Paris le 27 janvier 2026

WHITE PAPER

Construire une Intelligence Artificielle frugale : une trajectoire stratégique

Synthèse

La frugalité cible, tout en maintenant les performances et la satisfaction des besoins, la minimisation de l'empreinte environnementale d'un système ou d'une application, donc la minimisation de l'énergie et des matériels utilisés.

La frugalité va s'imposer comme un impératif pour l'Intelligence Artificielle non seulement via des contraintes de ressources énergétiques et d'impact environnemental, mais aussi, et c'est lié, tout simplement pour des raisons économiques.

Les grands modèles LLM, le plus souvent développés par des géants du numérique, existent mais ne sont plus l'enjeu stratégique déterminant. La bataille stratégique se joue désormais sur l'inférence, via des modèles plus petits et spécialisés selon les applications ou usages, en d'autres termes les métiers, y compris par exemple pour les objets connectés ou les systèmes embarqués.

La frugalité, incontournable, ne pourra résulter que d'une co-conception à double volet « matériel/logiciel », et même à triple volet en combinant les ressources humaines : leur intelligence et leur connaissance non seulement des technologies mais aussi des métiers, bref leur intelligence, bien réelle, est une nécessité première.

Qu'il s'agisse de survie économique des entreprises, de souveraineté des pays, ou d'éco-géostratégie européenne, la frugalité de l'IA n'est pas seulement un choix écologique ; c'est le fondement énergétique de toute compétitivité au siècle de l'IA.

Les acteurs n'auront pas le choix. Ils seront frugaux ou seront dépassés : ils devront s'imposer de définir leur trajectoire de frugalité, continûment évolutive, sous forme de feuille de route, c'est-à-dire de *trajectoire*.

Table des matières

1ère partie : LE TERRAIN DE JEU

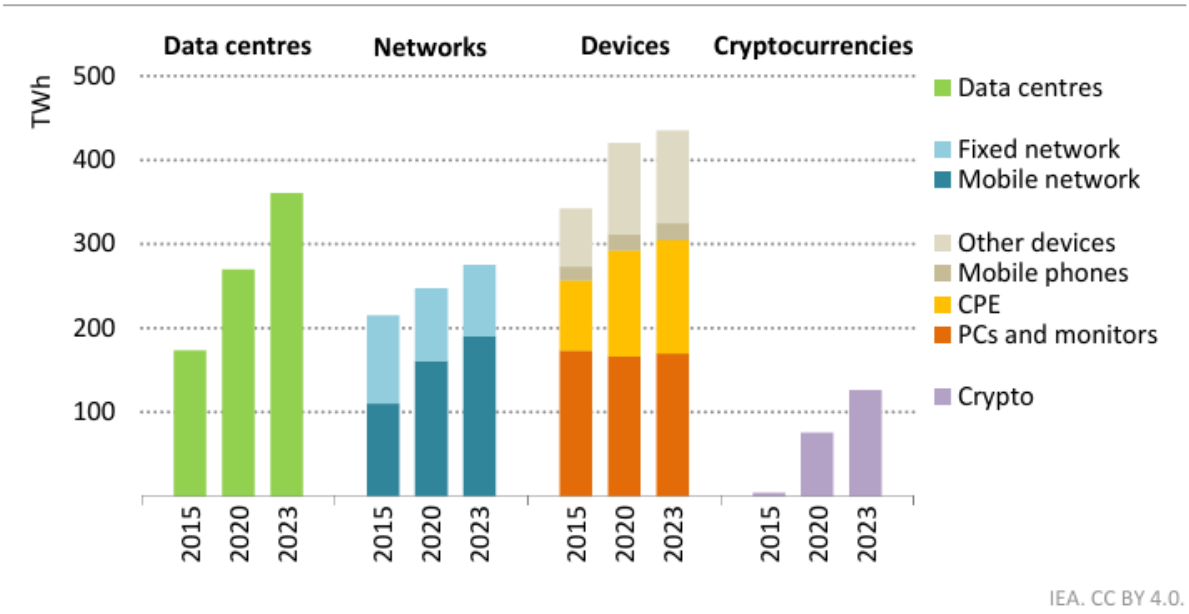
1 Consommations énergétiques

La frugalité cible, tout en maintenant les performances et la satisfaction des besoins, la minimisation de l'empreinte environnementale d'un système ou d'une application, donc la minimisation de l'énergie et des matériels utilisés : des kWh consommés, et des émissions de GES qui en découlent, mais aussi des matériels, dont la multiplication a un impact environnemental, notamment via l'extraction des matières premières.

L'Agence Internationale de l'Energie (IEA) publie divers rapports permettant d'appréhender des agrégats intéressants de consommations énergétiques, et d'estimer les tendances, en particulier en 2025 des travaux consacrés à l'IA [1]. L'Energy Institute publie chaque année sa Statistical Review of World Energy [2]. Nous pouvons ainsi établir la vue suivante.

L'énergie primaire mondiale en 2024 est de 592 Exajoules, soit environ 164 000 TWh, en croissance annuelle de 2%. La production d'électricité a été de l'ordre de 31 000 TWh, en croissance annuelle de 4%, correspondant à 19% de l'énergie primaire.[2]

Le secteur des technologies de l'information et de la communication (ICT) consomme de l'ordre de 1000 à 1100 TWh, soit moins de 4% de la production mondiale d'électricité. La répartition entre les composantes respectives des technologies ICT est celle de la figure 1 ci-dessous [1].



Energy use by data centres and cryptocurrencies have risen sharply since 2020, while devices and networks have seen slower growth

Notes: CPE = customer premises equipment, including routers and modems; PCs = personal computers, including laptops and desktops. Networks include core and access networks. Other devices include the Internet of Things and surveillance cameras.

Sources: IEA analysis based on data from Malmodin and Lundén (2018); IEA (2023); GSMA (2024) World Bank, (2024b); Malmodin, et al. (2024); Kamiya and Coroamă (2025); Cambridge Centre for Alternative Finance (2025), and company reports.

Figure 1 : Consommation électrique mondiale des data centers, réseaux, terminaux et minage de crypto-monnaies. 2015-2023

La consommation ne comprend pas, par définition, l'énergie consommée pour la fabrication des matériels. Cette énergie pour la fabrication est du même ordre de grandeur que celle de la consommation. C'est une donnée utile, même si elle reste un ordre de grandeur, car elle éclaire toute stratégie de frugalité. Il est stratégiquement aussi important de minimiser la consommation électrique que la multiplication des matériels ; c'est au niveau opérationnel que pourront être arbitrés les choix d'architecture, en regard des nombres d'équipements, de leur coût et de leur consommation d'énergie. Nous verrons que c'est un élément essentiel de l'approche de la frugalité, et de surcroit un terrain de jeu pour l'innovation.

La tendance croissante de la consommation d'électricité devrait se poursuivre, voire s'amplifier sur la période à venir jusqu'en 2030, les deux composantes principales étant les data centers et les terminaux.

S'agissant des data centers, déjà à 414 MWh en 2024, la croissance moyenne a été de l'ordre de 12% par an sur les 5 dernières années, et l'AIE anticipe une demande électrique doublant de 2025 à 2030 pour atteindre jusqu'à 900 à 1000 TWh par an, du fait de l'IA (scénario "IA partout"), ce qui correspondrait à 3% de l'électricité mondiale.

S'agissant des terminaux, la tendance va directement dépendre de la multiplication des terminaux, liée notamment à de nouveaux usages, via l'extension voire la généralisation de l'internet des objets (IoT). On estime d'ores et déjà à plusieurs dizaines de milliards le nombre d'objets connectés. Selon la nomenclature usuelle, tous ne relèvent pas du secteur ICT, rendant difficile la mise en cohérence des agrégats statistiques de l'IoT, du secteur ICT, et des autres secteurs. La difficulté subsiste, voire s'amplifie lorsqu'on traite l'IA. Les hypothèses examinées dans la littérature sont d'une grande variabilité ; qui plus est une part significative des consommations, gaspillée, perdue, pourrait être économisée par des procédures d'optimisation. Pour ce qui nous intéresse ici, en traitant de frugalité, nous considérerons que la tendance plausible, cohérente, est un doublement de la consommation pour l'agrégat terminaux, à l'instar de la tendance pour les data centers.

Ceci conduit à maintenir les deux composantes essentielles de consommation que sont les data centers et les terminaux. Les réseaux voient une progression moindre, la consommation n'étant pas proportionnelle à la croissance du trafic. A l'horizon 2030, la croissance par rapport à 2025 (300 TWh) devrait, en l'état des anticipations, être dans une fourchette de +20 à +40%.

Toutefois ceci ne tient pas compte de ce que pourrait être un IoT généralisé qu'on qualifierait d'"Internet of Everything". Dans une telle configuration, les architectures des systèmes vont nécessairement évoluer, vers des traitements locaux et des objets non pas autonomes mais coopératifs. Ceci imposerait d'adopter un protocole standard d'interconnexion à haut débit, et faible latence (exemple CXL, Compute Express Link, qui a recueilli l'héritage de l'ancien protocole Gen-Z).

Nous ne cherchons pas ici, parmi les consommations du secteur ICT, à identifier celles qui résulteraient directement de l'IA. De multiples commentaires, dont ceux de l'IEA, mettent en avant les grandes incertitudes d'une telle velléité ; en tendance, l'IA devient de plus en plus omniprésente, avec un effet d'éviction d'anciens usages, rendant peu significative toute analyse comparative discriminant l'IA dans les consommations du secteur ICT. Enfin, et c'est un argument essentiel sur lequel nous reviendrons au niveau même des entreprises, la consommation dépendra notamment des ressources disponibles. Sous l'effet de multiples contraintes, les outils et l'électricité disponibles seront en quantité limitée, tout comme les ressources financières. Cette limitation interagit avec les usages de l'IA, y compris la prise en compte de la frugalité.

2 L'Intelligence artificielle

Nous proposons ici, dans une perspective d'approche de la frugalité, une vue sur l'IA et les LLM. Elle reprend quelques éléments du rapport annuel "*HAI AI Index report"* de l'Université de Standford [3], reconnu pour son acuité, et des éléments d'actualité dont le rythme des évolutions est très soutenu.

LLM et training

Sur la seule année 2024, plus d'un nouveau LLM ou une nouvelle version de LLM considéré comme notablement performant a été en moyenne publié chaque mois ; et trois datasets ou autres types de modèles considérés comme des avancées notables ont été en moyenne distribués tous les deux mois, allant d'une banque de centaines de millions de prompts pour le LLM multilingue Aya, à des modèles de "text to image/video", "text to song" ou encore "vision-language".

2024, et plus encore 2025, est une année où l'IA rattrape l'humain dans presque tous les indicateurs de performance.

Notre intention ici n'est pas d'analyser en détail ce panorama, - une telle analyse n'aurait d'intérêt qu'en étant ciblée en regard de cas d'usages déterminés - mais de rappeler que l'IA ne saurait être une option pour les entreprises. C'est l'état des lieux dans lequel elles ont à exercer leur activité. Et celui-ci évolue vite. Ainsi, et cela pourrait se révéler être un levier majeur dans l'approche frugale, on assiste à une inflexion de tendance dans la taille des modèles. Alors qu'auparavant l'amélioration des performances était obtenue par l'augmentation de la taille des LLM et des bases de données d'entraînement (training), dans une sorte de course au gigantisme, 2024 a été un premier virage vers de plus petits modèles néanmoins très performants. Entre 2023 et 2025, on a observé des réductions de taille par 10 chaque année, à performances maintenues. L'asymptote de la tendance est logiquement indéterminée, et surtout dépendra des usages.

Des LLM compacts progressent, tels Llama 3.2 (90 B) et Llama 3.3 (70 B) par rapport à Llama 3.1 (405 B) chez Meta, ou Mistral-7B, qui avec 7 milliards de paramètres tient la comparaison avec des modèles dix fois plus larges, pour ne citer que deux exemples parmi bien d'autres connus ou à venir.

Les leviers techniques, multiples, pour réduire la taille des LLM sont notamment les suivants :

- la suppression de paramètres (élagage), à performances maintenues,
- la distillation des connaissances, dans laquelle on entraîne un modèle compact à reproduire le comportement du modèle plus gros,
- bien sûr les optimisations tant des architectures que des jeux de données d'entraînement,
- et la quantification. Avec des poids du modèle non plus de 32 bits, mais de 8 et même 4 bits, la taille du modèle et le temps calcul sont réduits. C'est une option qui peut s'avérer incontournable pour l'IA embarquée dans des terminaux.

En termes stratégiques, une phase est en train de s'achever. Celle de la croissance des modèles de training, le plus gros étant réputé devoir être le plus performant. Au plan matériel, il a conduit à la position dominante, quasi monopolistique des processeurs graphiques de NVIDIA.

Les grands acteurs, Google en tête mais les autres aussi (Grok d'Elon Musk, Amazon…) se lancent dans un "co-design" hardware-software, devant les libérer du "lock-in" avec NVidia.

La perspective la plus probable est que les grands modèles subsistent, au moins à terme rapproché. On annonce aux Etats-Unis des mega datacenters à 10 GWatts, et les premières initiatives de datacenters orbitaux émergent, motivées par une énergie solaire illimitée et un refroidissement naturel. Mais, hors le volet énergétique, là ne sera plus le champ d'innovations majeures.

La nouvelle phase qui s'ouvre est celle de modèles plus petits, qui permettront d'obtenir des résultats comparables ou suffisants à moindre coût, en étant frugaux. On devrait assister à une "explosion cambrienne" touchant les LLM et le training mais aussi et surtout l'inférence.

Au plan économique, on peut appréhender la situation en considérant que quelques géants ont consacré des ressources considérables pour créer, entraîner, leurs grands modèles LLM : c'est leur « investissement ». Mais ce ne sont pas les activités de « training » qui génèrent l'essentiel de leurs profits ; ce sont celles du domaine de l'inférence, des applications et usages, pour lesquelles certes leur investissement est d'évidence un atout, mais le jeu reste ouvert. C'est sur l'inférence que se joue, que doit se jouer la stratégie des entreprises.

<u>Inférence</u>

Côté inférence, une caractéristique de plus en plus prégnante est l'inférence en temps réel, à l'appui de décisions prises dans une logique d'immédiateté (typiquement avec une échelle de temps en millisecondes). Les applications telles que les aides à la conduite voire les véhicules autonomes, certains systèmes industriels, sont des exemples manifestes, parmi bien d'autres.

L’inférence à grande échelle voit ses coûts diminuer, mais ceux-ci restent importants. La frugalité devient une nécessité pour le passage à l'échelle (scaling), sans quoi les coûts trop importants ne seront pas compétitifs.

Les entreprises doivent donc évaluer le coût total d’usage (CTU), et arbitrer celui-ci en regard de la performance (qualité, latences etc.), et du budget possible, fonction lui-même du retour sur investissement.

La réduction de la taille des LLM ouvre la voie de l’amélioration de l’efficience et de la rapidité de l’inférence, via la quantification en 8 ou 4 bits, qui permet l’usage sur les terminaux avec de faibles latences, et les autres leviers précités pour les LLM compacts.

Le “co-design” hardware-software est un levier essentiel, comme nous le verrons : l’inférence photonique, ou sur des chipsets analogiques, par exemple, voit le jour. Mais l’optimisation de l’inférence reste un “réglage” adaptatif de la solution à déployer (architecture matérielle, logiciels, données, processus, etc.) en regard du besoin.

Sur le terrain que nous venons de décrire, qui est celui très rapidement évolutif dans lequel tout se joue, la frugalité n’est pas seulement une qualité souhaitable ; elle est une contrainte nécessaire, à toutes les échelles, pour que les usages restent abordables et compétitifs !

Nous l’exposerons sur quelques exemples applicatifs.

2ème partie : LES ENTREPRISES N’ONT PAS LE CHOIX

3 Intégration de l’IA

Toutes les entreprises, dans tous les secteurs, sont désormais confrontées à une exigence incontournable : intégrer l’IA.

L’IA optimise :

- la conception,
- la production,
- la maintenance,
- les opérations,
- la relation client,
- le commercial
- le pilotage.

et la gestion. L'entreprise qui n'utilise pas l'IA en interne perd en efficacité et en coût.

Les offres doivent intégrer l'IA :

- pour rester attractives,
- pour rester dans la course à l'innovation,
- pour être plus performantes
- pour anticiper les tendances
- pour répondre aux attentes des clients.

Une entreprise qui n'intègre pas l'IA dans ses produits se fait doubler, parfois en un seul cycle.

Une rupture est attendue : celle de l'IA embarquée et de l'explosion de l'internet des objets, qui verra l'IA quitter les data centers pour entrer dans les objets, les machines, les robots, les véhicules, les infrastructures. Des milliards d'objets pourraient devenir coopératifs.

Se produira là une expansion rapide ("explosion") des inférences… et de la valeur

La frugalité devient indispensable. Elle permet :

- des performances équivalentes ou meilleures,
- une baisse des coûts matériels et énergétiques,
- une réduction de l'empreinte environnementale,
- une meilleure robustesse,
- et une adoption plus large.

La frugalité ne doit pas être vue comme une restriction, mais comme un avantage compétitif dans un monde où l'énergie devient rare, et finalement un impératif.

Le train du training massif est passé, comme nous l'avons évoqué au §2.

En dehors des hyperscalers, la bataille du training massif est dépassée et la stratégie d'entreprise consiste à choisir :

- les bons modèles,
- les bonnes plateformes,
- les bons outils,
- les bons contrats cloud.

Un espace reste ouvert, et il est de portée considérable, le training embarqué, léger ou spécialisé pour :

- robots,
- véhicules,
- équipements industriels,
- micro-nœuds de calcul.

C'est un champ où l'Europe peut innover : modèles efficaces, co-design matériel/logiciel, optimisation extrême.

L'inférence devient le centre du jeu. Alors que l'entraînement est ponctuel, l'inférence est continue, scalable, directement liée aux usages métier. C'est dans l'inférence que se créent les marges, la valeur et les gains de productivité.

Les entreprises peuvent agir sur :

- la taille des modèles,
- leur efficacité,
- les techniques de quantification,
- la localisation des calculs,
- la réduction des flux,
- l'optimisation énergétique.

Ce sont des leviers accessibles, pragmatiques, rapides.

L'embarqué et "l'Internet of Everything" apparaissent comme un cas emblématique :

- Usages massifs,
- Architectures encore non figées,
- Besoin de faibles latences,
- Enjeux énergétiques majeurs,
- Explosion prévue des objets intelligents.

La bataille **n'est pas encore jouée** — et l'Europe et ses entreprises peuvent, doivent, s'y positionner.

Avant de traiter, en 3ème partie, de stratégies d'entreprise, nous souhaitons illustrer l'approche et l'appréhension de la frugalité, par trois exemples, des cas d'usage rencontrés de dimensions respectives différentes, de l'hyper groupe international à la PME start-up.

4 Exemple de cas d'usage dans un grand groupe de la tech

Le contexte est celui de Google qui n'arrive plus à poursuivre son "scaling" juste en achetant et utilisant de plus en plus de GPU : les MW disponibles et le coût de cette énergie deviennent une limitation. La stratégie suivie est alors une feuille de route technologique " full stack" notamment suivant trois grands axes:

- Des chipsets maison : les TPUs (Tensor Processing Unit), et maintenant les CPU Axion, de technologie Arm,
- Des data centers optimisés qui atteignent un PUE (Power Usage Effectiveness) de l'ordre de 1,1 et même moins, ce qui est une grande performance,
- Une Orchestration "carbon-aware" des workloads.

Plus précisément, voici la déclinaison.

Hardware

- **TPU v4 → v6 → Ironwood** : Google publie des gains massifs de performance par Watt entre générations ; TPU v4 annonce ~2,7× de gain de performance par Watt par rapport à v3, dans un stack combinant des chips plus efficaces, une meilleure PUE et une énergie plus "clean",

- **Ironwood (TPU v7)** est explicitement présenté comme doublement plus efficace que la génération précédente (Trillium / TPU v6) en performance par Watt, et ~30× plus efficace que la première génération,
- **Axion CPU** : le CPU Arm maison pour le cloud, annoncé avec +50 % de performance et +60 % d'efficacité énergétique par rapport aux x86 classiques pour les mêmes workloads,
- **Data centers** : la PUE moyenne publiée est autour de 1,09–1,12 sur l'ensemble de la flotte des data centers.

Software

- **IA pour le refroidissement** : des réseaux de neurones sont utilisés pour piloter le refroidissement et optimiser la PUE, avec des modèles capables de prédire la PUE à ±0,4 % près et d'ajuster en continu les consignes,
- **Schedulers "carbon-aware"** : Google décale dans le temps et l'espace les tâches flexibles vers les régions et les horaires où le mix électrique est le plus décarboné, via son système de *Carbon-Intelligent Compute Management* et ses API de scheduling.

Co-design hardware + software

- **TPU ⟷ TensorFlow / XLA / JAX :** les TPUs sont co-conçus avec le stack IA ; l'ISA (Instruction Set Architecture), les unités matricielles et la mémoire HBM (Mémoire à large bande passante) sont choisies en fonction des patterns de calcul des graphes TensorFlow, puis le compilateur XLA/JAX émet du code optimisé pour ces blocs (et inversement, les modèles sont adaptés vers des formes compatibles pour les multiplications matricielles),
- **Orchestration → conception DC** : les algorithmes de scheduling "carbon-aware" fixent les profils de charge, ce qui contraint la conception des data centers (refroidissement, distribution électrique, redondance). Inversement, les contraintes physiques (PUE, capacité par site) se reflètent dans la logique de scheduling et de placement des jobs.

Ces éléments techniques montrent que même un géant comme Google est contraint de suivre un co-design vertical (chips + data centers + Operating Systems + Scheduler) pour optimiser sa consommation et l'emploi des ressources énergétiques accessibles. La frugalité n'est pas une caractéristique souhaitable ("nice to have"), mais une condition nécessaire de scalabilité économique.

5 Exemple de cas d'usage chez Graphcore, PME britannique

Le contexte est celui de Graphcore, société "moyenne" (activité semiconducteurs pour l'IA, centaines d'employés, centaines de millions de revenus ou valorisation) qui existe et prospère grâce à son approche de co-design hardware /software pour l'efficacité.

Hardware.

L'IPU (Intelligence Processing Unit) est un processeur massivement parallèle, avec beaucoup de mémoire très proche du compute et un réseau interne très fin, conçu pour limiter les déplacements de données et maximiser l'utilisation des unités de calcul.

L'architecture réseau (IPU-Links, fabric interne) est pensée pour connecter des centaines / milliers d'IPU avec une latence faible et un bon ratio performance/W à l'échelle d'un data center.

Software

Poplar SDK : stack logiciel complet, co-conçue avec l'IPU, qui représente les modèles comme des graphes et gère la répartition fine des nœuds sur les cœurs, la mémoire, et les communications,

Le compilateur Poplar optimise les mouvements de données, la parallélisation multi-IPU et l'utilisation mémoire pour limiter les goulots d'étranglement énergivores (off-chip traffic).

Co-design hardware + software

Graphcore affirme que l'IPU et Poplar ont été co-conçus "from the ground up" pour le machine learning, et non "adaptés après coup" comme un GPU généraliste,

La structure de l'IPU (mémoire très fragmentée, grand nombre de tuiles (tiles)) n'est exploitable que parce que Poplar sait mapper automatiquement un graphe sur cette topologie ; et Poplar, en retour, impose des contraintes sur le design futur des IPU.

Une société "moyenne", qui n'a pas leurs moyens financiers, n'a aucune chance de rivaliser en force brute avec les géants. Son atout, et c'est en fait le seul, est le co-design algorithmique/architectural pour gagner en performance/Watt. La frugalité est ici un *avantage compétitif*, et pas du "greenwashing" marketing d'apparence !

6 Exemple de cas d'usage chez Green Waves technologies, TPE française

Le contexte est celui de Green Waves Technologies, petite société française qui fait du co-design extrême : le budget énergie de ses produits se compte en milliwatts, pas en mégawatts.

Hardware

GreenWaves (Grenoble) conçoit des processeurs RISC-V GAP8 / GAP9 pour l'IoT et les appareils hearables, focalisés sur l'ultra-low power : jusqu'à ~50 GOPS pour ~50 mW sur GAP9 selon la presse spécialisée,

L'architecture intègre un cluster de cœurs + accélération DSP /NN (Digital Signal Processor /Neural Network), avec des choix de process (GlobalFoundries 22FDX, etc.) orientés efficacité énergétique plutôt que fréquence maximale.

Software

GAP9 est associé à un flow logiciel complet : toolchain RISC-V, librairie DSP, toolchains de réseaux de neurones, etc., pour optimiser à la fois les filtres audios et les NNs sur cette architecture spécifique,

Les démonstrations (hearables, drones, vision embarquée) exploitent des kernels soigneusement optimisés pour rester dans quelques mW tout en offrant les fonctions de suppression de bruit, NN audio, vision, etc.

Co-design hardware +software

GreenWaves met en avant un design homogène où DSP, NN et cœur RISC-V sont pensés ensemble avec un *flow* logiciel intégré pour la co-conception,

Dans des cas comme GAP9Shield (nano-drones), la plateforme combine module hardware + stack logiciel optimisée pour gérer vision + ranging sous une enveloppe énergétique minuscule, typiquement celle d'une batterie très limitée.

Une petite entreprise, qui plus est startup, ne réussit que grâce à une co-conception serrée où chaque picojoule compte. La frugalité là est une condition d'existence même du produit, et pas un "bonus" du domaine de la RSE.

3ème partie : CONSOMMATION ÉNERGÉTIQUE ET STRATÉGIES

7 Frugalité dynamique

Les données de la 1ère partie et les exemples de la 2ème partie montrent que la consommation énergétique est pour l'IA dimensionnante et structurante.

Les usages géants et génériques doivent impérativement maîtriser leur consommation. Les applications intermédiaires doivent faire mieux que les génériques, en performance (typiquement FLOP/Watt), et donc cultiver la frugalité. Les usages de type IoT ou de faible empreinte doivent tenir dans leur contrainte énergétique, qui impose la frugalité.

Dans tous les cas, le passage à l'échelle ne tient économiquement que sous la contrainte énergétique, dont la tenue passe par la frugalité grâce, autant que faire se peut, au co-design hardware-software.

La frugalité ne doit pas être vue comme un point de fonctionnement, comme un état final à atteindre - une machine économe, un code efficient, pour moins consommer - mais comme une trajectoire, une feuille de route d'optimisation, vers un horizon qui reste toujours devant lorsqu'on progresse !

Par une veille suivie d'une co-conception hardware-software, un acteur obtient :

- une *accumulation* d'efficiences génération après génération,
- une *capacité d'adaptation* à chaque rupture technologique,
- une *résilience* face aux chocs énergétiques et économiques,
- une *barrière à l'entrée* impossible à combler pour les concurrents.

Chaque génération apprend de la précédente. Ce sont les boucles d'optimisation permanentes qui assurent la pérennité compétitive.

En fait, « Les acteurs ne sont pas en compétition sur des produits, ils sont en compétition sur des trajectoires d'optimisation. »

Les gagnants ne sont pas ceux qui ont :

- le meilleur chip,
- ou le meilleur compiler,
- ou le meilleur datacenter,

- ou les meilleurs algorithmes et codes.

Les gagnants sont ceux qui ont la meilleure trajectoire couplée :

- un pipeline d'optimisation hardware,
- un pipeline d'optimisation software,
- un pipeline d'optimisation architecturale,

qui fonctionnent en *boucle* d'amélioration permanente.

La frugalité devient une *force d'évolution* comparable à la sélection naturelle.

Ce n'est pas une "caractéristique". C'est la pression environnementale fondamentale du High Performance Computing et de l'IA modernes.

8 Comment faire : à l'échelle de l'entreprise

Chaque métier doit apprendre à mesurer, anticiper et réduire le coût énergétique total de ses choix techniques, car la consommation d'énergie devient un *coût direct*, un *risque opérationnel*, une *barrière à la scalabilité* et un *critère de compétitivité*.

a. Appréhender le coût complet

- CapEx (serveurs, stockage)
- OpEx (support, licences)

mais aussi :

- **Coût énergétique du modèle** (€/token, €/inférence, €/simulation)
- **Coût énergétique de la donnée** (transport, pré-processing, stockage)
- **Coût énergétique du cycle de vie** (réentraînement, réindexation).

b. Choisir intelligemment

Voici quelques exemples indicatifs :

- Banque : choisir un LLM 70B quand un modèle 3B suffirait = 20× de coût énergétique perdu,
- Industrie : choisir une simulation trop fine, ou mal maillée, ou mal couplée au solver = multiplication par 3 ou 10 de l'énergie consommée, pour rien,
- Santé : pipeline d'imagerie médicale : réduire la taille des batches, compresser les tensors, et adapter les convolutions au hardware = 2× à 4× d'économie.

c. Encourager des solutions "locales" ou spécialisées

- un modèle compressé, distillé, quantifié,
- un code adapté à la géométrie du hardware,
- un traitement distribué localisé,

- ou une architecture sure,

peut **diviser les coûts par 10**. Le *tuning* rejoint la stratégie métier.

d. La maîtrise de l'énergie est un avantage compétitif durable

au même titre que :

- l'efficacité organisationnelle,
- l'innovation produit,
- l'excellence opérationnelle.

Les exemples suivants sont démonstratifs.

- Scalabilité interne d'un cluster IA : Sans frugalité, augmenter la capacité de 3× implique 3× de serveurs, 3× d'énergie, 3× de cooling, 3× de support, 3× de facturation cloud. La frugalité permet une **scalabilité sub-linéaire** : +3× de capacité pour +1.4× de coûts,
- Avantage humain interne. Une équipe qui maîtrise le tuning hardware (choix du hardware, compréhension du silicium), le tuning software (optimisation codes) et le co-design (architecture + algorithmes + modèles), devient une **élite interne**, stratégique, protectrice,
- Diminution de la dépendance aux fournisseurs. Si l'entreprise sait optimiser ses modèles et/ou ses solvers, elle devient moins dépendante des prix des hyperscalers, elle peut négocier, elle peut internaliser certaines charges, elle stabilise sa roadmap. La **frugalité lui apporte** l'autonomie stratégique interne,
- Attractivité et rétention des talents. Les meilleures équipes techniques veulent travailler sur du co-design réel, sur du hardware innovant, sur de l'ingénierie fine, en "hard science". La **frugalité** structurée donne une **raison d'être** aux talents d'HPC/IA.

9 Comment faire : à l'échelle de la France et de l'UE

Deux enjeux essentiels à l'échelle du pays sont **la souveraineté et la compétitivité**. La frugalité devient une **nouvelle forme de souveraineté énergétique et numérique**.

a. Réduire la consommation et la dépendance énergétique

L'UE importe 55–60 % de son énergie. L'IA, le HPC, les data centers et la 5G vont augmenter la consommation de +20 à +40 % d'ici 2030. La frugalité, telle que nous la décrivons ici, sera impérative.

b. Réduire la dépendance technologique

Aujourd'hui, les sources stratégiques sont les suivantes :

- les GPUs viennent des USA,
- les chips photoniques viennent des USA et Chine,
- les mémoires avancées sont coréennes,

- les fonderies avancées sont à Taiwan, en Corée et aux USA.

Si l'UE ne maîtrise pas la performance/*W*, elle deviendra un **simple client captif**, jamais un acteur souverain.

c. Renforcer le tissu industriel

Les startups européennes du semi-conducteur, de l'edge, de l'embarqué, de l'IA frugale (SiPearl, GreenWaves, etc.) ne peuvent exister que si :

- on soutient la demande locale,
- on finance un pipeline d'itérations,
- on crée des sandboxes d'expérimentation,
- on industrialise le co-design décrit dans ce white paper.

Le soutien à l'innovation et à l'industrialisation n'est pas un luxe : c'est un multiplicateur de PIB !

10 Péril en la demeure

a. Explosion des coûts électriques

Sans frugalité, un seul hyperscale LLM peut consommer autant qu'une ville moyenne. Les entreprises européennes paieront 2× à 3× plus cher leurs services, deviendront moins compétitives et *in fine* verront fondre leurs marges, d'où moins d'emplois, moins d'innovation.

b. Déclassement technologique

Si les solutions les plus efficaces viennent de l'étranger, tous les secteurs (santé, finance, défense, industrie) deviennent dépendants d'acteurs non européens. Ainsi en est-il des GPU, de la lithographie DUV/EUV, des fonderies, des clouds hyperscale.

c. Captivité stratégique

Les modèles IA sont de plus en plus de type "fermé-source + opaque". Sans capacité pour les usagers d'optimiser eux-mêmes, il est impossible de vérifier les claims, de réduire les coûts et de créer des variantes adaptées aux industries européennes. On entre alors dans une forme de **colonisation numérique par l'énergie**.

d. Perte d'autonomie militaire et scientifique

L'IA militaire, médicale, énergétique, climatique, ne doit pas dépendre de serveurs non européens, de modèles non européens, de hardware non européen.

Un pays qui ne contrôle pas sa capacité de calcul **ne contrôle plus sa capacité de décision**. C'est une définition même de la **perte de souveraineté**.

11 En guise de conclusion

Nous sommes à une bascule du training vers l'inférence. Celle-ci va être en expansion type post big bang ! L'inférence sera distribuée, locale, fondue dans l'internet of Everything, avec des latences minimales. Ne pas l'appréhender est déjà prendre du retard.

Qu'il s'agisse de survie économique des entreprises, de souveraineté des pays, ou d'éco-géostratégie européenne, la frugalité de l'IA n'est pas seulement un choix écologique ; c'est **le fondement énergétique de toute compétitivité au siècle de l'IA**.

Dans ce qui peut être vu comme un paradigme, on pourrait assister à divers « effets rebond », connus depuis le 19ème siècle et le paradoxe de Jevons : plus la consommation d'une ressource est efficace, ...plus celle-ci est demandée, utilisée. Bref, la réduction de consommation d'un usage unitaire, réel, peut conduire à une augmentation de la consommation cumulée, par multiplication des usages.

En toute hypothèse, les acteurs n'auront pas le choix. Ils seront frugaux ou seront dépassés : ils devront s'imposer de définir leur trajectoire de frugalité, continûment évolutive, sous forme de feuille de route.

Souhaitons que ce white paper puisse aider le lecteur à en prendre conscience.

Références

[1] IEA (2025), *Energy and AI*, IEA, Paris

https://www.iea.org/reports/energy-and-ai, Licence: CC BY 4.0

[2] Energy Institute (2025) *Statistical Review of World Energy. 2025 74th edition*

https://www.energyinst.org/statistical-review

[3] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Toby Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, Sukrut Oak.

"*The AI Index 2025 Annual Report,*" AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025.

https://doi.org/10.48550/arXiv.2504.07139

Glossaire

CPU : Central Processing Unit

CTU : Coût total d'usage

CXL : Compute Express Link

GES : Gaz à effets de serre

GPU : Graphics Processor Unit

HBM : High Band Memory

IA : Intelligence Artificielle

ICT : Information and Communication Technologies

IEA : International Energy Agency

IoT : Internet of Things

IPU : Intelligence Processing Unit

ISA : Instruction Set Architecture

ISA : Instruction Set Architecture

LLM : Large language model

NN : Neural Network

PUE : Power Usage Effectiveness

TPU : Tensor Processing Unit

XLA/JAX : XLA (Accelerated Linear Algebra) est un compilateur Open Source pour le machine learning