



Immersion 2023

14 mois (en Immersion) :

De la mesure de la consommation électrique
À la gestion quotidienne de machines trempées...

Emmanuel Quémener

Ceux qui ont la chance de ne pas me connaître ... ne connaissent pas mon appétance...

... pour l'histoire (des sciences) !

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

- 1984 : The Last Starfighter
 - 27 minutes d'images synthétiques
 - $\sim 30 \cdot 10^9$ 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 $\sim 66\,000\,000$!



CBP

11/127



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

- Avec Cray X-MP :
 - 1982
 - 1 Gflops, 130 kW
- Avec HPE Frontier :
 - 2022
 - 1 Eflops, 21100 kW
 - 9248 nœuds
 - 36992 MI250X de 500 W
 - 94 % de Rpeak dans GPU
 - 87 % du total de la consommation annoncée...

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.



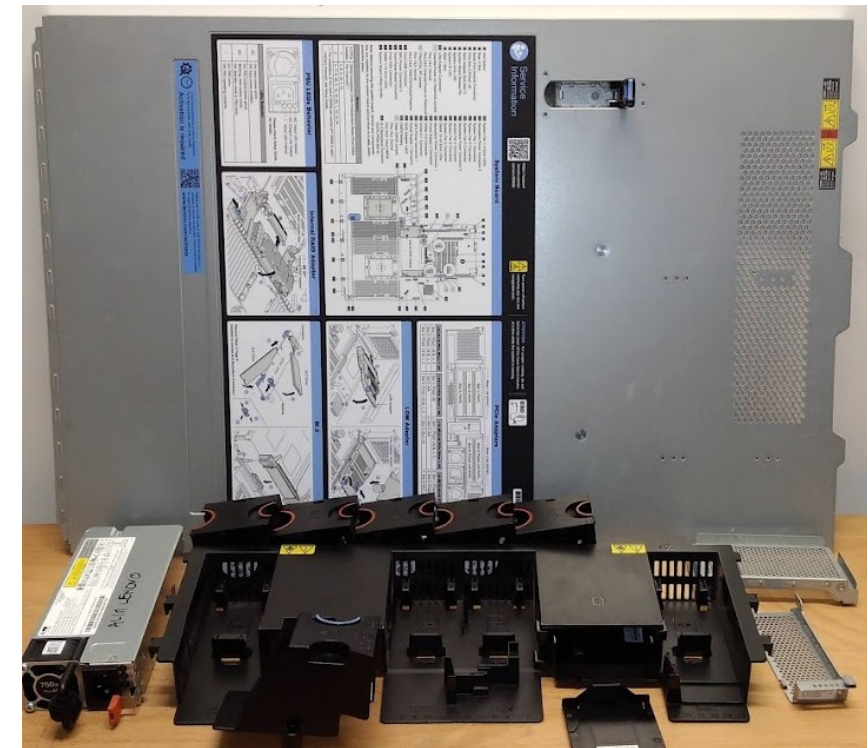
Retour aux « sources » : refroidir les CPU ou GPU dans un fluide?

Pourquoi cette étude à l'ENS de Lyon ?

- Pour couvrir tous les aspects de cette « transition » air vers huile:
 - Volet scientifique : efficacité de l'immersion, recyclage de la chaleur, ..
 - Volet technique : processus d'adaptation, de transformation des équipements, évolution composants
 - Volet opérationnel : processus d'exploitation quotidien, sécurité associée
- Et pourquoi l'ENS de Lyon alors ?
 - Pour la partie scientifique : de 2021 à 2022, intégration de l'étude au LIP, équipe Avalon
 - [1] T. Arabal, L. Betencour, E. Caron, and L. Lefevre. Setting up an experimental framework for analysing an immersion cooling system. In IEEE 34th International Symposium on Computer Architecture and High Performance Computing. SBAC-PAD 2022., Bordeaux, October 5-8 2022. To appear.
 - Pour la partie HP²C (*High Performance* ou *Hybrid Polymorph Computing*) : tout est internalisé
 - Pour le CBP : ressources « RADIS » (Reproductibles, Adaptables, Diverses, Interactives, Simples)
- Pour les 3 volets, depuis octobre 2022, uniquement CBP...

Préparation (initiale) pour l'immersion : 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 au 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
- Maintenant, un peu plus compliqué...



Des premières mesures plutôt déroutantes...

Sur serveurs Lenovo

- Une consommation comparable (modulo les variations de tensions d'entrée)



- Mais un « facteur de puissance » stable, en progression avec la charge

Même sur des machines « maison »

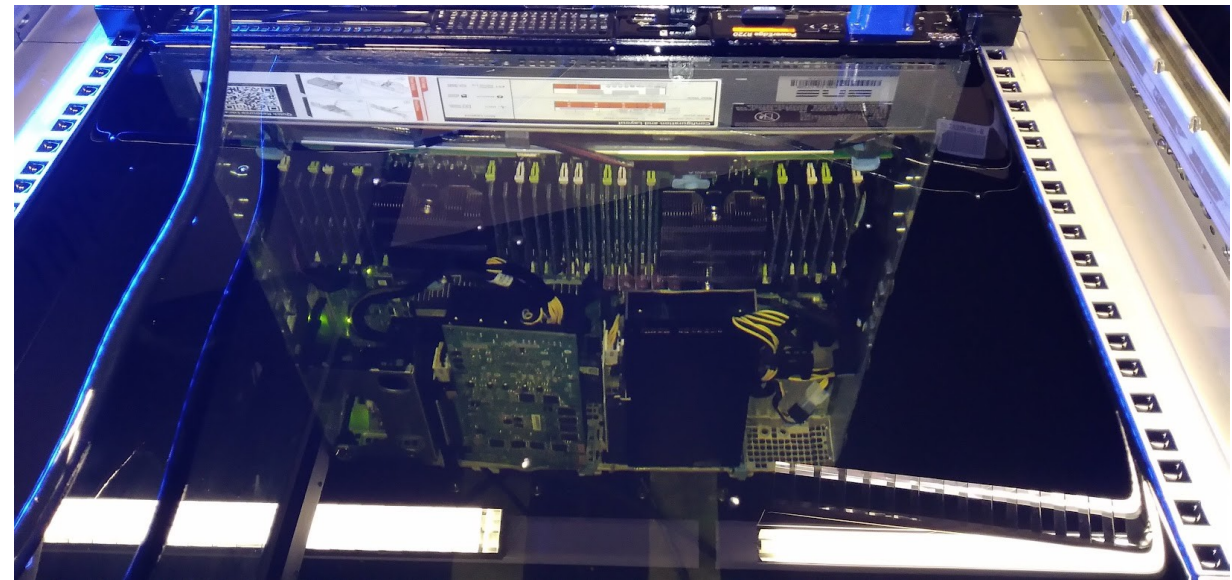
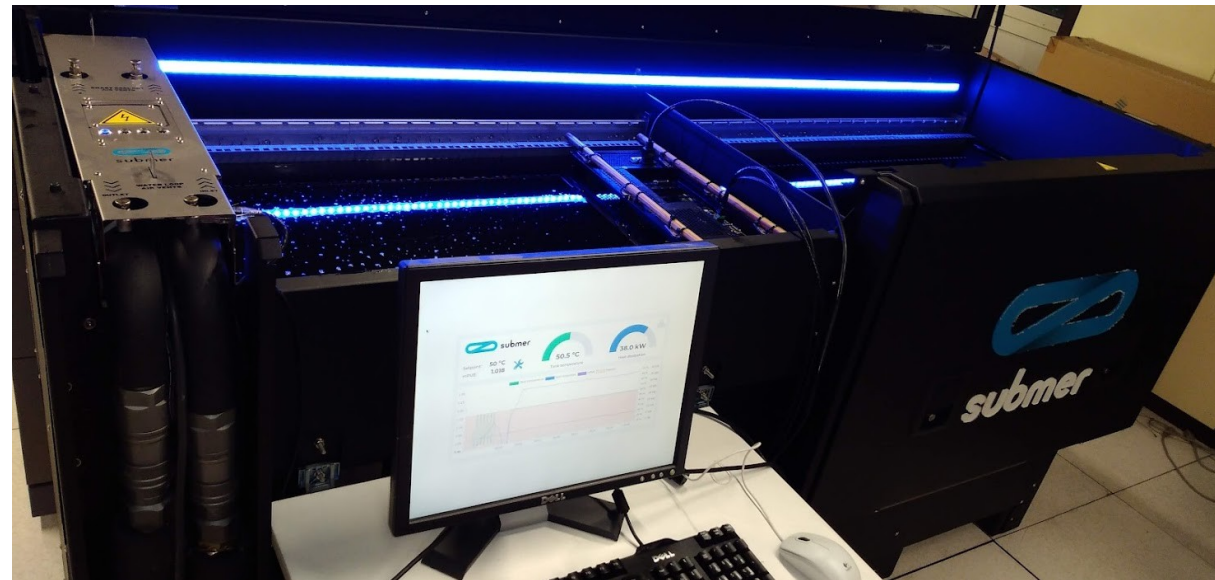
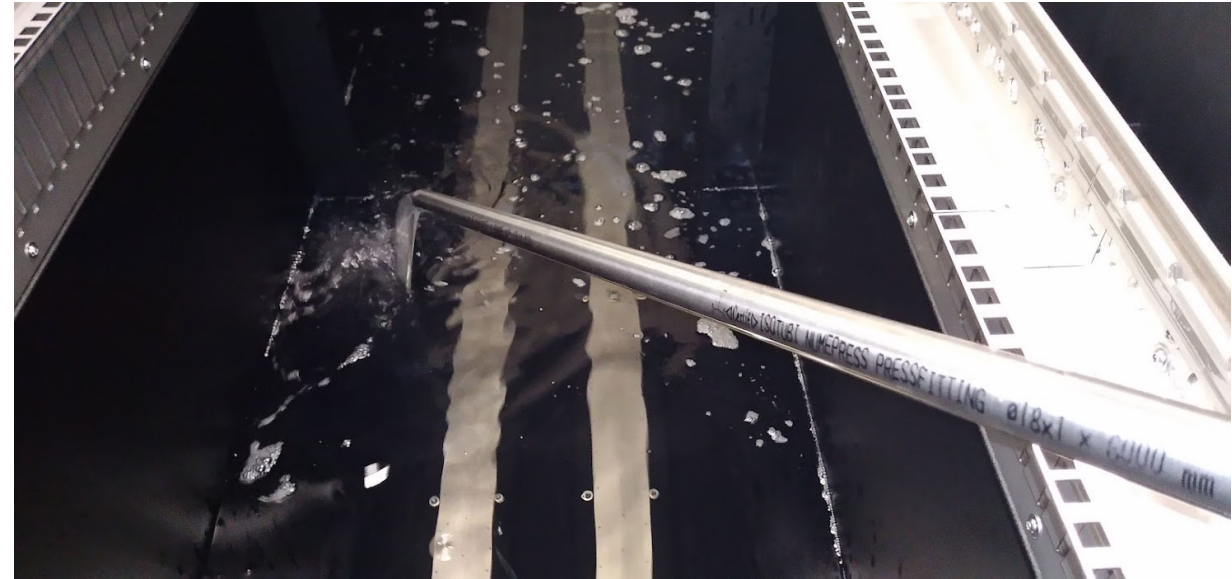
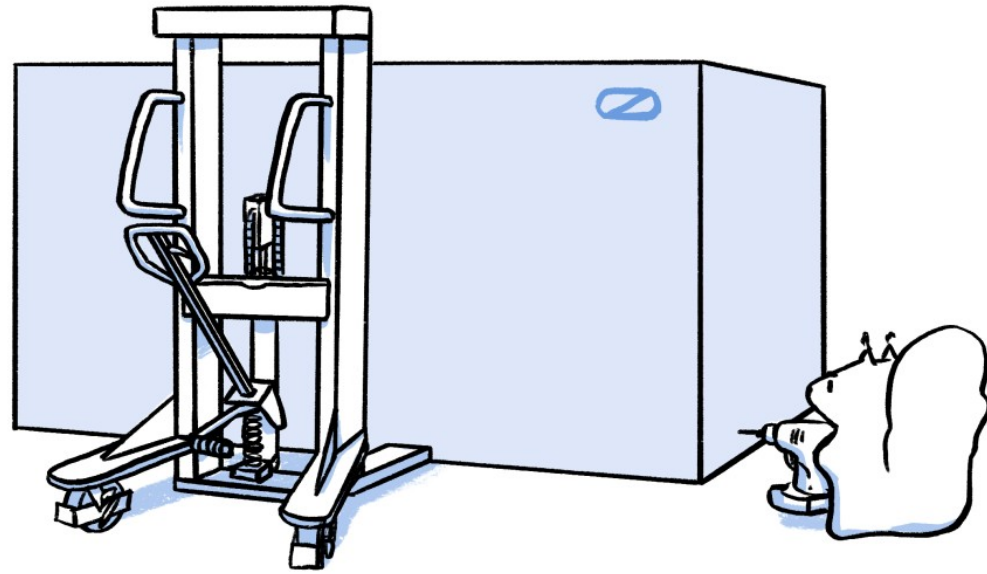
Sur les « Epyc à GPU »

- Une consommation toujours comparable (modulo les variations de tensions)



- Mais de grosses variations sur le « facteur de puissance » (et donc...)

Avec un bac Submer, premières analyses

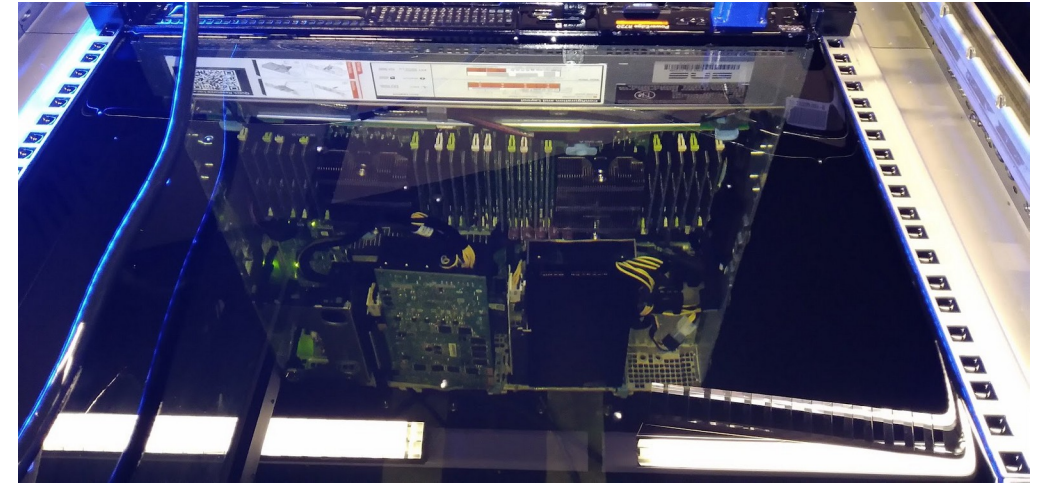


Nouveau bac, nouvelle huile, nouvelles machines, GPU et études !!!

Etudes sur GPU, CPU & RAM

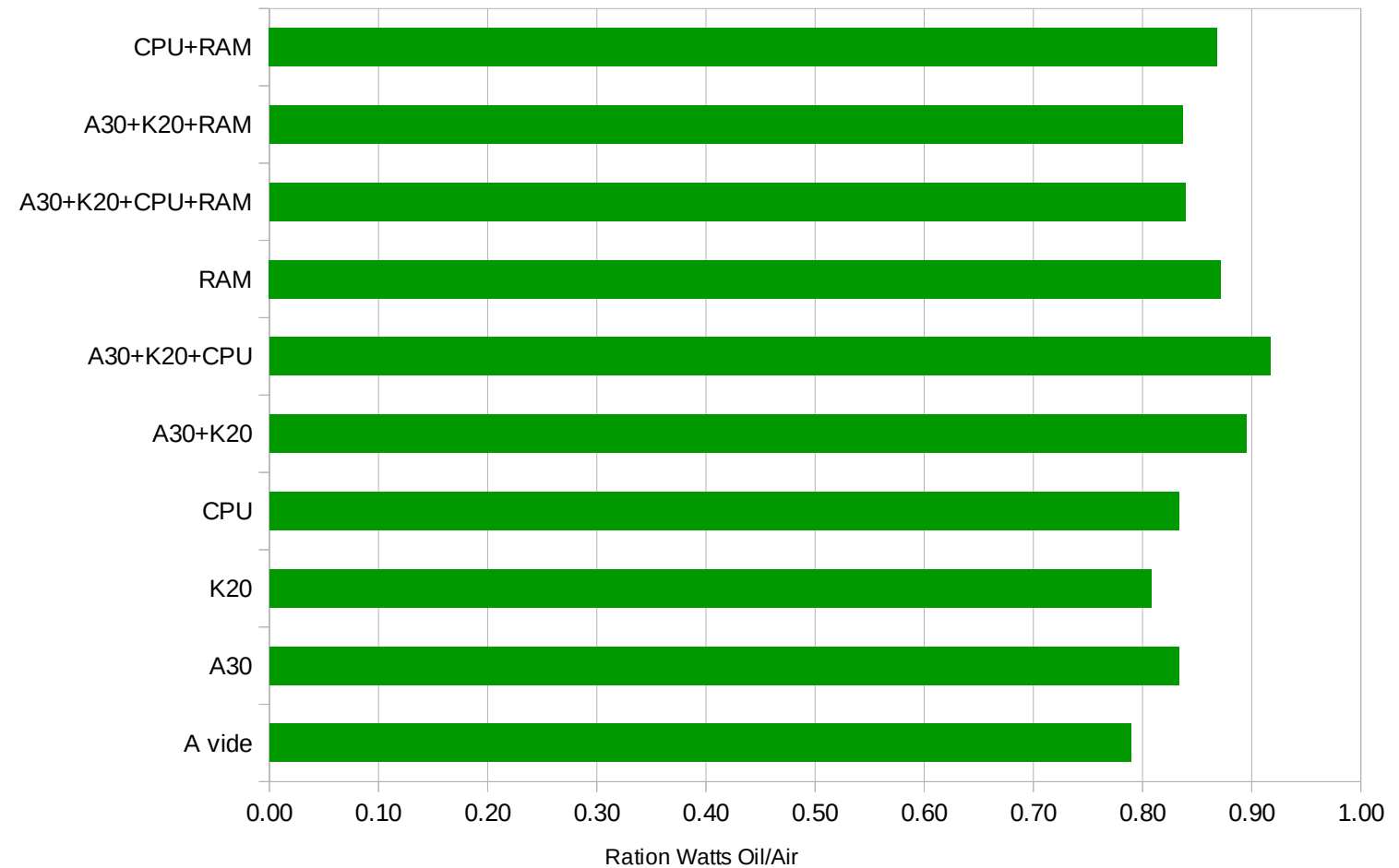
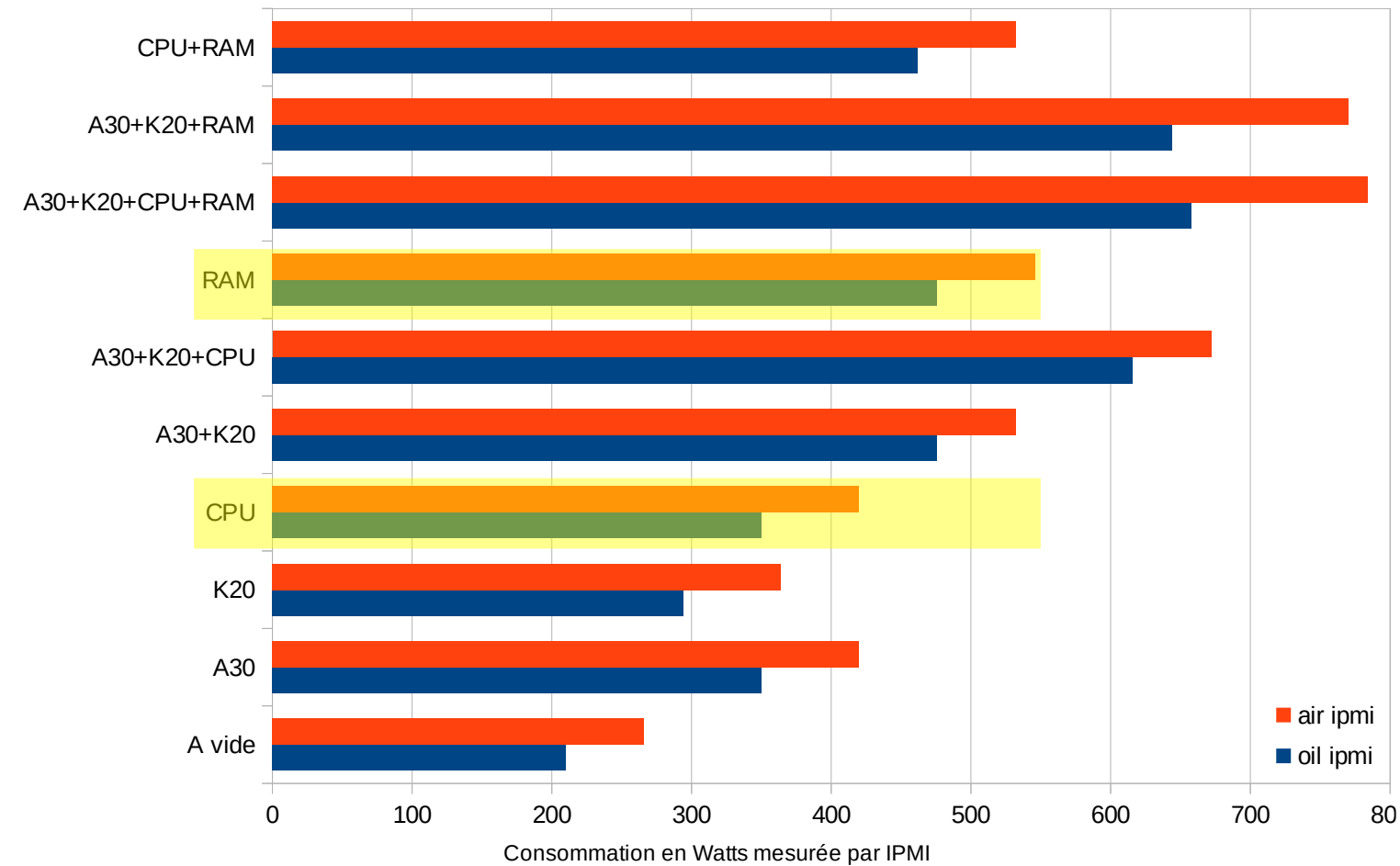
Premiers résultats intéressants (et intrigants)

- Banc d'essai : 2 Dell R720 équipés de...
 - CPU : 2 Intel E5-2670 8 coeurs, 95 W de TDP
 - RAM : 384 GB de RAM (24x16 GB)
 - Carte H310 (pour l'étude du stockage ;-)
 - 2 GPU : un Nvidia K20m ([Equip@Meso](#), 225 W) et un Nvidia A30 (prêt Nvidia, 165 W)
- Expérimentations : Nbody & PiDartDash pour CPU/GPU, mbw pour RAM
 - A vide... (réglage des BIOS en « performance » pour les CPU, boost possible, mode HT)
 - Sans la RAM : A30, K20m, CPU, A30+K20m, A30+K20m+CPU
 - Avec la RAM : mbw, mbw+A30+K20m, mbw+CPU, mbw+A30+K20m+CPU
- Métriques : performances, consommation (relève via IPMI)



Dans le bac Submer, pour les R720, ça donne ?

Ca consomme « moins » mais côté RAM...



- A vide, 21 % de moins.
- En charge, entre 8 % et 19 % de moins. Influence significative de la RAM !

Petit voyage d'un disque dur en immersion

La « minute de vérité »...

- Un HD de 2TB SAS « smart OK » installé
- Le 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 Sense: d7 00 10 08  
[ 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 0x80700 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 0x80700 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 08 00  
[ 8514.876828] blk_update_request: I/O error, dev sda, sector 3907029152 op 0x0:(READ) flags 0x80700 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 08 00  
[ 8534.875685] blk_update_request: I/O error, dev sda, sector 0 op 0x0:(READ) flags 0x80700 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 08 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 08 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 08 00  
[ 8616.898498] 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 08 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 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  
:
```


A ce jour, un bac « presque » plein : 22 nœuds + 2 commutateurs

- Des machines disparates (avec leur équivalent « air ») : ordre de trempage...

Dons ESR

MàD Industrie

MàD PSMN

- 1x Foundry Edgellon (pour l'interconnexion réseau du « sparse part »)
- 1x Dell R720 avec 2 Intel E5-2670 + 384 GB RAM + Nvidia K20m + Nvidia A30 (2022-10)
- 1x Dell PowerConnect 6248 (pour l'interconnexion du « dense part »)
- 4x HPE XL170R avec 2 Intel Gold 5218 + 192 GB RAM
- 4x Dell C6220 avec 2 Intel E5-2670 + 192 GB RAM
- 2x Intel S9200 avec 2 Intel Platinum 9242 + 384 GB RAM + NVME 1TB
- 8x Dell C8220 avec 2 Intel E5-2667v2 + 192 GB RAM
- 1x MadeInCBP avec 1 AMD Epyc 7252 + 64 GB RAM + Nvidia GTX 1080
- 1x MadeInCBP avec 1 Intel Gold 6226R + 128 GB RAM + Nvidia RTX 3080 + NVME 1TB (2023-03)
- 1x Lenovo avec 1 Intel Silver 4214 + 64 GB RAM + SSD 2TB



- Ca fait 41 sockets, 756 coeurs, 3712 GB RAM, 2 GPGPU Nvidia, 2 GPU Nvidia, 4 TB de stockage

Une année (enfin les 11 derniers mois) très riche en opérations diverses...

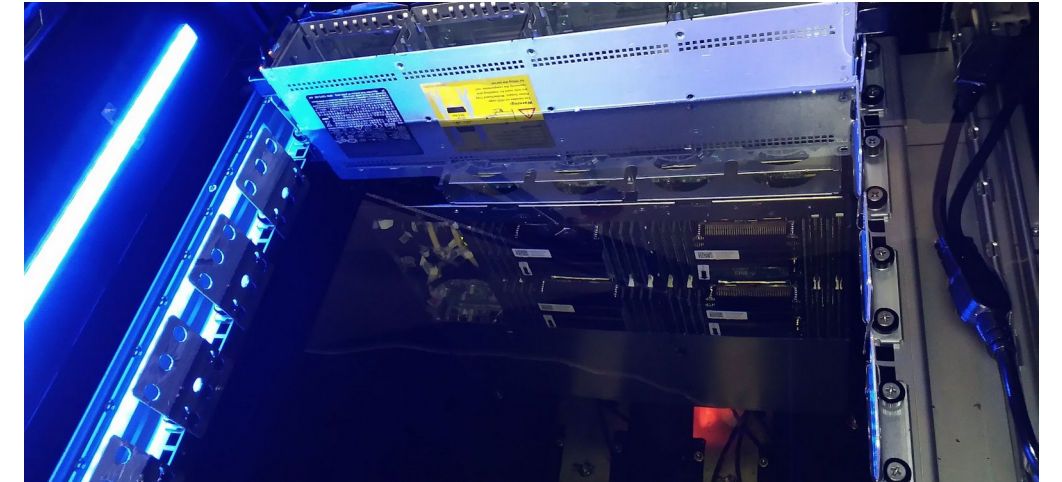
- Déclassement du bac « Voldemort » avec son huile : mai 2022
 - Fuite sur échangeur eau-huile, et ses effets délétères sur le matériel...
- Mise-en-œuvre bac Submer SmartpodXL : de juin à décembre 2022
 - Déballage, préparation plancher, adduction eau : de juin à fin septembre 2022
 - Installation, remplissage, « mise-à-niveau », adaptations 19" : d'octobre à décembre 2022
 - Exploitation avec portique, mise en œuvre EPI, sécurité : de décembre 2022 à aujourd'hui
- Expérimentations : suivant les 3 volets, mais cadre « science & références »
 - Préparation avant immersion des machines : plus compliqué...
 - Adaptation du bac pour immersion : pas assez profond...

Préparation des machines : la « dépose » ne suffit plus...

- 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
 - Modifier les cartes de supervision : « faire accepter » des ventilateurs absents
 - Fixer les nœuds verticalement : renforcer des crochets de retenue « inadaptés »
- Mais un problème récurrent : la profondeur limitée du SmartpodXL

Adaptation du bac SmartpodXL augmenter la profondeur admissible...

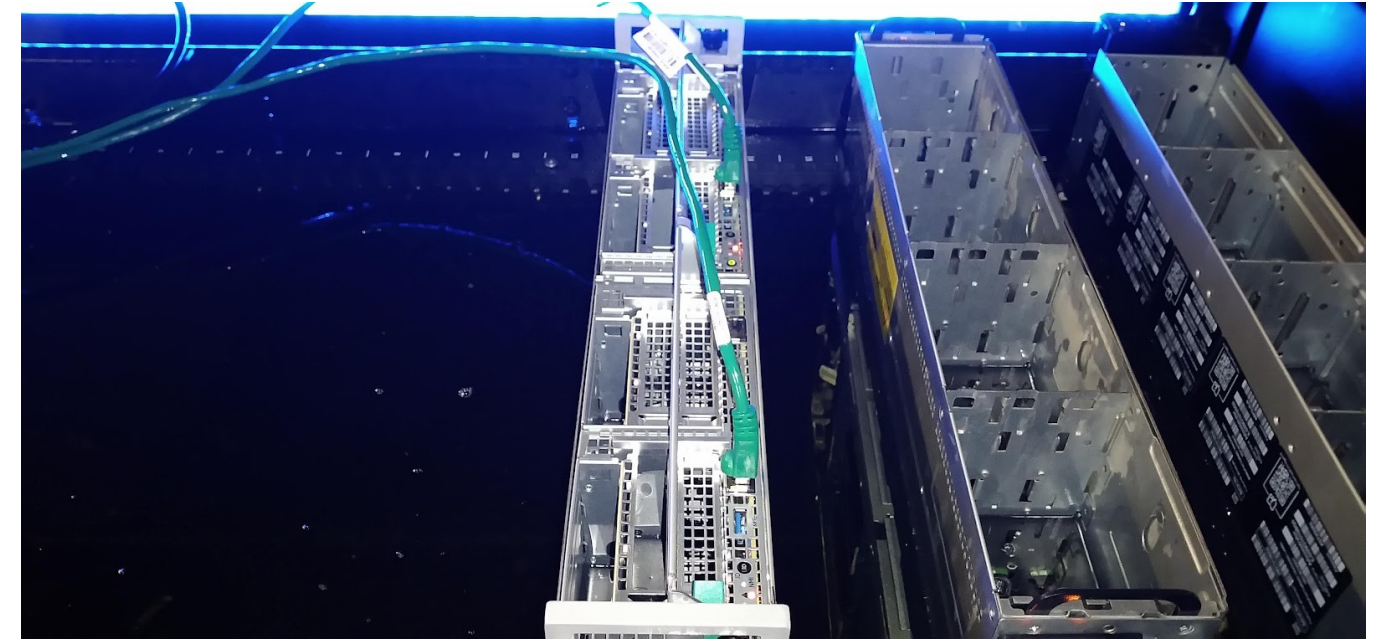
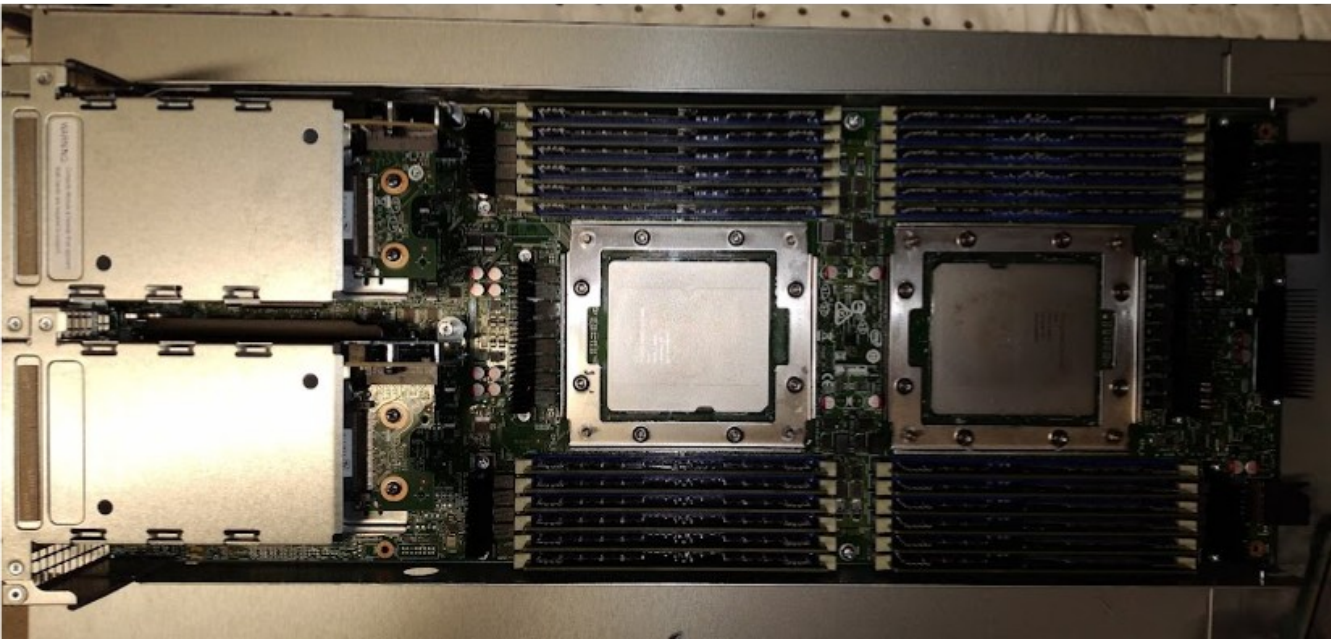
- De base, pas plus de 70 cm : seulement les Dell PowerEdge
- Exploitation de cales de 7 cm :
 - HPE XL170R : insertion OK, refroidissement OK
 - Dell C6220, Intel S9200, Dell C8000 : KO !
- Exploitation de cales de 12 cm :
 - Intel S9200 : insertion OK, refroidissement OK
 - Dell C6220 : insertion OK, refroidissement OK
 - Dell C8000 : insertion OK, refroidissement KO
- Développement nécessaire de « stratégie d'adaptation »...



Au chapitre des « bons clients » du futur...

Les serveurs Intel S9200 à Platinum 9242

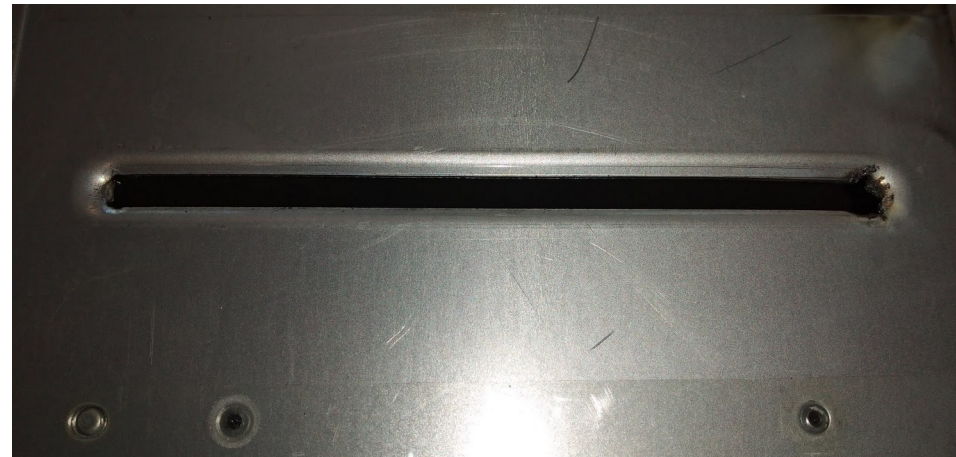
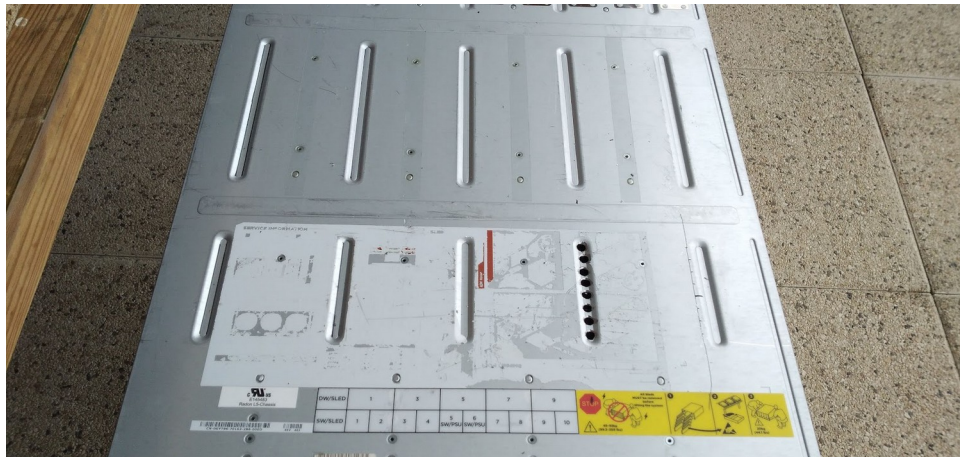
- Dans 2U, 4 sockets Platinum 9242 à la TDP de 350W (et 48 barrettes de RAM)



- Un cas « typique » des « futures » machines HPC :
 - Des processeurs très dissipatifs (+ de 300W par socket)
 - Des barrettes mémoire nombreuses et voraces (DDR5 : 10W par barrette)

« Customisation » des châssis : au chapitre des mauvaises « bonnes idées »

- Où on en est ? L'huile chauffe mais reste « confinée » dans le châssis..
- Où va-t-on ? Solution 1 : évacuation de l'huile chauffée.
- Comment ? Des petits trous ou des fentes...
 - Trous à la perceuse (mais des copeaux même avec les protections)...
 - Fentes à la découpeuse à plasma (mais avec des résidus de combustion)...

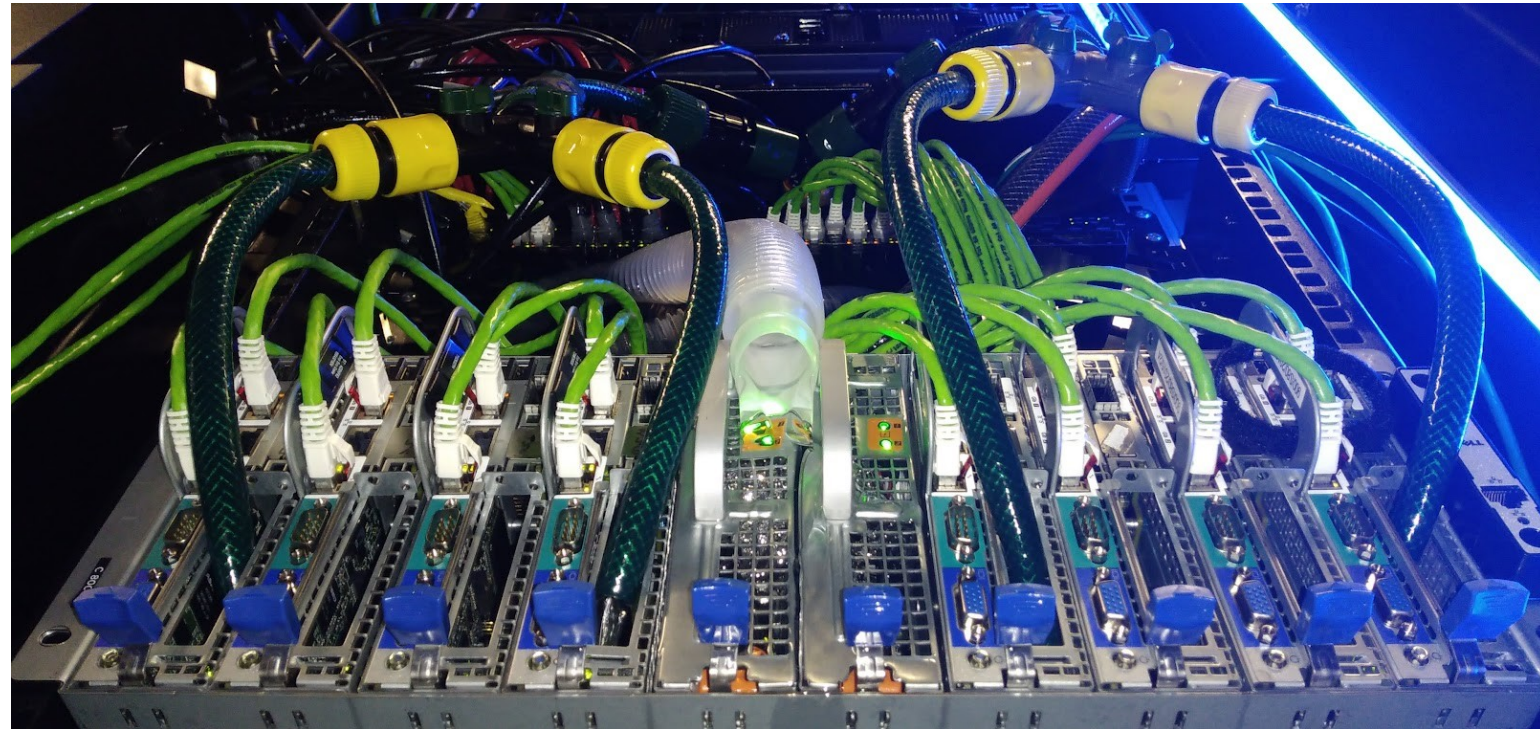


- En conclusion... Solution 1 : KO. Solution 2 = WIP

Customisation des chassis

Faire arriver « l'huile fraîche »

- Puisque l'huile ne peut pas s'évacuer par les côtés...
 - On en ajoute un maximum par le dessus !
- Mise en œuvre :
 - Une pompe immergée (à eau)
 - De la plomberie de jardinage
 - Une distribution sur 4 zones
- Résultats : ça marche...
 - Satisfaisant (mais pas optimal !)
 - Attente de boîtiers plus génériques mieux conçus pour l'immersion...



Comparaison des machines air & oil

Banc d'essai & comportement à la charge

- Hardware : Apollo XL170R avec Gold 5218, Intel S9200 avec Platinum 9242

- OS : SIDUS Debian Bullseye

- Tests :

- Au repos : stand by
- Pure calcul flottant : nbody
- Calcul Entier & flottant : pixpu
- Charge mémoire multithreadée : mbw

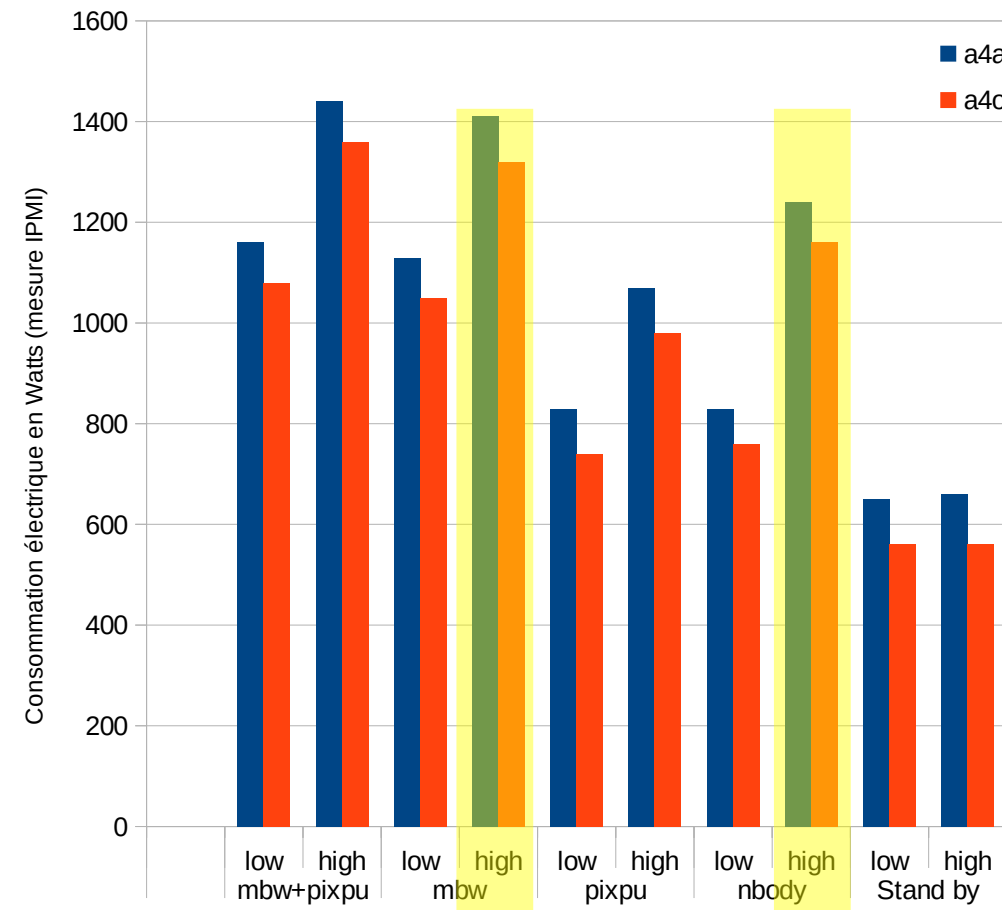
- Fréquences :

- Low, High, (Max)

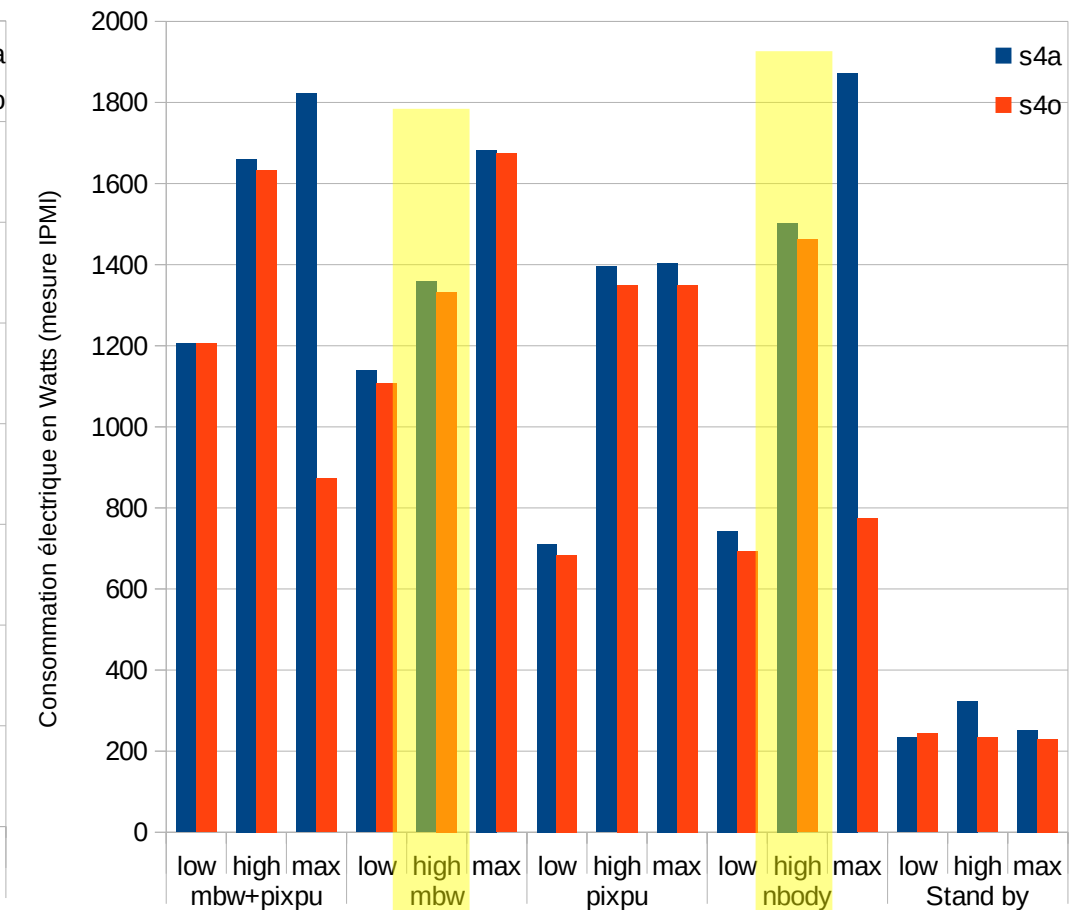
- Relève de consommation :

- Sensors, IPMI, Wattmètre

- L'immersion permet « toujours » de moins consommer...



Cas d'usage pour 4 Apollo 2000 dans l'air (a4a) ou dans l'huile (a4o)

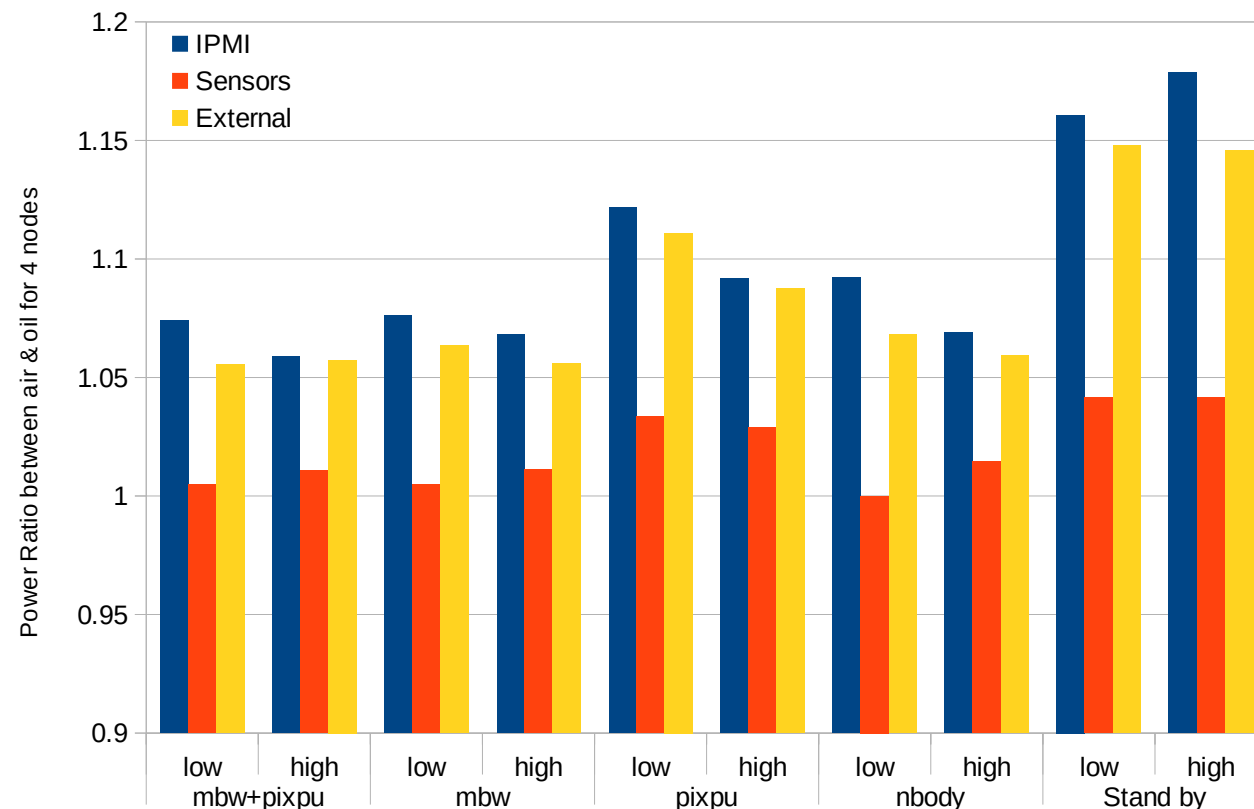


Cas d'usage pour 2 Intel S9200 dans l'air (s4a) ou dans l'huile (s4o)

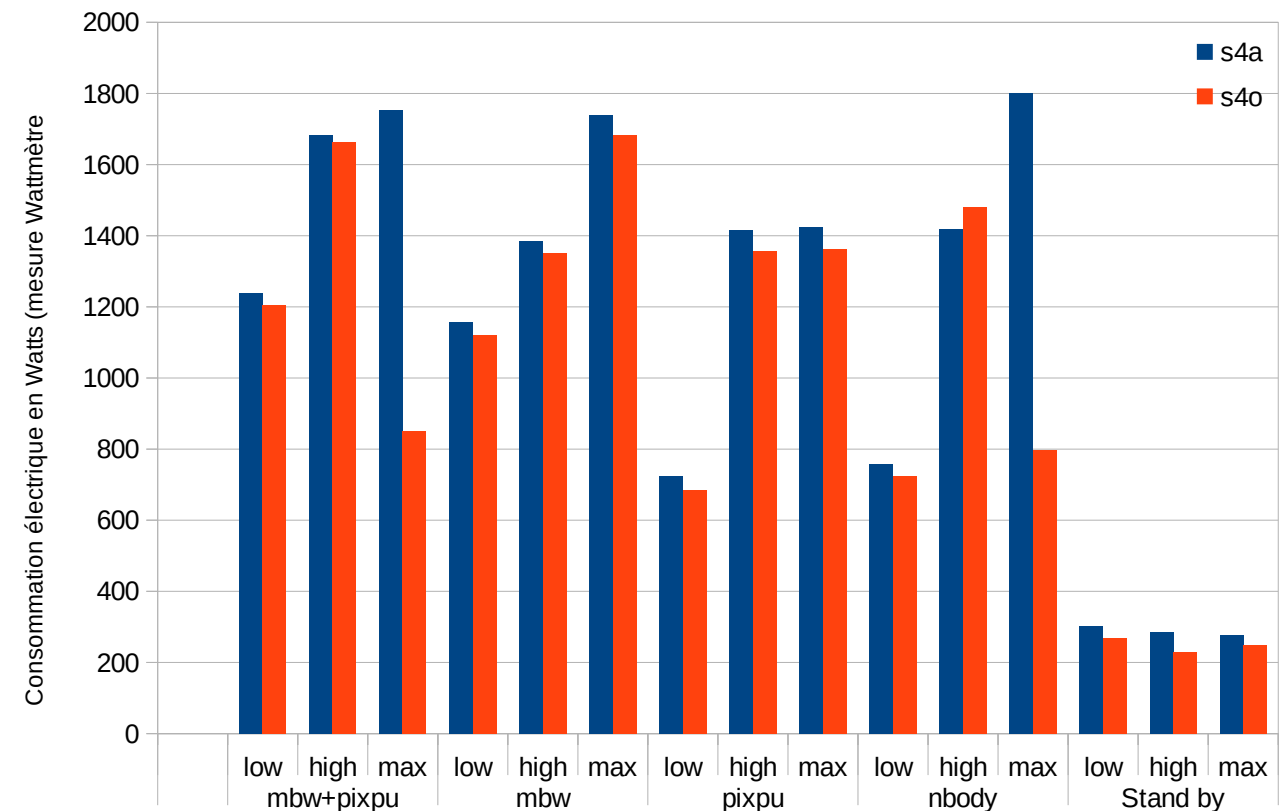
Comparaison air & oil

Consommer moins, mais combien ?

- Des ratio de consommation sont sensiblement différents en fonction de leur mesure, de l'ordre de 10 %...
- Dans les mesures de consommations, les mesures IPMI ou externes sont préférables



5 different use cases for the 2 extremal frequencies



Cas d'usage pour 2 Intel S9200 dans l'air (s4a) ou dans l'huile (s4o)

Retour sur le programme d'octobre 2022

- Evaluer plus précisément les consommations électriques :
 - Les pinces ampèremétriques disponibles apportent leur lot d'interrogations également...
 - Les wattmètres « simples » à déployer difficiles à interfacer (en masse)
- Evaluer des configurations avec des processeurs à beaucoup de coeurs
 - Seulement des systèmes à seulement 2x16 coeurs
- Evaluer plus de GPU
 - Des GPGPU : une montée en fréquence est-elle possible ?
 - Des GPU : une meilleure intégration (plus compacte) des séries 3000 et supérieures, ou AMD
- Appel à collaboration mais dans une approche scientifique (2 machines)

Et la suite ?

- Évaluation des performances d'un autre caloporteur diélectrique
 - Inertie thermique associée au refroidissement liquide sur CPU, GPU, RAM.
- Expérimentations sur les chassis à haute densité :
 - Distribution de l'huile par ailleurs efficace, mais quelle quantité nécessaire ?
- Écriture d'un cahier des charges pour le déplacement d'installations
 - Local inadapté pour ce genre de déploiement à échelle « industrielle ».
- Consolidation du volet « exploitation opérationnelle »
 - Écriture d'un livre blanc de « bonnes pratiques » sur l'immersion à l'étude...