Hype and trends

Arnaud Legrand, CNRS, University of Grenoble

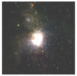
LIG laboratory, arnaud.legrand@imag.fr

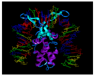
December 9, 2013

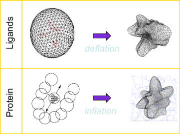
Outline

Some Particularly Challenging Computations

Science

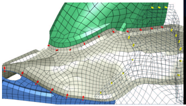




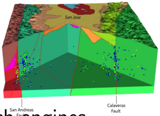


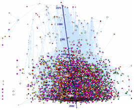
- Global climate modeling
- Astrophysical modeling
- Biology (genomics; protein folding; drug design)
- Computational Chemistry
- Computational Material Sciences and Nanosciences

Engineering



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



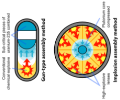


Business and Humanities

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

Defense

- Nuclear weapons – tested by simulations
- Cryptography



Courtesy of Martin Quinson (2011)

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

$\leadsto$ 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!)

Courtesy of Martin Quinson (2011)

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

Courtesy of Martin Quinson (2011)

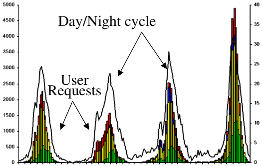
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\leadsto$ 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 $\leadsto$ High **Availability** Computing
- But load is very changing



Courtesy of Martin Quinson (2011)

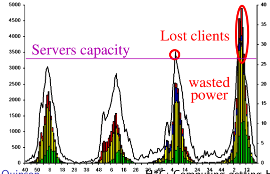
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage ⇝ 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 ⇝ High **Availability** Computing
- But load is very changing ⇝ Servers dimensioned for flash crowds



Courtesy of Martin Quinson (2011)

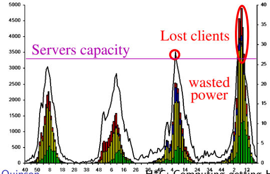
Why would Business need Computers

Initially, no need for performance

- Business computations seldom extend beyond ordinary rational arithmetic (unless when science is involved in business)
- Many desktop usage $\leadsto$ 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 $\leadsto$ High **Availability** Computing
- But load is very changing $\leadsto$ 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 (2011)

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: ≈ 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

⇒ Corporate Data Centers become as big as Scientific Supercomputers!

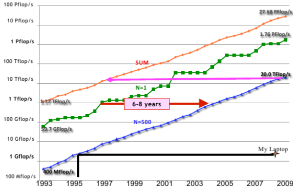
- . . . and share the same difficulties. The twins are technically reconciled ☺

Courtesy of Martin Quinson (2011)

How big are these machines?

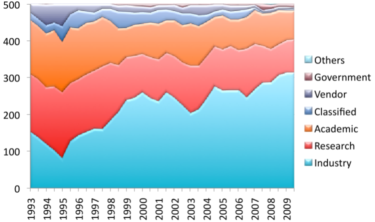
There is an International Ranking

- 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 (2011)

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

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
- This configuration has not been re-evaluated since and is thus now ranked 165

From http://perspectives.mvdirona.com

Outline

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 (2010)

How Virtualization Changed The Grid Perspective

Courtesy of Adrien Lèbre (2010)

xxx Computing

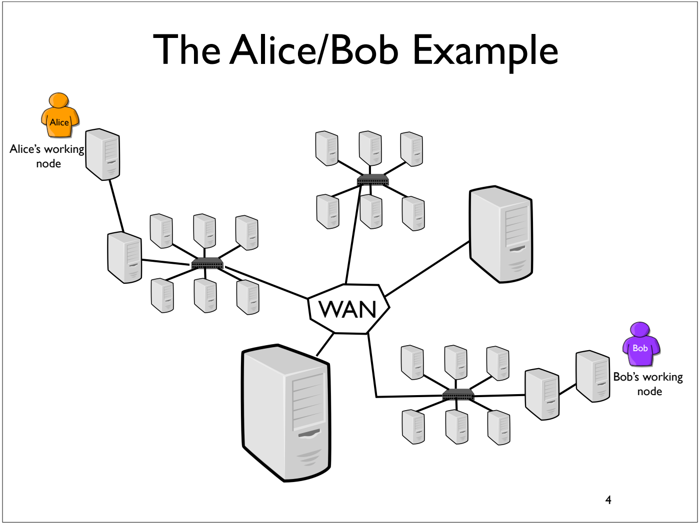
- xxx as Distributed
  (Cluster / Grid / Desktop / "Hive" / Cloud / Sky / ...)

- A common objective

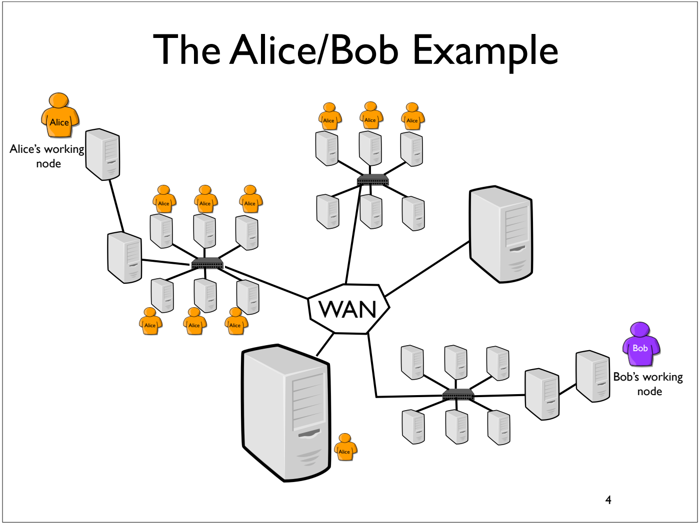
  provide computing resources (both hardware and software) in a flexible, transparent, secure, ... way

3

Courtesy of Adrien Lèbre (2010)

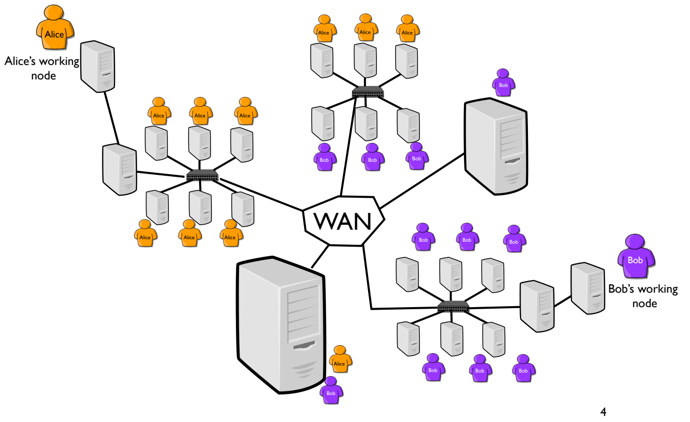


Courtesy of Adrien Lèbre (2010)

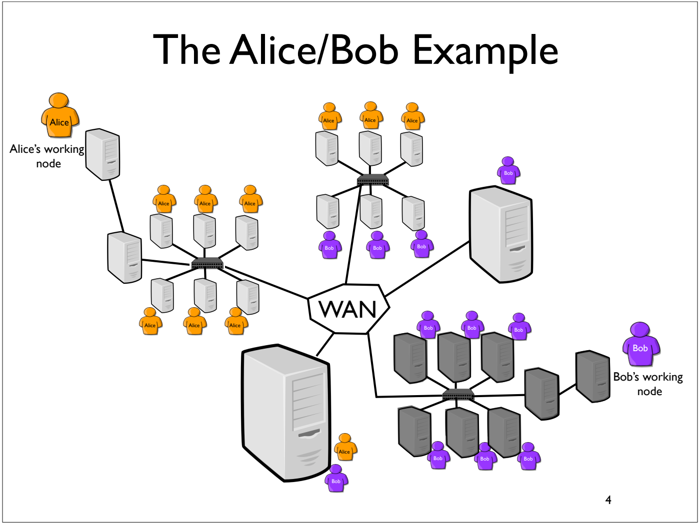


Courtesy of Adrien Lèbre (2010)

The Alice/Bob Example

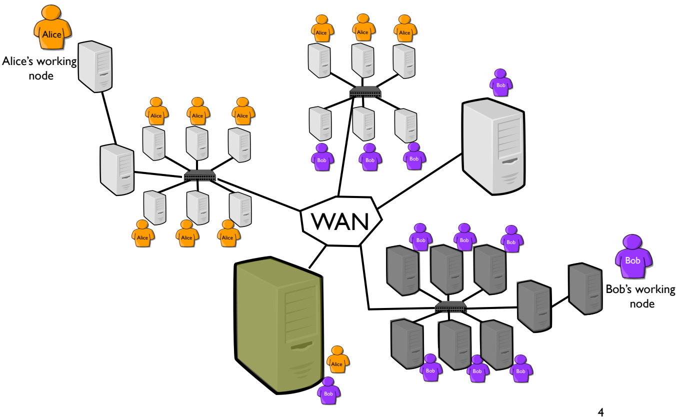


Courtesy of Adrien Lèbre (2010)

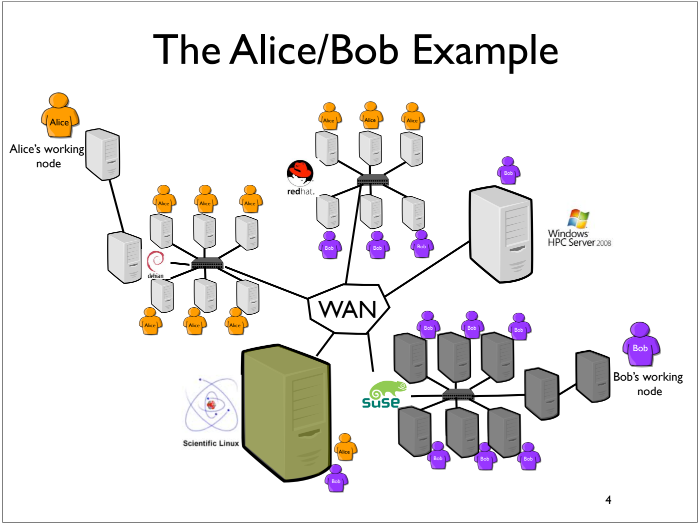


Courtesy of Adrien Lèbre (2010)

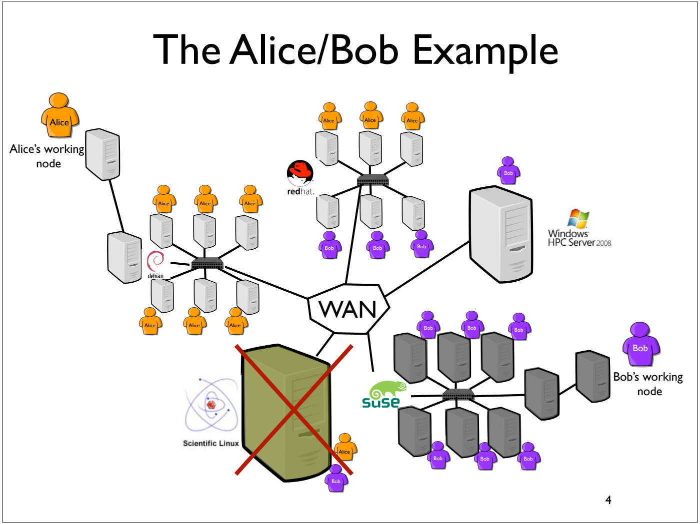
The Alice/Bob Example



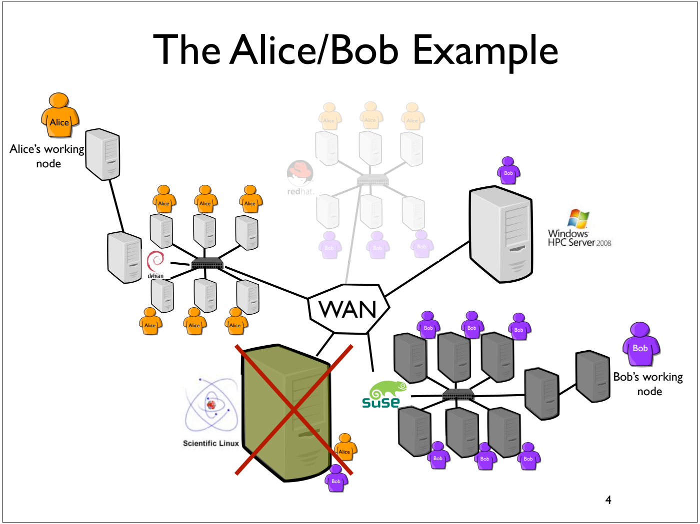
Courtesy of Adrien Lèbre (2010)



Courtesy of Adrien Lèbre (2010)

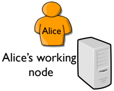


Courtesy of Adrien Lèbre (2010)

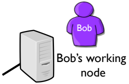


Courtesy of Adrien Lèbre (2010)

What a Grid!?!

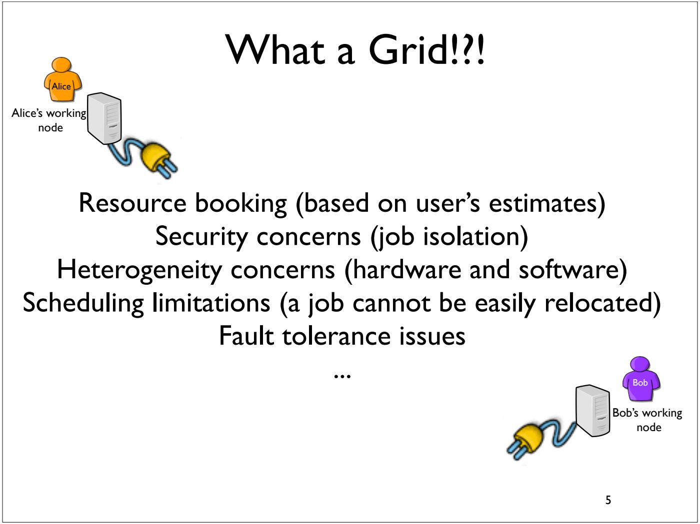


Resource booking (based on user's estimates)
Security concerns (job isolation)
Heterogeneity concerns (hardware and software)
Scheduling limitations (a job cannot be easily relocated)
Fault tolerance issues
...

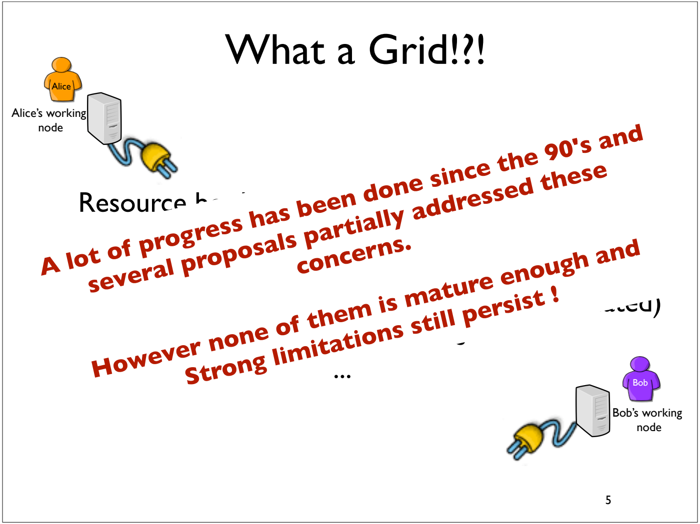


5

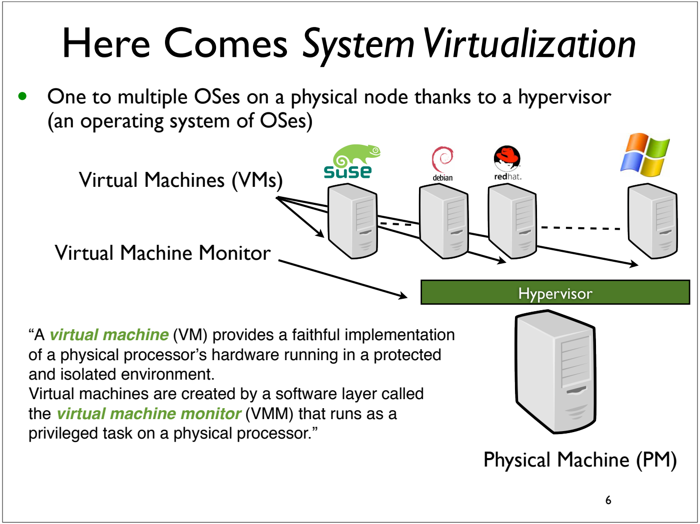
Courtesy of Adrien Lèbre (2010)



Courtesy of Adrien Lèbre (2010)



Courtesy of Adrien Lèbre (2010)



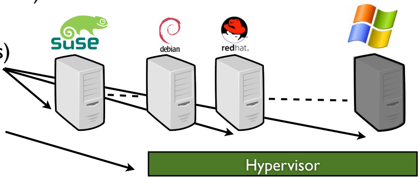
Courtesy of Adrien Lèbre (2010)

Here Comes *System Virtualization*

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

Virtual Machines (VMs)

Virtual Machine Monitor



"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 (2010)

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 (2010)

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)

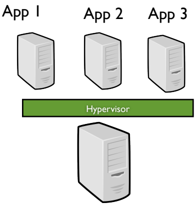
  Revival of System Virtualization approach (VmWare/Xen)

  Hard or soft partitioning of SMP/Numa Server

8

Courtesy of Adrien Lèbre (2010)

VM Capabilities

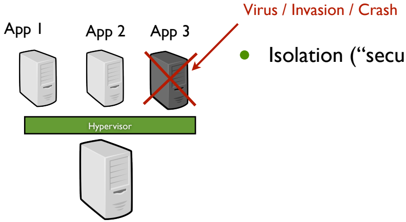


- Isolation ("security" between each VM)

9

Courtesy of Adrien Lèbre (2010)

VM Capabilities



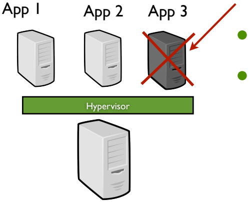
- Isolation ("security" between each VM)

9

Courtesy of Adrien Lèbre (2010)

VM Capabilities

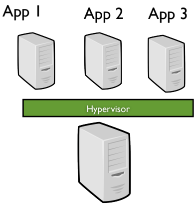
Virus / Invasion / Crash



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

Courtesy of Adrien Lèbre (2010)

VM Capabilities

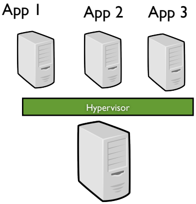


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)

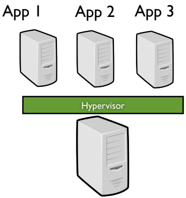
9

Courtesy of Adrien Lèbre (2010)

VM Capabilities



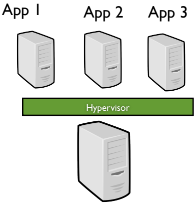
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume



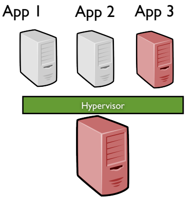
9

Courtesy of Adrien Lèbre (2010)

VM Capabilities



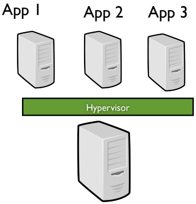
- Isolation (“security” between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)



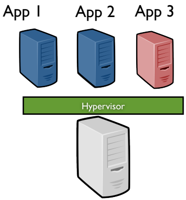
- Suspend/Resume

Courtesy of Adrien Lèbre (2010)

VM Capabilities

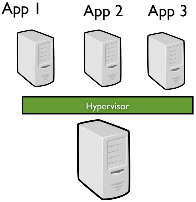


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume

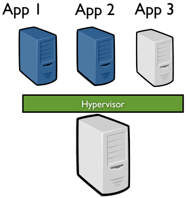


Courtesy of Adrien Lèbre (2010)

VM Capabilities

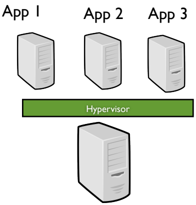


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume

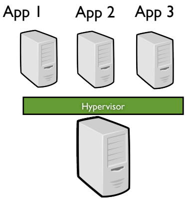


Courtesy of Adrien Lèbre (2010)

VM Capabilities



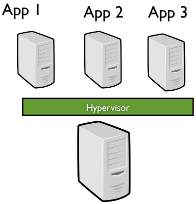
- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume



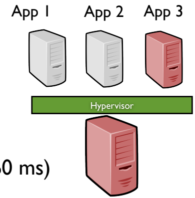
9

Courtesy of Adrien Lèbre (2010)

VM Capabilities



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

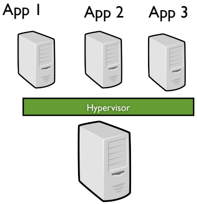


- Suspend/Resume
- Live migration
  (negligible downtime ~ 60 ms)

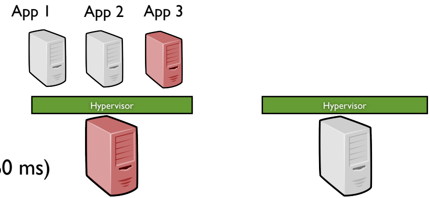
9

Courtesy of Adrien Lèbre (2010)

VM Capabilities

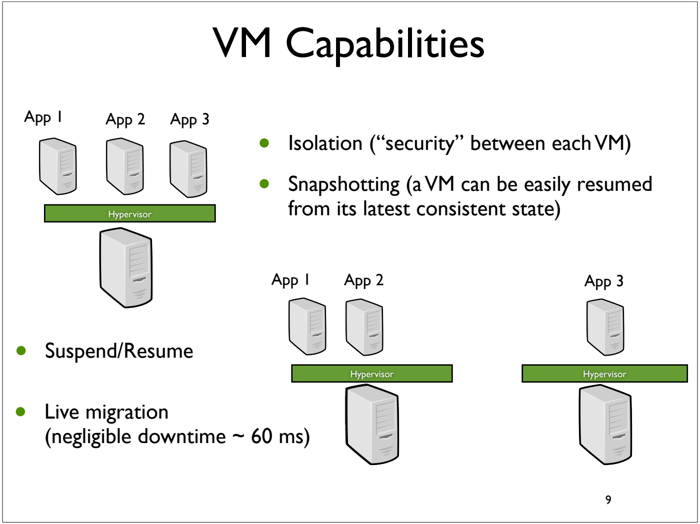


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume
- Live migration (negligible downtime ~ 60 ms)



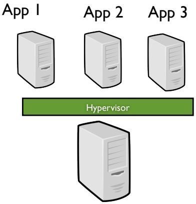
9

Courtesy of Adrien Lèbre (2010)

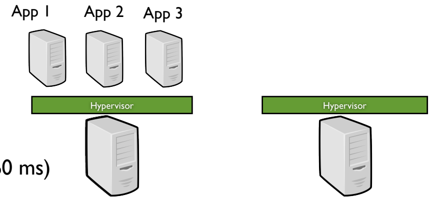


Courtesy of Adrien Lèbre (2010)

VM Capabilities

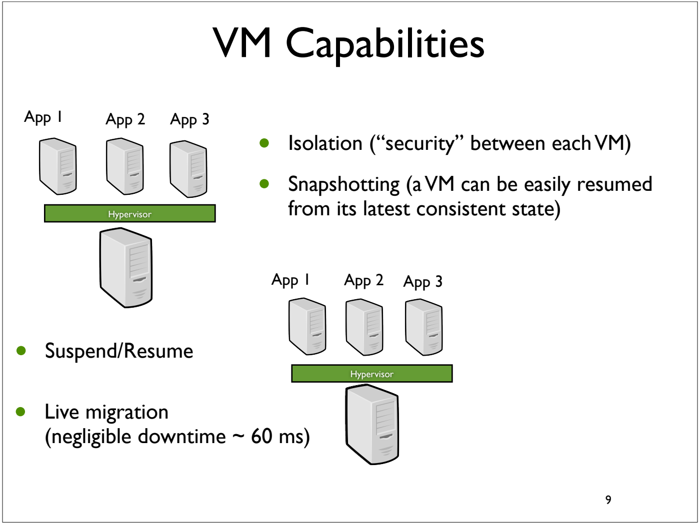


- Isolation ("security" between each VM)
- Snapshotting (a VM can be easily resumed from its latest consistent state)
- Suspend/Resume
- Live migration (negligible downtime ~ 60 ms)

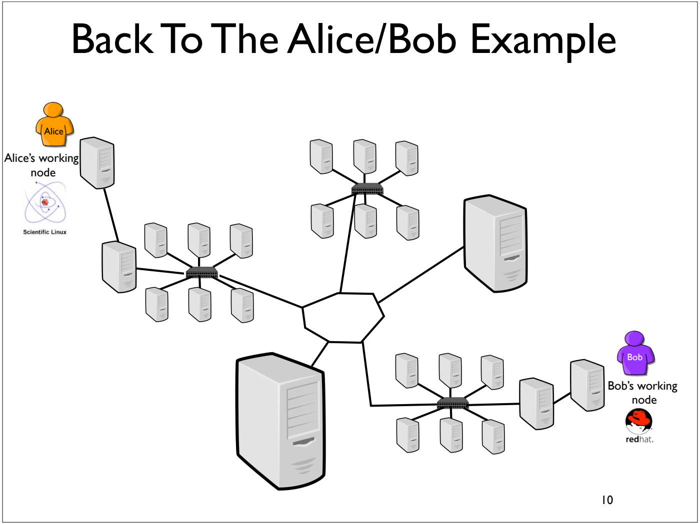


9

Courtesy of Adrien Lèbre (2010)

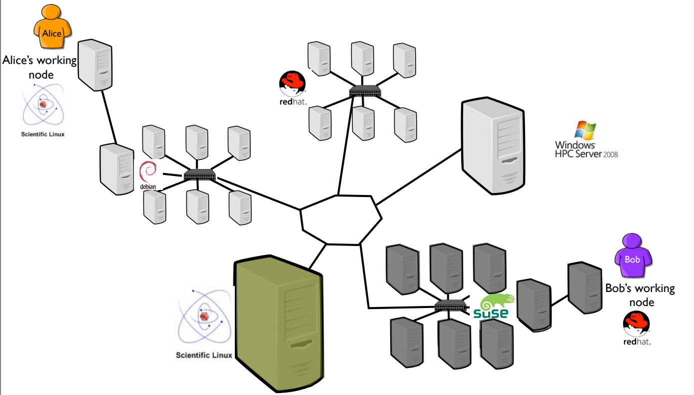


Courtesy of Adrien Lèbre (2010)



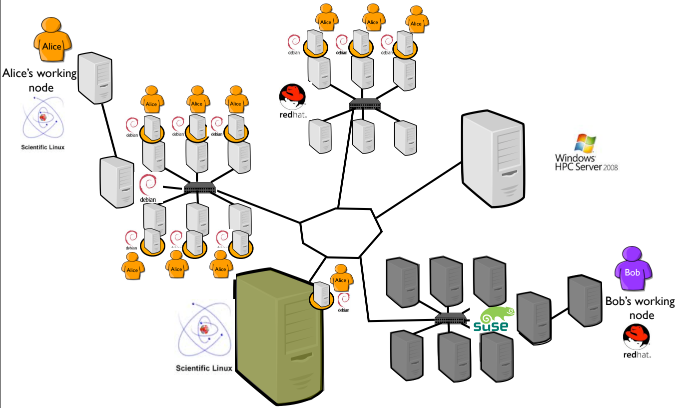
Courtesy of Adrien Lèbre (2010)

Back To The Alice/Bob Example



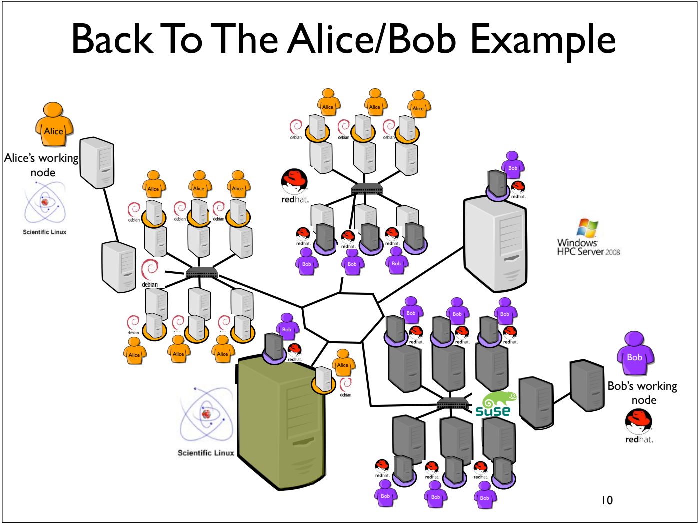
Courtesy of Adrien Lèbre (2010)

Back To The Alice/Bob Example

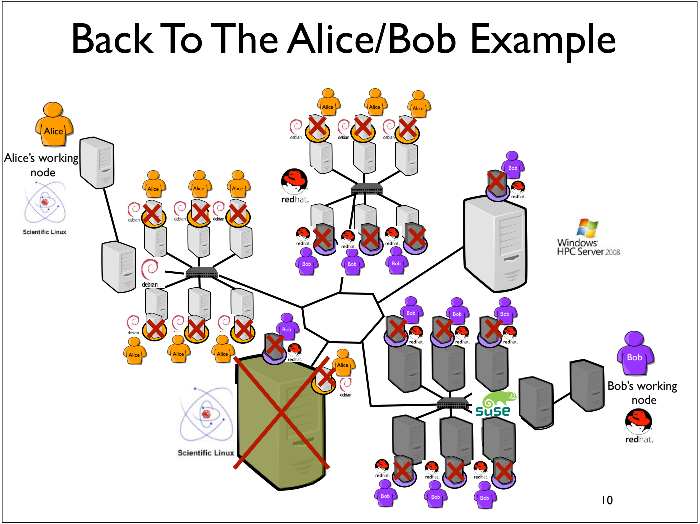


10

Courtesy of Adrien Lèbre (2010)

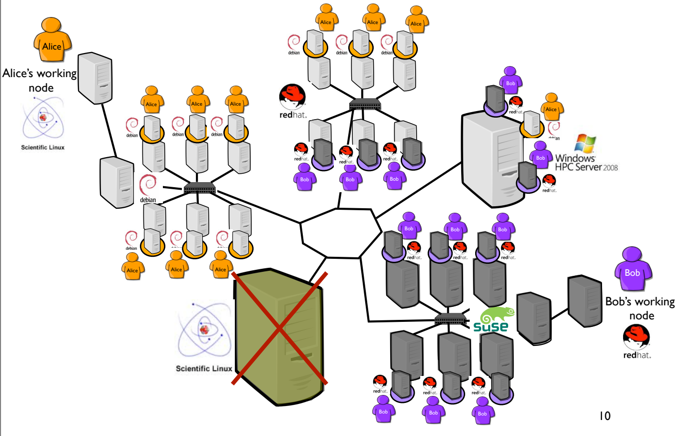


Courtesy of Adrien Lèbre (2010)

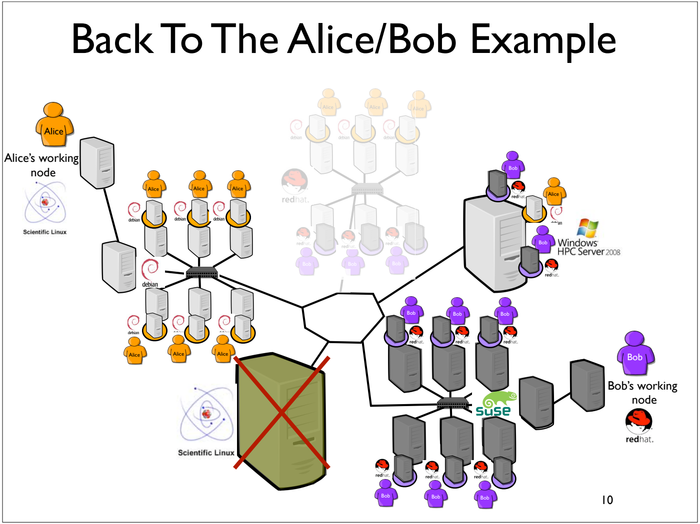


Courtesy of Adrien Lèbre (2010)

Back To The Alice/Bob Example

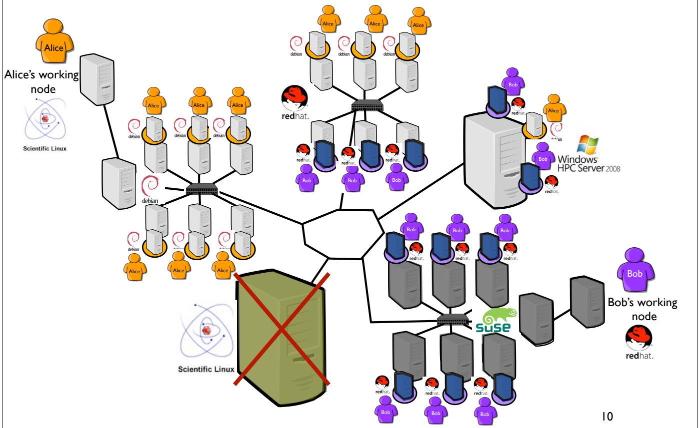


Courtesy of Adrien Lèbre (2010)



Courtesy of Adrien Lèbre (2010)

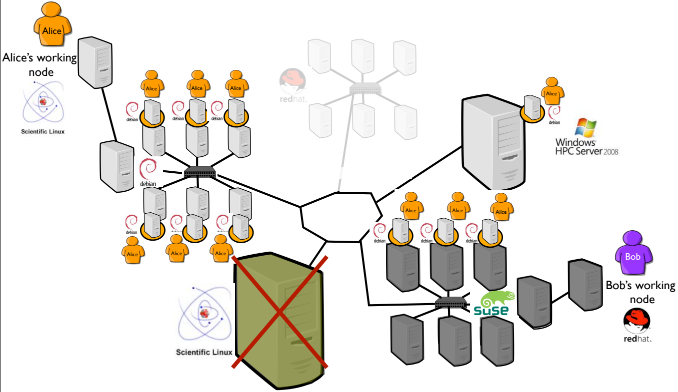
Back To The Alice/Bob Example



10

Courtesy of Adrien Lèbre (2010)

Back To The Alice/Bob Example



10

Courtesy of Adrien Lèbre (2010)

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 (2010)

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, 196
credits: I. Fos … hree Point Checklist

1961, Prof. John McCarthy

11

Courtesy of Adrien Lèbre (2010)

Focus on dynamical scheduling concerns

What can be done thanks to VM capabilities

Courtesy of Adrien Lèbre (2010)

Context

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

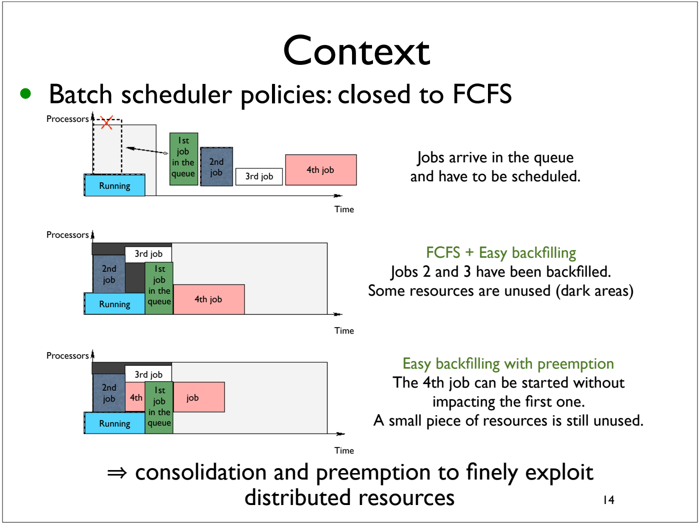
Based on user estimates (time/resources)
For a bounded amount of time
(e.g. 4 nodes for 2 hours)

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 (2010)



Courtesy of Adrien Lèbre (2010)

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 (2010)

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]

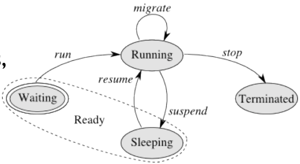
- Use case - energy concerns in Datacenters

16

Courtesy of Adrien Lèbre (2010)

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:



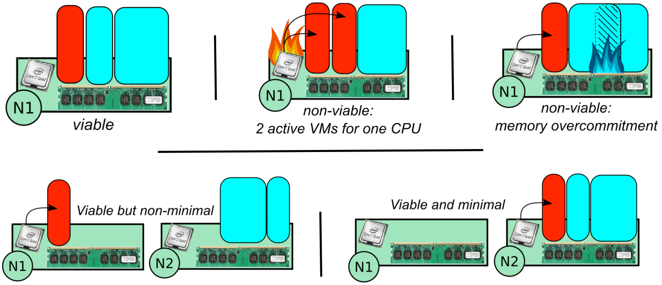
- A cluster-wide context switch (a set of VM context switches) enables to efficiently rebalance the cluster according to the: scheduler objectives / available resources / waiting vjobs queue

17

Courtesy of Adrien Lèbre (2010)

The Entropy Proposal

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

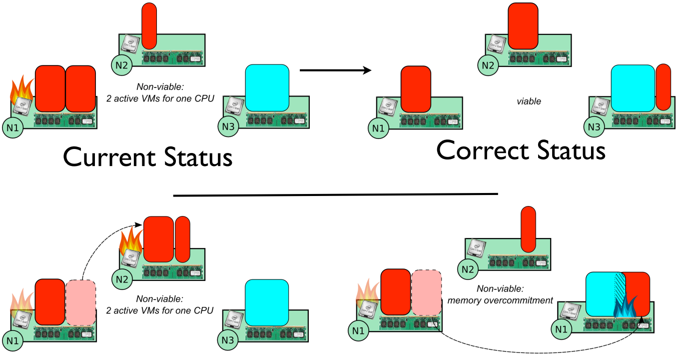


credits: F. Hermenier, Mines Nantes

18

Courtesy of Adrien Lèbre (2010)

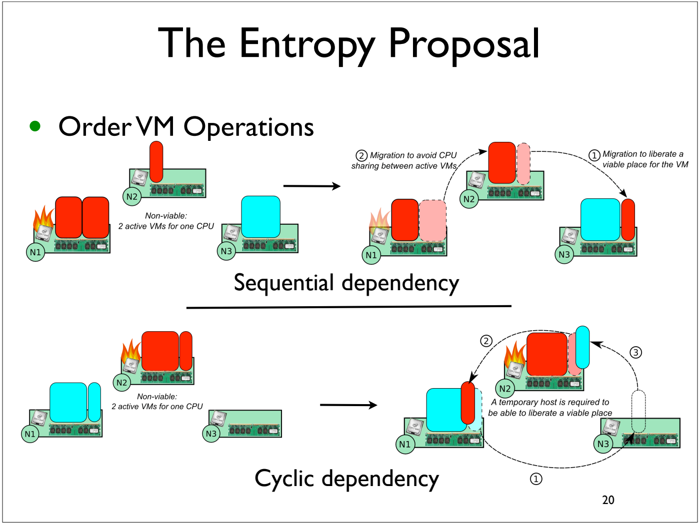
The Entropy Proposal



Non-viable manipulations

19

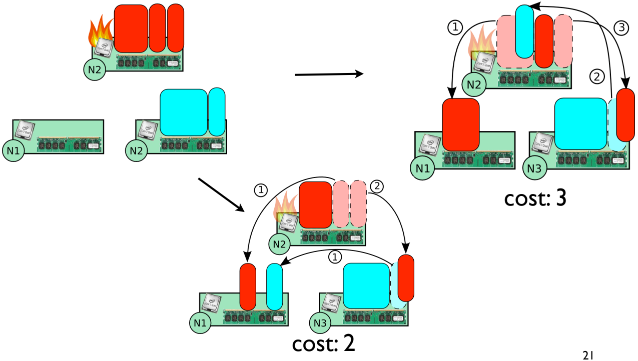
Courtesy of Adrien Lèbre (2010)



Courtesy of Adrien Lèbre (2010)

The Entropy Proposal

- Optimizing the reconfiguration process



21

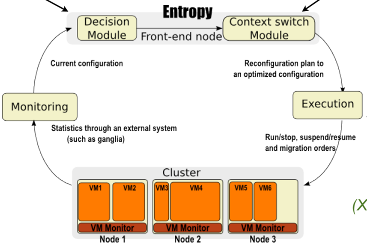
Courtesy of Adrien Lèbre (2010)

The Entropy Proposal

- The big picture: an autonomic model

Scheduling algorithm: select the jobs to run
(objectives/strategies defined by administrators)

Compute an efficient reconfiguration plan to reach the expected configuration
(through the Choco constraint solver)



Actions are done through *drivers*
(XEN XML-RPC API / ssh)

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

22

Courtesy of Adrien Lèbre (2010)

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 (2010)

Is consolidation really painless?

(12 January 2010) 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?

(12 January 2010) 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?

(12 January 2010) 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?

(12 January 2010) 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.

Cloud versus Cloud

Custom user environments!

On-demand access!

Elastic computing!

Isolation!

Capital expense -> operational expense!

Too complex: do I need to become a sys admin?

What is the best programming model, what are the tools I need to make effective use of them?

It costs too much! And what if Amazon raises prices?

Performance: especially I/O, especially Big Data!

12/1/13 NIMBUS *www.nimbusproject.org* 2

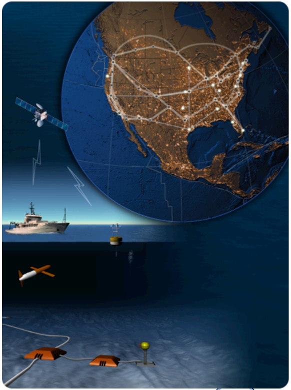
Courtesy of Kate Keahey (JLPC'13)

Cloud Storage Basics

- Ephemeral/Transient Storage
  - Local virtual disk attached to an instance
  - Persists only for the lifetime of an instance
  - Included in the cost of an instance
  - Varying capacity, e.g. 160 GB-48 TB on AWS
- Persistent attached storage
  - Block storage volumes that can be attached to an instance
  - Lifetime independent of a particular instance, can be mounted by many
  - Price based on space and time used
  - E.g., AWS Elastic Block Storage (EBS), Azure drives
- Storage Clouds
  - Data storage as binary objects (BLOBs)
  - Price differs based levels of service, e.g., access time or reliability, space used and time
  - E.g., AWS Simple Storage Service (S3), AWS Glacier, Azure BLOBs, Google Cloud Storage

Courtesy of Kate Keahey (JLPC'13)

Streaming Applications

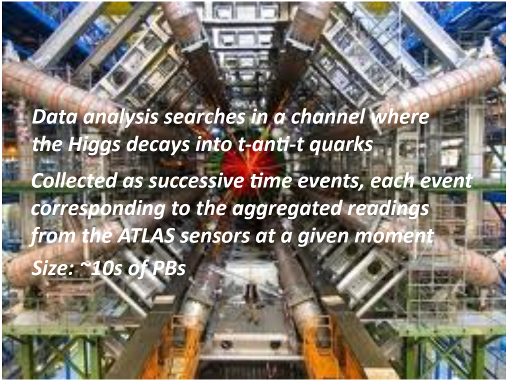


- Repeatedly apply an operation to a stream of data (time events)
- Examples:
  - Virtual Observatories: OOI, Forest project at ANL, IFC
  - Experiment processing: STAR, APS
- Requirements:
  - An "always-on" service
  - Real-time event-based data stream processing capabilities
  - Highly volatile need for data distribution and processing

12/1/13 NIMBUS *www.nimbusproject.org* 4

Courtesy of Kate Keahey (JLPC'13)

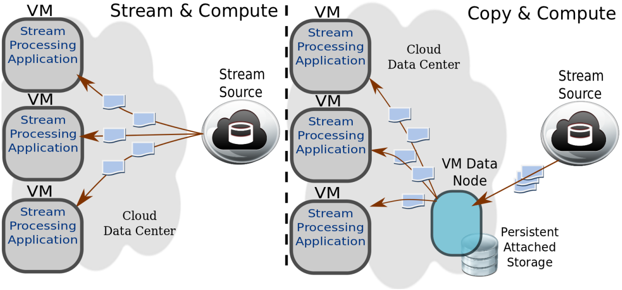
ATLAS Data Analysis



12/1/13 NIMBUS www.nimbusproject.org 5

Courtesy of Kate Keahey (JLPC'13)

Streaming Scenarios



Courtesy of Kate Keahey (JLPC'13)

Streaming Scenarios (2)

Stream&Compute (SC)

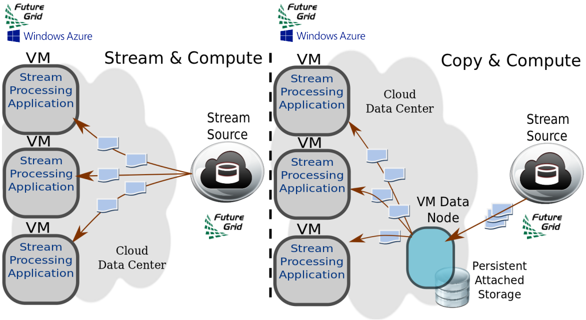
- Simpler model with fewer moving parts
- Potentially better response time
- Overlap computation and communication (potentially faster)
- Uses ephemeral storage (potentially cheaper)

Copy&Compute (CC)

- Independent of network saturation
- Persistent storage: less liable to data loss

Courtesy of Kate Keahey (JLPC'13)

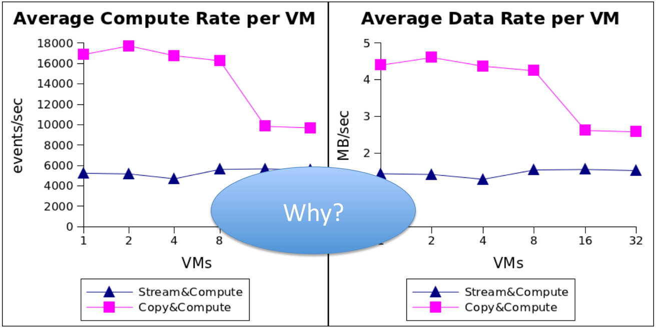
Experimental Configuration



- *Compute rate: events processed per time unit*
- *Data rate: amount of data acquired per time unit*

Courtesy of Kate Keahey (JLPC'13)

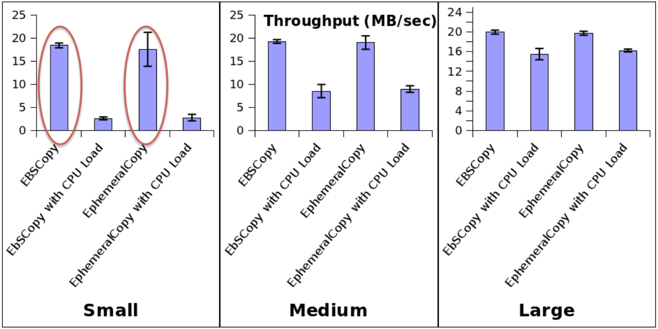
SC versus CC (Azure)



12/1/13 NIMBUS *www.nimbusproject.org* *10*

Courtesy of Kate Keahey (JLPC'13)

Data Throughput vs CPU Load



Courtesy of Kate Keahey (JLPC'13)

Cost

$$Total_{Cost} = \frac{TotalEvents}{CompRate_{Total}} * (N_{VMsData} + N_{VMsComp}) * VM_{Cost} + Storage_{Cost}$$

- Cost of instance: ~$0.1 per hour
- Cost of storage: ~$0.1 per 1GB month
- In our case (320M events & 5 GB attached storage)
  - Stream&Compute: $1.33
  - Copy&Compute: $0.48
  - Overall: SC is 2.77 times more expensive

Courtesy of Kate Keahey (JLPC'13)

Conclusions

- To stream or not to stream?
  - Not to stream!
  - Difference of ~4x in performance and ~3x in cost
- Amplification of virtualization performance trade-offs in the presence of remote traffic
- Hypervisor design
  - Need for controlled allocation of CPU to I/O processing
- *Paper: Tudoran et al., "Evaluating Streaming Strategies for Event Processing across Infrastructure Clouds", submitted to CCGrid*

Courtesy of Kate Keahey (JLPC'13)

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 has been #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

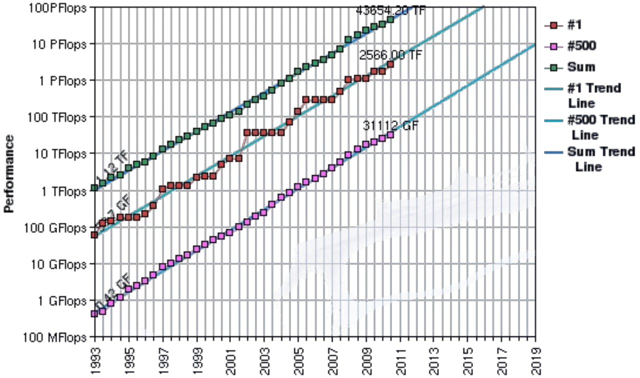
Tianhe-2 (MilkyWay-2)

- Simulation, analysis, and government security applications
- 16,000 computer nodes, each comprising two Intel Ivy Bridge Xeon processors and three Xeon Phi chips for a total of 3,120,000 cores
- 33.8 PFlops (Peak=54.9 PFlops)
- **17.8 GW**!!!

Tianhe-2 (MilkyWay-2)

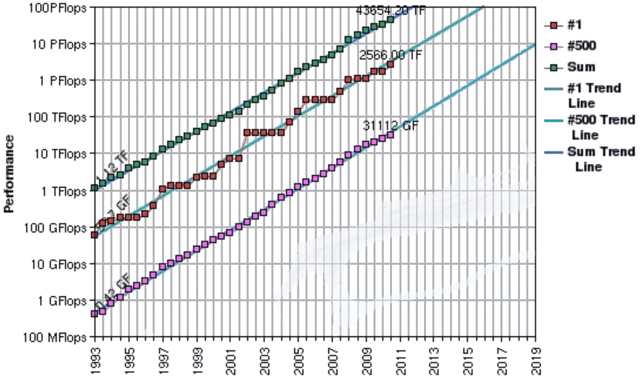
- Simulation, analysis, and government security applications
- 16,000 computer nodes, each comprising two Intel Ivy Bridge Xeon processors and three Xeon Phi chips for a total of 3,120,000 cores
- 33.8 PFlops (Peak=54.9 PFlops)
- **17.8 GW**!!!

Toward Exascale



- 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



- 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

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?

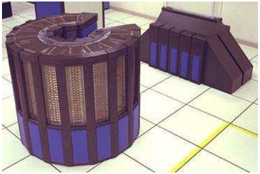
- In current systems the processors consume a lion's share of the energy approximately 43% or more.
- The remaining energy is used to power up the memories, the interconnection network, and the storage system.
- Furthermore, a significant fraction is wasted in power supply overheads, and in thermal dissipation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

In the beginning ... there were only supercomputers

- Built to order
  - Very few of them
- Special purpose hardware
  - Very expensive

- Control Data, Convex, ...
- Cray-1
  - 1975, 160 MFLOPS
    - 80 units, 5-8 M$
- Cray X-MP
  - 1982, 800 MFLOPS
- Cray-2
  - 1985, 1.9 GFLOPS
- Cray Y-MP
  - 1988, 2.6 GFLOPS

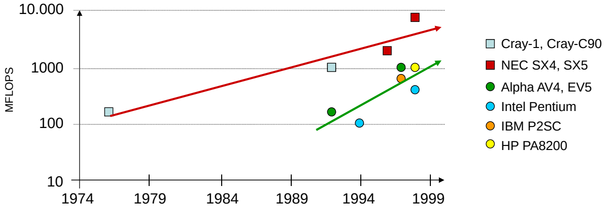
- Fortran+vectorizing compilers



3 HPC Advisory Council, Malaga September 13, 2012 MONT-BLANC

Courtesy of Alex Ramirez (2013)

The Killer Microprocessors



- Microprocessors killed the Vector supercomputers
  - They were not faster ...
  - ... but they were significantly cheaper and greener
- Need 10 microprocessors to achieve the performance of 1 Vector CPU
  - SIMD vs. MIMD programming paradigms

Courtesy of Alex Ramirez (2013)

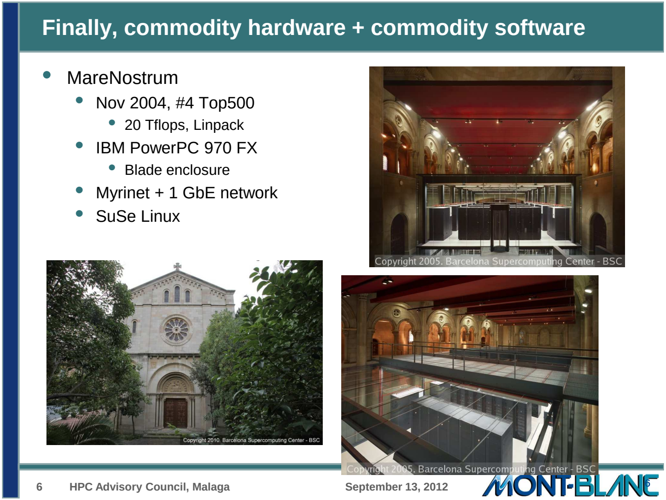
Then, commodity took over special purpose

- ASCI Red, Sandia
  - 1997, 1 Tflops (Linpack),
  - 9298 cores @ 200 Mhz
  - 1.2 Tbytes
  - Intel Pentium Pro
    - Upgraded to Pentium II Xeon, 1999, 3.1 Tflops

- ASCI White, LLNL
  - 2001, 7.3 TFLOPS
  - 8192 proc. @ 375 Mhz,
  - 6 Tbytes
  - (3+3) Mwats
  - IBM Power 3

Message-Passing Programming Models

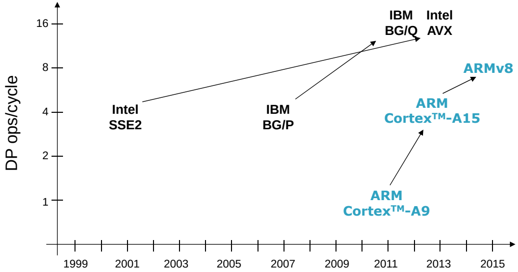
Courtesy of Alex Ramirez (2013)



Courtesy of Alex Ramirez (2013)

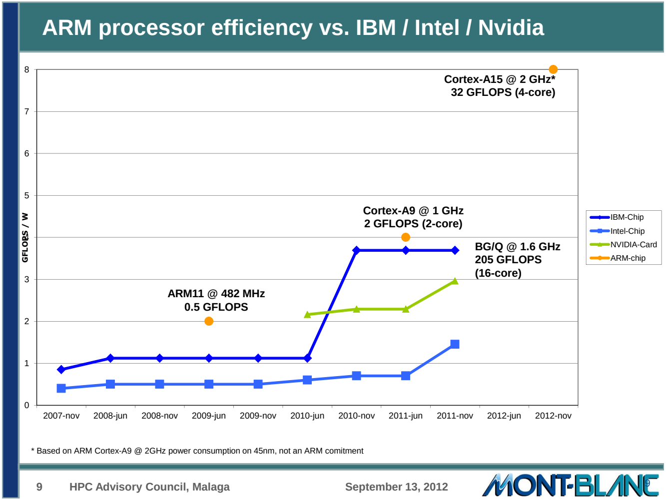
Courtesy of Alex Ramirez (2013)

ARM Processor improvements in DP FLOPS



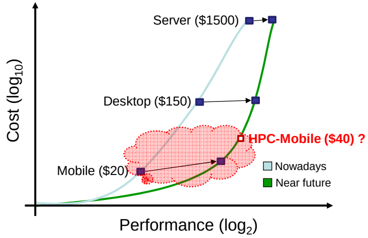
- IBM BG/Q and Intel AVX implement DP in 256-bit SIMD
  - 8 DP ops / cycle
- ARM quickly moved from optional floating-point to state-of-the-art
  - ARMv8 ISA introduces DP in the NEON instruction set (128-bit SIMD)

Courtesy of Alex Ramirez (2013)



Courtesy of Alex Ramirez (2013)

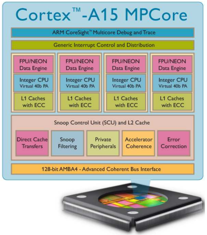
Are the “Killer Mobiles™" coming?

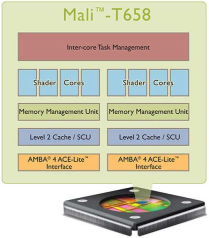


- Where is the sweet spot? Maybe in the low-end ...
  - Today ~ 1:8 ratio in performance, 1:100 ratio in cost
  - Tomorrow ~ 1:2 ratio in performance, still 1:100 in cost ?
- The same reason why microprocessors killed supercomputers
  - Not so much performance ... but much lower cost, and power

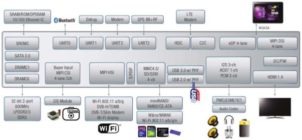
Courtesy of Alex Ramirez (2013)

Killer mobile™ example: Samsung Exynos 5450 *





- 4-core ARM Cortex-A15 @ 2 GHz
  - 32 GFLOPS
- 8-core ARM Mali T685
  - 168 GFLOPS*
- Dual channel DDR3 memory controller
- **All in a low-power mobile socket**



* Data from web sources, not an ARM or Samsung commitment

11 HPC Advisory Council, Malaga September 13, 2012 MONT-BLANC

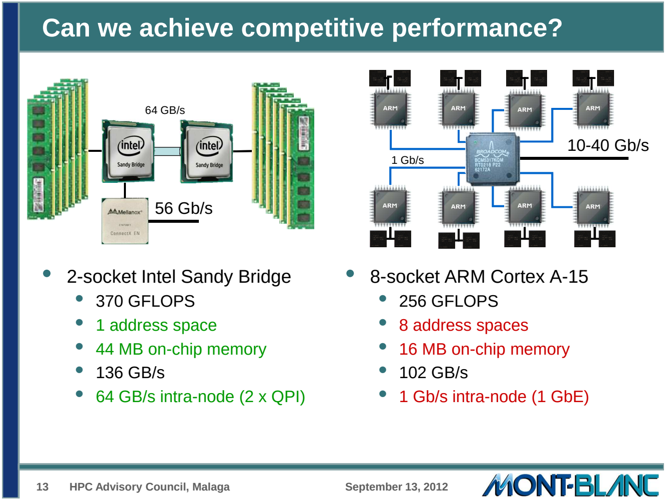
Courtesy of Alex Ramirez (2013)

Are we building BlueGene again?

- Yes ...
  - Exploit Pollack's Rule in presence of abundant parallelism
    - Many small cores vs. Single fast core
- ... and No
  - Heterogeneous computing
    - On-chip GPU
  - Commodity vs. Special purpose
    - Higher volume
    - Many vendors
    - Lower cost
  - Lots of room for improvement
    - No SIMD / vectors yet ...
  - Build on Europe's embedded strengths

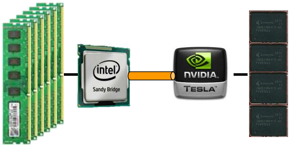
12 HPC Advisory Council, Malaga September 13, 2012 MONT-BLANC

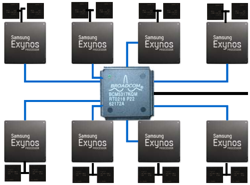
Courtesy of Alex Ramirez (2013)



Courtesy of Alex Ramirez (2013)

Can we achieve competitive performance?

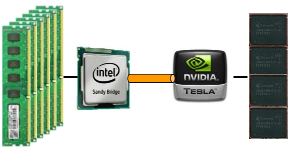


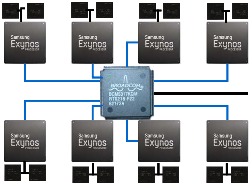


- Sandy Bridge + Nvidia K20
  - 1685 GFLOPS
  - 2 address spaces
  - 32 GB/s between CPU-GPU
    - 16x PCIe 3.0
  - 68 + 192 GB/s
- 8-socket Exynos 5450
  - 1600 GFLOPS
  - 16 address spaces
  - 12.8 GB/s between CPU-GPU
    - Shared memory
  - 102 GB/s

Courtesy of Alex Ramirez (2013)

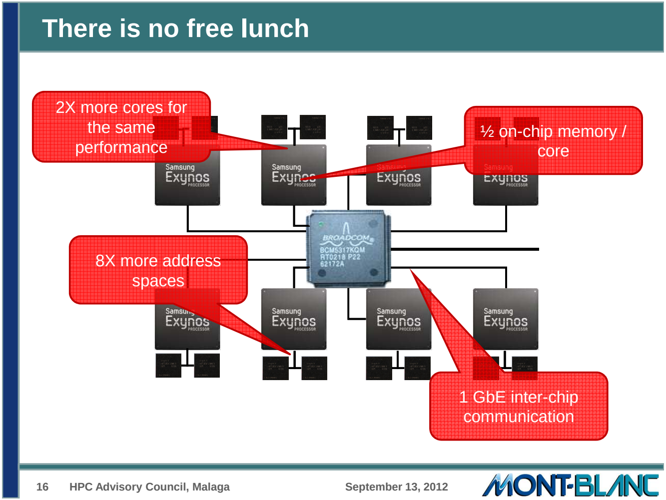
Then, what is so good about it?





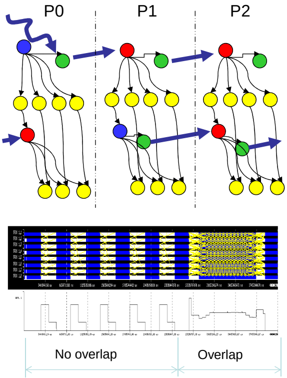
- Sandy Bridge + Nvidia K20
  - > $3000
  - > 400 Watt
- 8-socket Exynos 5450
  - < $200
  - < 100 Watt

Courtesy of Alex Ramirez (2013)



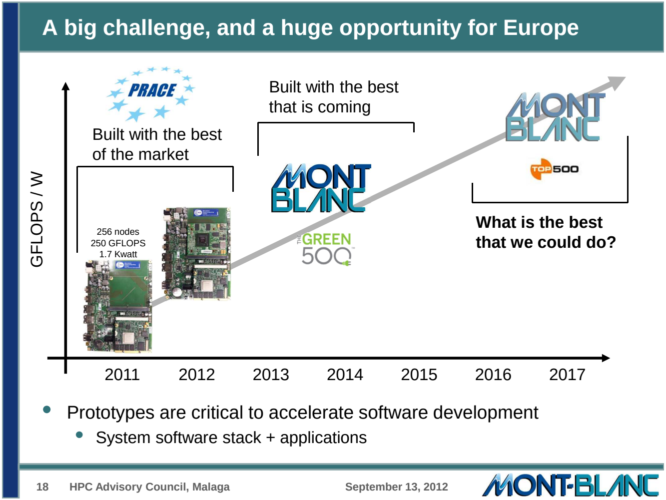
Courtesy of Alex Ramirez (2013)

OmpSs runtime layer manages architecture complexity



- Programmer exposed a simple architecture
- Task graph provides lookahead
  - Exploit knowledge about the future
- Automatically handle all of the architecture challenges
  - Strong scalability
  - Multiple address spaces
  - Low cache size
  - Low interconnect bandwidth
- Enjoy the positive aspects
  - Energy efficiency
  - Low cost

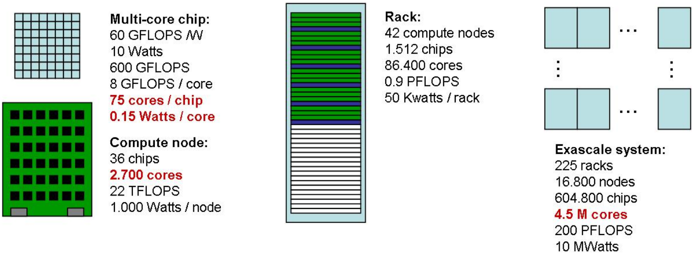
Courtesy of Alex Ramirez (2013)



Courtesy of Alex Ramirez (2013)

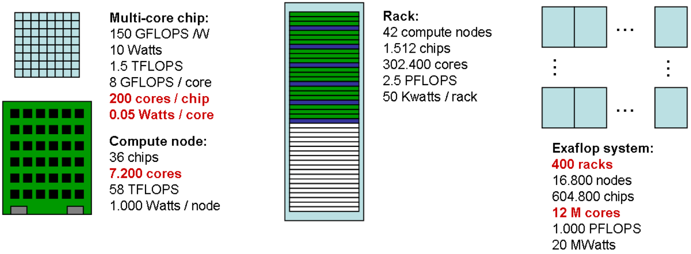
Possible Architecture: 200 PFlops with 10MWatt

- 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.



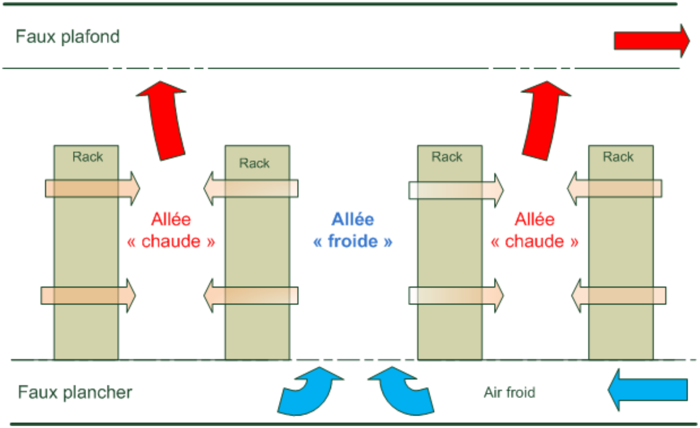
Projection for Exascale

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

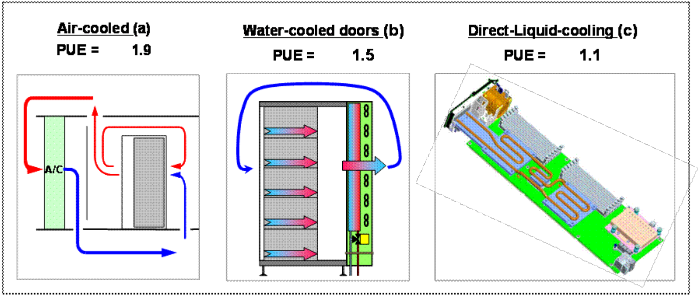


- Require new memory architecture, network, . . .

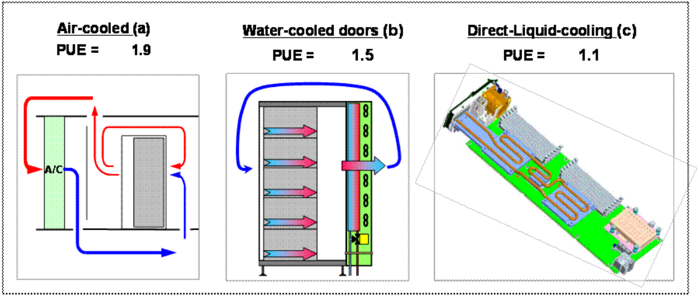
Thermal dissipation



Thermal dissipation



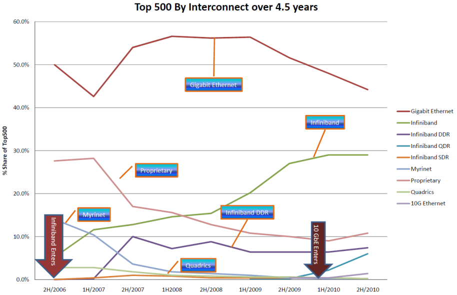
Thermal dissipation



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 more than a third 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

Toward Exascale

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,503.17	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	27.78
2	3,631.86	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
5	3,130.95	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
6	3,068.71	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.930GHz, Infiniband QDR, NVIDIA K20x	922.54
7	2,702.16	University of Arizona	iDataPlex DX360M4, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR14, NVIDIA K20x	53.62
8	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94
9	2,629.10	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	55.62
10	2,358.69	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2.000GHz, Infiniband FDR, Nvidia K20m	71.01

- Greenest supercomputer = 4.5GFlops/W
- Fastest supercomputer = 1.8GFlops/W
- Accelerators (GPU, Xeon Phi) are becoming more and more common.

The DEEP Exascale project

- DEEP: Dynamical Exascale Entry Platform
- one of the three Exascale projects funded by the EU: DEEP, CRESTA and MONTBLANC.
- It involves 16 partners from 8 different countries and is coordinated by the Jülich Supercomputing Centre.
- The project is a two-fold approach to the exascale challenge:

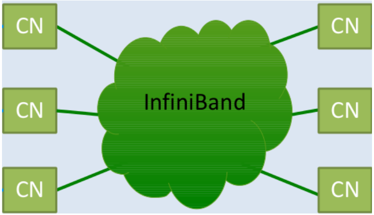
Hardware a novel supercomputing architecture: instead of adding accelerator cards to cluster nodes, an accelerator cluster, called Booster, will complement a conventional HPC system and increase its performance.

Software a matching software stack that includes programming models, libraries and performance tools adapted to its architecture.

- It will enable unprecedented scalability.

Courtesy of DEEP (2013)

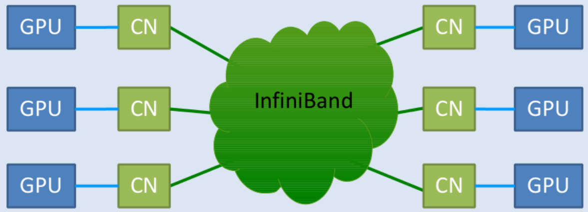
DEEP's architecture



Today's Clusters: CPU nodes connected by Infiniband

Courtesy of DEEP (2013)

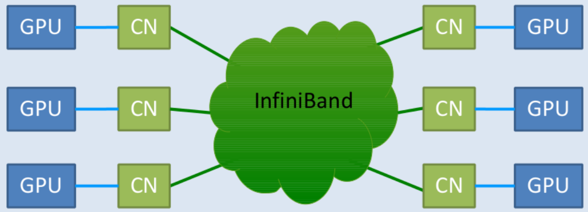
DEEP's architecture



Accelerators statically attached to CPU

Courtesy of DEEP (2013)

DEEP's architecture

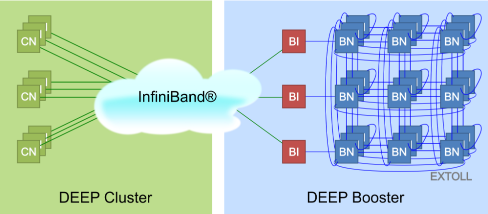


Accelerators statically attached to CPU

- Ideally: accelerator cluster + CPU cluster
- Problem: accelerators cannot run autonomously

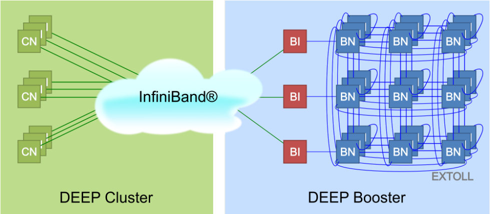
Courtesy of DEEP (2013)

DEEP's architecture



Courtesy of DEEP (2013)

DEEP's architecture

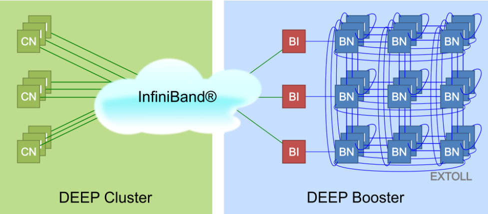


Booster: accelerator cluster

- Intel MIC processors
- EXTOLL network – developed at University of Heidelberg

Courtesy of DEEP (2013)

DEEP's architecture



Booster: accelerator cluster

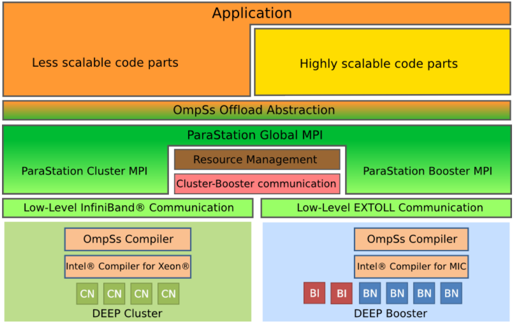
- Intel MIC processors
- EXTOLL network – developed at University of Heidelberg

Cluster

- Intel Xeon processors
- Infiniband connect by Mellanox

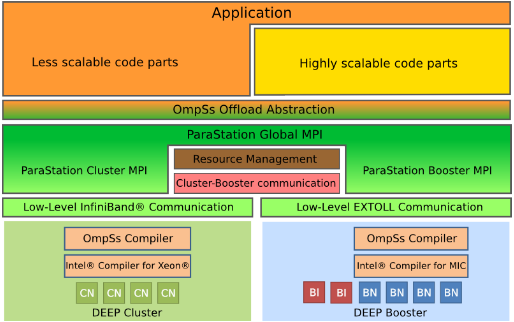
Courtesy of DEEP (2013)

DEEP's software



Courtesy of DEEP (2013)

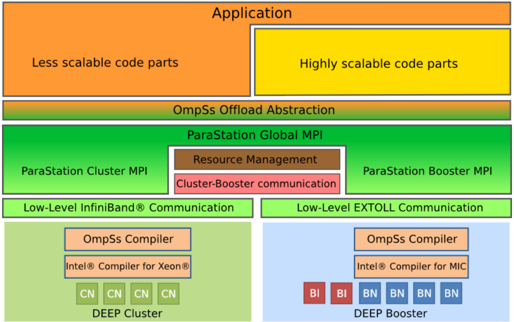
DEEP's software



OmpSs developed by the Barcelona Supercomputing Center, allows to decompose applications into tasks sent to the Cluster or the Booster efficiently.

Courtesy of DEEP (2013)

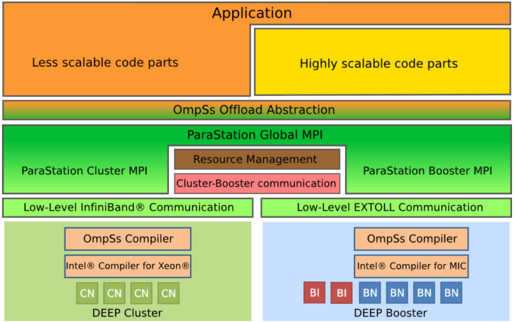
DEEP's software



Parastation MPI will allow to run traditional applications both on the Cluster and the Booster, for this Parastation MPI will be extended to support the Booster and its 3D torus topology.

Courtesy of DEEP (2013)

DEEP's software

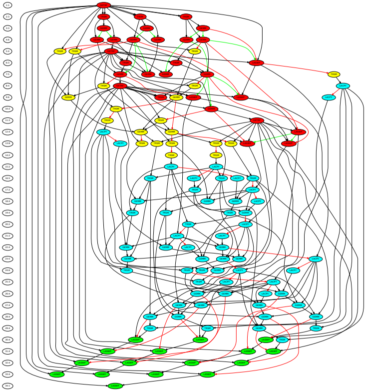


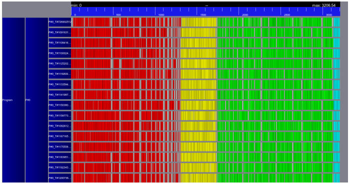
All software will be completely reworked in order to be optimized for DEEP.

Courtesy of DEEP (2013)

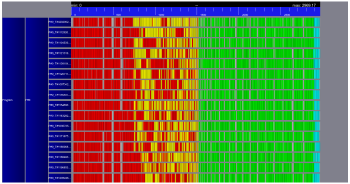
System software

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.

Performance tip!

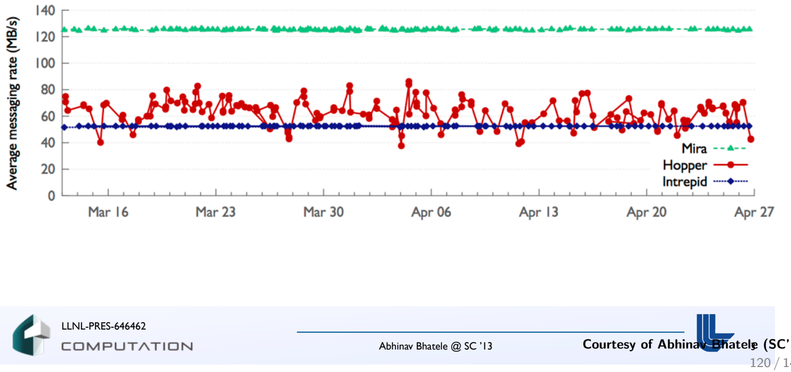
Cray machines

IBM machines

Courtesy of Abhinav Bhatele (SC'13)

Performance variability

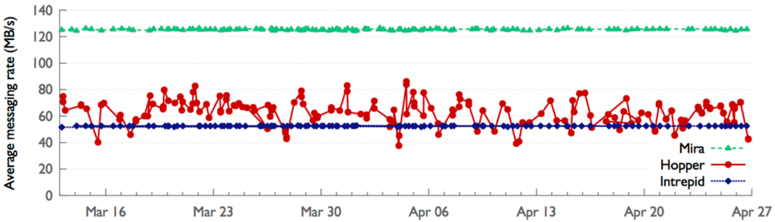
Average messaging rates for batch jobs running a laser-plasma interaction code



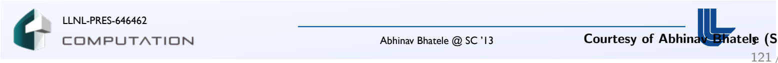
Courtesy of Abhinav Bhatele (SC'13)

Performance variability

Average messaging rates for batch jobs running a laser-plasma interaction code



$$\frac{\text{Total number of bytes sent on the network}}{\text{Time spent sending the messages}}$$



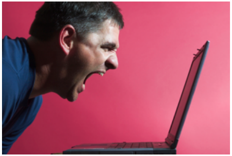
Courtesy of Abhinav Bhatele (SC'13)

Leads to several problems ...

- Individual jobs run slower:
  - More time to complete science simulations
  - Increased wait time in job queues
  - Inefficient use of machine time allocation/core-hours
- Overall lower throughput
- Increase energy usage/costs

Also affects software development

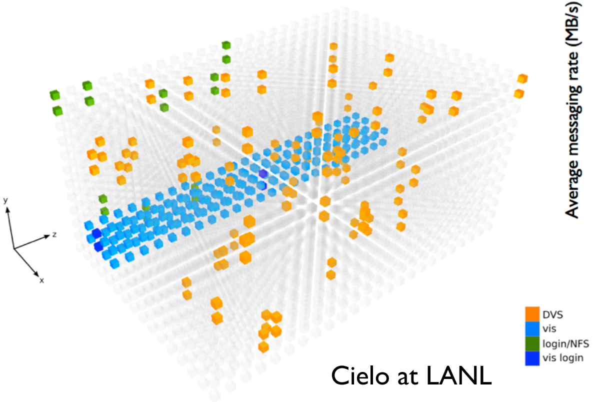
- Debugging performance issues
- Quantifying the effect of various software changes on performance
  - code changes
  - compiler/software stack changes
- Requesting time for a batch job
- Writing allocation proposals

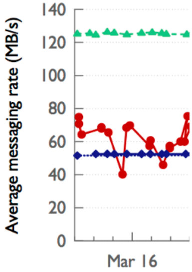


Setup: Machines

- Hopper: a Cray XE6 at LBNL
  - 2.1 GHz Opterons, 1.28 Petaflop/s
  - 3D Torus, 4 X, Z and 2 Y links, 9.4 GB/s
- Intrepid: an IBM Blue Gene/P at ANL
  - 0.85 GHz PowerPC 450, 0.56 Petaflop/s
  - 3D Torus, 0.425 GB/s
- Mira: an IBM Blue Gene/Q at ANL
  - 1.6 GHz PowerPC A2, 10 Petaflop/s
  - 5D Torus, 2 GB/s

Focus on Cray XE



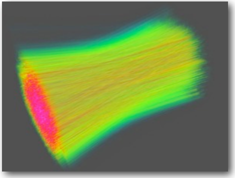


Cielo at LANL

Courtesy of Abhinav Bhatele (SC'13)

Setup: Application

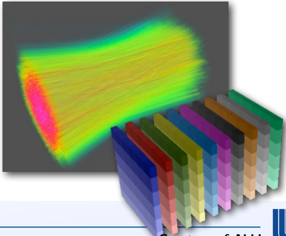
- pF3D: a highly scalable communication-heavy code
  - used to study laser beam and plasma interactions
- Balanced computation and communication across MPI processes
- 3D virtual process grid
  - 1D FFTs in X and Y direction
  - Send-receives in Z direction



Courtesy of Abhinav Bhatele (SC'13)

Setup: Application

- pF3D: a highly scalable communication-heavy code
  - used to study laser beam and plasma interactions
- Balanced computation and communication across MPI processes
- 3D virtual process grid
  - 1D FFTs in X and Y direction
  - Send-receives in Z direction



Data collection for the paper

- One or more runs on each machine every day: 512 nodes
- Information collected:
  - pF3D stats: messaging rate, time spent in different phases
  - queue status: running jobs and their placement
  - mpiP profiles: time spent in MPI operations

Data collection for the paper

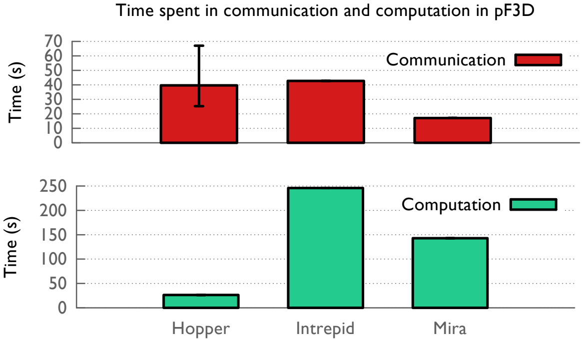
- One or more runs on each machine every day: 512 nodes
- Information collected:
  - pF3D stats: messaging rate, time spent in different phases
  - queue status: running jobs and their placement
  - mpiP profiles: time spent in MPI operations

Machine	Run No.	No. of nodes	No. of cores	No. of jobs	Period From	Period To	Process Topology	Domain $n_x \times n_y \times n_z$	(x,y) FFT msg. (kB)	Adv. msg. (kB)
Hopper		512	8,192	153	Mar, 2013	Apr, 2013	$32 \times 16 \times 16$	$128 \times 128 \times 8$	4, 8	384
Intrepid		512	2,048	102	Mar, 2013	Apr, 2013	$32 \times 16 \times 4$	$128 \times 128 \times 8$	4, 8	384
Mira		512	8,192	116	Mar, 2013	Apr, 2013	$32 \times 16 \times 16$	$128 \times 128 \times 8$	4, 8	384

LLNL-PRES-646462
COMPUTATION
Abhinav Bhatele @ SC '13

Courtesy of Abhinav Bhatele (SC'13)

pF3D characterization



LLNL-PRES-646462

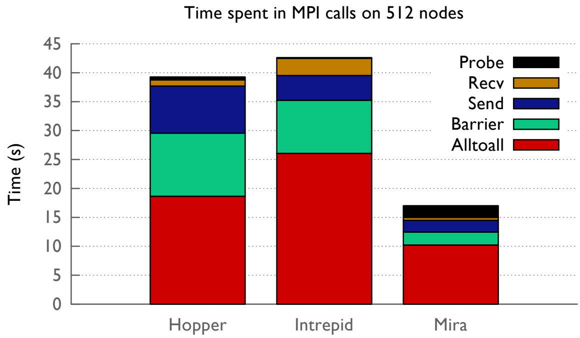
COMPUTATION

Abhinav Bhatele @ SC '13

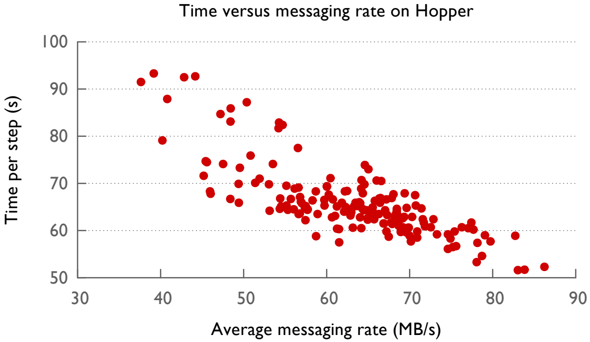
16

Courtesy of Abhinav Bhatele (SC'13)

pF3D characterization



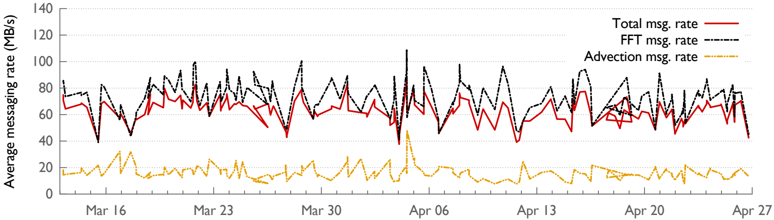
Communication in pF3D



LLNL-PRES-646462
COMPUTATION
Abhinav Bhatele @ SC '13

Communication in pF3D

Overall, FFT and advection messaging rates for pF3D on Hopper



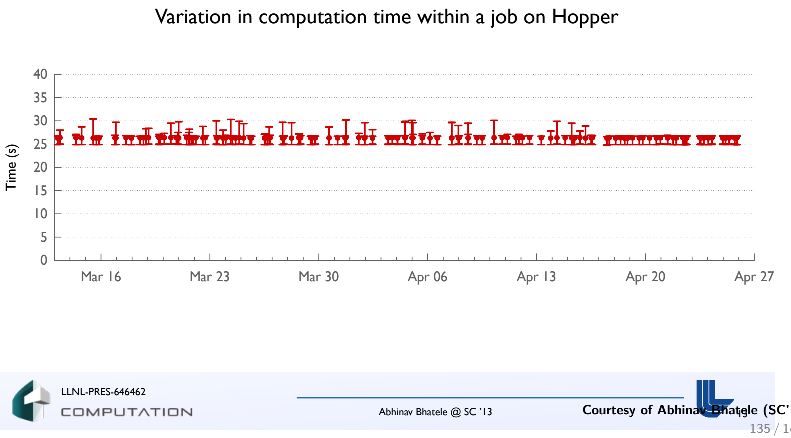
Courtesy of Abhinav Bhatele (SC'13)

Sources of variability

- Operating system noise (OS jitter)
  - OS daemons running on some cores of each node
- Placement/location of the allocated nodes for the job (Allocation shape)
- Contention for shared resources (Inter-job contention)
  - Sharing network links with other jobs

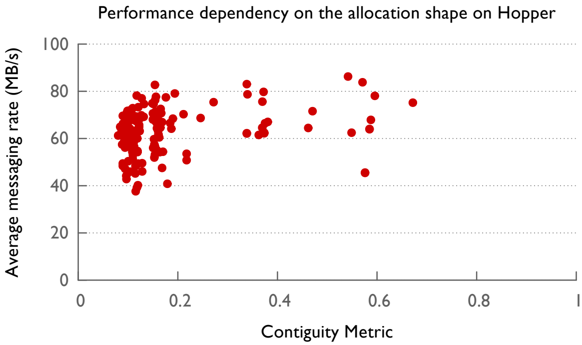
Courtesy of Abhinav Bhatele (SC'13)

OS jitter



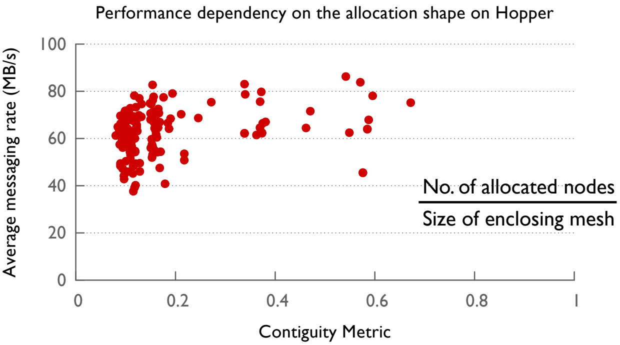
Courtesy of Abhinav Bhatele (SC'13)

Degree of fragmentation



Courtesy of Abhinav Bhatele (SC'13)

Degree of fragmentation



Courtesy of Abhinav Bhatele (SC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Overview

- What is running on Blue Waters?
- What are the issues and what to do about them?
  - Scalability
  - Runtime consistency
  - Other job interference
  - IO
  - Congestion Protection
  - Interrupts

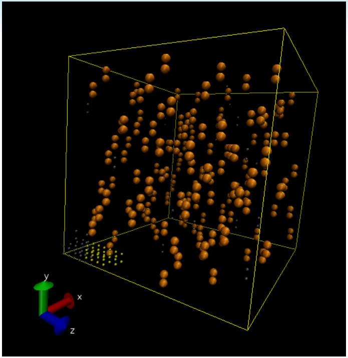
Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

IO

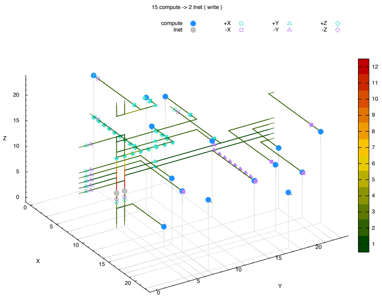
- LNETs scattered across the torus (orange colored geminis).
- Specific OSTs served by specific LNETs (not a full fat tree for the IB between OSTs and LNETs).
- IO is “topology sensitive”.



BLUE WATERS
SUSTAINED PETASCALE COMPUTING

NSF NCSA GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Routing of IO write



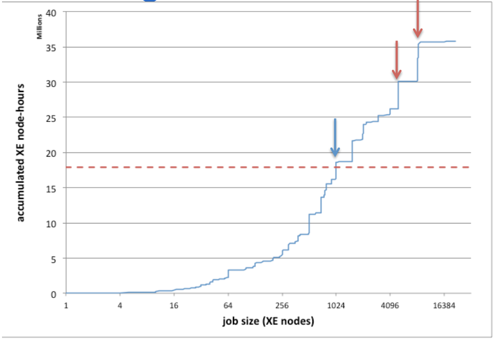
- 15 compute geminis (•) (30 nodes) writing to files served by a LNET pair (•).
- Color scale is the number of convergent routes on the link.

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

XE Usage in the last 3 months



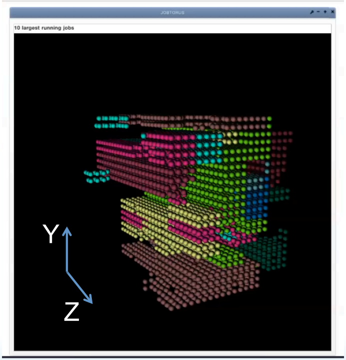
- 50% of usage is 1,024 nodes or larger.
- Two teams using 5,000 and 8,192 nodes.
- During Friendly User period, several teams sustained runs at full system.
- Nothing prevents users from submitting very large jobs and priority goes to larger jobs.
- Average expansion factor for large jobs < 10.

Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

TorusView of 10 largest running jobs



- Relatively compact allocations.
- Some scattered clustering.
- Lots of concave shapes.
- Not showing all the small jobs filling in the rest of the torus.

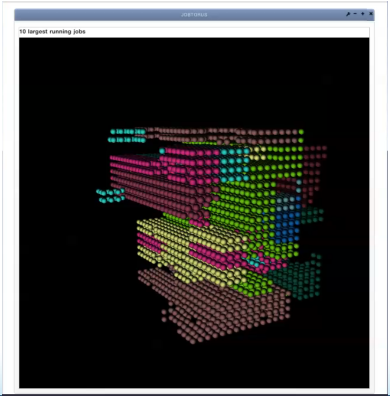
Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

NSF NCSA GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

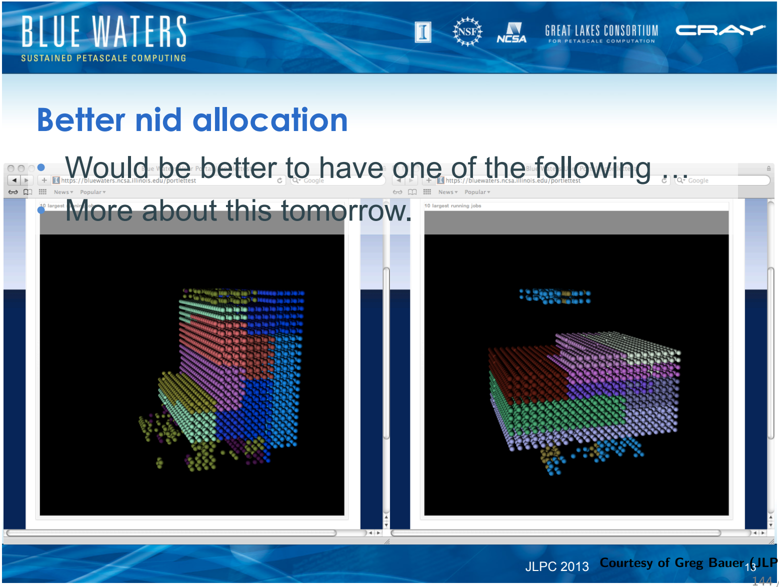
TorusView of 10 largest running jobs

- Allocations shift planes as the end of the Z direction is hit.
- Voids where larger job allocations wrap around smaller ones.



JLPC 2013 12

Courtesy of Greg Bauer (JLPC'13)



- Would be better to have one of the following ...
- More about this tomorrow.

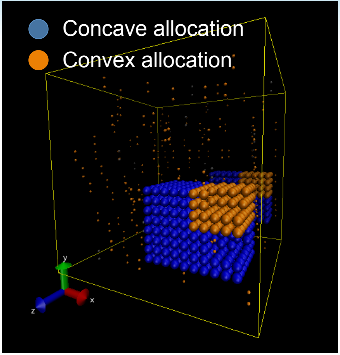
Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Impact of nid allocation

- Job – Job interaction
  - Analysis of key application communication intensity and sensitivity
  - 20% slowdown typical, 100% or more possible.



Communication	MILC	NAMD	NWCHEM	PSDNS	WRF
Intensive	2	2	3	2	1
Sensitive	2	3	1	2	1

1 – low 3 – high
as viewed by convex app.

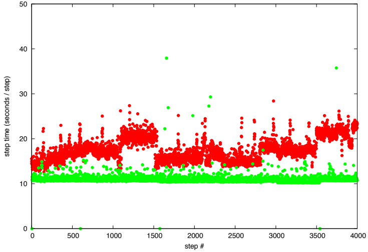
Courtesy of Greg Bauer (JLPC'13)

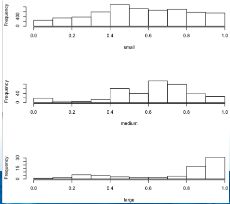
BLUE WATERS
SUSTAINED PETASCALE COMPUTING

NSF NCSA GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION CRAY

Impact of poor nid allocation - Consistency

- Two jobs (8,192 nodes) with nearly same nid allocation (s10_8972n). Red job affected by other workload communicating through the region.





- Run time variation - poor wallclock accuracy (padding wallclock).

Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS

SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Congestion Protection

- To avoid data loss, traffic injection is throttled for a period of time, when reaching a point where forward progress is stalling. Throttling is applied and removed until congestion is cleared.
- System monitors percentage of time that traffic trying to enter the network from the nodes and percentage of time network tiles are stalled.
- Fortunately not a common occurrence. It does happen, typically in bursts.
- Can happen with node-node (MPI, PGAS) or node-LNET (IO) traffic.
- Many-to-one and long-path patterns.
- Libraries and user can control node injection as a precaution.
- In CP reports, flit rates represent data arriving at the node from the interconnection network.

```
 Max
APID    Name                       Nodes   Flits/s   UID     Start      End
------------------------------------------------------------------------------------
2220460 Castro3d.Linux.             2048     31698   46466   16:00:45   19:41:40
2220462 Castro3d.Linux.             2048     81115   46466   16:01:05   19:37:03
2218386 namd2                       2000        --   43448   01:58:31   18:02:09
2220803 psolve                      2000     45732   47252   17:12:34   17:30:30
2218759 su3_rhmd_hisq_q             1536        --   12940   07:29:16
2219859 nwchem                      1000        --   32745   13:58:50   18:02:07
2220668 nwchem                      1000   4128749   32745   17:00:22   18:15:32
2219678 ks_spectrum_his              768        --   12940   11:30:04
2219512 namd2                        700        --   42864   10:35:55
…
…

============================================================
Top Bandwidth Applications
============================================================
0: apid 2218386 userid  43448 numnids  2000 apname           namd2 Kflits/sec: Total    3075
1: apid 2219859 userid  32745 numnids  1000 apname          nwchem Kflits/sec: Total    2743
2: apid 2220462 userid  46466 numnids  2048 apname      Castro3d.Linux. Kflits/sec: Total    2715
3: apid 2220460 userid  46466 numnids  2048 apname      Castro3d.Linux. Kflits/sec: Total    2691
4: apid 2219517 userid  42864 numnids   700 apname           namd2 Kflits/sec: Total    2271
5: apid 2219519 userid  42864 numnids   700 apname           namd2 Kflits/sec: Total    2073
6: apid 2218759 userid  12940 numnids  1536 apname      su3_rhmd_hisq_q Kflits/sec: Total    2071
7: apid 2219514 userid  42864 numnids   700 apname           namd2 Kflits/sec: Total    1762
8: apid 2220646 userid  12940 numnids   512 apname      ks_spectrum_his Kflits/sec: Total    1596
9: apid 2217219 userid  47296 numnids   500 apname          python Kflits/sec: Total    1389
…
============================================================
Congestion Candidate COMPUTE Nodes
============================================================
1: c17-0c1s0n1 (64051 flits/sec) (nid 18401; apid 2220473 userid 14394 numnids 32 apname numa_script.sh)
2: c9-0c0s1n0 (61950 flits/sec) (nid 23036; apid 2219894 userid 14394 numnids 32 apname numa_script.sh)
3: c10-1c0s3n2 (24438 flits/sec) (nid 5798; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)
4: c3-10c0s5n1 (24238 flits/sec) (nid 25867; apid 2219672 userid 35077 numnids 64 apname enzo.exe)
5: c12-1c0s2n2 (22544 flits/sec) (nid 8026; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)
6: c5-10c0s6n3 (20193 flits/sec) (nid 24813; apid 2219672 userid 35077 numnids 64 apname enzo.exe)
7: c12-1c0s2n0 (20161 flits/sec) (nid 8004; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)
8: c14-1c0s3n0 (19784 flits/sec) (nid 8120; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)
9: c10-1c0s2n1 (19273 flits/sec) (nid 5819; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)
10: c10-1c0s3n0 (17453 flits/sec) (nid 5816; apid 2219756 userid 14394 numnids 32 apname numa_script.sh)

============================================================
Top 100 Congestion Candidate Nodes (614 compute nodes: 134938785 flits/s, 590 service nodes: 1257373796 flits/
s)
============================================================
1: c20-10c0s3n0 4128749 flits/sec nid 12038; apid 2220668 userid 32745 numnids 1000 apname nwchem
2: c20-10c0s3n3 3396088 flits/sec nid 12057; apid 2220668 userid 32745 numnids 1000 apname nwchem
3: c21-11c1s1n2 3351520 flits/sec nid 15484; apid 2220668 userid 32745 numnids 1000 apname nwchem
4: c17-10c0s3n2 3233871 flits/sec nid 17894; apid 2220668 userid 32745 numnids 1000 apname nwchem
5: c21-11c1s1n3 2912123 flits/sec nid 15485; apid 2220668 userid 32745 numnids 1000 apname nwchem
6: c20-10c1s1n3 2739003 flits/sec nid 12067; apid 2220668 userid 32745 numnids 1000 apname nwchem
7: c20-10c1s1n2 2727704 flits/sec nid 12066; apid 2220668 userid 32745 numnids 1000 apname nwchem
8: c21-11c1s2n0 2629574 flits/sec nid 15524; apid 2220668 userid 32745 numnids 1000 apname nwchem
9: c15-11c1s4n0 2619990 flits/sec nid 19030; apid 2220668 userid 32745 numnids 1000 apname nwchem
10: c21-11c1s2n3 2604278 flits/sec nid 15483; apid 2220668 userid 32745 numnids 1000 apname nwchem
```


Courtesy of Greg Bauer (JLPC'13)

BLUE WATERS
SUSTAINED PETASCALE COMPUTING

GREAT LAKES CONSORTIUM FOR PETASCALE COMPUTATION

Congestion Protection Analysis

- Look at application to node relation.
- wrf listed as top application and the top 10 nodes are wrf nodes.
- nwchem running at same time (listed #4).
- The OVIS state of the network data should help here.