Cloud, Compute et Souveraineté : État des Lieux d'une Industrie en Rupture

Rétrospective et état des lieux du cloud, du silicium et de la souveraineté numérique.

Nicolas Masselot

Juin 2026

Table des matières

Brefs Objectifs

Ce document propose une rétrospective et un état des lieux du secteur du cloud et de l'infrastructure IA, à l'échelle mondiale, européenne et française. Il a été construit comme un carnet de connaissance personnel, destiné à centraliser ma compréhension du domaine dans un format clair et évolutif, et il sera mis à jour régulièrement pour suivre l'actualité.

Il s’est élargi au fur et à mesure de son écriture. Parti de l'économie du cloud, il descend progressivement vers les couches matérielles, le silicium, la mémoire, l'énergie, puis remonte vers les enjeux de souveraineté. J'ai cherché à le rendre accessible à un lecteur non technicien comme à quelqu'un du métier, en posant chaque notion avant de m'en servir. Cet objectif n'est pas toujours pleinement atteint : le sujet est si interconnecté que certains termes apparaissent une première fois avant d'être vraiment expliqués plus loin. Quand c'est le cas, un retour en arrière ou le glossaire d'ouverture permettent de lever l'ambiguïté.

Le document ne prétend pas à l'exhaustivité et ne constitue pas un livrable professionnel : il n'est pas sourcé de manière académique et certains concepts de base ne sont pas redéfinis. Son ambition est avant tout de structurer une réflexion personnelle sur les dynamiques de fond du secteur.

Vocabulaire utile (souligné dans le texte)

Bare Metal : serveur physique dédié à un seul client, sans couche de virtualisation intermédiaire. Le client dispose directement des ressources matérielles de la machine, ce qui améliore généralement les performances et la prévisibilité.

Machine virtuelle (VM) et hyperviseur : une machine virtuelle est une machine simulée logiciellement, qui tourne sur un serveur physique partagé entre plusieurs clients. L'hyperviseur est le logiciel qui découpe le serveur physique en plusieurs VMs indépendantes.

Noisy neighbor : phénomène où plusieurs clients partageant le même serveur physique via la virtualisation voient leurs performances dégradées par un client qui consomme excessivement les ressources communes. C'est un inconvénient de la mutualisation cloud.

Accélérateur IA : puce spécialisée dans l'entraînement ou l'inférence des modèles d'intelligence artificielle. Contrairement aux processeurs généralistes (CPU), les accélérateurs sont optimisés pour les calculs massivement parallèles nécessaires aux réseaux de neurones. Les GPU de NVIDIA constituent aujourd'hui les accélérateurs les plus répandus.

FLOPS : unité mesurant le nombre d'opérations mathématiques qu'un processeur ou un accélérateur peut effectuer chaque seconde.

L'Inférence est la phase d'utilisation d'un modèle d'IA déjà entraîné. Le modèle mobilise ses paramètres pour générer une réponse.

Machine de lithographie : équipement utilisé pour graver les circuits d'une puce sur une tranche de silicium (**wafer**). Ces machines projettent des motifs microscopiques qui déterminent l'emplacement des milliards de transistors composant les processeurs modernes. Les machines les plus avancées, dites EUV (Extreme Ultraviolet), sont produites exclusivement par le groupe néerlandais ASML.

Le rack-scale AI désigne la nouvelle approche, popularisée par NVIDIA avec ses systèmes GB200, qui consiste à concevoir et vendre l'unité de calcul non plus comme une puce ou un serveur individuel, mais comme une armoire (rack) entière préassemblée, intégrant des dizaines de GPU, leurs mémoires, leurs interconnexions, leur alimentation et leur refroidissement liquide.

CUDA (Compute Unified Device Architecture) est la plateforme logicielle propriétaire développée par NVIDIA depuis 2006 pour permettre aux chercheurs et développeurs d'utiliser ses GPU pour des calculs autres que graphiques. Son langage de programmation traduit le code en instructions exécutables sur les puces NVIDIA.

Modèle de fondation : grand modèle d'IA entraîné sur des données massives et conçu pour être adapté ensuite à de multiples tâches downstream (assistance, code, analyse, traduction).

Project Rainier est le projet d'Amazon AWS et Anthropic, déployé dans l'Indiana, regroupant plus de 500 000 puces Trainium2 d'Amazon et soutenu par un engagement de 100 milliards de dollars d'AWS sur dix ans.

Qualcomm est un géant américain des puces pour smartphones et modems cellulaires, qui

conçoit ses puces sans les fabriquer lui-même (**Fabless**). Fort de son savoir-faire en efficacité énergétique hérité du mobile, il tente depuis peu d'entrer sur le marché de l'inférence IA en datacenter.

Un Transformer est le type d'architecture de réseau de neurones sur lequel reposent les grands modèles de langage actuels, comme ChatGPT ou Claude. Introduit par Google en 2017, son principe clé est le **mécanisme d'attention**, qui permet au modèle de pondérer l'importance de chaque mot par rapport à tous les autres dans un texte, et donc de saisir le contexte.

Le CHIPS Act, est une loi américaine votée en 2022 de relocalisation des semi-conducteurs, prévoyant des subventions massives pour les foundries qui construisent des usines sur le sol américain. Ses bénéficiaires principaux sont TSMC (pour ses usines en Arizona), Intel, Micron et Samsung.

Le 18A est le nom du procédé de gravure le plus avancé d'Intel, celui qui doit le remettre au niveau de TSMC.

La scaling law désigne la loi empirique selon laquelle la performance d'un modèle de langage croît de manière prévisible avec la taille du modèle, la taille du dataset et la quantité de calcul investie dans l'entraînement. Cette loi a fondé toute la stratégie d'investissement de l'industrie entre 2020 et 2024 : empiler plus de paramètres, plus de données, plus de GPU.

Open weights : désigne un modèle dont les les « poids », c'est-à-dire les milliards de nombres qui définissent le comportement du modèle, sont publiés et téléchargeables. Llama (Meta), Mistral et DeepSeek sont les exemples les plus marquants de cette approche, qui s'oppose aux modèles purement propriétaires d'OpenAI ou d'Anthropic accessibles uniquement via API.

SMR (Small Modular Reactor) : Réacteur nucléaire de petite taille (50 à 300 MW) fabriqué en série en usine puis assemblé sur site, visant à réduire drastiquement les coûts et les délais face aux centrales géantes.

SecNumCloud est une qualification de sécurité délivrée par l'agence française de cybersécurité. Elle impose des critères stricts comme l'immunité aux lois extraterritoriales (CLOUD Act), le stockage des données en Europe et un contrôle capitalistique européen. Elle vise à protéger les données « critiques » françaises. Une dizaine d'acteurs sont qualifiés à ce jour, dont OVHcloud, Outscale (Dassault Systèmes), Cloud Temple et S3NS, tandis que d'autres comme NumSpot ou l'offre Bleu sont encore en cours de qualification.

L'AI Act est un règlement européen adopté en 2024 qui classe les systèmes d'IA selon leur niveau de risque et impose des obligations croissantes (transparence, audit, certification) selon ce niveau. Son application a été progressivement repoussée sous la pression de l'industrie, certaines obligations principales ne devant entrer en vigueur qu'en décembre 2027.

Kyutai est un laboratoire de recherche en IA à but non lucratif français. Son ambition est « open science » : publier librement ses recherches, sans pression de rentabilité immédiate. Il incarne le pari que l'Europe peut produire de la recherche fondamentale en IA hors des logiques propriétaires américaines.

L'**InvestAI Facility** est le mécanisme financier de l'Union européenne annoncé en 2025, doté

de 20 milliards d'euros, visant à financer jusqu'à cinq gigafactories d'IA sur le continent à l'horizon 2030.

SaaS, PaaS, IaaS : trois niveaux d'abstraction du cloud computing.

- En **IaaS** (Infrastructure as a Service), le fournisseur loue de l'infrastructure brute (serveurs virtuels, stockage, réseau) et le client gère lui-même son système d'exploitation, ses bases de données et ses applications.
- En **PaaS** (Platform as a Service), le fournisseur livre une plateforme prête à exécuter du code (base de données managée, **kubernetes** managé), et le client n'a qu'à déployer son application.
- En **SaaS** (Software as a Service), le fournisseur livre une application finie utilisable directement (Salesforce, Notion, Microsoft 365).

Conteneur : technologie qui permet d'emballer une application avec tout ce dont elle a besoin pour fonctionner, de manière à ce qu'elle tourne de façon identique sur n'importe quelle machine. Beaucoup plus léger et rapide qu'une machine virtuelle, le conteneur est devenu le format standard de déploiement des applications modernes.

Kubernetes : logiciel qui gère automatiquement des centaines ou milliers de conteneurs en production.

Partie I : L'Économie du Cloud et la Révolte contre les Hyperscalers

À l'origine, le cloud a d'abord été présenté comme un modèle flexible et économique : les entreprises pouvaient louer de la puissance informatique sans investir elles-mêmes dans des serveurs ou des data centers. Cela permettait de substituer des Opex souples aux Capex importants nécessaires à l'installation des serveurs.

Mais avec la généralisation du cloud, les hyperscalers ont progressivement construit des modèles tarifaires qui renforcent la dépendance des clients et maximisent leurs marges.

Les Frais d'Egress : le mécanisme de verrouillage

L'un des principaux mécanismes de verrouillage du cloud repose sur les « egress fees », les frais facturés lorsqu'un client veut récupérer ses données hors du cloud.

Les hyperscalers comme AWS rendent l'entrée des données (ingress) gratuite, afin d'attirer les entreprises vers leurs infrastructures. En revanche, transférer ces données vers Internet ou vers un autre cloud (egress) devient coûteux.

Cette asymétrie crée une forte dépendance : plus une entreprise stocke de données chez un hyperscaler, plus il devient cher et complexe d'en sortir.

Des acteurs comme **Cloudflare** ont tenté de dénoncer cette rente en proposant zéro frais d'egress, mais le vrai coup de grâce est venu de la régulation. Entré en vigueur en 2024, le **European Data Act** a imposé la libre circulation des données, forçant Google Cloud puis AWS à abolir les frais de sortie pour les clients souhaitant clôturer leurs comptes et migrer leurs charges de travail.

L'Émergence du FinOps, l'Alternative BYOC

La complexité volontaire des factures cloud (des fichiers de facturation contenant des millions de lignes) a donné naissance à la discipline du **FinOps** : des entreprises paient des consultants et des logiciels dédiés simplement pour comprendre ce qu'elles consomment. C'est l'aveu d'un échec d'alignement entre le fournisseur et le client.

Face à cette dépendance croissante, le modèle du **BYOC** (Bring Your Own Cloud) s'est développé. Des entreprises comme Snowflake ou Databricks proposent désormais à leurs clients d'utiliser leurs outils directement dans leur propre infrastructure cloud (AWS, Azure, Google Cloud), sans sortir les données vers les serveurs du fournisseur. Cela permet de garder davantage de contrôle sur les données et de limiter les frais d'egress.

Le Débat sur la Rapatriation (Bare Metal)

Pendant des années, quitter le cloud public pour revenir à des serveurs physiques était perçu

comme un retour en arrière. Plusieurs entreprises ont pourtant montré que ce choix pouvait être plus rationnel à grande échelle. La rapatriation peut prendre deux formes : louer des serveurs dédiés en **bare metal** chez un acteur comme OVHcloud, ou racheter ses propres serveurs et les exploiter en propre.

Le principal avantage du cloud public est la flexibilité : les entreprises peuvent augmenter ou réduire instantanément leurs ressources et déléguer toute la gestion de l'infrastructure. Mais cette flexibilité a un coût élevé. Pour des entreprises aux besoins de calcul importants et relativement stables, exploiter directement leurs propres serveurs revient souvent beaucoup moins cher que louer en permanence des ressources chez AWS, Azure ou Google Cloud.

Le retour au bare metal améliore aussi les performances. En supprimant les couches de **virtualisation** propres au cloud public, les applications gagnent en vitesse et ne subissent plus les ralentissements liés au **partage des ressources avec d'autres clients.**

Du Cloud d'abondance au Cloud de Pénurie

Pendant vingt ans, le cloud a vendu une ressource abondante et largement interchangeable : de la puissance de calcul classique et du stockage, disponibles à la demande. Le rapport de force était commercial : le fournisseur retenait son client par les frais de sortie, la complexité tarifaire et le verrouillage décrits plus haut.

La vague de l'IA a chamboulé le métier des acteurs du cloud. Entraîner et faire tourner un modèle exige des processeurs spécialisés, les GPU, qui se louent chez ces mêmes acteurs, à l'heure ou à la minute, exactement comme on louait hier un serveur web. Le métier reste de la location de capacité à distance. Mais un nouveau rayon a ouvert, et il a été pris d'assaut.

La ressource qui compte désormais n'est plus le calcul générique mais **l'accélérateur**, le GPU, et il est rare. Les investissements sont titanesques, puisque les quatre géants américains du cloud, les hyperscalers prévoient pour 600 milliards de dollars de dépenses d'infrastructure en 2026, après environ 390 milliards un an plus tôt.

Pire, les hyperscalers ne sont plus bridés par la demande mais par leur propre capacité d'offre. Microsoft accumule environ 80 milliards de dollars de commandes Azure qu'il ne peut honorer, des GPU dormant en stock faute d'électricité pour les alimenter. Sundar Pichai a de son côté admis que le chiffre d'affaires cloud de Google aurait été plus élevé s'il avait pu répondre à la demande.

Ce basculement change qui détient le pouvoir. Tant que la ressource était abondante, l'avantage allait à celui qui maîtrisait la relation commerciale, par les frais de sortie et le verrouillage. Maintenant qu'elle est rare, il va à celui qui contrôle le matériel : les puces et l'énergie pour les faire tourner. Ce qui décide du rapport de force, désormais, c'est l'accès au silicium.

Partie II : Sous le Capot de l'IA : mémoire, puces et interconnexions

Cette partie pose la grammaire technique des infrastructures d'IA modernes. Du mur mémoire aux compilateurs, elle présente les contraintes physiques et logicielles qui déterminent aujourd'hui les performances réelles des puces d'IA.

Le Mur Mémoire et la Bande Passante

Le critère suprême, lorsqu'on parle de GPU, a souvent été la puissance de calcul : combien d'opérations par seconde. La réalité est qu'une puce passe souvent son temps à attendre les données plutôt qu'à calculer. C'est le « mur mémoire » : les unités de calcul sont devenues si rapides que le facteur limitant n'est plus le calcul, mais la vitesse à laquelle on parvient à leur acheminer les données depuis la mémoire.

Quand la puce est bridée par la bande passante mémoire, elle est dite **memory-bound** ; quand elle est bridée par le calcul, on parle de **compute-bound**. Une puce peut afficher des **FLOPs** spectaculaires et rester inutilisée à 80 % si on n'arrive pas à la nourrir assez vite.

C'est exactement ce qui se joue dans **l'inférence** des grands modèles, qui se décompose en deux phases aux profils opposés.

- Le **prefill** (lecture et traitement du prompt), qui traite tous les tokens d'entrée en parallèle : ce sont de grosses multiplications de matrices, fortement compute-bound.

- Le **decode** (génération de la réponse token par token) : pour produire un seul token, le système doit relire l'intégralité des poids du modèle depuis la mémoire. Il recharge des gigaoctets de poids pour un travail de calcul minuscule, un travail profondément memory-bound.

  Pour ne pas tout recalculer à chaque token au fur et à mesure que la conversation s'allonge, le modèle garde en mémoire les calculs des tokens déjà traités (le **KV cache**), une réserve qui gonfle avec la longueur du contexte et dévore à son tour la bande passante.

HBM : le Rapprochement de la Mémoire et du Calcul

Pour nourrir les GPU modernes, il ne suffit plus d'ajouter davantage de mémoire : il faut la rapprocher physiquement du calcul.

L'industrie a répondu à ce problème avec la **HBM** (**High Bandwidth Memory**). L'idée est simple : au lieu de placer la mémoire loin du processeur, on la rapproche au maximum. Plusieurs couches de mémoire DRAM (mémoire vive) sont empilées verticalement les unes sur les autres et reliées par des milliers de connexions microscopiques traversant le silicium. Cette pile de mémoire est ensuite installée juste à côté du GPU (voir le schéma ci-dessous).

Les distances parcourues par les données deviennent beaucoup plus courtes et le nombre de

connexions beaucoup plus élevé. Là où une barrette DDR (classique) offre quelques dizaines de Go/s de bande passante, une pile HBM fournit plusieurs To/s.

Cette évolution a toutefois une conséquence importante : mémoire et processeur deviennent de plus en plus indissociables. Les futures générations de HBM doivent être conçues en coordination étroite avec les GPU eux-mêmes.

Le développement d'une puce d'IA dépend désormais simultanément de trois acteurs : le concepteur du processeur (NVIDIA, AMD), le fondeur chargé de l'assemblage (TSMC) et le fabricant de mémoire (SK Hynix, Samsung ou Micron).

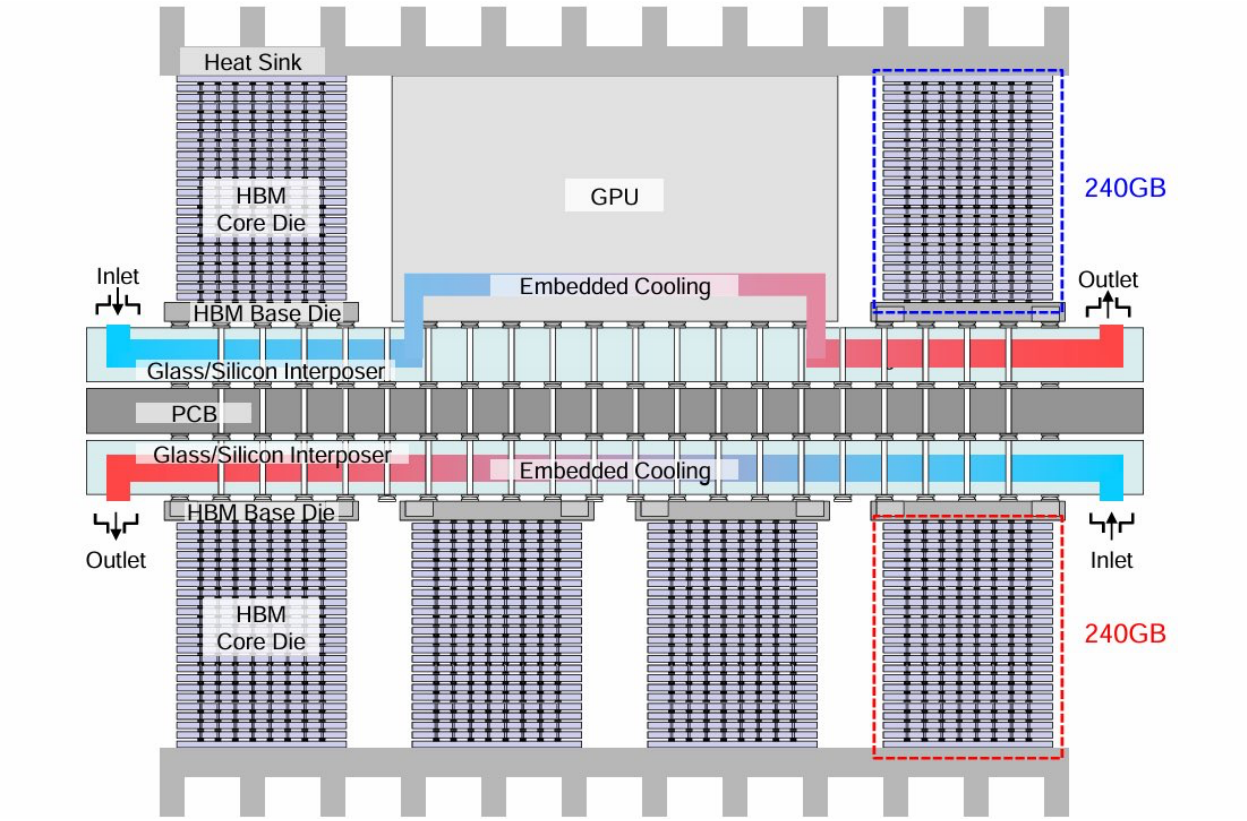
Packaging avancé : le nouveau verrou de l'industrie

Avec la HBM, les systèmes modernes ne se limitent plus à un GPU isolé : ils combinent plusieurs composants spécialisés : GPU, mémoire HBM et interconnexions ultra-rapides, assemblés dans un même ensemble. Cette étape d'intégration est appelée **packaging avancé**.

Cette évolution est renforcée par une contrainte physique fondamentale. Les GPU les plus avancés approchent déjà les limites maximales de taille imposées par les **machines de lithographie**. Il devient donc impossible de concentrer toujours plus de calcul et de mémoire sur une seule puce. L'industrie doit désormais assembler plusieurs composants distincts pour construire un système complet.

TSMC domine aujourd'hui ce marché grâce à sa technologie **CoWoS** (Chip-on-Wafer-on-Substrate), qui permet de connecter très étroitement GPU et HBM. Cette capacité d'assemblage est devenue aussi importante que la fabrication des puces elles-mêmes.

Le schéma ci-dessous illustre à quoi pourrait ressembler un package de prochaine génération :



Source : KAIST TERA Lab

Au centre, le GPU. Autour, les piles de mémoire HBM, posées au plus près du calcul. Chaque pile additionne les couches de mémoire (les **core dies**) au-dessus d'une puce de contrôle faisant le lien avec le reste du système (le **base die**).

Le tout repose sur **l'interposer**, la plaque qui relie le GPU à ses mémoires par des milliers de connexions. Pour évacuer la chaleur, le refroidissement n'est plus un simple radiateur posé sur le dessus : un liquide circule à l'intérieur même de l'ensemble, entrant froid (en bleu) et ressortant chaud (en rouge).

Interconnexion : Scale-up contre Scale-out

Un seul GPU ne suffit pas pour les grands modèles d'IA. Les modèles les plus avancés sont trop volumineux pour tenir dans la mémoire d'une seule puce et nécessitent donc de répartir le calcul sur de nombreux accélérateurs.

Deux approches complémentaires existent.

- Le **scale-up** consiste à relier étroitement un nombre limité de GPU afin qu'ils se comportent comme un seul système cohérent. L'objectif est de partager la mémoire et les

calculs le plus rapidement possible.

- Le **scale-out**_consiste à connecter entre eux de nombreux serveurs déjà équipés de plusieurs GPU. On passe alors d'un système unique à un véritable cluster pouvant regrouper des milliers, voire des dizaines de milliers d'accélérateurs.

La différence fondamentale est que le scale-up privilégie la vitesse des échanges entre un petit nombre de GPU, tandis que le scale-out privilégie la taille totale du système. Dans les deux cas, la qualité des interconnexions devient déterminante.

En effet, lorsqu'un modèle est réparti sur de nombreuses puces, une partie croissante du temps est consacrée non plus au calcul lui-même mais au déplacement des données. Les performances réelles d'un cluster dépendent donc autant de son réseau interne que de la puissance des GPU qui le composent.

C'est pourquoi **l'interconnexion** est devenue un champ de bataille stratégique. Dans les infrastructures d'IA modernes, déplacer efficacement les données est devenu presque aussi important que les calculer.

Cette logique de scale-up a une traduction physique : **le rack**. Autrefois simple armoire métallique où l'on empilait des serveurs indépendants, il est devenu l'unité de base du calcul IA, où des dizaines de GPU sont reliés si étroitement qu'ils se comportent comme un seul processeur géant.

La Pile Logicielle et les Compilateurs : le Vrai Fossé

Un GPU est une machine de calculs parallèles. Pour l'exploiter, il faut découper chaque calcul en milliers de petites tâches simultanées, les **kernels**, les répartir sur le matériel, gérer la hiérarchie de mémoire et organiser les transferts de données.

Le maillon décisif est le **compilateur** : la couche logicielle qui traduit un modèle, écrit par exemple dans PyTorch, en code optimisé pour une puce précise. C'est lui qui choisit les kernels, fusionne les opérations et organise l'accès à la mémoire. Un fabricant peut concevoir une puce dont la puissance de calcul et la bande passante sont excellentes : si son compilateur ne sait pas y exécuter efficacement n'importe quel modèle, la puce reste largement inexploitée. **Le facteur limitant est rarement le silicium, mais le logiciel qui en tire la performance.**

C'est le principal obstacle pour les concurrents de NVIDIA, et c'est pourquoi l'industrie cherche à se doter de couches logicielles portables, indépendantes d'un seul fabricant.

La Fin de la Loi de Moore : du Transistor au Système

Pendant cinquante ans, le progrès a suivi la **loi de Moore** : on est parvenu à doubler la densité de transistors tous les deux ans en les miniaturisant. Cette dynamique se heurte désormais à des limites physiques et économiques. Continuer à rétrécir exige la lithographie EUV d'ASML, à un coût démesuré et au prix d'une dépendance géopolitique extrême.

L'industrie déplace donc son axe d'optimisation du transistor vers le système : la performance vient moins de la finesse de gravure que de la façon d'assembler les puces et de faire circuler les

données entre elles (packaging avancé, chiplets, interconnexions ultra-rapides). Le centre de gravité passe du calcul brut au mouvement des données.

Ce basculement profite d'abord à ceux que l'on prive de gravure fine : faute de pouvoir miniaturiser, ils optimisent l'assemblage et la circulation des données. Le chinois Huawei, privé d'accès aux machines de gravure les plus avancées, en est l'exemple type.

Le même ressort joue côté logiciel, quand une architecture mieux pensée l'emporte sur la force brute : c'est ce qu'a réussi le laboratoire chinois DeepSeek, en obtenant avec des moyens de calcul réduits des résultats que l'on croyait réservés aux plus gros clusters.

Partie III : L'Hégémonie et l'Épopée NVIDIA

NVIDIA ne doit pas sa domination à une puce, mais à la maîtrise simultanée de tous les maillons de la chaîne : le logiciel qui rend ses puces exploitables, l'interconnexion qui les relie, la mémoire qui les alimente, le rack qui les assemble, et l'écosystème de développeurs qui rend l'ensemble incontournable. NVIDIA ne vend pas un accélérateur mais une plateforme complète.

Le Fossé CUDA : Pourquoi NVIDIA est une Entreprise Logicielle

L'avance de NVIDIA se situe d'abord sur la couche logicielle, où elle compte près de vingt ans d'avance. Dans les années 2000, Jensen Huang a pris une décision controversée : intégrer les composants **CUDA** dans chaque puce NVIDIA, y compris celles destinées aux joueurs. Ce pari a formé une génération entière de chercheurs et d'ingénieurs qui n'ont appris à programmer que sur CUDA.

Aujourd'hui, CUDA est un écosystème complet, développé en permanence en parallèle des puces. Sa force tient surtout à un ensemble de bibliothèques optimisées, comme **TensorRT-LLM**, qui contiennent des années de réglages accumulés et font tourner rapidement tout nouveau modèle dès sa sortie.

C'est aussi ce qui permet à des puces déjà anciennes, comme la série H100, de voir leurs performances doubler ou tripler par simple mise à jour logicielle, sans modification matérielle.

C'est ce qui explique qu'aucun concurrent ne comble ce fossé en quelques années : déplacer une charge de travail hors de CUDA suppose de réécrire et revalider des millions de lignes de code, de renoncer à des bibliothèques affinées pendant quinze ans, et de reformer des ingénieurs qui n'ont jamais connu autre chose.

On peut copier une puce, mais pas vingt ans d'écosystème logiciel ni la base d'ingénieurs formée à l'utiliser.

De Hopper à Blackwell : du Composant au Système

L'architecture Hopper, lancée en 2022, a donné la H100, standard absolu de l'entraînement des grands modèles : c'est sur des clusters de H100 que GPT-4, Claude et Llama 3 ont été entraînés. NVIDIA en a écoulé des millions avec des marges supérieures à 70 %.

La rupture conceptuelle est venue avec Blackwell (annoncée en mars 2024) : au lieu d'une puce monolithique, deux dies de calcul géants reliés par une interconnexion à très haute bande passante. Cette complexité d'assemblage reposait sur le packaging avancé « CoWoS-L » de TSMC (voir Partie II).

Un défaut de conception a créé une crise temporaire, ouvrant une brève fenêtre pour AMD et les puces maison des hyperscalers, et cet épisode a directement façonné le design de Rubin, pensé pour minimiser ces fragilités physiques. Blackwell, puis Blackwell Ultra, sont en production de masse depuis 2025.

L'essentiel est ailleurs que dans les téraflops : avec Blackwell, le produit cesse d'être une puce pour devenir un rack.

Le Rack comme Produit : GB200 NVL72 puis Vera Rubin

Avec Blackwell, NVIDIA a cessé de vendre des puces pour vendre des armoires entières. Le GB200 NVL72 en est le premier exemple : un rack complet qui réunit 72 GPU Blackwell et 36 processeurs Grace (le CPU maison de NVIDIA), reliés par une interconnexion ultra-rapide appelée **NVLink**, livré pré-assemblé et refroidi au liquide.

Vera Rubin, présentée en mars 2026, pousse la logique plus loin. Le premier rack tourne déjà chez Microsoft Azure, pour une disponibilité partenaires au second semestre 2026. NVIDIA revendique 3,6 exaflops d'inférence par rack, soit cinq fois la performance de Blackwell, et surtout un coût par token divisé par dix.

Au-delà des performances, Rubin illustre deux dynamiques de fond. La première est la dépendance à la mémoire : la HBM4 de Rubin est co-conçue entre NVIDIA, TSMC et les fabricants de mémoire, **ce qui fait de la pénurie de HBM le véritable goulot d'étranglement de l'industrie en 2026.**

La seconde est la capacité d'absorption. NVIDIA a intégré dans Rubin la technologie d'inférence de **Groq**, l'un de ses challengers : l'illustration parfaite de sa façon de neutraliser une menace en l'absorbant plutôt qu'en la combattant.

DGX Cloud et la Verticalisation : NVIDIA contre ses Propres Clients

NVIDIA ne se contente plus de vendre du matériel. Avec **DGX Cloud**, elle loue désormais sa puissance de calcul directement aux entreprises, entrant en concurrence frontale avec les hyperscalers (ses propres clients) qui voient d'un très mauvais œil cette captation de la valeur.

Cette tension est l'un des moteurs souterrains de tout le secteur : c'est précisément pour échapper à cette dépendance, et reprendre la marge que NVIDIA leur prélève, que les hyperscalers développent leurs propres puces (TPU, Trainium, Maia, MTIA). On y revient en Partie IV.

Le Pari Omniverse et l'Embodied AI (GR00T)

La thèse de Huang est que les LLMs ne sont qu'une première étape. Le véritable objectif de NVIDIA est l'« IA physique » : robots humanoïdes, machines autonomes et systèmes industriels capables d'agir dans le monde réel.

Pour cela, NVIDIA développe le **projet GR00T**, un **modèle de fondation** pour robots. **Omniverse**, ensuite, est une plateforme qui recrée des environnements virtuels physiquement réalistes, sortes de jumeaux numériques du monde, dans lesquels les robots s'entraînent. L'intérêt est d'apprendre dans la simulation pour éviter d'avoir à collecter des données réelles, lentes et coûteuses à produire.

Partie IV : Les Challengers de NVIDIA

Personne ne peut aujourd'hui reproduire l'ensemble de la plateforme NVIDIA. Ses concurrents ne cherchent donc pas à construire un « meilleur NVIDIA », mais ciblent un segment où les GPU généralistes sont trop coûteux ou insuffisamment optimisés.

L'entraînement des grands modèles reste le terrain le plus difficile à attaquer, car il exige simultanément du calcul, de la mémoire, des interconnexions performantes et une pile logicielle mature. L'inférence est plus accessible, ce qui explique pourquoi la plupart des challengers concentrent leurs efforts sur ce marché.

AMD et le Boulet Logiciel

AMD est le seul à affronter NVIDIA avec la même arme : un GPU généraliste, capable d'entraîner comme de faire tourner les modèles, et qui se programme de façon comparable. Ses puces, performantes et souvent moins chères que celles de NVIDIA, ne souffrent pas d'un déficit de silicium mais d'un boulet logiciel : la pile **ROCm** peine encore à rivaliser avec CUDA.

Là où l'écosystème de NVIDIA fait tourner n'importe quel modèle dès le premier jour, exploiter pleinement une puce AMD réclame un travail d'optimisation que seuls les hyperscalers peuvent se permettre.

Ce sont d'ailleurs eux qui portent son adoption, autant pour réduire leur dépendance à NVIDIA que pour le prix.

Les Hyperscalers et la Verticalisation

La menace la plus sérieuse vient des plus gros clients de NVIDIA. Pour cesser de lui abandonner leur marge, Google, Amazon, Microsoft et Meta conçoivent leurs propres puces, non pour les vendre mais pour s'équiper à moindre coût, surtout en inférence.

Google a dix ans d'avance avec ses **TPU**, au point de franchir une étape révélatrice : il annonce séparer une puce dédiée à l'entraînement d'une autre dédiée à l'inférence.

Amazon poursuit une stratégie à double voie, ses propres **Trainium** et du NVIDIA en parallèle, déployée à grande échelle dans le **supercalculateur qu'il bâtit pour Anthropic**.

Microsoft, avec **Maia**, et Meta, avec **MTIA**, suivent la même logique. La force de ces acteurs ne tient pas à la puissance d'une puce isolée, mais à leur capacité à en connecter des dizaines de milliers à très faible latence.

Ils ne conçoivent d'ailleurs pas ces puces seuls : **Broadcom**, et dans une moindre mesure **Marvell**, codéveloppent la plupart d'entre elles, à commencer par les TPU de Google, ce qui en fait l'un des grands gagnants discrets de la vague IA.

Famille 1 : Remplacer le GPU par une autre architecture

C'est le combat le plus difficile, car il vise le cœur du marché de l'entraînement, et il ne s'agit plus de faire un meilleur GPU mais d'en proposer une alternative radicale. L'exemple le plus marquant est **Cerebras**, qui grave une puce unique de la taille d'un wafer entier au lieu d'assembler des milliers de GPU. Le but est de supprimer la communication entre processeurs, qui ralentit les clusters classiques. C'est l'alternative la plus crédible sur les très grands modèles, mais son écosystème logiciel reste très inférieur à celui de CUDA.

Une autre approche, portée par **Tenstorrent** mise sur l'ouverture. Ses puces reposent sur le **RISC-V**, un jeu d'instructions ouvert et libre de droits, c'est-à-dire le langage de base qu'une puce sait exécuter, mais que personne ne possède. Là où les standards habituels exigent une licence et des royalties, le RISC-V permet à n'importe qui de concevoir sa puce sans rien payer ni demander d'autorisation.

Tenstorrent pousse la logique jusqu'au bout : son logiciel est entièrement public et l'entreprise licencie sa propriété intellectuelle à qui veut fabriquer ses propres puces sans dépendre de NVIDIA. C'est la stratégie anti-verrouillage incarnée, prisée des acteurs soucieux de souveraineté. Sa limite reste celle de tous : le logiciel, car peu de modèles tournent encore correctement sur ses puces.

Famille 2 : Les Spécialistes de l'Inférence

C'est le segment le plus prometteur, parce que l'inférence se prête bien à la spécialisation. Le cas le plus instructif est **Groq**, dont la puce d'inférence à très basse latence a suffisamment inquiété NVIDIA pour que celui-ci absorbe sa technologie et une partie de ses ingénieurs.

L'idée technique la plus intéressante vient de **d-Matrix**, qui rapproche le calcul de la mémoire pour réduire le coût du transport des données, principal poste de dépense de l'inférence. **Etched** va plus loin encore avec une puce capable d'exécuter uniquement des **Transformers** : la spécialisation est maximale et les gains potentiels élevés si l'architecture de référence reste.

Famille 3 : Les Rares Challengers de l'Entraînement

C'est le terrain le plus exigeant, et donc le moins peuplé. **MatX** est le plus ambitieux, avec une architecture conçue dès le départ pour les très grands modèles des laboratoires de pointe, mais l'entreprise est jeune et son silicium n'a pas encore été éprouvé à grande échelle. Cerebras reste là aussi le concurrent le plus crédible ; les autres se concentrent à court terme sur l'inférence.

Famille 4 : l'Edge & Physical AI : un autre marché

Cette famille ne cherche pas à déloger NVIDIA des data centers, mais à faire tourner l'IA directement sur les appareils, au plus près de l'utilisateur. C'est ce qu'on appelle **l'edge**, par opposition au cloud centralisé. Pour les téléphones, les voitures, ou les robots, envoyer les données vers un datacenter distant serait trop lent, trop coûteux ou trop sensible. L'enjeu n'y est donc pas la puissance brute mais de réagir en quelques millisecondes sans dépendre du réseau, et de consommer très peu d'énergie.

C'est un marché distinct, avec ses propres acteurs. L'européenne **Axelera** est la plus notable, et elle relie cette bataille à la souveraineté, en concevant en Europe des puces d'inférence à basse consommation. Elle y affronte surtout **Qualcomm**, qui transpose au datacenter et à la voiture son savoir-faire d'efficacité énergétique hérité du smartphone.

Pour NVIDIA, ces acteurs ne sont pas une menace directe, puisqu'ils ne touchent pas aux grands centres de calcul. Cependant, chaque tâche qui s'exécute sur un appareil est une tâche qui ne tourne pas dans le cloud, et donc un GPU de moins à louer.

Une Plateforme, donc une Cible Morcelée

Tous ces acteurs butent souvent sur le même obstacle : le logiciel. Les survivants seront ceux qui le contournent de l'une des trois seules façons possibles : posséder la pile de bout en bout, l'ouvrir pour la faire écrire par d'autres, ou rétrécir si fort la tâche que le problème logiciel s'évapore.

Partie V : La Géopolitique du Silicium

La fabrication des puces IA de pointe repose sur un triptyque indissociable : la maîtrise industrielle du fondeur taïwanais TSMC sur les nœuds 3 nm et 2 nm, l'optique et la lithographie de l'écosystème européen, et les équipements et logiciels américains qui conditionnent toute la chaîne.

ASML et Zeiss : le Monopole des Machines de Gravure

La fabrication des puces d'IA les plus avancées dépend entièrement des machines EUV (Extreme Ultraviolet) d'ASML. Le néerlandais repose lui-même sur l'allemand Carl Zeiss SMT, qui produit les miroirs et les optiques ultra-précis nécessaires à la lithographie. Le développement de cette technologie a nécessité des décennies de recherche, créant un écosystème monopolistique quasiment impossible à reproduire.

Cette domination industrielle est devenue un enjeu géopolitique majeur. Jusqu'en 2024, la direction d'ASML s'est opposée aux restrictions américaines visant la Chine, dénonçant une logique de guerre économique. Son successeur a dû appliquer une ligne beaucoup plus dure sous la pression directe de Washington, se traduisant en 2026 par des livraisons annulées et des restrictions drastiques sur la maintenance des machines de génération précédente déjà installées sur le sol chinois.

Les Équipementiers Américains : le Verrou Réel

Qu'une puce soit gravée au Japon, à Taïwan, en Corée ou en Europe, toutes les lignes de pointe nécessitent les équipements américains (**Applied Materials, Lam Research, KLA**) pour les étapes critiques de dépôt, de gravure et de métrologie. En combinant ces monopoles à l'usage de brevets américains par les acteurs étrangers, Washington verrouille la totalité de la chaîne de valeur mondiale.

Ce pouvoir structurel sert de levier à Washington, notamment dans sa politique d'interdictions de vente de puces et de machines les plus avancées à la Chine, appuyée par des sanctions contre ceux qui tentent de l'esquiver.

Sur la seule dernière année, le régulateur américain du commerce a infligé des centaines de millions de dollars d'amendes pour contournement, dont une retentissante contre l'équipementier Applied Materials, accusé d'avoir fait transiter ses machines vers la Chine via une filiale coréenne.

TSMC : le Fondeur Irremplaçable et le Risque Taïwanais

Avant ses challengers, un mot sur le maillon central : TSMC. Le taïwanais ne conçoit aucune puce, il se contente de les fabriquer pour les autres (NVIDIA, Apple, AMD). Graver à 3 puis 2 nanomètres avec des rendements industriels suppose des décennies de savoir-faire que ni Intel ni Samsung ne parviennent à rattraper.

Il en résulte une dépendance d'un genre rare : la quasi-totalité des puces IA de pointe de la planète sort des usines d'une seule entreprise, sur une seule île. C'est le paradoxe taïwanais, le « silicon shield » : Taïwan est à la fois le point le plus précieux et le plus vulnérable de toute la chaîne mondiale, à portée immédiate de la Chine.

Sous la pression de Washington, TSMC relocalise une partie de sa production aux États-Unis. Son site d'Arizona, devenu avec 165 milliards de dollars le plus gros investissement étranger de l'histoire du pays, grave déjà du 4 nm pour Apple et NVIDIA, et vise le 2 nm en fin de décennie. Cependant, les puces gravées en Arizona repartent encore à Taïwan pour y être assemblées, car le packaging avancé n'y est pas encore implanté.

Déplacer la gravure ne suffit pas à déplacer la dépendance ; l'île reste, pour quelques années encore, le seul endroit où une puce de pointe peut être construite de bout en bout.

Intel Foundry Services : le Pari du Retour

Entre 2020 et 2026, Intel a tenté de redevenir un acteur central après des années de retard sur TSMC : il est devenu un « fondeur » ouvert à des clients externes, et a relocalisé une partie de la production aux États-Unis grâce aux subventions du **CHIPS Act**.

L'exécution initiale fut difficile : Intel investissait massivement pendant qu'AMD captait la croissance du marché des serveurs. Le tournant intervient en 2026 : Apple et Intel concluent un accord préliminaire, rapporté en mai, pour qu'Intel fabrique une partie des puces d'Apple sur son **nœud 18A**.

Ce signal valide le redressement, d'autant qu'Intel a aussi attiré un investissement de 5 milliards de dollars de NVIDIA et des engagements d'Amazon, Microsoft et Google sur le 18A.

Samsung : le Géant Ambivalent de la Mémoire et de la Fonderie

Samsung occupe une position unique : à la fois fondeur concurrent de TSMC et deuxième producteur mondial de HBM.

Sur la fonderie, il a perdu du terrain : ses nœuds 3 nm souffrent de rendements insuffisants qui ont poussé des clients comme Qualcomm à retourner chez TSMC, le reléguant au second rang sur les nœuds les plus avancés.

Sur la mémoire, le coup est plus symbolique encore : SK Hynix a été sélectionné en priorité par NVIDIA pour la HBM4 de Vera Rubin, reléguant Samsung à un rôle de fournisseur complémentaire.

L'Architecture Législative Américaine et l'Endiguement de la Chine

Plutôt qu'un blocus total, l'administration américaine a choisi une voie intermédiaire : autoriser au compte-gouttes la vente à la Chine de certaines puces avancées, sous conditions. En janvier 2026, le BIS, le bureau du commerce qui délivre ces autorisations, a formalisé cette **approche au cas par cas**.

Chaque puce doit être testée avant export, les volumes sont plafonnés, et surtout chaque vente est frappée d'un tarif douanier de 25 % reversé au Trésor américain. Les puces les plus avancées, de génération Blackwell, restent quant à elles interdites.

Au même moment, le Congrès pousse dans le sens inverse. Une proposition de loi, **l'AI Overwatch Act**, sortie de commission en janvier 2026, lui donnerait le pouvoir de bloquer au cas par cas les ventes de puces avancées à la Chine que l'exécutif vient d'autoriser.

Elle est encore loin d'être adoptée, mais elle révèle la contradiction américaine : pendant que la Maison-Blanche fait payer l'accès, une partie du Congrès cherche à le fermer.

La Riposte Chinoise : Guerre des Minerais et Stratégie de la Quantité

Au moment même où Washington rouvre partiellement la porte, Pékin pousse ses entreprises à s'en détourner. En mai 2026, la Chine a publié pour la première fois une liste de puces IA « sûres et fiables » que ses administrations doivent privilégier : neuf modèles chinois y figurent, ceux de Huawei ou Cambricon par exemple, mais aucun de NVIDIA. L'objectif est une autonomie totale qui rendrait les sanctions sans effet.

Cet objectif s'appuie sur un écosystème très vertical. Huawei en est le centre, avec son accélérateur Ascend et sa **pile logicielle maison CANN**, pensée comme l'alternative à CUDA. La part de NVIDIA sur le marché chinois s'est effondrée après que Pékin a orienté ses administrations et ses géants du cloud vers le silicium national. Les analystes voient Huawei capter à lui seul autour de 60 % du marché chinois d'ici fin 2026.

La même quête d'indépendance descend jusqu'au jeu d'instructions, le langage de base qu'une puce exécute : plutôt que les standards occidentaux soumis à licence, la Chine investit le RISC-V, ouvert et libre de droits, via des projets comme **XiangShan**, soutenu par Alibaba et l'Académie des sciences.

Cette stratégie reste bridée par la fabrication. Privé des machines de gravure les plus avancées, le fondeur national **SMIC** est bloqué au nœud 7 nm. Faute de pouvoir miniaturiser, Huawei mise sur l'empilement de la logique et le raccourcissement des chemins du signal voir Partie II). La doctrine assumée est de compenser le déficit de puissance brute par la quantité et la saturation du marché intérieur. La HBM chinoise garde une génération de retard.

Mais Pékin conserve un levier redoutable : son quasi-monopole sur les terres rares et les minerais critiques de l'électronique (gallium, germanium), qu'il peut couper là où l'Occident est le plus vulnérable.

Partie VI : La Course aux Modèles de Fondation et l'Effondrement des Marges

La couche des modèles de fondation s'est rapidement structurée autour d'une poignée d'acteurs. En 2026, trois laboratoires dominent la course occidentale : OpenAI, Anthropic et Google DeepMind. Anthropic incarne l'accélération vertigineuse du secteur, avec un run-rate de revenus passé d'environ 1 milliard de dollars début 2025 à 47 milliards à la mi-2026.

Pour comprendre pourquoi le secteur a englouti de telles sommes dans le calcul brut, il faut remonter à une idée formulée par le chercheur Rich Sutton, la « **bitter lesson** ». Son constat, tiré de plusieurs décennies d'histoire de l'IA : à long terme, les méthodes qui se contentent d'exploiter toujours plus de puissance de calcul finissent presque toujours par battre les approches savantes pensées par des humains.

C'est le fondement théorique de la **scaling law** et de la course tout entière : si ajouter du compute gagne presque à coup sûr, alors dépenser des dizaines de milliards en GPU et en gigawatts devient parfaitement rationnel. C'est précisément cette doctrine que DeepSeek viendra fissurer.

Du Training à l'Inférence : Le Grand Basculement

Jusqu'en 2024, le compute IA était massivement dominé par l'entraînement : quelques dizaines de labs et d'entreprises dépensaient des milliards pour entraîner des modèles fondation sur des clusters de milliers de GPU, lors de campagnes longues et peu fréquentes.

Depuis 2025, le centre de gravité a basculé vers l'inférence. Des milliards de requêtes quotidiennes consomment désormais plus de compute que l'entraînement lui-même. La métrique centrale n'est plus le FLOP par seconde mais le coût par token. Ce sont désormais les besoins en inférence de masse, et non plus les campagnes d'entraînement, qui dictent les investissements énergétiques des hyperscalers et dimensionnent leurs data centers.

Le Choc DeepSeek : Quand l'Efficience Algorithmique Défie la Force Brute

En janvier 2025, le laboratoire chinois DeepSeek a provoqué une première onde de choc en prouvant qu'une architecture mieux pensée pouvait compenser un déficit en silicium. Sa technique clé, le « **mixture-of-experts (MoE)** » (**MOE**), n'active à chaque requête qu'une petite fraction des paramètres du modèle plutôt que sa totalité : moins de poids à relire en mémoire, donc un coût par token structurellement plus bas. C'est une réponse directe au mur mémoire décrit plus haut, pas une simple politique de prix agressive.

DeepSeek a prouvé qu'un modèle **open-weights** pouvait rivaliser avec les leaders américains pour une fraction minime du coût. Avec ses nouveaux modèles sortis en mai 2026, le laboratoire chinois a transformé l'exploit technique de 2025 en une guerre des prix totale. Son modèle de raisonnement affiche un prix par token près de trente fois inférieur à celui de Claude ou de GPT, pour des performances proches sur les tests de programmation.

L'Effondrement des Marges et la Crise de Rentabilité des Laboratoires Américains

En vendant le million de tokens à un prix dérisoire, DeepSeek détruit les marges du secteur. Pour OpenAI ou Anthropic, s'aligner sur ces tarifs est devenu financièrement intenable : leurs modèles coûtent beaucoup plus cher à tourner qu'ils ne rapportent.

Les modèles américains classiques doivent charger l'intégralité du contexte dans la mémoire ultra-coûteuse-coûteuse des GPU. À chaque requête traitée, OpenAI et Anthropic subventionnent à perte le coût de l'électricité et du matériel nécessaires pour générer la réponse. Sans une refonte logicielle totale visant à imiter l'efficience algorithmique chinoise, l'augmentation du volume d'utilisateurs creuse mécaniquement leurs pertes financières et rend leur dépendance au compute brut difficilement soutenable.

Face à cette pression simultanée sur les marges et sur la capacité de calcul, les laboratoires américains n'ont qu'une issue : sécuriser l'infrastructure à tout prix. C'est ce qui explique la verticalisation des géants et la course aux gigawatts qui structure désormais le secteur.

Partie VII : Infrastructure et Crise Énergétique

La Verticalisation Inédite des Géants

La course au calcul de frontière a brisé les échelles classiques de l'investissement technologique. Lancé début 2025, le **projet Stargate** est une coentreprise réunissant OpenAI, le japonais SoftBank et Oracle pour bâtir aux États-Unis un réseau de data centers géants dédiés à l'entraînement des modèles. Avec des engagements dépassant 100 milliards de dollars, il vise une concentration de puissance de calcul sans précédent.

Face à ce gigantisme, Amazon réplique par une intégration verticale totale avec son **Project Rainier**, le supercalculateur qu'il déploie dans l'Indiana pour Anthropic. L'enjeu pour Amazon est d'y faire tourner massivement ses propres puces maison, les Trainium, plutôt que du NVIDIA.

Les Neoclouds et le Risque de la Dette GPU

Pour répondre à la pénurie initiale de GPU, de nouveaux acteurs spécialisés comme **CoreWeave, Lambda Labs** ou **Crusoe** ont émergé. Leur modèle reposait sur une logique simple : emprunter massivement pour acheter des serveurs NVIDIA, puis louer cette puissance de calcul pour l'IA (par opposition aux hyperscalers généralistes). CoreWeave a poussé ce montage financier à l'extrême en utilisant directement ses puces H100 comme collatéral pour lever des milliards de dollars de dette.

En 2026, ce modèle montre ses limites systémiques. L'arrivée sur le marché des puces de nouvelle génération (Blackwell) a brutalement accéléré l'obsolescence des parcs de H100, tandis que l'efficience algorithmique à la DeepSeek a fait s'effondrer les prix de location de la puissance de calcul. Les Neoclouds qui n'ont pas sécurisé de contrats de très longue durée avec des hyperscalers se retrouvent pris au piège.

xAI, Colossus et la Philosophie de la Force Brute

xAI, le laboratoire d'IA d'Elon Musk, incarne une philosophie radicalement opposée à celle de DeepSeek : là où les ingénieurs chinois optimisent, Musk construit en force brute. Colossus, son supercalculateur installé à Memphis est passé de 100 000 à plus de 200 000 GPU NVIDIA en quelques mois.

Ce pari a produit un excès de capacité que Grok, le modèle de xAI, n'a jamais suffi à absorber. C'est dans ce contexte que s'est noué l'un des deals les plus improbables de l'histoire de l'IA. En mai 2026, Anthropic a signé avec xAI un contrat de 1,25 milliard de dollars par mois jusqu'en 2029, soit environ 40 milliards au total, pour louer l'intégralité de Colossus 1. La raison : Anthropic a connu une croissance de revenus et d'usage huit fois supérieure à ses prévisions en début d'année, créant une crise de compute importante.

En février 2026, Musk qualifiait publiquement Anthropic de « misanthropic and evil ». Quand l'accès au compute est en jeu, les rivalités idéologiques cèdent la place aux contrats.

Terafab : La Verticalisation Poussée à l'Extrême

Le projet le plus radical de 2026 a été dévoilé par Elon Musk en mars : Terafab, une coentreprise entre Tesla et SpacexAI. L'idée est une méga-fonderie qui réunirait sous un même toit toutes les étapes de la production, de la conception à la lithographie, la mémoire, le packaging et les tests, pour un coût annoncé autour de 25 milliards de dollars.

L'ambition affichée est de produire un térawatt de puissance de calcul par an, un volume sans commune mesure avec l'industrie actuelle, dont une large part est pensée pour fonctionner dans l'espace, sur des satellites alimentés en solaire.

En sous-main, c'est aussi un levier de négociation avec TSMC : en menaçant de s'en affranchir, le groupe de Musk cherche à améliorer sa position dans la file d'attente du fondeur. Le projet relève pour l'instant de l'annonce démesurée plus que du fait acquis, car bâtir une fonderie de pointe est l'un des exploits industriels les plus difficiles qui soient.

L'Énergie, nouvelle limite du calcul, et la Renaissance Nucléaire

L'explosion des besoins énergétiques de l'IA a provoqué un retournement stratégique majeur : la dépendance absolue au réseau électrique et la renaissance du nucléaire civil. Les data centers consomment désormais des volumes d'électricité si colossaux que les énergies renouvelables intermittentes (solaire, éolien) sont incapables de garantir la stabilité requise (24h/24).

Pour sécuriser leur croissance, les Big Tech financent directement la filière nucléaire par des contrats d'achat d'électricité à long terme. Microsoft a sécurisé plus de 800 MW par ses contrats. Google et AWS ont tous deux signé des accords historiques pour déployer des flottes de **SMR** à l'horizon 2030-2035.

La France et l'Atout Nucléaire : Le Compute au Prisme des Gigawatts

La compétition internationale s'est déplacée du terrain du silicium vers celui des réseaux électriques. L'accélération de l'inférence de masse et des modèles de raisonnement consomme des volumes d'électricité si gigantesques que l'accès immédiat à des mégawatts stables est devenu le principal avantage asymétrique.

Sur ce point, la carte mondiale se fracture. Alors que les hyperscalers américains rachètent des centrales nucléaires outre-Atlantique et que le Golfe déploie des parcs solaires couplés à des turbines à gaz, l'Europe continentale souffre de coûts énergétiques élevés et de réseaux saturés.

Dans cette guerre du gigawatt, la France a un avantage clé : son infrastructure nucléaire civile, stable et décarbonée, s'impose désormais comme sa principale arme d'attraction pour les centres de calcul européens, transformant sa politique énergétique en pilier de sa souveraineté numérique.

Partie VIII : L'Éveil Européen

L'Héritage des Échecs : D'Andromède à l'Illusion Gaia-X

L'histoire du cloud souverain européen est celle d'un contresens stratégique. Dès 2009, le projet français Andromède (scindé en Cloudwatt et Numergy) a échoué parce que l'État l'a pensé comme une infrastructure d'hébergement classique et passive, là où AWS construisait une plateforme misant sur l'automatisation par API et pensée pour les développeurs. Dix ans plus tard, la tentative collective Gaia-X s'est enlisée dans la bureaucratie : en intégrant les hyperscalers américains dans sa gouvernance, le projet a perdu sa substance, provoquant le départ de Scaleway en 2021.

Cette fracture historique sépare toujours l'Europe en deux camps en 2026 : d'un côté, les souverainistes durs, de l'autre, les pragmatiques des alliances hybrides.

Les Souverainistes Durs contre le « Cloud de Confiance »

Le camp de la souveraineté technologique intégrale est mené par **OVHcloud** et **Scaleway**. OVHcloud reste le champion industriel du bare metal, mais la transition vers l'IA exige des investissements massifs dans les puces NVIDIA. À l'inverse, Scaleway a fait le pari de l'hyper-spécialisation en achetant très tôt de grands clusters de GPU pour devenir le refuge des startups IA européennes. Dans le secteur public, c'est **NumSpot** qui pousse pour équiper les administrations et la santé, en visant la qualification **SecNumCloud**, le label de souveraineté délivré par l'État.

Ce modèle frontal s'oppose à la doctrine pragmatique du « **Cloud de Confiance** », incarnée par **Bleu** (Orange et Capgemini avec Microsoft) et **S3NS** (Thales avec Google). L'idée de départ : contrôler les technologies américaines en les exploitant sous licence, derrière une protection juridique européenne. Ces acteurs visent eux aussi SecNumCloud, et S3NS l'a déjà obtenue, preuve que la frontière entre les deux camps est plus poreuse qu'il n'y paraît.

En 2026, le bilan de ces alliances reste ouvert. Pour ses défenseurs, c'est une première étape pragmatique vers une souveraineté progressive ; pour ses critiques, une dépendance juridiquement atténuée mais sans rupture technologique réelle.

Le Cas Mistral AI : Souveraineté et Pragmatisme

L'évolution de Mistral AI illustre la position complexe du continent. La startup parisienne a mené un lobbying intense contre l'excès de régulation de **l'AI Act** au nom de la souveraineté, tout en s'appuyant sur un partenariat avec Microsoft et son infrastructure Azure : une dépendance assumée, le temps de grandir.

Depuis, elle a diversifié ses appuis. Son premier actionnaire est désormais l'équipementier ASML, avec environ 11 % du capital, et elle bâtit sa propre capacité de calcul en France, avec un cluster hébergé par Scaleway près de Paris.

Plutôt que de rivaliser par la force de calcul, l'écosystème français (autour de Mistral, de Station F et du laboratoire **Kyutai**) parie sur l'efficience : des modèles ouverts et ultra-optimisés, dont l'architecture légère vise à égaler les modèles fermés américains sans exiger les infrastructures démesurées de la Silicon Valley.

InvestAI, AION et l'Atout Nucléaire : L'Europe en 2026

Le véritable basculement opérationnel de 2026 est énergétique. Face aux difficultés logistiques du Royaume-Uni et de l'Allemagne, où les projets de mégacentres de données IA sont suspendus à cause du coût de l'électricité et de la saturation des réseaux, la France utilise son parc nucléaire civil comme un avantage compétitif.

C'est dans ce contexte que s'inscrit le programme européen **InvestAI Facility** (20 milliards d'euros) et l'émergence du consortium français **AION**. Porté par Scaleway et de grands industriels, AION ambitionne de construire en France une infrastructure de calcul de classe mondiale connectée directement aux centrales nucléaires d'EDF. L'ambition de la France est claire : devenir le hub énergétique et physique du compute européen.

Cette ambition vient de trouver sa traduction la plus spectaculaire. En mai 2026, au sommet Choose France, **SoftBank** a annoncé un investissement pouvant atteindre 75 milliards d'euros pour bâtir jusqu'à cinq gigawatts de data centers dédiés à l'IA, présenté comme le plus gros engagement jamais consenti dans une infrastructure numérique en Europe.

Partie IX : Le Choc Macro-Stratégique et l'Étau Juridique Global

L’Affaire du CLOUD Act

Ce débat sur la souveraineté découle d'un bras de fer historique. En 2013, la justice américaine exige l’accès à des courriels stockés par Microsoft sur des serveurs en Irlande. La firme refuse, protégeant des données hébergées à l’étranger. Pour trancher définitivement, le Congrès américain adopte en 2018 le CLOUD Act. Ce texte oblige toute entreprise technologique sous juridiction américaine (comme les hyperscalers) à livrer les données de ses utilisateurs aux autorités, quel que soit leur lieu de stockage physique dans le monde.

À l'ère de l'intelligence artificielle, ce mécanisme transforme l'infrastructure cloud en un outil d'ingérence directe, rendant caduque la simple idée d'un hébergement localisé si le fournisseur de services reste américain.

Le Choc du « Liberation Day » et la Rupture Transatlantique

Le basculement politique majeur est intervenu au printemps 2025. Lors du « Liberation Day », l'administration Trump a imposé une vague de tarifs douaniers massifs et indifférenciés sur les importations en provenance de Chine (34 %) et de l’Union européenne (20 %).

Au-delà du choc économique immédiat, ces mesures ont provoqué une rupture psychologique en Europe. Elles ont démontré le danger de dépendre d'infrastructures cloud contrôlées par des géants soumis aux décrets d'un gouvernement américain imprévisible. Cet épisode a légitimé, aux yeux des décideurs européens, la mise en place de politiques de préférence technologique.

Le Tempo Réglementaire : Ambitions et Réalités pour l'Écosystème Européen

Face à cet étau, la réponse de l'Union européenne s'est matérialisée par le déploiement de l'AI Act. Cependant, le résultat de ce tempo réglementaire est profondément asymétrique. Les géants américains disposent des ressources pour absorber le coût de la mise en conformité européenne, qu'ils considèrent simplement comme une taxe d'accès à un marché de 450 millions de consommateurs.

À l'inverse, ce fardeau réglementaire pèse de manière disproportionnée sur les startups et les PME européennes. Privées de capitaux massifs et n'ayant pas accès à un compute bon marché, elles doivent consacrer une part importante de leurs ressources à la mise en conformité plutôt qu'à la R&D. Devoir s'adapter en permanence à des règles en évolution ralentit leur croissance et réduit leur compétitivité face aux leaders mondiaux.

Conclusion : L'Infrastructure comme Destin

L'histoire du cloud, du compute et de la souveraineté entre 2006 et 2026 valide une thèse simple : la valeur durable ne réside pas dans le logiciel, mais dans les couches physiques qui le font tourner. Une application se copie en un week-end. Personne ne duplique en un week-end un rack de 72 GPU, une pile de mémoire HBM, une usine de gravure à 2 nm ou un réacteur nucléaire. Cette hyper-matérialisation du numérique a rebattu les cartes à toutes les échelles.

À l'échelle de l'entreprise, les hyperscalers avaient fait de la complexité tarifaire et des frais de sortie un instrument de capture. Mais l'IA a déplacé le combat : le pouvoir ne tient plus aux astuces commerciales, il tient à l'accès aux puces et à l'électricité, désormais si rares que les fournisseurs eux-mêmes en manquent.

À l'échelle de l'industrie, le pouvoir s'est logé dans les maillons les plus difficiles à répliquer. NVIDIA l'a prouvé : sa domination ne tient pas à une puce, mais à vingt ans d'écosystème logiciel et à la maîtrise du système entier. Ce quasi-monopole ne durera pas nécessairement. Le centre de gravité se déplace de l'entraînement vers l'inférence, là où l'avance logicielle compte moins et où des architectures spécialisées attaquent le coût par token plutôt que la puissance brute. Le silicium reste le terrain décisif, mais la question n'est plus seulement qui sait fabriquer la meilleure puce ; c'est qui saura faire tourner les modèles au prix le plus bas.

À l'échelle des États, l'accès au compute est désormais un enjeu de souveraineté au même titre que l'énergie ou le territoire, et il s'y confond : entraîner et faire tourner les modèles consomme tant d'électricité que la carte du pouvoir se redessine en gigawatts.

En juin 2026, trois blocs se disputent cette souveraineté. Les États-Unis dominent par la profondeur de leur capital, la densité de leur écosystème et le contrôle des nœuds critiques de la chaîne. La Chine résiste par la quantité, son efficience algorithmique et son levier sur les minerais critiques, en bâtissant une pile entière hors de portée de Washington. L'Europe cherche sa place. Elle tient pourtant un maillon irremplaçable de toute la chaîne, ASML et le monopole de la gravure, mais reste dépendante pour le reste, partagée entre des alliances hybrides et l'ambition de quelques acteurs comme Mistral, Scaleway ou le projet AION.

Cette réouverture vaut à tous les étages. Le verrou logiciel de NVIDIA pousse Tenstorrent à miser sur des architectures ouvertes ; sa puissance brute, taillée pour l'entraînement, laisse Cerebras l'attaquer sur l'inférence ; le monopole de la gravure occidentale pousse la Chine vers le RISC-V et ses propres usines. Chaque point de domination fait naître ceux qui veulent le briser, et c'est souvent la contrainte la plus dure qui accouche de la rupture, comme l'a montré DeepSeek.

L'avenir ne se jouera ni dans une directive bruxelloise ni dans un modèle de plus. Il se jouera dans la capacité à mobiliser des gigawatts, à maîtriser la gravure à 2 nm et à assumer une idée dérangeante : dans l'économie de l'IA, l'infrastructure n'est pas un moyen, elle est la fin.