General Purpose AI Policy Lab

PUISSANCE DE CALCUL

Chaîne d'approvisionnement et dépendances stratégiques pour le développement de l'IA

Juillet 2026

#6 - Note

General
Purpose AI
Policy Lab

RÉSUMÉ EXÉCUTIF

— CONTEXTE —

Le développement d'IA de frontière aux États-Unis et en Chine, plus performantes, générales et agentiques, expose la France et la communauté internationale à des problèmes de perte de contrôle irréversibles menaçant la sécurité nationale. Or ce développement repose sur une ressource physique centrale : la puissance de calcul. Les États-Unis dominant les deux voies d'accès à cette ressource (96 % de la production de GPU et 86 % de la capacité déployée), tout projet massif d'IA se fait *de facto* avec leur accord. Mais la chaîne de production repose aussi sur des goulets d'étranglement non américains, dont le monopole européen ASML–Zeiss–Trumpf sur la photolithographie EUV, technologie nécessaire à la fabrication de tout GPU de pointe. Ce goulet confère à l'Europe un levier stratégique concret et actionnable sur le développement mondial de l'IA.

Tom David, Président-Directeur Général
du General-Purpose AI Policy Lab

— QUESTIONS —

- Quel est le rôle de la puissance de calcul dans le développement de l'IA, et comment s'organise sa chaîne d'approvisionnement ?
- Quelles dépendances stratégiques cette chaîne d'approvisionnement crée-t-elle, et quels leviers confère-t-elle aux différentes puissances ?

— SYNTHÈSE —

Les performances des IA à usage général dépendent directement de la quantité de calcul investie dans leur entraînement, ce qui fait des GPU, où sont réalisés ces calculs, le facteur principal de leur développement.

La chaîne de production des GPU de pointe est marquée par des goulets d'étranglement difficilement substituables : monopole américain sur la conception, monopole de TSMC à Taïwan sur la fabrication, monopole européen ASML–Zeiss–Trumpf sur les équipements de photolithographie EUV, dépendances au Japon et en Corée du Sud.

Une filière chinoise domestique se développe en parallèle, actuellement en retard sur tous les maillons (~ 3–4 ans en conception, ~ 7 ans en fabrication, ~ 15–20 ans en photolithographie), mais un retard aujourd'hui ne permet pas de conclure sur la capacité de la Chine à le rattraper.

La répartition à l'échelle mondiale de la puissance de calcul, conditionnant la faisabilité de tout projet massif d'IA, détermine quels acteurs peuvent mener ou non de tels projets. Cette répartition est très inégale : **les États-Unis détiennent 86 % de la puissance de calcul mondiale**, la Chine environ 8 % et le reste du monde se partage les 6 % restants, **la France disposant de moins de 1 % du total.**

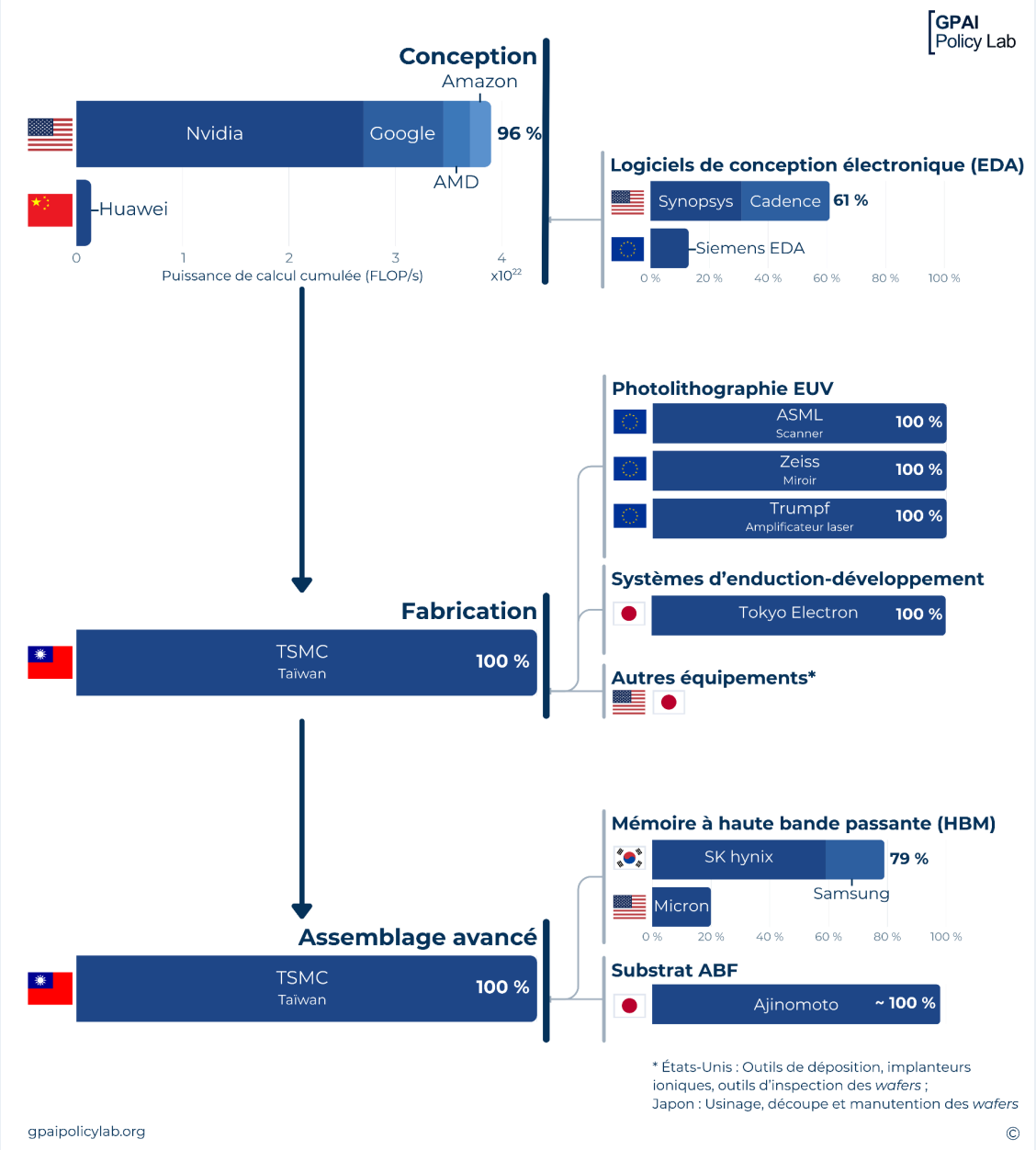

Figure A - Structure de la chaîne de production d'un GPU de pointe.

— IMPLICATIONS POUR LES ACTEURS FRANÇAIS ET EUROPÉENS —

Tout projet massif d'IA en Europe se fait *de facto* avec l'accord des États-Unis.

L'Europe a deux voies d'accès à la puissance de calcul (achat de GPU et location *cloud*) et les deux sont fortement concentrées aux États-Unis : 96 % de la production et 86 % de la capacité déployée.

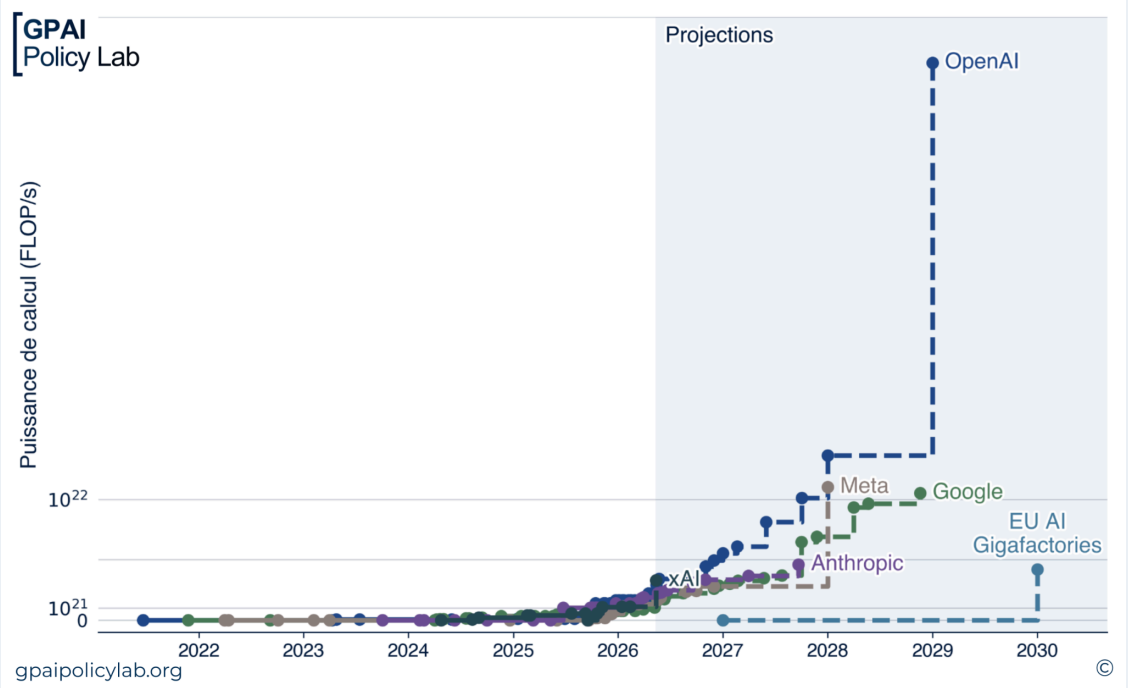

Figure B - Puissance de calcul des centres de calcul de frontière.

Les États-Unis disposent de mécanismes de contrôle sur ces deux voies d'accès. Ils peuvent restreindre :

1. l'accès *cloud* (effet immédiat) ;
2. les exportations de GPU (effet en ~ 3–5 ans) ;
3. les GPU déployés à distance (en discussion, horizon temporel incertain).

L'Europe, comme certaines puissances moyennes, dispose de leviers de gouvernance concrets sur le développement mondial de l'IA.

La puissance de calcul étant le facteur principal du développement des IA à usage général, les goulets d'étranglement de sa chaîne d'approvisionnement confèrent aux puissances qui les contrôlent un levier sur le développement de l'IA.

Bloquer un goulet permettrait de repousser les entraînements posant des problèmes pour la sécurité nationale, mais cet effet est temporaire car seule la filière principale serait ralentie, permettant à la filière chinoise de rattraper son retard et devenir la nouvelle frontière. Or cette perspective contrevient directement à la stratégie américaine de domination en IA, faisant de ces goulets des atouts diplomatiques face aux États-Unis.

Plusieurs goulets sont contrôlés par le Japon, la Corée du Sud et Taïwan, dont la dépendance à la protection sécuritaire américaine en limite l'usage, à moins de disposer de garanties de sécurité alternatives, par exemple dans le cadre d'une coalition de puissances moyennes. **Le triangle européen ASML–Zeiss–Trumpf, qui détient un monopole sur les équipements de photolithographie EUV, technologie nécessaire à tout GPU de pointe, ne dépend pas des États-Unis pour sa sécurité immédiate, ce qui en fait un levier actionnable et efficace, au moins à l'horizon de quelques années.**

General
Purpose AI
Policy Lab

Puissance de calcul

Chaîne d'approvisionnement et dépendances stratégiques pour le développement de l'IA

Cette note vise à répondre à deux questions :

- Quel est le rôle de la puissance de calcul dans le développement de l'IA, et comment s'organise sa chaîne d'approvisionnement ?
- Quelles dépendances stratégiques cette chaîne d'approvisionnement crée-t-elle, et quels leviers confère-t-elle aux différentes puissances ?

Les États-Unis et la Chine sont engagés dans une course au développement d'IA à usage général (*General-Purpose AI*, ou GPAI) de plus en plus performantes et agentiques, qui atteignent ou dépassent les compétences d'experts humains dans un nombre croissant de domaines[1]. Les conséquences de cette dynamique, à court comme à moyen terme, sont d'ampleur mondiale (accélération et amplification des cyberattaques[2], ou même des perspectives de perte de contrôle irréversible d'IA agentiques[3]), avec des implications majeures pour la sécurité nationale. Plus largement, les progrès de l'IA étant voués à restructurer les équilibres économiques et militaires, ils font de l'IA un sujet géopolitique incontournable. Or ce développement technologique repose sur une ressource physique centrale, dont le contrôle façonne les rapports de force : la puissance de calcul.

I. Quel est le rôle de la puissance de calcul dans le développement de l'IA, et comment s'organise sa chaîne d'approvisionnement ?

1. La puissance de calcul est le moteur du développement de l'IA

Les systèmes d'IA à usage général atteignent ou dépassent les performances d'experts humains dans plusieurs domaines, portés par une augmentation massive de l'échelle de leur entraînement (cf. Figure 1). Contrairement à un logiciel classique, programmé par des règles explicites, un système d'IA est entraîné. Ses performances dépendent de la quantité de calcul investie dans cet entraînement, qui croît avec la taille du modèle et le volume de données traitées.

Concrètement, les développeurs définissent un objectif d'entraînement, qui mesure la performance du modèle d'IA. À chaque itération, le modèle reçoit des données et produit une sortie qui est évaluée quantitativement. Ce signal est ensuite utilisé pour ajuster les paramètres internes du modèle, de manière à améliorer ses performances dans des conditions similaires à l'avenir. Au début de l'entraînement, les performances sont donc très mauvaises. Mais à mesure que le système est confronté à une grande variété de données, ses performances augmentent progressivement, non

[1] "Artificial General Intelligence (AGI) : Faisabilité technique et horizons temporels" GPAI Policy Lab (2025) ; "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI" GPAI Policy Lab (2026). https://gpaipolicylab.org/note-4.

[2] "Mythos et autres modèles-frontière : implications des progrès de l'IA pour la cyber en France et en Europe" *Campus Cyber* (2026). https://campuscyber.fr/mythos/.

[3] "Artificial General Intelligence (AGI) : Anticipation des objectifs et implications pour la sécurité et le contrôle" *GPAI Policy Lab* (2025). https://gpaipolicylab.org/note-2.

seulement sur les données d'entraînement, mais il généralise aussi, de manière étonnamment efficace, à des situations jamais rencontrées auparavant et requérant plus d'autonomie.

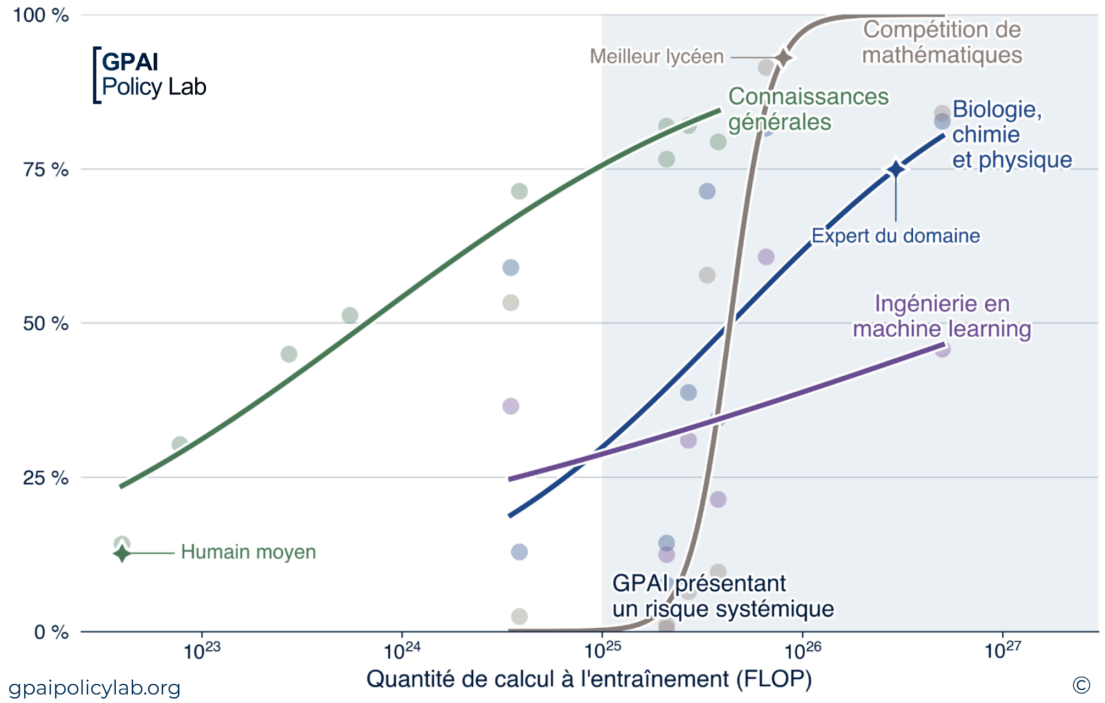

Figure 1 - Les performances des IA à usage général (GPAI) augmentent avec la quantité de calcul investie dans leur entraînement. Relation empirique entre la performance de systèmes d'IA de frontière, mesurée par le pourcentage de bonnes réponses sur différents benchmarks, en fonction de la quantité de calcul investie dans leur entraînement, mesurée en FLOP. Des repères de compétences humaines sont ajoutés. Les benchmarks sont : « MMLU »[4] pour les connaissances générales, « GPQA Diamond »[5] pour les connaissances en biologie, chimie et physique, « OTIS Mock AIME 2024-2025 »[6] pour les compétences en mathématiques et « WeirdML (v2) »[7] pour les compétences d'ingénierie en machine learning. Source : Epoch AI[8].

Chaque itération d'entraînement consiste en un assemblage de calculs informatiques élémentaires (additions, multiplications, etc.), appelés opérations. **Ces opérations sont réalisées par des GPU[9], le nom donné aux processeurs spécialisés dans les calculs nécessaires à l'entraînement de modèles d'IA**. La puissance de calcul d'un GPU se mesure par le nombre d'opérations qu'il peut effectuer par seconde (FLOP/s), et la taille d'un entraînement se mesure par la quantité totale d'opérations effectuées (FLOP)[10]. **Une plus grande puissance de calcul permet de réaliser de plus gros entraînements, donnant *a priori* des IA plus performantes.**

[4] Hendrycks et al. "Measuring Massive Multitask Language Understanding" *arXiv* (2020). 10.48550/arXiv.2009.03300.
[5] Rein et al. "GPQA: A Graduate-Level Google-Proof Q&A Benchmark" *arXiv* (2023). 10.48550/arXiv.2311.12022.
[6] Chen. "The OTIS Mock AIME" *Evan Chen* (2026). https://web.evanchen.cc/mockaime.html.
[7] Ihle. "WeirdML" *Håvard Tveit Ihle* (2025). https://htihle.github.io/weirdml.html.
[8] "AI Benchmarking Hub" *Epoch AI* (2026). https://epoch.ai/benchmarks.
[9] Le terme technique est « accélérateur IA » (*AI accelerator*). L'usage de GPU (*Graphics Processing Unit*) s'est néanmoins imposé car c'est la dénomination utilisée par Nvidia🇺🇸, leader mondial en conception d'accélérateur IA, héritée de leur fonction initiale de traitement graphique.
[10] Attention à ne pas confondre la quantité de calcul, mesurée en FLOP (nombre total d'opérations arithmétiques effectuées), et la puissance de calcul, mesurée en FLOP/s (nombre d'opérations effectuées par seconde), qui correspond donc à la vitesse à laquelle la quantité de calcul est effectuée. La distinction est analogue à celle entre la distance parcourue, exprimée en kilomètres, et la vitesse de déplacement, exprimée en kilomètres par heure, qui correspond donc à la vitesse à laquelle la distance est parcourue.

Encadré 1 (détails techniques) - Les performances des modèles d'IA dépendent de la quantité de calcul investie dans leur entraînement.

L'entraînement d'un modèle d'IA est un processus itératif reposant sur une « fonction objectif », une fonction mathématique définie par les développeurs. À chaque itération d'entraînement, le modèle est confronté à des données, et produit une sortie. La fonction objectif quantifie la performance du modèle sur ces données et ce signal est utilisé comme heuristique pour modifier les paramètres[11] internes du modèle, typiquement par descente de gradient ou une variante, afin d'augmenter statistiquement les performances du modèle dans une situation similaire.

Chaque itération augmente incrémentalement les performances du modèle. À mesure qu'il est confronté à une grande quantité et une grande variété de données, le modèle a de meilleures performances sur les données d'entraînement, mais également sur des données jamais rencontrées. Cette capacité de généralisation est caractéristique de l'entraînement des modèles d'IA.

La taille d'un modèle d'IA, mesurée par son nombre de paramètres, détermine le plafond de performance qu'il peut atteindre. L'entraînement, en ajustant progressivement ces paramètres à travers un grand nombre d'itérations, rapproche le système de ce plafond. Toutes choses égales par ailleurs, un modèle plus grand a donc un plafond de performance plus élevé.

Comme évoqué avant l'encadré, chaque itération d'entraînement consiste en un assemblage de calculs informatiques élémentaires (additions, multiplications, etc.), appelés opérations. Plus un modèle a de paramètres, plus chaque itération nécessite d'opérations pour les mettre à jour. La taille d'un modèle détermine donc non seulement son plafond de performance, mais aussi le coût en calcul de chaque itération d'entraînement.

La quantité totale de calcul nécessaire à un entraînement dépend donc de deux facteurs : la taille du modèle, qui détermine le coût de chaque itération, et le nombre d'itérations réalisées. **Cette quantité de calcul se mesure en nombre total d'opérations élémentaires, ou FLOP (pour *floating-point operations*)[12].** Au-delà des considérations théoriques exposées ici, on observe empiriquement une relation directe entre le nombre de FLOP investis dans l'entraînement d'un modèle et ses performances (cf. Figure 1).

L'entraînement d'un modèle se réduisant fondamentalement à l'exécution d'un grand nombre d'opérations, c'est notamment l'augmentation exponentielle de la puissance de calcul des GPU au fil du temps (cf. Figure 2a) qui rend possible l'augmentation de l'échelle des entraînements des modèles de frontière (cf. Figure 2b).

[11] aussi appelés « poids »

[12] Plusieurs définitions d'un FLOP coexistent selon le domaine considéré. Un nombre en virgule flottante (*floating-point*, FP) peut être encodé sur un nombre variable de *bits*, ce qui détermine sa précision. Un encodage sur moins de *bits* représente moins de chiffres significatifs (chiffres après la virgule) mais permet de stocker davantage de nombres dans une quantité donnée de mémoire. Ainsi, une mémoire de 64 *bits* peut contenir par exemple un seul nombre encodé sur 64 *bits*, deux nombres encodés sur 32 *bits*, ou encore huit nombres encodés sur 8 *bits*. En calcul haute performance (*high-performance computing*, HPC), les simulations physiques exigent une grande précision et utilisent typiquement un encodage sur 64 *bits* (FP64). En IA, la précision des paramètres importe moins, et il est de plus en plus courant de les encoder sur 8 *bits* (FP8).
Somala et al. "Widespread adoption of new numeric formats took 3-4 years in past cycles" *Epoch AI* (2025). https://epoch.ai/data-insights/training-precision.

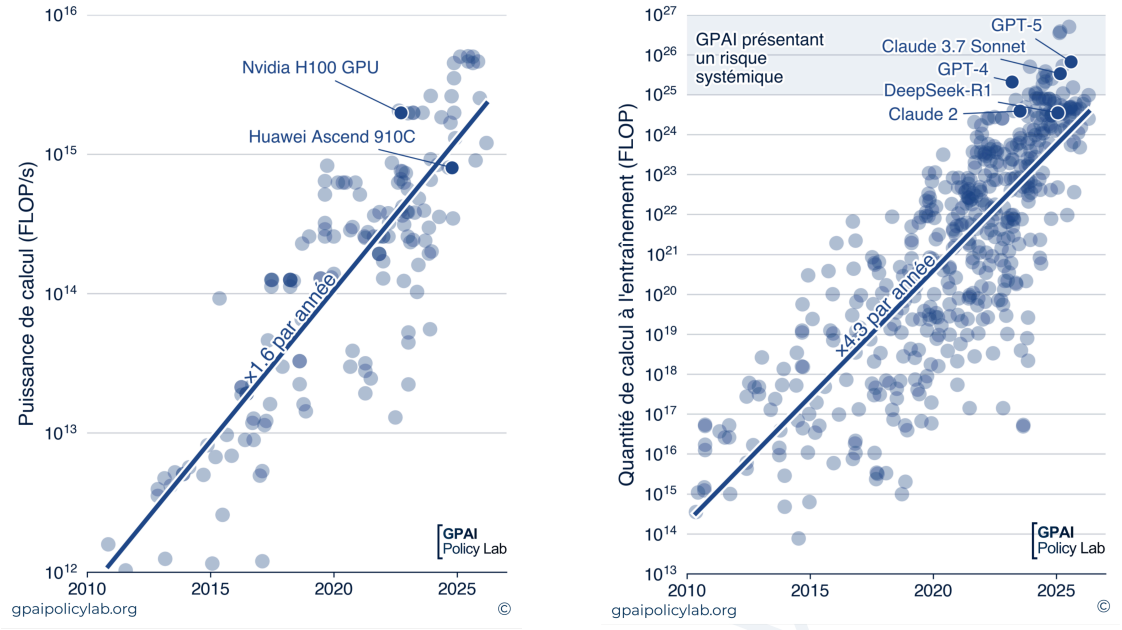

Figure 2 - Augmentation exponentielle de la puissance de calcul des GPU (a) et de l'échelle des entraînements des modèles de frontière (b). La puissance de calcul des GPU double tous les 18 mois et la quantité de calcul investie dans l'entraînement des modèles de frontière double tous les 6 mois. Source : Epoch AI[13].

2. La chaîne d'approvisionnement de la puissance de calcul est fortement concentrée

L'augmentation de l'échelle des entraînements illustrée sur la Figure 2b repose sur une production massive de GPU et l'infrastructure permettant leur exploitation. Ces GPU sont hébergés dans des centres de calcul, de gigantesques installations fournissant les ressources nécessaires à leur fonctionnement (alimentation électrique, refroidissement, connectivité, entre autres). **La chaîne d'approvisionnement de la puissance de calcul se structure ainsi en deux composantes, les GPU et les centres de calcul qui les hébergent, chacune marquée par des points de forte concentration.** Les résultats de cette section reposent sur des recherches documentaires variées car il n'existe à ce jour aucun mécanisme de suivi systématique de la production et de la distribution des GPU, ni de la répartition mondiale de la puissance de calcul.

A. Chaîne de production des GPU

La production de GPU de pointe repose sur une chaîne de production complexe, mondialement distribuée, et marquée par des goulets d'étranglement, c'est-à-dire des points de forte concentration et difficilement substituables (cf. Figure 4). **Les États-Unis ont un monopole sur la conception des GPU de pointe**, les acteurs principaux étant Nvidia🇺🇸, Google🇺🇸, AMD🇺🇸 et Amazon🇺🇸, cumulant plus de 95 % de la puissance de calcul IA produite dans le monde. **Leur fabrication est déléguée à TSMC🇹🇼 à Taïwan, qui dépend des machines de photolithographie EUV dont ASML🇪🇺 aux Pays-Bas détient le monopole.** Mais en parallèle, une **filière chinoise**

[13] "Data on Machine Learning Hardware" *Epoch AI* (2026). https://epoch.ai/data/machine-learning-hardware ; "Data on AI Models" *Epoch AI* (2026). https://epoch.ai/data/ai-models.

entièrement domestique, actuellement en retard sur l'ensemble des maillons par rapport à la filière principale, est en cours de développement (cf. Encadré 3).

Un GPU est un assemblage de plusieurs composants :

- **Puce de calcul** (*logic die*) : puce électronique où sont effectuées les opérations de calcul IA ;
- **Mémoire à haute bande passante** (*High Bandwidth Memory*, HBM) : mémoire à disposition de la puce de calcul, stockant les données d'entraînement et les paramètres du modèle d'IA. Un GPU contient typiquement plusieurs modules HBM ;
- **Interconnexions** : composants assurant la communication au sein du GPU (entre la puce de calcul et les modules de mémoire) et entre le GPU et le reste du serveur.

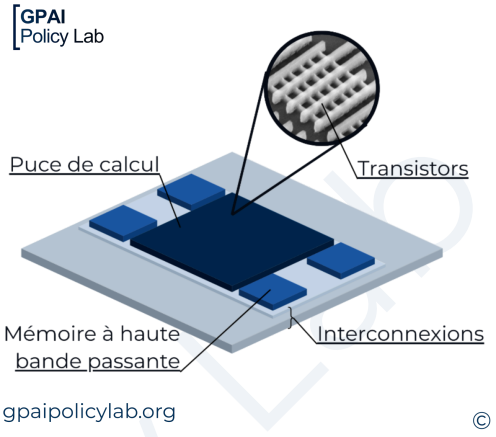

Figure 3 - Les composants d'un GPU

La chaîne de production peut être décomposée en trois grandes étapes (cf. Encadré 2), centrées sur la puce de calcul :

1. **Conception** (*design*) : conception de la puce de calcul et de son intégration avec les autres composants ;
2. **Fabrication** : gravure physique des circuits sur la puce de calcul ;
3. **Assemblage avancé** (*advanced packaging*) : intégration de la puce de calcul avec les autres composants (modules de mémoire, interconnexions) pour former le produit final.

Chacune des étapes de production dépend elle-même d'équipements, de logiciels et de matériaux en amont dont la production a des points de forte concentration, répartis principalement entre les États-Unis, Taïwan, le Japon et l'Europe (cf. Figure 4). Les autres processus de fabrication et leurs dépendances sont partagés par plusieurs autres acteurs, dont la Corée du Sud, la Chine, Singapour ou encore Israël, **la France n'ayant pas de place dominante dans cette filière**. Le Tableau 1 présente des points de concentration identifiés, ainsi que les acteurs et pays concernés, et l'Encadré 2 (détails techniques) détaille la chaîne de production des GPU. Faute de registre public cartographiant exhaustivement la chaîne de production, des points de concentration non identifiés à ce jour subsistent vraisemblablement.

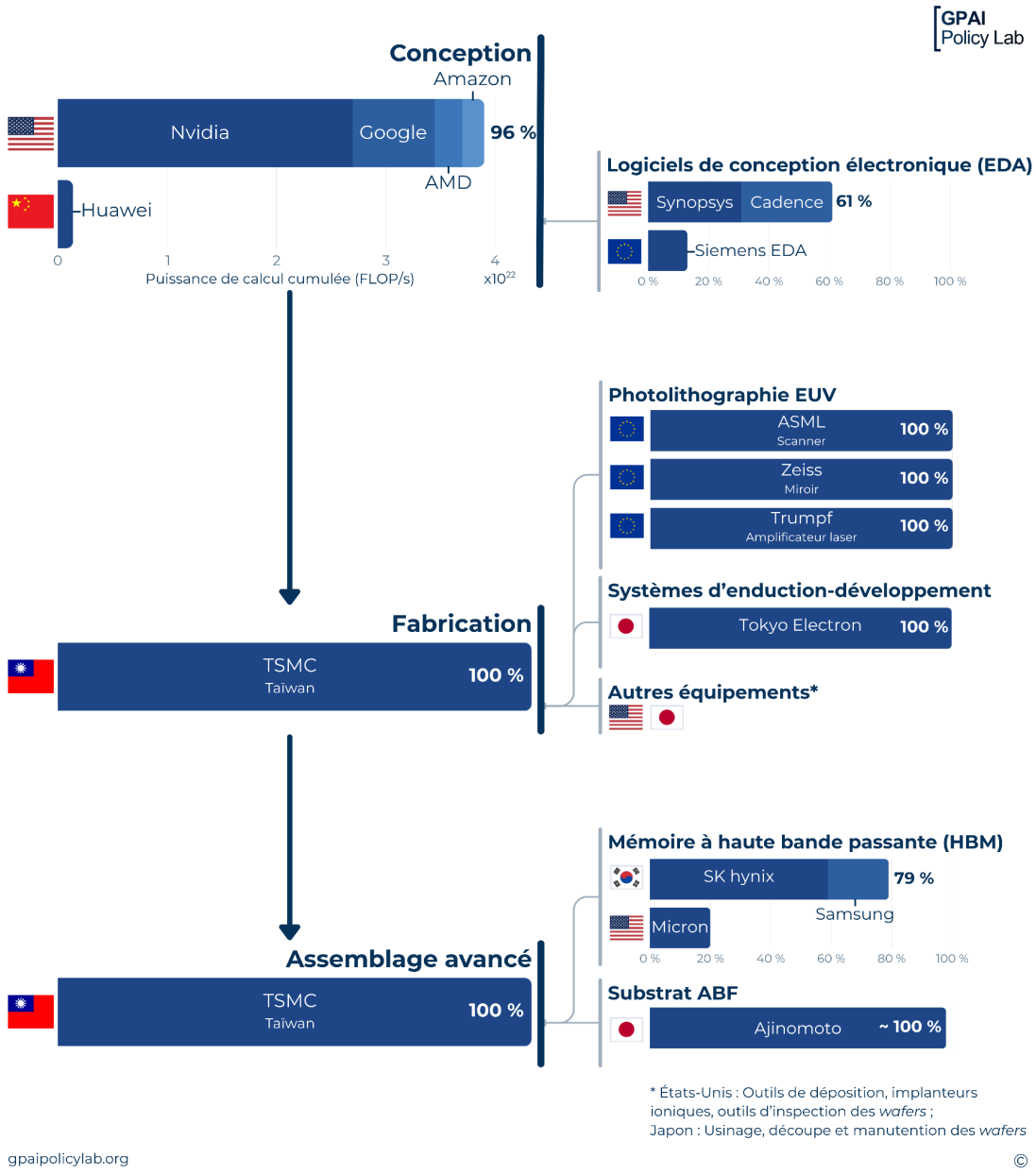

Figure 4 - Structure de la chaîne de production d'un GPU de pointe. Les données représentent les parts de marché* d'après le siège de la maison-mère, et non d'après la localisation physique de la production. *Excepté pour la conception, qui montre la puissance de calcul cumulée produite depuis 2022. Sources détaillées dans le Tableau 1.

Tableau 1 - Points de concentration dans la chaîne de production des GPU de pointe. Les données représentent les parts de marché* d'après le siège de la maison-mère, et non d'après la localisation physique de la production. *Excepté pour la conception, qui montre la puissance de calcul cumulée produite depuis 2022. Source principale : Emerging Technology Observatory[14].

	🇺🇸 USA	🇪🇺 EUROPE	🇯🇵 JAPON	🇹🇼 TAÏWAN	🇰🇷 CORÉE DU SUD
Conception[15]	**96 %**	—	—	—	—
Logiciels de conception électronique (EDA)[16]	**61 %**	13 %	—	—	—
Fonderie (puces de calcul)[17]	—	—	—	**~ 100 %**	—
Outils de déposition	**61 %**	12 %	15 %	—	3 %
Photolithographie EUV[18] • Scanners • Miroirs EUV • Amplificateurs laser	—	 **100 %** **100 %** **100 %**	—	—	—
Machines de photolithographie en immersion (DUV)	—	**99 %**	1 %	—	—
Systèmes d'enduction-développement (immersion et EUV)	—	—	**100 %**	—	—
Équipements d'usinage des *wafers*	—	—	**~ 100 %**	—	—
Équipements de découpe des puces (*dicing*)	1 %	—	**95 %**	—	1 %
Équipements d'inspection des *wafers*	**77 %**	7 %	9 %	—	—
Implanteurs ioniques	**90 %**	—	9 %	—	—
Systèmes de manutention de *wafers* et de photomasques	—	—	**86 %**	—	13 %
Assemblage avancé[19]	—	—	—	**~ 100 %**	—
Mémoire à haute bande passante (HBM)[20]	20 %	—	—	—	**79 %**
Substrat ABF[21]	—	—	**98 %**	—	—

[14] "Supply Chain Explorer: Semiconductors" *Emerging Technology Observatory* (2025). https://chipexplorer.eto.tech/.
[15] "Data on AI Chip Sales" *Epoch AI* (2026). https://epoch.ai/data/ai-chip-sales.
[16] "China Revenue at Risk as U.S. Curbs Slam EDA Giants: Impact on Synopsys, Cadence and More" *TrendForce* (2025). https://www.trendforce.com/news/2025/06/02/news-china-revenue-at-risk-as-u-s-curbs-slam-eda-giants-impact-on-synopsys-cadence-and-more/.
[17] "Taiwan – Semiconductors including chip design for AI" *International Trade Administration* (2025). https://www.trade.gov/country-commercial-guides/taiwan-semiconductors-including-chip-design-ai.
[18] Kleinhans et al. "Is the EU's Semiconductor Manufacturing Equipment a Strategic Chokepoint?" *Digital Power China* (2024). https://www.jpkleinhans.de/home/DPC2024_Lithography_Chapter.pdf.
[19] Patel et al. "AI Capacity Constraints - CoWoS and HBM Supply Chain" *SemiAnalysis* (2023). https://newsletter.semianalysis.com/p/ai-capacity-constraints-cowos-and.
[20] "Samsung Reportedly to Make Korean HBM4 Debut Tomorrow at SEDEX, SK hynix Also on Display" *TrendForce* (2025). https://www.trendforce.com/news/2025/10/21/news-samsung-reportedly-to-make-korean-hbm4-debut-tomorrow-at-sedex-sk-hynix-also-on-display/.
[21] Eskenazi et al. "Securing the Substrate Behind Every Chip: A U.S. Strategy for Ajinomoto Build-Up Film (ABF)" *Convergence Analysis* (2025). https://www.convergenceanalysis.org/fellowships/economics/securing-the-substrate-behind-every-chip-a-us-strategy-for-ajinomoto-build-up-film-abf.

Encadré 2 (détails techniques) - Cartographie de la chaîne de production des GPU de pointe[22]

A. Conception

Les entreprises qui développent les GPU se concentrent sur la conception de la puce de calcul et de son intégration dans l'assemblage final. Elles sont qualifiées de *fabless* (sans usine) car elles délèguent la fabrication à des entreprises spécialisées, appelées « fonderies ». **Plus de 95 % des GPU de pointe sont conçus par des entreprises américaines, Nvidia🇺🇸 étant le leader avec plus de 60 % de la puissance de calcul produite mondialement, suivi de Google🇺🇸, AMD🇺🇸 et Amazon🇺🇸, les quelques pour cent restants étant développés par Huawei🇨🇳 en Chine**. La conception repose sur l'utilisation de logiciels de conception électronique (*Electronic Design Automation*, EDA). Seules trois solutions complémentaires sont suffisamment avancées pour intégrer toutes les fonctionnalités nécessaires à la conception de puces de calcul avancées et de leur assemblage, développées par **les entreprises américaines Synopsys🇺🇸 et Cadence🇺🇸 d'une part, et l'entreprise européenne Siemens EDA🇪🇺**[23] **de l'autre.**

B. Fabrication

La puce de calcul est fabriquée par des fonderies spécialisées. La quasi-totalité des puces de calcul des GPU de pointe sont fabriquées par TSMC🇹🇼 à Taïwan. La puce est composée de milliards de transistors, des structures nanométriques qui effectuent les opérations élémentaires de calcul. La puissance de calcul d'un GPU se mesure par le nombre total d'opérations qu'il peut effectuer par seconde (FLOP/s), ce qui dépend à la fois du nombre de transistors et de leur cadence de fonctionnement. En pratique, les progrès en matière de puissance de calcul des GPU proviennent principalement de l'augmentation du nombre de transistors, rendue possible par leur miniaturisation, qui permet d'en intégrer davantage sur une même surface[24]. Cette miniaturisation est caractérisée par la taille de nœud[25], **les GPU de pointe nécessitant des tailles de nœud inférieures à environ 10 nm.**

La fabrication consiste à produire ces structures sur un *wafer*, une fine plaquette circulaire de silicium, par une succession de dépôts de matériaux, d'impressions de motifs par photolithographie et de gravures, répétés couche par couche. La photolithographie consiste à projeter de la lumière à travers un masque pour imprimer les motifs nanométriques des circuits sur le *wafer*. La voie industrielle dominante pour la fabrication des semi-conducteurs « avancés » nécessaires aux GPU de pointe est la photolithographie à l'ultraviolet extrême (EUV), dont la longueur d'onde est environ dix fois plus courte que celle de la photolithographie à l'ultraviolet profond (DUV), utilisée pour les semi-conducteurs à nœuds plus grands, dits « matures ». La longueur d'onde du DUV est en effet trop longue pour imprimer des structures inférieures à 20 nm en une seule exposition, ce qui impose des procédés de multi-exposition qui réduisent

[22] La description générale de la chaîne de production des semi-conducteurs s'appuie principalement sur trois sources : "Mapping the semiconductor value chain: Working towards identifying dependencies and vulnerabilities" *OECD Science, Technology and Industry Policy Papers* (2025). 10.1787/4154cdbf-en ; "Strengthening the Global Semiconductor Supply Chain in an Uncertain Era" *Semiconductor Industry Association and Boston Consulting Group* (2021). https://www.semiconductors.org/strengthening-the-global-semiconductor-supply-chain-in-an-uncertain-era/ ; Khan et al. "The Semiconductor Supply Chain: Assessing National Competitiveness" *Center for Security and Emerging Technology* (2021). https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/.
Les données de concentration de marché et de production proviennent du Tableau 1. Les informations plus spécifiques sont accompagnées de leur propre source.

[23] Anciennement Mentor Graphics🇺🇸, renommée après son rachat par Siemens🇪🇺 en 2017.

[24] Hobbhahn et al. "Predicting GPU performance" *Epoch AI* (2022). https://epoch.ai/blog/predicting-gpu-performance.

[25] La taille de nœud correspondait historiquement à la distance entre les transistors, mais désigne aujourd'hui une génération de technologie de fabrication sans être directement reliée à une dimension physique précise.

significativement les rendements et les cadences de production[26]. Une seule entreprise au monde produit les machines de photolithographie EUV nécessaires à la fabrication des puces de calcul des GPU de pointe : **ASML**, aux Pays-Bas. Ces machines dépendent de composants produits par deux entreprises monopolistiques allemandes : les miroirs haute précision de **Zeiss**[27] et les systèmes laser de haute puissance de **Trumpf**[28]. **La technologie de photolithographie EUV, nécessaire à la fabrication de toute puce de calcul de pointe, repose ainsi sur ces trois entreprises européennes,** qui l'ont co-développée au fil de plusieurs décennies[29].

La fabrication requiert aussi d'autres équipements spécialisés dont la production est fortement concentrée. Par exemple, les systèmes d'enduction-développement (*coater/developer systems*), qui appliquent et développent les couches de produits chimiques photosensibles sur le *wafer* avant et après la photolithographie. Seule **Tokyo Electron au Japon** produit ceux capables de répondre aux exigences en matière de précision et de cadence pour la production de masse des puces de pointe par TSMC. Plus généralement, la fabrication est un processus en salle blanche qui nécessite la coordination de centaines de machines différentes, et la maîtrise d'une fonderie comme TSMC réside dans sa capacité à optimiser l'ensemble du processus pour atteindre un haut niveau de fiabilité en production de masse[30]. Pour les puces de calcul des GPU de pointe, TSMC détient aujourd'hui un quasi-monopole sur cette capacité.

C. Assemblage avancé[31]

L'assemblage avancé (*advanced packaging*) intègre la puce de calcul avec d'autres composants, notamment plusieurs modules de mémoire à haute bande passante (*High Bandwidth Memory*, HBM) et des systèmes d'interconnexion (*interposer*, substrat ABF). Les modules HBM sont produits par des fabricants intégrés (*Integrated Device Manufacturer*), qui assurent à la fois conception et fabrication, contrairement aux *fabless*. Le marché des HBM de pointe est partagé entre la **Corée du Sud (SK hynix et Samsung) et les États-Unis (Micron).** La fabrication de HBM suit des étapes similaires à celle de la puce de calcul, mais les exigences de miniaturisation sont différentes[32]. Elle repose principalement sur la photolithographie DUV[33], dont le marché est partagé entre ASML et Nikon. TSMC est également le leader de l'assemblage avancé, avec sa technologie *Chip on Wafer on Substrate* (CoWoS[34]). Ce processus présente lui aussi une dépendance critique concentrée : les substrats ABF, un matériau permettant une densité de connexion très élevée entre les composants, dont la production est dominée par une seule entreprise japonaise, Ajinomoto[35].

[26] Dorsch. "Changes and Challenges Abound in Multi-patterning Lithography" *SEMI* (2015). https://www.semi.org/en/changes-and-challenges-abound-multi-patterning-lithography.

[27] "High-NA-EUV lithography: mirroring the future" *ZEISS SMT Magazine* (2024). https://www.zeiss.com/semiconductor-manufacturing-technology/smt-magazine/high-na-euv-lithography.html.

[28] "EUV Drive Laser" *TRUMPF*. https://www.trumpf.com/fr_INT/solutions/applications/lithographie-euv/euv-drive-laser/.

[29] "EUV Lithography: A European Joint Project" *ZEISS SMT Magazine* (2021). https://www.zeiss.com/semiconductor-manufacturing-technology/smt-magazine/euv-lithography-as-an-european-joint-project.html.

[30] "How a Semiconductor Factory Works" *Intel Newsroom* (2023). https://newsroom.intel.com/tech101/how-a-semiconductor-factory-works.

[31] Op. cit. "AI Capacity Constraints - CoWoS and HBM Supply Chain" ; Patel. "Advanced Packaging Part 1 – Pad Limited Designs, Breakdown Of Economic Semiconductor Scaling, Heterogeneous Compute, and Chiplets" *SemiAnalysis* (2021). https://newsletter.semianalysis.com/p/advanced-packaging-part-1-pad-limited.

[32] Low. "China is making strides in etching machines for memory" *The Substrate* (2026). https://www.the-substrate.net/p/china-is-making-strides-in-etching.

[33] Certaines couches des générations actuelles de HBM sont déjà fabriquées par photolithographie EUV, et les prochaines générations devraient en intégrer davantage.
Haley. "EUV's Future Looks Even Brighter" *Semiconductor Engineering* (2025). https://semiengineering.com/euvs-future-looks-even-brighter/.

[34] "CoWoS®" *TSMC* (2024). https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm.

[35] Op. cit. "Securing the Substrate Behind Every Chip: A U.S. Strategy for Ajinomoto Build-Up Film (ABF)".

Ces points de concentration constituent des goulets d'étranglement car premièrement la structure séquentielle de la chaîne fait qu'un blocage en un point affecte rapidement l'aval, et deuxièmement ils sont difficilement substituables.

La rapidité d'impact sur l'aval est illustrée par la dépendance des processus industriels vis-à-vis des équipements, logiciels et matériaux en amont, dont les délais d'impact varient. Par exemple, les machines de photolithographie EUV se dégradent en quelques mois en l'absence de maintenance régulière assurée par ASML[36], et les stocks de certains consommables de ces machines peuvent être de l'ordre de quelques mois, voire de quelques semaines[37].

Plusieurs facteurs contribuent à la difficulté de substitution de ces goulets.

- D'une part, la production de semi-conducteurs avancés repose sur des procédés d'ingénierie physique et chimique complexes. La technologie EUV repose par exemple sur des miroirs dont la précision nanométrique est telle que, si on les ramenait à l'échelle de la France métropolitaine, leur irrégularité maximale ne dépasserait pas l'épaisseur d'un cheveu[38].
- D'autre part, au-delà de la maîtrise technologique elle-même, le savoir-faire accumulé pour atteindre les rendements et les cadences de la production de masse constitue une barrière à part entière. Un acteur disposant de ressources importantes pourrait reproduire une technologie donnée, mais la reproduire au même débit et au même rendement serait un défi autrement plus difficile.
- Enfin, la chaîne de production des semi-conducteurs est caractérisée par un fort couplage entre ses maillons, les acteurs se sont co-développés pour optimiser l'ensemble du processus : ASML s'est coordonné avec Zeiss et Trumpf pour développer la technologie de photolithographie EUV[39], les machines d'enduction-développement de Tokyo Electron sont optimisées pour fonctionner en tandem avec celles d'ASML[40], les processus physiques et chimiques en amont sont optimisés pour ceux en aval, etc. Tout remplacement d'un point de concentration nécessiterait de le reproduire quasi à l'identique, ou de remplacer plusieurs maillons en parallèle.

Les tentatives de contournement de la Chine illustrent la difficulté à substituer ces goulets d'étranglement. Malgré des investissements massifs et une volonté politique forte, la Chine n'est pas parvenue à ce jour à atteindre l'autonomie sur la production de semi-conducteurs avancés, et le retard technologique de l'écosystème chinois par rapport aux leaders mondiaux demeure de plusieurs années, variant selon le maillon de la chaîne considéré (cf. Encadré 3 et Tableau 2 inscrit). **Toutefois, un retard aujourd'hui ne permet pas de conclure sur la capacité de la Chine à rattraper la filière principale ni à quel horizon temporel.**

[36] Op. cit. "Is the EU's Semiconductor Manufacturing Equipment a Strategic Chokepoint?".

[37] Tobin et al. "An Invisible Bottleneck: A Helium Shortage Threatens the Chip Industry" *The New York Times* (2026). https://www.nytimes.com/2026/03/27/business/helium-chips-iran-war.html.

[38] Op. cit. "High-NA-EUV lithography: mirroring the future".

[39] "PM: Werner-von-Siemens-Ring 2024 geht an Peter Kürz & Michael Kösters" *Stiftung Werner-von-Siemens-Ring* (2024). https://siemens-ring.de/werner-von-siemens-ring-kuerz-koesters/ ; "EUV Lithography – New Light for the Digital Age" *Deutscher Zukunftspreis* (2020). https://www.deutscher-zukunftspreis.de/en/team-1-2020.

[40] "Tokyo Electron Limited and ASML announce agreement" *ASML* (2011). https://www.asml.com/en/news/press-releases/2011/tokyo-electron-limited-and-asml-announce-agreement ; "Tokyo Electron to Collaborate with imec-ASML Joint High NA EUV Research Laboratory" *Tokyo Electron Limited* (2021). https://www.tel.com/news/topics/2021/20210608_001.html.

Encadré 3 - Efforts de la Chine pour l'autosuffisance en semi-conducteurs[41]

La Chine a engagé depuis 2014 des efforts conséquents pour réduire sa dépendance aux fournisseurs étrangers, issus de la filière principale, sur l'ensemble de la chaîne de production des semi-conducteurs. Cet effort s'est structuré autour de trois phases successives du *National Integrated Circuit Industry Investment Fund* (dit « *Big Fund* »), lancées en 2014, 2019 et 2024 respectivement. La troisième, dotée d'un capital initial de 344 milliards de yuans (environ 47,5 milliards USD), semble cibler spécifiquement des maillons identifiés comme critiques pour la production de puces avancées comme les GPU. Ces investissements viennent compléter un dispositif plus large comprenant des subventions publiques dédiées, des exemptions fiscales ciblées et le soutien direct à des entreprises identifiées comme champions nationaux (Huawei🇨🇳 pour la conception de GPU, Empyrean Technology🇨🇳 pour les logiciels de conception électronique (EDA), la fonderie SMIC🇨🇳 pour la fabrication, SMEE🇨🇳 et SiCarrier🇨🇳 pour les équipements de fabrication comme les machines de photolithographie et les systèmes d'enduction-développement, CXMT🇨🇳 pour la mémoire à haute bande passante (HBM), etc.).

Malgré ces efforts conséquents, la Chine n'est pas autonome sur la production de semi-conducteurs avancés, et le retard technologique de l'écosystème chinois par rapport aux leaders mondiaux demeure de plusieurs années, variant selon le maillon de la chaîne considéré (cf. Tableau 2). Il est toutefois à noter qu'un retard aujourd'hui ne permet pas de conclure sur l'évolution temporelle de ce retard, en particulier de prédire si la Chine parviendra à le rattraper, ni à quel horizon temporel. **Le retard générationnel indiqué n'est donc pas une projection du temps nécessaire pour que la filière chinoise rattrape la filière principale** sur ce maillon, mais correspond uniquement au nombre d'années écoulées depuis le moment où le leader mondial se trouvait au niveau technologique actuel du champion chinois.

L'analyse du Tableau 2 montre la préférence des acteurs chinois pour les ressources de la filière principale, lorsqu'ils peuvent y accéder. **Toutefois, les restrictions américaines à l'exportation contribuent à ce que la Chine développe en arrière-plan une seconde filière, par nécessité pour son propre développement technologique.** Cette filière chinoise est pour l'instant en retard sur l'ensemble des maillons, mais le fait qu'elle vise à être entièrement domestique la rendrait indépendante des contrôles à l'exportation de la filière principale.

Encore inférieure à la filière principale sur chaque étape, cette filière domestique chinoise constitue cependant l'alternative la plus avancée, et sa trajectoire de développement aura un impact sur les rapports de force entre les différents acteurs en amont de l'approvisionnement de la puissance de calcul IA, et donc sur la trajectoire du développement des IA de frontière.

[41] Janjeva et al. "China's Quest for Semiconductor Self-Sufficiency" *Centre for Emerging Technology and Security* (2024). https://cetas.turing.ac.uk/publications/chinas-quest-semiconductor-self-sufficiency ; Lee et al. "Mapping China's semiconductor ecosystem in global context: Strategic dimensions and conclusions" *Mercator Institute for China Studies* (2021). https://merics.org/en/report/mapping-chinas-semiconductor-ecosystem-global-context-strategic-dimensions-and-conclusions.

Tableau 2 - Filière chinoise de production des GPU.

ÉTAPE	CHAMPION CHINOIS	RETARD GÉNÉRA-TIONNEL[42]	COMMENTAIRE
Conception	Huawei	~ 3–4 ans	Le GPU Huawei🇨🇳 Ascend 910C approche la puissance de calcul du Nvidia🇺🇸 H100 (cf. Figure 2a).
Logiciels de conception électronique (EDA)	Empyrean Technology	—	Empyrean Technology🇨🇳 ne permet pas encore de concevoir des puces de pointe[43]. L'architecture DaVinci[44] de l'Ascend 910C a vraisemblablement été conçue avec les logiciels de la filière principale, avant les restrictions[45], puis via des contournements[46].
Fonderie	SMIC	~ 7 ans	SMIC🇨🇳 est au niveau technologique de TSMC🇹🇼 en 2018-2019, mais à rendement et cadence moindres, dus aux restrictions à la technologie EUV d'ASML🇪🇺[47]. En pratique, la majorité des puces de calcul de l'Ascend 910C ont été fabriquées par TSMC🇹🇼, en contournant les contrôles à l'exportation[48].
Photo-lithographie	SMEE	~ 15–20 ans	SMEE🇨🇳 commercialise des machines de photolithographie de production de masse au niveau d'ASML🇪🇺 du début des années 2000[49]. En pratique, la fonderie SMIC🇨🇳 utilise vraisemblablement les machines d'ASML🇪🇺 plutôt que celles de SMEE🇨🇳 pour fabriquer les puces de calcul[50].

[42] Le retard générationnel correspond au temps écoulé depuis que le leader mondial se trouvait au niveau technologique actuel du champion chinois, et non une projection du temps nécessaire pour rattraper ce retard, car il dépend aussi des vitesses de développement respectives.

[43] Empyrean Technology🇨🇳 couvre la conception de circuits logiques jusqu'à une taille de nœud de 7 nm, mais pas l'ensemble du processus requis pour la conception de puces de pointe.

Ezell. "How Innovative Is China in Semiconductors?" *Information Technology and Innovation Foundation* (2024). https://itif.org/publications/2024/08/19/how-innovative-is-china-in-semiconductors/.

[44] Liao et al. "DaVinci: A Scalable Architecture for Neural Network Computing" *2019 IEEE Hot Chips 31 Symposium* (2019). 10.1109/HOTCHIPS.2019.8875654.

[45] "Addition of Entities to the Entity List" *Federal Register* (2019). https://www.federalregister.gov/documents/2019/05/21/2019-10616/addition-of-entities-to-the-entity-list.

[46] King et al. "Synopsys Probed on Allegations It Gave Tech to Huawei, SMIC" *Bloomberg* (2022). https://www.bloomberg.com/news/articles/2022-04-13/synopsys-probed-on-allegations-it-gave-chip-tech-to-huawei-smic ; "SEMI And EDA Industry Leaders Unite To Combat Software Piracy" *SEMI* (2020). https://www.semi.org/en/news-media-press/semi-press-releases/esda-machine-certification-partnership.

[47] La fonderie SMIC🇨🇳 est capable de produire en masse une taille de nœud de 7 nm et produit le 5 nm à une échelle moindre. N'ayant pas accès aux machines EUV d'ASML🇪🇺, elle utilise des machines DUV d'ASML🇪🇺, combinées à la technique du *multi-patterning*, qui permet de réduire la taille du nœud, au détriment du débit et du rendement de la fabrication.

"TechInsights Finds SMIC 7nm (N+2) in Huawei Mate 60 Pro" *TechInsights*. https://www.techinsights.com/blog/techinsights-finds-smic-7nm-n2-huawei-mate-60-pro ; "SMIC N+3 Confirmed: Kirin 9030 Analysis Reveals How Close SMIC Is to 5nm" *TechInsights* (2025). https://www.techinsights.com/blog/smic-n3-confirmed-kirin-9030-analysis-reveals-how-close-smic-5nm ; Nenni. "A Brief History of TSMC Through 2025" *SemiWiki* (2025). https://semiwiki.com/semiconductor-manufacturers/tsmc/365049-a-brief-history-of-tsmc-through-2025/ ; Nie et al. "The Export Control Loophole Fueling China's Chip Production" *Center for a New American Security* (2025). https://www.cnas.org/publications/cnas-insights/cnas-insights-the-export-control-loophole-fueling-chinas-chip-production ; Op. cit. "Changes and Challenges Abound in Multi-patterning Lithography".

[48] Patel et al. "Huawei AI CloudMatrix 384 – China's Answer to Nvidia GB200 NVL72" *SemiAnalysis* (2025). https://newsletter.semianalysis.com/p/huawei-ai-cloudmatrix-384-chinas-answer-to-nvidia-gb200-nvl72.

[49] "China's SMEE Reportedly Wins RMB 110M Lithography Tool Contract Amid Domestic Push" *TrendForce* (2025). https://www.trendforce.com/news/2025/12/26/news-chinas-smee-reportedly-wins-rmb-110m-lithography-tool-contract-amid-domestic-push/ ; Hunt et al. "China's Progress in Semiconductor Manufacturing Equipment: Accelerants and Policy Implications" *Center for Security and Emerging Technology* (2021). https://cset.georgetown.edu/publication/chinas-progress-in-semiconductor-manufacturing-equipment/.

[50] Les restrictions en vigueur empêchent SMIC🇨🇳 d'acquérir les machines EUV d'ASML🇪🇺, mais ne concernent pas ses machines DUV, dont la résolution moins fine ne permet pas de produire des puces aussi avancées ni au même débit que celles de la fonderie TSMC🇹🇼.

Patel. "The Gaps In The New China Lithography Restrictions – ASML, SMEE, Nikon, Canon, EUV, DUV, ArFi, ArF Dry, KrF, and Photoresist" *SemiAnalysis* (2023). https://newsletter.semianalysis.com/p/the-gaps-in-the-new-china-lithography.

B. Les centres de calcul

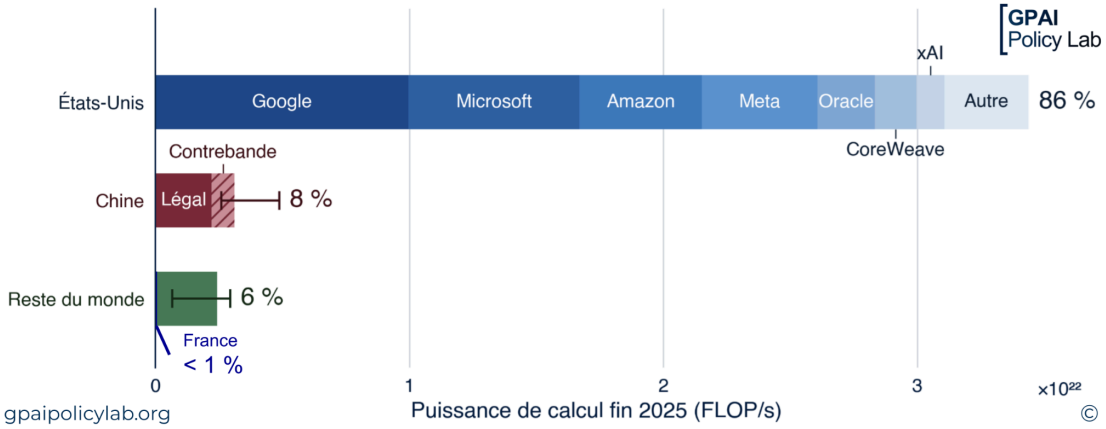

Figure 5 - Répartition mondiale de la puissance de calcul IA. Puissance de calcul cumulée à disposition[51] des principaux acteurs fin 2025. Le calcul de la puissance de calcul à disposition d'acteurs français est détaillé en annexe. Sources principales : Epoch AI[52] et ChinaTalk[53].

La puissance de calcul conditionnant la faisabilité de tout projet massif d'IA, sa répartition à l'échelle mondiale détermine quels acteurs peuvent mener de tels projets. Soit directement lorsque les infrastructures de calcul sont détenues par des acteurs développant l'IA, soit indirectement lorsqu'elles appartiennent à un fournisseur qui la propose comme service (*compute-as-a-service*), exerçant ainsi un levier sur les autres acteurs dépendant du fournisseur. La puissance de calcul IA est très inégalement répartie (cf. Figure 5), **les États-Unis concentrant 86 % du total mondial**, distribué principalement entre quelques acteurs majeurs du numérique (Google🇺🇸, Microsoft🇺🇸, Amazon🇺🇸, etc.). Un peu plus de la moitié de la part restante est détenue par des acteurs chinois, acquise légalement ou illégalement, les ~ 6 % restants étant répartis dans le reste du monde, dont moins de 1 % du total pour la France (cf. Annexe A).

Le développement d'IA de frontière mobilise des quantités de calcul telles qu'il est nécessaire de mobiliser des centaines de milliers, voire de millions de GPU de pointe en parallèle pour les effectuer dans des délais réalistes. Pour fournir une telle puissance de calcul cumulée, les GPU sont hébergés dans des centres de calcul, de gigantesques infrastructures pouvant compter jusqu'à des centaines de milliers de GPU, qui fournissent toutes les ressources nécessaires à leur exploitation (alimentation électrique, refroidissement, connectivité réseau, etc.). **Il n'existe à ce jour aucun registre public recensant exhaustivement les centres de calcul de frontière, leur localisation ou leur puissance de calcul.** Leur identification repose sur un ensemble de méthodes indirectes : analyse de permis de construire et de documents administratifs et juridiques, recoupement de sources ouvertes et imagerie satellite[54].

[51] L'attribution aux différents acteurs décrit la puissance de calcul qu'ils possèdent, sans critère juridique ou financier strict, mais ne décrit pas la distribution géographique. Par exemple, la puissance de calcul du centre de calcul « Microsoft Azure Pioneer-WEU » serait comptabilisée sous États-Unis/Microsoft, bien qu'elle soit localisée sur le territoire des Pays-Bas.

[52] "Data on AI Chip Owners" *Epoch AI* (2026). https://epoch.ai/data/ai-chip-owners.

[53] Zakaria. "How Much Compute Does China Have?" *ChinaTalk* (2026). https://www.chinatalk.media/p/how-many-chips-does-china-have.

[54] Les centres de calcul de cette échelle étant reconnaissables en particulier par la taille de leur système de refroidissement.
"Frontier Data Centers Documentation" *Epoch AI* (2026). https://epoch.ai/data/data-centers-documentation/methodology.

Encadré 4 - Le site de Stargate à Abilene (Texas) pour visualiser les ordres de grandeur d'un centre de calcul de frontière

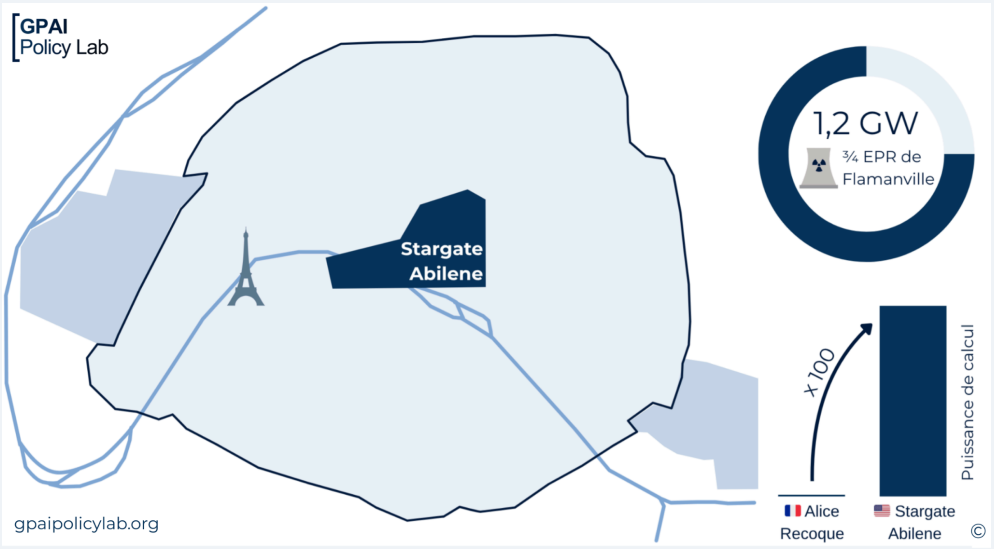

Figure 6 - Le site de Stargate à Abilene (Texas). Premier centre de calcul du projet Stargate. Visualisation de la taille du site comparée à Paris, de sa consommation électrique comparée au réacteur nucléaire EPR Flamanville-3 et de sa puissance de calcul comparée au futur supercalculateur exaflopique du Très Grand Centre de Calcul du CEA Alice Recoque.

Le site d'Abilene, au Texas, est le premier centre de calcul du projet Stargate[55], un programme américain d'investissement de 500 milliards de dollars sur quatre ans annoncé en janvier 2025 pour la construction d'infrastructures IA. **Ce premier site atteindra sa pleine capacité en juillet 2026, après 2 ans de construction pour un coût de 32 milliards de dollars**. Une fois complet, le site hébergera environ 300 000 GPU, consommera 1,2 GW de puissance électrique, et fournira une puissance de calcul totale de l'ordre de 10^{21} FLOP/s[56]. À titre de comparaison, sa consommation électrique représente environ les trois quarts d'un réacteur EPR comme celui de Flamanville de 1,6 GW[57], et **le supercalculateur Alice Recoque, futur supercalculateur exaflopique du très grand centre de calcul du CEA, n'atteindra au maximum qu'environ 1 % de la puissance de calcul du site d'Abilene[58].**

D'autres sites sont planifiés dans le cadre du projet Stargate, dont les localisations et capacités ne sont pas toutes publiques à ce jour.

[55] "Announcing The Stargate Project" *OpenAI* (2025). https://openai.com/index/announcing-the-stargate-project/.

[56] $2x10^{21}$ FP8 OP/s

"Frontier Data Centers" *Epoch AI* (2026). https://epoch.ai/data/data-centers.

[57] "L'EPR de Flamanville : unité de production n°3 de la centrale de Flamanville" *EDF*. https://www.edf.fr/centrale-nucleaire-flamanville3.

[58] Alice Recoque sera un supercalculateur exaflopique (> 10^{18} FP64 OP/s). Même en supposant que les GPU implémentent une technologie de type « *Rapid Packed Math* » d'AMD pour exécuter jusqu'à 8 opérations FP8 par cycle FP64, la puissance atteindrait ~ 10^{19} FP8 OP/s, soit deux ordres de grandeur de moins que le site d'Abilene.

"Exascale et Alice Recoque au TGCC" *CEA DAM*. https://www-hpc.cea.fr/fr/AliceRecoque.html.

Les principaux opérateurs de centres de calcul sont des *hyperscalers*, des entreprises fournissant un écosystème de services *cloud* (stockage de données, puissance de calcul, services réseau et applicatifs) au sein duquel la puissance de calcul spécialisée IA est un service parmi d'autres. Les « Big 3 », Google🇺🇸, Amazon🇺🇸 et Microsoft🇺🇸, disposent de plus de 50 % de la puissance de calcul IA mondiale (cf. Figure 5). Une seconde grande part de la puissance de calcul est détenue par les *neoclouds*, des entreprises plus récentes dont le leader est CoreWeave🇺🇸, spécialisées exclusivement dans la fourniture de puissance de calcul IA, sans offrir les autres services caractéristiques d'un *hyperscaler*[59]. Enfin, certaines entreprises achètent directement les GPU pour leur propre usage interne, sans fournir de service de *cloud*.

Les différentes entreprises développant des IA de frontière adoptent différentes combinaisons de ces stratégies : Google🇺🇸 développe ses propres GPU[60] opérés par son service de *cloud*, xAI🇺🇸 et Meta🇺🇸 exploitent leurs propres centres de calcul[61] contenant typiquement des GPU Nvidia🇺🇸[62]. OpenAI🇺🇸 et Anthropic🇺🇸 se fournissent via le service *cloud* des *hyperscalers*[63]. Tous peuvent compléter, par exemple avec des *neoclouds*, en fonction de leurs besoins en puissance de calcul[64].

La puissance de calcul IA exploitée sur le territoire chinois est divisée entre des GPU domestiques (principalement Huawei🇨🇳), des achats légaux de GPU américains (principalement Nvidia🇺🇸), et des approvisionnements illicites (via des pays tiers, des sociétés-écrans, etc.). En effet, en octobre 2022, les États-Unis ont instauré des contrôles à l'exportation qui restreignent l'accès des entités chinoises aux GPU de pointe, ainsi qu'à certains équipements pour leur fabrication[65]. Ce dispositif, dont les premières mesures sont antérieures à octobre 2022, a évolué à plusieurs reprises au gré de renforcements et d'ajustements ponctuels en sens inverse[66]. Globalement, les GPU que les entités chinoises peuvent acquérir légalement ont des performances moindres[67], afin de maintenir un retard entre les modèles IA de frontière chinois et ceux développés par les entreprises américaines. **Les volumes exacts acquis par contournement des contrôles à l'exportation restent par nature difficiles à établir, mais les estimations hautes suggèrent qu'ils pourraient représenter jusqu'à la moitié de la puissance de calcul IA disponible en Chine[68].**

---

[59] Patel et al. "AI Neocloud Playbook and Anatomy" *SemiAnalysis* (2024). https://newsletter.semianalysis.com/p/ai-neocloud-playbook-and-anatomy.

[60] appelés TPU pour *Tensor Processing Units*

[61] Op. cit. "Frontier Data Centers".

[62] Nvidia est vraisemblablement le fournisseur principal de Meta et xAI, au vu de sa position dominante sur le marché de conception des GPU (cf. Figure 4). Google, deuxième concepteur, n'a commencé à commercialiser ses TPU auprès de clients externes que depuis fin 2025. Le troisième concepteur, AMD, fournit également ces deux entreprises dans des proportions moindres.
Patel et al. "TPUv7: Google Takes a Swing at the King" *SemiAnalysis* (2025). https://newsletter.semianalysis.com/p/tpuv7-google-takes-a-swing-at-the ; "AMD Unveils Vision for an Open AI Ecosystem, Detailing New Silicon, Software and Systems at Advancing AI 2025" *AMD* (2025). https://www.amd.com/en/newsroom/press-releases/2025-6-12-amd-unveils-vision-for-an-open-ai-ecosystem-detai.html.

[63] "The next phase of the Microsoft OpenAI partnership" *OpenAI* (2026). https://openai.com/index/next-phase-of-microsoft-partnership/ ; "Expanding our use of Google Cloud TPUs and Services" *Anthropic* (2025). https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services ; "Anthropic and Amazon expand collaboration for up to 5 gigawatts of new compute" *Anthropic* (2026). https://www.anthropic.com/news/anthropic-amazon-compute.

[64] "CoreWeave Announces Agreement with OpenAI to Deliver AI Infrastructure" *CoreWeave* (2025). https://www.coreweave.com/news/coreweave-announces-agreement-with-openai-to-deliver-ai-infrastructure ; "CoreWeave Expands Agreement with OpenAI by up to $6.5B" *CoreWeave* (2025). https://www.coreweave.com/news/coreweave-expands-agreement-with-openai-by-up-to-6-5b ; "CoreWeave Announces Multi-Year Agreement With Anthropic" *CoreWeave* (2026). https://www.coreweave.com/news/coreweave-announces-multi-year-agreement-with-anthropic ; "CoreWeave and Meta Announce $21 Billion Expanded AI Infrastructure Agreement" *CoreWeave* (2026). https://www.coreweave.com/news/coreweave-and-meta-announce-21-billion-expanded-ai-infrastructure-agreement.

[65] "Implementation of Additional Export Controls: Certain Advanced Computing and Semiconductor Manufacturing Items; Supercomputer and Semiconductor End Use; Entity List Modification" *Federal Register* (2022). https://www.federalregister.gov/documents/2022/10/13/2022-21658/implementation-of-additional-export-controls-certain-advanced-computing-and-semiconductor.

[66] Sutter. "U.S. Export Controls and China: Advanced Semiconductors" *Congressional Research Service* (2025). https://www.congress.gov/crs-product/R48642.

[67] Le bridage des performances se fait notamment au niveau de la puissance de calcul totale du GPU, mais également via des limitations sur le débit des interconnections avec les modules HBM, avec d'autres GPU au sein du centre de calcul, etc.

[68] Juniewicz. "Diversion and resale: estimating compute smuggling to China" *Epoch AI* (2026). https://epoch.ai/blog/chip-smuggling.

II. Quelles dépendances stratégiques cette chaîne d'approvisionnement crée-t-elle, et quels leviers confère-t-elle aux différentes puissances ?

1. Tout projet massif d'IA se fait *de facto* avec l'accord des États-Unis

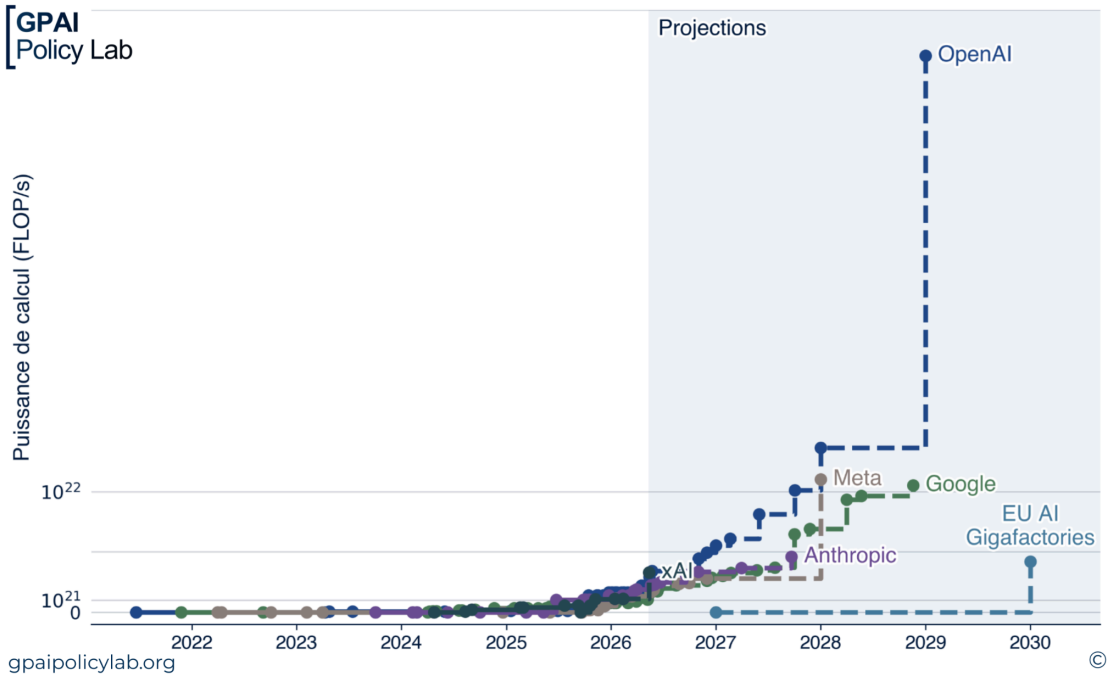

Figure 7 - Puissance de calcul des centres de calcul de frontière. Historique et projection de la puissance de calcul à disposition des entreprises de frontière localisée dans des centres de calcul de frontière, ainsi que le projet européen *AI Gigafactories*. Les données pour les entreprises de frontière, fondées sur l'identification de centres de calcul par analyse de documents administratifs et juridiques, recoupement de sources ouvertes, et imagerie satellite, sous-estiment la puissance de calcul totale des entreprises concernées. Pour Google🇺🇸 et Meta🇺🇸, elles ne couvrent qu'environ 10–20 % de leur parc total fin 2025 (cf. Figure 5). Pour le projet *AI Gigafactories*, on considère un scénario optimiste de 5 *gigafactories* de 100 000 GPU chacune, mises en service en 2030 après 3 ans de construction. Source : Epoch AI[69] et Commission européenne[70].

Comme détaillé en Section I.1, la puissance de calcul est une condition nécessaire à tout projet massif en matière d'IA. Elle est centrale non seulement pour le développement via l'entraînement, mais aussi pour l'exploitation une fois le modèle déployé, via l'inférence. La Figure 7 illustre l'évolution et les projections de la puissance de calcul totale à disposition des entreprises américaines développant des IA de frontière ainsi que le projet *AI Gigafactories* européen.

Comme détaillé en Section I.2, les acteurs qui ne sont ni américains ni chinois n'ont que deux voies d'accès à la puissance de calcul IA. La première consiste à acheter des GPU auprès de concepteurs comme Nvidia🇺🇸, concentrés à plus de 95 % aux États-Unis (cf. Figure 4). La seconde consiste à louer de la puissance de calcul via des services *cloud*, dominés par des entreprises américaines, le

[69] Op. cit. "Frontier Data Centers".

[70] "The AI Continent Action Plan" *European Commission* (2025). https://digital-strategy.ec.europa.eu/en/library/ai-continent-action-plan ; "Call for expression of interest in AI Gigafactories (AIGFs)" *European Commission* (2025). https://www.eurohpc-ju.europa.eu/document/download/47492db7-592e-4ad8-b672-9c822f94afa0_en.

reste du monde (c.-à-d. hors US et Chine) se partageant moins de 10 % de la puissance de calcul IA mondiale (cf. Figure 5).

Tout projet massif d'IA en Europe se fait donc *de facto* avec l'accord des États-Unis, qui disposent de plusieurs mécanismes concrets pour exercer ce contrôle.

Premièrement, l'accès *cloud* n'est pas un bien, mais un service, qui peut donc être soudainement suspendu ou restreint par le fournisseur sur décision des autorités.

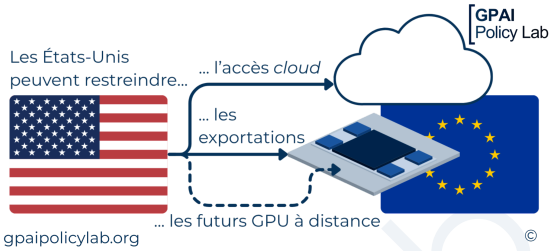

Figure 8 - Trois mécanismes de contrôle des deux voies d'accès à la puissance de calcul IA.

Deuxièmement, les contrôles à l'exportation permettent de restreindre l'accès aux nouvelles générations de GPU et aux équipements nécessaires à leur production. Les GPU déjà acquis restent fonctionnels, mais la puissance de calcul des nouvelles générations de GPU de pointe augmentant de manière exponentielle, des GPU non renouvelés deviennent obsolètes en environ 3 ans[71]. Ces contrôles sont activement utilisés, notamment à l'encontre de la Chine depuis 2018[72]. Les produits soumis à ces contrôles sont identifiés selon des critères techniques précis, comme la puissance de calcul, mesurée en FLOP/s[73], pour les GPU ou le débit de données pour les modules de mémoire HBM[74]. Le *Bureau of Industry and Security*, autorité américaine compétente, peut ensuite refuser une licence d'exportation en fonction de critères comme le pays de destination, l'entreprise cliente, ou l'usage final présumé. Par exemple, Nvidia🇺🇸 a dû développer des GPU bridés restant en dessous des seuils de contrôle spécifiquement pour pouvoir continuer à vendre sur le marché chinois. Ce dispositif de contrôle à l'exportation peut également s'appliquer à des entreprises hors États-Unis par la *Foreign Direct Product Rule*, dès lors que l'entreprise étrangère produit des biens à partir de technologies, logiciels ou équipements d'origine américaine. Cela couvre *de facto* la quasi-totalité des maillons critiques de la chaîne de production des semi-conducteurs vu la position dominante des États-Unis dans de multiples chaînons en amont (cf. Tableau 1). C'est par ce mécanisme que le gouvernement américain peut empêcher la fonderie taïwanaise TSMC🇹🇼 de produire des puces pour Huawei🇨🇳[75].

Troisièmement, des discussions sont en cours aux États-Unis sur l'intégration de mécanismes de contrôle à distance des GPU. Un texte de loi, le *Chip Security Act*, a été soumis aux deux chambres du Congrès américain en mai 2025[76], et est actuellement en attente de mise à l'ordre du jour. Le texte prévoit notamment deux échéances. Dans les six mois suivant sa promulgation, il exige l'implémentation de moyens de localisation des circuits intégrés soumis aux contrôles à l'exportation. Dans l'année suivant sa promulgation, il demande d'évaluer la nécessité et la faisabilité de « méthodes permettant de modifier les fonctionnalités de produits à circuits intégrés concernés [par les contrôles à l'exportation] qui ont été acquis de manière illicite ». À titre d'exemple, de telles méthodes permettraient au gouvernement américain d'imposer à Nvidia🇺🇸 de désactiver ou brider à

[71] La performance des GPU de pointe est multipliée par environ 1,6 chaque année (cf. Figure 2a). Après trois ans, un GPU non renouvelé se situe donc environ un demi-ordre de grandeur en dessous de la frontière, et après cinq ans, un ordre de grandeur. Rester à la frontière nécessite de les remplacer à cette échelle temporelle.

[72] Op. cit. "U.S. Export Controls and China: Advanced Semiconductors".

[73] L'unité exacte est la *Total processing performance* (TPP), qui est une implémentation particulière des FLOP/s.

[74] "The Commerce Control List." *Code of Federal Regulations* (2025). https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII/subchapter-C/part-774.

[75] Freifeld et al. "Exclusive: US ordered TSMC to halt shipments to China of chips used in AI applications" *Reuters* (2024). https://www.reuters.com/technology/us-ordered-tsmc-halt-shipments-china-chips-used-ai-applications-source-says-2024-11-10/.

[76] "S.1705 (Chip Security Act)" *Congress* (2025). https://www.congress.gov/bill/119th-congress/senate-bill/1705/text ; "H.R.3447 (Chip Security Act)" *Congress* (2025). https://www.congress.gov/bill/119th-congress/house-bill/3447/text.

distance des GPU déjà vendus et déployés dans les pays de son choix. Le mécanisme de localisation est actionnable dès aujourd'hui par mise à jour logicielle des puces Nvidia🇺🇸 déjà existantes[77]. De façon analogue, un mécanisme de modification de fonctionnalités à distance pourrait être mis en place rapidement par mise à jour logicielle[78]. Il nécessiterait à terme des modifications matérielles pour assurer sa robustesse[79], ce qui serait applicable uniquement à de futures générations de GPU, ajoutant un délai de quelques années dans la mise en application concrète de telles mesures. Ce mécanisme n'est donc pas déployé à ce jour, mais sa mise en œuvre ne se heurte à aucune impossibilité technique.

Encadré 5 - Les deux voies d'accès de l'Europe à la puissance de calcul IA, illustration par le cas de Mistral AI

Les besoins en puissance de calcul de Mistral AI illustrent les deux voies d'accès identifiées (le *cloud* et les GPU) et la dépendance vis-à-vis des États-Unis qui en résulte. Les données publiques sur l'entraînement des modèles de Mistral AI sont fragmentaires, mais suffisent à étayer ce constat.

La première voie d'accès passe par des services de *cloud*. L'essentiel[80] des entraînements de Mistral AI semble réalisé par cette voie, auprès de fournisseurs américains[81], pouvant restreindre cet accès de manière immédiate.

La seconde voie d'accès passe par l'achat de GPU. Mistral AI développe depuis 2025 sa propre offre de *cloud*, « *Mistral Compute* »[82]. L'entreprise acquiert pour cela ses propres GPU, destinés à être hébergés dans des centres de calcul en cours de développement en Europe[83]. Dans tous les cas identifiés, il s'agit de GPU Nvidia, restant donc soumis aux contrôles à l'exportation américains, ainsi qu'aux éventuels mécanismes de contrôle à distance intégrés aux futures générations de GPU lorsqu'ils renouvelleront ceux devenus obsolètes.

[77] Brass et al. "Location Verification for AI Chips" *Institute for AI Policy and Strategy* (2024). https://www.iaps.ai/research/location-verification-for-ai-chips.

[78] Petrie. "Near-Term Enforcement of AI Chip Export Controls Using A Firmware-Based Design for Offline Licensing" *arXiv* (2024). 10.48550/arXiv.2404.18308.

[79] Aarne et al. "Secure, Governable Chips: Using On-Chip Mechanisms to Manage National Security Risks from AI & Advanced Computing" *Center for a New American Security* (2024). https://www.cnas.org/publications/reports/secure-governable-chips.

[80] Deux exceptions marginales ont été identifiées, portant sur les deux premiers modèles de l'entreprise, de plus petite taille. Le supercalculateur italien Leonardo, membre du réseau EuroHPC JU, a contribué à l'entraînement de Mistral 7B (2023). Mixtral 8x7B (2023) a quant à lui été entraîné sur l'infrastructure du *neocloud* français Scaleway.

"Progress in Implementing the European Union Coordinated Plan on Artificial Intelligence (Volume 1): Italy" *OECD* (2025). https://www.oecd.org/en/publications/progress-in-implementing-the-european-union-coordinated-plan-on-artificial-intelligence-volume-1_6d530a88-en/italy_d24a6b76-en.html ; Branco. "Dans les entrailles de Nabu23, le supercalculateur d'entraînement d'IA le plus puissant d'Europe" *Les Numériques* (2024). https://www.lesnumeriques.com/intelligence-artificielle/dans-les-entrailles-de-nabu23-le-supercalculateur-d-entrainement-d-ia-le-plus-puissant-d-europe-a217434.html.

[81] Principalement CoreWeave, et possiblement Microsoft Azure

"Mistral AI Unlocks 2.5x Faster Training Speeds" *CoreWeave* (2025). https://www.coreweave.com/resources/case-studies/mistral-ai ; Boyd. "Microsoft and Mistral AI announce new partnership to accelerate AI innovation and introduce Mistral Large first on Azure" *Microsoft* (2024). https://azure.microsoft.com/en-us/blog/microsoft-and-mistral-ai-announce-new-partnership-to-accelerate-ai-innovation-and-introduce-mistral-large-first-on-azure/.

[82] Fabrion. "VivaTech : le patron de Nvidia dévoile une offre cloud avec Mistral AI et accélère en Europe" *Maddyness* (2025). https://www.maddyness.com/2025/06/11/vivatech-le-patron-de-nvidia-devoile-une-offre-cloud-avec-mistral-ai-et-accelere-en-europe/.

[83] Trois projets de centres de calcul associés à Mistral AI ont été identifiés en Europe : Eclairion (Bruyères-le-Châtel), Campus IA (Fouju) et EcoDataCenter (Suède).

Gooding. "Mistral AI raises $830m in debt financing for data center in Paris, France" *Data Centre Dynamics* (2026). https://www.datacenterdynamics.com/en/news/mistral-ai-raises-830m-in-debt-financing-for-data-center-in-paris-france/ ; Swinhoe. "MGX, Bpifrance, Nvidia, and Mistral AI plan 1.4GW Paris data center campus" *Data Centre Dynamics* (2025). https://www.datacenterdynamics.com/en/news/mgx-bpifrance-nvidia-and-mistral-ai-plan-14gw-paris-data-center-campus/ ; Swinhoe. "French AI firm Mistral signs $1.2bn deal to build EcoDataCenter facility in Sweden" *Data Centre Dynamics* (2026). https://www.datacenterdynamics.com/en/news/french-ai-firm-mistral-signs-12bn-deal-to-lease-ecodatacenter-facility-in-sweden/.

2. Les goulets de la chaîne de production sont des leviers de gouvernance de l'IA pour les puissances moyennes

La performance des GPAI dépend directement de la quantité de calcul investie dans leur entraînement (cf. Section I.1). **Gouverner la puissance de calcul revient donc à agir sur le facteur principal de leur développement.** Cette centralité se reflète dans les dispositifs réglementaires en vigueur. Par exemple, l'EU AI Act utilise le seuil de 10^{25} FLOP pour classifier un modèle d'IA en GPAI présentant des risques systémiques et déclencher les obligations associées[84], et le SB-53 californien utilise un seuil de 10^{26} FLOP pour qu'un modèle d'IA soit considéré « de frontière »[85].

Cette centralité de la puissance de calcul est modulée par les progrès logiciels[86] du développement de l'IA. L'amélioration des méthodes d'entraînement, la production de données synthétiques et/ou de meilleure qualité et l'évolution des architectures réduisent progressivement la quantité de calcul requise pour atteindre un niveau de performance donné[87]. Par conséquent, les dispositifs fondés sur des seuils de quantité de calcul nécessitent un réajustement périodique à la baisse pour conserver leur efficacité.

Plusieurs des goulets d'étranglement de la chaîne décrite en Section I.2 sont détenus par des acteurs non américains. Les monopoles et les autres points de forte concentration identifiés dans le Tableau 1 sont rappelés ici dans le Tableau 3.

Situés en amont de la production des GPU, ils constituent des étapes dont dépendent les maillons en aval, y compris ceux sous contrôle américain. **En dépit de la position dominante des États-Unis identifiée en Section II.1, les puissances qui contrôlent ces segments disposent ainsi d'un levier sur la gouvernance de la puissance de calcul, et donc de l'IA.**

Tableau 3 - Points de concentration dans la chaîne de production des GPU (cf. Tableau 1).

ACTEUR	MAILLON
🇪🇺 Europe	Photolithographie EUV
🇹🇼 Taïwan	Fabrication des puces de calcul
	Assemblage avancé
🇯🇵 Japon	Systèmes d'enduction-développement
	Substrat ABF
	Usinage, découpe et manutention des *wafers*
🇰🇷 Corée du Sud	Mémoire à haute bande passante

Cependant, deux filières coexistent aujourd'hui dans la production mondiale de GPU. La filière principale, à la frontière technologique, repose sur une chaîne de production géographiquement distribuée, mais globalement dominée par les États-Unis. La seconde filière est le résultat de la volonté de la Chine d'intégrer cette chaîne de production sur son territoire. Elle est l'alternative la plus avancée, bien que chaque maillon ait actuellement un retard significatif comparé au leader mondial correspondant.

[84] "EU Artificial Intelligence Act" *European Commission* (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
[85] "SB-53 (Transparency in Frontier Artificial Intelligence Act)" *California Legislature* (2025). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53.
[86] Parfois appelés, un peu abusivement, « progrès algorithmiques », car cela comprend les améliorations des algorithmes, mais également tous les autres progrès non matériels.
[87] Ho. "The least understood driver of AI progress" *Epoch AI* (2026). https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress.

Dans ce contexte, **agir sur un maillon ralentirait la progression des nouvelles générations de GPU** issus de la filière principale et infléchirait la progression de la taille des entraînements d'IA de frontière, donc de leurs performances générales. Par conséquent, cela rallongerait les horizons temporels de développement d'IA posant des problèmes pour la sécurité nationale et internationale. Cependant, un tel ralentissement n'empêcherait pas à long terme les entraînements de plus grande échelle, car il réduirait l'avance de la filière principale sur la filière chinoise, et une fois cette dernière parvenue à l'autonomie et à rattraper son retard, le blocage n'aurait plus d'effet. **Un tel levier serait donc efficace pour gagner du temps, mais ne permettrait pas de garantir durablement l'absence de conséquences irréversibles majeures liées à l'IA.**

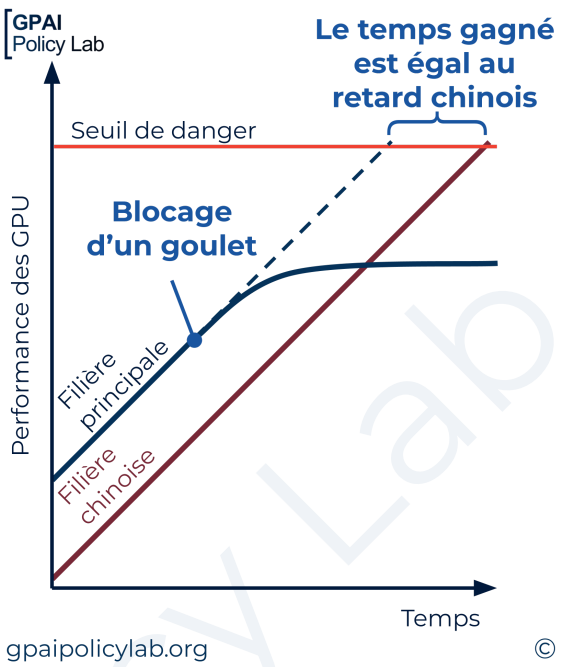

Figure 9 - Conséquence du blocage d'un goulet. Actionner un levier (blocage d'un goulet) permet de gagner du temps, jusqu'à ce que la filière chinoise rattrape son retard sur la filière principale.

Une telle situation où la filière chinoise rattrape la filière principale contreviendrait à la stratégie américaine de domination mondiale, et en particulier de la Chine, en matière d'IA[88]. Au-delà du temps qu'ils permettent de gagner, **ces leviers constituent donc, pour les pays qui les détiennent, des atouts diplomatiques leur permettant de faire valoir leurs intérêts dans les négociations face aux États-Unis.**

Cependant, utiliser ces leviers aurait un coût diplomatique conséquent, et les acteurs les mobilisant s'exposeraient à des rétorsions américaines, perspective renforcée par une administration Trump dont les modalités d'action en matière de politique internationale rompent avec les pratiques antérieures.

Or le Japon et la Corée du Sud, qui contrôlent plusieurs goulets de la filière principale (cf. Tableau 3), dépendent du parapluie nucléaire américain pour leur sécurité face à la Chine et à la Corée du Nord, deux puissances nucléarisées, sécurité obtenue dans le cadre d'alliances militaires en échange du stationnement de bases militaires américaines sur leur territoire[89]. Taïwan dépend également de la protection américaine, mais selon un schéma moins formalisé sans traité de défense ni bases permanentes[90]. Seuls, ces pays ne pourraient probablement pas utiliser ces leviers, mais une coalition de puissances moyennes apportant des garanties de sécurité alternatives pourrait les rendre actionnables.

En comparaison, les leviers européens sont actionnables et efficaces. La production mondiale de GPU, issus de la filière principale ou de la filière chinoise, dépend des machines de photolithographie

[88] "Winning the Race: America's AI Action Plan" *The White House* (2025). https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.
[89] "U.S. Extended Deterrence and Regional Nuclear Capabilities" *Congressional Research Service* (2026). https://www.congress.gov/crs-product/IF12735.
[90] Ibid. "U.S. Extended Deterrence and Regional Nuclear Capabilities".

et leur maintenance par l'entreprise néerlandaise ASML, machines qui dépendent des acteurs allemands Zeiss pour leurs miroirs et Trumpf pour leurs amplificateurs lasers. Malgré des efforts conséquents pour développer des alternatives, **le triangle ASML–Zeiss–Trumpf a un monopole sur cette technologie critique qui durera au moins quelques années, et ces deux pays ne dépendent pas des États-Unis pour leur sécurité immédiate.**

Conclusion

La performance des GPAI dépend directement de la quantité de calcul investie dans leur entraînement, ce qui fait de la puissance de calcul le facteur principal de leur progression. L'entraînement repose sur des GPU de pointe, des semi-conducteurs spécialisés dont la chaîne de production est géographiquement distribuée mais marquée par plusieurs points de forte concentration difficilement substituables. **Les États-Unis y occupent une place dominante**, avec un quasi-monopole sur la conception des GPU porté par des entreprises comme Nvidia, Google, AMD et Amazon. **D'autres goulets sont toutefois détenus par des acteurs non américains**, comme TSMC à Taïwan pour la fabrication, des équipements et des matériaux critiques produits au Japon, et surtout **le triangle européen ASML–Zeiss–Trumpf qui détient un monopole sur les machines de photolithographie EUV, technologie nécessaire à la fabrication de GPU de pointe. Une seconde filière moins avancée se développe en Chine**, alimentée notamment par les contrôles à l'exportation imposés par les États-Unis. Les GPU et maillons de production de cette seconde filière sont moins performants que leurs équivalents de la filière principale, mais l'écart pourrait se réduire.

Une fois produits, les GPU sont regroupés par centaines de milliers dans des centres de calcul qui fournissent l'infrastructure nécessaire à leur exploitation, pour l'entraînement des GPAI et pour leur déploiement. **Ce segment d'infrastructure est également très concentré et dominé par des acteurs américains** (Google, Amazon, Microsoft).

Tout projet massif d'IA se fait, *de facto*, avec l'accord des États-Unis. Les deux voies d'accès à la puissance de calcul, l'achat de GPU et la location de services *cloud*, sont toutes deux dominées par des acteurs américains. Les États-Unis disposent de moyens concrets pour exercer ce contrôle (suspension immédiate des services *cloud*, contrôles à l'exportation des GPU et des équipements de production, avec extension extraterritoriale via la *Foreign Direct Product Rule* et potentiellement des mécanismes de désactivation à distance dans de futures générations de GPU). **Les goulets non américains dans la chaîne de production confèrent néanmoins à certaines puissances moyennes un levier de gouvernance sur la puissance de calcul**, car bloquer un goulet en amont agirait sur les dépendances en aval et donc infléchirait le développement des GPU.

L'activation de ces leviers permet de repousser les entraînements de plus grande taille, ceux les plus à même de mener à des scénarios de perte de contrôle irréversibles. Mais cette stratégie n'est pas robuste sur le long terme car en ralentissant la filière principale, elle réduirait le retard de la filière chinoise, qui cherche à intégrer l'ensemble de la chaîne de production sur son territoire. Une fois l'autonomie atteinte, ces leviers ne permettront plus d'agir sur les GPU de pointe qui seraient alors chinois.

Une telle situation irait à l'encontre des ambitions américaines de domination mondiale en IA, faisant donc de ces leviers des arguments de négociation pour les puissances moyennes vis-à-vis des États-Unis. Toutefois, leur actionnement aurait un coût diplomatique fort, amplifié pour les pays

bénéficiant de la protection américaine, comme le Japon, la Corée du Sud ou Taïwan, suggérant que ces pays seraient plus à même d'actionner leurs leviers dans le cadre d'une coalition de plusieurs puissances moyennes. Par contraste, les leviers européens (ASML🇪🇺 et ses dépendances) sont actionnables immédiatement et infléchiraient la progression des GPU, au moins à l'horizon de quelques années.

Ces constats prennent toute leur portée à la lumière des enjeux de sécurité rappelés en introduction. L'entraînement des GPAI au-delà d'un certain seuil de capacité produit des comportements qu'aucun mécanisme de contrôle connu ne permet de corriger de manière fiable. **Poursuivre la trajectoire de développement actuelle expose la France et la communauté internationale à des perspectives de perte de contrôle irréversibles[91] avec des implications majeures pour la sécurité nationale.** Ces enjeux se posent à des horizons temporels courts, potentiellement dès la fin de la décennie[92], notamment via l'automatisation de la recherche en IA par l'IA[93]. Les mesures de sécurité mises en place par les entreprises de frontière sont insuffisantes pour prévenir ces scénarios[94]. À supposer qu'une dynamique collective s'engage pour les prévenir, celle-ci reposerait vraisemblablement sur une coordination internationale du développement des GPAI de frontière. Une telle coordination est conditionnée à la mise en place de mécanismes de vérification permettant à chaque acteur de s'assurer que les autres respectent leurs engagements. La nature physique et mesurable de la puissance de calcul, ainsi que la forte concentration de son exploitation, rendent une telle vérification techniquement faisable, selon une logique analogue à celle qui encadre la vérification des matières fissiles par l'AIEA dans le domaine nucléaire. Nos travaux sur les mécanismes de vérification[95] détaillent ainsi plusieurs scénarios techniques actionnables à différents horizons temporels, allant de mesures indirectes réalisables dès aujourd'hui à des approches fondées sur des preuves cryptographiques permettant une vérification fine sans compromettre la confidentialité des secrets industriels et étatiques.

Suite possible à cette note

1. Quels goulets d'étranglement supplémentaires apparaissent dans la chaîne d'approvisionnement de la puissance de calcul lorsque ses dépendances sont cartographiées plus finement ? Peut-on estimer leur rapidité d'impact sur l'aval et leur durabilité, notamment face à la filière chinoise ?
2. Comment l'activation de ces goulets module-t-elle les modèles de menace de perte de contrôle, notamment celui découlant de l'automatisation de la R&D en IA ?
3. En quoi l'organisation de la production de la puissance de calcul donne-t-elle à la France et à ses alliés des leviers pour développer des mécanismes de vérification permettant d'obtenir des garanties de sécurité dans le cadre d'une coordination internationale du développement d'IA de frontière ?

---

[91] Maier et al. "Take Goodhart Seriously: Principled Limit on General-Purpose AI Optimization" *arXiv* (2025). 10.48550/arXiv.2510.02840.

[92] Op. cit. "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI" ; Andréoletti et al. "Position: The Window for Preventive Action Before AI Saturates Most Cognitive Benchmarks is Closing" *HAL* (2026). https://hal.science/hal-05500135.

[93] "Scénarios d'automatisation de la R&D en IA : Conséquences sur l'évolution des capacités et les risques émergents" *GPAI Policy Lab* (à paraître).

[94] "Frontier AI Safety Frameworks : Approches et limites face aux scénarios de loss of control" *GPAI Policy Lab* (2025).

[95] "Garanties de contrôle de l'IA et mécanismes de vérification : Implications et options pour la France" *GPAI Policy Lab* (2026) ; Peigné-Lefebvre et al. "Zero knowledge verification for frontier AI training is possible" *GPAI Policy Lab* (2026). https://docs.google.com/document/d/1pec8fzSma4zPIbpETToZXPv-SSJoR2fn.

Annexes

A. Estimation d'une borne haute de la puissance de calcul IA à disposition d'acteurs français

On considère la puissance de calcul à disposition d'acteurs français, c'est-à-dire d'une part les acteurs publics, et d'autre part les acteurs privés dont le siège de la maison-mère est en France. Les acteurs publics disposant de puissance de calcul IA relèvent principalement de la recherche scientifique et de la défense. Dans les deux cas, des informations publiques permettent d'estimer le total à leur disposition. En revanche, les acteurs privés français ne communiquent généralement pas ce type d'information. En l'absence de registre public des centres de calcul (cf. Section I.2), il est donc impossible d'estimer précisément cette puissance de calcul, et donc le total à disposition d'acteurs français.

Au lieu d'estimer la puissance de calcul IA à disposition d'acteurs français, on détermine une borne haute, informée par les données du secteur public. Seule la puissance des GPU est comptabilisée, les CPU étant négligeables en comparaison pour les calculs d'IA.

Les acteurs publics français disposant d'une puissance de calcul significative relèvent principalement de la recherche scientifique et de la défense. La majorité de cette puissance de calcul est coordonnée par le GENCI (Grand Équipement National de Calcul Intensif), qui structure l'écosystème de calcul intensif français sur trois niveaux[96]. Le premier est l'écosystème européen *European High-Performance Computing Joint Undertaking* (EuroHPC JU), au sein duquel la France ne dispose pas encore de puissance de calcul hébergée sur son territoire (le supercalculateur Alice Recoque sera le premier). Le deuxième est l'écosystème national, composé du CINES (Centre Informatique National de l'Enseignement Supérieur), de l'IDRIS (Institut du Développement et des Ressources en Informatique Scientifique) et du TGCC (Très Grand Centre de Calcul du CEA). Le troisième est l'écosystème régional MesoNET. À cela s'ajoutent les ressources propres du CEA (Commissariat à l'énergie atomique et aux énergies alternatives) et de l'AMIAD (Agence ministérielle pour l'intelligence artificielle de défense).

L'écosystème régional et les autres acteurs publics disposant de puissance de calcul (le *cluster* Grid'5000 de l'Inria[97], les supercalculateurs Belenos et Taranis de Météo-France[98], etc.) contribuent de manière marginale. On estime leur apport combiné à environ 10 % de la puissance de calcul des centres principaux.

[96] "Notre écosystème" *GENCI*. https://www.genci.fr/connaitre-genci/notre-ecosysteme
[97] "Grid'5000 : l'expérimentation informatique à grande échelle" *Inria* (2017). https://www.inria.fr/fr/grid5000-lexperimentation-informatique-grande-echelle
[98] "Les supercalculateurs de Météo-France" *Météo-France* (2025). https://meteofrance.com/actualites-et-dossiers/comprendre-la-meteo/les-supercalculateurs-de-meteo-france.

Tableau 4 - Puissance de calcul IA à disposition d'acteurs publics français.

CENTRE DE CALCUL	SUPERCALCULATEUR	PUISSANCE DE CALCUL (FLOP/s)
Centre Informatique National de l'Enseignement Supérieur (CINES)	Adastra[99]	$7{,}7 \cdot 10^{17}$
Institut du Développement et des Ressources en Informatique Scientifique (IDRIS)	Jean Zay[100]	$3{,}4 \cdot 10^{18}$
	Dalia[101]	$3{,}8 \cdot 10^{17}$
Très Grand Centre de Calcul du CEA (TGCC)	Joliot-Curie[102]	$1{,}6 \cdot 10^{16}$
Centre de Calcul Recherche et Technologie (CCRT)	Topaze[103]	$1{,}9 \cdot 10^{17}$
Direction des Applications Militaires du CEA (CEA-DAM)	EXA1[104]	$9{,}4 \cdot 10^{17}$
Agence ministérielle pour l'intelligence artificielle de défense (AMIAD)	ASGARD[105]	$5{,}1 \cdot 10^{18}$
Autre (MesoNET, Inria, Météo-France, etc.)	—	10 % du dessus = $1{,}1 \cdot 10^{18}$
TOTAL		$1{,}2 \cdot 10^{19}$

Les acteurs privés sont moins transparents sur la puissance de calcul dont ils disposent. Les principaux sont les fournisseurs de services *cloud* français (OVHcloud, Scaleway, Sesterce), auxquels s'ajoutent des entreprises disposant de centres de calcul pour leur propre usage (Atos, TotalEnergies, EDF, etc.). D'autres entreprises du numérique, comme Mistral AI ou FlexAI, ne possèdent pas d'infrastructure de calcul qui leur est propre à ce jour.

Pour estimer la puissance de calcul à leur disposition, on s'appuie sur le nombre de *data centers* associés à ces acteurs recensés sur la plateforme Data Center Map[106], qui en identifie 54 (cf. Tableau 5). En comparaison des sept centres de calcul publics identifiés dans le Tableau 4, en supposant que la variation de la puissance de calcul entre les centres de calcul soit similaire dans les deux groupes, on estime la puissance de calcul IA à disposition des acteurs privés à l'aide d'une règle de trois, ce qui donne ($1{,}2 \cdot 10^{19}$ / 7 · 54 =) $9{,}3 \cdot 10^{19}$ FLOP/s.

[99] "Centre Informatique National de l'Enseignement Supérieur (CINES)" *GENCI*. https://www.genci.fr/centre-informatique-national-de-lenseignement-superieur-cines

[100] "Institut du Développement et des Ressources en Informatique Scientifique (IDRIS)" *GENCI*. https://www.genci.fr/institut-du-developpement-et-des-ressources-en-informatique-scientifique-idris

[101] "Présentation de Dalia" *IDRIS*. http://www.idris.fr/docs/dalia/dalia-presentation/

[102] "Très Grand Centre de calcul du CEA (TGCC)" *GENCI*. https://www.genci.fr/tres-grand-centre-de-calcul-du-cea-tgcc

[103] "CCRT - MOYENS DE CALCUL" *CEA*. https://www-hpc.cea.fr/fr/CCRT-moyens.html

[104] Trueman. "Eviden delivers second EXA1 supercomputer to France's CEA" *Data Centre Dynamics* (2024). https://www.datacenterdynamics.com/en/news/eviden-delivers-second-exa1-supercomputer-to-frances-cea/

[105] Lagneau. "Intelligence artificielle : Le ministère des Armées a inauguré le supercalculateur le plus puissant d'Europe" *Zone Militaire* (2025). https://www.opex360.com/2025/09/05/intelligence-artificielle-le-ministere-des-armees-a-inaugure-le-supercalculateur-le-plus-puissant-deurope/

[106] *Data Center Map*. https://www.datacentermap.com/

Tableau 5 - Nombre de *data centers* appartenant à des acteurs privés français recensé par Data Center Map[107].

Organisation	Nombre de *data centers*
OVHcloud	34
Scaleway	5
Sesterce	1
Atos	10
TotalEnergies	1
EDF	3
TOTAL	54

En additionnant la contribution des acteurs publics et des acteurs privés, on obtient un total de $(1{,}2 \cdot 10^{19} + 9{,}3 \cdot 10^{19} =) \ 1{,}1 \cdot 10^{20}$ FLOP/s. Le total mondial étant d'environ $4 \cdot 10^{22}$ FLOP/s (cf. Figure 5), cette estimation indique que la puissance de calcul IA à disposition d'acteurs français est en dessous de 1 % du total mondial d'un facteur 4 environ.

[107] Ibid. *Data Center Map*.