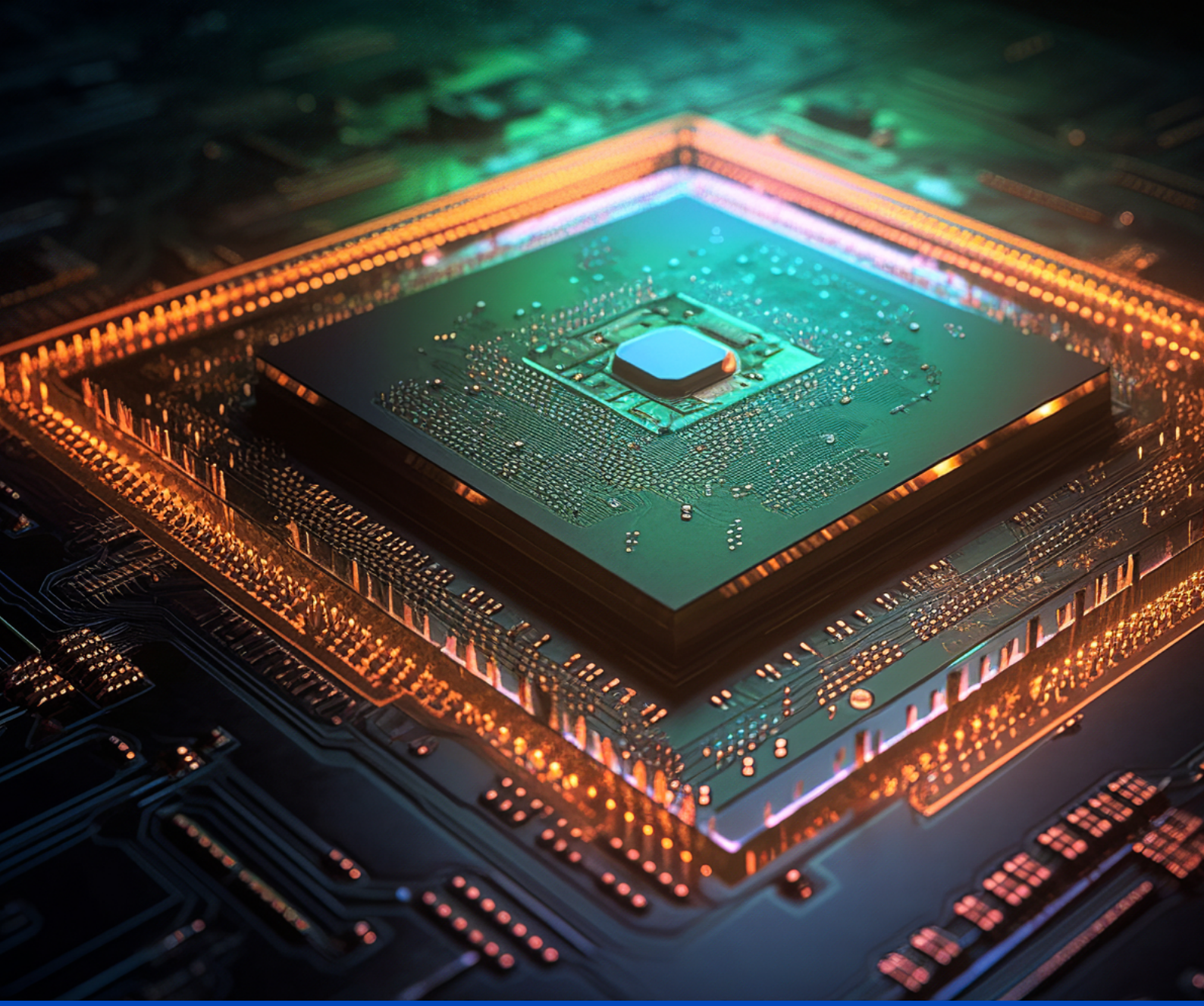


L’argument en faveur de l’infrastructure IA

WisdomTree®

WisdomTree.eu
+44 (0) 207 448 4330

L’argument en faveur de l’infrastructure IA : la base physique d’un avenir numérique

L’infrastructure d’intelligence artificielle (IA) devient une couche fondamentale de l’économie mondiale. À mesure que l’intelligence des modèles progresse et que les agents IA commencent à prendre en charge une part des tâches numériques, la demande en capacité de calcul augmente rapidement.

La valeur ne se situe plus uniquement au niveau des GPU[1] ou des modèles, mais dans l’écosystème informatique qui les entoure et les rend possibles. L’IA évolue vers un système physique reposant sur l’énergie, le matériel et des systèmes et chaînes d’approvisionnement de plus en plus complexes. Cela entraîne une expansion durable dans les domaines suivants :

Semi-conducteurs et fabrication

Réseaux

Centres de données et énergie

1 GPU : Graphics Processing Unit. Initialement conçus pour le rendu graphique, les GPU sont désormais les principaux processeurs utilisés pour l’entraînement et l’inférence IA grâce à leurs capacités de traitement parallèle.

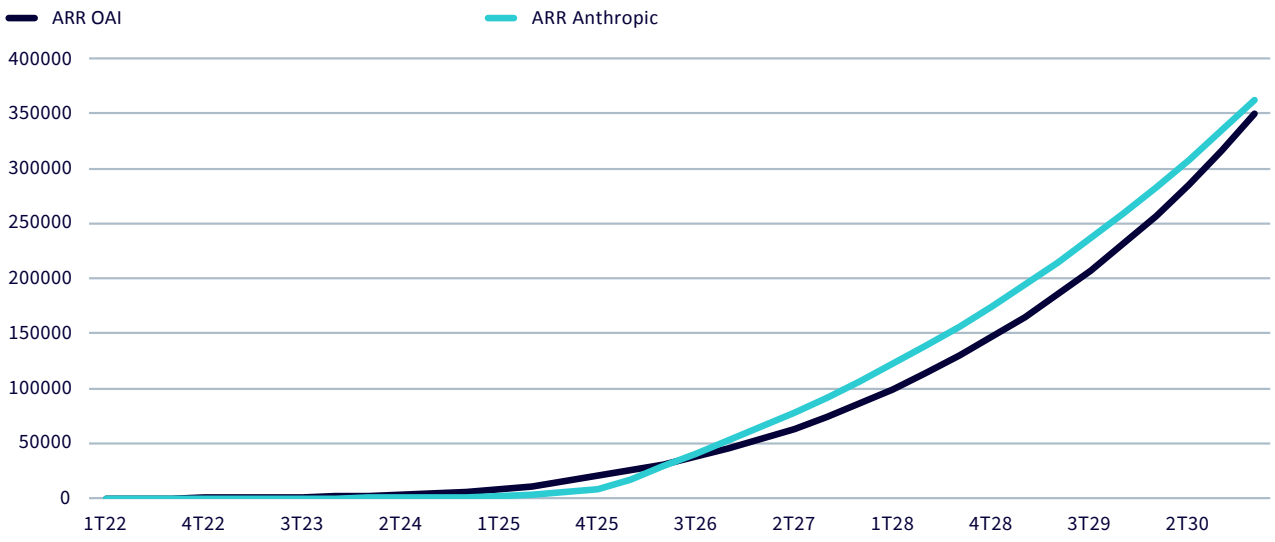
Demande d’inférence et montée en puissance du calcul

La caractéristique déterminante de ce cycle est la nature de la demande. La demande en capacité de calcul IA est désormais visible et persistante dans les usages professionnels comme grand public. L’entraînement d’un modèle nécessite une capacité de calcul immense sur une période donnée, tandis que l’inférence est continue.

Au niveau le plus élémentaire, cette demande se mesure en tokens, unités de texte, de code ou de données traitées par les modèles IA. Chaque requête, flux de travail ou tâche exécutée par un agent génère des tokens, chacun nécessitant du calcul pour être traité. À mesure que l’usage augmente, la génération de tokens croît proportionnellement, créant une demande structurelle en capacité de calcul.

Cela se reflète dans la croissance des principaux laboratoires de modèles IA, les « producteurs de tokens IA ».

Figure 1 : Revenus annualisés des principaux fournisseurs de LLM[2]



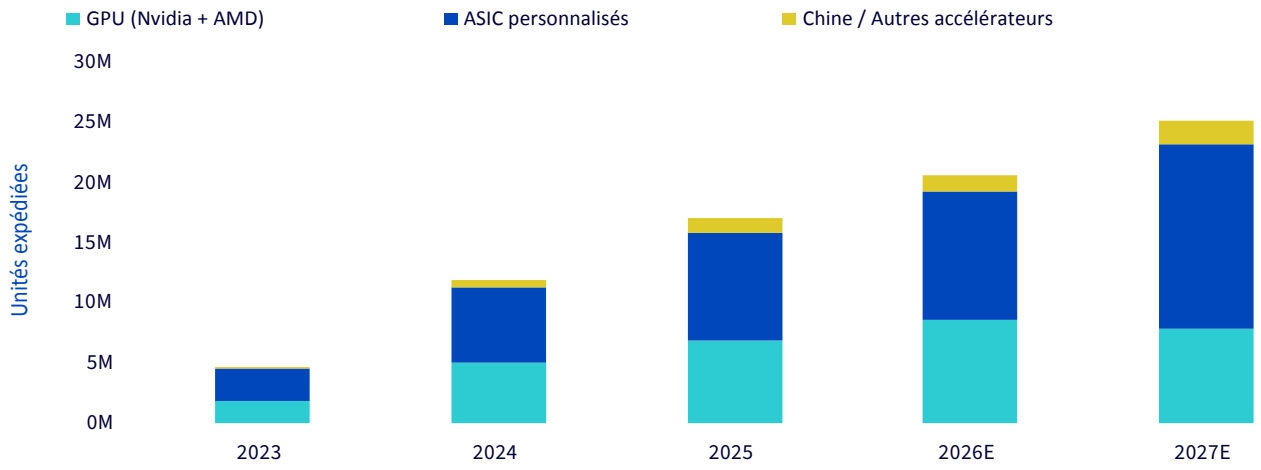
Source : SemiAnalysis, avril 2026. **Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.**

2 LLM : Large Language Model. Type de modèle IA entraîné sur de vastes ensembles de textes pour générer, comprendre et raisonner sur le langage. Exemples : la série GPT d’OpenAI et Claude d’Anthropic.

Les revenus récurrents annualisés (ARR)[3] des fournisseurs de modèles IA, tels qu'OpenAI et Anthropic, sont passés de zéro à plus de 30 milliards de dollars en quelques années seulement. Le rythme d'adoption dépasse les cycles technologiques[4] précédents et suit une courbe de croissance exponentielle, les prévisions de SemiAnalysis atteignant plus de 300 milliards de dollars d'ARR d'ici 2030.

Les hyperscalers[5] et les fournisseurs de neocloud[6] fournissent une grande partie de la capacité de calcul utilisée par ces laboratoires IA, tout en développant leurs propres modèles concurrents. Les deux groupes engagent des centaines de milliards en dépenses d'investissement pour accroître leurs capacités. Il s'agit de décisions prospectives avec des horizons pluriannuels, axées sur le déploiement de puces de plus en plus spécialisées pour une efficacité maximale. Cela se traduit par de nouveaux designs de puces et une augmentation des commandes auprès des partenaires de fabrication.

Figure 2 : Expéditions d'unités XPU[7]



Source : SemiAnalysis Accelerator Model, avril 2026. **Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.**

3 ARR : Annualised Recurring Revenue. Mesure normalisée des revenus récurrents et prévisibles sur une base annuelle, couramment utilisée pour suivre la croissance des entreprises technologiques par abonnement.

4 ChatGPT a été largement reconnu comme l'application grand public ayant connu la croissance la plus rapide de l'histoire, atteignant 100 millions d'utilisateurs actifs mensuels en seulement deux mois.

5 Les hyperscalers sont de grands fournisseurs de services cloud exploitant d'immenses infrastructures de centres de données à l'échelle mondiale.

6 Les neoclouds sont des fournisseurs de services cloud spécialisés dans l'IA, proposant des solutions GPU-as-a-Service haute performance pour l'entraînement et l'inférence.

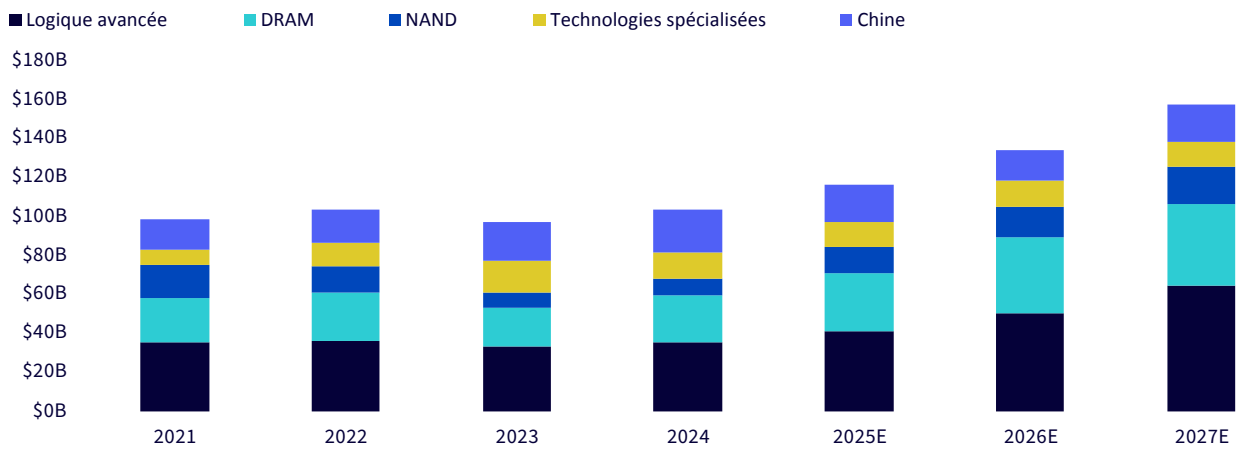
7 XPU : Terme générique désignant tout processeur accélérateur utilisé dans les charges de travail IA, incluant GPU, ASIC personnalisés et autres siliciums spécialisés.

Les expéditions d'accélérateurs IA devraient passer d'environ 5 millions d'unités en 2023 à plus de 25 millions en 2027, soit une multiplication par cinq sur la période. Cela reflète à la fois le besoin croissant de soutenir les charges de calcul IA et l'élargissement de l'écosystème au-delà d'un fournisseur ou d'une architecture unique.

Montée en puissance des semi-conducteurs et goulets d'étranglement dans la chaîne d'approvisionnement

Répondre à cette demande en puces nécessite une expansion majeure de la fabrication de semi-conducteurs. Aujourd'hui, les fonderies n'ont pas la capacité suffisante pour produire les volumes requis par les clients. Accroître cette capacité implique de construire de nouvelles installations et d'investir dans les équipements spécialisés, appelés Wafer Fabrication Equipment (WFE), utilisés à chaque étape de la transformation d'une tranche de silicium brute en puce finie.

Figure 3 : Dépenses mondiales en WFE[8] par application



Source : SemiAnalysis WFE Model, février 2026. **Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.**

Les dépenses mondiales en WFE devraient approcher 160 milliards de dollars d'ici 2027, reflétant un changement structurel vers une expansion des capacités tirée par l'IA. Malgré cette hausse, des contraintes majeures persistent, notamment en lithographie avancée[9] et en capacité de salles blanches[10], ce qui continue de limiter les volumes de production.

8 WFE : Wafer Fabrication Equipment. Machines spécialisées utilisées pour fabriquer des puces semi-conductrices, incluant les systèmes de lithographie, de dépôt, de gravure et d'inspection.

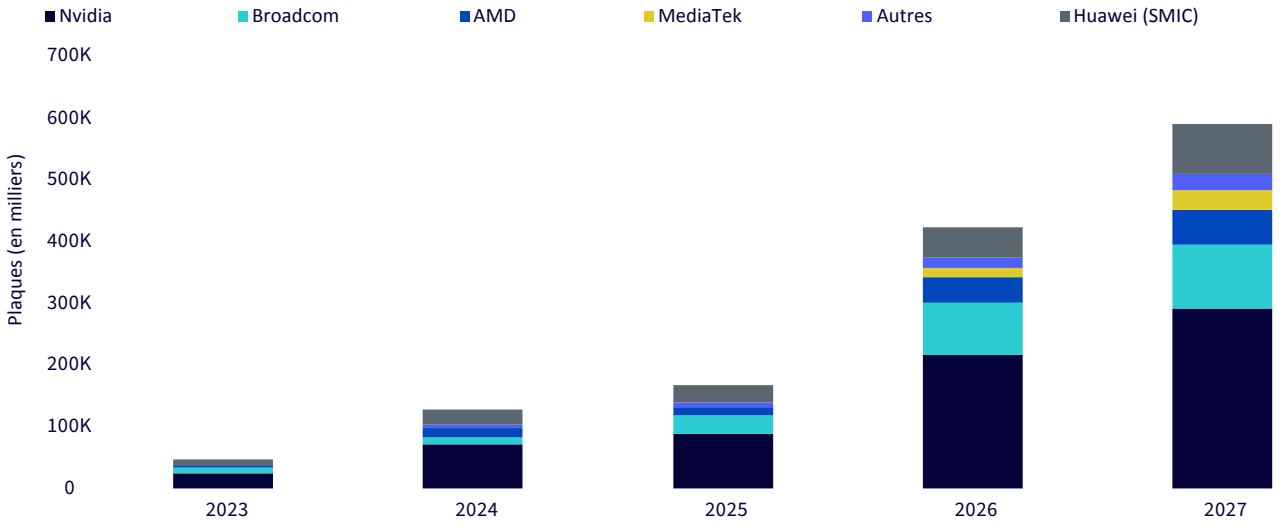
9 Lithographie : Processus consistant à transférer des motifs de circuits sur une tranche de silicium à l'aide de lumière. En particulier, la lithographie EUV (ultraviolet extrême) est essentielle pour fabriquer les puces IA les plus avancées et demeure l'une des étapes les plus capitalistiques de la fabrication de semi-conducteurs.

10 Une salle blanche est un espace utilisé dans la fabrication de semi-conducteurs, caractérisé par des niveaux extrêmement faibles de particules, de variations de température, d'humidité et d'autres formes d'instabilité environnementale. De petites particules ou des écarts aux conditions idéales peuvent ruiner une tranche de silicium ou même un lot entier de puces.

Parallèlement, les puces elles-mêmes deviennent plus complexes. Chaque nouvelle génération d’accélérateurs IA nécessite davantage de mémoire et une bande passante accrue pour suivre la hausse des besoins en calcul, tout en respectant des contraintes similaires d’espace et d’énergie.

Cette complexité croissante signifie que les composants ne peuvent plus simplement être placés côte à côte dans un boîtier classique. Ils doivent plutôt être intégrés de manière étroite grâce à des techniques d’assemblage avancées comme le CoWoS (Chip-on-Wafer-on-Substrate), qui empilent et connectent plusieurs chiplets en une seule unité. En conséquence, l’assemblage avancé est devenu l’un des goulets d’étranglement les plus immédiats de la chaîne d’approvisionnement.

Figure 4 : Expéditions de tranches CoWoS[11] par société



Source : SemiAnalysis WFE Model, février 2026. Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.

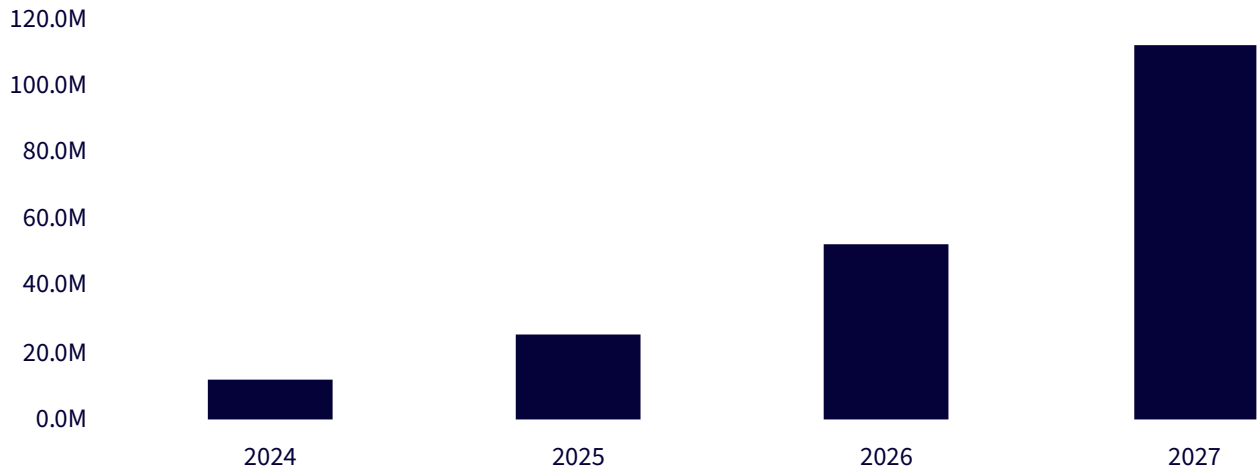
Les expéditions de tranches CoWoS devraient être multipliées par plus de dix entre 2023 et 2027, mais la demande continue de dépasser l’offre. Les concepteurs de puces réservent désormais des capacités d’assemblage plusieurs années à l’avance. Les entreprises disposant de ces capacités spécialisées occupent un point de passage critique, l’accès à l’assemblage avancé déterminant directement qui peut produire les puces les plus avancées.

11 CoWoS : Chip-on-Wafer-on-Substrate. Technologie avancée d’assemblage développée par TSMC permettant d’intégrer plusieurs puces sur un même substrat, essentielle pour les accélérateurs IA haute performance.

Contraintes au niveau des réseaux et des systèmes

Les accélérateurs ne sont pas les seules puces importantes. À mesure que les systèmes IA montent en échelle, des milliers de processeurs doivent fonctionner ensemble, et l’infrastructure réseau qui les relie devient tout aussi essentielle. Les données doivent circuler entre les puces à des vitesses et volumes considérables, ce qui rend les puces réseau, les émetteurs-récepteurs optiques et les équipements de commutation essentiels au système. Si le réseau ne suit pas, même les accélérateurs IA les plus puissants restent inutilisés.

Figure 5 : Demande en émetteurs-récepteurs IA



Source : SemiAnalysis, avril 2026. Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.

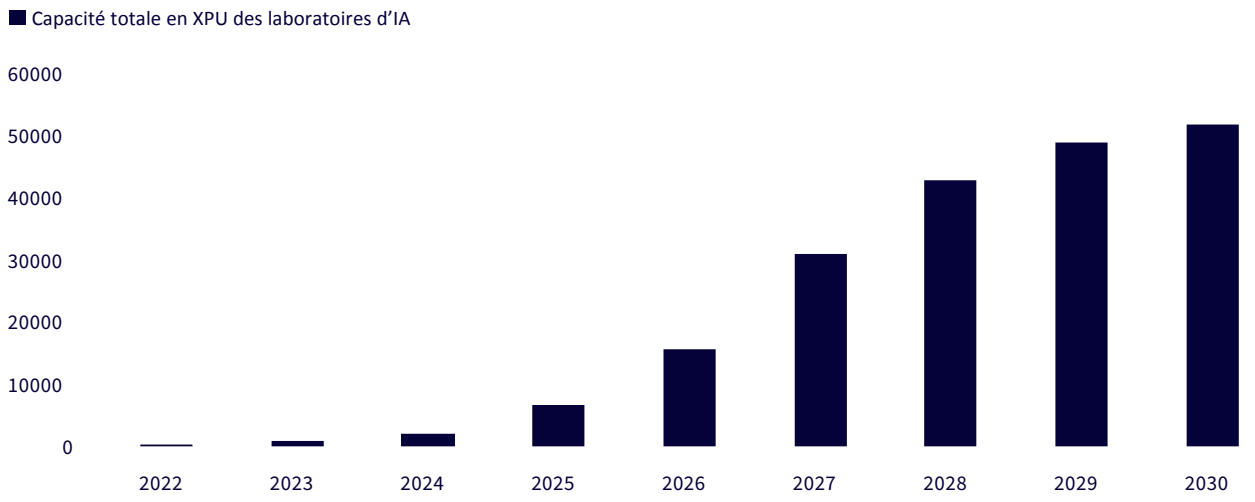
Ce qui relie souvent ces processeurs à haute vitesse dans les centres de données, ce sont les émetteurs-récepteurs optiques 800G[12]. La demande pour ces composants devrait passer de 25 millions d’unités en 2025 à 112 millions en 2027, tandis que les liaisons de nouvelle génération 1,6T montent également en puissance rapidement. Cette croissance reflète le fait que le réseau devient un déterminant de plus en plus important de la performance et du coût global des systèmes.

Demande énergétique et expansion des centres de données

Même si les puces et les réseaux sont disponibles, ils doivent pouvoir fonctionner quelque part. Chaque accélérateur nécessite de l’énergie, du refroidissement et un espace physique en centre de données pour fournir des tokens aux utilisateurs finaux. L’ampleur de cette demande devient tangible lorsqu’on l’observe sous l’angle de la consommation d’énergie, les charges d’inférence IA entraînant une hausse durable de la demande énergétique.

12 800G : Désigne les émetteurs-récepteurs optiques 800 gigabits par seconde, la norme haut débit actuelle utilisée dans les centres de données IA pour transférer les données entre serveurs et commutateurs. Les émetteurs-récepteurs 1,6T (1 600 Gb/s) de nouvelle génération commencent à être déployés pour les clusters IA à très haut débit.

Figure 6 : Capacité XPU des fournisseurs de LLM (MW[13])



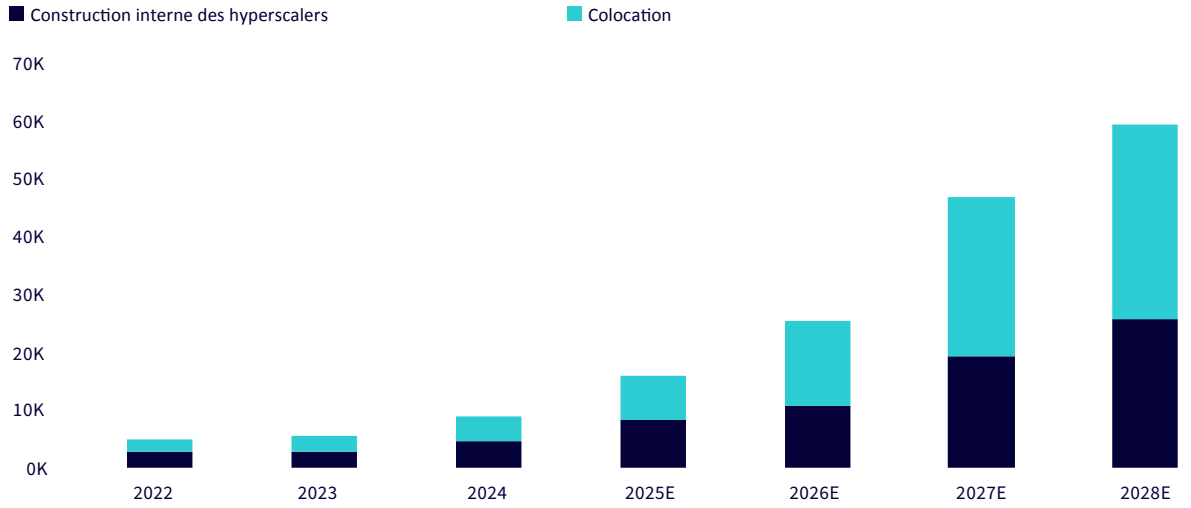
Source : SemiAnalysis AI Datacentre Industry Model, avril 2026. Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.

La demande énergétique liée à l’inférence devrait dépasser largement 50 000 MW d’ici 2030. À cette échelle, l’IA ne constitue plus une charge marginale pour les infrastructures énergétiques traditionnelles. Elle devient un moteur de demande énergétique à l’échelle du système et, par conséquent, ces infrastructures se comportent davantage comme des processus industriels. Cela se reflète dans le terme « usines d’IA », utilisé par Jensen Huang, CEO de NVIDIA, pour décrire les centres de données IA modernes.

La hausse des besoins en calcul et en énergie impose une expansion rapide de la capacité des centres de données. La seule manière de répondre à cette demande est de mettre en service davantage d’usines d’IA.

13 MW = megavatio.

Figure 7 : Ajouts mondiaux de capacité en MW dans les centres de données (hors Chine)



Source : SemiAnalysis AI Datacentre Industry Model, avril 2026. Les prévisions ne constituent pas un indicateur de performance future et tout investissement comporte des risques et des incertitudes. Les performances passées ne préjugent pas des performances futures et tout investissement peut perdre de la valeur.

Les ajouts de capacité des centres de données devraient passer d’environ 16 000 MW en 2025 à près de 60 000 MW en 2028. Les hyperscalers construisent leurs propres installations et s’associent aux neoclouds et à d’autres opérateurs de centres de données pour mettre rapidement de nouvelles capacités en ligne.

Toutefois, les infrastructures électriques et de réseau fonctionnent sur des horizons pluriannuels et deviennent la contrainte déterminante pour la vitesse de déploiement des nouvelles capacités. Pour contourner cette limite, les opérateurs se tournent de plus en plus vers la production d’énergie sur site — turbines à gaz, systèmes motorisés, piles à combustible — privilégiant la rapidité plutôt que la dépendance au réseau. En définitive, la capacité à déployer l’IA à grande échelle dépend directement de la disponibilité des infrastructures physiques, et les centres de données ainsi que l’énergie devraient constituer un goulet d’étranglement majeur au cours de la prochaine décennie.

Le cycle d’investissement dans l’infrastructure IA est en cours

L’IA évolue vers un système physique, et l’infrastructure nécessaire pour la soutenir couvre les semi-conducteurs et leurs chaînes d’approvisionnement, les réseaux, ainsi que les centres de données et les systèmes énergétiques. La demande s’accélère à mesure que l’IA passe en production dans le monde entier, tandis que les contraintes à chaque niveau de la pile deviennent plus visibles.

L’IA alimente l’un des plus importants cycles d’investissement de l’histoire du secteur technologique. La demande s’accumule, mais le rythme de déploiement dépend non seulement de cette demande, mais surtout de la vitesse à laquelle l’infrastructure physique peut être construite. Cela prolonge la durée du cycle et positionne l’infrastructure IA comme un thème d’investissement majeur pour la décennie à venir.

Informations importantes

Communications commerciales publiées dans l’Espace économique européen (« EEE ») : Ce document est publié et approuvé par WisdomTree Ireland Limited, une société autorisée et réglementée par la Central Bank of Ireland.

Communications commerciales émises dans des juridictions en dehors de l’EEE : Ce document est publié et approuvé par WisdomTree UK Limited, une société autorisée et réglementée par la Financial Conduct Authority du Royaume-Uni.

WisdomTree Ireland Limited et WisdomTree UK Limited sont toutes les deux désignées comme « WisdomTree » (le cas échéant). Notre Politique sur les conflits d’intérêts et notre Inventaire sont disponibles sur demande.

Les informations figurant dans ce document sont fournies à titre informatif et ne constituent pas une offre de vente, ou une sollicitation d’offre d’achat de titres ou d’actions. Ce document ne doit pas être utilisé comme fondement d’une décision d’investissement. La valeur des investissements peut fluctuer et vous êtes susceptible de perte tout ou partie du montant investi. La performance passée ne constitue pas nécessairement une indication des performances futures. Toute décision d’investissement doit être fondée sur les informations figurant dans le prospectus approprié et sur des conseils indépendants en matière d’investissement, fiscaux et juridiques.

L’application des réglementations et lois fiscales peut souvent conduire à des interprétations différentes. Tous les points de vue ou opinions exprimés dans cette communication représentent les points de vue de WisdomTree et ne doivent pas être interprétés comme des conseils réglementaires, fiscaux ou juridiques. WisdomTree ne donne aucune garantie ou représentation quant à l'exactitude des vues ou opinions exprimées dans cette communication. Toute décision d'investissement doit être fondée sur les informations contenues dans le prospectus approprié et après avoir sollicité des conseils indépendants en matière d'investissement, fiscaux et juridiques.

Ce document n’est pas et ne doit en aucun cas être interprété comme une publicité ou une offre publique d’actions ou de titres aux États-Unis ou dans toute province ou tout territoire des États-Unis. L’introduction, la transmission et la distribution (directes ou indirectes) de l’original ou d’une copie de ce document sont interdites aux États-Unis.

Bien que WisdomTree s’efforce d’assurer l’exactitude du contenu de ce document, WisdomTree ne peut en garantir l’exactitude. Les fournisseurs de données tiers sollicités pour obtenir les informations contenues dans le présent document ne donnent aucune garantie ou représentation de quelque sorte en rapport avec ces données. Lorsque WisdomTree exprime ses propres opinions concernant le produit ou l’activité du marché, ces opinions sont susceptibles de changer. WisdomTree, ses affiliés et leurs dirigeants, directeurs, partenaires ou employés respectifs déclinent toute responsabilité pour toute perte directe ou indirecte découlant de l’utilisation de ce document ou de son contenu.

WisdomTree.eu
+44 (0) 207 448 4330