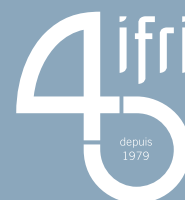




FÉVRIER
2025



IA, centres de données et énergie

Quelles tendances ?

Laure DE ROUCY-ROCHEGONDE
Adrien BUFFARD

Centre
géopolitique des
technologies

Centre énergie
et climat

L’Ifri est, en France, le principal centre indépendant de recherche, d’information et de débat sur les grandes questions internationales. Créé en 1979 par Thierry de Montbrial, l’Ifri est une fondation reconnue d’utilité publique par décret du 16 novembre 2022. Elle n’est soumise à aucune tutelle administrative, définit librement ses activités et publie régulièrement ses travaux.

L’Ifri associe, au travers de ses études et de ses débats, dans une démarche interdisciplinaire, décideurs politiques et experts à l’échelle internationale.

Les opinions exprimées dans ce texte n’engagent que la responsabilité des auteurs.

ISBN : 979-10-373-1041-5

© Tous droits réservés, Ifri, 2025

Couverture : © Shutterstock.com

Comment citer cette publication :

Laure de Roucy-Rochegonde et Adrien Buffard, « IA, centres de données et énergie. Quelles tendances ? », *Notes de l’Ifri*, Ifri, février 2025.

Ifri

27 rue de la Procession 75740 Paris Cedex 15 – FRANCE

Tél. : +33 (0)1 40 61 60 00 – Fax : +33 (0)1 40 61 60 60

E-mail : accueil@ifri.org

Site internet : ifri.org

Auteurs

Laure de Roucy-Rochegonde est directrice du Centre géopolitique des technologies de l'Ifri depuis février 2024. Elle était précédemment chercheuse au Centre des études de sécurité de l'Ifri depuis mars 2020, où elle a travaillé sur les applications militaires de l'intelligence artificielle, la conflictualité normative et la maîtrise des armements. Docteure en science politique, elle est également chercheuse associée au Centre de recherches internationales (CERI, Sciences Po/CNRS). En octobre 2024 est paru son premier ouvrage, intitulé *La Guerre à l'ère de l'Intelligence artificielle : quand les machines prennent les armes* (PUF).

Parallèlement à ses activités de recherche, elle enseigne l'éthique de la guerre et la maîtrise des armements à Sciences Po Paris et à l'Université Paris 2 Panthéon-Assas. Elle est par ailleurs titulaire d'un master de politiques publiques et d'un *bachelor* de Sciences Po Paris, au cours duquel elle a passé une année au département de *War Studies* de King's College à Londres.

Adrien Buffard est étudiant en master International Energy Transitions à Sciences Po Paris. Il a effectué un stage au Centre énergie et climat de l'Ifri au cours de son parcours universitaire. Actuellement en recherche à EDF R&D sur des sujets de géoéconomie de l'énergie, il a précédemment travaillé au ministère de la Transition écologique sur la planification et la mise en concurrence de l'éolien en mer, à l'ambassade de France en Lituanie sur la dépendance énergétique des pays baltes à la Russie, et en tant qu'analyste financier à Bloomr. Il est diplômé du Collège universitaire de Sciences Po et licencié ès lettres modernes auprès de l'Université Sorbonne Nouvelle.

Résumé

Le secteur des technologies de l'information et de la communication représente aujourd'hui 9 % de la consommation mondiale d'électricité, les centres de données (*data centers*) 1 à 1,3 % et l'intelligence artificielle (IA) moins de 0,2 %. La consommation d'énergie des centres de données est restée stable autour de 200 térawattheures (TWh) par an jusqu'en 2019, grâce aux gains continus en efficacité qui ont compensé l'envolée des besoins d'instances de calcul et de capacités de stockage. La demande électrique des *data centers* est alors devenue exponentielle et a atteint 460 TWh en 2022 (cryptomonnaies incluses), malgré l'utilisation de puces et la conception d'algorithmes plus efficaces ainsi que l'augmentation de la taille moyenne des *data centers* qui ont permis de réaliser des économies d'échelle. La demande croissante d'énergie du *cloud* d'abord, et maintenant de l'IA (10 % de la demande d'électricité des *data centers* aujourd'hui), a exacerbé cette tendance.

Si la consommation d'électricité des centres de données est en plein essor au niveau mondial (elle devrait se situer entre 1 000 et 2 000 TWh en 2030), elle reste un moteur modeste de la croissance mondiale de la demande finale d'électricité, loin derrière la climatisation et l'électrification des industries. Le déploiement de l'IA offre un large éventail d'opportunités pour les secteurs de l'industrie et de l'énergie, allant de gains d'efficacité à une meilleure modélisation de la fusion nucléaire, en passant par les réseaux intelligents. La demande d'électricité explose ainsi dans certaines régions où les centres de données sont largement concentrés. Aux États-Unis, où se trouvent plus de la moitié des centres de données du monde, ceux-ci pourraient représenter jusqu'à 13 % de la consommation totale d'électricité en 2030 (contre 4 % en 2024), soit 560 TWh de consommation à cette date. En Europe, les besoins en IA devraient représenter 4 à 5 % de la demande totale d'électricité d'ici là (contre 2 à 3 % en 2024).

Certaines incertitudes demeurent autour des effets futurs de l'IA, étant donné qu'elle doit faire face à de nombreux défis qui pourraient entraver son déploiement et son utilisation dans les années à venir. Les files d'attente pour la connexion aux réseaux des centres de données s'élargissent et renforcent la pression sur des réseaux déjà sous contraintes aux États-Unis et en Europe. Pour subvenir à l'explosion de la demande électrique, des capacités de production pilotables supplémentaires sont nécessaires. Les *data centers* pourraient ainsi constituer un moteur de la hausse de production de gaz, en particulier aux États-Unis où ils profiteront de l'agenda trumpiste. De même, la concentration du marché des puces dans le contexte d'une guerre commerciale qui s'intensifie pourrait ralentir

le déploiement de l'IA. Les autres défis liés au déploiement de l'IA comprennent le manque de flexibilité, l'approvisionnement en eau et la formation des travailleurs.

Si les GAFAM – pour Google (Alphabet), Apple, Facebook (Meta), Amazon et Microsoft – jouent un rôle clé dans le déploiement des centres de données, le déploiement rapide de l'IA met en péril leurs objectifs climatiques, étant donné que la croissance de la demande électrique et des besoins en infrastructures dépasse la vitesse d'atténuation de leur empreinte carbone. Toutefois, avec un nombre record de PPAs dans les énergies renouvelables, l'énergie nucléaire, les petits réacteurs modulaires (SMR) – ainsi qu'une flotte importante de véhicules électriques à chargement bidirectionnel – les GAFAM deviennent des acteurs énergétiques de plus en plus importants et pourraient jouer un rôle clé dans la gestion de la demande électrique, en particulier si les centres de données s'avèrent flexibles. De leur côté, les géants des hydrocarbures examinent attentivement l'irruption de l'IA qui les pousse à reconsidérer leurs investissements dans les énergies renouvelables et à envisager l'emploi de cette technologie pour décarboner leurs activités plus rapidement.

À l'avenir, les *data centers* seront probablement encore plus concentrés dans certains endroits. Les États-Unis resteront le premier pays accueillant des *data centers*, tandis qu'en Europe ce sont les pays où l'énergie est abondante et bon marché (Scandinavie et France) ou ceux où se trouvent les industries, la finance ou des aménagements fiscaux (Allemagne, Royaume-Uni et Irlande) qui les accueilleront. L'annonce des projets prométhéens de Stargate aux États-Unis (500 milliards de dollars) et les 109 milliards d'euros promis par la France augurent la construction de gigantesques centres de données dont les capacités atteindront les cinq gigawatts (GW), ce qui devrait inciter les géants de la *tech* à considérer d'autant plus l'utilisation de centrales nucléaires traditionnelles pour leurs *data centers*. Le Sommet pour l'action sur l'IA, qui s'est tenu début février 2025 à Paris, a rappelé l'importance de concevoir une IA « durable » et compatible avec la transition énergétique.

Executive summary

The information and communication technologies sector today accounts for 9% of global electricity consumption, data centers for 1-1.3%, and artificial intelligence (AI) for less than 0.2%. The energy consumption of data centers was stable at around 200 terawatt-hour (TWh) per year until 2019, thanks to continued efficiency gains that totally offset skyrocketing computational and storage needs. However, efficiency gains then started to plateau and power usage effectiveness completely stalled, while the electricity needs of power data centers began rising exponentially to reach 460 TWh in 2022 (cryptocurrencies included), despite more efficient chips, algorithmic improvements and bigger data centers that have led to economies of scale. The growing energy demands of cloud services, first, and now AI workloads (10% of today's data centers' electricity demand), have exacerbated this trend. In the future, hyperscale data centers will gain shares amongst all kinds of data centers, and AI will probably account for around 20% of data centers' electricity demand by 2030.

If data center electricity consumption is booming at the global level (expected to be between 1,000 and 2,000 TWh in 2030), it remains a modest driver of the global growth in final electricity demand, far behind air conditioning and the electrification of industries. Since AI deployment offers a wide range of opportunities for the industry and energy sectors, ranging from efficiency gains to fusion monitoring and smart grids, it is pushing up the electricity demand in some regions where it is widely concentrated. In the US, where more than half of the world's data centers are located, they could make up to 13% of the total electricity consumption in 2030 (compared with 4% in 2024), representing 560 TWh of consumption then. In Europe, AI needs should account for 4 to 5% of total electricity demand by then (up from 2-3% in 2024).

Some uncertainties over AI's future impacts remain, as the technology must cope with numerous challenges that will likely hamper its full deployment and usage in the years ahead. Grid queues are rising, and data centers are reinforcing the pressure on already strained grids in the US and Europe. To cope with the surging electricity demand, additional dispatchable capacity generation is needed, and data centers may be a long tailwind for gas production, especially when coupled with Trump's "drill, baby, drill" agenda. Likewise, the chip market concentration amidst a deepening trade war may slow down AI deployment. Other concerns and challenges encompass the lack of flexibility, water supply, and workers' skills.

Big Tech is playing a key role in the rollout of data centers. However, the fast deployment of AI is jeopardizing firms' climate targets as electricity growth and the need for better infrastructure are outpacing their green agenda. Still, with record power purchase agreements (PPAs) in renewables, nuclear power, and small modular reactor (SMR) start-ups –in addition to a large fleet of bidirectional charging electric vehicles–, Big Tech companies are increasingly becoming important energy actors and could further play a key role in providing demand response services, especially if data centers proved to be flexible. Meanwhile, Big Oil looks carefully at the AI irruption, reconsidering its renewables investments and willing to take advantage of the technology to decarbonize faster and further its drilling operations.

In the future, data centers will most likely be even more concentrated in some places. The US will remain their favorite host, whilst European data centers will prefer either countries with cheap and abundant power (Scandinavia and France) or where industries, finance and tax incentives are located (Germany, the UK, and Ireland). With the geopolitical competition on AI capabilities and megaprojects such as 500 billion dollars Stargate or France's 109 billion euros announced investments, immense data centers will be built with capacities reaching 5 GW, providing incentives for Big Tech to look closer to traditional nuclear power plants. Leaders are raising awareness of the environmental impacts of AI, as the AI Action Summit in early February 2025 in Paris highlighted, insisting on the need for sustainable AI.

Sommaire

INTRODUCTION	8
LA MONTÉE EN PUISSANCE DES CENTRES DE DONNÉES	10
Une demande électrique en expansion depuis que les gains en efficacité se sont stabilisés	10
Des centres de données plus grands et plus tournés vers l'IA	14
IA ET CENTRES DE DONNÉES : UNE RÉVOLUTION BINAIRE	16
Une consommation énergétique prééminente	16
IA et centres de données : catalyseurs d'opportunités	20
Les défis et contraintes auxquels les centres de données font face ..	22
PROSPECTIVE DES TENDANCES DOMINANTES	27
Entre ambition technologique et dilemme énergétique : les GAFAM et les géants des hydrocarbures	27
L'avenir des centres de données	30
CONCLUSION	34

Introduction

Le lendemain de son investiture, Donald Trump a annoncé le lancement de ce que le fondateur d'OpenAI, Sam Altman, a décrit comme « le projet le plus important de notre époque ». Baptisée Stargate, cette initiative vise à investir 500 milliards de dollars au cours des quatre prochaines années dans le développement de l'intelligence artificielle (IA). Soutenue par trois géants de la technologie – le spécialiste du *cloud* Oracle, la société d'investissement japonaise SoftBank et la désormais omniprésente société mère de ChatGPT, OpenAI – Stargate a pour objectif de « construire l'infrastructure physique et virtuelle qui alimentera la prochaine génération d'IA » tout en créant « plus de 100 000 emplois » aux États-Unis.

Au cœur de ce projet d'envergure se trouve la construction de centres de données. Les technologies de l'IA reposent entièrement sur la capacité à capturer, stocker et traiter des données, rendant les centres de données indispensables. Selon Larry Ellison, président d'Oracle, Stargate se concentrera principalement sur l'établissement de ces centres à travers les États-Unis. Cela marque un tournant pour OpenAI, qui s'appuyait jusqu'à présent sur l'infrastructure fournie par son partenaire Microsoft, mais qui sera désormais en mesure de tirer parti de ses propres *data centers*. Dans un rapport publié en septembre 2024, la société de Sam Altman a exhorté le gouvernement américain à soutenir la création de centres de données géants de cinq gigawatts (GW), dont le coût global est estimé à 100 milliards de dollars (Md\$).

Trump semble avoir tenu compte de cet appel. Le nouveau président a annoncé son intention de déclarer un « état d'urgence » pour lever les obstacles réglementaires à la construction de centres de données malgré les inquiétudes concernant leurs effets sur l'environnement. Il a également suggéré que Stargate produise sa propre électricité, sans donner plus de détails sur la manière d'y parvenir. Selon certaines estimations, la consommation d'énergie des centres de données pourrait tripler d'ici 2028, un chiffre impressionnant.

Cependant, quelques jours seulement après cette annonce très médiatisée aux États-Unis, plusieurs entreprises technologiques américaines ont vu le cours de leurs actions s'effondrer à Wall Street. Nvidia, la société cotée en Bourse ayant la plus grande valeur au monde, a perdu près de 600 milliards de dollars de capitalisation boursière le lundi 27 janvier 2025. La raison ? Le succès fulgurant de DeepSeek, un concurrent chinois de ChatGPT, qui est rapidement devenu l'application d'IA la plus téléchargée sur de multiples plateformes dans le monde entier. DeepSeek est également présenté comme étant beaucoup plus économe en

énergie que ses concurrents. Sa formation n'aurait nécessité que 5,6 millions de dollars – bien que ce chiffre soit en réalité tronqué – contre 100 millions de dollars pour ChatGPT-4, et aurait utilisé des ordinateurs équipés de seulement 2 000 puces Nvidia, alors que ChatGPT a eu besoin de 16 000 puces plus perfectionnées.

Cette réduction drastique des besoins en calcul permet non seulement de faire chuter les coûts, mais pourrait également entraîner une baisse significative de la demande énergétique des centres de données. Si DeepSeek venait à remplacer des modèles tels que ceux d'OpenAI, la consommation d'énergie globale s'en trouverait fortement réduite. La montée en puissance de DeepSeek a d'ailleurs provoqué une vague de ventes d'actions dans le secteur de l'énergie, parallèlement à la baisse dans le secteur technologique. Le géant américain Constellation Energy, qui avait signé en 2024 un contrat de fourniture d'électricité avec Microsoft, a vu son cours chuter de plus de 20 % sur le Nasdaq.

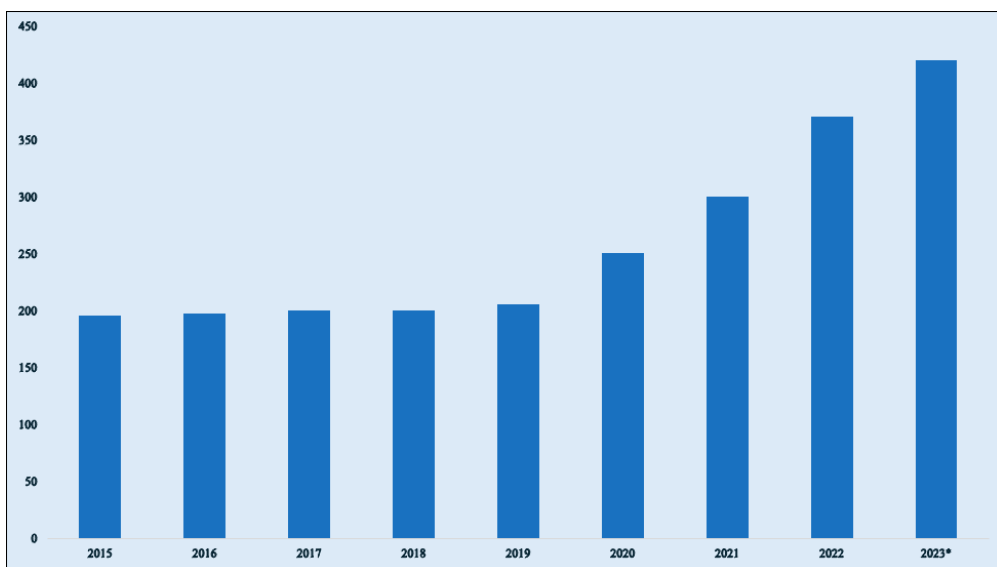
Alors que le Sommet pour l'action sur l'IA, qui s'est tenu à Paris les 10 et 11 février 2025, a mis en avant l'urgence de développer une « IA écologiquement soutenable », il devient désormais essentiel d'évaluer l'impact réel de l'intelligence artificielle et des centres de données sur le secteur énergétique — et de déterminer si les réseaux électriques sont capables d'absorber une telle hausse de la demande. La présente *Note* entend répondre à ces interrogations.

La montée en puissance des centres de données

Stabilisation des gains en efficacité et expansion de la demande électrique

En 2022, le secteur des technologies de l'information et de la communication représentait 9 % de la consommation mondiale d'électricité et les centres de données 1 à 1,3 %¹. Jusqu'en 2018, la consommation d'électricité des *data centers* est restée constante, la consommation mondiale annuelle d'électricité plafonnant à environ 200 TWh (*Schéma 1*) malgré l'augmentation du nombre et de la taille moyenne des centres.

Schéma 1. Demande électrique mondiale des centres de données (TWh) (cryptomonnaies exclues), 2015-2023



Sources : IEA, Cisco, Goldman Sachs, GECF.

Entre 2010 et 2018, la demande d'électricité des *data centers* n'a augmenté que de 6 %, alors que leur charge de travail a bondi de 550 % et que la capacité mondiale installée de stockage de données a été multipliée par 26 (*Schéma 2*). Ce découplage résulte des gains d'efficacité. En effet, l'efficacité du stockage des données a été multipliée par neuf au cours de la période 2010-2018, avec un taux de croissance annuel composé (TCAC) de 20 %. Le « ratio moyen d'efficacité énergétique » (*power usage*

1. « World Energy Outlook 2024 », Agence internationale de l'énergie (AIE), 2024, p. 186.

effectiveness, abrégé PUE²⁾) a été considérablement réduit, passant de 2,50 en 2010 à 1,58 en 2018, grâce à d’importantes améliorations réalisées sur les systèmes de refroidissement et sur les composants auxiliaires, tels que les systèmes de sécurité, l’éclairage et les blocs d’alimentation. Les systèmes de refroidissement et les composants auxiliaires représentaient respectivement 40 % et 20 % de la consommation d’électricité des centres de données en 2010³, alors qu’ils n’en représentent plus que 25 % et 12 % aujourd’hui, tandis que l’équipement informatique seul représente 60 à 65 % de la consommation totale des *data centers*. La stagnation du PUE laisse augurer une probable envolée de la consommation électrique agrégée des centres de données à l’avenir.

Schéma 2. Tendance des variables de consommation des centres de données, 2010-2018

La demande en services a augmenté...	Changement relatif entre 2010 et 2018 (en multiples)
Capacité de stockage installée mondiale (exaoctets)	26
Trafic du protocole internet mondial (zettaoctets/an)	11
Charges de travail et instances de calcul des centres de données (millions)	6.5
Nombre de serveurs installés dans le monde (millions)	1.3
... tout comme l’efficacité énergétique	Changement relatif entre 2010 et 2018 (en multiples)
PUE moyen	0.75
Intensité énergétique des serveurs (Wh par unité de calcul)	0.24
Nombre moyen de serveurs par charge de travail	0.19
Consommation moyenne du stockage (kWh/téraoctet)	0.11

Sources : E. Masanet et al., « *Recalibrating Global Data Center Estimates* », Sciences Magazine, 28 février 2020, p. 984-986 ; Berkeley Lab.

Pourtant, à partir de 2018, la demande d’électricité liée aux centres de données est montée en flèche. En 2022, les centres de données ont consommé environ 350 TWh, sans compter l’énergie utilisée pour le minage de crypto-monnaies, estimée à environ 110 TWh⁴. La même année, ces centres de données auraient consommé un total de 508 TWh d’électricité s’ils avaient fonctionné sans interruption et à capacité maximale⁵.

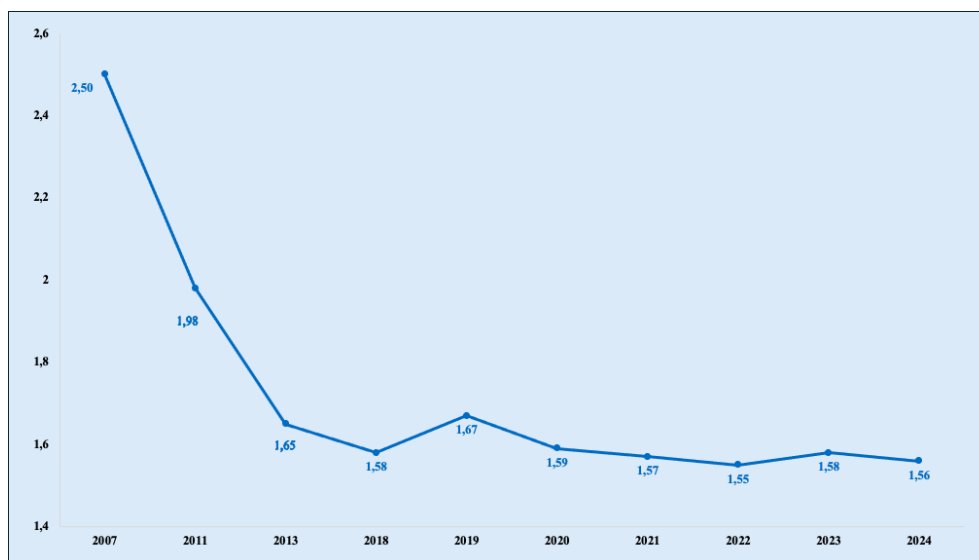
Si des améliorations substantielles de l’efficacité technique ont été réalisées au cours de la dernière décennie, conduisant à une consommation globale modérée des centres de données, le rythme des gains en efficacité

2. Le PUE est calculé en divisant l’énergie totale consommée par un centre de données (climatisation et éclairage inclus) par l’énergie utilisée uniquement pour les calculs et le stockage. Un minimum théorique de 1 reviendrait à dire toute l’énergie utilisée l’est pour les activités « utiles » des centres de données.
3. « Analyzing Artificial Intelligence and Data Center Energy Consumption », Electric Power Research Institute, 2024.
4. « Data Centers and Data Transmission Networks », AIE, 2 février 2025, disponible sur : www.iea.org (fondé sur l’analyse de Cambridge Bitcoin Electricity Consumption Index).
5. J. Saul et al., « AI Is Already Wreaking Havoc on Global Power Systems », Bloomberg, 21 juin 2024.

du PUE s'est stabilisé et cette tendance à la stagnation devrait se poursuivre⁶.

Cela a entraîné une croissance soudaine et importante. En effet, le PUE des centres de données a complètement stagné ces dernières années, s'établissant à 1,58 en 2023, soit le même niveau qu'en 2018 (*Schéma 3*), même si dans les centres de données les plus modernes, dotés de technologies avancées, malgré le besoin accru de refroidissement dû à des GPU plus sophistiqués, le PUE peut être ramené à 1,2. Dans certains pays avancés, comme la France, le PUE moyen est même de 1,7⁷ alors que 1,2 devrait être le plafond standard. Cela signifie que les améliorations apportées aux systèmes de refroidissement et aux composants auxiliaires ont atteint un plateau ces dernières années.

Schéma 3. PUE mondial moyen des data centers, 2007-2024



Source : Uptime Institute.

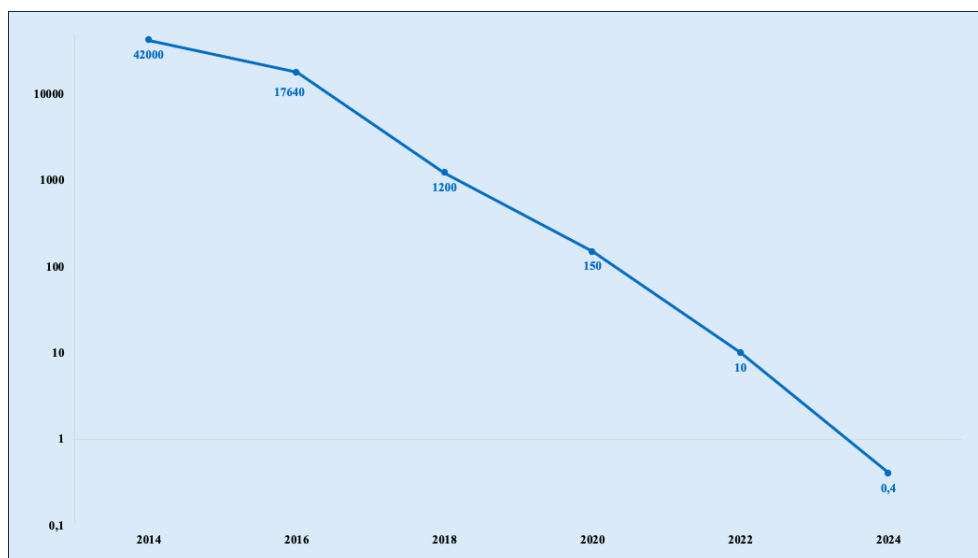
Contrairement au PUE, les puces (l'élément principal des centres de données, tant pour le stockage que pour l'entraînement de l'IA et les inférences) se sont continuellement améliorées et ont joué un rôle clé dans la réduction de la consommation d'énergie. Grâce notamment à l'émergence des *chipelets* – ces micropuces qui ont permis aux fabricants de diviser les grandes puces monolithiques en composants modulaires plus petits et plus efficaces – l'efficacité énergétique des puces spécialisées pour l'IA s'est considérablement améliorée, à un rythme exponentiel. Chaque année, la consommation par token a été divisée par quatre, ce qui correspond à un facteur d'amélioration de 100 000 en seulement dix ans (*Schéma 4*). Toutefois, des puces plus efficaces peuvent également accroître

6. E. Masanet *et al.*, « Recalibrating Global Data Center Estimates », *Sciences Magazine*, 28 février 2020, p. 984-986; D. Bizo, « Global PUEs: Are they Going Anywhere? », Uptime Institute, 4 décembre 2023, disponible sur : <https://journal.uptimeinstitute.com> ; « World Energy Outlook 2024 », *op. cit.*, p. 186.

7. « Avis d'experts – Les data centers ou centres de données », Ademe, Novembre 2024.

la demande en raison de l'effet de rebond, certains pensant même que l'IA pourrait devenir un nouvel exemple du paradoxe de Jevons⁸. Il est donc peu probable que les améliorations de l'efficacité énergétique compensent entièrement la demande croissante de services des *data centers*⁹.

Schéma 4. Efficacité énergétique des puces spécialisées pour l'IA de Nvidia (J/Token), 2014-2024



Source : NVIDIA.

Nota bene : Ce graphique utilise une échelle logarithmique. Les années 2020, 2022 et 2024 représentent respectivement les puces Nvidia A100, H100 et B100.

Après le PUE et l'efficacité des puces électroniques, la taille des centres de données est le dernier élément clé à considérer pour juger de l'efficacité générale des *data centers*. La concentration de grands centres de données a fortement contribué à réduire la consommation d'électricité nécessaire aux opérations informatiques, grâce aux économies d'échelle. Les centres de données de petite taille (0,5 à 2 MW), y compris les centres de données périphériques (*edge data centers*) situés à proximité de l'utilisateur final, représentent environ la moitié de tous les centres de données, mais seulement 10 % de leur capacité de consommation électrique¹⁰. Les centres de données de grande taille peuvent être détenus et exploités sur place par des entreprises (20-30 % de la charge totale) ou être des *hyperscalers* (60-70 % de la charge totale), d'immenses installations capables d'augmenter rapidement leurs opérations pour répondre aux vastes besoins informatiques des géants du *cloud* tels qu'Amazon AWS, Google Cloud et Microsoft Azure.

8. La directrice de Microsoft Satya Nadella a publié « Jevons paradox strikes again! » sur X (ex-Twitter) le 27 janvier 2025.

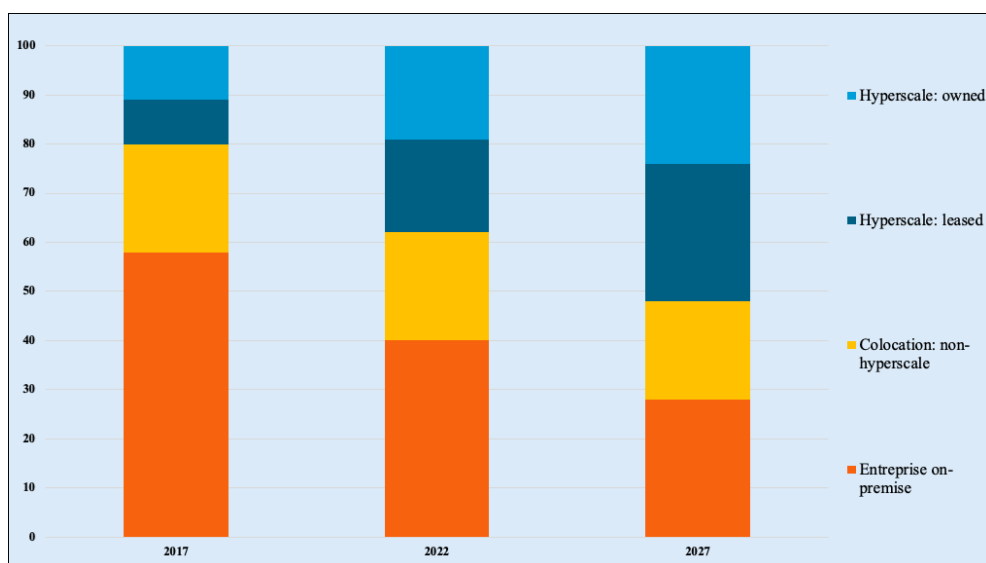
9. « World Energy Outlook 2024 », *op. cit.*, p. 186-189.

10. « Analyzing Artificial Intelligence and Data Center Energy Consumption », *op. cit.*

Des centres de données plus grands et plus tournés vers l'IA

Les centres de données sont de plus en plus grands, tant en taille qu'en capacité. Au cours de la dernière décennie, les entreprises ont réorienté leurs investissements informatiques vers le *cloud* et se sont montrées réticentes à l'idée d'investir dans leurs propres *data centers* ; les principaux fournisseurs de *cloud* ont donc rapidement mis en place d'immenses réseaux mondiaux de *hyperscalers*. La croissance des *hyperscalers* a été alimentée par le développement rapide de services numériques plus orientés vers le consommateur, tels que les réseaux sociaux, le commerce électronique et les jeux en ligne. La superficie moyenne des centres de données dans le monde est aujourd'hui d'environ 38 300 mètres carrés (m²), soit près de cinq fois plus qu'en 2010¹¹. En 2022, au niveau mondial, 37 % des centres de données étaient des *hyperscalers* et 23 % des grands *data centers* en colocalisation, sans être toutefois des *hyperscalers*. En effet, la part des centres de données sur site diminue (Schéma 5), malgré la croissance prévue du marché mondial des centres de données périphériques de 22,1 % à un taux de croissance annuel composé (TCAC) jusqu'en 2030¹². Cette tendance devrait se poursuivre. Environ 70 % de la croissance des centres de données devrait être assurée directement ou indirectement (via le *cloud*, par exemple) par les *hyperscalers* d'ici à 2030.

Schéma 5. Évolution tendancielle de la capacité des centres de données, 2017-2027



Source : Synergy Research Group.

11. DC Bite, cité dans J. Saul *et al.*, « AI Is Already Wreaking Havoc on Global Power Systems », *op. cit.*

12. « Analyzing Artificial Intelligence and Data Center Energy Consumption », *op. cit.*

Il est important de souligner que l'augmentation de la demande pour les centres de données a d'abord été liée à l'utilisation croissante du *cloud*, et ce, bien au-delà du déploiement de l'IA qui l'a poussée plus tard (bien que l'IA puisse avoir aussi recours au *cloud*). En effet, l'IA ne représente actuellement que 10 à 15 % de la consommation d'énergie des centres de données, soit moins de 0,2 % de la demande mondiale d'électricité, même si ces chiffres sont appelés à grandir. Selon certaines estimations, la demande d'électricité des centres de données augmentera de 160 % d'ici à 2030 et, l'IA devrait représenter environ 19 % de la demande totale d'électricité des centres de données dès 2028¹³.

13. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », Goldman Sachs, 14 mai 2024, disponible sur : www.goldmansachs.com.

IA et centres de données : une révolution binaire

Une consommation énergétique proéminente

À l'échelle mondiale, la demande énergétique des centres de données augmente rapidement. L'Agence internationale de l'énergie (AIE) estime que leur consommation d'électricité (crypto-monnaies comprises) dépassera les 800 TWh en 2026, contre 460 TWh en 2022¹⁴. Dans les scénarios de forte croissance, elle pourrait même dépasser les 1 000 TWh à la fin de l'année prochaine¹⁵. Selon S&P, la consommation pourrait même atteindre 1 300 TWh en 2026. À l'horizon 2030, les prévisions s'échelonnent de 1 100 TWh¹⁶ à 2 000 TWh¹⁷, ce qui représenterait respectivement 3-4 % et 6-8 % de la consommation finale d'électricité dans le monde.

Les écarts conséquents sur les estimations de la future quantité exacte d'électricité consommée par les centres de données sont dus aux incertitudes concernant les modalités d'utilisation de l'IA, son déploiement et son adoption par le monde professionnel. En effet, l'IA a une consommation électrique qui dépend fortement de l'entraînement (à partir de zéro ou de modèles pré-entraînés), du nombre de paramètres et, surtout, du type de tâches d'IA effectuées (*Schéma 6*). Une demande de texte ChatGPT ne nécessite généralement que 0,0005 Wh d'énergie, tandis que la génération d'une image utilise 0,5 Wh, soit près de 1 000 fois plus (quoi que cela représente encore 2 000 fois moins que la consommation d'énergie d'un cycle de machine à laver). Toutefois, l'énergie nécessaire pour exécuter une simple requête ChatGPT reste dix fois supérieure à celle d'une recherche Google, et l'on estime que l'utilisation de l'IA générative telle que ChatGPT dans chaque recherche Google entraînerait une consommation annuelle d'électricité de 29,2 TWh, soit l'équivalent de la production annuelle de quatre à cinq réacteurs nucléaires¹⁸. Si ce chiffre semble élevé, il reste inférieur aux centaines de TWh supplémentaires qui, selon les prévisions, seront nécessaires dans les années à venir. La croissance proviendra également d'autres utilisations, principalement de la reconnaissance vocale

14. « Electricity 2024 – Analysis and Forecast to 2026 », AIE, janvier 2024, p. 31.

15. *Ibid.*

16. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », *op. cit.*

17. Schneider Electric Global Analysis.

18. J. Anderson *et al.*, « Power of AI: Wild Predictions of Power Demand from AI Put Industry on Edge », S&P Global, 16 octobre 2023, disponible sur : www.spglobal.com.

en direct, du sous-titrage d’images et de la génération de vidéos. En ce qui concerne l’entraînement de l’IA, il consomme beaucoup d’énergie, mais au niveau agrégé, il ne représente qu’une petite partie de l’énergie consommée par l’IA, surtout dans le cas de modèles largement déployés comme ChatGPT, dont on estime que l’entraînement de la 4^e version a nécessité 50 GWh¹⁹ (la production d’un réacteur nucléaire français fonctionnant à pleine capacité pendant deux jours).

Pourtant, dans l’ensemble, le traitement de données en *data center* ne joue qu’un rôle moteur faible dans la croissance de la demande électrique mondiale. D’ici 2030, les centres de données représenteront seulement 5 % de la croissance de la demande totale d’électricité dans le monde, une source faible de croissance de la demande électrique à égalité avec le dessalement de l’eau, et loin derrière le déploiement des véhicules électrique ou de la climatisation, qui chacun y contribueront trois fois plus. L’électrification de l’industrie jouera même un rôle six fois plus important dans l’augmentation de la demande électrique mondiale²⁰.

Schéma 6. Demande énergétique par type de requête IA, 2024

Type de requête	Energie moyenne requise (fondé sur les 20 principaux modèles <i>open source</i>)	Multiplieur (Classification de texte comme référence)
Génération d’image	0.477 Wh	954
Captation d’image	0.109 Wh	218
Reconnaissance vocale	0.025 Wh	50
Réponse à une question	0.001 Wh	2
Classification de texte	0.0005 Wh	1

Source : Sasha Luccioni et al. (2024).

Les centres de données sont concentrés et tendent à l’être de plus en plus avec le temps. Ces dernières années, la plupart des nouveaux centres de données ont été installés aux États-Unis. Alors qu’ils accueillent un peu plus d’un tiers des 8 000 centres de données en 2020, les États-Unis en hébergeaient plus de la moitié (5 381 sur 10 655 centres de données dans le monde) en mars 2024²¹. En outre, leur concentration est également visible au sein même des États-Unis : 15 États représentent 80 % de la capacité des centres de données. Toutefois, si les États-Unis dominent largement le secteur (51 %), l’Europe joue également un rôle important en accueillant environ 15 % des centres de données du monde : l’Allemagne (521), le Royaume-Uni (514) et la France (315) sont les trois premiers, et l’Irlande,

19. T. B. Minde, « Generative AI Does Not Run on Thin Air », Research Institute of Sweden, 10 octobre 2023, disponible sur : www.ri.se.

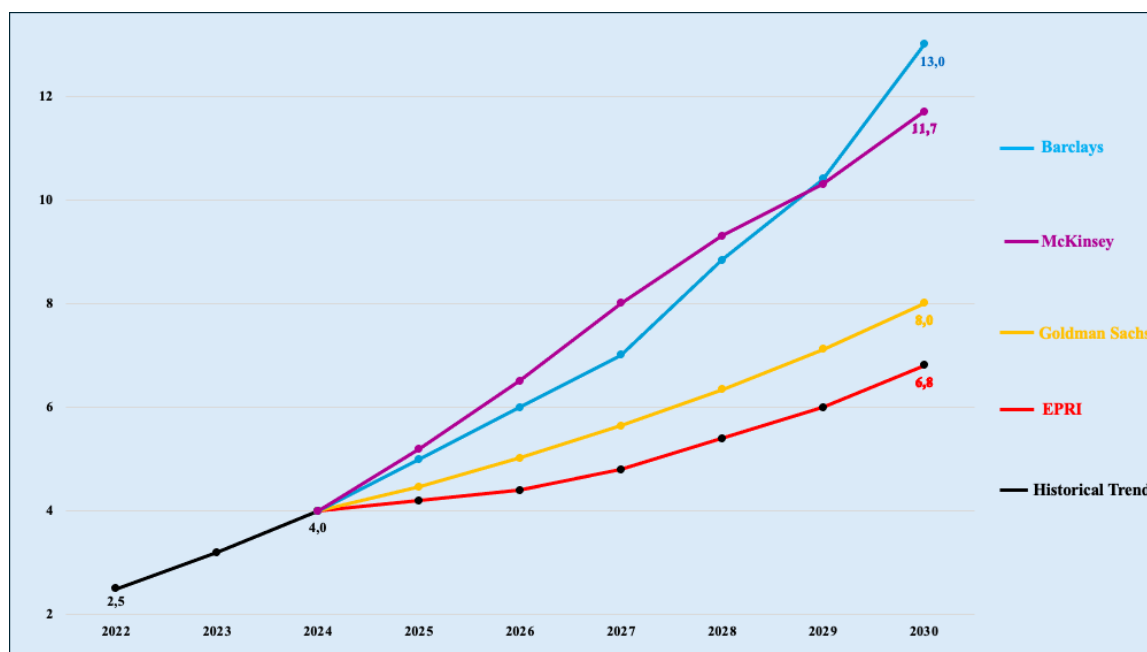
20. « World Energy Outlook 2024 », *op. cit.*, p. 187.

21. « Analyzing Artificial Intelligence and Data Center Energy Consumption », *op. cit.*

contrairement à une idée reçue, est loin derrière dans le classement (entre la 15^e et la 20^e place, avec moins de 100 centres de données). Les autres acteurs principaux sont la Chine, qui accueille 4 % des centres de données dans le monde (449 centres de données en mars 2024, ce qui la place au 4^e rang mondial), et le Canada, qui en héberge 3 % (336 centres de données, ce qui le place au 5^e rang mondial).

Ainsi, l'importance de la demande d'électricité liée aux *data centers* varie en fonction de leur emplacement. Aux États-Unis, les centres de données représentent aujourd'hui 4 % de la consommation totale d'électricité²², et le secteur a déjà dépassé les 10 % de la consommation d'électricité dans au moins cinq États américains. En Virginie, les *data centers* ont consommé un quart de l'électricité totale en 2023. Si les centres de données de l'Union européenne (UE) ont utilisé 55 TWh d'électricité, soit 2 % de leur consommation totale d'électricité²³, dans certains pays spécifiques comme l'Irlande, les centres de données ont déjà dépassé la barre impressionnante des 20 % de la demande nationale d'électricité. En France, ils représentent 42 % de la charge numérique totale, soit environ 10 % de la consommation française d'électricité²⁴.

Schéma 7. Projections de la consommation électrique des centres de données aux États-Unis, 2022-2030



Source : Ifri, à partir de EPRI, Goldman Sachs, Barclays Research, McKinsey.

22. *Ibid.*

23. P. Bertoldi et G. Kamiya, « Energy Consumption in Data Centres and Broadband Communication Networks in the EU », Bureau des publications de l'Union européenne, 2024.

24. « Avis d'experts – Les data centers ou centres de données », *op. cit.*

Les prévisions à moyen terme divergent aussi par pays. Les États-Unis devraient être le marché à la croissance la plus rapide pour les centres de données, passant de 25 GW de capacité et 175 TWh de consommation annuelle en 2023 à plus de 80 GW de capacité et 560 TWh en 2030^{25,26}. D'ici 2030, les centres de données devraient ainsi représenter entre 7 et 13 % de la consommation d'électricité finale des États-Unis (*Schéma 7*). Ils seront même le premier moteur de la croissance de la demande d'électricité. Entre 2024 et 2030, la demande d'électricité des centres de données aux États-Unis devrait augmenter d'environ 400 TWh (presque la consommation annuelle d'électricité de la France) à un taux de croissance annuel composé (CAGR) d'environ 23 %²⁷. Ainsi, sur les 2,4 % de croissance de la consommation d'électricité aux États-Unis au cours de la période 2022-2030, 0,9 % sera lié aux centres de données²⁸.

En Europe, les effets sur le réseau seront moins importants. Selon McKinsey, les centres de données dans l'UE, en Norvège, en Suisse et en Grande-Bretagne devraient atteindre 35 GW de capacité et 150 TWh de consommation annuelle d'ici 2030²⁹, contre 10 GW et 62 TWh aujourd'hui. Les centres de données européens devraient ainsi représenter 5 % de la consommation totale d'électricité du continent en 2030³⁰ (contre 2 % aujourd'hui). À cette date, les besoins en électricité des *data centers* européens pourraient même correspondre à la consommation totale actuelle du Portugal, de la Grèce et des Pays-Bas réunis (220 TWh), selon Goldman Sachs, voire 287 TWh dans le scénario de forte demande de Beyond Fossil Fuels³¹. Bien que ces derniers scénarios soient probablement surévalués, un tel déploiement en Europe des centres de données reviendrait à nécessiter de capter pour leurs seuls besoins 20 % des projets d'énergies renouvelables supplémentaires développés par l'UE d'ici 2030. Toutefois, certaines analyses européennes sont plus prudentes quant à la croissance de la demande en Europe, et la Commission européenne prévoit que les centres de données de l'UE consommeront 98,5 TWh en 2030, soit moins que les Pays-Bas³². En ce qui concerne l'Irlande, les centres de données pourraient consommer un tiers de l'électricité du réseau en 2026,

25. A. Green *et al.*, « How Data Centers and the Energy Sector Can Sate AI's Hunger for Power », McKinsey, 17 septembre 2024, disponible sur : www.mckinsey.com.

26. « AI Revolution: Meeting Massive AI Infrastructure Demands », Barclays Research, 16 janvier 2025.

27. A. Green *et al.*, « How Data Centers and the Energy Sector Can Sate AI's Hunger for Power », *op. cit.*

28. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », *op. cit.*

29. A. Granskog *et al.*, « The Role of Power in Unlocking the European AI Revolution », McKinsey, 24 octobre 2024, disponible sur : www.mckinsey.com.

30. *Ibid.*

31. J. McArdle et P. Terras, « System Overload: How New Data Centres Could Throw Europe's Energy Transition Off Course », Beyond Fossil Fuels, février 2025.

32. Commission Delegated Regulation (EU) 2024/1639 of 14 March 2024 on the first phase of the establishment of a common Union rating scheme for data centres, p. 1.

si aucune politique de restriction supplémentaire d'accès au réseau n'est mise en place d'ici-là³³.

IA et centres de données : catalyseurs d'opportunités

L'IA est censée offrir un large éventail de possibilités pour l'économie. L'IA générative pourrait contribuer à créer entre 2 600 et 4 400 Md\$ de valeur économique à l'échelle mondiale. Mais pour atteindre un quart de ce potentiel d'ici à la fin de la décennie, entre 50 et 60 GW de centres de données supplémentaires seraient requis rien qu'aux États-Unis³⁴.

En ce qui concerne l'industrie en général, les gains d'efficacité à court terme induits par l'IA pourraient s'élever à 30 %, selon le directeur technique *data center* de Siemens. Mais ces gains qui changent la donne doivent tenir compte de la nécessité absolue d'éduquer et de former les travailleurs aux utilisations et aux possibilités de l'IA, ce qui prendra probablement du temps.

Concernant le secteur de l'énergie, l'IA pourrait accélérer le déploiement de certaines technologies révolutionnaires. Associée au calcul à haute performance (HPC), l'IA pourrait contribuer à l'étude de la stabilisation et le comportement du plasma de fusion, à l'anticipation des instabilités du plasma nucléaire et permettre d'ajuster les contrôles en l'espace de quelques millisecondes. Cela pourrait rendre les réactions de fusion plus stables et permettre un contrôle en temps réel. L'IA pourrait également être utilisée pour améliorer l'efficacité des panneaux solaires et la performance des batteries, qui nécessitent une immense puissance de calcul, ainsi que pour améliorer le fonctionnement des réacteurs à fission nucléaire grâce à la maintenance prédictive. Les réseaux intelligents, la gestion de l'eau, l'énergie géothermique et le captage, l'utilisation et le stockage du carbone (CCUS) devraient également bénéficier grandement de l'IA et du calcul intensif.

L'IA pourrait également aider les grandes compagnies pétrolières à réduire leurs émissions de gaz à effet de serre (GES). Adnoc, l'entreprise phare d'hydrocarbures de Abou Dabi, prévoit de déployer l'IA agentique pour agir de manière autonome dans toutes ses opérations. Adnoc affirme avoir alloué 23 milliards de dollars au développement d'une technologie à faible émission de carbone utilisant l'IA agentique et se dit « très optimiste » quant à la possibilité d'atteindre ses objectifs de réduction des émissions de carbone plus tôt que prévu grâce à cette technologie. Toutefois, le rôle actif de l'IA dans la décarbonation des grandes compagnies pétrolières n'a pas encore été mis en évidence.

33. « Electricity 2024 – Analysis and Forecast to 2026 », *op. cit.*, p. 32-33.

34. A. Green *et al.*, « How Data Centers and the Energy Sector can Sate AI's Hunger for Power », *op. cit.*

En outre, la réutilisation de la chaleur générée par les centres de données est actuellement limitée. Le potentiel de récupération de la chaleur résiduelle des *data centers* en France est estimé à environ 1 TWh par an³⁵, ce qui représente environ 6 % de l'énergie totale consommée par les centres de données en 2024.

Il existe de grandes incertitudes quant au potentiel de flexibilité des centres de données. Aux yeux des auteurs, il apparaît que si en théorie les opérateurs de centres de données peuvent rendre leurs installations plus flexibles pour les réseaux électriques, cela est assez rare dans la pratique. La première raison est qu'en ce qui concerne le stockage et le *cloud* (~50 % de la charge actuelle des centres de données), il est difficile de varier ou de déplacer la demande d'électricité. En ce qui concerne l'IA (~10 % de la charge totale des centres de données), l'entraînement et l'inférence (c'est-à-dire les requêtes, utilisations de l'IA) doivent être analysés séparément. En ce qui concerne l'entraînement, une fois qu'une phase est lancée, elle ne peut être interrompue jusqu'à ce qu'elle soit terminée, même si l'intensité des calculs (par exemple, avec le HPC combiné) peut être légèrement modulée. La seule façon de s'adapter au réseau serait donc de lancer les phases de formation en dehors des heures de pointe. Concernant l'inférence maintenant, il faut différencier l'IA générale comme les grands modèles de langage (LLM) ou les IA largement utilisées comme ChatGPT (le *chatbot*), et l'IA spécifique conçue pour certaines tâches précises et donc pour des entreprises particulières. Dans le premier cas, il est peu probable que la demande soit flexible (c'est-à-dire qu'elle puisse être déplacée ou reprogrammée). La nécessité d'une réponse immédiate exclut toute reprogrammation en dehors des heures de pointe, en particulier dans le cas de la reconnaissance vocale en direct. De plus, il semble que ces demandes soient traitées de manière centralisée dans des *hyperscalers* et qu'elles ne soient que rarement transférées vers des *edge data centers* connectés à des réseaux électriques moins sollicités (jusqu'à 10-15 % des fois). En revanche, pour la seconde (IA spécifique), il est possible qu'elle corresponde au cycle de production quotidien des énergies renouvelables. Il est en effet imaginable qu'une utilisation généralisée de tâches spécifiques (comme les demandes d'IA) se fasse pendant les heures de travail lorsque les panneaux solaires photovoltaïques reçoivent la lumière du jour, et que les besoins soient réduits lorsque les travailleurs dorment et n'utilisent pas cette technologie. Il est également plus facile de faire évoluer ces solutions, car elles nécessitent moins d'infrastructures, ce qui les rend plus flexibles. La même logique – utilisation de jour, arrêt de nuit – pourrait être appliquée aux *software as a service* (SaaS) et aux serveurs d'entreprise, qui représentent ensemble environ 20 % de la charge des centres de données d'aujourd'hui.

35. « Avis d'experts – Les data centers ou centres de données », *op. cit.*

Les défis et contraintes auxquels les centres de données font face

Bien qu'ils offrent un large éventail d'opportunités dans tous les secteurs, l'IA et les centres de données devront tous deux surmonter plusieurs défis systémiques. Nous pouvons énumérer quatre facteurs susceptibles d'entraver leur déploiement : 1) les contraintes de réseau et les files d'attente ; 2) l'immense besoin d'investissements dans des infrastructures adaptées et les capacités de production d'énergie ; 3) la guerre commerciale dans un marché très concentré ; et 4) le manque de main-d'œuvre qualifiée en matière d'IA.

L'expansion rapide des centres de données met déjà certains réseaux électriques sous pression. L'exemple le plus notable est celui de l'Irlande, où l'opérateur public d'électricité a imposé un moratoire sur les centres de données de Dublin au début de 2022 et a fixé des conditions pour la connexion de nouveaux centres au réseau électrique, dont une préférence pour ceux qui produisent leur propre électricité. L'opérateur irlandais a des plans pour développer son réseau afin d'accueillir dans le futur d'autres *data centers*, mais les restrictions sur les nouveaux raccordements importants devraient se poursuivre jusqu'en 2028. Les files d'attente pour la connexion aux réseaux électriques se multiplient dans le monde entier. Dans l'ouest de Londres, historiquement considéré comme un *hub* pour les centres de données, les nouvelles installations doivent attendre jusqu'en 2030 pour se connecter au réseau³⁶. En Suède, particulièrement dans le sud, la demande de connexion est si forte – notamment poussée par les renouvelables – que les *data centers* peuvent avoir à attendre des années. Ainsi, les GAFAM se livrent parfois à des guerres d'enchères pour obtenir des sites de centres de données bénéficiant d'un accès facile à l'électricité. Aux États-Unis, lorsqu'une organisation membre du réseau électrique du nord-est des États-Unis a demandé à l'autorité de régulation de l'énergie (FERC) d'augmenter la capacité électrique d'un centre de données d'Amazon de 300 à 480 MW, la FERC a rejeté la demande, arguant qu'elle « pourrait avoir des conséquences majeures sur la fiabilité du réseau ». En outre, compte tenu de la concurrence potentielle avec les énergies renouvelables pour les connexions au réseau, des goulets d'étranglement sont de plus en plus susceptibles d'apparaître. Cela est d'autant plus probable que la construction d'un centre de données ne prend qu'un à deux ans, alors que les délais de mise en place et d'amélioration du réseau électrique sont beaucoup plus longs (quatre ans pour un gros transformateur électrique ; huit à dix ans pour une ligne de transport souterraine ou aérienne³⁷). Les conclusions concernant les files d'attente

36. Kao Data, cité dans J. Saul *et al.*, « AI Is Already Wreaking Havoc on Global Power Systems », *op. cit.*

37. A. Green *et al.*, « How Data Centers and the Energy Sector can Sate AI's Hunger for Power », *op. cit.*

doivent être tempérées dans une certaine mesure, car les centres de données demandent généralement plusieurs sites et connexions au réseau avant de choisir où s'implanter, ce qui augmente partout les délais de connexion au réseau.

Des investissements dans les réseaux sont donc nécessaires, bien au-delà de ce qui était auparavant convenu pour s'adapter à la part croissante des sources d'électricité renouvelables dans les bouquets électriques. Aux États-Unis bien sûr, mais aussi en Europe. Même si son estimation est probablement excessive, Goldman Sachs considère que l'Europe a besoin de dépenser plus de 800 Md€ pour ses réseaux de transmission et de distribution d'électricité au cours de la seule décennie à venir³⁸. Ce chiffre est bien supérieur aux 584 Md€ de la feuille de route de la Commission européenne pour renforcer le réseau³⁹. En ce qui concerne les autres investissements nécessaires, l'augmentation de la capacité de production d'électricité sûre ou pilotable devient un sujet clé aux États-Unis. Le réseau électrique nord-américain est confronté à des « défis critiques en matière de fiabilité », car la production d'électricité ne parvient pas à suivre le rythme croissant de demandes de connexion liées à l'IA⁴⁰. Certains États sont déjà sous pression lorsqu'ils examinent leurs prévisions de croissance de la charge de pointe, et la moitié du réseau américain risque de souffrir de pénuries d'électricité au cours des dix prochaines années, selon la North American Electric Reliability Corporation⁴¹. Si, à la charge de pointe actuelle de la Géorgie (16,5 GW), nous additionnons les 69 centres de données qui demandent actuellement l'accès au réseau de cet État, la charge de pointe future doublera pour atteindre 31 GW. Cet écart est stupéfiant et l'État pourrait ne pas réussir à créer des capacités électriques supplémentaires d'une telle ampleur, même avec le gaz comme solution de transition. Pour rappel, la tendance actuelle aux États-Unis est d'ajouter 6 GW de production de gaz à cycle combiné (CCGT) chaque année, dans tous les États combinés. Jusqu'en 2030, 47 GW de nouvelles capacités de production d'énergie thermique (environ 50 Md\$) sont nécessaires pour soutenir l'expansion des centres de données⁴². Cela suppose donc un problème de capacités additionnelles de production électrique, d'autant plus que ce chiffre a été estimé avant l'annonce des 500 Md\$ d'investissement dans Stargate, qui nécessitera probablement environ 25 GW d'énergie ininterrompue pour fonctionner⁴³, ce qui représente

38. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », *op. cit.*

39. « An EU Action Plan for Grids », Communication de la Commission européenne 2023/757, 11 novembre 2023.

40. « 2024 Long-Term Reliability Assessment », North American Electricity Reliability Corporation, 2024.

41. *Ibid.*

42. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », *op. cit.*

43. En mars 2024, Sam Altman a laissé entendre prévoir un investissement de 100 milliards de dollars pour un centre de données de 5 GW appelé « Stargate ». Maintenant que nous savons que Stargate représentera un investissement de 500 milliards de dollars, nous avons multiplié le chiffre précédent par 5.

cinq années de CCGT supplémentaires aux États-Unis, soit l'équivalent de vingt réacteurs nucléaires Westinghouse AP1000. Les États-Unis auront donc probablement des difficultés à faire face à cette demande supplémentaire d'électricité, car l'énergie nucléaire ne peut être construite en quelques années, d'autant plus que moins de dix réacteurs Westinghouse ont été mis en service dans le monde à l'heure actuelle. La même analyse de Goldman Sachs a reproduit ses estimations de tendances américaines pour l'Europe⁴⁴ et considère que l'IA conduira l'Europe à augmenter sa consommation d'électricité de 50 % entre 2023 et 2033. Si pour les États-Unis l'augmentation vertigineuse de la consommation électrique des *data centers* semble raisonnable, elle nous le semble bien moins pour l'Europe.

La pénurie de main-d'œuvre est un obstacle supplémentaire au déploiement de l'IA et des *data centers*, en particulier la pénurie grandissante de travailleurs qualifiés dans l'électricité, essentiels à l'exécution de ces projets. Les estimations de McKinsey prévoient une pénurie potentielle de 400 000 travailleurs aux États-Unis, sur la base des prévisions de construction de centres de données et d'expansion d'actifs comparables nécessitant des compétences similaires, tels que la fabrication de semi-conducteurs et les méga-usines de batteries.

De plus, l'IA est un marché très concentré, notamment en ce qui concerne sa composante technologique clé : les puces – et, plus particulièrement, les unités graphiques (abrégiées GPU pour *graphic processing unit*). Deux chiffres sont éloquentes à ce sujet. Alors que Taiwan Semiconductor Manufacturing Company (TSMC) possède/compte plus de 90 % du marché de la production de puces liées à l'IA⁴⁵, en matière de ventes, c'est Nvidia qui détient environ 88 % du marché des GPU⁴⁶. De même, les acheteurs de la génération Hopper (la plus récente et performante) de Nvidia sont rares, et Microsoft fait la course en tête. En 2024, l'entreprise a acheté environ 500 000 puces Hopper, soit presque autant que Meta, Amazon et Google réunis⁴⁷. Si la construction des centres de données ne prend qu'un à deux ans (même si cette durée augmente), il faut trois à cinq ans pour mettre en place une usine de fabrication de puces, ce qui signifie que le marché déjà concentré sera probablement confronté à des pénuries dans les années à venir, si la production ne se diversifie pas ou ne s'accélère pas. Toutefois, un nombre croissant de fabricants de puces entrent sur le marché et se spécialisent dans un aspect (entraînement ou inférence) de la technologie de l'IA. Mais quand on sait qu'un seul centre de données Stargate de 5 GW aura besoin de 2 millions de GPU pour

44. « AI Is Poised to Drive 160% Increase in Data Center Power Demand », *op. cit.*

45. « World Energy Outlook 2024 », *op. cit.*, p. 186-189.

46. R. Dow, « Shipments of Graphics Add-in Boards Decline in Q1 of 24 as the Market Experiences a Return to Seasonality », John Peddie Research, 6 juin 2024, disponible sur : www.jonpeddie.com.

47. G. Smith *et al.*, « FirstFT: Microsoft Buys Twice as Many Nvidia AI Chips as its Rivals in 2024 », *Financial Times*, 18 décembre 2024, disponible sur : www.ft.com.

fonctionner⁴⁸, il est peu probable que les nouveaux venus sur le marché comblent l'écart entre la demande monumentale et l'offre limitée.

Dans ce contexte, et contrairement aux attentes, le retour de Donald Trump à la Maison-Blanche et la guerre commerciale annoncée entre les États-Unis, l'UE et la Chine risquent de freiner le développement et le déploiement de l'IA dans le monde. En effet, cette guerre technologique et commerciale devrait pousser les différentes puissances à entraver la circulation des technologies et pratiques. La compétition pour les puces a, en ce sens, déjà lieu, et ce depuis 2022. À l'époque, les États-Unis ont mis en place une interdiction d'exportation de puces vers la Chine, notamment pour les technologies de Nvidia. Ainsi, les GPU de pointe tels que les A100 et H100 de Nvidia ne peuvent plus depuis être vendus aux entreprises chinoises. Pourtant, la Chine avait anticipé ce mouvement et stocké des puces avant que les États-Unis n'instaurent des contrôles à l'exportation⁴⁹. En outre, la Chine compte trois développeurs de logiciels pour deux aux États-Unis, et il n'existe actuellement aucune restriction sur les entités basées aux États-Unis liées ou détenues par des entreprises technologiques chinoises qui accèdent à des puces d'IA haut de gamme par l'intermédiaire de centres de données situés aux États-Unis. Les entreprises chinoises utilisent également le manque de puces pour maîtriser des technologies, ainsi que pour approcher et recruter des acteurs et des développeurs clés dans la Silicon Valley⁵⁰.

Outre les facteurs susceptibles de faire augmenter (ou baisser) la demande d'électricité liée à l'IA, le boom des centres de données doit faire face à deux autres problèmes liés à l'énergie : le prix et l'approvisionnement en eau.

L'explosion de la demande en centres de données, combinée aux investissements importants mentionnés précédemment, augmente la probabilité que les clients voient le prix de l'énergie augmenter. Aux États-Unis, la décision de la FERC de ne pas donner accès au réseau, mentionnée plus haut, a également été justifiée par les possibles « conséquences majeures sur les coûts pour les consommateurs ». Pour reprendre l'exemple du Nebraska, les coûts liés à l'augmentation de la production d'électricité pour soutenir les nouvelles installations de Google et Meta sont répercutés sur les habitants, avec une augmentation des tarifs estimée à environ 2,5 à 3 % par an⁵¹. En Europe, les prix de gros de l'électricité en Irlande ont été en

48. « Infrastructure Is Destiny: Economic Returns on US Investment in Democratic AI », OpenAI, septembre 2024.

49. S. Rai *et al.*, « China AI Startup Stockpiled 18 Months of Nvidia Chips Before Ban », Bloomberg, 10 novembre 2023, disponible sur : www.bloomberg.com.

50. E. Olcott, « Chinese Tech Groups Build AI Teams in Silicon Valley », *Financial Times*, 18 novembre 2024, disponible sur : www.ft.com.

51. D. Lee, « Sam Altman's Energy "New Deal" Is Good for AI. What About Americans? », Bloomberg, 17 octobre 2024, disponible sur : www.bloomberg.com.

moyenne un tiers plus élevé cette année que dans le reste du continent. Il est indéniable qu'une partie de ces prix plus élevés sont dus à la consommation de plus de 20 % de l'électricité par les centres de données, qui dissocient l'augmentation de la demande de la croissance plus lente de la capacité d'approvisionnement en électricité.

Le refroidissement est également nécessaire pour empêcher les puces de mal fonctionner, ce qui a toujours nécessité une consommation d'eau importante. Des progrès ont été réalisés en matière d'efficacité de l'utilisation de l'eau. Entre la première génération de centres de données qu'elle possède, au début des années 2000, et la génération actuelle, en 2023, Microsoft a réduit l'intensité de sa consommation d'eau (eau consommée par kilowattheure) de plus de 80 %. Cependant, les centres de données ont toujours besoin de grandes quantités d'eau pour refroidir les dispositifs de stockage. Aux États-Unis, l'eau utilisée par les centres de données équivaut à celle d'une ville moderne de 3 millions d'habitants. La demande en eau de l'industrie des centres de données suscite déjà de vives inquiétudes dans certains états comme la Virginie. Des solutions alternatives existent, mais elles nécessitent en moyenne 5 % d'énergie en plus⁵².

52. « 2024 Environmental Sustainability », Microsoft, 2024.

Prospective des tendances dominantes

Entre ambition technologique et dilemme énergétique : les GAFAM et les géants des hydrocarbures

Les GAFAM sont à la tête du boom de l'IA avec des produits tels que ChatGPT (OpenAI/Microsoft), Gemini (Google), Alexa (Amazon) et Siri (Apple). Toutes ces entreprises ont annoncé leur intention de faire fonctionner leurs centres de données entièrement avec de l'énergie décarbonée et de se fixer des objectifs de neutralité climatique. Amazon et Microsoft prévoient d'alimenter leurs centres de données uniquement avec des énergies propres d'ici à cette année, et Google d'ici à 2030. Amazon souhaite atteindre la neutralité climatique d'ici à 2040. Pourtant, pour l'instant, la voie à suivre reste floue et l'IA met en péril les objectifs climatiques de ces grandes entreprises technologiques. Microsoft a récemment admis que ses efforts en matière d'IA compromettaient son objectif de longue date d'être neutre en carbone d'ici à 2030, ses émissions de CO₂ ayant bondi de 30 % l'année dernière. Cependant, l'entreprise ne souhaite pas inverser la tendance comme en témoigne son président qui affirme que « les bienfaits de l'IA pour le monde l'emporteront sur son impact environnemental ».

En outre, les centres de données des GAFAM sont presque déjà optimaux en ce qui concerne leur efficacité. Google affirme que le PUE de son parc mondial de centres de données était de 1,1 en 2023. Meta annonce 1,08 pour la même année. En 2023, les centres de données opérationnels de Google affichaient en moyenne une efficacité d'utilisation de l'eau (WUE pour *water usage efficiency*) de 0,18 litre/kWh, tandis que celle des centres de données était de 1,8 litre en moyenne pour 1 kWh⁵³. Cela soulève des inquiétudes quant à la marge de manœuvre des GAFAM pour continuer à réduire leurs émissions de carbone alors que la demande pour leurs services est en très forte croissance. En outre, si les émissions liées aux centres de données des GAFAM sont faibles grâce aux certificats d'énergie renouvelable, la prise en compte des émissions des centres de données de

53. « Data Center Water Cooling », Sunbird, disponible sur : www.sunbirddecim.com.

Google, Microsoft, Meta et Apple selon le mix moyen de leur lieu de production multiplie leurs émissions de gaz à effet de serre par 7,62⁵⁴.

Les centres de données sont en concurrence avec l'industrie ainsi qu'avec les secteurs du chauffage, du refroidissement et du transport pour sécuriser leur approvisionnement en électricité, ce qui fait grimper les prix de l'électricité. Parmi les dix premières entreprises acheteuses d'énergie propre avec des contrats d'achat d'électricité (abrégé PPA pour *power purchase agreement*) en 2023, on trouve Amazon, Meta, Alphabet et Microsoft. Amazon a même été le premier acheteur d'électricité solaire et éolienne au monde. Cette tendance devrait se poursuivre dans les années à venir, car ces entreprises ont besoin de sécuriser l'approvisionnement en électricité pour le développement de leurs activités. Au cours de la période 2024-2025, les 86 projets éoliens et solaires soutenus par Meta et répartis dans 24 états américains et 74 pays ajouteront 9 800 MW d'énergie renouvelable aux systèmes locaux.

Outre les contrats d'achat d'électricité pour les énergies renouvelables, les GAFAM s'intéressent également à l'énergie nucléaire, qu'il s'agisse des petits réacteurs modulaires (SMR) ou des centrales traditionnelles : Sam Altman, président d'OpenAI, préside également la société Oklo, spécialisée dans les SMR. Amazon, Microsoft et Google ont tous signé des partenariats avec des opérateurs nucléaires pour répondre à leurs besoins croissants en électricité. Amazon est également sur le point d'acquérir une participation dans la *start-up* de SMR X-energy. Pour les développeurs de SMR, les PPA conclus avec ces géants sont plus avantageux que ce que le marché de gros de l'électricité peut leur offrir, ce qui pousse les GAFAM à investir massivement le secteur. En ce qui concerne les centrales nucléaires existantes, l'opération la plus connue a été la réouverture de la centrale nucléaire de Three-Mile Island pour Microsoft, qui a permis à Constellation de conclure le PPA le plus important de son histoire. De même, la division *cloud computing* d'Amazon a accepté de dépenser 650 millions de dollars (M\$) pour acquérir un campus de centres de données connecté à la centrale nucléaire de Talen Energy Corp, vieille de 40 ans, située sur la rivière Susquehanna en Pennsylvanie.

Les développements majeurs dans le domaine de l'énergie nucléaire sont plus susceptibles de se matérialiser dans le futur (au-delà de dix ans) que dans le présent. Cela n'empêche pas les grandes entreprises technologiques d'appeler à une renaissance du nucléaire. Open AI, dans un document intitulé « Infrastructure Is Destiny », appelle à la construction de plusieurs centrales électriques dans différents états des États-Unis, afin de supporter des centres de données de 5 GW pour un coût d'environ

54. I. O'Brien, « Data Center Emissions Probably 662% Higher Than Big Tech Claims: Can it Keep Up the Ruse? », *The Guardian*, 25 septembre 2024.

100 Md\$⁵⁵. Quelques mois plus tard, ce projet a reçu un nom : Stargate. Avec plusieurs centres de données de 5 GW, il incite encore plus les GAFAM à développer et à acquérir de grandes centrales nucléaires, car les SMR ne suffiront pas pour des projets aussi titanesques. En France, les énergéticiens ont entendu cet appel. Alors que Emmanuel Macron annonçait le 9 février 2025 à l'occasion de l'ouverture du Sommet pour l'action sur l'IA un investissement de 109 Md€ dans l'IA dans les années à venir, le géant national EDF lançait un appel à manifestation d'intérêt auprès d'entreprises du numérique à qui ils promettent de fournir quatre sites industriels prêts à l'emploi, connectés au réseau électrique et alimentés par 2 GW de centrales nucléaires disponibles⁵⁶.

Conformément à ces tendances, les États-Unis mettent déjà en œuvre des politiques visant à répondre aux besoins des GAFAM, mais avec des résultats paradoxaux. La Virginie, l'État qui compte le plus grand nombre de centres de données dans le pays, a perdu 750 M\$ de recettes fiscales en raison des incitations liées aux centres de données en 2023. Dans l'Illinois, 468 M\$ de subventions pour les centres n'ont permis de créer que 339 postes, soit environ 1,4 M\$ par emploi⁵⁷. Ce programme de réduction des impôts et de facilitation rapide risque probablement d'être renforcé sous la nouvelle administration Trump. En effet, le lendemain de son retour à la Maison-Blanche, Donald Trump a annoncé un investissement de 500 Md\$ sur quatre ans, avec un engagement immédiat de 100 Mds\$ pour une co-entreprise appelée Stargate, réunissant OpenAI, le leader des logiciels Oracle et la banque d'investissement japonaise SoftBank, afin de renforcer la domination des États-Unis dans le domaine de l'IA. Les premiers centres de données sont déjà en construction au Texas.

Tous ces éléments peuvent conduire à un rôle inévitable des GAFAM dans l'avenir des systèmes électriques, en particulier si la question de la flexibilité qu'ils promettent est résolue par la R&D. Si les centres de données s'avèrent flexibles, avec des accords d'achat d'électricité record dans les énergies renouvelables, l'énergie nucléaire, les *start-ups* SMR – en plus d'une grande flotte de véhicules électriques à charge bidirectionnelle –, les GAFAM pourraient devenir des acteurs importants dans la gestion de la demande électrique et joueraient un rôle indispensable dans les équilibres électriques et les questions politiques liées au réseau.

Au début du mois de novembre 2024, Abou Dabi a accueilli un sommet sur l'énergie et l'IA au cours duquel les grandes compagnies pétrolières ont reconnu que « l'essor rapide de l'intelligence artificielle incite fortement les

55. « Infrastructure Is Destiny: Economic Returns on US Investment in Democratic AI », OpenAI, septembre 2024.

56. « EDF accompagne les entreprises du numérique pour le développement de nouveaux centres de données en France », Communiqué de presse EDF, 10 février 2025, disponible sur : www.edf.fr.

57. J. Whiton, « New Data on Data Center Subsidies, Same Old Problems », Good Jobs First, 25 juillet 2023, disponible sur : <https://goodjobsfirst.org>.

plus grandes compagnies pétrolières du monde à accroître leurs investissements dans les énergies renouvelables⁵⁸ ». On dit même que Shell et BP, qui s'étaient retirées des énergies renouvelables au cours des deux dernières années pour se recentrer sur leurs activités principales dans le domaine du pétrole et du gaz, envisagent de réinvestir dans les énergies renouvelables.

Les grandes sociétés pétrolières se félicitent également de l'énorme croissance de la demande d'électricité des centres de données, qui entraîne une relance des centrales électriques à combustibles fossiles, notamment à gaz avec ou sans captation de carbone. L'IA devrait donner un coup de boost à la demande de gaz naturel, qui est le combustible le plus susceptible de combler le vide laissé par l'absence de pilotabilité des sources d'énergie renouvelables, et ce dans un contexte où le charbon continue d'être progressivement éliminé des bouquets électriques en Occident⁵⁹. Dans certains pays, dont l'Irlande, l'Arabie saoudite et la Malaisie, l'énergie nécessaire pour faire fonctionner à plein régime tous les centres de données qu'ils prévoient de construire dépasse l'offre disponible d'énergie renouvelable. La plupart des centres de données de Malaisie ne sont pas encore en service, mais si l'on tient compte de tout ce qui est en construction, la quantité d'électricité utilisée par les seuls centres de données dépassera largement la production totale d'énergie renouvelable du pays en 2022, dernière année pour laquelle des données sont disponibles⁶⁰.

L'avenir des centres de données

Les centres de données poursuivront leur expansion et augmenteront leur consommation d'électricité. En outre, ils continueront à être de plus en plus concentrés, malgré le déploiement des centres de données périphériques. Il n'y a aucune raison pour que les États-Unis perdent leur domination (ou leur dépendance) en matière de centres de données. Grâce à la rapidité des permis de construire, aux réductions d'impôts et à la déréglementation promises par le président Trump et son ministère de l'Efficacité gouvernementale (DOGE) mené par Elon Musk, il est peu probable que la tendance actuelle s'inverse. La « main invisible » du marché agira et les centres de données seront de plus en plus concentrés dans les 15 premiers états, en particulier au Texas et en Virginie où la consommation d'électricité des centres de données en 2030 pourrait atteindre 120 TWh par an dans chaque état (soit l'équivalent de la consommation actuelle de tous les centres de données américains), selon le scénario de croissance le plus élevé de l'Electric Power Research Institute

58. M. Moore, « "AI Huge Power Gives Oil Major Incentives to Invest in Renewables", says Adnoc Boss », *Financial Times*, 4 novembre 2024, disponible sur : www.ft.com.

59. « AI revolution: Meeting Massive AI Infrastructure Demands », Barclays Research, 16 janvier 2025.

60. J. Saul *et al.*, « AI Is Already Wreaking Havoc on Global Power Systems », *op. cit.*

(EPRI). En Virginie, les centres de données pourraient représenter la moitié de la consommation d'électricité de l'état.

En Europe, la demande d'électricité pour les centres de données augmentera dans deux types de pays⁶¹. Le premier type est celui des pays disposant d'une énergie abondante et bon marché provenant de sources nucléaires, hydroélectriques, éoliennes ou solaires (les pays nordiques, l'Espagne et la France). Le second type de pays comprendra les pays qui sont des centres financiers et technologiques et qui offrent des allègements fiscaux ou d'autres incitations pour attirer les *data centers* (Allemagne, Royaume-Uni et Irlande). En effet, concernant l'IA, ce sont les consommateurs qui décideront de l'emplacement des centres de données, à savoir là où se trouvent l'industrie et la finance. Et la proximité des utilisateurs est importante dans ces secteurs très compétitifs et modernes, car des retards de quelques millisecondes peuvent avoir des coûts de productivité conséquents à long terme.

Les centres de données de périphérie (*edge data centers*) connaîtront une croissance régulière d'environ 20 % au cours des cinq prochaines années. Ils serviront principalement utilisés pour les villes intelligentes (*smart cities*) et le *trading*. Toutefois, leur déploiement sera trop lent pour compenser les contraintes de manque de fiabilité des réseaux causées par les centres de données à grande échelle. En effet, le déploiement de la 5G réduit le compromis entre la distance et la latence, de sorte que l'IA non spécialisée et les centres de données peuvent désormais être situés n'importe où dans le monde, ce qui encourage à la fois la colocation et les *hyperscalers* pour les économies d'échelle qu'ils permettent de réaliser.

Les centres de données vont ainsi continuer à grossir. Alors que d'énormes centres de données d'une capacité proche de 100 MW sont déjà en service, Meta a annoncé en janvier 2025 qu'elle prévoyait de construire un centre de données de 2 GW, qui serait presque aussi grand que la ville de Paris et nécessiterait au moins deux réacteurs nucléaires pour être alimenté à pleine capacité. Cette tendance dépasse même la croissance tendancielle de 0,5 ordre de grandeur par an, déjà difficile à suivre, des *data centers* spécialisés dans l'IA⁶². Si cette tendance se poursuit, nous pourrions assister au fonctionnement d'un centre de données de 10 GW en 2028, qui nécessiterait à lui seul l'équivalent de la consommation électrique moyenne du Portugal. Étant donné que le premier centre de données Stargate de 5 GW devrait également fonctionner en 2028, les centres de données sont sur la bonne voie pour poursuivre leur croissance exponentielle basée sur cette tendance.

Toutefois, les contraintes qui pèsent sur les réseaux peuvent conduire à de nouvelles formes de politiques et à une utilisation plus démocratique de l'IA. Chris Lehane, responsable de la politique d'OpenAI, a lancé l'idée

61. « AI Is Poised to Drive 160 % Increase in Data Center Power Demand », *op. cit.*

62. L. Aschenbrenner, « Situational Awareness: The Decade Ahead », juin 2024.

qu'en échange d'incitations et de facilitations réglementaires pour la connexion au réseau, « [il lui semblait] très raisonnable de demander que 20 ou 25 % du calcul soit mis à la disposition [du] système d'éducation publique, [des] universités publiques, afin de créer un pôle d'IA en rapport avec les industries déjà [présentes] dans un État⁶³ ». Il reste à voir si cette proposition se concrétisera dans les prochaines années.

Concernant les centres de données, l'un des principaux sujets pour 2025 sera l'IA frugale et la sobriété. Le déploiement d'une IA frugale et spécifique est une caractéristique essentielle pour atténuer l'explosion de la demande d'électricité des *data centers* liés à l'IA. Nous entendons par IA frugale moins de paramètres dans les modèles (grâce au *pruning* notamment, c'est-à-dire à la suppression des données non pertinentes ou redondantes pour améliorer l'efficacité) et moins d'entraînement. Cela peut être réalisé principalement en utilisant des modèles de préformation au lieu de modèles formés de bout en bout, produisant le même résultat avec beaucoup moins d'entraînement et donc d'énergie gaspillée.

En ce qui concerne la sobriété, la France a accueilli le Sommet pour l'action sur l'IA les 10 et 11 février 2025 – le troisième sommet international sur l'IA, après le Sommet sur la sécurité de l'IA à Bletchley Park (Royaume-Uni) en novembre 2023 et un autre sommet à Séoul en mai 2024. Parmi les trois principaux sujets abordés figurent la durabilité de l'IA et une initiative politique visant à répondre à la plus grande prise de conscience des externalités négatives de l'IA sur l'environnement. Le ministère français de la Transition écologique et la présidence ont cherché à aborder quatre aspects de la durabilité de l'IA : l'IA frugale ; l'IA spécifique ; l'évaluation environnementale ; le recyclage et l'économie circulaire.

Si les États-Unis ont une avance absolue dans l'infrastructure des centres de données, la *start-up* chinoise DeepSeek a lancé fin janvier 2025 son LLM, qui a complètement bouleversé le monde de l'IA. En promettant d'être aussi « intelligent » que ChatGPT, DeepSeek a rebattu les cartes du grand jeu géopolitique de l'IA⁶⁴. Non seulement il prouve que la Chine peut et va rivaliser avec les États-Unis pour la suprématie de l'IA (ce qui a conduit certains à parler d'un « moment Spoutnik »), mais il indique également que la qualité peut être obtenue par la sobriété, tant au niveau du financement que de la conception. DeepSeek affirme n'avoir utilisé que 2 000 puces pour entraîner la troisième version de son modèle, soit beaucoup moins que les 16 000 puces sur lesquelles ChatGPT a été entraîné, bien que ce chiffre soit discutable et ait suscité le scepticisme dans le monde de l'IA. Le modèle DeepSeek ne nécessite apparemment qu'une

63. D. Lee, "Sam Altman's Energy 'New Deal' Is Good for AI. What About Americans?", *Bloomberg*, 17 octobre 2024.

64. B. Pajot, « Un "moment DeepSeek" ? », *Briefings de l'Ifri*, Ifri, 26 mars 2025, disponible sur : www.ifri.org.

petite quantité d'énergie pour fonctionner, et sa phase de pré-entraînement n'a coûté que 6 millions de dollars. Cependant, ce chiffre, largement diffusé par les médias, ne représente qu'une faible partie du coût total du développement de ce modèle. On estime que les dépenses totales en capital de l'entreprise se sont élevées à 1,6 Md\$, dont plus de 500 M\$ en dépenses matérielles (notamment en GPU⁶⁵). De même, la baisse des coûts des inférences de l'IA suit la tendance d'une amélioration du calcul algorithmique entre quatre et dix par an. Depuis la publication de ChatGPT, les coûts par instance de calcul ont ainsi été divisés par 1 200. L'IA nous réserve des prodiges et des mystères qui restent à découvrir.

65. D. Patel *et al.*, « DeepSeek Debates: Chinese Leadership on Cost, True Training Cost, Closed Model Margin Impacts », SemiAnalysis, 31 janvier 2025, disponible sur : <https://semianalysis.com>.

Conclusion

Le Sommet pour l'action sur l'IA a été marqué par une vague d'annonces concernant les investissements dans l'IA, avec 109 Md€ engagés en France et 200 Md€ dans l'ensemble de l'UE. Ces investissements soutiendront notamment la construction de centres de données sur le sol européen. Or, selon un rapport de McKinsey, la demande énergétique des centres de données de l'UE et du Royaume-Uni réunis pourrait atteindre jusqu'à 150 TWh d'ici 2030, soit l'équivalent d'un tiers de la consommation finale d'électricité de la France en 2024. Par conséquent, le sommet a également servi de plateforme pour discuter des principaux défis environnementaux et énergétiques liés à l'essor de ces technologies.

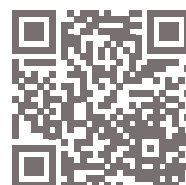
Bien que cette question ne fasse pas explicitement partie des cinq thèmes principaux du sommet, elle est apparue comme une préoccupation sous-jacente dans la plupart des discussions tenues à Paris les 10 et 11 février 2025. La promotion du développement d'une IA « durable » est apparue comme l'une des trois priorités clés pour l'action future. Le sommet avait également pour objectif de déclencher une dynamique mondiale visant à rendre l'IA plus respectueuse de l'environnement. Cependant, la demande massive d'énergie liée au déploiement à grande échelle de l'IA reste un défi fondamental, qui a donné lieu à plusieurs annonces importantes.

Tout d'abord, la France, le Programme des Nations unies pour l'environnement (PNUE) et l'Union internationale des télécommunications ont lancé une coalition pour une IA « écologiquement durable ». Cette coalition soutiendra et amplifiera les initiatives visant à réduire l'empreinte énergétique et environnementale de l'IA, en veillant à ce que sa croissance s'aligne sur des objectifs plus larges en matière d'énergie et de climat. En outre, l'Agence internationale de l'énergie a présenté le premier Observatoire mondial sur l'IA et l'énergie, conçu pour mieux anticiper les besoins énergétiques des modèles d'IA et des centres de données, optimiser les systèmes énergétiques et tirer parti de l'IA pour réduire les émissions de carbone et réaliser des innovations de rupture dans les applications énergétiques. Enfin, le PNUE publiera, en mars 2025, un ensemble de lignes directrices visant à orienter les investissements publics et privés vers des centres de données économes en énergie.

Cette *Note* vise à éclairer ce débat en présentant des prévisions crédibles sur la demande d'énergie liée à l'IA et sur les centres de données qu'elle requiert. Elle a mis en évidence la faible probabilité que les gains d'efficacité en cours compensent entièrement l'augmentation de la consommation d'électricité due à l'expansion de l'utilisation des centres de

données, en raison de l'effet de rebond. L'IA entraînera une hausse de la demande d'électricité, mais il est peu probable qu'elle provoque une consommation d'électricité exponentielle inarrêtable, car sa croissance sera probablement limitée par plusieurs facteurs : les contraintes et les files d'attente du réseau, les capacités supplémentaires de production d'électricité, la guerre commerciale sur un marché des puces très concentré, et son taux d'adoption, qui dépend des délais de formation de la main-d'œuvre et des modes spécifiques d'intégration de l'IA dans les industries.

En fin de compte, la question reste de savoir si l'IA deviendra la prochaine force motrice de l'industrie. L'importance de ses effets sur le système énergétique dépendra en grande partie du rythme auquel les entreprises et les travailleurs adopteront cette technologie.



27 rue de la Procession 75740 Paris cedex 15 – France

Ifri.org