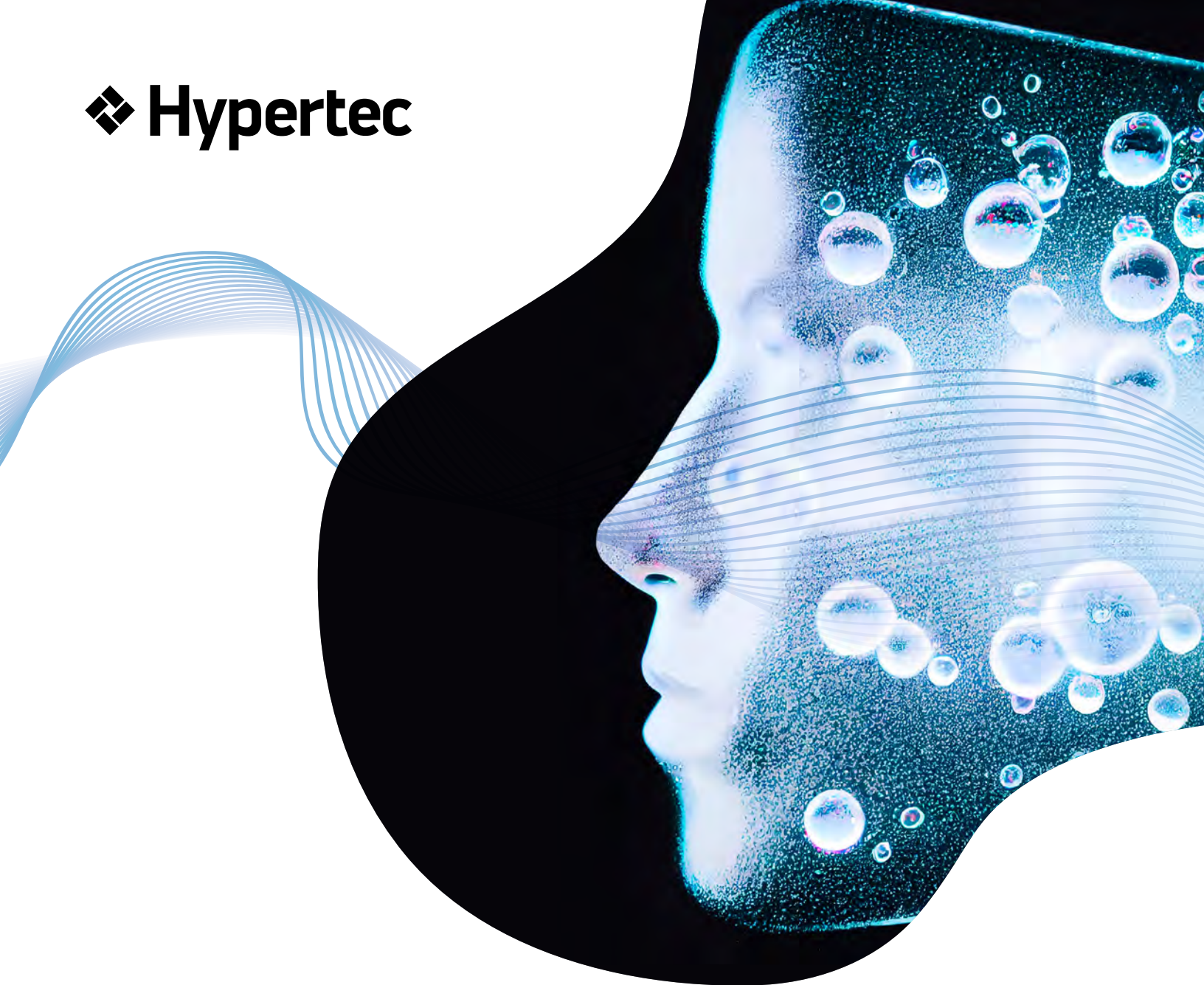




Découvrir l'avantage de l'immersion pour l'IA

Une approche optimale pour la création d'infrastructures de nouvelle génération

Rapport spécial

Table des matières

SOMMAIRE	3
INTRODUCTION	4
LA VÉRITÉ SUR LE REFROIDISSEMENT PAR AIR	6
L'efficacité de l'utilisation énergétique (PUE)	7
L'efficacité de l'utilisation de l'eau (WUE)	7
L'environnement, le social et la gouvernance (ESG)	7
Le contrôle des coûts	7
UN REFROIDISSEMENT ÉMERGENT POUR L'IA	8
Les échangeurs de chaleur de « porte arrière »	9
Direct sur puce	9
Le refroidissement par immersion	10
UNE PLONGÉE EN PROFONDEUR DANS LE REFROIDISSEMENT PAR IMMERSION	12
Il permet une meilleure utilisation de la capacité énergétique	12
Il élimine l'espace vide entre les serveurs	13
Il supprime totalement le bruit des ventilateurs et des souffleries	13
Il réduit les taux de défaillance des composants	13
Il augmente la durée de vie du matériel	14
Il prend en charge plusieurs générations de processeurs sans nécessiter de modifications ultérieures	14
Il permet à l'infrastructure de l'IA de vivre en dehors du centre de données	14
LE RETOUR SUR INVESTISSEMENT DU REFROIDISSEMENT PAR IMMERSION	16
Le refroidissement par immersion se compare favorablement aux autres technologies de refroidissement par liquide	16
Le refroidissement par immersion réduit les coûts d'exploitation actuels et futurs	16
Le refroidissement par immersion offre une valeur à long terme	17
Le refroidissement par immersion est respectueux de l'environnement	17
CONCLUSION	18
RÉFÉRENCES	21

Sommaire

La croissance explosive de l'IA exerce une pression sans précédent sur les opérateurs de centres de données.

En moins de deux ans, les opérateurs qui n'étaient pas en mesure d'alimenter et de refroidir plus de 10 kW par rack sont contraints à utiliser des serveurs qui consomment **jusqu'à 10,2 kW — chacun** — et sont conscients que l'industrie est sur la bonne voie pour consommer **2 kW par connecteur dans un court laps de temps**. Ce changement majeur dans la conception des serveurs, conçus pour prendre en charge l'IA, oblige les dirigeants, les ingénieurs et les administrateurs des centres de données à revoir en profondeur les technologies dont ils auront besoin pour réussir, qu'ils le veuillent ou non.

Pour faire face à ce changement, les organisations sont obligées de repenser leurs technologies de refroidissement. L'époque des unités de conditionnement d'air pour salle informatique (CRAC) et celles des traitements de l'air pour salle informatique (CRAH) commence à s'estomper. Cela s'explique en partie par le fait que le refroidissement par air ne peut refroidir en toute sécurité qu'un maximum de 40 kW de serveurs par rack, alors que l'infrastructure d'IA exige des densités de 80 kW ou plus. Mais il y a une autre raison, moins évidente, qui pousse les entreprises à repenser leurs technologies de refroidissement. Selon l'[Uptime Institute](#), l'amélioration de l'efficacité énergétique est devenue une préoccupation majeure, et le refroidissement par air est réputé pour son gaspillage d'énergie. Les opérateurs de centres de données, quelle que soit leur taille, doivent améliorer leur efficacité énergétique - et les technologies de refroidissement alternatives font partie de leur approche.

À la lecture de ce document, vous explorerez l'impact de l'IA sur les centres de données d'aujourd'hui. Vous étudierez en profondeur les technologies qui exercent une pression sur les opérateurs de centres de données, explorerez les alternatives au refroidissement par air et terminerez par une évaluation des raisons pour lesquelles une technologie de refroidissement particulière est une technologie optimale pour une infrastructure d'IA de pointe.

Après avoir lu ce document, vous aurez une meilleure compréhension des technologies de refroidissement pour les centres de données de la prochaine génération, et vous découvrirez une organisation qui pourrait devenir le partenaire essentiel de vos ambitions en matière d'IA.



Introduction

Comme le savent tous ceux qui travaillent dans le secteur des centres de données, l'arrivée de l'IA a provoqué une onde de choc dans notre secteur, bouleversant d'importantes hypothèses relatives à la conception et à l'économie des centres de données.

Répondant à des centaines de cas d'utilisation, elle s'infiltré dans de [nombreuses industries](#) même si elle nécessite une infrastructure ressemblant davantage à un supercalculateur qu'à une infrastructure infonuagique traditionnelle. Construits avec des centaines ou des milliers de GPU haute performance, des CPU, des clés USB et une infrastructure réseau intensive, les clusters d'entraînement à l'IA possèdent une architecture que la plupart des centres de données ne peuvent pas prendre en charge avec les conceptions et les technologies des centres de données traditionnels.

Puisque l'IA exige l'infrastructure de calcul la plus puissante possible, on assiste à une sorte de course contre la montre entre de nombreux acteurs clés - des sociétés de colocation aux fournisseurs de silicium en passant par les fournisseurs de services infonuagiques à grande échelle (hyperscalers) - pour créer la prochaine infrastructure d'IA la plus innovante qui soit afin de réaliser leurs ambitions en matière d'IA - qu'il s'agisse de résoudre un cas d'utilisation unique ou d'accaparer le marché.

Cependant, un problème bloque la voie vers l'innovation en matière d'IA. L'IA est à l'origine d'une tempête parfaite en termes de consommation d'énergie des centres de données... et de production de chaleur.

Pourquoi?

Quelques changements technologiques sont à l'origine de cette tempête.

Pour répondre aux besoins de calcul exigeants de l'IA, les centres de données sont témoins de l'augmentation du nombre de processeurs spécialisés connus sous le nom de xPU (par exemple, GPU, TPU). Contrairement à l'architecture standard de l'infonuagique avec des serveurs contenant un ou deux processeurs x86, les serveurs destinés à l'IA contiennent des processeurs x86 et plusieurs xPU. Les xPU excellent dans les tâches de traitement parallèle, ce qui accélère considérablement la formation et l'inférence de l'IA par rapport aux CPU traditionnels.



Pour atteindre les niveaux de performance souhaités, les serveurs d'IA disposent de quatre, six, huit ou dix processeurs - un ou deux CPU et plusieurs xPU. Ils nécessitent également des grandes quantités de mémoire, des clés USB et une infrastructure de réseau étendue qui dépasse de loin toutes les conceptions d'infrastructure que nous avons vues, à l'exception des superordinateurs.

Si vous comparez cette image avec le nombre relativement limité de silicium, de mémoire, de stockage et de réseaux utilisés dans les serveurs d'infrastructure en nuage, vous pouvez constater qu'une quête agressive dans l'amélioration des performances conduit à une augmentation continue de la production de chaleur dans les serveurs. Plus de composants = plus de chaleur.

Pour compliquer les choses, depuis un certain temps, la conception de circuits intégrés en silicium, suivant de près l'augmentation des performances d'une génération à l'autre, a été contrainte à une consommation d'énergie encore plus importante. Depuis des années, chaque nouvelle génération de processeurs s'accompagne d'une augmentation de la consommation d'énergie et de la puissance de consommation thermique (TDP), c'est-à-dire la quantité maximale de chaleur qu'un processeur est censé dissiper.

Il n'y a pas si longtemps, un processeur x86 typique avait un TDP de 300W et un GPU, de 150W. Qu'en est-il aujourd'hui? Voici un tableau des spécifications de puissance thermique de calcul (TDP) pour certains des GPU d'IA les plus populaires.

NOM	TDP
Intel Gaudi 2	600W
AMD Instinct MI300	600W
AMD Instinct MI300A	Pic de 760W
NVIDIA H200	700W

Pouvez-vous imaginer un serveur équipé de dix processeurs de 700 W, soit un serveur dont la puissance serait de 7 kW? Étonnamment, le TDP de ce processeur et la consommation d'énergie du serveur sont presque obsolètes. La prochaine génération de processeurs Intel CPU/GPU consommera 1500W. Le dernier serveur NVIDIA DGX B200 nécessite jusqu'à 14,3 kW de puissance.

Cette augmentation explosive du nombre de processeurs, associée à l'augmentation du TDP, est une véritable tempête pour la gestion thermique des centres de données. Des dizaines de milliers de centres de données ont été conçus pour supporter une puissance et une chaleur de 10 à 15 kW par bâti. Aujourd'hui, on leur demande de supporter 10 à 14 kW par serveur - et de prévoir une augmentation constante de la densité et du TDP.

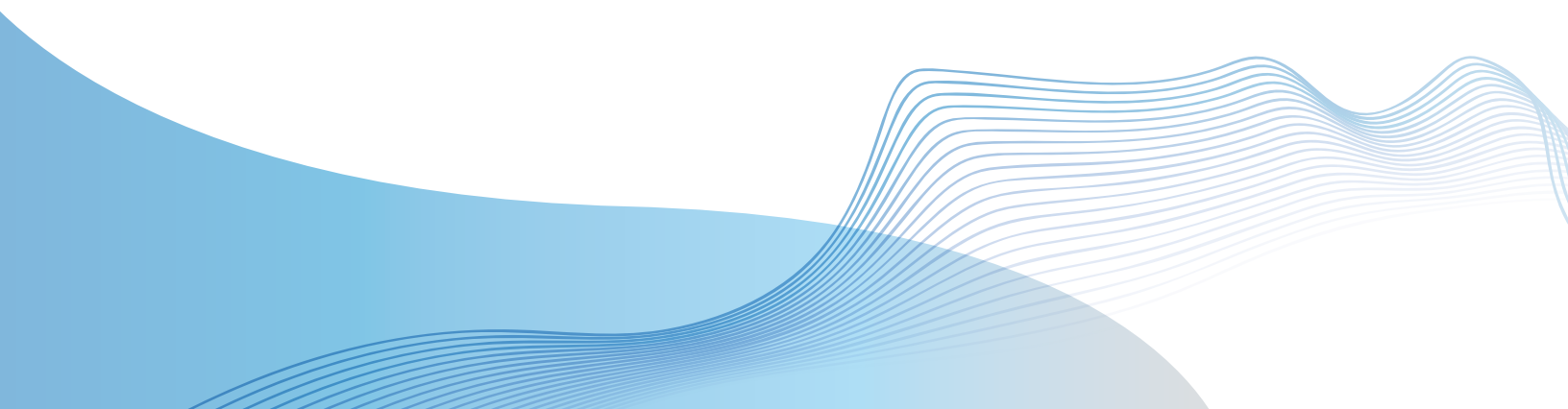
Cette évolution met en évidence un défi fondamental: le refroidissement par air conventionnel ne peut tout simplement plus suivre.



La vérité sur le refroidissement par air

Il est vrai que pendant des décennies, le refroidissement par air a été la pierre angulaire dans la gestion thermique des centres de données, mais le regroupement de six, huit, voire dix processeurs à TDP élevé dans un serveur 2U crée des goulots d'étranglement thermiques que le **refroidissement par air** traditionnel a du mal à résoudre. Les systèmes de refroidissement par air ont également du mal à suivre le rythme de l'augmentation constante de la chaleur générée par la prolifération des composants à TDP élevé, soulevant des inquiétudes quant à l'étranglement thermique et à la possibilité d'une défaillance prématurée de l'équipement.

Pire encore, le refroidissement par air est intrinsèquement inefficace, et l'utilisation des ressources devient de plus en plus une priorité absolue à mesure que les organisations sont invitées à faire plus avec moins. Les opérateurs de centres de données cherchent différents moyens de se concentrer sur l'efficacité opérationnelle, la durabilité et le contrôle des coûts. Eux-mêmes, leurs clients et leurs investisseurs recherchent des moyens de s'améliorer:



L'efficacité de l'utilisation énergétique (PUE)

L'efficacité de l'utilisation énergétique (PUE) est utilisée pour mesurer l'efficacité énergétique d'un centre de données. Bien qu'imparfaite, c'est la meilleure méthode de mesure dont dispose l'industrie. La plupart des centres de données sont aux prises avec un indicateur PUE sous-optimal, entraînant une augmentation des coûts énergétiques et de l'impact sur l'environnement. Selon certaines études, le PUE moyen se situe à 1,58, ce qui signifie que, pour 1 000 W dépensés pour l'informatique, 580 W sont gaspillés. Le refroidissement par air est l'une des principales causes d'un PUE sous-optimal. L'énergie dépensée à cet effet, que ce soit par des ventilateurs, des souffleurs ou des compresseurs, est de l'énergie qui n'est pas consacrée à la capacité de calcul.

L'efficacité de l'utilisation de l'eau (WUE)

Le refroidissement par air fonctionnant avec des réfrigérants est une solution coûteuse, qui incite les fournisseurs de centres de données à adopter des systèmes de refroidissement par évaporation, qui consomment de grandes quantités d'eau tout en polluant l'eau rejetée avec des contaminants. Le gaspillage de l'eau représente un problème à la fois pratique, en raison des réserves d'eau limitées, et politique. Les fournisseurs de centres de données ont besoin de modèles qui réduisent ou éliminent complètement l'utilisation de l'eau.

L'environnement, le social et la gouvernance (ESG)

Les réfrigérants utilisés pour le refroidissement par air font l'objet d'un examen réglementaire et, parallèlement, les organisations s'efforcent de plus en plus d'améliorer leur profil ESG en réduisant les gaspillages et leur empreinte carbone. Des centres de données plus efficaces sont indispensables pour les entreprises qui cherchent à minimiser leur impact sur l'environnement.

Le contrôle des coûts

La demande croissante de centres de données conçus pour l'IA s'accompagne d'une augmentation des coûts liés à la modernisation des systèmes d'alimentation et de refroidissement dans les installations plus anciennes, ainsi que des coûts de construction de nouvelles installations. Les organisations ont besoin de technologies qui atténuent les perturbations et les aident à mettre en place une infrastructure d'IA le plus rapidement possible tout en optimisant les coûts.

Un refroidissement émergent pour l'IA

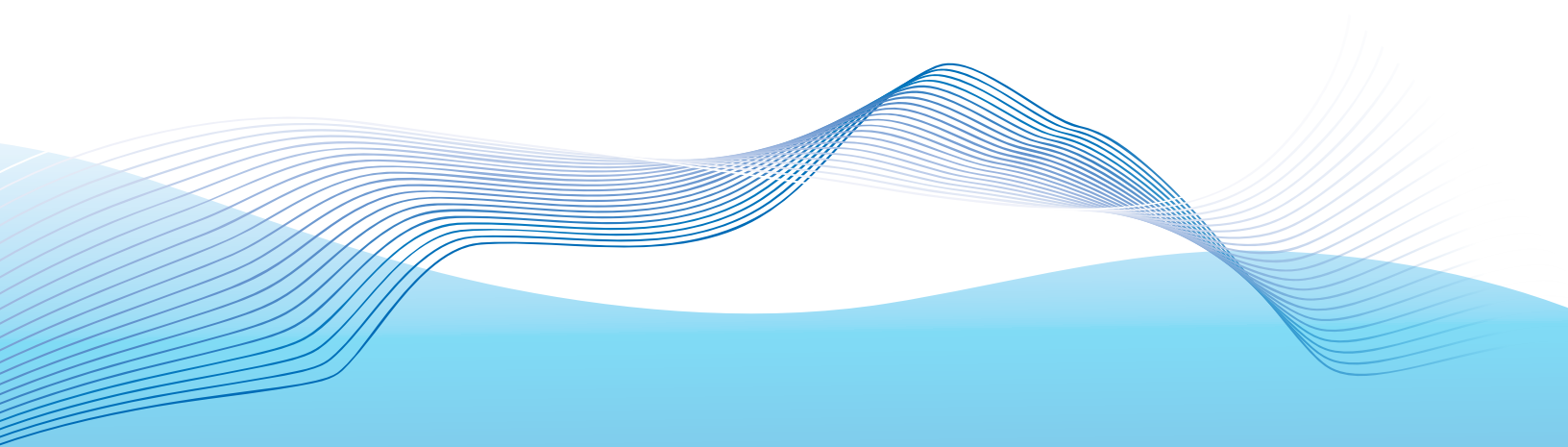
Pour toutes ces raisons, les leaders de l'industrie estiment que, en ce qui concerne l'infrastructure d'IA en particulier, [les technologies de refroidissement par air deviennent obsolètes](#), de même que les centres de données qui les utilisent. Les grandes entreprises, telles que Meta ou Amazon, ont été contraintes de [repenser radicalement la conception de leurs centres de données](#) pour demeurer à la fine pointe de l'infrastructure de l'IA.

Comment repensent-ils leur conception?

Le refroidissement par liquide. Il est indispensable à la construction et à l'utilisation d'une infrastructure d'IA à grande échelle, parce qu'il offre la juste combinaison entre la capacité de refroidissement, les avantages environnementaux et les avantages en termes de coûts. Ayant fait ses preuves depuis des décennies avec les superordinateurs, la croissance de l'IA implique que le refroidissement par liquide devienne rapidement une technologie courante qui supprime le refroidissement par air. Contrairement à l'air, le refroidissement liquide peut prendre en charge des baies de 50 kW, 100 kW et même 150 kW, à grande échelle, avec des avantages substantiels en termes d'efficacité, de coût et de durabilité. Les analyses de marché indiquent que le marché du refroidissement par liquide dans les centres de données croît à un taux de croissance annuel composé ([CAGR](#)) [de plus de 25%](#) parce qu'il est demandé pour tous les types d'infrastructures intensives des centres de données, y compris l'infrastructure de l'IA.

Les technologies de refroidissement par liquide se présentent sous différentes formes. Certaines sont utilisées conjointement avec le refroidissement par air, d'autres le remplacent. Les technologies de refroidissement par liquide occupent une place de plus en plus importante dans le secteur des centres de données, et les fournisseurs se livrent une concurrence féroce, chacun cherchant à prouver qu'il est la solution unique pour l'infrastructure des centres de données d'IA.

Dans l'ensemble, l'industrie s'accorde à dire que le refroidissement par liquide est la voie de l'avenir. Les fournisseurs, depuis [Intel](#) jusqu'à [NVIDIA](#) fabriquent des puces qui requièrent un refroidissement par liquide. Les principales entreprises de colocation telles que [Digital Realty](#) et [CyrusOne](#) déploient rapidement différentes formes de refroidissement par liquide. Il existe une convergence de consensus dans l'industrie, et il est important de comprendre les avantages et les inconvénients des différentes technologies. Examinons les trois principales technologies utilisées pour contrôler le refroidissement de l'infrastructure de l'IA.





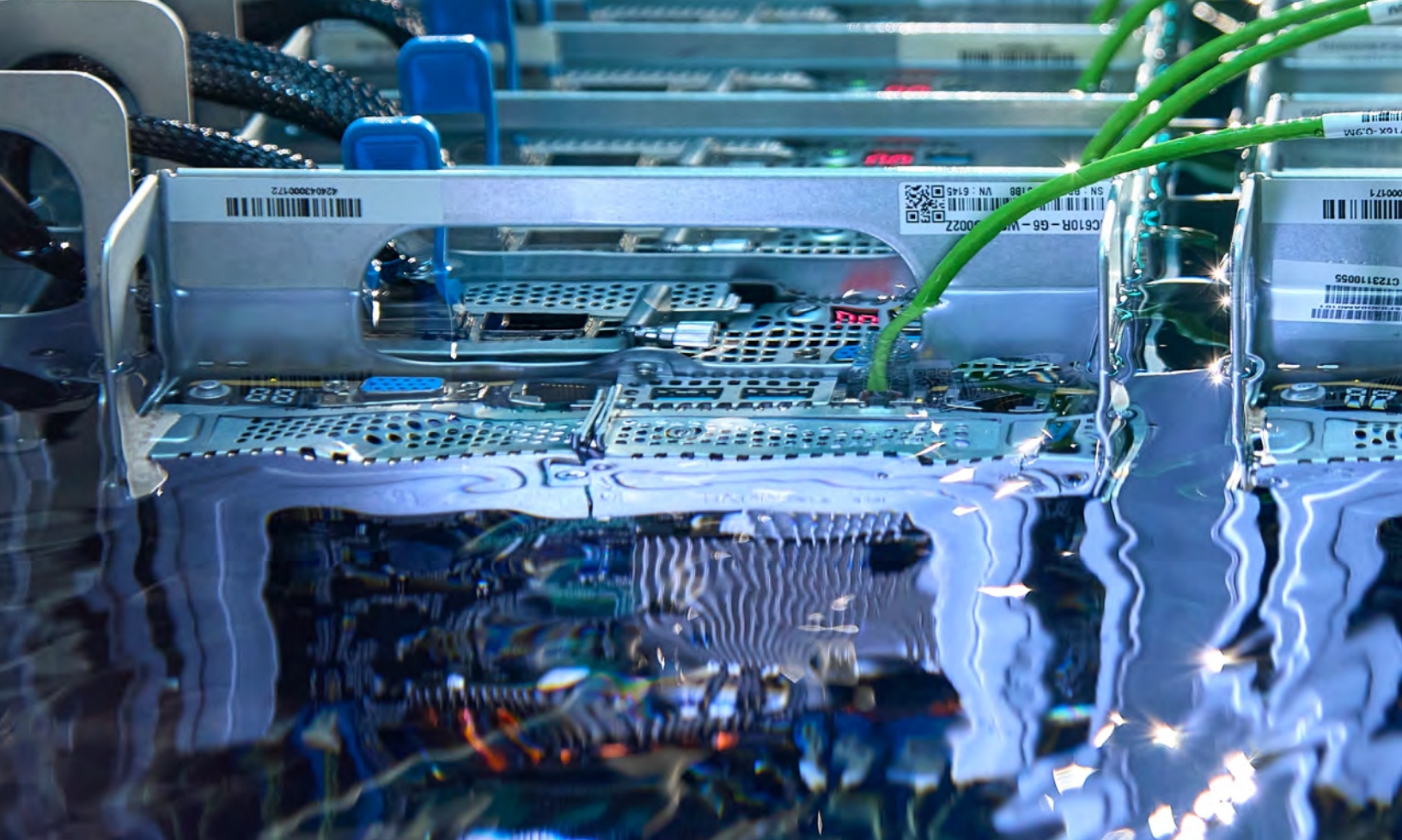
Les échangeurs de chaleur de porte arrières

Les échangeurs de chaleur de porte arrière (RDHx), installés derrière les baies de serveurs des centres de données, sont des dispositifs de refroidissement par liquide qui éliminent directement la chaleur du flux d'air d'échappement. Ils offrent plusieurs avantages, notamment une meilleure efficacité énergétique grâce à des températures d'eau réfrigérée plus élevées, ce qui permet d'éliminer les refroidisseurs. Ils peuvent prendre en charge des capacités de refroidissement plus élevées nécessaires aux charges informatiques chaudes et denses de l'IA. Ils nécessitent moins de maintenance que les unités de traitement de l'air (CRAC) et sont facilement évolutifs pour ajouter progressivement du refroidissement là où cela est nécessaire, réduisant ainsi la nécessité de procéder à des rénovations plus importantes. Cependant, les systèmes RDHx nécessitent une infrastructure de plomberie supplémentaire, présentent un rapport coût-efficacité insuffisant pour les baies de faible densité et peuvent entraîner des fuites d'air, réduisant ainsi l'efficacité, s'ils ne sont pas correctement scellés. Bien qu'ils soient bénéfiques pour les environnements à haute densité en tant que transition vers l'utilisation de liquide pour refroidir un ou deux bâtis, leur adéquation à long terme et à grande échelle dépend de facteurs tels que l'agencement du centre de données, l'infrastructure de distribution du liquide de refroidissement et les besoins futurs prévus en termes de capacité.

Direct sur puce

Ces produits font circuler un liquide de refroidissement directement vers les processeurs et les composants générateurs de chaleur dans les serveurs, ce qui permet un transfert de chaleur supérieur dans les environnements refroidis par air. En éliminant la chaleur à la source par le biais de plaques froides ou de micro-canaux, cette technologie permet d'augmenter la densité et la fréquence des processeurs sans étranglement thermique, de réduire la consommation d'énergie de refroidissement, de réduire la densité d'encombrement des serveurs et d'allonger la durée de vie des composants. Toutefois, le refroidissement direct sur la puce implique certaines complexités et certains risques.

Les serveurs peuvent être endommagés par des fuites. Le refroidissement de plusieurs CPU, puces mémoire et GPU dans un seul châssis nécessite des tuyaux de distribution de liquide de refroidissement complexes vers tous les composants, entraînant une augmentation des points de défaillance. De nombreuses solutions exigent une qualité d'eau très élevée. Enfin, tout produit directement intégré à la puce nécessite des dépenses futures en ingénierie de R-D qui devront être imputées aux clients, étant donné que chaque nouvelle puce nécessite la conception d'une plaque de refroidissement unique. Au fur et à mesure que l'utilisation du refroidissement direct sur puce passe de quelques dizaines de serveurs à des centaines ou des milliers de serveurs dans un centre de données, les risques augmentent de manière substantielle.



Le refroidissement par immersion

Enfin, nous en arrivons au refroidissement par immersion. Bien qu'il s'agisse d'une solution pour les équipements informatiques, notamment les superordinateurs, elle devient rapidement une référence en raison de sa valeur inhérente. Le refroidissement par immersion. Contrairement à d'autres techniques de refroidissement par air. L'immersion, monophasée ou biphasée, entoure toute l'infrastructure informatique de manière permanente avec chaque composant. Les centres de données refroidis par immersion sont généralement fabriqués à partir de matériaux résistants à la corrosion comme l'acier inoxydable.

1

Les réservoirs sont remplis de fluide diélectrique et l'équipement informatique est immergé dans le fluide.

2

Les réservoirs sont équipés d'un système de refroidissement en boucle fermée qui pompe le liquide de refroidissement chaud vers un échangeur de chaleur, qui peut dissiper la chaleur dans l'air ou dans l'eau.



ution de niche, utilisée dans les environne-
infrastructure en raison de sa simplicité et
technologies, il ne repose pas du tout sur le
structure informatique et reste en contact
sont construits avec des réservoirs, généra-
ble ou des polymères spécialisés.

3

Les réservoirs disposent souvent de systèmes de surveillance permettant de suivre des paramètres tels que la température du liquide, les débits et les performances des composants.

À première vue, le refroidissement par immersion est assurément différent des autres technologies.

En quoi est-il pertinent pour l'infrastructure d'IA et offre-t-il des avantages par rapport à d'autres systèmes de refroidissement liquide?

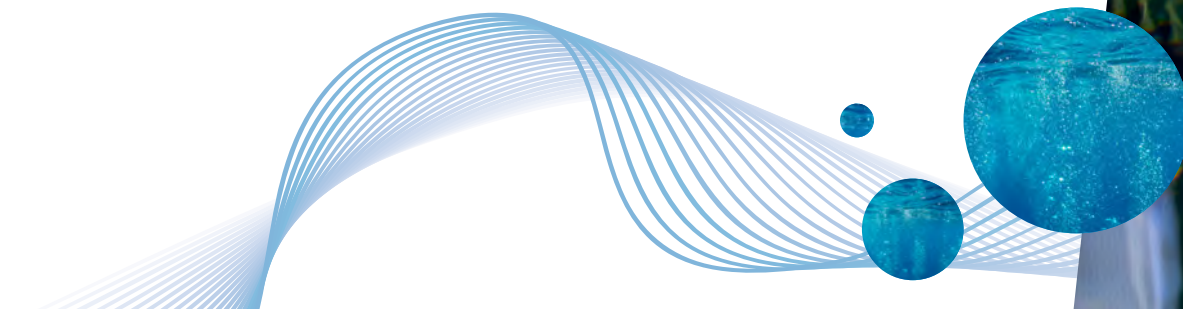
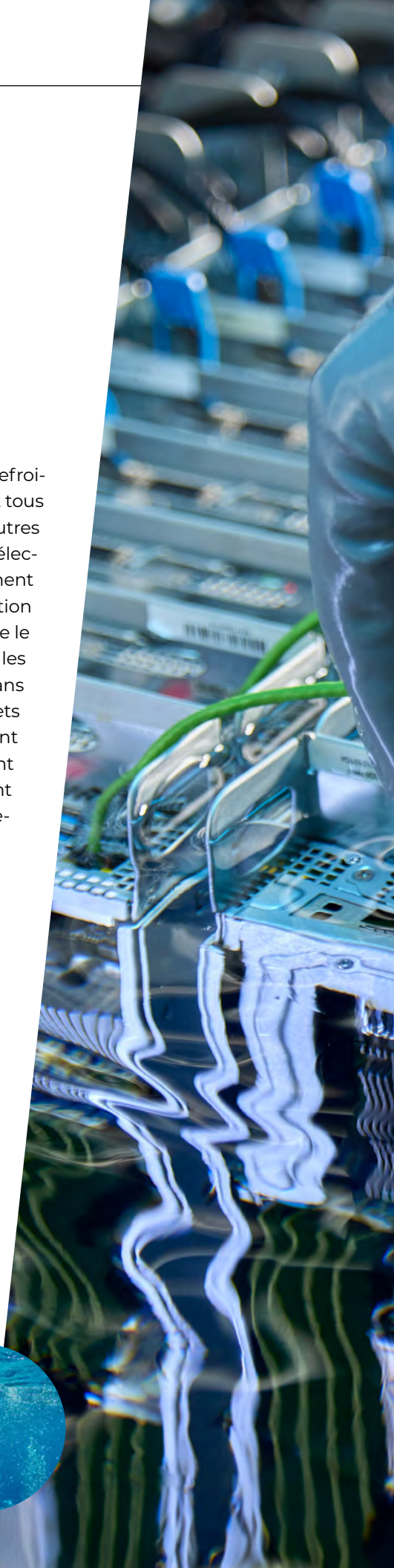
Une plongée en profondeur dans le refroidissement par immersion

Contrairement aux autres technologies de refroidissement par liquide, le refroidissement par immersion pour les serveurs permet d'immerger directement tous les composants, y compris les CPU, les GPU, les modules de mémoire et les autres composants générateurs de chaleur, dans un liquide de refroidissement diélectrique. Ce contact direct entre les composants et le liquide de refroidissement permet une évacuation très efficace de la chaleur, constituant une amélioration significative par rapport à d'autres méthodes de refroidissement car il élimine le besoin de mécanismes intermédiaires de transfert de chaleur, notamment les dissipateurs thermiques, les échangeurs de chaleur et les ventilateurs, qui, dans le cas d'infrastructures chaudes et denses, peuvent contribuer à des goulets d'étranglement thermiques et à des inefficacités. Le liquide de refroidissement utilisé dans les environnements refroidis par immersion élimine efficacement la chaleur des composants les plus denses et les plus puissants, permettant des densités de calcul plus élevées et des fréquences d'horloge potentiellement plus élevées sans risque de surchauffe ou d'étranglement thermique.

Le refroidissement par immersion offre plusieurs autres avantages essentiels.

Il permet une meilleure utilisation de la capacité énergétique

Le refroidissement par immersion offre une efficacité énergétique supérieure. En immergeant directement les serveurs dans un liquide diélectrique, la chaleur est dissipée plus efficacement. Cela se traduit par une meilleure utilisation de la capacité énergétique disponible, car moins d'énergie est gaspillée pour le refroidissement. L'immersion permet aux centres de données de réduire leur consommation d'énergie pour la même capacité de calcul, ou d'augmenter la capacité de calcul dans une enveloppe énergétique limitée. Ces gains d'efficacité maximisent le rendement des dépenses énergétiques.





Il élimine l'espace vide entre les serveurs

Les centres de données traditionnels refroidis par air souffrent souvent d'une utilisation inefficace de l'espace. De nombreuses installations ne peuvent pas prendre en charge l'infrastructure de calcul de l'IA à grande échelle en raison des limitations de refroidissement, ce qui les oblige à répartir les serveurs sur de nombreux bâtis, laissant ainsi des espaces vides. Les centres de données conventionnels gaspillent également de l'espace en raison des allées chaudes et froides nécessaires et des dégagements pour la circulation de l'air. Le refroidissement par immersion évite ces limitations et ces contraintes, en permettant de regrouper les serveurs de manière plus dense dans le même espace. Cette densité plus élevée se traduit par une puissance de calcul accrue par mètre carré, ce qui en fait une solution idéale pour les charges de travail de l'IA qui nécessitent des ressources de calcul massives ou dans des situations où l'espace est limité. En maximisant l'utilisation de l'espace disponible, le refroidissement par immersion permet également aux centres de données de moduler leurs capacités avec plus d'efficacité, réduisant ainsi le besoin de recourir à des agrandissements coûteux ou à de nouvelles installations.

Il élimine totalement le bruit des ventilateurs et des souffleries

Les centres de données conventionnels refroidis par air ont la réputation d'être bruyants en raison des ventilateurs et des souffleries nécessaires au refroidissement. Il est fréquent que les niveaux sonores des centres de données dépassent les 90 décibels (dBa). Ce bruit ne crée pas seulement un environnement de travail désagréable, mais contribue également à augmenter le niveau de stress et peut, potentiellement causer des dommages auditifs au personnel. Le refroidissement par immersion, en revanche, se déroule de manière silencieuse, car le liquide de refroidissement dissipe la chaleur sans qu'il soit nécessaire de recourir à un mouvement d'air mécanique. Ce fonctionnement plus silencieux améliore non seulement les conditions de travail du personnel du centre de données, mais réduit également le risque de problèmes liés au bruit qui pourraient avoir une incidence sur les équipements sensibles ou les espaces voisins.

Il réduit les taux de défaillance des composants

Les paramètres environnementaux sont importants dans les centres de données. Les températures élevées, l'humidité et les particules peuvent contribuer à augmenter le nombre de défaillances des composants. Le refroidissement par immersion atténue ces risques en fournissant aux serveurs un environnement contrôlé et isolé, en les protégeant des contaminants externes et en réduisant le stress thermique. En maintenant des températures optimales pour leur fonctionnement et en éliminant l'exposition à l'humidité et aux particules, le refroidissement par immersion peut réduire de manière significative les taux de défaillance des composants, ce qui se traduit par une amélioration de la fiabilité et du temps de fonctionnement.

Il augmente la durée de vie du matériel

Une combinaison des contraintes thermiques réduites et d'un environnement contrôlé dans les systèmes de refroidissement par immersion peut prolonger la durée de vie utile des composants matériels. En maintenant des températures optimales pour leur fonctionnement et en minimisant l'exposition aux contaminants, le refroidissement par immersion peut ralentir la dégradation des composants, leur permettant de fonctionner en toute fiabilité pendant de plus longues périodes. Cette durée de vie prolongée se traduit par une réduction des taux de rafraîchissement, diminuant ainsi le coût total de possession (TCO) et minimisant l'impact environnemental associé aux remplacements fréquents de matériel.

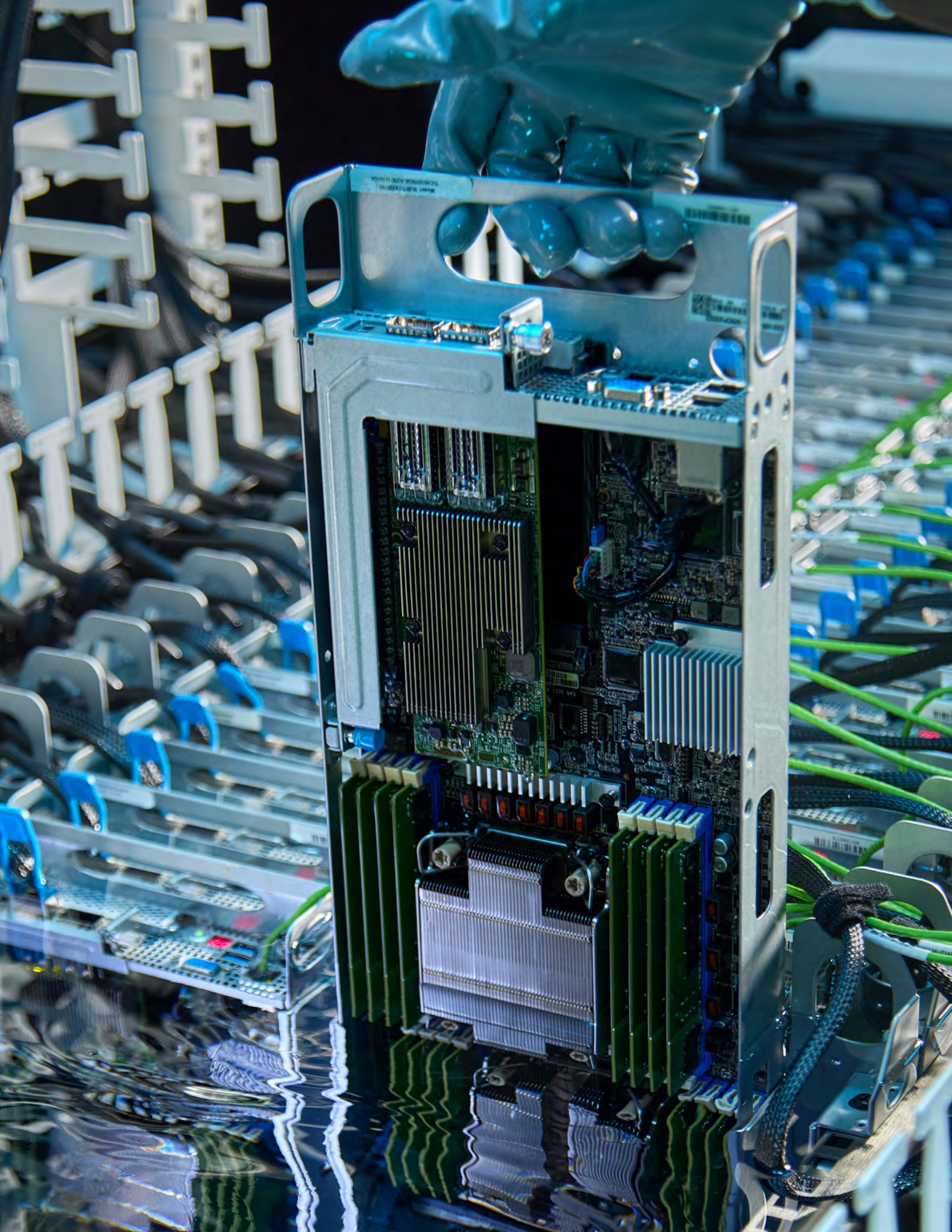
Il prend en charge plusieurs générations de processeurs sans nécessiter de modifications ultérieures

À mesure que les charges de travail de l'IA continuent d'évoluer et exigent une puissance de calcul toujours plus grande, le refroidissement par immersion offre une solution à l'épreuve du temps. Contrairement aux systèmes de refroidissement par air traditionnels ou aux technologies de refroidissement par liquide telles que D2C, pouvant nécessiter d'importants ajustements ou remplacements pour s'adapter aux nouvelles générations de processeurs, les systèmes de refroidissement par immersion peuvent s'adapter plus facilement à l'évolution des exigences matérielles. Le liquide de refroidissement peut dissiper efficacement la chaleur provenant d'une vaste gamme d'architectures de processeurs et d'enveloppes d'alimentation, et permet aux centres de données d'intégrer de manière transparente du nouveau matériel sans devoir procéder à des révisions coûteuses de l'infrastructure.

Il permet à l'infrastructure de l'IA de vivre en dehors du centre de données

Contrairement aux environnements matériels refroidis par air et par liquide qui nécessitent une infrastructure massive comprenant des refroidisseurs et une distribution de liquide de refroidissement, le refroidissement par immersion permet aux organisations de réorganiser l'emplacement de leur infrastructure. Il est désormais courant pour les organisations de déployer une nouvelle infrastructure d'intelligence artificielle à l'aide de **conteneurs mobiles et modulaires** refroidis par immersion. D'autres sont en mesure de déployer la capacité de leur centre de données refroidi par immersion dans toutes sortes d'endroits, des armoires réseau aux entrepôts, sans se soucier des contraintes fonctionnelles, environnementales et opérationnelles posées par le refroidissement conventionnel. Ces déploiements refroidis par immersion, qu'ils aient lieu dans des climats défavorables, à la périphérie d'une grande ville ou sur le stationnement d'une installation de colocation, offrent aux entreprises technologiques un nouveau moyen de mettre rapidement en œuvre une infrastructure d'IA sans le coût et la complexité d'une modernisation de l'espace existant.

En résumé, la faculté qu'a le refroidissement par immersion de prendre en charge des densités plus élevées, d'améliorer l'efficacité du refroidissement, d'allonger la durée de vie du matériel, de réduire le bruit, d'assurer la pérennité et d'offrir des options de site en fait une solution irréfutable pour construire l'infrastructure haute performance et évolutive requise pour les charges de travail de l'IA. Les leaders du marché sont d'accord. [Intel](#), [Microsoft](#), et d'autres organisations telles que l'[Open Compute Project](#) ont validé le refroidissement par immersion pour l'infrastructure de l'IA.



Le retour sur investissement (ROI) du refroidissement par immersion

Mais offre-t-il un retour sur investissement mesurable?

Il le fait, de quatre manières significatives.

01

Le refroidissement par immersion se compare favorablement aux autres technologies de refroidissement liquide

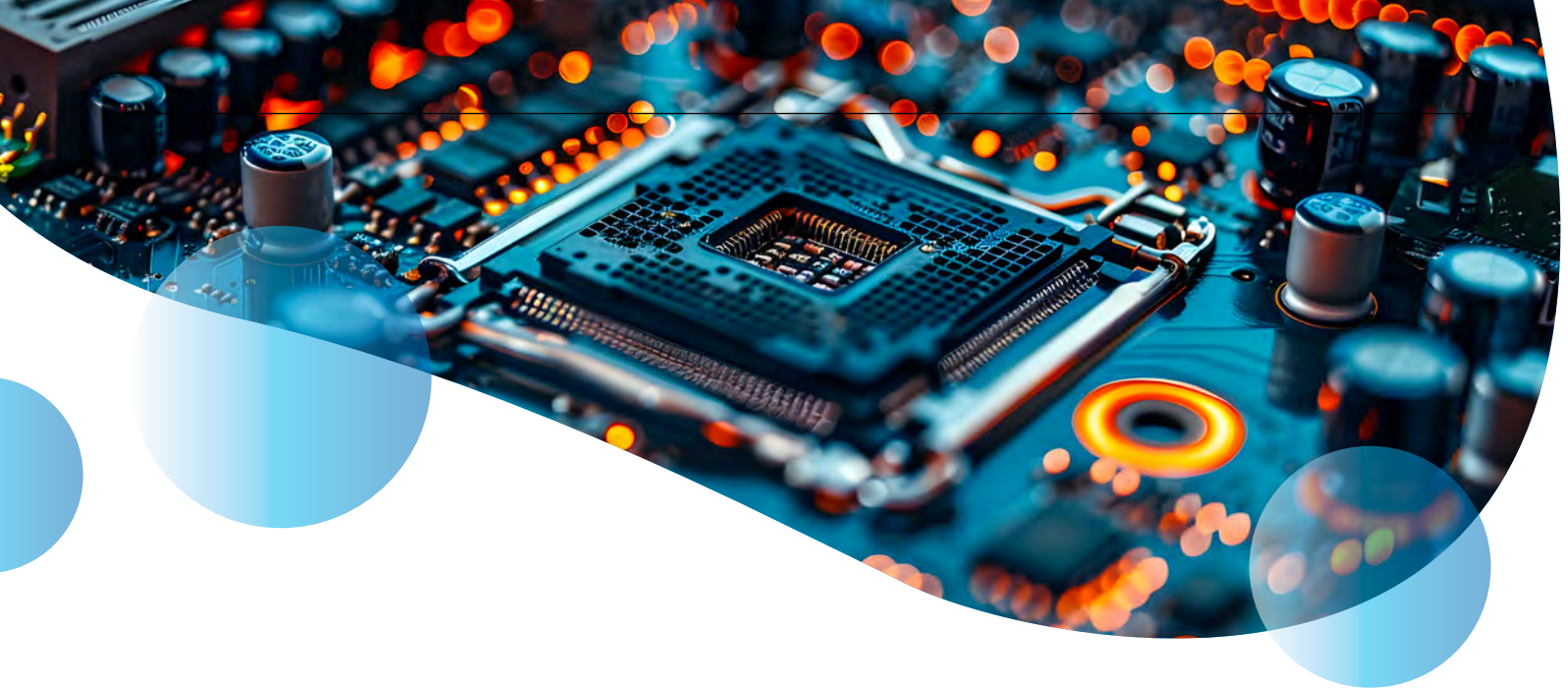
Si le refroidissement D2C offre une excellente gestion thermique pour l'informatique de haute performance, il s'accompagne de coûts initiaux et continus considérables. La nature complexe des systèmes D2C nécessite une infrastructure spécialisée et une maintenance continue. Les générations suivantes de matériel serveur s'accompagnent de dépenses supplémentaires liées aux composants de refroidissement direct sur puce. Le refroidissement par immersion, en revanche, utilise des composants plus simples et facilement disponibles sans ajouter de coûts aux serveurs - en fait, il peut réduire les coûts des serveurs en éliminant les composants inutiles tels que les ventilateurs et les boîtiers de châssis.

02

Le refroidissement par immersion réduit les coûts d'exploitation actuels et futurs

Les coûts d'exploitation représentent une dépense substantielle pour tout opérateur de centre de données, et pour beaucoup d'entre eux, les coûts sont déjà hors de contrôle en raison de la demande de nouvelles infrastructures et des impacts macroéconomiques tels que l'inflation. Le refroidissement par immersion permet d'atténuer l'augmentation des coûts d'exploitation :

- a.** Le refroidissement par immersion élimine le besoin de centrales de traitement d'air (AHU) coûteuses et énergivores, de réfrigérants onéreux et fortement réglementés, ainsi que de réseaux de gaines complexes nécessitant une maintenance.
- b.** En revanche, en transférant directement la chaleur des serveurs au liquide de refroidissement, il réduit la complexité de l'infrastructure de refroidissement et permet d'économiser de l'énergie. La réduction des factures d'électricité et la diminution des points de défaillance permettent de réduire les coûts d'exploitation.
- c.** De plus, en maintenant un refroidissement plus constant sur les serveurs, le refroidissement par immersion réduit le stress thermique et l'usure. Cela se traduit par une durée de vie plus longue pour le matériel d'IA coûteux, réduisant ainsi les coûts de remplacement.



- d. Pour couronner le tout, les serveurs immergés dans le fluide diélectrique sont protégés de la poussière et de l'humidité, responsables de la plupart des pannes matérielles. Cela se traduit par une réduction des temps d'arrêt pour la maintenance et les réparations, ce qui améliore la rentabilité des opérations.
- e. Enfin, le refroidissement par immersion permet de regrouper plus étroitement les serveurs grâce à sa capacité supérieure d'évacuation de la chaleur. Cela se traduit par une maximisation de l'espace au sol du centre de données destiné aux charges de travail d'IA, ouvrant des possibilités de sélection de sites alternatifs, de réduction des coûts immobiliers et d'amélioration de la valeur au mètre carré.

03

Le refroidissement par immersion offre une valeur à long terme

Le refroidissement par immersion offre une solution supérieure pour une valeur à long terme. Contrairement à la conception qui doit être revue pour chaque nouveau processeur, le refroidissement par immersion fonctionne avec tous les types de processeurs et d'autres composants. Il offre aux opérateurs de centres de données une solution à l'épreuve du temps pouvant prendre en charge de nombreuses générations de serveurs, même les plus chauds.

04

Le refroidissement par immersion est respectueux de l'environnement

Les installations veulent améliorer leur profil de durabilité, pour des raisons éthiques et économiques. La réduction de l'espace, la réduction de l'utilisation de l'énergie et la réduction des coûts des matériaux dans les environnements refroidis par immersion se traduisent par des opérations à l'empreinte carbone réduite. Le refroidissement par immersion peut offrir un indicateur d'efficacité énergétique (PUE) aussi bas que 1,02, et le refroidissement par immersion monophasé peut utiliser des [fluides de refroidissement renouvelables et respectueux](#) de l'environnement qui sont écologiques et sécuritaires, réduisant ainsi les coûts futurs d'assainissement de l'environnement. Ces avantages permettent aux exploitants de centres de données d'éviter les coûts matériels et immatériels liés à la surveillance réglementaire.

CONCLUSION

L'IA est en train de briser les anciens paradigmes des centres de données parce que, pour le dire simplement, elle ne peut pas être réalisée à grande échelle avec les technologies de refroidissement traditionnelles.

Le refroidissement par immersion représente l'avenir du centre de données, et chez Hypertec, notre approche de la construction de l'infrastructure informatique basée sur l'immersion signifie que nous sommes à l'avant-garde de ce changement.





Pourquoi Hypertec est-il devenu un spécialiste de l'immersion ?

Il y a plusieurs années, avant même la croissance explosive de la demande en matière d'IA, nous avons constaté que les processeurs devenaient **plus chauds que jamais**, que les pressions exercées par les organismes de réglementation et les initiatives ESG rendaient plus difficile et plus coûteuse l'utilisation **à grande échelle du refroidissement par air à base de réfrigérant**, et que les entreprises souhaitaient des solutions de refroidissement offrant une **performance, une résilience, une durabilité et une valeur supérieures**.

L'immersion était la solution logique à tous ces défis. Ensuite, OpenAI et d'autres ont appuyé sur l'accélérateur de la croissance d'infrastructures chaudes et denses qu'il n'était tout simplement pas possible de refroidir à l'air.

Chez Hypertec, nous sommes les premiers et les seuls à utiliser le refroidissement par immersion. Toute notre R-D est consacrée à l'amélioration de la performance, de la fiabilité et de la valeur des serveurs conçus pour l'immersion. Ayant récemment lancé notre **6ième génération**, nous avons une grande expérience dans la construction de matériel exclusif pour les environnements d'immersion. Cette feuille de route s'est traduite par des déploiements chez les clients, dans un large éventail d'environnements et de cas d'utilisation, qui offrent une valeur que d'autres ne peuvent pas égaler.

Le passage au refroidissement par immersion se traduit par des économies d'énergie de plus de 50 %, ce qui réduit considérablement la consommation d'énergie et l'empreinte carbone. En outre, le refroidissement par immersion réduit considérablement l'espace nécessaire; deux réservoirs de refroidissement par immersion peuvent remplacer jusqu'à 20 bâtis de serveurs traditionnels. Cela se traduit par une réduction de 60 % de l'espace physique et une diminution significative des coûts opérationnels.

En outre, les économies d'espace substantielles réalisées grâce au refroidissement par immersion libèrent de l'espace au sol précieux pour des infrastructures supplémentaires ou d'autres opérations critiques. La réduction de l'espace et des besoins en énergie améliore l'efficacité globale, permettant aux centres de données de gérer une puissance de calcul cinq fois supérieure sans avoir besoin d'unités CRAC supplémentaires, car la charge thermique de l'air est éliminée.

En outre, les GPU ne fonctionnent pas à leur pleine capacité en raison d'une surchauffe, souvent causée par une circulation d'air inadéquate dans le boîtier de l'ordinateur ou par des ventilateurs défectueux. Sans une ventilation adéquate, le GPU ne peut pas se refroidir efficacement. En revanche, le refroidissement par immersion absorbe la chaleur plus efficacement que l'air. Il élimine les points chauds, réduit la résistance thermique entre le GPU et le système de refroidissement et assure une distribution uniforme de la température. Par conséquent, les GPU fonctionnent à 100 % de leur capacité, offrant des performances remarquables et ininterrompues.

À l'avenir, l'évolution du silicium de nouvelle génération posera des défis de plus en plus importants aux responsables des centres de données. Étant donné que ces processeurs avancés exigent des densités de puissance plus élevées et génèrent plus de chaleur, les méthodes de refroidissement traditionnelles deviendront de plus en plus inadéquates. Il est urgent d'explorer et d'adopter des solutions de refroidissement innovantes, comme le refroidissement par immersion, pour relever efficacement ces défis. En envisageant et en mettant en œuvre ces options dès maintenant, les centres de données peuvent avoir une longueur d'avance et s'assurer qu'ils sont bien préparés à faire face aux exigences croissantes des technologies futures.

En résumé, le refroidissement par immersion se présente comme une solution transformatrice dans le paysage des centres de données. En offrant une efficacité énergétique, une optimisation de l'espace et des performances matérielles, il ouvre la voie à des centres de données prêts pour l'avenir. L'adoption de cette technologie permettra non seulement d'augmenter les capacités de calcul, mais aussi de contribuer de manière significative aux efforts de développement durable.

Pour en apprendre plus concernant tout ce qui différencie Hypertec, visitez notre site hypertec.com.



Références

Donnellan, D. (n.d.). *Executive summary: Uptime Institute Global Data Center Survey 2023*.

Future of Data Center cooling. (n.d.). Digital Realty. Retrieved June 4, 2024, from <https://www.digitalrealty.com/resources/articles/future-of-data-center-cooling>

Intel dives into the future of cooling. (n.d.). Intel. Retrieved June 4, 2024, from <https://www.intel.com/content/www/us/en/newsroom/news/intel-dives-into-future-of-cooling.html>

Miller, R. (n.d.). *Meta Plans Shift to Liquid Cooling for its Data Center Infrastructure*. Data Center Frontier. Retrieved June 4, 2024, from <https://www.datacenterfrontier.com/cooling/article/11436915/meta-plans-shift-to-liquid-cooling-for-its-data-center-infrastructure>

Moss, S. (n.d.). *Nvidia's CEO confirms upcoming system will be liquid cooled*. DCD. Retrieved June 4, 2024, from <https://www.datacenterdynamics.com/en/news/nvidias-ceo-confirms-next-dgx-will-be-liquid-cooled/>

NVIDIA DGX B200 datasheet. (n.d.). NVIDIA. Retrieved June 4, 2024, from <https://resources.nvidia.com/en-us-dgx-systems/dgx-b200-datasheet>

Patrizio, A. (2023, November 16). *High CPU temps are here to stay*. Network World. <https://www.networkworld.com/article/1247693/high-cpu-temps-are-here-to-stay.html>

Products. (n.d.). Hypertec. Retrieved June 4, 2024, from <https://hypertec.com/immersion-cooling/>

Project, O. C. (2023). *Investigation of immersion cooling for data center hardware: Progress, opportunities and challenges* [Video]. In YouTube. <https://www.youtube.com/watch?v=M7y1c2PQLlw>

Thompson, S. (2024, May 19). *Intel's new Falcon shores GPU to draw massive 1500W power*. eTeknix. <https://www.eteknix.com/intels-new-falcon-shores-gpu-to-draw-massive-1500w-power/>

Ware, A. (2024, February 19). *Intel advances liquid cooling solutions for datacenters, saving water and energy*. HPCwire. <https://www.hpcwire.com/2024/02/19/intel-advances-liquid-cooling-solutions-for-datacenters-saving-water-and-energy/>

Why data centers could hit obsolescence sooner than you think. (2024, March 20). <https://www.us.jll.com/en/trends-and-insights/investor/why-data-centers-could-hit-obsolescence-sooner-than-you-think>

Zissou: New datacenter, server, and software architectures for immersion-cooled systems. (2021, June 28). Microsoft Research. <https://www.microsoft.com/en-us/research/project/zissou-new-datacenter-server-and-software-architectures-for-immersion-cooled-systems/>

(N.d.). Retrieved June 4, 2024, from <https://www.cyrusone.com/resources/blogs/latest-liquid-cooling-tech-makes-solid-impact-at-cyrusones-lebanon-facility>

CE. (n.d.). Retrieved June 4, 2024, from <https://www.carrierenterprise.com/hvac-news/what-to-know-about-the-r410-refrigerant-phase-out>

Catherine. (2023, June 29). *Hypertec Launches TRIDENT iC Series 6th Gen Immersion-Born Servers*. Hypertec. <https://hypertec.com/hypertec-launches-new-ciara-trident-ic-series-immersion-born-servers/>

Akshat V., November 18, 2023, *Fix Graphics Card Overheating* [High GPU, VRAM, VRM Temperature], <https://graphicscardhub.com/graphics-card-overheating/>



Dossier spécial sur le refroidissement par immersion



Contactez-nous



+ 1 514.745.4540



hypertec.com/immersion-cooling/



9300, Route Transcanadienne
St-Laurent, Québec, H4S 1K5
Canada