Fiabilisation de l'infrastructure du datacentre et du calculateur du PMCS2I

Amélioration de l'efficacité énergétique

M.Camier[1,4], Q.Ringes[2], L.Pouilloux[1,4], G.Capiod[2], D.Magane[3], N.Grosjean[4], A.Cadiou[1,4]

1. Pôle de calcul PMCS2I
2. DSI École Centrale
3. DirPAT École Centrale
4. Laboratoire de Mécanique des Fluides et d'Acoustique

JCAD 2022 - mercredi 12 octobre 2022

Plan

Le datacentre de l'École Centrale de Lyon

Amélioration de l'efficacité énergétique et des capacités d'hébergement

Fiabilisation du datacentre et des moyens de calcul

Le datacentre de l'École Centrale de Lyon

Pourquoi un datacentre à l'École Centrale de Lyon ?

École d'Ingénieurs généralistes et de spécialités (ENISE)

- 4000 étudiants et 1100 personnels (EC, C, ITA, . . .)
- 6 départements d'enseignements
- 6 UMR CNRS (LTDS, **LMFA**, Ampère, INL, ICJ, LIRIS)
- des plateformes et bancs d'essais pour la recherche fondamentale et partenariale

Pourquoi un datacentre à l'École Centrale de Lyon ?

École d'Ingénieurs généralistes et de spécialités (ENISE)

- 4000 étudiants et 1100 personnels (EC, C, ITA, . . .)
- 6 départements d'enseignements
- 6 UMR CNRS (LTDS, **LMFA**, Ampère, INL, ICJ, LIRIS)
- des plateformes et bancs d'essais pour la recherche fondamentale et partenariale

En 2013, besoins exprimés en informatique : web, mail, enseignement, données, calcul

- regrouper les salles informatiques ($\sim$ 10) des départements et labos
- héberger les serveurs des UMR
- transférer le calculateur du pôle PMCS2I (192 cœurs, 2010)

Pourquoi un datacentre à l'École Centrale de Lyon ?

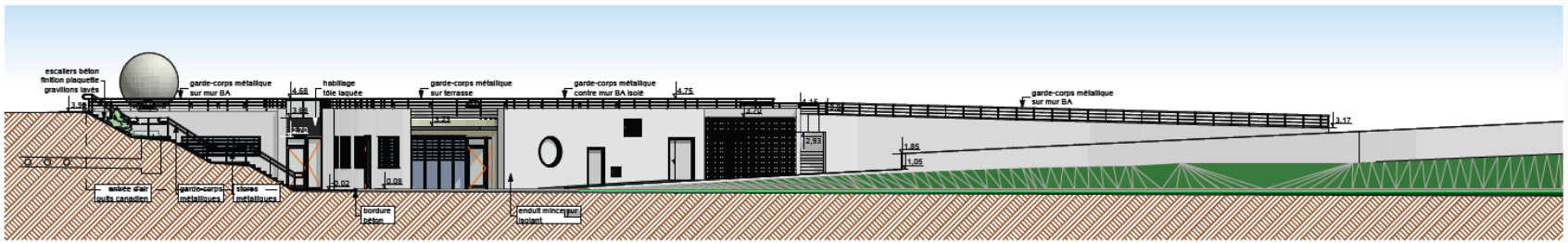
École d'Ingénieurs généralistes et de spécialités (ENISE)

- 4000 étudiants et 1100 personnels (EC, C, ITA, . . .)
- 6 départements d'enseignements
- 6 UMR CNRS (LTDS, **LMFA**, Ampère, INL, ICJ, LIRIS)
- des plateformes et bancs d'essais pour la recherche fondamentale et partenariale

En 2013, besoins exprimés en informatique : web, mail, enseignement, données, calcul

- regrouper les salles informatiques ($\sim$ 10) des départements et labos
- héberger les serveurs des UMR
- transférer le calculateur du pôle PMCS2I (192 cœurs, 2010)

$\Rightarrow$ Construction d'un bâtiment HQE pour la DSI, comprenant un DC mutualisé



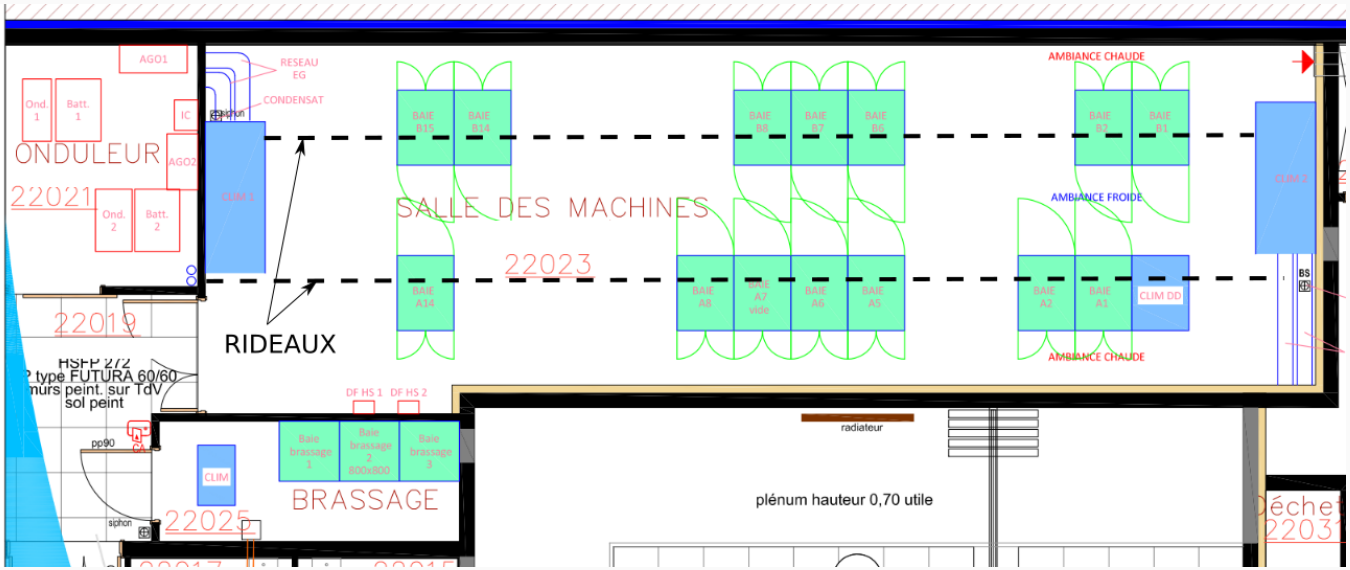
Le datacentre à la livraison en 2015

Les points positifs

- grande surface (68 m^2)
- 2 onduleurs + groupe électrogène
- couplage climatisation/chauffage

Les points négatifs

- pas de cloisonnement thermique
- pas de redondance froid
- pas de prise 32A pour le HPC



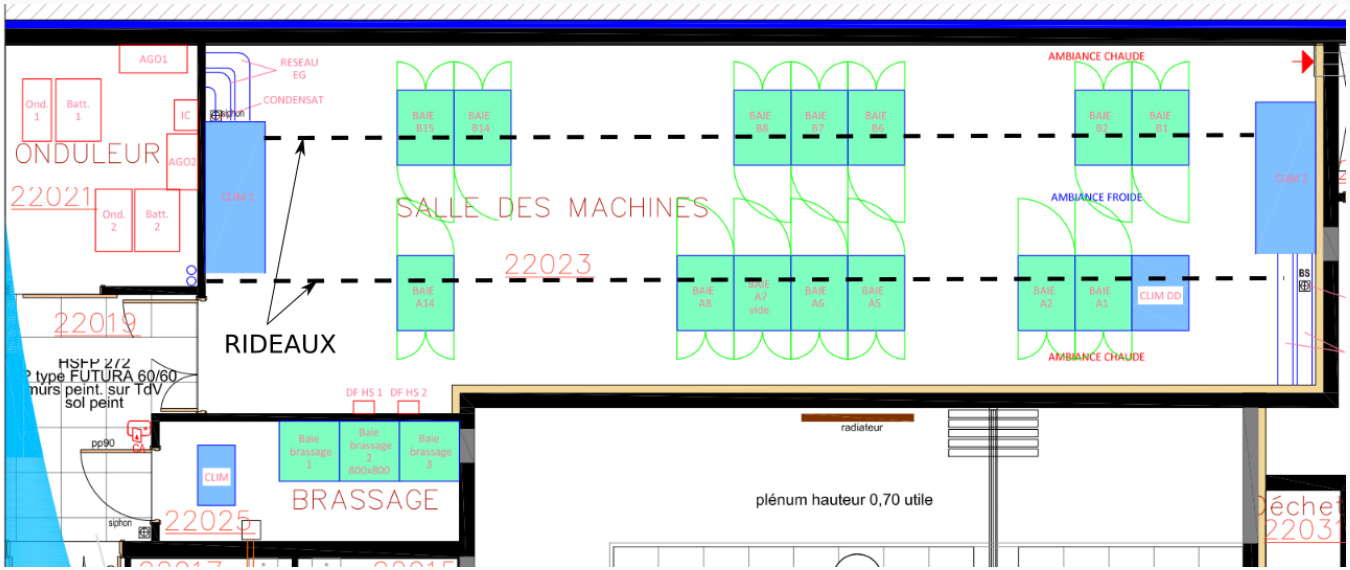
Le datacentre à la livraison en 2015

Les points positifs

- grande surface (68 m^2)
- 2 onduleurs + groupe électrogène
- couplage climatisation/chauffage

Les points négatifs

- pas de cloisonnement thermique
- pas de redondance froid
- pas de prise 32A pour le HPC



Audit externe pour amélioration dès la livraison

- mise en place de rideaux à lanières PVC
- ajout d'une climatisation DD 25 kW pour panne/maintenance
- ajout de prises 32 A

Politique de fonctionnement et remplissage de la salle machine

Politique du datacentre

- pas de facturation de l'hébergement pour les UMR de l'établissement
- obligation de PDUs monitorables
- procédure d'extinction déclenchable par la DSI de l'École
- mutualisation des ressources de calcul au niveau du Pôle

Politique de fonctionnement et remplissage de la salle machine

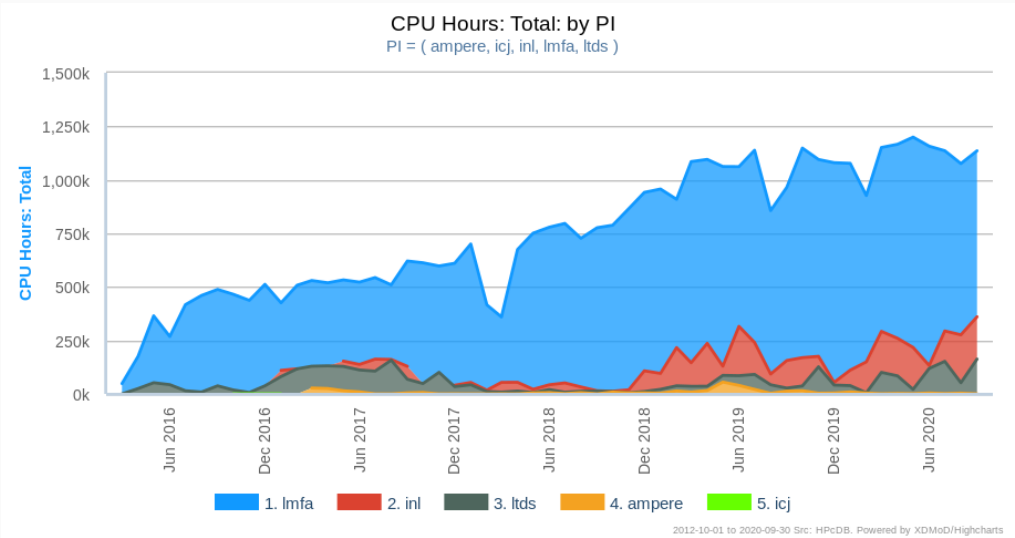
Politique du datacentre

- pas de facturation de l'hébergement pour les UMR de l'établissement
- obligation de PDUs monitorables
- procédure d'extinction déclenchable par la DSI de l'École
- mutualisation des ressources de calcul au niveau du Pôle

Migration successive des serveurs des entités du campus

- 2015 : DSI, Ampère, INL, LTDS
- 2016 : PMCS2I (768 cœurs)
- 2020 : LMFA

2016-2020 $\Rightarrow$ ajout au fil de l'eau de serveurs de calcul ($\sim$ 1000 cœurs), de stockage et d'hyperviseurs.



Trois éléments majeurs en 2020

- Intégration du datacentre dans le projet lyonnais CINELYS
- Réception de financement du CPER CIDRA pour le calcul (500 k€)
- COVID : augmentation des besoins pour l'enseignement à distance

Trois éléments majeurs en 2020

- Intégration du datacentre dans le projet lyonnais CINELYS
- Réception de financement du CPER CIDRA pour le calcul (500 k€)
- COVID : augmentation des besoins pour l'enseignement à distance

Discussion avec le GDS EcoInfo, nouvel audit externe capacitaire et groupe de travail au niveau de l'établissement (DR, DSI, DPAT, PMCS2I) pour

- accroître les capacités d'hébergements (refroidissement, puissance électrique)
- améliorer l'efficacité énergétique (PUE < 1.3 comme cible)
- fiabiliser les systèmes environnementaux et les matériels hébergés

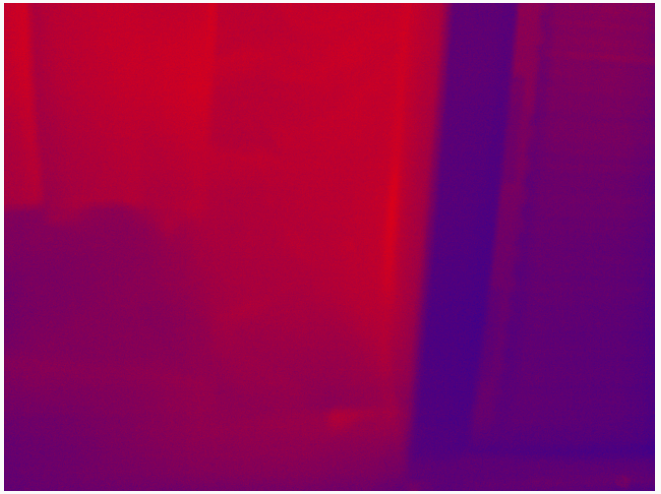
Amélioration de l'efficacité énergétique et des capacités d'hébergement

État initial en 2016

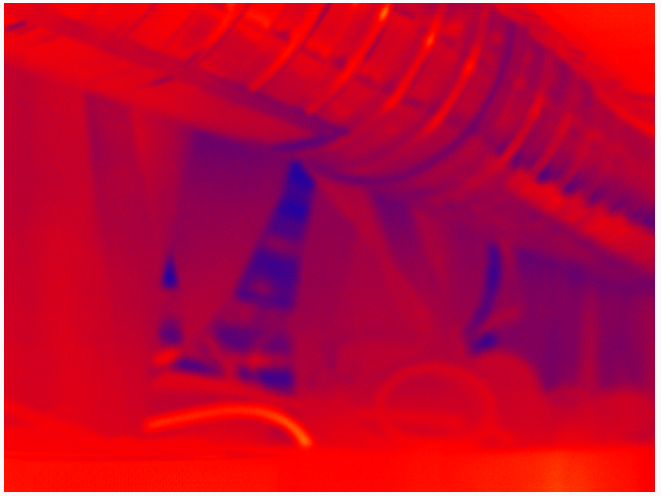
Salle serveur en 2016

État initial en 2016

Salle serveur en 2016



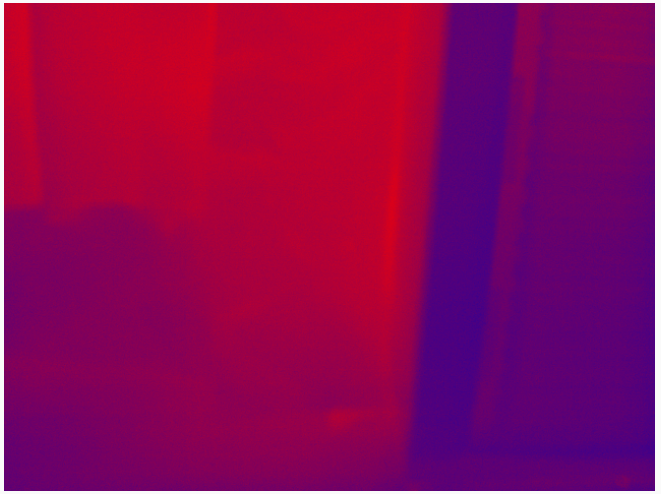
Allée froide : rideaux à gauche vs baie à droite



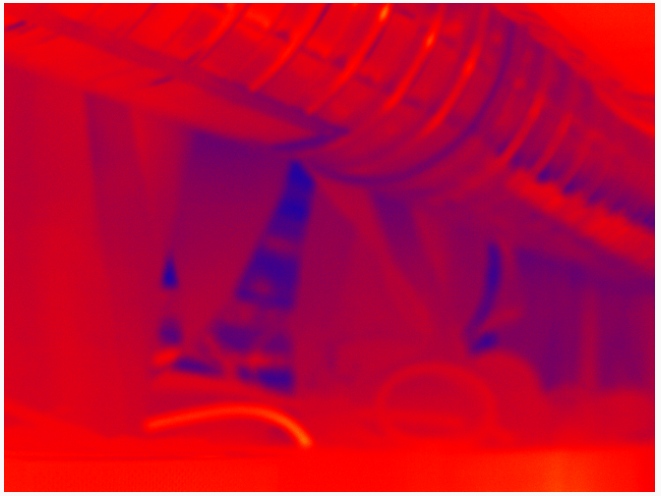
Allée chaude : perte au-dessus des baies

État initial en 2016

Salle serveur en 2016



Allée froide : rideaux à gauche vs baie à droite



Allée chaude : perte au-dessus des baies

⇒ Besoin de travailler sur l'urbanisation et le confinement !

Modification de l'urbanisation et le confinement

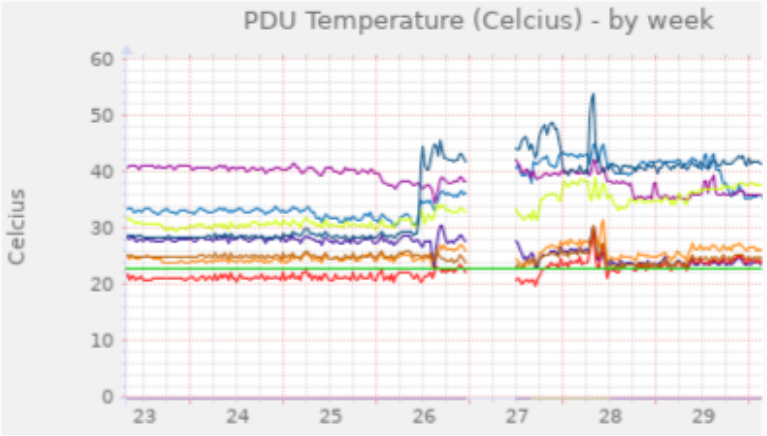
Salle serveur en 2021

Coût de la solution

- baies : 10 k€
- confinement : 20 k€
- onduleur : 18 k€

Modification de l'urbanisation et le confinement

Salle serveur en 2021



Suivi de la température dans la salle

Coût de la solution

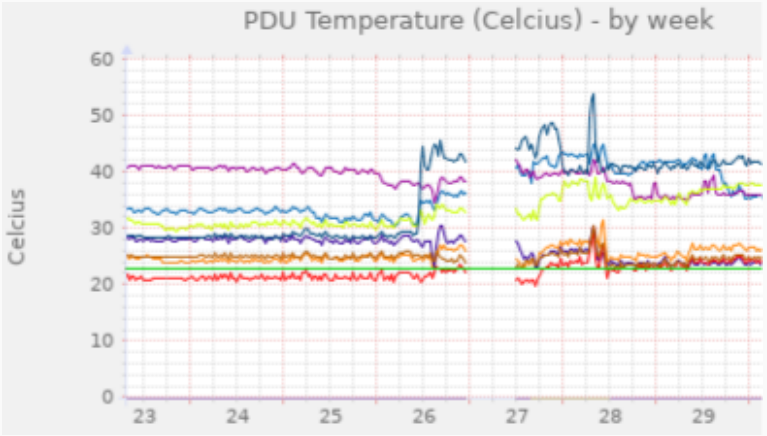
- baies : 10 k€
- confinement : 20 k€
- onduleur : 18 k€

Effet dans la salle

- réduction de la zone à refroidir
- homogénéisation de la température dans l'allée froide

Modification de l'urbanisation et le confinement

Salle serveur en 2021



Suivi de la température dans la salle

Coût de la solution

- baies : 10 k€
- confinement : 20 k€
- onduleur : 18 k€

Effet dans la salle

- réduction de la zone à refroidir
- homogénéisation de la température dans l'allée froide

⇒ Mise en route des moyens de calcul supplémentaires (+1700 cœurs)

Expérimentations dans la salle serveur

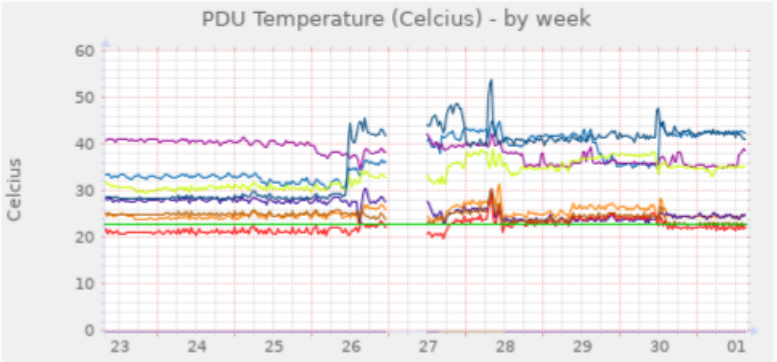
Estimation de l'écoulement (Emmanuel Jondeau)

Tentative de contrôle du flux d'air

Expérimentations dans la salle serveur

Estimation de l'écoulement (Emmanuel Jondeau)

Tentative de contrôle du flux d'air

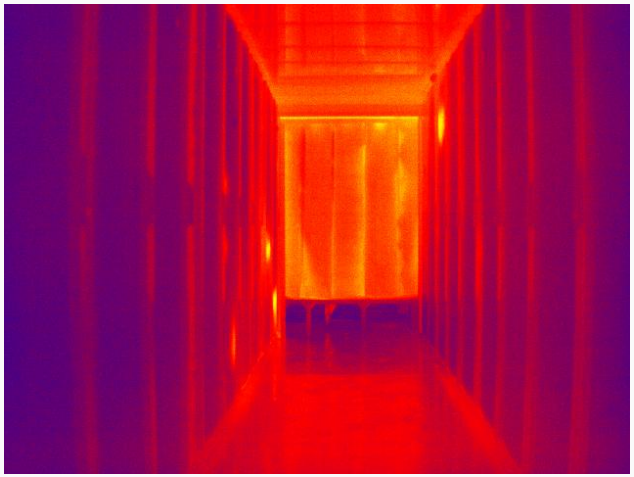


Effet du carton et des rideaux

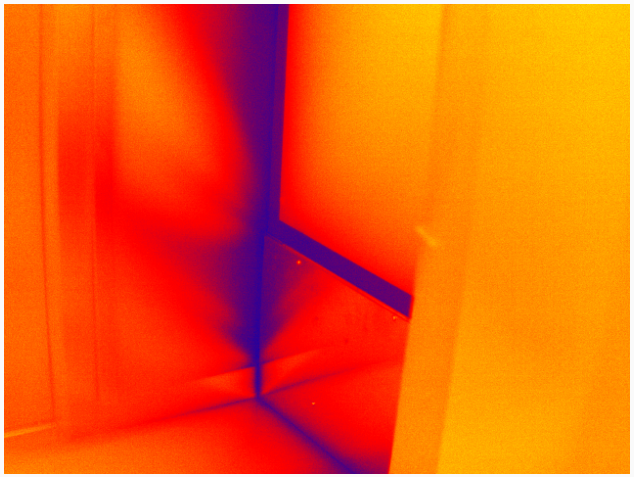
Expérimentations dans la salle serveur

Estimation de l'écoulement (Emmanuel Jondeau)

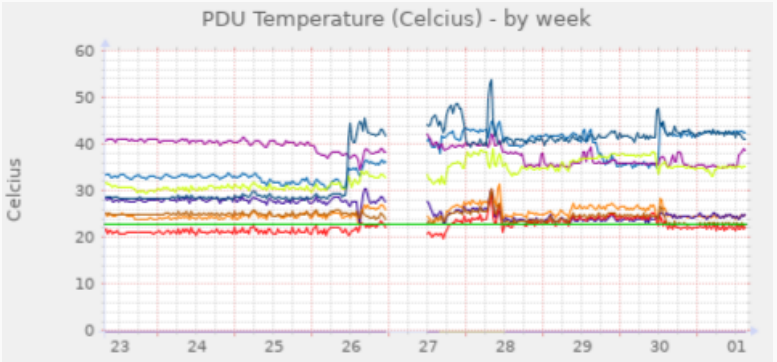
Tentative de contrôle du flux d'air



Allée froide



Allée chaude : fuite autour des climatisations

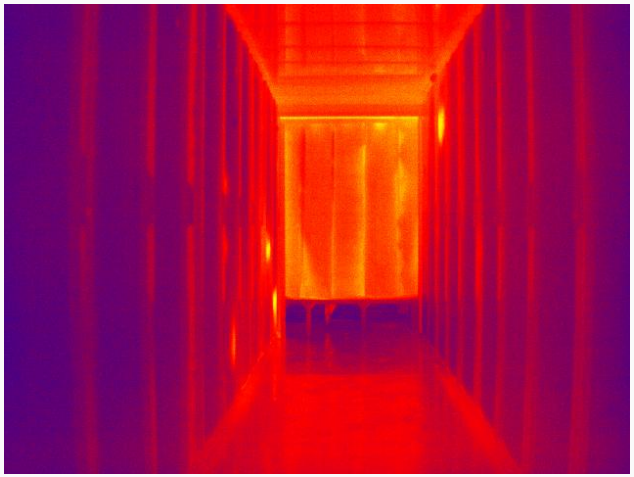


Effet du carton et des rideaux

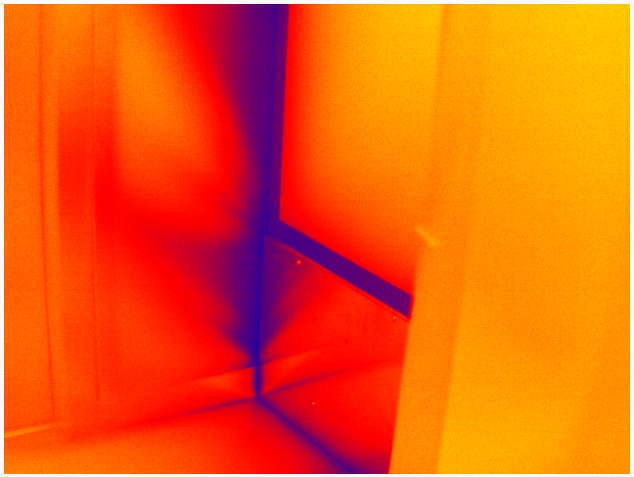
Expérimentations dans la salle serveur

Estimation de l'écoulement (Emmanuel Jondeau)

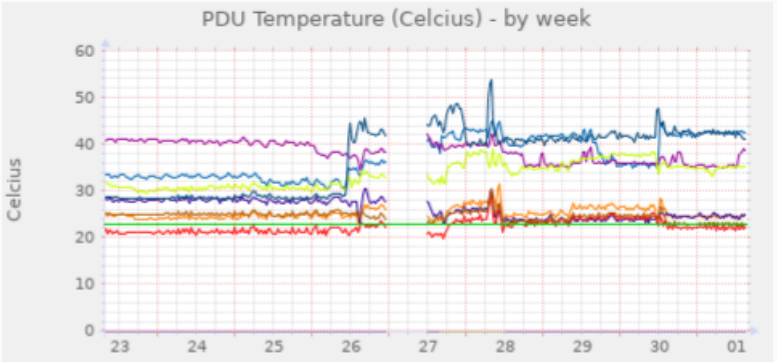
Tentative de contrôle du flux d'air



Allée froide



Allée chaude : fuite autour des climatisations



Effet du carton et des rideaux

⇒ Construction d'une solution professionnelle . . .

Fin de la réalisation du confinement

Tunnel froid

Passe-câble

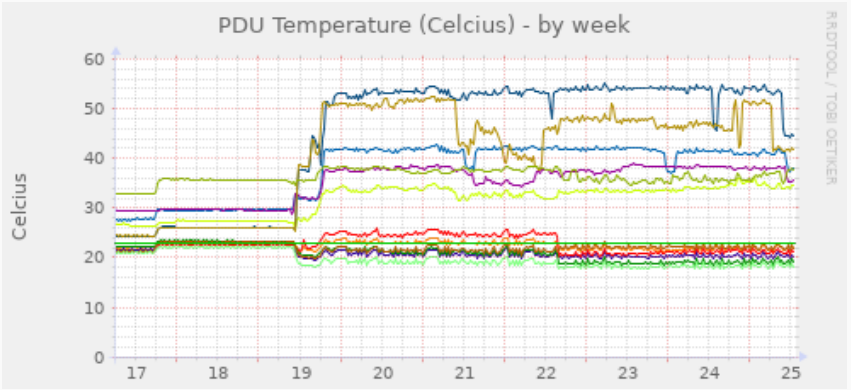
Déflecteur

Fin de la réalisation du confinement

Tunnel froid

Passe-câble

Déflecteur



Effet du tunnel

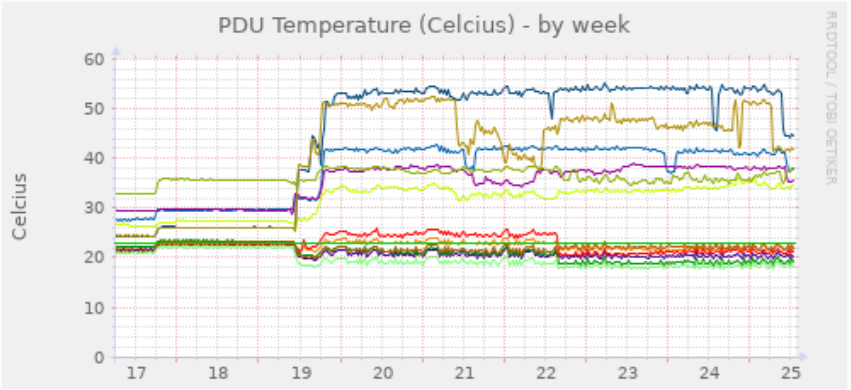
Effet porte fermée

Fin de la réalisation du confinement

Tunnel froid

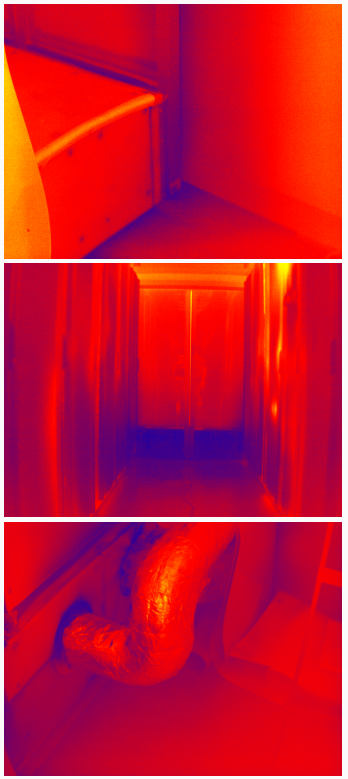
Passe-câble

Déflecteur



Effet du tunnel

Effet porte fermée



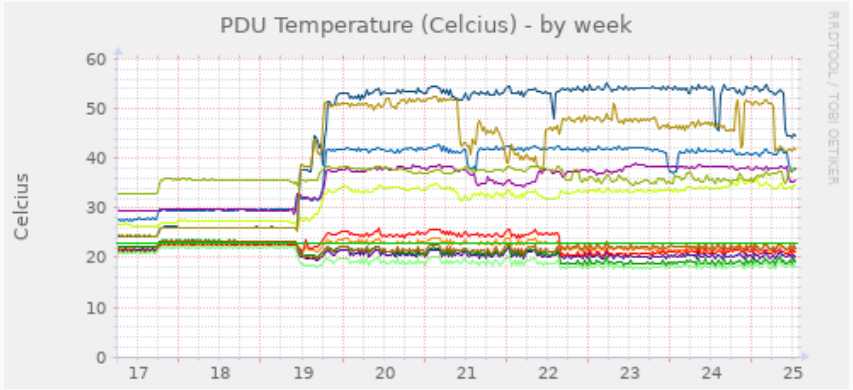
Validation en caméra thermique

Fin de la réalisation du confinement

Tunnel froid

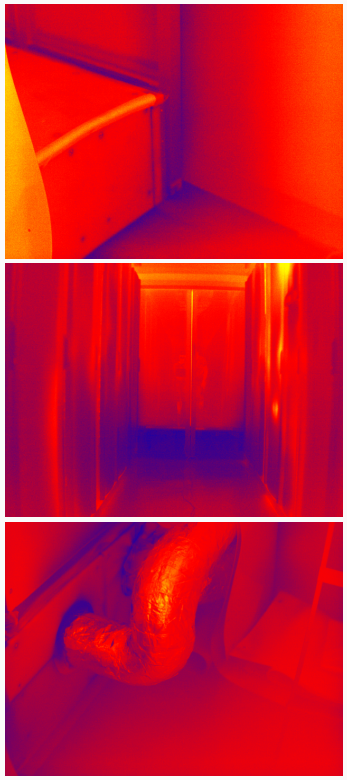
Passe-câble

Déflecteur



Effet du tunnel

Effet porte fermée



Validation en caméra thermique

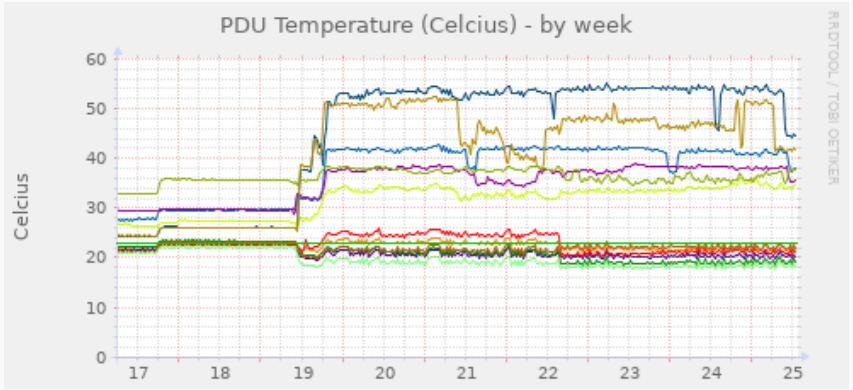
- tunnel : 1000 € - Alexandre Azouzi (LMFA)
- découpe des portes/déflecteur/passe-câble : 100 € - Farid Hamoudi (DirPAT)

Fin de la réalisation du confinement

Tunnel froid

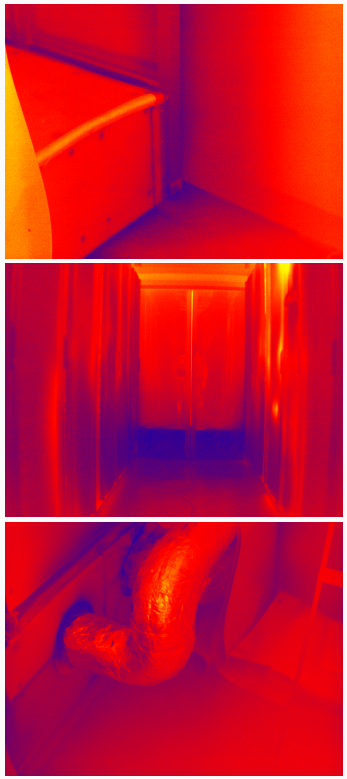
Passe-câble

Déflecteur



Effet du tunnel

Effet porte fermée



Validation en caméra thermique

- tunnel : 1000 € - Alexandre Azouzi (LMFA)
- découpe des portes/déflecteur/passe-câble : 100 € - Farid Hamoudi (DirPAT)

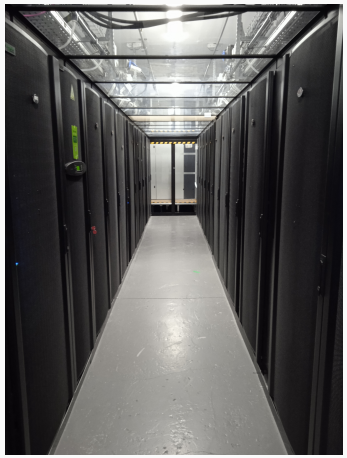
$\Rightarrow$ changement des consignes du groupe froid de 8 à 12°C ! !

Bilan et perspectives

Bilan

Coût urbanisation + onduleur : 43 k€ investissement + personnels

	2016	2020	2022
Cœurs PMCS2I	768	1704	3480
Stockages	300 To	1.5 Po	4 Po
P IT (kW)	14	32	90
P Cooling (kW)	13.5	20	40
PUE	~ 2	~ 1.6	$\sim$1.45



Économie réalisée : 14 kW sur la climatisation $\sim$ 10 MWh par mois

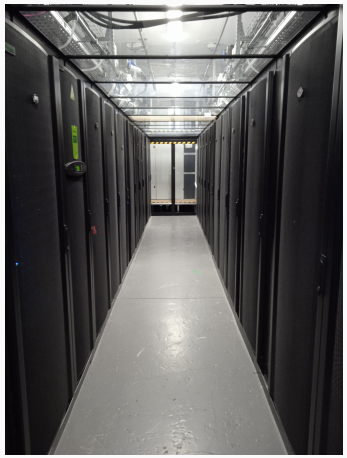
$\Rightarrow$ 10 k€/an au tarif janv 2022 ; 100 k€/an au tarif janv 2023 ?

Bilan et perspectives

Bilan

Coût urbanisation + onduleur : 43 k€ investissement + personnels

	2016	2020	2022
Cœurs PMCS2I	768	1704	3480
Stockages	300 To	1.5 Po	4 Po
P IT (kW)	14	32	90
P Cooling (kW)	13.5	20	40
PUE	~ 2	~ 1.6	$\sim$1.45



Économie réalisée : 14 kW sur la climatisation $\sim$ 10 MWh par mois

$\Rightarrow$ 10 k€/an au tarif janv 2022 ; 100 k€/an au tarif janv 2023 ?

Perspectives

- supervision automatisée de la consommation hors IT
- installation d'un second groupe froid + réseau inrow
- mise à niveau des onduleurs
- définition d'un modèle économique

Fiabilisation du datacentre et des moyens de calcul

Procédures d'arrêts automatisées

Problématique d'hébergement

- électrique
- groupe froid et climatisation

Procédures d'arrêts automatisées

Problématique d'hébergement

- **électrique**
- groupe froid et climatisation

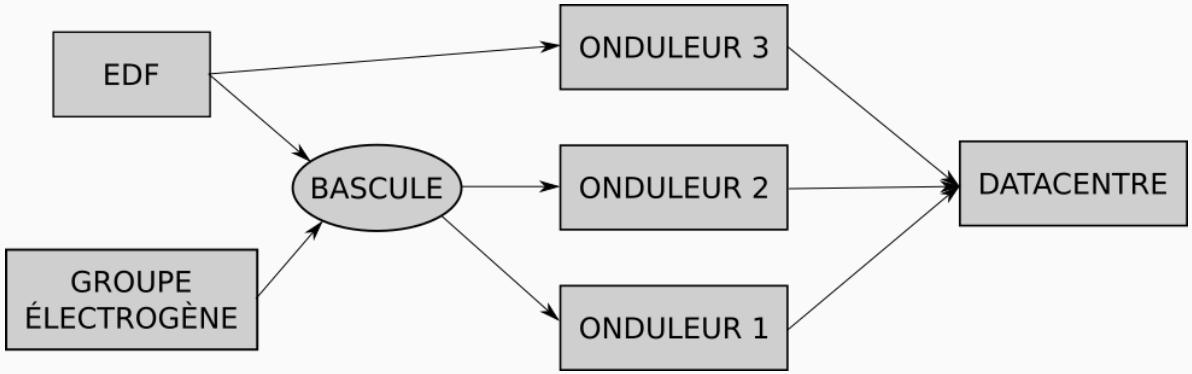


Schéma électrique

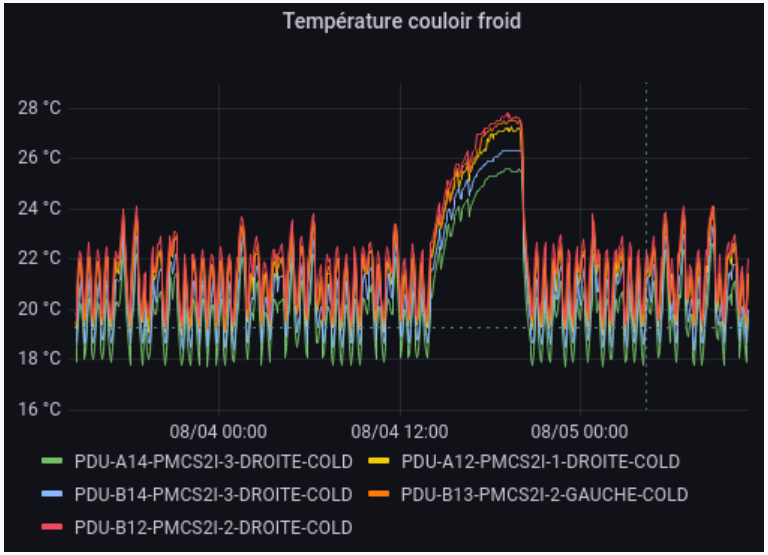
Utilisation de `apcupsd`

- interrogation via SNMP sur le nouvel onduleur
- topologie en arbre au niveau du cluster

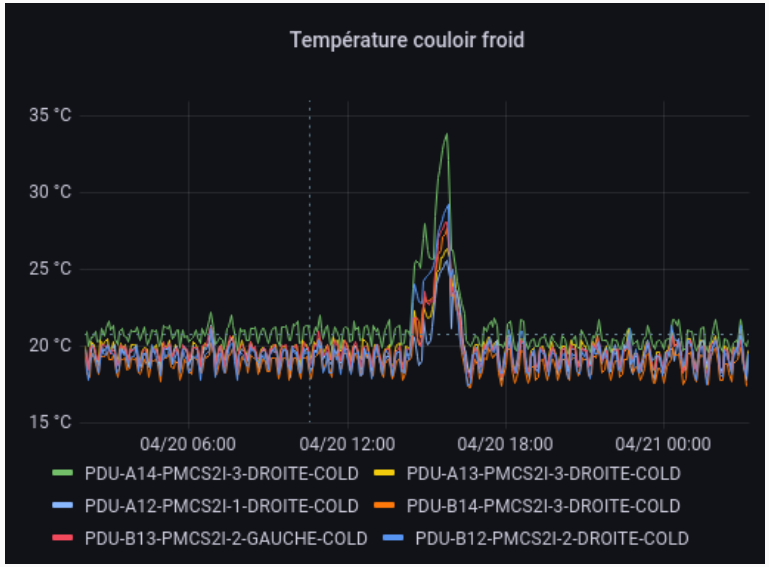
Procédures d'arrêts automatisées

Problématique d'hébergement

- électrique
- **groupe froid et climatisation**



Panne groupe froid



Panne climatisation

- script d'extinction via cron
- différenciation de la panne
- SNMP sur PDUs

Solution de supervision mutualisée

Besoin

- interface centralisé entre les acteurs (DSI, DirPAT, PMCS2I)
- historique, sans perte
- alertes via mail/chat

Solution de supervision mutualisée

Besoin

- interface centralisé entre les acteurs (DSI, DirPAT, PMCS2I)
- historique, sans perte
- alertes via mail/chat

Existant

- 2 munin (1 DSI, 1 PMCS2I)
- historique bloqué à 1 an, avec pertes d'informations (stockage RRD)
- pas d'alertes configurées

Solution de supervision mutualisée

Besoin

- interface centralisé entre les acteurs (DSI, DirPAT, PMCS2I)
- historique, sans perte
- alertes via mail/chat

Existant

- 2 munin (1 DSI, 1 PMCS2I)
- historique bloqué à 1 an, avec pertes d'informations (stockage RRD)
- pas d'alertes configurées

Solution

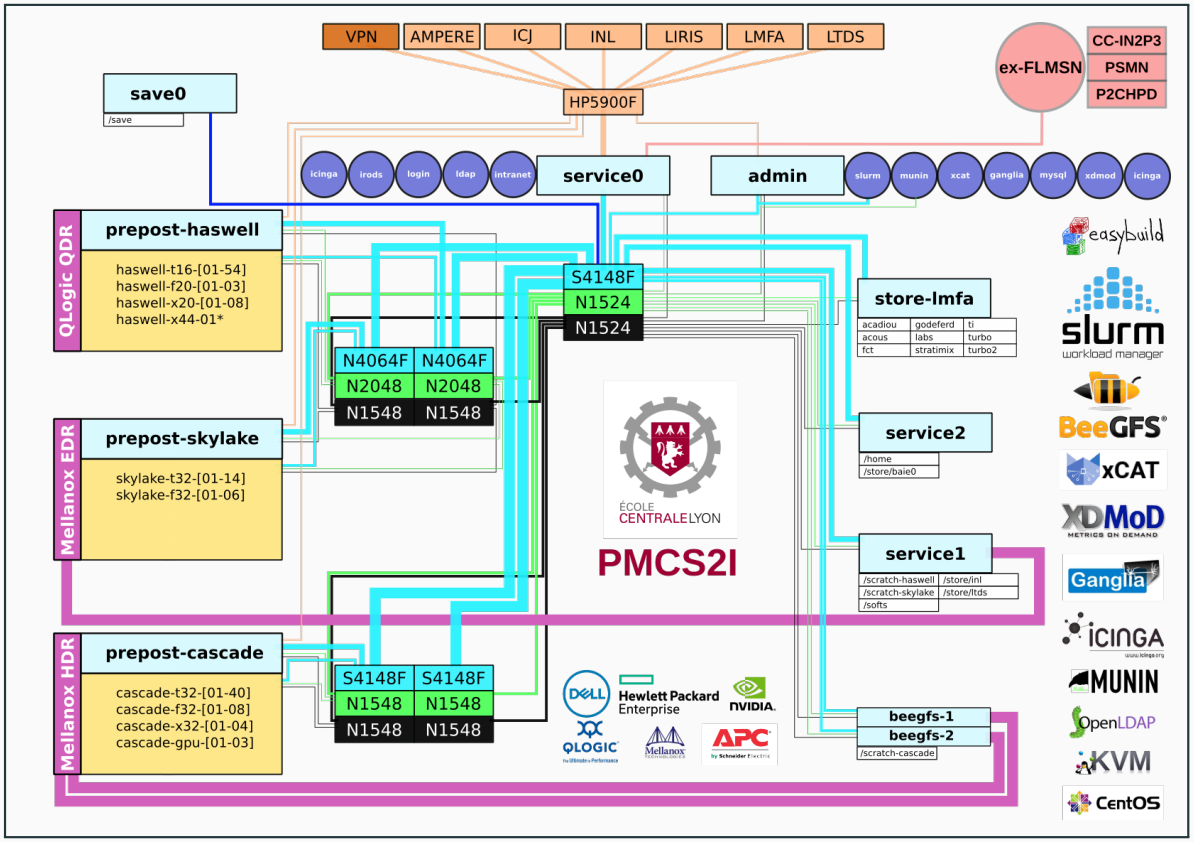
- `icinga + grafana`
- `InfluxDB` : open source time series database
- interface centralisée et partagée
- alertes via `Rocketchat`/mail/xmpp
- intégration continue via `gitlab-ci`

Interface centralisée

- prises de décision et validation des évolutions
- extension au monitoring du calculateur

Architecture des moyens de calcul du PMCS2I

- 3480 coeurs (1128 haswell, 640 slylake, 1712 cascade)
- 25 To RAM
- 6 GPUs RTX6000
- 2 Po données
- 40 To stockage haute perf (BeeGFS)
- 3 serveurs pré-post-traitement
- réseau 10 Gb/s
- 3 îlots Infiniband
- 2 M d'h par mois
- 256 utilisateurs sur 4 laboratoires



Cluster de calcul

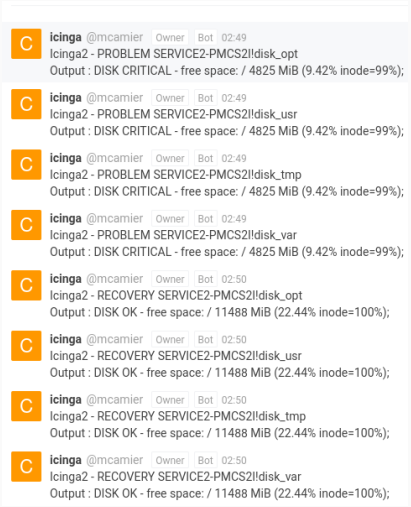
- architecture : 4 réseaux, 4 stockage, 3 queues de calcul, 300 logiciels
- détection des problèmes par les utilisateurs
- objectif de remonter les problèmes en amont

Supervision de l'infrastructure de calcul

Écriture de check spécifiques au cluster (icinga)

- sur tous les serveurs : consommation énergétique via iDRAC/iLO
- sur les noeuds : point de montage, InfiniBand, slurmd, hyperthreading
- sur les prepost : service GDM, OpenSM
- service SLURM : service slurmctld, munge - node status
- service beegfs : status, quota

alertes via RocketChat → détecter et corriger des problèmes **avant les utilisateurs**



Supervision de l'infrastructure de calcul

Interface centralisant les informations (grafana)

- usage des coeurs, nombre de jobs, d'utilisateurs
- usage des machines de pré-post production
- flux réseau
- usage des partitions

Détection des usages des prepost

Politique d'usage des machine de prepost

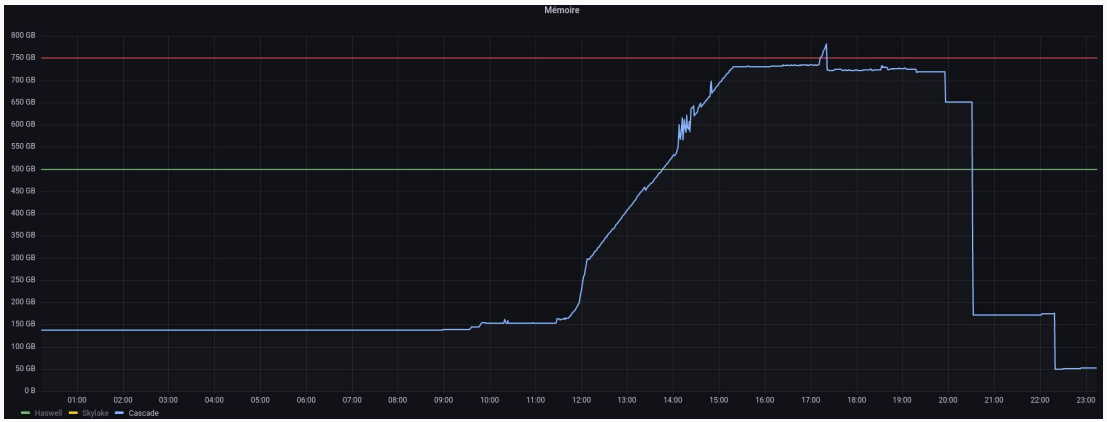
- pas de soumission $\rightarrow$ machine partagée
- bonne pratique : pas plus de 50% CPU, pas plus de 50% RAM, pendant 2h
- accès SSH ou X2Go (+ VirtualGL)

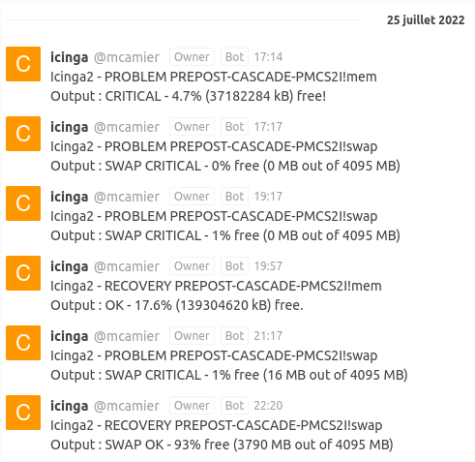
Détection des usages des prepost

Politique d'usage des machine de prepost

- pas de soumission → machine partagée
- bonne pratique : pas plus de 50% CPU, pas plus de 50% RAM, pendant 2h
- accès SSH ou X2Go (+ VirtualGL)

→ usage réel : les utilisateurs font comme ils veulent





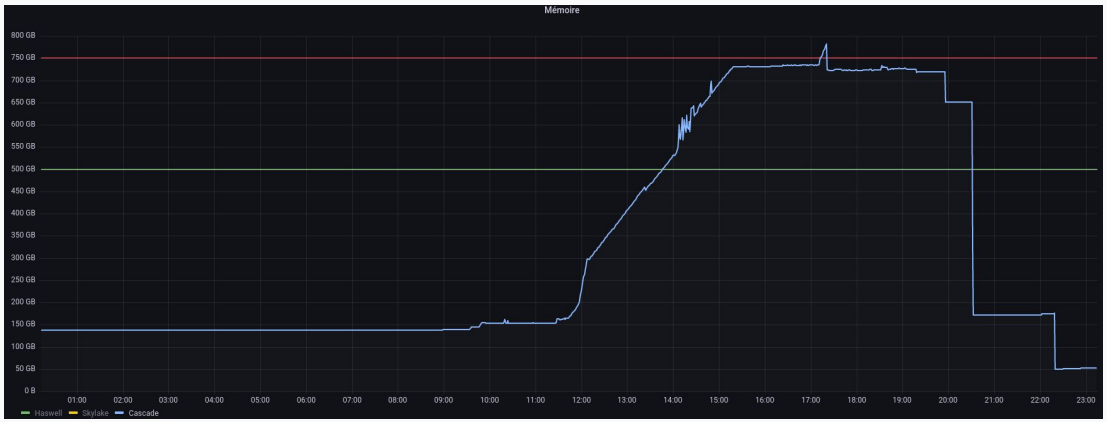
Usage mémoire d'une prepost

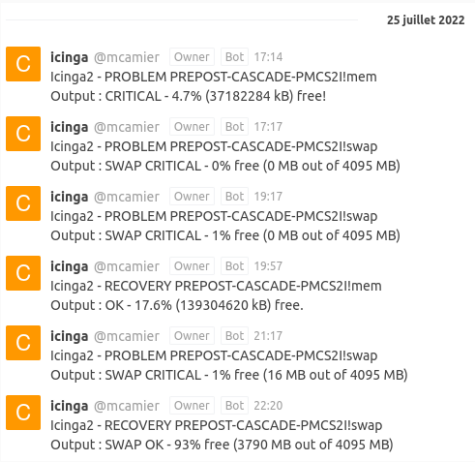
Détection des usages des prepost

Politique d'usage des machine de prepost

- pas de soumission → machine partagée
- bonne pratique : pas plus de 50% CPU, pas plus de 50% RAM, pendant 2h
- accès SSH ou X2Go (+ VirtualGL)

→ usage réel : les utilisateurs font comme ils veulent





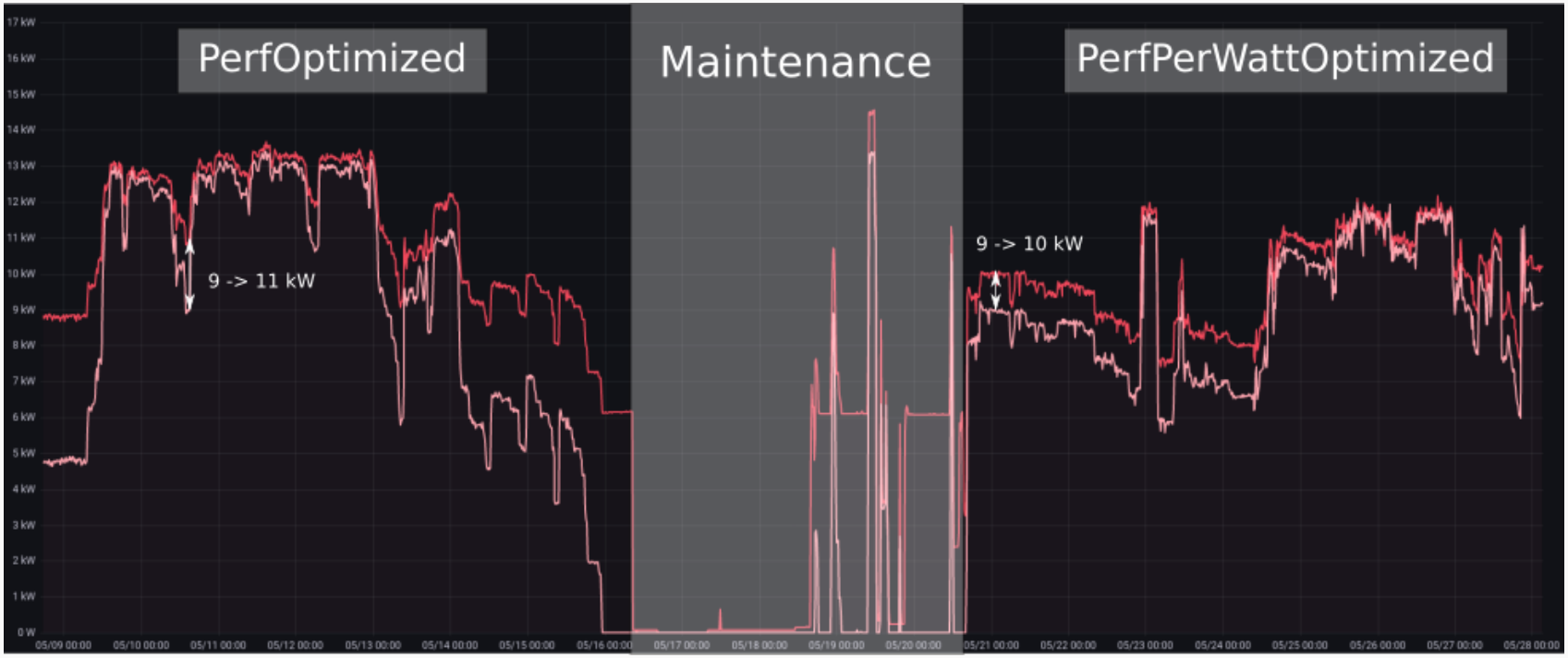
Usage mémoire d'une prepost

→ permet de détecter, puis de prévenir l'utilisateur de sa mauvaise utilisation

Amélioration de l'efficacité énergétique

Le monitoring permet de quantifier les décisions prises

- ajustement des consignes de climatisations
- réduction de la consommation par un paramètre BIOS



Écart entre noeuds utilisés et noeuds allumés avant et après le changement de paramètre BIOS

Synthèse et Travail en cours

Résultats obtenus

- une salle supervisée, dont on contrôle les flux
- un cluster sous surveillance... dont on peut piloter le coût énergétique

Outil de supervision

- ajout de nouveaux check au fur et à mesure des besoins
- envoi automatique aux utilisateurs concernés par un mauvais usage

Fiabilisation des nœuds et des modules/logiciels

- validation de la performance d'un noeud dans le prolog SLURM
- benchs automatique avec ReFrame ?