Conférence Aramis 2023

Consommer moins ou consommer mieux ?

Quelles approches simples adopter ?

Emmanuel Quémener

Des nouveaux usages… … aux nouveaux contextes !

- L'IT a une part croissante (de consommation) :
  - Parce qu'on s'en sert de plus en plus (pas une discipline épargnée)
  - Parce que les outils sont plus gourmands (Machine Learning en tête)
- Mais un contexte de rentrée 2022 unique :
  - Emergence (tardive ?) d'une « conscience écologique »
  - Augmentation drastique du coût de l'énergie (électricité en tête x10)
  - Risques de coupure électrique cet hiver
- Questionnement tous azimuts de direction & laboratoires
- Comment aborder le problème ?

Une approche ... « intelligente » ?
« Moyens » de ses « ambitions »

Qu'est-ce que « l'intelligence » ?

- Pour un « latin » : un « état »
  - Intelligence : capacité d'abstraction dans la résolution de problèmes
- Pour un « anglo-saxon » : 3A pour un « objectif »
  - **Appréhension** : capacité à récupérer les informations
  - **Analyse** : capacité à analyser les informations collectées
  - **Action** : capacité à mettre en œuvre des processus
- Dans mon cas : à défaut de la « latine », prenons l'autre.

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 Potiential (unité kgCO2eq)
  - **PE** : Consommation de ressources énergétiques
    - Primary Energy (unité MJ)

Quelle empreinte carbone ? Finalement, 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)
- Le matériel est (partiellement) contrôlable par l'OS...

Quelle empreinte carbone ? Avant, pendant, après...

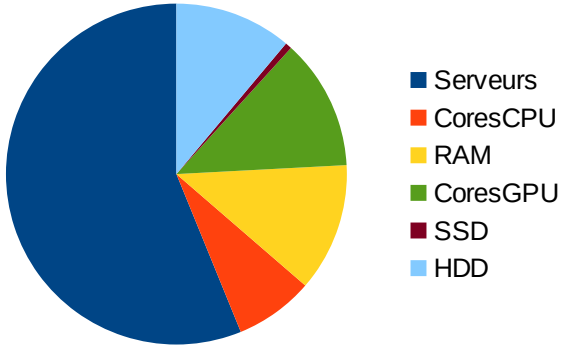
- **Avant** : sa fabrication (et son transport)
- **Pendant** : son exploitation (et sa maintenance)
- **Après** : son recyclage (et son transport, stockage)
- Contexte particulier en France : électricité décarbonée...
- Des « informations » contradictoires (avec l'expérience) :
  - Venant des constructeurs : configurateur Dell
  - Venant de la « communauté ESR » : guides EcoInfo, ...
- En fait, au CBPsmn, impression du « contre-exemple »

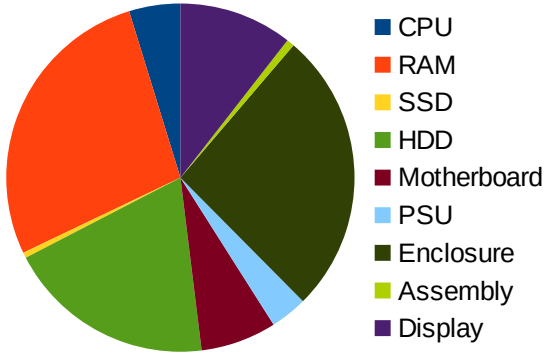
Quelle empreinte carbone ? Interrogations sur la « littérature »

- Petite expérience : *établir un devis sur Matinfo5*
  - Empreintes carbone identiques quel que soit : le modèle CPU, la RAM
- Littérature plus « consistante, cohérente, pertinente » :
  - https://boavizta.org/blog/empreinte-de-la-fabrication-d-un-serveur
  - Approche 2 : « facteurs d'émission arbitraires par composant »
    - servergwp(kgCO2eq) = 900(kgCO2eq) + cpuunits(unit) x 100(kgCO2eq/unit) + ramsize(GB) x 150/128(kgCO2eq/GB) + ssdunits(unit) x 100(kgCO2eq) + hddunits(unit) x 50(kgCO2eq) + gpuunits(unit) x 150(kgCO2eq/unit)
  - Approche 3 : « vers une formule de calcul d'impact multicritère » basé sur les semiconducteurs
    - server = cpu + ram + ssd + hdd + motherboard + psu + enclosure + assembly
    - cpu = cpuunits x ((cpucoreunits x cpudiesize + 0,491) x cpu_die + cpu_base)
    - ram = ramunits x ((ramsize / ramdensity) x ram_die + ram_base)
    - ssd = ssdunits x ((ssdsize / ssddensity) x ssd_die + ssd_base)
    - hdd = hddunits x hdd_unit
    - psu = psuunits x psuunitweight x psu_weight
    - enclosure = rack ou enclosure = blade x 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 CO_2 à la fabrication





- Exploitation : 300W H24 7/7
  - 40 tonnes CO_2 par an soit 4 françaises « moyennes »
  - Mais 200 utilisatrices différentes chaque mois !
- Ratio : 1 pour 12 à 1 pour 6… Classique en IT !

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 (increvable)
  - « data view » : laptop
- A moindre coût :
  - Étendre la capacité : distribution de + de 7TiB de RAM DDR3
  - Consolider 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

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

Action #2 : recyclage/requalifier Re-* des vieilles machines

- Requalifier des machines pour :
  - Une salle de formation
  - Des clusters de formation

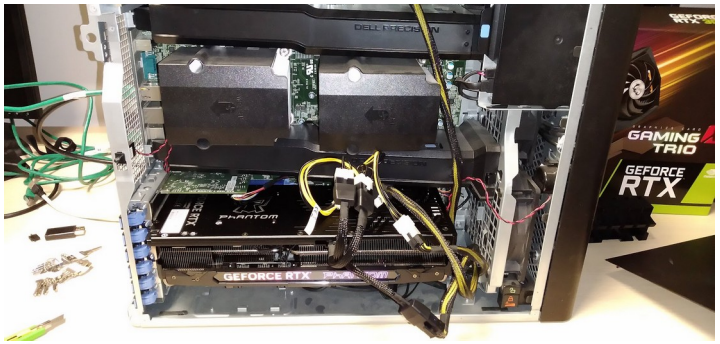



- Exploiter des machines comme « hôte GPU »
  - Anciennes stations de travail : MacPro, câbles & contrôleurs...
  - Anciens nœuds de cluster : Supermicro, câbles & « rehausseurs »...

Actions #3 : détourner approche progressive, mode KISS

Truisme : chauffer en récupérant le « déchet chaleur »

- Pas nouveau (ancien chauffage de véhicules « thermiques »)
- Contexte favorable : interdiction des chauffages d'appoint...



Action #4 : cycles arrêt/redémarrage sur 3 salles & 50 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 d'une année académique : indisponibilité rare…
  - Vieillissement du matériel à évaluer

Appréhension : récupérer les infos… En fait, pas si simple !

- Récupérer « localement » : échelle du composant
  - Via l'OS 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 COU	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@platinum4oil1:~#
```

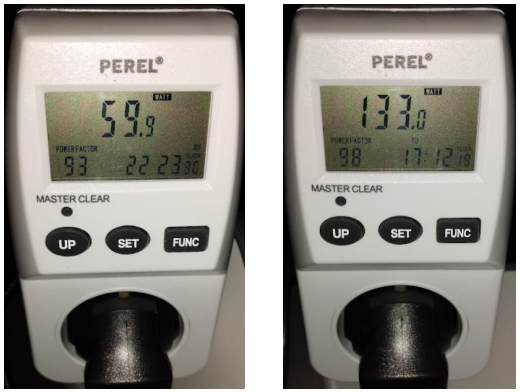
```
numa@casimir: ~
File  Edit  View  Search  Terminal  Help
Core 14:        +39.0°C  (high = +88.0°C, crit = +98.0°C)
Core 15:        +41.0°C  (high = +88.0°C, crit = +98.0°C)

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@apollo4oil1:~# 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@apollo4oil1:~#
```

```
numa@casimir: ~
File  Edit  View  Search  Terminal  Help
root@platinum4oil1:~# ipmitool sensor | grep Watts | awk -F'|' '{ print $1" "$2"
 "$3 }'
PS1 Input Power     225.000      Watts
PS3 Input Power     9.000        Watts
root@platinum4oil1:~#
```

Wattmètre



Pince ampèremétrique

Mesurer la consommation approche locale (et interrogation...)

Aucune généricité : une avalanche de cas particuliers

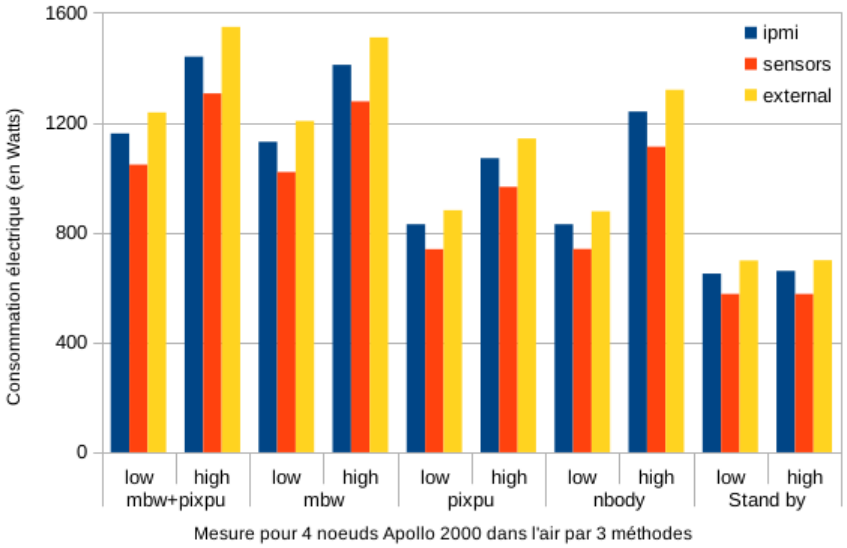
- Via « sensors » de l'OS, Cloud@CBP : 41/148
- Via « ipmitool » de l'OS, Cloud@CBP : 56/148

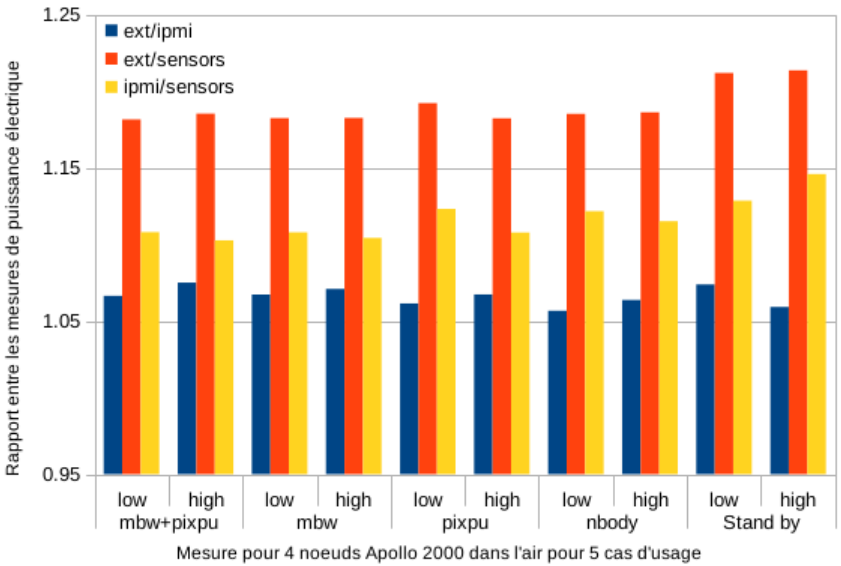
CPU Power	20.000
MB power	116.000
Mem Power	1.000
PS 1 Input	340.000
PS1 Input Power	8.000
PS 1 Output	0.000
PS 2 Input	330.000
PS2 Input Power	320.000
PS 2 Output	0.000
PS3 Input Power	225.000
PSU1 AC In Pwr	60.000
PSU1 DC Out Pwr	45.000
PSU1 PIN	210.000
PSU1 POUT	160.000
PSU2 AC In Pwr	0.000
PSU2 DC Out Pwr	0.000
PSU2 PIN	na
PSU2 POUT	na
Pwr Consumption	168.000
PWR_CPU1	80.000
PWR_GB_GPU0	75.000
PWR_GB_GPU1	63.000
PWR_GB_GPU2	70.000
PWR_GB_GPU3	73.000
PWR_PIN_PSU1	472.000
PWR_POUT_PSU1	288.000
Sys Fan Pwr	4.000
Sys Power	70.000
System Level	217.000

- Et quelle « pertinence » offrir à ces mesures ?
  - Exploitation d'une mesure « externe »...

Mesures « locales »
Quel comportement à la charge ?

Expérimentation : basse/haute fréquences, 5 cas d'usage





- Des mesures à appréhender avec précautions
  - La mesure OS via sensors la plus « instable »
  - Le ratio MesureWattmètre/MesureIPMI le plus « stable »
- Solution : MesureIPMI (compensée au besoin…)
- Mais le trouple Wattmètre/Webcam/Led reste une solution systématique...

Mesurer la consommation : échelle « globale »

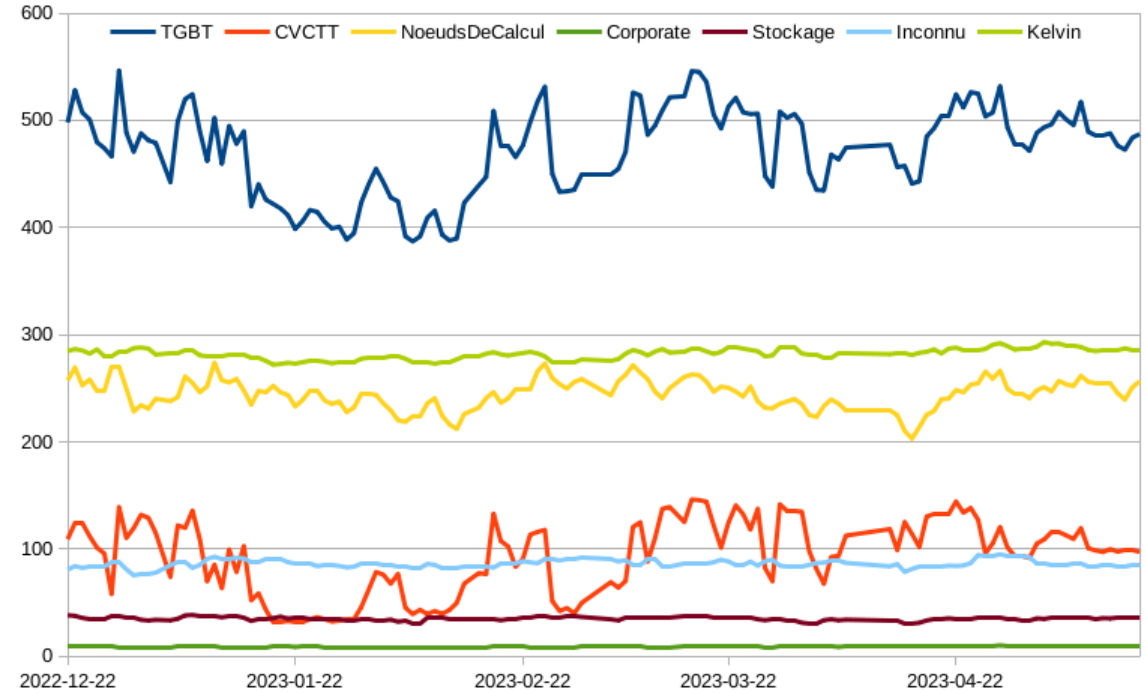
Passer d'un « framebuffer » web à une BDD Sqlite

- Ouvrir (manuellement) session x2go ouvrant un navigateur
- Convertir une capture d'écran en texte injectable dans une BDD
  - La commande « scrot » pour la capture
  - La commande « convert » pour le suréchantionnage
  - La commande « tesseract » pour l'OCR
  - La commande « sqlite3 » pour l'insertion

```
MYDB=/home/sing/ConsommationSingDB.sqlite
EPOCH=$(date "+%s")
[ "$USER" == "root" ] && cp /home/sing/.Xauthority /root
[ ! -d "/tmp/snapshots4$USER" ] && mkdir -p /tmp/snapshots4$USER
cd /tmp/snapshots4$USER
MYDISPLAY=$(ps aux | grep x2goagent | grep -v grep | grep sing | awk '{ print $NF }')
scrot -D $MYDISPLAY '%Y%m%d-%H%M%S-%s.png' -e 'convert $f -crop 488x240+516+344 -resize 480% -negate -sharpen 0x1 png:- | tesseract - text_$f 2>/dev/null ; rm $f ; echo text_$f.txt >LastFile ' ;
DATA=$(cat $(cat LastFile) | grep -Po '[^;\ kW]+(?=\ kW)' | tr '\n' ':' )
if [ $(echo $DATA | wc -w) -eq 1 ] && [[ ! "$DATA" == *"::"* ]]
then
    if [ $(cat $(cat LastFile) | wc -l) -lt 15 ]
    then
        SQL=$(echo $DATA | awk -v epoch="$EPOCH" -F':' '{ print "INSERT INTO JourEnCours VALUES ("epoch","$1","$3","$5","$7","$9","$11","$13");" }')
        sqlite3 $MYDB "$SQL"
        SQL=$(echo $DATA | awk -v epoch="$EPOCH" -F':' '{ print "INSERT INTO Cumul VALUES ("epoch","$2","$4","$6","$8","$10","$12","$14");" }')
        sqlite3 $MYDB "$SQL"
    else
        SQL=$(echo $DATA | awk -v epoch="$EPOCH" -F':' '{ print "INSERT INTO JourEnCours VALUES ("epoch","$1","$2","$3","$4","$5","$6","$7");" }')
        sqlite3 $MYDB "$SQL"
        SQL=$(echo $DATA | awk -v epoch="$EPOCH" -F':' '{ print "INSERT INTO Cumul VALUES ("epoch","$8","$9","$10","$11","$12","$13","$14");" }')
        sqlite3 $MYDB "$SQL"
    fi
else
    echo "Failed sanity check of data occurs : check the website..."
    echo "Failed sanity check of data occurs : check the website..." | mutt -s "Access problem to SING website" emmanuel.quemener@ens-lyon.fr
fi
```

Analyse du DataCenter la statistique (et son pentacle)

- Période du 22/12/2022 au 17/05/2023 : 5 mois…



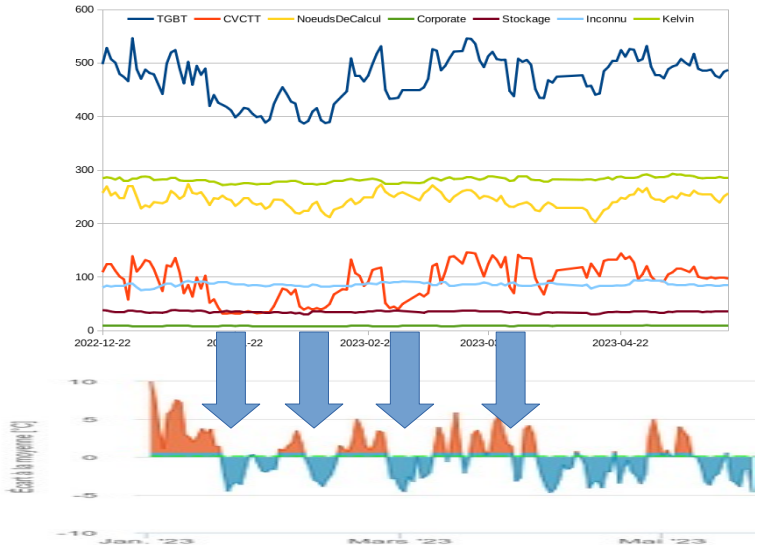
Légende sommaire...

- **TGBT** : départ électrique
- **CVCTT** : climatisation
- **NoeudsDeCalcul** : éponyme...
- **Corporate** : machines DSI
- **Stockage** : machines labos
- **Inconnu** : “non ondulé” et plus
- **Kelvin** : température à Lyon

- Moyenne : 478 kW, Médiane : 471 kW, Max-Min : 159 kW
- Une forte variabilité, mais quelle est sa « nature » ?

Analyse du DataCenter quelques analyses complémentaires

- Les plus gros postes de « consommation » :
  - Nœuds de calcul : de 46 % à 60 %
  - Climatisation (part PUE>1) : entre 8 % et 30 % du total (25 %)
  - « Inconnu » : de 16 % à 22 %
  - Stockage : de 6 % à 9 %
- Des corrélations (de Pearson) intéressantes :
  - 90 % de corrélation entre TGBT et CVCTT :
  - 80 % de corrélation entre température moyenne à Lyon et CVCTT
  - 47 % de corrélation entre NDC & Inconnu, 33 % de corrélation entre Stockage & Inconnu
- Des pistes d'améliorations ?
  - Limiter le CVCTT (améliorer le PUE), mais hors champ « utilisateurs »
  - Limiter la consommation du HPC (essentiellement NœudsDeCalcul)
- Avec la « nouvelle donne » : limiter sa consommation « instantanée »...

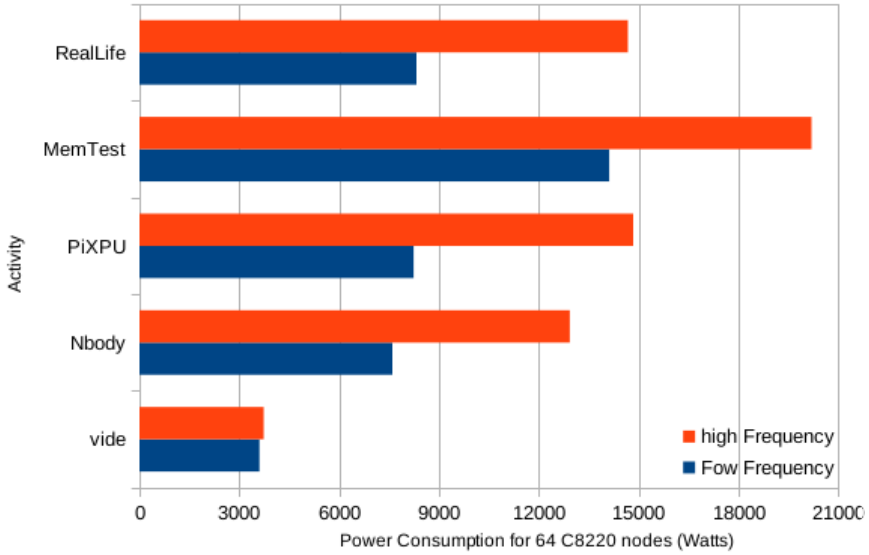


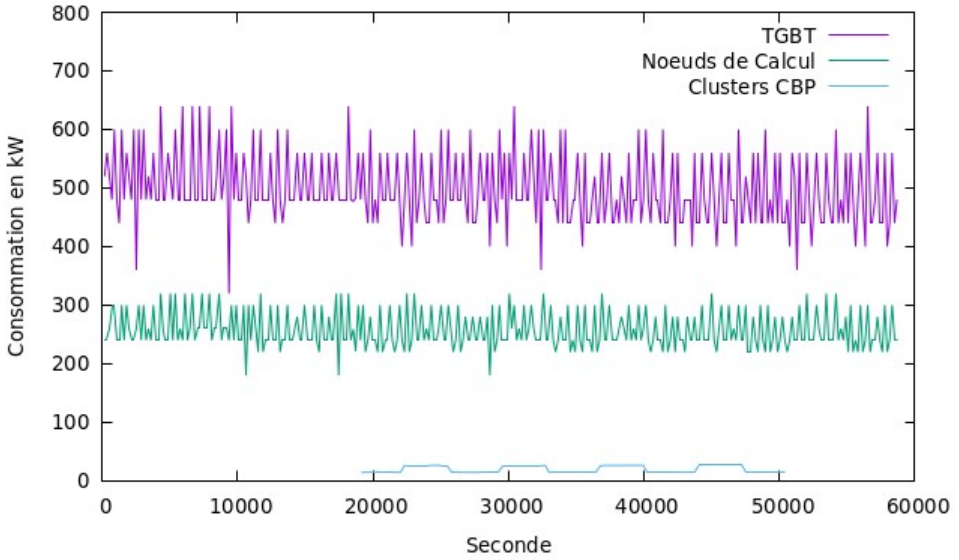
Actions DataCenter
Limiter sans gréver le HPC

- Des pistes :
  - Arrêter les machines, les redémarrer à la demande
  - Passer les machines en « suspend »
  - Passer en « on-demand » les fréquences
  - Forcer les fréquences sur les extrêmes
- Des obstacles :
  - Des redémarrages « peu prédicibles » (au moins en durée)
  - Des comportements en « suspend » plutôt chaotiques
  - Du « on-demand » efficace sur l'ALU, mais peu sur l'IO
- Forcer les fréquences : du simple & du systématique !

Actions 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

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
- En France, cas particulier d'une électricité décarbonée
  - Renouveler son parc par un moins énergivore ?
    - C'est externaliser son empreinte carbone à l'extérieur
  - Attitude simple et systématique :
    - Exploiter le plus longtemps possible les composants, même les boîtiers !
    - Consolider les performances par des extensions « judicieuses »
    - Immerger les équipements pour l'air dans l'huile : projet Immersion

Appel aux dons !!!
Computhèque comme sanctuaire

- Qui pourrait me fournir les composants suivants :
  - Carte contrôleur IDE sur port ISA 16 bits, disquettes 5.24 pouces
- Autrement, la computhèque du CBP accueille :
  - Tout équipement informatique le plus ancien possible :
    - Les vieux 8 bits des années 1980 : Sinclair ZX, Commodore, Oric, etc...
    - Les vieux PC avec des cartes ISA : 80286, 80386, ...
    - Les vieux périphériques : SCSI, scanner, disques durs, lecteurs de bande, etc...
  - Tout équipement informatique un peu exotique :
    - Machines de technologie : Dec Alpha 21264, HPPA, Sun...
- Merci pour votre générosité : james.mylq@ens-lyon.fr