N°52 / JANVIER 2023

CHOCS

REVUE SCIENTIFIQUE
ET TECHNIQUE
DE LA DIRECTION
DES APPLICATIONS
MILITAIRES

cea

Calcul haute performance
En route vers l'exascale

www-dam.cea.fr

N°52 / JANVIER 2023

CHOCS

EN COUVERTURE. **Photographie du réseau d'interconnexion BXI haute performance du calculateur Tera 1000-2 développé dans le cadre de la R&D menée en commun par le CEA – DAM et Atos.**

CHOCS
CEA – DAM
Institut supérieur des études nucléaires de défense (ISENDé)
Bruyères-le-Châtel
91297 Arpajon Cedex
Tél. : 33 (0)1 69 26 76 98
Fax : 33 (0)1 69 26 70 05

Email : chocs@cea.fr

Brochure imprimée sur papier écogéré
ISSN : 1157-741X
Dépôt légal à parution

ISENDé
INSTITUT SUPÉRIEUR DES ÉTUDES NUCLÉAIRES DE DÉFENSE

sommaire

02

Panorama et enjeux des processeurs de calcul pour l'exascale

Auteurs

P. Carribault
CEA – DAM,
centre DAM Île-de-France
Université Paris-Saclay, CEA,
Laboratoire d'informatique
haute performance pour
le calcul et la simulation
(LiHPC), Bruyères-le-Châtel

G. Colin de Verdière
CEA – DAM,
centre DAM Île-de-France

Abstract

Overview and challenges of computing processors for the exascale

The computational power of new generation supercomputers has been increasing, namely due to the continuous progress in terms of processors. The advances over the last few decades have made it possible to put forward a greater number of operations per second in order to design machines able to beat the petaflops milestone in 2009 (10^{15} operations per second). And in order to take the next step, the exascale, or scale of the exaflop (10^{18} operations per second), the integrated circuit and chip designers are currently facing two major challenges: energy efficiency and proper balance between peak performance and the performance sustained on existing parallel applications. This article presents two directions in the development of processors, helping to meet these challenges and points out the choices made at the time by the CEA – DAM designers concerning the Tera 1000 supercomputer. These choices are faced with the directions described, as well as with current and next generation machines to come all over the world.

L'augmentation de la puissance de calcul des nouvelles générations de supercalculateurs est possible notamment grâce à la constante amélioration des processeurs. L'évolution de ces derniers, depuis plusieurs décennies, a permis de proposer un plus grand nombre d'opérations effectuées par seconde pour concevoir des machines dépassant le **pétaflops** en 2009 (10^{15} opérations par seconde). Pour franchir le prochain palier de l'**exascale**, échelle de l'**exaflops** (10^{18} opérations par seconde), les concepteurs de circuits intégrés ou de puces font face à deux enjeux majeurs : l'efficacité énergétique et le compromis entre les performances crête et les performances soutenues sur les applications **parallèles** existantes. Cet article expose deux pistes d'évolution de processeurs pour répondre à ces enjeux et met en perspective les choix de conception faits par le CEA – DAM pour le supercalculateur **Tera 1000**. Ces choix sont confrontés aux pistes d'évolution décrites ainsi qu'aux machines actuelles et futures dans le monde.

Depuis plusieurs décennies, les supercalculateurs ont augmenté leur puissance de calcul dans une enveloppe énergétique en constante augmentation. Les performances de ces machines, se mesurant en capacité d'exécuter des opérations élémentaires sur des nombres réels, ou flop/s (*floating-point operations per second*), ont franchi plusieurs étapes majeures ces dernières années. Comme le montre la **figure 1**, les supercalculateurs ont atteint en 2009 la barre symbolique du pétaflops, soit 10^{15} opérations par seconde, tandis que la prochaine étape majeure est l'exascale, échelle de l'exaflops (10^{18} opérations par seconde). Cette tendance n'est possible que par l'amélioration des composants des supercalculateurs. Bien que ces derniers contiennent beaucoup d'autres éléments indispensables pour la haute performance, la partie calcul est quasi intégralement effectuée par le processeur, faisant de lui un élément crucial.

Dans le cadre du **programme Simulation**, le CEA – DAM s'appuie sur des supercalculateurs pour mener à bien ses missions. Mais, comparées aux conceptions d'origine, les structures internes et externes des dernières générations de processeurs changent radicalement ; il est donc nécessaire de comprendre leur évolution afin de choisir les types de processeurs les plus adaptés pour assurer l'exécution des codes du CEA – DAM et préparer l'avenir. Cet article présente

dans la première partie l'évolution des architectures de calcul et introduit dans la deuxième les choix de processeurs qui ont été faits pour le supercalculateur Tera 1000, construit par la société française Atos-Bull ; enfin, dans la dernière partie, il expose les pistes futures des processeurs.

Évolution des architectures de calcul

Les processeurs n'ont cessé d'évoluer en profitant des avancées technologiques en matière de conception de circuits intégrés. En effet, l'amélioration de la finesse de gravure et du procédé de leur conception a permis d'assembler plus de transistors sur une même surface sous contrainte de l'enveloppe énergétique et de la dissipation de chaleur. Cette place supplémentaire disponible peut être utilisée de différentes manières. Il existe néanmoins un compromis entre l'augmentation des performances crête des processeurs, nombre maximum d'opérations à virgule flottante par seconde, et les performances soutenues, performances réelles des applications. Tout cela doit, bien évidemment, se faire en conservant la maîtrise de l'énergie dissipée (voir **encadré 1** et **figure 2**). De plus, ce compromis a aussi un impact sur la partie logicielle du supercalculateur : est-ce que cet aspect est transparent ou non pour la **pile logicielle** et pour l'utilisateur ? Le concepteur de processeurs fait face au choix de mettre en avant l'architecture ou la microarchitecture du processeur (voir **encadré 2**). Deux pistes d'évolution se dégagent alors : se diriger vers un processeur à **cœurs** dits polyvalents en mettant en avant la microarchitecture ou bien développer un processeur à cœurs dits spécialisés en proposant une architecture étendue. Ces deux pistes sont exposées ci-après.

Piste 1 : processeur à cœurs polyvalents

La première piste consiste, lors d'un changement de génération de puces, à dédier la place disponible pour de nouveaux transistors à des mécanismes privilégiant la microarchitecture pour augmenter les performances de façon transparente. Ce faisant, il est possible d'espérer une bonne efficacité de toutes les applications, mais sans pousser les performances crête à leur maximum. Cette catégorie de processeurs regroupe, par exemple, les puces de la gamme Intel Xeon (Nehalem, Haswell, Skylake, etc.) ou AMD EPYC (Naples, Rome et Milan).

Dans cette approche, les performances du processeur de calcul peuvent croître grâce à plusieurs paramètres. Il est possible d'augmenter la fréquence du processeur : cette dernière influence directement la vitesse de travail et de traitement et donc améliore les performances, mais cette direction est limitée du fait de la consommation énergétique (la puissance dissipée par un processeur est sensiblement proportionnelle à sa fréquence et au carré de sa tension de fonctionnement 2). Il est également possible de jouer sur la gestion interne des instructions, soit en permettant de traiter plus d'opérations en même temps (notion de *superscalaire*), soit en découpant les instructions en plus petits morceaux pour exécuter plus rapidement les opérations en cascade (architecture dite en *pipeline*).

La multiplication des transistors dans la microarchitecture peut aussi influencer grandement les performances soutenues du processeur, c'est-à-dire la capacité de traitement réellement obtenue durant l'exécution d'applications parallèles. Ainsi, il existe plusieurs pistes, comme

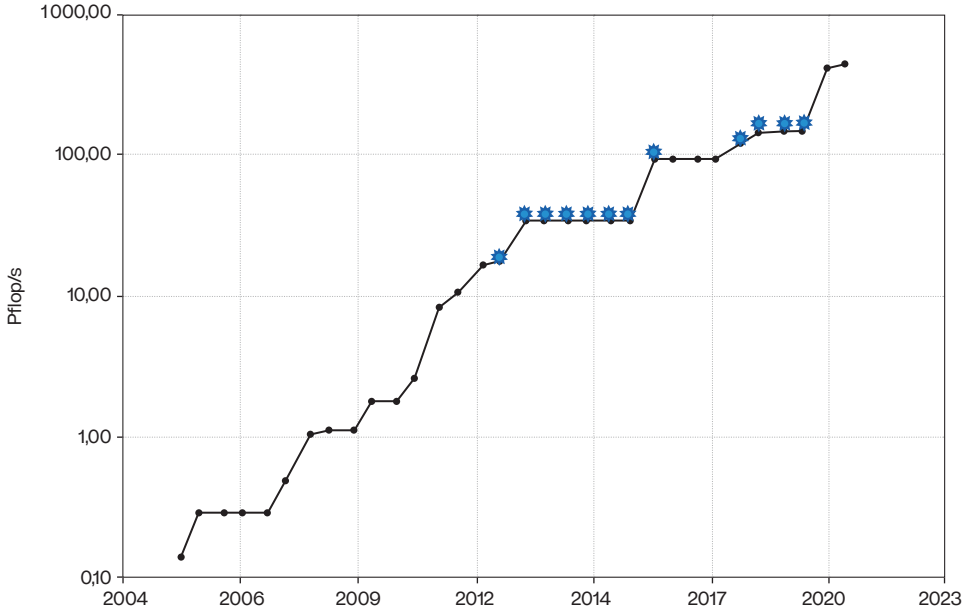

figure 1
Évolution des performances des supercalculateurs les plus puissants au Top500 1 depuis 2005. Ce graphique montre la puissance obtenue en exécutant l'application *Linpack* sur la machine la plus puissante au monde en fonction de la date. La puissance est mesurée en pétaflops (1 Pflop/s = 10^{15} opérations à virgule flottante par seconde) : 1 000 Pflop/s correspondent alors à l'exascale, c'est-à-dire à 1 exaflops (1 Eflop/s = 10^{18} opérations à virgule flottante par seconde). Les supercalculateurs désignés par une étoile bleue s'appuient sur des processeurs additionnels spécialisés nommés GPU (*Graphics Processing Unit*).

ENCADRÉ 1

Enjeu de la consommation : puissance de calcul et efficacité énergétique

Un des défis majeurs relatifs à l'évolution des performances des processeurs de calcul concerne la consommation électrique et donc la capacité à pouvoir exécuter des opérations sur des nombres réels en dépensant le moins d'énergie possible. En effet, une consommation trop importante peut impliquer plusieurs problèmes : une difficulté d'alimentation du processeur, un coût trop important de l'énergie globale sur l'ensemble du supercalculateur et une complexité à dissiper la chaleur dégagée pouvant conduire à la destruction du circuit. Il est donc souhaitable de disposer de processeurs aux performances de calcul augmentées tout en maintenant leur consommation électrique. C'est pour cette raison qu'une métrique a été développée pour évaluer cette capacité. Comme le montre la **figure 2**, les puces sont présentées avec un indicateur défini par le ratio de la performance rapportée à l'unité de puissance électrique et exprimé en Gflops/W. Plus ce ratio est élevé, plus le processeur peut effectuer un nombre important d'opérations sur des nombres réels en consommant peu d'énergie. Cet indicateur devient alors critique pour sélectionner un processeur capable de fournir assez de puissance de calcul efficace et mener à l'exascale au sein d'un supercalculateur. Même si tous les constructeurs tendent à concevoir des processeurs possédant une meilleure efficacité énergétique, le ratio atteint par la puce de Fujitsu ARM A64FX permet d'assembler une machine performante avec une consommation électrique maîtrisée. Pour des supercalculateurs atteignant l'exascale, des technologies de type accélérateur (cœurs spécialisés) permettent de repousser les limites de ce facteur pour en rester à une consommation énergétique acceptable.

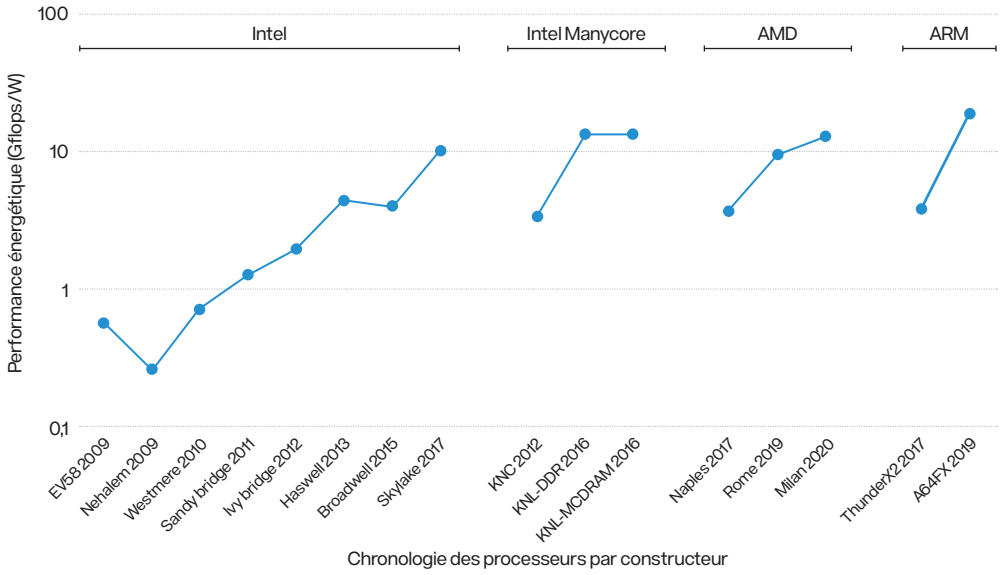

figure 2
Évolution de la performance énergétique, exprimée en Gflops/W (voir encadré 1), des processeurs de calcul d'Intel, d'AMD et d'ARM. Ce graphique illustre la capacité des processeurs à effectuer plus d'opérations arithmétiques pour une consommation électrique constante. La tendance générale met en avant une meilleure efficacité énergétique démontrant les progrès technologiques accomplis, ces derniers devant être poursuivis pour construire une machine de classe exaflopique dans une enveloppe énergétique raisonnable.

l'augmentation de la taille mémoire cache (consistant à garder dans une mémoire ultrarapide proche du processeur les éléments auxquels on a accédé récemment et fréquemment), l'amélioration de la prédiction de branchement (revenant à ne pas arrêter l'exécution des instructions en attente de la résolution d'un branchement de type *if* ou *while/for*) et l'élargissement de l'exécution dans le désordre (gestion automatique des dépendances d'instructions à l'exécution).

Néanmoins, cette piste de processeur à cœurs polyvalents peut atteindre rapidement ses limites et rendre difficile l'assemblage d'un supercalculateur de type exascale, car les contraintes énergétiques nécessaires pour maintenir le côté polyvalent sont trop importantes. C'est pour cette raison que la seconde piste a également été étudiée.

Piste 2 : processeur à cœurs spécialisés

La seconde piste consiste à utiliser les transistors supplémentaires d'une nouvelle génération de puces pour augmenter les performances crête du processeur de façon drastique, quitte à limiter l'amélioration des performances soutenues sur certaines applications. Se reposant souvent sur un travail collaboratif entre le développeur de code et la pile logicielle, cette approche permet une évolution importante des capacités de calcul dans une enveloppe

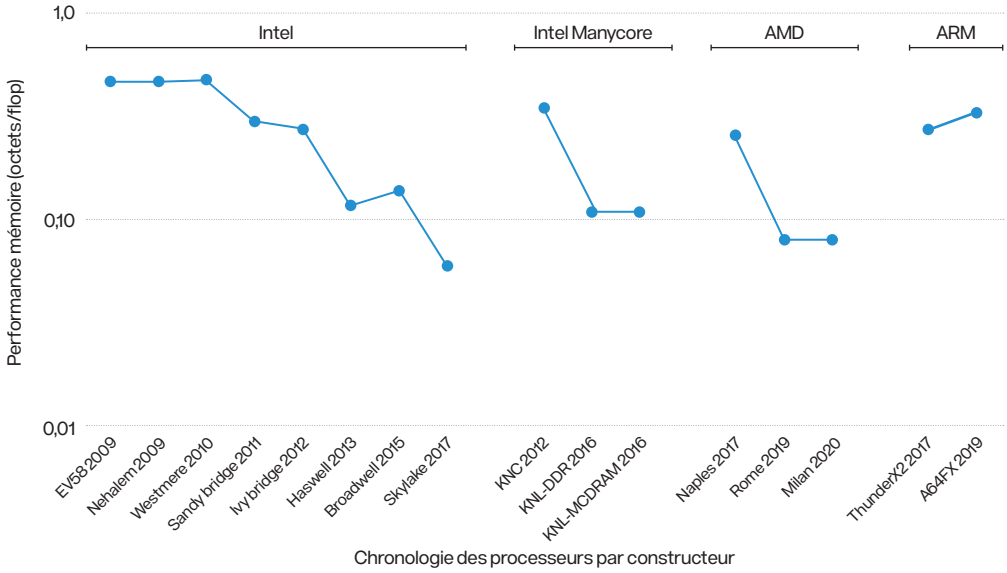

figure 3

Évolution des performances de la mémoire par rapport aux performances de calcul des processeurs. Ce graphique illustre le nombre d'octets qu'il est possible de transférer depuis la mémoire pendant l'exécution d'une opération à virgule flottante. Lorsque ce ratio est grand, il indique alors que le processeur peut facilement effectuer des calculs sur des données venant de la mémoire, sans risquer de créer une pause pour attendre ces données. Sur cette figure, cette valeur a tendance à baisser, indiquant un écart qui se creuse entre les capacités de traitement de la mémoire et les capacités de calcul. Il devient ainsi nécessaire de basculer sur des technologies de mémoire type HBM (comme le processeur A64FX) pour retrouver des ratios équivalents aux processeurs des générations précédentes. L'axe secondaire des ordonnées permet de resituer chaque processeur dans le temps.

énergétique maîtrisée en augmentant le nombre d'opérations par watt consommé (Gflops/W, voir **encadré 1**). Les tendances principales des processeurs de cette catégorie regroupent : (i) l'augmentation du nombre d'unités de calcul à travers un accroissement significatif du nombre de cœurs et un élargissement des unités vectorielles (voir **encadré 3**) ; (ii) une simplification et une très forte réplication des cœurs de calcul permettant de maintenir ce rythme d'augmentation. Les processeurs dits ***manycore***, comme Intel Xeon Phi **KNL** (*Knights Landing*) [3] ou Fujitsu A64FX [4] font partie de cette catégorie, au même titre que les accélérateurs de type **GPU** NVIDIA [5] ou GPU AMD [6]. Ces changements matériels doivent alors être exploités par l'application soit directement dans le code source, soit indirectement à l'aide de bibliothèques ou du compilateur : les unités vectorielles peuvent être utilisées à travers un jeu d'instructions spécifiques **SIMD** (*Single Instruction Multiple Data*) tandis que les nombreux cœurs fonctionneront de façon simultanée si l'application est massivement parallèle.

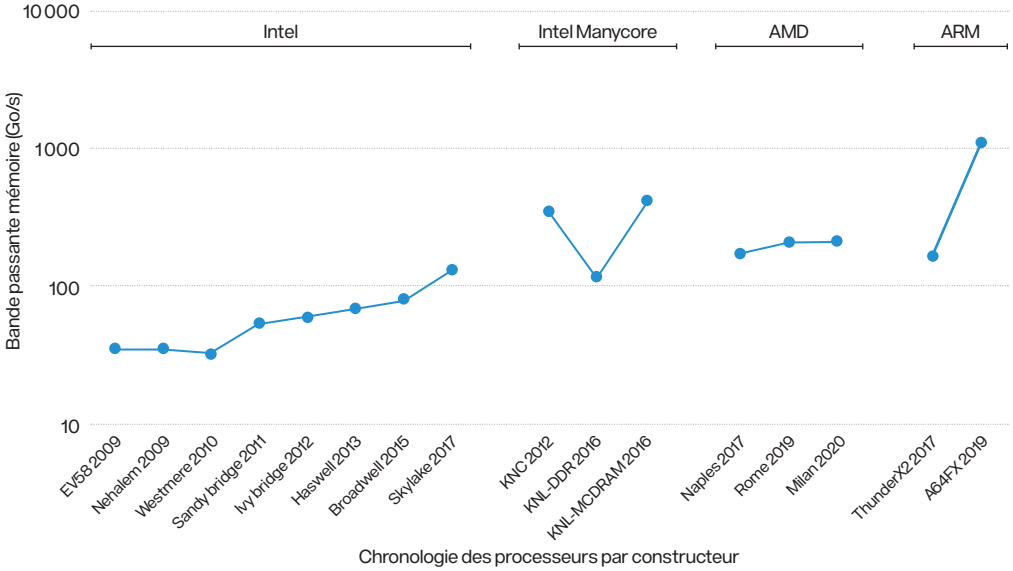

figure 4

Évolution de la bande passante mémoire des différentes générations de processeurs. Ce graphique illustre la quantité de données transférable vers le processeur par seconde. Tandis que ce débit reste faible sur des technologies de processeurs polyvalents, les générations de processeurs spécialisés s'appuient sur des mémoires à haute bande passante permettant d'atteindre jusqu'à 1 To/s.

Les termes spécialisés et les sigles, définis dans le glossaire en fin de numéro, apparaissent en vert

Afin de pouvoir alimenter ces nombreuses unités de calcul, il est nécessaire d'avoir un débit mémoire important (**figure 3**). En effet, les données sont stockées dans la mémoire principale des processeurs se reposant classiquement sur la technologie DDR DRAM (*Double Data Rate Dynamic Random-Access Memory*) qui offre un bon compromis entre latence, bande passante et capacité de stockage. Plusieurs technologies existent pour proposer une mémoire avec une grande bande passante plus importante, mais une capacité réduite : la *Graphics DDR* ou GDDR est présente dans les cartes graphiques d'ancienne génération, la *Multi-Channel DRAM* ou MCDRAM est utilisée dans le processeur Intel Xeon Phi KNL tandis que la gamme *High-Bandwidth Memory* ou HBM est exploitée dans le processeur Fujitsu A64FX et sur les cartes graphiques de nouvelle génération (**figure 4**).

Processeurs de Tera 1000

La première partie a mis en avant les défis principaux concernant l'évolution des processeurs de calcul et deux pistes possibles. Le supercalculateur Tera 1000 repose sur deux types de puces, représentant chaque piste évoquée précédemment. Cette partie décline les **nœuds** de calcul des machines **Tera 1000-1** et **Tera 1000-2**.

Processeur à cœurs polyvalents : Tera 1000-1

La première partition du supercalculateur Tera 1000 se nomme Tera 1000-1. Elle est fondée sur un processeur se situant dans la première piste contenant des cœurs polyvalents. Tera 1000-1 se compose de 2 200 nœuds de calcul comportant chacun deux processeurs Intel Xeon Haswell à 16 cœurs et de 128 Go de mémoire. Cette machine permet d'assurer la continuité du précédent calculateur en production, **Tera 100**, qui comportait des nœuds de calcul à base de processeurs Intel Xeon Nehalem octo-cœurs. La gamme de processeurs Intel Xeon fait partie intégrante de la catégorie polyvalente, car elle propose un nombre de cœurs par puce réduit (16 dans le cas de Tera 1000) avec une fréquence proche du maximum des

ENCADRÉ 2

Enjeu des performances : architecture et microarchitecture

Les processeurs de calcul proposent plusieurs fonctions pour assurer les buts suivants : garantir le bon fonctionnement de l'exécution d'une application par le respect de la sémantique séquentielle des instructions et essayer de réduire son temps d'exécution en exploitant et maximisant le **parallélisme**. Pour ce faire, les architectes ont à leur disposition deux solutions : l'architecture et la microarchitecture. Choisir de dédier une partie du processeur (nouveaux transistors disponibles) à l'une ou l'autre solution relève alors du défi des performances.

La microarchitecture regroupe toutes les fonctionnalités qui se situent à l'intérieur du processeur et fonctionnent de façon automatique et autonome sans l'intervention de l'utilisateur. Toutes les avancées technologiques qui impliquent la microarchitecture permettent une amélioration du comportement du processeur (par exemple, augmentation des performances sur des applications parallèles) sans modification de la séquence d'instructions à exécuter. Cette approche permet une évolution transparente des unités de calcul. Parmi les mécanismes les plus connus, la fréquence du processeur est celle qui a vu une grande évolution jusqu'à la fin des années 2000. En effet, augmenter la fréquence du processeur consiste à accélérer le rythme d'exécution du processeur, exécutant ainsi plus d'instructions par seconde, quel que soit le type d'instruction. Mais il existe plusieurs autres évolutions majeures dans le cadre de la microarchitecture, comme la mise en place d'une hiérarchie mémoire à travers des caches ou l'usage de prédicteurs de branchements.

L'architecture d'un processeur relève de tout ce qui est visible de l'extérieur et contrôlable par un utilisateur *via*, par exemple, des instructions spécifiques. Dans cette catégorie, plusieurs mécanismes augmentant les performances de façon importante sont à noter. En effet, l'apparition de plusieurs cœurs de calcul au sein d'une même puce relève de l'architecture, car ces différentes entités doivent être contrôlées avec des flux d'instructions différents et indépendants, ce qui est à la charge de la pile logicielle et de l'utilisateur. Une autre évolution majeure de l'architecture relative aux performances de calcul concerne l'ajout d'instructions optimisées pour les calculs à virgule flottante. Par exemple, la possibilité d'exécuter, en même temps, une addition et une multiplication à virgule flottante (FMA, pour *floating-point fused multiply-add*) permet d'améliorer les performances, mais cela doit être exploité explicitement à travers une instruction dédiée. L'ajout d'une extension du jeu d'instructions avec des manipulations de vecteurs est aussi une évolution architecturale pour augmenter les performances accessibles par les processeurs modernes.

processeurs de cette génération (2,3 GHz, avec un pic jusqu'à 3,6 GHz en fonction de la charge des cœurs) et un ensemble de mécanismes dédiés à l'optimisation de l'exécution des applications parallèles. Les performances crête de cette machine ne sont pas maximisées, mais les performances soutenues sur les codes de production sont assurées. Les unités de calcul sont alimentées par une mémoire classique de type DDR, aidée par plusieurs niveaux de cache.

Processeur à cœurs spécialisés : Tera 1000-2

Pour compléter le spectre d'évolution des processeurs de calcul, la seconde partition du supercalculateur Tera 1000 s'appuie sur une architecture dite spécialisée. Tera 1000-2 propose des processeurs Intel Xeon Phi KNL (*Knights Landing*) avec un grand nombre de cœurs (**figure 5**). Cette approche représente un compromis entre un processeur polyvalent type Intel Xeon Haswell et une architecture dédiée comme un GPU NVIDIA. En effet, le choix s'est porté sur un processeur manycore, car il propose 64 cœurs de calcul avec une microarchitecture plus simple que Tera 1000-1 (cœur de type Atom issu du monde de l'embarqué) et un jeu d'instructions reposant sur des unités de calcul vectoriel à 512 bits. Pour exploiter pleinement un nœud de calcul de la partition Tera 1000-2, il est nécessaire d'exprimer un degré de parallélisme important qui permettra d'alimenter à la fois tous les cœurs de calcul et les différentes unités de calcul vectoriel de chaque cœur. Dans

ENCADRÉ 3

Hiérarchie de l'architecture et du parallélisme

Pour proposer des performances toujours plus importantes, les supercalculateurs intègrent de plus en plus d'unités de calcul regroupées de façon hiérarchique (**figure E3**). Ainsi, une telle machine est composée de nœuds de calcul qui peuvent communiquer par un réseau d'interconnexion (réseau **BXI**, voir l'**article 3**). L'exemple sur la figure illustre deux nœuds reliés par un tel réseau à travers une carte réseau ou NIC (*Network Interface Controller*). Au sein d'un de ces nœuds, toutes les unités de calcul partagent une même vision virtuelle de la mémoire et peuvent donc s'échanger des données directement par l'intermédiaire de cette mémoire commune. Ces nœuds sont organisés en processeurs (polyvalents ou spécialisés) qui comportent plusieurs cœurs capables d'exécuter des flux d'exécution ou ***threads*** différents de façon simultanée. Cette structure fait ainsi apparaître deux niveaux de parallélisme : entre les nœuds et entre les cœurs de calcul au sein de chacun des nœuds. Mais ces niveaux ne sont pas les seuls. Ainsi, à l'intérieur d'un cœur de processeur, il existe plusieurs types d'unités capables d'exécuter des opérations à virgule flottante : scalaires et vectorielles. Tandis que les unités scalaires sont exploitées automatiquement par le processeur, les unités vectorielles demandent des instructions spécifiques afin d'exprimer l'exécution d'une même opération sur des données différentes. Cela représente alors un dernier niveau de parallélisme à exploiter afin d'obtenir de meilleures performances. Cette exploitation passe par l'utilisation de modèles de programmation spécifiques (voir l'**article 5**).

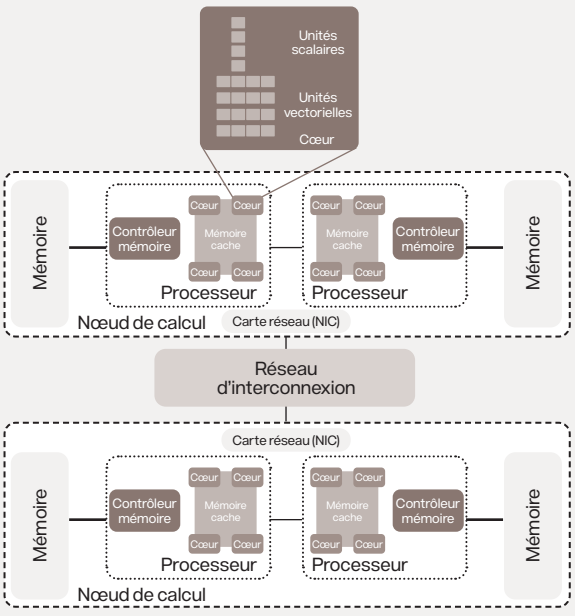

figure E3

Exemple d'organisation d'un supercalculateur avec deux nœuds de calcul reliés par un réseau d'interconnexion à travers une carte réseau ou NIC (*Network Interface Controller*). Chaque nœud possède deux processeurs (polyvalents ou spécialisés) qui comportent chacun quatre cœurs de calcul. Tous les cœurs d'un nœud ont accès à la même zone mémoire située de part et d'autre des processeurs. Chaque cœur expose des unités de calcul scalaires et vectorielles, ces dernières permettant d'appliquer la même opération sur plusieurs données de façon simultanée. Ce schéma illustre ainsi différents niveaux de parallélisme : entre les nœuds de calcul, entre les cœurs à l'intérieur d'un nœud et au sein des unités de calcul vectorielles dans les cœurs.

Les termes spécialisés et les sigles, définis dans le glossaire en fin de numéro, apparaissent en vert

ce contexte, au moins 64 flots d'instructions indépendants devront s'exécuter simultanément et, à chaque cycle d'horloge, chacun de ces flots devra émettre 8 opérations élémentaires indépendantes sur des nombres réels double précision. Cette borne représente un minimum, car l'architecture d'un tel processeur ne permet pas forcément de garder toutes ces unités de calcul occupées. En effet, il devient indispensable d'exprimer plus de flux d'exécution ou *threads* que de cœurs afin de pallier les moments de pause.

Enfin, pour mettre en œuvre cette quantité importante de calcul, il est nécessaire d'alimenter le processeur en données. En effet, si les opérations sont possibles en parallèle, les données doivent être rapatriées rapidement de la mémoire afin de mener à bien ces calculs. Lorsque celles-ci sont déjà dans des registres ou dans un niveau de cache proche, leur accès est alors très rapide (durant le même cycle d'horloge ou quelques cycles maximum). Mais quand il faut aller chercher les données en mémoire principale, il se crée alors un goulot d'étranglement. Le processeur Intel Xeon Phi propose alors un niveau mémoire supplémentaire à capacité réduite (16 Go) qui possède une bande passante plus importante que la mémoire principale classique. Il est alors possible de charger et modifier plus de données à la fois, permettant d'alimenter efficacement le processeur pour les calculs de type vectoriel et de réduire notablement les pauses des cœurs.

Tera 1000-2 met l'accent sur l'amélioration des capacités de calcul avec une consommation énergétique maîtrisée. Relevant de la seconde piste évoquée dans la première partie, la difficulté d'exploitation est alors mise sur la pile logicielle et le développement d'applications parallèles afin d'exprimer assez de parallélisme à plusieurs niveaux pour permettre d'atteindre des performances importantes (**multithreading** et **vectorisation**). Cette machine représente un compromis d'anticipation des évolutions entre les architectures polyvalentes (Tera 100) et les futures générations de processeurs spécialisés (GPU) envisagées pour les futurs calculateurs exascale comme le décrit la troisième partie.

État de l'art et perspectives

Plusieurs supercalculateurs reposent déjà sur des processeurs de la catégorie spécialisée pour proposer une très grande puissance de calcul. Cette section explore les machines actuelles (mondiales et européennes) et futures les plus puissantes afin d'analyser leurs choix en matière de processeurs. Il sera alors possible de mettre en perspective ces machines avec les choix effectués pour Tera 1000, et plus particulièrement Tera 1000-2.

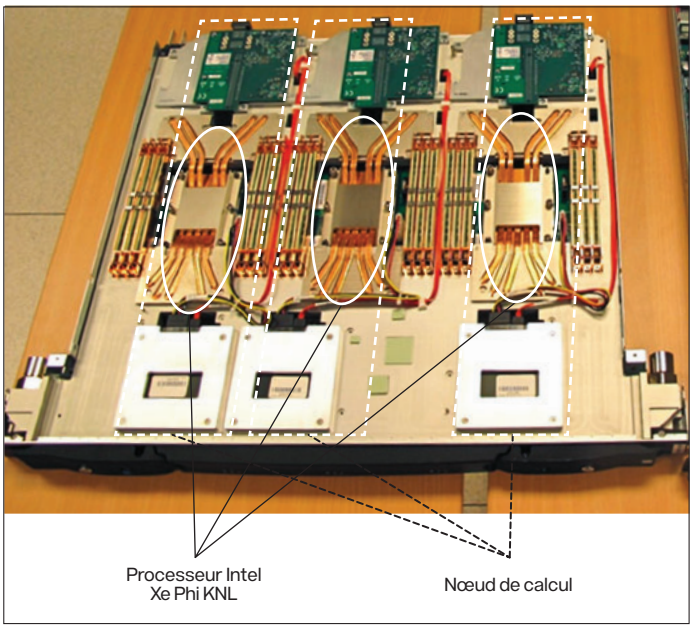

figure 5

Photographie d'une lame de calcul Tera 1000-2 comprenant trois nœuds de calcul. Le supercalculateur Tera 1000-2 contient plus de 8 500 nœuds de calcul composés chacun d'un processeur Intel Xeon Phi KNL avec 64 cœurs et deux mémoires distinctes : 192 Go de mémoire classique DDR et 16 Go de mémoire à haute bande passante MCDRAM (figure 4). Elle s'appuie sur le réseau d'interconnexion Atos-Bull BXI (*Bull eXascale Interconnect*) pour assurer les communications de données entre les nœuds de calcul.

Fujitsu Fugaku (Riken, Japon)

Le premier supercalculateur classé au Top500 en novembre 2020 est Fugaku, une machine construite par Fujitsu et exploitée par le centre de calcul japonais du Riken. Totalisant une puissance de calcul supérieure à un demi-exaflops (comme l'illustre la figure 1), ce supercalculateur utilise un processeur manycore ARM Fujitsu A64FX [4] composé de 64 cœurs de calcul, d'un jeu d'instructions vectorielles de largeur 512 bits et une mémoire à haute bande passante HBM. C'est grâce à un effort considérable d'intégration que le constructeur de supercalculateurs et de processeurs Fujitsu a pu concevoir une telle puce. Mais la puissance de calcul est obtenue au détriment de mécanismes microarchitecturaux permettant l'optimisation automatique d'applications parallèles existantes. Il est alors nécessaire d'exprimer un grand niveau de parallélisme pour exploiter pleinement la puissance de calcul proposée. Cette machine repose sur un processeur de la seconde piste (processeurs à cœurs spécialisés), correspondant au choix fait pour la machine Tera 1000-2.

Atos JUWELS Booster Module (Jülich, Allemagne)

Toujours dans le cadre du haut du classement Top500 du mois de novembre 2020, la première machine européenne est installée en Allemagne,

au centre de calcul Jülich, et est construite par Atos. Dans une partition nommée Booster, la coopération entre un processeur de calcul généraliste (**CPU**) et une carte graphique dédiée au calcul (GPU) est mise en avant. Un nœud de calcul allie deux processeurs AMD EPYC 7402 (Rome) de 24 cœurs chacun avec quatre cartes GPU NVIDIA A100. Ce supercalculateur repose sur une architecture de nœuds de calcul hétérogène dont la majeure partie de la puissance de calcul (plusieurs dizaines de pétaflops) provient des processeurs spécialisés dérivés des cartes graphiques. Cette machine se range dans la piste spécialisée préfigurant les architectures à venir.

HPE Frontier (ORNL) et HPE El Capitan (LLNL)

Les premières machines exaflopiques annoncées seront construites par l'entreprise HPE (Cray à l'origine) et installées dans les laboratoires de recherche nationaux aux États-Unis. La première nommée Frontier a été mise en service durant l'année 2022 dans les locaux d'Oak Ridge National Laboratory (ORNL). Elle exploite une architecture hétérogène portée par le constructeur AMD grâce à des CPU de type EPYC et des GPU de la même firme. Le DOE (Department of Energy) prévoit le développement d'un autre supercalculateur de la même classe basé sur une technologie similaire : El Capitan (Lawrence Livermore National Laboratory, LLNL). Ces supercalculateurs reposent sur la seconde piste d'évolution des capacités de calcul à travers l'exploitation de processeurs spécialisés grâce à : (i) des CPU contenant plusieurs dizaines de cœurs et des unités de calcul vectoriel et (ii) des GPU comprenant plusieurs milliers d'unités de calcul très simples.

HPE Aurora (Argonne National Laboratory, États-Unis)

Un autre supercalculateur exaflopique se situera à Argonne (Illinois) et utilisera la prochaine technologie d'Intel en matière de processeur généraliste (CPU Xeon) et de processeur spécialisé de type GPU. Cette machine repose sur une architecture hétérogène pour proposer une puissance de calcul supérieure à 1 exaflops. La partie dite généraliste à base de CPU Intel Xeon Sapphire Rapids comportera plusieurs dizaines de coeurs de calcul avec un jeu d'instructions vectorielles AVX512. Cette partie se rapproche alors de la solution Fugaku et du choix de Tera 1000-2. Mais la majeure partie des performances sera fournie par les GPU Ponte Vecchio qui posséderont plusieurs milliers d'unités de calcul très simples, laissant la pile logicielle et le développeur d'applications chargé d'exploiter correctement ces unités. Ici, la piste des processeurs spécialisés est suivie avec un choix visant à minimiser la consommation électrique par rapport aux performances délivrées (Gflops/W).

À travers un état de l'art des machines les plus puissantes en 2021 et au-delà, il apparaît que celles-ci s'appuient sur la piste des processeurs spécialisés afin de proposer une puissance de calcul importante dans une enveloppe énergétique raisonnable. Par conséquent, le choix fait pour Tera 1000-2 prend tout son sens.

Conclusion

Proposer des supercalculateurs avec une puissance de calcul toujours plus importante requiert des évolutions majeures des unités de calcul. Concentrées au niveau des processeurs, ces dernières ont évolué en deux pistes principales répondant ainsi à deux enjeux majeurs concernant l'efficacité énergétique et l'utilisation des transistors : conception de puces dites polyvalentes et puces dites spécialisées. Les choix faits pour la machine Tera 1000 suivent ces pistes en proposant deux supercalculateurs afin de permettre une continuité de la production au CEA – DAM par rapport à la machine précédente (Tera 100) et de préparer les prochaines générations de supercalculateurs grâce à une étape intermédiaire. Cette étape a démontré qu'une utilisation efficace de puces spécialisées nécessite un travail important au niveau des applications. Un passage en revue des machines actuelles les plus puissantes et des futurs supercalculateurs de classe exaflopique démontre que ces choix sont pertinents et permettront de préparer les applications pour les prochaines années.

RÉFÉRENCES

1 **J. Dongarra, P. Luszczek,** « TOP500 », dans *Encyclopedia of parallel computing*, Springer (2011).

2 **T. Kidd,** « Why P scales as C*V^2*f is so obvious », https://software.intel.com/content/www/us/en/ develop/blogs/why-p-scales-as-cv2f-is-so-obvious-pt-2-2.html (2009).

3 **A. Sodani,** « Knights landing (KNL): 2nd Generation Intel® Xeon Phi processor », *2015 IEEE Hot Chips 27 Symposium (HCS)*, p. 1-24, doi: 10.1109/HOTCHIPS.2015.7477467 (2015).

4 **M. Sato *et al.*,** « Co-design for A64FX manycore processor and Fugaku », *Proc. of SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, p. 1-15, doi: 10.1109/SC41405.2020.00051 (2020).

5 **I. Buck,** GPU computing with NVIDIA CUDA, SIGGRAPH '07: ACM SIGGRAPH (2007).

6 **J. Choquette, W. Gandhi, O. Giroux, N. Stam, R. Krashinsky,** « NVIDIA A100 tensor core GPU: performance and innovation », *IEEE Micro*, **41,** p. 29-35, doi: 10.1109/MM.2021.3061394 (2021).

Les termes spécialisés et les sigles, définis dans le glossaire en fin de numéro, apparaissent en vert

N°52 / JANVIER 2023

CHOCS

Revue scientifique
et technique de la Direction
des applications militaires

N°52 / JANVIER 2023

CHOCS

REVUE SCIENTIFIQUE ET TECHNIQUE
DE LA DIRECTION DES APPLICATIONS MILITAIRES

cea

www-dam.cea.fr