HPC Performance and Energy Efficiency

Overview and Trends

June 9th, 2015
SMAI 2015 Congress
Les Karellis (Savoie)

Dr. Sébastien Varrette
Parallel Computing and Optimization Group (PCOG)

http://hpc.uni.lu

Outline

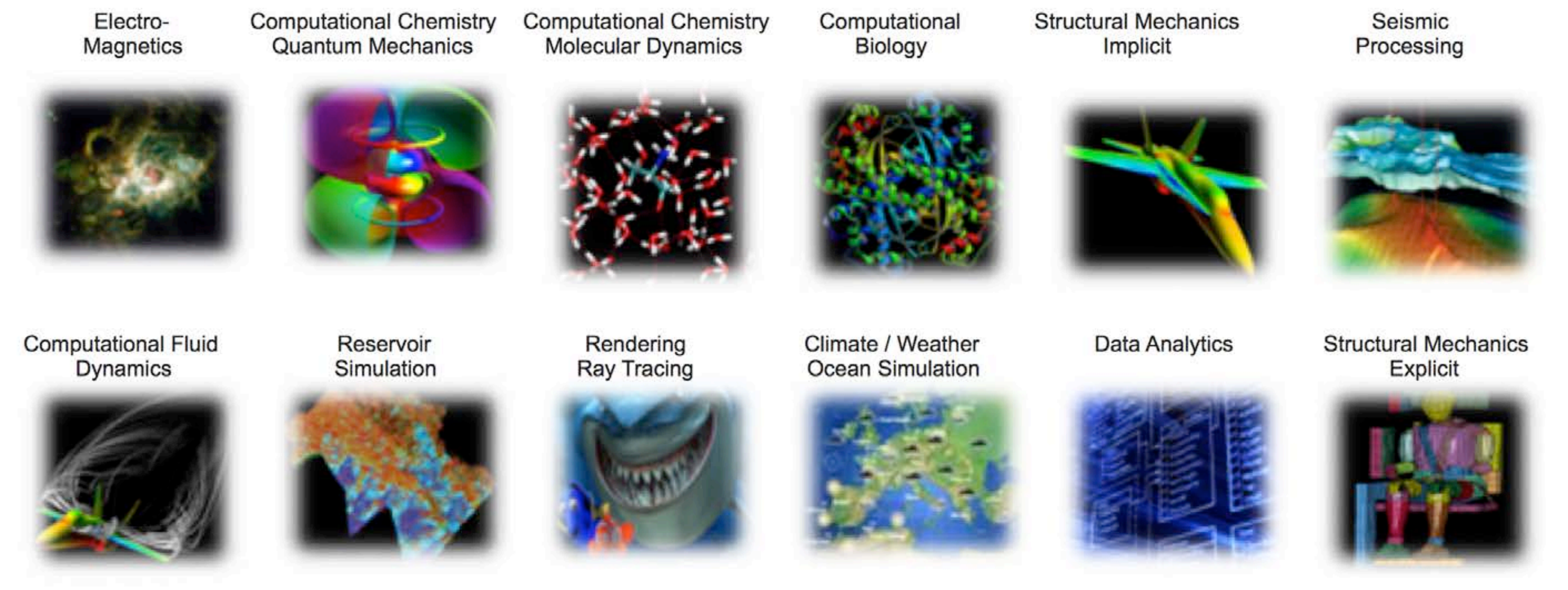
- Introduction & Context
- HPC Data-Center Trends: Time for DLC
- HPC [Co-]Processor Trends: Go Mobile
- Middleware Trends: Virtualization, RJMS
- Software Trends: Rethinking Parallel Computing
- Conclusion

Introduction and Context

UNIVERSITÉ DU LUXEMBOURG

HPC at the Heart of our Daily Life

- **Today...** R&D, ***Academia***, Industry, Local Collectivities



- … **Tomorrow**: digital health, nano/bio techno…

Performance Evaluation of HPC Systems

- Commonly used metrics
  - ✓ FLOPs: raw compute capability
  - ✓ GUPS: memory performance
  - ✓ IOPS: storage performance
  - ✓ bandwidth & latency: memory operations or network transfer
- Energy Efficiency
  - ✓ Power Usage Effectiveness (PUE) in HPC data-centers
    - Total Facility Energy / Total IT Energy
  - ✓ Average system power consumption during execution (W)
  - ✓ Performance-per-Watt (PpW)

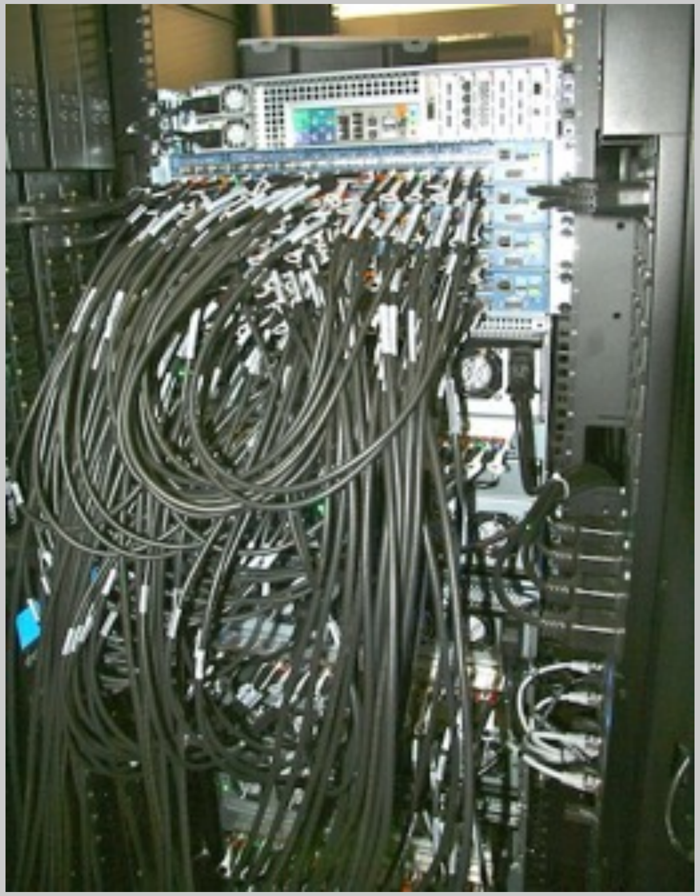
Ex (in Academia): The UL HPC Platform

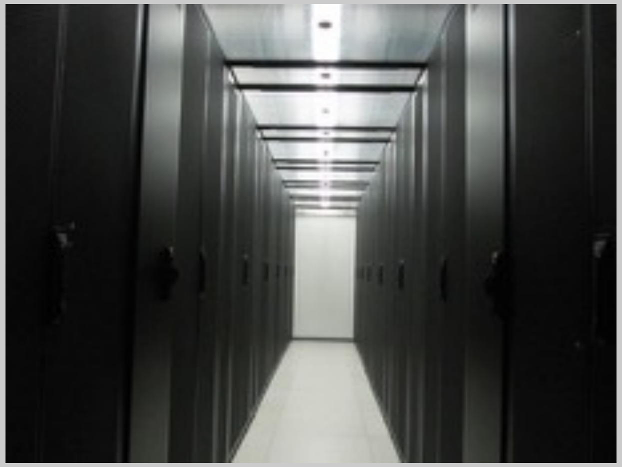
`http://hpc.uni.lu`

- 2 geographical sites, 3 server rooms
- 4 clusters, ~281 users
  - ✓ **404** nodes, **4316 cores** (**49.92 TFlops**)
  - ✓ Cumul. *shared* raw storage: **3,13 PB**
  - ✓ Around **197 kW**
- > **6,21 M€** HW investment so far
- Mainly **Intel**-based architecture
- Mainly Open-Source software stack
  - ✓ Debian, SSH, OpenLDAP, Puppet, FAI...

Ex (in Academia): The UL HPC Platform

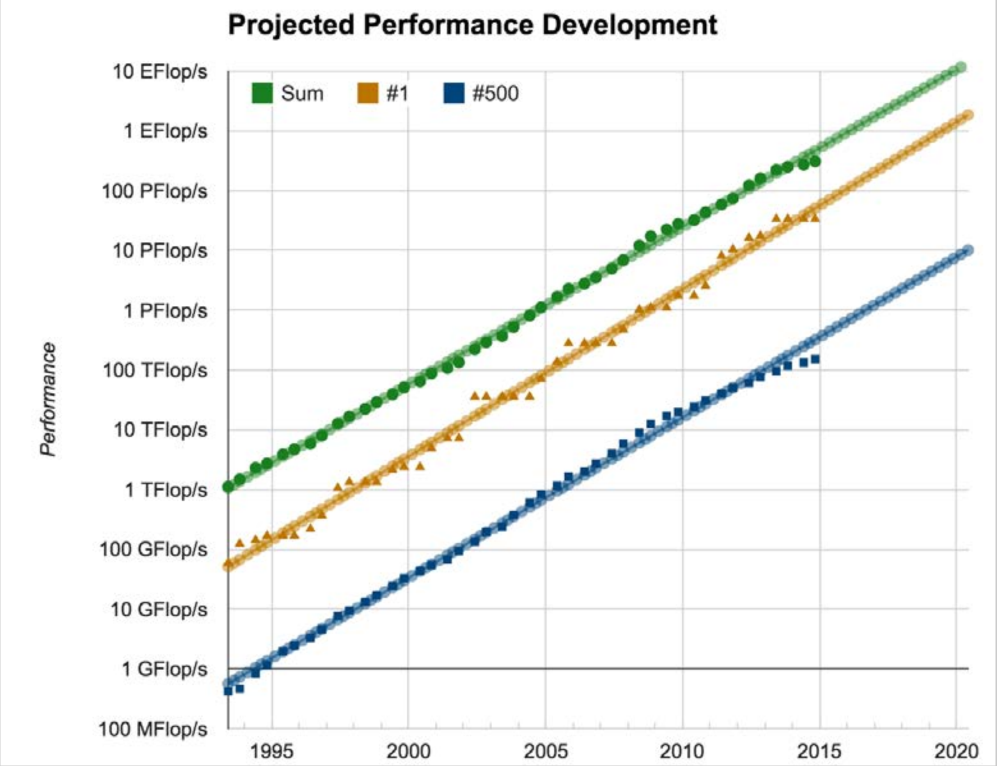
http://hpc.uni.lu





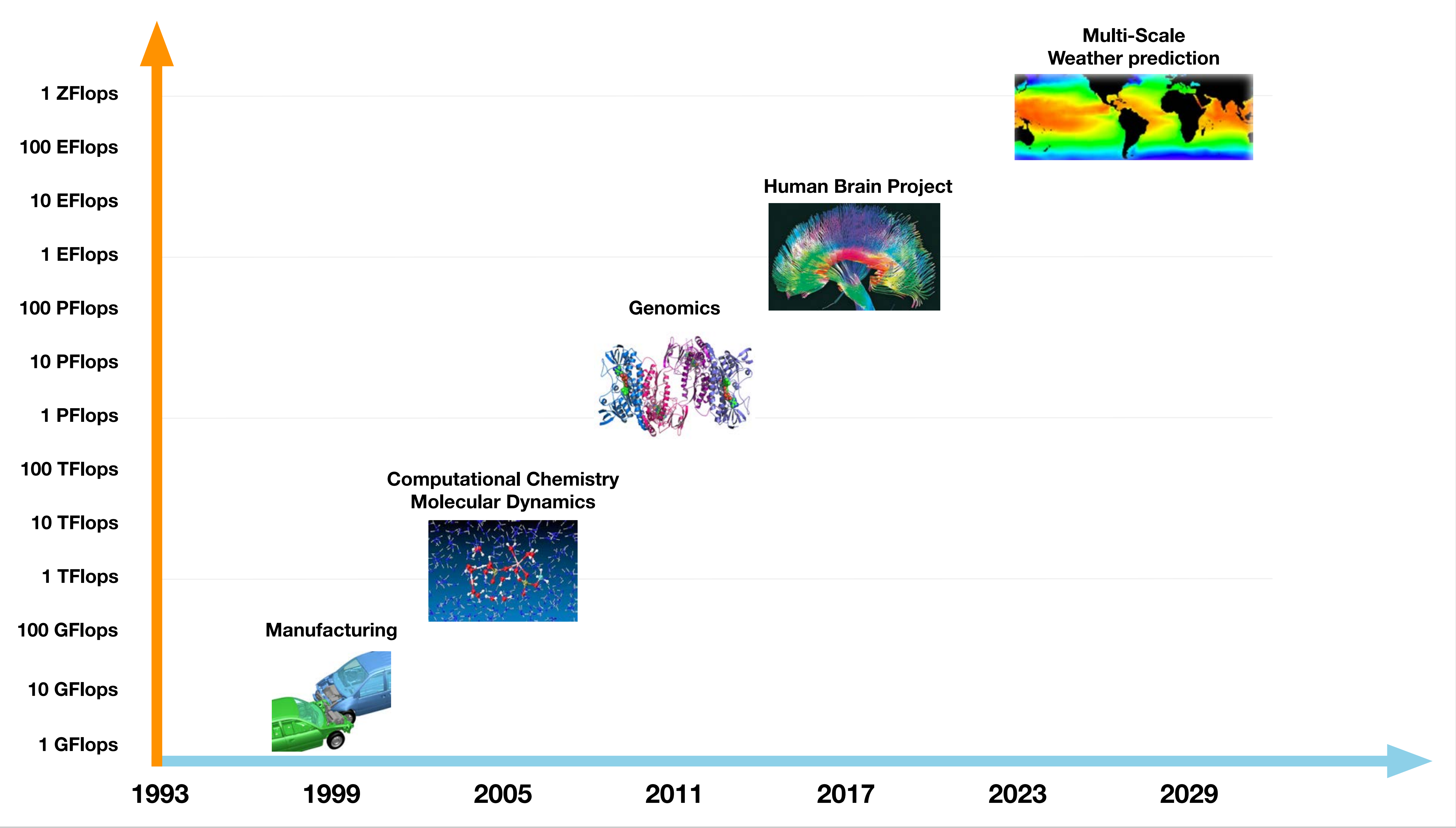
General HPC Trends

- Top500: world's 500 most powerful computers (since 1993)
  - ✓ Based on High-Performance LINPACK (HPL) benchmark
  - ✓ Last list [Nov. 2014]
    - #1: Tianhe-2 (China): 3,120,000 cores
      - **33.863 PFlops… and 17.8 MW**
    - Total combined performance:
      - 309 PFlops
      - 215.744 MW over 258 systems (which provided power information)

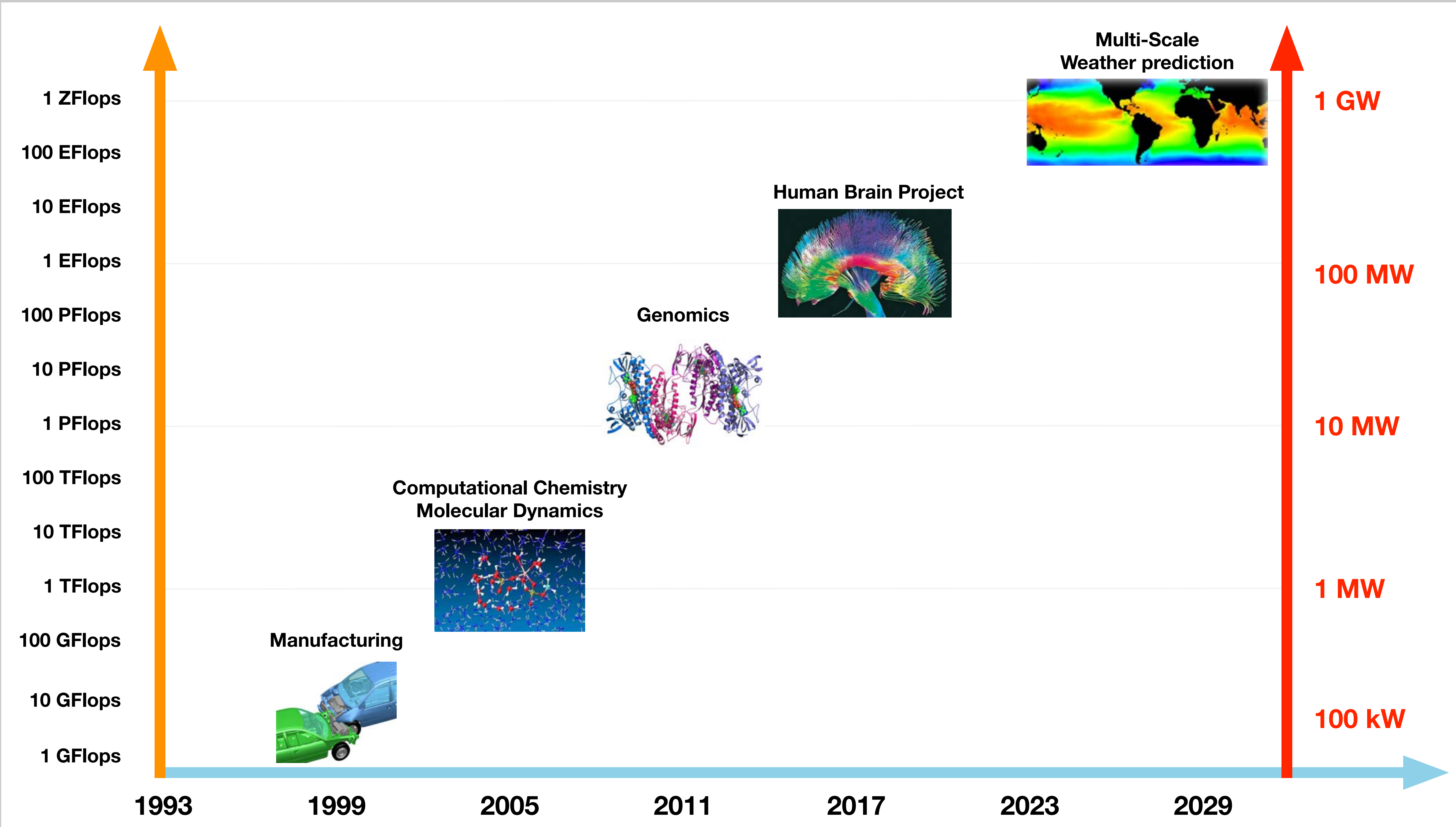


- Green500: Derive PpW metric from Top500 (MFlops/W)
  - ✓ #1: L-CSC GPU Cluster (#168): 5.27 GFlops/W
- Other Benchmarks: HPC{C,G}, Graph500…

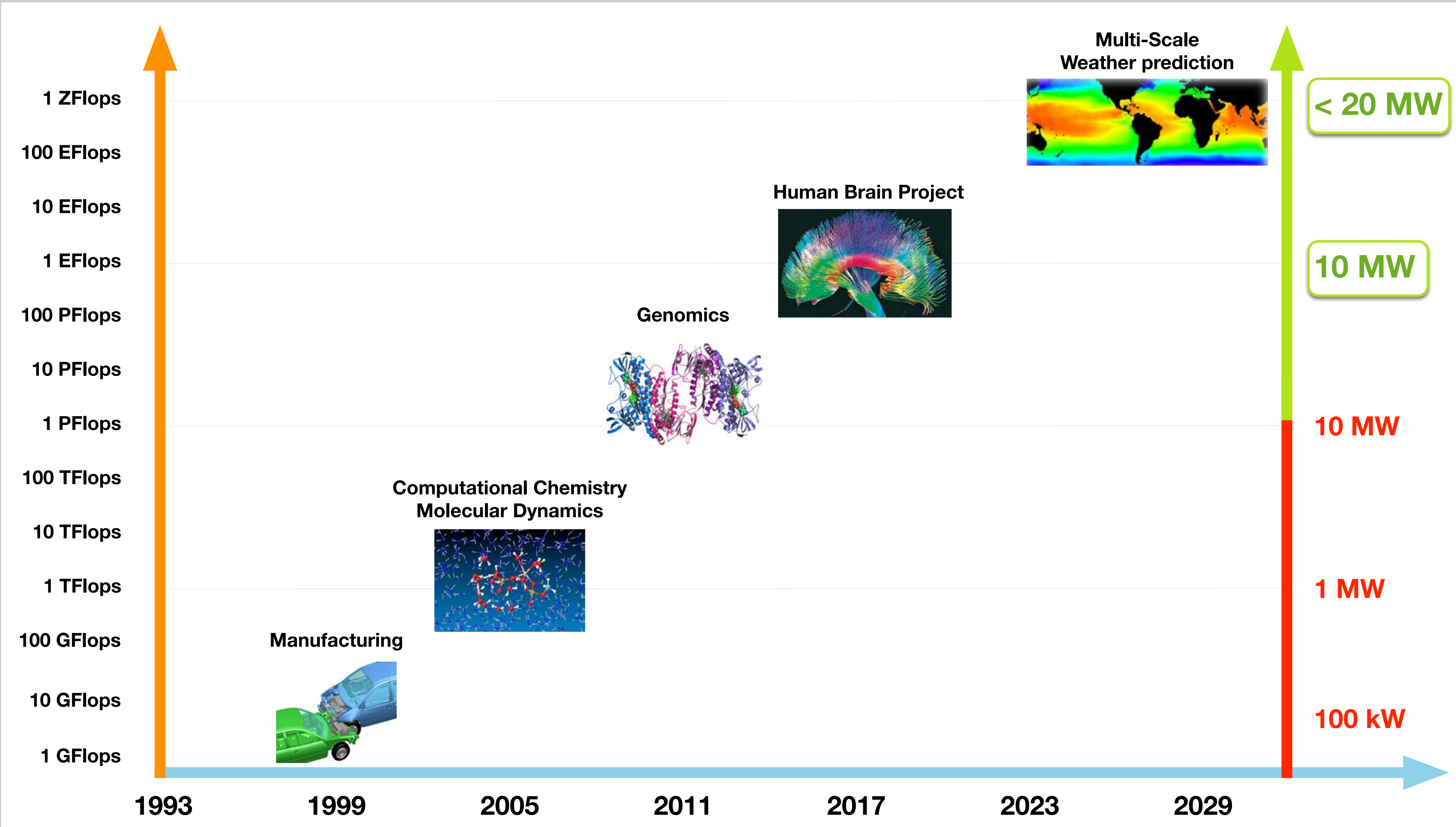
Computing Needs Evolution



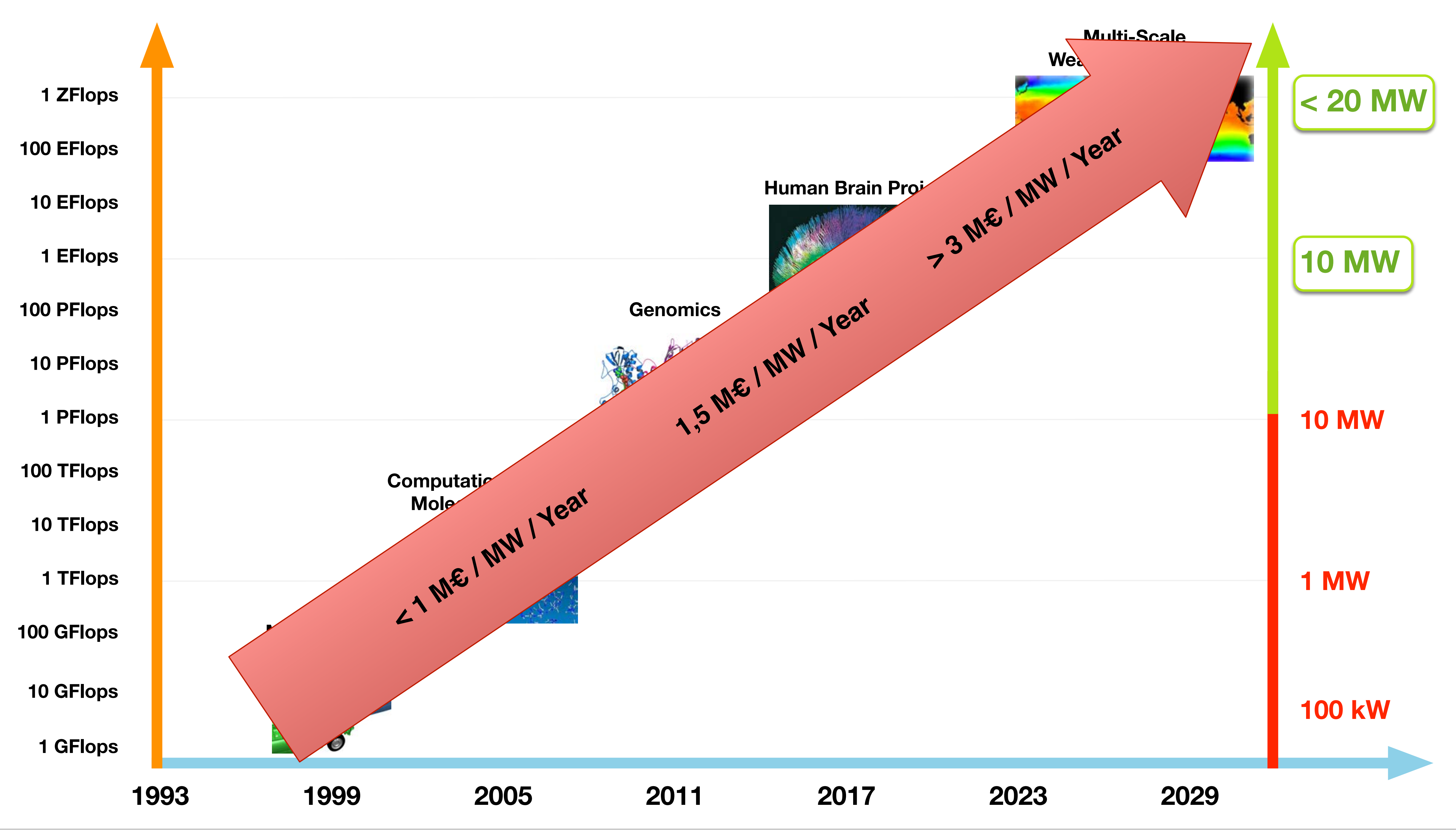
~~Computing~~ Power Needs Evolution



~~Computing~~ Less Power Needs Evolution



The Budgetary Wall



Energy Optimization paths toward Exascale

- H2020 Exascale Challenge: 1 EFlops in 20 MW
  - ✓ Using today's most energy efficient TOP500 system: 189MW

new [co-]processors, interconnect…

Hardware

Virtualization, RJMS…

Middleware

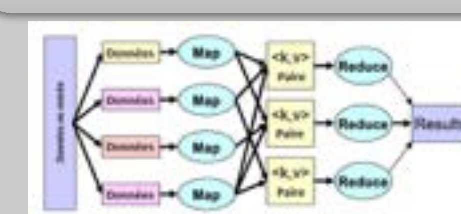
New programming/execution models

Data-center

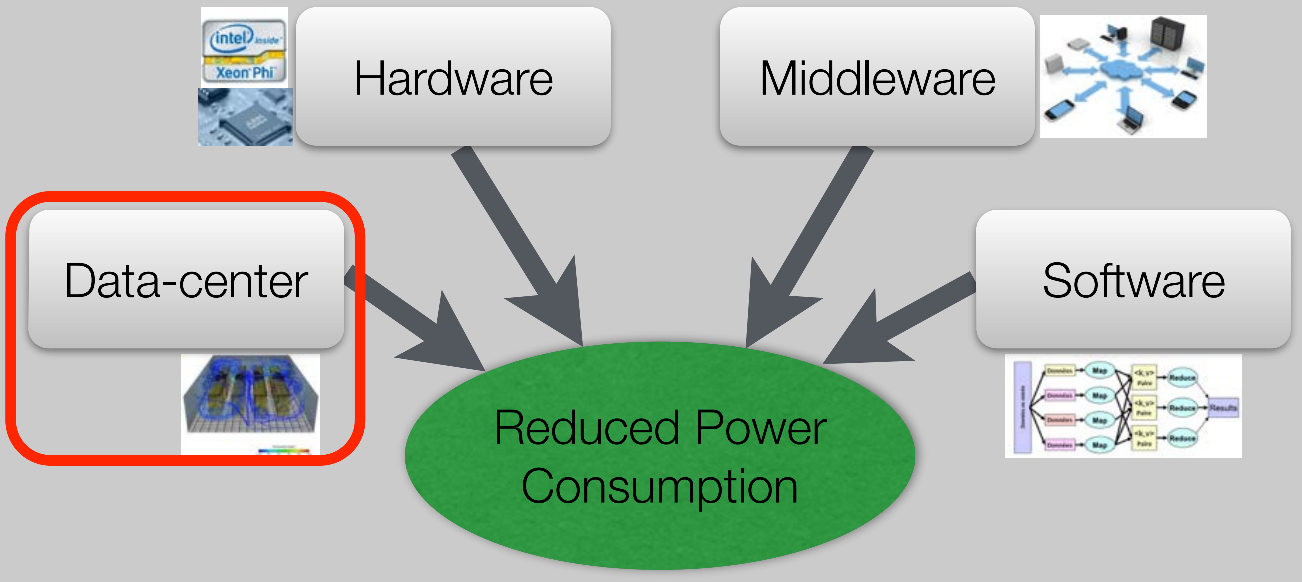
Software

PUE optim.
DLC…

Reduced Power Consumption

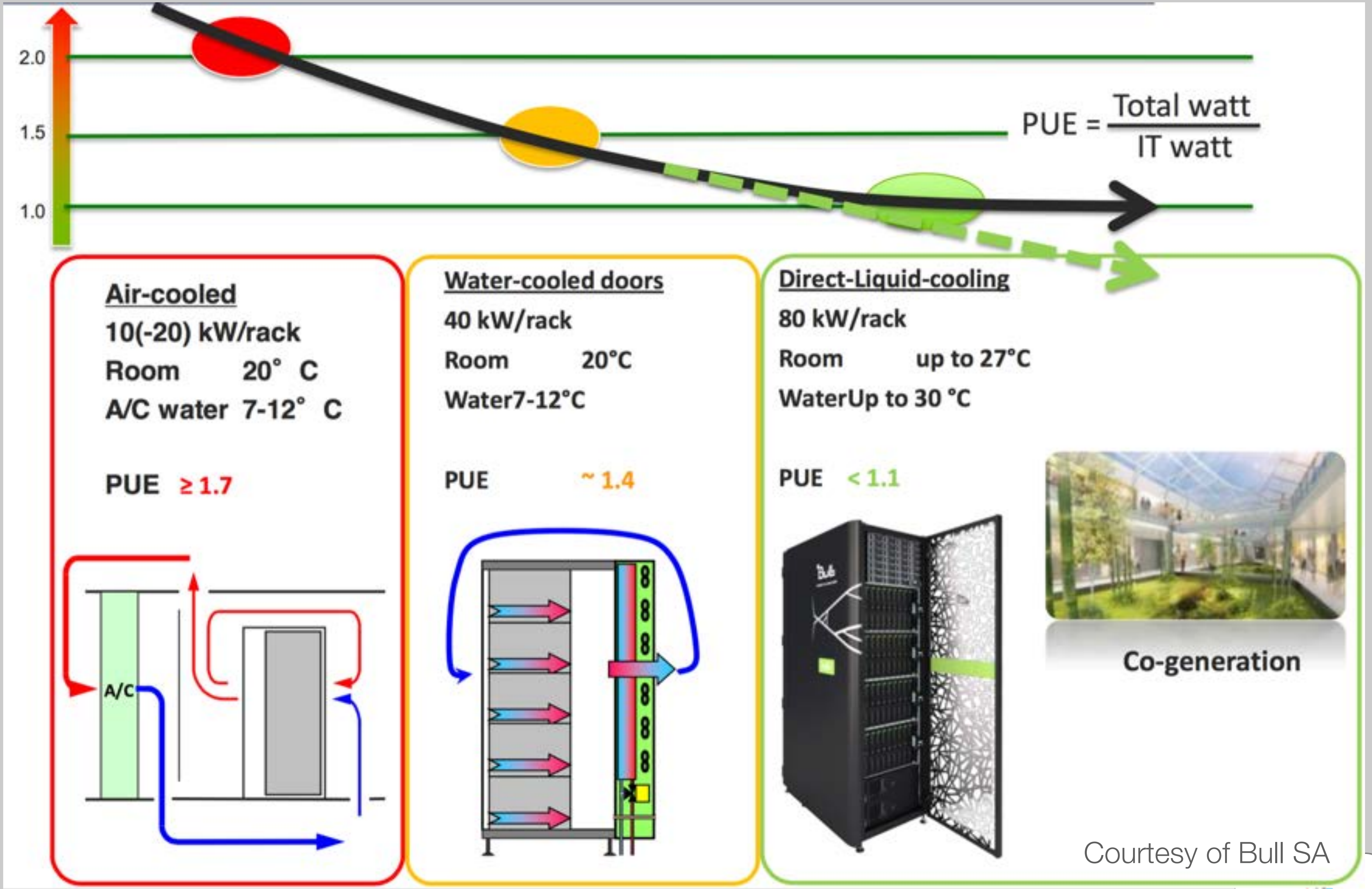


HPC Data-Center Trends: Time for DLC



UNIVERSITÉ DU LUXEMBOURG

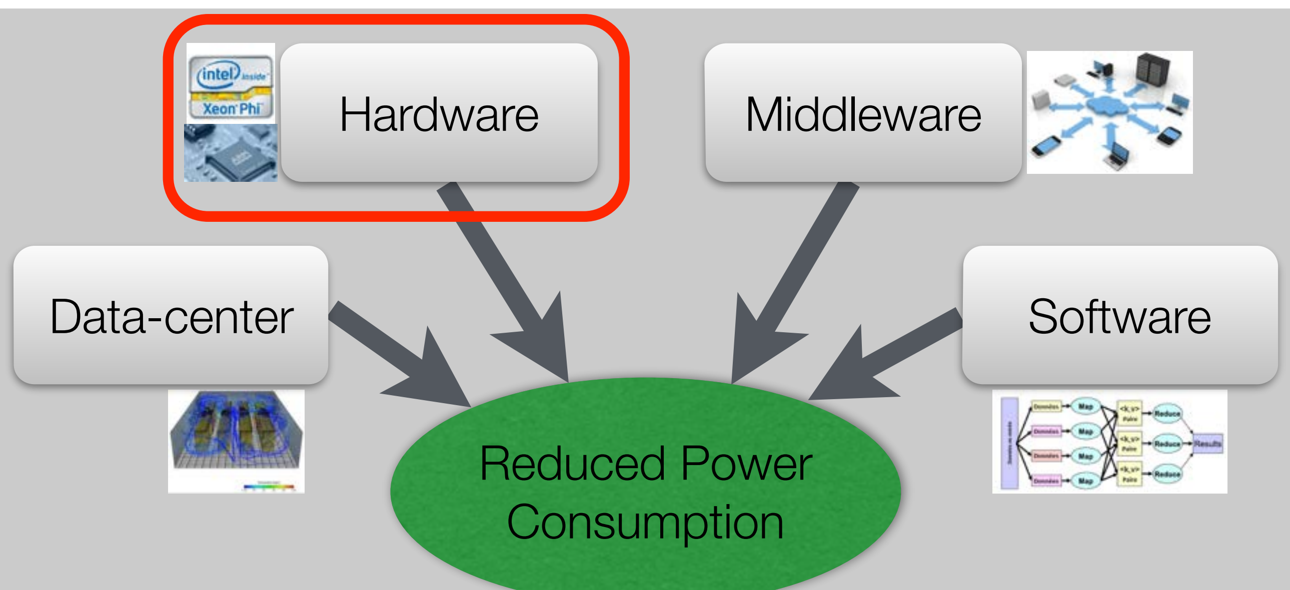
Cooling and PUE



Cooling and PUE

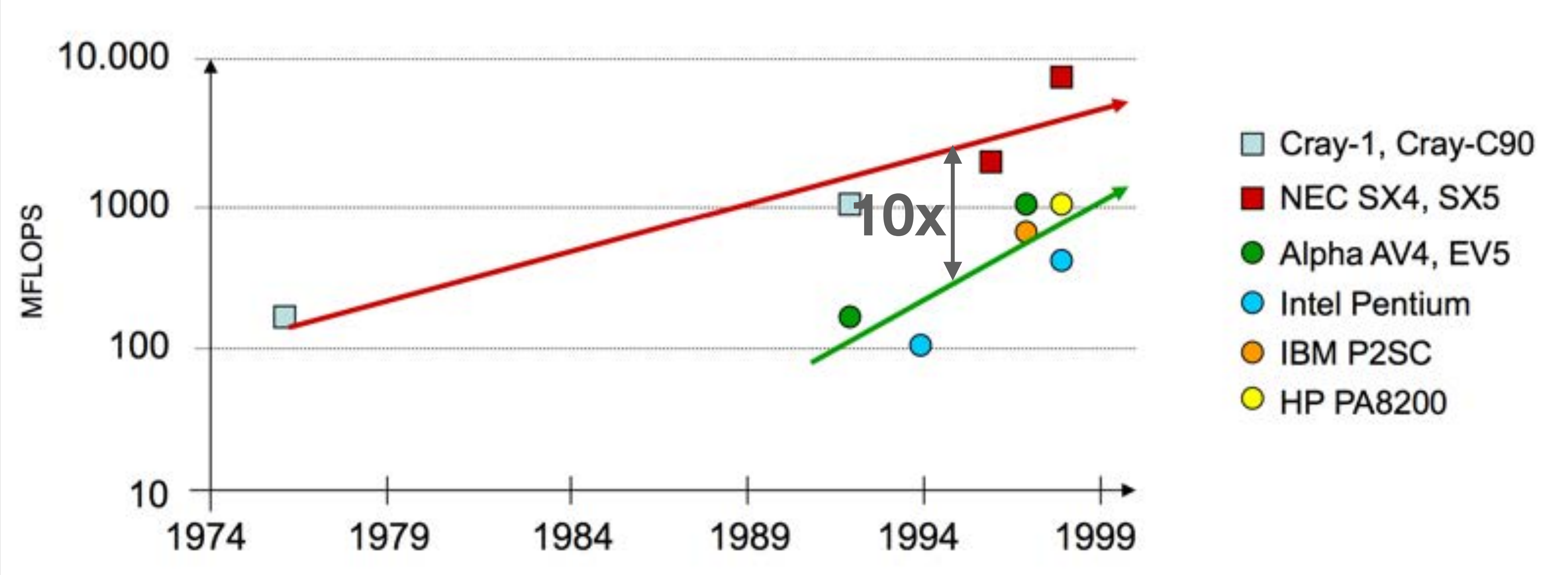
- Direct immersion: the CarnotJet example (PUE: 1.05)

HPC [Co-]Processor Trends: Go Mobile



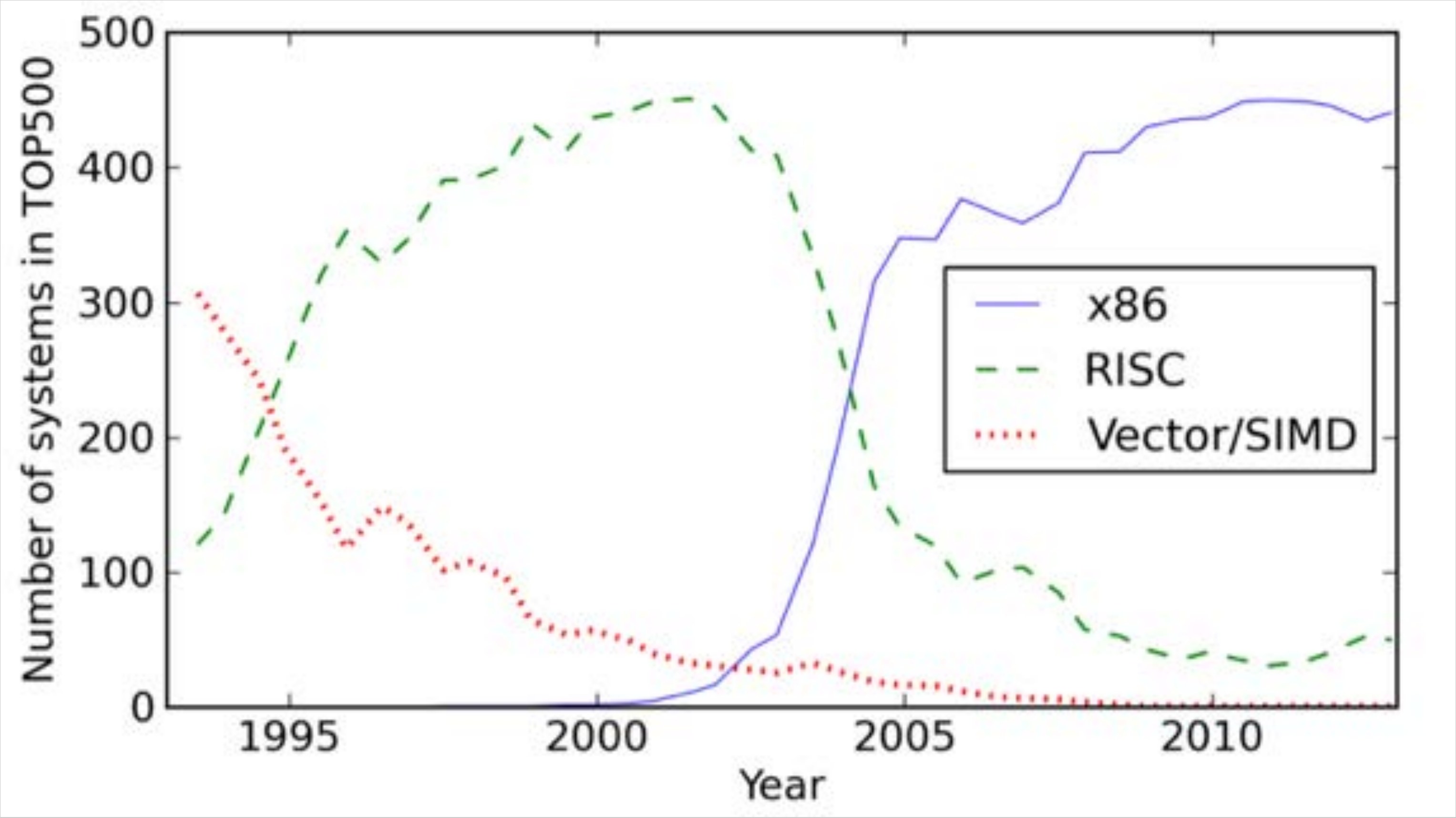
Back to 1995: vector vs. micro-processor

- Microprocessors ~10x slower than one vector CPU
  - ✓ … thus not faster… But cheaper!



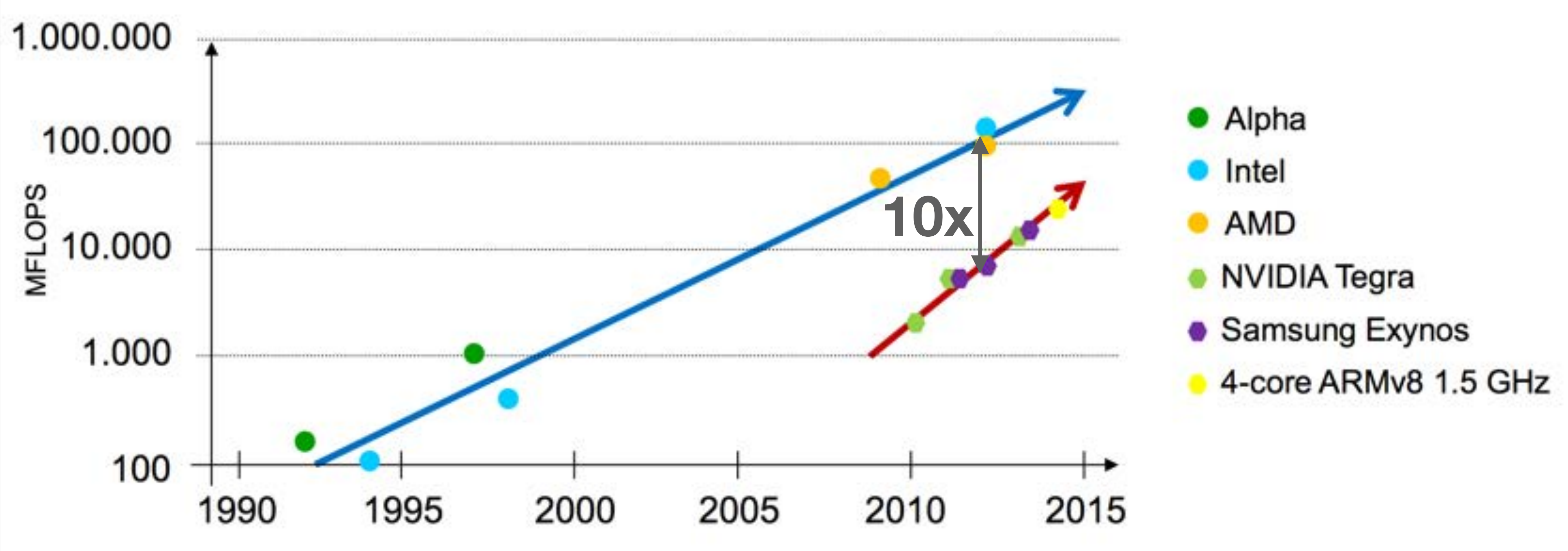
Back to 1995: vector vs. micro-processor

- Microprocessors ~10x slower than one vector CPU
  - ✓ … thus not faster… But cheaper!



How about now?

- Mobile SoCs ~10x slower than one microprocessor
  - ✓ … thus not faster… But cheaper!

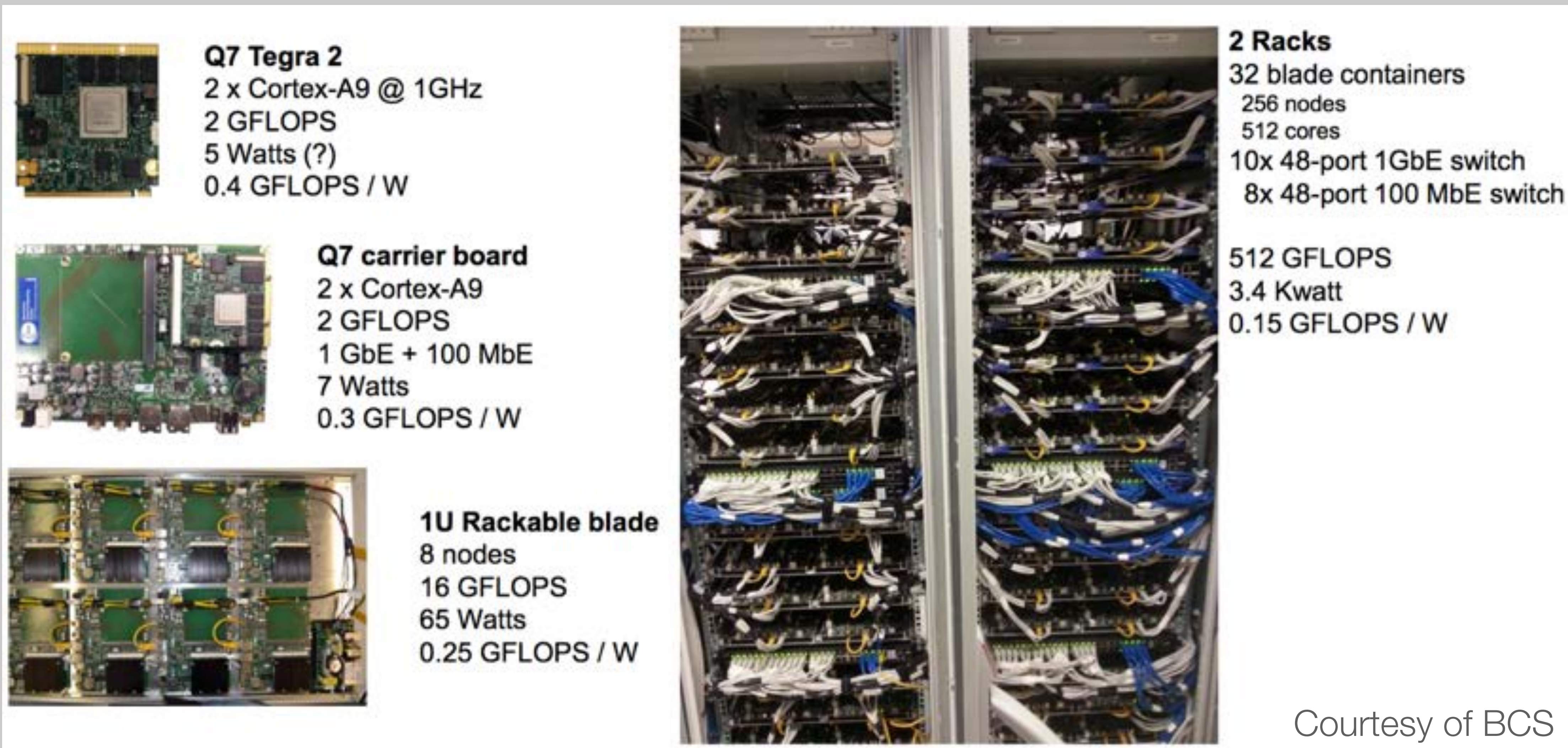


  - ✓ the “already seen” pattern?

- Mont-Blanc project: build an HPC system from embedded and mobile devices

Mont-Blanc (Phase 1) project outcomes

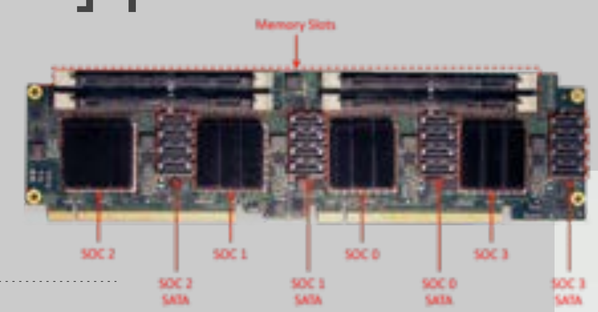
- (2013) Tiribado: the first ARM HPC multicore system

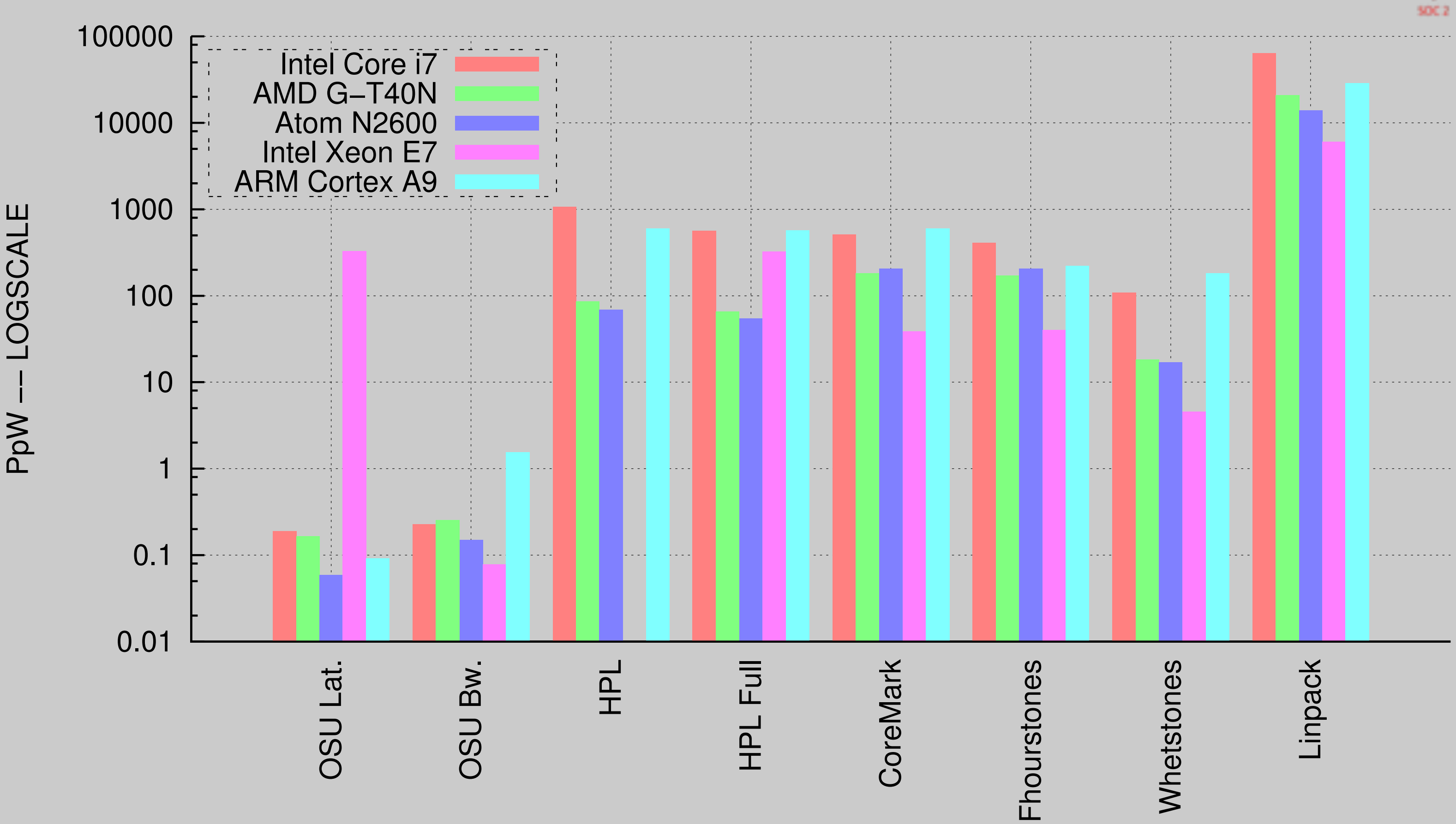


0,15 GFlops/W

The UL HPC `viridis` cluster (2013)

- 2 encl. (96 nodes, 4U), 12 calxeda boards per enclosure
  - ✓ 4x ARM Cortex A9 @ 1.1 GHz [4C] per Calxeda board
    - 2x300W, “10” GbE inter-connect



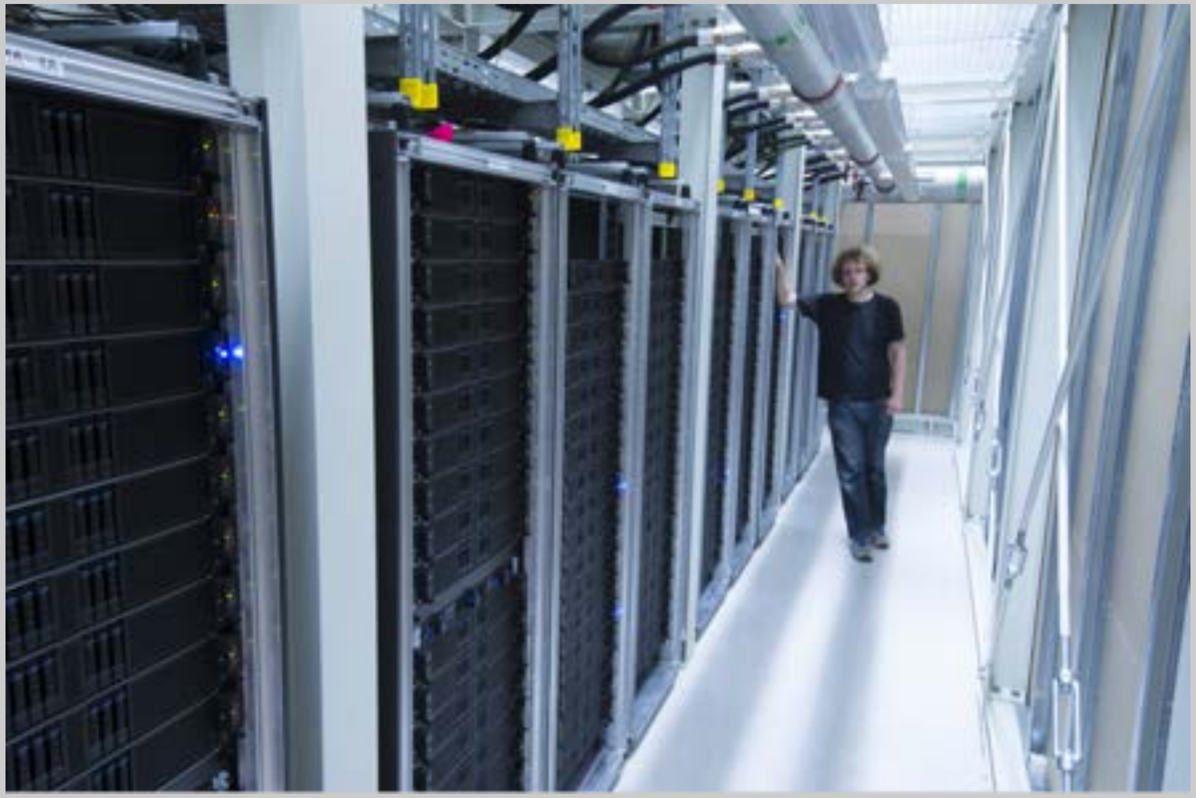


0,513 GFlops/W

[EE-LSDS’13] M. Jarus, S. Varrette, A. Oleksiak, and P. Bouvry. Performance Evaluation and Energy Efficiency of High-Density HPC Platforms Based on Intel, AMD and ARM Processors. In Proc. of the Intl. Conf. on Energy Efficiency in Large Scale Distributed Systems (EE-LSDS’13), volume 8046 of LNCS, Vienna, Austria, Apr 2013.

Commodity vs. GPGPUs: L-CSC (2014)

- The German L-CSC cluster (Frankfurt) (2014)
- Nov 2014: 56 (out of 160) nodes, on each:
  - ✓ 4 GPUs, 2 CPUs, 256 GB RAM
  - ✓ #168 on Top 500 (1.7 PFlops)
  - ✓ #1 on Green 500



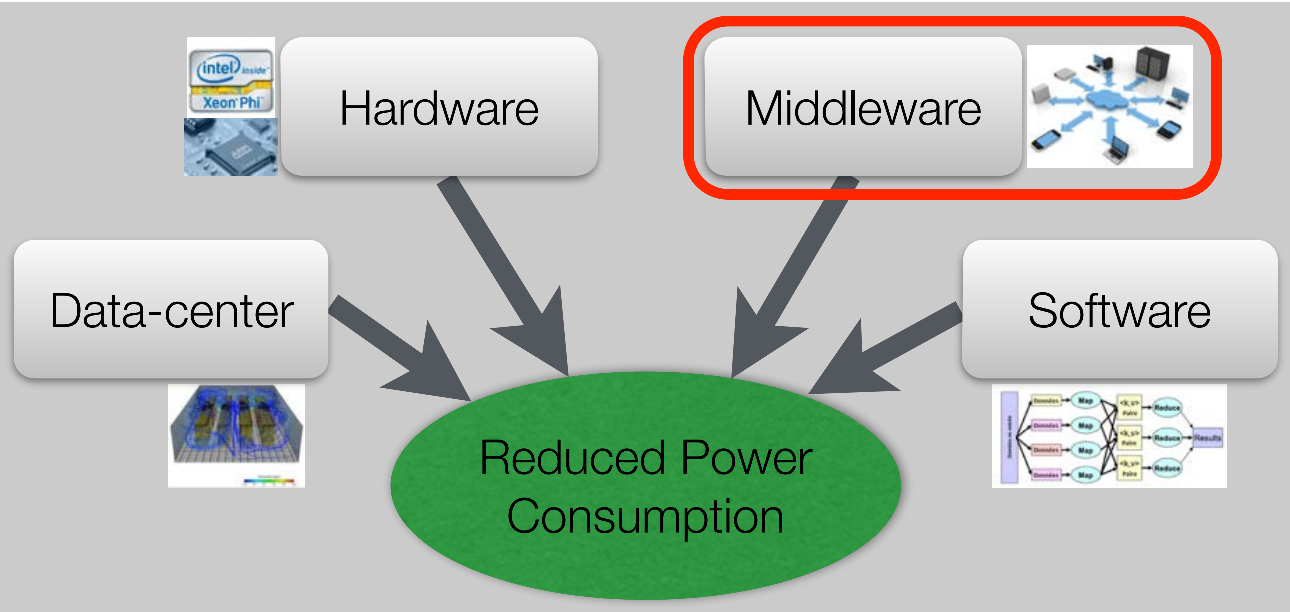
5,27 GFlops/W

Mobile SoCs and GPGPUs in HPC

- Very fast development for Mobile SoCs and **GP**GPUs
- Convergence between both is foreseen
  - ✓ CPUs inherits from GPUs multi-core with vector inst.
  - ✓ GPUs inherits from CPUs cache-hierarchy
- In parallel: large innovation in other embedded devices
  - ✓ Intel Xeon Phi co-processor
  - ✓ FPGAs etc.

Objective: 50 GFlops/W

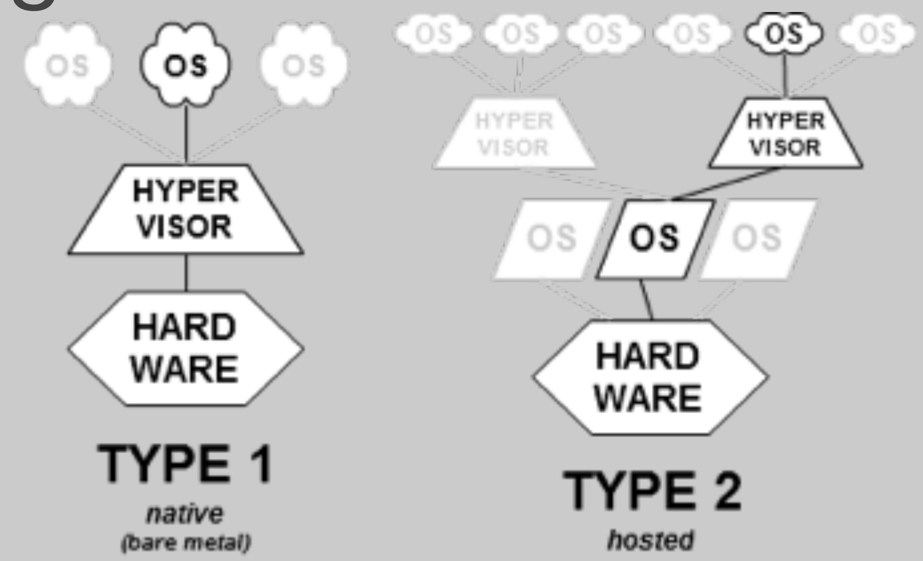
Middleware Trends: Virtualization, RJMS



UNIVERSITÉ DU LUXEMBOURG

Virtualization in an HPC Environment

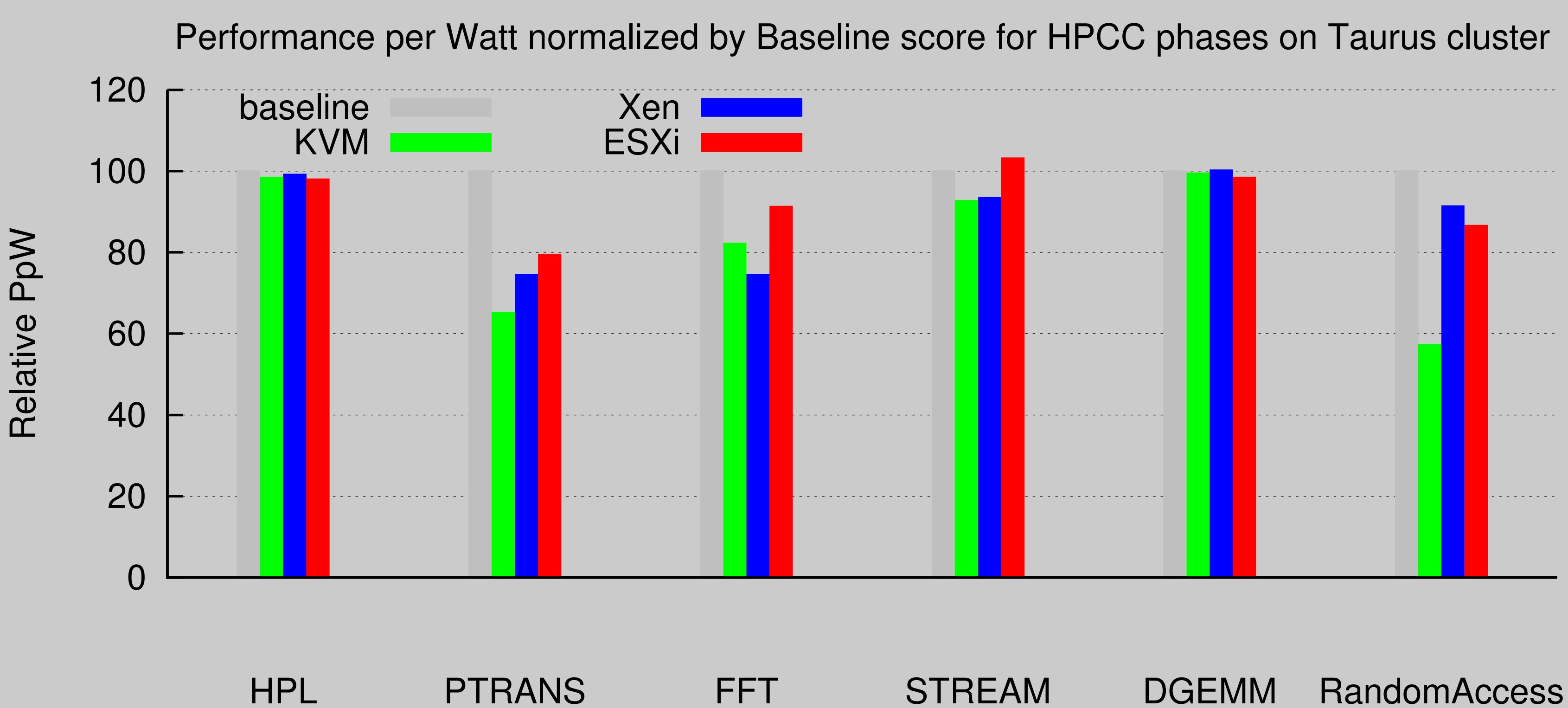
- **Hypervisor:** Core virtualization engine / environment
  - ✓ Type 1 adapted to HPC workload
  - ✓ *Performance Loss:* > 20%

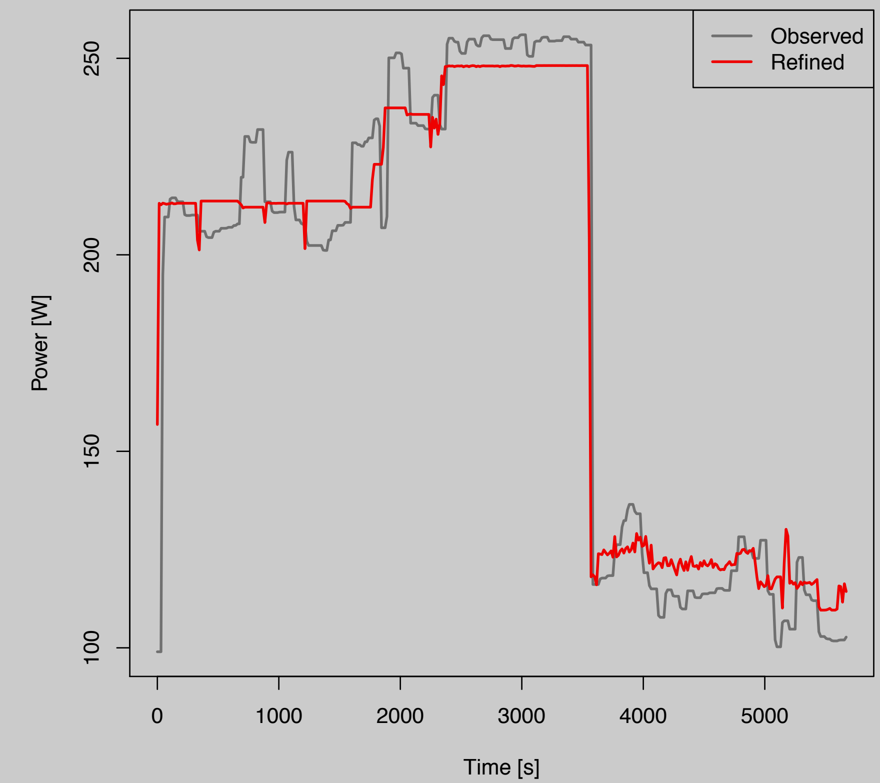


Xen, VMWare (ESXi), KVM   Virtualbox

Virtualization in an HPC Environment

- **Hypervisor:** Core virtualization engine / environment
  - ✓ Type 1 adapted to HPC workload
  - ✓ *Performance Loss:* > 20%

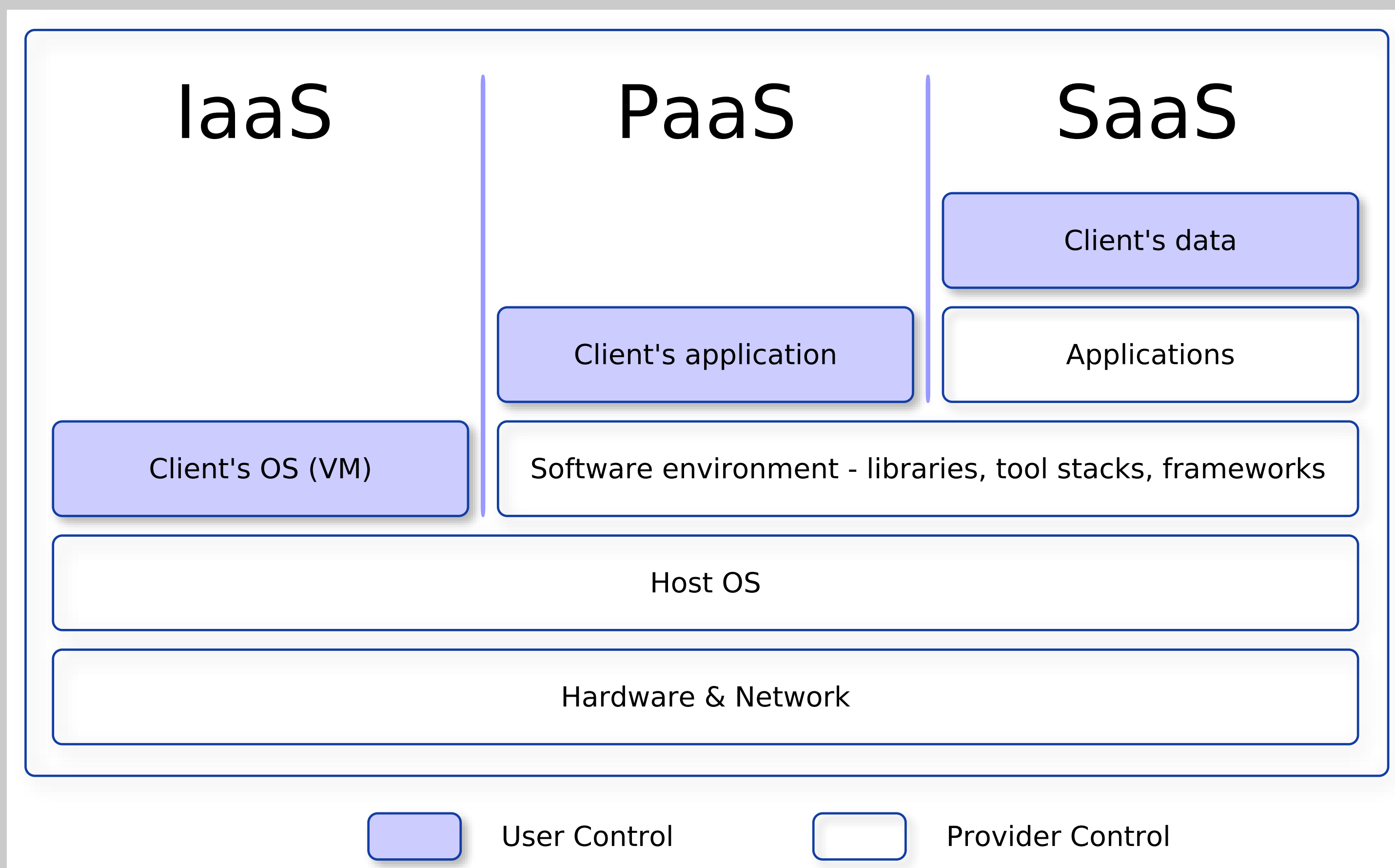




[CCPE'14] M. Guzek, S. Varrette, V. Plugaru, J. E. Pecero, and P. Bouvry. ***A Holistic Model of the Performance and the Energy-Efficiency of Hypervisors in an HPC Environment***.
Intl. J. on Concurrency and Computation: Practice and Experience (CCPE), 26(15):2569–2590, Oct. 2014.

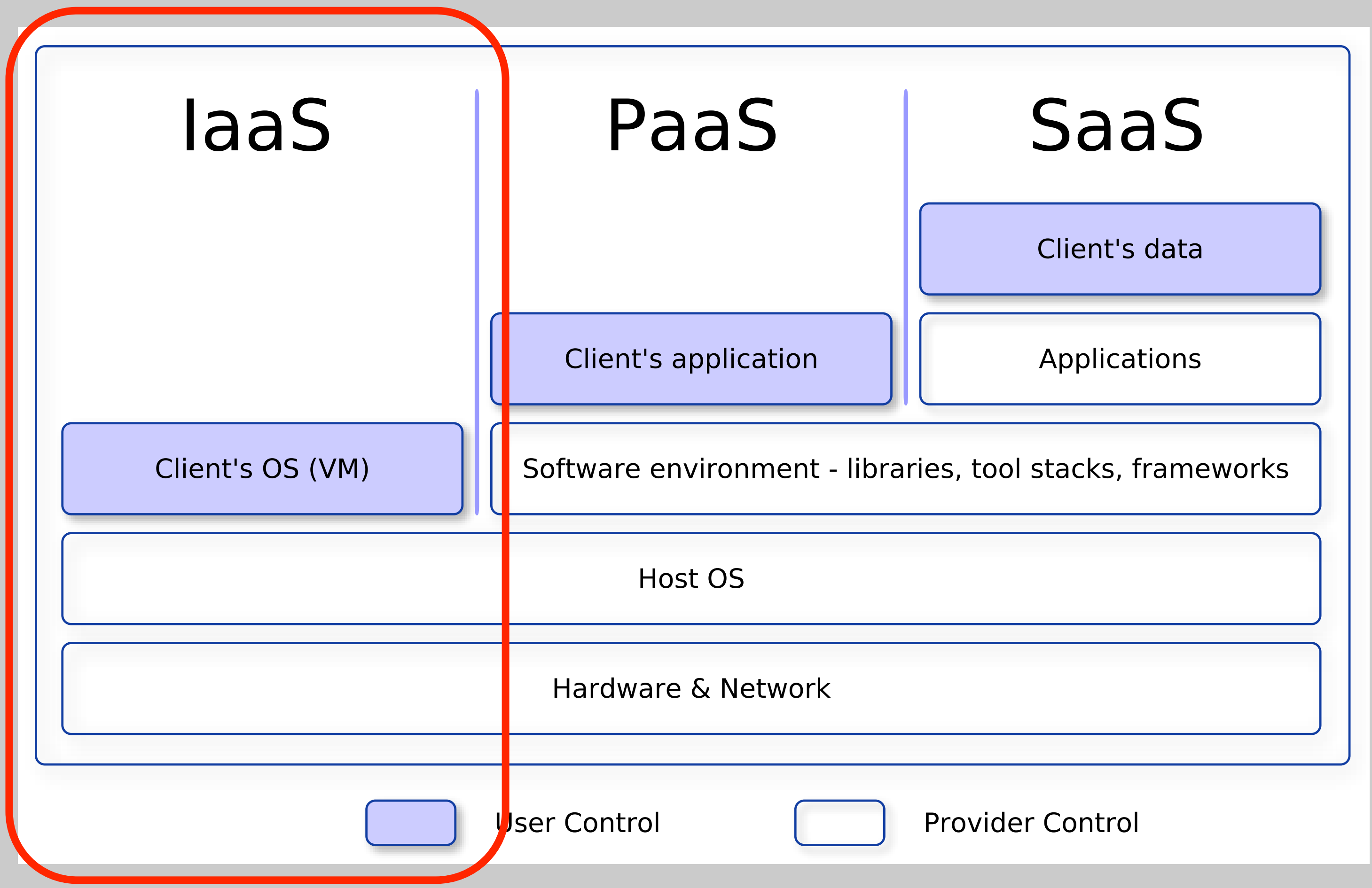
Cloud Computing vs. HPC

- World-widely advertised as THE solution to all problems
- Classical taxonomy:
  - ✓ {Infrastructure,Platform,Software}-as-a-Service
  - ✓ Grid'5000: Hardware-as-a-Service



Cloud Computing vs. HPC

- World-widely advertised as THE solution to all problems
- Classical taxonomy:
  - ✓ {Infrastructure,Platform,Software}-as-a-Service
  - ✓ Grid'5000: Hardware-as-a-Service



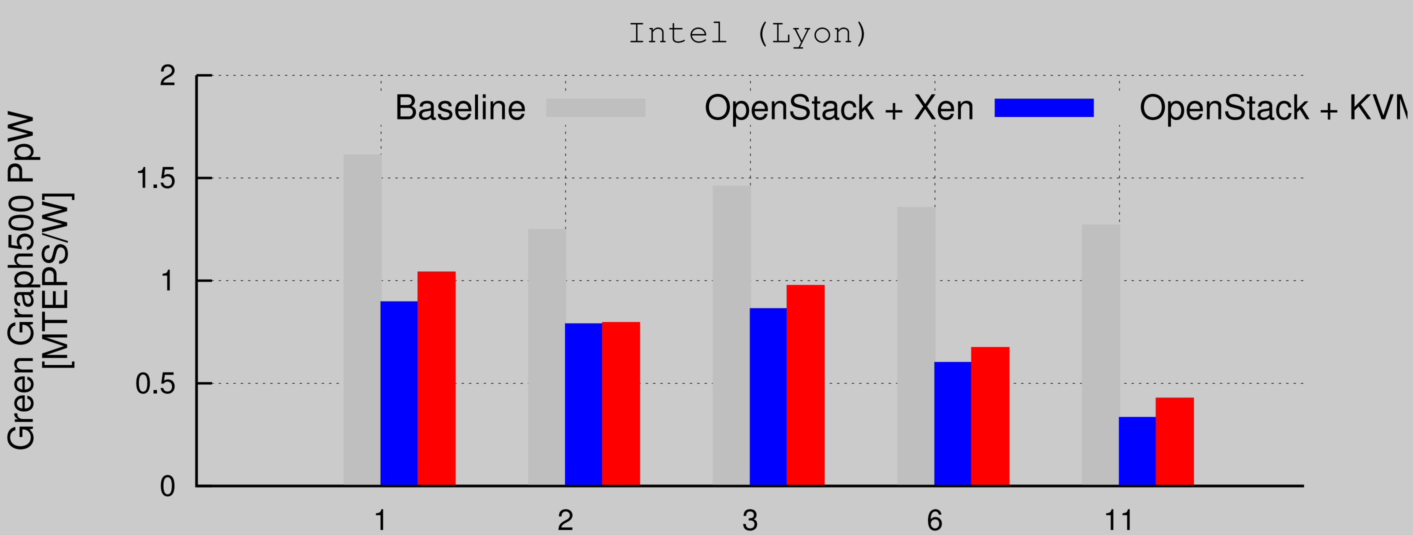
Cloud Middleware for HPC Workload

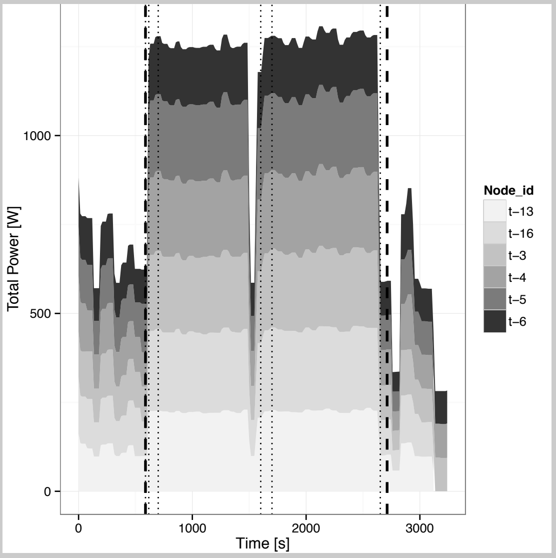
Middleware:	vCloud	Eucalyptus	OpenNebula	OpenStack	Nimbus
License	Proprietary	BSD License	Apache 2.0	Apache 2.0	Apache 2.0
Supported Hypervisor	VMWare/ESX	Xen, KVM, VMWare	Xen, KVM, VMWare	Xen, KVM, Linux Containers, VMWare/ESX, Hyper-V,QEMU, UML	Xen, KVM
Last Version	5.5.0	3.4	4.4	8 (Havana)	2.10.1
Programming Language	n/a	Java / C	Ruby	Python	Java / Python
Host OS	VMX server	RHEL 5, ESX Debian, Fedora, CentOS 5, openSUSE-11	RHEL 5, Debian, Fedora, CentOS 5,openSUSE-11	Ubuntu, ESX Debian, RHEL, SUSE, Fedora	Ubuntu, Debian, RHEL, SUSE, Fedora
Guest OS	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris
Contributors	VMWare	Eucalyptus systems, Community	C12G Labs, Community	Rackspace, IBM, HP, Red Hat, SUSE, Intel, AT&T, Canonical, Nebula, others	Community

Cloud Middleware for HPC Workload

Middleware:	vCloud	Eucalyptus	OpenNebula	OpenStack	Nimbus
License	Proprietary	BSD License	Apache 2.0	Apache 2.0	Apache 2.0
Supported Hypervisor	VMWare/ESX	Xen, KVM, VMWare	Xen, KVM, VMWare	Xen, KVM, Linux Containers, VMWare/ESX, Hyper-V,QEMU, UML	Xen, KVM
Last Version	5.5.0	3.4	4.4	8 (Havana)	2.10.1
Programming Language	n/a	Java / C	Ruby	Python	Java / Python
Host OS	VMX server	RHEL 5, ESX Debian, Fedora, CentOS 5, openSUSE-11	RHEL 5, Debian, Fedora, CentOS 5,openSUSE-11	Ubuntu, ESX Debian, RHEL, SUSE, Fedora	Ubuntu, Debian, RHEL, SUSE, Fedora
Guest OS	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris	Windows (S2008,7), openSUSE,Debian,Solaris
Contributors	VMWare	Eucalyptus systems, Community	C12G Labs, Community	Rackspace, IBM, HP, Red Hat, SUSE, Intel, AT&T, Canonical, Nebula, others	Community

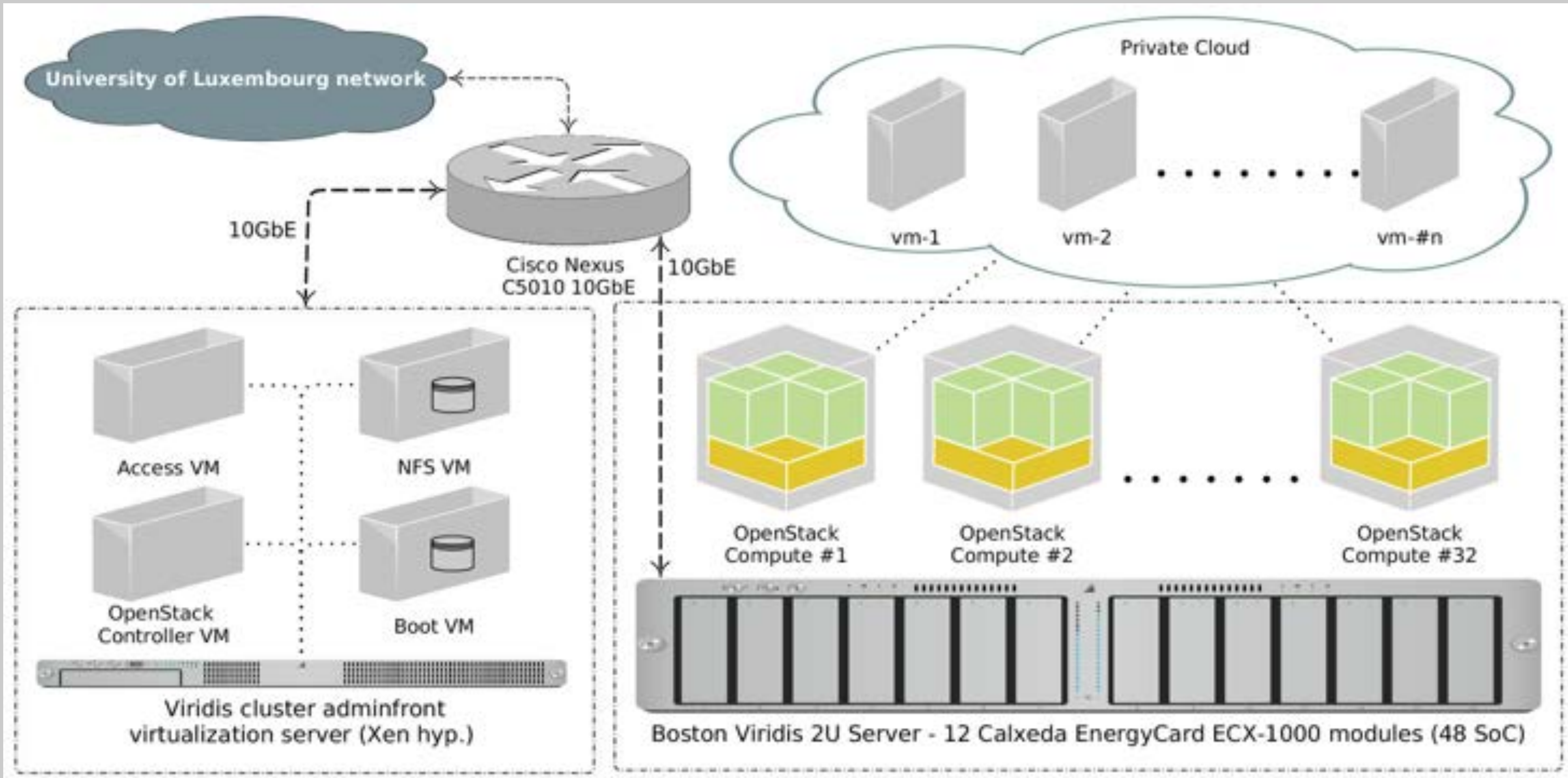
	Avg. Performance drop				Avg. Energy-efficiency drop	
	HPL	**STREAM**	**RandomAccess**	**Graph500**	**Green500**	**GreenGraph500**
OpenStack+Xen	41.5%	19%	89.7%	21.6%	56.5%	42%
OpenStack+KVM	58.6%	7.2%	67.5%	23.7%	38.5%	40%





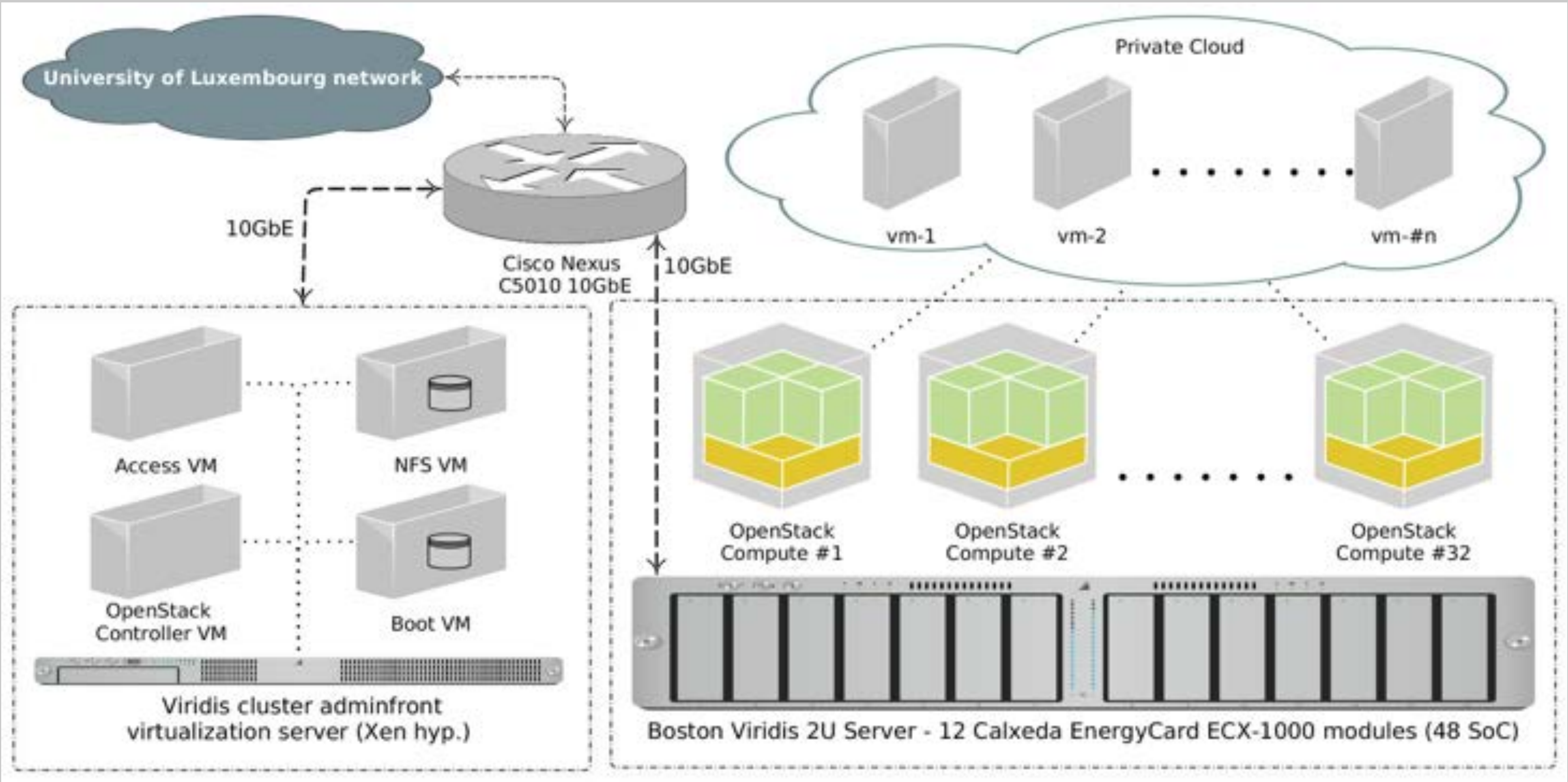
[ICPP'14] S. Varrette, V. Plugaru, M. Guzek, X. Besseron, and P. Bouvry. ***HPC Performance and Energy-Efficiency of the OpenStack Cloud Middleware***. In Proc. of the 43rd IEEE Intl. Conf. on Parallel Processing (ICPP-2014), Heterogeneous and Unconventional Cluster Architectures and Applications Workshop (HUCAA'14), Sept. 2014. IEEE.

Cloud IaaS (OpenStack) on Mobile SoCs



[CloudCom'14] V. Plugaru, S. Varrette, and P. Bouvry. **Performance Analysis of Cloud Environments on Top of Energy-Efficient Platforms Featuring Low Power Processors.** In Proc. of the 6th IEEE Intl. Conf. on Cloud Computing Technology and Science (CloudCom'14), Singapore, Dec. 15–18 2014.

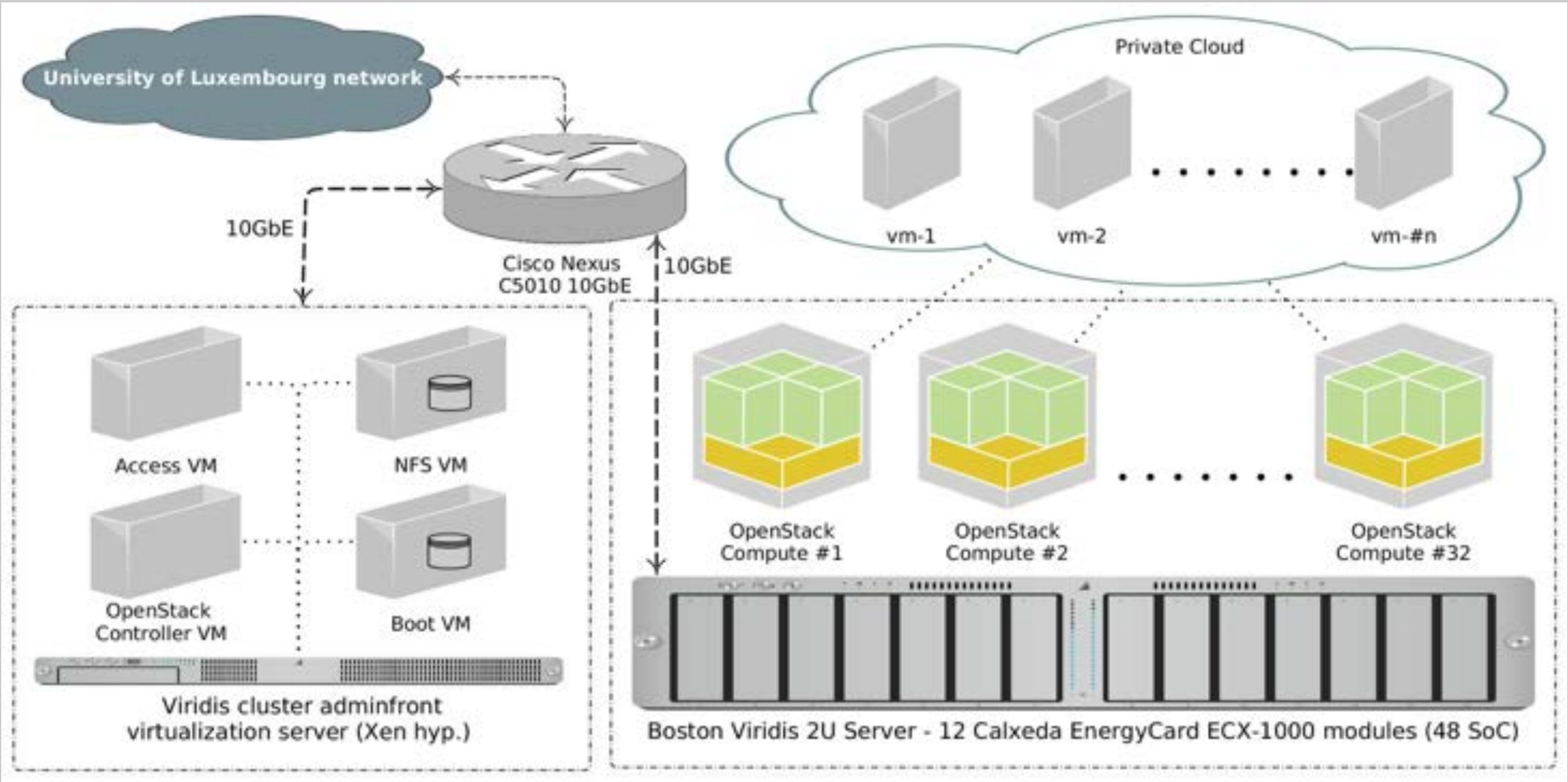
Cloud IaaS (OpenStack) on Mobile SoCs



	Avg. Performance drop				Avg. Energy-efficiency drop – Green500
	HPL	PTRANS	FFT	RandomAccess	
OpenStack 1VM/host	20.5%	56%	47%	25.2%	17.7%
OpenStack 2VM/host	24%	65.6%	56%	38.2%	23.5%

[CloudCom'14] V. Plugaru, S. Varrette, and P. Bouvry. **Performance Analysis of Cloud Environments on Top of Energy-Efficient Platforms Featuring Low Power Processors.** In Proc. of the 6th IEEE Intl. Conf. on Cloud Computing Technology and Science (CloudCom'14), Singapore, Dec. 15–18 2014.

Cloud IaaS (OpenStack) on Mobile SoCs



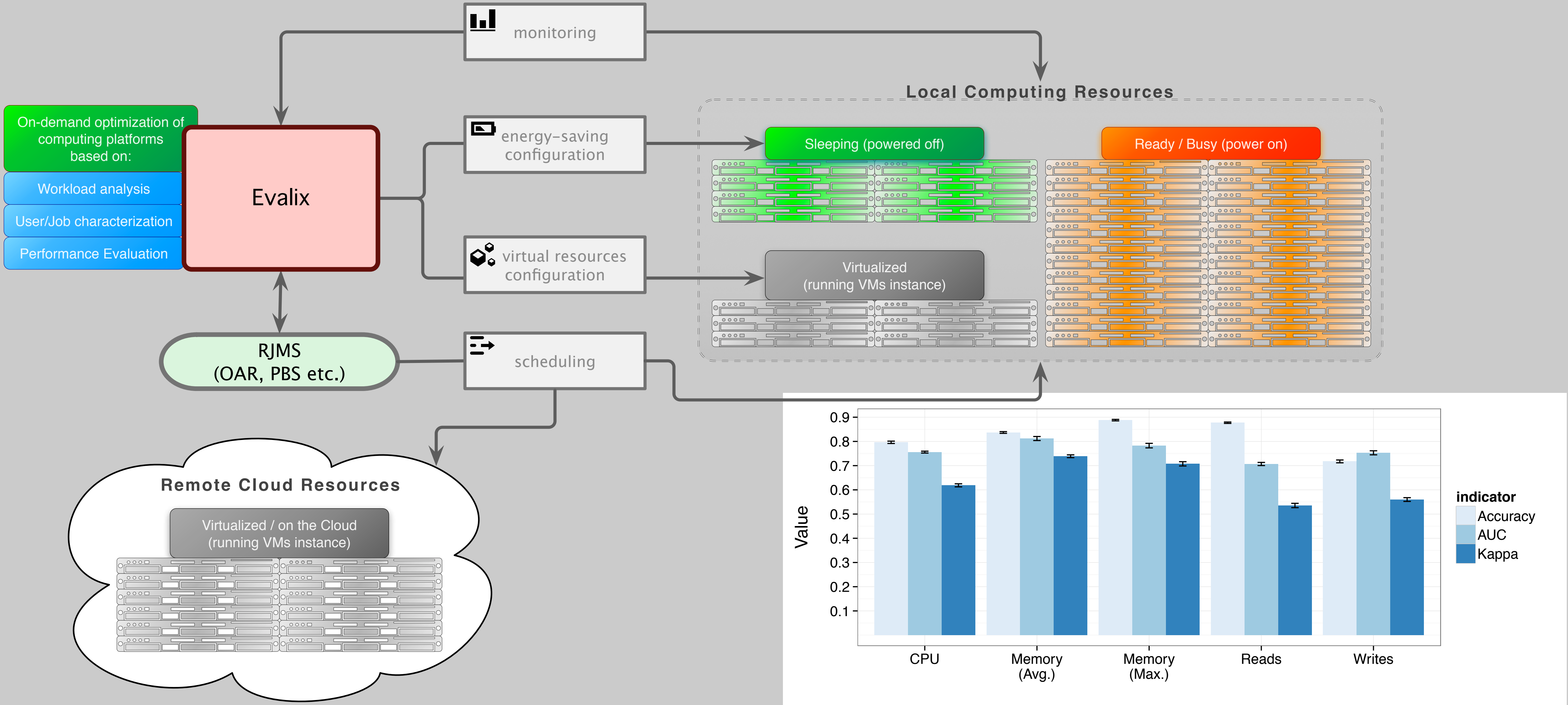
	Avg. Performance drop				Avg. Energy-efficiency drop – Green500
	HPL	PTRANS	FFT	RandomAccess	
OpenStack 1VM/host	20.5%	56%	47%	25.2%	17.7%
OpenStack 2VM/host	24%	65.6%	56%	38.2%	23.5%

Configuration	PpW	G500 Rank
Viridis Baseline	513.53 MFlops/W	**204**
Viridis OpenStack/LXC 1VM/host	371.76 MFlops/W	**234**
Viridis OpenStack/LXC 2VM/host	333.94 MFlops/W	**239**

[CloudCom'14] V. Plugaru, S. Varrette, and P. Bouvry. **Performance Analysis of Cloud Environments on Top of Energy-Efficient Platforms Featuring Low Power Processors.** In Proc. of the 6th IEEE Intl. Conf. on Cloud Computing Technology and Science (CloudCom'14), Singapore, Dec. 15–18 2014.

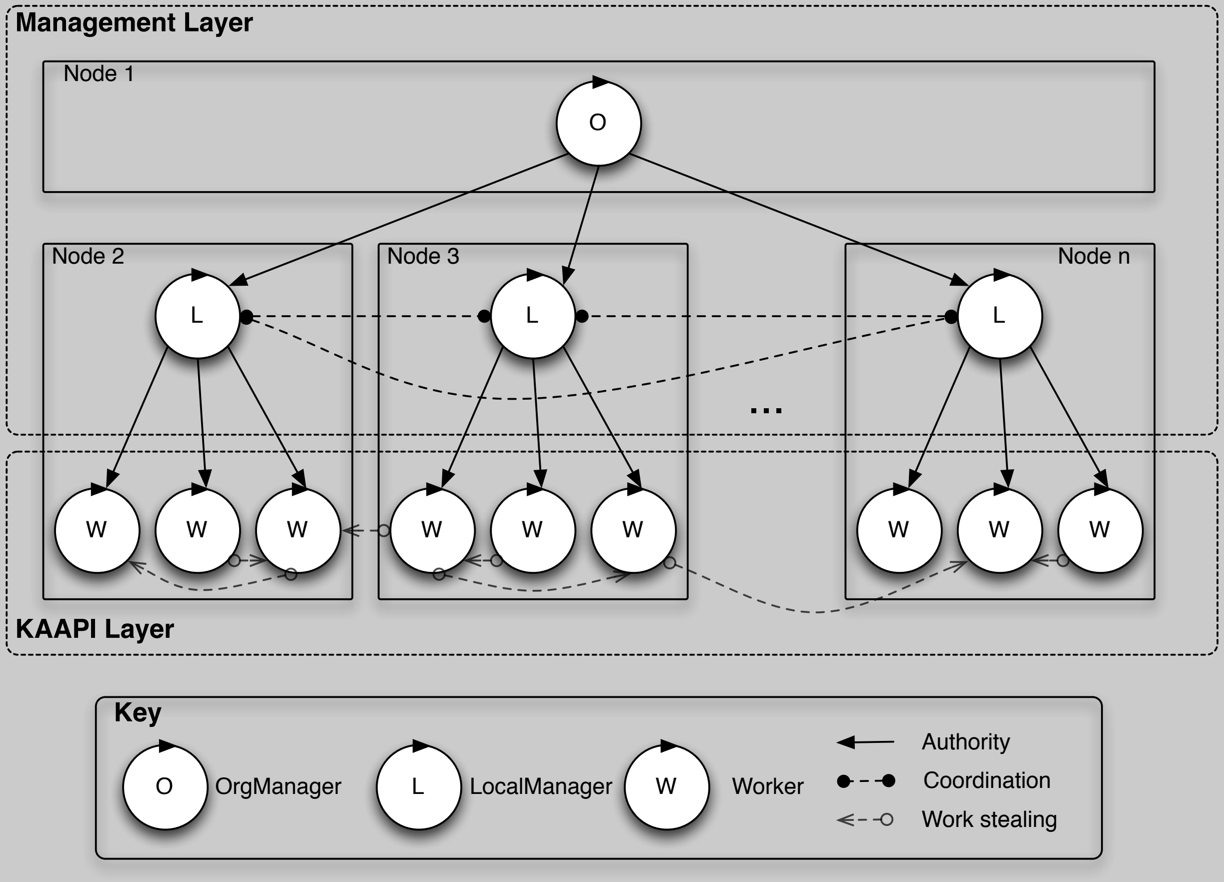
Virtualization, RJMS and HPC

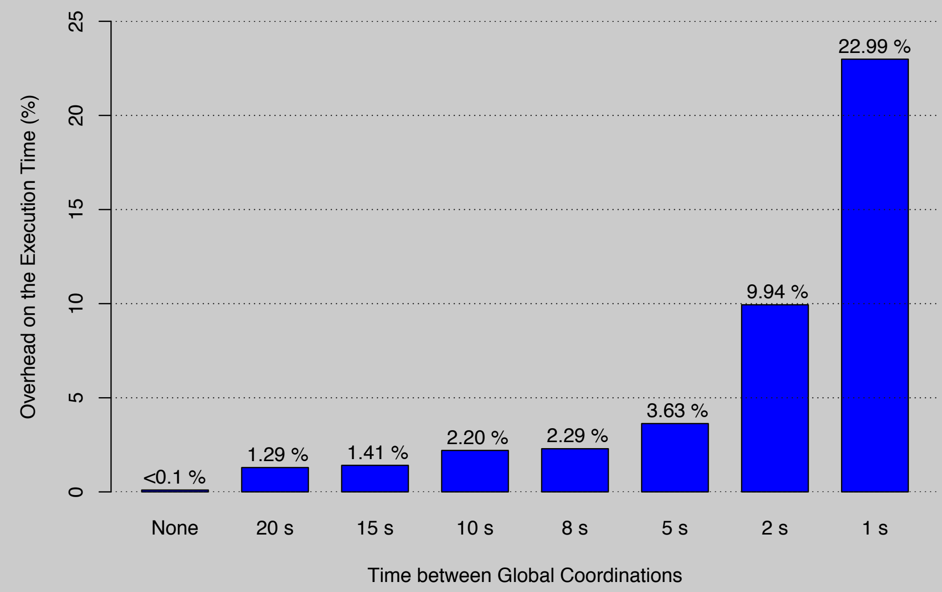
- Virtualization not suitable for pure HPC performance
  - ✓ YET not all workloads running on HPC are pure-parallel

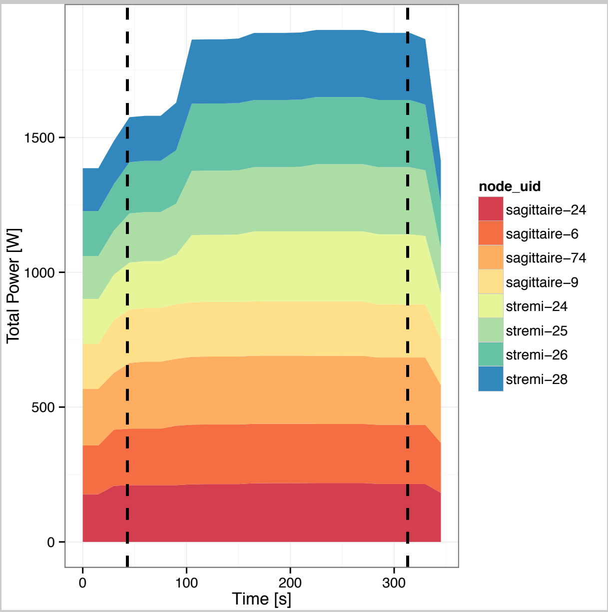


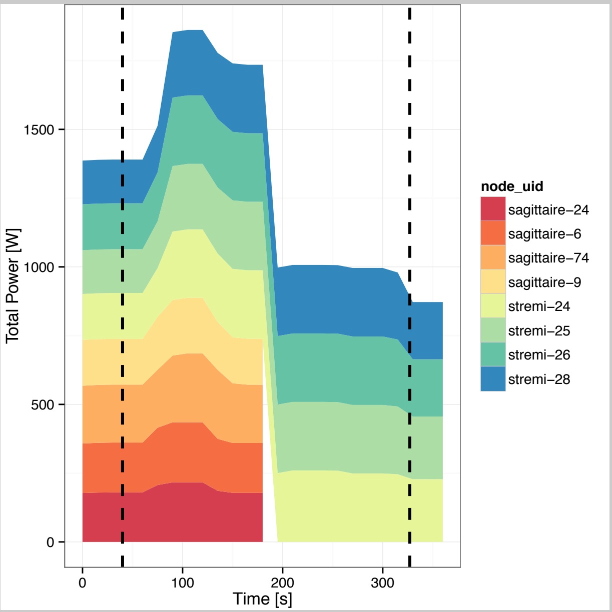
[JSSPP'15] J. Emeras, S. Varrette, M. Guzek, and P. Bouvry. **Evalix: Classification and Prediction of Job Resource Consumption on HPC Platforms.** In Proc. of the 19th Intl. Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP'15), part of IPDPS 2015, Hyderabad, India, May 25–2919 2015. IEEE Computer Society.

- Multi-Agent System (MAS) for energy aware executions



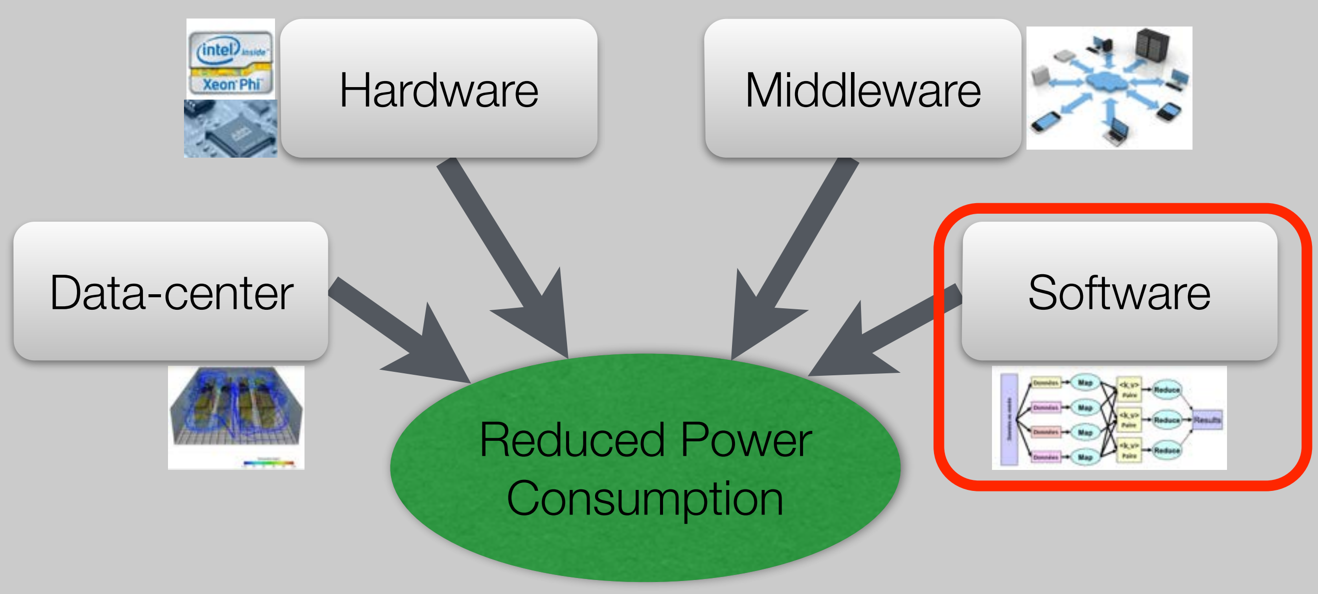






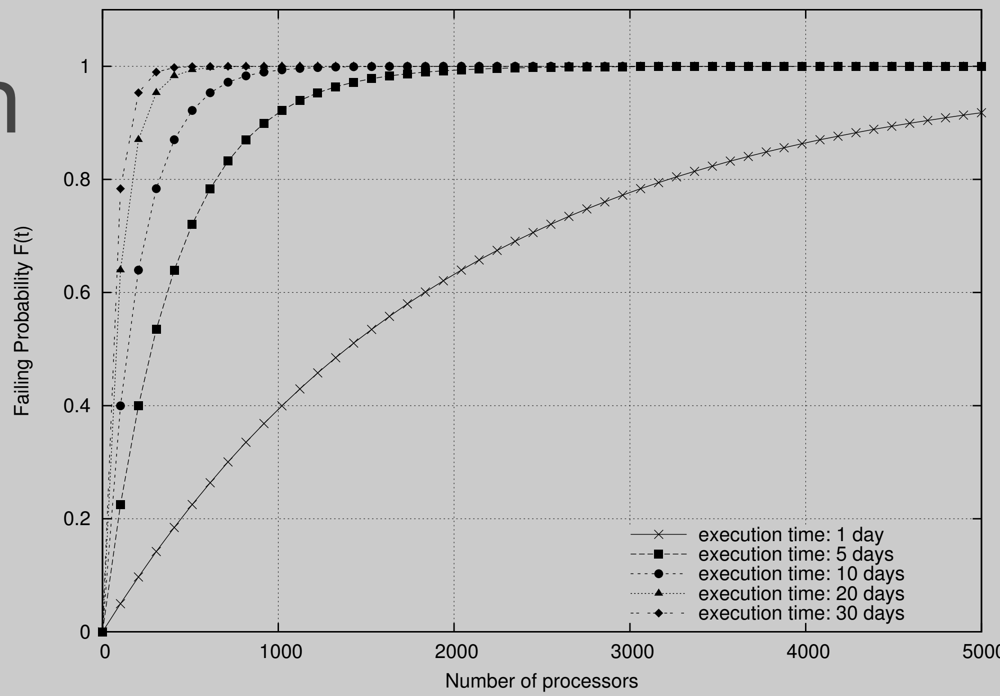
[ISSPIT'14] M. Guzek, X. Besseron, S. Varrette, G. Danoy, and P. Bouvry. **ParaMASK: a Multi-Agent System for the Efficient and Dynamic Adaptation of HPC Workloads.** In Proc. of the 14th IEEE Intl. Symp. on Signal Processing and Information Technology (ISSPIT'14), Noida, India, Dec. 2014. IEEE Computer Society

Software Trends: Rethinking Parallel Computing



Why is Exascale different for Software?

- Extreme power constraints, leading to:
  - ✓ clock rate similar to today's systems
  - ✓ heterogeneous computing elements. Ex: IBM Power Cell
  - ✓ Memory per {core | Flops} will be **smaller**
  - ✓ Moving data will be expansive (time and power)
- HW↦SW Fault detection/correction
  - ✓ becomes programmer's job
- Extreme Scalability
  - ✓ 10^8 - 10^9 concurrent threads
  - ✓ Performance is likely to be variable
    - static decomposition will not scale



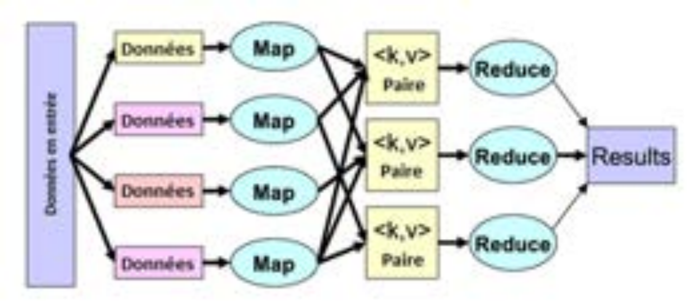
HPC Applications Compatibility Roadmap

Application	Traditional (x86_64)	Traditional +GPU	Energy efficient ARMv7	CC	(C)ompute/(D)ata intensive
Synthetic benchmarks					
HPCC	✓	*TBI*	✓	✓	C+D
HPCG	✓	*TBI*	✓	✓	C+D
Graph500	✓	*TBI*	✓	✓	C+D
Finite Element Analysis, Computational Fluid Dynamics software					
LS-DYNA	✓	*TBI*	*TBI*	✓	C+D
OpenFOAM	✓	*TBI*	*TBI*	✓	C+D
Molecular dynamics applications					
AMBER	✓	✓	*TBI*	✓	C+D
NAMD	✓	✓	*TBI*	✓	C+D
Bio-informatics applications					
GROMACS	✓	✓	✓	✓	C+D
ABySS	✓	×	✓	✓	C+D
mpiBLAST	✓	× alt.: GPU-BLAST	✓	✓	D
MrBayes	✓	× alt.: GPU MrBayes	✓	✓	C
Materials science software					
ABINIT	✓	✓	✓	✓	C+D
QuantumESPRESSO	✓	✓QE-GPU	✓	✓	C+D
Data analytics and machine learning benchmarks					
HiBench/Hadoop	✓	*TBI*	✓	✓	D

Rethinking Parallel Computing

- Today's execution model might be obsolete
  - ✓ Von Neumann machine
    - Program Counter, Arithmetic Logic Unit (ALU), addressable memory
  - ✓ Classic vector machine, GPUs w. collec. of threads (Warps)
- Plan change in the execution model:
  - ✓ no assumption on performance regularity
    - not unpredictable but imprecise
  - ✓ synchronization is costly: don't make it desirable
  - ✓ Memory operation are costly: move operations to data?
  - ✓ Represent key HW operations, beyond simple ALU
    - Remote update (RDMA), Remote atomic op. (compare & swap)
    - Execute short code sequence (active messages, parcels…

Challenges for Programming Models



- Probably successful: MPI, Map-Reduce
- Still pending challenges for exascale:
  - ✓ provide a way to coordinate resource allocation
  - ✓ clean way to share data with consistent memory models
  - ✓ Mathematical Model Guidance
    - continuous representation, possibly adaptative
    - lossy (within accuracy limits) yet preserving essential properties
  - ✓ Manage code by Abstract Data Structure Language (ADSL)
  - ✓ Adaptative with a multi-level approach
    - lightweight, locally optimized vs. intra node vs. regional
    - may rely on different programming models

Conclusion

- Still a long way to go ;)

new [co-]processors, interconnect…

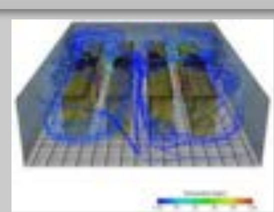
Hardware

Virtualization, RJMS…

Middleware

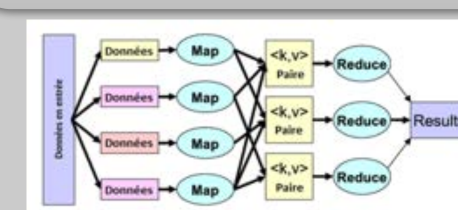
Data-center

PUE optim.
DLC…



New programming/execution models

Software



Reduced Power Consumption

- Questions?