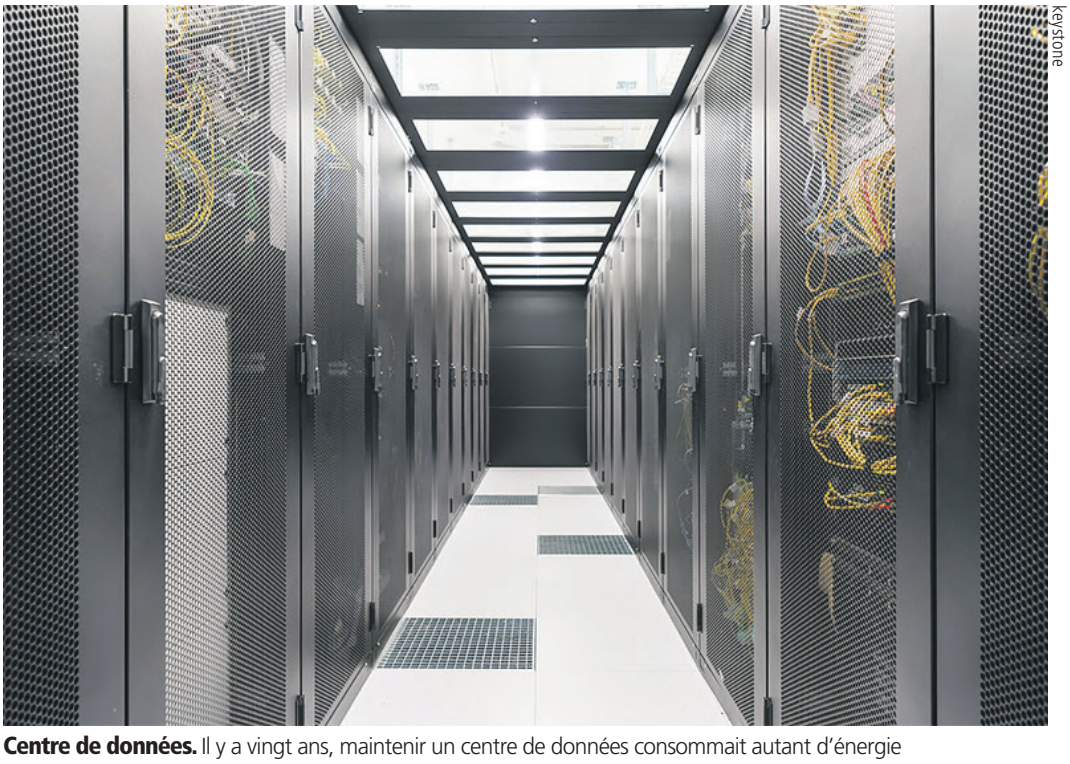
L'IA s'attaque au défi de la consommation d'énergie

INTELLIGENCE ARTIFICIELLE. Les acteurs du secteur cherchent à optimiser le refroidissement des centres de données, un système particulièrement énergivore, ou encore l'efficacité des puces elles-mêmes.

Grâce à de nouvelles techniques de refroidissement, des puces plus performantes ou l'évolution de la programmation, le secteur de l'intelligence artificielle (IA) cherche à limiter sa consommation d'énergie, dans un contexte de croissance frénétique.

L'infrastructure IA repose sur les centres de données, qui devraient peser, selon l'Agence internationale de l'énergie (AIE), environ 3% des besoins mondiaux en électricité d'ici à 2030, soit le double de la proportion actuelle. Mardi, Donald Trump devait se rendre en Pennsylvanie pour annoncer, selon plusieurs médias, quelque 70 milliards de dollars d'investissements dans cet Etat pour l'IA et les infrastructures énergétiques.

«Il y a plusieurs façons de traiter le problème», avance Mosharaf Chowdhury, professeur à l'Université du Michigan. «Vous pouvez augmenter les sources d'énergie», voie dans laquelle sont aussi engagés les poids lourds de l'IA, «ou réduire la demande» d'électricité à capacité équivalente, dit-il. Pour l'universitaire, des solutions «malignes» peuvent être trouvées à tous les niveaux de la chaîne de l'IA



keystone

Centre de données. Il y a vingt ans, maintenir un centre de données consommait autant d'énergie que les serveurs. Aujourd'hui, cela ne représente plus que 10% de leur consommation, selon le cabinet Arup.

Refroidissement optimisé

Selon Gareth Williams, du cabinet de conseil Arup, aujourd'hui, l'énergie nécessaire à maintenir un centre de données représente 10% de ce que consomment les serveurs eux-mêmes, contre 100% il y a 20 ans.

Cette réduction est à mettre au crédit, entre autres, de la généralisation du refroidissement liquide, en lieu et place de la ventilation classique, qui va jusqu'à faire circuler des fluides directement dans les serveurs.

«Tous les gros l'utilisent, maintenant», observe Gareth Williams, «c'est devenu incontournable».

Les puces Nvidia ont multiplié par plus de 100 la consommation d'une armoire à serveurs par rapport à ce qu'elle représentait il y a 20 ans.

De ce fait, le liquide peut monter à des températures nettement plus élevées que précédemment, selon Gareth Williams, mais cela rend, paradoxalement, le refroidissement plus facile au contact de l'air extérieur, du fait de la différence de température.

«Faire moins d'argent»

Autre évolution, les centres de données sont maintenant équipés de capteurs, mis à profit par l'IA pour contrôler la température non plus à l'échelle d'un site mais par «micro-zones» et «optimiser la consommation d'eau et d'électricité» de manière anticipée, selon Pankaj Sachdeva, du cabinet McKinsey.

Le laboratoire de Mosharaf Chowdhury a mis au point des algorithmes pour évaluer précisément la quantité d'électricité nécessaire à chaque puce pour fonctionner avec, à la clé, un gain de 20% à 30%.

L'équipe de Yi Ding, professeure à l'Université de Purdue, a, elle, démontré que l'on pouvait prolonger la vie des puces les plus performantes pour l'IA, les GPU (graphics processing unit) ou cartes graphiques, «sans amoindrir les performances», dit-elle à l'AFP.

«Mais il est difficile de convaincre les fabricants de semi-conducteurs de faire moins d'argent» en incitant les consommateurs à utiliser les mêmes équipements plus longtemps, glisse l'universitaire.

Le match se joue aussi au niveau de la programmation et de l'entraînement des grands modèles d'IA générative.

En janvier, le chinois DeepSeek a ainsi présenté son modèle d'IA générative R1 aux performances similaires à celles des grands acteurs américains bien que développé avec des GPU moins puissants.

Les ingénieurs de la start-up y sont parvenus notamment en programmant plus précisément les cartes graphiques. Ils ont aussi quasiment sauté une étape d'entraînement du modèle, jugée indispensable jusqu'ici.

Augmentation de la demande

Pour autant, malgré ces percées technologiques, «on ne pourra pas réduire la consommation totale d'énergie, à cause du paradoxe de Jevons», prédit Yi Ding. L'économiste britannique William Stanley Jevons a proclamé ainsi qu'une utilisation plus efficace d'une ressource limitée fait mécaniquement augmenter la demande, car son coût diminue.

«La consommation d'énergie va continuer de monter», avertit Yi Ding, malgré tous les efforts pour la limiter, «mais peut-être moins rapidement». (afp)