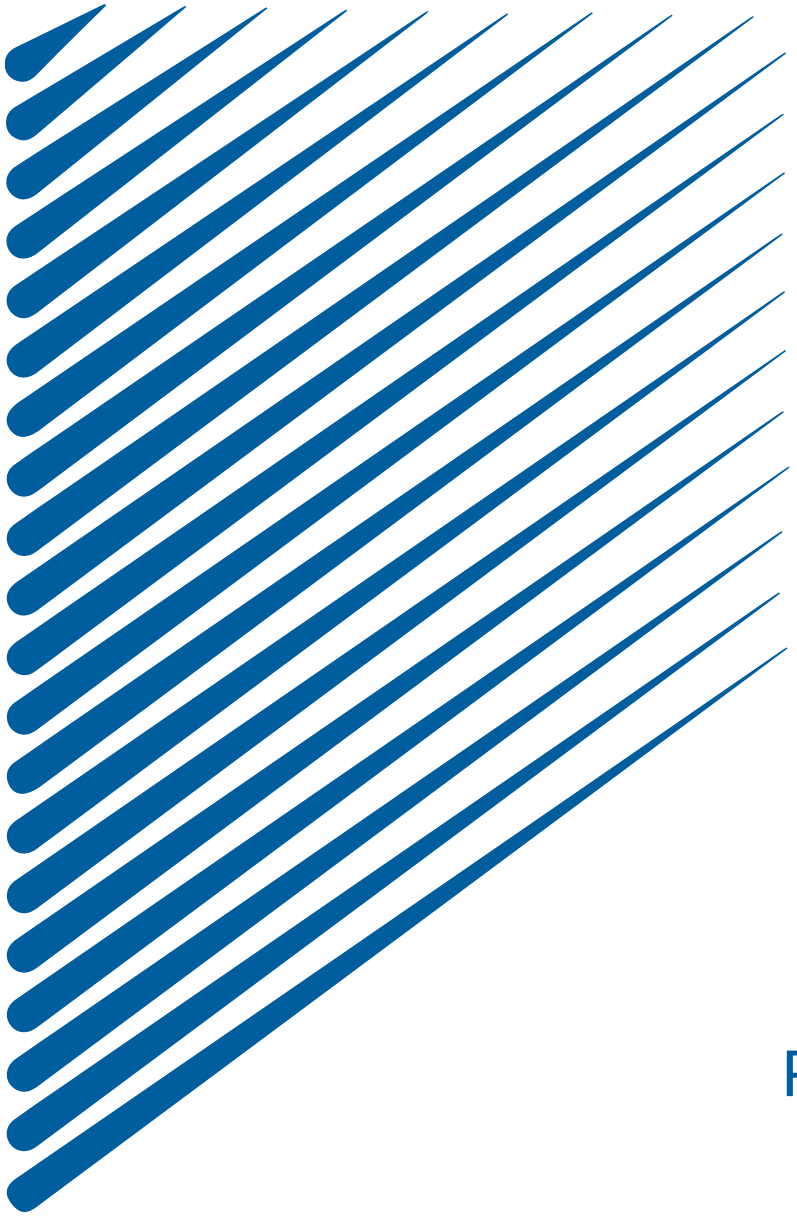


MINES PARIS | PSL

Vers une IA durable :
outils et techniques pour l'efficacité énergétique des systèmes informatiques modernes

Petr Dokladal

Roblex Nana Tchakouté, Claude Tadonki, Youssef Mesri

UN EXERCICE D'ECHAUFFEMENT

Combien font 15x17 (… sans calculette) ?

Pendant les 15 s à effectuer la multiplication votre cerveau a consommé 315 J.
(300 J pour « rester en vie » + 15 J pour ce calcul)

Le budget énergétique d'une multiplication sur un CPU moderne :

Opération	Coût Energétique [pJ]
32-bit Integer (INT32) Addition	0,1 pJ
32-bit Float (FP32) Addition	0.9 pJ
32-bit Float (FP32) Multiplication	3.7 pJ
32-bit DRAM Read	~16 à >100 pJ (LPDDR ou DDR3/DDR4)

Note: La consommation énergétique dépend de la technologie de gravure. Les gains en dessous de 3nm ne sont plus linéaires (augmentation de pertes liées aux effets quantiques).

ÉVOLUTION DU MACHINE LEARNING, DU DEEP LEARNING VERS L'IA

1943 – Neurone artificiel - le premier modèle mathématique (McCulloch et Pitts)

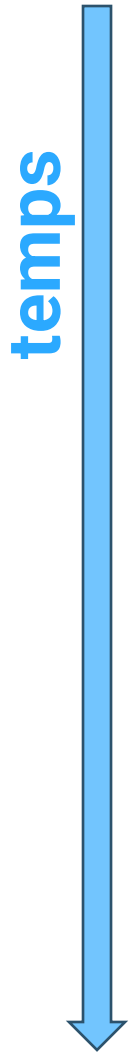
1957 – Perceptron - un algorithme d'apprentissage supervisé inspiré du fonctionnement des neurones biologiques (Rosenblatt)

1986 – Rétropropagation - algorithme de rétropropagation relance l'intérêt pour les réseaux de neurones multicouches, permettant l'apprentissage plus complexe (Rumelhart, Hinton & Williams)

< hiver de l'IA de 1990 à 2000 >

2012 – Révolution du deep learning – (Hinton et ses étudiants) remportent le concours ImageNet avec AlexNet, un réseau de neurones convolutifs profond, marquant le début de l'ère du deep learning moderne.

2017 – **Transformer** - L'article "Attention is All You Need" (Vaswani et al.) introduit le modèle Transformer, qui révolutionne le traitement du langage naturel et devient la base des modèles comme BERT, GPT, etc.



1999 – NVIDIA introduit le terme "**GPU**" avec la GeForce 256 → Début de l'accélération graphique dédiée.
2006 – Lancement de **CUDA** par NVIDIA → Permet aux GPU de servir au calcul parallèle général, essentiel pour l'IA.
2017 – Introduction des **Tensor Cores** avec l'architecture Volta → Accélération massive des calculs de matrices pour le deep learning.
2020 – Architecture **Ampere** (NVIDIA A100) → GPU optimisé pour l'entraînement de modèles massifs (ex. GPT-3)

QUE REPRÉSENTE UNE TÂCHE IA EN TERMES DE CALCUL ET … D'ÉNERGIE

YOLOv5 – Détection d'objets en temps réel - Détection d'objets dans des images (ex. voitures, personnes, animaux).

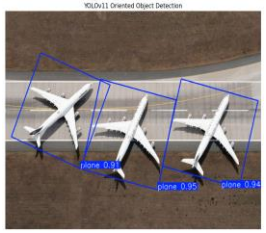
•**Taille du modèle** : ~7 millions de paramètres (YOLOv5s).

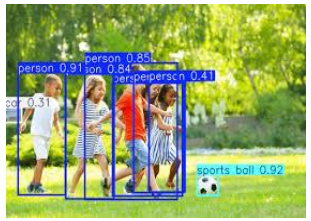
•**Temps d'entraînement** : Sur un dataset comme COCO (~120k images), avec un GPU NVIDIA V100 : **quelques heures à 1 jour**

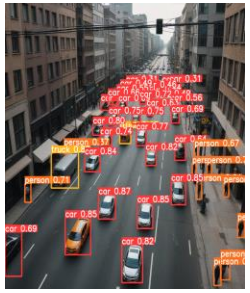
•**Temps d'inférence** : ~10 ms par image sur GPU (100 FPS)

•**Consommation énergétique estimée** :

- Entraînement complet : **~1 à 5 kWh** (selon GPU et durée).
- Inférence : **~0.01 Wh par image**.







GPT-3 – Génération de texte - Compréhension et génération de texte (ex. rédaction, traduction, dialogue).

•**Taille du modèle** : **175 milliards de paramètres**.

•**Temps d'entraînement** :

- Sur des clusters de centaines de GPU pendant plusieurs semaines.
- Estimation : **~50 GWh** (équivalent à 3 jours d'électricité d'une métropole).

•**Temps d'inférence** :

- ~0.5 à 1 seconde pour générer 100 tokens sur GPU A100.

•**Consommation énergétique estimée** :

- Inférence : **~0.02 à 0.05 Wh par requête** (selon longueur).

MINES PARIS | PSL

A L'ÉCHELLE PLANÉTAIRE 1/3 ?

La **moyenne d'émissions de CO_2 par kilowattheure (kWh)** d'électricité produite aux États-Unis est **0,367 kg CO_2 par kWh** environ (l'**U.S. Energy Information Administration (EIA)**, 2023)

Ca donne une estimation **18 t de CO_2 pour UN SEUL entraînement de GPT-3.**

L'entraînement optimisé de chatGPT-3 aurait généré 626 t de CO_2.

Pourqui ? … car on répète l'entraînement beaucoup de fois pour choisir les meilleurs valeurs d'hyperparamètres.

MINES PARIS | PSL

Strubell et al, (2019), https://arxiv.org/abs/1906.02243

A L'ÉCHELLE PLANÉTAIRE 2/3 ?

Aujourd'hui, on estime 700M d'utilisateurs quotidiens de GPT … (sans tenir compte d'autres modèles)

Chat GPT n'est pas le pire … la génération T2V *Sora* (OpenAI), *Veo* (DeepMind), *Vibes* (META) est bien plus énergivore encore.

Exemple [1] :

Génération de vidéo clip ~5s (81 images), 1 280 x 720 pixels	Energie
WAN2.1–T2V–14B (14 milliards de paramètres)	415 Wh
WAN2.1–T2V–1.3B (1,3 milliard de paramètres)	90 Wh

Consommation quadratique fonction du temps et de la résolution de la vidéo.

[1] Delavache, et al. 2025, https://arxiv.org/pdf/2509.19222

A L'ÉCHELLE PLANÉTAIRE 3/3 ?

40 Md tonnes CO2 ont été produites en 2024 à l'échelle planétaire.

L'impact de l'IA est estimé aujourd'hui à :

- 3 à 4 % des émissions de gaz à effet de serre (GES) dans le monde [1, 2], et
- 2,5 % de l'empreinte carbone de la France [3]

Dépassant les émissions du transport aérien (~ 2%).

Remarque:

Il n'y a pas que la consommation énergétique … il y a aussi consommation de l'eau (refroidissement).

[1] https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/
[2] https://theshiftproject.org, Pour Une sobriété Numérique (rapport 2025)
[3] ARCEP, Grands-Dossiers-Thematiques-Transverses (rapport 2025)

QUE PROPOSONS-NOUS ?

1. Permettre aux développeurs d'optimiser leur code
2. Optimiser l'exécution de grosses tâches de calcul sur les supercalculateurs

Thèse de Roblex Nana Tchakouté 2023-2025
(présentée le 5 décembre prochain)

une des premières thèses du TTI.5

Centres de recherche impliqués :

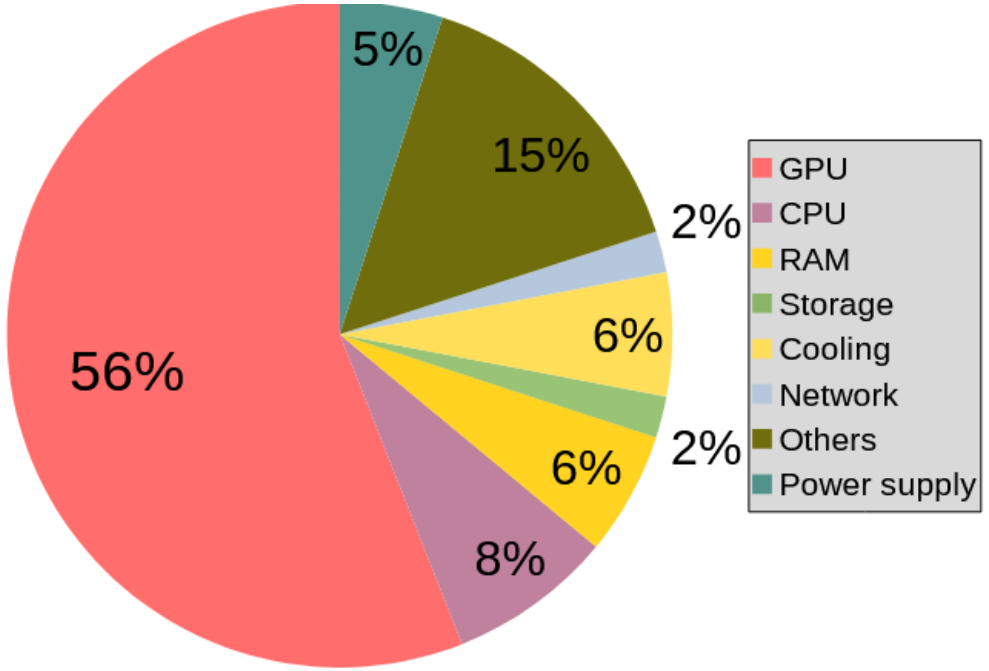
Claude Tadonki (CRI)

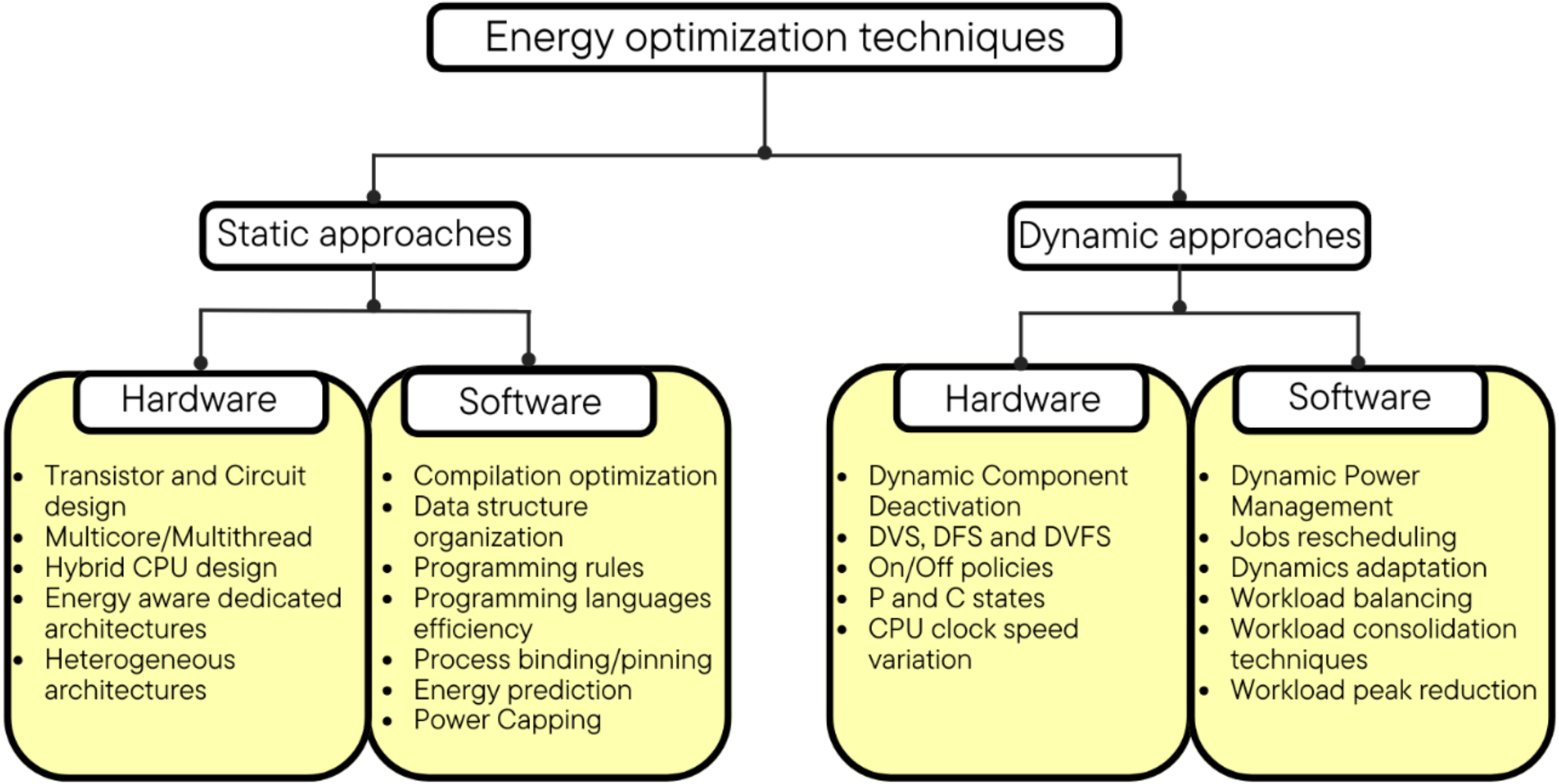
Youssef Mesri (CEMEF)

Petr Dokladal (CMM)

Objectifs : Réduire la consommation énergétique sans impacts sur les performances ni la dégradation du service.

CONSOMMATION D'UN NŒUD DE CALCUL : RÉPARTITION PAR COMPOSANT

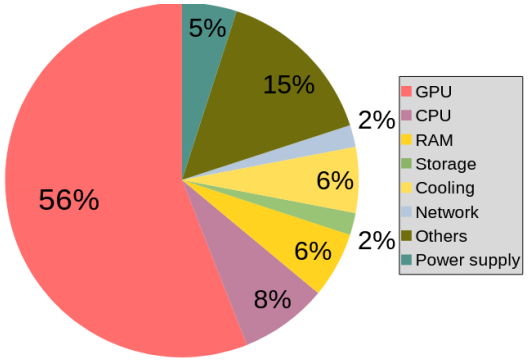


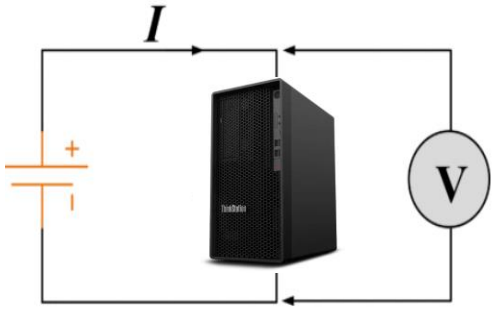

Energy optimization techniques
Static approaches
Dynamic approaches
Hardware
Transistor and Circuit design
Multicore/Multithread
Hybrid CPU design
Energy aware dedicated architectures
Heterogeneous architectures
Software
Compilation optimization
Data structure organization
Programming rules
Programming languages efficiency
Process binding/pinning
Energy prediction
Power Capping
Hardware
Dynamic Component Deactivation
DVS, DFS and DVFS
On/Off policies
P and C states
CPU clock speed variation
Software
Dynamic Power Management
Jobs rescheduling
Dynamics adaptation
Workload balancing
Workload consolidation techniques
Workload peak reduction

EA2P – NOUVEL OUTILS DE MESURE D'ENERGIE

On a besoin d'une mesure plus précise qu'un wattmètre :

1. Déclinée par composant
2. En résolution temporelle (~1ms)





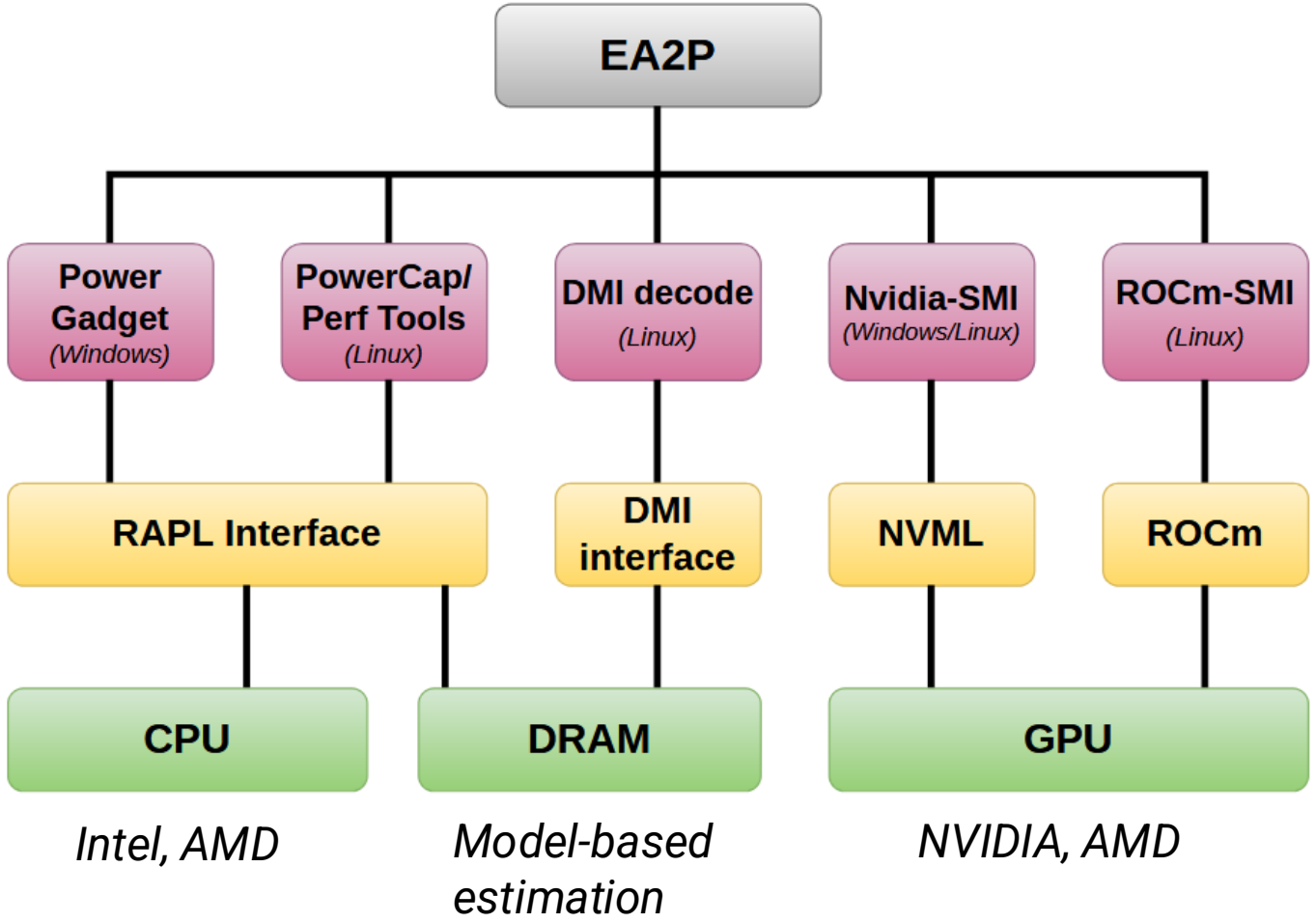
EA2P : Un profileur énergétique multiplateforme pour Python (applications IA)

- Un profileur énergétique multi-composant, mesurant la consommation du **CPU**, du **GPU** et **estimant celle de la RAM**.
- Basé sur Python, facile à intégrer dans les workflows **IA** ou **HPC**.
- S'appuie sur des **API matérielles existantes** (interfaces **RAPL**, **Linux Perf**, **Nvidia-SMI**, **ROCm-SMI**)
- Utilise un **modèle analytique** pour estimer la consommation énergétique de la RAM sur les plateformes ne disposant pas de capteurs dédiés.
- **Validé expérimentalement** pour sa **flexibilité**, sa **précision relative** et son **utilité** dans divers scénarios de profilage.

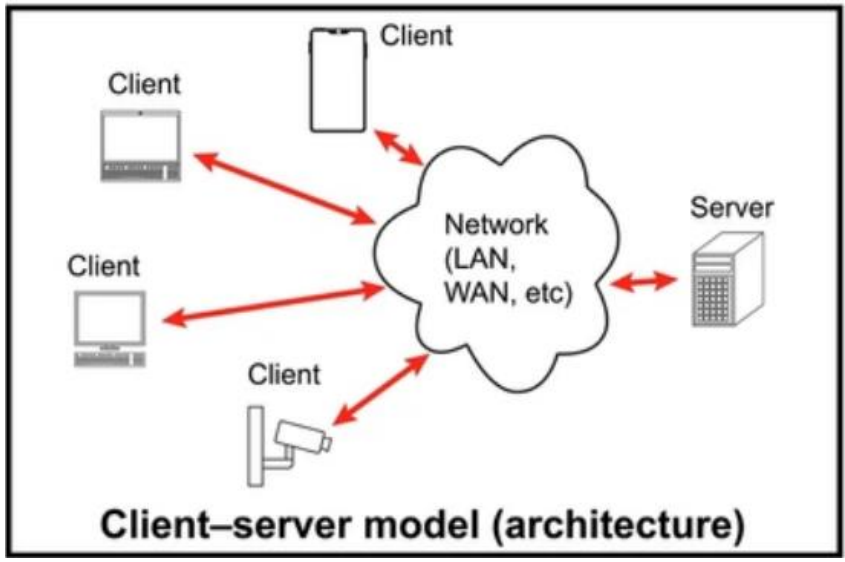
EA2P ARCHITECTURE

Forces :
- Agrégation des informations des différents composants
- Calcul hybride naturellement supporté
- Multi-processing naturellement supporté
- Déploiement facile

Limitations :
- Echantillonnage limité à ~1ms



STRATÉGIES D'OPTIMISATION ET DE GESTION DE COMPROMIS DANS UN DATA CENTER



Dans un data center, nombre de tâches entrent en concurrence. Historiquement, leur exécution est basée sur le principe Premier-Venu=Premier-Servi.

Avec l'avènement de l'IA de nouvelles spécificités apparaissent :

1. Hétérogénéité du HW (**avant** : CPU ; **nouvellement** : GPU)
2. Hétérogénéité de tâches
   avant : modélisation (phénomènes physiques – météo, mécanique, méca-fluide, …) ;
   nouvellement : tous types de data (texte, images, réseaux sociaux, modélisation ….)
3. Criticité mixed (par ex. : prédiction météo : toutes les 6h !!! en dépit d'un entraînement d'un nouveau modèle de fondation planifié

NOMBREUX CRITÈRES A SATISFAIRE 1/2

✅ **Performance**

Débit : Maximiser le nombre de tâches terminées par unité de temps.

Temps moyen de complétion (JCT) : Réduire le délai entre soumission et fin d'exécution.

⚠️ Souvent en tension avec l'équité et l'efficacité énergétique.

🌱 **Énergie**

Objectif : Réduire la consommation totale et l'empreinte carbone.

Indicateur : *Energy-Delay Product (EDP)* — équilibre entre performance et sobriété.

Exemples : GreenFlow, GREEN.

⚖️ **Équité**

Allocation équitable : Répartition des ressources selon les droits/priorités.

Équité d'expérience : Temps de complétion équitable entre utilisateurs.

Conflits fréquents entre équité et performance.

NOMBREUX CRITÈRES A SATISFAIRE 2/2

⏱️ **Qualité de Service (QoS)**

Objectif : Respect des délais critiques (SLAs).

Indicateur : Taux de dépassement de deadline.

⚠️ Peut réduire le débit global et l'équité.

🔄 **Flexibilité & Robustesse**

Flexibilité : Adaptation aux charges et imprévus.

Robustesse : Maintien des performances malgré les perturbations.

⚠️ Tension entre optimisation locale et résilience globale.

📏 **Stabilité**

Objectif : Planification prévisible et minimisation des interruptions.

⚠️ Trop de flexibilité nuit à la reproductibilité et à l'expérience utilisateur.

LEVIERS D'OPTIMISATION DU SCHEDULER

🧩 Hétérogénéité

Choix du matériel (ex. H100 vs A100) impacte performance, consommation et équité.

Décision critique : quel job sur quelle architecture ?

🔀 Malléabilité

Ajustement dynamique des ressources :

- **Nombre de GPUs (k)** : plus de GPUs = exécution plus rapide mais plus de consommation.
- **Taille de batch (b)** : meilleure efficacité GPU, mais contraintes mémoire et convergence.

Le scheduler choisit le meilleur combo *(k, b, h)* selon l'objectif.

⏸️ Préemption

Interruption d'un job peu prioritaire pour libérer des ressources. **Nécessaire** pour respecter les délais (SLAs).

⚠️ Coût important : perte de travail non sauvegardé → gaspillage énergétique. À utiliser avec parcimonie

EAS-SIM : SIMULATEUR POUR LE CO-DESIGN DES SCHEDULERS IA

🎯 Pourquoi un simulateur ?

Coût élevé : les tests réels mobilisent des milliers d'heures GPU.

Temps long : une semaine simulée = une semaine réelle.

Variabilité : les systèmes physiques sont non déterministes (bruit OS, réseau, température).

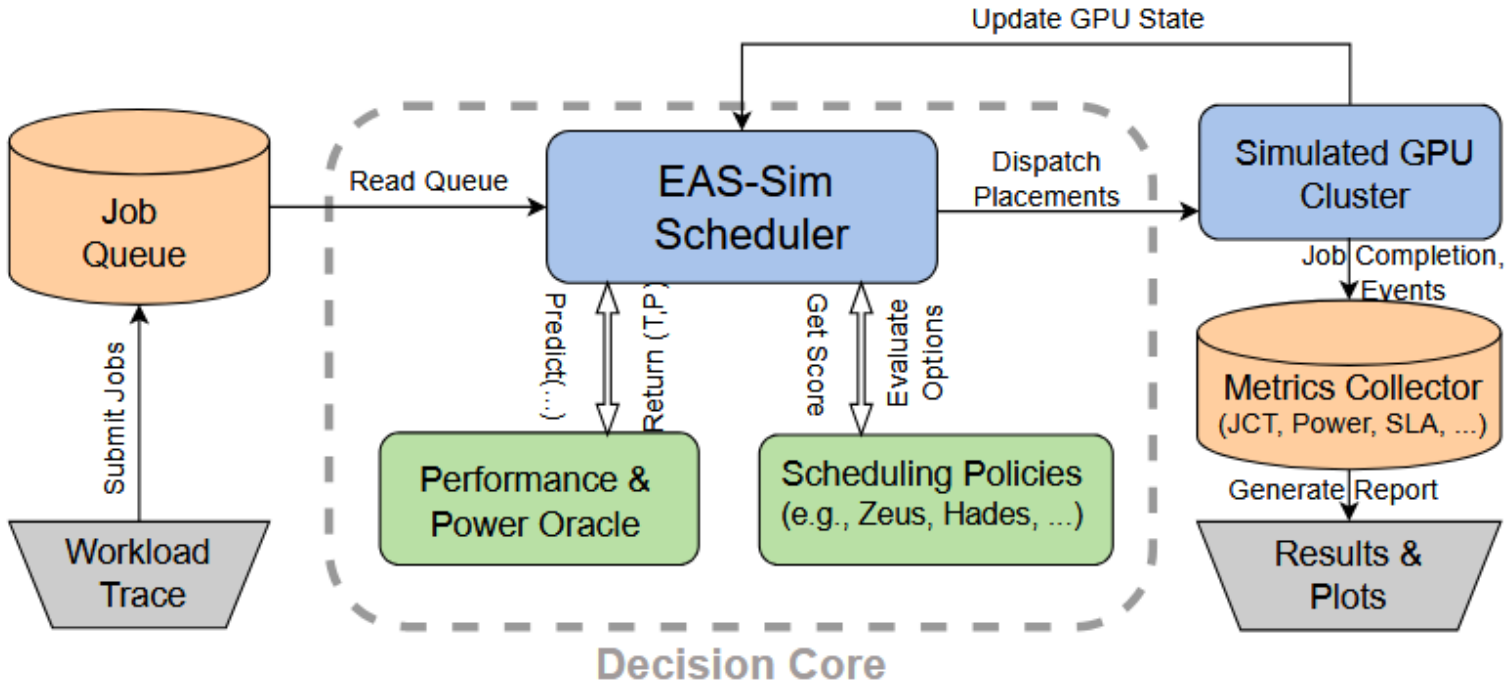


Figure : Architecture du simulateur Eas-Sim

🛠️ Solution : EAS-Sim

Simulation à événements discrets.

Rapide, extensible, haute fidélité.

Permet l'exploration efficace de politiques multi-objectifs (performance, énergie, équité…).

APPROCHES DE SCHEDULING EXISTANTES

⚡ **Pollux *(Performance)***

Objectif : maximiser le débit.

Heuristique : Shortest Job First (SJF) avec adaptation malleable.

Score basé sur le temps d'exécution estimé, ajusté par un facteur d'âge.

⚖️ **Themis *(Équité)***

Objectif : équité d'allocation entre utilisateurs.

Score modulé par la consommation historique (GPU-seconds).

Priorise les utilisateurs sous-consommateurs.

⏱️ **Chronos *(SLA / criticité mixte)***

Objectif : respecter les deadlines.

Priorité selon le temps restant avant échéance.

Jobs urgents classés séparément des jobs best-effort.

➡️ Ces heuristiques servent de base pour injecter la conscience énergétique dans les politiques de scheduling.

ENERGY-DELAY PRODUCT (EDP)

Définition : EDP = P × T² (P = puissance moyenne, T = durée d'exécution)

🎯 **Pourquoi l'EDP ?**

Pénalise les configurations :

- Trop rapides mais énergivores.
- Sobre mais trop lentes.

Favorise les compromis : **efficacité énergétique + performance**.

✅ **Avantages**

Réduction de la fragmentation du cluster.
Meilleur débit global.
Métrique robuste pour l'optimisation multi-objectifs.

POLITIQUES NOUVELLES DE SCHEDULING ÉNERGÉTIQUES

- **Zeus *(Performance + Énergie)***

  Version énergétique de Pollux.

  Minimise EDP pour chaque configuration malléable.
- **Hades *(Équité + Énergie)***

  Version énergétique de Themis.

  Score = EDP / (consommation historique + 1)
- **AuraChronos *(SLA + Énergie + Préemption)***

  Priorité aux jobs urgents (slack).

  Backfill optimisé via EDP (au lieu du temps brut).
- **Charon *(Contrainte de puissance)***

  Allocation conditionnée au respect d'un budget énergétique global (**Pcap**).

PROTOCOLE D'ÉVALUATION

Durée simulée : 10 heures de fonctionnement du cluster

Répétabilité : Moyenne sur 10 simulations indépendantes

Environnement : Cluster hétérogène avec 16 GPUs :

- 8 × A100 (400W, Ampere)
- 8 × H100 (700W, Hopper)

📦 Workload

3 modèles DL variés (arrivant en suivant le processus Poissonien $\lambda = 0.005$ jobs/s) :

- Vision Transformer (ViT-Base, FP32)
- ALBERT v2 (bfloat16)
- ResNet-50 (FP32)

RESULTATS : UNE NOUVELLE FRONTIÈRE ÉNERGIE-PERFORMANCE

⚡ **Zeus vs Pollux**

Pollux : 10.9 jobs/h pour ≈57 kWh

Zeus : 11.0 jobs/h pour ≈51 kWh

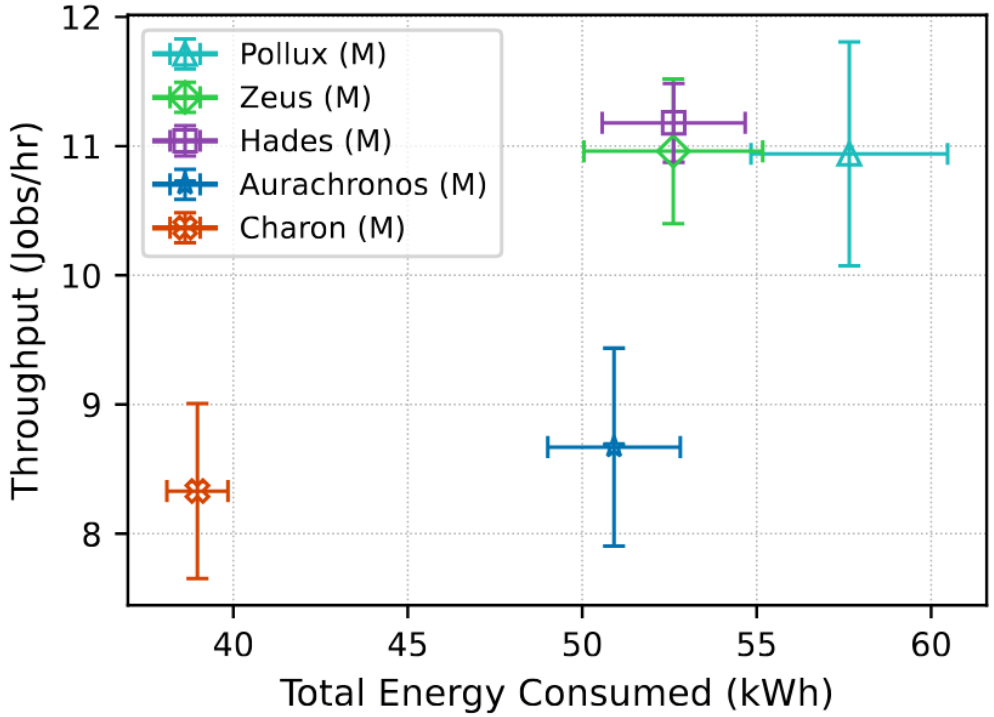
➡️ **−10.5% d'énergie consommée**, sans perte de performance

🧠 **Pourquoi ?**

Zeus évite les configurations trop énergivores (ex. H100 inutiles)

Meilleur "eco-routing" des jobs → moins de fragmentation

F**rontière de Pareto** repoussée : meilleure performance + sobriété



BILAN DE LA THÈSE

🎯 **Objectif**

Optimiser l'**efficacité énergétique** des applications IA sur infrastructures modernes.

🧰 **Contributions majeures**

EA2P : outil de profilage énergétique multi-plateforme.

Modèles analytiques : prédiction fine de la consommation (≈4% d'erreur).

EAS-Sim : simulateur haute fidélité pour le co-design de politiques de scheduling.

Politiques multi-objectifs : Zeus, Hades, AuraChronos, Charon.

✅ **Résultats clés**

−10% d'énergie avec Zeus, sans perte de performance.

AuraChronos : SLA garanti, 20× moins de deadlines manquées.

EDP : métrique robuste pour concilier performance et sobriété.

PERSPECTIVES DE RECHERCHE

🔧 Amélioration des modèles

Intégrer les interférences entre jobs.

Calibration dynamique en ligne.

🌐 Co-optimisation globale

Coordination entre OS, hyperparamètres et techniques de compression.

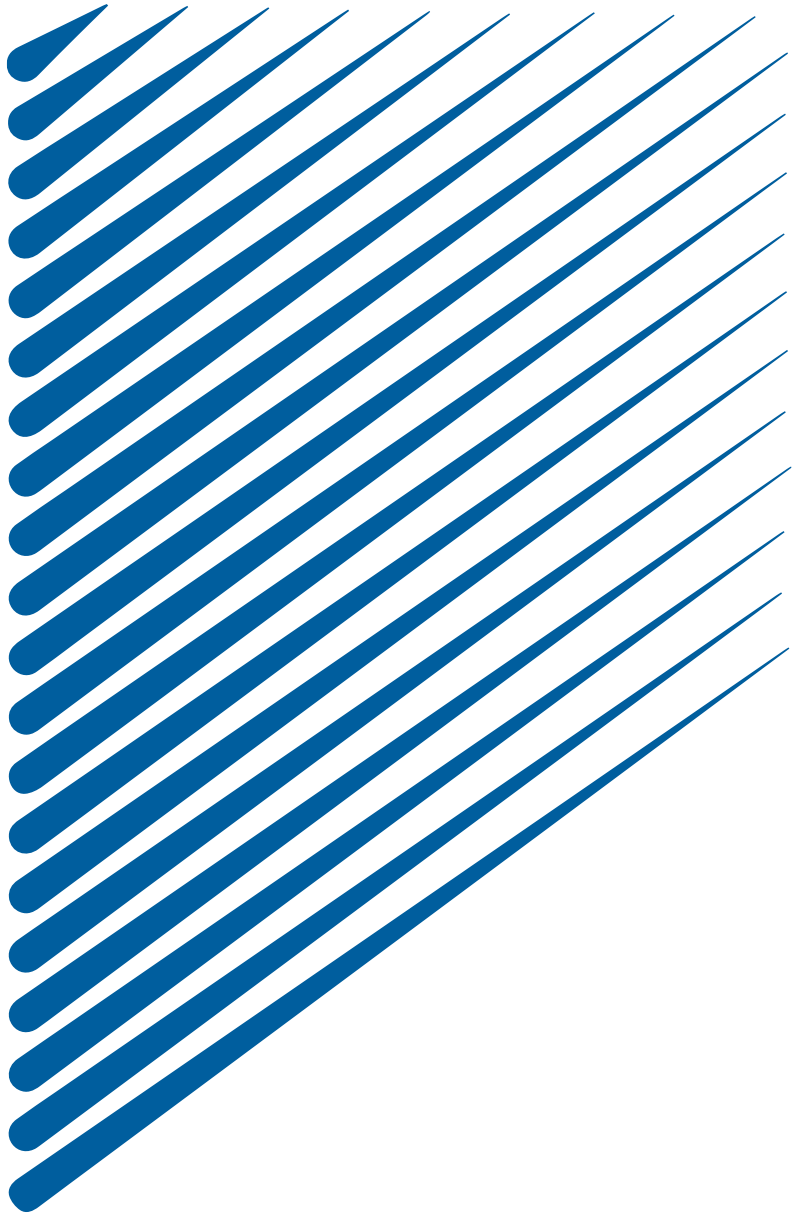
🌍 Extension au continuum Edge–HPC–Cloud

Prise en compte des contraintes réseau, batterie, thermique.

🧠 IA pour Green AI

Utiliser l'apprentissage par renforcement pour des politiques adaptatives.

➡️ Une approche rigoureuse et intégrée pour une IA durable et responsable.



MINES PARIS | PSL

Merci de votre attention

Petr.Dokladal@minesparis.psl.eu