Hype and trends

Arnaud Legrand, CNRS, University of Grenoble

LIG laboratory, arnaud.legrand@imag.fr

November 28, 2011

Outline

1 Virtualization

2 Toward Exascale

HPC and Clouds: Twins separated at birth

Context: Next Generation Systems

- High Performance Computing: scientific computation
- Clouds: computer system underlying most IT services in few years

Question: How do these systems relate to each other?

- Science and Business are rather different processes
- Yet, the computers are the same

Goal of the presentation

- Compare these system kinds (and some others)
- Understand what Performance means in each community
- Hopefully nurture each community with new ideas (cross-fertilization)

Disclaimers

- I'm by no way a business expert: I'm a civil servant, shielded from the reality
- But I'm a computer scientist specialist of distributed computing systems

Science and Business

There is no "vs." in the title

- Scientific Organizations rely on business processes: Management, Projects
- Businesses may rely on science: Pharmaceutic, Engineering
- Let's focus on problem requirements, not the ones of the organizations

Running a business

- Governed by own rules (amongst which, relevant laws)
- Focus: selling products
- Others: Reporting; Decision making; Interact with uncontrolled organizations
- Lot of small unrelated transactions, no "big problem" (beside maintaining state)

Doing Science

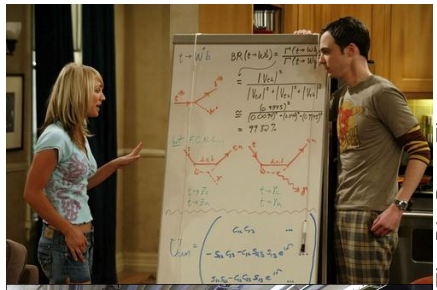
- Governed by much more immutable laws (that are partially unknown)
- Goal: *Predict* the outcome of an element combination
- Focus: Understand the laws allowing to compute the response of elements
- "Problems" here means "Business Opportunities" there (my speech is biased)

How does Science work?

Proposed theories remain valid until proved false (or better proposed)

Classical approaches in science and engineering

1. Theoretical work: equations on a board
2. Experimental study on an scientific instrument


The Big Bang Theory

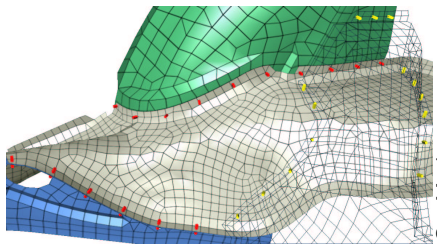

Large Hardron Collider

That's not always desirable (or even possible)

- Some phenomenons are intractable theoretically
- Experiments too expensive, difficult, slow, dangerous

The third scientific way: *Computational Science*

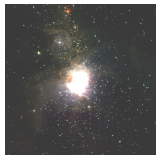
3. Study in silico using computers
   Modeling / Simulation of the phenomenon
   Data Mining also used sometimes

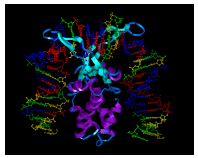

Car Mesh

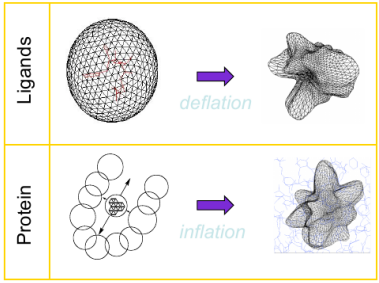
That is why we have High Performance Computing systems

Courtesy of Martin Quinson

Some Particularly Challenging Computations

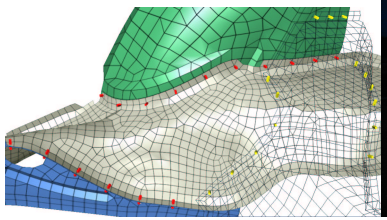






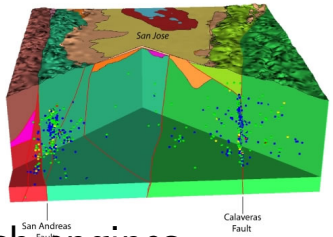
Science

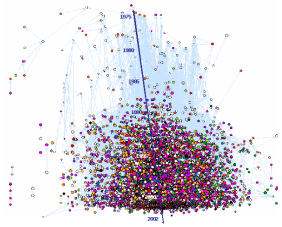
- Global climate modeling
- Astrophysical modeling
- Biology (genomics; protein folding; drug design)
- Computational Chemistry
- Computational Material Sciences and Nanosciences



Engineering

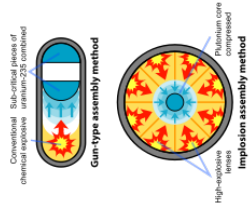
- Crash simulation
- Semiconductor design
- Earthquake and structural modeling
- Computation fluid dynamics (airplane design)
- Combustion (engine design)





Business and Humanities

- Financial and Economic modeling
- Transaction processing, web services and search engines
- Social Networking



Defense

- Nuclear weapons – tested by simulations
- Cryptography

Courtesy of Martin Quinson

Performance in Scientific Computations

Scientific Problems are Large

- The finer the Mesh, the better the Prediction: need more points for quality
  Forecast prediction: hundreds of km: one day ahead; 1 week ahead: kilometers
- Some intrinsically large problems (cosmology, atom studies, etc)

We want the result quickly

- Need to run numerous experiments to find the one invalidating the theory

$\rightsquigarrow$ Computer systems devoted to science: the biggest existing ones

- Large amount of interconnected processing units
- High bandwidth, low latency Networks (never rely on the Internet!)

Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\rightsquigarrow$ the business uses computers without relying on them
- Computer systems distributed iff the company is: interconnect business units

And then came the Internet

- Some company relying on the Internet emerged (eBay, amazon, google)
- Computers naturally play a central role in their business plan
- Cannot afford to loose clients $\rightsquigarrow$ High **Availability** Computing

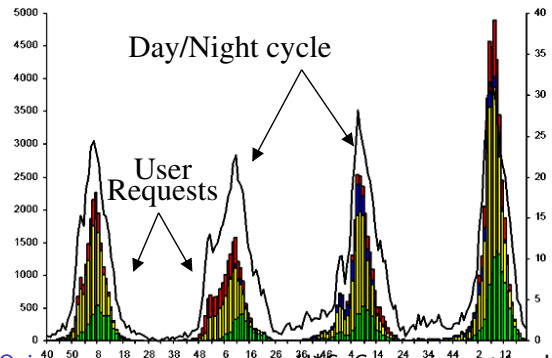
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\rightsquigarrow$ the business uses computers without relying on them
- Computer systems distributed iff the company is: interconnect business units

And then came the Internet

- Some company relying on the Internet emerged (eBay, amazon, google)
- Computers naturally play a central role in their business plan
- Cannot afford to loose clients $\rightsquigarrow$ High **Availability** Computing
- But load is very changing



Courtesy of Martin Quinson

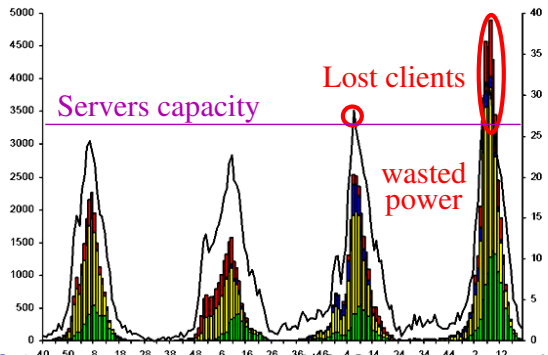
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\rightsquigarrow$ the business uses computers without relying on them
- Computer systems distributed iff the company is: interconnect business units

And then came the Internet

- Some company relying on the Internet emerged (eBay, amazon, google)
- Computers naturally play a central role in their business plan
- Cannot afford to loose clients $\rightsquigarrow$ High **Availability** Computing
- But load is very changing $\rightsquigarrow$ Servers dimensioned for flash crowds



Courtesy of Martin Quinson

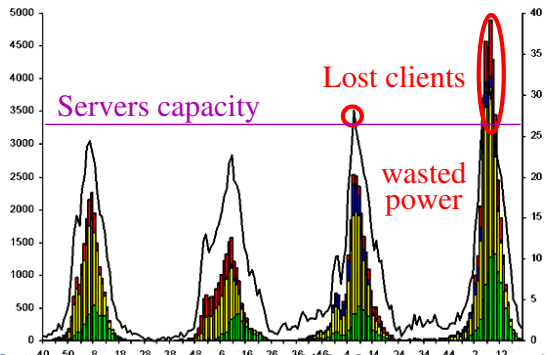
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\rightsquigarrow$ the business uses computers without relying on them
- Computer systems distributed iff the company is: interconnect business units

And then came the Internet

- Some company relying on the Internet emerged (eBay, amazon, google)
- Computers naturally play a central role in their business plan
- Cannot afford to loose clients $\rightsquigarrow$ High **Availability** Computing
- But load is very changing $\rightsquigarrow$ Servers dimensioned for flash crowds



Amazon idea

- Rent unused power to others!
- Computers better amortized
  Buy bigger ones, loose no client
- Infrastructure as a Service (IaaS)
- Highly Cost-Efficient Computing

Courtesy of Martin Quinson

Here Come the Clouds

Client Incitatives

- IT maintenance burden assumed by external specialists
- Pay only used power: rent a server 1h, send computations *in the cloud*, enjoy
  This is called Elastic Computing
- The created need revealed very profound: everyone wants it now
- Clients even want to rent OS+apps (PaaS) or software (SaaS)

Virtualization

- Installing an OS: $\approx$ one hour. Not flexible enough.
- Rent virtual machines instead: overprovisionning and other optimizations

The Data Centers Growth

- Scale allows Cost Cuttings, as always. Motivation for big DC already existed
- Clouds removes the wastes due to over-dimensioning

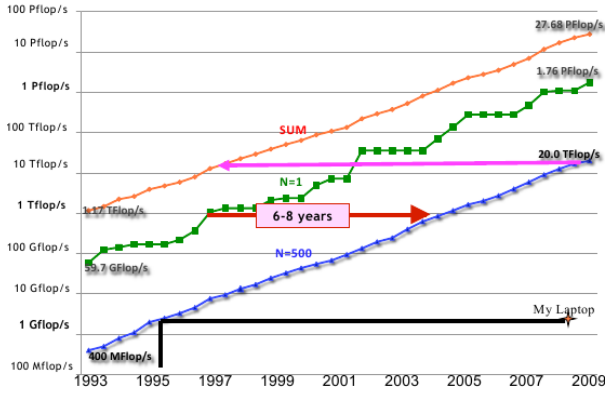
$\Rightarrow$ Corporate Data Centers become as big as Scientific Supercomputers!

- ...and share the same difficulties. The twins are technically reconciled ☺

How big are these machines?

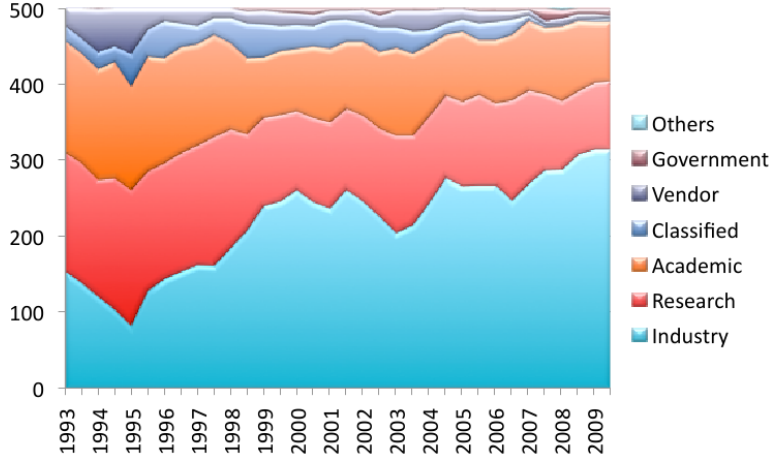
There is an International Ranking

- TOP500: updated twice a year since 1993
- Computational power growth: Exponential
- My laptop is a 10 years old supercomputer! (and my phone is a 10 years old desktop)



Machine usage

- 60% used by the industry
- The industry does science for sure
- But the increase is now due to clouds
- Some of this machines are classified HPC and Cloud don't need to argue: The big players are intelligences :)



Courtesy of Martin Quinson

42: The Answer to the Ultimate Question of Life, the Universe, and Everything

On Monday November 14th 2011, the Top 500 Supercomputer list was updated.

Rank	Site	Computer/Year	Vendor	Cores	Rmax	Rpeak	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011	Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010	NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009	Cray Inc.	224162	1759.00	2331.00	6950.0
...							
42	Amazon Web Services United States	Amazon EC2 Cluster, Xeon 8C 2.60GHz, 10G Ethernet / 2011	Self-made	17024	240.09	354.10	??

From http://perspectives.mvdirona.com

42: The Answer to the Ultimate Question of Life, the Universe, and Everything

On Monday November 14th 2011, the Top 500 Supercomputer list was updated.

Rank	Site	Computer/Year	Vendor	Cores	Rmax	Rpeak	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011	Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010	NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009	Cray Inc.	224162	1759.00	2331.00	6950.0
...							
42	Amazon Web Services United States	Amazon EC2 Cluster, Xeon 8C 2.60GHz, 10G Ethernet / 2011	Self-made	17024	240.09	354.10	??

- Virtualization Tax is Now Affordable.

  When Cray 1 supercomputer was announced in 1976, it didn't even use virtual memory. It was believed at the time that only real-mode memory access could deliver the performance needed. Now virtual memory in a guest operating system running under a hypervisor.

From http://perspectives.mvdirona.com

42: The Answer to the Ultimate Question of Life, the Universe, and Everything

On Monday November 14th 2011, the Top 500 Supercomputer list was updated.

Rank	Site	Computer/Year	Vendor	Cores	Rmax	Rpeak	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011	Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010	NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009	Cray Inc.	224162	1759.00	2331.00	6950.0
...							
42	Amazon Web Services United States	Amazon EC2 Cluster, Xeon 8C 2.60GHz, 10G Ethernet / 2011	Self-made	17024	240.09	354.10	??

- Virtualization Tax is Now Affordable.
- Commodity Networks can Compete with IB, Myrinet, etc.:

This is the only Top500 entrant below number 128 on the list that is not running either Infiniband or a proprietary, purpose-built network. This result at #42 is an all Ethernet network showing that a commodity network, if done right, can produce industry leading performance numbers.
What's the secret? 10Gbps directly the host is the first part. The second is full non-blocking networking fabric (clos network) where all systems can communicate at full line rate at the same time.

From http://perspectives.mvdirona.com

42: The Answer to the Ultimate Question of Life, the Universe, and Everything

On Monday November 14th 2011, the Top 500 Supercomputer list was updated.

Rank	Site	Computer/Year	Vendor	Cores	Rmax	Rpeak	Power
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect / 2011	Fujitsu	705024	10510.00	11280.38	12659.9
2	National Supercomputing Center in Tianjin China	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 / 2010	NUDT	186368	2566.00	4701.00	4040.0
3	DOE/SC/Oak Ridge National Laboratory United States	Cray XT5-HE Opteron 6-core 2.6 GHz / 2009	Cray Inc.	224162	1759.00	2331.00	6950.0
...							
42	Amazon Web Services United States	Amazon EC2 Cluster, Xeon 8C 2.60GHz, 10G Ethernet / 2011	Self-made	17024	240.09	354.10	??

- Virtualization Tax is Now Affordable.
- Commodity Networks can Compete with IB, Myrinet, etc.:
- Anyone can own a Supercomputer for an hour:
  You can have a top50 supercomputer for under $2,600/hour

From http://perspectives.mvdirona.com

Outline

1 Virtualization

2 Toward Exascale

Dynamic scheduling of virtual machines, scalability and fault tolerance are still the issues!

Adrien Lèbre, Flavien Quesnel
ASCOLA Research Group
Ecole des Mines de Nantes

Courtesy of Adrien Lèbre

How Virtualization Changed The Grid Perspective

Courtesy of Adrien Lèbre

xxx Computing

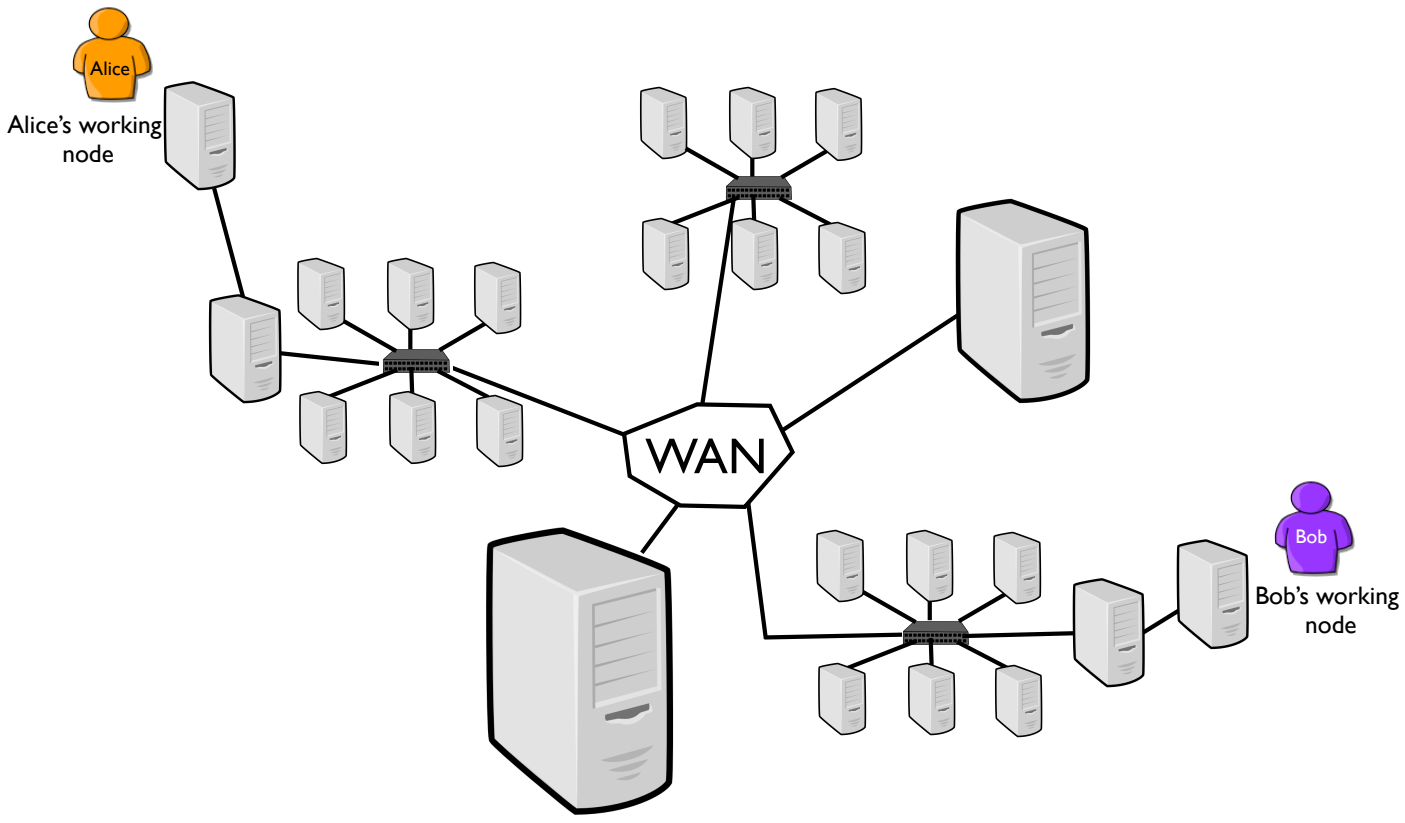
- xxx as Distributed
  (Cluster / Grid / Desktop / "Hive" / Cloud / Sky / ...)
- A common objective

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

3

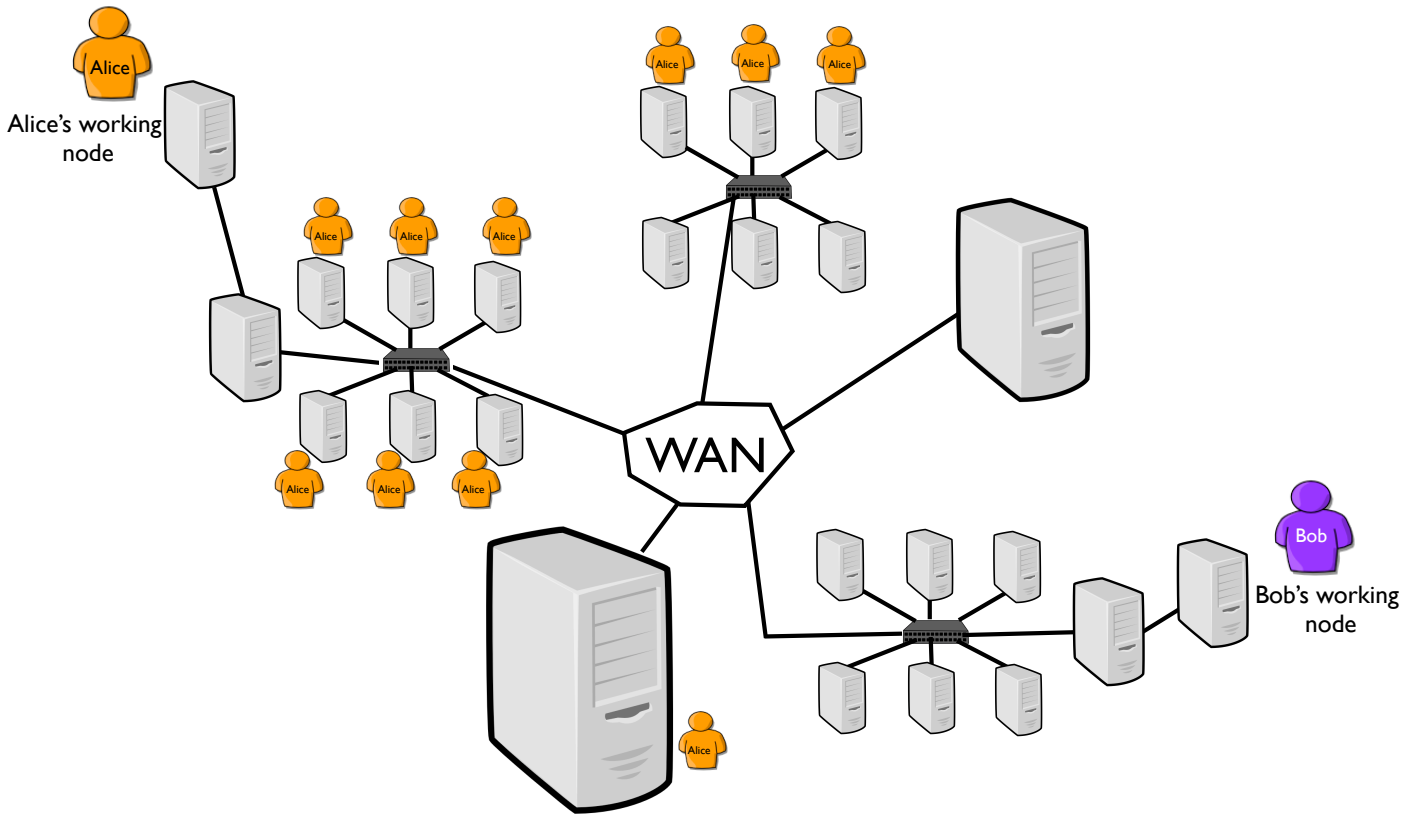
Courtesy of Adrien Lèbre

The Alice/Bob Example



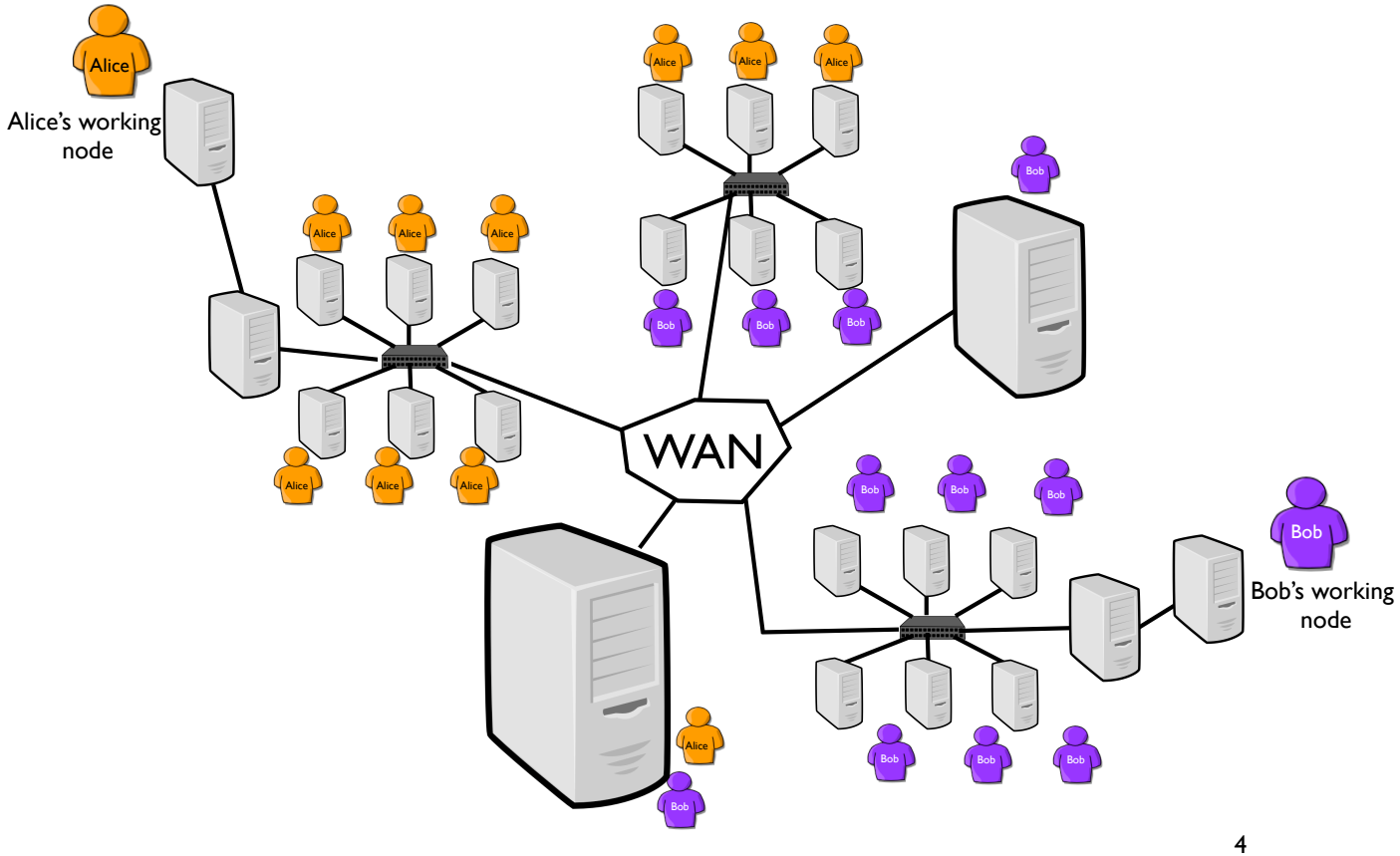
Courtesy of Adrien Lèbre

The Alice/Bob Example

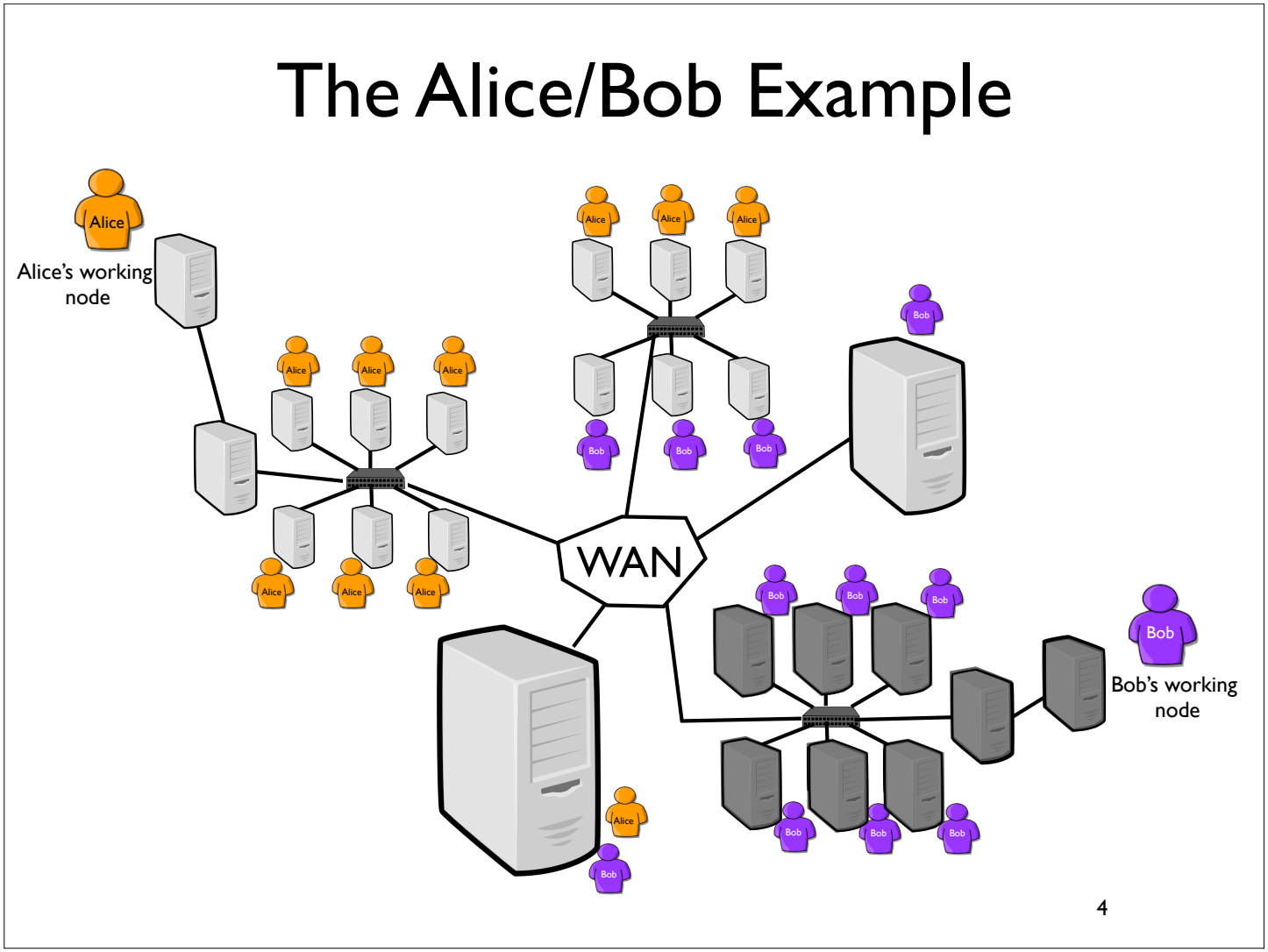


Courtesy of Adrien Lèbre

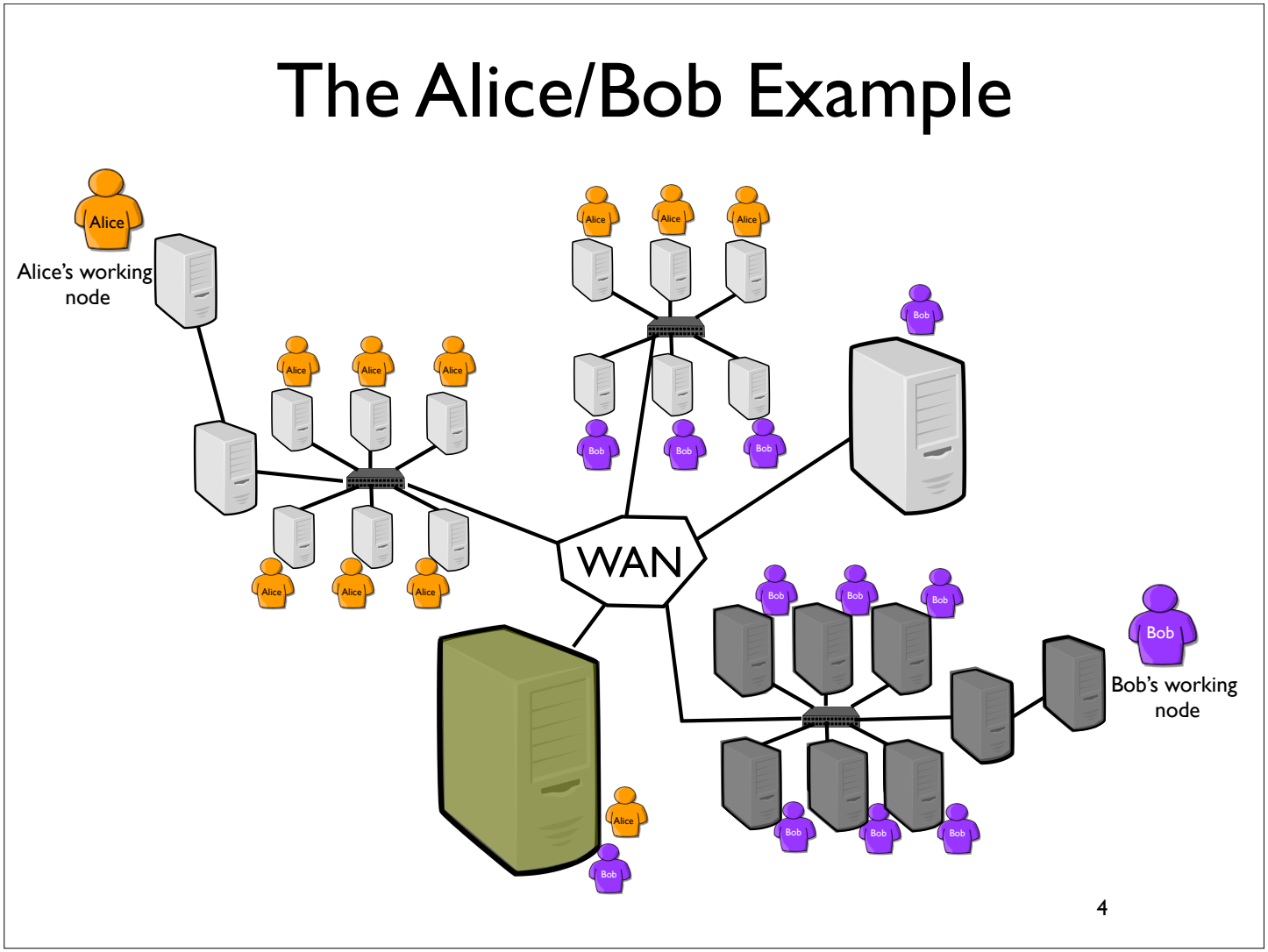
The Alice/Bob Example



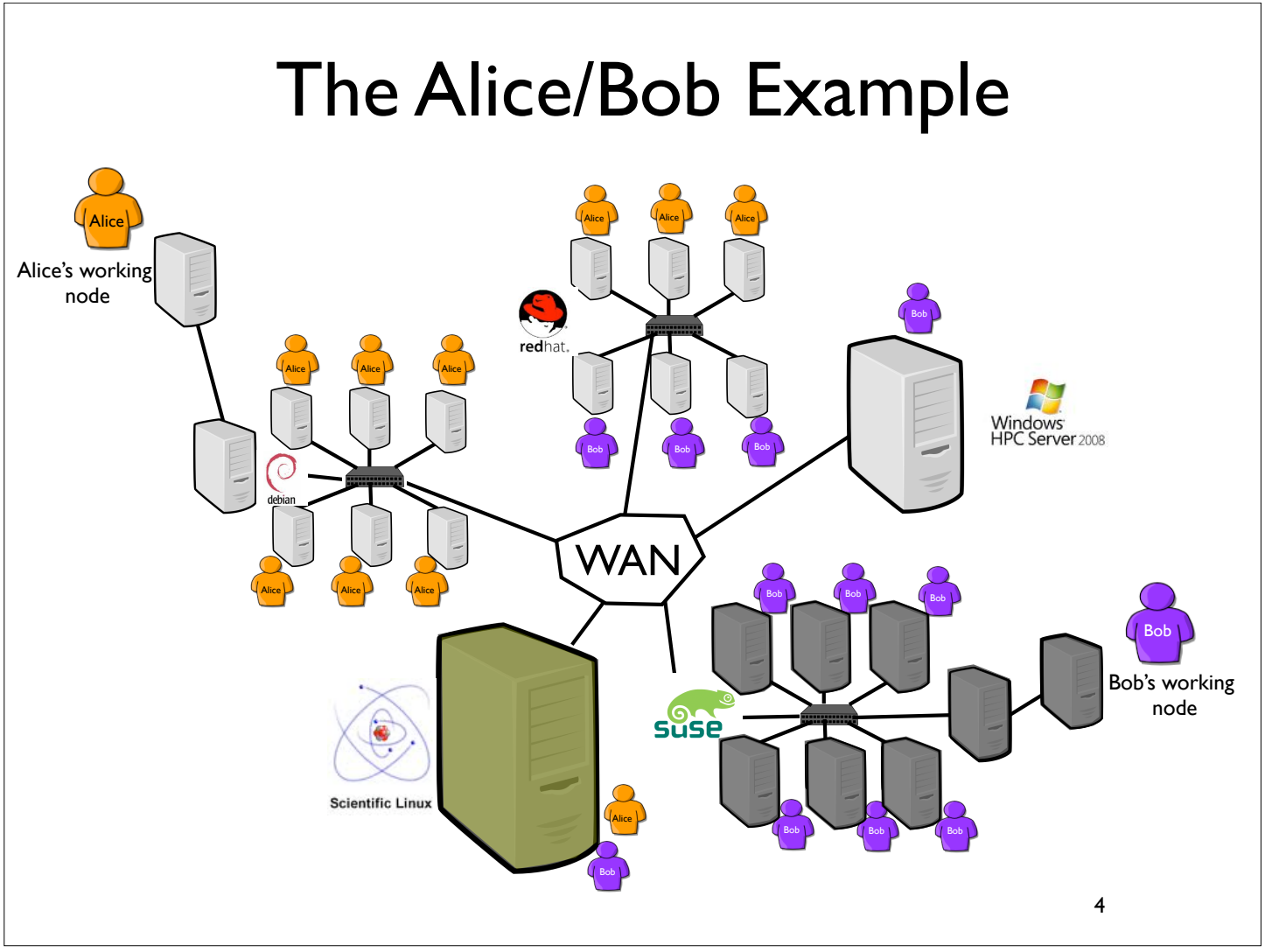
Courtesy of Adrien Lèbre



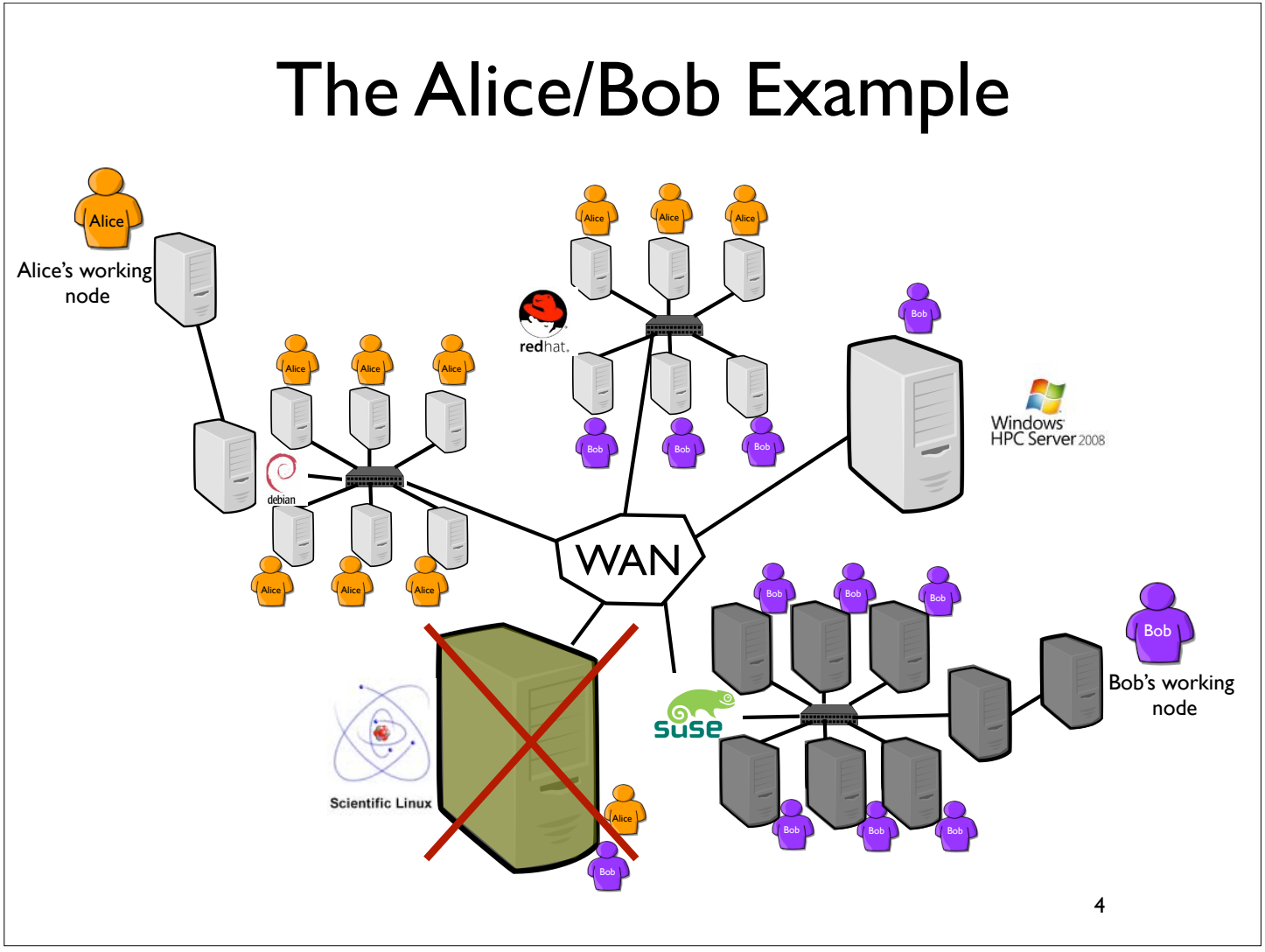
Courtesy of Adrien Lèbre



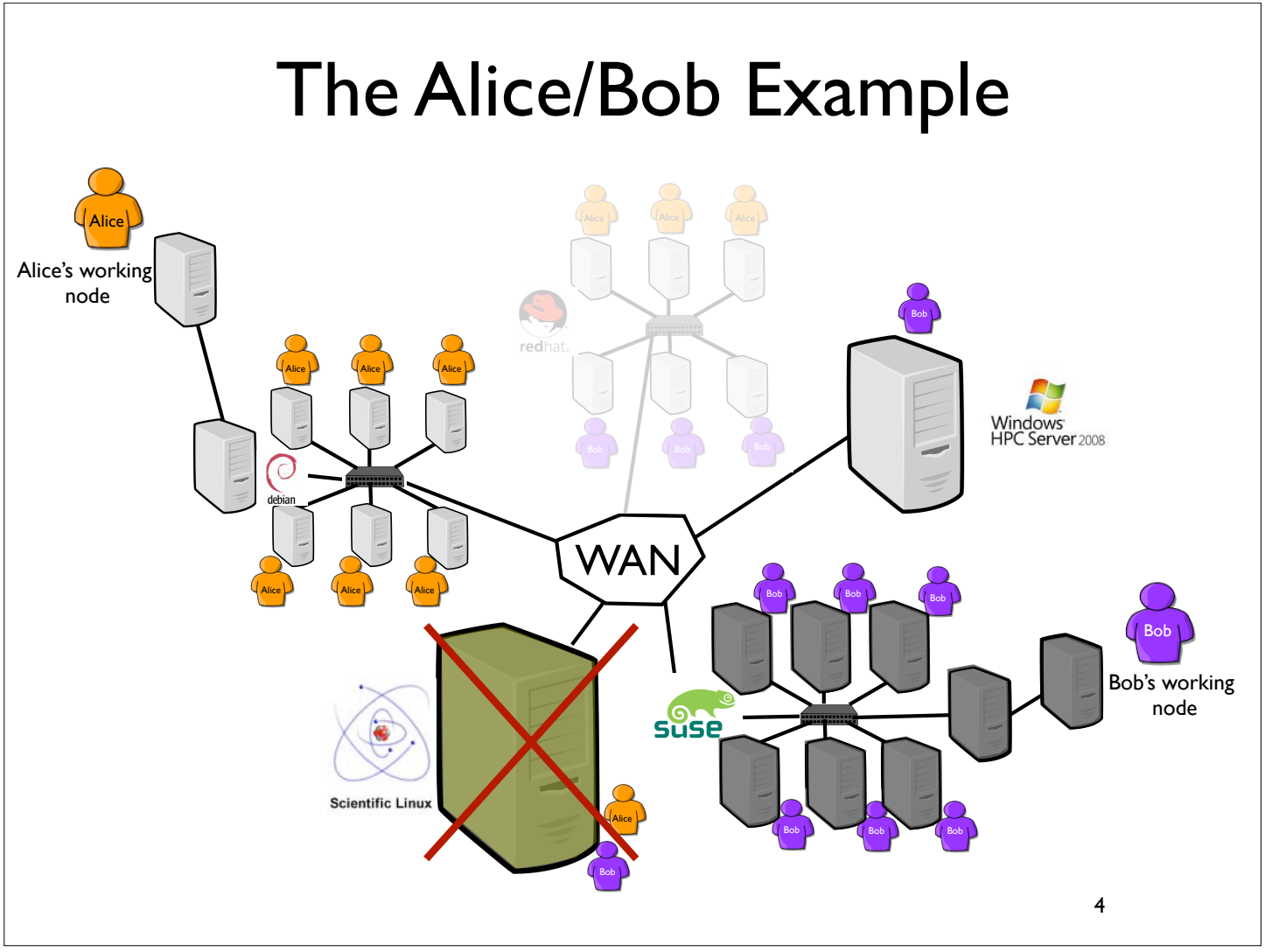
Courtesy of Adrien Lèbre



Courtesy of Adrien Lèbre

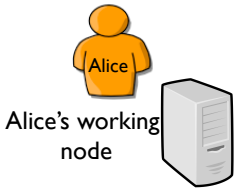


Courtesy of Adrien Lèbre

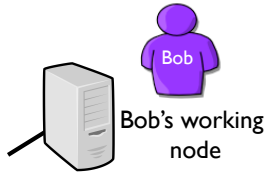


Courtesy of Adrien Lèbre

What a Grid!?!



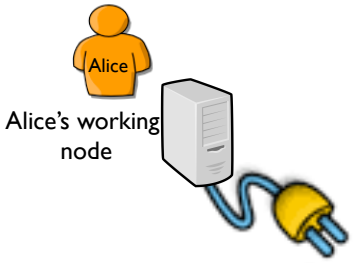
Resource booking (based on user's estimates)
Security concerns (job isolation)
Heterogeneity concerns (hardware and software)
Scheduling limitations (a job cannot be easily relocated)
Fault tolerance issues
...



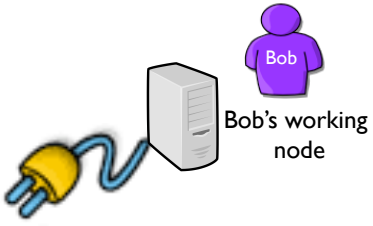
5

Courtesy of Adrien Lèbre

What a Grid!?!



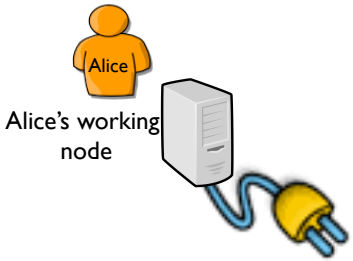
Resource booking (based on user's estimates)
Security concerns (job isolation)
Heterogeneity concerns (hardware and software)
Scheduling limitations (a job cannot be easily relocated)
Fault tolerance issues
...



5

Courtesy of Adrien Lèbre

What a Grid!?!

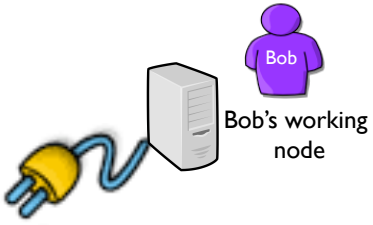


Resource b

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 !

...

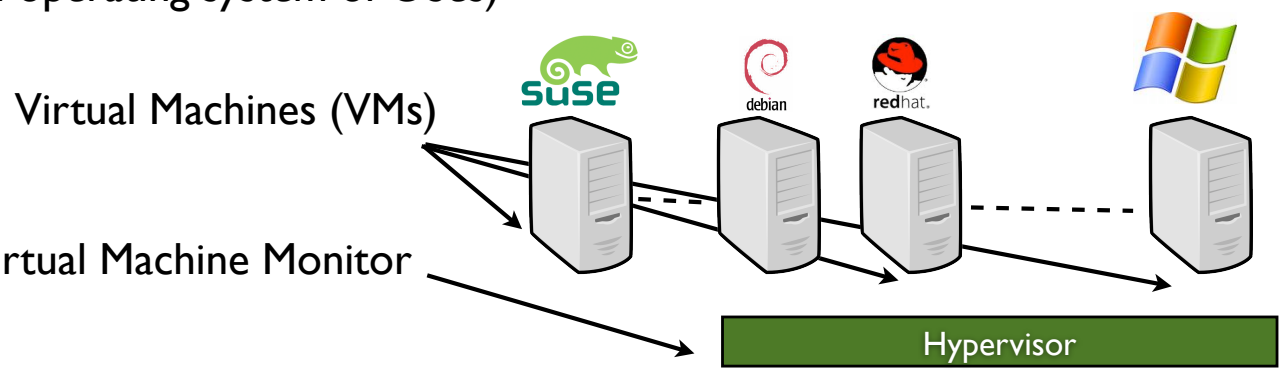


5

Courtesy of Adrien Lèbre

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."

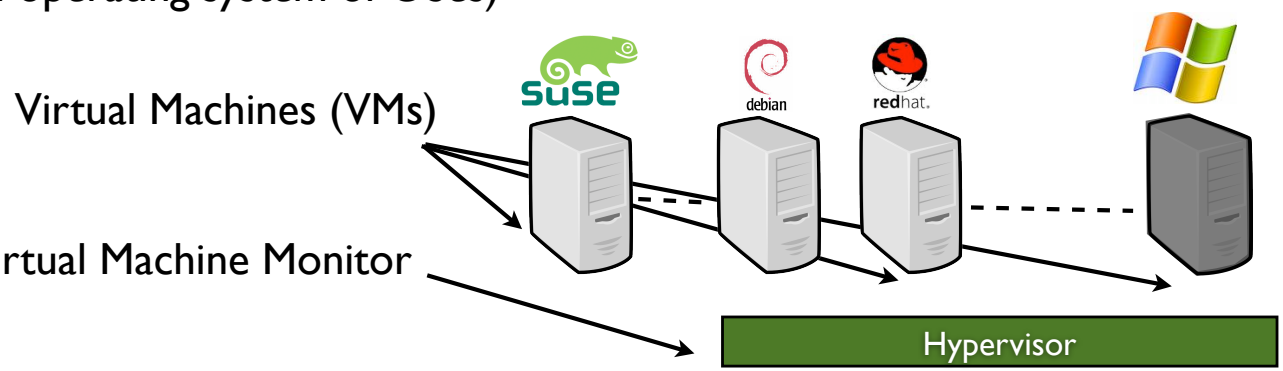
Physical Machine (PM)

6

Courtesy of Adrien Lèbre

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)

6

Courtesy of Adrien Lèbre

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

7

Courtesy of Adrien Lèbre

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)

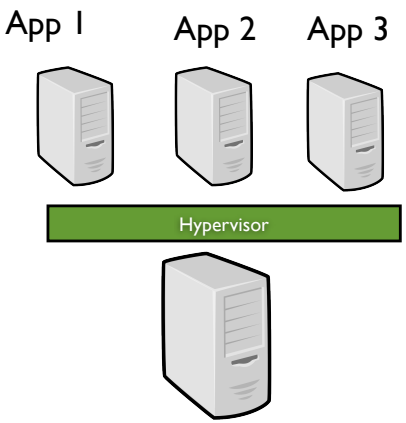
  Revival of System Virtualization approach (VmWare/Xen)

  Hard or soft partitioning of SMP/Numa Server

8

Courtesy of Adrien Lèbre

VM Capabilities

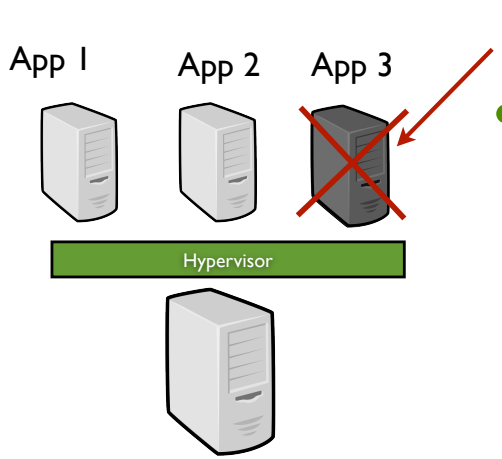


- Isolation ("security" between each VM)

9

Courtesy of Adrien Lèbre

VM Capabilities

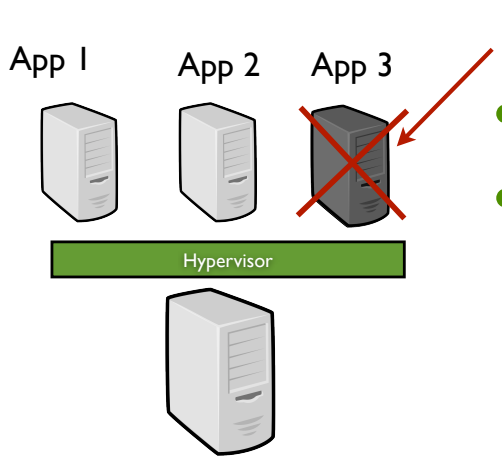


- Isolation ("security" between each VM)

9

Courtesy of Adrien Lèbre

VM Capabilities

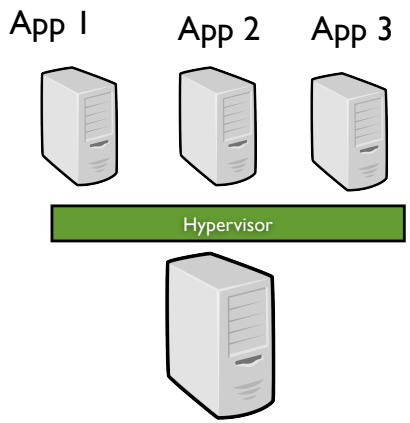


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)

9

Courtesy of Adrien Lèbre

VM Capabilities

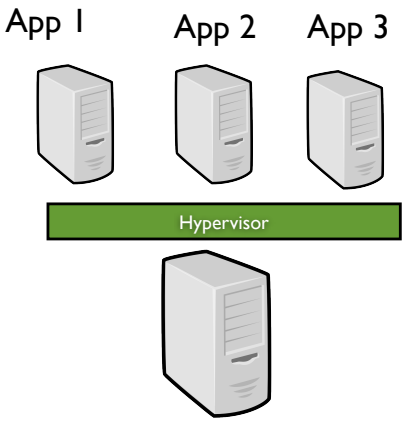


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)

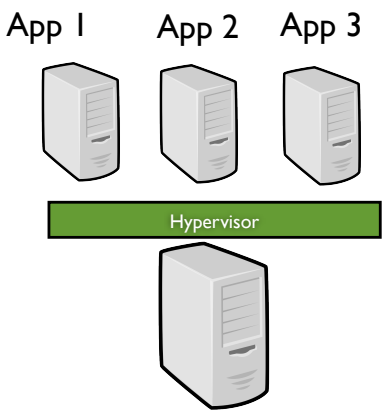
9

Courtesy of Adrien Lèbre

VM Capabilities



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

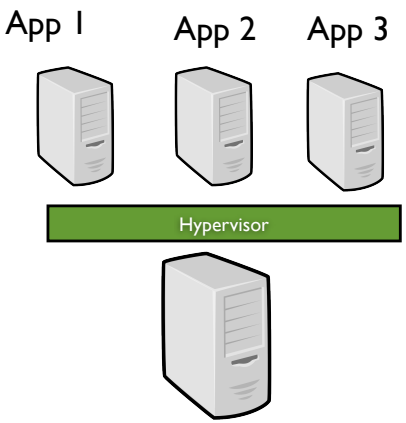


- Suspend/Resume

9

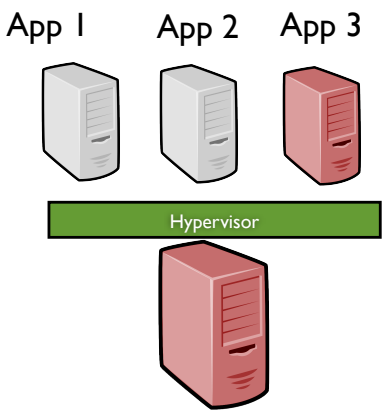
Courtesy of Adrien Lèbre

VM Capabilities



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

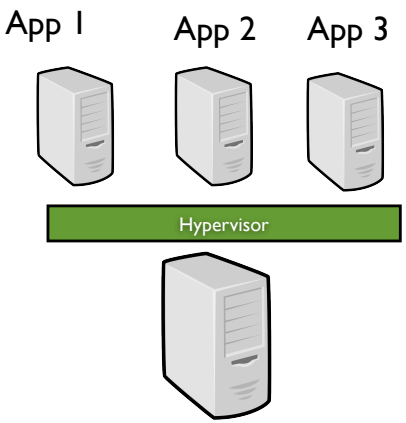
- Suspend/Resume



9

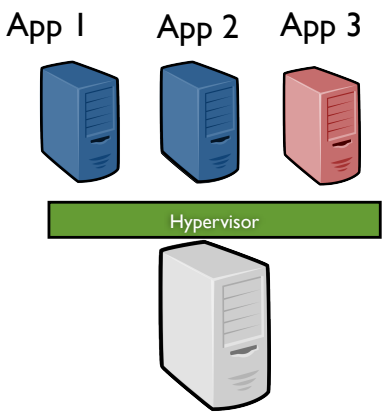
Courtesy of Adrien Lèbre

VM Capabilities



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

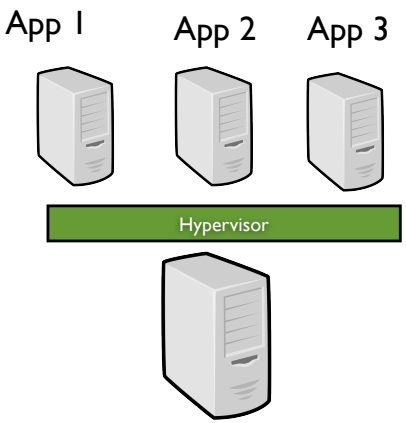
- Suspend/Resume



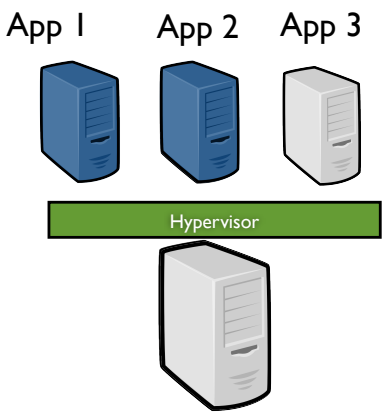
9

Courtesy of Adrien Lèbre

VM Capabilities



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

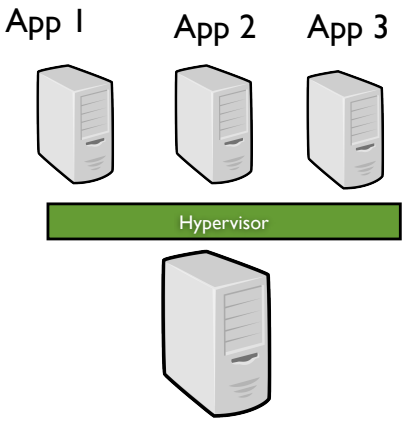


- Suspend/Resume

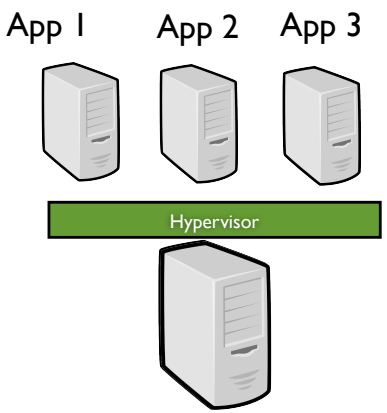
9

Courtesy of Adrien Lèbre

VM Capabilities



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

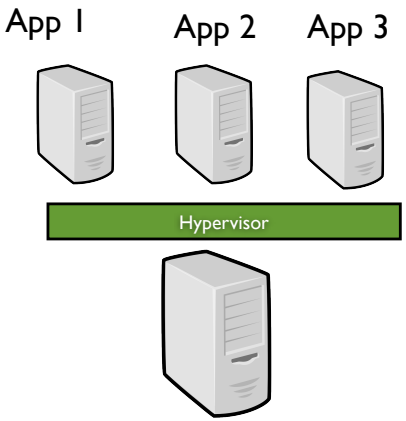


- Suspend/Resume

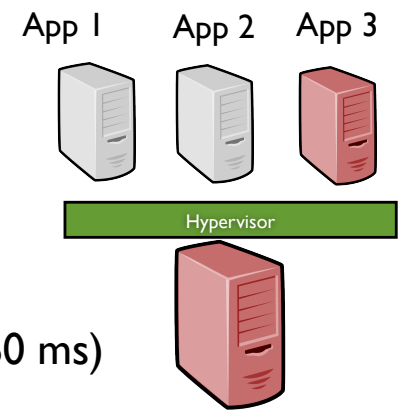
9

Courtesy of Adrien Lèbre

VM Capabilities



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

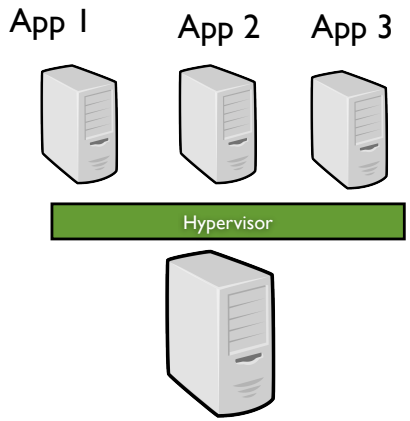


- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)

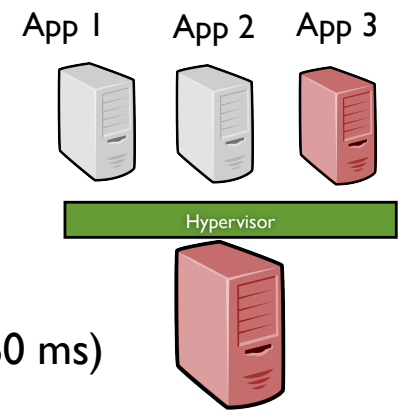
9

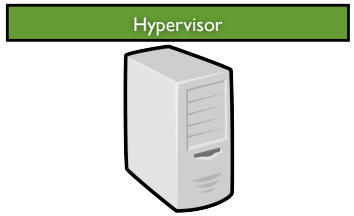
Courtesy of Adrien Lèbre

VM Capabilities



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



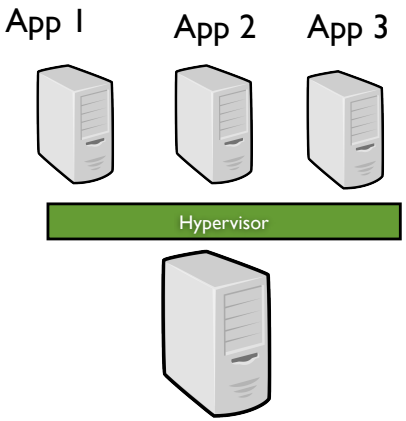


- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)

9

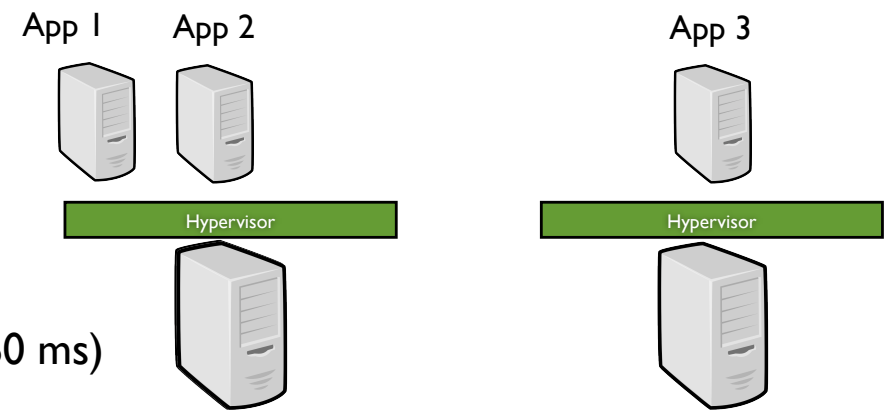
Courtesy of Adrien Lèbre

VM Capabilities



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

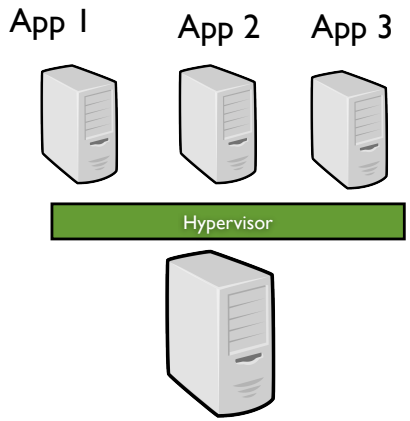
- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)



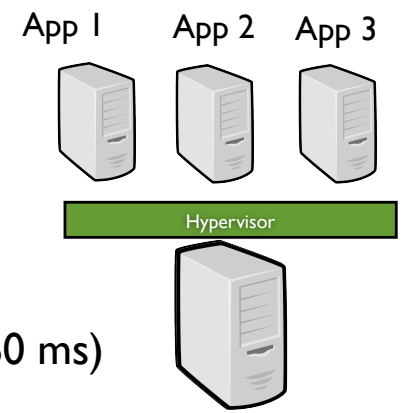
9

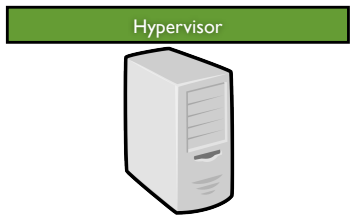
Courtesy of Adrien Lèbre

VM Capabilities



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



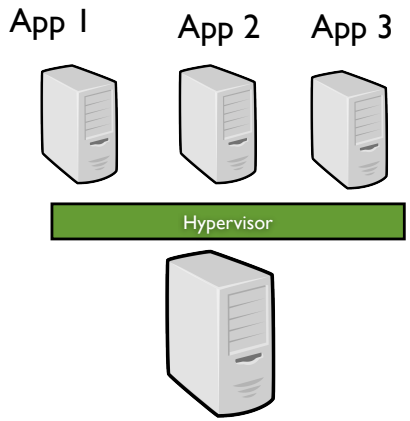


- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)

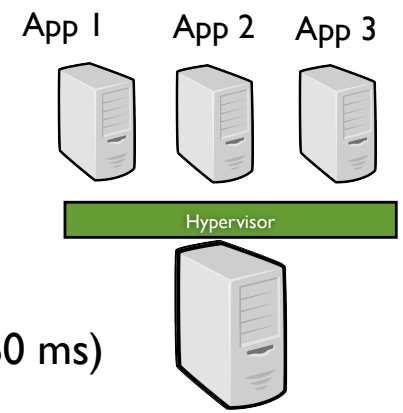
9

Courtesy of Adrien Lèbre

VM Capabilities



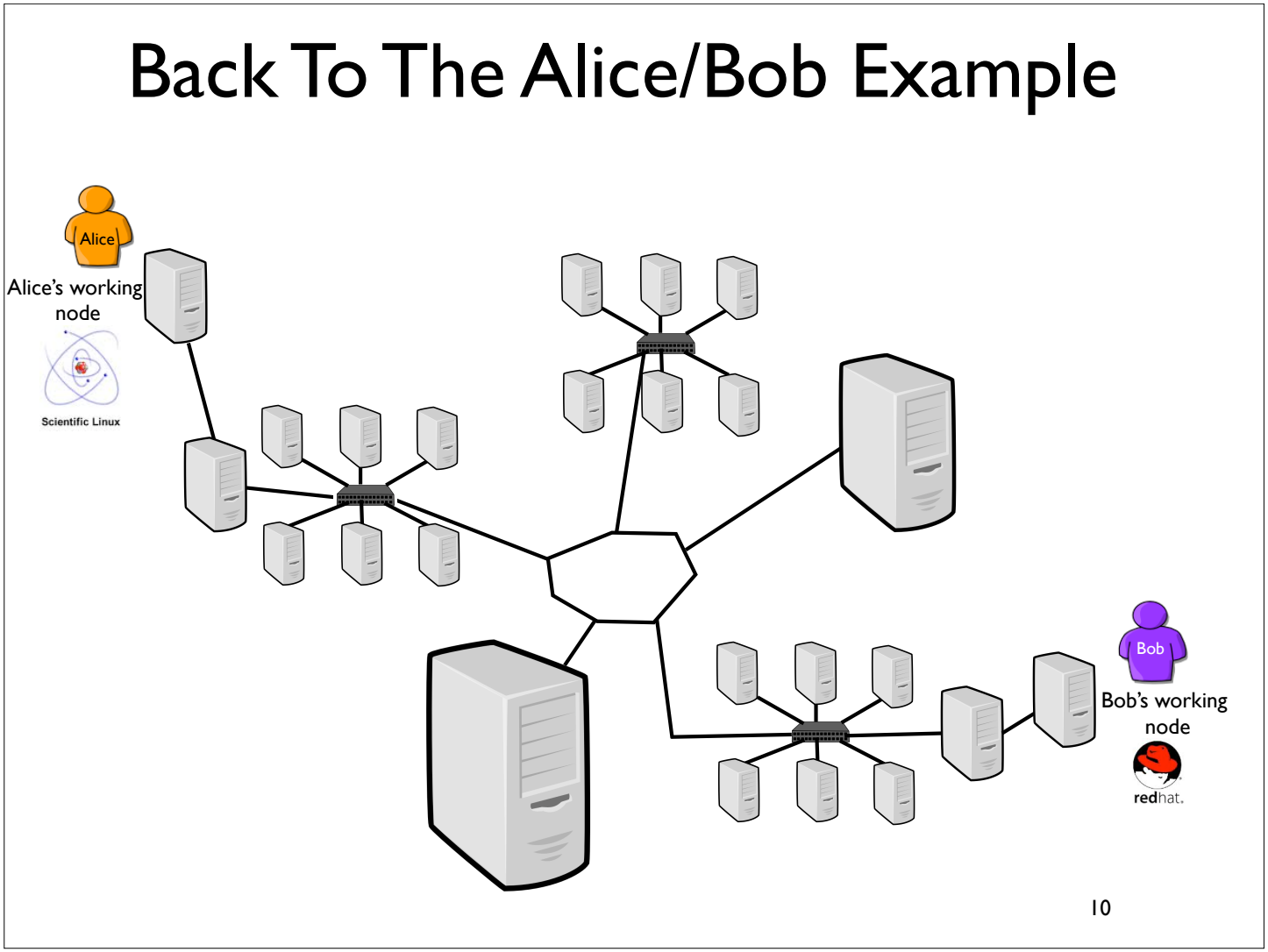
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)



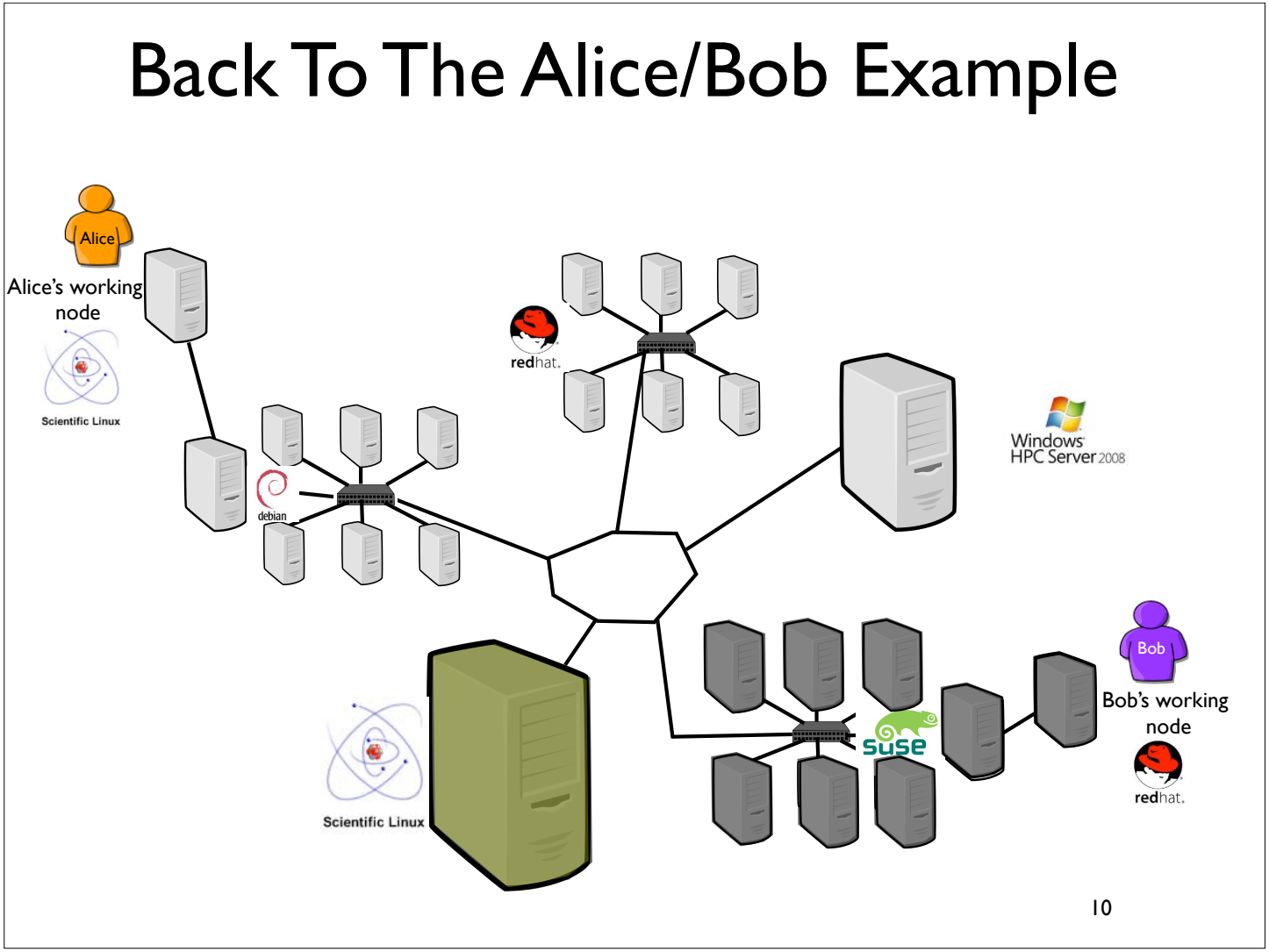
- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)

9

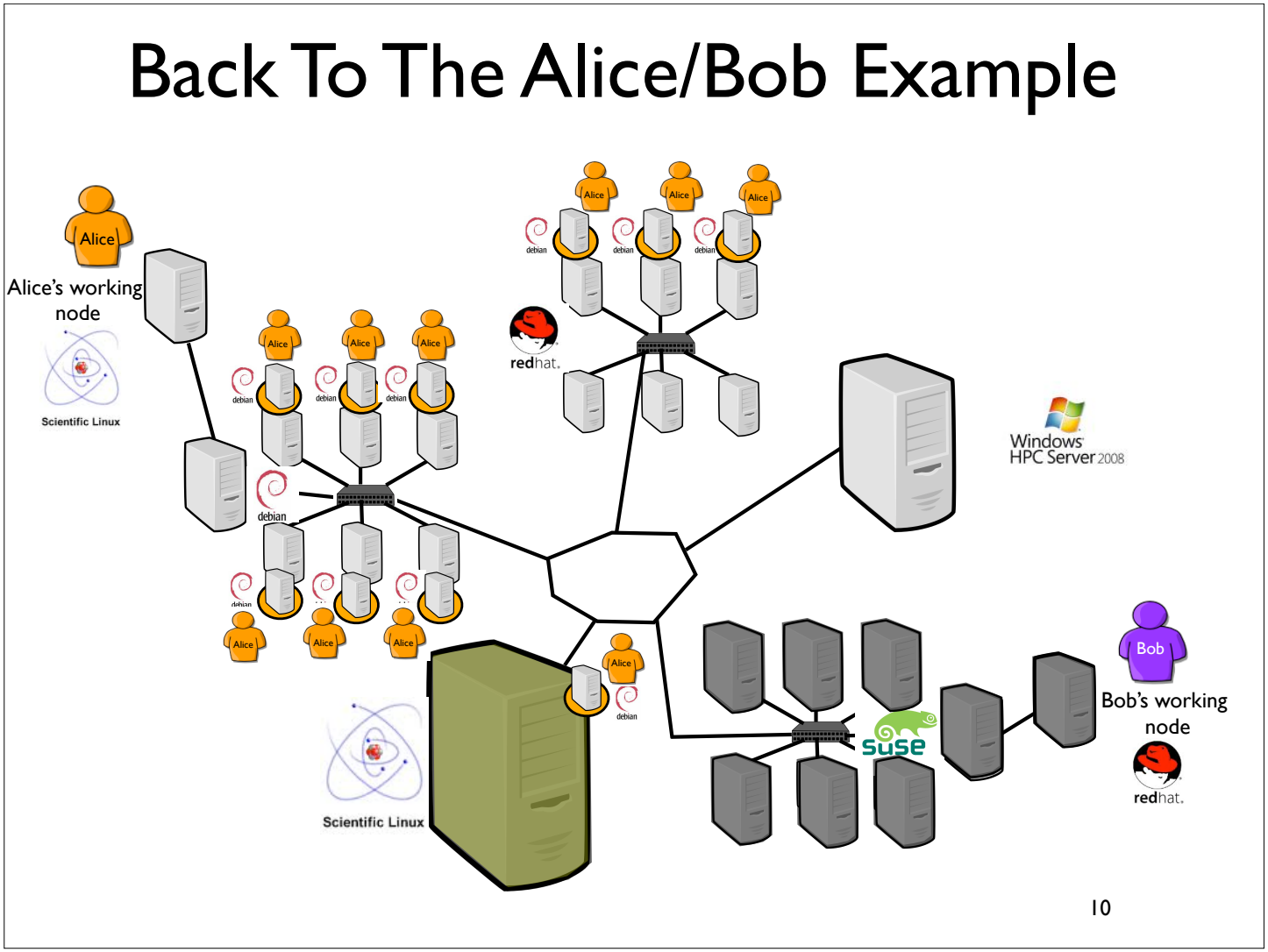
Courtesy of Adrien Lèbre



Courtesy of Adrien Lèbre

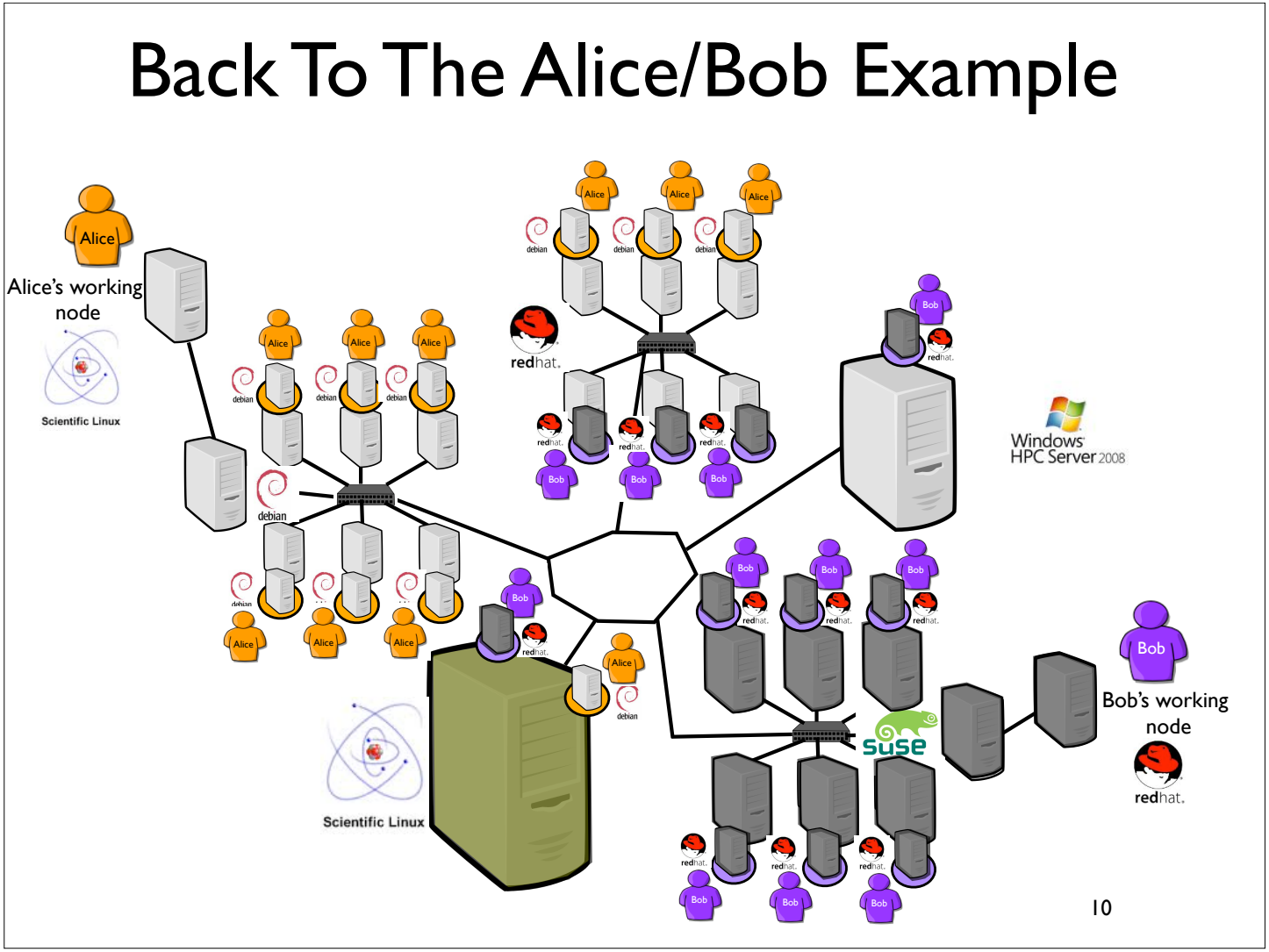


Courtesy of Adrien Lèbre



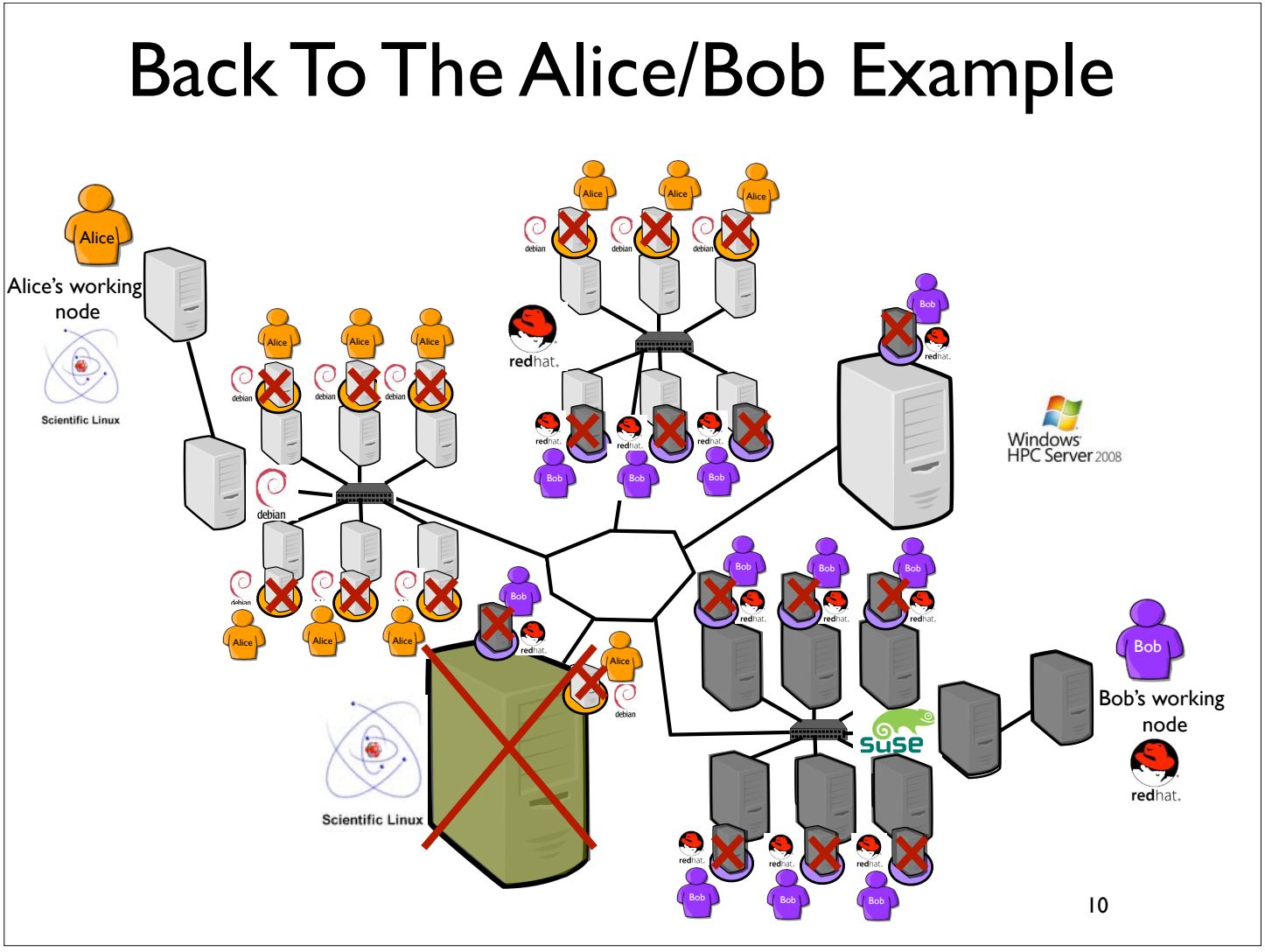
Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



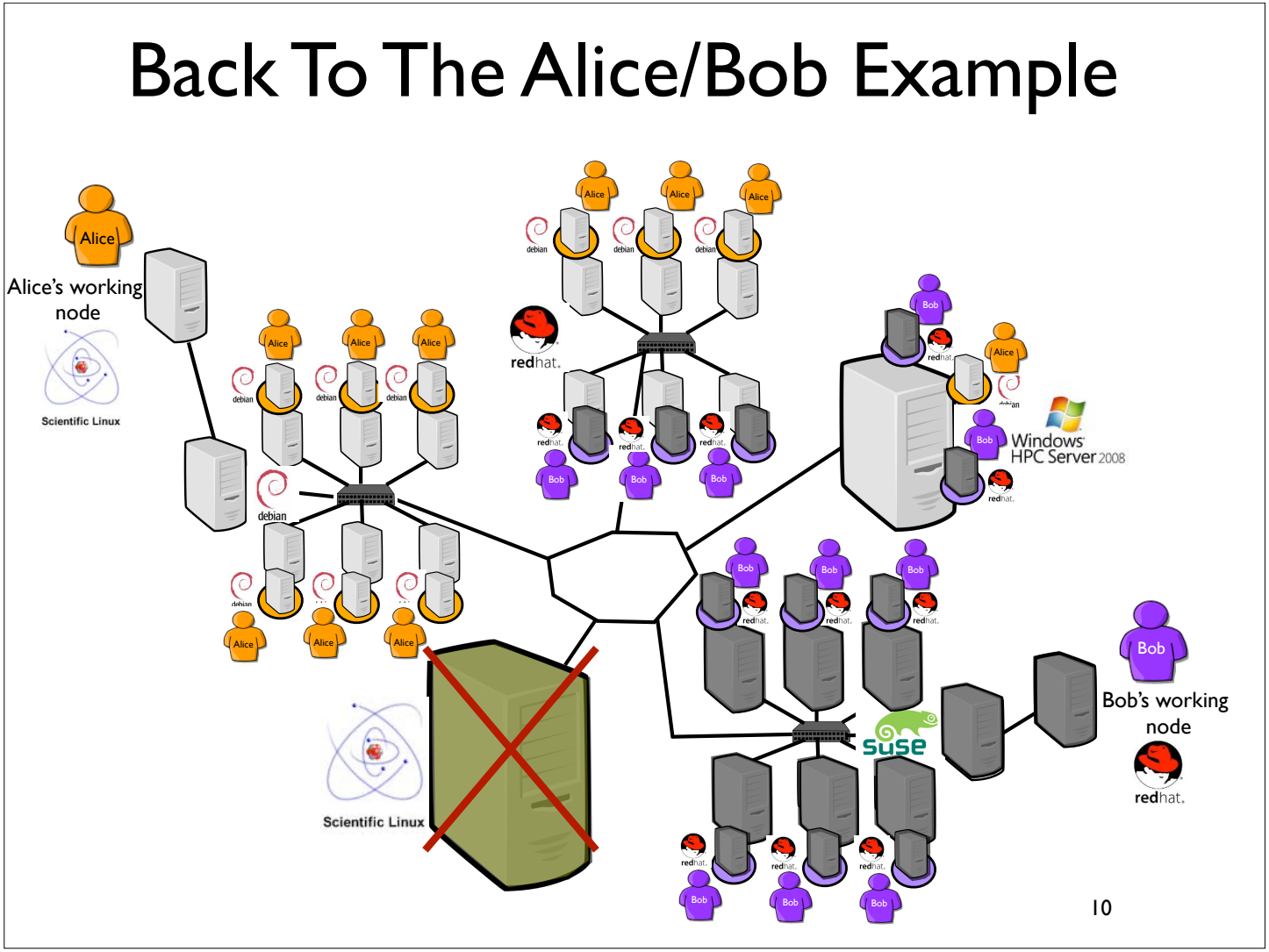
Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



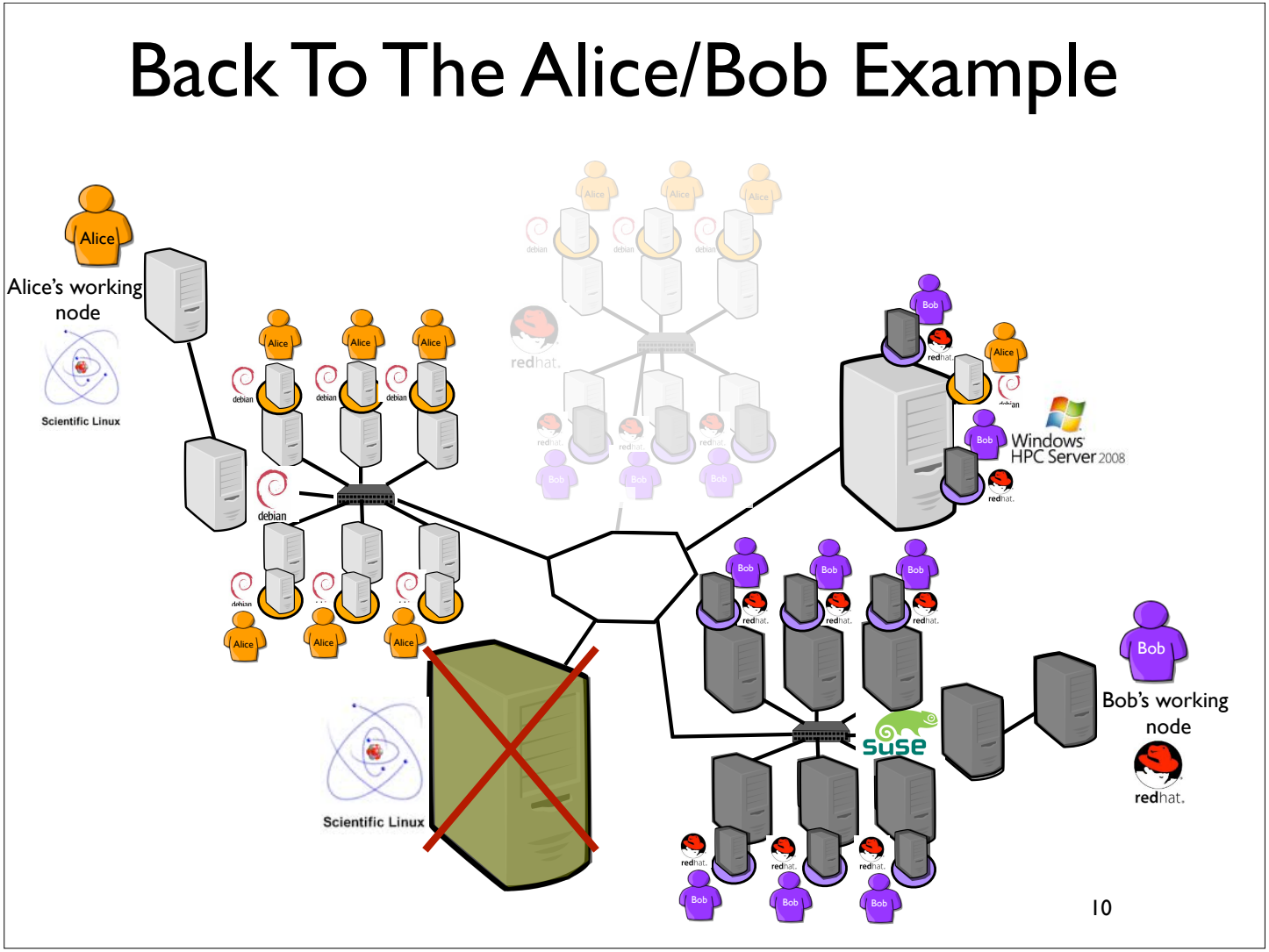
Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



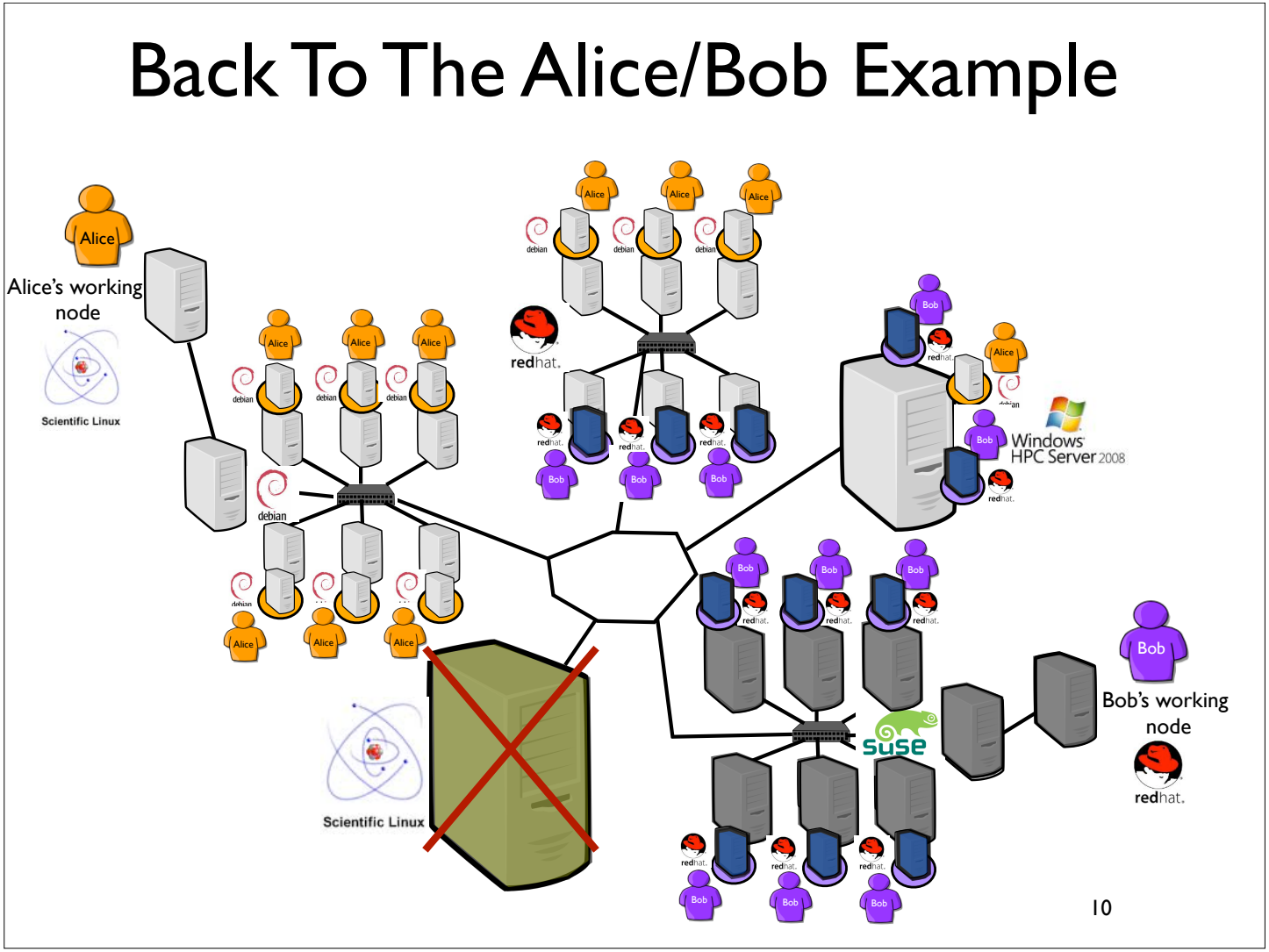
Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



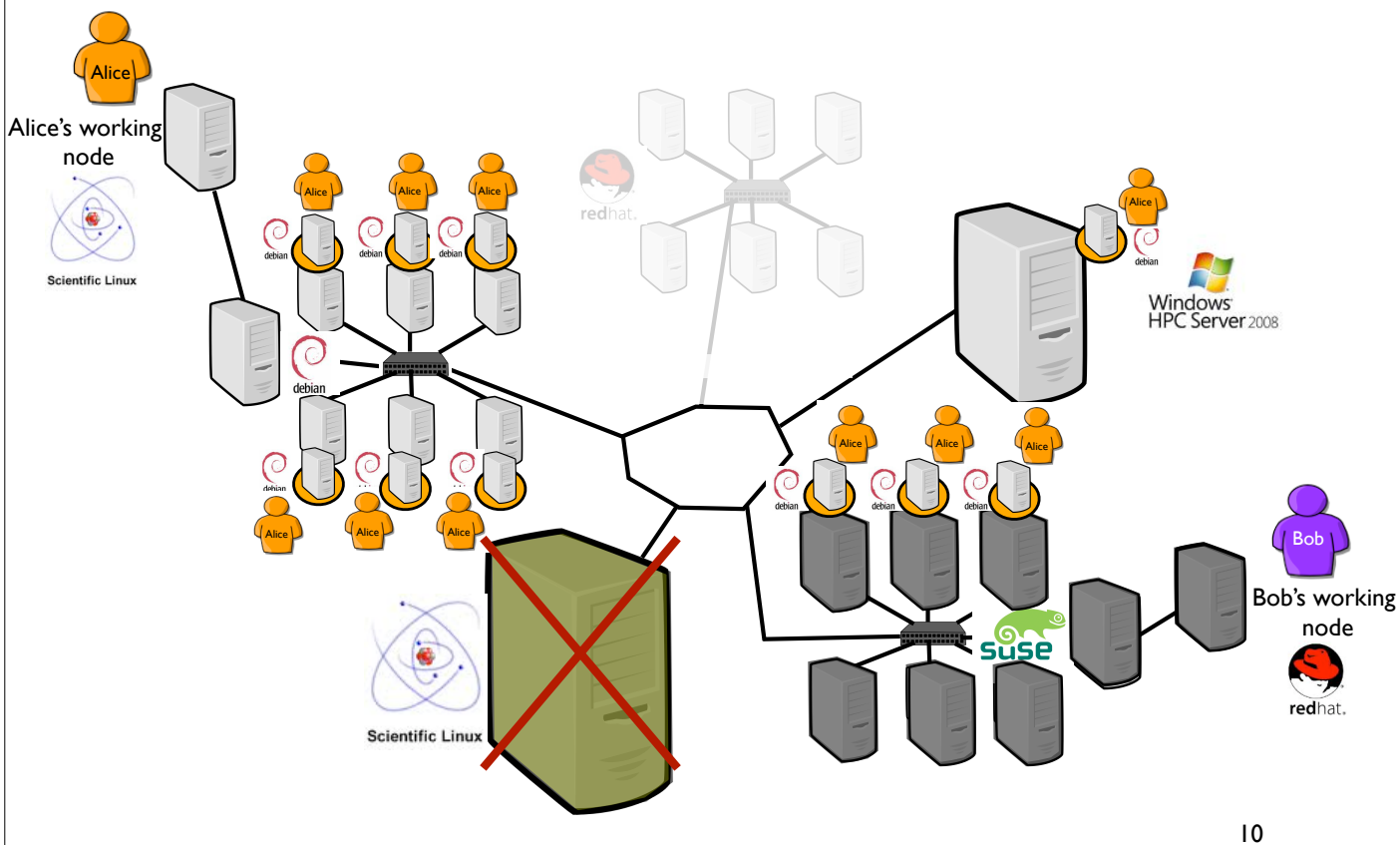
Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



Courtesy of Adrien Lèbre

Back To The Alice/Bob Example



Courtesy of Adrien Lèbre

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"

11

Courtesy of Adrien Lèbre

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, A Three Point Checklist

1961, Prof. John McCarthy

11

Courtesy of Adrien Lèbre

Focus on dynamical scheduling concerns

What can be done thanks to VM capabilities

Courtesy of Adrien Lèbre

Context

Job scheduling strategies for clusters/grids:
static allocation of resources / "user-intrusive"

Based on user estimates (time/resources)
For a bounded amount of time
(e.g. 4 nodes for 2 hours)

Resources are reassigned at the end of the slot without considering real needs of applications
(in the worst case, running applications can be simply withdrawn from resources, i.e. G5K best effort mode)

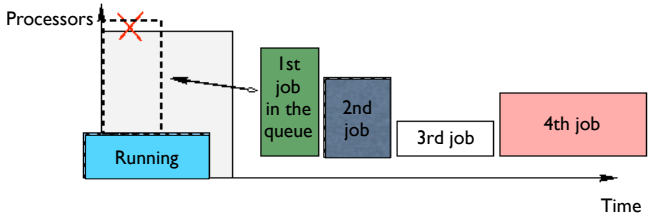
$\Rightarrow$ Coarse-grain exploitation of the architecture

13

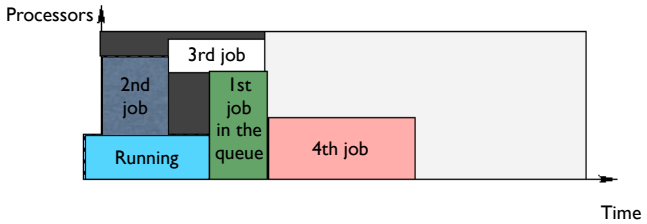
Courtesy of Adrien Lèbre

Context

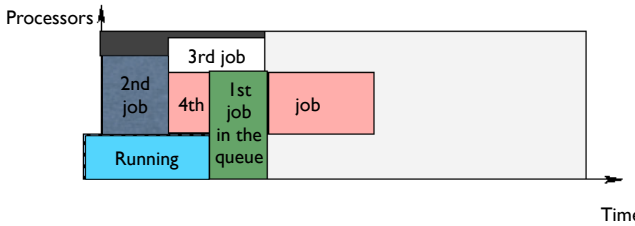
- Batch scheduler policies: closed to FCFS



Jobs arrive in the queue
and have to be scheduled.



FCFS + Easy backfilling
Jobs 2 and 3 have been backfilled.
Some resources are unused (dark areas)



Easy backfilling with preemption
The 4th job can be started without
impacting the first one.
A small piece of resources is still unused.

⇒ consolidation and preemption to finely exploit distributed resources

14

Courtesy of Adrien Lèbre

Consolidation and Preemption

- Few schedulers include preemption mechanisms based on checkpointing solutions:
  -  Strongly middleware/OS dependent
  -  Still not consider application resource changes
- SSI approaches include both consolidation and preemption of processes:
  -  Strongly middleware/OS dependent
  -  SSI developments are tedious (most of them have been given up)
- Exploit all VM capabilities (start/stop - suspend/resume - migrate)

15

Courtesy of Adrien Lèbre

Consolidation and Preemption

- The Entropy proposal

  F. Hermenier, Ph.D. in CS (University of Nantes / 2009)
  Use of Live migration capability to finely exploit cluster resources [Hermenier et al. 09]

  Generalization: the Cluster-Wide Context Switch concept [Hermenier et al. 10]

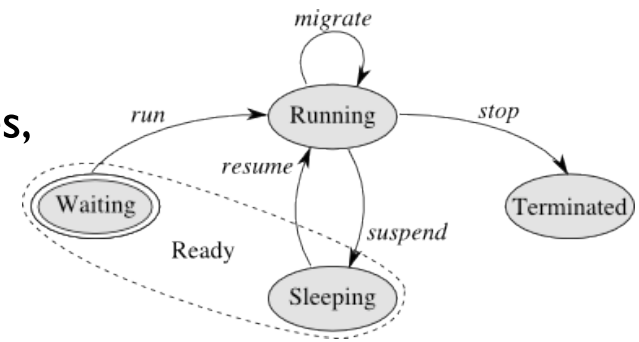
- Use case - energy concerns in Datacenters

16

Courtesy of Adrien Lèbre

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:



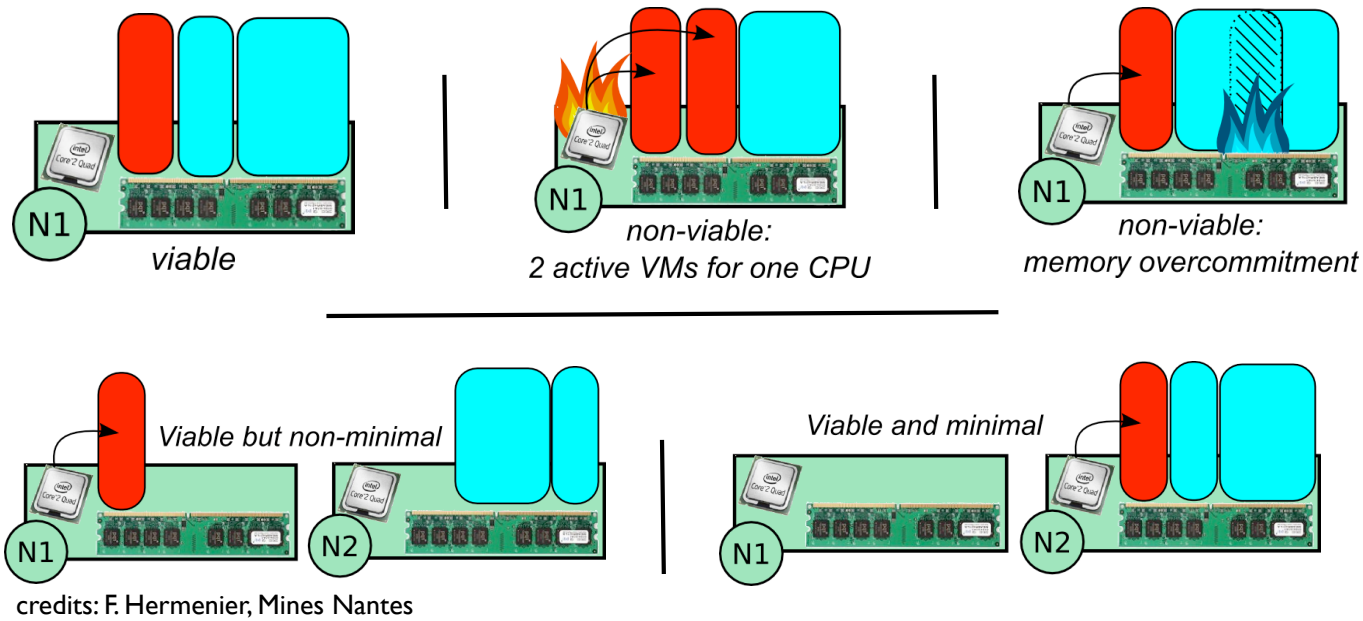
- A cluster-wide context switch (a set of VM context switches) enables to efficiently rebalance the cluster according to the: scheduler objectives / available resources / waiting vjobs queue

17

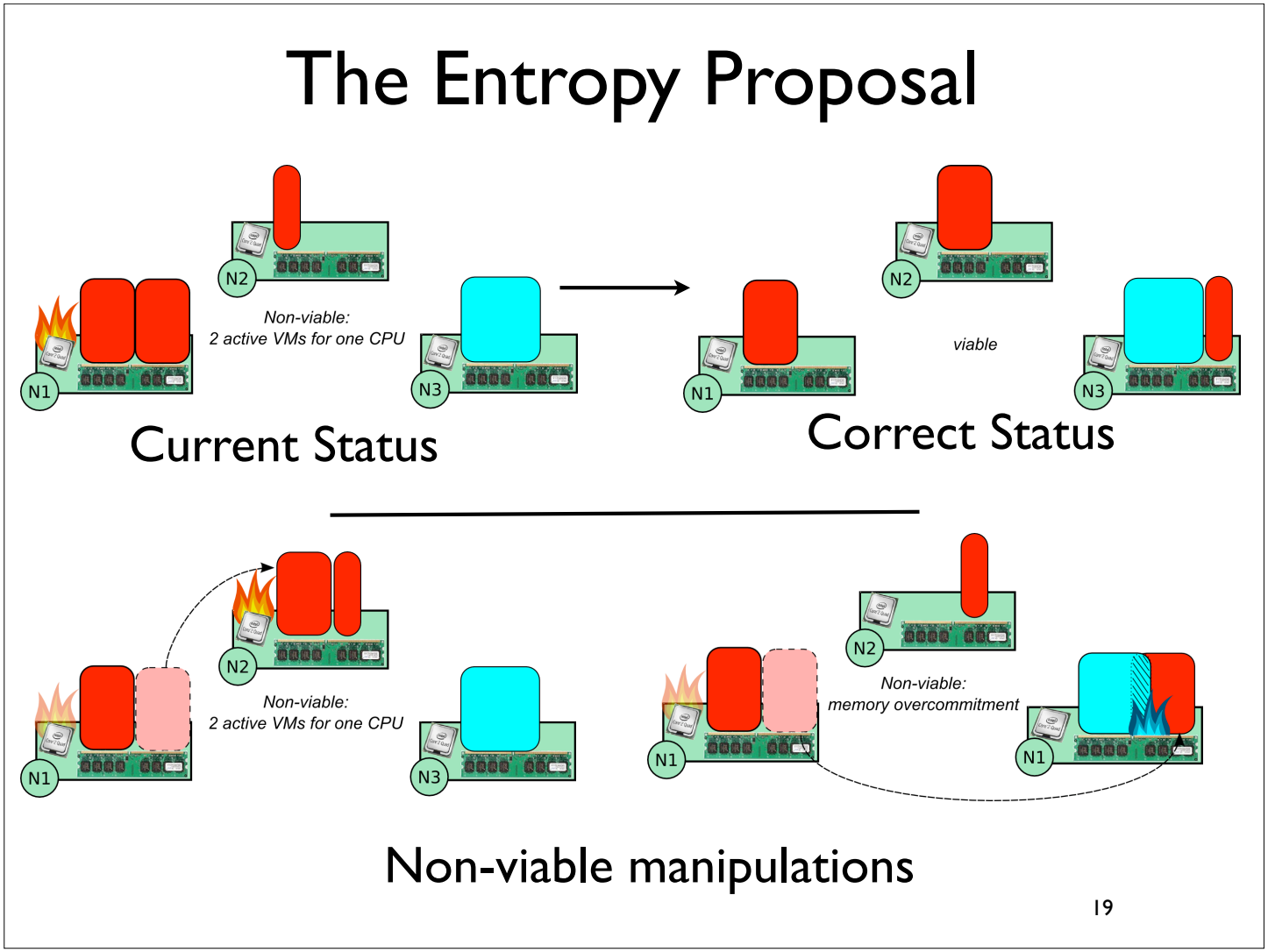
Courtesy of Adrien Lèbre

The Entropy Proposal

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



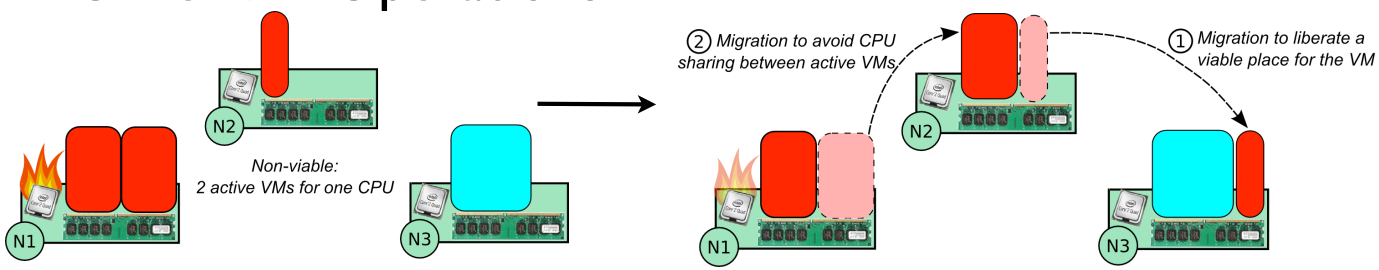
Courtesy of Adrien Lèbre



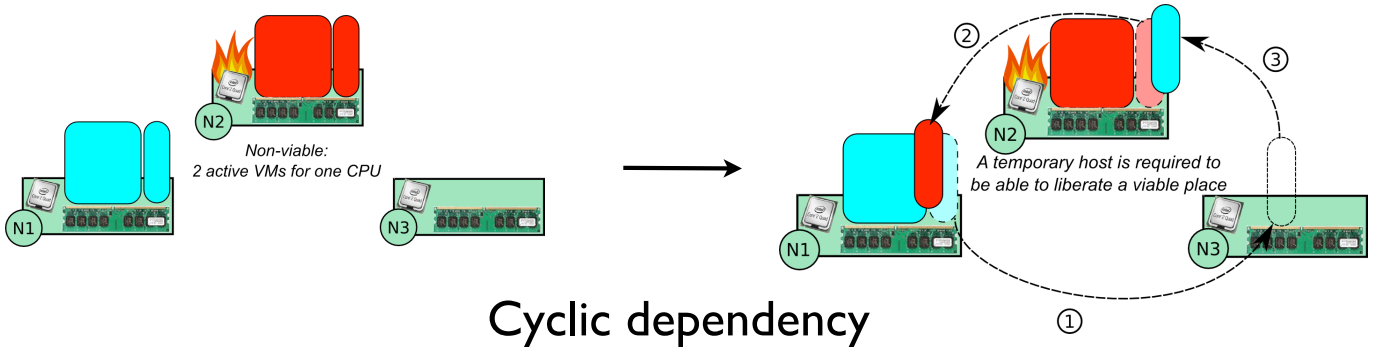
Courtesy of Adrien Lèbre

The Entropy Proposal

- Order VM Operations



Sequential dependency



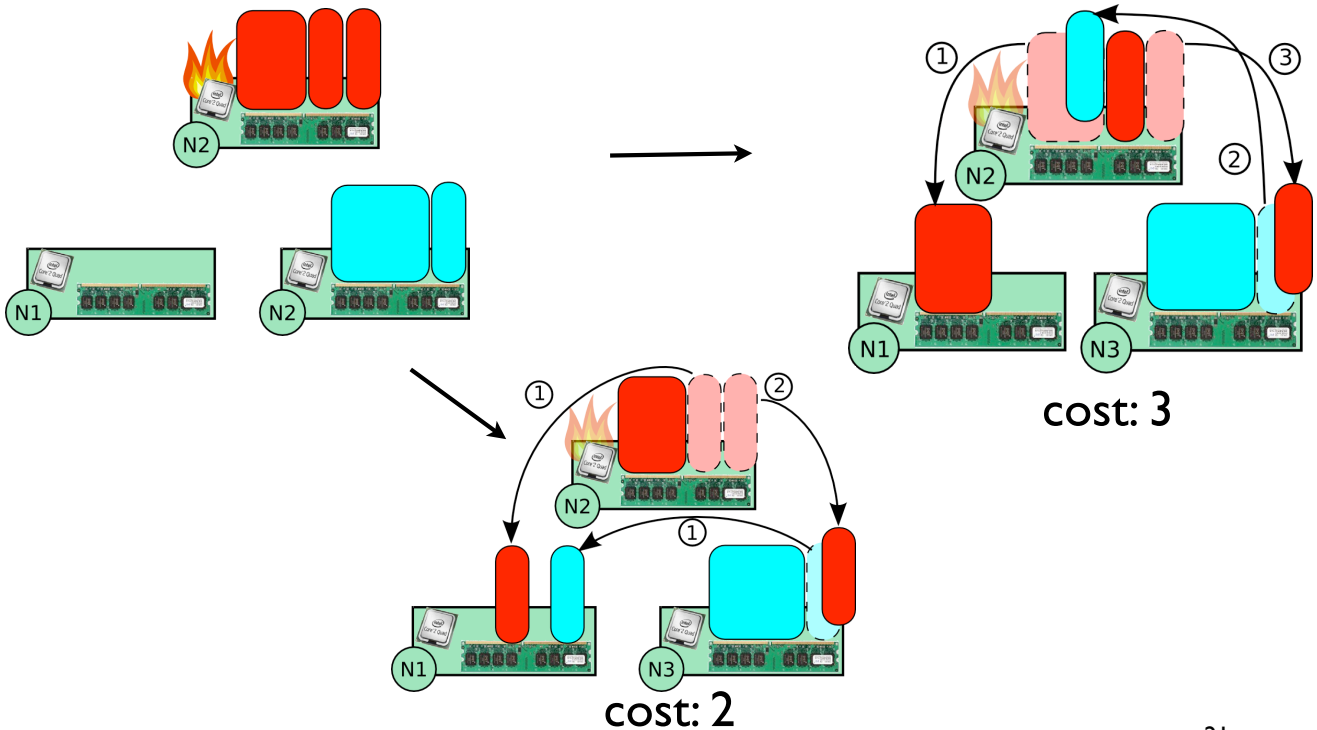
Cyclic dependency

20

Courtesy of Adrien Lèbre

The Entropy Proposal

- Optimizing the reconfiguration process

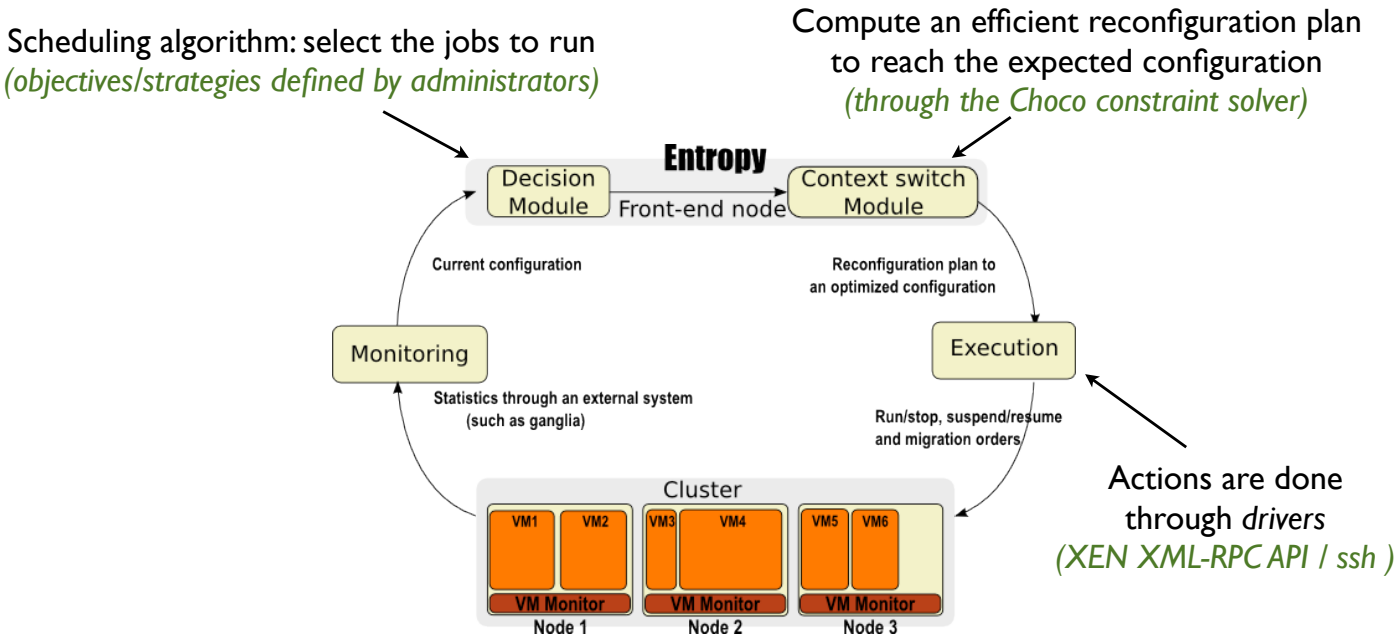


21

Courtesy of Adrien Lèbre

The Entropy Proposal

- The big picture: an autonomic model



- http://entropy.gforge.inria.fr, irc.freenode.net #entropy,

22

Courtesy of Adrien Lèbre

The Entropy Proposal

- To sum up

An autonomic framework to make the implementation of vjobs scheduling policies easier

Strength: composition of constraints
Developed since 2006 (ANR SelfXL / MyCloud, ANR Emergence, 10 persons)

"Prix de la croissance verte numérique" in 2009

Scalability of both computation and execution of the reconfiguration plan

Work in progress

Performance/scalability/...

23

Courtesy of Adrien Lèbre

Is consolidation really painless?

From
http://alan.blog-city.com/has_amazon_ec2_become_over_subscribed.htm

- Amazon in the early days was fantastic.

  Instances started up within a couple of minutes, they rarely had any problems and even their SMALL INSTANCE was strong enough to power even the moderately used MySQL database. For a good 20 months, all was well in the Amazon world, with really no need for concern or complaint.

Is consolidation really painless?

From http://alan.blog-city.com/has_amazon_ec2_become_over_subscribed.htm

- Amazon in the early days was fantastic.
- Neighborhood isn't what it use to be

  Noisy Neighbors: *A quick termination and a new spin up would usually, through the laws of randomness, have us in a quiet neighborhood where we could do what we needed.*

  As time went on, and our load increased, the real usefulness of the SMALL instances, soon disappeared with us pretty much writing off any real production use of them. This is a shame, as many of our web servers are not CPU intensive, just I/O instensive.

  Moving up to the "High-CPU Medium Instance" as our base image has given us some of that early-pioneer feeling that we are indeed getting the intended throughput that we expect from an instance.

Is consolidation really painless?

From http://alan.blog-city.com/has_amazon_ec2_become_over_subscribed.htm

- Amazon in the early days was fantastic.
- Neighborhood isn't what it use to be
- The commute is such a drag

However, in the last month of two, we've even noticed that these "High-CPU Medium Instance" have been suffering a similar fate of the Small instances.

In normal circumstances, a ping between two internal nodes within Amazon is around the 0.3ms level, with the odd ping reporting a whopping 7ms ever 30 or so packets.

When our instances appear to be dying or at least shaky, then this network latency jumps up to a whopping 7241ms.

Under extreme load, the virtual operating system is not able to process the network queue.

Is consolidation really painless?

From
http://alan.blog-city.com/has_amazon_ec2_become_over_subscribed.htm

- Amazon in the early days was fantastic.
- Neighborhood isn't what it use to be
- The commute is such a drag
- Different road surfaces

In one particular "fire fighting mode", we spent an hour literally spinning up new instances and terminating them until we found ourselves on a node that actually responded to our network traffic.
Not all the Amazon instances are equal in terms of the underlying hardware, and depending on which processor you get allocated can make a huge difference to the performance of your running instance.
So not only should we check for the CPU we are running on, we now must also take note of the network performance before we can safely push an instance into production.

This is not what cloud computing is all about.

Recap

- Virtualization changed the grid perspective because it solved many of heterogeneity, isolation and fault tolerance issues.
- Virtualization changed classical batch scheduling issue because preemption helps.
- Now, focus on consolidation and energy minimization.

Recap

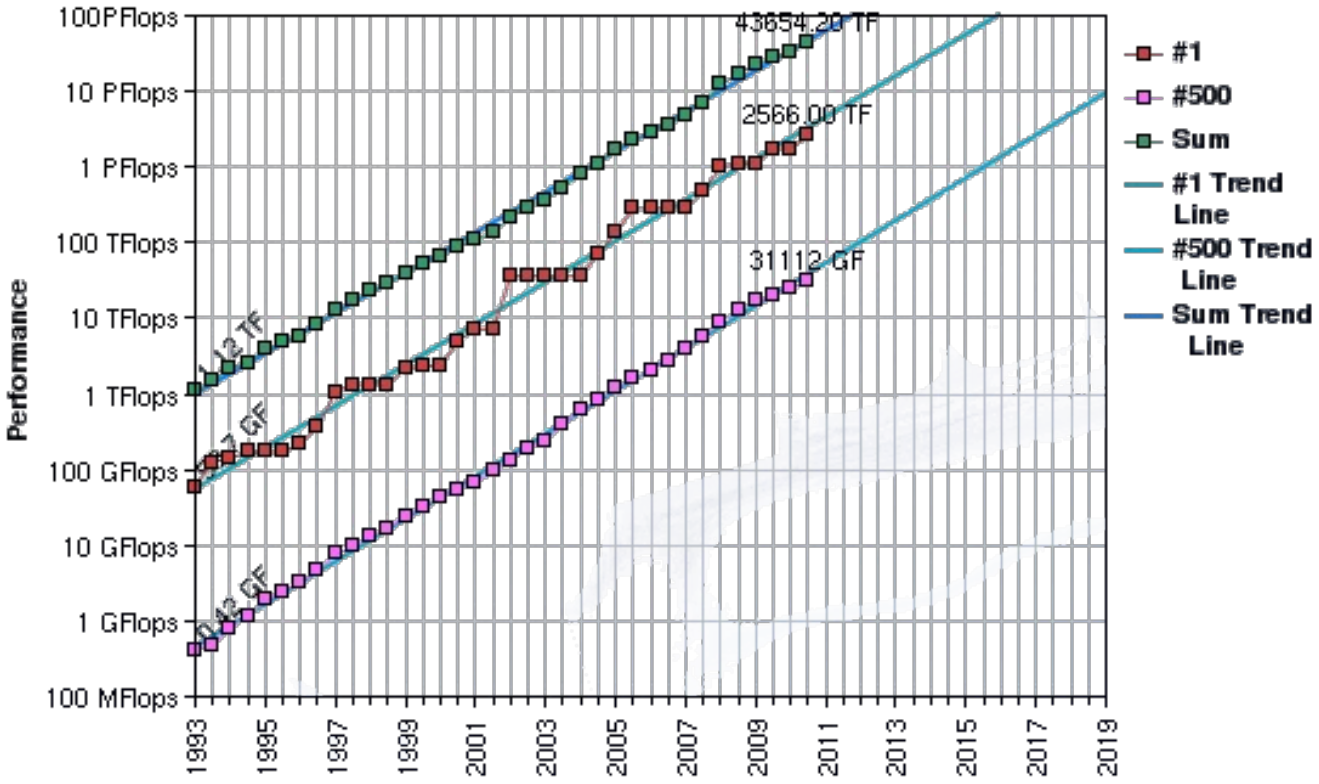
- Virtualization changed the grid perspective because it solved many of heterogeneity, isolation and fault tolerance issues.
- Virtualization changed classical batch scheduling issue because preemption helps.
- Now, focus on consolidation and energy minimization.
- Remember EC2 is now #42 on Top500:
  - Commodity Networks can Compete with IB, Myrinet, etc.
  - No power consumption was reported...
    - The worldwide demand for data center power in 2005 was equivalent to the output of about 17 1,000-megawatt power plants (1% of world electricity consumption in 2005).
    - Google continuously uses enough electricity to power 200,000 homes.
      The average energy consumption on the level of a typical user, is about 180 watt-hours a month (a 60-watt light bulb for three hours).
      http://www.nytimes.com/2011/09/09/technology/google-details-and-defends-its-use-of-electricity.html
    - Partly because of the 2008 recession, power consumption by data centers hasn't grown at expected rates.
      http://www.nytimes.com/2011/08/01/technology/data-centers-using-less-power-than-forecast-report-says.html?_r=1

Outline

1 Virtualization

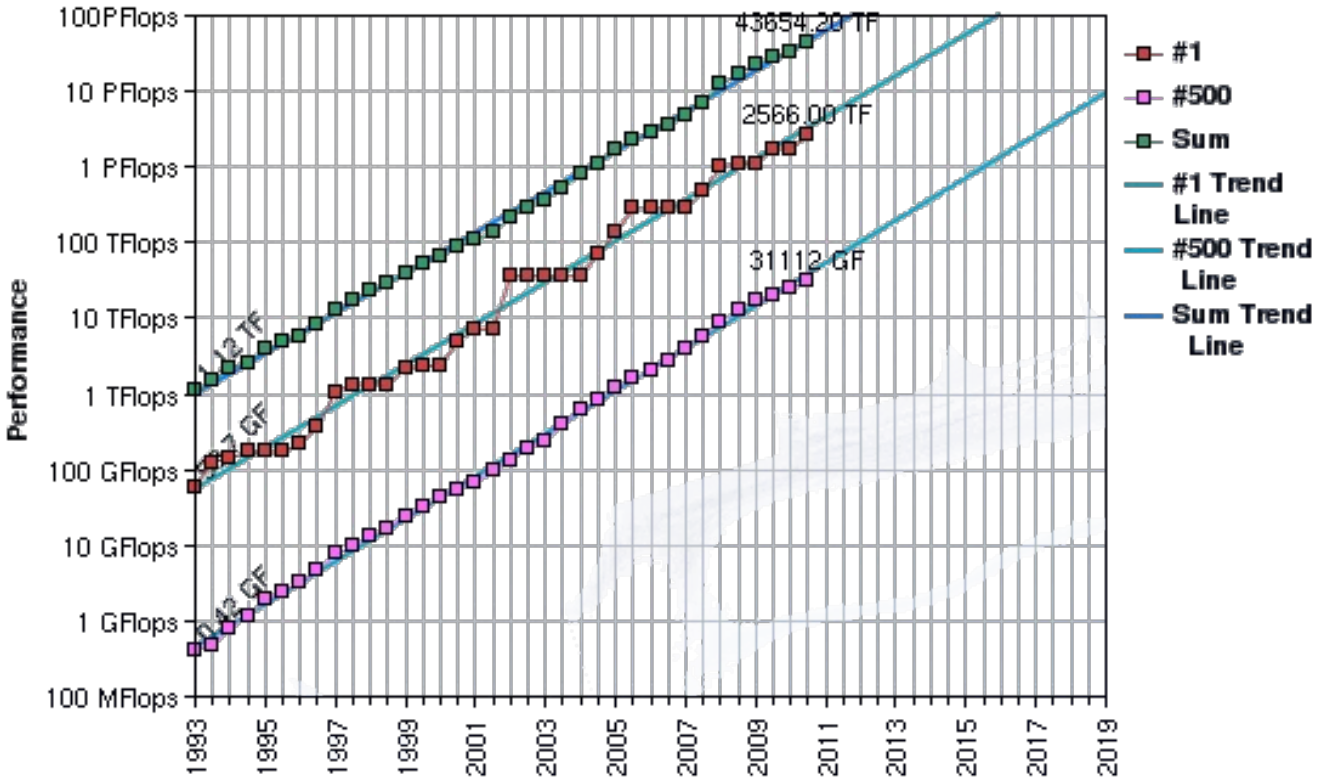
2 Toward Exascale

Toward Exascale: The Mont-Blanc proposal



- Exponential improvements at the rate of one order of magnitude every 3 years: One petaflops was achieved in 2008, one exaflops is expected in 2020.

Toward Exascale: The Mont-Blanc proposal



- Exponential improvements at the rate of one order of magnitude every 3 years: One petaflops was achieved in 2008, one exaflops is expected in 2020.
- Based on a 20 MW power budget, which is already very high, this requires an efficiency of 50 GFLOPS/Watt.

Toward Exascale: The Mont-Blanc proposal

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	2026.48	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
2	2026.48	IBM Thomas J. Watson Research Center	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	85.12
3	1996.09	IBM - Rochester	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	170.25
4	1988.56	DOE/NNSA/LLNL	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	340.50
5	1689.86	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype 1	38.67
6	1378.32	Nagasaki University	DEGIMA Cluster, Intel i5, ATI Radeon GPU, Infiniband QDR	47.05
7	1266.26	Barcelona Supercomputing Center	Bullx B505, Xeon E5649 6C 2.53GHz, Infiniband QDR, NVIDIA 2090	81.50
8	1010.11	TGCC / GENCI	Curie Hybrid Nodes - Bullx B505, Nvidia M2090, Xeon E5640 2.67 GHz, Infiniband QDR	108.80
9	963.70	Institute of Process Engineering, Chinese Academy of Sciences	Mole-8.5 Cluster, Xeon X5520 4C 2.27 GHz, Infiniband QDR, NVIDIA 2050	515.20
10	958.35	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	1243.80

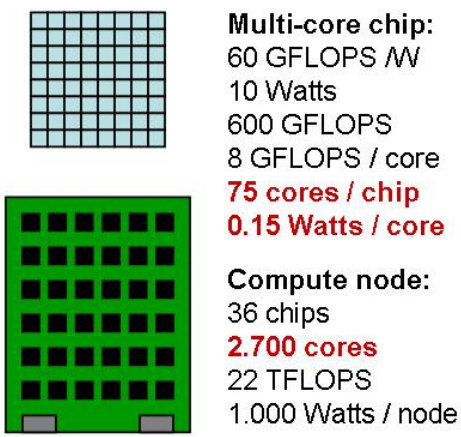
- Exponential improvements at the rate of one order of magnitude every 3 years: One petaflops was achieved in 2008, one exaflops is expected in 2020.
- Based on a 20 MW power budget, which is already very high, this requires an efficiency of 50 GFLOPS/Watt.
- However, the current leader in energy efficiency (IBM BlueGene/Q) achieves only 2.0 GFLOPS / Watt. Thus, a 25$\times$ improvement is required.

Where does the power go?

- In current systems the processors consume a lion's share of the energy approximately 43% or more.
- The remaining energy is used to power up the memories, the interconnection network, and the storage system.
- Furthermore, a significant fraction is wasted in power supply overheads, and in thermal dissipation (cooling), which do not contribute to performance at all.

Possible Architecture: 200 PFlops with 10MWatt

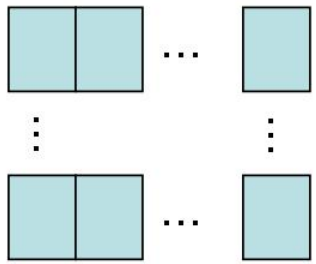
- On a power envelope of 10 Watts, this implies that each multi-core chip must achieve 600 GFLOPS of peak performance.
- If we assume 8 GFLOPS processors (2 GHz, 4 operations per cycle), this requires 75 cores per chip, consuming 0.15 Watts / core.
- As a reference, the current dual-core ARM Cortex A9 consumes 1.9 Watts at 2 GHz and uses 6.7 mm22. The 800 Mhz version consumes only 0.5 Watts and 4.6 mm2. That is, 0.25 Watts per processor, quite close to the target 0.15 Watts required.
- We are much closer to the target in this direction, than using today's high-end processors.



Multi-core chip:
60 GFLOPS /W
10 Watts
600 GFLOPS
8 GFLOPS / core
75 cores / chip
0.15 Watts / core

Compute node:
36 chips
2.700 cores
22 TFLOPS
1.000 Watts / node

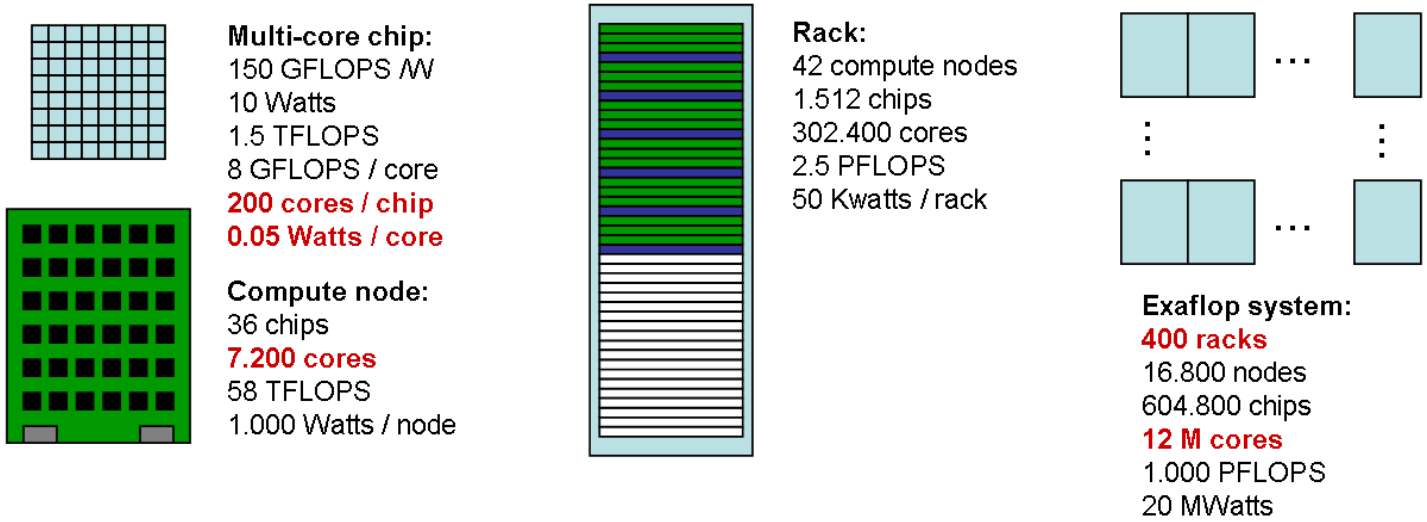
Rack:
42 compute nodes
1.512 chips
86.400 cores
0.9 PFLOPS
50 Kwatts / rack



Exascale system:
225 racks
16.800 nodes
604.800 chips
4.5 M cores
200 PFLOPS
10 MWatts

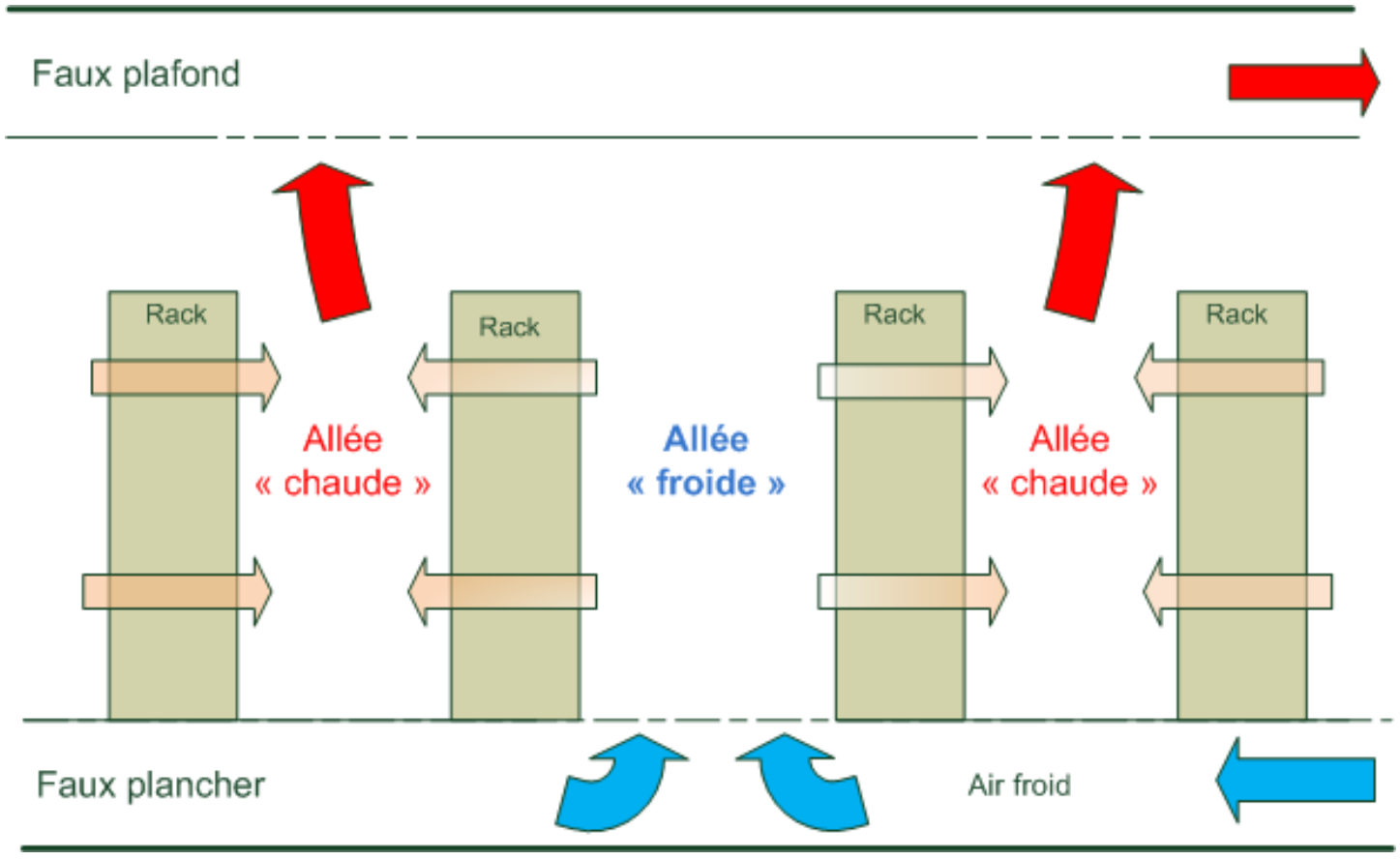
Projection for Exascale

- 1000 Pflops / 20MWatt = 10GFlops / Watt $\rightsquigarrow$ 200 8 Gflops core / chips and 0.05Watt/ core!!!

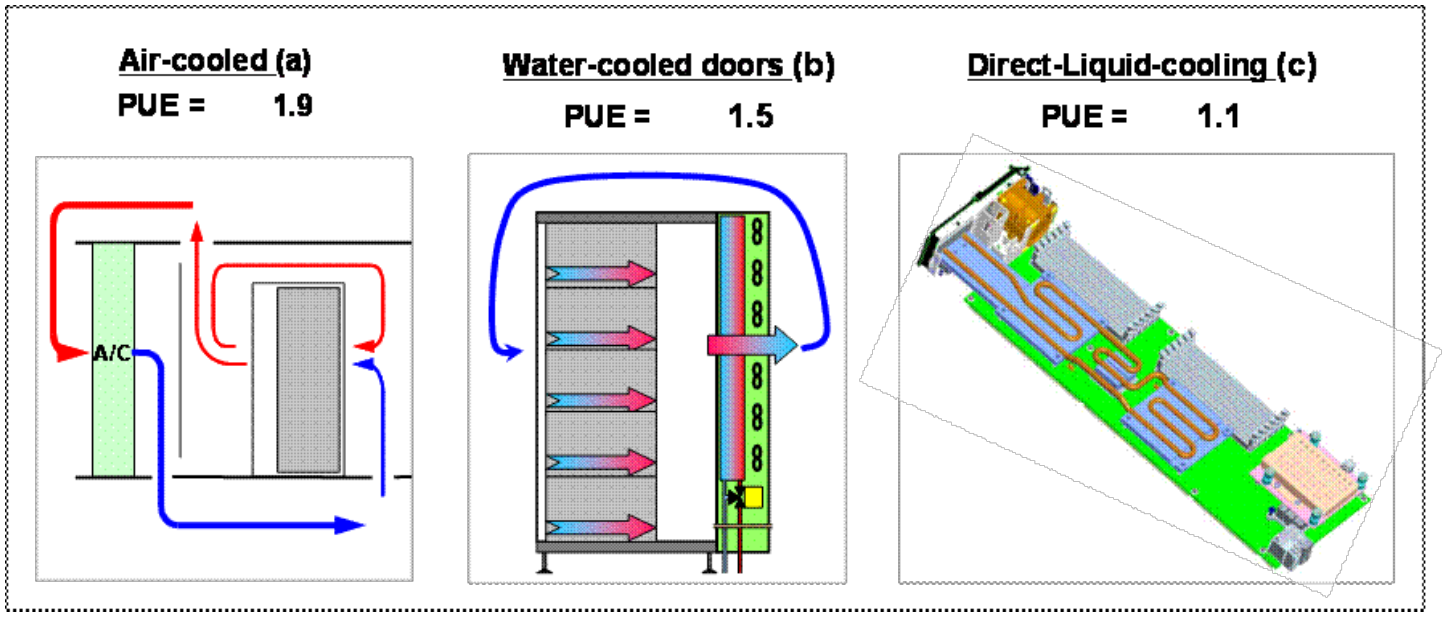


- Require new memory architecture, network, ...

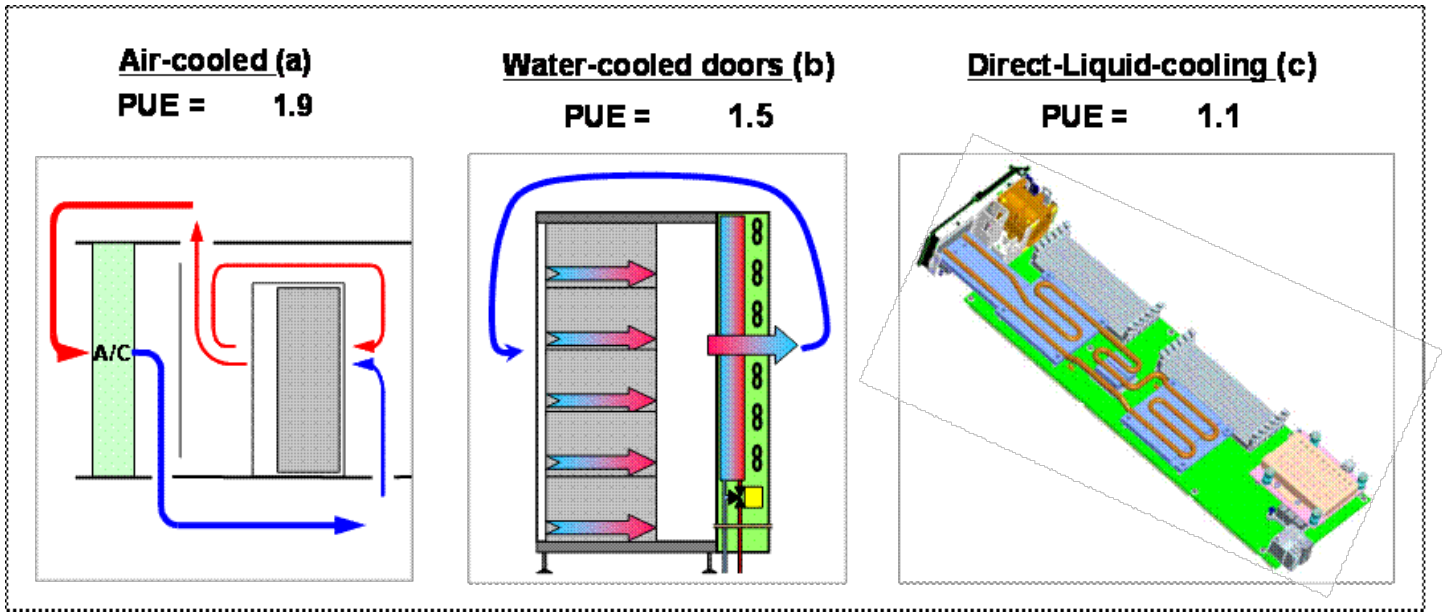
Thermal dissipation



Thermal dissipation



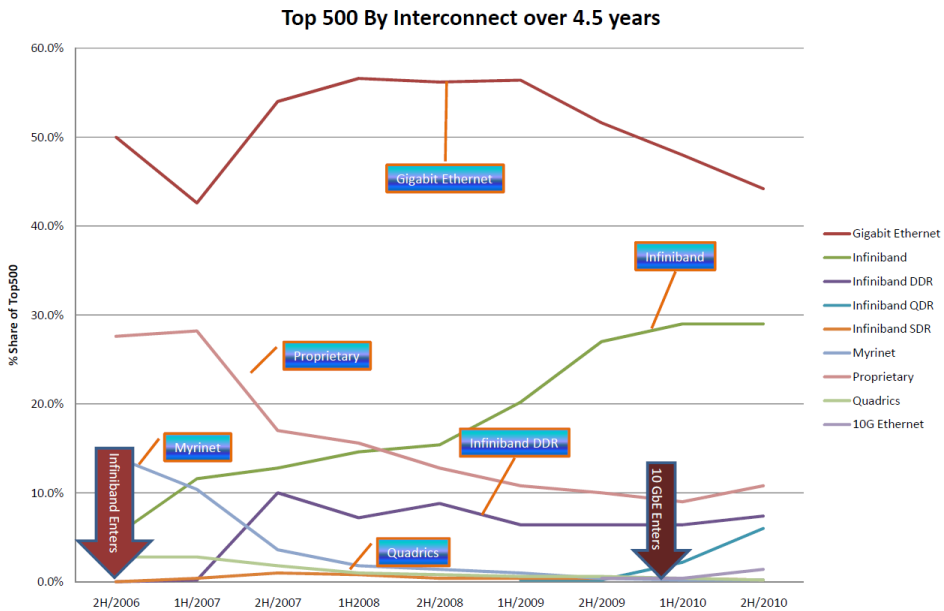
Thermal dissipation



The use of low-power embedded technologies will have significant implications on the thermal characteristics of the system, which will require re-evaluating these cooling methods, and maybe proposing new ones.

Interconnect

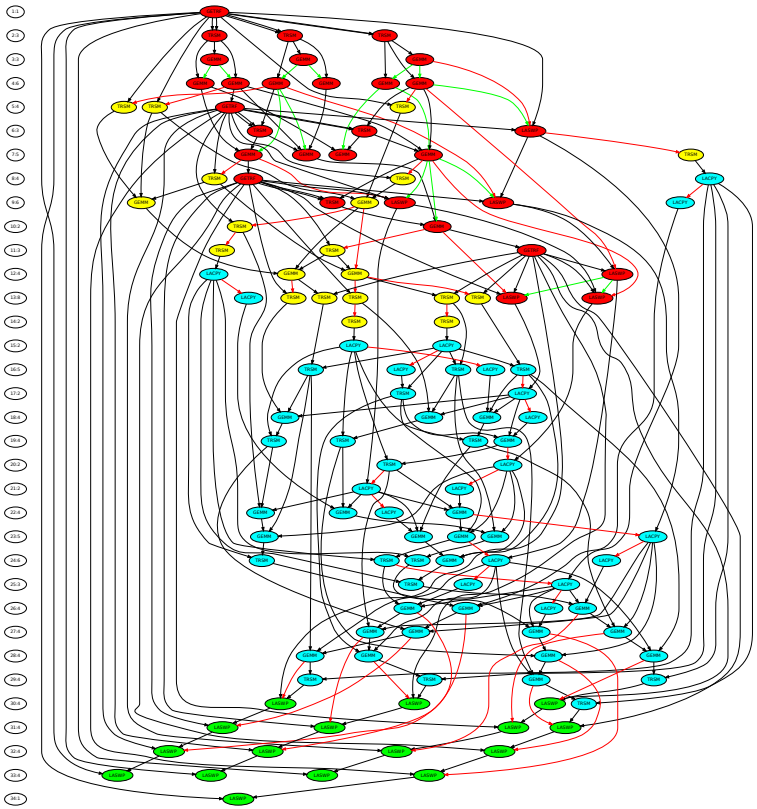
- Current HPC systems are characterized by either the large scale integration of low-power embedded devices, or clusters of commodity x86 servers (with increasing use of GPU acceleration).
- The interconnect for such systems are either based on proprietary technology or on widely available switch network technology such as Infiniband or Ethernet.

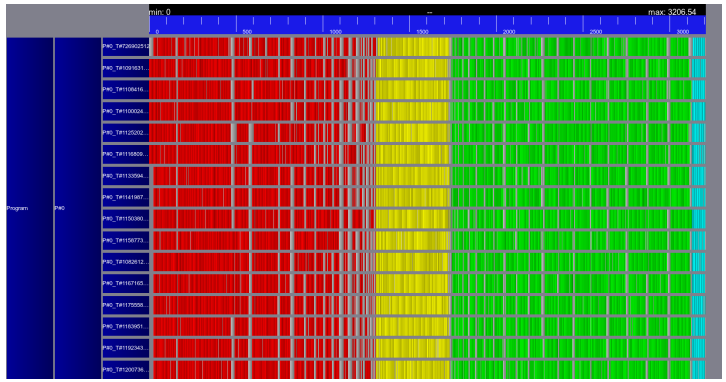


- For the majority of HPC cluster systems in the Top100, the network of choice is Infiniband primarily due to the performance and price.
- For the majority of systems within the Top500 the dominant interconnect is Ethernet.
- Ethernet is a standard low-power interface that can be used as the method of interconnection.
- Storage and Network Convergence

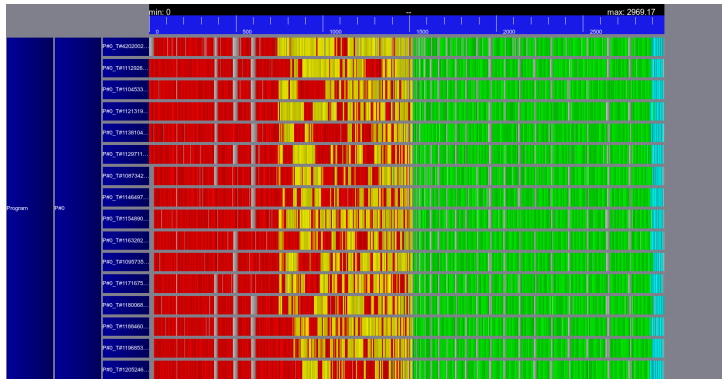
System software

Classical MPI programs have static load balancing and synchronizations. Hence, they exhibit load imbalance when scale increases.





(a) With synchronization between each stage.



(b) With interleaved stages.

Figure 5: Execution traces of the DGETRI routine with a 5000-by-5000 matrix and NB = 250 on a 16-cores architecture.

"High Performance Matrix Inversion Based on LU Factorization for Multicore Architectures". Dongarra, Faverge, Ltaief, Luszczek. 4th Workshop on Many-Task Computing on Grids and

Failure management.

As the size of new supercomputers scales to tens of thousands of sockets, the mean time between failures (MTBF) is decreasing to just several hours and long executions need some kind of fault tolerance method to survive failures$\rightsquigarrow$ a lot of attention on failure management.

A challenging problem

On August 8, 2011, NCSA announced that IBM had terminated its contract to provide hardware for Blue Waters, and would refund payments to date.

Cray Inc. has won a $188 million contract with the University of Illinois to build the supercomputer for the Blue Waters project.

The building was designed using complex fluid dynamic models to optimize the cooling system. Energy efficiency at the data center is estimated to be in the 85%-90% range, far superior to the 40% efficiency typically seen in large data centers.

Conclusion and Take-home message

Grid Computing

- Infrastructure for *computational science:* lot of sequential simulation jobs
- Main issues: compatibility, virtual organizations (trust and accountability mgmt)
- Performance is throughput (want the campaign done not only one element)

Cloud Computing

- Infrastructures underlying the corporate IT is getting bigger
- Main issue: keep up with the load, even when facing flash crowd effects
- Performance is manageability (cost effectiveness) + availability, response time

High Performance Computing and Exascale

- Have the world's biggest computer, to lead IT world's research
- Main issues: do the biggest possible parallel simulations [justify the investment]
- Performance is to simulate big enough and fast enough

High * Computing is converging

- With * being one of: Performance, Throughput, Cost-effective or Availability
- They were split at birth because of differences between Science and Business
- Good News: we will work more and more together in the future!