Hewlett Packard Enterprise

Brochure

ARMOIRE HPE CRAY EX REFROIDIE À L'EAU POUR LES SYSTÈMES À GRANDE ÉCHELLE

Check if the document is available in the language of your choice.

Ces 20 dernières années, la plupart des solutions de calcul intensif étaient conçues de manière similaire : un grand cluster de serveurs à évolutivité horizontale, basé sur un des principaux fournisseurs de processeurs et connecté via une interconnexion aux normes de l'industrie. Cette approche a bien fonctionné, offrant les meilleures performances possibles pour relever des défis dans les domaines des sciences et de l'ingénierie, comme la météo et le climat, la dynamique des fluides numériques, la conception mécanique, la sécurité, les applications de défense, etc.

Mais la prochaine décennie apporte de nouvelles charges de travail et une gamme remarquable de solutions de traitement qui remettent en question le statu quo du calcul haute performance (HPC). La croissance incessante des données associée à une nécessité commerciale mondiale pour la transformation numérique génère des charges de travail plus nombreuses et plus volumineuses. Les charges de travail de modélisation et de simulation traditionnelles fusionnent avec l'IA, l'analytique et l'internet des objets (IoT) pour créer d'importants workflows stratégiques pour l'entreprise

Simultanément, de nombreux fournisseurs de x86, Arm®, processeurs graphiques et systèmes de portes logiques programmables (FPGA) devraient proposer des offres très attractives. Par conséquent, il devient de plus en plus difficile de prédire quelles architectures de traitement à cœur offriront la meilleure valeur. Cela signifie que les utilisateurs HPC auront besoin de systèmes capables :

1. De prendre en charge plusieurs plateformes de calcul
2. D'être mis à niveau vers de nouvelles architectures à mesure qu'elles deviennent disponibles

Ces facteurs de confluence ont imposé une refonte complète de l'architecture du calcul intensif. Ils sont également le signal d'un point d'inflexion technologique important et d'une nouvelle ère pour le HPC. Plus qu'une question de vitesse, l'ère de l'exascale représente un nouvel ensemble de fonctionnalités pour un nouvel ensemble de charges de travail qui transforme chaque domaine d'investigation.

L'architecture du calcul haute performance HPE Cray est la réponse de HPE à ces nouvelles demandes. Il s'agit d'une conception entièrement nouvelle. Le calcul, les interconnexions, le logiciel et le stockage ont été repensés et remaniés pour répondre aux prérequis actuels et futurs du système pour les charges de travail du HPC, de l'intelligence artificielle (IA) et convergées. Conçu autour des données, il permet d'exécuter simultanément des charges de travail diverses et rapides. Les innovations relatives au matériel et au logiciel limitent les goulets d'étranglement du système en termes de traitement, de mouvement de données et d'I/O. Elles permettent d'éliminer la distinction entre les clusters et les superordinateurs et offrent une interconnexion enrichie de logiciels et de systèmes dans différents formats. Différentes architectures de processeurs et d'accélérateurs ainsi qu'un choix de technologies d'interconnexion de systèmes, y compris notre nouvelle interconnexion HPE Slingshot, peuvent être utilisées.

Armoire refroidie à l'eau HPE Cray EX

Pour les clients qui exigent le meilleur en termes de performances, de densité et d'efficacité pour les systèmes à grande échelle, le superordinateur HPE Cray est disponible sous forme d'une armoire refroidie à l'eau qui prend en charge le refroidissement direct à l'eau de tous les composants dans une configuration lame compacte.

Cette architecture d'armoire HPE Cray EX comporte de nombreuses fonctionnalités innovantes qui prennent en charge les processeurs et processeurs graphiques à forte consommation énergétique (plus de 500 W). Cela permet de réduire considérablement les prérequis en matière de câblage d'interconnexion et les dépenses opérationnelles. Cette infrastructure refroidie à l'eau offre également une architecture système beaucoup plus compacte et permet de réduire l'utilisation de câbles d'interconnexion optiques au profit de câbles électriques moins coûteux.

De plus, l'infrastructure HPE Cray EX a été conçue avec soin pour prendre en charge de nombreuses architectures de processeur et options d'accélérateurs tout en restant compatible avec les prochaines générations de processeurs, de processeurs graphiques et de technologies d'interconnexion pour au moins dix ans.

L'architecture en lames pour le calcul et le réseau est essentielle à la flexibilité de l'armoire HPE Cray EX. Elle permet de combiner différentes technologies de processeurs et de processeurs graphiques, mais aussi d'offrir un chemin de mise à jour simple vers les processeurs et fonctionnalités d'interconnexion de prochaines générations. Cruciale pour de nombreux utilisateurs, la compatibilité physique, logicielle et du réseau des différentes lames de processeurs et de processeurs graphiques permet de retarder les décisions en matière de choix de plateforme de calcul, puisqu'elles sont toutes conçues pour se connecter à la même infrastructure.

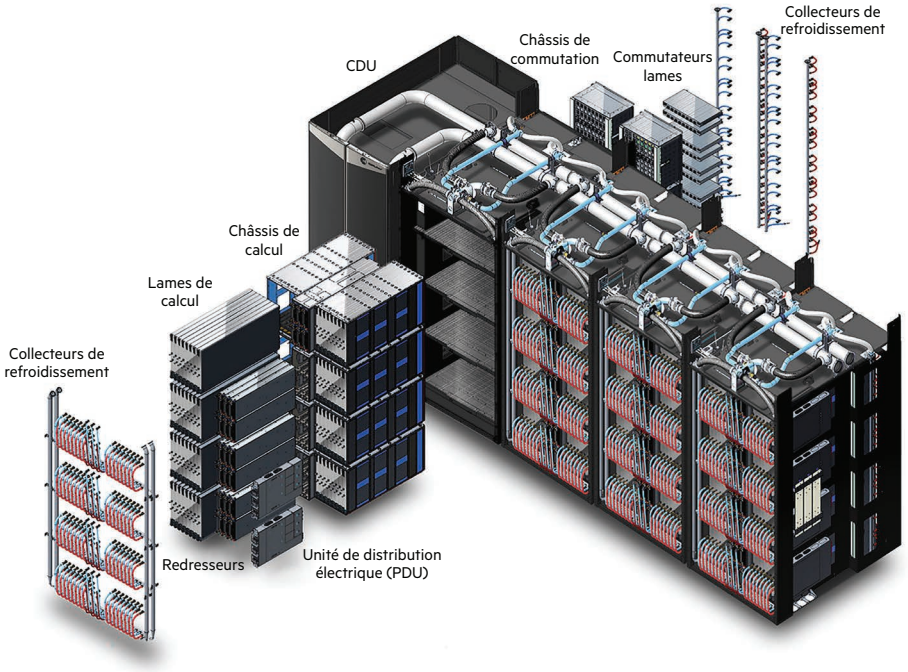

FIGURE 1. Vue éclatée de l'armoire HPE Cray EX

Architecture de l'armoire HPE Cray EX

Les modules de base de l'architecture de calcul et de commutation de l'armoire refroidie à l'eau sont les suivants :

- Châssis de calcul : Le châssis de calcul est un assemblage mécanique abritant jusqu'à huit lames de calcul. Chaque armoire HPE Cray EX contient huit châssis de calcul, permettant d'obtenir jusqu'à 64 lames de calcul et jusqu'à 512 processeurs par armoire. Les lames de calcul sont alignées à la verticale et insérées dans le châssis de calcul par l'avant de l'armoire.
- Lame de calcul : La lame de calcul est un module du châssis de calcul qui englobe les éléments de calcul HPE Cray EX (processeurs, connexions du fabric, cartes à circuits imprimés et composants de refroidissement et d'alimentation). La lame de calcul initiale contient quatre nœuds AMD EPYC 7002 à double socket.
- Châssis de commutation : Le châssis de commutation est un assemblage mécanique qui abrite jusqu'à huit lames de commutateur d'interconnexion HPE Slingshot. Chaque armoire HPE Cray EX contient huit châssis de commutation HPE Slingshot, permettant d'obtenir jusqu'à 64 commutateurs par armoire. Les lames de commutateur sont alignées à l'horizontale et insérées dans le châssis de commutation par l'arrière de l'armoire.
- Lame de commutateur : La lame de commutateur contient les composants en silicium du commutateur de fabric HPE Slingshot, une carte à circuit imprimé avec des connexions pour les lames de calcul et tous les composants de refroidissement et d'alimentation. HPE Cray EX prend en charge jusqu'à huit lames de commutateur par châssis de commutation permettant d'obtenir jusqu'à 16 connexions du fabric par lame de calcul (deux connexions du fabric par connecteur physique).

Chaque armoire contient huit châssis de calcul et huit châssis de commutation conçus pour des connexions directes du fabric depuis les lames de commutateur vers les lames de calcul sans câblage ou panier intermédiaire. La lame de commutateur est refroidie directement à l'eau tout comme les ports réels du commutateur, ce qui permet de dissiper une quantité importante de chaleur en cas d'utilisation de câbles optiques actifs. Le châssis de calcul contient jusqu'à huit lames de calcul et le châssis de commutation contient jusqu'à huit lames de commutateur.

Le châssis de calcul et le châssis de commutation sont fixés ensemble, comme illustré à la figure 2. Les connexions directes entre les lames de commutateur et les lames de calcul passent par des connecteurs sur chaque lame, via l'espace ouvert d'une frame entre le châssis de calcul et le châssis de commutation. Étant donné que les lames de calcul sont alignées de façon verticale et que les lames de commutateur sont alignées de façon horizontale, chaque lame de commutateur peut se brancher directement sur chaque lame de calcul.

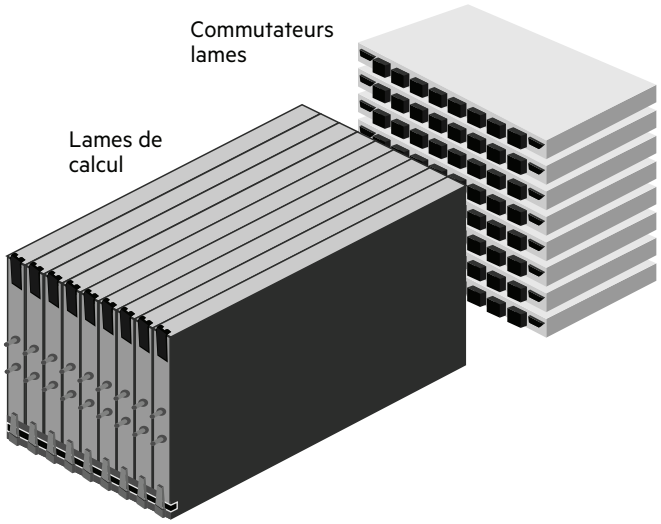

FIGURE 2. Interface de la lame de calcul et de la lame de commutateur

Détail de la lame de calcul HPE Cray EX425

La lame de calcul initiale refroidie à l'eau (HPE Cray EX425) est équipée de quatre serveurs AMD EPYC 7002 à double processeur. Parmi les futurs produits proposés, des lames de calcul basées sur d'autres architectures de processeurs et des lames avec processeurs graphiques sont prévues. Le format de ces futures lames sera compatible et elles seront dotées de fonctionnalités d'interconnexion similaires.

La figure 3 illustre la lame de calcul HPE Cray EX425. Tous les composants sont refroidis à l'eau.

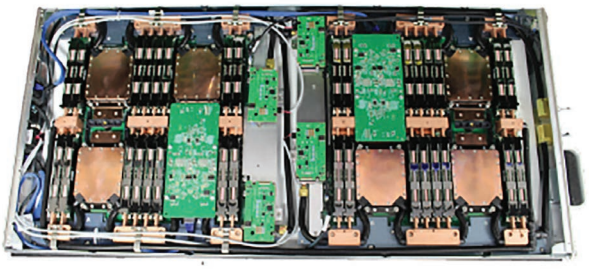

FIGURE 3. Lame de calcul HPE Cray EX425

L'armoire HPE Cray EX n'est pas équipée de ventilateur. (Voir la section Alimentation et Refroidissement pour plus d'informations.)

Détail de la lame de commutateur

La lame de commutateur contient les composants en silicium du commutateur de fabric HPE Slingshot, une carte à circuit imprimé avec des connexions pour les lames de calcul et tous les composants de refroidissement et d'alimentation. HPE Cray EX prend en charge jusqu'à huit lames de commutateur par châssis de commutation permettant d'obtenir jusqu'à huit connexions du fabric par lame de calcul.

HPE SLINGSHOT DANS UNE ARMOIRE HPE CRAY EX

L'interconnexion HPE Slingshot est un nouveau réseau de calcul haute performance, haute vitesse et spécialement conçu avec un commutateur personnalisé de 64 ports qui fournit une bande passante de 12,8 To/s.

Connexions réseau de la lame de calcul

Les nœuds de calcul des lames de calcul HPE Cray EX communiquent avec le fabric réseau via une carte mezzanine avec une interface PCIe vers les processeurs. La carte mezzanine pour la lame de calcul à quatre nœuds AMD EPYC 7002 contient les connexions HPE Slingshot pour deux nœuds. Les lames de calcul prenant en charge l'injection simple (une connexion réseau par nœud) et l'injection double (deux connexions réseau par nœud), les cartes mezzanine sont déployées en groupes de deux ou quatre par lame de calcul pour la prise en charge de l'injection simple et double. Les futures lames de calcul pourront prendre en charge davantage de ports d'injection.

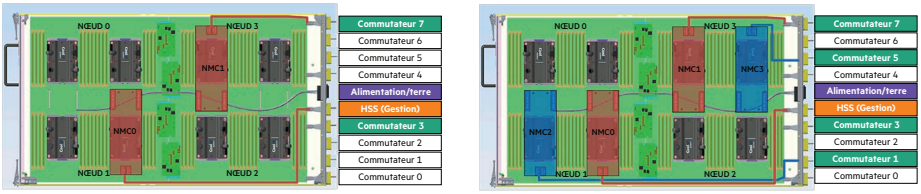

FIGURE 4. Connexions entre le nœud de calcul et la lame de commutateur

Connexion réseau de la lame de commutateur HPE Slingshot

Le commutateur refroidi à l'eau HPE Slingshot est doté de 16 connexions internes, qui font l'interface avec les lames de calcul (deux par lame de calcul) et 48 connexions externes pour les connexions commutateur à commutateur. Dans le cas de la lame de calcul à quatre nœuds AMD EPYC 7002, deux lames de commutateur sont nécessaires pour un port d'injection simple HPE Cray EX425 vers chaque nœud. (Quatre connexions vers quatre nœuds par lame de calcul ; 32 connexions vers 32 nœuds par châssis). La double injection a besoin de quatre lames de commutateur.

SCHÉMA 5. Lame de calcul HPE Cray EX425

SCHÉMA 6. Lame de commutateur HPE Slingshot

Création du réseau

Le réseau refroidi à l'eau est généralement conçu autour de groupes de 16 commutateurs avec une connexion entre chaque à l'intérieur de chaque groupe, bien que d'autres configurations soient prises en charge. Ces groupes sont ensuite associés à une topologie directe capable de s'adapter à des centaines d'armoires et à des centaines de milliers de nœuds et qui offre une communication de terminal à terminal en trois sauts pour tous les terminaux. En raison de la densité du système dans une infrastructure refroidie à l'eau, il est possible d'utiliser des câbles électriques pour toutes les communications entre commutateurs du groupe, les connexions optiques n'étant nécessaires que pour les communications groupe à groupe. Cela permet de réduire le coût de la solution réseau globale et d'améliorer la fiabilité.

Exemple de réseau refroidi à l'eau

- Groupe de 16 commutateurs
- 2 commutateurs par châssis pour l'injection simple vers 32 nœuds de calcul (8 lames de calcul)
- 16 commutateurs par armoire pour l'injection simple vers 256 nœuds de calcul (64 lames de calcul)
- Peut évoluer à des centaines d'armoires avec une connexion en 3 sauts maximum de n'importe quel terminal à n'importe quel autre

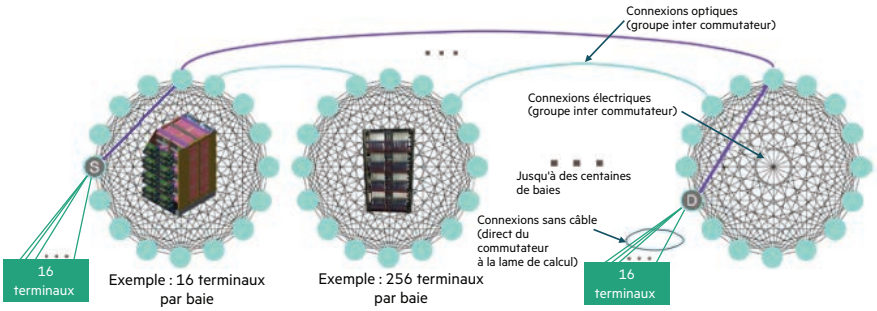

FIGURE 7. Exemple de topologie directe dans les commutateurs HPE Slingshot

ALIMENTATION ET REFROIDISSEMENT

Une armoire refroidie à l'eau pouvant prendre en charge jusqu'à 300 kW d'alimentation, un soin particulier a été porté à l'alimentation et au refroidissement du HPE Cray EX.

Alimentation : Chaque armoire est dotée d'une série d'unités de distribution électrique (PDU) et de redresseurs qui convertissent l'alimentation CA triphasée de 480 ou 400 V entrante en alimentation CC de 380 V pour la distribution aux lames de calcul et de commutateur individuelles. Une série de convertisseurs CC à CC sur les lames de calcul et de commutateur convertissent l'alimentation CC à 380 V entrante en CC 48 V d'abord puis dans les tensions CC adéquates pour les différents composants. L'armoire HPE Cray EX prend en charge une alimentation entrante par le dessus ou par le dessous.

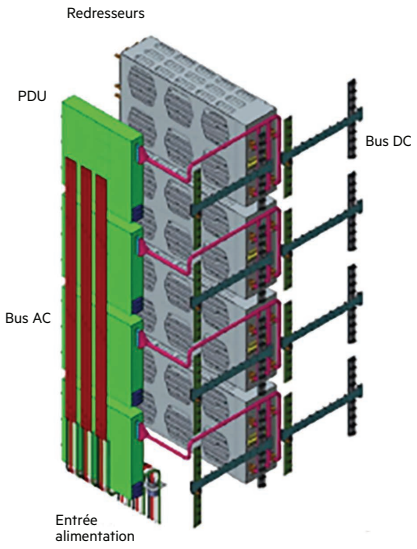

FIGURE 8. PDU refroidi à l'eau

Refroidissement : L'armoire HPE Cray EX et tous les composants sont complètement refroidis par les boucles refroidies à l'eau qui circulent dans l'infrastructure de calcul. L'unité de distribution de refroidissement (CDU) refroidit le liquide et capte la chaleur du système via un échangeur de chaleur avec l'eau du datacenter.

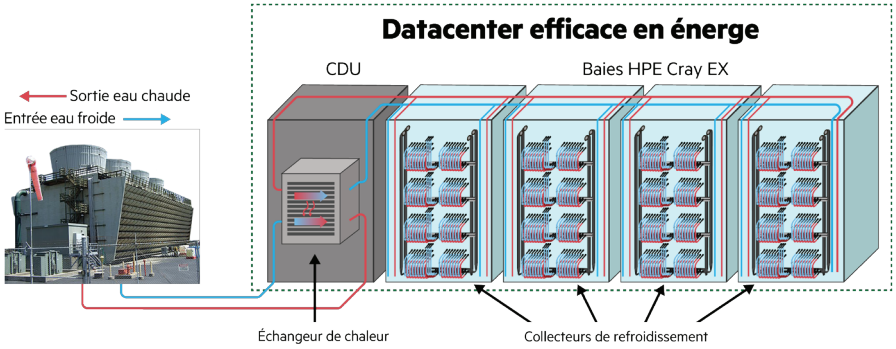

FIGURE 9. Flux de refroidissement à l'eau HPE Cray EX

La boucle de refroidissement globale est une boucle fermée dont l'origine est dans la CDU. Une CDU peut prendre en charge jusqu'à quatre armoires refroidies à l'eau. La CDU maintient le liquide de refroidissement à la température spécifiée et capte la chaleur par un mécanisme de transfert de chaleur vers l'eau du datacenter. La CDU a besoin d'une arrivée d'eau à une température inférieure à 32 ˚C, ce qui permet de s'affranchir de refroidisseur dans de nombreux environnements et de réduire davantage la consommation d'énergie. La technologie de refroidissement à l'eau tiède du datacenter qui est effectivement déployée dépend de l'environnement et varie en fonction du climat.

Le refroidissement à l'eau circule vers les lames et composants individuels des armoires refroidies à l'eau à travers une série de collecteurs qui distribuent le liquide de refroidissement depuis la conduite primaire qui va de la CDU aux lames et commutateurs individuels puis rapporte le liquide chauffé à la conduite de retour vers la CDU pour être refroidie. L'armoire est dotée de collecteurs de refroidissement à l'avant des lames de calcul et à l'arrière des lames de commutateurs. De plus, les structures de refroidissement à l'eau refroidissent les redresseurs qui sont également connectés à la conduite primaire.

Les connexions depuis et vers les lames de calcul et de commutateur sont rapides à connecter, ne gouttent pas et permettent de retirer une lame pour l'entretien sans avoir à arrêter l'ensemble du système.

Des plaques froides captent la chaleur des processeurs directement. Les cartes mezzanine NIC, lorsqu'elles sont présentes au-dessus des processeurs, sont également refroidies par les mêmes plaques froides du processeur.

Les modules de séparation prennent des fluides de la boucle de refroidissement et fournissent le même débit aux tubes capillaires qui dirigent le liquide vers le champ DIMM. La figure 10 illustre la lame de calcul avec les structures de plaque DIMM froide ainsi que les modules de séparation. Le refroidissement de la lame de commutateur est similaire.

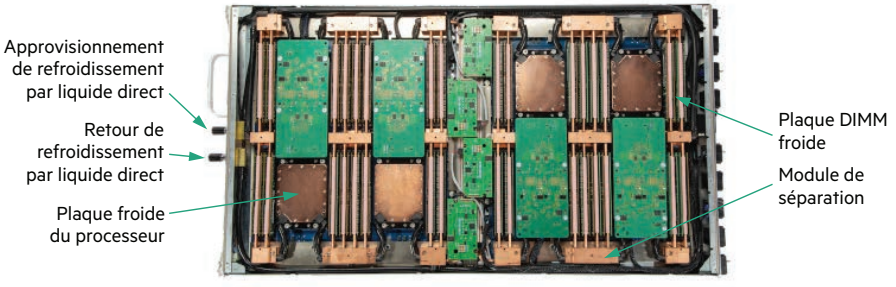

FIGURE 10. Flux de refroidissement à l'eau HPE Cray EX

Grâce à l'efficacité du refroidissement à l'eau par rapport à l'air, le budget pour l'alimentation et le refroidissement d'un superordinateur refroidi à l'eau peut être significativement inférieur à celui d'une installation de taille similaire refroidie à l'air. De plus, une armoire refroidie à l'eau pouvant accepter une eau de datacenter jusqu'à 32 °C, cette solution offre davantage de flexibilité pour le choix de la technologie de refroidissement de l'eau du datacenter (par exemple, suppression du refroidisseur).

CONCLUSION

L'architecture refroidie à l'eau HPE Cray EX offre l'expérience de calcul haute performance HPE Cray dans un format hautement intégré et flexible avec des performances élevées, une évolutivité, une efficacité et une valeur possibles pour les technologies actuelles et futures de processeurs et de processeurs graphiques.

- **Performance :** Le refroidissement à l'eau direct prend en charge les éléments de traitement de qualité connectés au réseau haut débit HPE Slingshot pour les charges de travail du HPC et de l'IA. Associée à l'environnement de programmation et d'exécution HPE Cray, l'armoire refroidie à l'eau fournit des performances élevées au niveau du nœud et du système.
- **Évolutivité :** Évoluez vers des centaines d'armoires et des centaines de milliers de nœuds. Un réseau sans câbles dans le châssis permet de réduire la quantité de câbles externes et de câbles optiques par rapport à d'autres fabricants.
- **Coût total de possession :** Économisez sur les coûts d'exploitation relatifs à l'utilisation de l'eau et de l'électricité sur toute la durée de vie du produit.
- **Flexibilité :** L'infrastructure refroidie à l'eau, flexible et pourtant hautement intégrée, offre un large choix de plateformes de calcul, des solutions réseau pouvant être mises à niveau, une compatibilité future et la possibilité de retarder les décisions sur le choix du processeur.
- **Fiabilité :** Une utilisation minimale de câbles, l'absence de pièces mobiles (ventilateurs), des blocs d'alimentation et des solutions de refroidissement très fiables qui réduisent le risque de surchauffe, contribuent à la fiabilité globale de la plateforme. Associé à HPE Cray System Management, l'armoire HPE Cray EX améliore la disponibilité du système pour les systèmes de grande taille par rapport aux solutions refroidies à l'air de taille similaire.

POUR EN SAVOIR PLUS

hpe.com/fr/fr/compute/hpc/supercomputing/cray-exascale-supercomputer.html

Hewlett Packard Enterprise

a50002389FRE, novembre 2020