SIDUS
Extreme Deduplication & Reproducibility

***S**ingle **I**nstance **D**istributing **U**niversal **S**ystem*

A Swiss Knife for Scientific Computing and elsewhere

Starting with a tiny joke...

- How do you call people speak 3 languages ?
  - Trilingual people !
- How do you call people speak 2 languages ?
  - Bilingual people !
- How do you call people speak 1 language ?
  - French people !

I'm french :

If I twist your eardrums, I apologize...

My story with Debian (I'm not a DevDeb)
19 years ago Debian, in 1996...

- Everything started in 1996, I got married (2 times...)
  - On 10th of August with Noëlle
  - On 17th of June with Debian
- Nineteen years later in 2015 :
  - 3 wonderful children & separated for 3 years...
  - Several hundred on Debian systems installed (or deployed) !
    - As Workstations on PC and Laptops
    - As servers, routers (ATM un 2001, GE in 2002, 10G in 2007), VPN gateways
    - As nodes of scientific computing
    - As chrooted systems on (pull the vomit bag under your seat) RedHat distro
    - **In my universe, most Debian systems running are started, not installed...**

What SIDUS is NOT...
First : SIDUS is not SIDIOUS !

Darth SIDIOUS alias Palpatine

Sidus : “constellation” in Latin

Difference between Sidious & SIDUS : IO

From SIDUS to SIDIOUS when I/O problems ? Let's see...

What SIDUS is NOT...
But what SIDUS shares with them !

What SIDUS is NOT !

- **LTSP :** *Linux Terminal Server Project*
  - One server, simplified administration of clients
- **FAI, Kickstart, Debian Installer Preseed :**
  - *« And the machine replaces human during installation process »*
- **LiveCD by network :**
  - A ISO image distributed by network

What SIDUS shares with them

Boot PXE, TFTP, NFSroot, **AUFS**

Two mainly properties of SIDUS Reproducibility in space/time

- **Uniqueness of configuration**
  - Two SIDUS clients : the same OS bit by bit !
- **Local resources exploited**
  - Processors & RAM (& extra...) exploited : client ones !
- **Reproducibility?** For an unchanged SIDUS
  - Time stability (for a defined client)
    - Two consecutive boots on a defined machine offers exactly the same system
  - Space stability (for two or more different clients)
    - Two clients starting at one run exactly the same system

SIDUS in 7 Questions : CQQCOQP or W5H2

In French, analytical method is CQQCOQP :

- Comment, Quoi, Qui, Combien, Où, Quand, Pourquoi ?

In English (Globish) HWWHWWW (wolf howling ?)

- How, What, Who, How much, Where, When, Why ?

Simple method to describe something

- Very used in journalism not forget elements
- Very useful in project management

SIDUS in 7 Questions : CQQCOQP (Tell me) Why (yyyyyy) ?

Why ?

- To uniform de facto all the « clients »
- To limit administration tasks to a unique one
- To compare materials with a done base
- To get it back fluids (Watts & BTU)
- To streamline workstation use
- To investigate storage resources under anesthesia
- To make sure of reproducibility on OS & its applications

SIDUS in 7 Questions : CQQCOQP
For What? For Whom?

For What ?

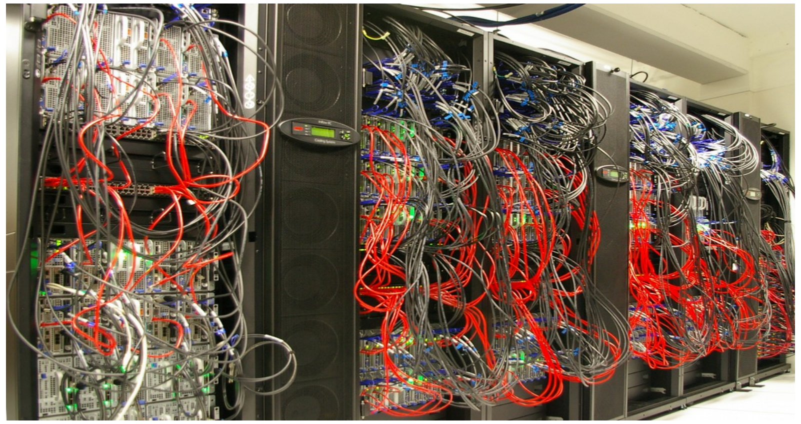
- Cluster nodes in HPC
- Self-service workstations
- Graphical workstations
- IT experimental benchs
- *Compute On My Own Device*

For Whom ?

- Researcher in computing science
- Engineer in computing science
- Computer room administrator
- Teacher using complex tools
- RSIS

SIDUS in 7 Questions : CQQCOQP, keep going... Where & When ?

- **Centre Blaise Pascal**, ENS-Lyon : computer room
  - 12 Neoware in 2010Q1, 24 stations 2013Q4
- **Centre Blaise Pascal**, ENS-Lyon : cluster
  - 24 nodes in 2010Q1, 76 permanent nodes in 2014Q2
- **Centre de calcul PSMN**, ENS-Lyon
  - 100 nodes 2012Q2, **480** nodes 2015Q1
  - All Equip@Meso
- Laboratories, ENS-Lyon
  - Chimie, **IGFL**, LBMC, UMPA, RDP
- École de physique des Houches
  - Editions from 2011 to 2015



SIDUS in 7 questions : CQQCOQP, the last line ! How does it work ?

- AUFS : *Another Union File System*
  - Aggregate File Systems in one : LiveCD trick
  - 4 steps:
    1. Mount NFSroot with OS on a first folder
    2. Create TMPFS on a second folder
    3. Glue with AUFS of the 2 previous folders
    4. Offer the resulted folder as the root of OS
  - Behavior of a normal Read/Write Filesystem
  - On reboot, every modification disappears
- One prerequisite : chroot for initial installation & administration

SIDUS in 7 questions : CQQCOQP, the end !
How to install : SIDUS in 7 steps (from Lenny to Wheezy)

1) Creation by Debootstrap of a new root exported by NFS

2) Creation of a “umbilical cord” with the host

- Mount of /proc /sys /dev/shm folders

3) Installation (& purge of specific unwanted packages)

4) Adaptation to local environment (timezone, keyboard, locales, filer, auth)

5) Creation of booting sequence with AUFS

- **Copy of the rootaufs file in /etc/initramfs/scripts/init-bottom**
- **Launch update-initramfs -k all -u**

6) Importation of kernel & specific initrd to TFTP server

7) Release the “umbilical cord” with the host

Migration to Debian Jessie
Limitations & evolutions

- **Goal : provide a technical base for hypervisors**
  - Context : on DHCP boot time, initramfs provides a physical interface, not a bridge
  - Solution : activate at boot time a bridge, impossible to initramfs-tools, change to dracut
- **Goal : improve integrity & privacy for users data (on unsecure networks)**
  - Context : NFSv4 needs to open to widely the NFS home directories
  - Solution : Kerberos is a too complex solution, CIFS with Posix extensions & pam_mount
- **Goal : keep working rootaufs in Jessie**
  - Context : up to Wheezy, on rootaufs in init-bottom folder was necessary, it brokes on Jessie
  - Solution : change deeply initramfs-tools scripts or migrate to dracut
- **Goal : limit the number of SIDUS instances on specific materials**
  - Context : to One Nvidia, One AMD/ATI, One 64 bits & One 32 bits for VirtualBox
  - Solution : customize at boot time the configuration of graphical boards

SIDUS in 7 questions : CQQCOQP, the end !
How to install : SIDUS in 7 steps for Jessie

1) Creation by **Debootstrap** of a new root exported by NFS
2) Creation of a “umbilical cord” with the host
   - Mount of /proc /sys /dev/shm folders
3) Installation (& purge of specific unwanted packages)
4) Adaptation to local environment (timezone, keyboard, locales, filer, auth)
5) **Creation of booting séquence with AUFS**
   - **Modify the broken aufs-mount.sh from Dracut to SIDUS one**
   - **Change DHCP lease time to forever in dclient-script.sh script**
   - **Hack to work around boot problems (like autofs conflicts with dbus)**
   - **Launch to create initrd : dpkg-reconfigure dracut**
6) Importation of kernel & specific initrd to TFTP server
7) Release the “umbilical cord” with the host

SIDUS in 7 questions : the end of howl ! How to administrate ?

- One limitation : the `/proc` must be unique..
  - Great vigilance for processes which go inside
    - Manipulation of Java, compilation with optimization, installation
- The **Good** :
  - Jail in chroot, set of “umbilical cord”
  - Classical operations, unset the cord
- The **Bad :**
  - Jail in chroot, classical operations directly (work 90% of time)
- The **Ugly (?)** :
  - Machine booting NFSroot in Read/Write and administration as WS

SIDUS in 7 questions : the end of Howl !
How does it cost ?

- An ideal network : Gigabit Ethernet for client, 10G server (local HD)
  - But it works on 100 Mb/s network !
- An ideal server : 4 CPU, 16 GB RAM, 10G, SSD
  - But it was working with a Sunfire v(eau)40z on 330 nodes in PSMN !
- An ideal client : all clones
  - But it works on 16 types of machines in PSMN, 10 types in CBP
- An ideal (alias motivated) integrator/administrator : ;-)
  - Deployed by L. Taulelle with rushes of documentation : PSMN
  - Déployed by T. Bellembois via on-line documentation : IGFL

Demonstration ?
SIDUS in real live

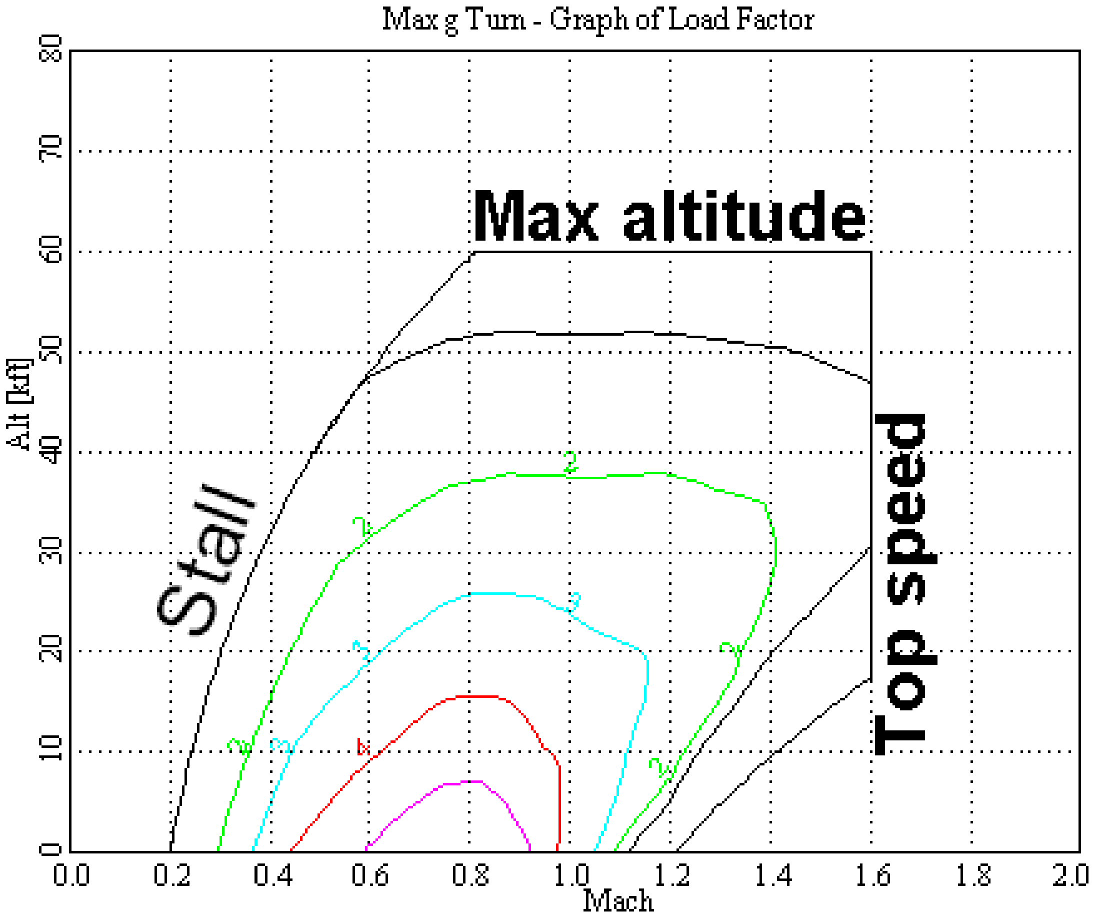
- Connection to my job via x2go to a gateway
- Connection to a Hypervisor with onboard GPGPU device
- Launch of a virtual machine booting by PXE
- Selection of the right image of SIDUS (Cuda one)
- Boot time (stresssss....)
- Login
- Analysis of CPU/Memory/Devices
- Launch of a DGEMM directly inside the virtual machine under SIDUS
- Enjoy !

SIDUS in reproducibility
Storage & Parallelism applications

- Emergence of Parallelism Domains to define hardware/software
- Pertinence of GlusterFS as distributed *scratch* in HPC context
  - Influence of BIOS on performance and variability
- Comparison of GPU & high level parallelism influence
  - Variability as discriminant factor between GPU
- Execution variability in « *Closed Embarrassing Parallelism* »
  - Difficulty of wall time estimation in compute time & influence of locality

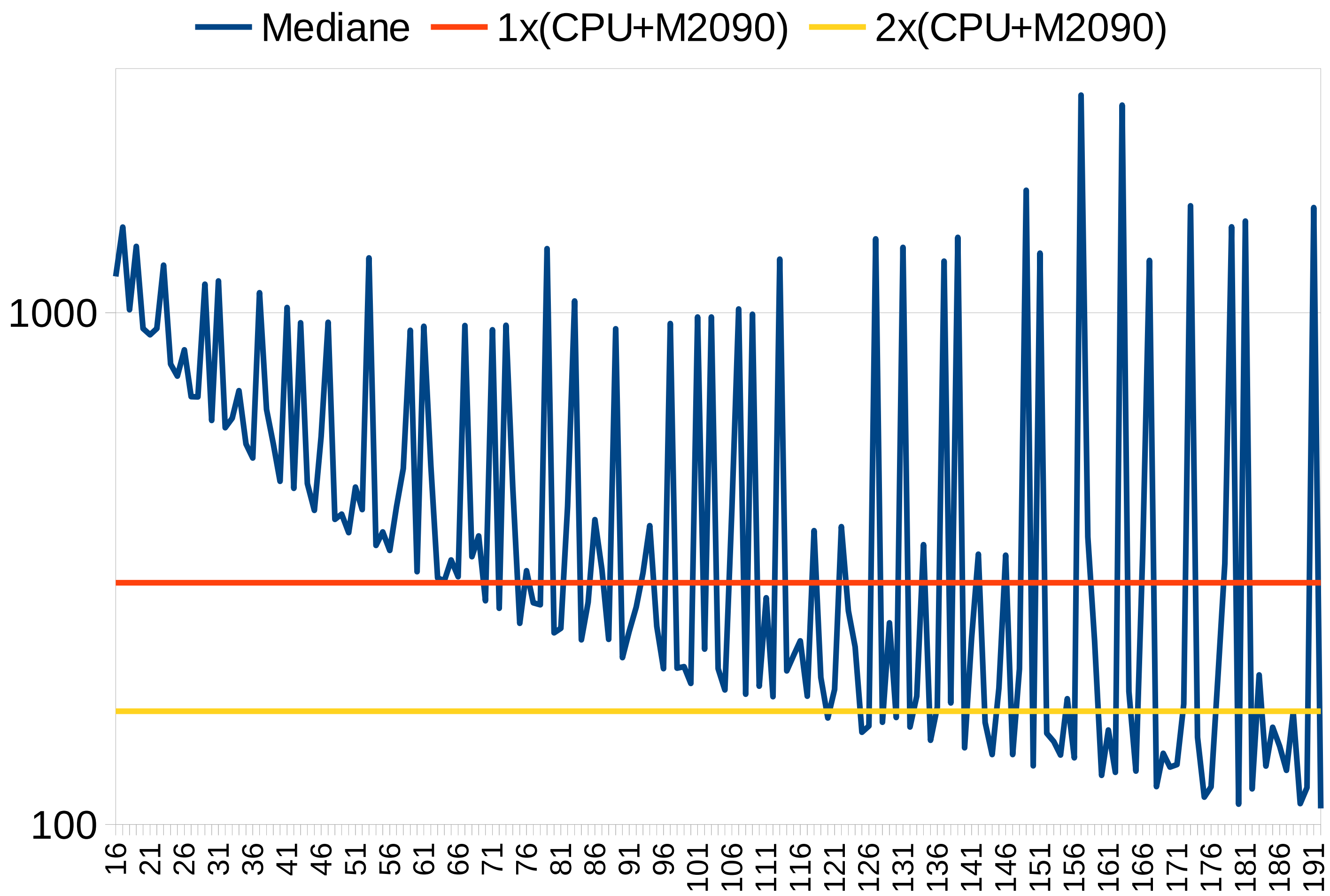
From flight envelop to parallelism envelop
The end of duality : “working/not working”

Flight Envelop



speed/altitude/G-force

Parallelism envelop



parallelism/memory/CPU/GPU

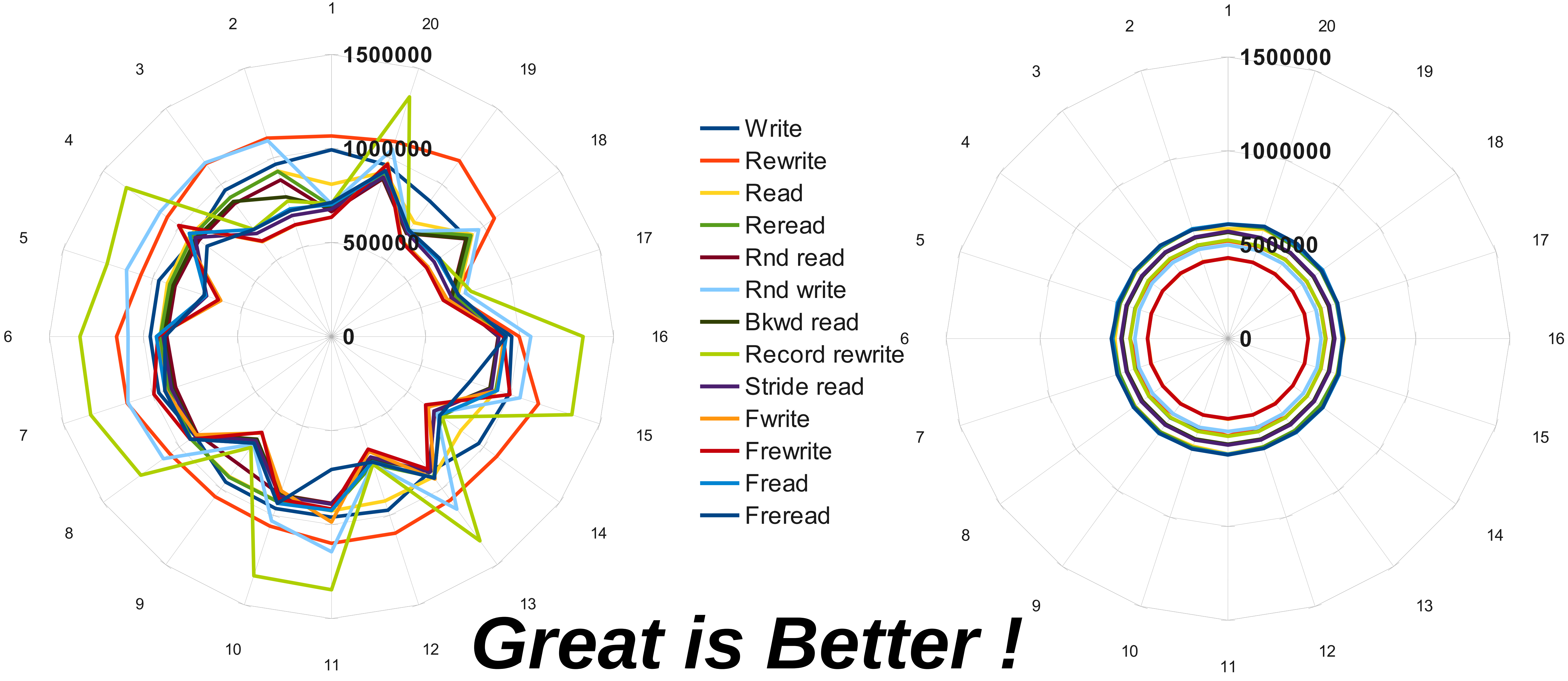
Variability in silicon space/time : Which maneuverability ?

- **Time**: same machine, different time ?
- **Space**: same time, different machines ?
- The solutions :
  - Restore of a identical OS image
    - Replicator, SystemImager, MondoRescue, ...
    - Kadeploy sur Grid'5000
    - Boot iSCSI with Back Office Snapshot (sur LVM, ZFSonLinux, BtrFS)
  - Installation with same automatic process :
    - FAI, Kickstart, Debian-Installer Preseed
  - **SIDUS : *Single Instance Distributing Universal System***

Lack of reproducibility ?
Example illustration : GlusterFS/IOZone

- **Objective** :
  - Evaluation of GlusterFS as High Performance /scratch
- **Experimental bench** : 20 nodes + infrastructure
  - 20 nodes Sandy Bridge 2x8 cores with 64 GB of RAM
  - A system **SIDUS** Debian Wheezy
  - Interconnection InfiniBand FDR 56 Gb/s
  - **No disk latency : RamDisk BRD/Ext2 & TMPFS of 60 GB**
  - 10 pairs GlusterFS : 1 server on RamDisk, 1 client
  - IOZone3 use : 13 tests of read/write
  - 20 experiences on a statistical and representative set

Day #1 : launch of tests & first surprises ! On I/O data rates



From node 11 to node 1

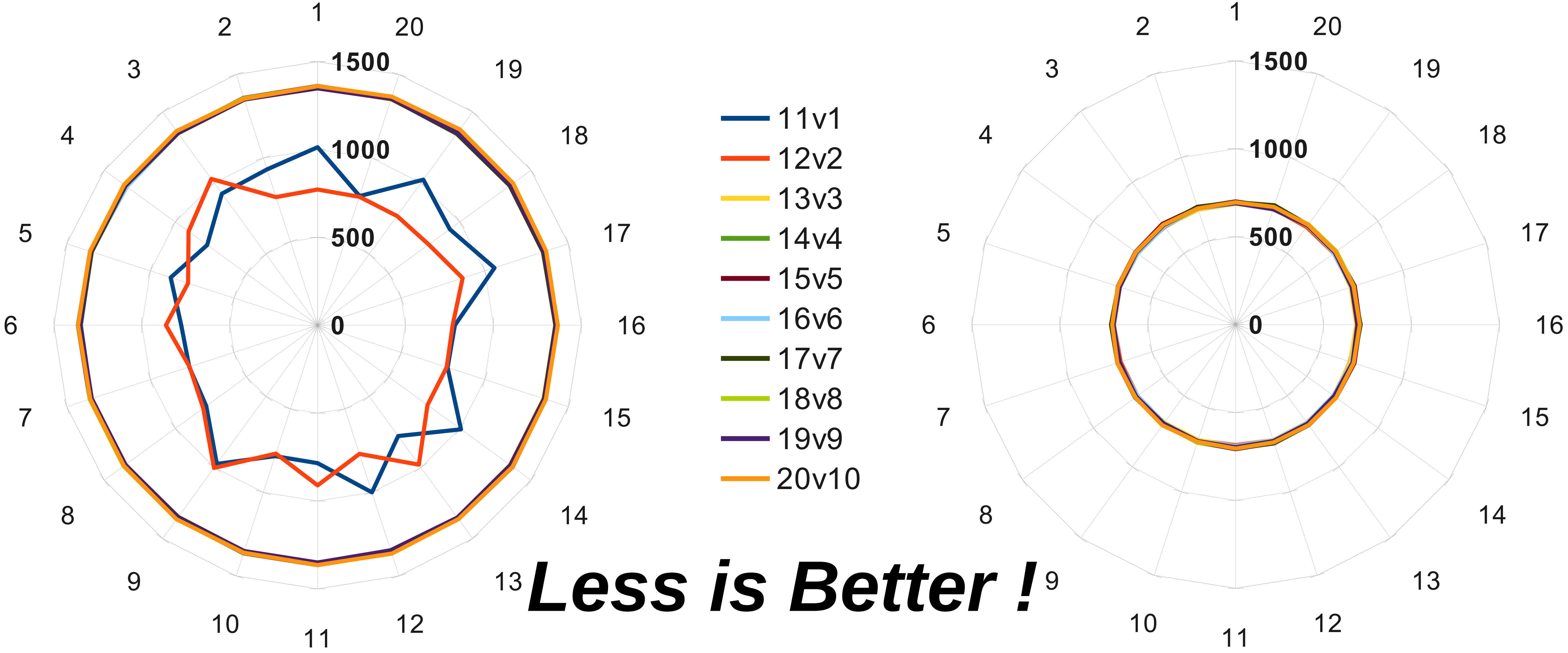
From node 12 to node 2

Great is Better !

Days #1 & #2 : modification & new tests
On elapsed executions (*User Time*)

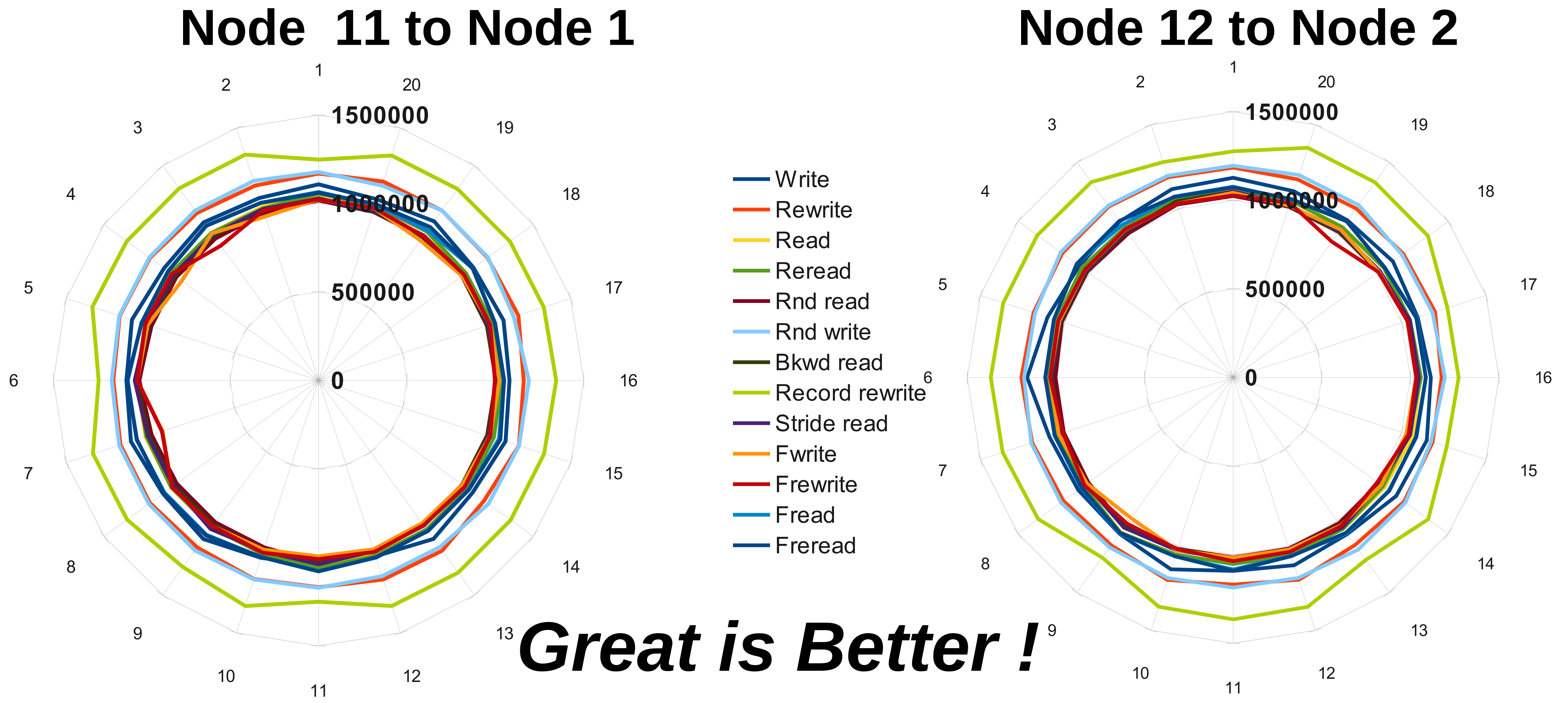
For the 10 couples **before**...

For the 10 couples, **after**!



Less is Better !

Day #2 : and on the 2 couples of first day ?
On I/O data rates



Great is Better !

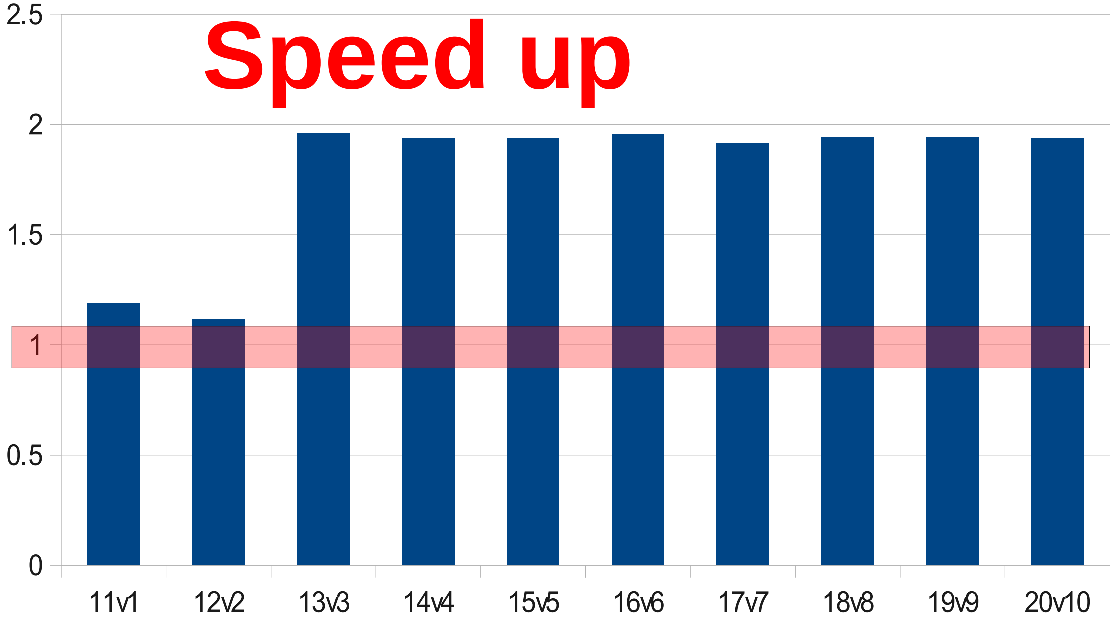
What miracle days #1 & #2 ?

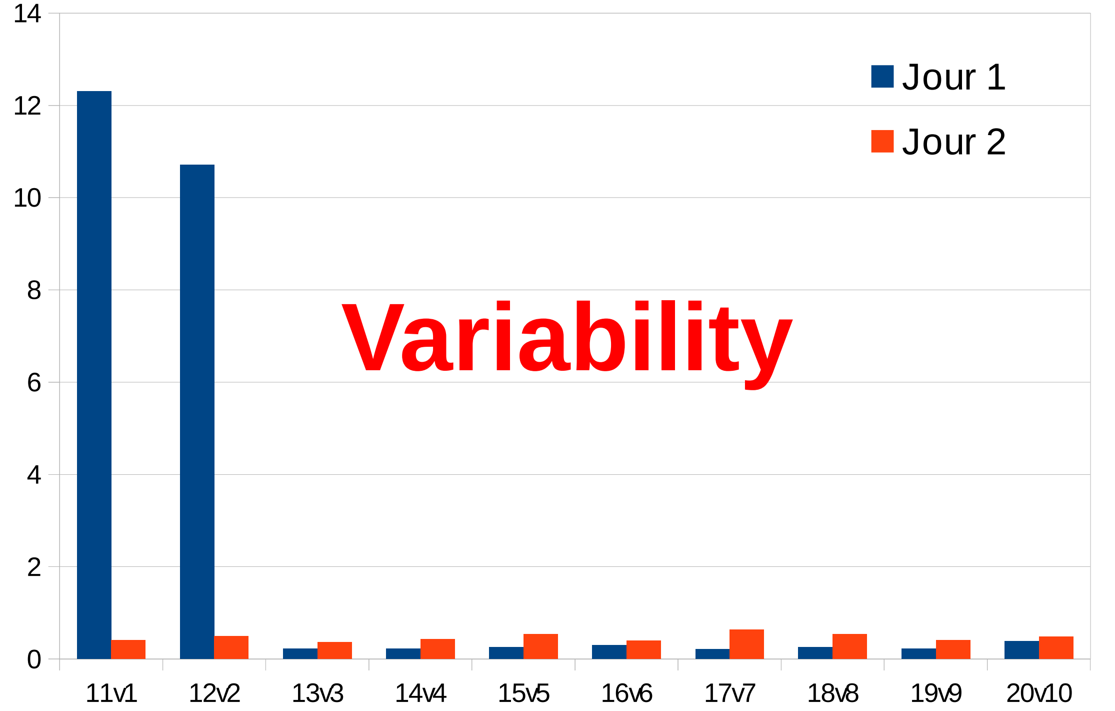
Two questions : How...

- ... multiply by 2 the speed ?
- ... divide by 20/30 the variability ?

The answer :

- Optimize the network ? No
- Optimize the kernels of OS ? No
- Tune the BIOS ? **YES !!!**
  - BIOS for 1 & 2 in Max Performance
  - BIOS for 3 to 20 by default
- Solution : BIOS in Max Perf !

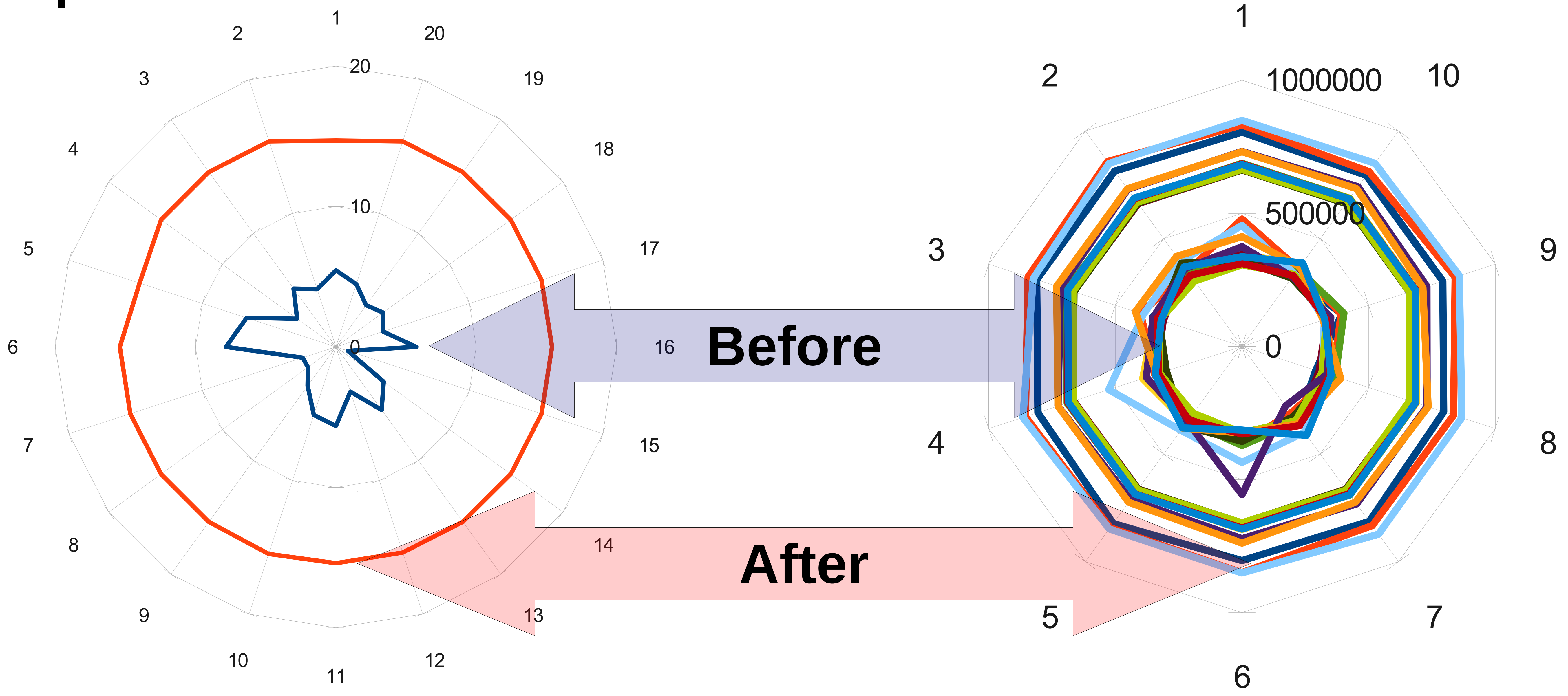




The no-reproducibility reproducible ?
On Equip@Meso

Iperf client/server with IB

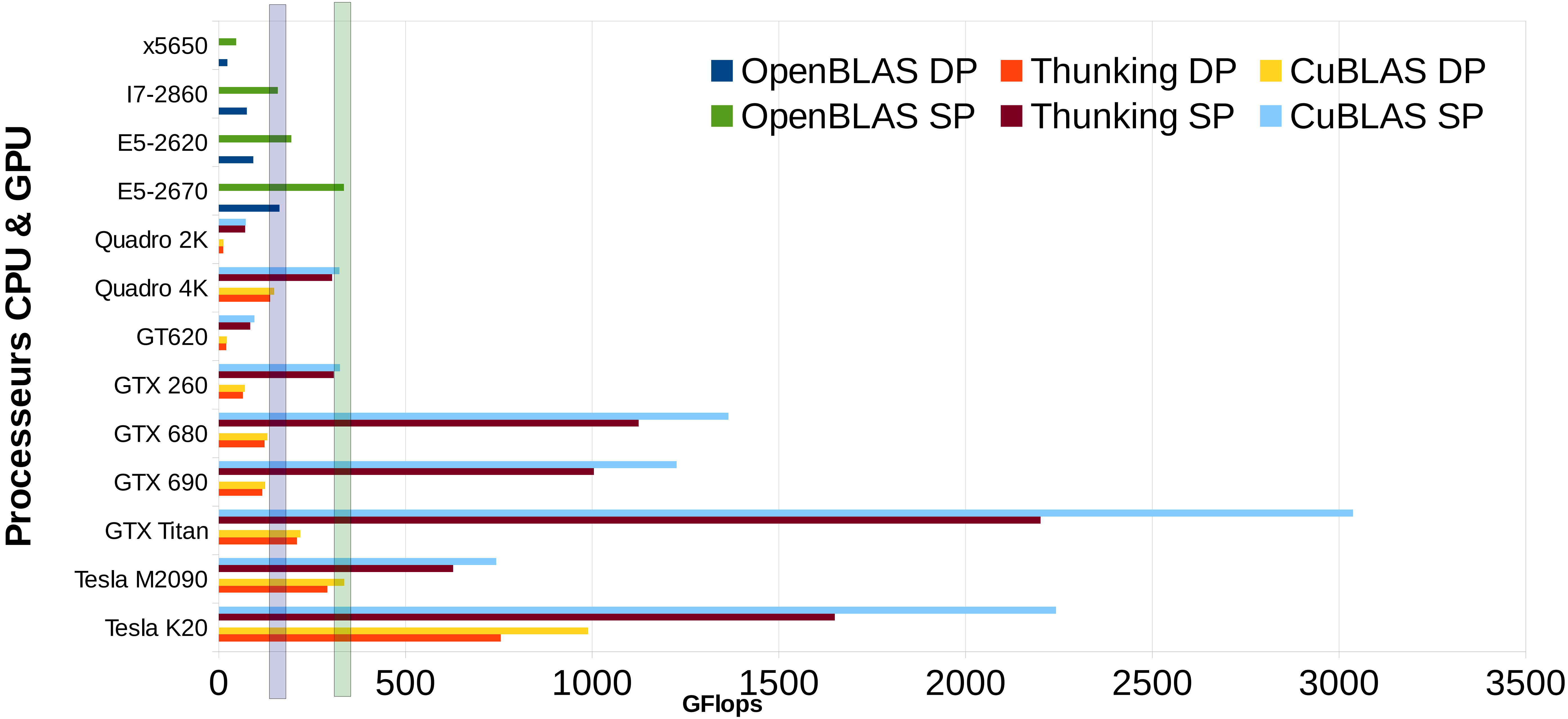
Iozone3 on GlusterFS



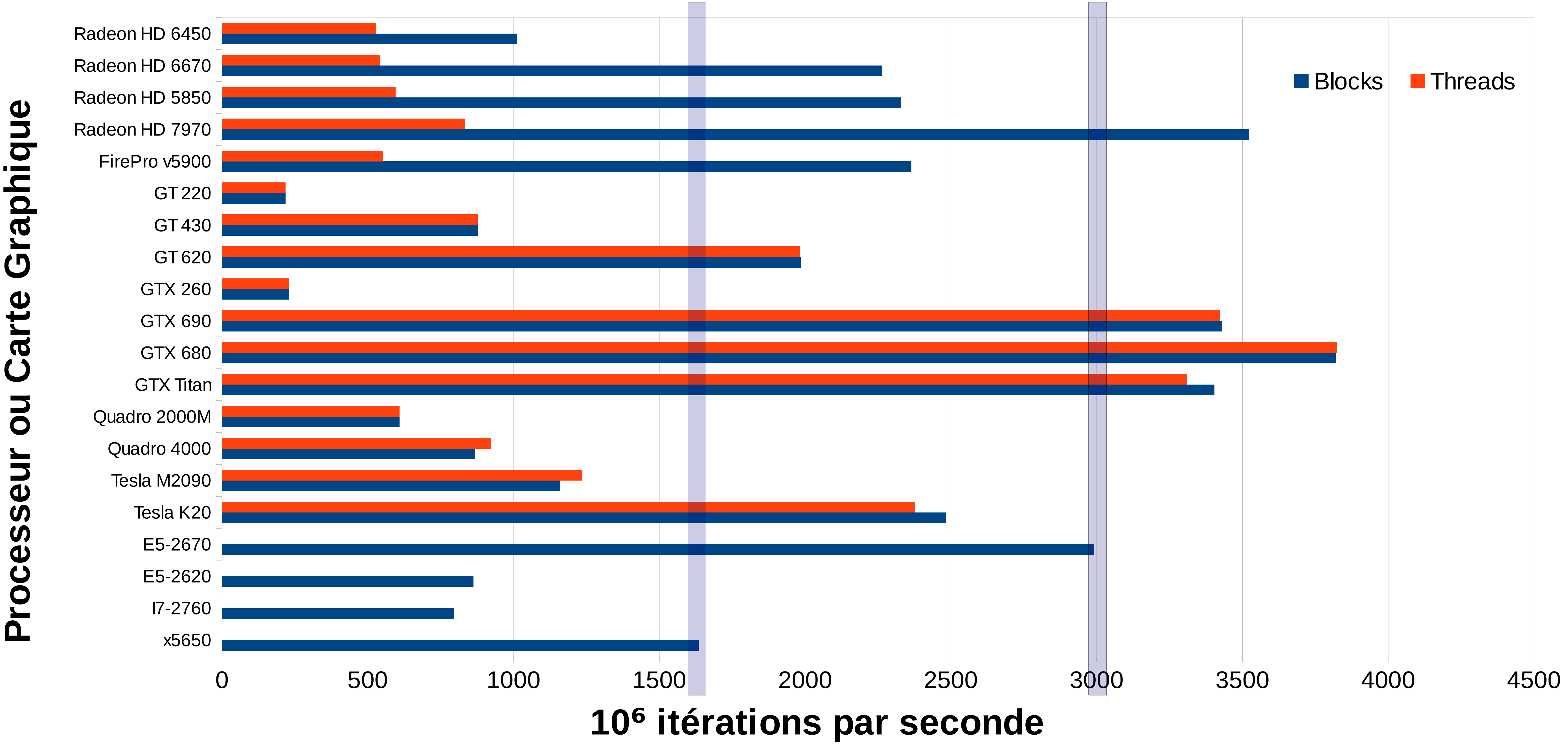
Comparison de GPU
What GPU for my applications ?

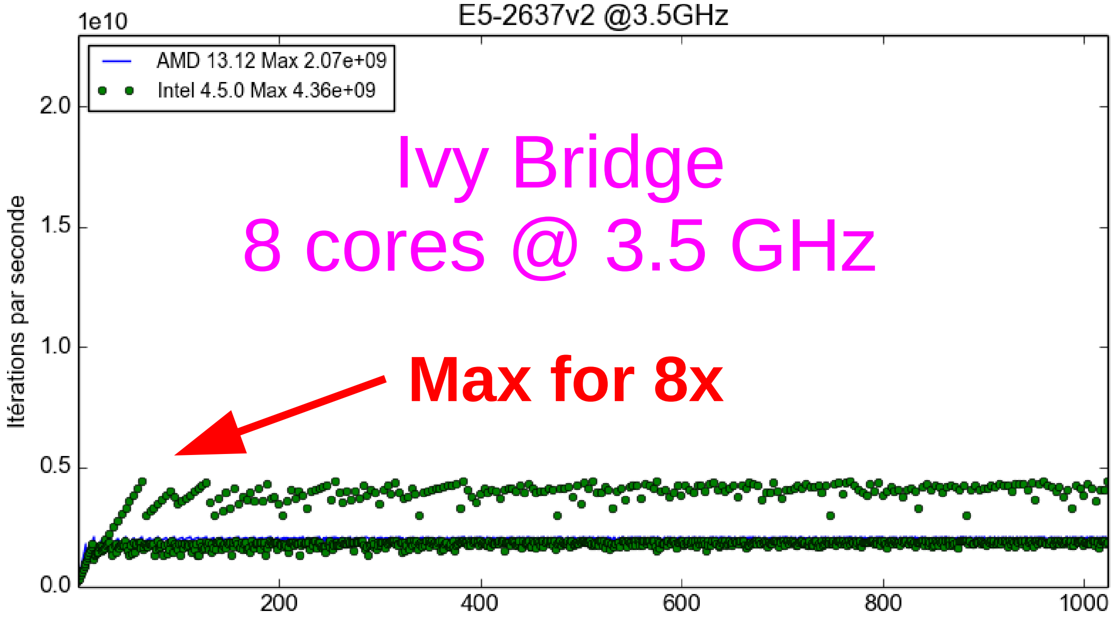
- **Objective** :
  - Evaluate performances of GPU (what choice pour which application)
- **Experimental platform** : 28 nodes/workstations
  - 20 types of graphical boards on 28 machines of 5 different types
  - First price, huge *gamer*, GPGPU, AMD/ATI/Nvidia
  - System **SIDUS** Debian Wheezy (with 2 instances...)
  - Pi Monte Carlo (load on RNG) : ALU solicitation “CPU band”
  - Explore from 1 to 1024 Blocks/Threads
  - Comparison between CPU, GPU & Manycore Phi

Why Pi Monte Carlo test ?
Free BLAS (and Linpack) hegemony !



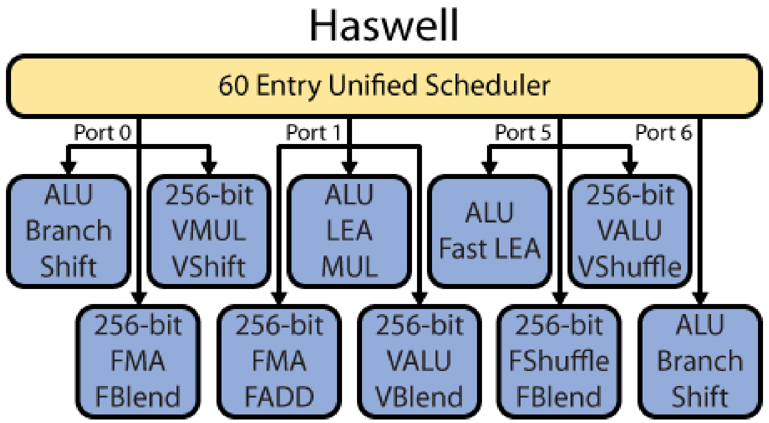
Comparison CPU/GPU : Pi by MC

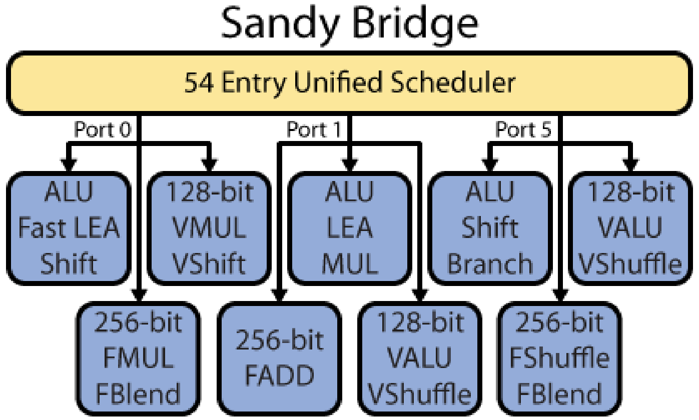


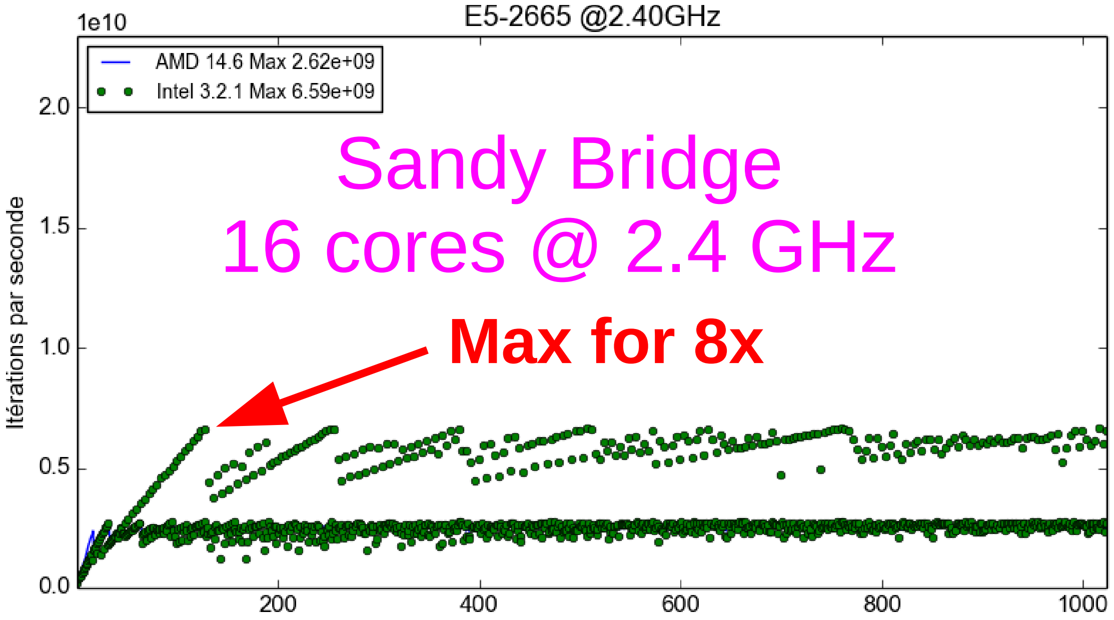

E5-2637v2 @3.5GHz

Ivy Bridge
8 cores @ 3.5 GHz

Max for 8x

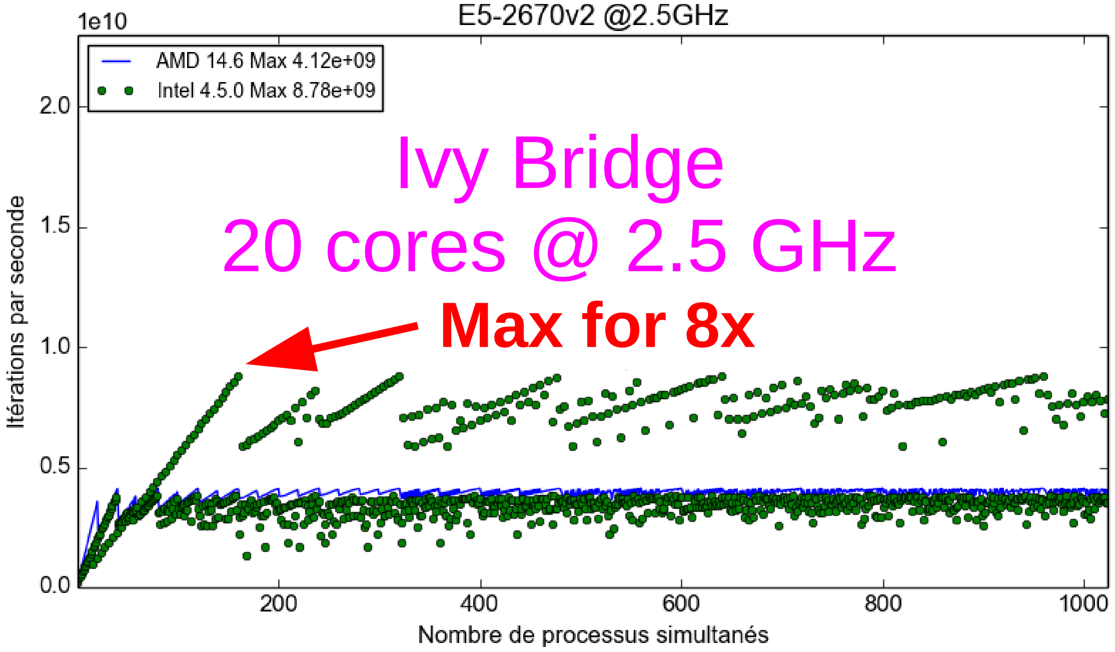

Haswell


Sandy Bridge


E5-2665 @2.40GHz

Sandy Bridge
16 cores @ 2.4 GHz

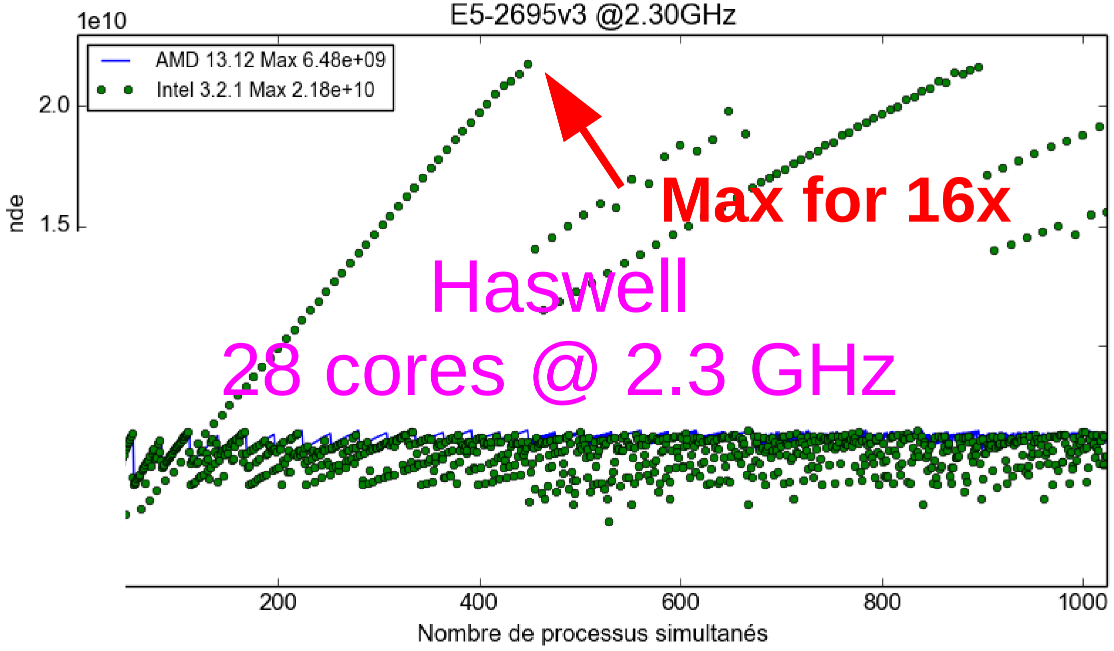
Max for 8x


E5-2670v2 @2.5GHz

Ivy Bridge
20 cores @ 2.5 GHz

Max for 8x

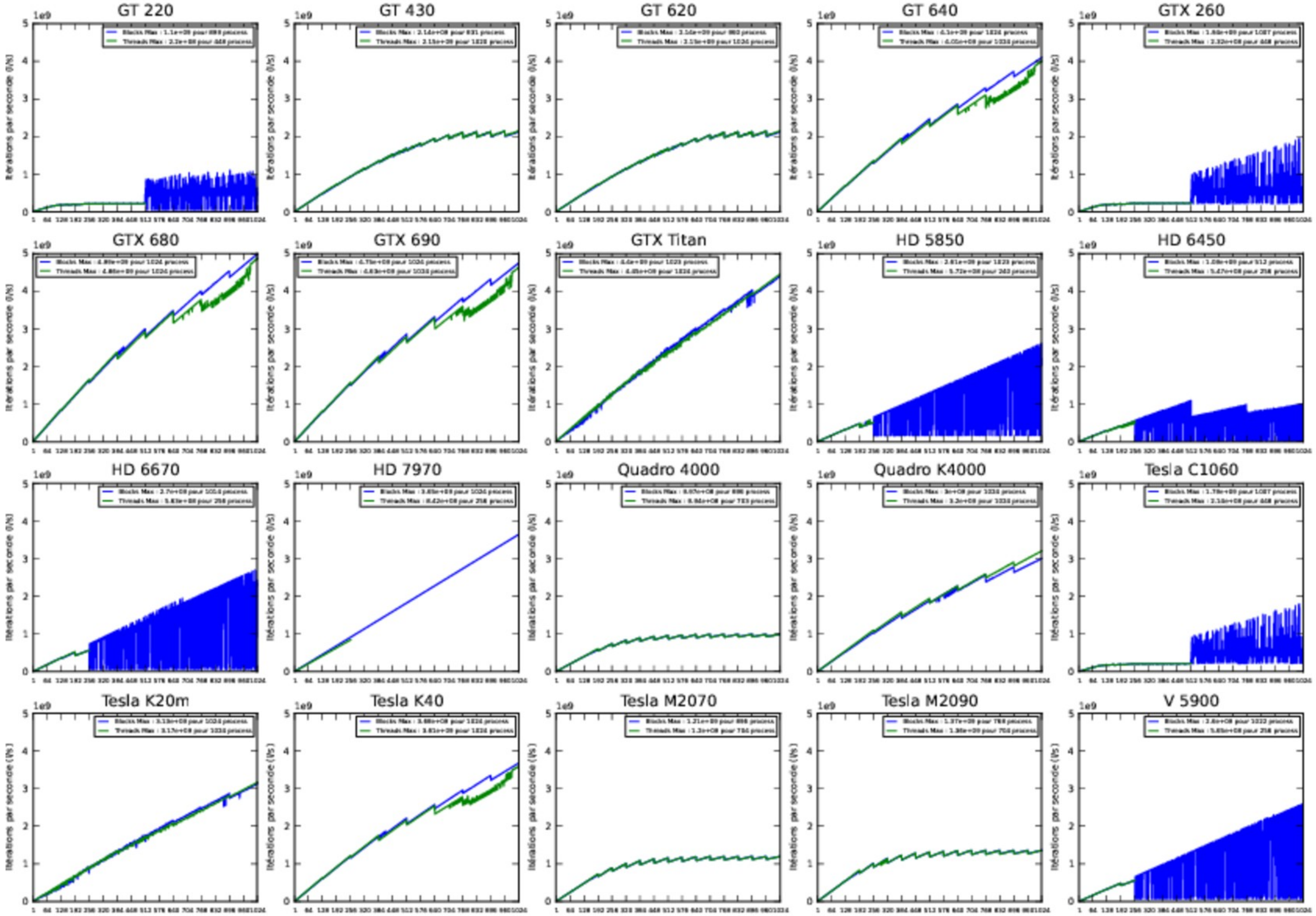
- Intel x2 à x3 vs AMD
- Period of 4
- Maximum Performance :
  - x8 for (Sandy|Ivy)Bridge
  - X16 for Haswell


E5-2695v3 @2.30GHz

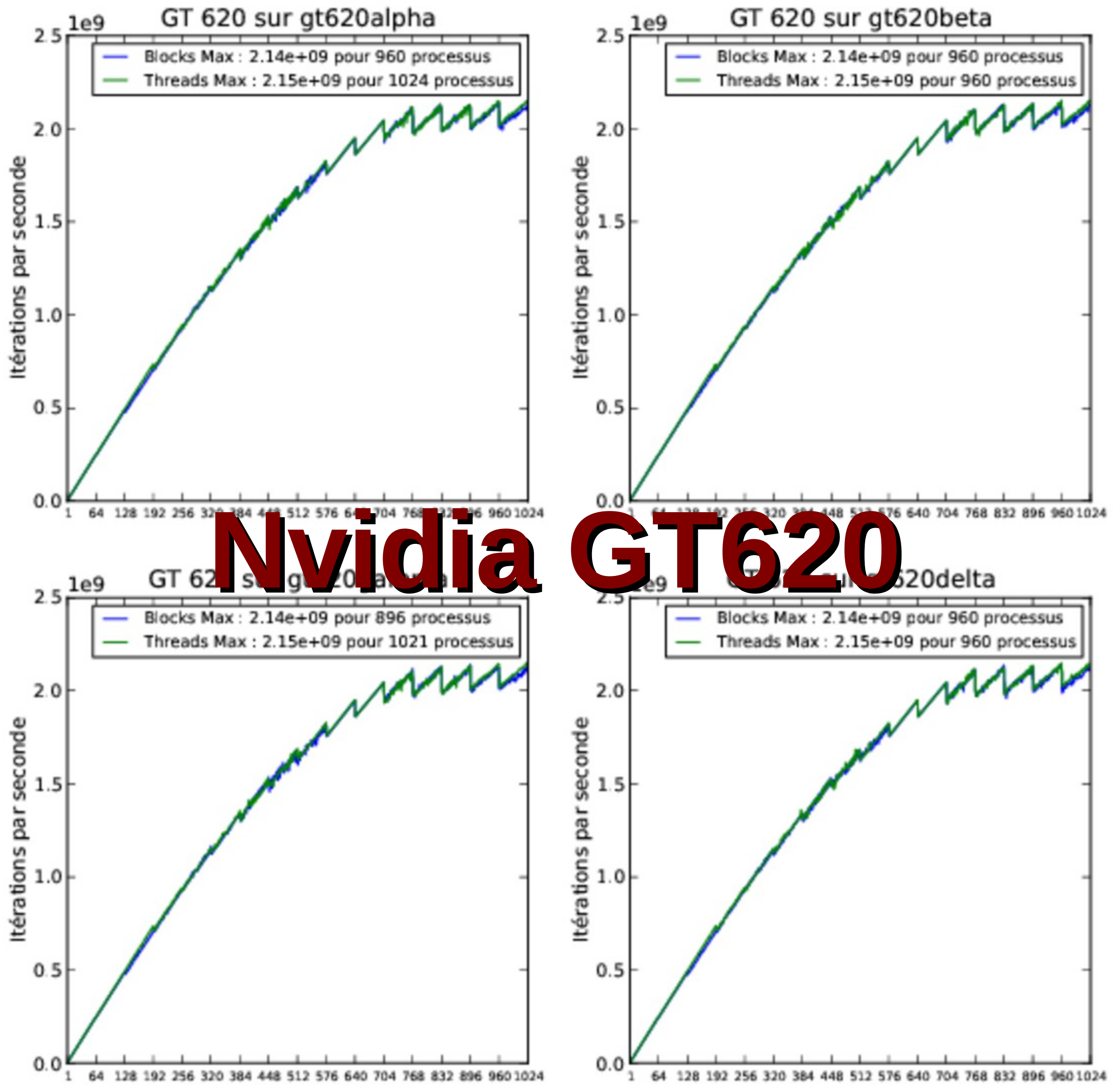
Max for 16x

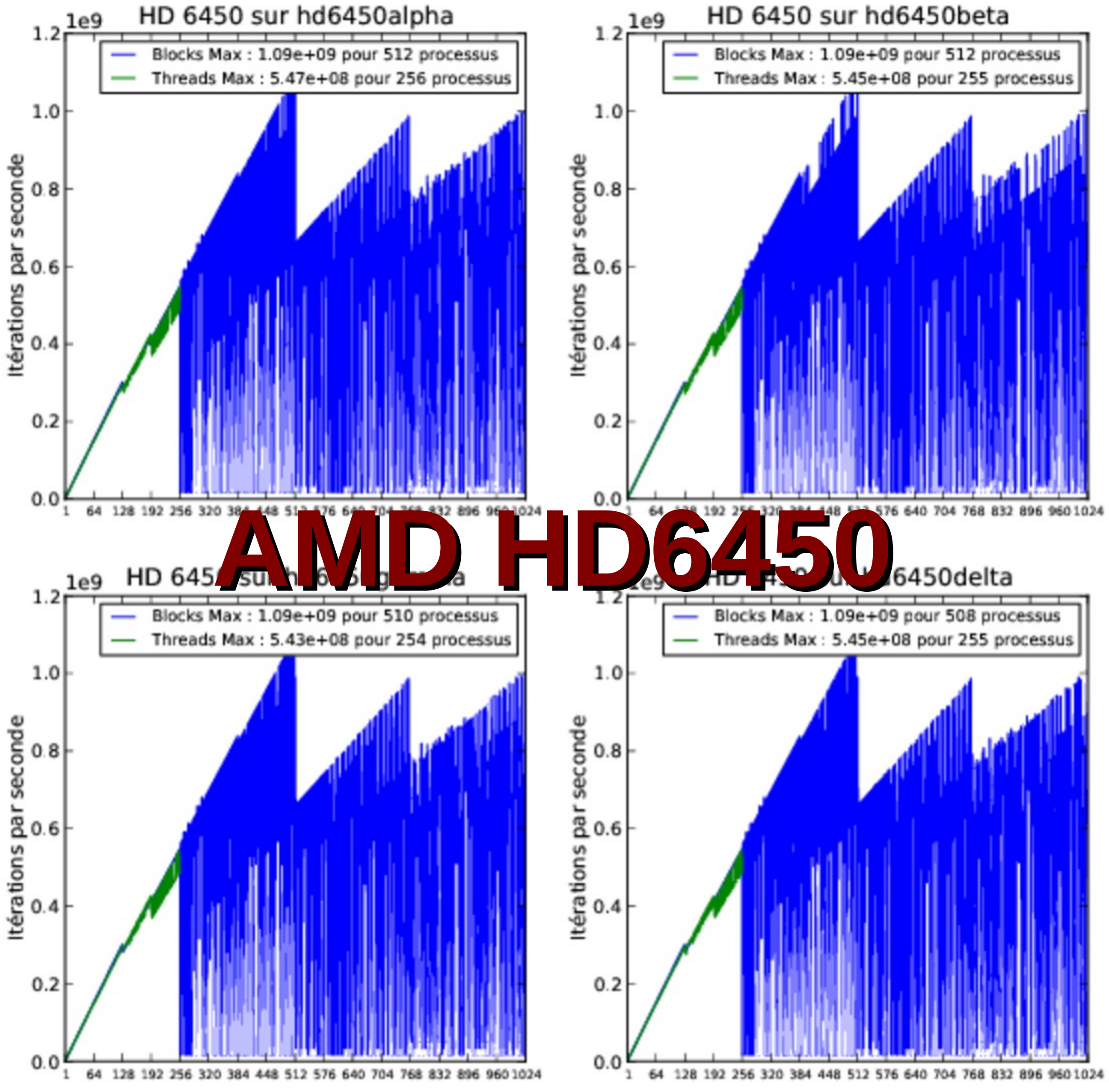
Haswell
28 cores @ 2.3 GHz

Visual Comparison of performances Parallelism envelop of GPUs

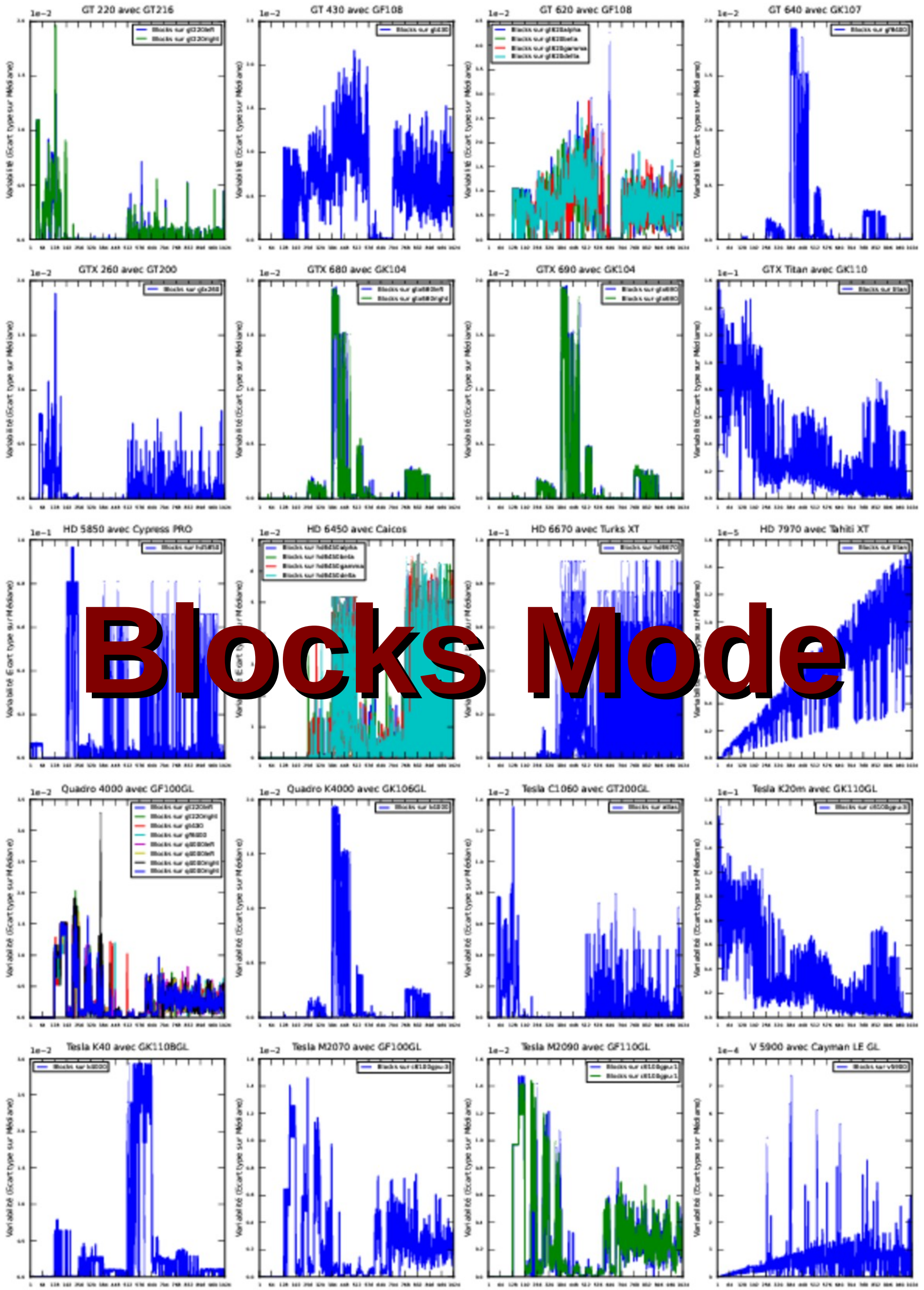


Visual comparison of performances

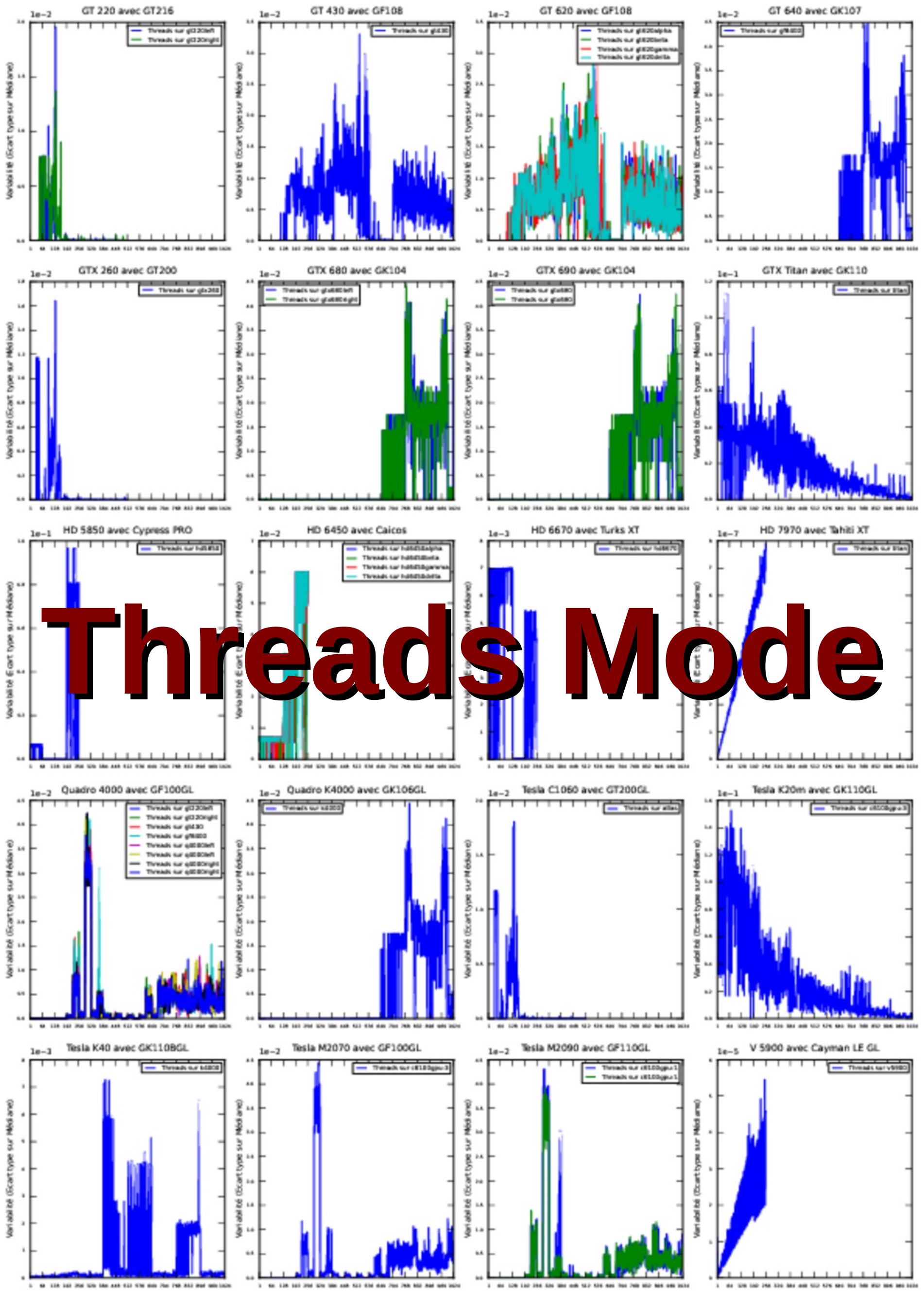

Nvidia GT620


AMD HD6450

Visual comparison of performances
The variability as discriminant factor !

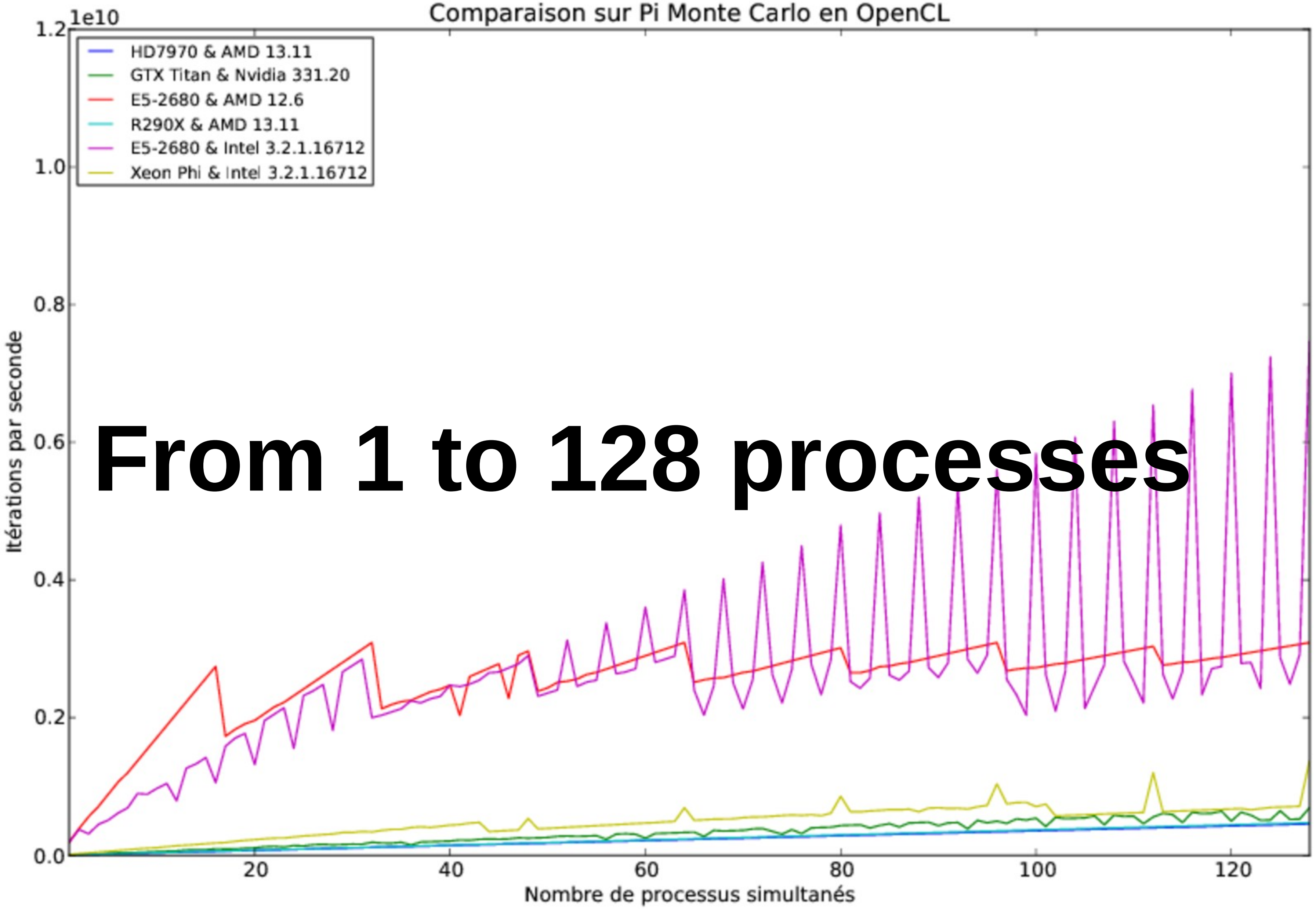


Blocks Mode



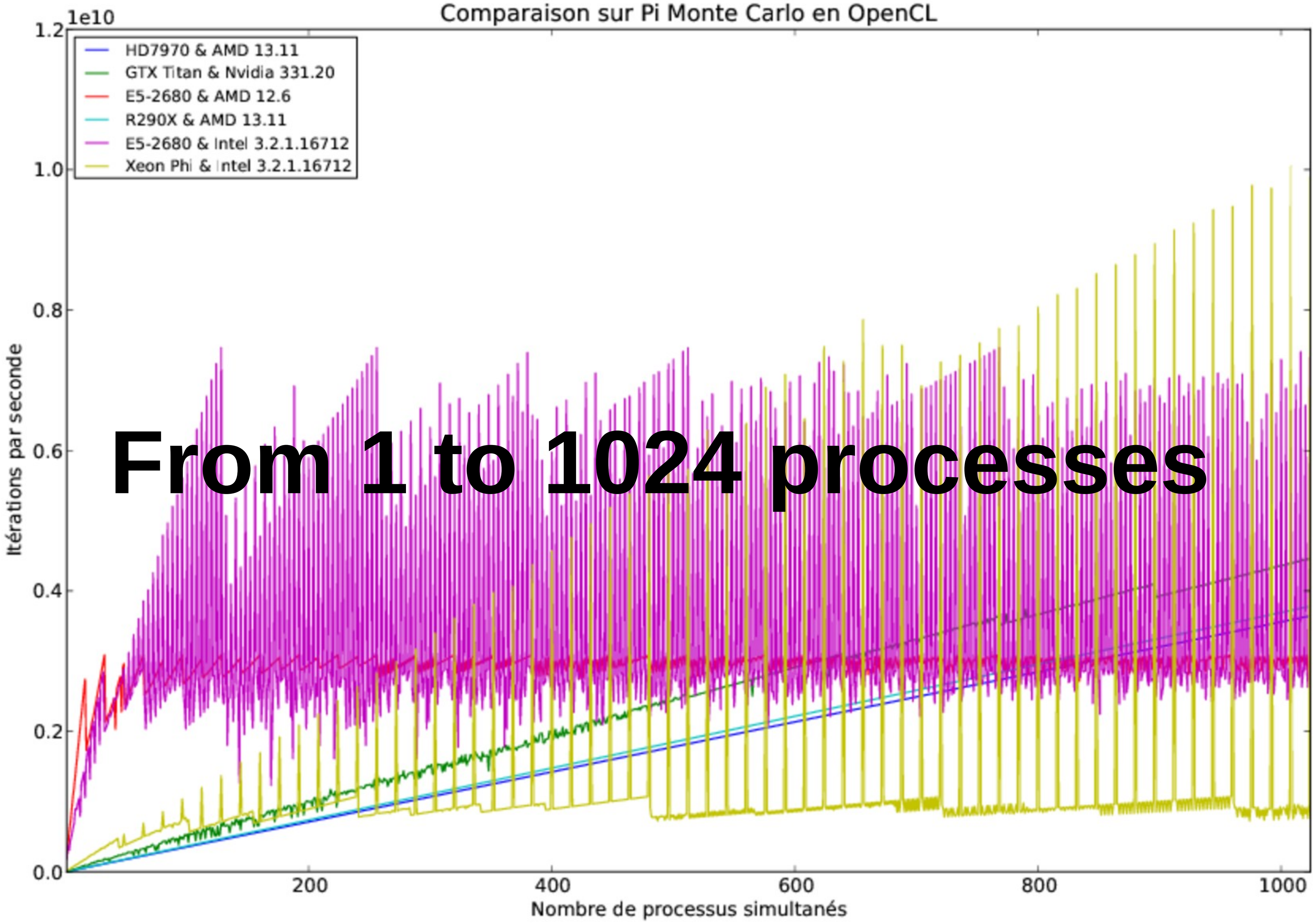
Threads Mode

Exploration of large parallelism domains Between CPU/GPU & accelerator !



From 1 to 128 processes

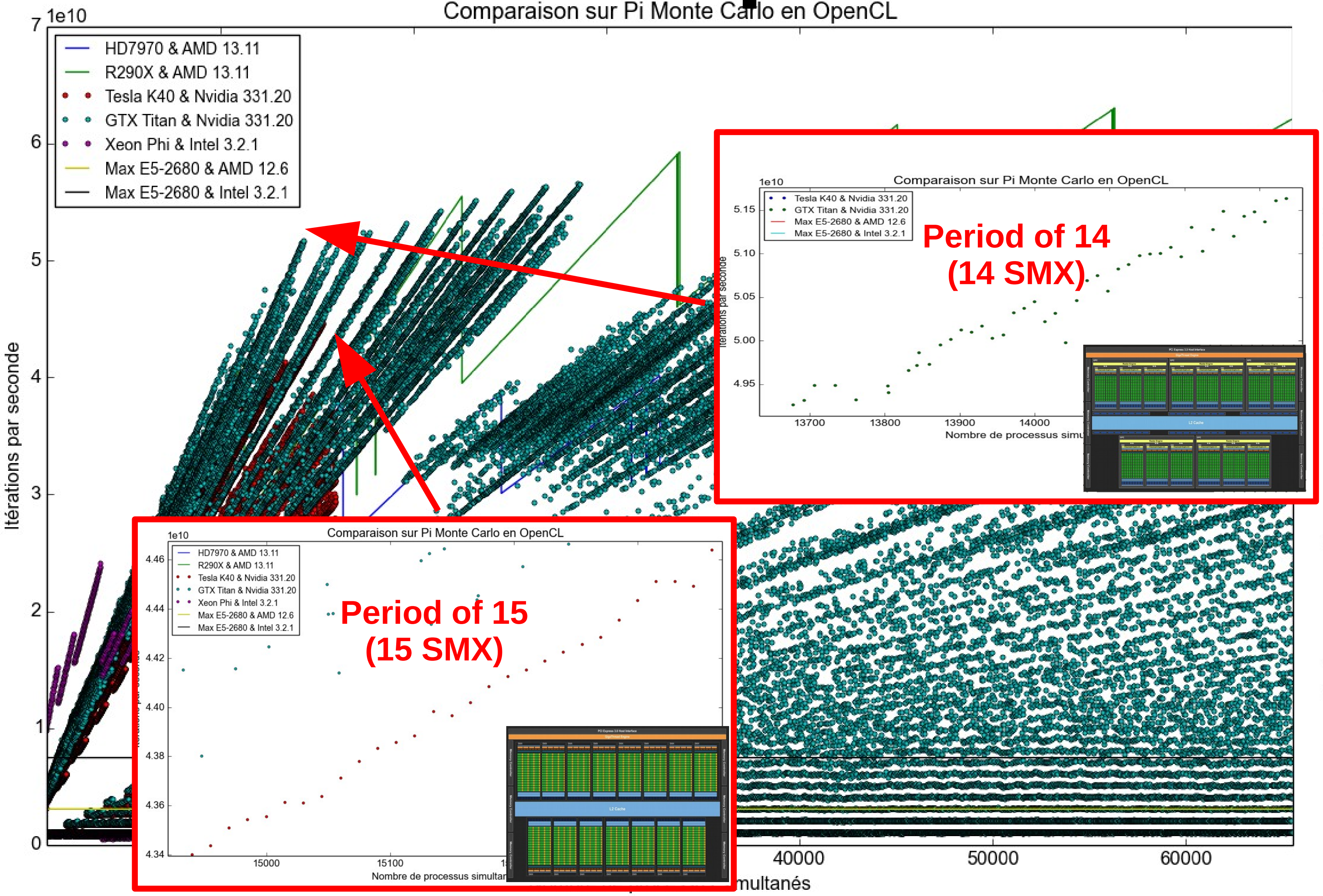
Exploration of large parallelism domains Between CPU/GPU & accelerator !



From 1 to 1024 processes

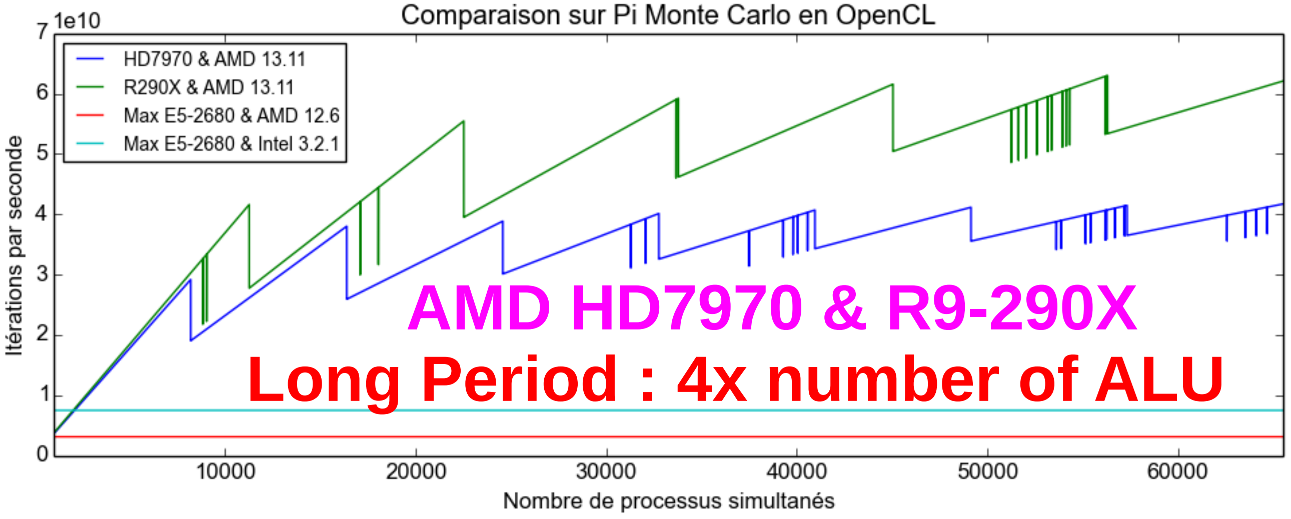
Exploration of large parallelism domains Between CPU/GPU & accelerator !

1024 to 65536 processes

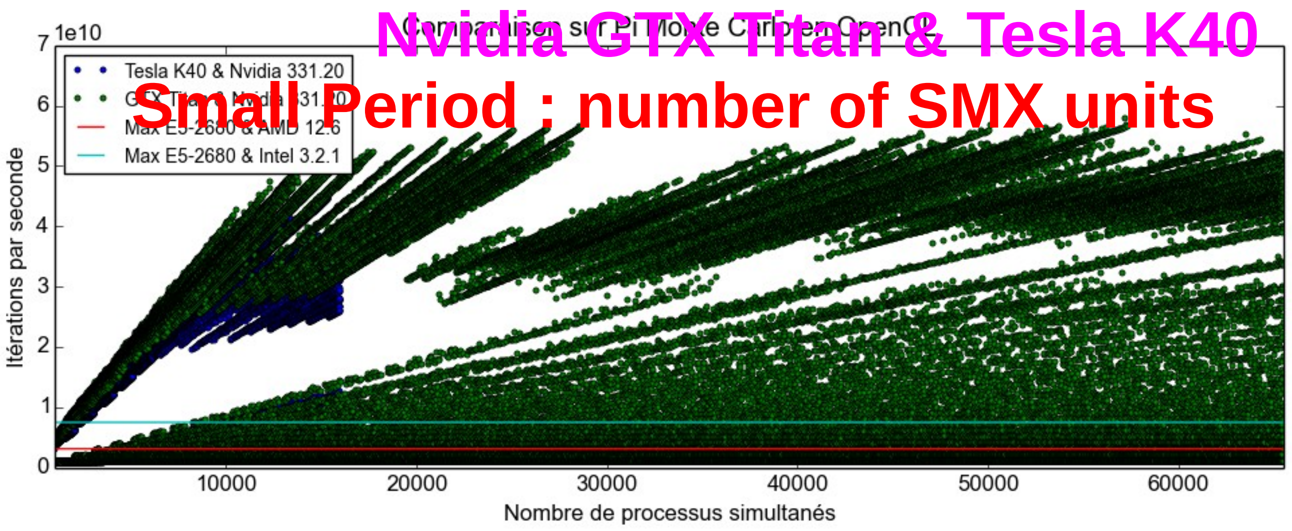


Period of 14
(14 SMX)

Period of 15
(15 SMX)



AMD HD7970 & R9-290X
Long Period : 4x number of ALU



Nvidia GTX Titan & Tesla K40
Small Period : number of SMX units

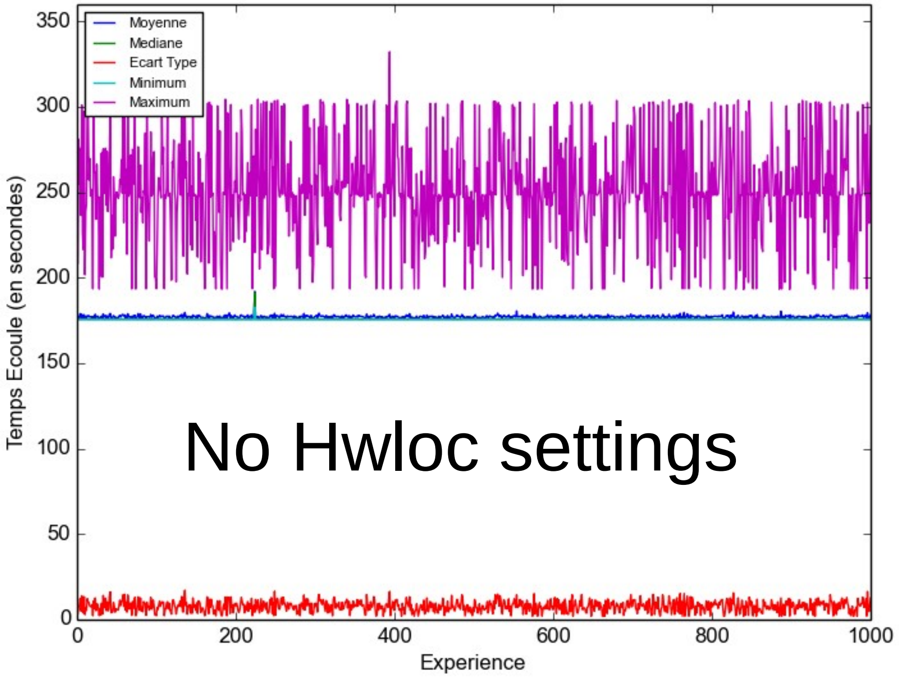
Exploration of GPU Conclusion...

- For each application its carte !
  - ALU number, RAM size, Simple Precision/Double Precision
- Parallelism degree > 500 to get (at least) GPU>CPU
- Manycore Xeon Phi between CPU et GPU
- Nvidia scheduler “strange”
  - Excellent detector for prime numbers > 1024
- AMD not to forget (linearity & DP performance)

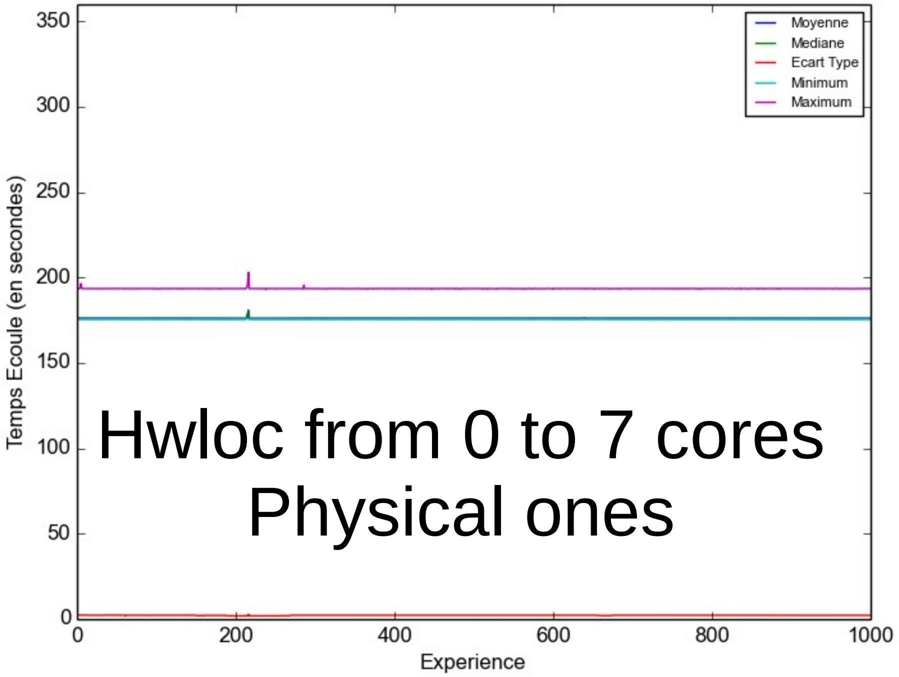
Behavior of cluster nodes
What variability in « *Embarrassing //lelism* »

- **Objective :**
  - Evaluate the scalability in MPI, what statistic to get (Average, Max, Median) ?
- **Experimental bench :**
  - 48 nodes bi-sockets 4-cores R410, interconnection Infiniband
  - Unique System SIDUS
  - Code Pi Monte Carlo distribution in MPI (10^{14} iterations)
  - Launch by *mpirun -np 384*
  - Setting locality with hwloc-bind as argument of mpirun
  - 1000 simulations

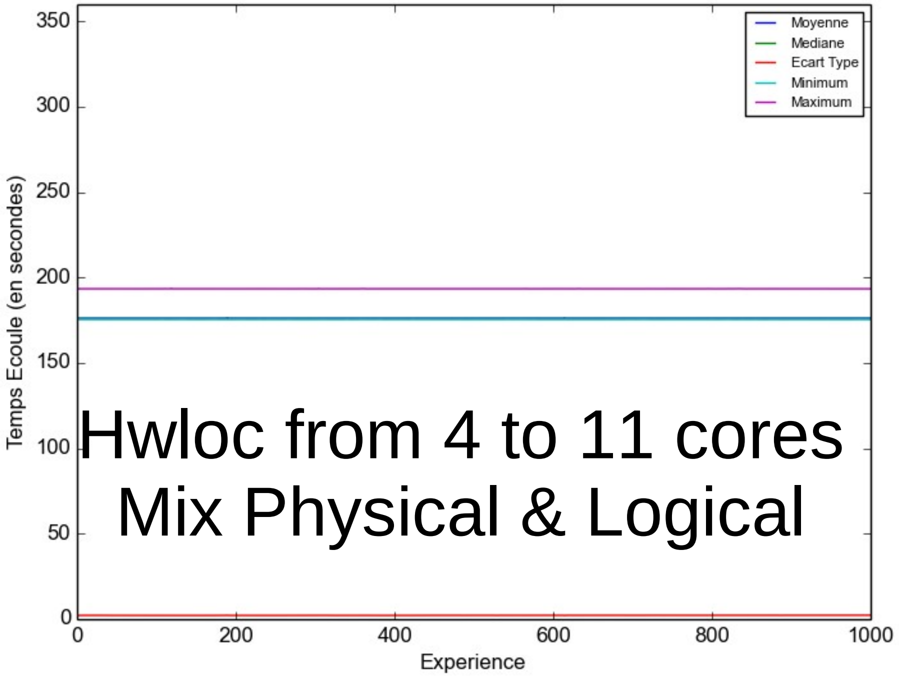
Influence of locality on large MPI deployment
1000 runs : statistics
Average/Median/Stdev/Min/Max



No Hwloc settings

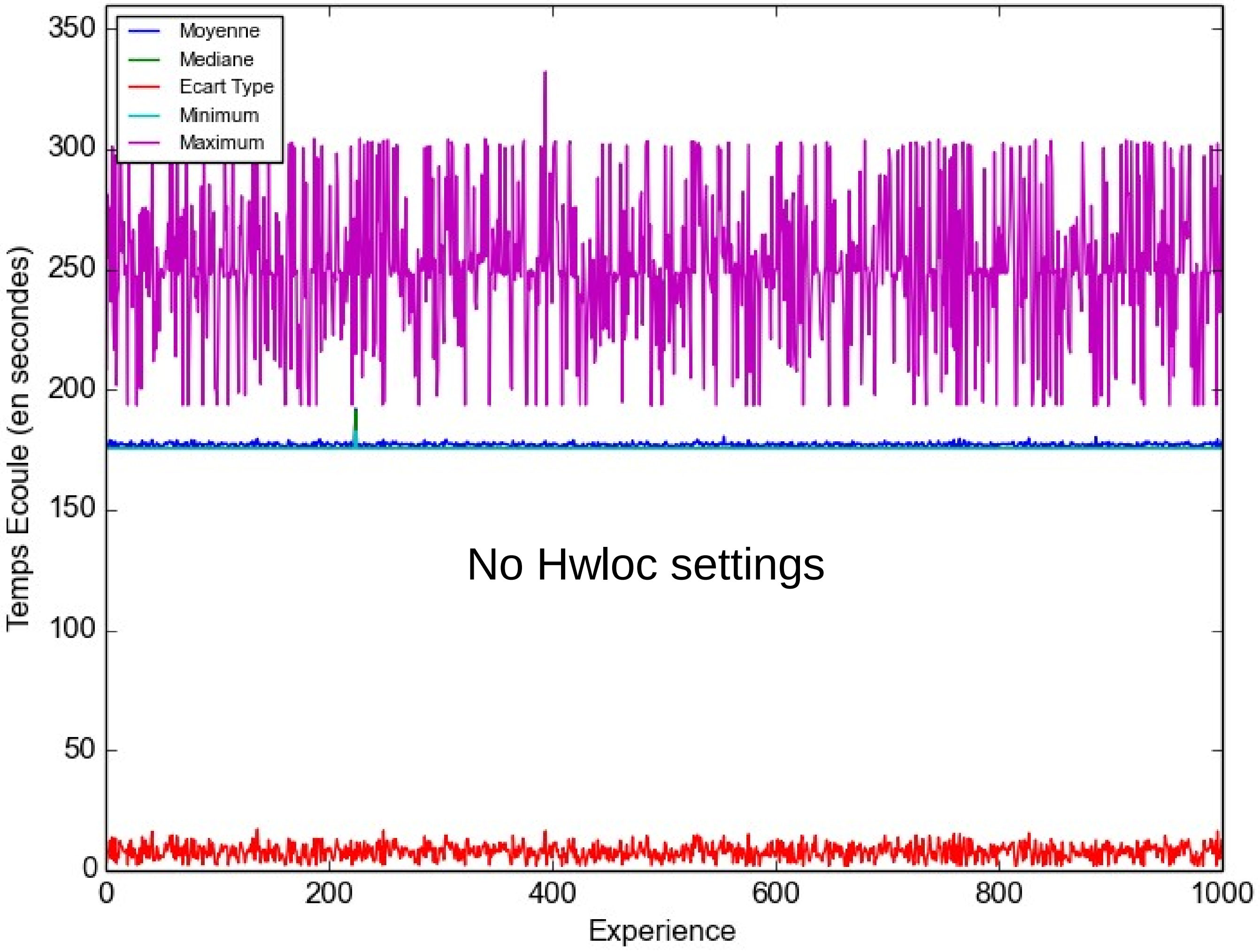


Hwloc from 0 to 7 cores
Physical ones

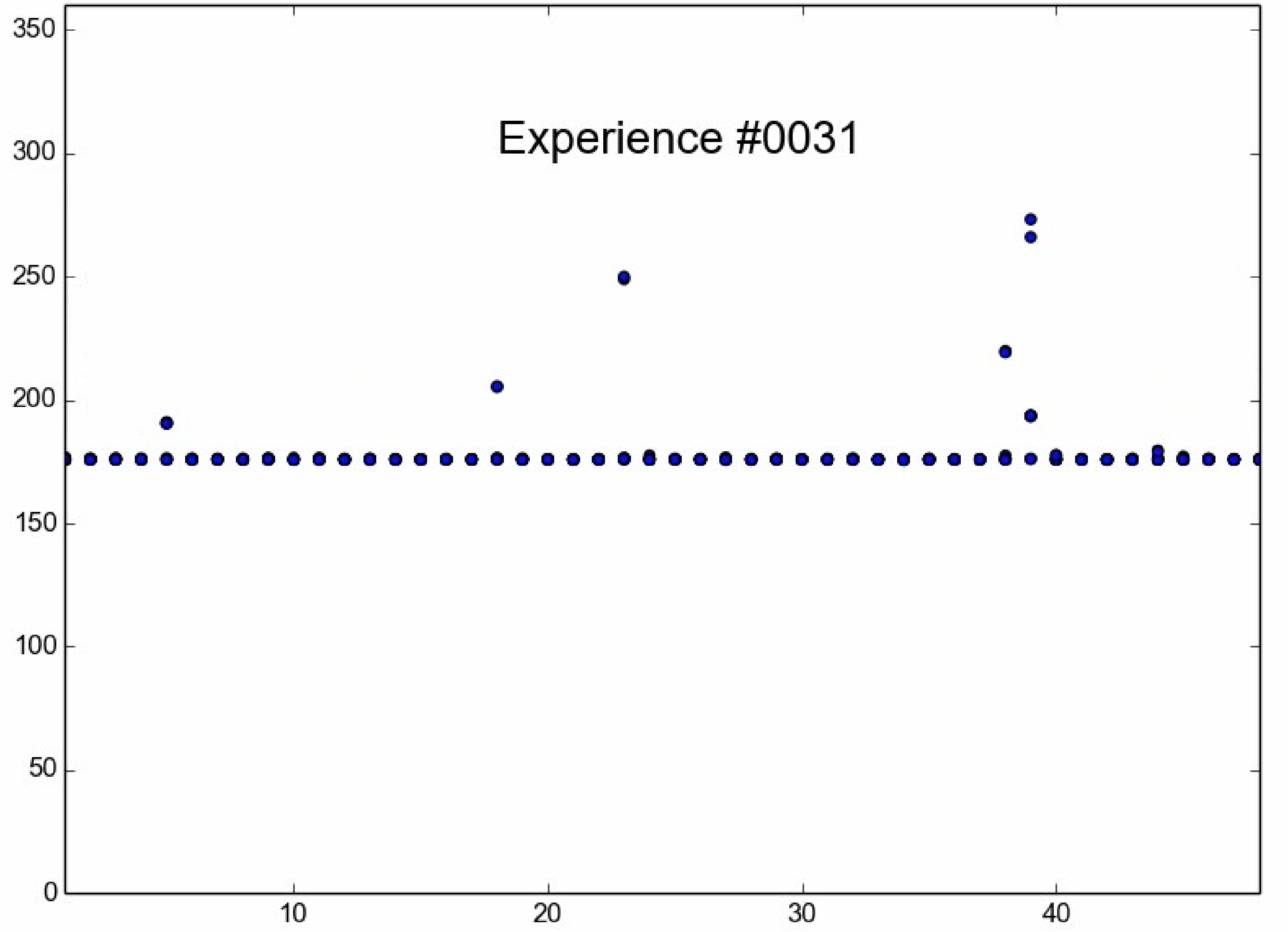


Hwloc from 4 to 11 cores
Mix Physical & Logical

1000 runs, No Locality precised
Large variability of elapsed time on ranks



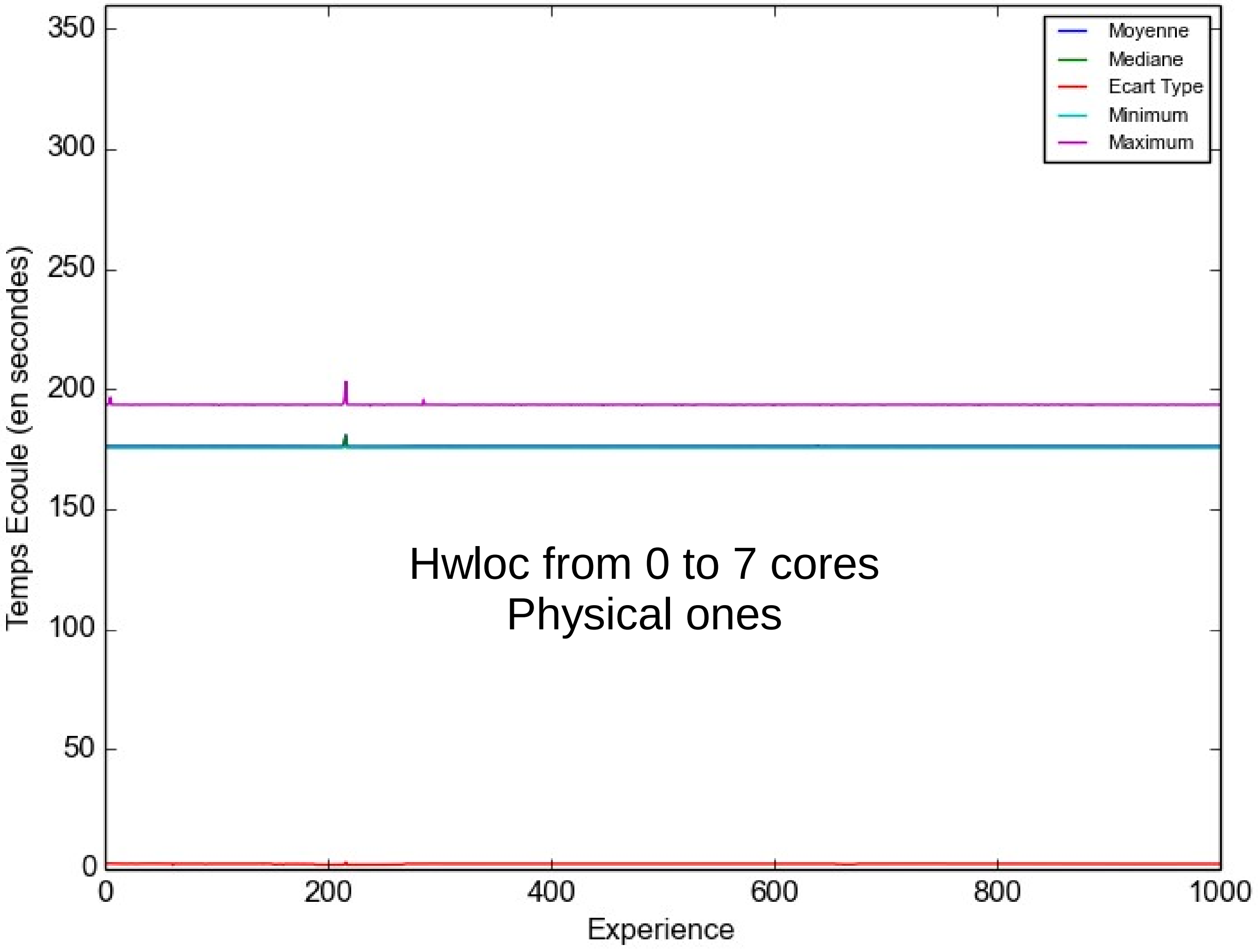
No Hwloc settings



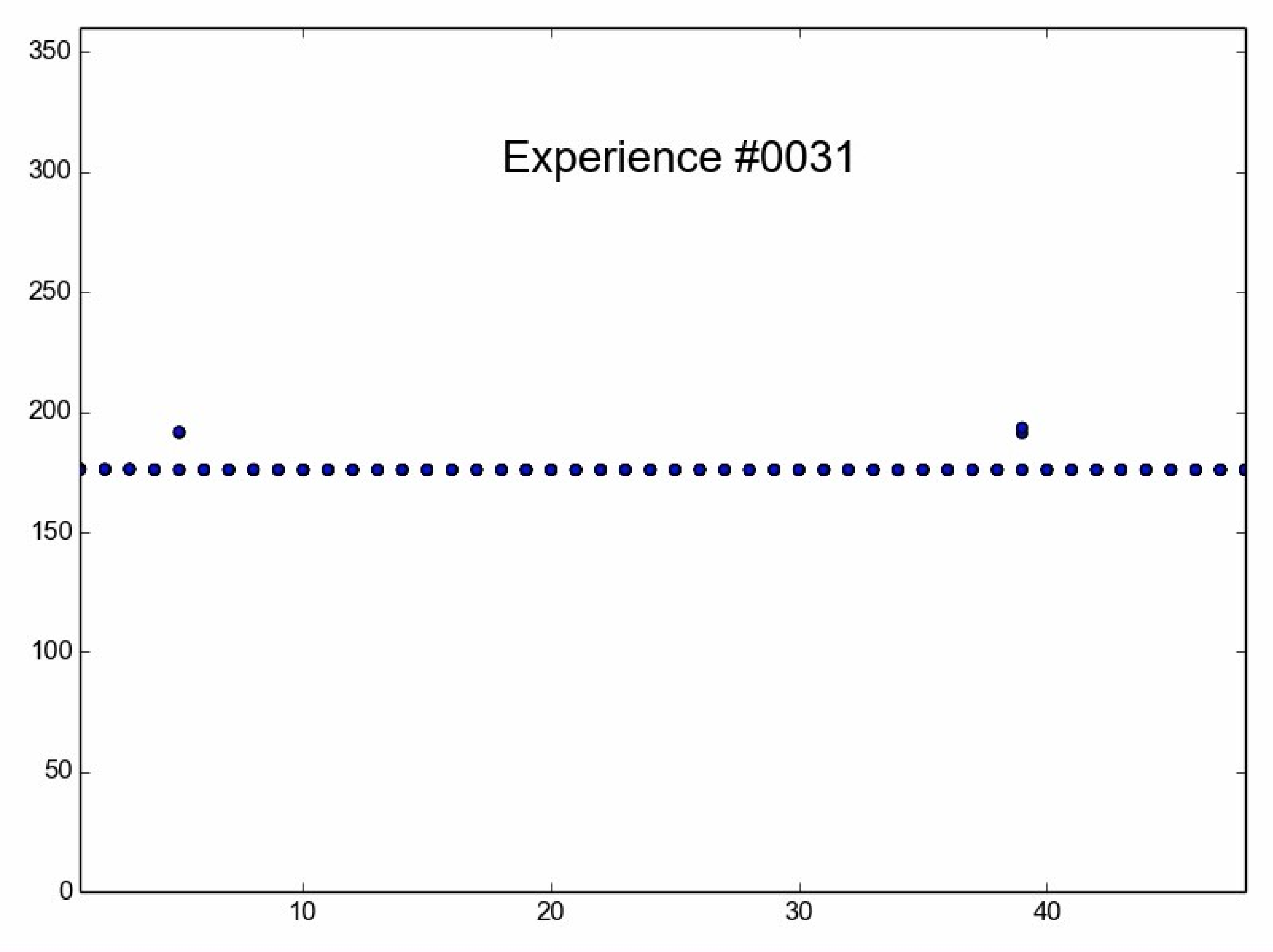
Experience #0031

Nodes from #1 to #48, 8 cores by node

1000 runs, Locality set from 0 to 7 cores
Variability fall down of elapsed time on ranks



Hwloc from 0 to 7 cores
Physical ones



Nodes from #1 to #48, 8 cores by node

Future of SIDUS

- Valuate
  - Scientific Computing, Scientific IT
  - Park management, Learning on demand
- Simplify installation & administration
  - Dedicate a machine with Read/Write access to Instance
  - Offer a direct SSH connection to instance & classical operations
  - Use Debian Preseed to simplify installation process
- Deploy Meso/Grille
- SIDUS *everywhere*
  - Launch SIDUS out of site via VPN

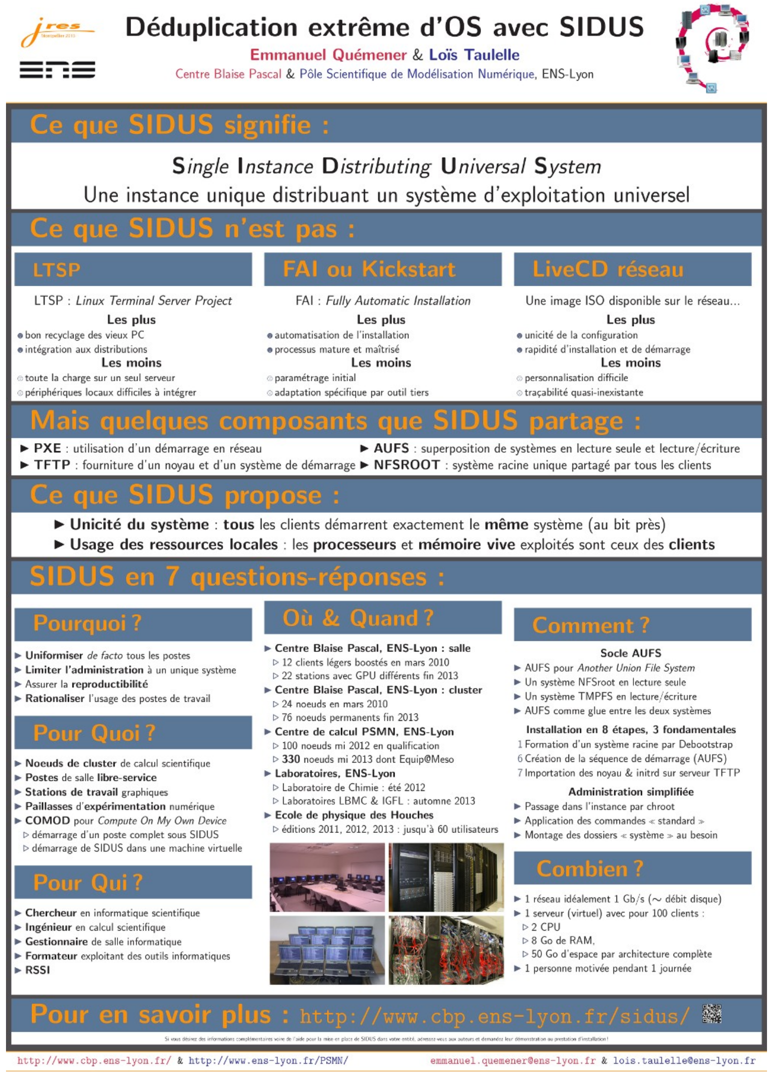
Much more information ?

http://www.cbp.ens-lyon.fr/sidus/

Linux Journal 11/2013

LINUX JOURNAL
Since 1994: The Original Magazine of the Linux Community
SPONSORED BY
New Relic.
NOVEMBER 2013 | ISSUE 235 | www.linuxjournal.com
SYSTEM ADMINISTRATION
SIDUS for Extreme Operating System Deduplication
GlusterFS Install, Benchmark and Optimize
Integrate Applications with the Python-Based ZATO PLATFORM
+ OPENSTACK SPECIAL REPORT
A LOOK AT SOLIDRUN'S CuBox Pro
PLAY CLASSIC CONSOLE GAMES ON A Raspberry Pi
Rails 4's NEW POSTGRESQL FEATURES
LJ235-Nov2013.indd 1
10/24/13 10:43 AM

Poster JRES 2013



Déduplication extrême d'OS avec SIDUS

Emmanuel Quémener & Loïs Taulelle

Centre Blaise Pascal & Pôle Scientifique de Modélisation Numérique, ENS-Lyon

Ce que SIDUS signifie :

Single **I**nstance **D**istributing **U**niversal **S**ystem

Une instance unique distribuant un système d'exploitation universel

Ce que SIDUS n'est pas :

LTSP

LTSP : *Linux Terminal Server Project*

Les plus
- bon recyclage des vieux PC
- intégration aux distributions

Les moins
- toute la charge sur un seul serveur
- périphériques locaux difficiles à intégrer

FAI ou Kickstart

FAI : *Fully Automatic Installation*

Les plus
- automatisation de l'installation
- processus mature et maîtrisé

Les moins
- paramétrage initial
- adaptation spécifique par outil tiers

LiveCD réseau

Une image ISO disponible sur le réseau...

Les plus
- unicité de la configuration
- rapidité d'installation et de démarrage

Les moins
- personnalisation difficile
- traçabilité quasi-inexistante

Mais quelques composants que SIDUS partage :

- **PXE** : utilisation d'un démarrage en réseau
- **TFTP** : fourniture d'un noyau et d'un système de démarrage
- **AUFS** : superposition de systèmes en lecture seule et lecture/écriture
- **NFSROOT** : système racine unique partagé par tous les clients

Ce que SIDUS propose :

- **Unicité du système** : **tous** les clients démarrent exactement le **même** système (au bit près)
- **Usage des ressources locales** : les **processeurs** et **mémoire vive** exploités sont ceux des **clients**

SIDUS en 7 questions-réponses :

Pourquoi ?
- **Uniformiser** *de facto* tous les postes
- **Limiter l'administration** à un unique système
- Assurer la **reproductibilité**
- **Rationaliser** l'usage des postes de travail

Pour Quoi ?
- **Noeuds de cluster** de calcul scientifique
- **Postes** de salle **libre-service**
- **Stations de travail** graphiques
- **Paillasses** d'**expérimentation** numérique
- **COMOD** pour *Compute On My Own Device*
  - démarrage d'un poste complet sous SIDUS
  - démarrage de SIDUS dans une machine virtuelle

Pour Qui ?
- **Chercheur** en informatique scientifique
- **Ingénieur** en calcul scientifique
- **Gestionnaire** de salle informatique
- **Formateur** exploitant des outils informatiques
- **RSSI**

Où & Quand ?
- **Centre Blaise Pascal, ENS-Lyon : salle**
  - 12 clients légers boostés en mars 2010
  - 22 stations avec GPU différents fin 2013
- **Centre Blaise Pascal, ENS-Lyon : cluster**
  - 24 noeuds en mars 2010
  - 76 noeuds permanents fin 2013
- **Centre de calcul PSMN, ENS-Lyon**
  - 100 noeuds mi 2012 en qualification
  - **330** noeuds mi 2013 dont Equip@Meso
- **Laboratoires, ENS-Lyon**
  - Laboratoire de Chimie : été 2012
  - Laboratoires LBMC & IGFL : automne 2013
- **Ecole de physique des Houches**
  - éditions 2011, 2012, 2013 : jusqu'à 60 utilisateurs

Comment ?

Socle AUFS
- AUFS pour *Another Union File System*
- Un système NFSroot en lecture seule
- Un système TMPFS en lecture/écriture
- AUFS comme glue entre les deux systèmes

Installation en 8 étapes, 3 fondamentales

1 Formation d'un système racine par Debootstrap
6 Création de la séquence de démarrage (AUFS)
7 Importation des noyau & initrd sur serveur TFTP

Administration simplifiée
- Passage dans l'instance par chroot
- Application des commandes « standard »
- Montage des dossiers « système » au besoin

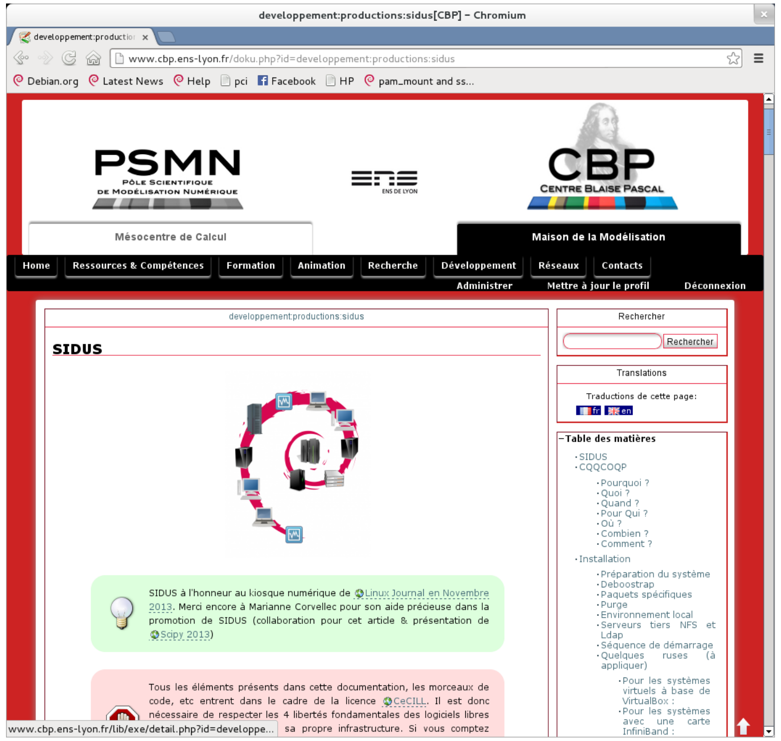
Combien ?
- 1 réseau idéalement 1 Gb/s (~ débit disque)
- 1 serveur (virtuel) avec pour 100 clients :
  - 2 CPU
  - 8 Go de RAM,
  - 50 Go d'espace par architecture complète
- 1 personne motivée pendant 1 journée

Pour en savoir plus : http://www.cbp.ens-lyon.fr/sidus/

Si vous désirez des informations complémentaires voire de l'aide pour la mise en place de SIDUS dans votre entité, adressez-vous aux auteurs et demandez leur démonstration ou prestation d'installation !

http://www.cbp.ens-lyon.fr/ & http://www.ens-lyon.fr/PSMN/ emmanuel.quemener@ens-lyon.fr & lois.taulelle@ens-lyon.fr

Web Site CBP



SIDUS

SIDUS à l'honneur au kiosque numérique de Linux Journal en Novembre 2013. Merci encore à Marianne Corvellec pour son aide précieuse dans la promotion de SIDUS (collaboration pour cet article & présentation de Scipy 2013)

Tous les éléments présents dans cette documentation, les morceaux de code, etc entrent dans le cadre de la licence CeCILL. Il est donc nécessaire de respecter les 4 libertés fondamentales des logiciels libres … sa propre infrastructure. Si vous comptez

Table des matières

As conclusion... Back to Past

- Few decades before French Revolution (1789)
  - Corporatism : Hide knowledge to people
  - Two Scientists : Diderot & D'Alembert
  - *Encyclopédie du savoir, des sciences, des techniques*
- In 2015, in software environment
  - Corporatism : Close software to people, jailed infrastructures
  - Few developers (~1000)
  - Debian Distribution : Encyclopedia of working software

Ending with a tiny joke !

How do you call people speak 3 languages ?

  - Trilingual people !
- How do you call people speak 2 languages ?
  - Bilingual people !
- How do you call people speak 1 language ?
  - French people !
- Yes, it's the same joke, but...
  - I'm aleady french (no change in 45 minutes) :
  - Questions ? Please speak sllllooooowwwwly !

Iconographie

- http://en.wikipedia.org/wiki/Antikythera_mechanism
- http://www.nasa.gov/centers/dryden/news/FactSheets/FS-008-DFRC.html
- http://en.wikipedia.org/wiki/Antikythera_mechanism
- http://congrex.nl/icso/Papers/Session%2014a/FCXNL-10A02-1977297-1-BERGERON_ICSO_PAPER%20.pdf
- http://upload.wikimedia.org/wikipedia/commons/8/8b/Babbage_Difference_Engine.jpg
- http://www.earsel.org/Advances/2-1-1993/2-1_22_Harger.pdf
- http://en.wikipedia.org/wiki/File:NewmarkAnalogueComputer.jpg
- ...