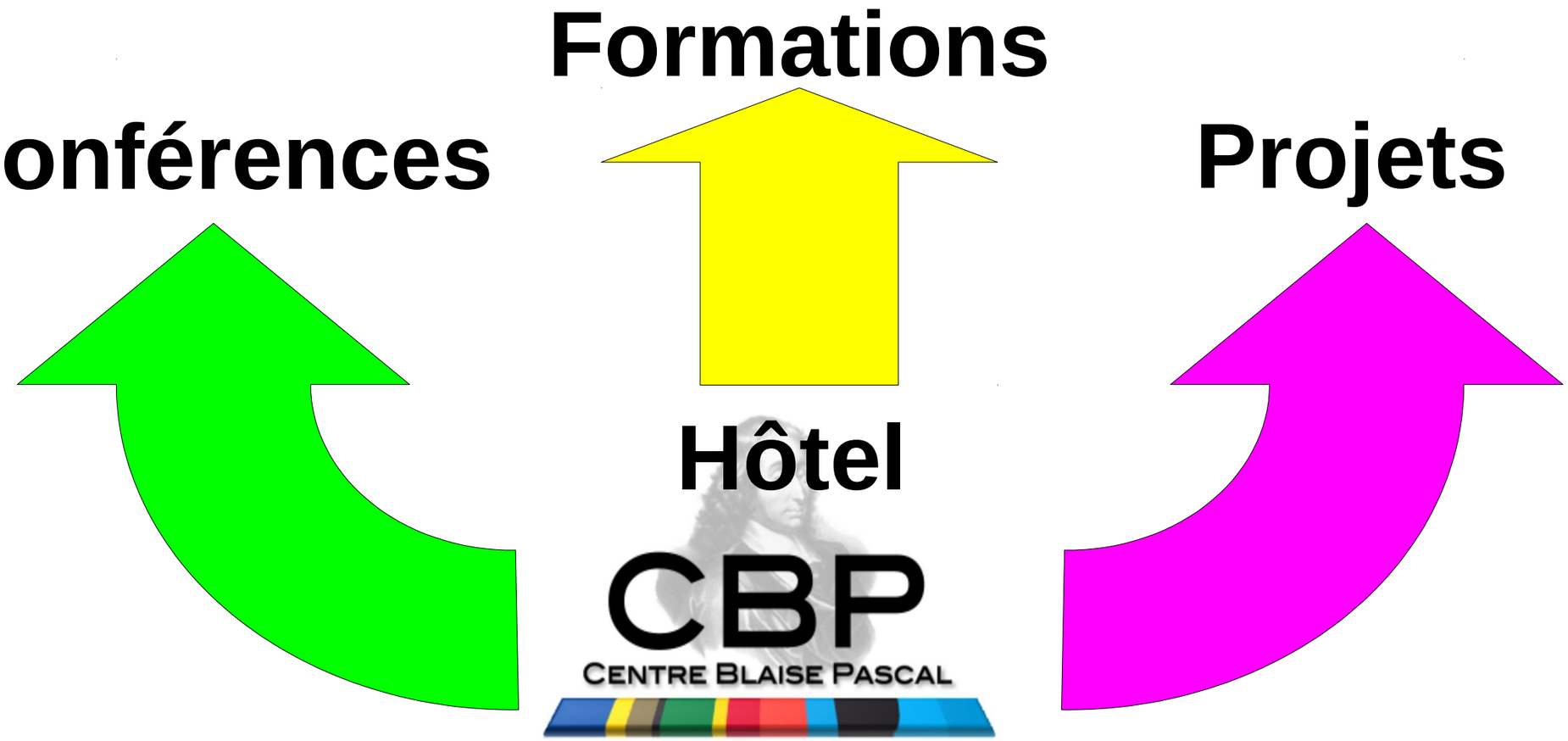
SIDUS
Déduplication extrême d'OS & reproductibilité

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

Une Instance Unique Distribuant un Système d'Exploitation Universel

Un couteau suisse pour le calcul scientifique & ailleurs.

Qu'est-ce que le Centre Blaise Pascal ?
Hôtel à projets & Maison de la modélisation
Avant tout, « outil » de recherche...



« Catalyser » l'informatique scientifique : CBP
Maison de la Simulation, Plate-forme expérimentale
Centre d'essais

- Nasa X29
  - Cellule de F5
  - Moteur de F18
  - Servos de F16
- Études
  - Flèche inversée
  - Incidence >50°
  - « Fly-By-Wire »

Le CBP (via son pilote d'essais) : réutilise, met à disposition et explore...

Ce que SIDUS n'est pas...
Premièrement : SIDUS n'est pas SIDIOUS !

Darth SIDIOUS

alias Palpatine

Sidus : “constellation” en Latin

Différence entre Sidious & SIDUS : IO (Input/Output)

De SIDUS à SIDIOUS avec des problèmes I/O ?

Nous allons voir !

Ce que SIDUS n'est pas...
Mais ce que SIDUS partage avec eux !

Ce que SIDUS n'est PAS !

- **LTSP :** *Linux Terminal Server Project*
  - Un serveur avec la “charge”, administration simplifiée du client
- **FAI, Kickstart, Debian Installer Preseed :**
  - *« Lorsque la machine remplace l'humain...*
    dans la procédure d'installation... »
- **LiveCD par réseau :**
  - Une image ISO distribuée par le réseau

Mais ce que SIDUS partage avec eux

Boot PXE, TFTP, NFSroot, (**AUFS)**

Ce dont SIDUS se rapproche
Ce que c'est finalement

De quels autres projets ça se rapproche...

- FaDDeF : http://projets.mathrice.org/faddef/
- DRBL : http://drbl.org/ (à confirmer...)

Et concrètement, c'est quoi ?

- Ce n'est pas un logiciel !
- Pas de paquet à installer, tout existe déjà !
- Juste un ensemble de commandes à appliquer...

C'est une approche :

- partager simplement & efficacement un OS
- *Avec SIDUS, je n'installe plus, je démarre les machines !*

Les deux principales propriétés de SIDUS
Reproductibilité dans l'espace-temps

- **Unicité de configuration**
  - Deux clients SIDUS : le même OS au bit près !
- **Exploitation des ressources locales**
  - Processeurs & RAM (& extra...) exploités : ceux des clients !
- **Reproductibilité ?** Pour un SIDUS inchangé
  - Stabilité dans le temps (pour un client défini)
    - Deux démarrages consécutifs sur une même machine : même système
  - Stabilité dans l'espace (pour deux ou plus clients différents)
    - Deux clients démarrant simultanément : même système
  - Deux machines **ne peuvent pas ne pas avoir** le même OS !

SIDUS en 7 Questions : CQQCOQP

Exemple de méthode analytique avec CQQCOQP :

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

La méthode la plus simple pour décrire quelque chose...

- Très utile en journalisme (théoriquement...)
- Très utile en gestion de projets (pour le pragmatique...)

SIDUS en 7 Questions : CQQCOQP
(Ben) Pourquoi ?

Pourquoi ?

- Pour uniformiser de facto tous les clients
- Pour limiter l'administration d'un parc à un poste
- Pour comparer des matériels différents / base unique
- Pour récupérer des “fluides” (Watts & BTU)
- Pour rationaliser l'usage des stations de travail
- Pour investiguer du stockage sous anesthésie
- Pour s'assurer de la reproductibilité OS & applications

SIDUS en 7 Questions : CQQCOQP
Pour Quoi ? Pour Qui ?

Pour Quoi ?

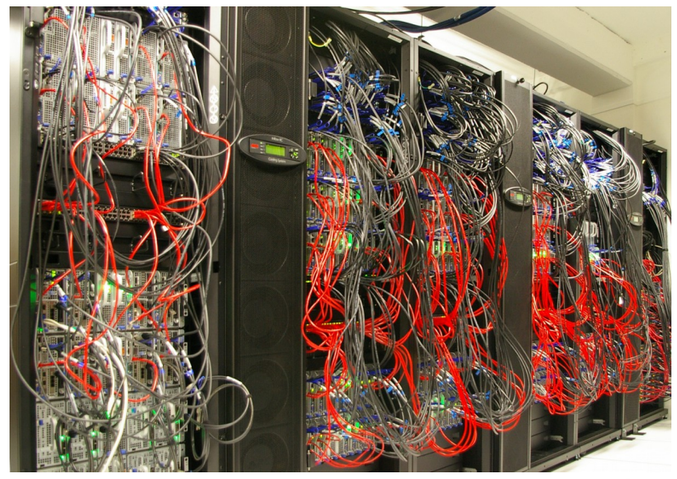
- Nœuds de cluster en HPC
- Postes de travail
- Stations graphiques
- Paillasses numériques
- Poste *COMOD*
  - *Vous connaissez BYOD ?*
    - *Mon matos pour bosser*
  - *Compute On My Own Device*
    - *Mon matos pour calculer !*

Pour Qui ?

- Ingénieur en informatique
- Administrateur de salle info
- Chercheur en informatique
- Professeur d'outils complexes
- Responsable en Sécurité des Systèmes d'Information
- What else ?

SIDUS en 7 questions : CQQCOQP Où & Quand ?

- **Centre Blaise Pascal**, ENS-Lyon : salle informatique
  - 12 Neoware en 2010Q1, 24 stations 2015Q2
- **Centre Blaise Pascal**, ENS-Lyon : cluster
  - 24 nœuds in 2010Q1, 76 nœuds en 2015Q2
- **Centre de calcul PSMN**, ENS-Lyon
  - 100 nœuds 2012Q2, **480** nœuds 2015Q2
  - Tout Equip@Meso
- Laboratoires, ENS-Lyon
  - Chimie, **IGFL**, LBMC, UMPA, RDP
- École de physique des Houches
  - Editions de 2011 à 2015



SIDUS en 7 questions : CQQCOQP
(Dis) Comment (ça marche) ?

- AUFS : *Another Union File System*
  - Agrégation de systèmes de fichiers : ruse de LiveCD
  - 4 étapes :
    1. Monter en NFSroot en lecture seule l'OS sur un point de montage
    2. Créer un système de fichiers temporaire TMPFS sur un second
    3. Lier les deux précedents dossiers avec AUFS
    4. Offrir le dossier résultant comme racine de l'OS
  - Comportement d'un système de fichier en Lecture/Écriture
  - Au redémarrage toute modification disparaît
- Un prérequis : chroot pour l'installation initiale (& administration)

SIDUS en 7 questions : CQQCOQP, la fin !
Comment installer : SIDUS en 7 étapes (de Etch à Wheezy)

Pour des serveurs DNS, DHCP, TFTP, NFS bien configurés...

1) Création par Debootstrap d'une nouvelle racine exportée par NFS
2) Création d'un “cordon ombilical” avec l'hôte
   - Montage des dossiers /proc /sys /dev/shm
3) Installation (& purge de certains paquets non sollicités)
4) Adaptation à l'environnement local
   - Fuseau horaire, clavier, localisation, serveur de fichiers utilisateurs, authentification.
5) Création d'une séquence de démarrage avec AUFS
   - Copie du script rootaufs file in /etc/initramfs/scripts/init-bottom
   - Lancement de update-initramfs -k all -u
6) Importation du noyau & du initrd spécifique sur le serveur TFTP
7) Décrochage du “cordon ombilical” avec l'hôte

Migration vers Debian Jessie
Limitations & évolutions

- **Objectif : fournir un socle pour des hyperviseurs**
  - Contexte : au démarrage par DHCP, initramfs avec interface physique, pas un bridge
  - Solution : activer le bridge au démarrage, impossible avec initramfs-tools, passage à dracut
- **Objectif : améliorer la confidentialité des données utilisateurs (réseau...)**
  - Contexte : NFSv4 nécessite une ouverture trop lâche aux dossiers utilisateurs
  - Solution : Kerberos est trop contraignant, CIFS avec les extensions Posix & pam_mount
- **Objectif : conserver le fonctionnement de rootaufs en Jessie**
  - Contexte : jusqu'à Wheezy, rootaufs opérationnel dans init-bottom, ça casse en Jessie
  - Solution : changement profond de initramfs-tools ou passage à dracut
- **Objectif : limiter le nombre d'instances SIDUS pour matos spécifiques**
  - Contexte : De une Nvidia, une AMD/ATI, une 64 bits & une 32 bits pour VirtualBox
  - Solution : configuration à la volée au démarrage pour les cartes graphiques

SIDUS en 7 questions : CQQCOQP, la fin !
Comment installer : SIDUS en 7 étapes pour Jessie

Pour des serveurs DNS, DHCP, TFTP, NFS bien configurés...

1)Création par Debootstrap d'une nouvelle racine exportée par NFS

2)Création d'un “cordon ombilical” avec l'hôte

- Montage des dossiers /proc /sys /dev/shm

3)Installation (& purge de certains paquets non sollicités)

4)Adaptation à l'environnement local

- Fuseau horaire, clavier, localisation, serveur de fichiers utilisateurs, authentification.

5)Création d'une séquence de démarrage avec AUFS

- **Modification du aufs-mount.sh cassé de Dracut par le SIDUS**
- **Changement du bail DHCP passé à « forever » dans le script dclient-script.sh**
- **Quelques ruses pour éviter des problèmes au démarrage (ex autofs/dbus)**

6)Importation du noyau & du initrd spécifique sur le serveur TFTP

7)Décrochage du “cordon ombilical” avec l'hôte

SIDUS en 7 questions : la fin ! Comment administrer ?

- Une limitation : le `/proc` doit être unique..
  - Grande vigilance sur les processus
    - Manipulation of Java, compilation with optimization, installation
- La truand :
  - Passage par le chroot,
  - Opérations classiques marchant à 90%
- La brute :
  - Passage par le chroot,
  - Etablissement du “cordon ombilical”
  - Opérations classiques
  - Libération du “cordon ombilical”
- La bonne :
  - Démarrage d'une machine avec accès en Lecture/Ecriture du NFSroot
  - Opérations comme sur une machine standard

SIDUS en 7 questions : la tout fin ! Combien ça coûte ?

- Un réseau idéal : Gigabit Ethernet au client, 10G au serveur
  - Mais ça tourne correctement sur un réseau 100 Mb/s !
- Un serveur idéal : 4 CPU, 16 GB RAM, 10G, SSD
  - Mais cela fonctionnait pour 330 nœuds sur un v(eau)40z au PSMN !
- Un client idéal : toutes les machines identiques
  - Mais cela fonctionne pour 16 types de machines au PSMN, 10 au CBP (et 30 cartes graphiques différentes)
- Un intégrateur & administrateur idéal (alias motivé) : ;-)
  - Deployé par L. Taulelle avec des rushs de documentation : PSMN
  - Déployé par T. Bellembois avec la documentation en ligne : IGFL

Démonstration locale

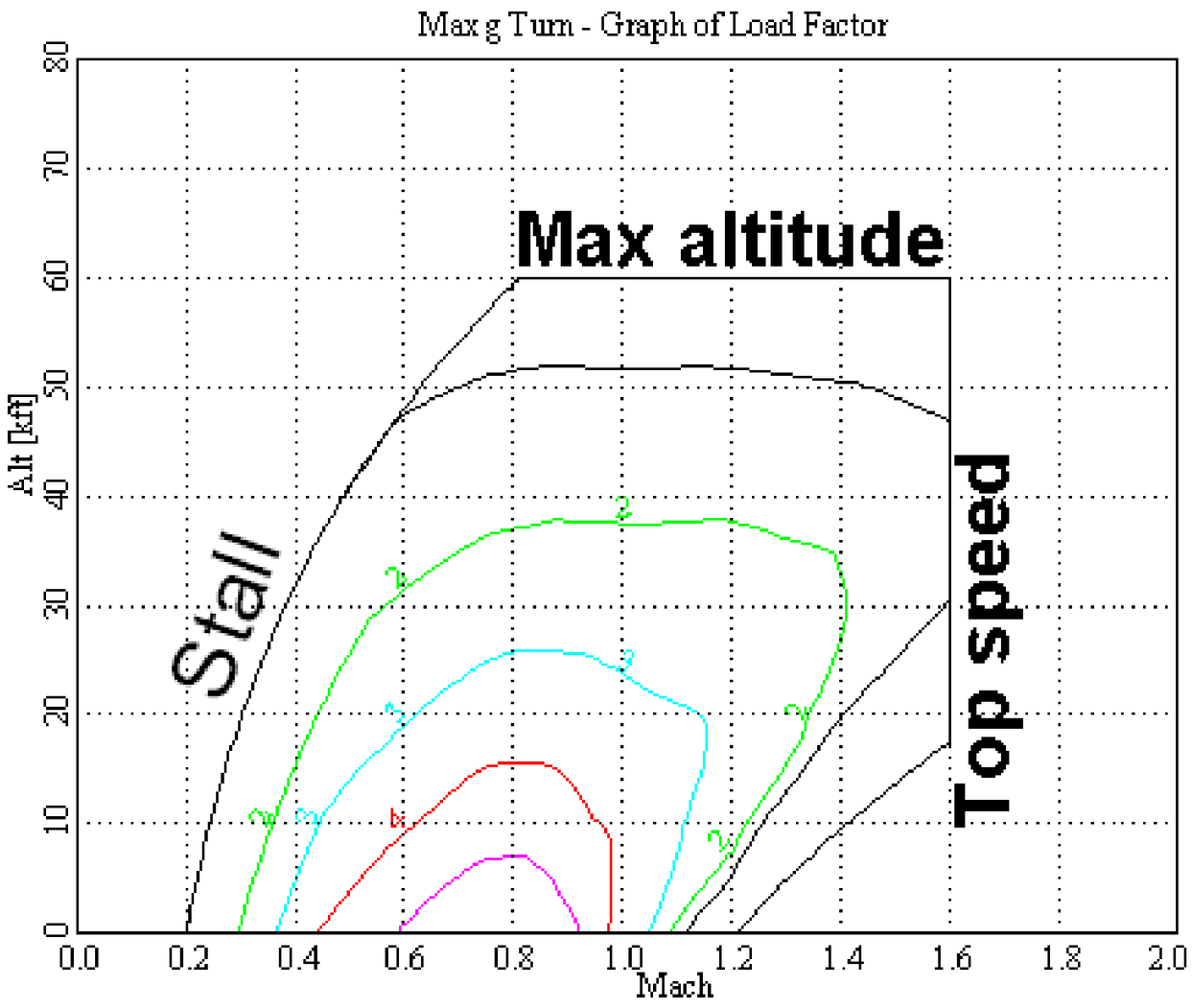
- Démarrage de l'hôte de SIDUS
  - Passerelle, serveurs DNS, DHCP, TFTP, NFS
- Démarrage d'un client local sur LiveCD
  - Pas de Stellarium
- Démarrage d'un client local sous SIDUS
  - Démonstration de Stellarium
- Démarrage d'un client distant sous SIDUS
  - Démonstration de Stellarium

SIDUS en reproductibilité
Applications sur le Stockage & Parallélisme

- Emergence de domaines de parallélisme pour logiciels & matériels
- Pertinence de GlusterFS comme scratch distribué en HPC
  - Influence du BIOS sur la performance et la variabilité variability
- Comparison de GPU & influence dans les hauts niveaux de parallélisme
  - La Variabilité comme facteur discriminant entre des GPU différents.
- Variabilité d'exécution dans la distribution par MPI
  - Difficulté de l'estimation de temps d'exécution & influence de la localité

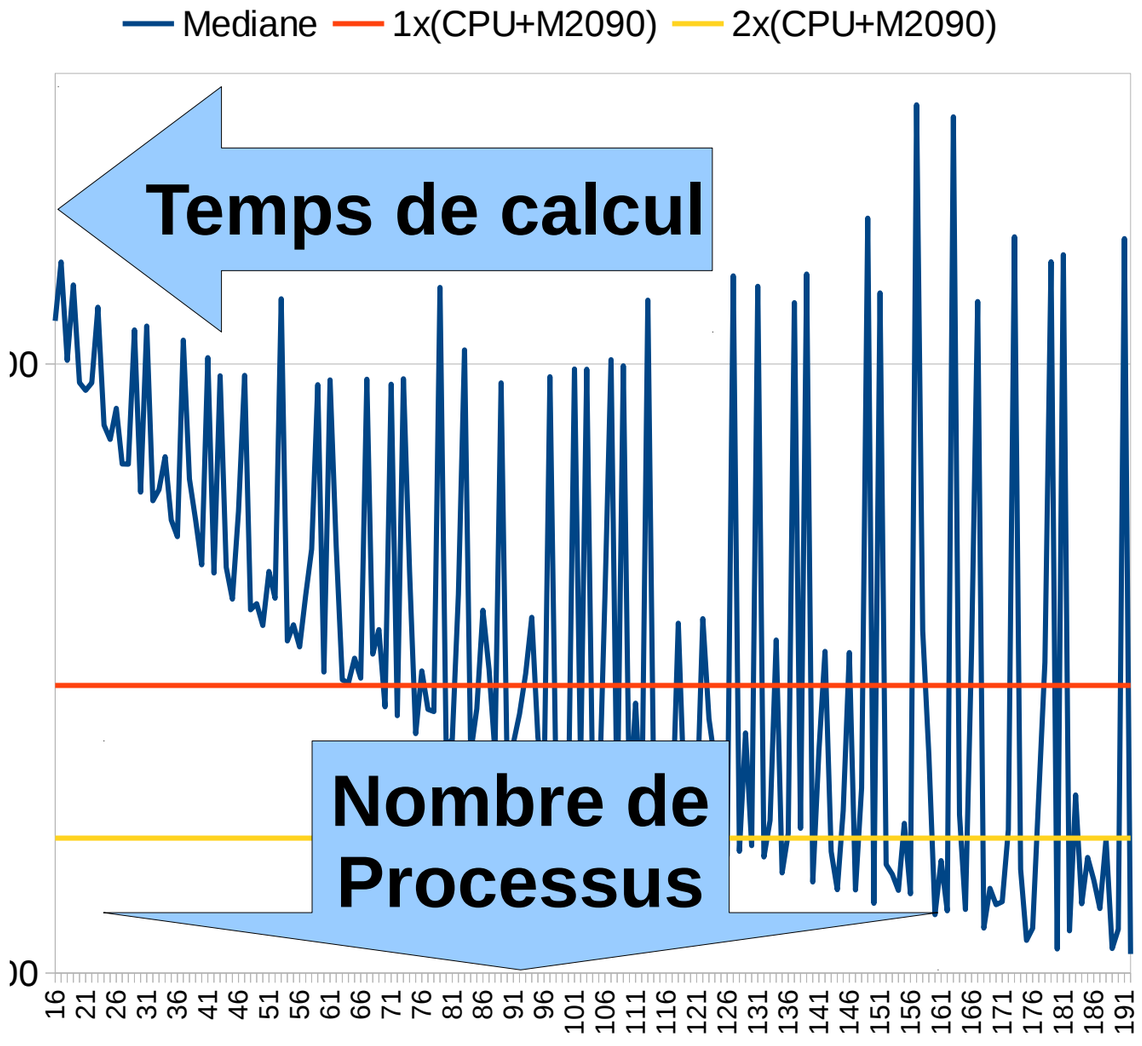
De l'enveloppe de vol à l'enveloppe de parallélisme La fin de la dualité : « marche/marche pas »

Enveloppe de vol



Vitesse/altitude/Force G

Enveloppe de parallélisme



Parallélisme/Mémoire/GPU

Variabilité sur silicium dans l'espace-temps
Quelle manœuvrabilité ?

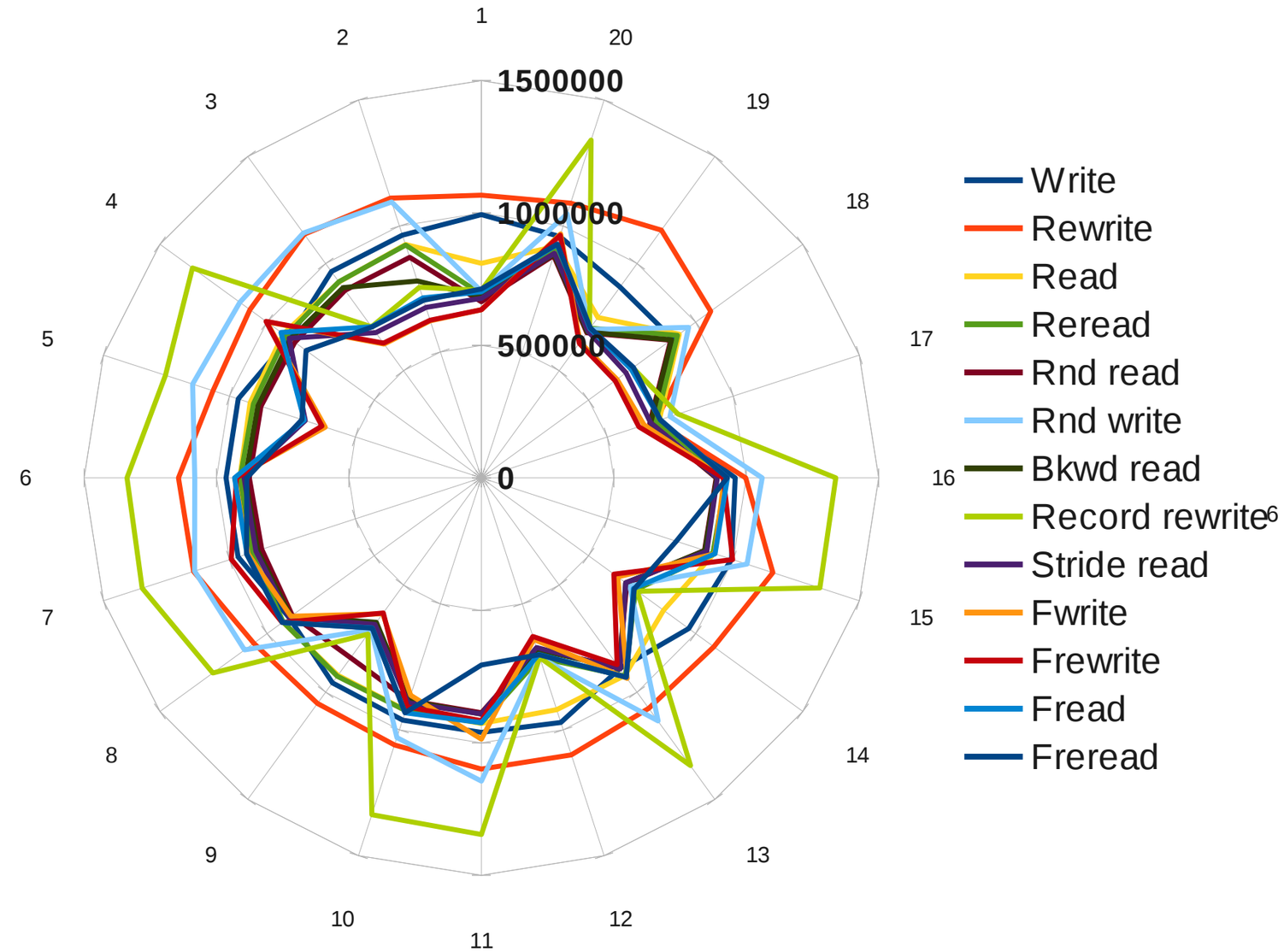
- **Temps** : même machine, instants différents ?
- **Espace** : même instant, différentes machines ?
- Les solutions :
  - Restauration par déploiement de l'image d'un OS
    - Replicator, SystemImager, MondoRescue, ...
    - Kadeploy sur Grid'5000
    - Boot iSCSI avec cuisine d'instantanés (sur LVM, ZFSonLinux, BtrFS)
  - Installation avec la même procédure d'installation :
    - FAI, Kickstart, Debian-Installer Preseed
  - **SIDUS : *Single Instance Distributing Universal System***

Défaut de reproductibilité ?
Exemple d'illustration : GlusterFS/IOZone

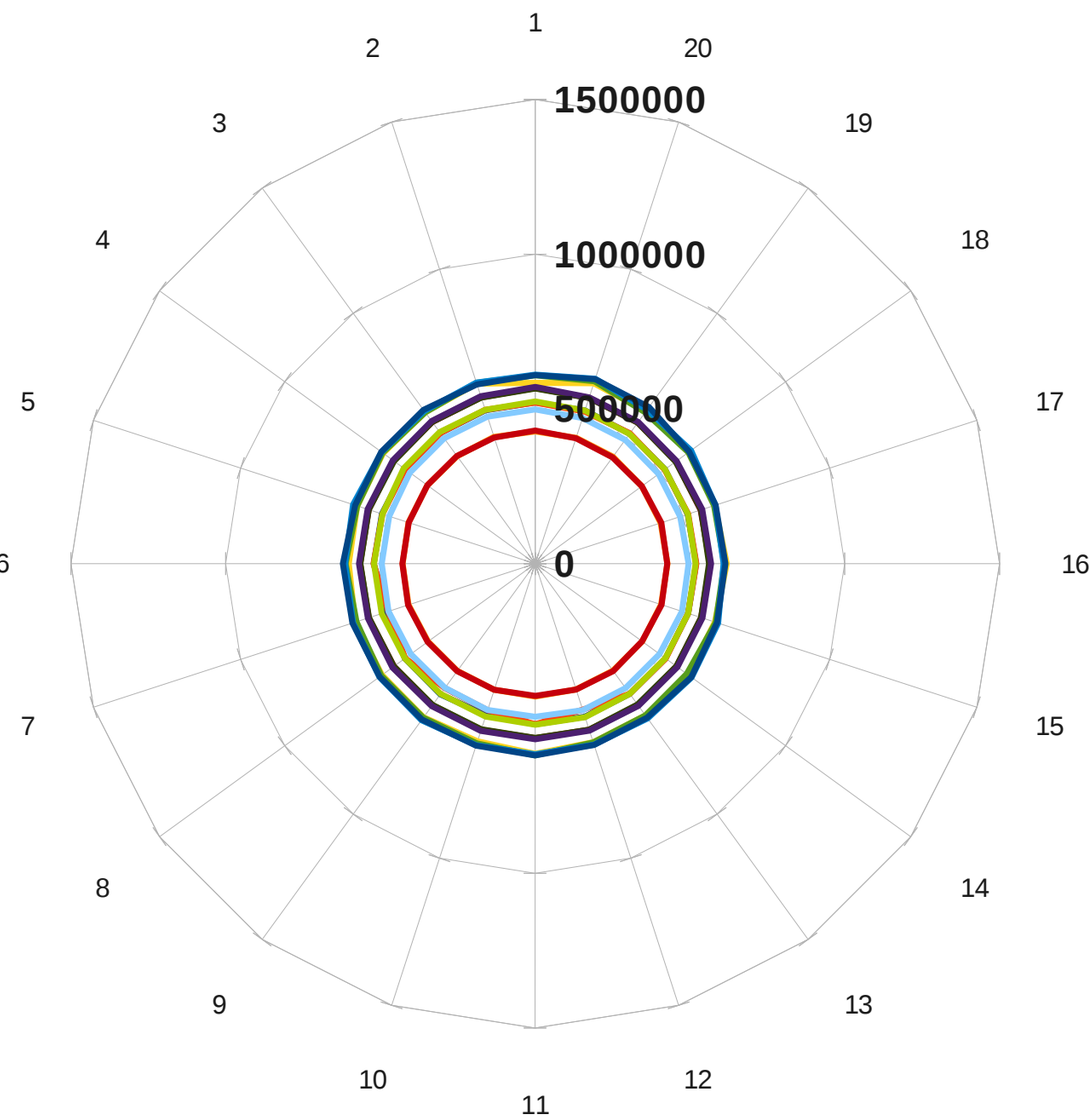
- **Objectif** :
  - Évaluation de GlusterFS comme scratch Haute Performance
- **Banc d'essai** : 20 nœuds + infrastructure
  - 20 nœuds Sandy Bridge 2x8 cœrs avec 64 GB of RAM
  - Un système **SIDUS** Debian Wheezy
  - Une Interconnexion en InfiniBand FDR 56 Gb/s
  - **Pas de latence disque : RamDisk BRD/Ext2 & TMPFS de 60 GB**
  - 10 paires GlusterFS : 1 serveur avec RamDisk, 1 client
  - Utilisation de IOZone3 : 13 tests de lecture/écritures
  - 20 expériences pour un ensemble statistiquement representatif

Jour #1 : lancement du test & premières surprises !
Sur les vitesses de transfert I/O

Du nœud 11 au nœud 1



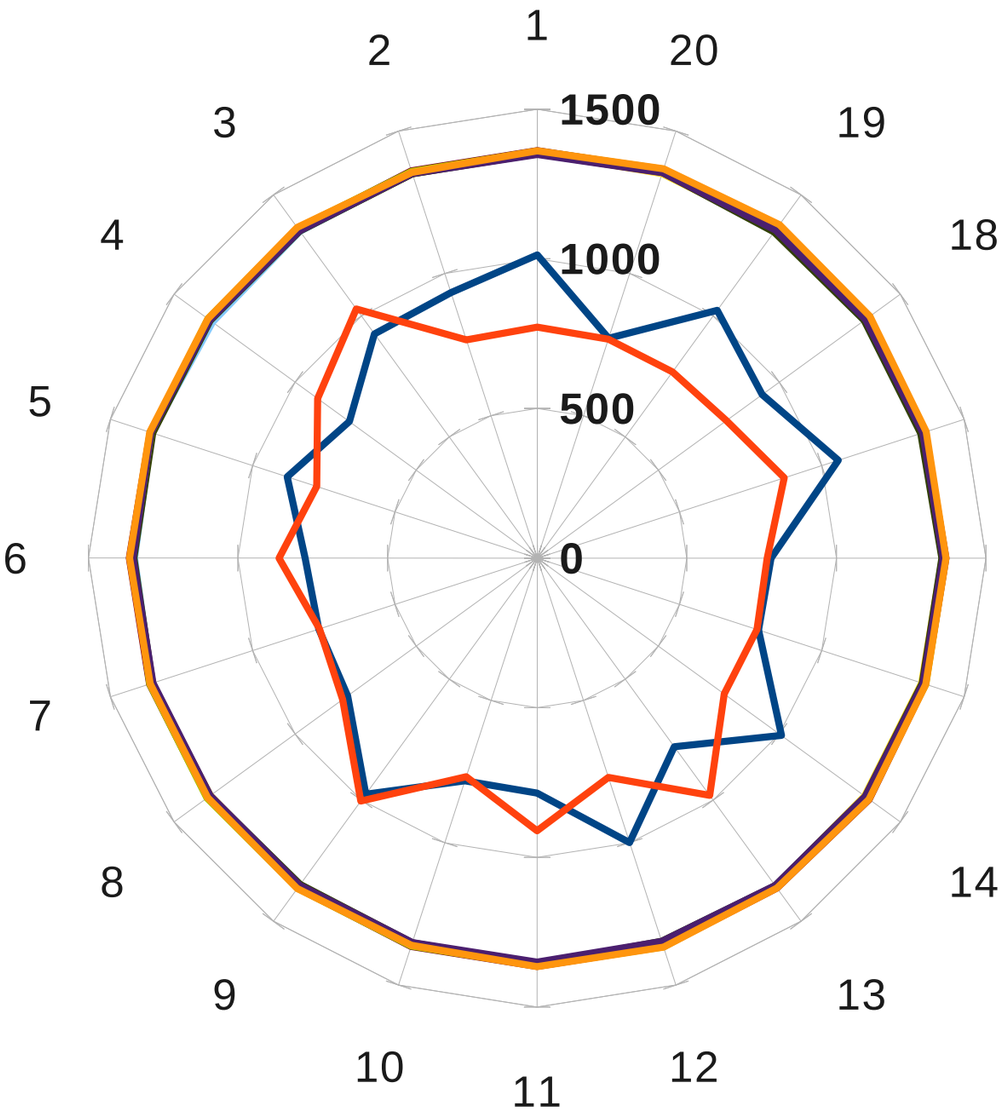
Du nœud 12 au nœud 2

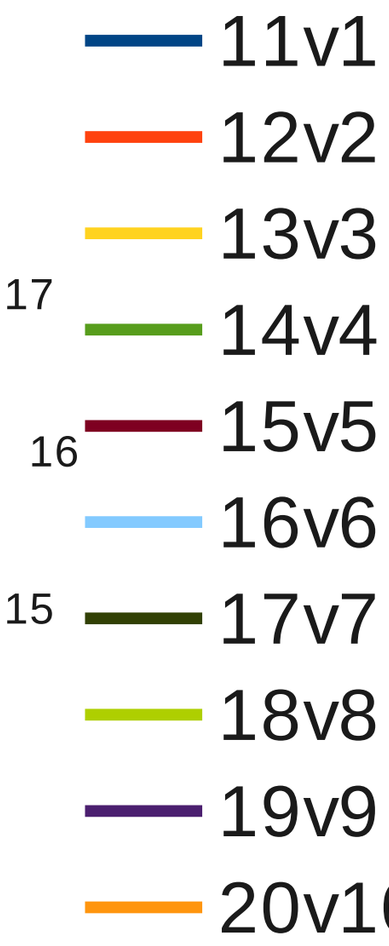


Plus c'est à l'extérieur, mieux c'est !

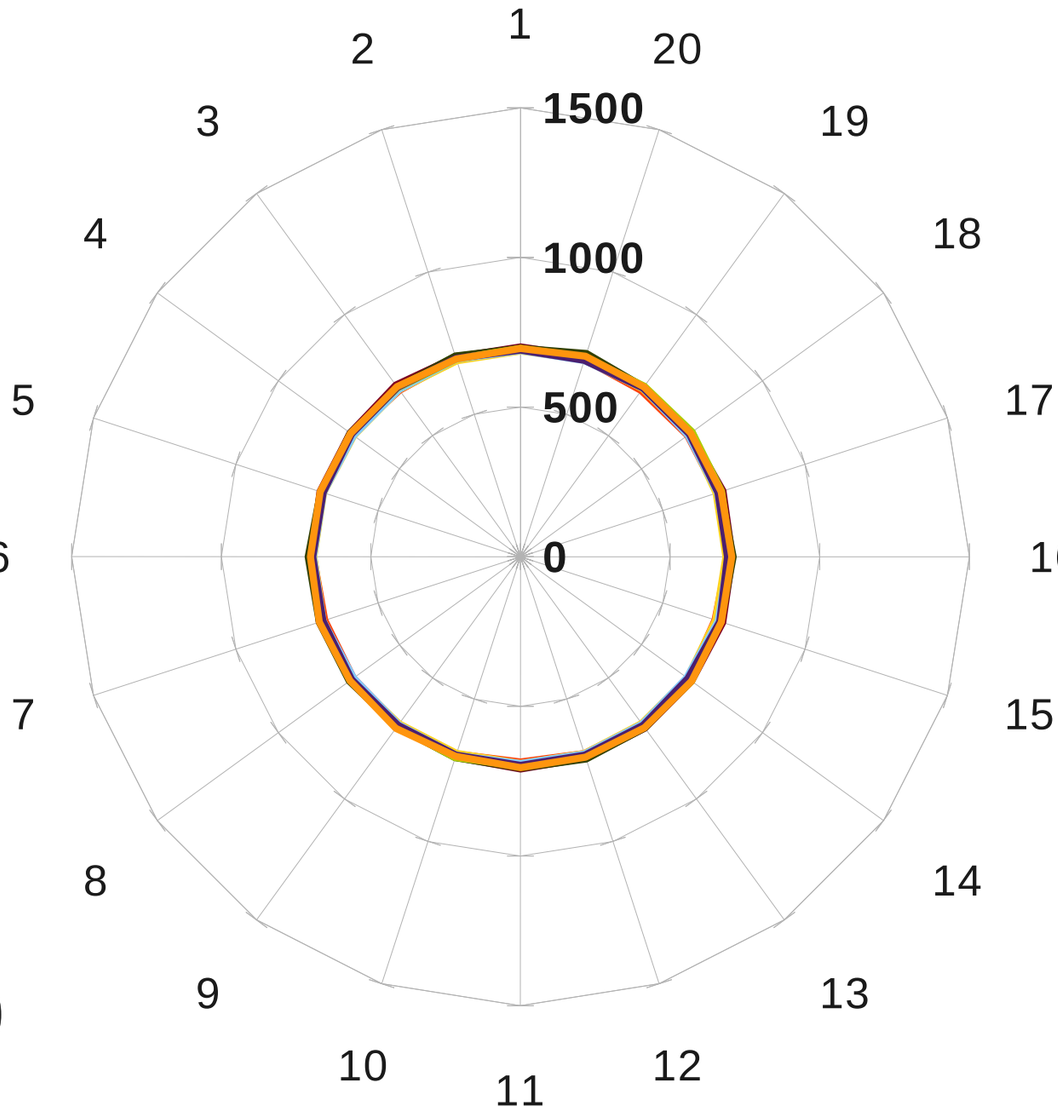
Jours #1 & #2 : modification & nouveaux tests
Sur les temps d'exécutions (*User Time*)

Pour les 10 couples **avant**...





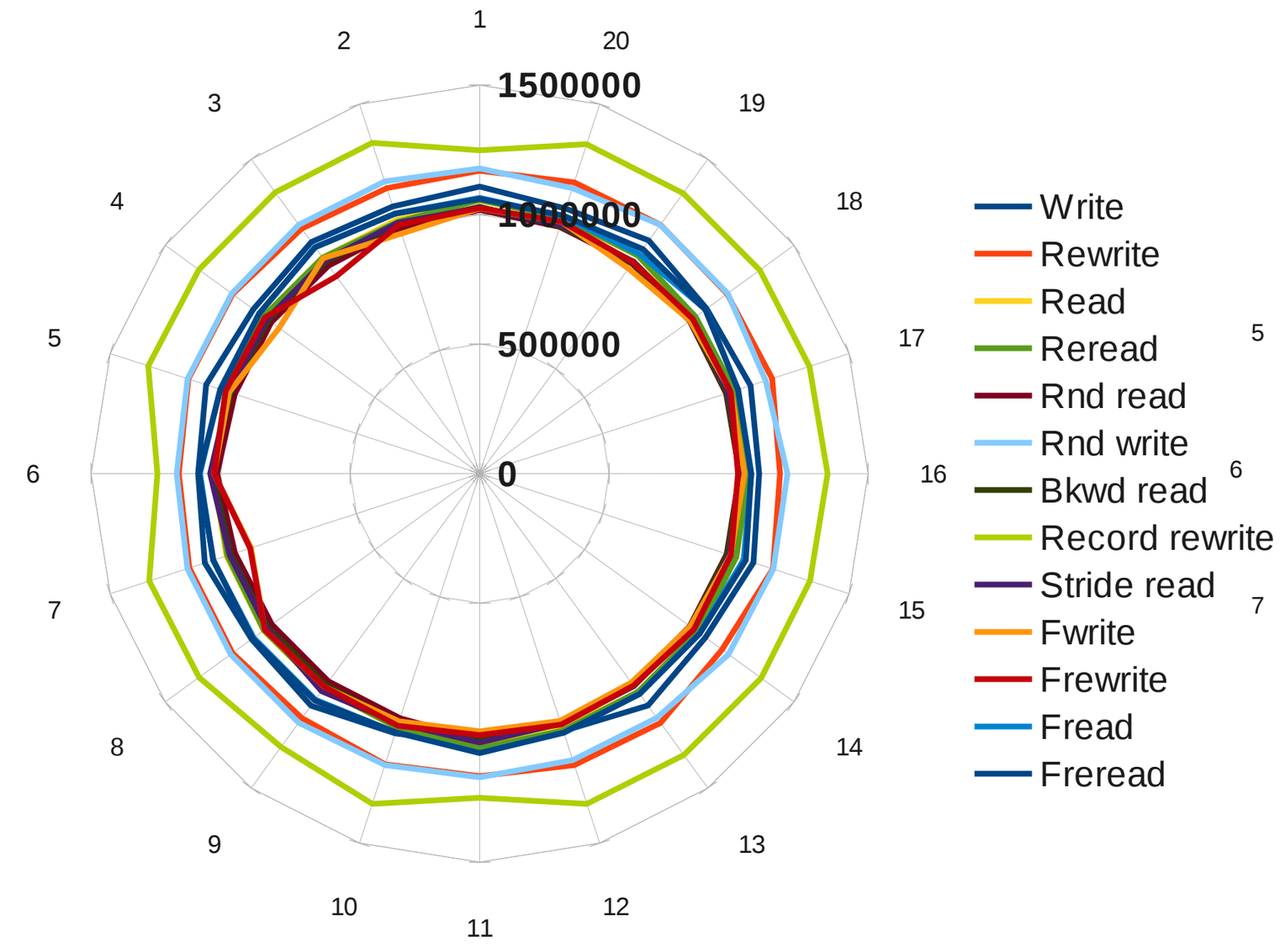
Pour les 10 couples, **après**...



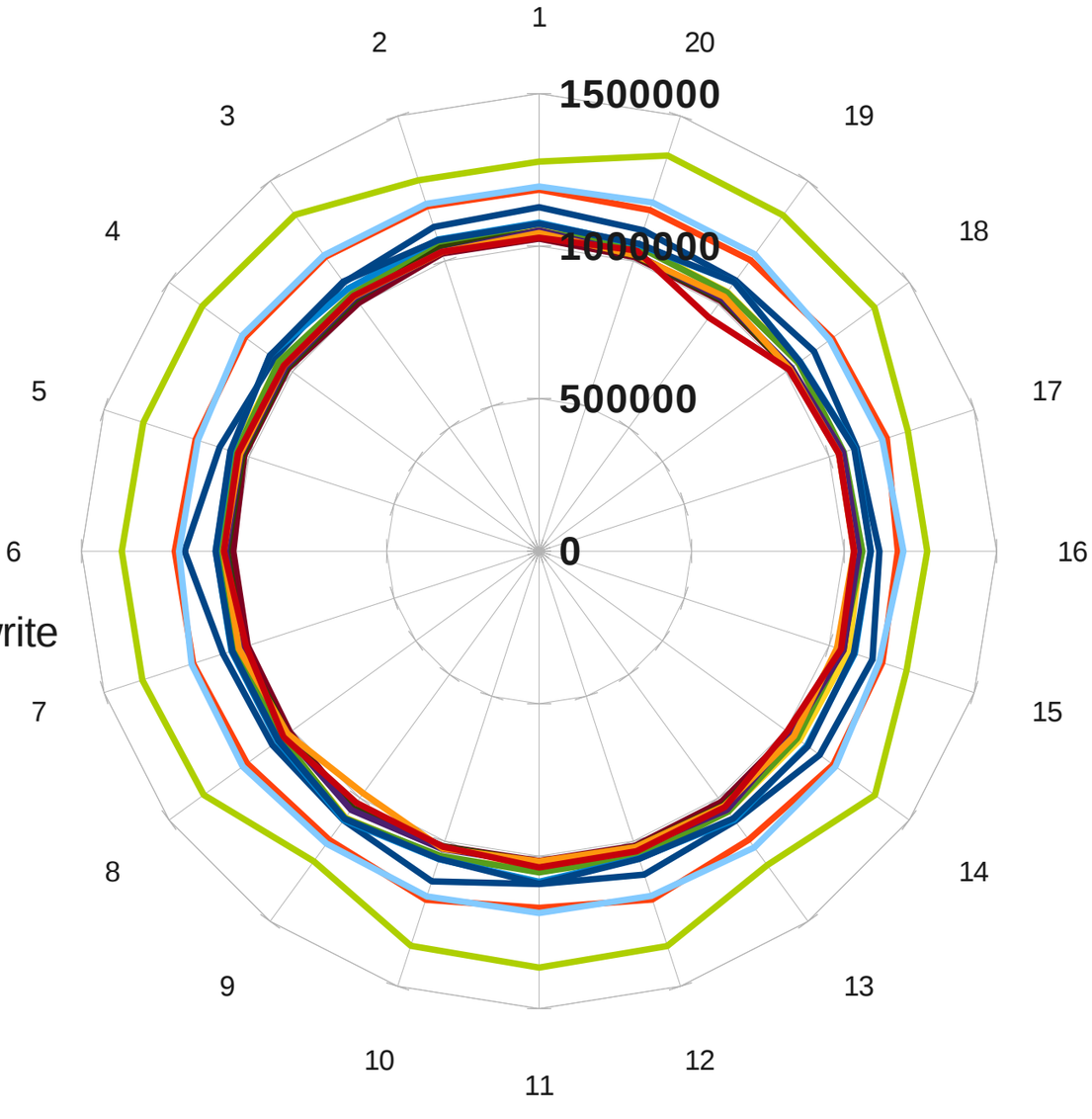
Plus c'est vers le centre, mieux c'est !

Jour #2 : lancement du test & premières surprises !
Sur les vitesses de transfert I/O

Du nœud 11 au nœud 1



Du nœud 12 au nœud 2



Plus c'est à l'extérieur, mieux c'est !

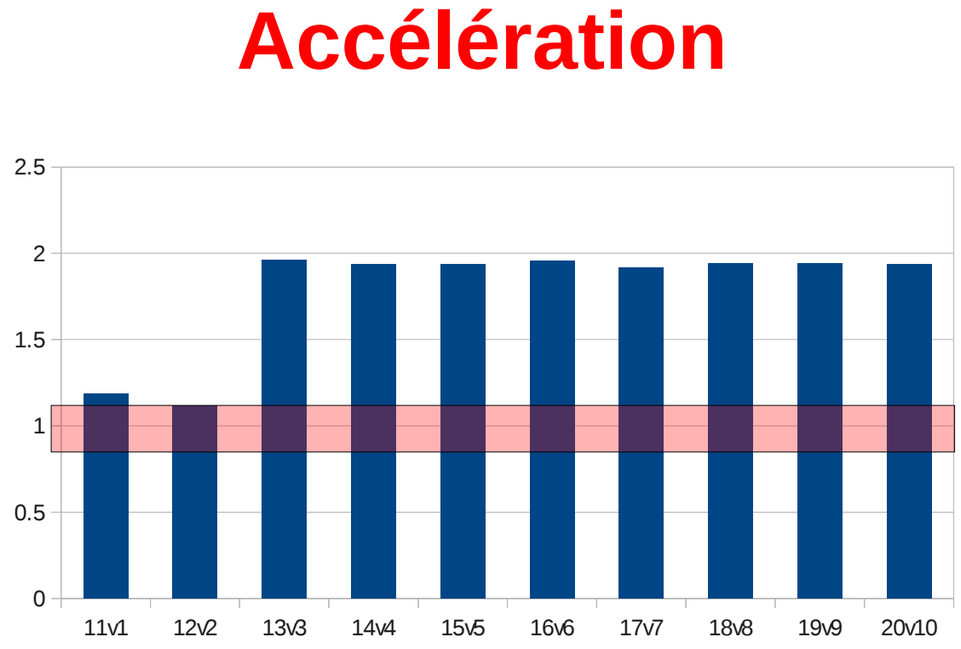
Quel miracle entre jours #1 & #2 ?

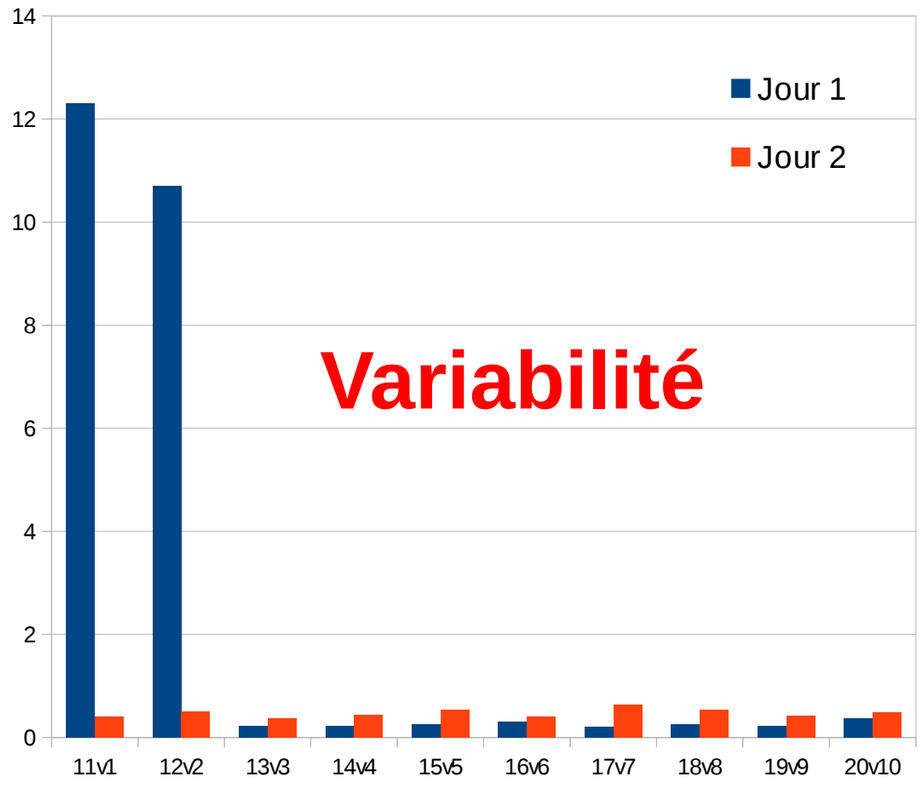
Deux questions : Comment...

- ... multiplier par 2 la vitesse ?
- ... diviser par 20/30 la variabilité ?

La réponse :

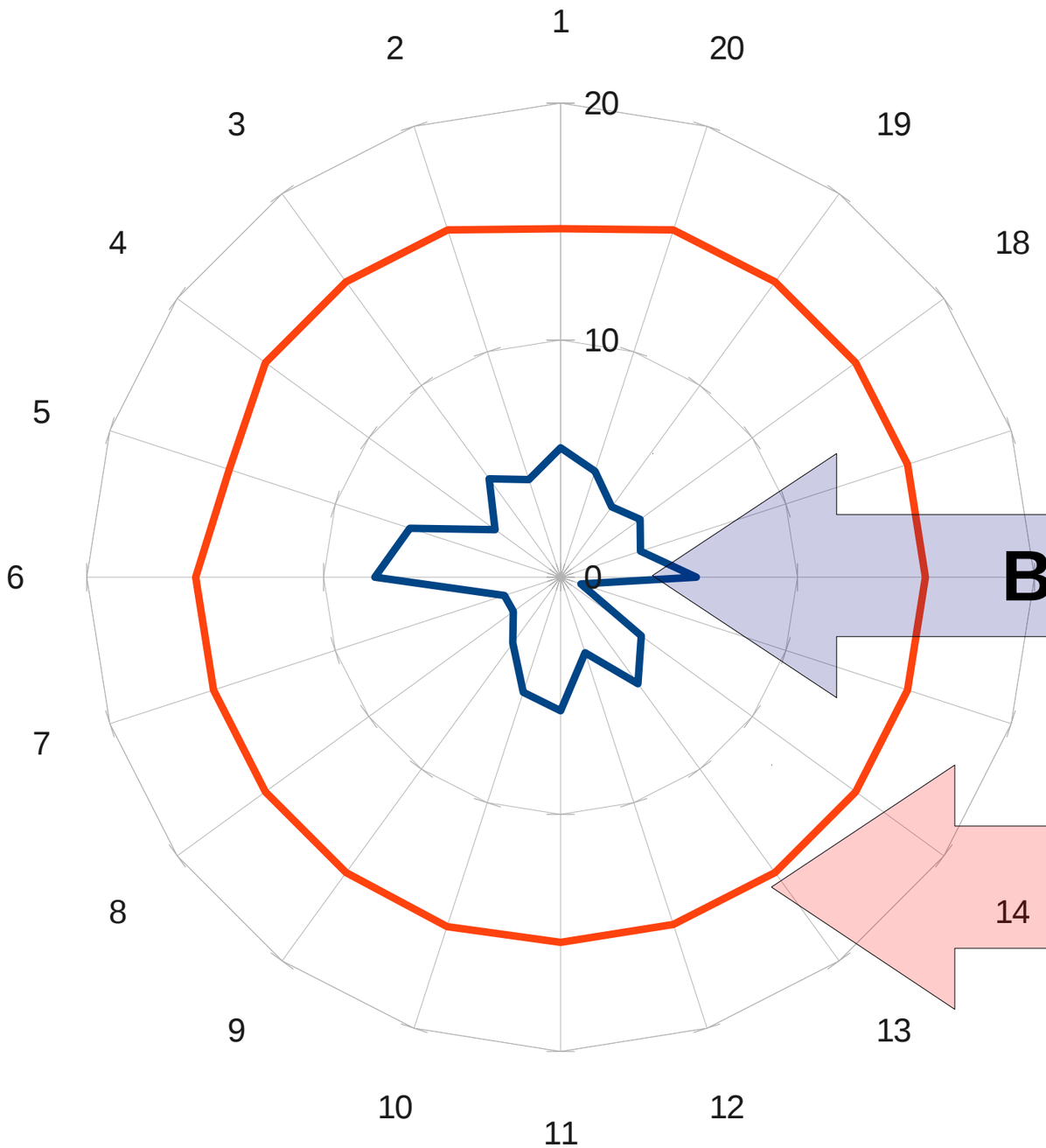
- Optimiser le réseau ? Non
- Optimiser les noyaux d'OS ? Non
- Optimiser le BIOS ? OUI **!!!**
  - BIOS sur 1 & 2 en Max Performance
  - BIOS de 3 à 20 en default
- Solution : BIOS en Max Perf !



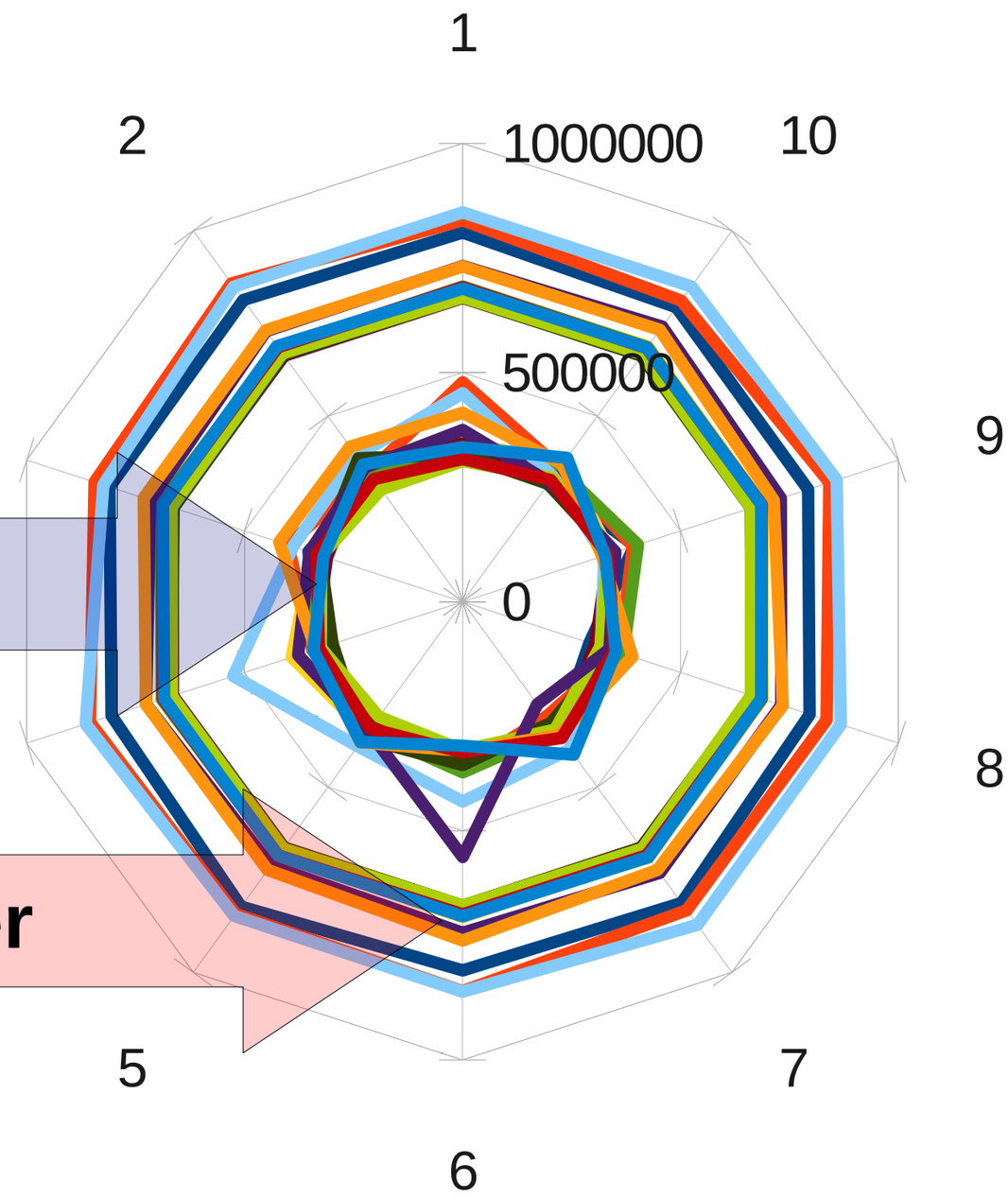


The no-reproducibility reproducible ?
On Equip@Meso

Iperf client/server with IB



Iozone3 on GlusterFS



Before

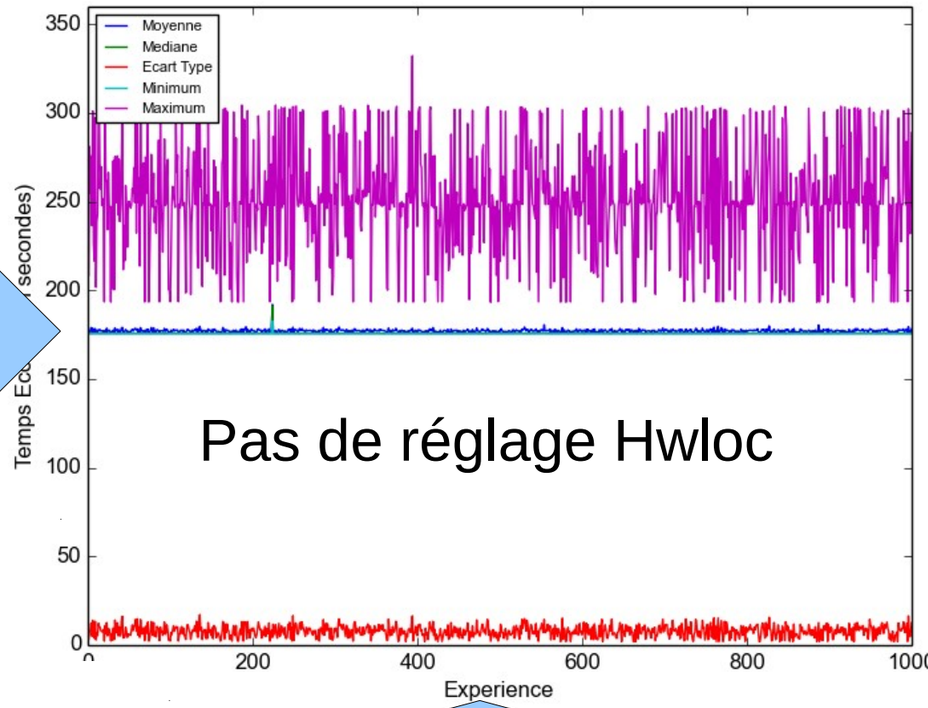
After

Comportement de nœurs de cluster Quelle variabilité en parallélisme massif

- **Objectif:**
  - Évaluation de la scalabilité en MPI, quelle statistique à extraire ?
- **Banc expérimental :**
  - 48 nœuds bi-sockets 4-cœurs R410, interconnexion Infiniband
  - Système unique SIDUS
  - Code Pi Monte Carlo avec distribution en MPI (10^{14} itérations)
  - Lancement par *mpirun -np 384*
  - Réglage de la localité avec hwloc-bind comme argument de mpirun
  - 1000 simulations

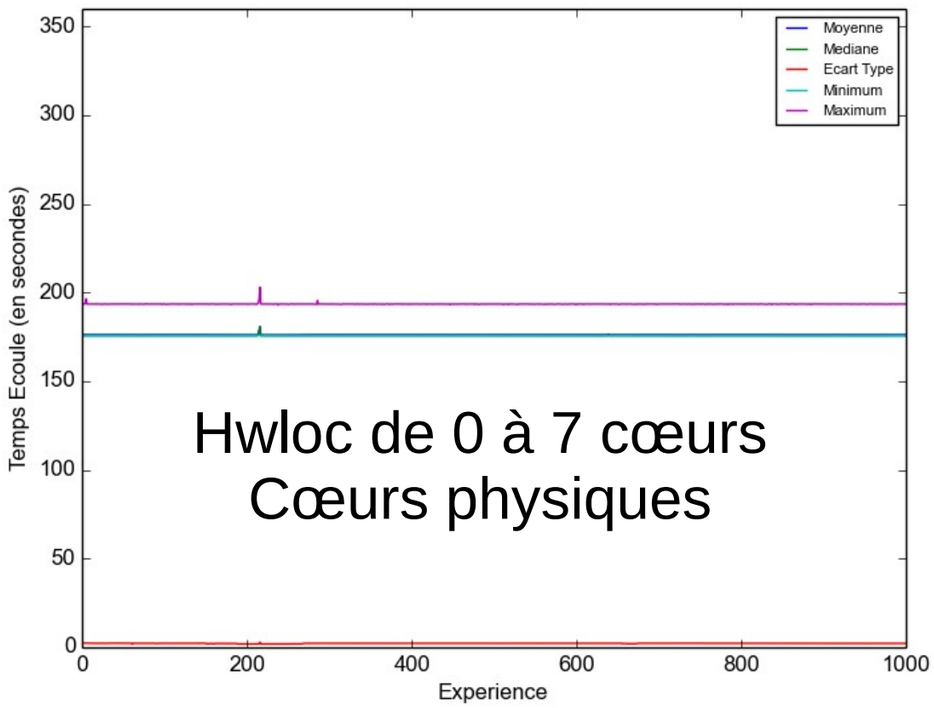
Influence de la localité sur large déploiement MPI
1000 runs : statistiques extraites
Moyenne/Médiane/Écart Type/Min/Max

Temps de calcul
Pour le même boulot !

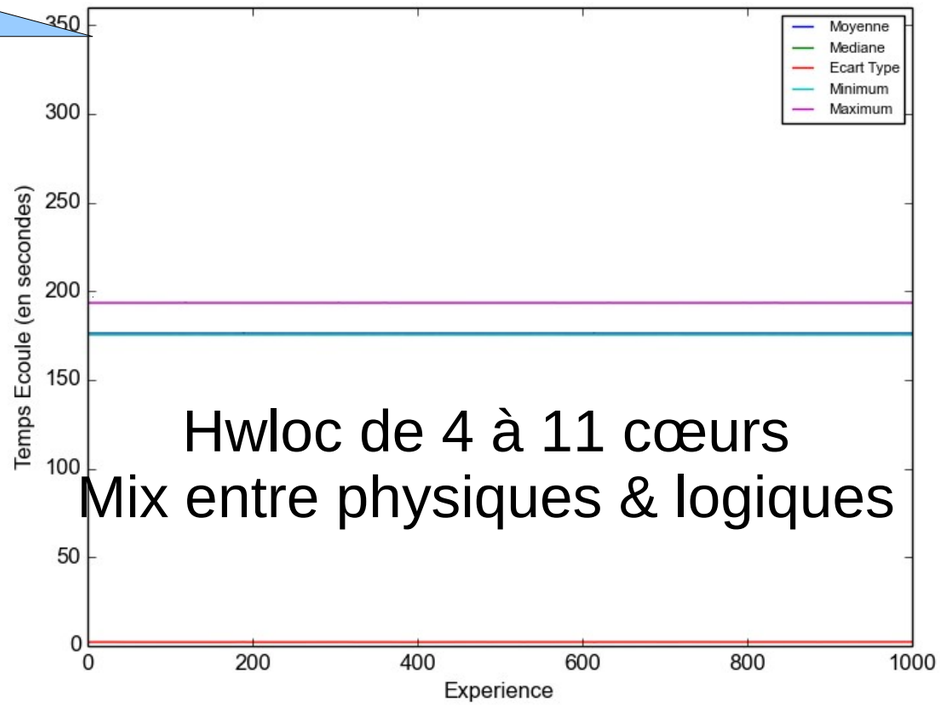


Pas de réglage Hwloc

Indice de L'unité de calcul

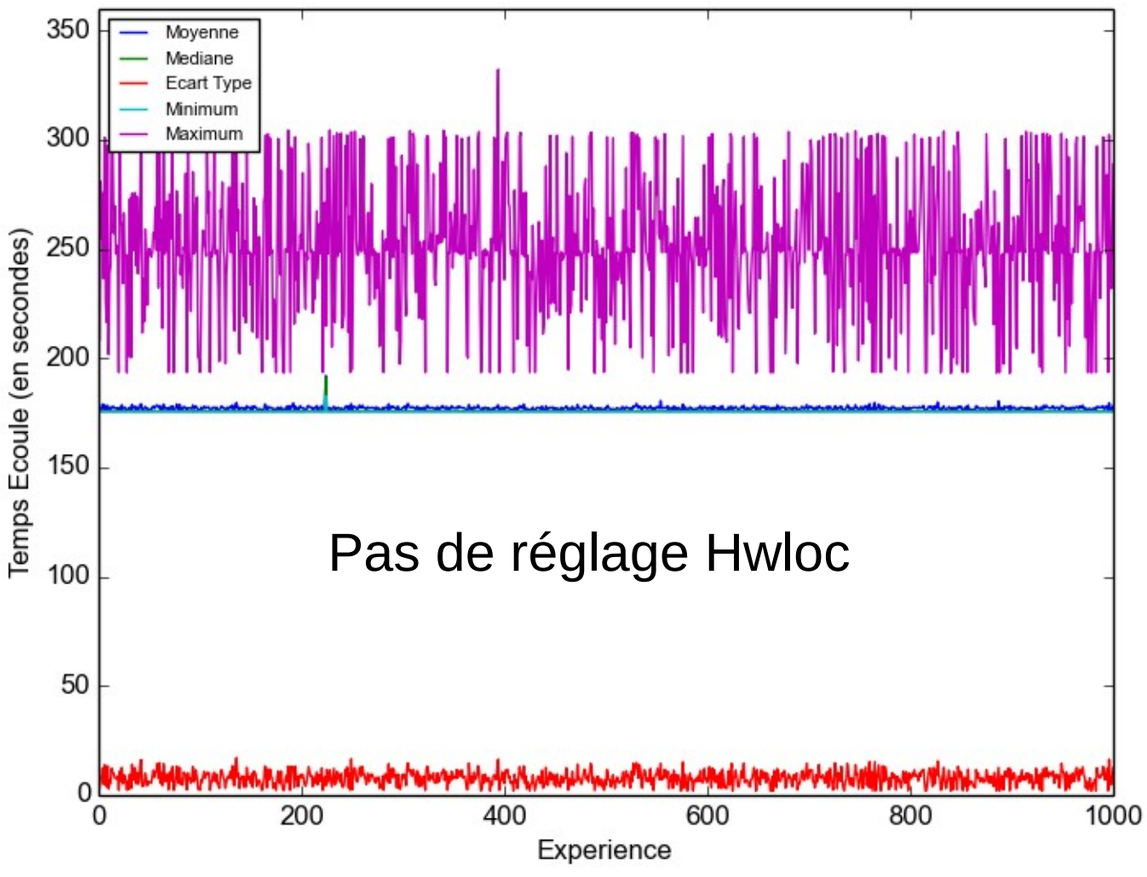


Hwloc de 0 à 7 cœurs
Cœurs physiques

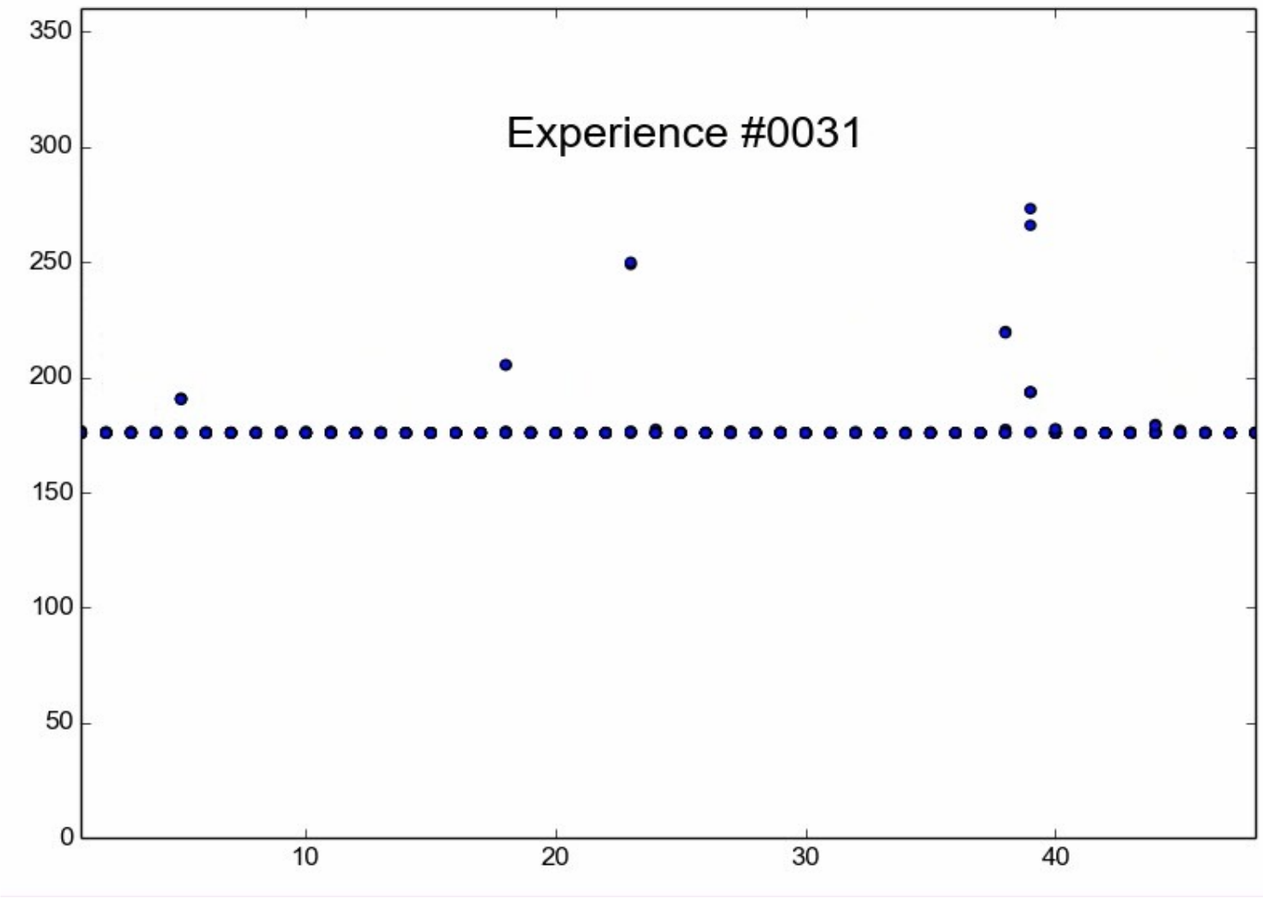


Hwloc de 4 à 11 cœurs
Mix entre physiques & logiques

1000 runs, localité de 0 à 7 cœurs
Grosse variabilité sur les 384 cœurs

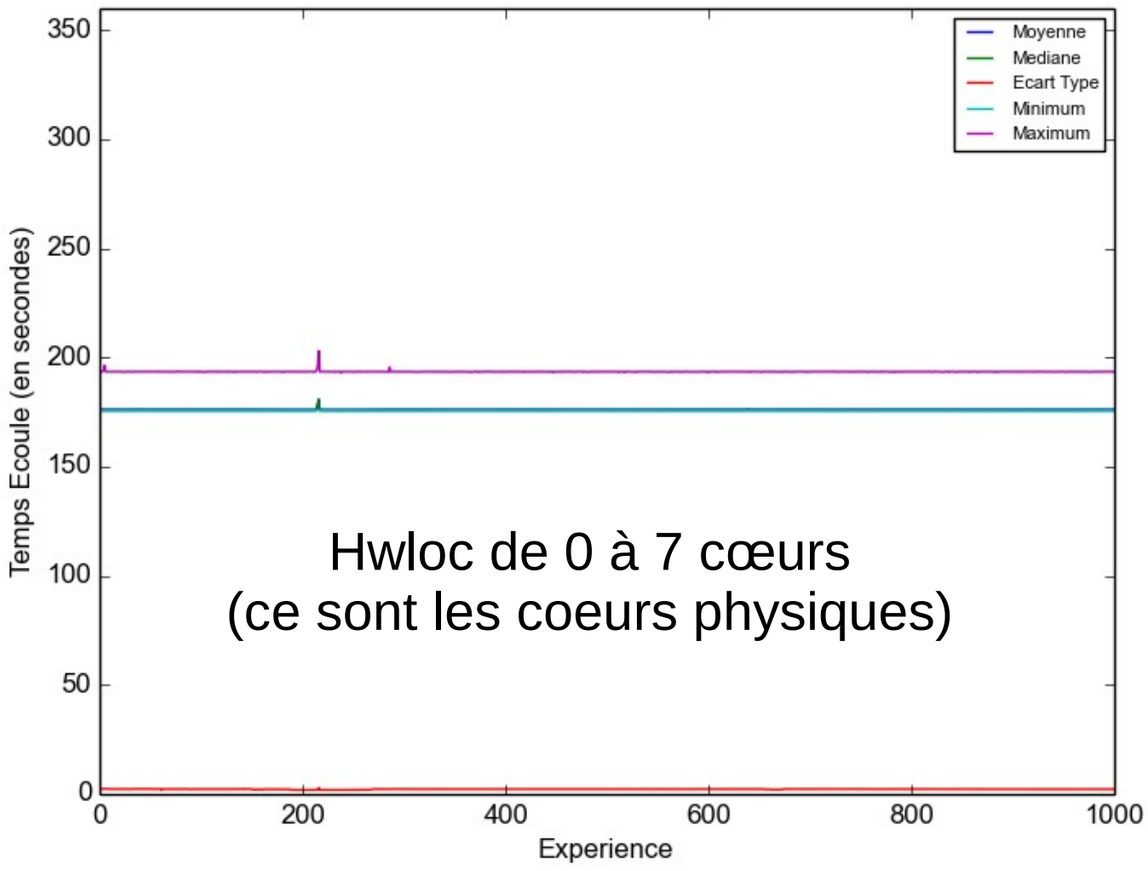


Pas de réglage Hwloc

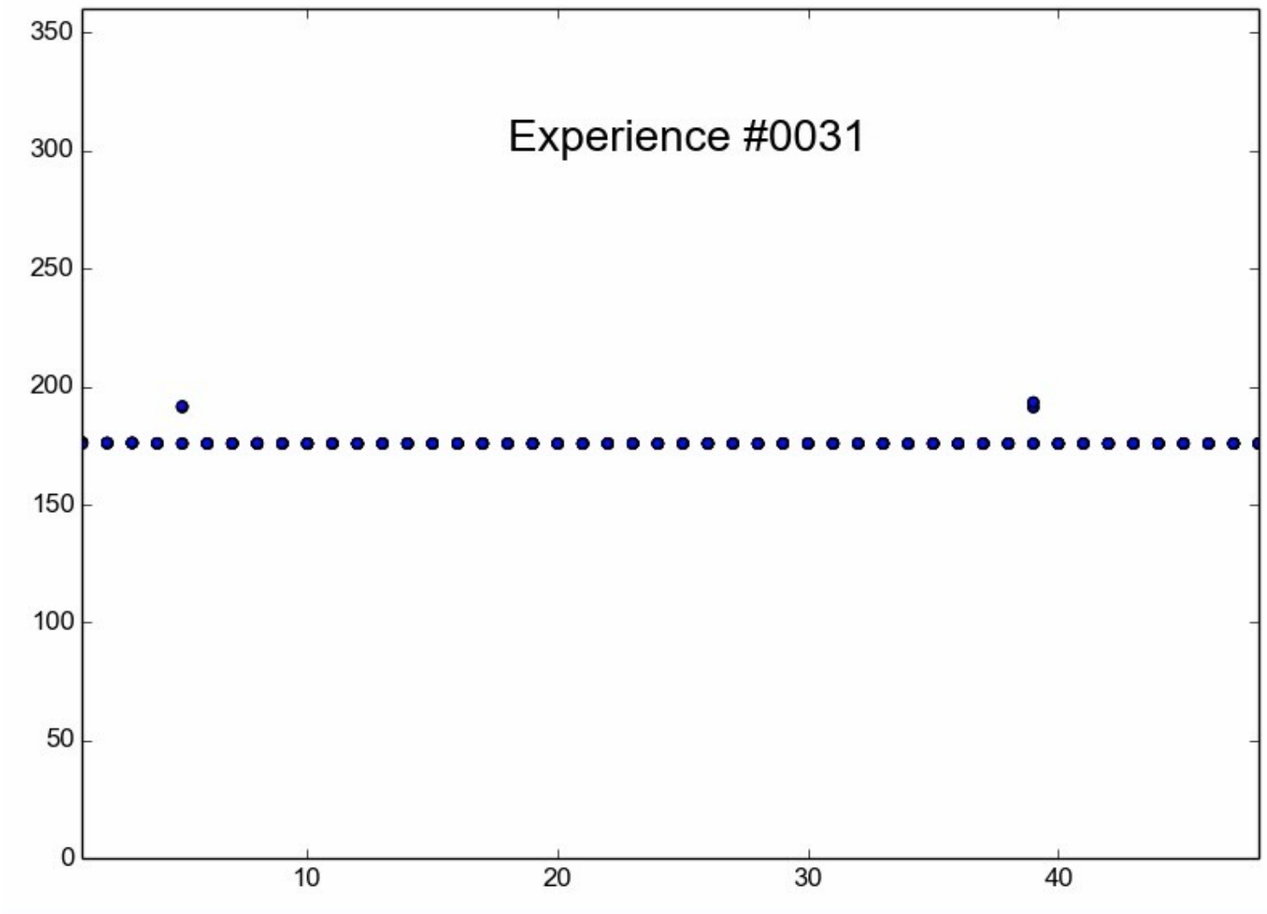


Des nœuds #1 à #48, 8 cœurs par nœud

1000 runs, localité de 0 à 7 cœurs
La variabilité disparaît sur les 384 cœurs



Hwloc de 0 à 7 cœurs
(ce sont les coeurs physiques)



Des nœuds #1 à #48, 8 cœurs par nœud

Futur de SIDUS

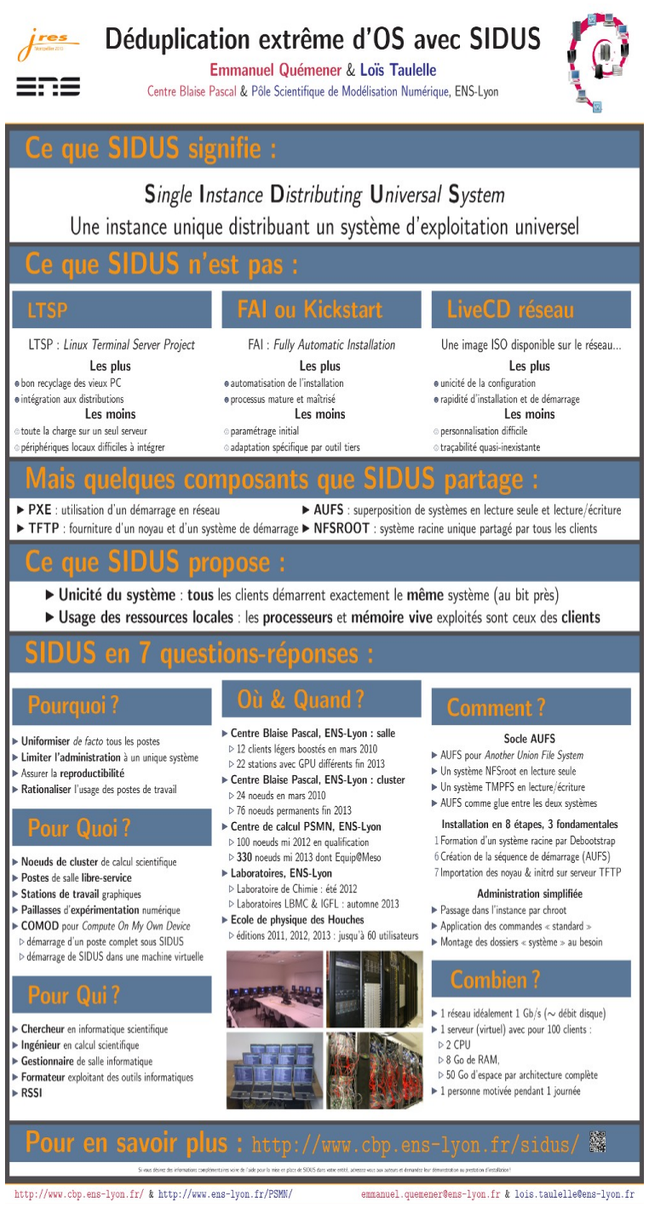
- Valorisation
  - Calcul scientifique, infrastructure de calcul scientifique
  - Gestion de parc, enseignement à la demande
- Simplification de l'installation & administration
  - Dédier une machine en Lecture/Écriture pour l'administration
  - Offrir une connexion SSH sur l'instance pour les opérations classiques
  - Utiliser Debian Preseed pour le processus d'installation
- Déploiement sur Mésocentre ou Grille
- SIDUS *de partout*
  - Lancer SIDUS de l'extérieur par VPN

D'autres informations ?

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

Linux Journal 11/2013

Poster JRES 2013



Site Web CBP

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
- ...