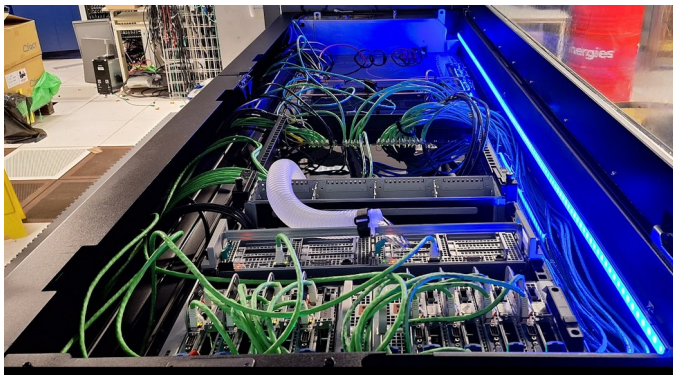


Du calcul numérique et son empreinte environnementale en général... à sa mesure et sa réduction Au CBPsmn en particulier...



Emmanuel Quémener



Ce que ce ne sera pas...

Pas de « reprise » de Labos 1.5 !



Estimation de l'empreinte carbone d'une heure de calcul sur un cœur CPU ou sur un GPU

Méthodologie

Par Labos 1point5 [Janvier 2024]

Sommaire

Introduction	3
Estimation de l'empreinte carbone d'un service de calcul	4
Estimation des unités fonctionnelles	11
Incertitudes	13
Annexes	15
Références	21

Dans cette version de la méthodologie, il a été décidé d'exclure la fin de vie, compte tenu de son très faible impact relatif compte tenu des circonstances de traitement des équipements électriques et électroniques (DEEE) professionnels en France. La récupération de la chaleur fatale est exclue car hors périmètre.

Justement, on va s'intéresser à ces éléments !

Inéquation impossible ?

Moins consommer & mieux servir...

- Nécessité de placer des « nombres » sur des « faits »
- Indicateurs de « consommation » :
 - Croissance (infinie dans un environnement fini :-/)
 - **ADP** : potentiel d'épuisement des ressources abiotiques :
 - Abiotic Depletion Potential (unité kgSbeq)
 - **PRG** : Potentiel de Réchauffement Global ou « empreinte carbone »
 - GWP : Global Warming Potential (unité kgCO₂eq)
 - **PE** : Consommation de ressources énergétiques
 - Primary Energy (unité MJ)

Quelle empreinte carbone ?

Finalelement, une vieille idée...

- Analyse « comptable », « **3 coûts** » pour tout « service »
 - **Coût d'entrée** : appropriation, développement, intégration, ...
 - **Coût d'exploitation** : MCO, évolutions réglementaires, sanitaires, ...
 - **Coût de sortie** : remplacement, abandon, délégation, ...
- Pour du matériel (et son usage) : même combat !
 - **Avant** : sa fabrication (et son transport)
 - **Pendant** : son exploitation (et sa maintenance)
 - **Après** : son recyclage (et son transport, stockage)
- Pendant : consommation électrique (et chaleur fatale)...

De la « résistance » salubre... à la « chaleur fatale » de l'électronique

Expérience commune de la « résistance chauffante »

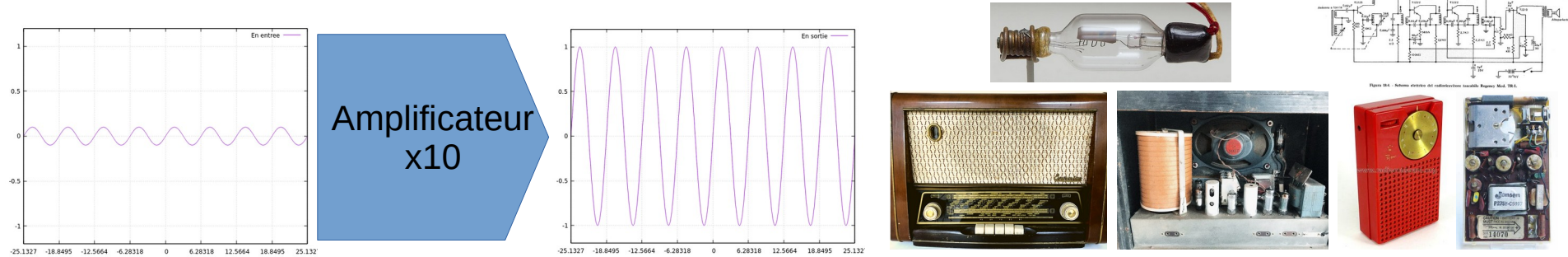


Courant électrique dans résistance donne chaleur...

Pourquoi cette « chaleur fatale » ?

Histoire d'un grand détournement...

- Au début, « amplifier un signal » (un signal radio...)



- Sauf que « lampe » ou « transistor » sont « physiques » :

- Un mode « linéaire » (le plus possible)
- Un mode « bloqué » : rien en sortie
- Un mode « saturé » : la même valeur non nulle

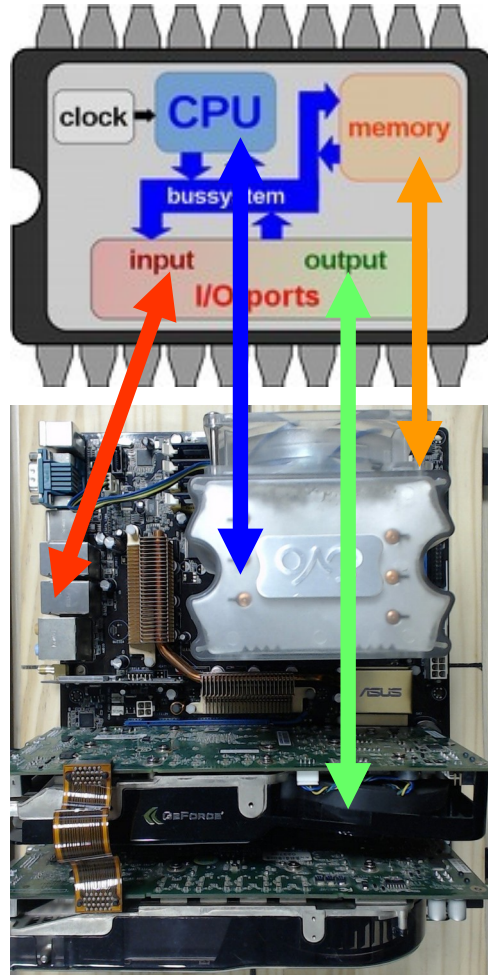
Amplificateur "idéal"

Interrupteur Commandé
Porte Logique !

Enfin un cas où les extrêmes sont intéressants !

Mais pourquoi ça « chauffe » alors ?

Combinaison de X facteurs !



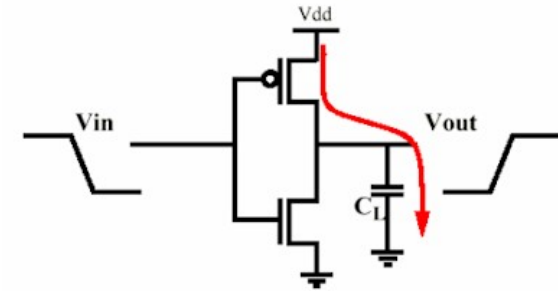
- Les transistors ne sont pas seuls :
 - D'autres composants : résistances, capacités, self, etc. tous dissipatifs...
- Les transistors ne « commutent » pas instantanément...
 - Et pendant leur « transition », ils se comportent « comme les autres »
- Les transistors sont nombreux :
 - Dans les portes logiques (de 2 à 6) par « fonction » logique
 - Dans la mémoire SRAM (6/bit), DRAM (1/bit),
 - Dans le stockage SSD ou NVMe (6/bit)
 - Processeur récent : 8 milliards
 - 8 GB de DRAM : 64 milliards
 - NVMe ou SSD de 1TB : 48 milliards...
- Les transistors « commutent » vite !
 - Horloge : plusieurs milliards de fois par seconde

Pour un peu fixer les idées...

La consommation électrique IT

- Grandeur et décadence de la fréquence

- Entre 1981 et 1999 : de 4 MHz à 400 MHz x100 en ~20 ans
- Entre 1999 et 2004 : de 400 MHz à 3 GHz x~10 en 5 ans
- Entre 2004 et 2009 : de 3 GHz à 2 GHz x0.66 en 5 ans
- Entre 2009 et 2024 : de 2 GHz à entre 0.8 GHz et 5 GHz !



- Thermal Design Power : enveloppe thermique de dissipation maximale

- $TDP = \frac{1}{2} C V^2 f$ avec C = Capacitance, f = fréquence, V = tension (fonction de f !)
- Capacitance = Finesse² . Nb Transistors . Constante de Mylq (~ 0.015)

- TDP pour un processeur : jusqu'à 350 W (sur 12 cm²)

- Densité de chaleur d'une plaque à induction !



- TDP devient le facteur limitant de puissance (de traitement)

Quelle empreinte carbone ?

Interrogations sur la « littérature »

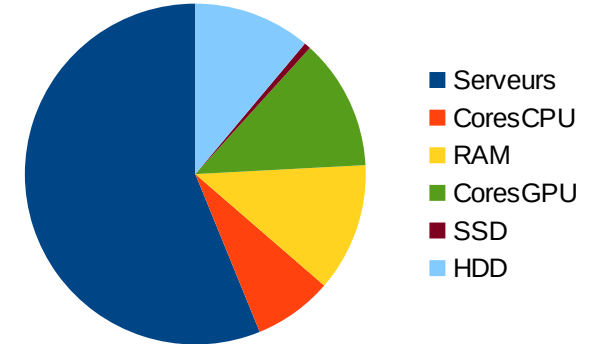
- Petite expérience : *établir un devis sur Matinfo5*
 - Empreintes carbone identiques quel que soit : le modèle CPU, la RAM
- Littérature plus « pertinente, cohérente et consistante » :
 - <https://boavizta.org/blog/empreinte-de-la-fabrication-d-un-serveur>
 - Approche 2 : « facteurs d'émission arbitraires par composant »
 - $\text{servergwp}(\text{kgCO}_2\text{eq}) = 900(\text{kgCO}_2\text{eq}) + \text{cpuunits}(\text{unit}) \times 100(\text{kgCO}_2\text{eq}/\text{unit}) + \text{ramsize}(\text{GB}) \times 150/128(\text{kgCO}_2\text{eq}/\text{GB}) + \text{ssdunits}(\text{unit}) \times 100(\text{kgCO}_2\text{eq}) + \text{hddunits}(\text{unit}) \times 50(\text{kgCO}_2\text{eq}) + \text{gpuunits}(\text{unit}) \times 150(\text{kgCO}_2\text{eq}/\text{unit})$
 - Approche 3 : « vers une formule de calcul d'impact multicritère » basé sur les semiconducteurs
 - $\text{server} = \text{cpu} + \text{ram} + \text{ssd} + \text{hdd} + \text{motherboard} + \text{psu} + \text{enclosure} + \text{assembly}$
 - $\text{cpu} = \text{cpuunits} \times ((\text{cpucoreunits} \times \text{cpudiesize} + 0,491) \times \text{cpu_die} + \text{cpu_base})$
 - $\text{ram} = \text{ramunits} \times ((\text{ramsize} / \text{ramdensity}) \times \text{ram_die} + \text{ram_base})$
 - $\text{ssd} = \text{ssdunits} \times ((\text{ssdsize} / \text{ssddensity}) \times \text{ssd_die} + \text{ssd_base})$
 - $\text{hdd} = \text{hddunits} \times \text{hdd_unit}$
 - $\text{psu} = \text{psuunits} \times \text{psuunitweight} \times \text{psu_weight}$
 - $\text{enclosure} = \text{rack ou enclosure} = \text{blade} \times \text{bladeenclosure}/16$

Analyse : CBP comme « pollueur »

Fabrication & Exploitation

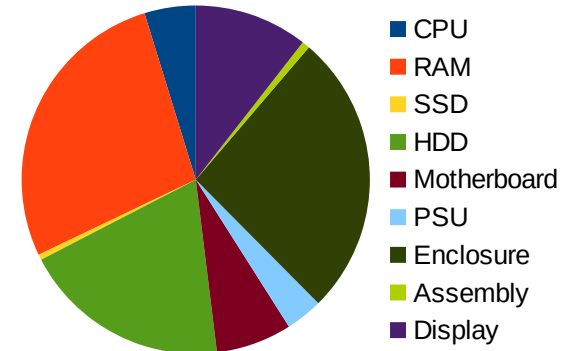
- Fabrication : 303 machines

- Moins de 10 % sous garantie
- 5820 coeursCPU, 6435 coeursGPU,
- 50 TiB RAM, 4 PB, 1100 HDD, 33 SSD
- 485 ou 252 tonnes CO₂ à la fabrication



- Exploitation : 303 avec 300W H24 7/7

- 40 tonnes CO₂ par an soit 4 françaises « moyennes »
- Mais 400 utilisatrices différentes chaque mois !



- Ratio : 1 pour 12 à 1 pour 6... En gros, jamais l'égalité...

Actions à envisager

- D'abord réduire les coûts d'entrée (ou de sortie)
 - En exploitant le maximum les machines déjà fabriquées
- Puis réduire les coûts d'exploitation :
 - En exploitant la chaleur fatale directement
 - En améliorant l'évacuation la chaleur fatale
- Mais, avant tout « plan de bataille », « intelligence » :
 - Connaître sa consommation à petite échelle (machine ou composant)
 - Connaître sa consommation à grande échelle (Data Center)

Appréhension : récupérer les infos...

En fait, pas si simple !

- Récupérer « localement » : échelle du composant
 - Via le système d'exploitation, directement : sensor
 - Via l'IPMI et l'OS : « ipmitool » ou mieux « ipmi-sensors »
 - Via un wattmètre (et une webcam)
 - Via une pince ampèremétrique
- Récupérer « globalement » : échelle du Data Center
 - Un site web authentifié (en Java)
 - « page Web » = « framebuffer »

SOUS SOL	JOUR EN COUR	CUMUL
GENERAL TGBT	5450 kWh	20130286 kW-hr
GENERAL ECLAIRAGE	0 kWh	1694 kW-hr
GENERAL CVC	27 kWh	118004 kW-hr
GENERAL CVC TT	1057 kWh	3990965 kW-hr
GENERAL TGHQ NDC	2899 kWh	9464892 kW-hr
GENERAL TGHQ CORPORA	116 kWh	826766 kW-hr
GENERAL TGHQ STOCKAGE	399 kWh	2046484 kW-hr

Mesurer la consommation : échelle « locale », approches...

OS : « sensors »

```
numa@casimir: ~  
File Edit View Search Terminal Help  
Core 60:      +35.0°C (high = +91.0°C, crit = +101.0°C)  
Core 61:      +37.0°C (high = +91.0°C, crit = +101.0°C)  
  
power_meter-acpi-0  
Adapter: ACPI interface  
power1:      180.00 W (interval = 1.00 s)  
root@platinum4o11:~#
```

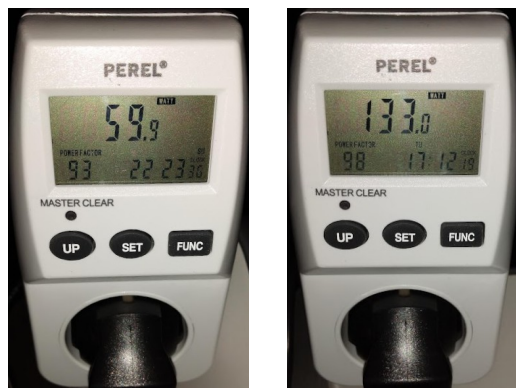
```
numa@casimir: ~  
File Edit View Search Terminal Help  
Core 14:      +39.0°C (high = +88.0°C, crit = +98.0°C)  
Core 15:      +41.0°C (high = +88.0°C, crit = +98.0°C)  
  
power_meter-acpi-0  
Adapter: ACPI interface  
power1:      136.00 W (interval = 300.00 s)  
coretemp-isa-0000
```

OS : « ipmitool sensor »

```
numa@casimir: ~  
File Edit View Search Terminal Help  
root@apollo4o11:~# ipmitool sensor | grep W | awk -F'|' '{ print $1" "$2" "$3 }'  
  
PS 1 Input      570.000      Watts  
PS 2 Input      0.000        Watts  
PS 1 Output     0.000        Watts  
PS 2 Output     0.000        Watts  
root@apollo4o11:~#
```

```
numa@casimir: ~  
File Edit View Search Terminal Help  
root@platinum4o11:~# ipmitool sensor | grep Watts | awk -F'|' '{ print $1" "$2" "$3 }'  
  
PS1 Input Power 225.000      Watts  
PS3 Input Power 9.000        Watts  
root@platinum4o11:~#
```

Wattmètre

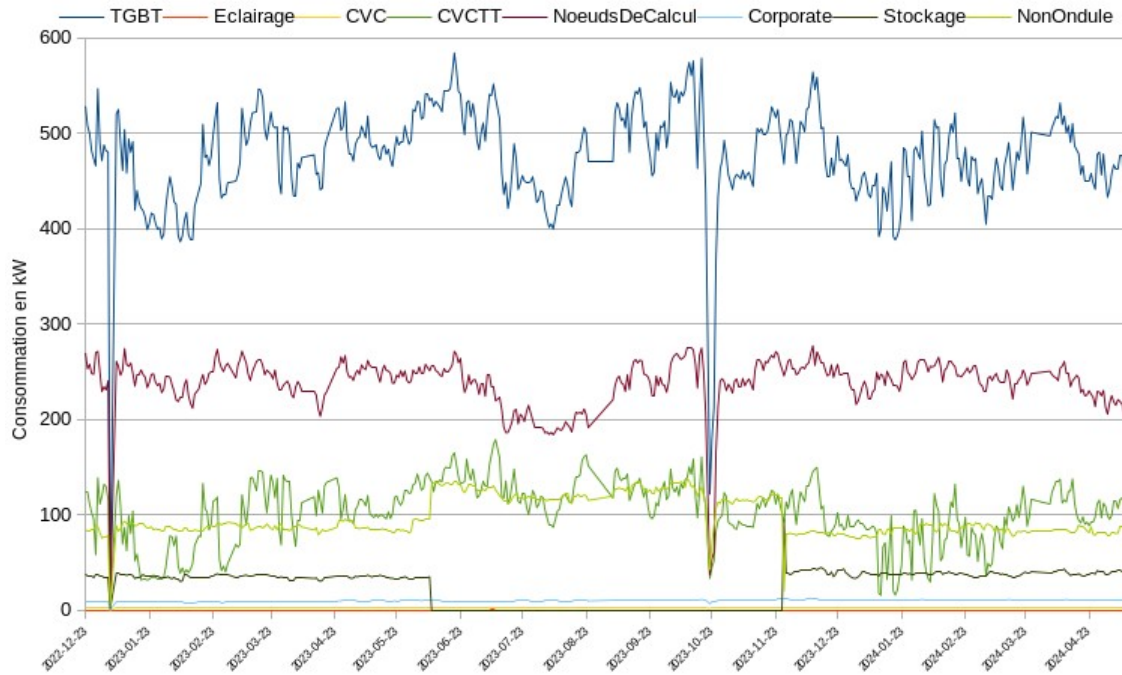


Pince ampèremétrique



Analyse du DataCenter la statistique (et son pentacle)

Période du 23/12/2022 au 13/05/2024 : plus de 500 jours

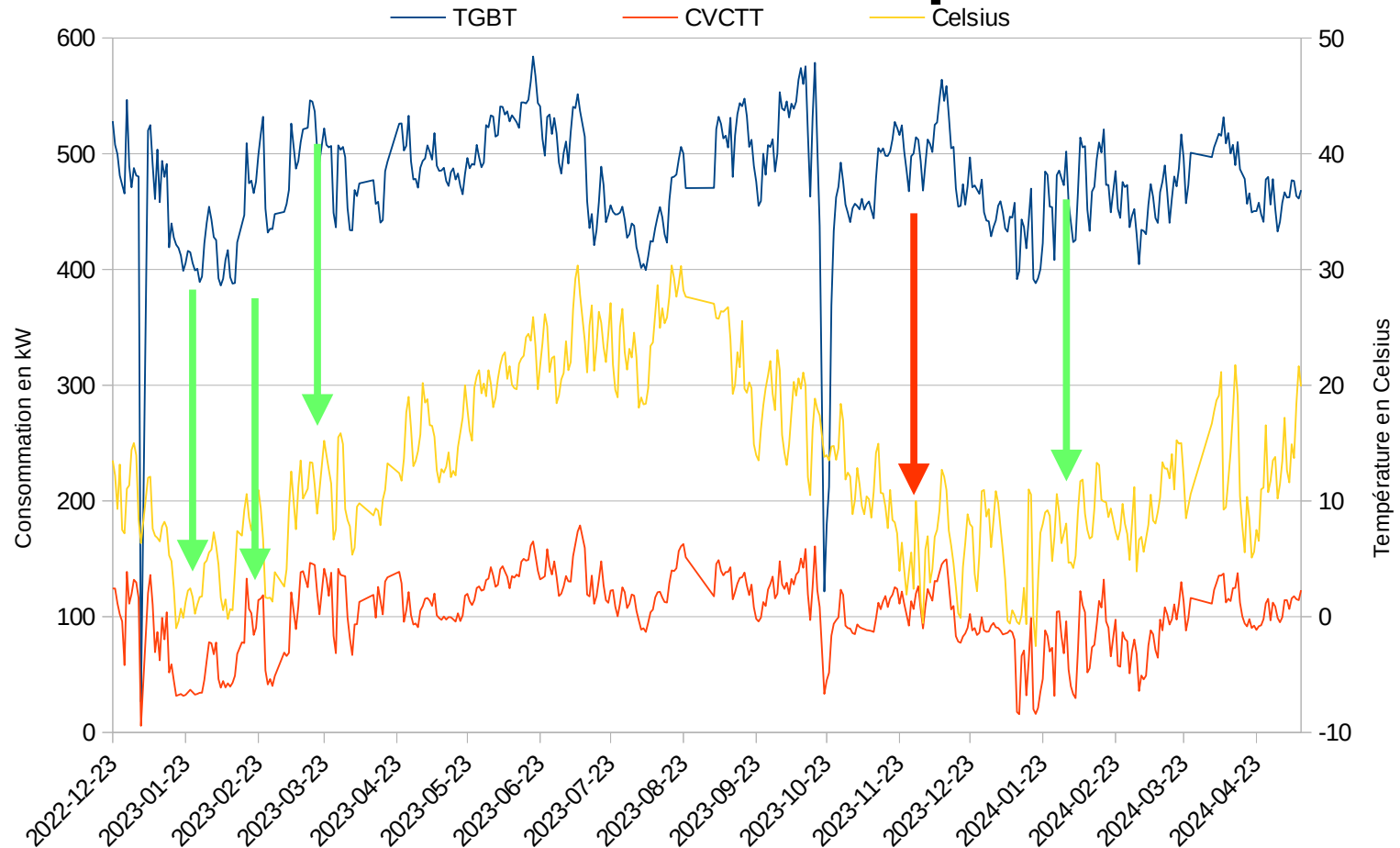


Légende sommaire...

- **TGBT** : arrivée électrique
- **CVCTT** : climatisation
- **NoeudsDeCalcul** : éponyme...
- **Corporate** : machines DSI
- **Stockage** : machines labos
- **Non ondulé** : tout sauf

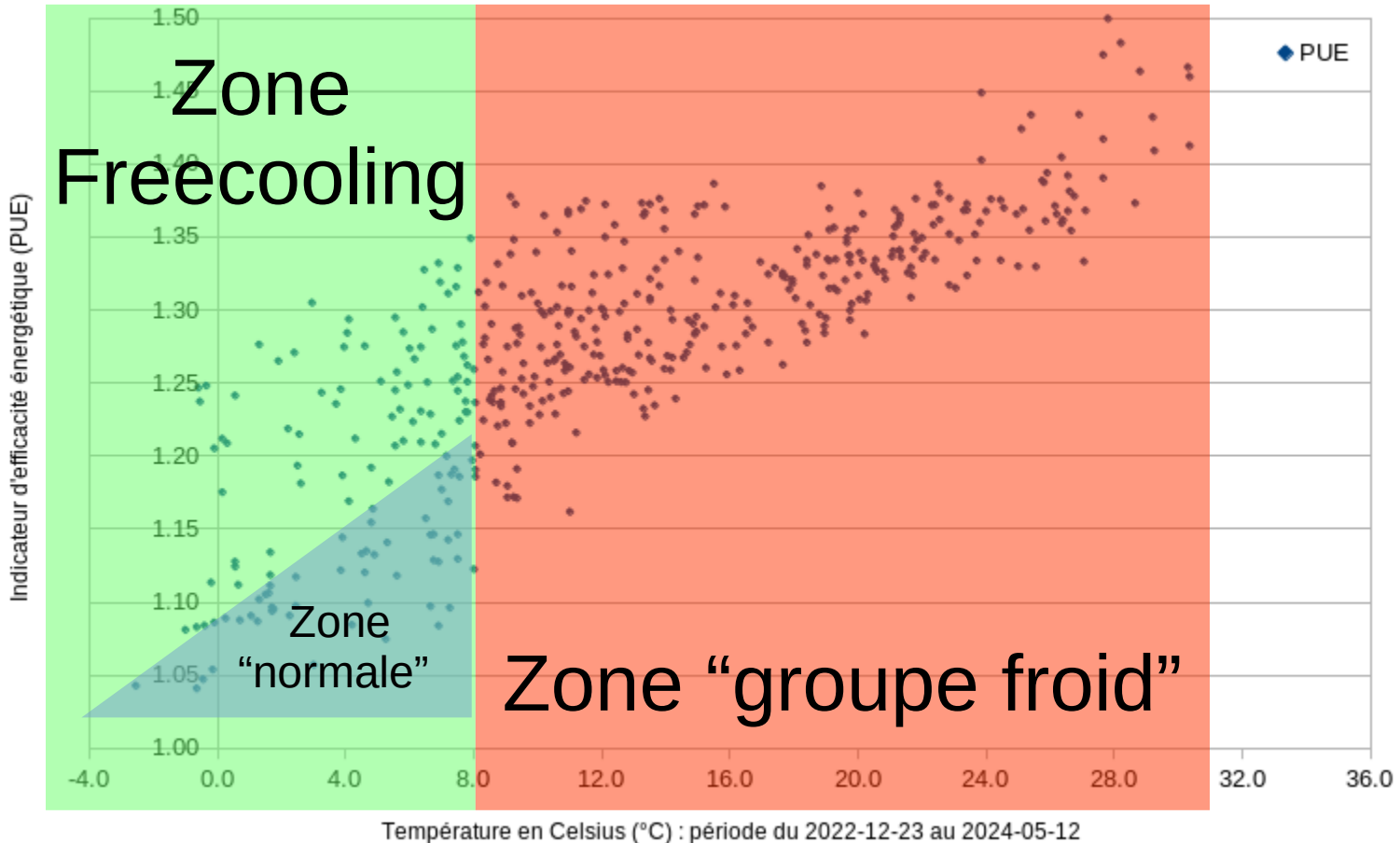
- Moyenne : 478 kW, Médiane : 476 kW, Max-Min ~200 kW
- Une forte variabilité, mais quelle est sa « nature » ?

Consommation Data Center Relation avec la température ?



Une forte relation mais quelle influence sur le PUE ?

Mais qu'en est-il du PUE sur cette période de 500 jours ?



DataCenter « état de l'art » (PUE 1.29) mais perfectible !

Analyse du DataCenter

Mais que faire en « exploitation » ?

- Des pistes d'améliorations :
 - Améliorer le PUE, mais hors champ « utilisateurs »
 - Limiter la consommation des serveurs consommateurs (à GPU)
 - Limiter la consommation du centre de calcul (600 machines)
 - Dépayser des machines dans les bureaux
 - Éteindre les machines anciennes
- Avec la « nouvelle donne » :
 - Explosion du prix le jour : entre 6h et 22h, x8 du prix
 - Risques de coupure électrique...

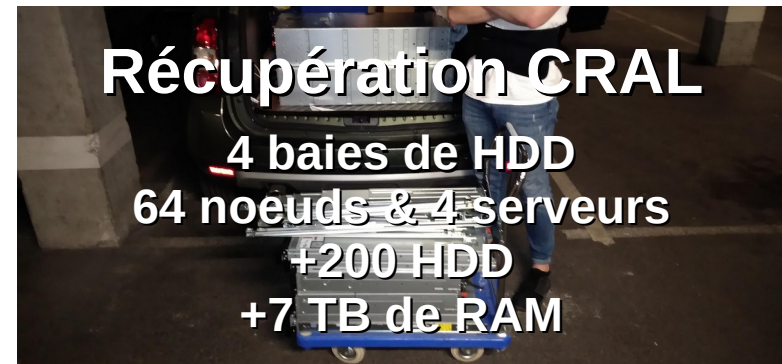
Action : pour consommer « moins », achetons moins (de neuf)...

- Tout dépend de sa destination :
 - « data storage » : serveur de fichiers
 - « data crunch » : station de travail ou nœud, CPU, GPU, RAM
 - « data send » : équipement réseau (incroyable)
 - « data control » : laptop, smartphone
- A moindre coût :
 - Étendre la capacité : distribution de 512 barrettes de RAM DDR3
 - Consolider l'existant avec quelques composants « neufs » :
 - Serveurs « projets » et « scratch » : RAM de 192 à 384GB, HDD de 4 à 16TB
 - Laptops ou iMac de 2 à 16GB de RAM, changement HDD par SSD
 - Serveur d'équipe de laboratoire : de l'occasion sauf HDD de 20TB (200TB net : 5k€)

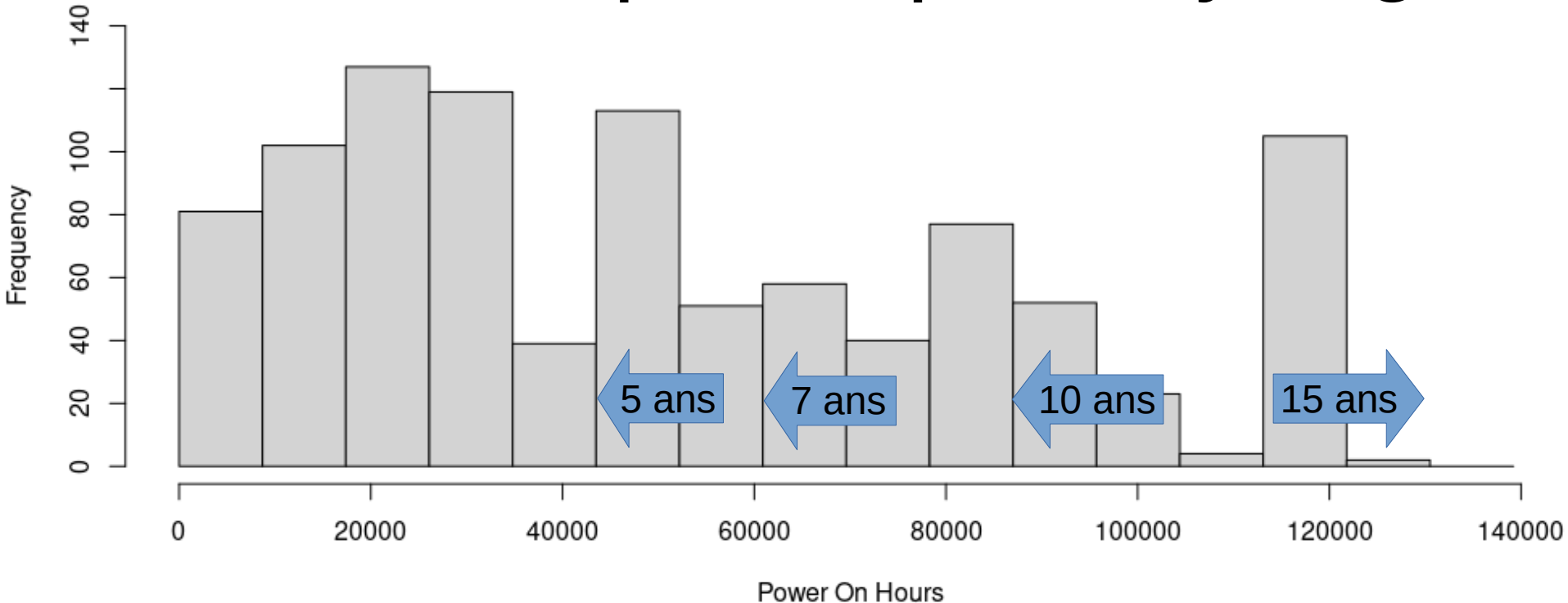
Action #1 « Avant & Après » : privilégier les « cycles courts »

« Les déchets des uns sont les ressources des autres. »

- Constats (implacables) :
 - Pas de fabrication « locale »
 - Ressources inexploitées à proximité
- Quelles actions « Avant » :
 - récupération, requalification,
 - démontage, détournement,
 - achat d'occasion chez broker
- Quelles actions « Après » :
 - Cession des machines inexploitées



Et les disques durs (50kgCO₂/HD) Réaffectation plutôt que recyclage...



- Des disques plus « résistants » que la garantie,
 - Donc la garantie n'est pas une DLC (Date Limite de Consommation)
 - Une garantie n'est pas une garantie de fonctionnement (loin de là !)

Action #1 : les « cycles courts » généralisables ?

Oui, une question de volonté mais il faut :

- Disposer de systèmes « résilients » (solutions propriétaires exclues)
- Avoir de quoi tester rapidement les arrivées de matériel
- Stocker le matériel de secours dans des endroits « sûrs »
- Connaître les usages pour adapter le matériel
- Privilégier les systèmes d'exploitation libres

Action #2 : recyclage/requalifier

Re-* des vieilles machines

- Requalifier des machines pour :

- Une salle de formation
- Des clusters de formation



- Exploiter des machines comme « hôte GPU »

- Anciennes stations de travail : MacPro, câbles & contrôleurs...
- Anciens nœuds de cluster : Supermicro, câbles & « rehausseurs »...



Action #3 : cycles arrêt/redémarrage sur 4 salles & 64 machines...

- Objectif : éteindre les machines des salles « hors cours »
- Opérations : allumage à 7h30, extinction à 19h30
- Craintes : vieillissement prématuré des HDD
 - Critères : cycle marche/arrêt, variation température, température max
- Adaptations :
 - Dans le BIOS, activation WoL & désactivation autres modes...
- Retour sur presque 2 années : indisponibilité rare...
 - Vieillesse du matériel à évaluer

Actions #4 : relocaliser le matériel déployer des « AnchiAles »

Truisme : chauffer les bureaux en récupérant la « chaleur fatale »

- Pas nouveau (ancien chauffage de véhicules « thermiques »)
- Contexte favorable : interdiction des chauffages d'appoint...
- Passage de 6 machines à 18 machines entre 2022 et 2023



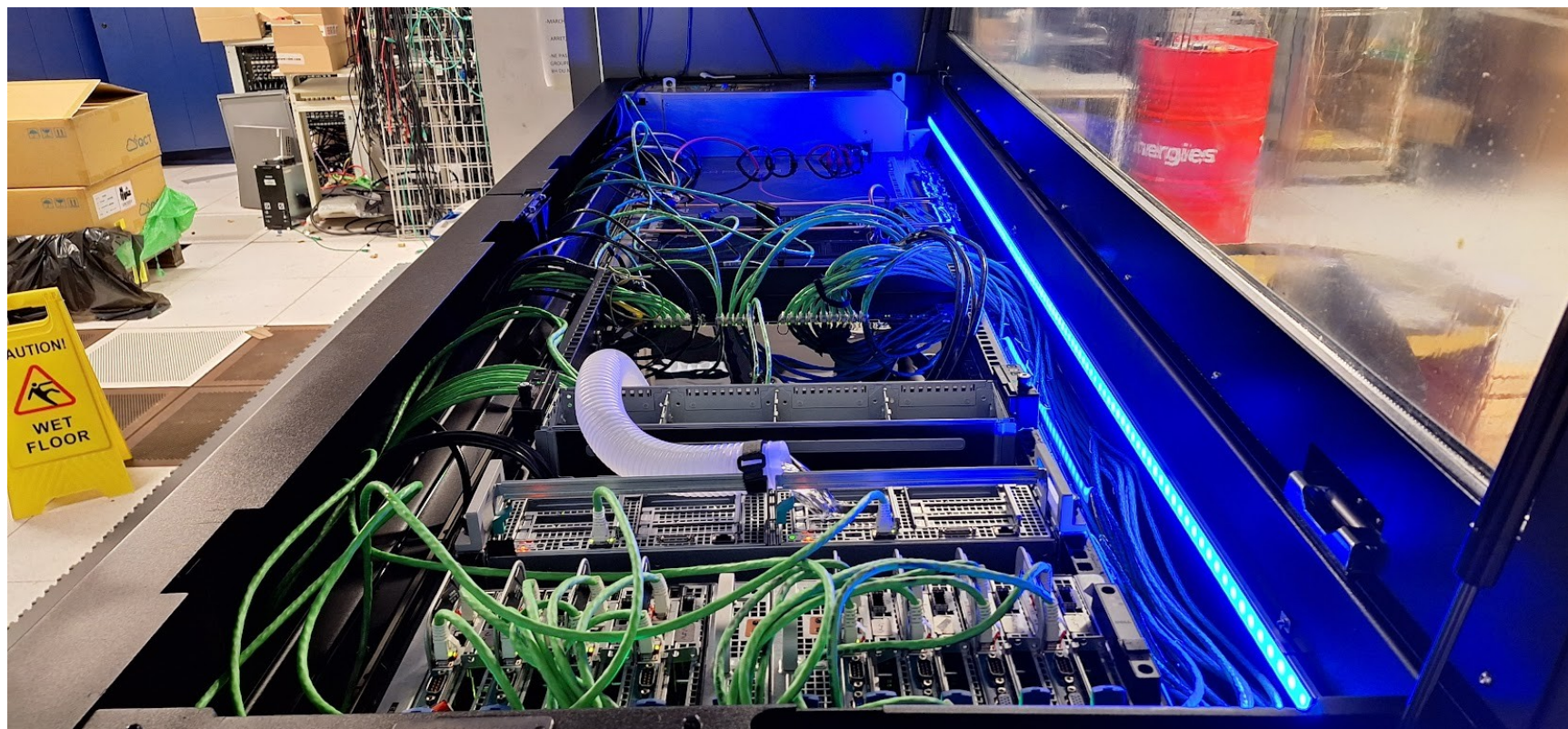
Action #4 : déploiement AnchlAles

Quel bilan ? Généralisation ?

- Plus de demandes que d'offres en 2023 :
 - Montage « rapide » de machine un WE
- La chaleur « ventilée » offre un confort meilleur
 - Une machine « à vide » (de 100W) offre un confort de radiateur de 1kW
- La généralisation exige une « modification » d'approche
 - Des équipements « plus » génériques (salle machine, bureau, ...)
 - Une gestion de la transhumence entre les saisons
 - Un placement estival dans des zones « hors d'eau »
 - Des bacs d'immersion « individuels » à diffuser dans les locaux...

Action #5 : immerger les machines

Limiter la PUE à ~ 1 voire < 1



Action #5 : immerger les machines

Pas « spécialement » neuf...

Il y a une génération (humaine)...
Un film de série B en 1984

- 1984 : The Last Starfighter
 - 27 minutes d'images synthétiques
 - ~ $30 \cdot 10^9$ opérations par image
 - Utilisation d'un Cray X-MP (130 kW)
 - 68 jours (en fait, 1 année nécessaire)



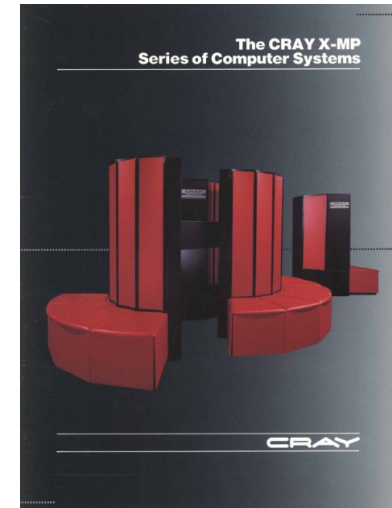
- 2020 : RTX 3090 (350 W)
 - 33 secondes
 - Comparaison RTX 3090 / Cray
 - Performance : 178 000 !
 - Consommation ~ 66 000 000 !



Emmanuel QUEMENER CC BY-NC-SA
December 6, 2021

CBP

11/127



The dense concentration of components requires special cooling techniques to overcome the accompanying problems of heat dissipation. A proven, patented cooling system using liquid refrigerant maintains the necessary internal system temperature, contributing to high system reliability and minimizing the need for expensive room cooling equipment.

Pourquoi plus rien (ou presque) en 40 ans ?

Immerger les machines : Périmètre d'action digne de l'école

- Pour couvrir tous les aspects de cette « transition » air vers huile:
 - Volet scientifique : efficacité de l'immersion, recyclage de la chaleur, ..
 - Volet technique : adaptation, transformation des équipements, évolution composants
 - Volet opérationnel : exploitation quotidien, sécurité associée
- Et pourquoi l'ENS de Lyon alors ?
 - Pour la partie scientifique : 2021-2022, intégration de l'étude au LIP, équipe Avalon
 - Pour la partie HPC ou Cloud : tout est internalisé
 - Pour le CBP : ressources « RADIS » (Reproductibles, Adaptables, Diverses, Interactives, Simples)
- Pour les 3 volets, depuis octobre 2022, uniquement Cloud@CBP

Préparation pour l'immersion v1: supprimer « tout ce qui bouge »

- Dans l'air : évacuer les « calories » :
 - On multiplie la surface de contact : x200 pour un radiateur
 - On diffuse la chaleur par conduction ou convection (caloduc)
 - On chasse avec un ventilateur l'air chauffé :
 - Dans un serveur : ventilateur de 4 à 6 cm, rotation de 3000 à 20000 tours/min
 - Dans une station : 8 à 16 cm, rotation de 500 à 1000 tours/min
- Pour préparer les machines :
 - Supprimer la pâte thermique : processeur et radiateur en contact direct
 - Supprimer les ventilateurs (là en le conservant dans l'alimentation)
 - Supprimer les « guides » plastiques

L'immersion sans modification

Sans ventilateur, ça démarre & ça tourne... ici...

- Au repos : oil 74 W, air 68 W
 - Températures : oil 24°C, air 25°C
- En charge : oil 142 W, air 142 W
 - Températures : oil 36°C, air 38.5 °C
 - Mais en IPMI et AC : 140 W pour les deux
 - Mais en IPMI et DC : oil 115 W, air 120 W
- Côté performance :
 - 1 % de différence...



Bref, ça fonctionne mais pas de gain (ou presque)...

Mais pas pour tous les composants

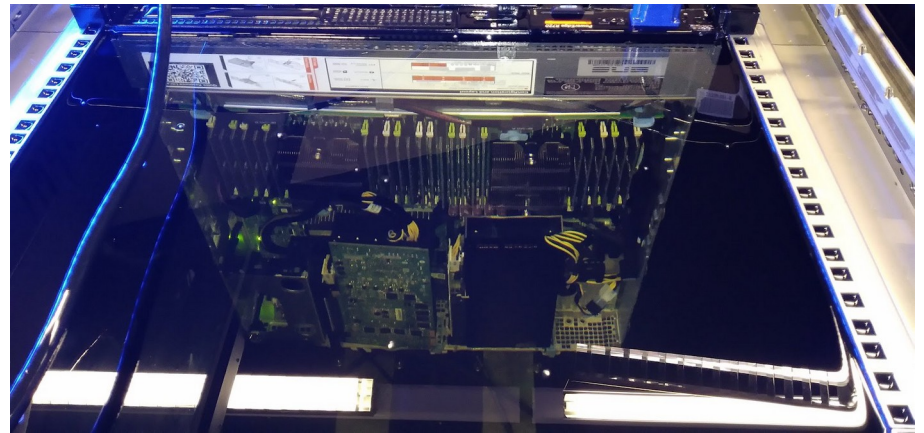
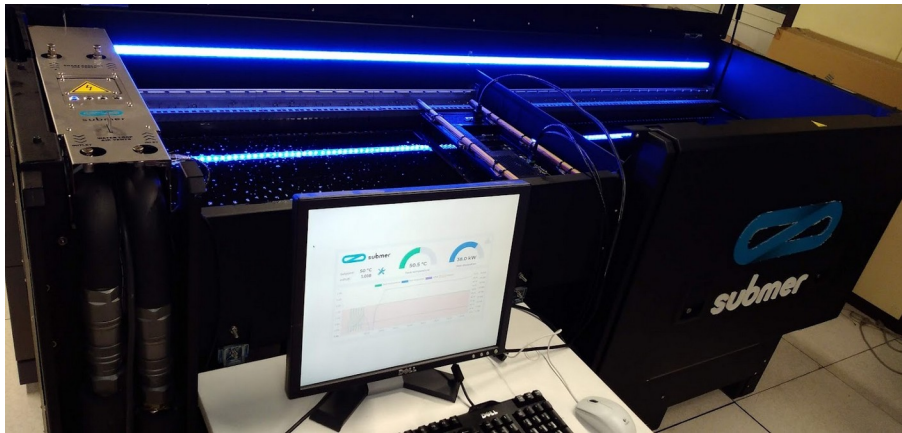
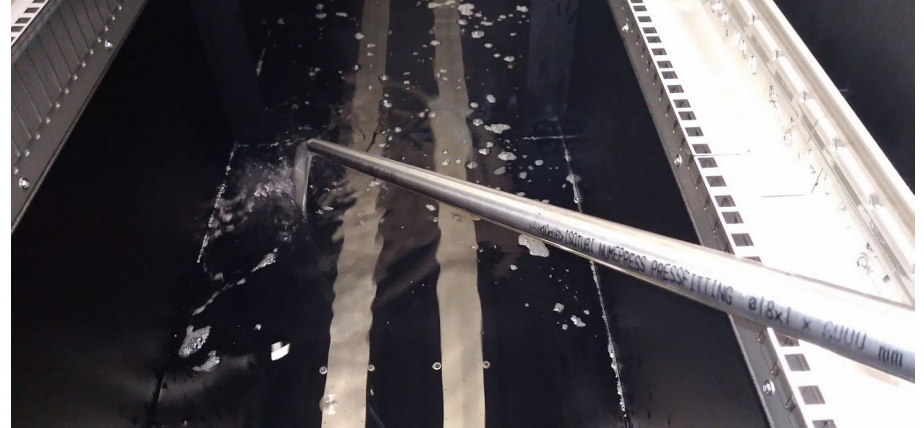
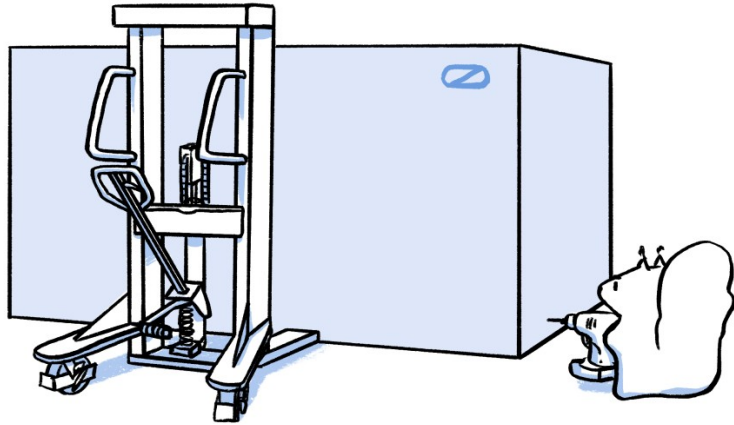
Petit voyage d'un disque dur en immersion

- Un HD de 2TB SAS « smart OK » installé
- Le serveur R720 est démarré
- Après 40 secondes, HD de 2TB détecté
- Un badblocks invasif est lancé
- Après 2h21'3" première erreur I/O
- Après 2h22'22", seconde salve d'erreurs
 - « erreur de positionnement mécanique »...
- Expertise médico-légale en cours...

```
numa@casimir: ~  
File Edit View Search Terminal Help  
[ 24.424473] sd 0:0:0:0: [sda] 3907029168 512-byte logical blocks: (2.00 TB/1.82 TiB)  
[ 24.425404] sd 0:0:0:0: [sda] Write Protect is off  
[ 24.425455] sd 0:0:0:0: [sda] Mode Senses: d7 00 10 00  
[ 24.426709] sd 0:0:0:0: [sda] Write cache: enabled, read cache: enabled, supports DPO and FUA  
[ 24.586604] sd 0:0:0:0: [sda] Attached SCSI disk  
[ 8463.931829] sd 0:0:0:0: [sda] tag#20 FAILED Result: hostbyte=DID_ERROR driverbyte=DRIVER_OK cmd_age=12s  
[ 8463.931934] sd 0:0:0:0: [sda] tag#20 CDB: Read(10) 28 00 38 84 40 e0 00 01 00 00  
[ 8463.932006] blk update request: I/O error, dev sda, sector 948191456 op 0x0:(READ) flags 0x00700 phys_seg 14 prio class 0  
[ 8463.932125] sd 0:0:0:0: [sda] tag#21 FAILED Result: hostbyte=DID_ERROR driverbyte=DRIVER_OK cmd_age=11s  
[ 8463.934396] sd 0:0:0:0: [sda] tag#21 CDB: Read(10) 28 00 38 84 41 e0 00 01 00 00  
[ 8463.936627] blk update request: I/O error, dev sda, sector 948191712 op 0x0:(READ) flags 0x00700 phys_seg 1 prio class 0  
[ 8514.872160] sd 0:0:0:0: [sda] tag#15 FAILED Result: hostbyte=DID_ERROR driverbyte=DRIVER_OK cmd_age=12s  
[ 8514.874534] sd 0:0:0:0: [sda] tag#15 CDB: Read(10) 28 00 e8 e0 88 a0 00 00 00 00  
[ 8514.876820] blk update request: I/O error, dev sda, sector 3907029152 op 0x0:(READ) flags 0x00700 phys_seg 1 prio class 0  
[ 8534.871117] sd 0:0:0:0: [sda] tag#14 FAILED Result: hostbyte=DID_ERROR driverbyte=DRIVER_OK cmd_age=12s  
[ 8534.873456] sd 0:0:0:0: [sda] tag#14 CDB: Read(10) 28 00 00 00 00 00 00 00 00 00  
[ 8534.875685] blk update request: I/O error, dev sda, sector 0 op 0x0:(READ) flags 0x00700 phys_seg 1 prio class 0  
[ 8542.744970] sd 0:0:0:0: [sda] tag#15 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=7s  
[ 8542.747193] sd 0:0:0:0: [sda] tag#15 Sense Key : Hardware Error [current]  
[ 8542.749335] sd 0:0:0:0: [sda] tag#15 Add. Sense: Mechanical positioning error  
[ 8542.751452] sd 0:0:0:0: [sda] tag#15 CDB: Read(10) 28 00 00 00 00 00 00 00 00 00  
[ 8542.753529] blk update request: critical target error, dev sda, sector 0 op 0x0:(READ) flags 0x0 phys_seg 1 prio class 0  
[ 8542.755606] Buffer I/O error on dev sda, logical block 0, async page read  
[ 8583.595997] sd 0:0:0:0: [sda] tag#3 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=8s  
[ 8583.598111] sd 0:0:0:0: [sda] tag#3 Sense Key : Hardware Error [current]  
[ 8583.600101] sd 0:0:0:0: [sda] tag#3 Add. Sense: Mechanical positioning error  
[ 8583.602068] sd 0:0:0:0: [sda] tag#3 CDB: Read(10) 28 00 00 00 00 00 00 00 00 00  
[ 8583.604007] blk update request: critical target error, dev sda, sector 0 op 0x0:(READ) flags 0x0 phys_seg 1 prio class 0  
[ 8583.605987] Buffer I/O error on dev sda, logical block 0, async page read  
[ 8616.894326] sd 0:0:0:0: [sda] tag#0 FAILED Result: hostbyte=DID_ERROR driverbyte=DRIVER_OK cmd_age=152s  
[ 8616.896447] sd 0:0:0:0: [sda] tag#0 CDB: Read(10) 28 00 38 84 40 e0 00 00 00 00  
[ 8616.898490] blk update request: I/O error, dev sda, sector 948191456 op 0x0:(READ) flags 0x0 phys_seg 1 prio class 0  
[ 8616.900601] Buffer I/O error on dev sda, logical block 118523932, async page read  
[ 8646.464430] sd 0:0:0:0: [sda] tag#2 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=29s  
[ 8646.466591] sd 0:0:0:0: [sda] tag#2 Sense Key : Hardware Error [current]  
[ 8646.468637] sd 0:0:0:0: [sda] tag#2 Add. Sense: Mechanical positioning error  
[ 8646.470696] sd 0:0:0:0: [sda] tag#2 CDB: Read(10) 28 00 00 00 00 00 00 00 00 00  
[ 8646.472757] blk update request: critical target error, dev sda, sector 0 op 0x0:(READ) flags 0x0 phys_seg 1 prio class 0  
[ 8646.474883] Buffer I/O error on dev sda, logical block 0, async page read  
[ 8669.149348] sd 0:0:0:0: [sda] tag#2 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=126s  
[ 8669.151483] sd 0:0:0:0: [sda] tag#2 Sense Key : Hardware Error [current]  
[ 8669.153392] sd 0:0:0:0: [sda] tag#2 Add. Sense: Mechanical positioning error  
[ 8669.155070] sd 0:0:0:0: [sda] tag#2 CDB: Read(10) 28 00 00 00 00 00 00 00 01 00 00  
[ 8669.156884] blk update request: critical target error, dev sda, sector 0 op 0x0:(READ) flags 0x0 phys_seg 32 prio class 0  
[ 8692.157705] sd 0:0:0:0: [sda] tag#0 FAILED Result: hostbyte=DID_OK driverbyte=DRIVER_SENSE cmd_age=2s  
[ 8692.160016] sd 0:0:0:0: [sda] tag#0 Sense Key : Hardware Error [current]  
[ 8692.162239] sd 0:0:0:0: [sda] tag#0 Add. Sense: Mechanical positioning error
```

Bref, la durée de vie d'un disque dur « classique » est pas ouf...

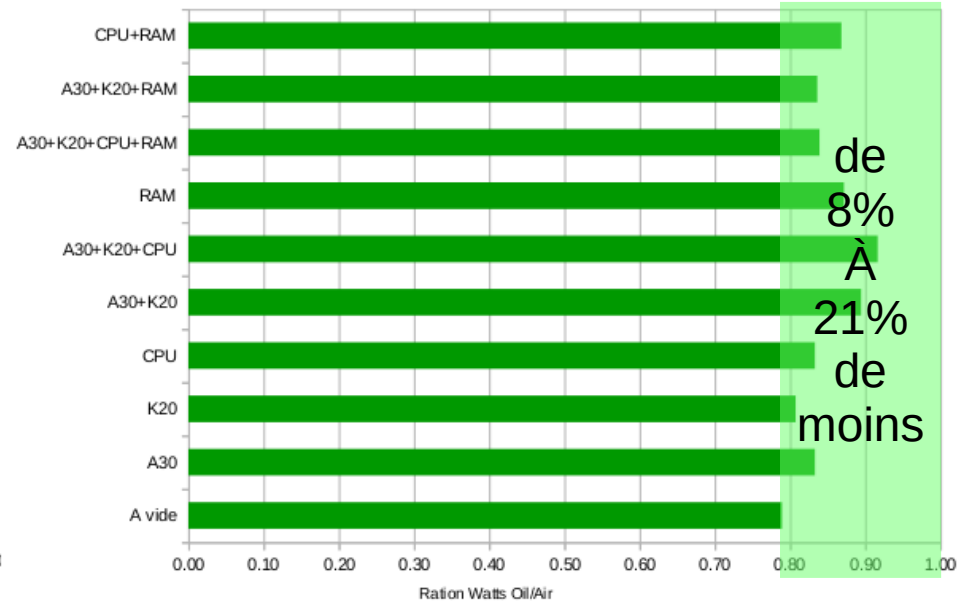
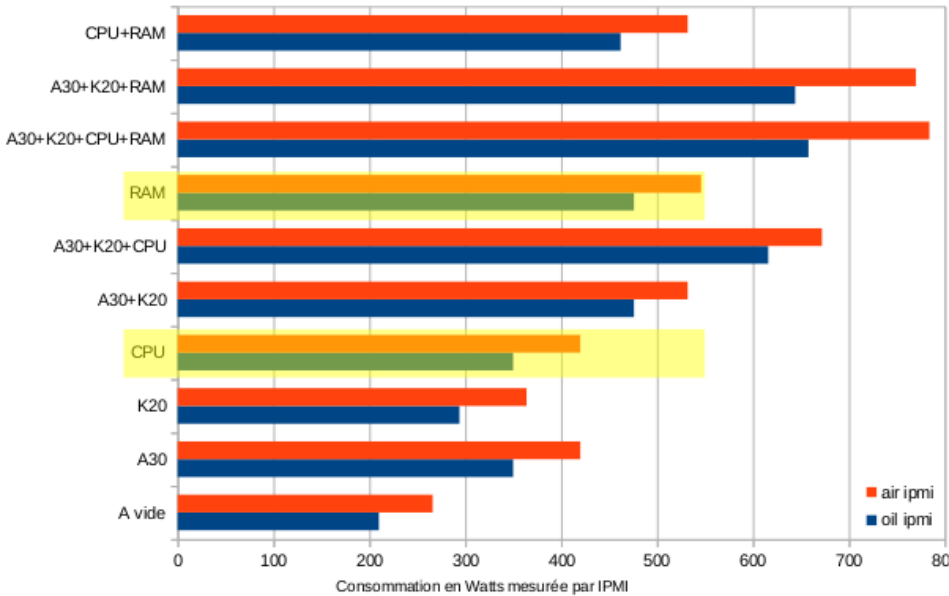
Après un couple « bac-huile » à oublier, retour aux sources...



Une installation un peu plus compliquée...

Et est-ce que ça consomme moins?

Expérience sur machine à GPU



- Des gains à l'immersion significatifs avec :
 - à vide, 21 % en moins
 - en charge, entre 8 % et 19 % de moins
- Mais une « grosse » sensibilité au « cas d'usage »...

Préparation de machines v2 : des difficultés diverses...

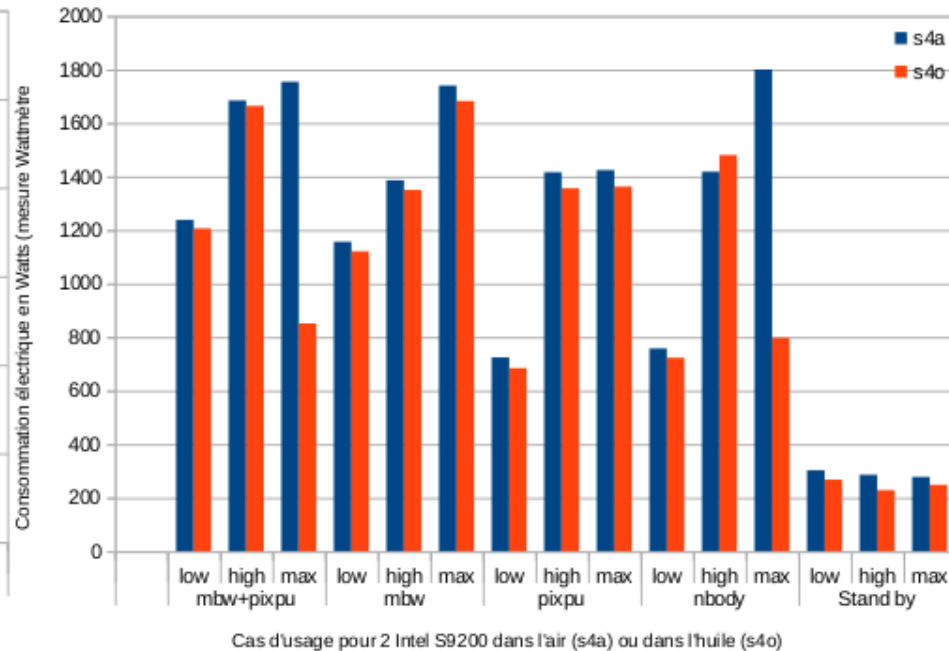
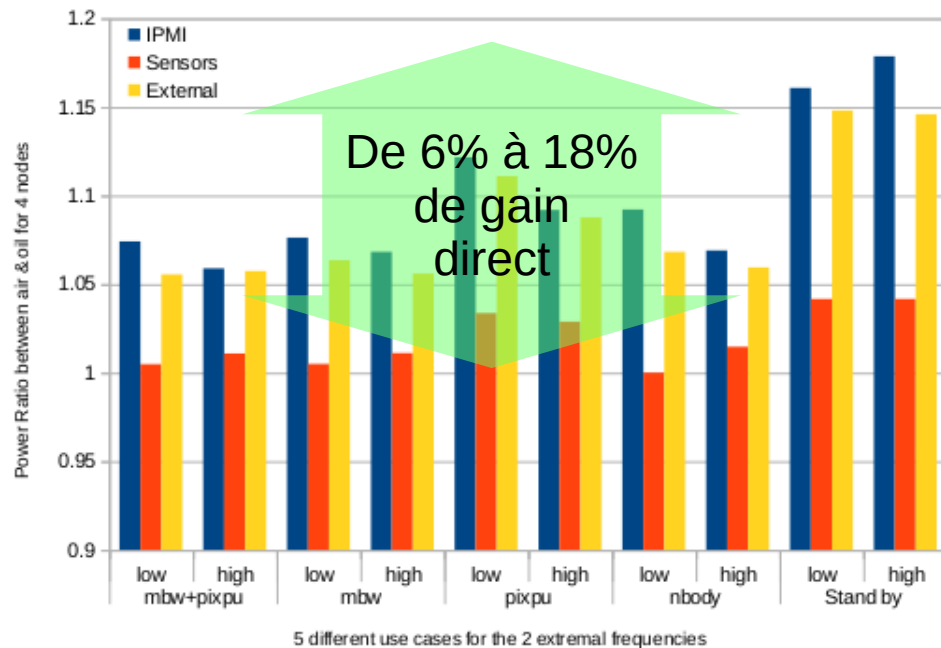
- L'an dernier : « supprimer » comme seules « actions »
 - Supprimer la pâte thermique : processeur et radiateur en contact direct
 - Supprimer les ventilateurs (là en le conservant dans l'alimentation)
 - Supprimer les « guides » plastiques
- Maintenant : « adapter » comme « exigence cardinale »
 - Retourner les ventilateurs : mauvaise orientation du flux
 - Fixer les nœuds verticalement : renforcer des crochets de retenue « inadaptés »
 - Modifier les cartes de supervision : « faire accepter » des ventilateurs absents
 - Modifier les consignes de fonctionnement : « forcer » le fonctionnement...

Mais problème récurrent : la profondeur limitée du bac...

Et côté consommation sur CPU ?

Des résultats convergents...

- Des ratio de consommation :
 - sensiblement différents en fonction de leur mesure, de l'ordre de 10 %...
 - Gain à l'immersion surtout pour des machines « à vide »

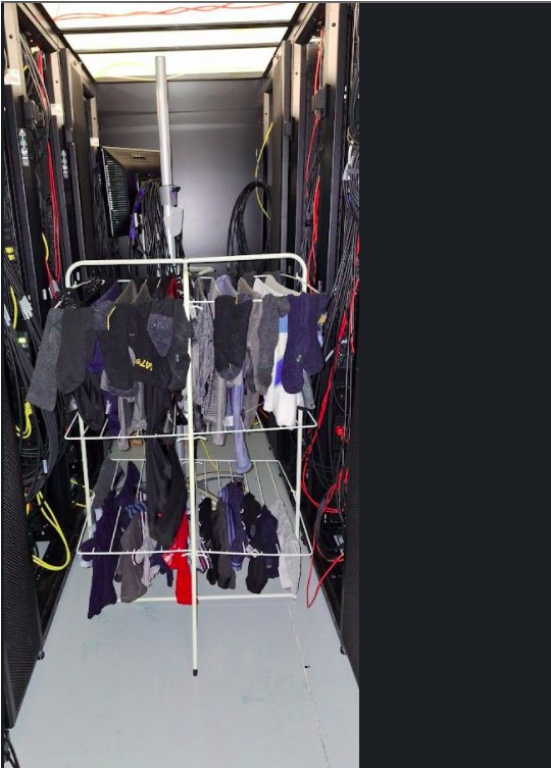


Action #5 : avantage de l'immersion

Des avantages aux inconvénients...

- Constats d'exploitation après 2 ans :
 - Réduction de consommation « directe » de 10 % à 20 %
 - Température de l'huile jusqu'à 55°C sans pertes de performance
 - Bonne tenue des équipements (un taux de panne comparable)
- Avantages :
 - Refroidissement de tous les composants sans distinction de nature (barrettes de RAM)
 - Adaptation possible de l'existant ou exploitation de « matériel générique »
 - Exploitation de la chaleur fatale plus facile (eau à 50°C)
 - Exploitation d'huile « bio-sourcée » (huile de friture désoxygénée)
- Inconvénients :
 - En plus des « risques » électriques, des « risques » chimiques
 - Frilosité des intégrateurs depuis quelques années

Action 6# : dans le Data Center Améliorer le refroidissement... En séchant son linge !



Emmanuel Quemener • Vous
Pilote d'essais informatiques et couple alph...
3 mois •

Quand on parle de récupération de la **#ChaleurFatale** dans le **#DataCenter**, on pense d'abord à une installation complexe intégrant généralement une pompe à chaleur. Pourtant, quiconque exploite massivement un data center (je m'occupe de plus de 200 machines dans celui de l'**École normale supérieure de Lyon**), sait qu'il existe un problème cardinal à l'exploitation du caloporteur "air" : il est d'autant plus efficace qu'il est humide et les data center sont généralement "trop" secs, avec un taux d'humidité frôlant les 20%, alors qu'il faudrait que cela soit plutôt 50%. Une manière simple, instantanée, de relever ce taux d'humidité "utilement" est de mettre simplement son linge à sécher ! Je vous i...voir plus

114 8 commentaires · 6 republications

J'aime Commenter Partager

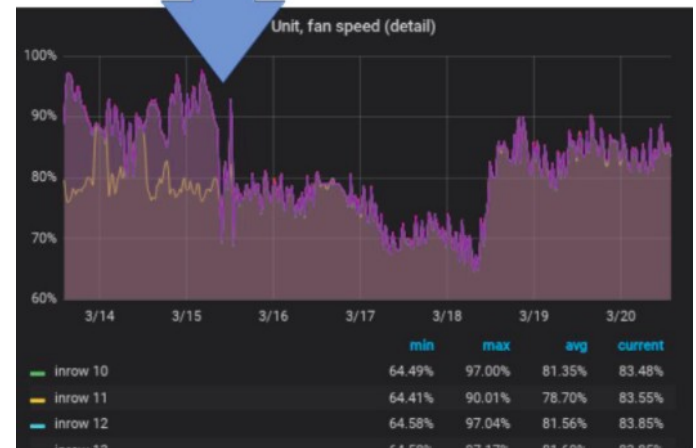
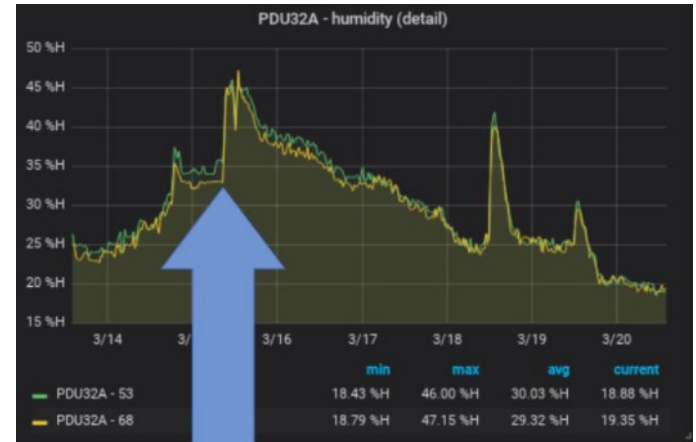
10 110 impressions Voir les statistiques

Ajouter un commentaire...

Les plus pertinents

John Morelle • 1er
Responsable d'Affaires HPC & AI chez ...
3 mois ...

Méthode expérimentale très sympathique tout en demeurant finalement efficace dans sa démonstration ! Par déduction, + les personnes sont grandes » et ...voir plus

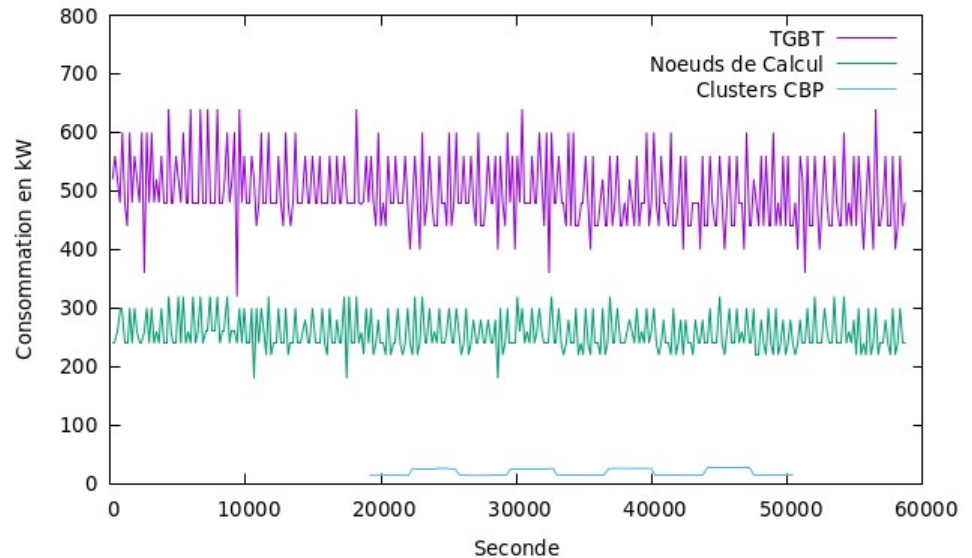
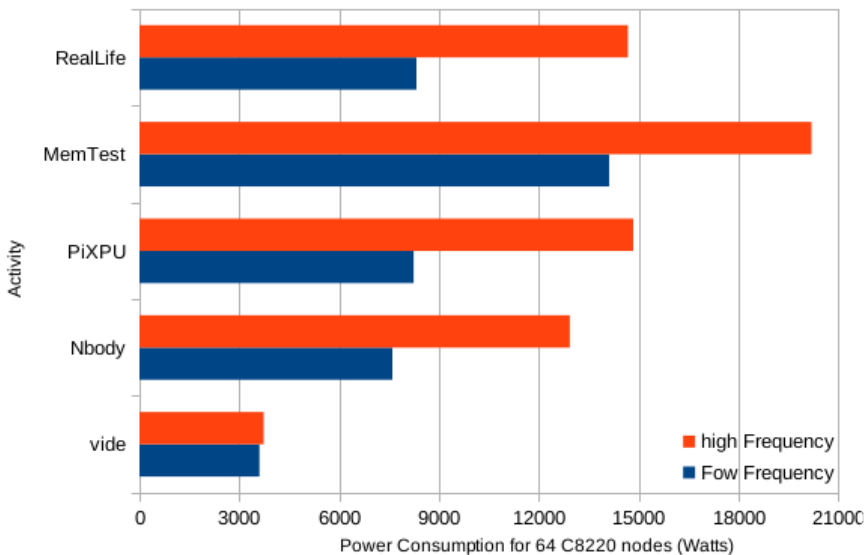


Un exemple de win-win

Action #7 : dans le DataCenter

Limiter sans gréver le HPC

- Un prérequis : laisser à l'OS le contrôle
 - Et donc paramétrer le BIOS avec la fréquence en « OS Control »
- Expériences habituelles + cas d'usage sur 64 nœuds

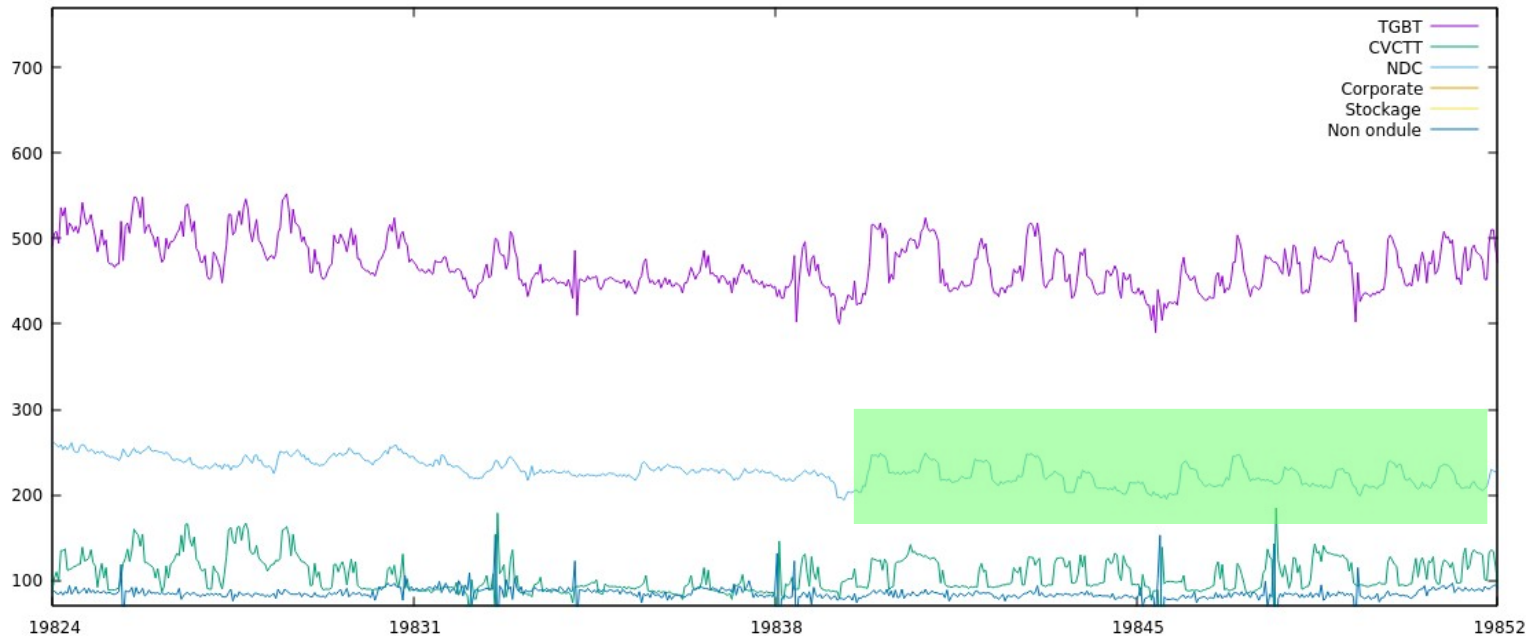


- Pas grand-chose à l'échelle des 800 nœuds du DC...
 - Généralisation en cours sur le Mésocentre PSMN

Action #8 sur PSMN

approche multiple

- Limiter la fréquence le jour de 6h à 22h (divisée par 2)
- Ne pas exécuter de nouvelles tâches le jour



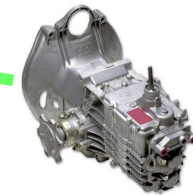
- De l'ordre de 20 % de consommation en moins...

En conclusion

Consommer moins ou mieux ?

- Déjà, la modulation de consommation est possible :
 - Par une extinction des installations « diurnes » (salles de cours)
 - Par une modulation de la fréquence pour les H24 7/7
- Consolider les installations ou les approches
 - Améliorer le PUE du Data Center existant : pompe à chaleur...
 - Créer un Data Center « en immersion » et exploiter la « chaleur fatale »
 - Distribuer des AnchiAles avec machines « immergées »
 - Renouveler son parc par un moins énergivore ?
 - C'est externaliser son empreinte carbone à l'extérieur

Au-delà du recyclage traditionnel, un état d'esprit : petit exemple...



Recycler palettes ou ventilateurs de machines...