ANALYSIS



L'IA, industrie d'infrastructure

La fin de l'abondance numérique

Novembre 2026

L'économie numérique a longtemps été appréhendée comme une économie du logiciel. La compétitivité reposait avant tout sur la capacité à concevoir de meilleurs algorithmes, attirer un nombre croissant d'utilisateurs et exploiter les données à grande échelle. Dans ce modèle, les ressources informatiques demeuraient largement invisibles : elles étaient consommées comme une commodité, selon des coûts relativement prévisibles et rarement déterminants dans les choix stratégiques. Les infrastructures rendaient cette croissance possible, sans pour autant en constituer le principal facteur de différenciation.

En 2025, l'Agence internationale de l'énergie [1] indiquait que les dépenses d'investissement de cinq des plus grandes entreprises technologiques ont dépassé **400 milliards de dollars** et que l'essentiel de ces investissements était désormais dirigé vers les **centres de données et les infrastructures nécessaires au développement de l'intelligence artificielle**, au point de dépasser les investissements mondiaux consacrés à la production de pétrole et de gaz. Ces investissements devraient encore progresser de **75 % en 2026**.

En mars 2025, **BlackRock**, **Microsoft**, **MGX**, rejoints ensuite par **NVIDIA** et **xAI**, ont lancé un partenariat mondial destiné à financer de nouvelles infrastructures de calcul et les réseaux électriques associés. En parallèle, les plus grands gestionnaires d'actifs considèrent désormais **l'approvisionnement en électricité comme l'un des principaux facteurs limitants du développement de l'IA**, estimant que la demande des centres de données pourrait nécessiter près de **150 gigawatts** de capacités électriques supplémentaires d'ici la fin de la décennie.



Entre 2019 et 2025, les performances des supercalculateurs dédiés à l'IA ont doublé tous les neuf mois, tandis que leurs besoins en puissance électrique et le coût de leur matériel doublaient environ chaque année. Le plus grand système identifié en 2025 mobilisait déjà **200 000 accélérateurs**, représentait près de **7 milliards de dollars** de matériel et nécessitait **300 mégawatts** de puissance électrique. Si cette trajectoire se poursuit, les prochaines générations de supercalculateurs pourraient se mesurer en **gigawatts** plutôt qu'en capacité de calcul.

Ces évolutions traduisent un déplacement progressif des critères de compétitivité. La performance ne dépend plus uniquement des algorithmes ou de la qualité des modèles, mais aussi de la capacité à mobiliser de l'énergie, du capital, des semi-conducteurs et des infrastructures de calcul. Comprendre l'IA suppose dès lors de s'intéresser non seulement aux modèles eux-mêmes, mais également aux ressources qui rendent leur développement possible : **l'IA est-elle encore une industrie du logiciel, ou devient-elle progressivement une industrie d'infrastructures ?**

Source

[1] https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary?

I. Le calcul devient le facteur limitant à la performance

Pendant plusieurs décennies, les progrès de l'économie numérique ont principalement reposé sur des innovations logicielles. Les capacités de calcul continuaient naturellement de progresser, mais elles constituaient rarement une contrainte majeure. La puissance informatique suivait une trajectoire suffisamment prévisible pour accompagner les besoins des logiciels.

L'intelligence artificielle modifie progressivement cet équilibre. L'entraînement comme l'exploitation des modèles les plus avancés reposent désormais sur une ressource qui n'est plus abondante : **le calcul**. Celui-ci ne se résume pas au nombre de GPU disponibles. Il résulte d'une chaîne beaucoup plus large qui associe semi-conducteurs, centres de données, capacités de refroidissement, alimentation électrique, réseaux, foncier et capitaux nécessaires pour financer l'ensemble de ces infrastructures.

Les indicateurs utilisés pour mesurer les capacités de calcul ont eux-mêmes évolué. La référence est progressivement passée du nombre d'opérations réalisées par seconde (FLOPS) au nombre de GPU mobilisés, avant de s'intéresser aujourd'hui à la puissance électrique disponible. Ce glissement n'est pas anecdotique. Il traduit le fait que la capacité de calcul dépend de moins en moins des seules performances des processeurs et de plus en plus de la possibilité de mobiliser des infrastructures industrielles de grande ampleur.

Les ordres de grandeur illustrent cette transformation. Les infrastructures des principaux fournisseurs américains atteignent déjà plus de **2 gigawatts de puissance installée chacun**, soit l'équivalent de deux réacteurs nucléaires. À titre de comparaison, les capacités actuellement mobilisées par des acteurs européens comme Mistral seraient de l'ordre de **100 mégawatts**, soit près de vingt fois moins. Dans le même temps, **le coût d'entraînement des modèles les plus avancés est désormais estimé à plusieurs centaines de millions de dollars** et pourrait rapidement dépasser le milliard.

Mais cette montée en puissance révèle surtout que le véritable facteur limitant ne réside plus uniquement dans les processeurs graphiques. Selon l'Agence internationale de l'énergie, l'ensemble de la chaîne de valeur est désormais sous tension : production de semi-conducteurs avancés, mémoire à haute bande passante (HBM), transformateurs électriques, turbines, raccordements au réseau ou encore capacités de financement figurent désormais parmi les principaux **goulots d'étranglement** susceptibles de ralentir le développement des capacités de calcul.

L'électricité constitue à elle seule une illustration de ce changement. Alors que les générations précédentes de centres de données pouvaient être déployées relativement rapidement, les délais nécessaires au développement des infrastructures énergétiques se comptent désormais en années. L'Agence internationale de l'énergie estime qu'il faut généralement entre **quatre et huit ans** pour construire une nouvelle ligne de transport d'électricité dans les économies avancées, tandis que les délais d'approvisionnement en transformateurs ont doublé en quelques années. Autrement dit, **la vitesse de développement de l'intelligence artificielle dépend désormais de plus en plus du rythme, beaucoup plus lent, auquel peuvent être déployées des infrastructures physiques**.

Toujours selon l'Agence internationale de l'énergie, la densité de puissance des serveurs dédiés à l'IA a été multipliée par **onze entre 2020 et 2025** et pourrait encore quadrupler d'ici 2027. À cet horizon, un seul rack de serveurs pourrait atteindre une puissance de pointe comparable à celle consommée par **plus de soixante foyers**. Là où les générations précédentes de serveurs étaient principalement limitées par leur puissance de calcul, les nouvelles générations sont désormais contraintes par leur alimentation électrique, leur refroidissement et leur intégration dans les centres de données.

Enfin, le calcul devient également un enjeu financier. Les investissements nécessaires au développement des infrastructures sont désormais tels que, selon l'Agence internationale de l'énergie, ils ne peuvent plus être financés uniquement sur les bilans des grandes entreprises technologiques. Leur développement dépend de plus en plus des marchés de capitaux et des anticipations de retour sur investissement. **Le calcul cesse ainsi d'être une simple ressource informatique pour devenir un actif industriel** dont la disponibilité dépend autant de l'accès au capital que des avancées technologiques.

II. La rareté du calcul crée de nouveaux arbitrages

Comme l'énergie dans l'industrie ou les capacités logistiques dans le commerce mondial, le calcul conditionne désormais une partie croissante de la compétitivité des acteurs de l'intelligence artificielle. Ce déplacement des contraintes prépare un changement plus profond : lorsque le calcul devient une ressource rare, il ne suffit plus de chercher à en produire davantage. Il devient également nécessaire d'en optimiser l'allocation, le coût et l'usage. C'est précisément ce qui fait émerger de nouveaux arbitrages économiques et technologiques. Cette évolution marque une rupture avec l'économie traditionnelle du logiciel. Une fois développé, un logiciel pouvait être reproduit à grande échelle pour un coût marginal très faible. À l'inverse, chaque requête adressée à un modèle d'intelligence artificielle continue de mobiliser des infrastructures, de l'énergie et des capacités de calcul. **Les choix technologiques deviennent ainsi progressivement des arbitrages économiques**.

Deux trajectoires se dessinent aujourd'hui : la première repose sur **l'augmentation continue des ressources mobilisées** pour l'entraînement et l'inférence des modèles, tandis que la seconde vise à **améliorer leurs performances à puissance de calcul constante**. L'industrie progresse ainsi simultanément par l'accroissement des capacités de calcul et par l'optimisation de leur utilisation.

Les progrès récents de certains acteurs chinois illustrent cette seconde approche. Les architectures de type *Mixture of Experts*, utilisées notamment par DeepSeek, n'activent qu'une partie des paramètres du modèle à chaque requête. L'objectif n'est plus seulement de produire davantage de calcul, mais de produire davantage d'intelligence avec une quantité de calcul donnée. L'avantage compétitif dépend alors autant de l'efficacité du calcul que de son volume. Lorsque le calcul devient une ressource contrainte, la capacité à l'utiliser efficacement acquiert une valeur économique croissante. Les arbitrages se déplacent alors du seul terrain technologique vers celui de l'allocation des ressources.

Cette tension apparaît également du côté des utilisateurs. Au Crédit Agricole, les coûts unitaires d'inférence ont diminué d'environ **75 % en dix-huit mois**, une baisse qui pourrait laisser penser que le coût du calcul est appelé à disparaître.

L'expérience observée est pourtant différente. Rapportés à des volumes de l'ordre de **60 milliards d'interactions**, ces mêmes coûts peuvent conduire à des montants annuels jugés insoutenables, même pour des banques comme le Crédit Agricole. La baisse du coût unitaire du calcul ne réduit donc pas nécessairement la facture globale lorsque les usages progressent encore plus rapidement.

Le calcul introduit enfin des arbitrages qui étaient jusqu'ici largement étrangers à l'industrie du logiciel : le choix d'une architecture d'inférence peut désormais dépendre du coût de l'électricité, de la disponibilité des infrastructures ou de la localisation des centres de données. L'optimisation d'un système d'IA ne consiste donc plus uniquement à sélectionner le modèle le plus performant, mais à trouver le meilleur compromis entre performance, coût, consommation énergétique et disponibilité des ressources. Or cette logique conduit à arbitrer entre performance informatique et coût des ressources mobilisées. Dans certains cas, une architecture moins performante peut être privilégiée lorsque le coût de l'électricité ou des capacités de calcul qu'elle requiert est plus faible.

À mesure que le calcul devient un facteur de production, les décisions relatives aux systèmes d'intelligence artificielle se rapprochent ainsi de celles observées dans les industries lourdes, intensives en capital. Les économies d'échelle deviennent plus importantes. Les investissements initiaux augmentent fortement. L'accès aux ressources conditionne une partie de la compétitivité. Les arbitrages technologiques doivent également tenir compte de paramètres longtemps restés en arrière-plan dans l'industrie du logiciel : coût de l'énergie, utilisation des infrastructures ou rendement du calcul disponible.

Conclusion – L'IA devient un sujet de pilotage économique

À mesure que l'intelligence artificielle passe de l'expérimentation au déploiement à grande échelle, les organisations qui déploient l'IA à grande échelle sont ainsi conduites à arbitrer entre les fonctionnalités offertes par les modèles qui restent évidemment un critère de choix, mais elles s'apprécient désormais également au regard des ressources nécessaires à leur exploitation, tels que la consommation en calcul, leur coût d'exploitation, la disponibilité des infrastructures, l'efficacité énergétique et la soutenabilité économique des solutions retenues.

Ces évolutions ne transforment pas l'IA en industrie énergétique ou en industrie des semi-conducteurs. Elles invitent toutefois à la regarder autrement : moins comme une simple extension de l'économie du logiciel que comme une activité dont les performances reposent de plus en plus sur la maîtrise de ressources physiques rares.

Contact



Jean Barrere
Partner
jean.barrere@accuracy.com
+33 6 45 87 34 73



Les associés et les professionnels d'Accuracy sont disponibles pour discuter de vos besoins et concevoir une solution adaptée avec les experts compétents.