FONDATIONS INFORMATIQUES (2/11)

Les cartes graphiques : gaming et IA

Un composant né pour afficher des jeux vidéo est devenu, presque par accident géométrique, le moteur de l'intelligence artificielle mondiale. La même puce se scinde aujourd'hui en deux destins, l'un pour vos loisirs, l'autre pour les modèles qui transforment l'économie.

Septembre 2026
Classification : Non sensible
Lecture : 18 min

Fondations informatiques : L'ordinateur et ses composants · **Cartes graphiques** ← · Systèmes d'exploitation & virtualisation · Serveurs & baies · NAS & SAN · Équipements réseau · Protocoles Internet · Cybersécurité, les bases · Bases de données · Algorithmique · Projets informatiques

L'ESSENTIEL

Un processeur classique excelle à faire une chose compliquée à la fois, très vite, en jonglant entre des tâches différentes. Une carte graphique fait l'inverse : elle fait la même opération simple, des milliers de fois, en même temps. Un processeur de serveur compte aujourd'hui quelques dizaines à quelques centaines de cœurs puissants et polyvalents ; une carte graphique en compte plusieurs milliers, chacun minuscule, chacun incapable de grand-chose seul, mais tous capables d'exécuter en parallèle exactement le même calcul.

Ce détail d'architecture, pensé dans les années 1990 pour calculer la couleur de millions de pixels à l'écran trente fois par seconde, s'est révélé être exactement ce dont l'intelligence artificielle avait besoin quinze ans plus tard : l'apprentissage d'un réseau de neurones se résume, mathématiquement, à des multiplications de matrices, la même famille d'opération répétée à une échelle massive que le rendu d'un jeu vidéo. Personne n'avait conçu la carte graphique pour l'IA ; l'IA s'est simplement révélée être, structurellement, le même problème que l'affichage d'un jeu.

En 2026, cette parenté originelle a produit une fourche assumée à l'intérieur d'une même famille de puces. L'architecture Blackwell de Nvidia, lancée fin 2024, se décline en une version grand public, gravée chez TSMC, associée à de la mémoire GDDR7 et à un bus PCIe de cinquième génération, et une version datacenter, gravée sur un procédé différent chez le même fondeur, associée à de la mémoire empilée HBM3E bien plus coûteuse et à un bus PCIe de sixième génération. La carte qui affiche vos jeux et celle qui entraîne un modèle de langage descendent de la même architecture, mais ne partagent plus grand-chose une fois sur l'étagère.

La domination qui en résulte est écrasante et documentée : Nvidia contrôle entre 75 et 87 % du marché mondial des accélérateurs d'IA selon les sources et les trimestres de 2026, et jusqu'à 90 à 94 % du marché des cartes graphiques grand public. Son chiffre d'affaires datacenter a atteint 193,7 milliards de dollars sur l'exercice 2026. Et pourtant, son propre fondateur a déclaré lors de sa conférence annuelle disposer d'un trillion de dollars de commandes fermes jusqu'en 2027, avant d'ajouter sans détour que l'entreprise allait être en rupture d'approvisionnement.

COMPRENDRE : À QUOI ÇA SERT, COMMENT ÇA MARCHE, OÙ ÇA S'IMBRIQUE

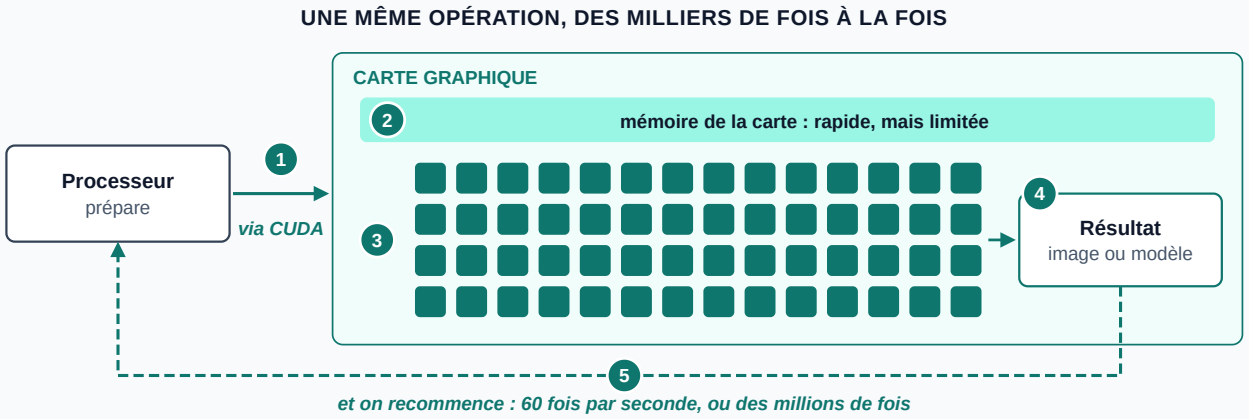
Pour le lecteur non spécialiste, environ quatre minutes. Initiés : passez directement aux chiffres qui suivent.

À QUOI ÇA SERT

Une carte graphique fait la même opération simple sur des milliers de données à la fois : colorer les pixels d'une image, ou multiplier les tableaux de nombres d'un modèle d'IA. Là où le processeur excelle dans les tâches variées et enchaînées, elle excelle dans le travail massivement répétitif, ce qui en a fait le moteur du jeu vidéo, puis de l'intelligence artificielle.

Image mentale : *le processeur est un chef étoilé qui enchaîne des plats différents ; la carte graphique, une brigade de milliers de commis qui épluchent chacun une pomme de terre au même instant. Pour un banquet de mille couverts identiques, la brigade gagne ; pour un menu à la carte, le chef.*

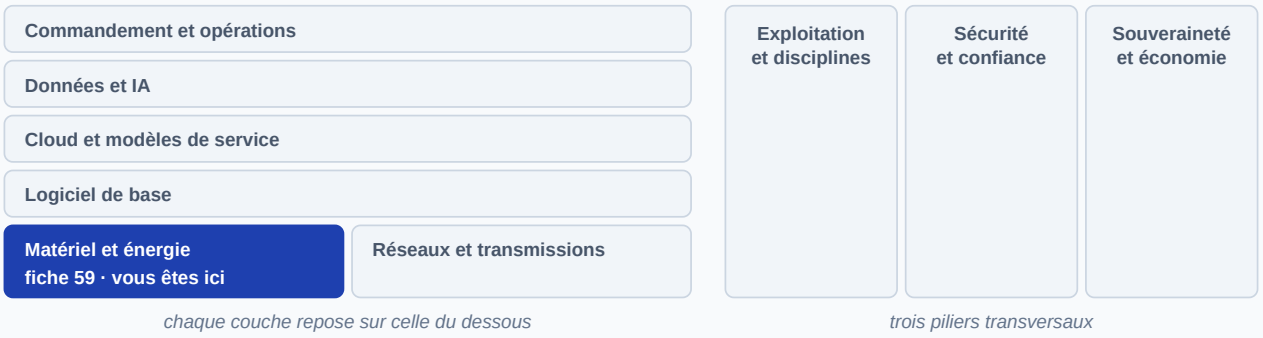
COMMENT ÇA MARCHE



1. Le processeur prépare le travail et l'envoie à la carte ; le programme est écrit pour une plateforme logicielle, le plus souvent CUDA, celle de NVIDIA, d'où une dépendance logicielle autant que matérielle.
2. Les données, textures d'une image ou paramètres d'un modèle, sont copiées dans la mémoire propre de la carte, très rapide mais limitée : sa taille borne ce que la carte peut traiter d'un coup.
3. Des milliers de cœurs simples appliquent la même opération à des milliers de données en même temps, un pixel ou un nombre chacun.
4. Les résultats sont assemblés : une image s'affiche, ou le modèle ajuste ses paramètres d'un petit pas.
5. La boucle recommence, soixante fois par seconde pour un jeu, des millions de fois pour entraîner un modèle ; pour les plus gros, des milliers de cartes travaillent ensemble, et le réseau qui les relie devient aussi décisif que les cartes elles-mêmes.

OÙ ÇA S'IMBRIQUE DANS L'ENSEMBLE

La carte graphique est le moteur de la couche « données et IA » : les modèles de langage et les agents de la collection n'existent qu'à travers elle, et sa rareté explique à la fois la densité électrique des baies d'IA et l'économie des clouds spécialisés.



Repose sur : 58 · L'ordinateur et ses composants ; 61 · Serveurs et baies ; 27 · Énergie, hardware & durabilité

Sert à : 02 · LLM : Grands modèles de langage ; 01 · IA Agentique ; 39 · Confidential Computing pour l'IA

À lire avec : 25 · Économie du cloud (opérateur) ; 67 · Algorithmique et développement logiciel

LES CHIFFRES QUI RECADRENT LE DÉBAT (SEPTEMBRE 2026)

75 à 87 %

Part de marché de Nvidia sur les accélérateurs d'IA en datacenter en 2026 selon les sources, en repli depuis un pic à 87-94 % en 2024, mais toujours archi-dominante.

Silicon Analysts, IDC, 2026

4 millions

Développeurs utilisant CUDA dans le monde, la couche logicielle propriétaire de Nvidia jugée « excellente » par les principaux frameworks d'IA, contre « correcte » pour les alternatives ouvertes.

Analyse sectorielle, 2026

193,7 Md$

Chiffre d'affaires datacenter de Nvidia sur l'exercice fiscal 2026, avec des marges brutes de 79 à 88 %, contre 64 à 68 % pour AMD.

NVIDIA 10-K, FY2026

« Nous allons être en pénurie »

La déclaration du fondateur de Nvidia à sa conférence annuelle 2026, après avoir annoncé un trillion de dollars de commandes fermes jusqu'en 2027.

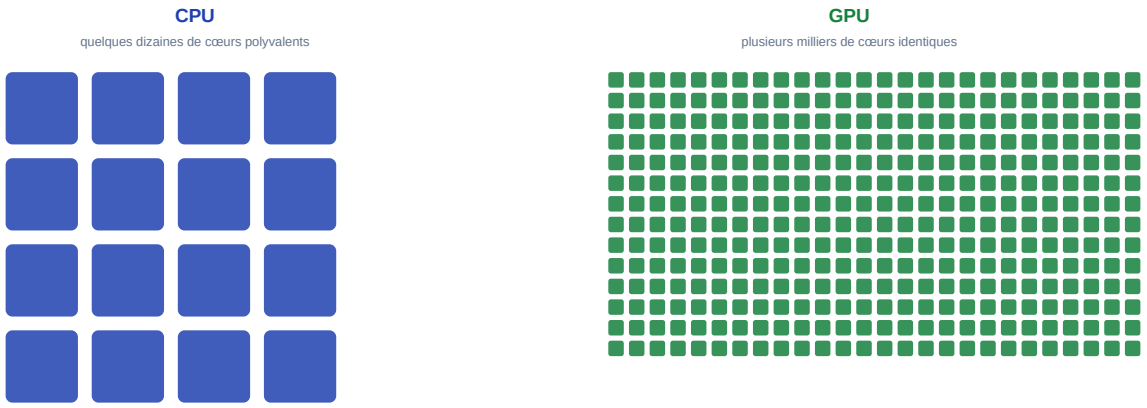
GTC 2026

Lecture C-level : aucun autre composant de cette collection ne concentre autant de pouvoir de marché sur un seul acteur, à la fois par le matériel et par le logiciel qui l'exploite. Un discours qui évoque le GPU sans nommer cette double dépendance passe à côté du sujet réel.

CPU CONTRE GPU : LE SCHÉMA C-LEVEL

La différence ne se voit nulle part mieux que dans le nombre et la taille des cœurs de calcul. Ce n'est pas une nuance technique, c'est toute la raison d'être de la carte graphique.

PEU DE CŒURS PUISSANTS, OU DES MILLIERS DE CŒURS SIMPLES



optimisé pour des tâches diverses et séquentielles

optimisé pour une seule opération répétée en parallèle, à grande échelle

LE RENDU D'UN PIXEL ET L'ENTRAÎNEMENT D'UN RÉSEAU DE NEURONES SONT, MATHÉMATIQUEMENT, LE MÊME PROBLÈME

Un CPU gagne en flexibilité ce qu'il perd en débit ; un GPU gagne en débit ce qu'il perd en flexibilité. L'IA a hérité du second parce que la multiplication de matrices, cœur du calcul neuronal, est exactement l'opération que des milliers de cœurs simples savent répéter le mieux.

LA FOURCHE BLACKWELL : UNE ARCHITECTURE, DEUX DESTINS

Depuis fin 2024, la même génération de puce se scinde délibérément en deux familles de produits, jamais interchangeables.

1

`Le grand public : GDDR7, PCIe 5.0`

Les cartes grand public de la gamme GeForce RTX embarquent de la mémoire GDDR7, moins coûteuse et suffisante pour l'affichage, avec en 2026 l'apparition du rendu neuronal, une technique combinant modèles d'IA et rendu traditionnel pour améliorer la performance et la qualité d'image des jeux.

Le gaming n'a jamais quitté l'IA : il l'utilise désormais pour améliorer ses propres images, la boucle s'est refermée.

2

`Le datacenter : HBM3E, PCIe 6.0`

Les puces destinées à l'entraînement et à l'inférence de modèles massifs embarquent de la mémoire empilée à très haute bande passante, HBM3E, bien plus onéreuse au gigaoctet mais indispensable pour nourrir des modèles dont les paramètres se comptent en centaines de milliards.

C'est cette variante, jamais la version grand public, qui équipe les fermes de calcul consacrées à l'IA.

3

`Le verrou logiciel, plus solide que le verrou matériel`

CUDA, la couche logicielle propriétaire de Nvidia, compte quatre millions de développeurs et reçoit la note maximale de compatibilité de la part des principaux cadriciels d'IA, quand les alternatives ouvertes reçoivent des notes seulement correctes à bonnes.

Un client qui change de fournisseur de GPU ne change pas seulement de matériel, il réécrit une partie de son code et accepte un risque de performance, un verrou logiciel qui survit à n'importe quelle amélioration purement matérielle de la concurrence.

SANS COMPLAISANCE : LA DOMINATION A DES FISSURES, PAS ENCORE DES BRÈCHES

POINT DE VIGILANCE	CE QU'IL FAUT SAVOIR
AMD progresse réellement, sans renverser le rapport de force	Le dernier accélérateur d'AMD atteint 80 à 90 % de la performance de son équivalent Nvidia à un coût inférieur, avec parfois de meilleurs résultats en inférence pure ; sa part de marché reste néanmoins limitée à 5 à 8 % du chiffre d'affaires mondial.
Le silicium propriétaire des hyperscalers grignote la marge, pas le cœur du marché	Les puces conçues en interne par les plus grands opérateurs cloud représentent déjà 12 à 18 % du calcul d'IA déployé, mais restent des actifs internes non vendus sur le marché ouvert, sans concurrencer directement Nvidia sur le segment souverain ou d'entreprise où CUDA reste la norme.
La pénurie n'est pas un problème de production, c'est un problème de demande	Le fondateur de Nvidia n'annonce pas une rupture de capacité industrielle, il annonce que la demande, portée par un trillion de dollars de commandes fermes, dépasse ce que l'industrie entière peut produire ; une pénurie de succès, pas d'échec.

SOUVERAINETÉ : LA DOUBLE DÉPENDANCE, MATÉRIELLE ET LOGICIELLE

DIMENSION	ENJEU 2026
La dépendance matérielle rejoint celle déjà posée pour le CPU	– Comme pour le processeur, les puces graphiques les plus avancées sortent des mêmes usines taïwanaises ; changer de fournisseur de GPU ne change rien à cette concentration de la fabrication physique.
La dépendance logicielle est propre au GPU, et plus difficile à déplacer	– CUDA n'est fabriqué nulle part, il s'apprend et s'écrit ; cette dépendance ne se règle ni par un accord commercial ni par une usine construite ailleurs, seulement par des années d'investissement dans un écosystème logiciel alternatif.
Les alternatives existent, mais restent un marché de niche assumé	± AMD et les puces des hyperscalers prouvent qu'une alternative technique fonctionne ; aucune n'a encore atteint l'échelle qui permettrait à un client souverain de bâtir une infrastructure critique sans un plan de repli vers Nvidia.

La thèse souveraine de cette fiche. Le GPU cumule les deux dépendances que le reste de cette collection documente séparément ailleurs, celle du silicium, partagée avec le CPU et concentrée chez les mêmes fondeurs, et celle du logiciel, propre à l'IA et bien plus difficile à déplacer parce qu'elle vit dans les compétences de quatre millions de développeurs plutôt que dans une chaîne d'approvisionnement. Un acteur souverain qui ne travaille que la première dépendance, diversifier ses fournisseurs de puces, sans jamais investir dans la seconde, la maturité d'un écosystème logiciel alternatif, n'aura résolu que la moitié la plus facile du problème.

SIX CAS RÉELS À CITER, ET LEUR LEÇON

CAS	CE QUI S'EST PASSÉ	LEÇON POUR UN COMITÉ DE DIRECTION
Le rendu neuronal sur les cartes grand public, 2026	Nvidia généralise sur ses cartes de jeu une technique combinant intelligence artificielle et rendu traditionnel pour améliorer la performance et la qualité d'image.	La frontière entre gaming et IA ne se referme jamais complètement ; elle se redessine en permanence dans les deux sens.
Le trillion de dollars de commandes annoncé à GTC 2026	Le fondateur de Nvidia annonce ce volume de commandes fermes jusqu'en 2027, avant de reconnaître publiquement une situation de pénurie.	Une entreprise en position dominante peut annoncer sans détour ses propres limites de production, un signe de rapport de force plutôt que de faiblesse.
AMD et ses contrats de plusieurs gigawatts avec deux géants de la tech	Le concurrent principal de Nvidia signe des accords de plusieurs gigawatts de capacité avec deux des plus grands utilisateurs mondiaux d'IA.	Une part de marché à un chiffre peut coexister avec des contrats individuels d'une échelle industrielle majeure ; les deux mesures racontent des histoires différentes.
Les puces internes des hyperscalers, 12 à 18 % du calcul déployé	Les plus grands opérateurs cloud déploient leurs propres puces d'IA en interne, réduisant d'autant leur dépendance à l'achat externe.	La plus efficace des stratégies de réduction de dépendance reste la conception propriétaire, accessible seulement aux acteurs disposant du capital et de l'échelle nécessaires.
Les notes de compatibilité CUDA contre ROCm et oneAPI	Les principaux cadriciels d'intelligence artificielle notent la compatibilité CUDA comme excellente, contre correcte à bonne pour les alternatives ouvertes d'AMD et d'Intel.	Un écart de maturité logicielle, même en voie de réduction, continue de peser sur chaque décision d'architecture technique au quotidien.
La marge brute de 84 à 88 % sur les puces datacenter	Nvidia dégage des marges très supérieures à celles de ses concurrents sur ses produits les plus avancés, malgré une demande qui dépasse déjà l'offre.	Une position dominante se traduit directement en capacité d'investissement disproportionnée dans la génération de puce suivante, un cercle qui s'auto-renforce.

TROIS DÉBATS DE FOND, ET LA POSITION DÉFENDABLE

1 · Faut-il investir dans une alternative à CUDA, ou l'accepter comme un standard de fait ?

Pour investir : dépendre d'un unique écosystème logiciel propriétaire expose à un risque stratégique que la performance seule du matériel concurrent ne suffit jamais à compenser.

Contre : l'écart de maturité reste tel que réécrire une infrastructure critique sur un écosystème encore jugé correct plutôt qu'excellent introduit un risque opérationnel immédiat, contre un bénéfice de souveraineté incertain à long terme.

Position défendable : investir dans la portabilité du code et la compatibilité multi-plateforme dès la conception, sans migrer prématurément une charge critique, pour garder l'option ouverte sans payer le prix de la bascule avant que l'écosystème alternatif ne soit réellement prêt.

2 · Le silicium propriétaire des hyperscalers annonce-t-il la fin de la domination Nvidia ?

Pour : une part croissante du calcul d'IA mondial échappe déjà à l'achat de GPU Nvidia, un mouvement structurel qui ne peut que s'accentuer.

Contre : ce silicium reste un actif interne non commercialisé, inaccessible à tout acteur qui n'a pas l'échelle d'un hyperscaler, un marché qui se referme plutôt qu'il ne s'ouvre à de nouveaux entrants.

Position défendable : cette tendance réduit le marché adressable de Nvidia chez les plus grands acteurs, sans offrir aucune option équivalente aux clients de taille moyenne ou souveraine, qui restent, eux, entièrement dépendants du marché ouvert.

3 · La pénurie de GPU doit-elle inquiéter un client souverain, ou le rassurer sur la demande ?

Pour l'inquiétude : une demande qui dépasse l'offre pousse mécaniquement les prix à la hausse et allonge les délais pour tout acheteur, souverain ou non.

Pour le rassurance : une pénurie de succès confirme la valeur stratégique de l'actif, justifiant un investissement anticipé plutôt qu'une attente d'une détente des prix qui pourrait ne jamais survenir.

Position défendable : la pénurie doit inciter à sécuriser des engagements de capacité à long terme dès maintenant, exactement comme le font déjà les plus grands acteurs du marché, plutôt qu'à parier sur un retournement de situation à court terme.

MATRICE DES RISQUES DES CARTES GRAPHIQUES

RISQUE	CRITICITÉ	MANIFESTATION
Confondre carte grand public et carte datacenter	Élevé	Dimensionner un projet d'IA sur les caractéristiques d'une carte GeForce grand public, alors que la mémoire et la bande passante nécessaires exigent la variante datacenter, bien plus coûteuse.
Ne traiter que la dépendance matérielle	Critique	Diversifier ses fournisseurs de puces sans jamais investir dans la portabilité logicielle, laissant intacte la dépendance à un écosystème propriétaire.
Ignorer la pénurie dans la planification budgétaire	Élevé	Bâtir un calendrier de déploiement d'IA sans intégrer les délais et les hausses de prix qu'une demande mondiale dépassant l'offre impose déjà.
Présenter le silicium des hyperscalers comme une alternative accessible	Moyen	Suggérer à un client de taille moyenne qu'une puce interne de type TPU ou Trainium constitue une option d'achat réelle, alors qu'elle reste un actif propriétaire non commercialisé.
Sous-estimer la maturité réelle des alternatives ouvertes	Moyen	Ne pas réévaluer régulièrement l'écart de compatibilité entre CUDA et ses alternatives, un écart qui se réduit chaque année sans avoir encore disparu.

LES SIX DÉCISIONS À TRANCHER CE TRIMESTRE

Q1 **Notre discours distingue-t-il systématiquement carte grand public et carte datacenter ?**
Vérifier que toute proposition technique nomme la variante mémoire et bus réellement requise pour la charge visée.

Q2 **Investissons-nous dans la portabilité logicielle, pas seulement dans la diversité matérielle ?**
Intégrer un critère de compatibilité multi-plateforme dans tout choix d'architecture d'IA dès la conception.

Q3 **Notre planification budgétaire intègre-t-elle la pénurie annoncée par le marché lui-même ?**
Sécuriser des engagements de capacité à moyen terme plutôt que de parier sur une détente des prix.

Q4 **Distinguons-nous silicium propriétaire des hyperscalers et alternative de marché ouvert ?**
Ne jamais présenter les puces internes des plus grands opérateurs comme une option d'achat accessible à un client tiers.

Q5 **Suivons-nous l'évolution réelle de la maturité de ROCm et des alternatives ouvertes ?**
Réévaluer cette veille chaque trimestre plutôt que de se fier à une évaluation figée.

Q6 **Notre offre documente-t-elle les deux dépendances, matérielle et logicielle, séparément ?**
S'assurer qu'aucune communication ne traite le sujet du GPU comme un problème d'approvisionnement matériel seul.

CINQ QUESTIONS QU'ON VOUS POSERA, ET CE QUI EST ATTENDU

Q1 **Pourquoi une carte graphique est-elle devenue le moteur de l'IA ?**
Attendu : parce que l'entraînement d'un réseau de neurones repose sur des multiplications de matrices, la même famille d'opérations que le calcul de pixels à grande échelle, un problème que des milliers de cœurs simples exécutés en parallèle résolvent mieux que quelques cœurs polyvalents.

Q2 **Quelle est la différence entre une carte grand public et une carte datacenter ?**
Attendu : la mémoire, GDDR7 moins coûteuse pour le grand public contre HBM3E à très haute bande passante pour le datacenter, et le bus d'entrée-sortie, deux variantes d'une même architecture jamais interchangeables.

Q3 **Qu'est-ce que CUDA, et pourquoi est-ce un verrou plus solide que le matériel ?**
Attendu : la couche logicielle propriétaire de Nvidia, utilisée par quatre millions de développeurs et jugée la plus compatible par les principaux cadriciels d'IA, un verrou qui survit à toute amélioration purement matérielle de la concurrence.

Q4 **Le silicium propriétaire des hyperscalers menace-t-il la position de Nvidia ?**
Attendu : il réduit le marché adressable chez les plus grands acteurs sans offrir d'alternative accessible aux clients de taille moyenne, qui restent entièrement dépendants du marché ouvert.

Q5 **Pourquoi Nvidia annonce-t-elle une pénurie malgré sa position dominante ?**
Attendu : la demande, portée par des commandes fermes de l'ordre du trillion de dollars, dépasse la capacité de production de l'industrie entière, une pénurie de succès plutôt qu'un échec industriel.

RED FLAGS DANS LES DISCOURS SUR LES CARTES GRAPHIQUES

« Une carte graphique grand public suffit pour notre projet d'IA »

La mémoire GDDR7 des cartes grand public n'offre ni la capacité ni la bande passante nécessaires aux modèles de taille significative, réservées à la variante datacenter en HBM3E.

« Nous avons diversifié nos fournisseurs de GPU, nous sommes souverains »

Sans investissement dans la portabilité logicielle, la dépendance à l'écosystème CUDA demeure entière, quel que soit le nombre de fournisseurs matériels.

« Les puces internes de Google ou d'Amazon sont une alternative accessible »

Ces puces restent des actifs propriétaires non commercialisés, jamais une option d'achat pour un client tiers.

« AMD a rattrapé Nvidia »

Une performance technique à 80-90 % ne s'est pas encore traduite en part de marché comparable, qui reste à un chiffre face à une domination toujours écrasante.

« La pénurie de GPU va bientôt se résorber »

Aucun signal du marché en 2026 ne confirme cette hypothèse ; les commandes fermes annoncées s'étendent au contraire jusqu'en 2027.

GLOSSAIRE C-LEVEL : 12 TERMES DES CARTES GRAPHIQUES

GPU	*Graphics Processing Unit.* Processeur massivement parallèle, conçu à l'origine pour le rendu graphique, composé de milliers de cœurs simples exécutant la même opération simultanément.
Cœur (core, GPU)	Unité de calcul élémentaire d'une carte graphique, bien plus simple qu'un cœur de processeur central, mais présente en très grand nombre.
Multiplication de matrices	Opération mathématique au cœur à la fois du rendu graphique et de l'apprentissage des réseaux de neurones, expliquant la convergence entre gaming et IA.
Blackwell	Architecture de puce Nvidia lancée fin 2024, déclinée en une variante grand public et une variante datacenter aux caractéristiques mémoire distinctes.
GDDR7	Génération de mémoire graphique utilisée sur les cartes grand public, moins coûteuse et suffisante pour l'affichage.
HBM3E	*High Bandwidth Memory.* Mémoire empilée à très haute bande passante, utilisée sur les puces datacenter, indispensable aux modèles d'IA les plus volumineux.
CUDA	Plateforme logicielle propriétaire de Nvidia permettant de programmer ses GPU, principal verrou d'écosystème du marché de l'IA.
ROCm	Plateforme logicielle ouverte d'AMD, alternative à CUDA, jugée en 2026 moins mature sur les principaux cadriciels d'IA.
Rendu neuronal (neural graphics)	Technique combinant modèles d'IA et rendu traditionnel pour améliorer la performance et la qualité d'image d'un jeu vidéo.
Silicium propriétaire (custom silicon)	Puce d'IA conçue en interne par un opérateur cloud pour son propre usage, non commercialisée sur le marché ouvert.
Accélérateur d'IA	Désignation générique d'une puce, GPU ou autre, spécialisée dans le calcul massivement parallèle requis par l'intelligence artificielle.
TCO (Total Cost of Ownership)	Coût total de possession d'une infrastructure, incluant matériel, énergie et exploitation, déjà détaillé pour le GPU dans la fiche Économie d'un cloudeur de cette collection.

La seule question vraiment stratégique à poser à votre COMEX :

« Notre stratégie GPU traite-t-elle la dépendance logicielle avec le même sérieux que la dépendance matérielle ? *Le silicium se diversifie, se négocie, parfois même se construit en interne. Le code, lui, s'apprend et se réécrit, un investissement humain que ni la diplomatie industrielle ni le contrat d'approvisionnement ne remplacent. Ignorer cette asymétrie, c'est résoudre la moitié la plus facile du problème et se croire arrivé.* »

FONDATIONS INFORMATIQUES · 2/11

Du composant au logiciel qui l'orchestre

Cette fiche a traité le composant le plus disputé de l'informatique actuelle. La fiche suivante, Systèmes d'exploitation et virtualisation, traite la couche logicielle qui orchestre tous les composants matériels déjà posés, CPU, RAM, stockage, GPU, et qui rend possible l'existence même du cloud.

Sources principales (septembre 2026)

1. NVIDIA Corp, Form 10-K, FY2026 : rendu neuronal, chiffre d'affaires datacenter.
2. Wikipedia, Blackwell (microarchitecture) : GDDR7/HBM3E, PCIe 5.0/6.0, fabrication TSMC.
3. CommandLinux, AI GPU Market Share NVIDIA vs AMD vs Intel: 2026 Statistics : parts de marché, silicium propriétaire des hyperscalers.
4. Silicon Analysts, AMD vs NVIDIA AI GPU Market Share 2026 : trillion de dollars de commandes, citation GTC 2026, marges comparées.
5. CoderCops, AI Chip Wars 2026 : CUDA 4 millions de développeurs, notes de compatibilité par cadriciel.
6. Axis Intelligence, Nvidia GPU Market Share 2026 : part de marché grand public, Jon Peddie Research.
7. Fiches Économie d'un cloudeur et Confidential Computing pour l'IA de la collection.

Briefing synthétisé pour audience C-level · Informations arrêtées au 19 septembre 2026, à revérifier · Ce document est une note de synthèse, non un conseil technique ou opérationnel.