AMD CDNA™ 2 ARCHITECTURE

Propulser la recherche la plus importante de l'humanité avec L'accélérateur **HPC et IA** le plus puissant au monde

Introduction

L'histoire des GPU est une histoire d'évolution - des conceptions extrêmement spécialisées et câblées aux accélérateurs entièrement programmables. Ceux-ci

Les accélérateurs sont fondamentalement optimisés pour un débit massif et peuvent exécuter des langages de programmation standard tels que C ++ dans le cadre d'un écosystème logiciel riche. Cette combinaison s'est avérée incroyablement convaincante pour le calcul scientifique et l'apprentissage automatique, où le débit de calcul a permis d'innover et de découvrir d'énormes possibilités dans une grande variété d'applications

Les architectures d'accélérateurs précédentes ont constamment amélioré les performances et l'efficacité tout en devenant de plus en plus programmables.

Cependant, l'architecture AMD CDNA™ 2 fait passer ce chemin évolutif au niveau supérieur, atteignant plus de x 4 amélioration des performances par rapport à l'architecture AMD CDNA de génération précédente, avec un débit vectoriel de pointe FP64 de 47,9 TFLOP/s, pour permettre des niveaux exascale des performances avec une programmation inégalée dans les systèmes hétérogènes

L'architecture AMD CDNA 2 représente un bond en avant majeur par rapport à la génération précédente en améliorant la technologie Matrix Core pour le HPC et l'IA, piloter les capacités de calcul pour les données à virgule flottante à double précision et une variété de primitives de multiplication matricielle.

Il se concentre également sur l'amélioration de la communication et de la mise à l'échelle des accélérateurs, en tirant parti de l'Infinity Fabric™ unique d'AMD pour permettre un module multi-puces qui fournit 1,8x la densité de calcul par rapport à la génération précédente et permet une meilleure connectivité au sein d'un seul système. Enfin, l'architecture AMD CDNA 2 permet à des accélérateurs tels que l'AMD Instinct™ MI250X de fonctionner comme un homologue complet au sein d'un système informatique en offrant une cohérence du cache avec certains processeurs EPYC optimisés de 3e génération, offrant une rampe d'accès rapide et simple pour les codes CPU à exploiter la puissance des accélérateurs. Ces nouvelles fonctionnalités sont toutes déverrouillées de manière transparente par la plate-forme logicielle ouverte ROCm™ d'AMD, qui offre un ensemble riche d'outils pour les clients qui portent ou développent des applications de pointe.

Présentation de l'architecture AMD CDNA 2

L'architecture AMD CDNA™ 2 s'appuie sur les énormes forces de base de l'architecture AMD CDNA originale pour offrir un bond en avant dans les performances et la convivialité du système lors de l'utilisation d'une technologie de processus similaire. L'architecture AMD CDNA est un excellent point de départ pour une plateforme de calcul. Cependant, pour offrir des performances exascale, l'architecture a été remaniée avec des améliorations à presque tous les aspects des unités de calcul à l'interface mémoire, avec un accent particulier sur l'amélioration radicale des interfaces de communication pour une évolutivité totale du système.

L’architecture AMD CDNA 2 est conçue avant tout pour les applications de calcul scientifique et d’apprentissage automatique les plus exigeantes.Il alimente la nouvelle génération de produits AMD Instinct™ MI200 qui ciblent des solutions allant des systèmes simples compacts jusqu’aux plus grands supercalculateurs exascale au monde avec des modèles de programmation uniques et hautement différenciés . La figure 1a illustre la DMLA Matrice de calcul graphique Instinct MI200 (GCD). L’AMD Instinct™ MI200 est construit sur des technologies d’emballage avancées, permettant à deux GCD d’être intégré dans un seul boîtier dans le format OAM (OCP Accelerator Module) des produits MI250 et MI250X, comme illustré à la figure1b. Chaque GCD est construit sur l’architecture AMD CDNA 2.

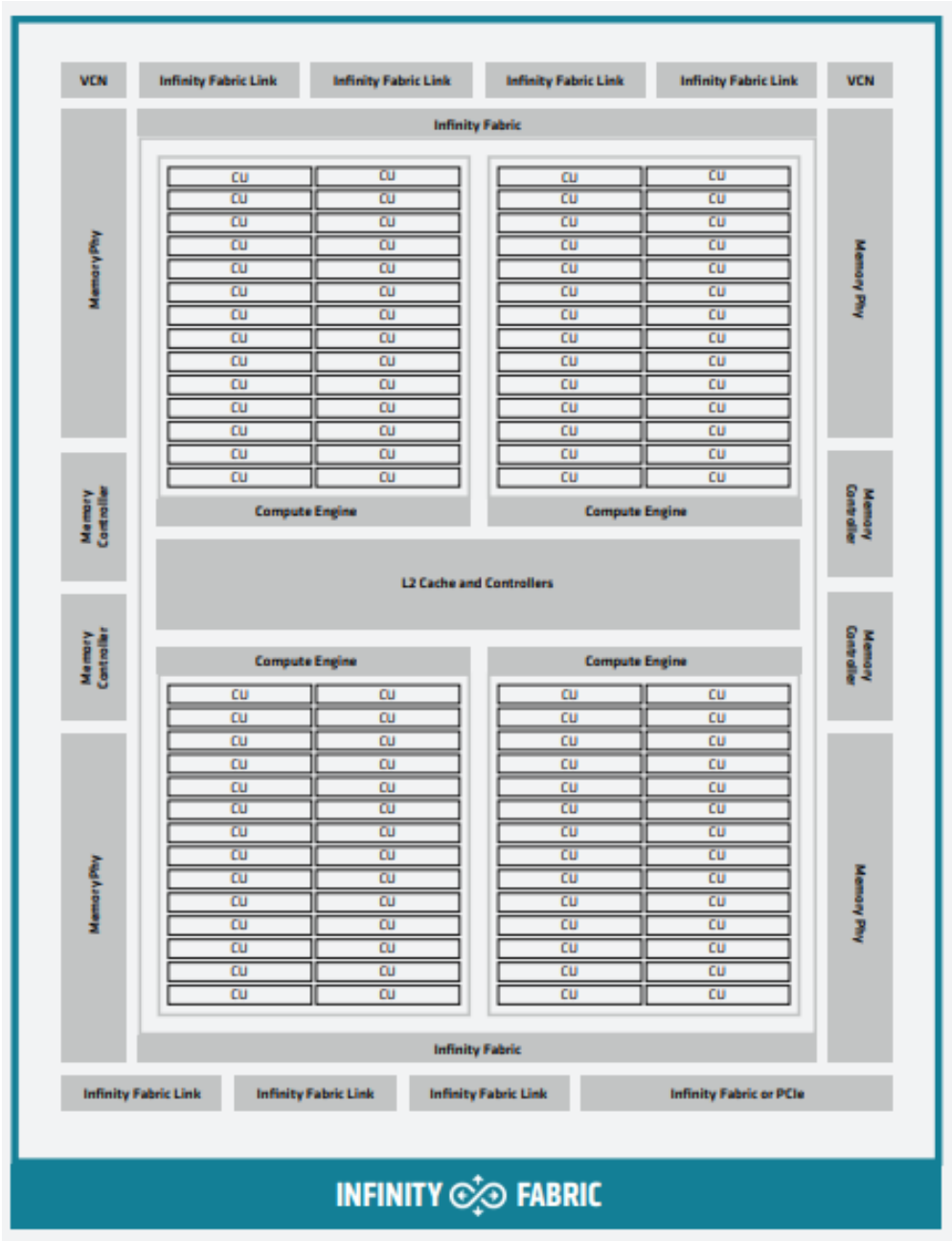

Figure 1a – Schéma fonctionnel de la matrice de calcul graphique AMD Instinct™ MI200 GCD)

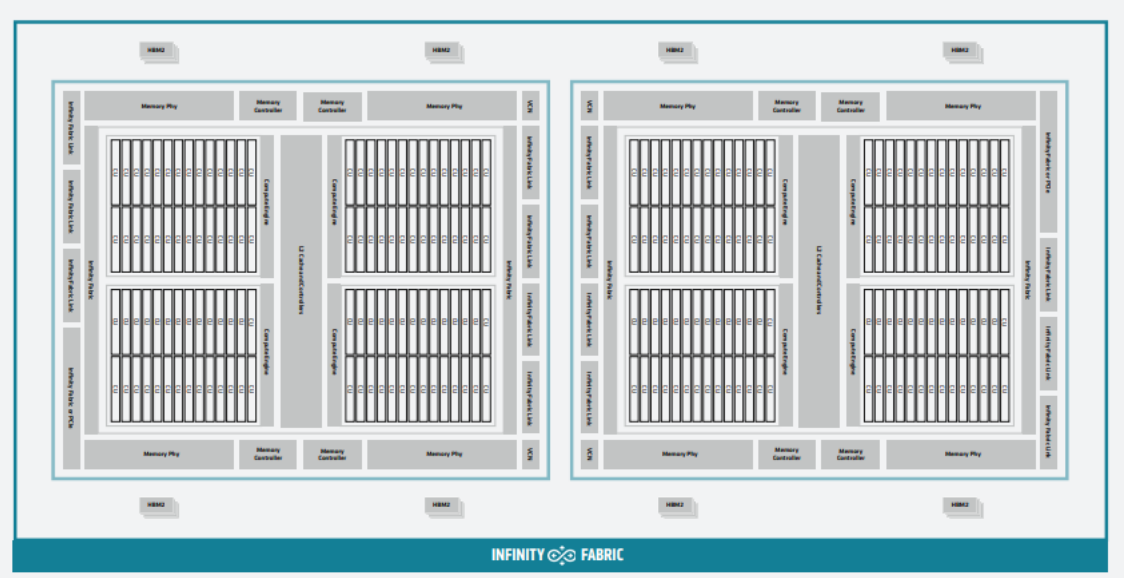

Figure 1b – Schéma fonctionnel du module multi-puce AMD Instinct™ MI200 (AMD Instinct™ MI250/MI250X) dans l'accélérateur de facteur de forme OAM qui comprend deux matrices de calcul graphique (GCD) comme illustré

Les trois fonctions critiques de tout processeur sont le calcul, la mémoire et la communication. Chaque AMD CDNA™ 2 GCD possède plusieurs blocs dédiés à ces fonctions et sont connectés avec sur plaquette ''on-die fabric'' à grande vitesse. Cependant, pour offrir des performances exascale, cela seul est pas assez. L'une des innovations cruciales de l'architecture AMD CDNA 2 consiste à utiliser l'Infinity Fabric unique d'AMD pour étendre l'on-die fabric à travers le package afin que chaque GCD apparaisse comme un GPU dans un système de mémoire partagée. Connecter deux GCD ensemble de cette manière double les ressources, créant ainsi un bloc de construction informatique plus grand en plus des nombreuses autres améliorations

L'architecture AMD CDNA 2 a plusieurs incarnations différentes offrant à la fois une implémentation personnalisée qui utilise AMD Infinity Fabric™ pour l'interface avec un processeur AMD EPYC™ optimisé de 3e génération pour une plate-forme de supercalculateur HPE/Cray spécifique ainsi qu'une implémentation disponible qui s'appuie sur PCI-Express® pour s'interfacer avec les processeurs hôtes . Chaque GCD inclut un processeur de commandes qui obtient des commandes au niveau de l'API à partir du processeur hôte et les traduit en travail qui peut être généré sur différentes parties de l'AMD CDNA 2 architecture.

L'une des innovations fondamentales de l'architecture AMD CDNA de la génération précédente a été l'introduction de la technologie Matrix Core dans les unités de calcul (CU) pour augmenter le débit de calcul en mettant l'accent sur les types de données utilisés dans l'apprentissage automatique. Le noyau matriciel.la technologie de l'architecture AMD CDNA 2 s'appuie sur cette base et a été améliorée pour prendre en charge un plus large éventail de types de données et Applications avec un accent particulier sur le calcul scientifique avec les données du 64e PC. En outre, semblable à la génération précédente, la baie CU

est partitionné en quatre moteurs de nuanceur qui exécutent les noyaux de calcul générés par le processeur de commandes. Le résultat net est que les accélérateurs amd instinct de la série MI200 peuvent fournir jusqu'à un débit théorique de pointe de 47,9 TFLOP / s à double précision, soit 4,2 fois la génération précédente

Chaque GCD dispose également de 2x Video Codec Next (VCN), un bloc logique qui fournit des capacités d'encodage et de décodage sur les entrées et les sorties flux de données. Il s'agit d'une logique particulièrement utile pour les charges de travail de formation Machine Learning pour la détection d'objets qui s'entraînent sur l'image et les données vidéo. . Les blocs VCN prennent en charge H.264/AVC, HEVC, VP9 et JPEG pour le décodage, ainsi que H.264/AVC et HEVC pour l'encodage

Communication et mise à l'échelle

À bien des égards, les améliorations les plus critiques apportées à l'architecture AMD CDNA™ 2 concernent les capacités de communication de chaque GCD au sein de l'appareil AMD MI200 et en particulier dans les capacités uniques offertes par la technologie AMD Infinity Fabric™. La génération précédente s'est appuyée sur PCI-Express standard pour se connecter au processeur hôte et offrait trois liens **AMD Infinity Fabric**™ se connectant à d'autres GPU. Dans l'exemple de topologie HPC phare illustré à la figure 2a. l'architecture AMD CDNA 2 développe les capacités de communication à un niveau différent avec quatre types différents d'interfaces spécialisées à des fins différentes: Infinity Fabric in-package, Infinity inter-package Liens Fabric, liens Infinity Fabric cohérents vers le processeur hôte et lien PCIe en aval, Le tissu Infinity dans l'emballage, cohérent Infinity Fabric et les liens PCIe en aval sont tous nouveaux et déverrouillent les capacités système uniques illustrées à la figure 2a. Dans le plus Topologies d'apprentissage automatique traditionnelles et phares, illustrées à la figure 2b-c, les GPU sont connectés au processeur hôte ;via PCIe mais toujours bénéficier du nombre accru de liens GCD-to-GCD Infinity Fabric dans le périphérique GPU ainsi que de l'inter-package liens externes Infinity Fabric.

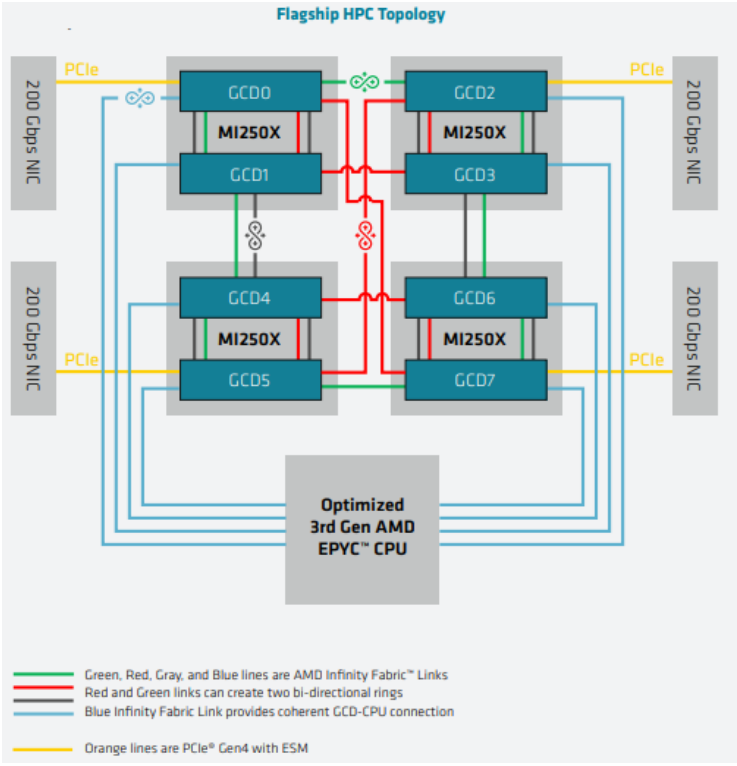

figure 2a – Schéma fonctionnel d'un nœud HPC phare construit à l'aide de l'accélérateur AMD Instinct™ MI250X et optimisé 3rd processeur AMD EPYC™ de génération

Topologie HPC/ML grand public

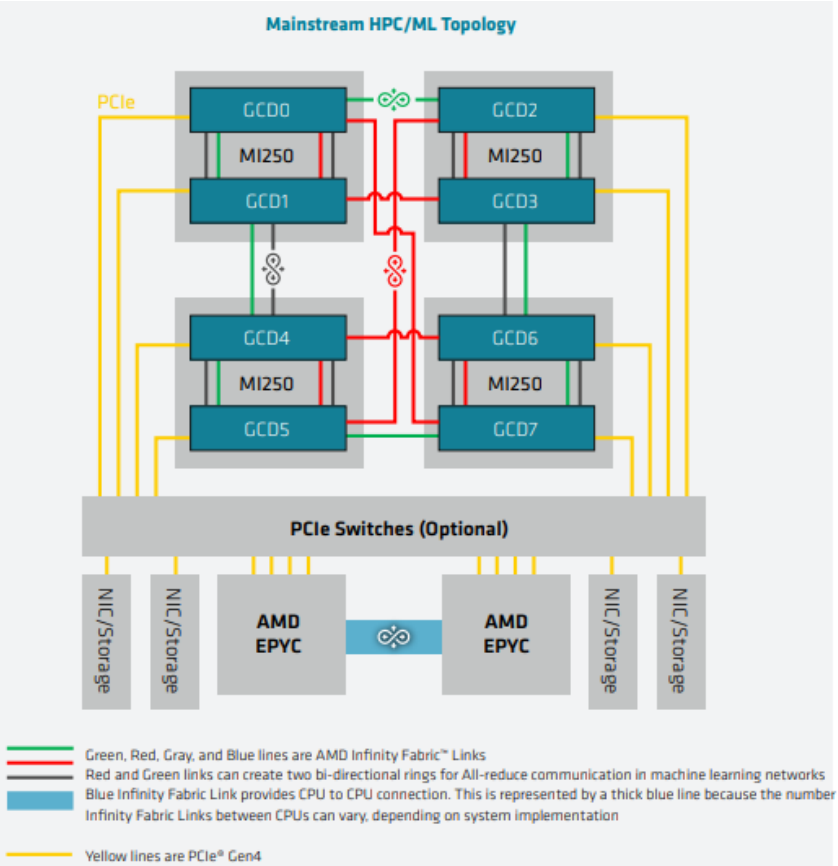

Figure 2b – Schéma fonctionnel d'un nœud HPC/ML grand public construit à l'aide des accélérateurs AMD Instinct™ MI250 et AMD Processeurs EPYC™

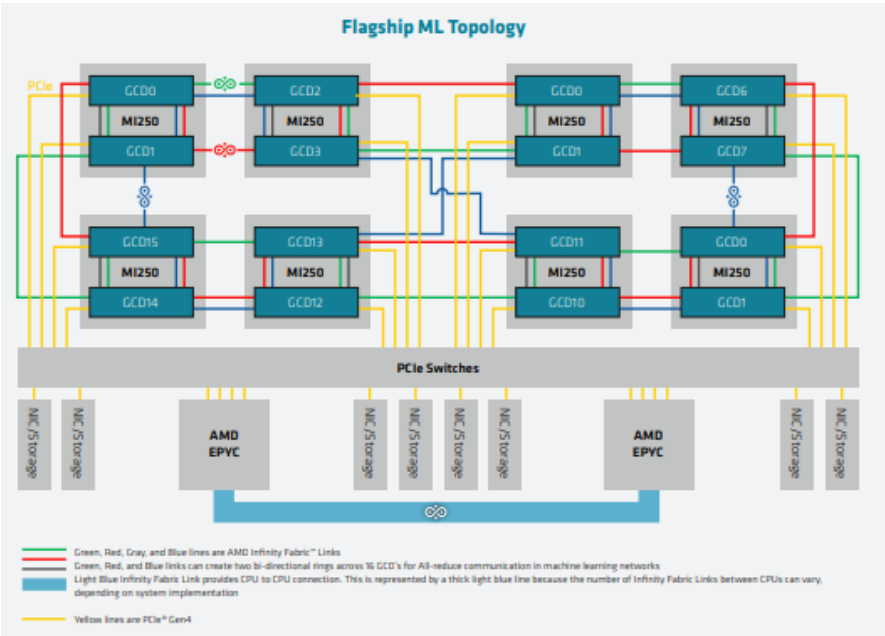

Figure 2c – Schéma fonctionnel d'un nœud optimisé pour le ML construit à l'aide des accélérateurs AMD Instinct™ MI250 avec trois logiques anneaux reliant les accélérateurs et les processeurs AMD EPYC™

L'interface AMD Infinity Fabric intégrée à l'emballage est l'une des innovations clés de la famille AMD CDNA™ 2, connectant les 2 GCD au sein duMI250 ou MI250X. Il tire parti des distances extrêmement courtes entre les GCD dans le boîtier pour fonctionner à 25 Gbps et à puissance extrêmement faible, offrant une bande passante bidirectionnelle maximale théorique allant jusqu'à 400 Go/s entre les GCD

Les 8 liens externes AMD Infinity Fabric™ pour GPU P2P ou E/S sur les accélérateurs AMD Instinct™ MI250 (ou MI250X) délivrent jusqu'à 800 Go/s de la bande passante théorique totale fournissant jusqu'à 235 % des performances théoriques de la bande passante GPU P2P (ou E/S) de la génération précédente Produits de calcul GPU AMD Instinct9™

L'interface hôte cohérente est un autre aspect nouveau de l'architecture AMD CDNA 2, qui permet la cohérence avancée de la mémoire. La couche physique est implémentée sous la forme d'un lien Infinity Fabric à 16 voies. Logiquement, le lien peut se comporter comme une interface Infinity Fabric lorsqu'il est couplé avec un processeur AMD EPYC™ de 3e génération optimisé, permettant des capacités uniques de cohérence du cache. Lorsque vous êtes connecté à un autre serveur X86 Processeurs . L'interface hôte cohérente est un autre aspect nouveau de l'architecture AMD CDNA 2, qui permet la cohérence avancée de la mémoire. La couche physique est implémentée sous la forme d'un lien Infinity Fabric à 16 voies. Logiquement, le lien peut se comporter comme une interface Infinity Fabric lorsqu'il est coupléavec un processeur AMD EPYC™ de 3e génération optimisée, permettant des capacités uniques de cohérence du cache. Lorsque vous êtes connecté à un autre serveur X86Processeurs

La dernière interface est une liaison ESM PCIe 4.0 en aval qui fonctionne jusqu'à 25 Gbps. Contrairement à l'interface hôte, cette interface en aval est couplée à un complexe racine PCIe, qui peut piloter des périphériques d'E/S connectés au GPU. Cette capacité est cruciale pour la vision de l'informatique exascale, qui repose sur des CPU et des GPU agissant sur un pied d'égalité dans un système. Avec l'introduction d'une cohérence totale entre le CPUet le GPU, ces deux appareils agiront comme des pairs pour le calcul. Cette liaison d'E/S en aval leur permet de se connecter à la fois à un haut débit réseau et agir en tant que paires à part entière dans le contexte de la communication.

AMD CDNA - tableau du shader

Comme AMD CDNA, dans AMD CDNA™ 2, le processeur de commandes reçoit les commandes API et les transforme en tâches de calcul. Le calcul des tâches sont gérées par les quatre moteurs de calcul asynchrones (ACE), qui distribuent des fronts d'onde du shader de calcul aux unités de calcul.

Les unités de calcul AMD CDNA 2 adoptent généralement une approche évolutive et s'appuient sur les bases solides des générations précédentes en tant qu' illustré à la figure 3. L'architecture AMD CDNA originale a été dérivée de l'architecture GCN antérieure et a introduit la notion de matrice

en tant que citoyen de première classe en ajoutant la technologie Matrix Core et la prise en charge de nouveaux types de données. AMD CDNA 2 double la mise en cette approche et améliore plusieurs autres aspects.

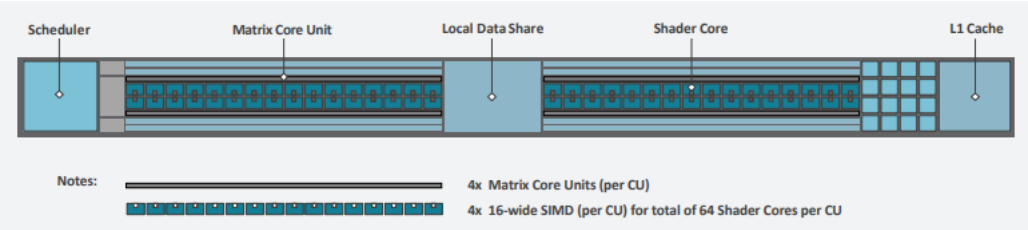

Figure 3 – Schéma fonctionnel conceptuel d'une unité de calcul améliorée (CU) avec vue SIMD de l'AMD CDNA™ 2 architectures

L'architecture AMD CDNA 2 intègre 112 unités de calcul physiques par GCD, divisées en quatre baies ; les produits initiaux comprennent 108 (pour AMD Instinct™ MI250) ou 110 (pour AMD Instinct™ MI250X) CU actifs par GCD. Comme illustré à la figure 3, chaque CU contient des fichiers de registre et pipelines optimisés pour les instructions scalaires, vectorielles et matricielles. Les instructions vectorielles fonctionnent avec des fronts d'onde contenant 64 éléments de travail et la plupart des instructions de double précision sont exécutées sur quatre cycles à l'aide d'un ensemble de quatre pipelines de 16 de large chacun. La matrice des instructions extraient les données du fichier de registre vectoriel, mais ont leurs propres chemins de données spécialisés qui tirent parti de la réutilisation implicite des données dans la multiplication matricielle pour réduire le nombre d'accès au registre pour un calcul donné afin d'améliorer à la fois les performances et l'efficacité.Les données peuvent être partagées entre les voies d'un front d'onde à l'aide d'instructions de permute ou du magasin de données local (LDS).

Les unités de calcul AMD CDNA 2 ont été optimisées avec soin pour améliorer les performances du calcul scientifique et fournir plus de débit cohérent pour les types de données utilisés dans l'apprentissage automatique. Historiquement, les GPU ont été optimisés pour les opérations en virgule flottante à précision unique et les opérations de double précision se déroulent à une vitesse inférieure, allant de la demi-vitesse à aussi lente qu'un seizième. Pour améliorer les applications de calcul scientifique, le pipeline vectoriel AMD CDNA 2 a été réglé de manière à ce que le fonctionnement sur des données à double précision plus larges soit le même taux que celui de la précision simple, avec 64 opérations de multiplication-ajout fondu (FMA) par horloge. Profiter de cela pour l'amélioration, les pipelines vectoriels peuvent également exécuter des opérations sur des valeurs de précision uniques compressées, doublant ainsi le débit à 128opérations FMA de précision par cycle.

Le partage de données local est conçu pour transmettre explicitement des données au sein d'une unité de calcul. Ce flux de communication crée une opportunité naturelle pour les opérations atomiques distribuées à très haut débit. Dans l'architecture AMD CDNA 2, les unités d'exécution atomique dans le LDS ont été amélioré pour augmenter le débit pour FP64 min, max, et ajouter des opérations atomiques, qui sont couramment utilisées dans le calcul scientifique, par exemple,

Computation	MI100 (FLOPS/CLOCK/CU)	MI250X (FLOPS/CLOCK/CU)	MI100 (Peak)	MI250X (Peak)	MI200 Peak Speedup
MI200 Matrix FP64 vs. MI100 Vector FP64	64	256	11.5 TFLOPS	95.7 TFLOPS	8.3x
MI200 Vector FP64 vs. MI100 Vector FP64	64	128	11.5 TFLOPS	47.9 TFLOPS	4.2x
MI200 Matrix FP32 vs. MI100 Matrix FP32	256	256	46.1 TFLOPS	95.7 TFLOPS	2.1x
MI200 Packed FP32 vs. MI100 Vector FP32	128	256	23.1 TFLOPS	95.7 TFLOPS	4.2x
MI200 Vector FP32 vs. MI100 Vector FP32	128	128	23.1 TFLOPS	47.9 TFLOPS	2.1x
MI200 Matrix FP16 vs. MI100 Matrix FP16	1024	1024	184.6 TFLOPS	383 TFLOPS	2.1x
MI200 Matrix BF16 vs. MI100 Matrix BF16	512	1024	92.3 TFLOPS	383 TFLOPS	4.2x
MI200 Matrix INT8 vs. MI100 Matrix INT8	1024	1024	184.6 TOPS	383 TOPS	2.1x

Tableau 1 – Comparaison générationnelle des formats numériques et du débit de pointe entre MI250X (OAM) et MI100 (PCIe).

Technologie AMD CDNA 2 Matrix Core

La technologie **Matrix Core** de l'architecture AMD CDNA™ 2 a également été améliorée, en mettant l'accent sur le calcul haute performance. La technologie AMD CDNA 2 Matrix Core prend désormais en charge les données à double précision, ce qui est essentiel pour de nombreuses applications de calcul scientifique.

La multiplication matrice-matrice est l'une des primitives importantes qui peuvent être exploitées dans les noyaux HPC. Sa mise en œuvre accélérée peut accélérer l'exécution des applications HPC, y compris l'important **Linpack** haute performance (HPL). Effectuer la multiplication matricielle à l'aide des instructions générales FMA64 sont moins efficaces, dépenser beaucoup d'énergie pour les accès aux fichiers de registre pour chaque opérande. En fin de compte, cette énergie,l'utilisation limite les performances maximales possibles dans un TDP donné**.**

AMD CNDA 2 introduit un ensemble d'instructions de multiplication matricielle spécifiquement pour la précision FP64 avec une microarchitecture simplifiée. Nouveau. les instructions réalisent une multiplication matricielle basée sur des blocs pour les tailles de blocs matriciels fixes de 16x16x4 et 4x4x4 (MxNxK) et sont à l'échelle de l'onde opérations où l'entrée des données de bloc de matrice de sortie sont réparties sur les voies d'un front d'onde. L'ensemble de l'entrée est lu à partir des registres une fois et réutilisé plusieurs fois pendant le calcul pour une réduction substantielle de la puissance.

Les instructions de multiplication de matrice FP64 peuvent fournir deux fois le débit par rapport à l'utilisation d'instructions vectorielles FP64, tout en fournissant également aider à améliorer l'efficacité énergétique. Le résultat net est une amélioration correspondante de 4X du FP64 TFLOP/s par rapport au MI100. Ces instructions peuvent être utilisées dans les bibliothèques fournies par AMD pour accélérer les calculs d'algèbre linéaire et devraient démontrer un maximum de débit FP64

pour MI200. Comme l'illustre le tableau 1 ci-dessus, cela quadruple le débit de calcul pour fp64 par rapport au précédente génération.

De plus, la technologie AMD CDNA 2 Matrix Core a amélioré les performances du bfloat16 afin qu'elle offre un débit équivalent à FP16

AMD CDNA package FP32

Une autre première dans l'architecture AMD CDNA™ 2 est Package FP32, qui exécute deux instructions vectorielles de composants sur les opérandes FP32 pour les opérations FMA, FADD et FMUL. Ces nouvelles instructions doublent le débit vectoriel FP32 par horloge et par CU pour ces opérations. Les instructions du package FP32 reposent sur le fait que les opérandes vectoriels sont adjacents et alignés sur des registres pairs et appliquent le même arrondi et la même 'dénormalisation' modes pour les deux opérations ; Pour faciliter le placement d'opérandes scalaires dispersés ensemble dans un tel vecteur, l'architecture prend en charge une instruction de déplacement compressée qui accède à deux registres scalaires et les copie dans une paire de registres d'emplacement adjacents. Le résultat de cette opération de déplacement peut devenir une entrée d'une opération mathématique emballée d'une manière relativement simple. L'exemple de code ci-dessous montre les modifications nécessaires pour tirer parti du Package FP32.

Original

```
float vxi = 0.0f, vyi = 0.0f, vzi = 0.0f;
  for (int j = hipThreadIdx_x; j < count1; j += hipBlockDim_x) {
    float dx = xx1[j] - xxi;
    float dy = yy1[j] - yyi;
    float dz = zz1[j] - zzi;
    float dist2 = dx*dx + dy*dy + dz*dz;
    if (dist2 < fsrrmax2) {
      float rtemp = (dist2 + rsm2)*(dist2 + rsm2)*(dist2 + rsm2);
      float f_over_r = massi*mass1[j]*(1.0f/sqrt(rtemp) - (ma0 +
dist2*(ma1 + dist2*(ma2 + dist2*(ma3 + dist2*(ma4 + dist2*ma5))))));

      vxi += fcoeff*f_over_r*dx;
      vyi += fcoeff*f_over_r*dy;
      vzi += fcoeff*f_over_r*dz;
    }
  }
```

Modified to use Packed FMA32

```
float2 vxi = 0.0f, vyi = 0.0f, vzi = 0.0f;
  for (int j = hipThreadIdx_x; j < count1; j += 2*hipBlockDim_x) {
    float2 dx = {xx1[j] - xxi, xx1[j+ hipBlockDim_x] - xxi};
    float2 dy = {yy1[j] - yyi, yy1[j+ hipBlockDim_x] - yyi};
    float2 dz = {zz1[j] - zzi, zz1[j+ hipBlockDim_x] - zzi};
    float2 dist2 = dx*dx + dy*dy + dz*dz;
    if (dist2 < fsrrmax2) {
      float2 rtemp = (dist2 + rsm2)*(dist2 + rsm2)*(dist2 + rsm2);
      float2 f_over_r = massi*mass1[j]*(1.0f/sqrt(rtemp) - (ma0 +
dist2*(ma1 + dist2*(ma2 + dist2*(ma3 + dist2*(ma4 + dist2*ma5))))));

      vxi += fcoeff*f_over_r*dx;
      vyi += fcoeff*f_over_r*dy;
      vzi += fcoeff*f_over_r*dz;
    }
  }
```

Figure 4 – Exemple de code montrant les modifications nécessaires pour tirer parti du Package FP32**.**

La plate-forme logicielle ouverte AMD ROCm active AMD CDNA 2

La clé de l'informatique accélérée pour le HPC et le ML est une pile logicielle et un écosystème qui déverrouille facilement les capacités des logiciel développeurs et clients. La pile AMD ROCm™, illustrée à la figure 5, fournit un ensemble d'outils open source et faciles à utiliser qui sont construits autour des normes de l'industrie et permettre la création de logiciels portables bien optimisés pour tout, des simples programmes de poste de travail aux applications exascale massives.

Les principes derrière AMD ROCm sont assez simples. Premièrement, l'informatique accélérée exige l'égalité entre les deux processeurs et accélérateurs. Bien qu'ils se concentrent sur différentes charges de travail, ils devraient travailler ensemble efficacement et avoir un accès égal à des ressources telles que mémoire. Deuxièmement, un riche écosystème de bibliothèques logicielles et d'outils devrait permettre d'utiliser des logiciels portables et un code performant qui peut en tirer parti de nouvelles capacités. Enfin, une approche open-source permet aux fournisseurs, aux clients et à l'ensemble de la communauté, ainsi qu'à l'amplification de l'investissement d'AMD.

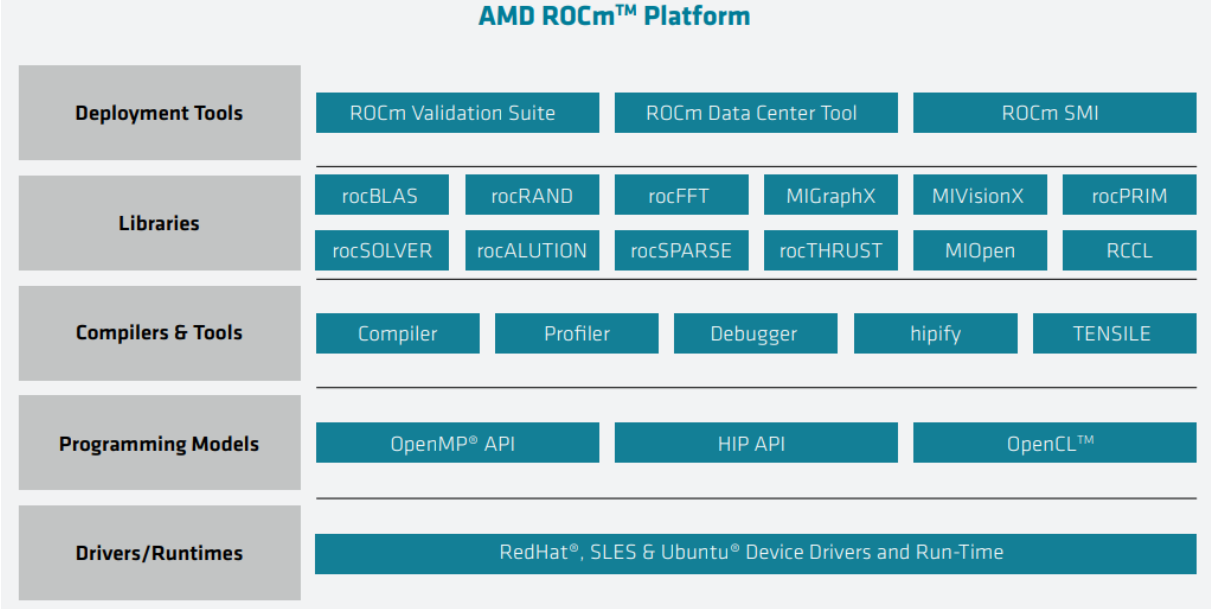

Figure 5 – La pile ROCm open source d'AMD comprend les outils dont les développeurs ont besoin pour créer des performances élevées applications pour le calcul scientifique et l'apprentissage automatique.

L'écosystème **AMD ROCm** est crucial pour mettre les capacités de l'architecture AMD CDNA™ 2 entre les mains des développeurs, les fournisseurs, les clients et l'ensemble de la communauté. Par exemple, la bibliothèque **rocBLAS** a incorporé les nouvelles instructions pour la matrice FP64 multiplication et package FP32 les vecteurs FP32 afin que les développeurs qui fonctionnent avec des bibliothèques de niveau supérieur obtiendront d'excellentes performances dès le jour Un. À un niveau inférieur, le compilateur et le runtime ROCm peuvent tirer parti de ces mêmes fonctionnalités pour générer des fichiers binaires hautes performances pour du code personnalisé et un ensemble plus diversifié d'applications au-delà de l'algèbre linéaire ; À un niveau encore plus élevé, l'Infinity Hub d'AMD (https://www.amd.com/en/technologies/infinity-hub) contient des applications HPC et ML conteneurisées prêtes à l'emploi et prenant en charge les derniers Accélérateurs de la série MI200.

Dans le même temps, les capacités uniques de l'architecture AMD CDNA 2 - en particulier la cohérence du cache, permettent de simplifier les applications et, offrant des performances encore plus élevées. Par exemple, certaines parties de NWChemEx utilisent une mémoire unifiée cohérente; portage sur non cohérent les processeurs et les accélérateurs pourraient ajouter de la complexité, y, introduire de nouveaux bogues et retarder généralement le déploiement de l'application. AMD MI250Xaccélérateur avec le processeur AMD EPYC™ optimisé de 3e génération dans une configuration cohérente avec le cache peut grandement améliorer la productivité

Pour d'autres applications, la communication fine améliore les performances. Par exemple, le HACC simulateur de cosmologie, construit un arbre 'structure de données' pour suivre les particules et leurs voisins les plus proches. Ces relations sont ensuite utilisées pour calculer la gravitation gravitationnelle à N corp forces entre les particules. Bien que le calcul des forces du corps N soit un ajustement parfait pour un accélérateur, la construction de l'arbre est mieux faite sur un processeur capable de gérer le code ramifié avec une localité de données difficile. Un accélérateur cohérent en cache comme le MI250X peut lire la particule pour commencer le travail de frai pendant que le processeur AMD EPYC™ optimisé de 3e génération construit simultanément d'autres parties de l'arbre,éviter la synchronisation explicite et les blocages, et réduire les copies coûteuses de l'arbre

entier. La sérialisation fine rendue possible par la cohérence du cache peut améliorer l'utilisation et les performances par rapport à un accélérateur non cohérent. Étant donné que les accélérateurs cohérents de cache sont relativement nouveaux, l'AMD MI250X couplé au processeur AMD EPYC optimisé de 3e génération sera une plate-forme cruciale pour l'industrie explorer les possibilités et comprendre les avantages de la cohérence du cache.

Conclusion

L'ère de l'informatique **exascale** repoussera les limites de la découverte humaine dans le calcul scientifique et l'apprentissage automatique, débloquant de nouvelles innovations qui profiteront au monde entier. Les systèmes informatiques hétérogènes sont indéniablement la clé de ce bond en avant en informatique performance. Les GPU sont passés de simples pipelines matériels à fonction fixe à des accélérateurs entièrement programmables et à usage général qui sont un élément critique de tout système hétérogène.

L'architecture AMD CDNA™ 2 est une étape clé pour les accélérateurs et les systèmes hétérogènes - une avancée qui ouvrira les possibilités de calcul exascale. L'architecture AMD CDNA 2 améliore considérablement le débit de calcul, offrant près de 48 TFLOP/s de crête calcul théorique à double précision dans un seul accélérateur. Cette amélioration générationnelle jusqu'à 4 fois supérieure dans compute1 est atteint grâce à des améliorations aux unités de calcul et des innovations plus radicales dans la communication qui permettent de presque doubler les performances grâce aux AMD Technologie **Infinity Fabric**™ et emballage multi-puces et amélioration de l'évolutivité des nœuds complets.

Ces mêmes améliorations radicales permettent aux accélérateurs tels que l'AMD Instinct™ MI250X de non seulement partager de la mémoire avec un Processeur AMD EPYC™ de 3e génération, mais pour offrir une cohérence totale du cache et agir comme un élément entièrement homologue d'un système hétérogène. Il s'agit d'un énorme pas en avant dans la programmation qui permet également des optimisations uniques et crée un banc d'essai pour l'industrie. L'avantage de la cohérence du cache sur de nombreuses charges de travail différentes. Mieux encore, les avantages de l'architecture AMD CDNA 2 seront facilement disponible pour les fournisseurs, les clients et l'ensemble de la communauté via l'écosystème open source AMD ROCm™.

Sigles:

ACE - Asynchronous Compute Engine	ACE - Moteur de calcul asynchrone
AI - Artificial Intelligence	IA - Intelligence artificielle
AVC - Advanced Video Coding	AVC - Codage vidéo avancé
CPU - Central Processing Unit	CPU - Unité centrale de traitement
ESM - Extended Speed Mode	ESM - Mode vitesse étendue
FMA - Fused Multiply-Add	FMA - Fusion Multiply-Add
GCD - Graphics Compute Die	GCD - GPU Graphics Compute Die -
GPU - Graphics Processing Unit	Unité de traitement graphique
HEVC - High Efficiency Video Coding	HEVC - Codage vidéo haute efficacité
HPC - High Performance Computing	HPC - Calcul haute performance

LDS - Local Data Store

ML- Machine Learning

OAM - Open Compute Project Accelerator Module

PCIe - PCI-Express

SIMD - Single Instruction, Multiple Data

TDP - Thermal Design Power

TFLOPS - Trillions Floating Point Operations per Second

TOPS - Trillions Operations per Second

VCN - Video Codec Next

LDS - Magasin de données local

ML- Apprentissage automatique

OAM - Open Compute Project Accelerator Module

PCIe - PCI-Express

SIMD - Instruction unique, données multiples

TDP - Puissance de conception thermique

TFLOPS - Trillions d’opérations en virgule flottante par seconde