

Une pénurie de mémoire RAM et stockage SSD sans précédent

Comité de suivi Matinfo

Version 1.0 du 18/09/2026

1 Diffusion de ce document

Ce document est destiné à toute personne faisant partie d'un établissement de l'enseignement supérieur et de la recherche français. Il n'a pas vocation à sortir de ce cadre.

2 Une augmentation de prix sans précédent pour la mémoire

Depuis septembre 2025, les prix de la mémoire de type DRAM et Flash se sont envolés vers des sommets encore jamais atteints. Les figures 1 (barrette de 16 Go de DDR5) et 2 (galette de base pour fabriquer p. ex. les disques SSD) parlent d'elles-mêmes et vous montrent les proportions de l'augmentation. La mémoire DRAM est utilisée dans tous les ordinateurs en tant que mémoire vive. La mémoire Flash de type NAND est utilisée dans tous les SSD des ordinateurs modernes pour le stockage de masse.

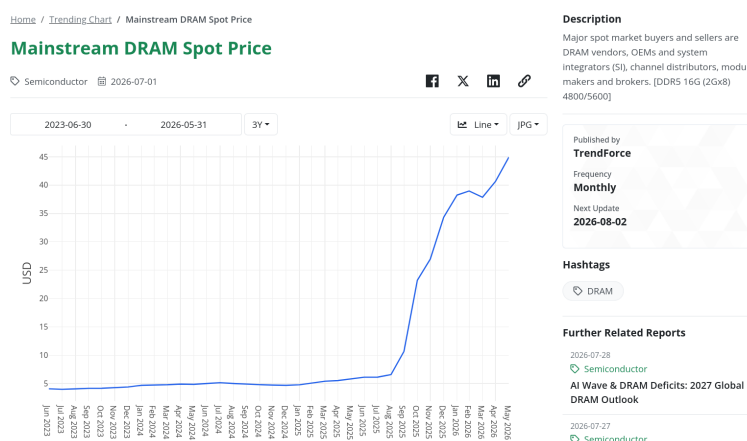


FIGURE 1 – Évolution du prix de la mémoire DRAM Pour un module de 16 Go de mémoire DRAM de type DDR5 4800/5600 : le prix était stable depuis 2023 et à partir de septembre 2025, il augmente brutalement et continûment pour arriver à un coefficient de x 9 (+800 %) en mai 2026. Source : <https://datatrack.trendforce.com/>

Avec le processeur, la mémoire vive et le stockage sont les trois plus gros postes du coût total d'un ordinateur. Vous comprendrez ainsi aisément la répercussion de ces augmentations sur les prix de commercialisation, aussi bien pour un poste de travail individuel que pour un serveur. Mais comment et pourquoi en sommes-nous arrivés là ?

3 Cocktail explosif : IA générative et loi de l'offre/demande débridée

Plus un jour ne se passe sans que l'on évoque l'intelligence artificielle (IA) et plus précisément l'IA générative (IAg). Les grands modèles de langage, ou de génération d'image ou de son, sont en train de chambouler la technique et la société. Les géants USA (OpenAI, Anthropic, Google, Meta, X,

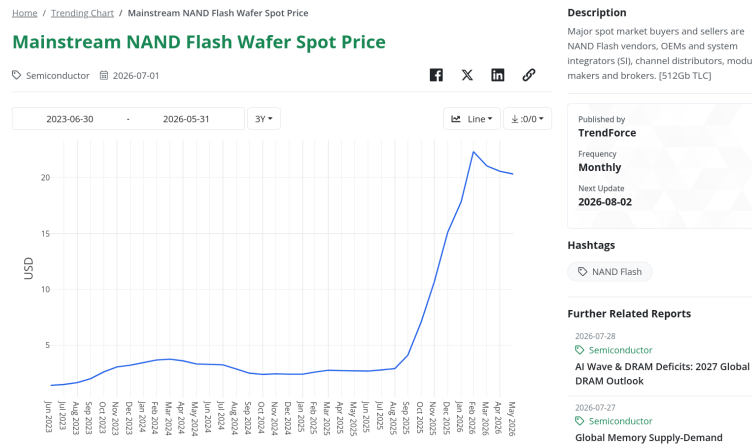


FIGURE 2 – Évolution du prix de 512 Go de mémoire flash NAND TLC Pour un ensemble de 512Go de mémoire flash type TLC utilisé dans les SSD : le prix était stable depuis 2023 et à partir de septembre 2025, il augmente brutalement et continûment pour arriver à un coefficient de plus de x 20 (+2000 %) en février 2026. Depuis, il semble y avoir un plateau. Source : <https://datatrack.trendforce.com/>

etc.) ou chinois ont bien vu l'intérêt économique et stratégique d'être des acteurs incontournables à l'échelle mondiale. Ils investissent donc des sommes colossales dans des serveurs dimensionnés pour l'IA et dans les datacenters qui vont autour. Une IA qui donne de « bons » résultats a dû être entraînée sur un corpus de données le plus grand possible (et de la meilleure qualité possible, mais c'est une autre histoire), d'où un besoin énorme de stockage. Toujours pour avoir de « bons » résultats, les modèles d'inférence (qui répondent au prompt) ont un besoin là aussi énorme de mémoire dans les composants spéciaux nommés GPU. Précisons que ces GPU utilisent de la mémoire spéciale très rapide, la High Bandwidth Memory (HBM), qui est une sorte d'empilement en 3D des mémoires DRAM traditionnelles que nous connaissons dans nos ordinateurs.

En résumé, chaque géant de l'IA se rue sur les GPU (garnis en mémoire HBM) et le stockage dans des proportions... gigantesques.

Un point important est que la chaîne d'approvisionnement de ces composants comporte peu d'acteurs mondiaux. Pour les GPU, il n'y a actuellement concrètement que Nvidia, qui lui-même dépend quasiment que du seul constructeur TSMC (Taïwan). Pour la mémoire RAM, Flash et surtout la HBM, il y a trois acteurs : Samsung, SKHynix (Corée du Sud) et Micron (USA). Ces fabricants de mémoire ont besoin d'outils de lithographie très sophistiqués que seul ASML (Pays-Bas) sait construire.

Il y a ainsi peu d'offre et une demande énorme. À chaque maillon de la chaîne, le vendeur vend à l'acheteur le plus offrant et monte ses prix pour faire du profit. Chaque acheteur a tendance à surenchérir pour pouvoir être devant ses concurrents. Et voilà comment les prix s'envolent.

Pour l'anecdote, les profits de Samsung sont tels que les employés ont menacé de faire grève¹ pour avoir leur part du gâteau et pour ne pas voir toute cette manne partir chez les actionnaires et dirigeants.

Les trois « fondeurs » de mémoire ont complètement réorienté leurs moyens de production de mémoire DRAM pour ne fabriquer quasi exclusivement que de la HBM pour vendre aux géants de l'IA. La DRAM traditionnelle de nos ordinateurs et serveurs devient ainsi une denrée rare... et donc chère, que s'arrachent au plus offrant les HP, Dell, Lenovo et consorts.

Une illustration de cette réorientation : Micron a complètement arrêté la commercialisation de sa gamme grand-public Crucial^{2,3}, bien connue des gamers et des geeks qui assemblent leur PC. On pourrait se dire « Mais si la demande est si forte, pourquoi ne pas augmenter l'offre ? ». Ce n'est pas si facile, car construire une usine de très haute technologie nécessite des investissements de plusieurs milliards d'euros et plusieurs années. Lors du Covid-19, ces fabricants avaient grandement investi face à la demande d'ordinateurs portables pour le télétravail et s'étaient retrouvés à terme

1. Samsung : une prime record pour les employés (Capital)
2. Crucial quitte le marché (Numerama)
3. Adieu Crucial (Hardware and co)

face à un marché saturé et une demande tarie. Chat échaudé craint l'eau froide et les investisseurs ne veulent pas risquer de perdre des milliards si d'aventure l'engouement pour l'IAg s'avérait être une bulle spéculative qui explose.

Il y a également des répercussions en cascade sur d'autres composants tels que les disques durs rotatifs (HDD) : la demande en stockage est telle que l'offre en SSD ne suit pas et les géants de l'IA se reportent sur les HDD, en surenchérissant au détriment des constructeurs traditionnels de PC et de serveurs. À terme, il semblerait que les processeurs et les GPU soient eux aussi touchés. Le fait que toute l'offre soit « détournée » vers la demande énorme d'un seul type d'acteur, prêt à surenchérir sur le concurrent, ne laisse que des reliquats aux autres acteurs et, pour les pièces dont on a la chance de pouvoir bénéficier, les délais s'allongent considérablement, passant de plusieurs semaines à plusieurs mois.

4 Quelles conséquences pour la recherche ?

Les établissements d'enseignement supérieur et de recherche (ESR) ont la possibilité de faire leurs achats de matériel informatique, postes de travail, pilotage d'expérience ou serveurs, via l'accord-cadre Matinfo⁴. Matinfo, par son effet de volume (250 établissements de l'ESR sont adhérents) permet en temps normal de bénéficier de tarifs et de services très avantageux. Cet accord-cadre comporte un ensemble de clauses juridiques qui respectent le code des marchés publics. En particulier, il y a des clauses d'encadrement des augmentations de prix et de partage de risque entre clients et fournisseurs.

Mais dans le contexte actuel où les augmentations dépassent les maximums qui avaient été pourtant largement dimensionnés, on en arrive à des situations de blocage. En effet, la pente de l'augmentation des prix est telle qu'il est très difficile de trouver un juste milieu pour les deux parties pour un bordereau de prix unitaires (BPU) qui doit être valable quelques mois, durée minimale nécessaire pour permettre le déroulement du processus d'achat : devis, commande, réception et paiement. Il faut à la fois que les fournisseurs ne vendent pas à perte et que l'ESR ne paie pas en argent public des sommes exagérées au-delà du « juste » prix.

À l'heure de la rédaction de ces lignes (mi-septembre 2026), par rapport au début du marché Matinfo 6 (octobre 2025), soit en 12 mois, les augmentations oscillent entre +100 % et +300 % selon les lots.

Les blocages temporaires des BPU Matinfo 6 et l'augmentation inédite et imprévisible perturbent fortement les activités de recherche qui se basent directement ou indirectement sur du matériel informatique. Les demandes de budget doivent se baser sur des devis qui ne veulent plus rien dire au moment du potentiel achat.

L'effet de pénurie a aussi ses répercussions par effet domino. Toujours à l'heure de la rédaction de ces lignes, pour les serveurs, les barrettes de DRAM de 16 et 32 Go ne sont plus commandables et il faut se rabattre, tant qu'il est possible d'en avoir, vers des barrettes de 64 Go. Dans un cas de serveur de calcul haute performance, on souhaite souvent peupler l'intégralité des bancs mémoire des processeurs à des fins de parallélisation, p. ex. on souhaite avoir 16 x 16 Go de DRAM plutôt que 4 x 64 Go de DRAM. Si pour un besoin de 16 x 16 Go on est obligé d'acheter 16 x 64 Go au prix fort, cela change totalement la donne budgétaire et l'achat sera probablement impossible.

5 Quelles conséquences plus généralement ?

Au-delà de notre sphère professionnelle de l'ESR, tout le monde va être ou est déjà touché dans sa sphère personnelle. Les smartphones et les consoles de jeu sont eux-aussi des ordinateurs qui ont besoin de mémoire.

Les « professionnels de la profession » qui se hasardent à faire des pronostics annoncent que tout 2027 va être encore difficile et qu'il est probable que 2028 le soit aussi. On lit et on entend également que lorsque les prix baisseront, ils ne retrouveront probablement pas les niveaux de début 2025. L'âge d'or de la mémoire bon marché ne serait plus qu'un souvenir. Il va falloir attendre d'être en 2029 pour constater qui aura eu raison, mais en attendant bien malin qui saura faire un bon pronostic.

4. <https://matinfo.fr>

Une action de groupe (class action) a été lancée aux USA fin juin 2026⁵,⁶ qui accuse les trois grands fondeurs de mémoire de s'être entendus sur les prix et d'avoir artificiellement augmenté la pénurie. Il sera intéressant de lire les conclusions du tribunal états-unien.

6 Le gigantisme des datacenters pour l'IAg

En cette fin d'article, Nous aimerions amener les personnes lectrices à considérer les proportions des moyens mis en œuvre par les géants de l'IAg.

Un article⁷ sur le site de l'IEEE décrit le gigantisme des datacenters. Il est difficile de le résumer et j'encourage les lecteurs et lectrices motivés à le lire (c'est en anglais). Je vais ici me limiter à énumérer des chiffres en guise d'éléments de réflexion.

Meta construit en Louisiane le datacenter Hyperion⁸,⁹. L'investissement est de 10-15 milliards de \$. La puissance électrique consommée en continu est de 5 GW. Rappelons que l'EPR de Flamanville, c'est 1,6 GW. Le datacenter comporte 41 000 baies (les armoires qui contiennent des serveurs). Chaque baie contient 72 GPU NVidia de type GB200, contient 13 To de HBM, consomme en continu 120 kW et pèse 1,5 t. Au total, cela fait plus de 3 millions de GPU. Ce datacenter nécessite la construction de 3 centrales électriques à gaz.

OpenAI construit au Texas le datacenter Stargate¹⁰ qui héberge 450 000 GPU NVidia de type GB200. Il consomme en continu une puissance électrique de 300 MW fournie par tout un ensemble de turbines à gaz.

XAI construit à Memphis (Tennessee) le datacenter Colossus 2¹¹ qui contient 550 000 GPU. L'article ne parle pas du stockage, c'est-à-dire du nombre de disques SSD utilisés et du volume total par datacenter que cela représente.

Globalement, les demandes pour les turbines à gaz pour entraîner les génératrices électriques sont telles que le délai est de 7 ans pour en avoir une neuve et le marché de l'occasion pour les turbines aviation (reconditionnement de turboréacteurs d'avion) est en plein essor.

7 Conclusion

Le collectif Matinfo met tout en œuvre pour trouver une issue à cette situation exceptionnelle et pour que l'accord-cadre retrouve un fonctionnement plus fluide dans les meilleurs délais.

Article rédigé (sans IA) par le Cosu Matinfo

5. Entente présumée sur les prix de la RAM (Clubic)
6. Pénurie de RAM : ont-ils orchestré la crise ? (Numerama)
7. 5 GW datacenter (IEEE)
8. Image du datacenter Meta Hyperion (IEEE)
9. Image du datacenter Meta Hyperion (Fortune)
10. Image du datacenter Stargate OpenAI
11. Image du datacenter Colossus 2 XAI