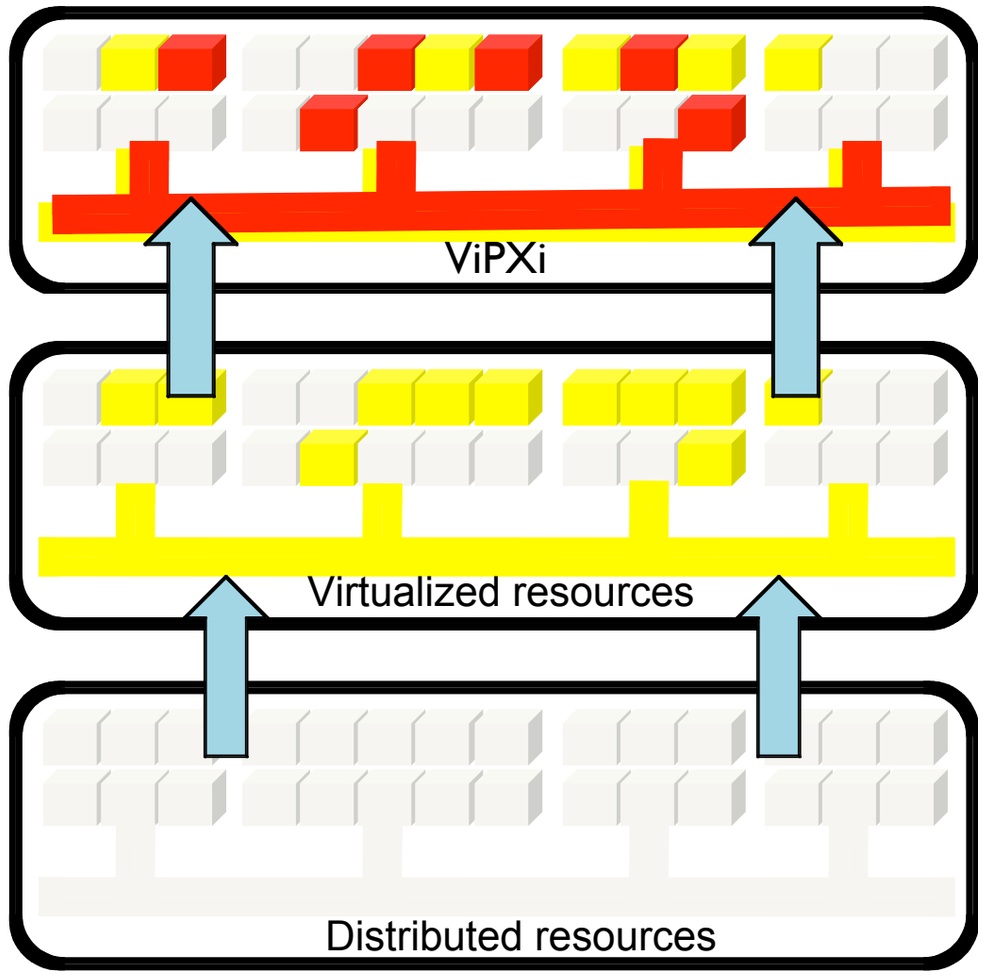
credits: G. Koslovski, INRIA Reso

The HIPerNet Proposal

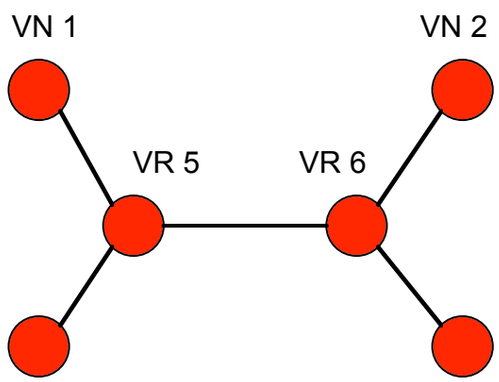
- HIPerNet is a software solution virtualizing a physical infrastructure and orchestrating the virtual infrastructures composed and provisioned over it.
- At the lowest level, the HIPerSpace manager plays the role of an "infrastructure hypervisor".

  It pilots a set of virtualized and exposed computing network resources (resources that can be managed by the HIPerNet framework) .

  At the highest level, HIPerNet manages ViPXis (VIrtual Private eXecution Infrastructures)



- A VipXi defines a federation of virtual capacities interconnected by a virtual network.



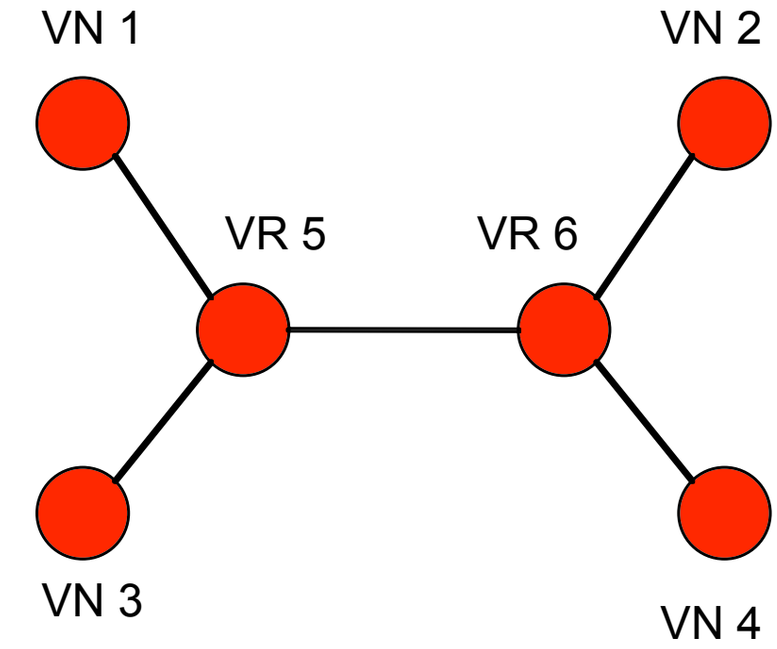
credits: G. Koslovski, INRIA Reso

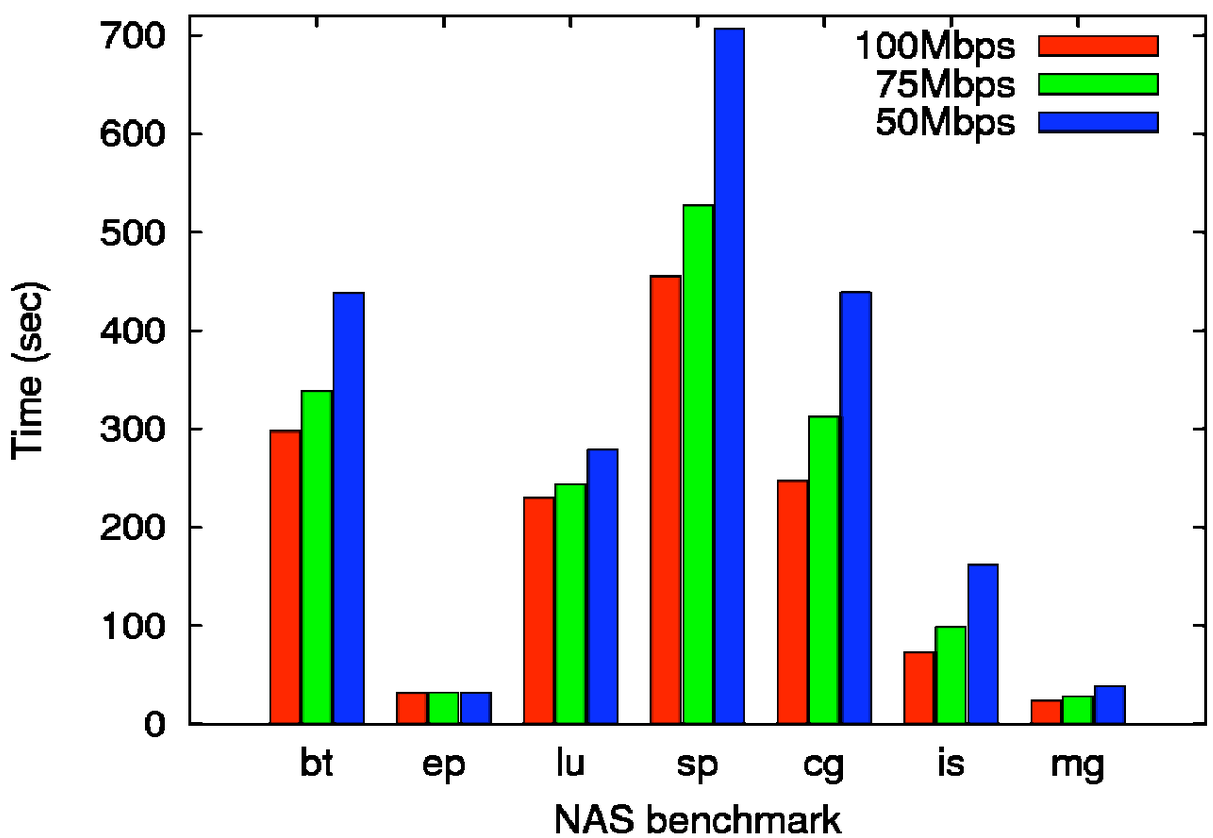
The HIPerNet Proposal

- ViPxi example

  Specification of three ViPXis

  Same resource set and topology:
  4 virtual nodes: VN1, VN2, VN3, and VN4
  2 virtual routers: VR5 and VR6
  Different links configuration





	VN X - VR Y	VR 5 - VR 6
ViPXi 1	100 Mbps	200 Mbps
ViPXi 2	75 Mbps	150 Mbps
ViPXi 3	50 Mbps	100 Mbps

The HIPerNet Proposal

- To sum up

Creates and manages confined virtual infrastructures, exploiting both resources and network virtualization

Execution using HIPerNet framework is straightforward.
All complexity is hidden to the user

More features are being integrated
(performance measurement, monitoring, GUI, …)

HIPerNet is being industrialized by the LYaTiss INRIA"s spinoff

Awarded by OSEO (emergence 2009, creation-developpement 2010)
Winner of French Tech Tour in Sillicon Valey (June 9th, 2010)

LYaTiss
software & services
for virtual infrastructures

Preemption/Migration Capabilities

- The Entropy proposal

  F. Hermenier, Ph.D. in CS (University of Nantes / 2009)
  Use of Live migration capability to finely exploit cluster resources

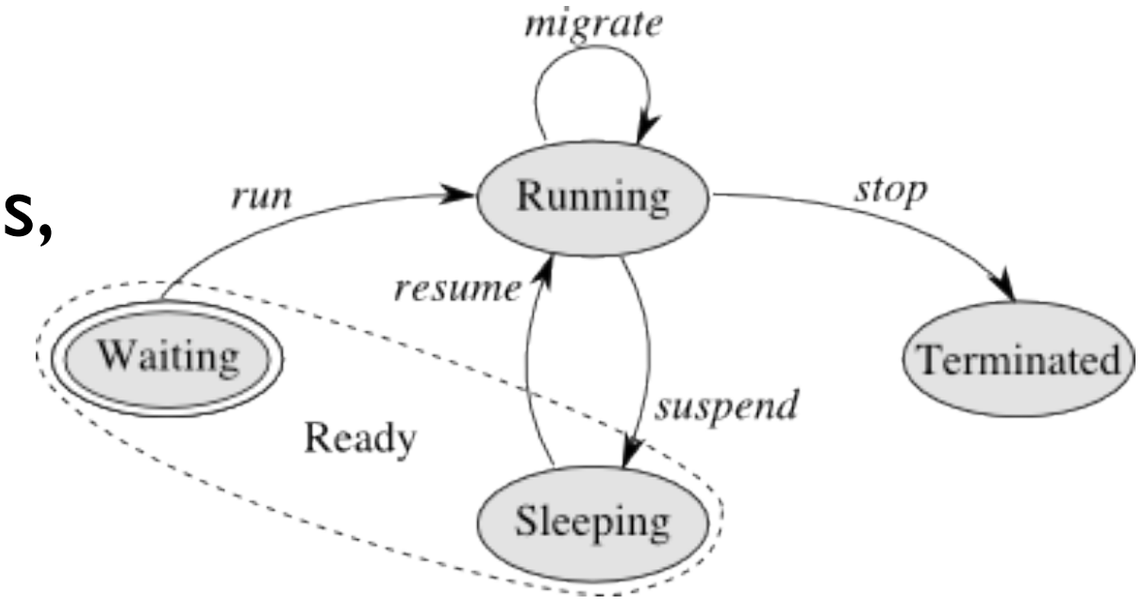
  Generalization: the Cluster-Wide Context Switch concept (Hermenier and all, 2010)

- Use case - energy concerns in Datacenters

Cluster-Wide Context Switch

- General idea: manipulate vJobs instead of jobs (by encapsulating each submitted job in one or several VMs)

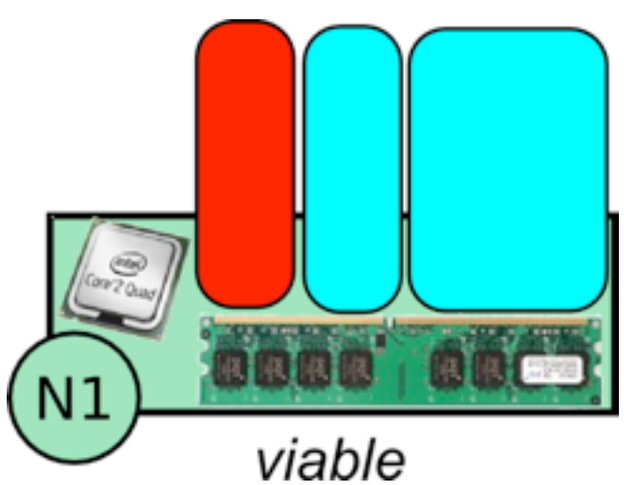
- In a similar way of usual processes, each vjob is in a particular state:



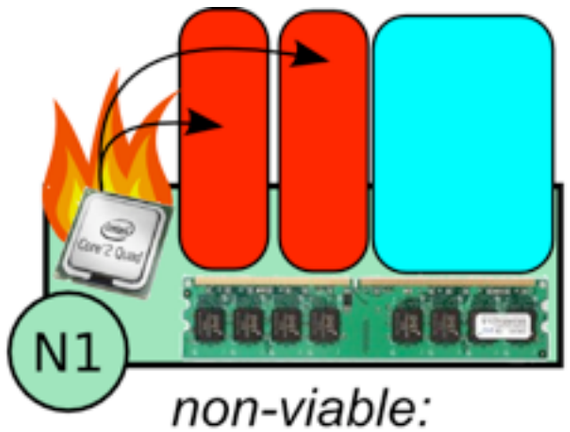
- A cluster-wide context switch (a set of VM context switches) enables to efficiently rebalance the cluster according to the: scheduler objectives / available resources / waiting vjobs queue

The Entropy Proposal

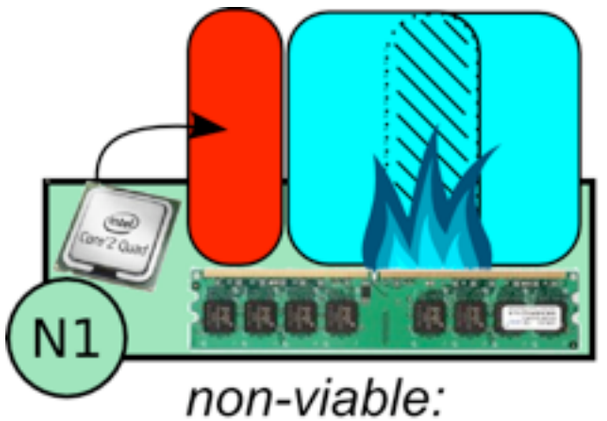
- To finely exploit resources (efficiency and energy constraints)
- Find the "right" mapping between VM needs and resources provided by PM



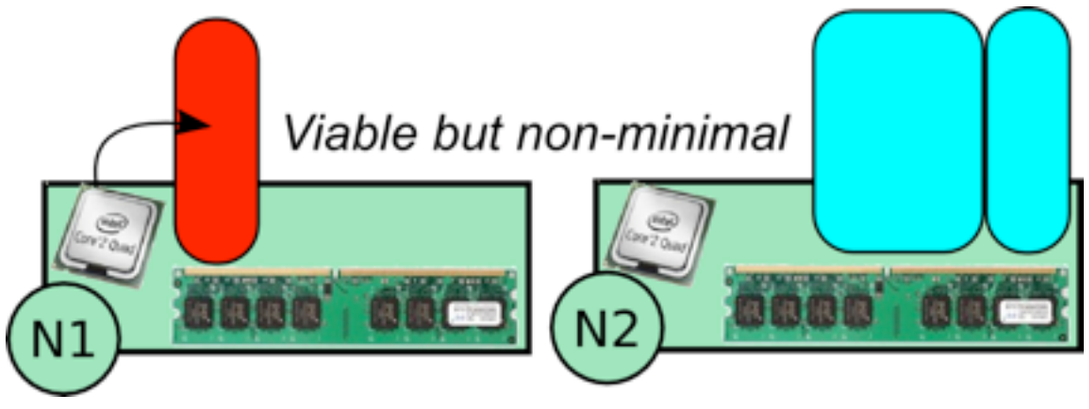
viable

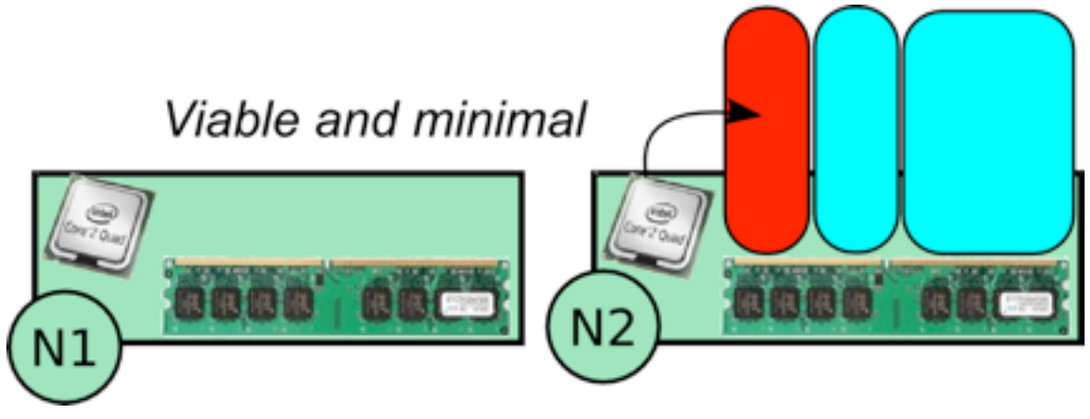


non-viable:
2 active VMs for one CPU



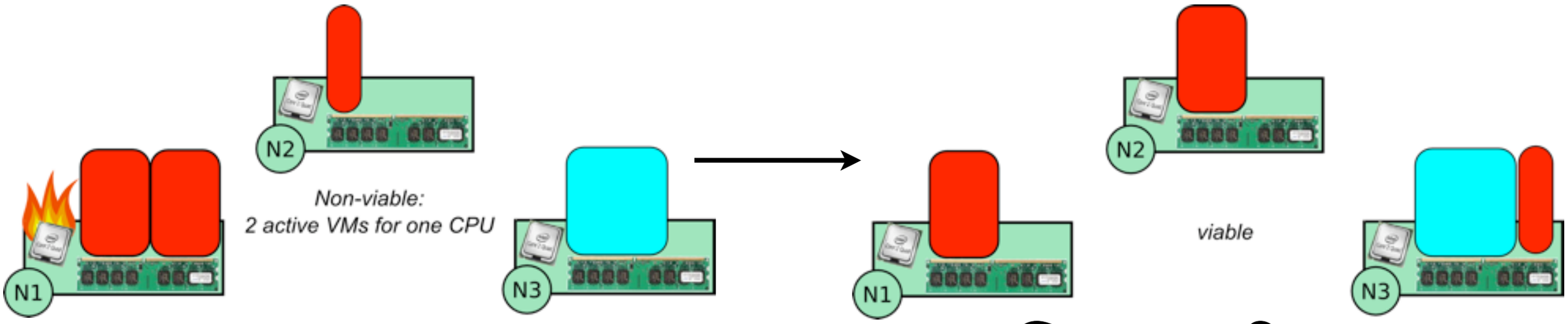
non-viable:
memory overcommitment





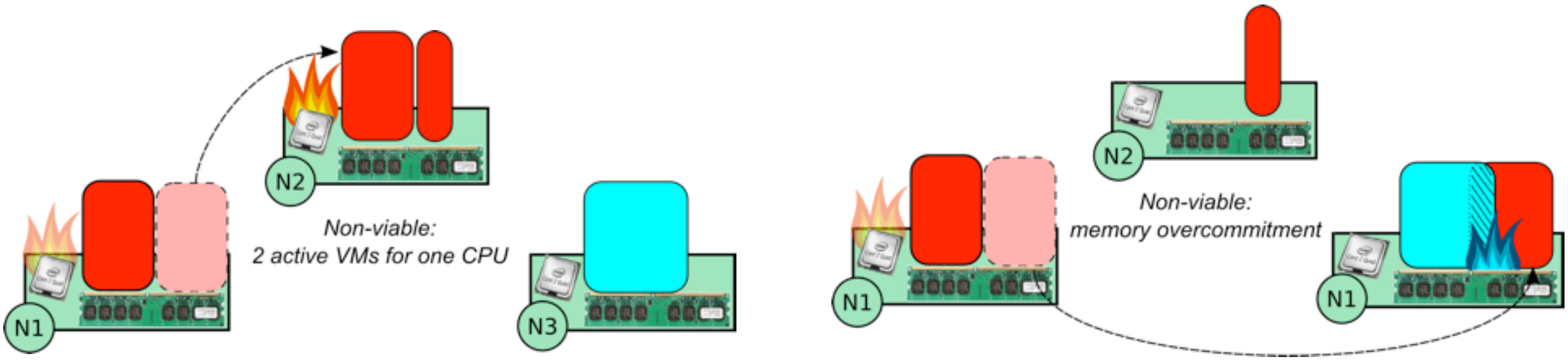
credits: F. Hermenier, Mines Nantes

The Entropy Proposal



Current Status

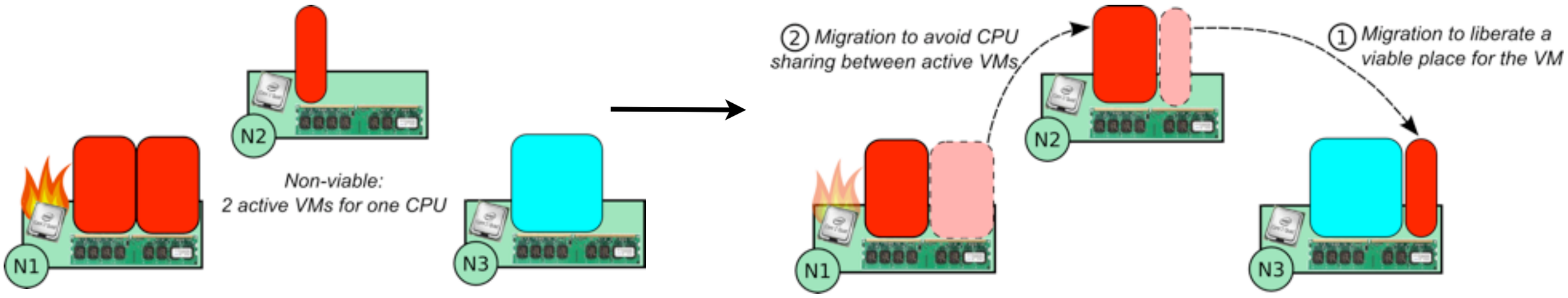
Correct Status



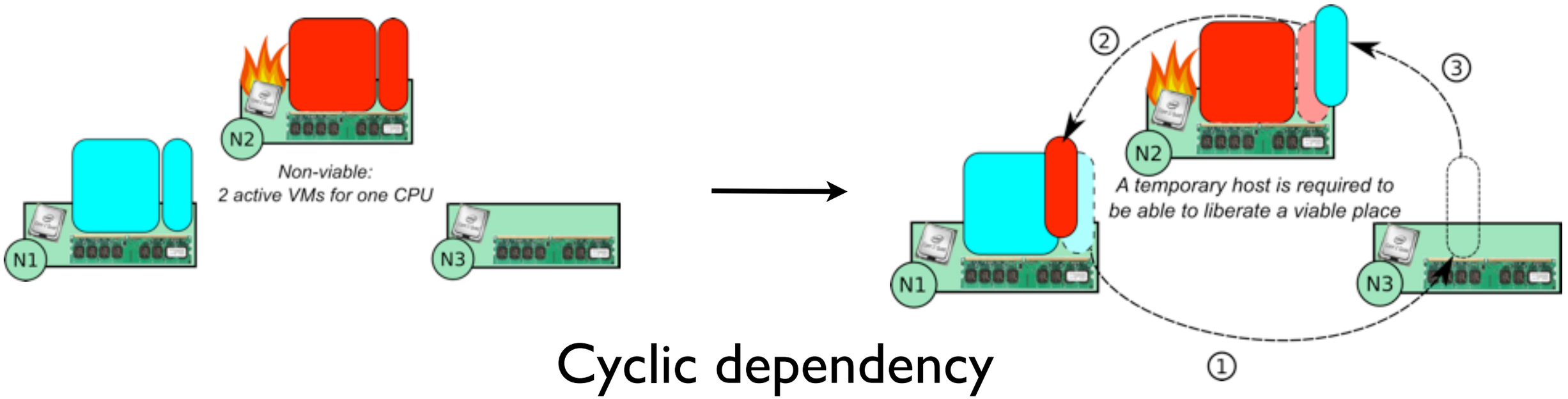
Non-viable manipulations

The Entropy Proposal

- Order VM Operations



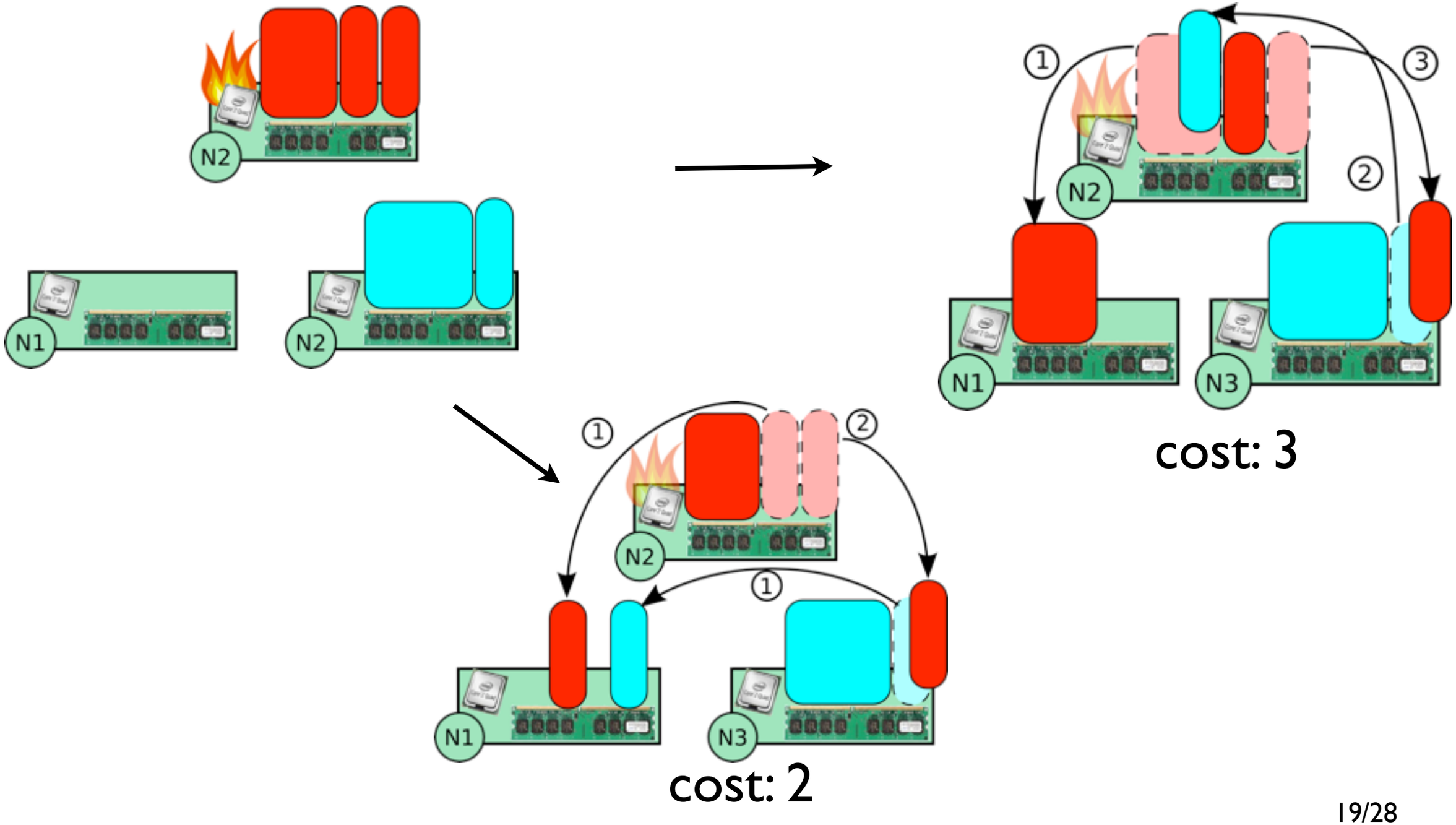
Sequential dependency



Cyclic dependency

The Entropy Proposal

- Optimizing the reconfiguration process



The Entropy Proposal

- To sum up

An autonomic framework to perform advanced scheduling policies of vjobs

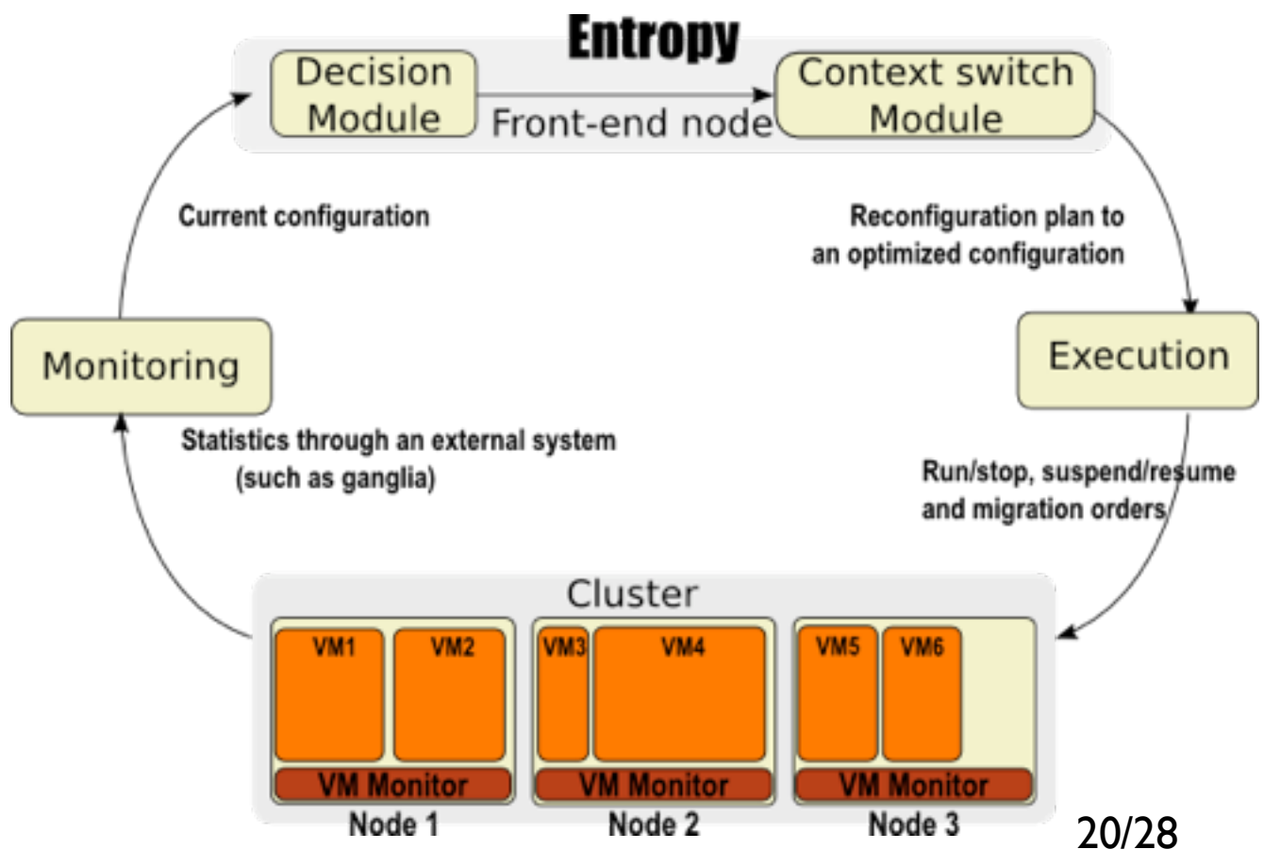
Developed since 2006 (ANR SelfXL / MyCloud, ANR Emergence, 10 persons)

"Prix de la croissance verte numérique" in 2009

Cluster scale

Work in progress

Performance/scalability



Snapshotting capability

- The Saline proposal

  J. Gallard, Phd Student, INRIA Myriads team/XtreemOS project (started in 2008)

  Improving “transparent dynamicity” in Grid usage thanks to VM capabilities
  (can we provide a SSI-like solution thanks to virtualization ?)

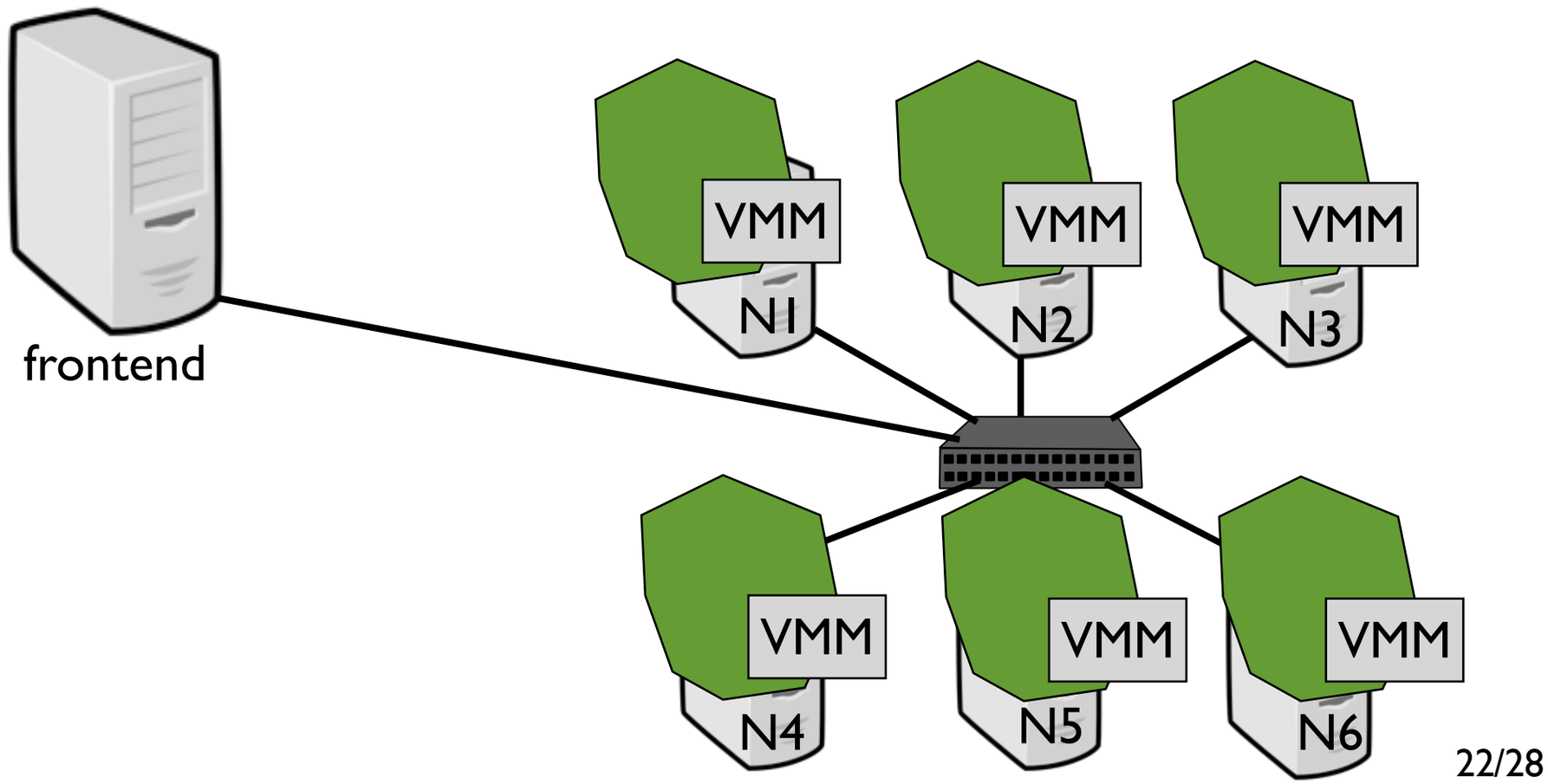
- Use case - management of best effort jobs in Grid

The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

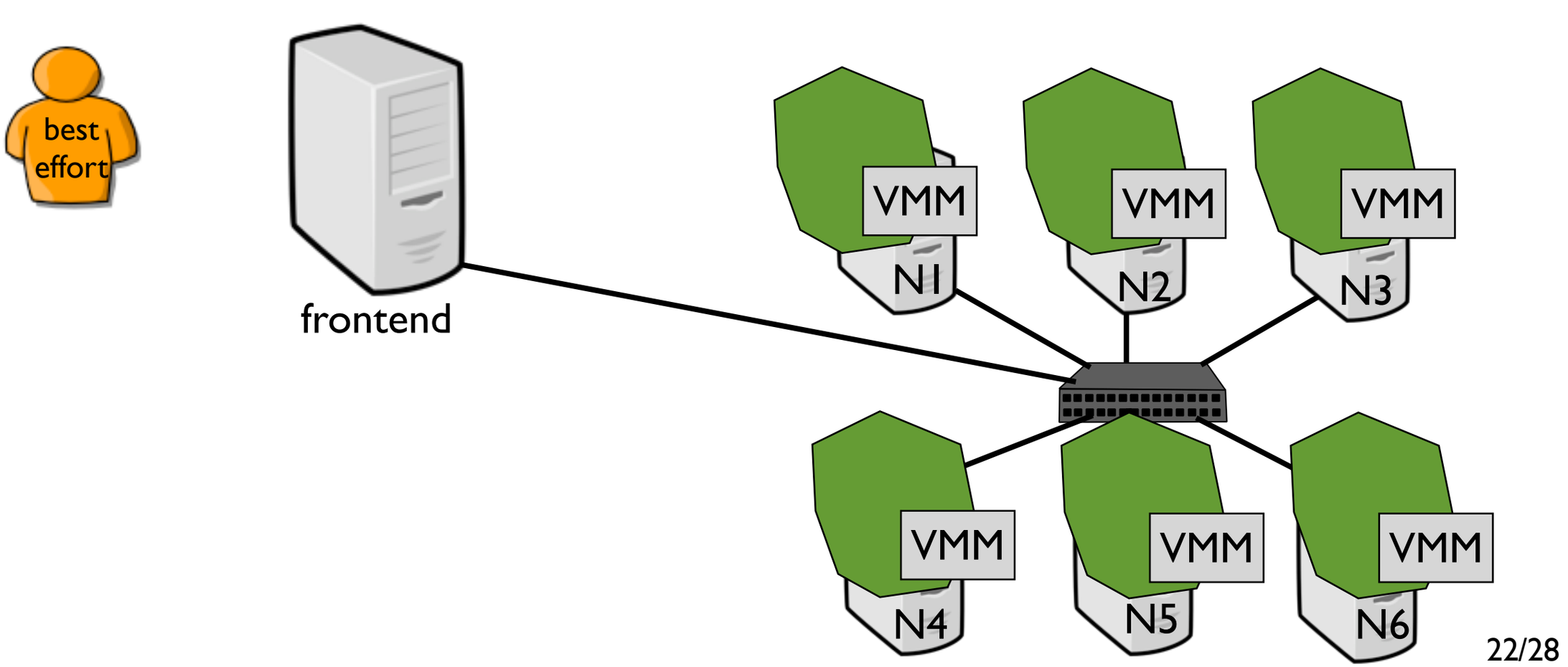


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

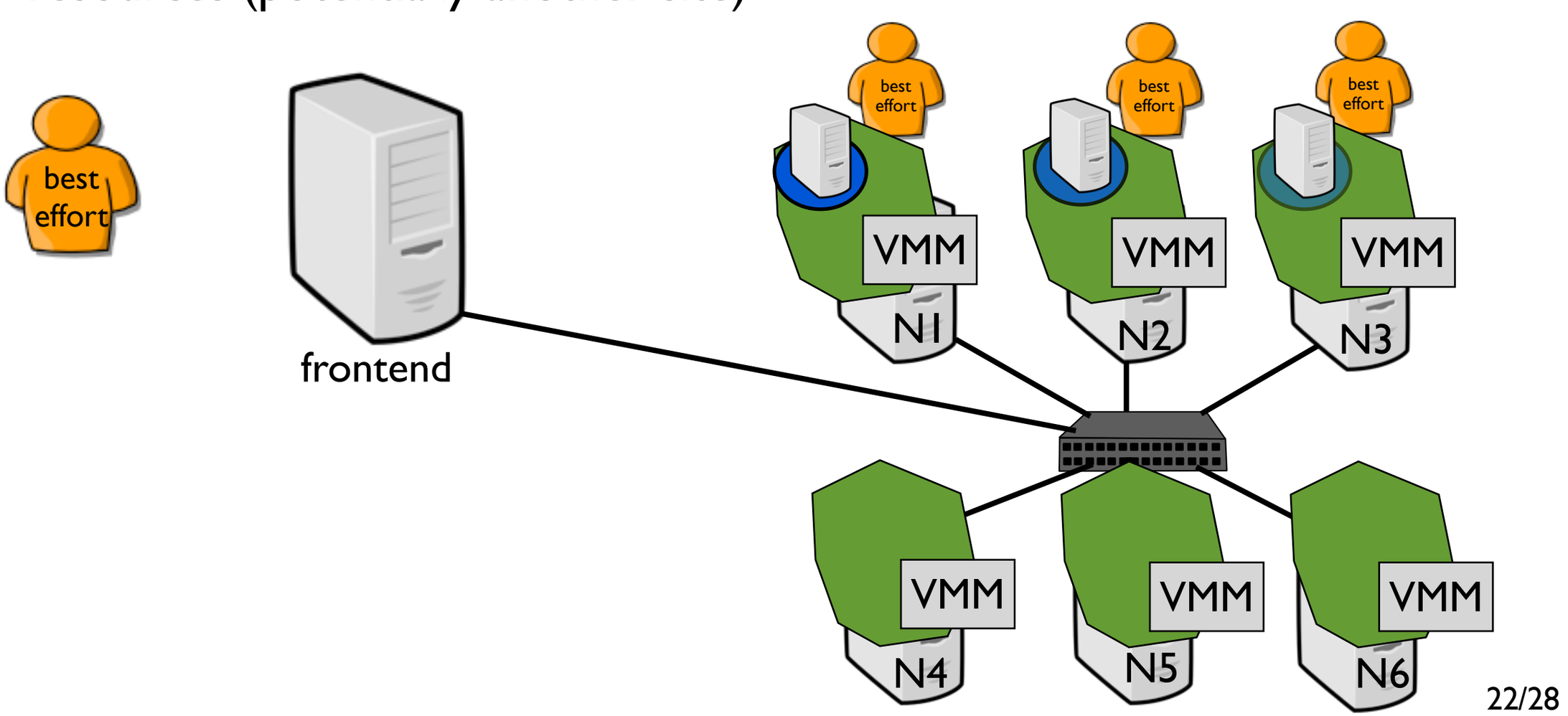


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

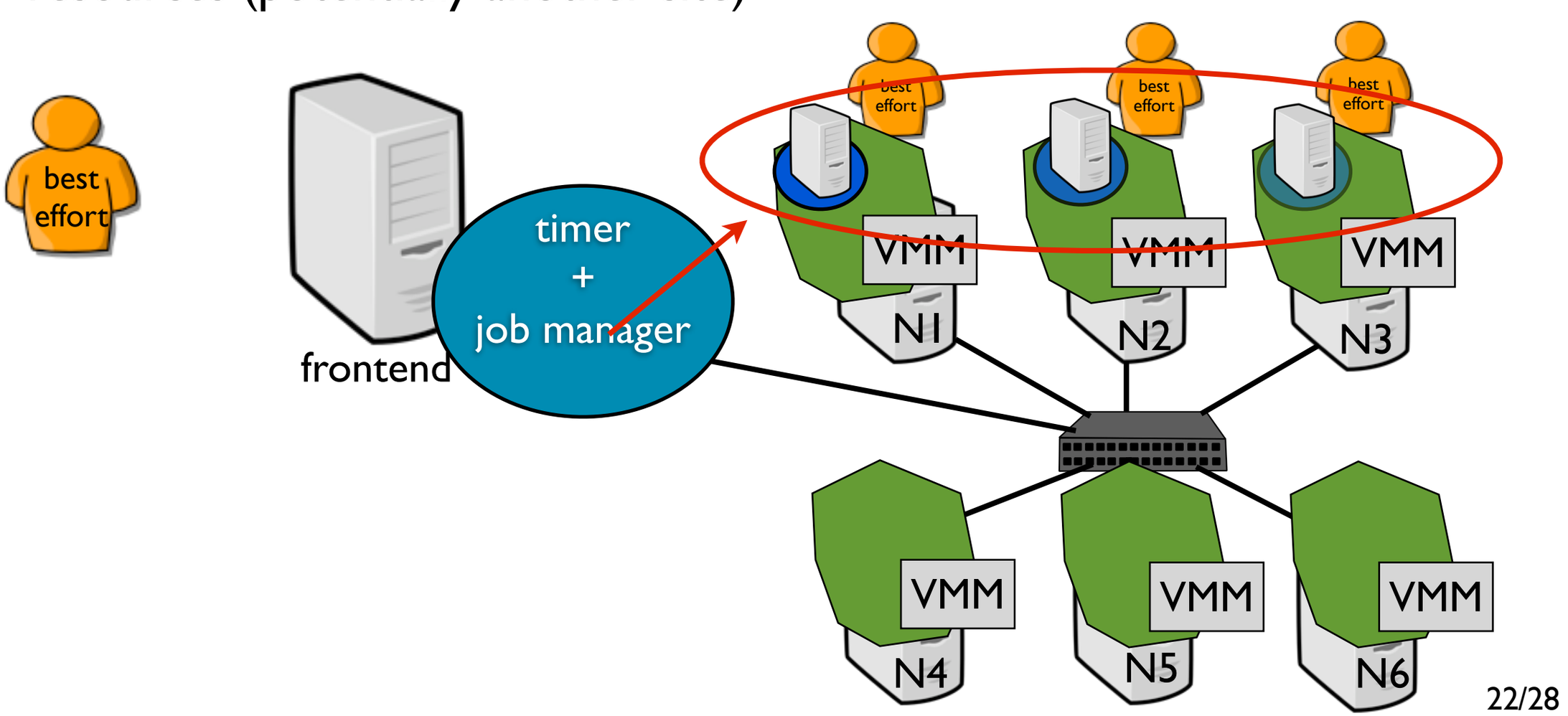


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

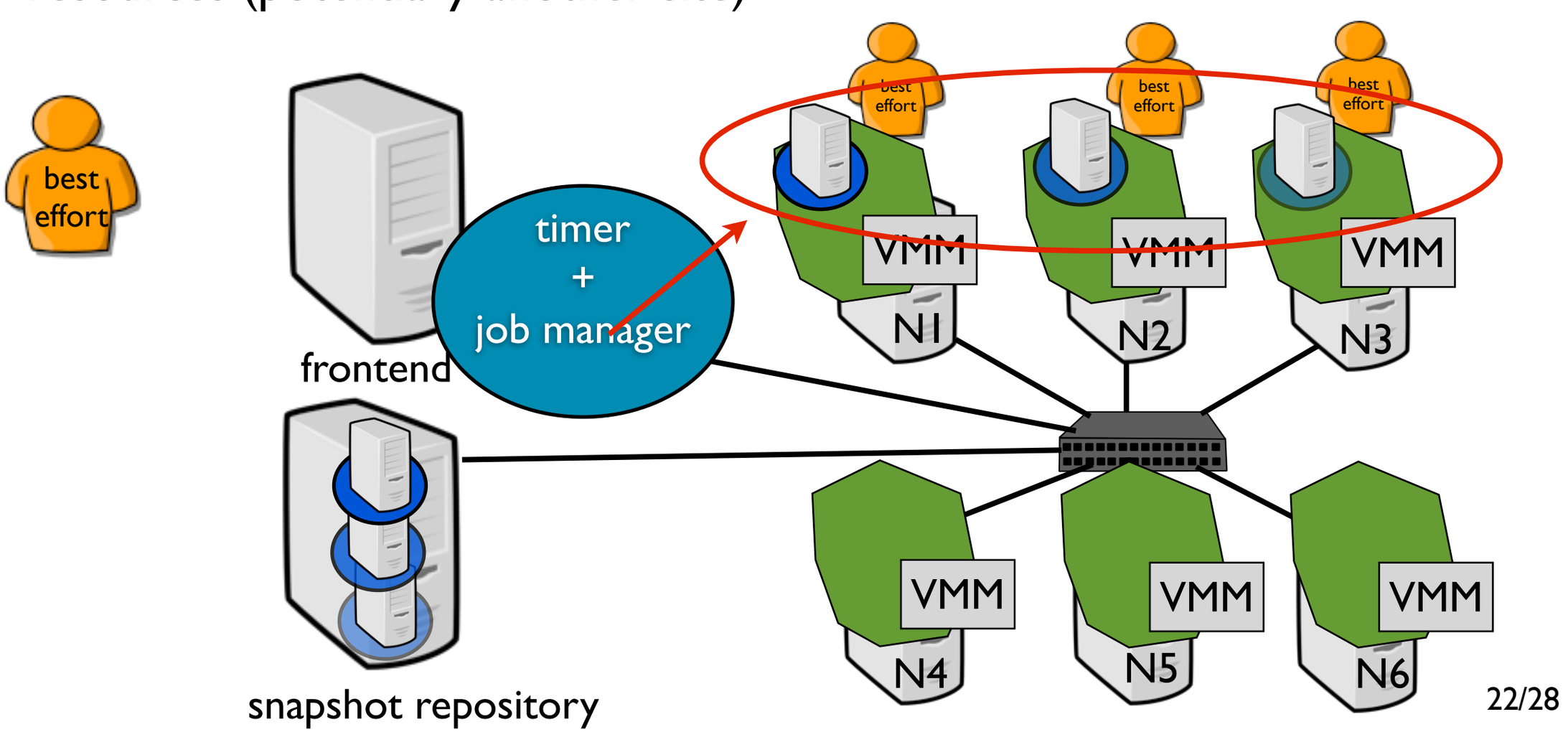


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

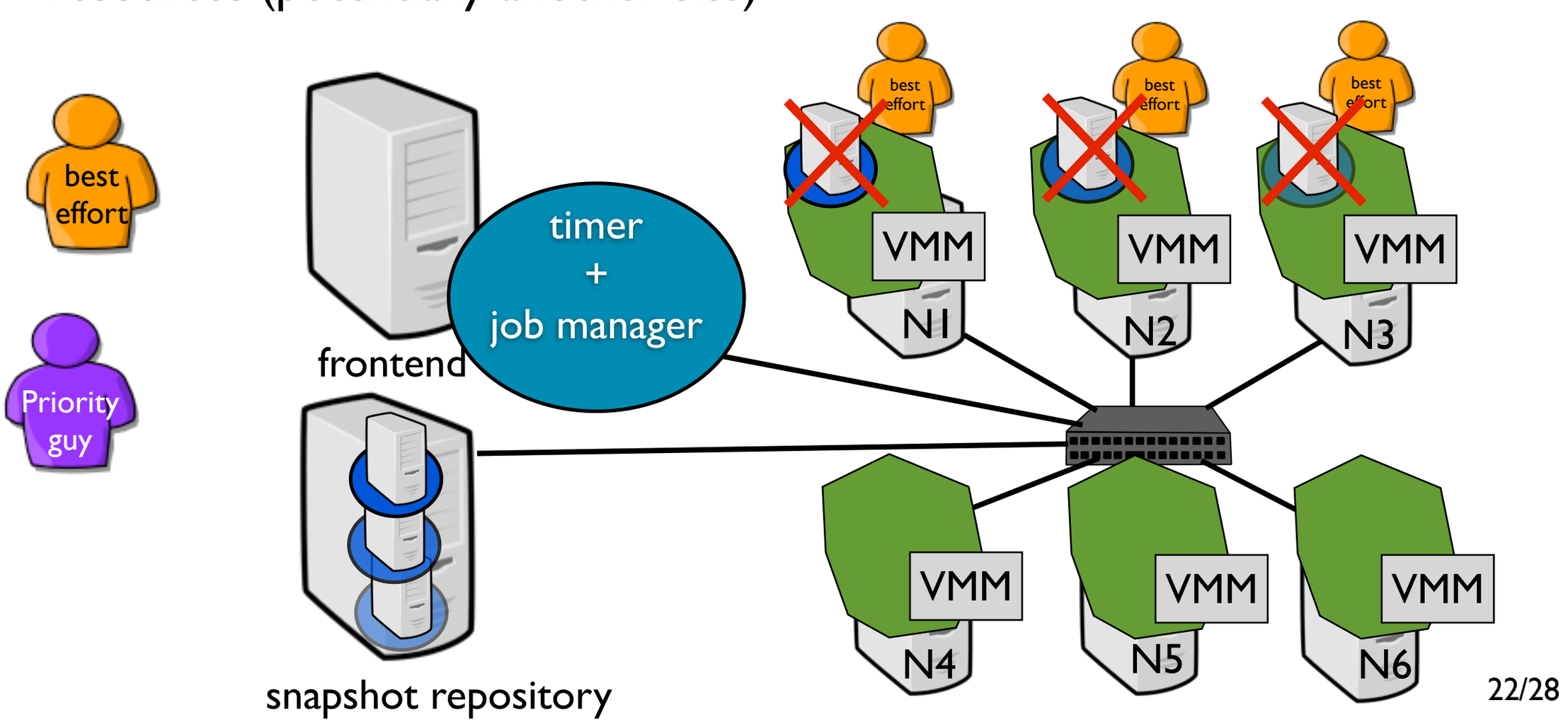


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

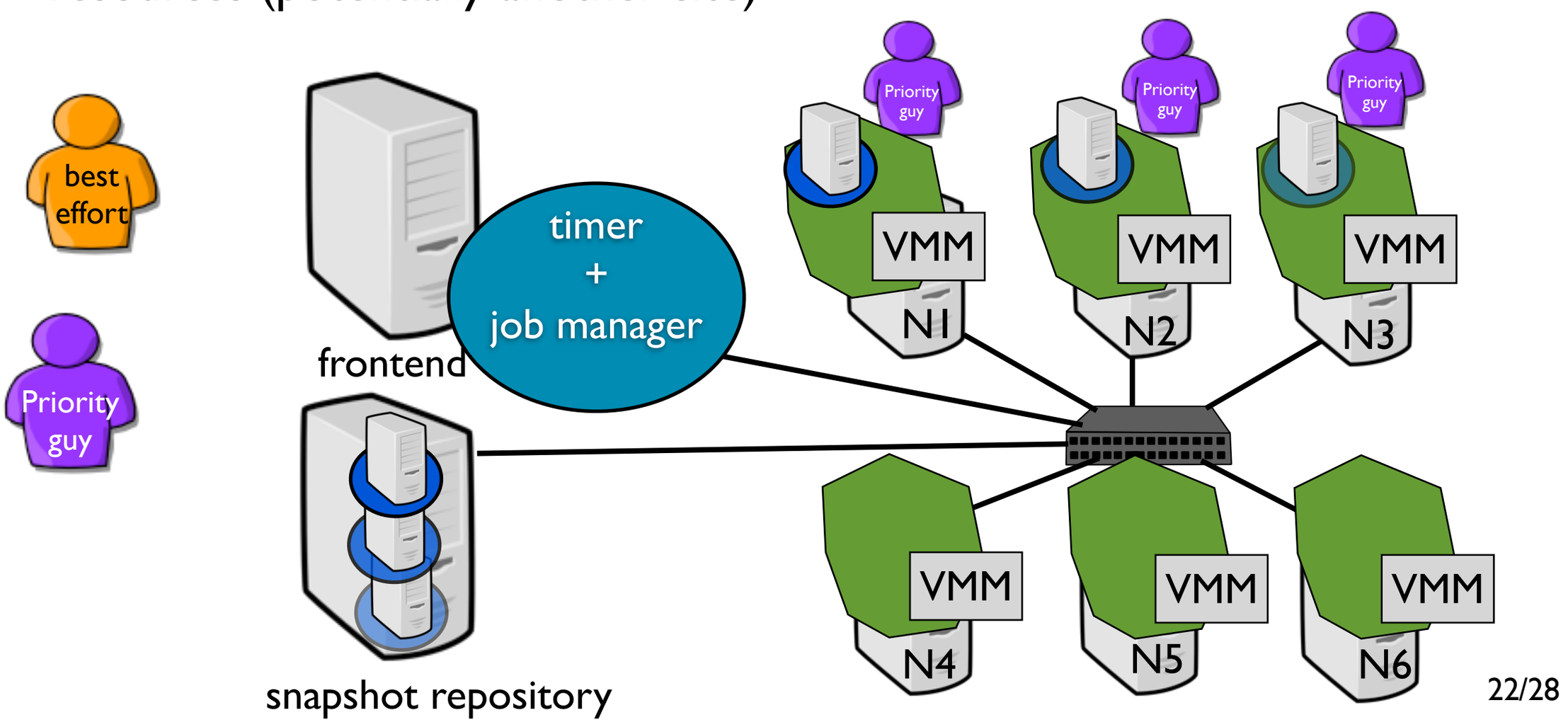


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)

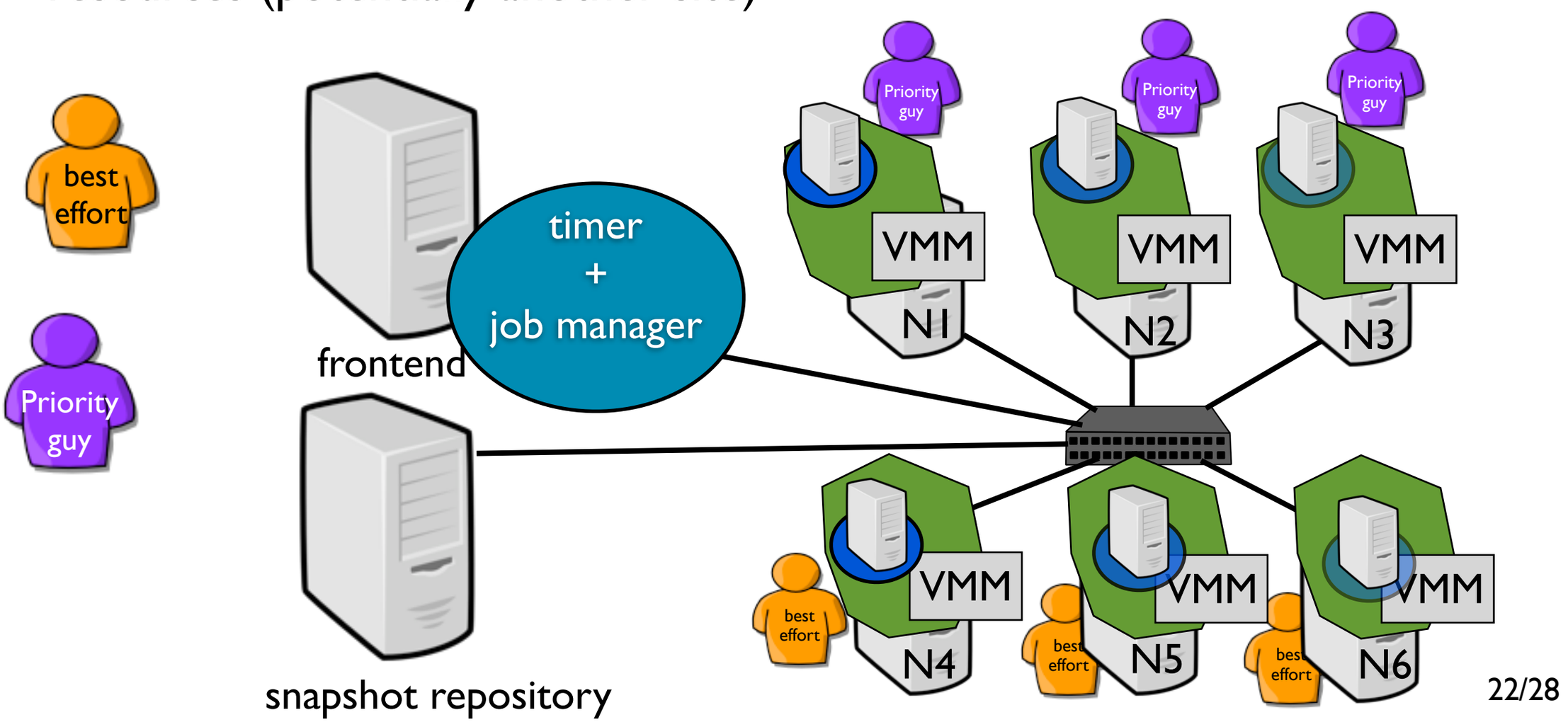


The Saline Proposal

- Overview: manage jobs dynamically through the whole Grid

  A job is executed by a virtual cluster
  (several VMs distributed on one site)

  Exploit VM capabilities to temporarily suspend or relocate the job to other resources (potentially another site)



The Saline Proposal

- A simple approach

  Theoretically speaking !

  In practice: several challenges

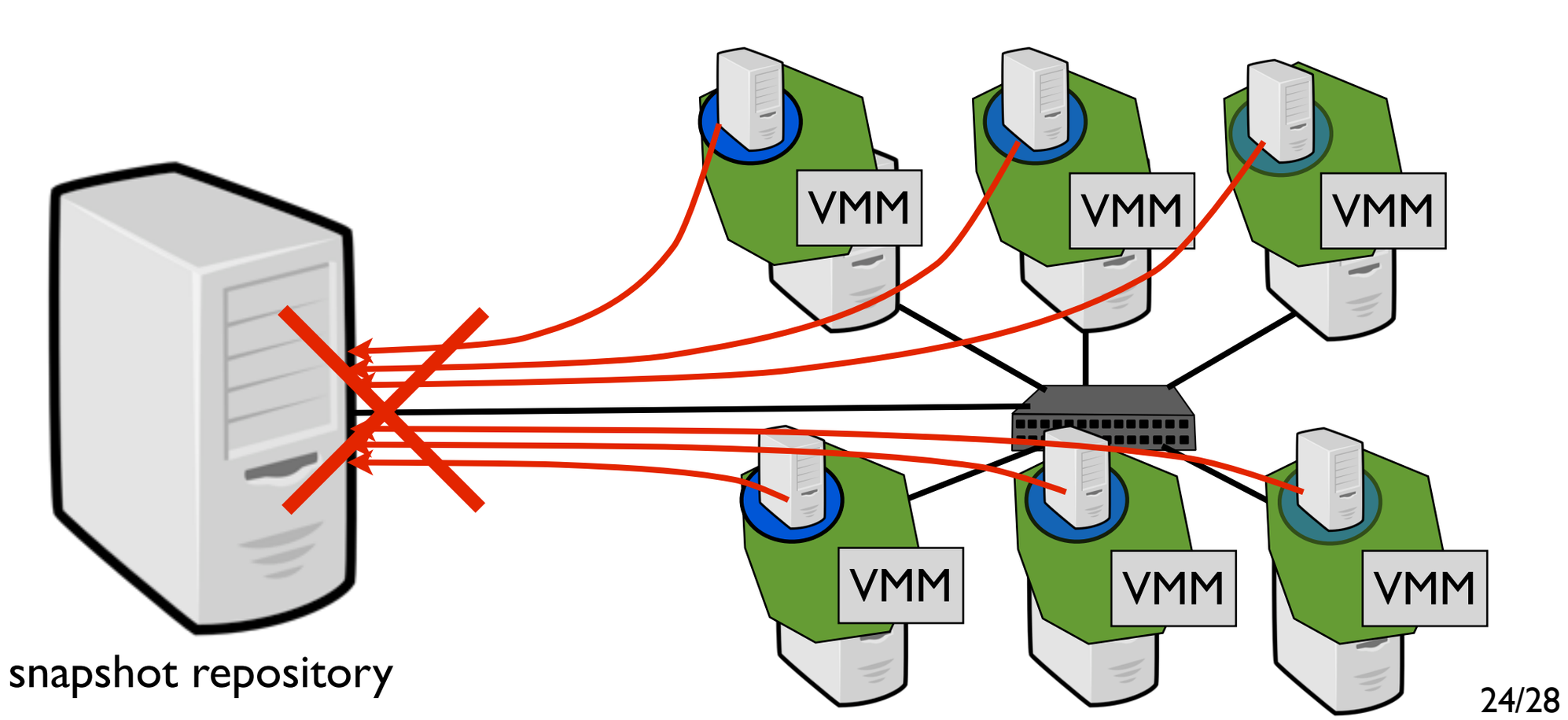
  Images and snapshots management (at cluster/Grid level)

  Network and IP addressing issues (especially during grid-wide resumes)
  VMs of one job should be on the same subnet
  MAC/IP conflicts at grid level

The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)

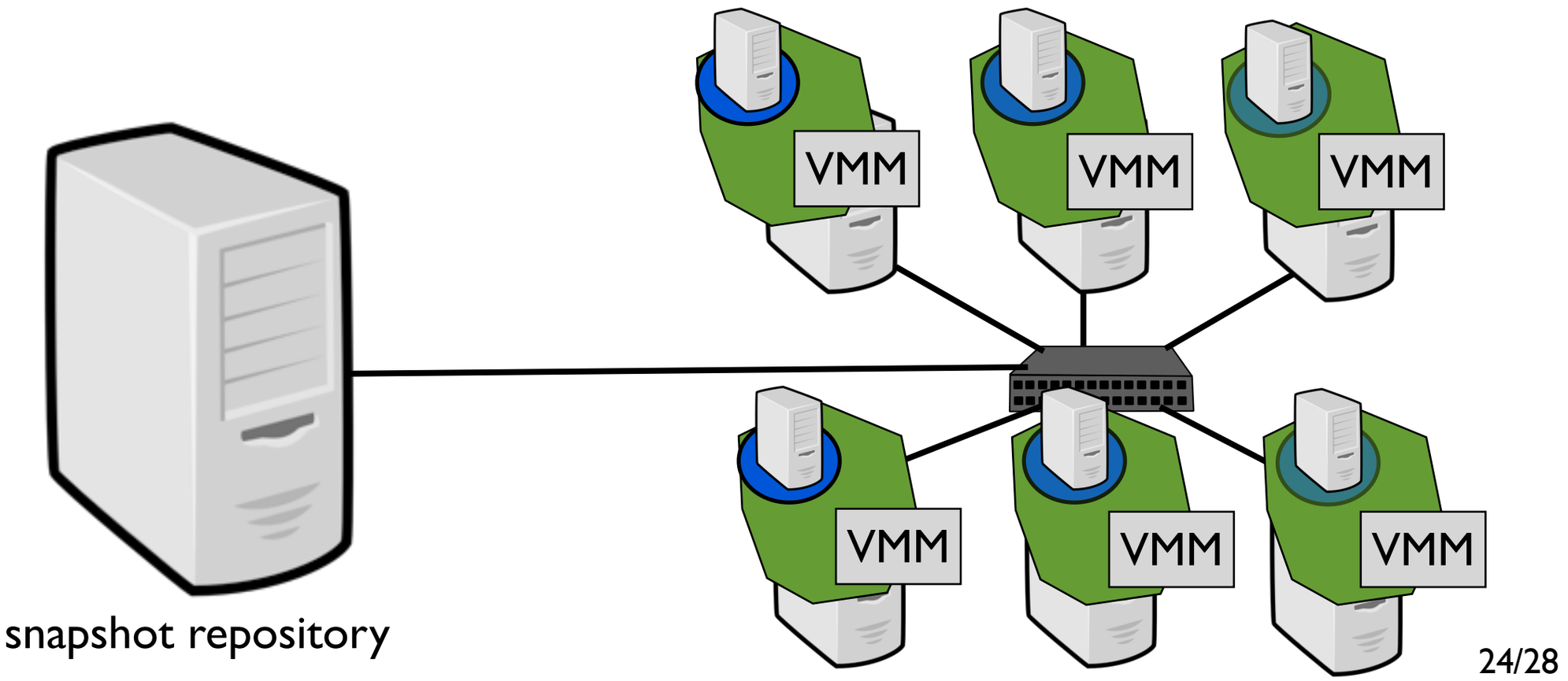


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster

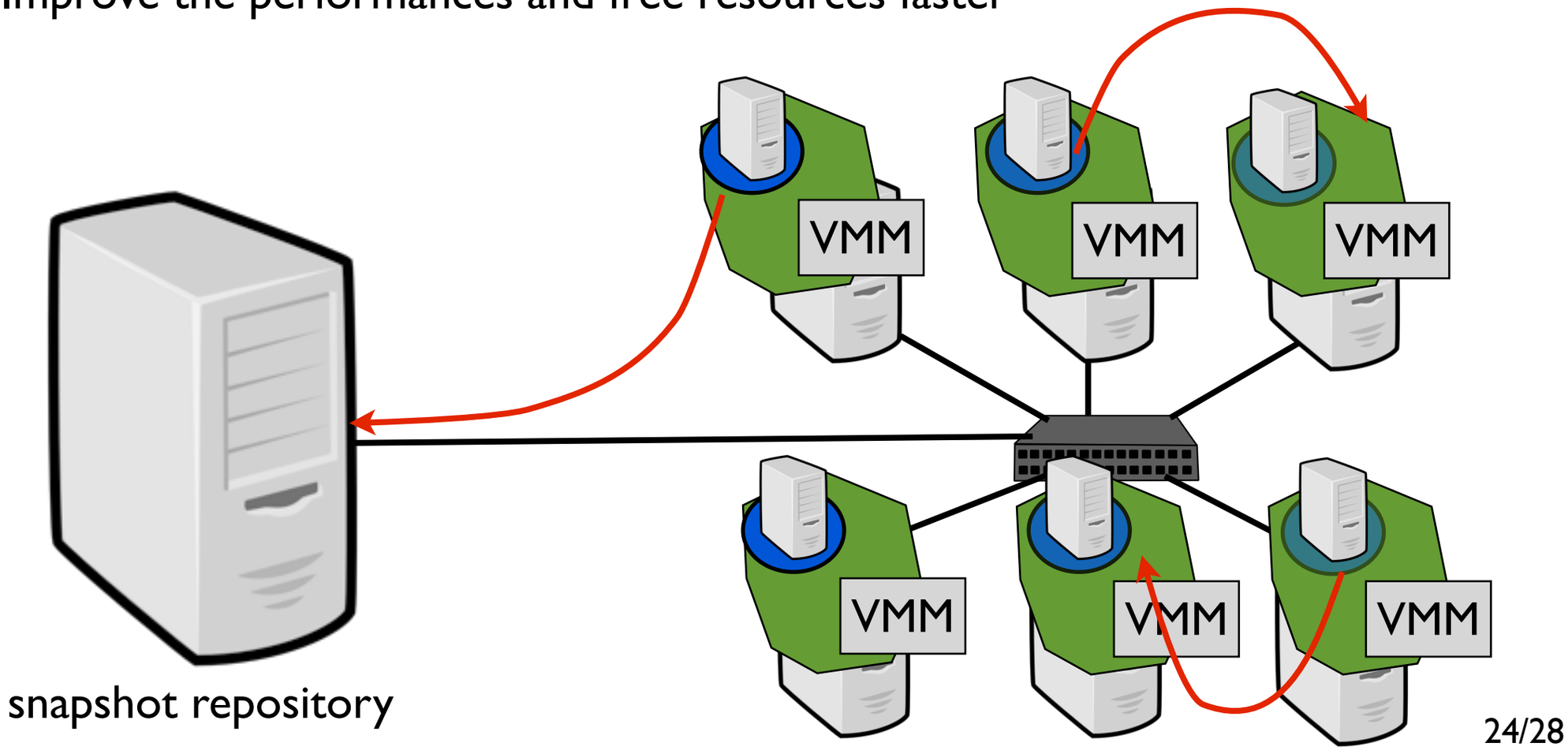


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

  Recursively bring all the snapshots on the subset of nodes
  Improve the performances and free resources faster

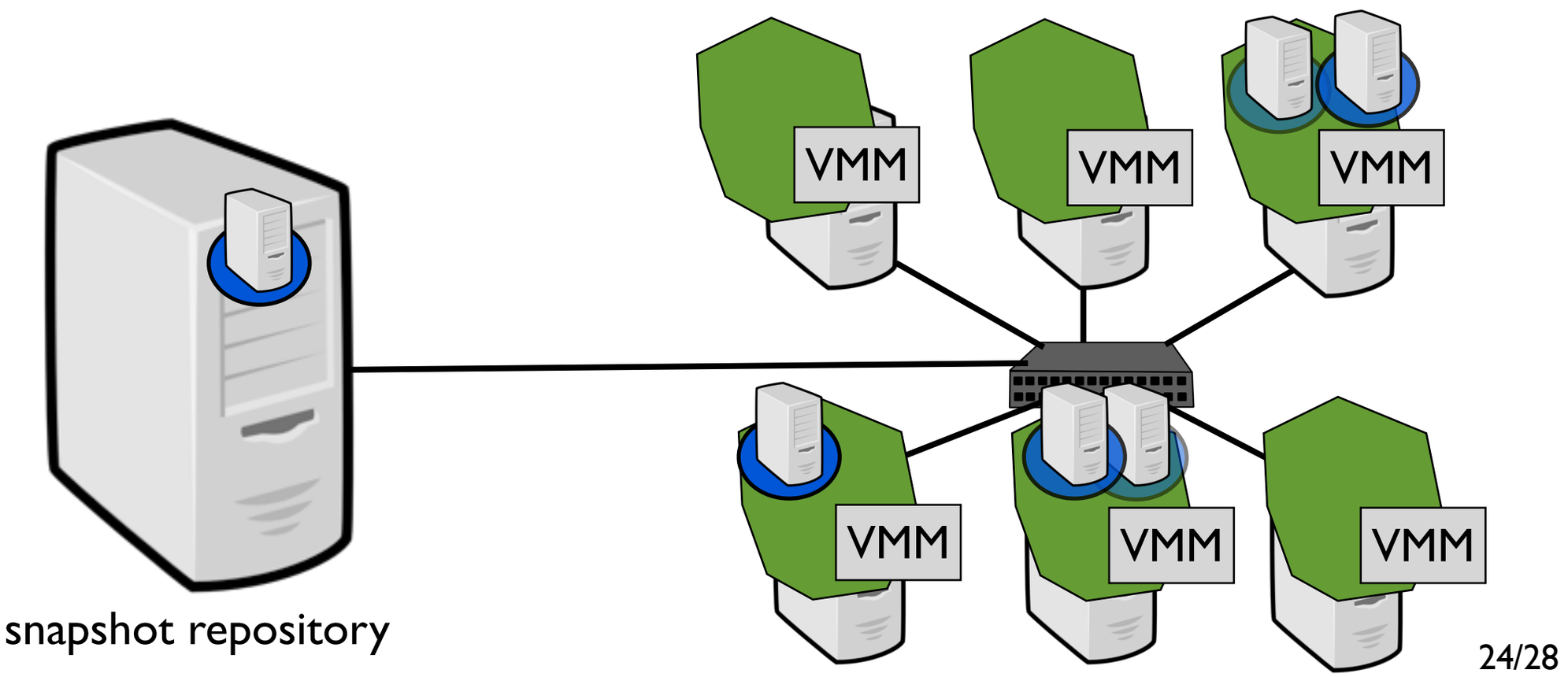


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster

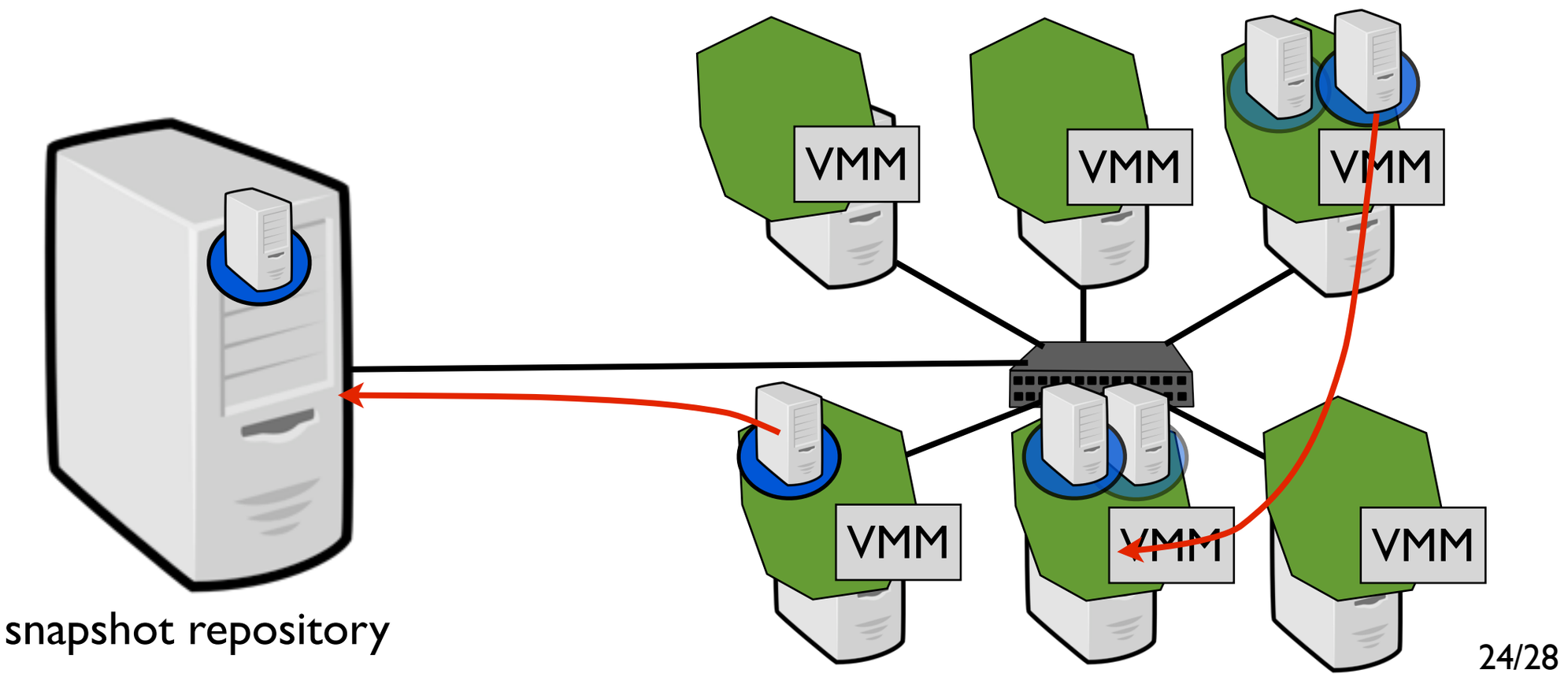


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster

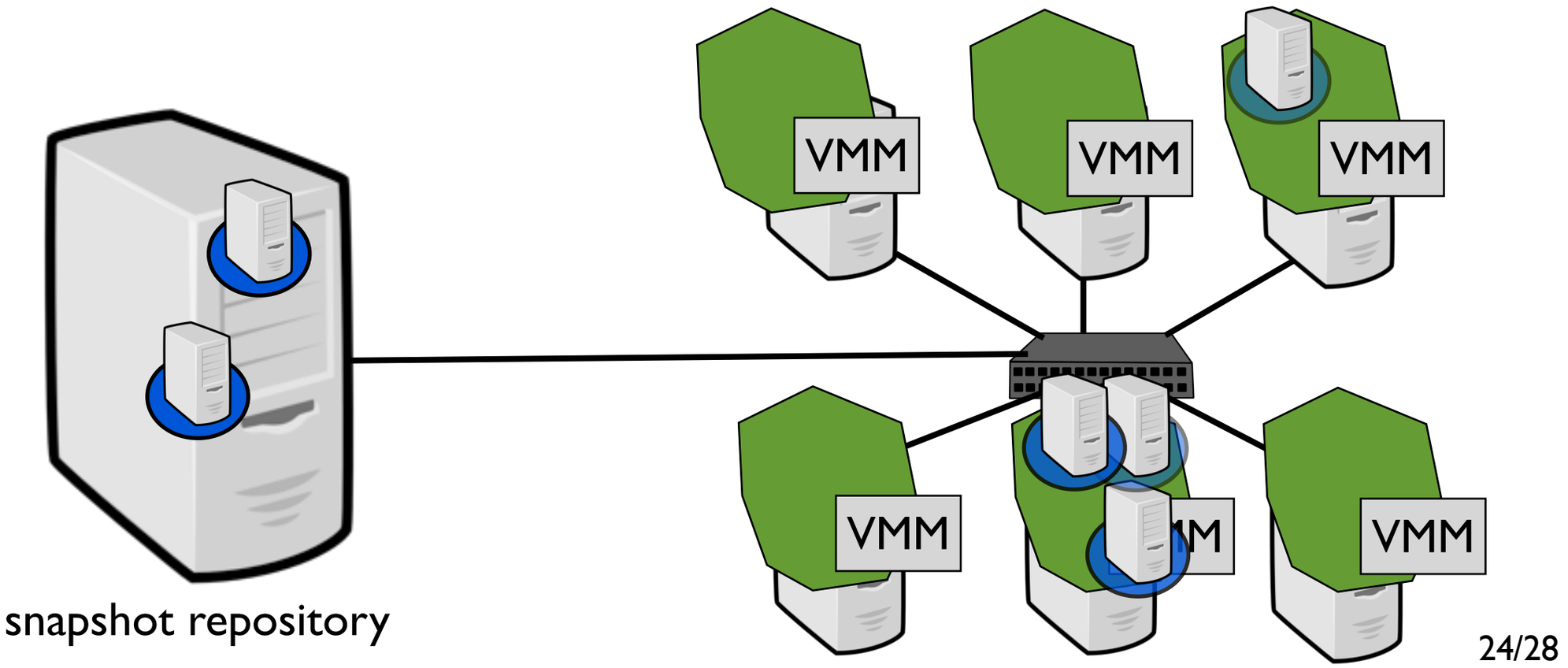


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster

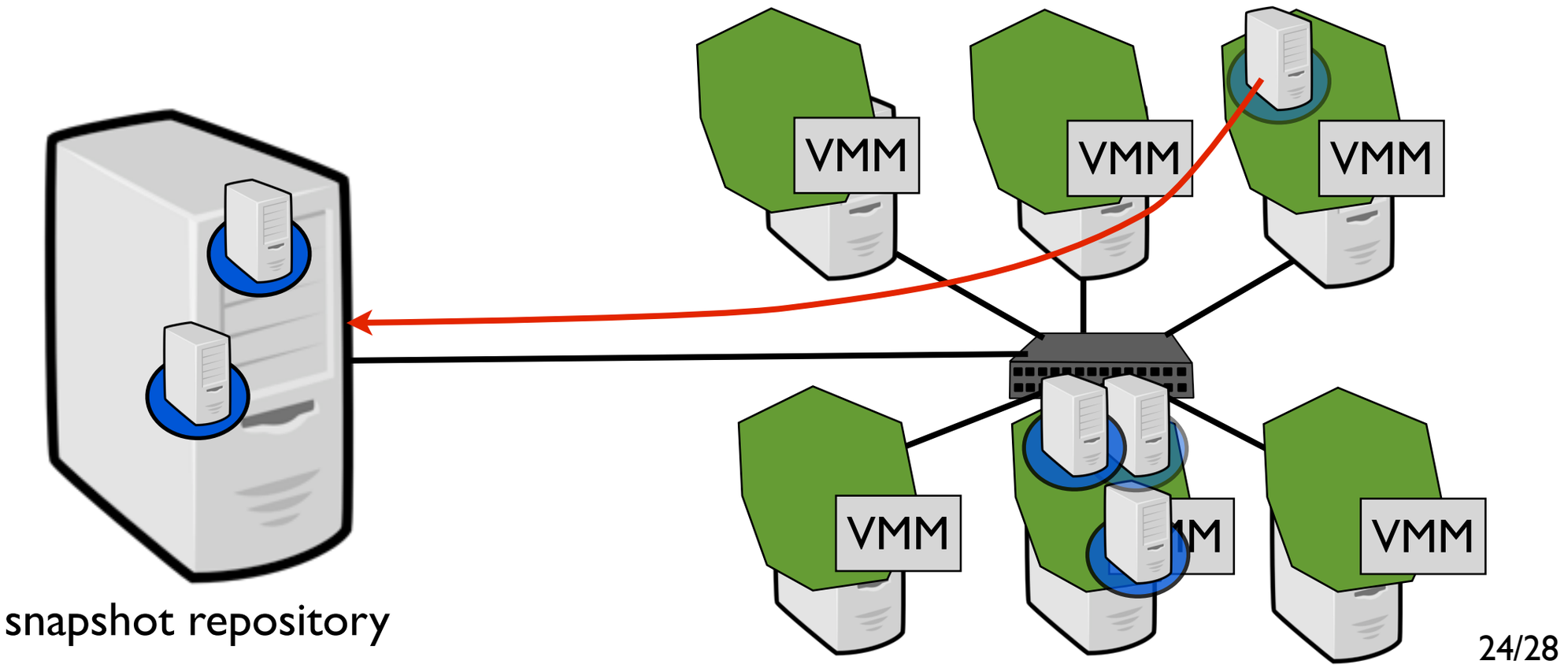


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster

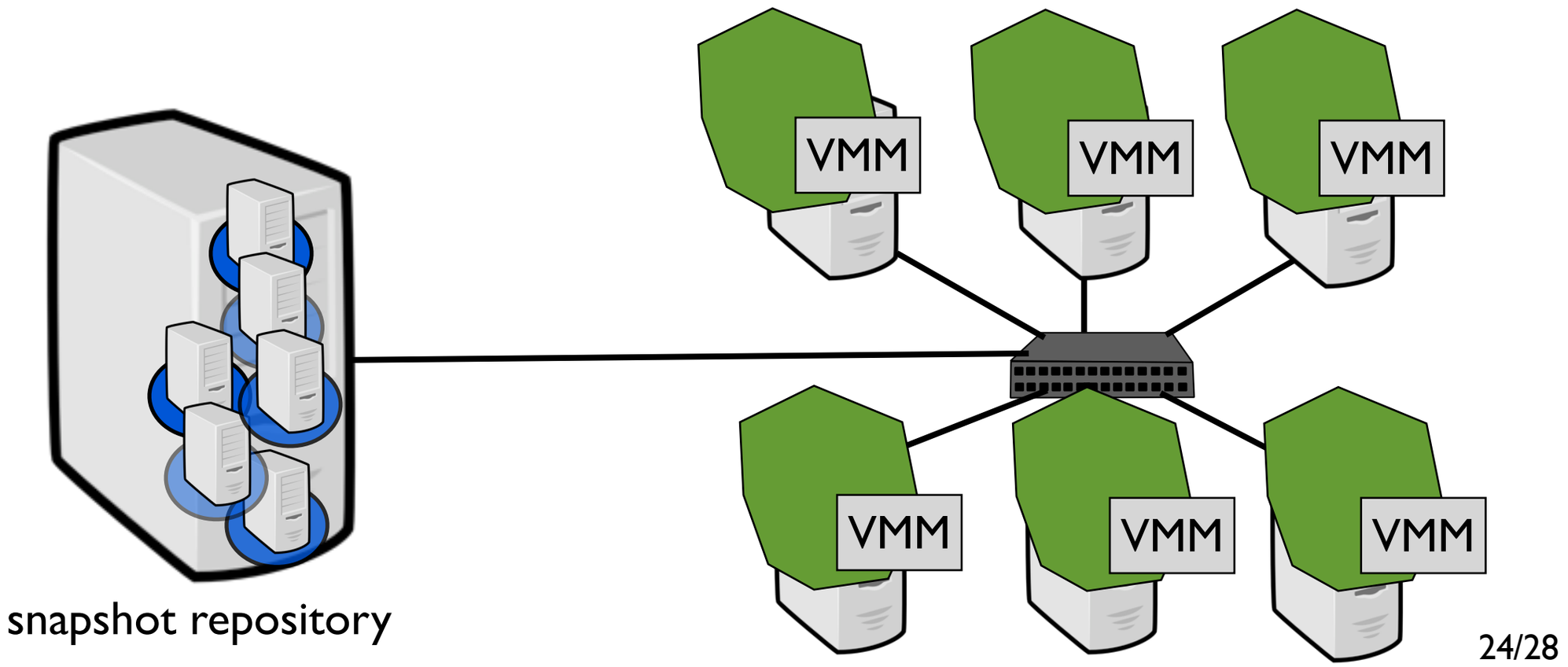


The Saline Proposal

- Snapshot management: periodically save all snapshots

  From all nodes to a snapshot repository (n to 1: scalability issue)
  Proposal: schedule the data transfer to limit the congestion (`kget+`)

    Recursively bring all the snapshots on the subset of nodes
    Improve the performances and free resources faster



The Saline Proposal

- To sum up

Thanks to VM capabilities, users can exploit Grids in a more transparent and dynamic way

Grid snapshot/resume achieved

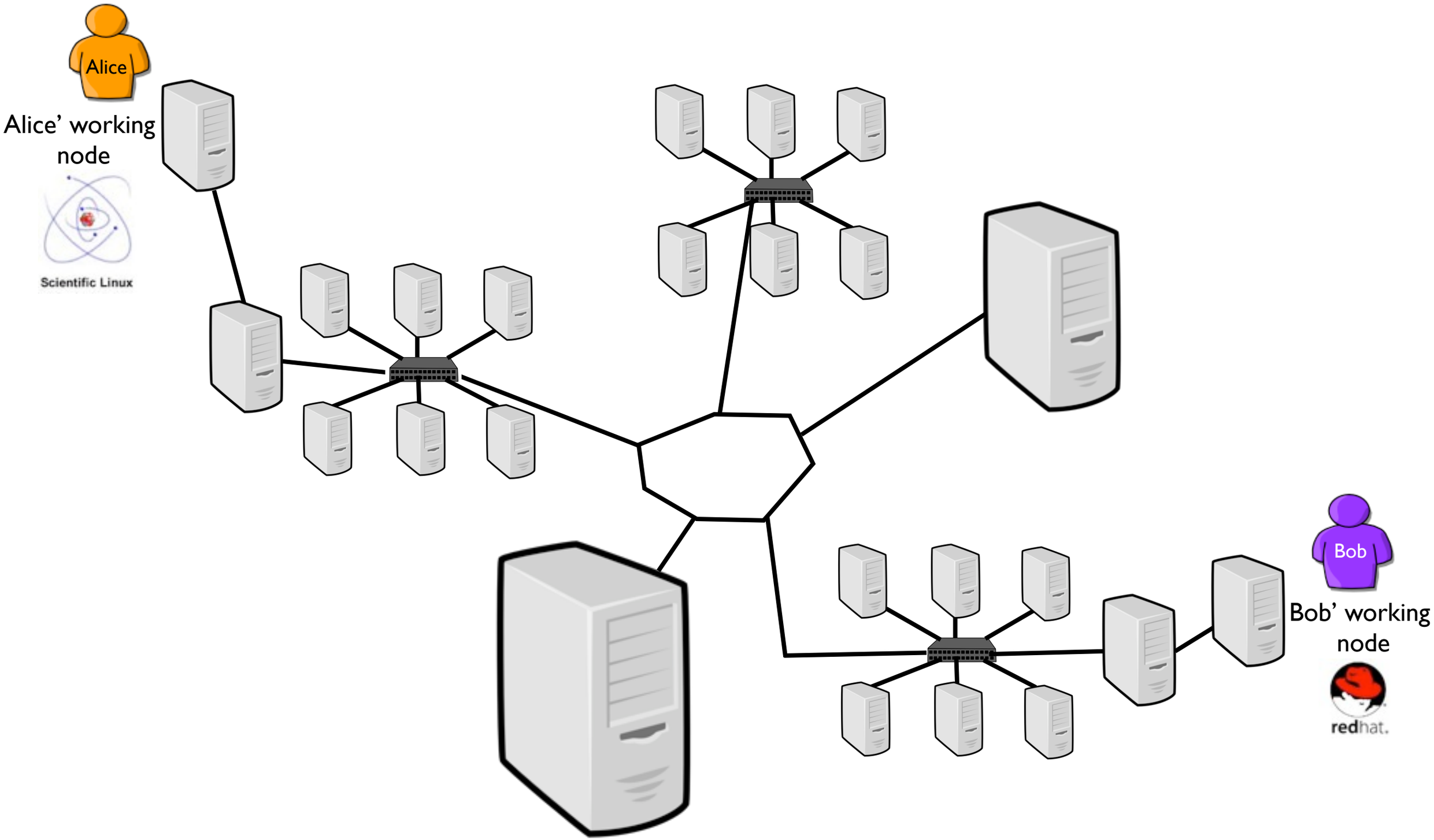
Work in progress

Improve the integration between batch scheduler and job manager (use live migration when possible, improve scheduling decisions)

Split a virtual cluster between two or more sites

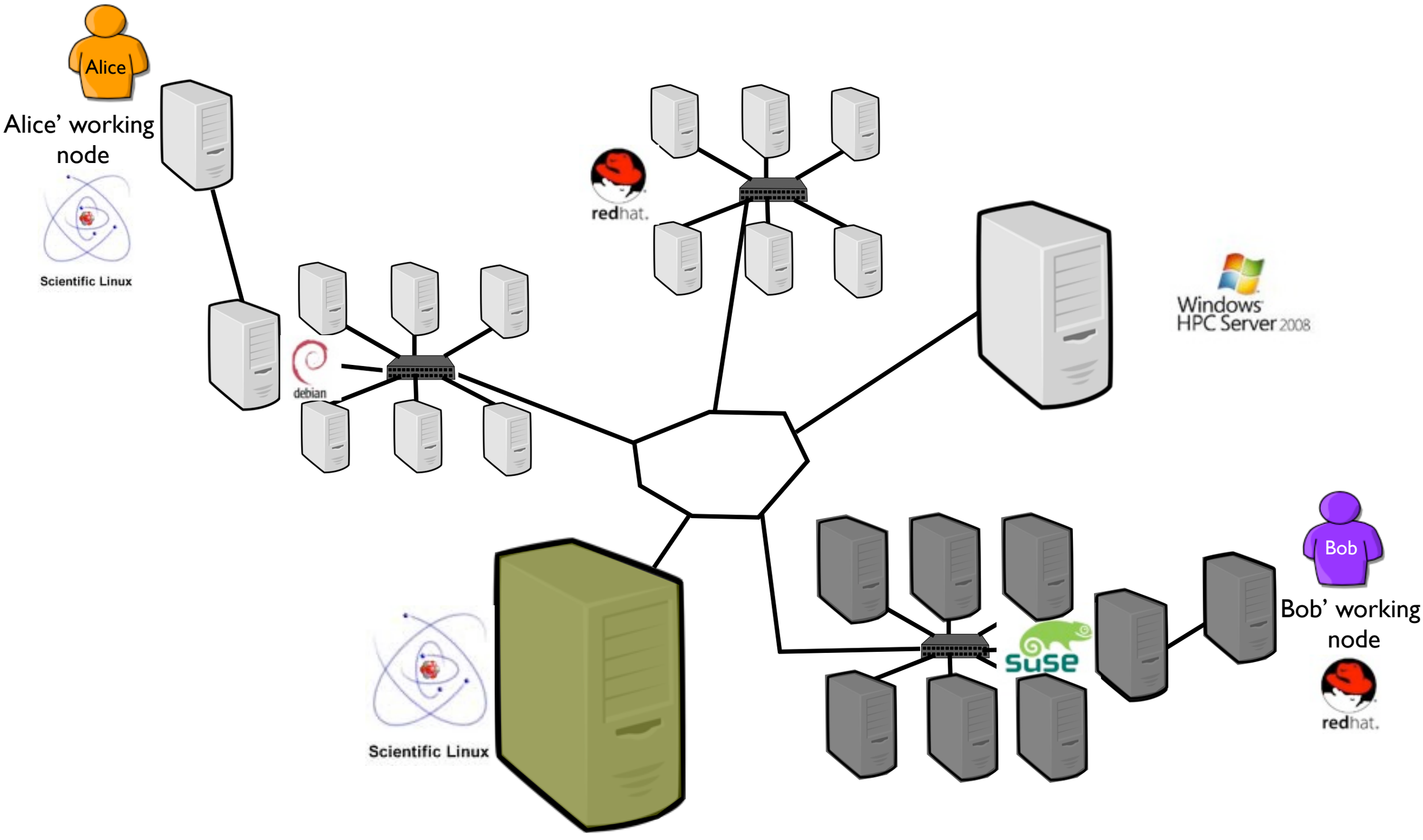
Conclusion

The Alice/Bob Example



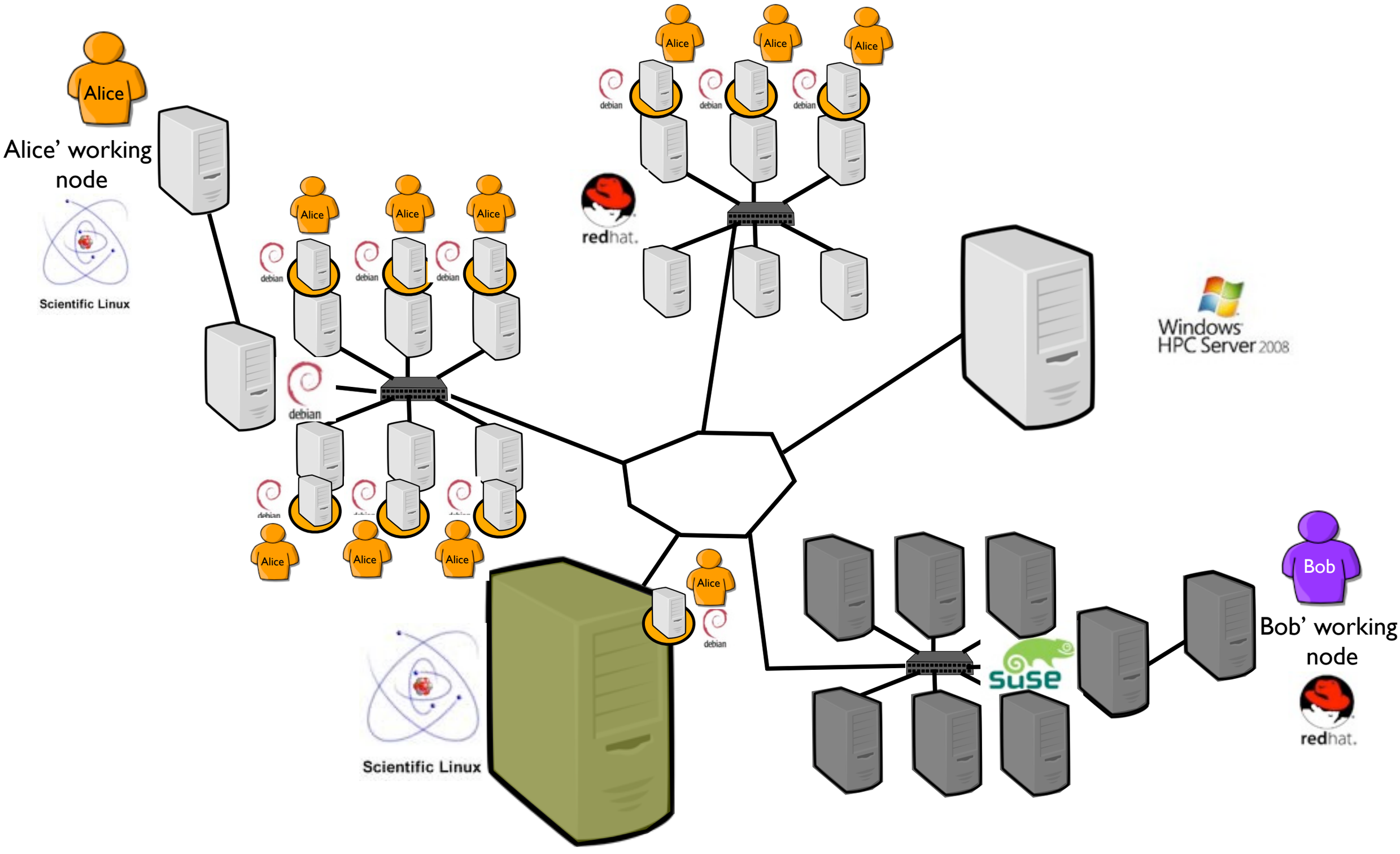
Conclusion

The Alice/Bob Example



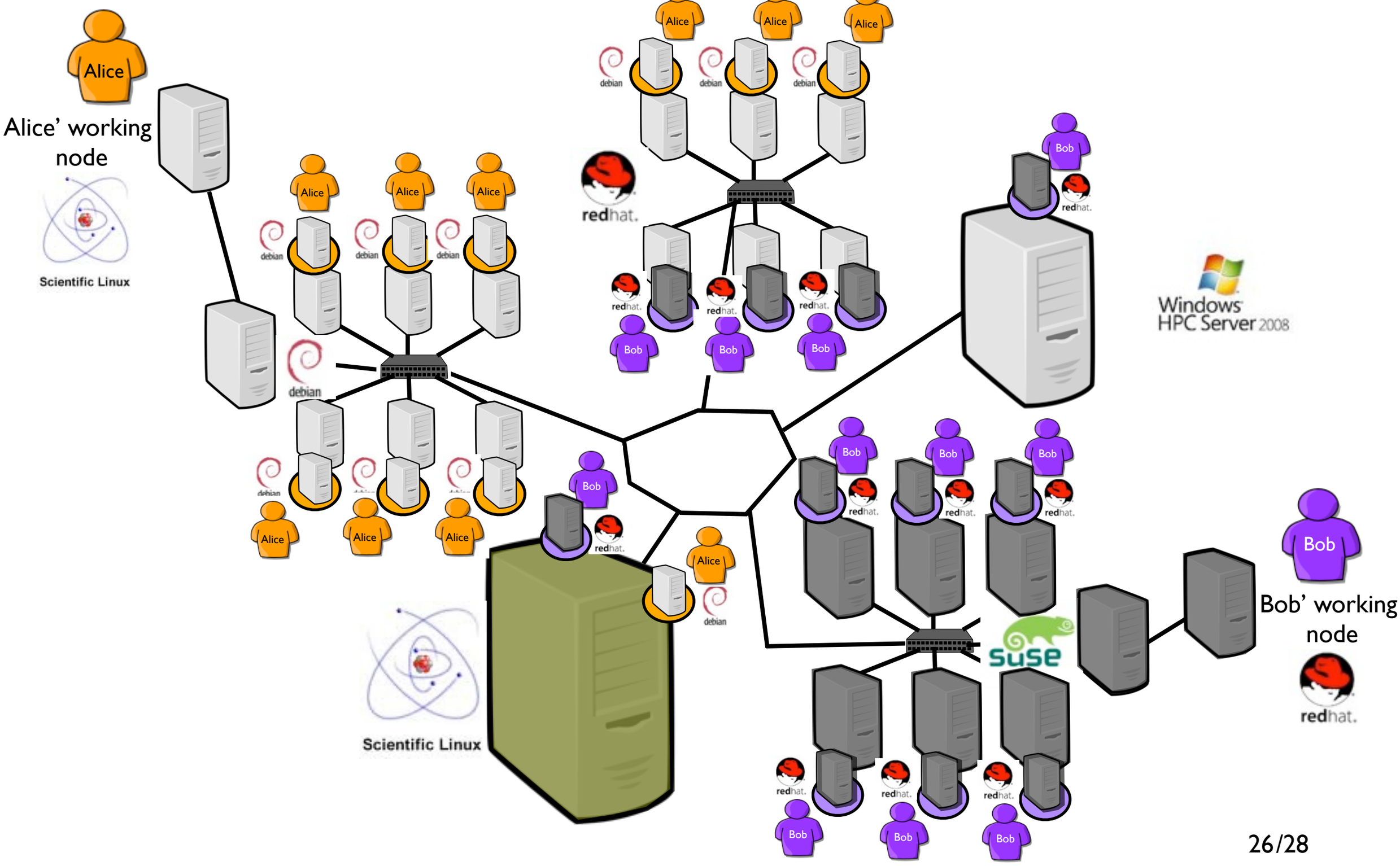
Conclusion

The Alice/Bob Example



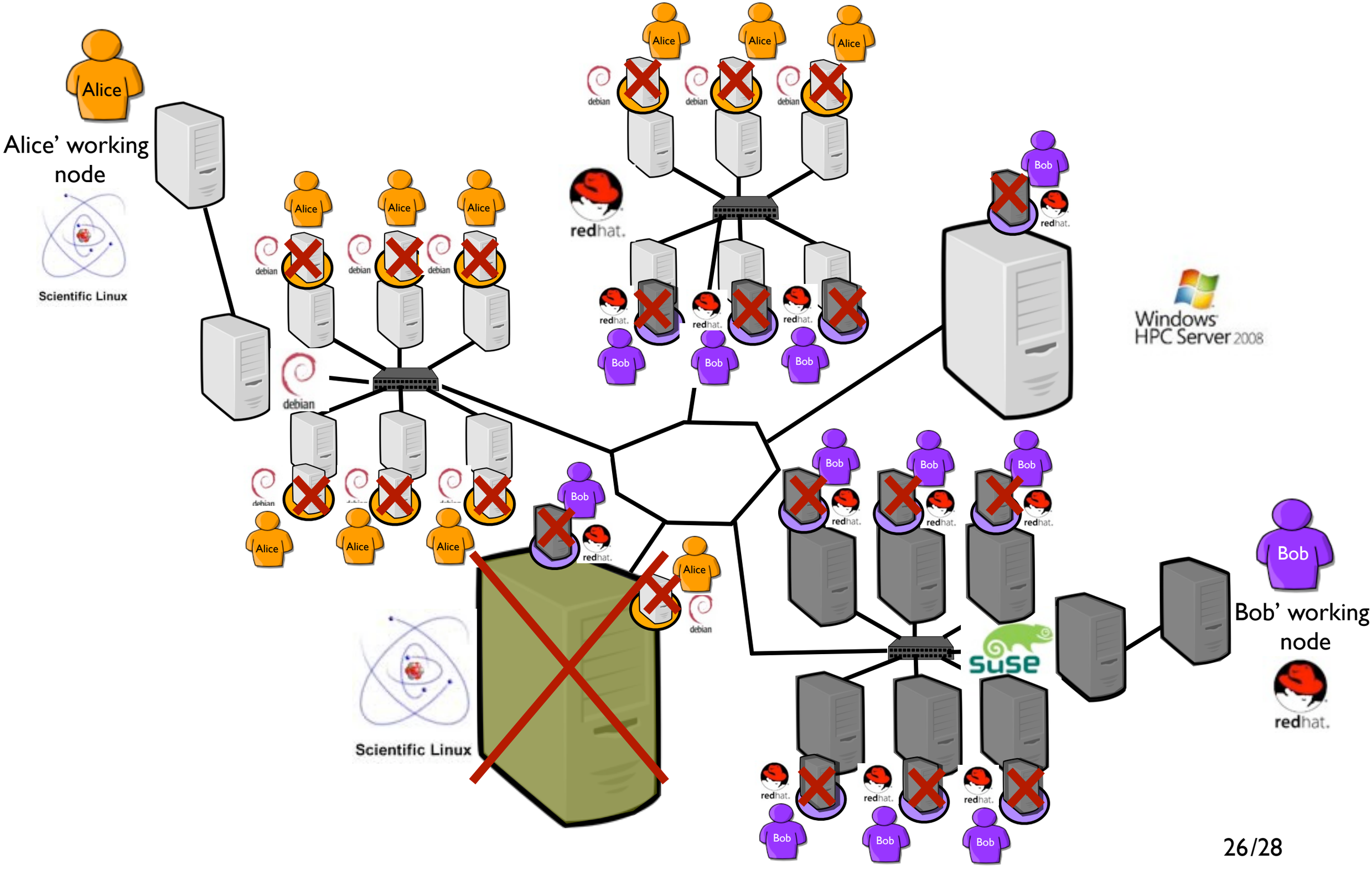
Conclusion

The Alice/Bob Example



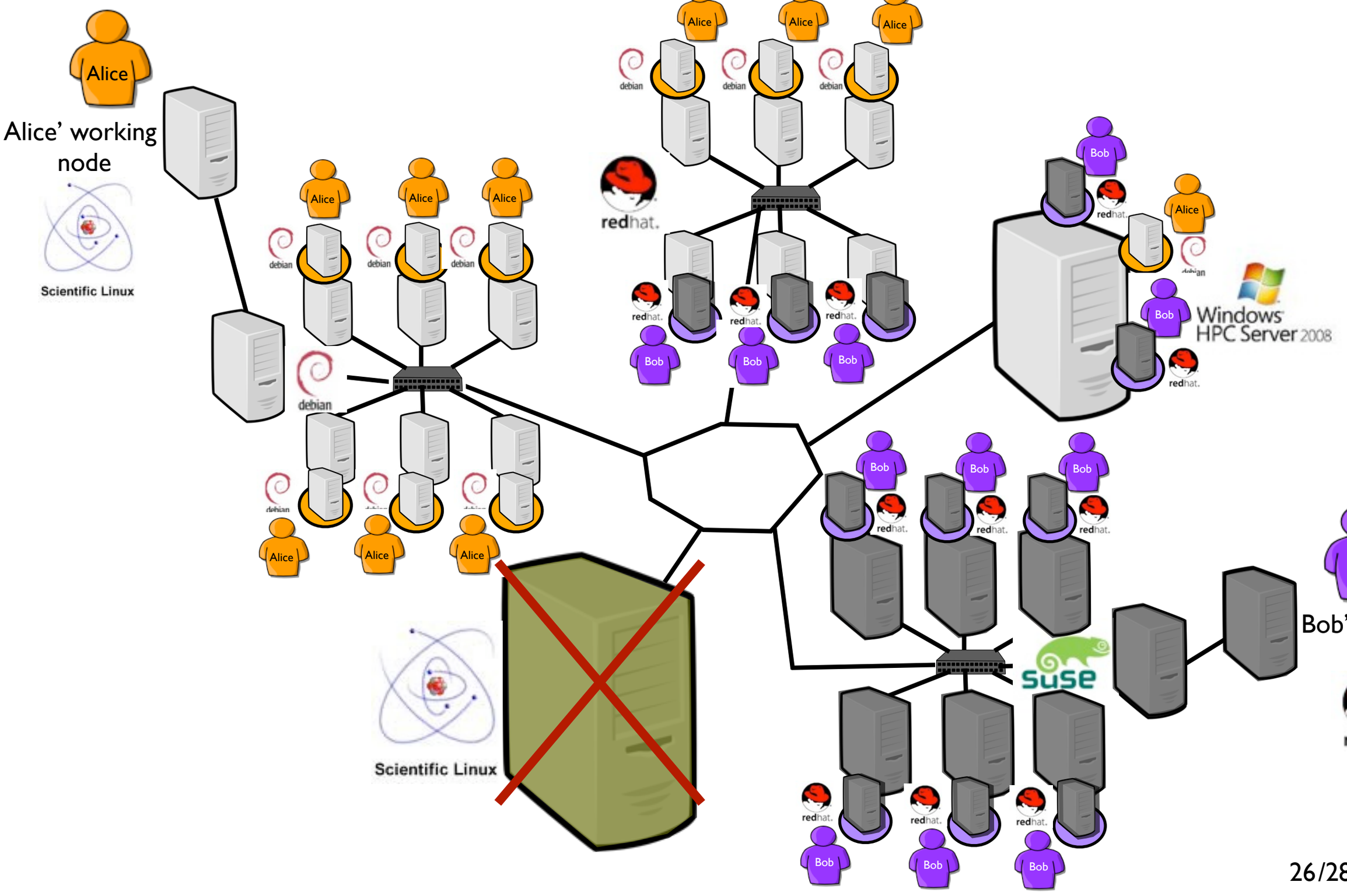
Conclusion

The Alice/Bob Example



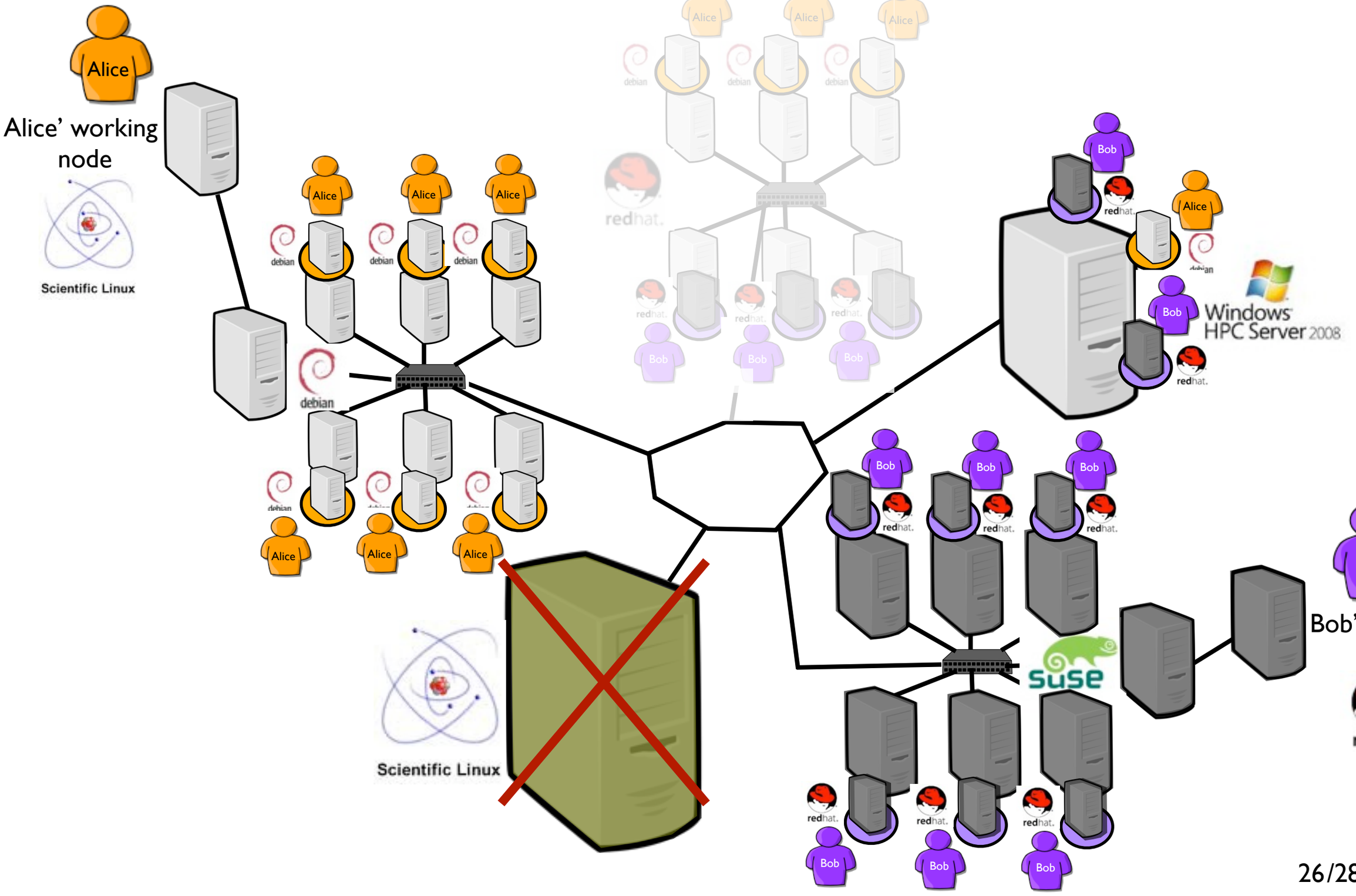
Conclusion

The Alice/Bob Example



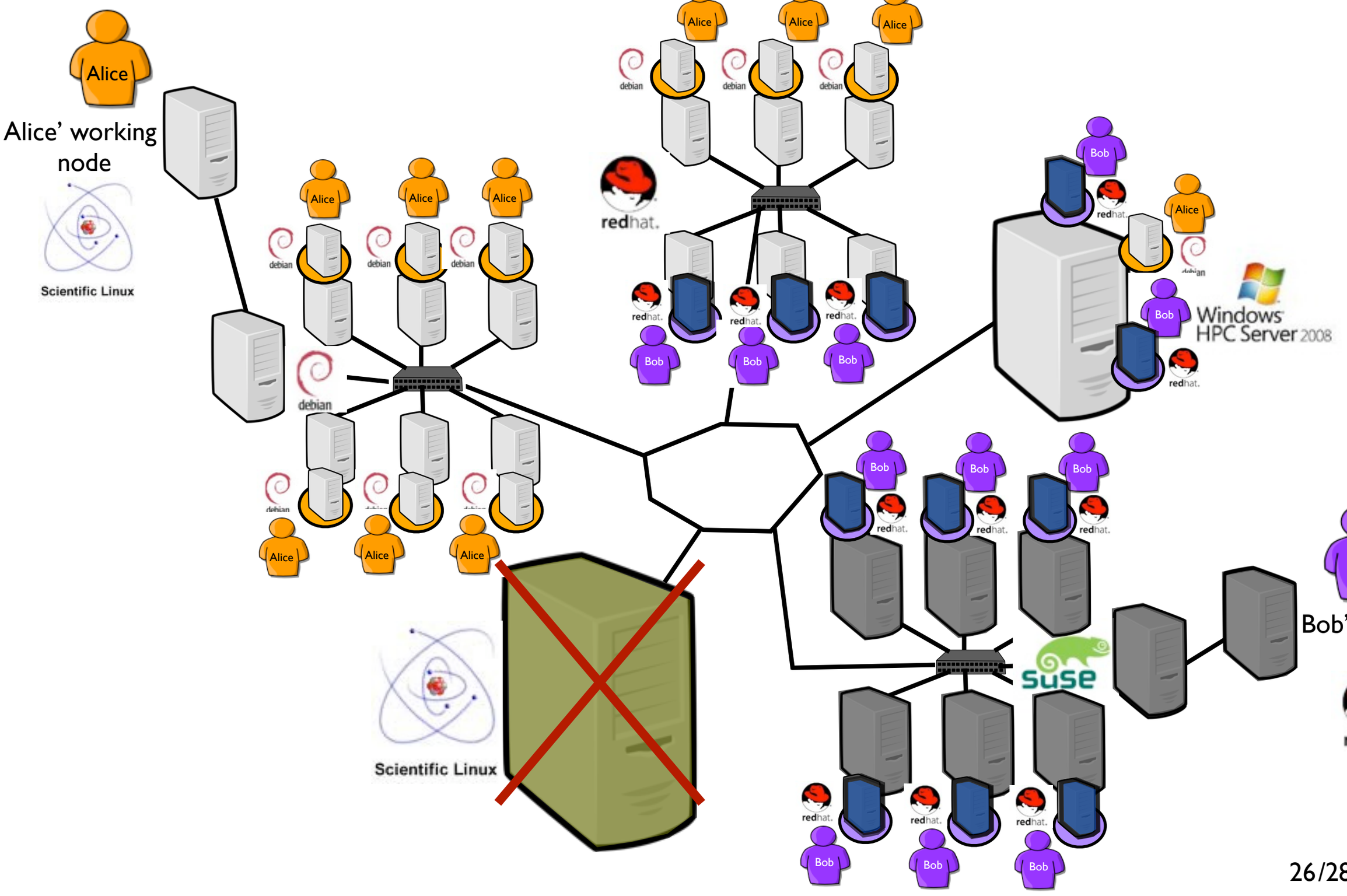
Conclusion

The Alice/Bob Example



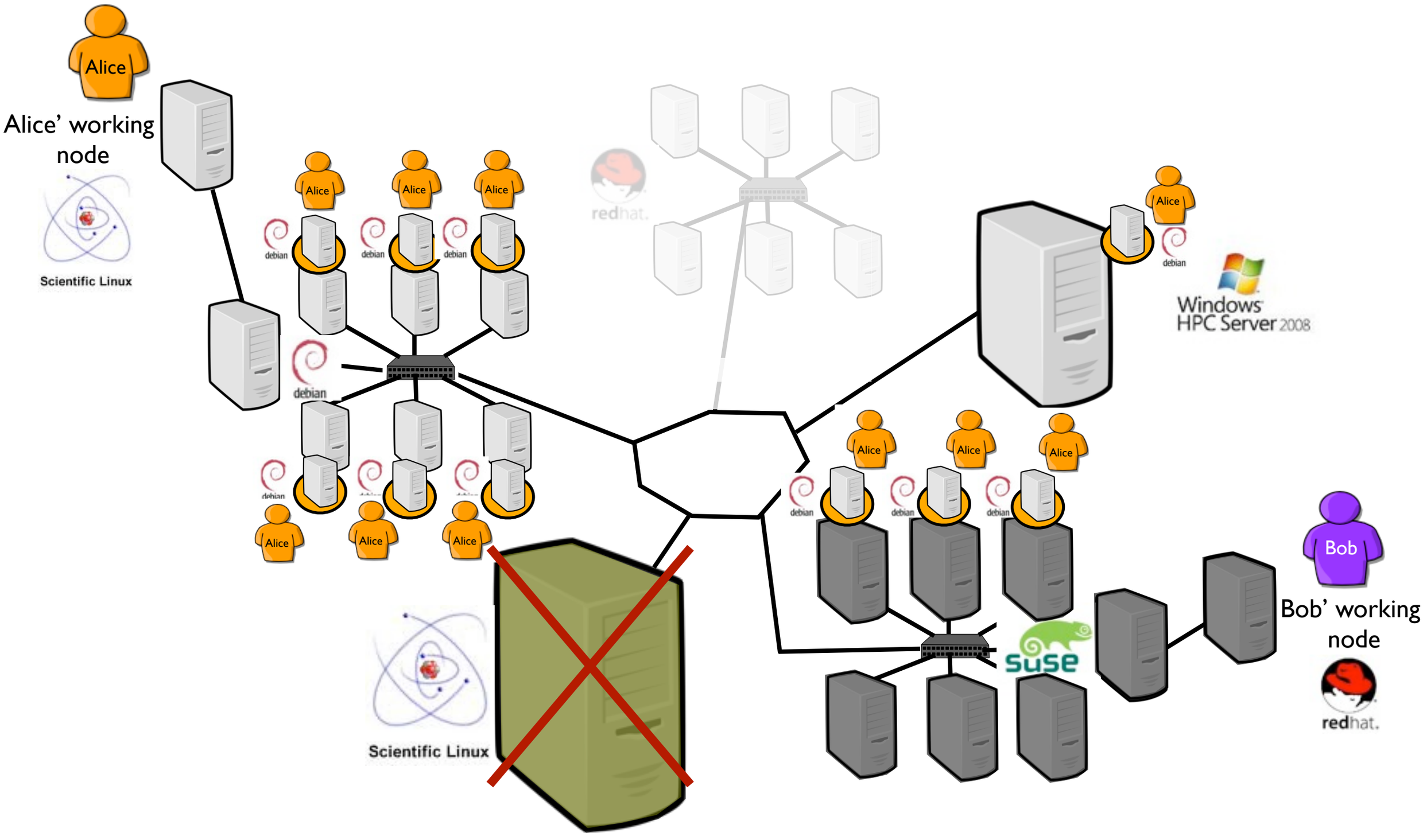
Conclusion

The Alice/Bob Example



Conclusion

The Alice/Bob Example



Conclusion

What a Grid !?!

- System Virtualization provides mature "techniques for organizing computer systems resources to provide extraordinary system flexibility" Golberg, 1974
- Leveraging previous works is mandatory
- System Virtualization leads to new concerns

Storage (intersite management of VM images/VM migrations)
Network IP management (suspend/resume)

Scalability (More VMs, More PMs, More VMs ...)

Advanced policies for vjobs management
Coordination between users and administrators expectations

Conclusion

xxxx Computing

- xxxx as Utility

  “We will probably see the spread of *computer utilities,* which, like present electric and telephone utilities, will service individual homes and offices across the country”

Conclusion

xxxx Computing

- xxxx as Utility

"We will probably see the spread of *computer utilities,* which, like present electric and telephone utilities, will service individual homes and offices across the country"

Len Kleinrock, 1969

credits: I. Foster 2002 What is the C ...t Checklist

1961 !

Thank you / Questions ?

How Virtualization Changed The Grid Perspective

Adrien Lèbre
Ecole des Mines de Nantes

"Des grilles aux Clouds, nouveaux problèmes et nouvelles solutions"
13 December 2010, Ecole Normale Supérieure de Lyon

Bibliography

- **Survey of Virtual Machine Research**
  By Goldberg, 1974

- **Virtual Machines, versatile platforms for systems and processes,**
  By J.E. Smith, R. Nair, Editions Elsevier, 2005

- **HiperCal ANR**, http://hipcal.lri.fr/
  Specifying and provisioning Virtual Infrastructures with HIPerNET
  By Fabienne Anhalt, Guilherme Koslovski, and Pascale Vicat-Blanc Primet. ACM International Journal of Network Management, special issue on Network Virtualization and its Management, 2010

- Entropy: http://entropy.gforge.inria.fr
  Entropy: a consolidation manager for clusters,
  By Hermenier and all, proceedings of the 2009 international conference on Virtual execution environments
  Cluster-Wide Context Switch of Virtualized Jobs
  By Hermenier and all, INRIA Research Report, May 2009

- **Saline: Improving Best-Effort Job Management in Grids**
  By Jérôme Gallard, Adrien Lèbre, Christine Morin, INRIA Research Report 7055, Sept 2009

- Further informations on the **Grid'5000 Virtualization Working Group**:
  https://www.grid5000.fr/mediawiki/index.php/Virtualization_Working_Group

How Virtualization Changed The Grid Perspective

Adrien Lèbre
Ecole des Mines de Nantes

"Des grilles aux Clouds, nouveaux problèmes et nouvelles solutions"
13 December 2010, Ecole Normale Supérieure de Lyon