Empreinte carbone des heures de calculs : limites et paradoxes

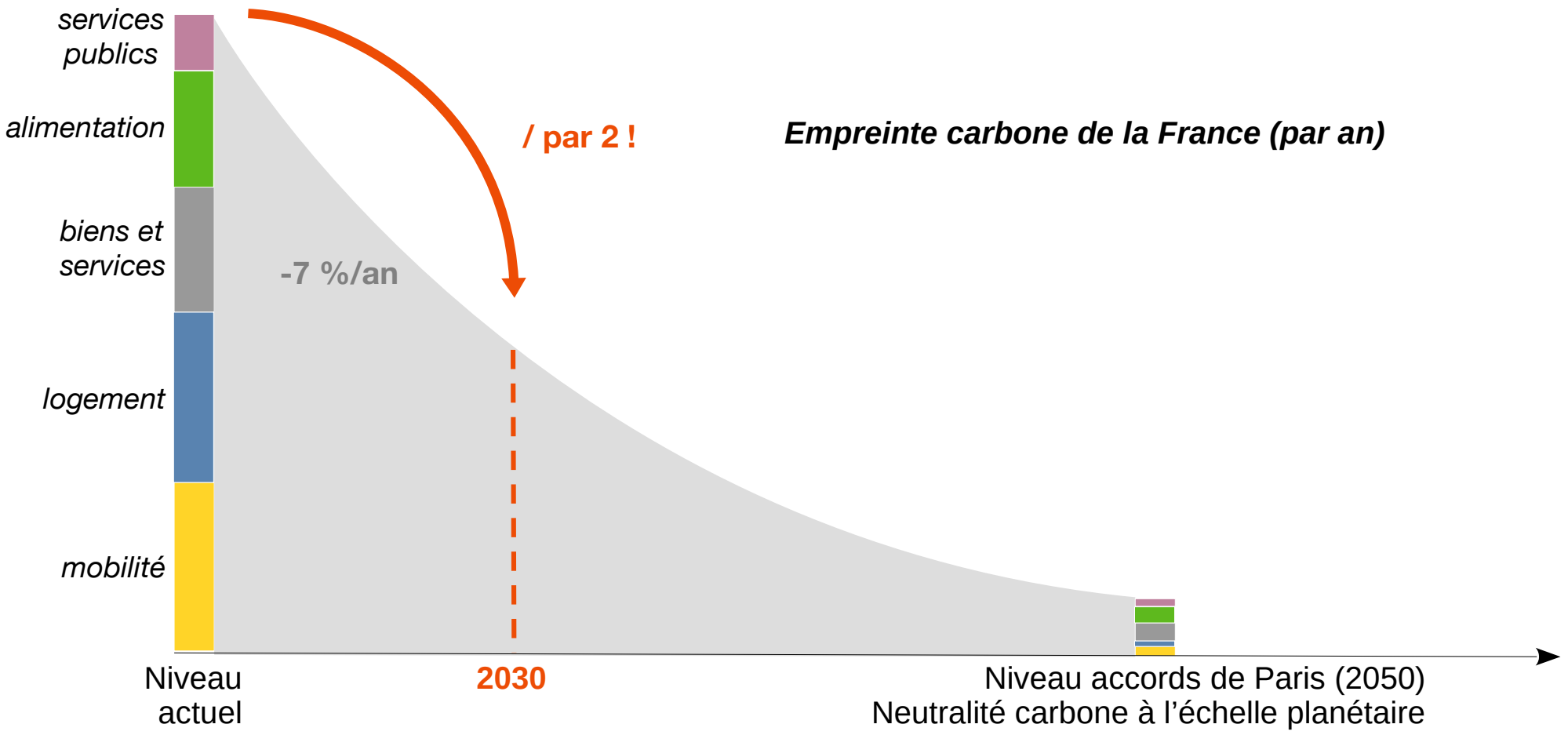
Gaël Guennebaud

Empreinte carbone des heures de calculs – limites et paradoxes*

Gaël Guennebaud (*gael.guennebaud@inria.fr*)
2022 – 10 – 21

POUR UNE INFORMATIQUE ÉCO-RESPONSABLE

Enjeu climatique : pas plus de +2° ?

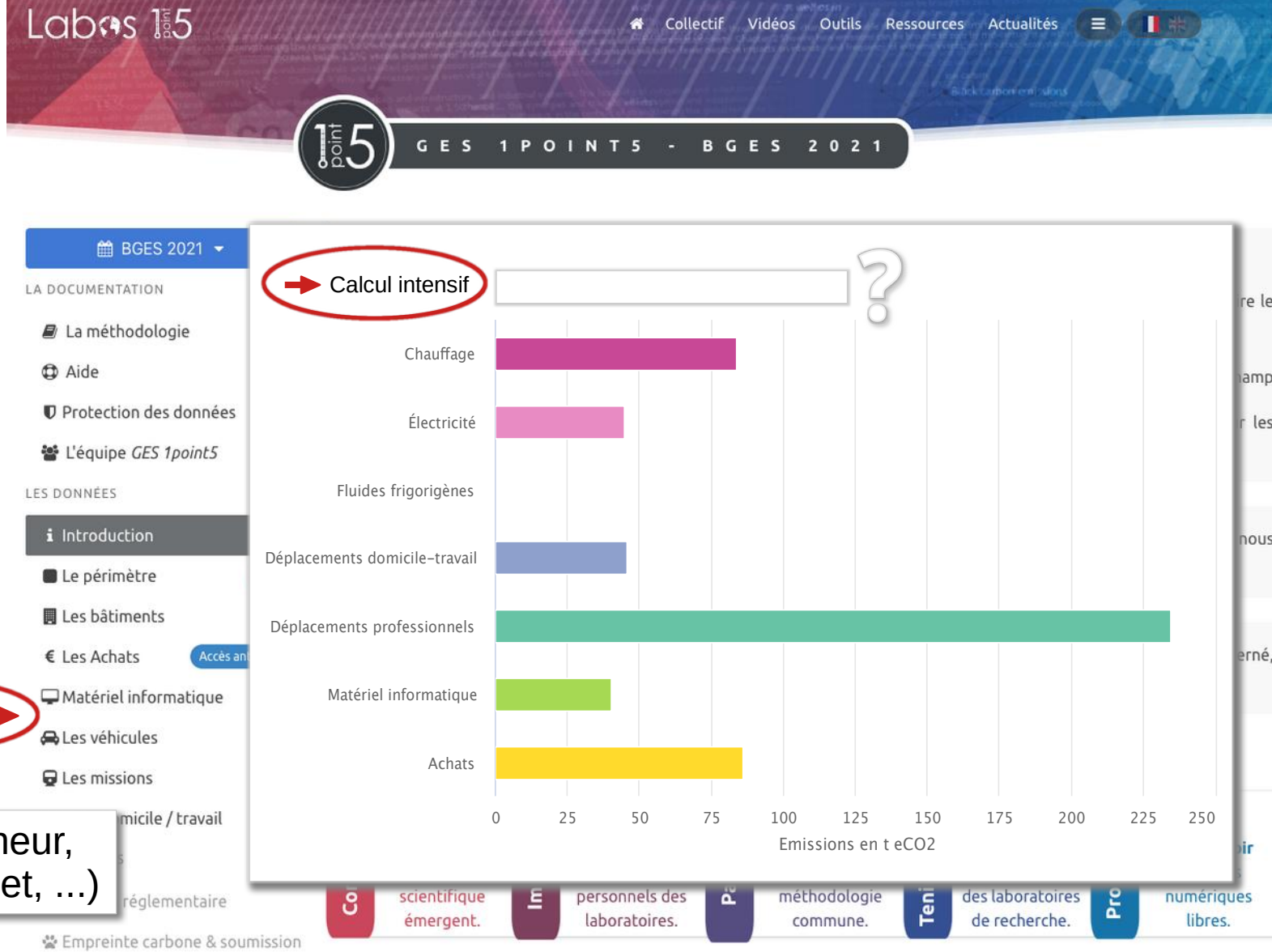


Estimer l'empreinte carbone ?

Bilan GES des **laboratoires**

Calcul intensif

(Empreinte d'un chercheur, d'une équipe, d'un projet, ...)



Estimer l'empreinte carbone ?

Empreinte des **publications** (en IA)

Difficile à exploiter car dépend
- du périmètre,
- de la méthodologie,
- de nombreuses hypothèses...

the best performing model chosen based on the validation set performance. The experiments were implemented in PyTorch and trained on a single Nvidia GTX 1080 graphics processing unit (GPU) with 8GB memory. The development and training of all models in this work was **estimated** to produce **71.9 kg of CO2eq**, equivalent to 601.5 km travelled by car as **measured** by **Carbontracker**[14] (Anthony et al., 2020).

Patch-based Medical Image Segmentation using Matrix Product State Tensor Networks, Selvan et al., Journal of Machine Learning for Biomedical Imaging, 2022. [arxiv]

Ex. d'outils : CodeCarbon, Experiment impact tracker, Carbontracker, EnergyScope, Scaphandre, Green Algorithms, etc.

« Mesurer » pour **réduire** ?

Sensibiliser les « consommateurs »…

et donc **espérer** enclencher des mécanismes de sobriété…

→ Est-ce suffisant ?

→ Est-ce vraiment nécessaire ?

- Peut ralentir l'action
- Peut amener à des actions contre-productives *(tout comme l'absence de mesure des impacts !)*

Deux grandes approches

Bottom-up

- Basée consommation pendant le job
  - → compteurs matériel/logiciel
  - → périmètre (très) limité

→ plutôt pour analyse de code que bilan GES

Deux grandes approches

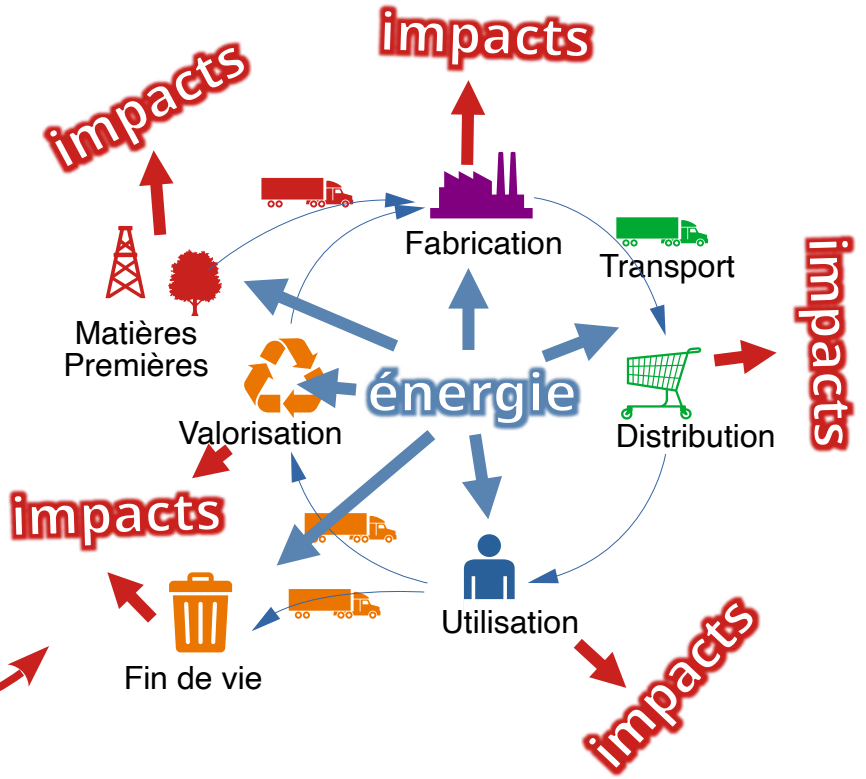
Top-down

- Basée ACV et allocation de l'empreinte totale

→ x gCO_2e / heure.coeur

→ empreinte = x x nb heures.coeurs

Reflet de la vérité terrain ?



+ Justice climatique & sociale

Crise climatique

Effondrement de la biodiversité

+ Épuisement des ressources

Changement d'utilisation des sols

Eau

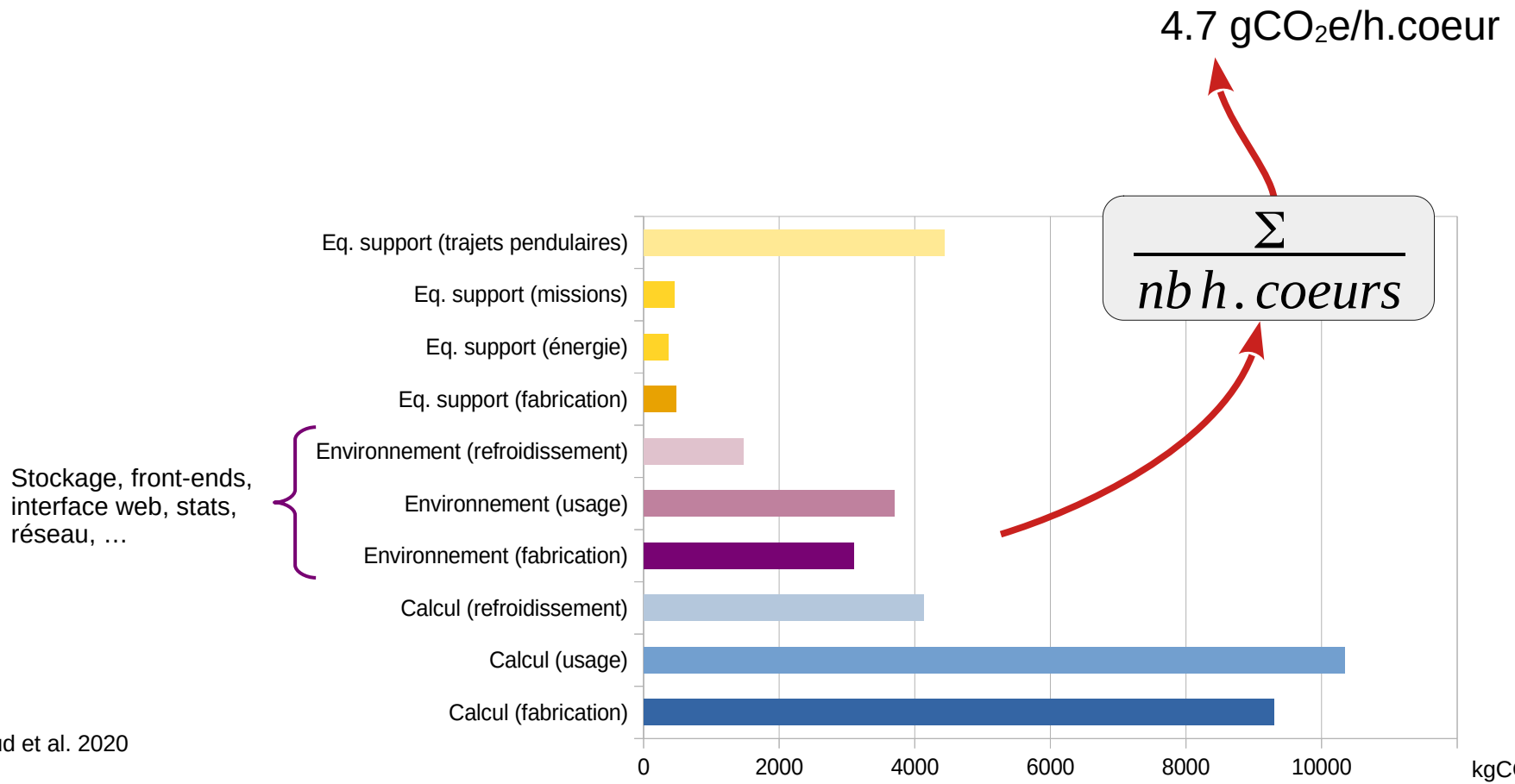
Pollutions

Limite #0 : beaucoup d'autres impacts

Ex. fabrication d'un serveur
1300-4000 kgCO_2e

- Épuisement des ressources
  - Tension sur les métaux entre **ENR**, mobilité élec, TIC
- Pollutions des sols, de l'eau
  - Déchets des mines
- Questions éthiques
  - Travail enfants
  - Appropriation des terres
  - Conflits d'accès à l'eau
- Tensions géopolitiques, conflits armés

« x gCO_2e / heure.coeur » ?



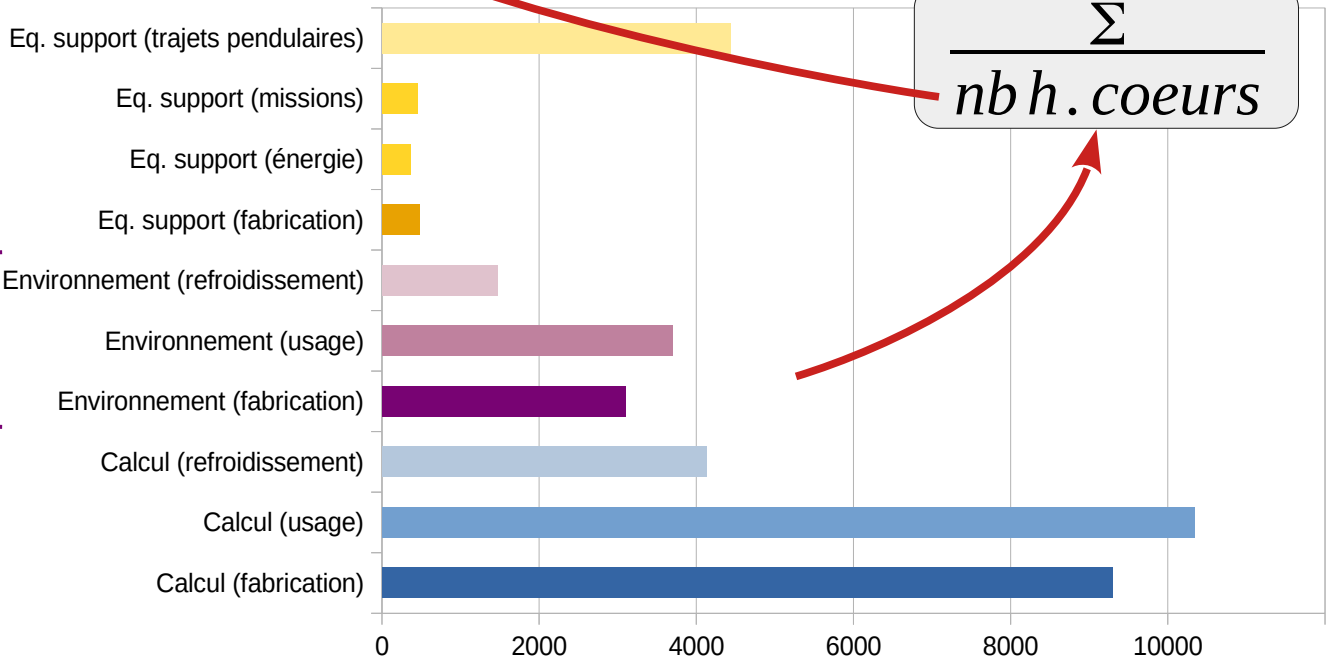
GRICAD, Berthoud et al. 2020

Limite #1 : normalisation

Réellement connu à la fin de l'année (ou de la période)
→ lié au taux d'usage

4.7 gCO_2e/h.coeur

$$\frac{\Sigma}{nb\,h.coeurs}$$



Stockage, front-ends, interface web, stats, réseau, …

GRICAD, Berthoud et al. 2020

Limite #2 : périmètre

...
Déploiement / maintenance
Fuites fluide frigorigène
Baies, cables, etc. (fabrication)
Refroidissement (fabrication)
Bâtiments (fabrication)

+ Impacts indirects & connexes

- Conception/dév. des codes
- Acquisition/gestion des données
- Exploitation des résultats
- ...

Eq. support (trajets pendulaires)
Eq. support (missions)
Eq. support (énergie)
Eq. support (fabrication)
Environnement (refroidissement)
Environnement (usage)
Environnement (fabrication)
Calcul (refroidissement)
Calcul (usage)
Calcul (fabrication)

0 2000 4000 6000 8000 10000 kgCO₂e/an

Stockage, front-ends, interface web, stats, réseau, …

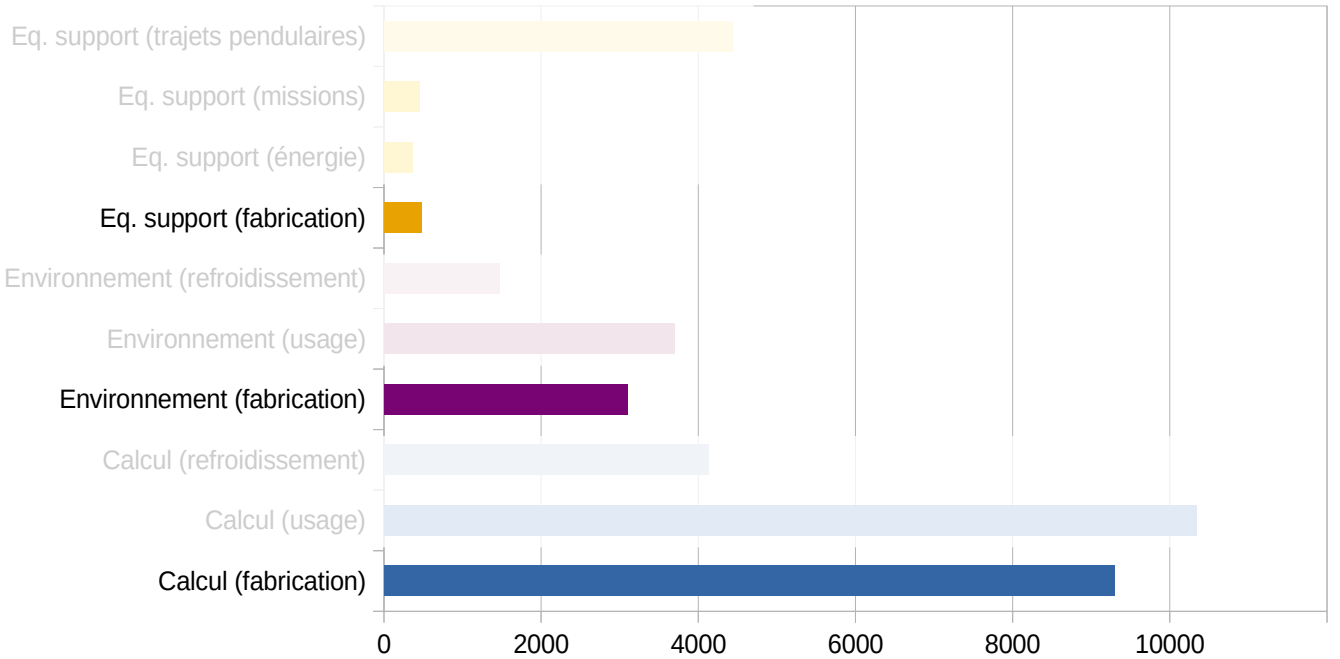
Part comptabilisable par Carbontracker

Limite #3 : *troncation*

Ex. fabrication des équipements info

- Construction
  - des engins
  - des usines
  - des bureaux
- R&D
- Marketing
- Etc.

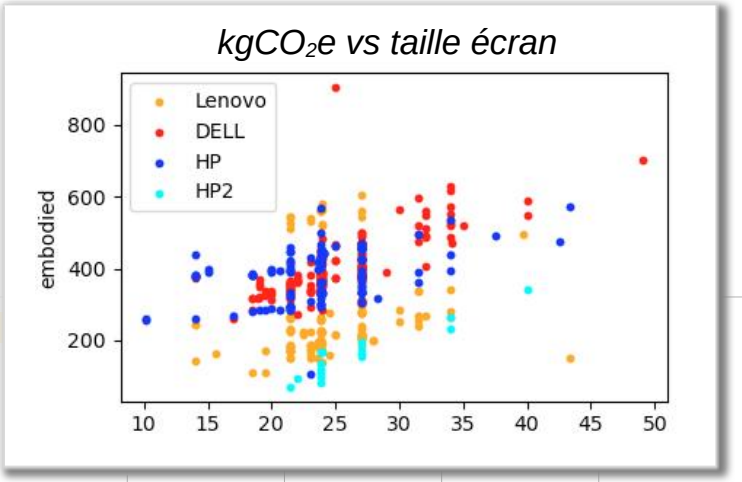
→ +40 % ?

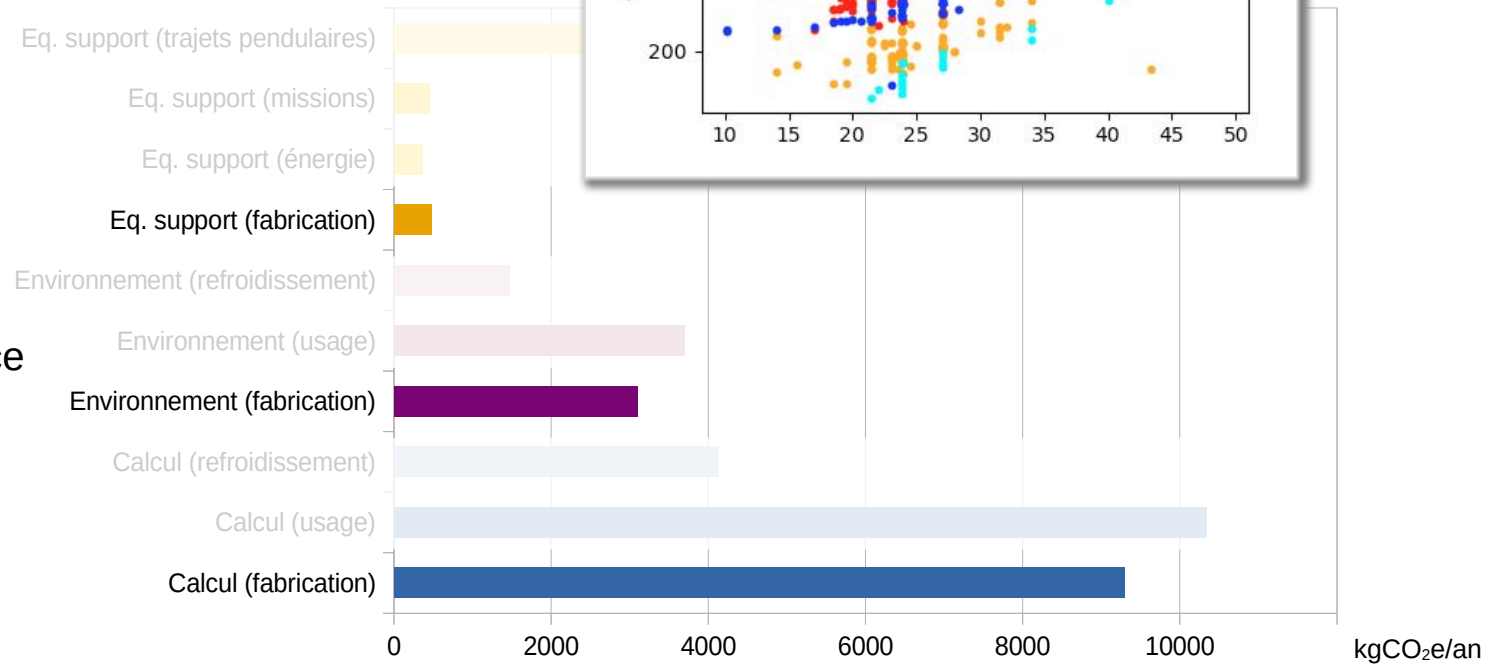


Limite #4 : incertitudes

Ex. fabrication des équipements info

- Configuration
  - CPUs, SSDs, GPUs
  - ...
- Intrinsèque
  - Manque de données
  - Manque de transparence
- Méthodologique





Limite #4 : incertitudes

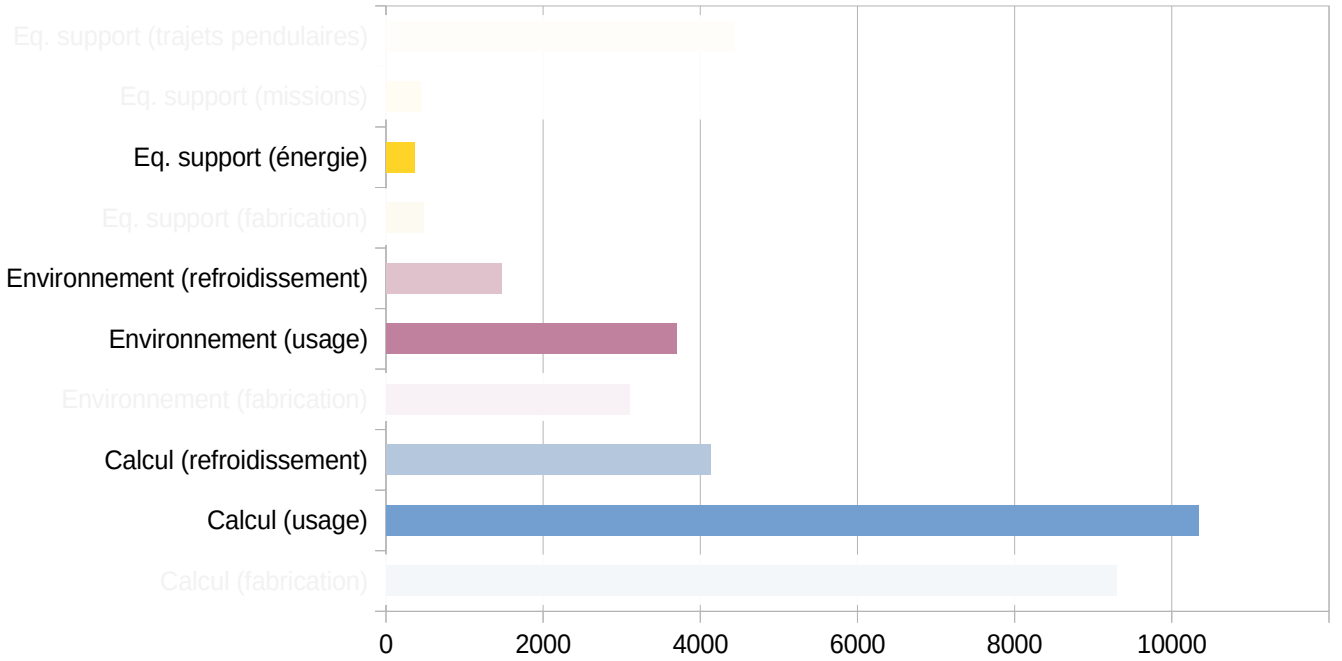
(la seule « vraie » mesure)

Ex. conso kWh élec → $kgCO_2e$?

$$= consommation\_mesurée \cdot FE_{élec}$$

- Intrinsèque (amonts)
- Méthodologique
  - Granularité :
    - Moyenne de l'année
    - Moyenne du mois
    - Moyenne du jour
    - Moyenne instantanée *(ex. CarbonTracker)*
  - Moyenne ?

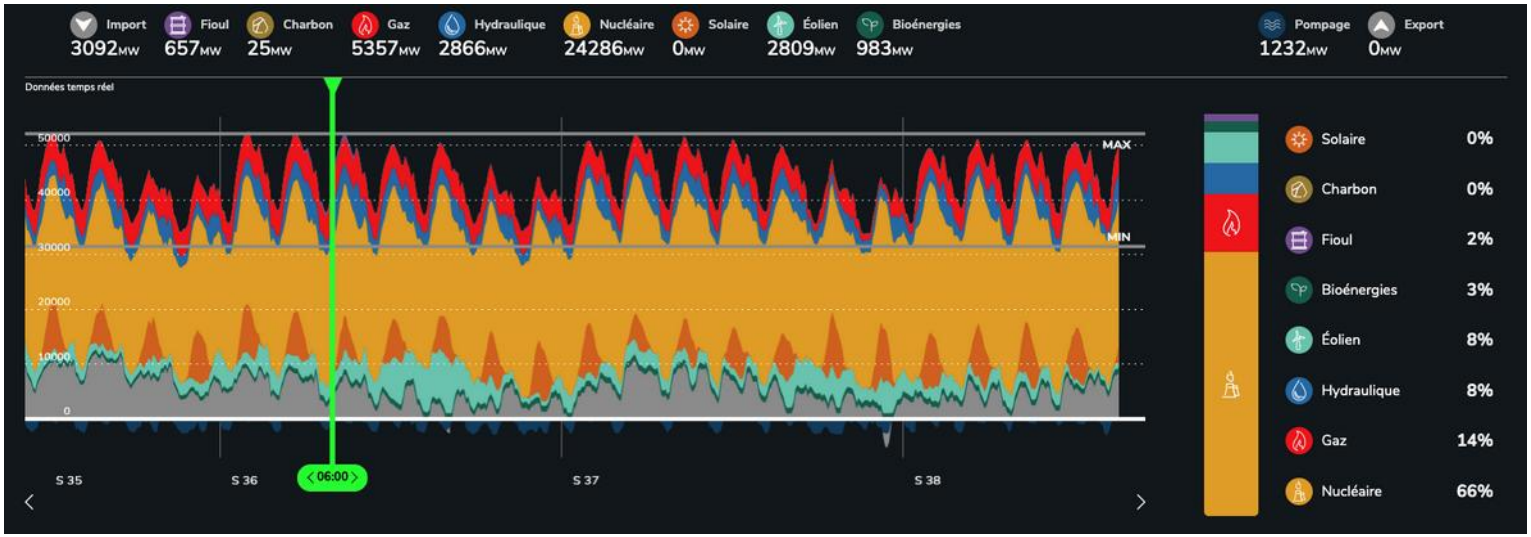
→ Paradoxe #1



Paradoxe #1 : FE moyen vs marginal

Si période avec centrale à gaz à 50 %

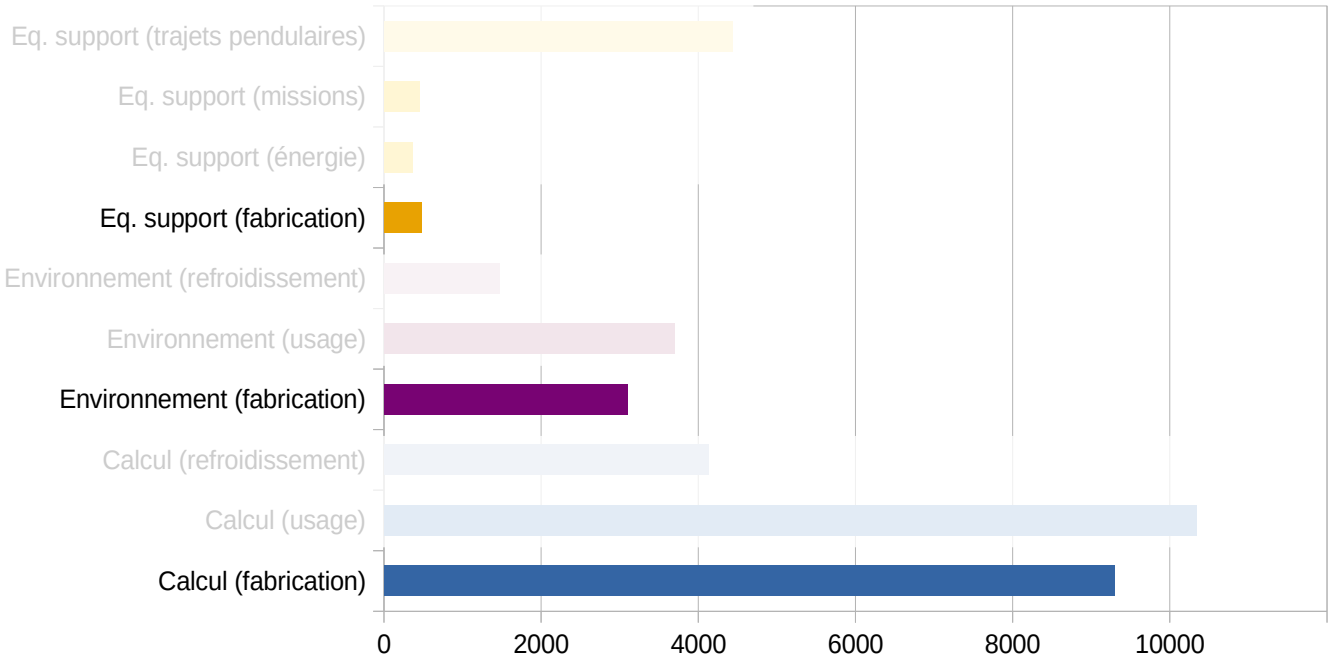
- Moyenne de : **85** gCO_2e/kWh
- Baisse de la conso → baisse de la prod. par gaz → **500** gCO_2e évités par kWh évités
- Hausse de la conso → hausse de la prod. par gaz → **500** gCO_2e/kWh ?



Limite #5 : durée de vie

$$E_{h.coeur} = \frac{E_{fabrication}}{lifetime \cdot nb_{h.coeurs}} + \cdots$$

Réellement connue
à sa fin de vie !
(ici anticipé à 7ans)



Paradoxe #2
faire durer

$$E_{h.coeur} = \frac{E_{fabrication}}{lifetime \cdot nb_{h.coeurs}} + \frac{E_{élec}}{nb_{h.coeurs}} + \cdots$$

(++ sur $lifetime$ et $E_{élec}$)

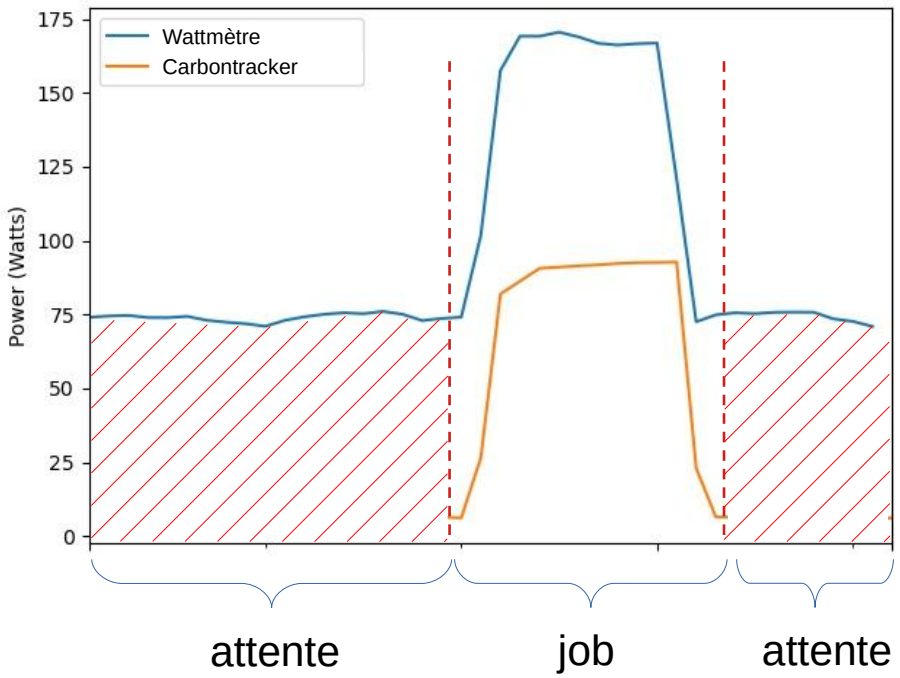
Faire durer une tranche → bien ?

- Impacts fabrication ←?→ Impacts usage
  - → si continue de répondre à **un vrai besoin** → OK ?
  - → si délaissé* → gabegie énergétique

*(*sans extinction auto des noeuds)*

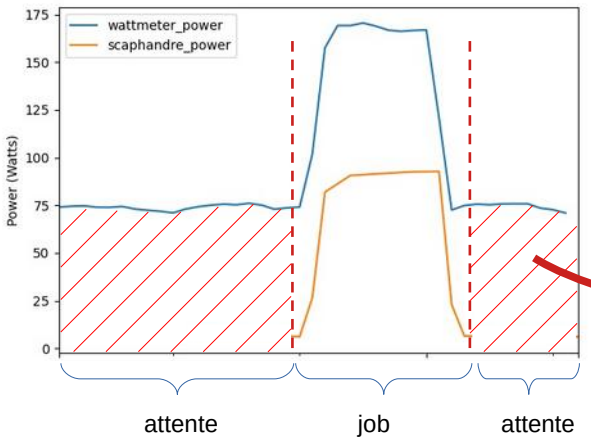
Limite #6 : non proportionnalité

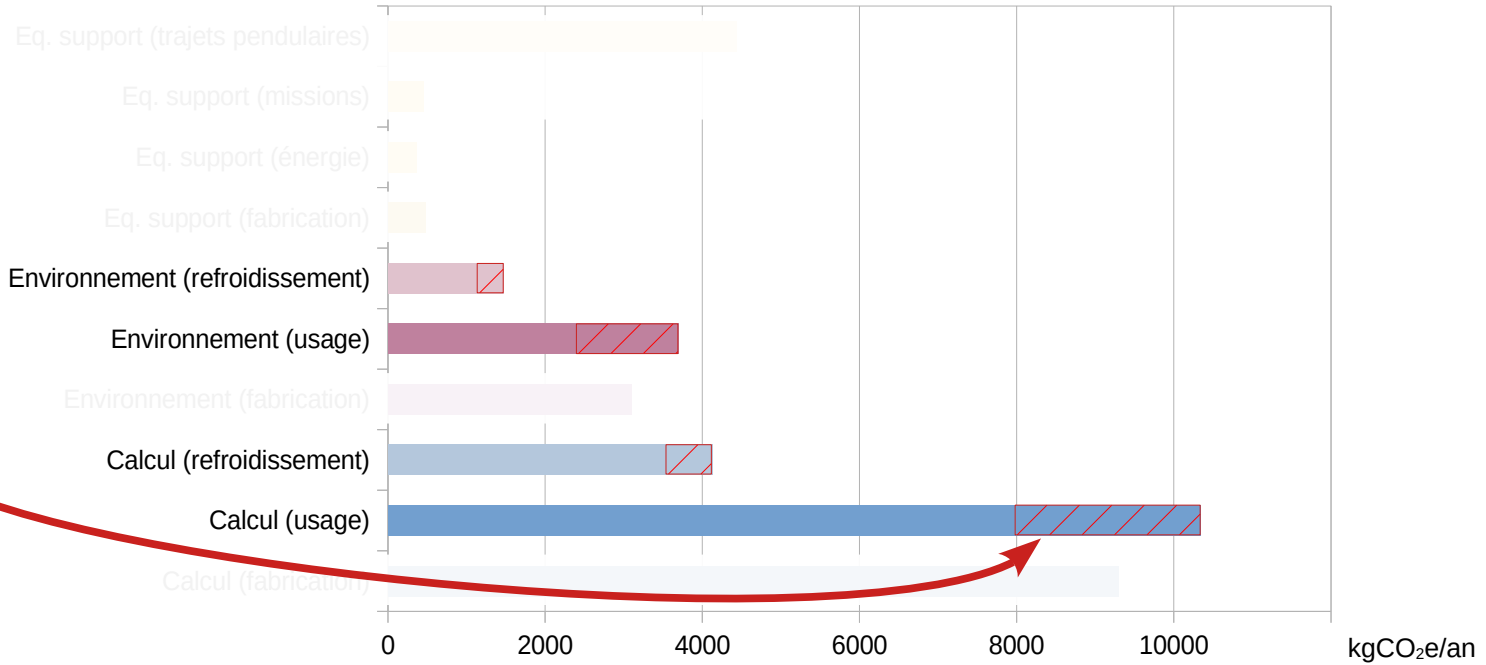
Zéro job = beaucoup d’énergie « gâchée »



Limite #6 : non proportionnalité

Zéro job = beaucoup d'énergie « gâchée »



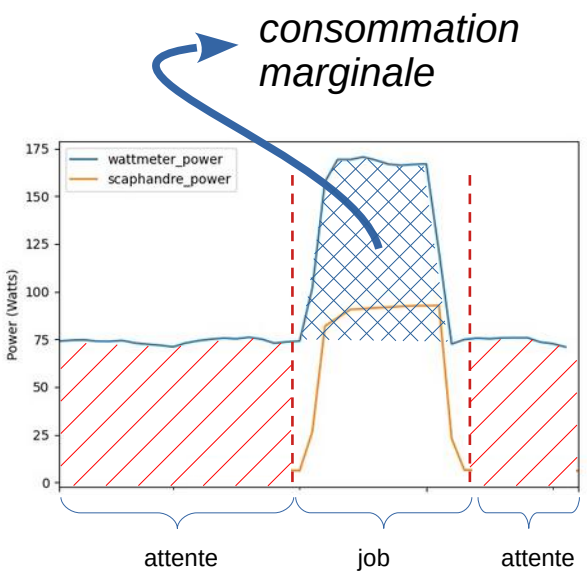


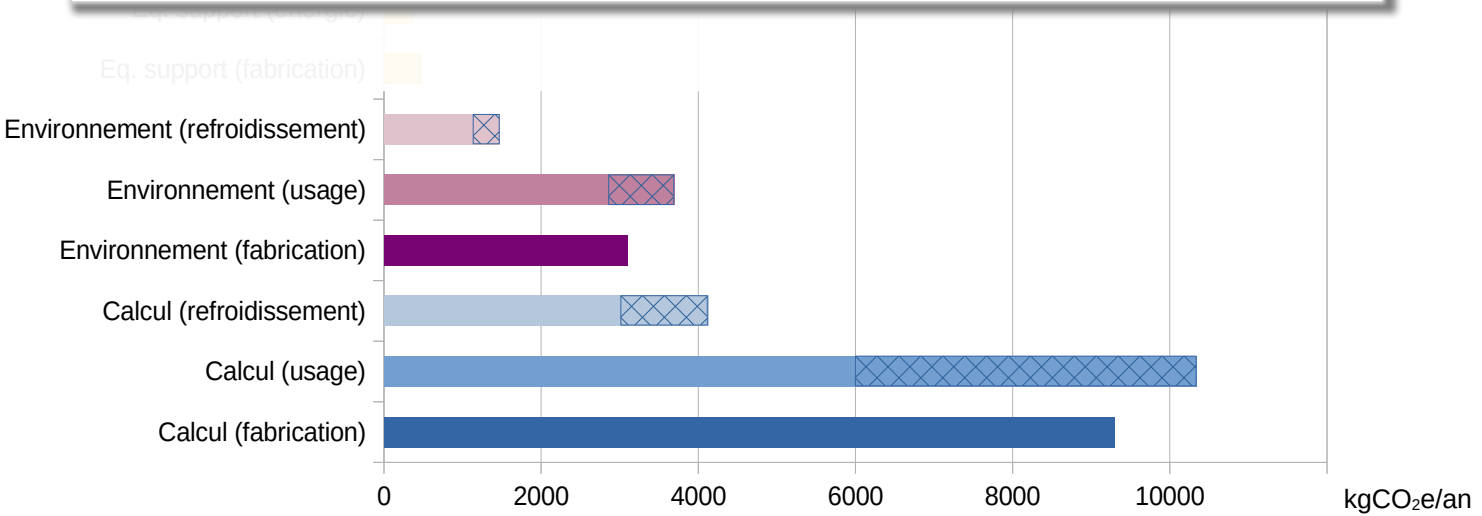
Limite #6 : non proportionnalité

Zéro job = beaucoup d'énergie « gâchée »

4.7 gCO_2e/h.coeur

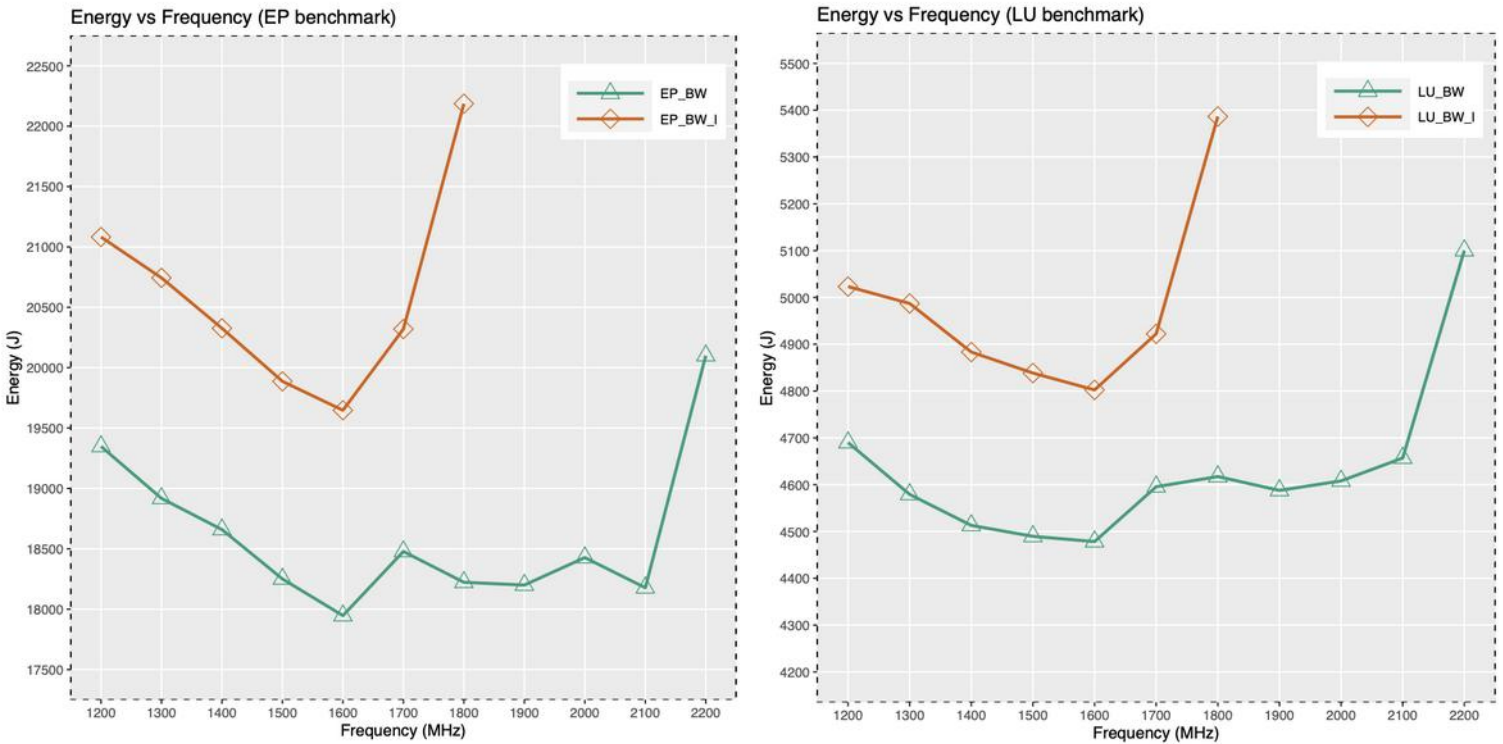
- N heures supplémentaires → Nx4.7 gCO_2e supplémentaires
- N heures évitées → Nx4.7 gCO_2e évitées





Limite #6 : non proportionnalité

Low power processors consume less energy.



BW_l: Xeon E5-2630L v4 (Broadwell) -> low power processor (orange)
BW: Xeon E5-2630 v4 (Broadwell) (green)

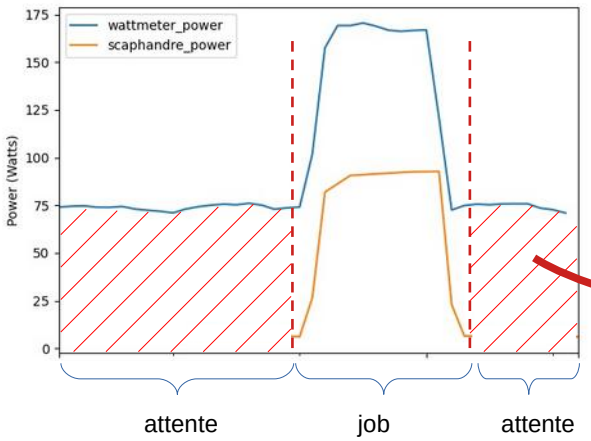
[ISCC 2021]

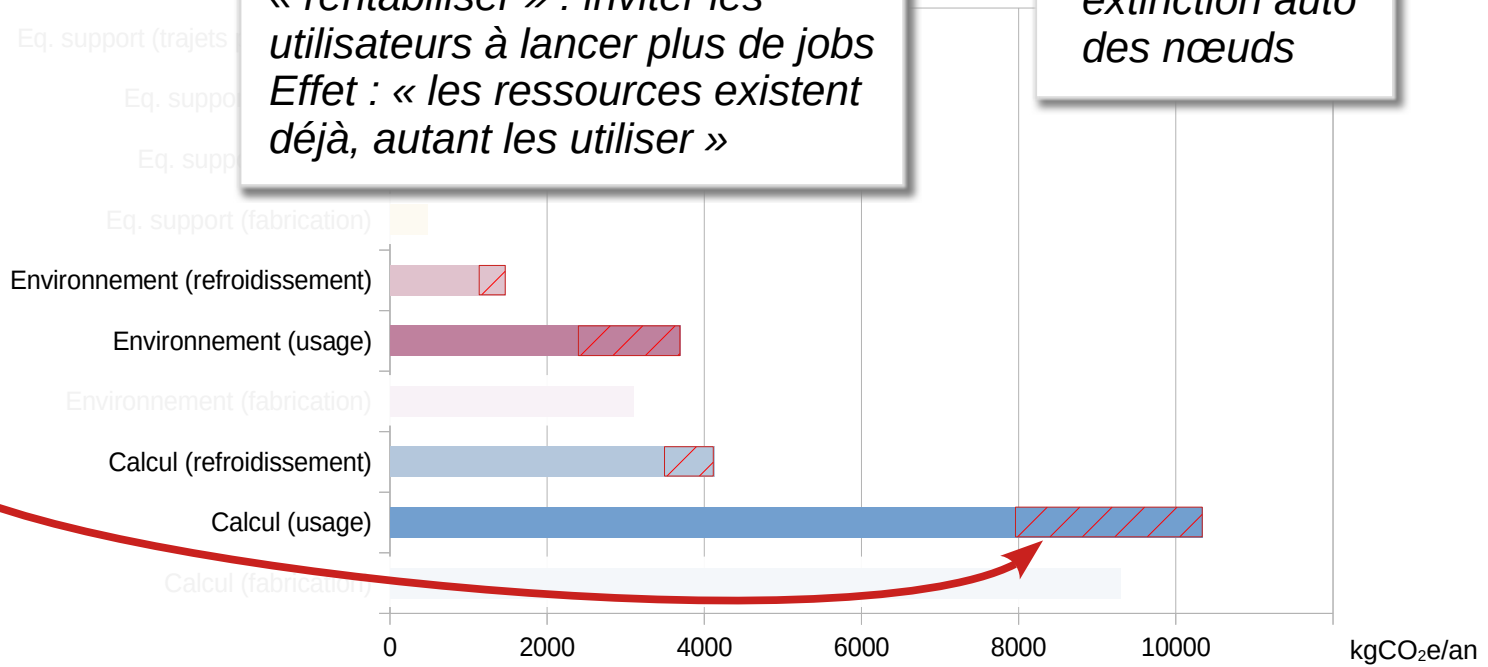
Paradoxe #3 : les coûts irrécupérables

Zéro job = beaucoup d'énergie « gâchée »

« rentabiliser » : inviter les utilisateurs à lancer plus de jobs
Effet : « les ressources existent déjà, autant les utiliser »

extinction auto des nœuds





Paradoxe #4
tau d'usage

$$E_{h.coeur} = \frac{E_{fabrication}}{lifetime \cdot nb_{h.coeurs}} + \frac{E_{élec}}{nb_{h.coeurs}} + \cdots$$

($E_{h.coeur}$ ↘ ; $E_{fabrication}$ ↗ ; $E_{élec}$ ↗↗ ; $nb_{h.coeurs}$ ++ ; $nb_{h.coeurs}$ ++)

Baisser les *gCO_2e/h.coeur* c'est mieux, non ?

- Contre exemple : tau d'usage ++
  - → baisse des gCO_2e/h.coeur
  - → augmentation de la consommation globale garantie !
- Effets **indirects** : tau d'usage ++
  - → serveurs saturés
    - → ajouts de nouveaux serveurs → impacts ++
  - → réduit la possibilité d'extinction auto → ~~impacts--~~

Paradoxe #5 : PUE

Baisse du **PUE** ↔ baisse des kWh, non ?

- Contre exemple : extinction auto des nœuds

  → augmentation du PUE

  → mais diminution de la consommation globale garantie !

$$E_{élec}^{--} = E_{info}^{--} + E_{clim+autre}^{-} = \underbrace{\left(\frac{E_{info}^{--} + E_{clim+autre}^{-}}{E_{info}^{--}} \right)}_{\text{PUE}^{+}} \cdot E_{info}$$

Paradoxe #6 : compétitivité

Moore/Koomey's laws

→ Les *$gCO_2e/h.coeur$* diminuent au cours du temps

Cluster A (2019)	VS	Cluster B (2022)
3.7 gCO_2e/h.coeur		2.1 gCO_2e/h.coeur

Mais solution plus efficace ?

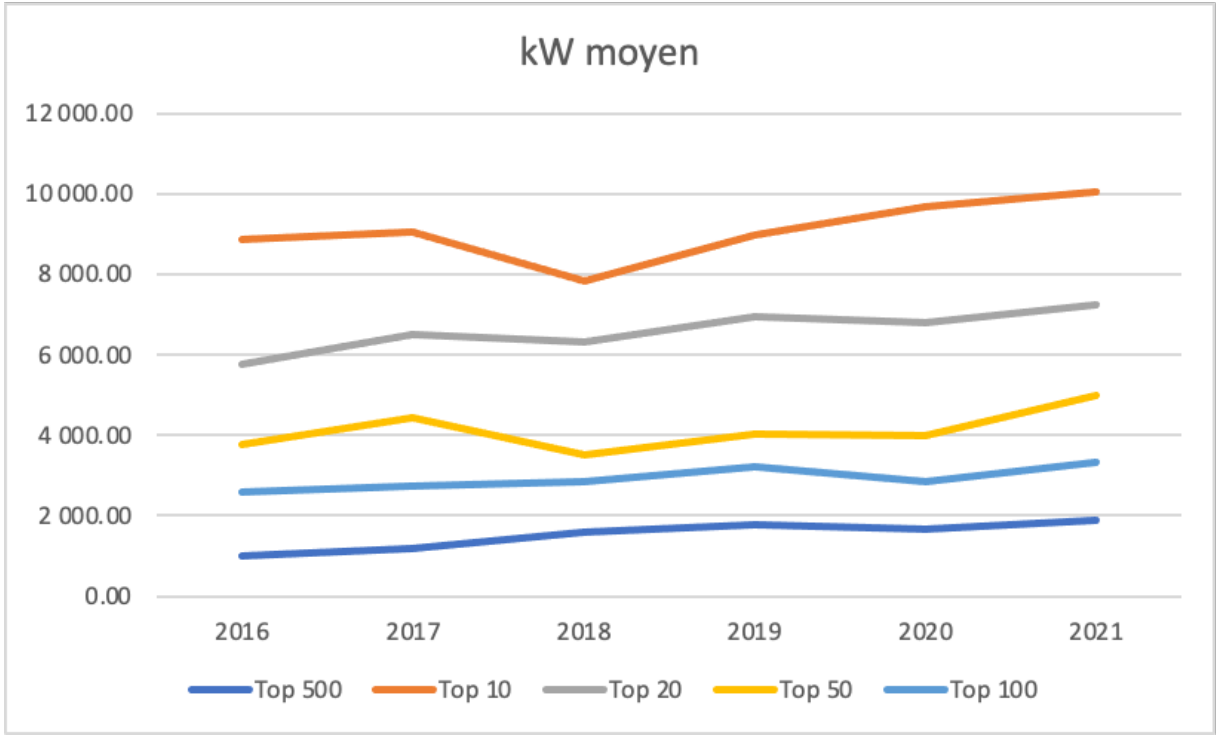
→ Obsolescence prématurée

→ Renouvellement ++ → impacts ++

Paradoxe #7 : efficacité énergétique

(aka paradoxe de Jevons)

W/Gflops en baisse constante, et pourtant :



+ fabrication

Conclusion

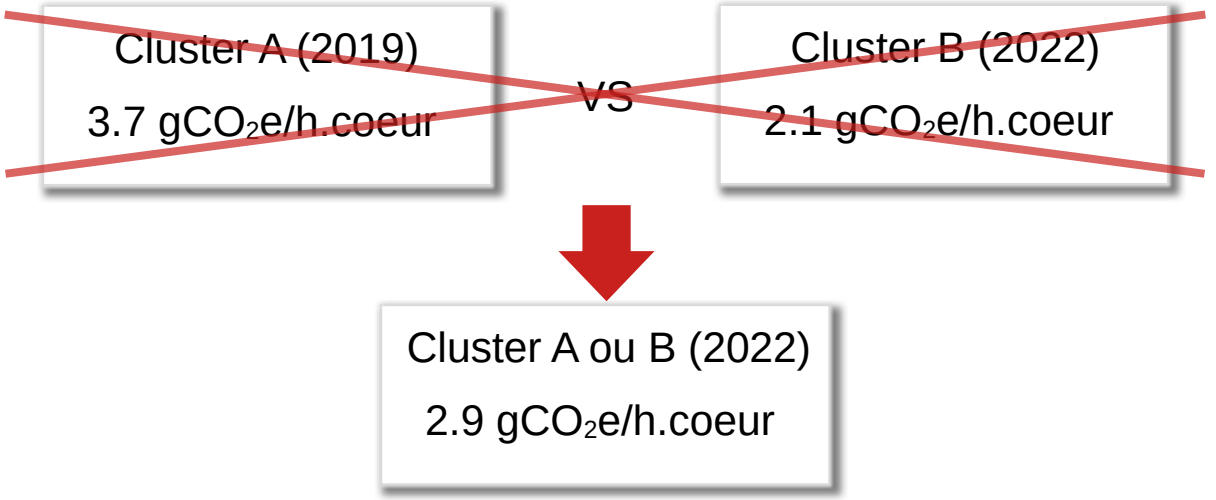
« ~~*4.7 $gCO_2e/h.coeur$*~~ » → qu'en penser ?

- → rester humble, la *vérité terrain* n'existe pas
  (question mal posée)
- → **pas de proportionnalité**
- → mélange des choux et des carottes
- → danger des indicateurs **d'efficacité**
  (PUE, x gCO_2e/unité, …)
- → individualise les responsabilités

Une voie de sortie : penser global et collectif

Étape 1 : « service mutualisé »

Retour sur notre dilemme :



→ approche choisie par le GENCI *(quid de la transparence)*
→ approche envisagée dans GES1.5 à une échelle plus large *(avec variantes)*

Étape 2 : penser global

gCO2e/h.coeur → **total** en *tCO2e/an, GWh/an, ...*

pour les infra calcul de la recherche en France

- Gros des impacts **à la prise de décision** d'installer un nouveau cluster (ou tranche)
  - Impliquer les utilisateurs dans cette prise de décision
  - Débats éclairés *(estimations des impacts des différents postes)*
- Réintroduire les limites du monde réel