DSI – Calculateur Haute Performance

Cursus – André-Guy Bruneau M.Sc. IT – Octobre 2025

1. Introduction : L'Impératif Stratégique pour le Calcul Hétérogène Post-Exascale

1.1 Le Contexte : Convergence de la Simulation, de l'IA et du Calcul Quantique

Le paysage du calcul haute performance (HPC) est en pleine mutation. Historiquement cantonné au domaine de la simulation numérique pour la science et l'ingénierie, le HPC est aujourd'hui le moteur de révolutions technologiques majeures, notamment l'intelligence artificielle (IA) à grande échelle et l'émergence de paradigmes de calcul non classiques.[1] Les défis scientifiques et industriels les plus complexes de notre époque — qu'il s'agisse de la découverte de nouveaux médicaments, de la modélisation précise du changement climatique, de la conception de nouveaux matériaux ou de la garantie de la sécurité nationale — ne peuvent plus être résolus par une seule approche computationnelle.[4] Ils exigent une convergence profonde et native entre la simulation traditionnelle, l'analyse de données massives, l'apprentissage automatique et, de plus en plus, le calcul quantique.[6] Ce Dossier Système d'Information (DSI) présente la conception architecturale d'un supercalculateur de nouvelle génération conçu spécifiquement pour répondre à cette convergence, en intégrant de manière cohérente huit types de processeurs spécialisés au sein d'une plateforme unifiée.

1.2 Les Défis Fondamentaux : Surmonter le Mur Énergétique et le Goulot d'Étranglement des Données

La progression vers des puissances de calcul toujours plus élevées se heurte à des barrières physiques fondamentales. La fin de la loi de mise à l'échelle de Dennard vers 2004 a mis un terme à l'augmentation exponentielle des fréquences d'horloge des processeurs, nous confrontant directement au "mur de l'énergie".[7] Les agences de financement, comme le Département de l'Énergie américain, imposent désormais des contraintes strictes sur la consommation électrique des systèmes exaflopiques, typiquement de l'ordre de 20 à 30 mégawatts (MW), afin de garantir leur viabilité opérationnelle et environnementale.[8] Parallèlement, un second défi majeur est apparu : le goulot d'étranglement des données. Dans les architectures modernes, le coût énergétique et la latence associés au déplacement des données entre la mémoire et les unités de calcul dépassent désormais ceux des opérations arithmétiques elles-mêmes.[7] La maximisation de la localité des données et la minimisation des transferts sont devenues des préoccupations architecturales de premier ordre.[9]

1.3 Philosophie de Conception : Une Architecture Modulaire et Spécialisée pour des Charges de Travail Diversifiées

Face à ces contraintes, l'approche monolithique et homogène, où un supercalculateur est constitué de milliers de nœuds identiques, montre ses limites. Une architecture "taille unique" ne peut répondre efficacement à la diversité croissante des charges de travail. La philosophie de conception qui sous-tend ce projet est donc celle de la **spécialisation hétérogène**.[10] Le principe est simple mais puissant : chaque tâche, ou segment de tâche, d'un workflow applicatif doit être exécutée sur le type de processeur qui offre le meilleur rapport performance/watt pour cette tâche spécifique.[12] Cette approche permet de s'attaquer simultanément au mur de l'énergie, en utilisant des processeurs plus efficaces, et

au goulot d'étranglement des données, en rapprochant le calcul spécialisé des données pertinentes. L'architecture proposée est la matérialisation de cette philosophie, intégrant des processeurs généralistes (CPU), des accélérateurs massivement parallèles (GPU, TPU), des logiques reconfigurables (FPGA), des unités de traitement de données (DPU), des accélérateurs d'inférence (NPU), et des processeurs émergents probabilistes (TSU) et quantiques (QPU) au sein d'un ensemble cohérent et orchestré.

La transition vers une hétérogénéité profonde n'est pas une simple évolution, mais un véritable changement de paradigme. Le calcul scientifique ne se mesure plus uniquement en opérations en virgule flottante par seconde (FLOPS). Les limites physiques du scaling des processeurs traditionnels ont forcé l'industrie à innover dans des directions radicalement différentes, donnant naissance à une taxonomie de processeurs spécialisés.[7] Les GPU et TPU excellent en opérations tensorielles (TOPS) pour l'IA [14], les DPU sont optimisés pour les opérations d'entrée/sortie de données (IOPS) et le traitement de paquets [15], et des technologies de rupture comme les TSU introduisent le calcul probabiliste natif.[13] Le développement indépendant de ces puces par des acteurs majeurs (NVIDIA, Google) et des startups (Extropic) n'est pas une coïncidence ; il reflète la reconnaissance que la nature même du calcul s'est diversifiée. Par conséquent, l'architecture d'un supercalculateur de pointe doit être conçue non plus comme une collection monolithique de nœuds identiques, mais comme un "système de systèmes" intégré, capable de répondre à cette nouvelle pluralité computationnelle.

Cette complexité matérielle croissante déplace inévitablement le principal défi de l'ingénierie vers la couche logicielle. Construire le matériel le plus rapide n'est plus le seul objectif ; le défi majeur est de développer le logiciel capable de l'exploiter intelligemment. La gestion de clusters hétérogènes est déjà un problème NP-difficile, comme en témoigne l'existence de plateformes d'orchestration sophistiquées telles que Lenovo LiCO ou Adaptive Computing Moab.[16] En passant de deux ou trois types de processeurs à huit, dont certains sont non déterministes (TSU, QPU), la complexité de l'ordonnancement et de la gestion des ressources explose. Les recherches actives sur les modèles de programmation unifiés et les workflows hybrides confirment que l'abstraction de cette complexité est un enjeu crucial pour la communauté.[18] L'avantage compétitif de ce supercalculateur ne résidera donc pas seulement dans sa configuration matérielle unique, mais de manière prépondérante dans sa capacité à mapper dynamiquement et efficacement les workflows sur cette configuration via un orchestrateur intelligent, potentiellement lui-même basé sur l'IA.[21] Le succès de cette architecture dépendra à 80% de la robustesse et de l'intelligence de sa pile logicielle.

2. Analyse Approfondie des Charges de Travail Cibles

Pour justifier la complexité et le coût d'une architecture hétérogène à huit processeurs, il est impératif de définir précisément les classes de problèmes qu'elle vise à résoudre. Cette section caractérise quatre domaines d'application stratégiques, chacun présentant des exigences computationnelles distinctes qui légitiment l'intégration de processeurs spécialisés.

2.1 Calcul Scientifique et Simulation Numérique

Le calcul scientifique traditionnel reste une charge de travail fondamentale pour le HPC. Il consiste principalement à résoudre numériquement des systèmes d'équations, souvent des équations aux dérivées partielles, qui modélisent des phénomènes physiques complexes.[23] Ces charges de travail peuvent être classées en fonction de leurs modèles de communication, ce qui a un impact direct sur les exigences architecturales.[24]

- **Charges Fortement Couplées (Tightly-coupled) :** Ces applications sont caractérisées par des communications inter-processus fréquentes, à faible latence et synchronisées, généralement implémentées via le protocole MPI (Message

Passing Interface). Chaque nœud de calcul travaille sur une partie du problème (par exemple, un sous-domaine d'une grille de simulation) et doit échanger régulièrement des données de frontières avec ses voisins. La performance globale est limitée par le nœud le plus lent et la latence du réseau. Les exemples typiques incluent la dynamique des fluides numérique (CFD), la modélisation météorologique et climatique, et les simulations de traitement de données sismiques.[24] Ces charges de travail exigent un réseau d'interconnexion à très haute bande passante (supérieure à 100 Gbit/s) et à latence ultra-faible, tel qu'InfiniBand avec prise en charge de l'accès direct à la mémoire à distance (RDMA).[24]

- **Charges Faiblement Couplées (Loosely-coupled) :** Souvent qualifiées de "gênantes parallèles" (embarrassingly parallel), ces applications consistent en un grand nombre de tâches indépendantes qui nécessitent une communication minimale, voire nulle, entre elles. La performance est principalement déterminée par le débit de calcul brut de l'ensemble du système. Ces charges de travail sont intrinsèquement plus tolérantes aux pannes, car la défaillance d'une tâche n'affecte généralement pas les autres. Les exemples incluent les simulations de Monte Carlo pour l'analyse de risques financiers, le séquençage et l'analyse génomique à haut débit, et le rendu d'images de synthèse (CGI) où chaque image peut être calculée indépendamment.[24]

2.2 Intelligence Artificielle à Grande Échelle

L'intelligence artificielle, et plus particulièrement l'apprentissage profond, est devenue une charge de travail prédominante dans le HPC, avec des exigences distinctes de celles de la simulation traditionnelle.[2]

- **Entraînement de Modèles de Fondation :** L'entraînement de modèles de très grande taille (LLM, modèles de diffusion) avec des centaines de milliards, voire des trillions de paramètres, représente l'une des charges de travail les plus exigeantes. Ce processus implique un parallélisme de données et de modèles à une échelle massive, nécessitant une communication inter-accélérateurs à très haute vitesse et une énorme capacité mémoire pour stocker les poids, les gradients et les états de l'optimiseur.[30] Les calculs sont majoritairement des multiplications de matrices denses, souvent effectuées en précision mixte (par exemple, FP16, BF16, et plus récemment FP8) pour accélérer le débit des cœurs Tensor spécialisés et réduire l'empreinte mémoire.[33] Des interconnexions comme NVIDIA NVLink sont essentielles pour ce type de charge de travail.
- **Inférence à Haut Débit et Basse Latence :** Une fois un modèle entraîné, son déploiement pour des applications en temps réel (inférence) présente un ensemble différent de défis. Pour des services interactifs comme les agents conversationnels, la latence de réponse est une métrique critique. Pour des applications à grande échelle comme la détection de fraude ou la recommandation de contenu, le débit (inférences par seconde) et le coût par inférence sont les facteurs dominants.[27] L'inférence bénéficie souvent de calculs en plus basse précision (par exemple, INT8) pour maximiser le débit et l'efficacité énergétique.

2.3 Analyse de Données Haute Performance (HPDA)

L'HPDA représente la convergence du HPC et de l'analyse Big Data. L'objectif est d'appliquer la puissance de calcul parallèle des supercalculateurs pour extraire des informations exploitables de jeux de données massifs (pétaoctets et au-delà) en temps quasi réel.[9]

- **Défis :** Les systèmes HPDA doivent faire face aux "3V" du Big Data : le **Volume** (quantité de données), la **Vélocité** (vitesse à laquelle les données sont générées et doivent être traitées, souvent en streaming), et la **Variété** (formats de données hétérogènes, structurés et non structurés).[35] Contrairement aux simulations HPC traditionnelles où les données sont souvent générées en interne, les charges de travail HPDA impliquent l'ingestion et le traitement de données provenant de sources externes. La localité des données est donc primordiale pour éviter que les opérations d'entrée/sortie (I/O) ne deviennent le principal goulot d'étranglement, éclipsant le temps de calcul.[9]

- **Exemples :** L'analyse en temps réel de données de séquençage génomique pour la médecine de précision, le traitement de vastes ensembles de données de marché pour l'analyse de risques financiers, et l'analyse de flux de données provenant de capteurs IoT pour la maintenance prédictive dans l'industrie.[26]

2.4 Calcul Hybride Quantique-Classique

À l'ère actuelle des processeurs quantiques bruités de taille intermédiaire (NISQ), l'atteinte d'un "avantage quantique" pratique repose sur des modèles de calcul hybrides. Dans ce paradigme, le processeur quantique (QPU) n'est pas un ordinateur autonome mais fonctionne comme un co-processeur ou un accélérateur spécialisé pour un système HPC classique.[40]

- **Workflow Typique (Algorithmes Variationnels) :** Le workflow le plus courant pour les algorithmes quantiques variationnels implique une boucle itérative entre le processeur classique et le QPU. Le processeur classique (CPU, assisté parfois par un GPU) exécute une boucle d'optimisation. À chaque itération, il : 1) prépare un circuit quantique avec un ensemble de paramètres ; 2) envoie ce circuit au QPU pour exécution ; 3) récupère les résultats des mesures quantiques ; 4) calcule une fonction de coût basée sur ces résultats ; et 5) utilise un algorithme d'optimisation classique pour proposer un nouvel ensemble de paramètres visant à minimiser ce coût.[41]
- **Exemples :** Les algorithmes emblématiques de cette approche sont le *Variational Quantum Eigensolver* (VQE), utilisé en chimie quantique pour trouver l'état fondamental de molécules, et le *Quantum Approximate Optimization Algorithm* (QAOA), appliqué à des problèmes d'optimisation combinatoire.[41]

2.5 Matrice de Correspondance Charges de Travail-Processeurs

Le tableau suivant synthétise l'adéquation de chaque type de processeur de notre architecture pour les différentes charges de travail identifiées. Il sert de justification fondamentale à la conception hétérogène du système.

Charge de Travail	**CPU**	**GPU**	**FPGA**	**DPU**	**NPU**	**TPU**	**TSU**	**QPU**
Simulation (Fortement Couplée)	Support	**Primaire**	Secondaire	Support	N/A	N/A	N/A	N/A
Simulation (Faiblement Couplée)	Secondaire	**Primaire**	Secondaire	N/A	N/A	N/A	N/A	N/A
Entraînement IA (Grands Modèles)	Support	**Primaire**	N/A	Support	N/A	Secondaire	N/A	N/A
Inférence IA (Basse	Support	Secondaire	Secondaire	Support	**Primaire**	Secondaire	N/A	N/A

Latence)								
Analyse de Données (HPDA)	Secondaire	Secondaire	Secondaire	**Primaire**	Secondaire	N/A	N/A	N/A
Calcul Hybride (VQE/QAOA)	**Primaire**	Secondaire	**Primaire**	N/A	N/A	N/A	N/A	**Primaire**
IA Générative (Échantillonnage)	Support	Secondaire	N/A	N/A	N/A	N/A	**Primaire**	N/A

*Légende : **Primaire** = Rôle principal, optimal pour la tâche ; **Secondaire** = Rôle de support ou efficace pour des sous-tâches spécifiques ; **Support** = Rôle d'infrastructure ou de facilitation ; **N/A** = Non applicable ou non optimal.*

Ce tableau met en évidence qu'aucun processeur n'est universel. Le GPU est dominant pour la simulation et l'entraînement IA, mais le CPU reste indispensable pour l'orchestration, le DPU pour la gestion des données, le NPU pour l'inférence efficace, et les processeurs émergents (FPGA, TSU, QPU) ouvrent la voie à de nouvelles capacités pour le calcul hybride et l'IA de nouvelle génération.

3. Architecture Détaillée des Nœuds de Calcul Hétérogènes

Le nœud de calcul est l'unité fondamentale du supercalculateur. Sa conception hétérogène est le cœur de notre proposition architecturale. Chaque composant a été sélectionné pour son rôle spécifique, avec des spécifications de pointe et après une analyse comparative rigoureuse.

3.1 Le Processeur Central (CPU) : Orchestration et Tâches Séquentielles

- **Rôle Architectural :** Dans un nœud aussi complexe, le CPU n'est plus le principal moteur de calcul, mais le "chef d'orchestre". Ses fonctions critiques sont : l'exécution du système d'exploitation (Linux) et de l'hyperviseur, la gestion des pilotes pour l'ensemble des sept autres co-processeurs, l'orchestration des flux de données entre les composants via le bus PCIe et les interconnexions spécialisées, et l'exécution des portions de code intrinsèquement séquentielles ou de contrôle qui ne peuvent être parallélisées (conformément à la loi d'Amdahl).[13] Il est également essentiel pour piloter les boucles d'optimisation classiques dans les workflows hybrides quantiques-classiques.[40]
- Sélection et Justification : AMD EPYC 9754 **-** La sélection se porte sur le processeur AMD EPYC 9754. La justification repose sur deux critères prépondérants pour son rôle d'orchestrateur. Premièrement, sa densité de cœurs (128 cœurs physiques) est sans équivalent sur le marché, offrant une capacité massive de traitement parallèle pour gérer simultanément les multiples files d'attente, interruptions et contextes logiciels générés par les accélérateurs.44 Deuxièmement, il offre un nombre très élevé de voies PCIe 5.0, ce qui est une condition sine qua non pour connecter l'ensemble des sept accélérateurs sans créer de goulot d'étranglement au niveau de l'interconnexion interne du nœud.44

- **Spécifications Techniques (AMD EPYC 9754) :**
  - **Cœurs / Threads :** 128 cœurs / 256 threads (architecture "Zen 4c")
  - **Fréquence :** 2.25 GHz (base) / 3.1 GHz (boost maximum)
  - **Cache L3 :** 256 Mo
  - **Mémoire :** 12 canaux DDR5-4800, supportant jusqu'à 6 To par socket
  - **Interconnexion :** Jusqu'à 160 voies PCIe 5.0 (en configuration 2P)
  - TDP (Thermal Design Power) : 360 W (configurable de 320 W à 400 W)[44]
- **Analyse Comparative :** Le principal concurrent est l'Intel Xeon Platinum 8490H. Bien qu'il s'agisse d'un processeur très performant pour les charges de travail traditionnelles, ses 60 cœurs le placent en net désavantage pour la tâche d'orchestration massive requise ici.[47] Dans ce contexte architectural, le débit total de gestion des tâches parallèles (nombre de cœurs) et la bande passante d'E/S (voies PCIe) sont des métriques plus importantes que la performance brute d'un cœur unique.[50]

Caractéristique	AMD EPYC 9754	Intel Xeon Platinum 8490H
Cœurs / Threads	128 / 256	60 / 120
Fréquence (Base/Boost)	2.25 / 3.1 GHz	1.9 / 3.5 GHz
Cache L3	256 Mo	112.5 Mo
Mémoire	12 canaux DDR5-4800	8 canaux DDR5-4800
Voies PCIe (Version/Nombre)	5.0 / jusqu'à 160 (2P)	5.0 / 80
TDP	360 W	350 W

3.2 L'Accélérateur Parallèle Massif (GPU) : Entraînement IA et Simulations Numériques

- **Rôle Architectural :** Le GPU est le moteur principal pour les charges de travail nécessitant un parallélisme de données massif. Son architecture, composée de milliers de cœurs, est optimisée pour deux domaines clés : les calculs matriciels en précision mixte (FP16/BF16/FP8) qui sont au cœur de l'entraînement des modèles d'IA, et les calculs en double précision (FP64) requis pour la simulation scientifique de haute précision.[52]
- Sélection et Justification : NVIDIA H200 - Le choix se porte sur le NVIDIA H200. Cette décision est motivée par une combinaison de facteurs : une performance de premier plan en FP64 pour la science, des cœurs Tensor de dernière génération pour l'IA, et surtout une capacité mémoire HBM3e (141 Go) et une bande passante mémoire (4.8 To/s) exceptionnelles, qui sont des facteurs limitants pour l'entraînement des modèles de fondation les plus volumineux.54 De plus, l'écosystème logiciel NVIDIA CUDA, avec ses bibliothèques (cuDNN, cuBLAS), ses outils de profilage et sa large adoption par la communauté scientifique et IA, représente un avantage décisif en termes de maturité, de

performance et de productivité des développeurs.53

- **Spécifications Techniques (NVIDIA H200) :**
  - **Performance FP64 :** ~60 TFLOPS
  - **Mémoire :** 141 Go HBM3e
  - **Bande Passante Mémoire :** 4.8 To/s
  - **Interconnexion GPU-GPU :** NVLink 4ème génération, 900 Go/s
  - Précisions IA : FP8, FP16, BF16, INT8[54]
- **Analyse Comparative :** L'AMD Instinct MI300X est un concurrent extrêmement performant, offrant une performance FP64 brute potentiellement supérieure (~81.7 TFLOPS) et une capacité mémoire plus grande (192 Go HBM3).[55] Cependant, l'écosystème logiciel ROCm d'AMD, bien qu'en progression rapide, n'a pas encore atteint la maturité, l'étendue et le niveau d'optimisation de l'écosystème CUDA. Pour un système de production à grande échelle visant une compatibilité maximale avec les codes existants, la maturité logicielle de NVIDIA constitue un avantage stratégique qui l'emporte sur un avantage potentiel en performance brute.

Caractéristique	**NVIDIA H200**	**AMD Instinct MI300X**
Performance FP64 (pic)	~60 TFLOPS	Jusqu'à 81.7 TFLOPS
Performance FP16 (avec sparsité)	Compétitive	Jusqu'à 5.3 PFLOPS
Capacité Mémoire	141 Go HBM3e	192 Go HBM3
Bande Passante Mémoire	4.8 To/s	5.3 To/s
Bande Passante Interconnect	900 Go/s (NVLink)	896 Go/s (Infinity Fabric)
Écosystème Logiciel	Très Mature (CUDA)	En développement (ROCm)

3.3 Le Tissu Reconfigurable (FPGA) : Accélération Matérielle sur Mesure et Faible Latence

- **Rôle Architectural :** Le FPGA (Field-Programmable Gate Array) est l'incarnation de la flexibilité matérielle. Sa logique programmable permet de créer des circuits numériques sur mesure, optimisés pour un algorithme spécifique. Il excelle dans les domaines où les architectures rigides des CPU et GPU sont inefficaces : le traitement de flux de données en temps réel avec une latence déterministe et ultra-faible (par exemple, traitement de paquets réseau, trading financier), l'accélération d'algorithmes non standards (par exemple, en bio-informatique ou en cryptographie), et un rôle émergent crucial : le contrôle en temps réel et la correction d'erreurs des systèmes quantiques hybrides, où la synchronisation à la nanoseconde est indispensable.[57]
- Sélection et Justification : AMD (Xilinx) Versal Premium VP1902 **-** Ce SoC adaptatif est choisi pour sa capacité logique massive, la plus élevée de la gamme Versal, et sa connectivité de pointe. Ses 18.5 millions de cellules logiques permettent de prototyper et d'accélérer des systèmes numériques extrêmement complexes. Son grand nombre de

transceivers à très haute vitesse (jusqu'à 112 Gbps) est essentiel pour ingérer et traiter des flux de données à haut débit ou pour s'interfacer directement avec des systèmes de test et mesure ou des composants quantiques.61

- **Spécifications Techniques (AMD Versal Premium VP1902) :**
  - **Cellules Logiques Système :** 18.5 millions
  - **LUTs (Tables de correspondance) :** 8,460 K
  - **Mémoire sur puce :** 239 Mb Block RAM, 619 Mb UltraRAM
  - **Moteurs DSP :** 6,864
  - **Transceivers :** Jusqu'à 160, incluant 32 GTM (jusqu'à 112 Gbps PAM4) et 128 GTYP (jusqu'à 32.75 Gbps)
  - Processeurs intégrés : Dual-core Arm Cortex-A72, Dual-core Arm Cortex-R5F[61]
- **Analyse Comparative :** L'Intel Stratix 10 GX 2800 est un concurrent direct en termes de capacité logique brute.[63] Cependant, l'architecture Versal est plus qu'un simple FPGA ; c'est un SoC (System-on-Chip) adaptatif qui intègre des cœurs de processeurs Arm, des moteurs d'IA et un réseau sur puce (NoC) programmable. Cette intégration en fait une plateforme plus complète et plus puissante pour les applications de co-traitement complexes où la logique programmable doit interagir étroitement avec le logiciel de contrôle.

Caractéristique	AMD Versal VP1902	Intel Stratix 10 GX 2800
Cellules Logiques (K)	18,507	~2,753
LUTs (K)	8,460	933
Mémoire Bloc (Mb)	239	229
Moteurs DSP	6,864	5,760
Transceivers (Nombre total)	160	96
Vitesse Max Transceiver	112 Gbps	28.3 Gbps

3.4 L'Unité de Traitement de Données (DPU) : Déchargement des Infrastructures Réseau et Stockage

- **Rôle Architectural :** Le DPU est le "troisième pilier" de l'infrastructure du centre de données, agissant comme un processeur frontal pour gérer les tâches liées aux données. Son rôle est de décharger le CPU des fonctions d'infrastructure lourdes et répétitives, telles que la gestion de la pile réseau (virtualisation, pare-feu, équilibrage de charge), la virtualisation du stockage (NVMe-oF), et l'application de politiques de sécurité (chiffrement, isolation).[64] Dans notre architecture, il est également crucial pour le pré-traitement des données à la volée (par exemple, décompression, formatage) avant de les acheminer vers les accélérateurs d'IA, minimisant ainsi la charge sur le CPU et la latence de la pipeline de données.[66]
- Sélection et Justification : NVIDIA BlueField-3 B3220 **-** Le DPU NVIDIA BlueField-3 est sélectionné en raison de sa

synergie profonde avec l'écosystème NVIDIA. Il intègre des cœurs Arm performants et une panoplie d'accélérateurs matériels programmables via le SDK DOCA.68 Son avantage décisif dans notre architecture est sa prise en charge de la technologie GPUDirect Storage. Cette fonctionnalité permet au DPU de transférer des données depuis le réseau ou un stockage distant directement vers la mémoire des GPU, en contournant complètement le CPU et la mémoire système. Ce chemin de données direct est fondamental pour alimenter efficacement les GPU dans les charges de travail HPDA et d'entraînement IA à grande échelle.66

- **Spécifications Techniques (NVIDIA BlueField-3 B3220) :**
  - **Cœurs de Calcul :** 16 cœurs Arm v8.2 A78
  - **Mémoire :** 32 Go DDR5
  - **Connectivité Réseau :** 2 ports 200 Gb/s (Ethernet ou InfiniBand NDR200)
  - **Interface Hôte :** PCIe 5.0 x16
  - Accélérateurs : Moteurs pour chiffrement (IPsec/TLS), virtualisation (SR-IOV), stockage (NVMe-oF), RDMA (RoCE), GPUDirect.[68]
- **Analyse Comparative :** L'AMD Pensando Salina 400 est une alternative puissante, offrant une connectivité 400GbE et un moteur de traitement de paquets P4 hautement programmable, ce qui le rend très flexible pour les tâches de réseau définies par logiciel.[70] Cependant, pour une architecture dont le principal moteur de calcul est le GPU NVIDIA, l'intégration native et les optimisations de performance offertes par la combinaison BlueField-3 et GPUDirect créent une synergie architecturale supérieure, justifiant son choix malgré un débit de port potentiellement plus faible sur le modèle B3220.

Caractéristique	**NVIDIA BlueField-3 B3220**	**AMD Pensando Salina 400**
Cœurs de Calcul	16 x Arm A78	16 x Arm Neoverse N1
Mémoire	32 Go DDR5	Jusqu'à 128 Go DDR5
Débit Réseau	2 x 200 Gb/s	2 x 400 Gb/s (un seul actif via PCIe)
Moteur de Traitement	Accélérateurs matériels fixes + programmables	Moteur P4 programmable
Capacités de Déchargement	Réseau, Stockage, Sécurité	Réseau, Stockage, Sécurité
Intégration Écosystème	Excellente (GPUDirect, DOCA)	Bonne (P4 standard)

3.5 L'Unité de Traitement Neuronal (NPU) : Inférence IA Basse Consommation et en Temps Réel

- **Rôle Architectural :** Le NPU est un circuit intégré spécifique à une application (ASIC) conçu exclusivement pour l'inférence de réseaux de neurones avec une efficacité énergétique maximale (TOPS/Watt). Il est optimisé pour les calculs en basse précision (principalement INT8) et une architecture de mémoire qui minimise les mouvements de données pour les opérations d'IA courantes.[15] Son rôle dans le nœud est double : 1) exécuter des tâches d'inférence à très faible latence pour des applications en temps réel, et 2) fournir une capacité d'analyse IA "toujours active" sur des flux de données (par exemple, des données de monitoring de simulation) sans consommer les précieuses ressources des GPU.[73]
- Sélection et Justification : Architecture de type Qualcomm AI Engine (Snapdragon X2 Elite) **-** Le marché des NPU pour serveurs est encore émergent. La conception s'appuiera sur l'intégration d'un NPU de classe "client" haut de gamme via une carte d'extension PCIe. Le NPU Hexagon de Qualcomm, intégré dans la puce Snapdragon X2 Elite Extreme, est actuellement le leader en termes de performance brute annoncée, avec 80 TOPS en INT8.74 Cette performance élevée le rend idéal pour l'analyse en temps réel de multiples flux de données ou pour des tâches d'inférence complexes à faible latence.
- **Spécifications Techniques (Qualcomm AI Engine - basé sur X2 Elite Extreme) :**
  - **Performance NPU :** 80 TOPS (INT8)
  - **Architecture :** Qualcomm Hexagon NPU
  - **Composants :** Inclut des unités scalaires, vectorielles et tensorielles
  - Précisions supportées : Optimisé pour INT8 et autres formats de basse précision[74]
- **Analyse Comparative :** L'Intel NPU 3720, intégré dans les processeurs Arrow Lake, offre une performance de 13 TOPS.[78] Bien que son intégration dans l'écosystème x86 soit un avantage, sa performance brute est nettement inférieure à celle de la solution Qualcomm. Pour une architecture visant des capacités d'analyse en temps réel de pointe, le choix du NPU le plus performant est primordial.

Caractéristique	**Qualcomm AI Engine (X2 Elite)**	**Intel NPU 3720 (Arrow Lake)**
Performance (TOPS INT8)	80 TOPS	13 TOPS
Efficacité (TOPS/Watt)	Leader du marché (données non spécifiées)	Optimisé pour basse consommation
Précisions Supportées	INT8, FP16	INT8, FP16
Interface	Intégré dans SoC (à adapter pour PCIe)	Intégré dans CPU

3.6 L'Unité de Traitement Tensoriel (TPU) : Opérations Tensorielles à Très Grande Échelle

- **Rôle Architectural :** Le TPU est un autre type d'ASIC, développé par Google, spécifiquement optimisé pour les opérations tensorielles à très grande échelle. Son architecture, basée sur des "systolic arrays", est conçue pour effectuer des multiplications de matrices massives avec un débit et une efficacité énergétique extrêmes. Il est particulièrement performant pour l'entraînement et l'inférence de modèles d'apprentissage profond développés avec les frameworks de Google, TensorFlow et JAX.[14]
- Sélection et Justification : Google TPU v5p
  Le TPU v5p représente le summum de la performance pour les charges de travail optimisées pour l'écosystème Google. Il offre une puissance de calcul brute (459 TFLOPS BF16 par puce), une grande capacité de mémoire HBM à haute bande passante, et une topologie d'interconnexion 3D Torus qui garantit une excellente scalabilité pour l'entraînement de modèles distribués sur des milliers de puces.82 Son intégration dans le nœud via PCIe permet de dédier une ressource spécialisée et ultra-performante aux workflows qui en tireront le meilleur parti.
- **Spécifications Techniques (Google TPU v5p) :**
  - **Performance par Puce :** 459 TFLOPS (BF16) / 918 TOPS (INT8)
  - **Mémoire par Puce :** 95 Go HBM2e
  - **Bande Passante Mémoire :** 2,765 Go/s
  - **Interconnexion Puce-à-Puce (ICI) :** 4,800 Gbps
  - Architecture du TensorCore : 4 unités de multiplication de matrices (MXU)[83]
- **Intégration et Hétérogénéité Logicielle :** L'intégration d'un TPU, intimement lié à l'écosystème logiciel de Google (JAX, TensorFlow), au sein d'un système dont le principal accélérateur (GPU) est dominé par l'écosystème NVIDIA (CUDA), représente un défi d'intégration logicielle significatif.[81] Cependant, cette stratégie est validée par des acteurs industriels de premier plan comme Anthropic, qui utilisent une approche multi-cloud et multi-accélérateurs (GPU et TPU) pour entraîner leurs modèles.[86] Cette décision implique que notre supercalculateur doit être hétérogène non seulement au niveau matériel, mais aussi au niveau logiciel. La plateforme d'orchestration (décrite à la section 5) devra être capable de gérer des environnements conteneurisés distincts et d'orchestrer les flux de données entre des sous-systèmes exécutant des piles logicielles différentes. C'est un défi de complexité, mais il garantit que chaque composant d'un workflow peut être exécuté sur l'architecture matérielle et logicielle la plus performante pour lui.

3.7 L'Unité d'Échantillonnage Thermodynamique (TSU) : Calcul Probabiliste pour l'IA Générative

- **Rôle Architectural :** Le TSU représente une rupture technologique. C'est un processeur probabiliste et non déterministe. Contrairement aux processeurs classiques qui exécutent des instructions arithmétiques, le TSU est conçu pour **échantillonner directement** à partir de distributions de probabilités complexes.[87] De nombreux algorithmes d'IA générative, comme les modèles de diffusion, reposent fondamentalement sur des étapes d'échantillonnage stochastique. Le TSU vise à exécuter ces étapes de manière native et avec une efficacité énergétique potentiellement des milliers de fois supérieure à celle d'un GPU qui doit simuler le hasard par des calculs déterministes.[88]
- Sélection et Justification : Extropic TSU (basé sur le prototype X0) **-** L'intégration de cette technologie, bien qu'émergente, positionne notre architecture à l'avant-garde de la recherche sur l'IA éco-énergétique. Le TSU fonctionnera comme un co-processeur spécialisé, sur lequel les GPU déchargeront les étapes d'échantillonnage coûteuses des workflows d'IA générative.89 Le prototype X0 d'Extropic a validé les principes de base de cette

approche.90

- **Spécifications Conceptuelles :**
  - **Architecture :** Le TSU est un réseau de circuits probabilistes appelés "p-bits". Un p-bit est un circuit dont la tension de sortie fluctue aléatoirement entre un niveau haut (1) et un niveau bas (0), la probabilité de se trouver dans un état ou l'autre étant programmable via une tension de contrôle.[87]
  - **Fonctionnement :** En connectant un grand nombre de p-bits dont les probabilités dépendent de l'état de leurs voisins, le TSU implémente matériellement l'algorithme d'échantillonnage de Gibbs. Cela lui permet de tirer des échantillons de modèles énergétiques (EBMs), qui sont une représentation mathématique de distributions de probabilités complexes.[88]

3.8 Le Processeur Quantique (QPU) : Co-processeur pour Avantage Quantique Spécifique

- **Rôle Architectural :** Le QPU est un accélérateur destiné à une classe très spécifique de problèmes pour lesquels les algorithmes quantiques offrent un avantage de complexité, potentiellement exponentiel, par rapport aux meilleurs algorithmes classiques connus. Dans notre architecture, le QPU n'est pas un nœud de calcul autonome, mais une ressource accessible via le réseau et intégrée dans des workflows hybrides orchestrés par les processeurs classiques.[40]
- Sélection et Justification : IBM Heron (supraconducteur) et IonQ Forte (ions piégés)
  Le domaine du matériel quantique est en évolution rapide, avec plusieurs technologies de qubits concurrentes. Pour maximiser la polyvalence et l'accès aux avancées de la recherche, l'architecture doit prévoir l'intégration des deux approches les plus matures : les qubits supraconducteurs et les ions piégés.
  - **IBM Heron (révision r2)** offre une échelle plus grande (156 qubits), une intégration profonde avec l'écosystème logiciel open-source Qiskit, et une topologie de connectivité "heavy-hex" qui est optimisée pour les codes de correction d'erreurs.[92]
  - **IonQ Forte** (36 qubits) offre une connectivité "all-to-all" entre ses qubits, un avantage significatif pour les algorithmes qui nécessitent des intrications complexes entre des qubits non adjacents. Il revendique également des fidélités de porte parmi les plus élevées de l'industrie.[94]

Caractéristique	**IBM Heron r2**	**IonQ Forte**
Nombre de Qubits	156	36
Technologie de Qubit	Supraconducteur (Transmon)	Ion piégé (Ytterbium)
Connectivité	Réseau "Heavy-Hex"	All-to-All
Fidélité Porte 2Q (typique)	~99.6% - 99.9%	~99.6%
Écosystème Logiciel	Qiskit	Multiples (Qiskit, Cirq, etc.)

L'intégration se fera via des interfaces standardisées, permettant à l'orchestrateur de soumettre des tâches à l'un ou l'autre des QPU en fonction des exigences de l'algorithme.

4. Tissu d'Interconnexion et Architecture Mémoire Unifiée

Une architecture hétérogène aussi complexe ne peut fonctionner efficacement sans un tissu d'interconnexion sophistiqué qui permet une communication à haute bande passante et faible latence entre tous les composants. La stratégie mémoire doit également dépasser le modèle traditionnel de mémoire locale pour créer un espace d'adressage unifié et cohérent.

4.1 Communication Intra-Nœud et Inter-Nœuds

Trois technologies d'interconnexion complémentaires formeront le système nerveux du supercalculateur :

- **NVIDIA NVLink & NVSwitch :** Cette technologie propriétaire est la solution de choix pour la communication GPU-GPU. La 4ème génération de NVLink, utilisée avec les GPU H200, offre une bande passante bidirectionnelle de 900 Go/s par GPU.[96] Au sein d'un nœud, NVLink connecte directement les GPU entre eux, créant un pool de mémoire unifié pour des tâches comme l'entraînement de grands modèles d'IA. À l'échelle du rack et au-delà, les NVSwitch permettent de construire un tissu "all-to-all" non bloquant entre des centaines de GPU, ce qui est essentiel pour le parallélisme de modèle à très grande échelle.[97]
- **Compute Express Link (CXL) :** CXL est un protocole standard ouvert, construit sur la couche physique PCIe, qui révolutionne la communication entre le CPU et les périphériques. Son rôle est double et fondamental pour notre architecture.[100] Premièrement, via le protocole CXL.cache, il permet aux accélérateurs (FPGA, DPU, NPU) d'accéder à la mémoire du CPU de manière cohérente, comme s'il s'agissait de leur propre mémoire, éliminant des copies de données coûteuses.[101] Deuxièmement, via CXL.mem, il permet de créer des pools de mémoire désagrégée. Des boîtiers d'extension mémoire peuvent être partagés entre plusieurs nœuds, ce qui augmente considérablement l'utilisation de la mémoire et permet d'allouer dynamiquement de grandes capacités aux nœuds qui en ont besoin.[101] Bien que CXL introduise une latence supplémentaire par rapport à la DRAM locale (de l'ordre de 200 ns [103]), des études ont montré que les charges de travail GPU, par exemple, peuvent tolérer des latences de l'ordre de la microseconde sans dégradation significative des performances pour certaines applications.[104]
- **Réseau Principal (InfiniBand/Ethernet) :** Pour la communication à grande échelle entre les nœuds, deux options seront déployées en parallèle. Un réseau InfiniBand NDR (400 Gb/s) avec support RDMA restera la norme pour les applications HPC traditionnelles fortement couplées, où la latence la plus faible possible est critique.[27] En parallèle, un réseau Ethernet à 400 Gb/s avec support RoCE (RDMA over Converged Ethernet) sera déployé pour les charges de travail d'IA, de cloud et d'analyse de données, offrant une interopérabilité et une flexibilité accrues.[31] Le DPU BlueField-3, avec sa double connectivité, pourra s'interfacer avec les deux réseaux.

L'interconnexion n'est pas simplement un ensemble de câbles ; c'est une hiérarchie de communication. NVLink forme le "cluster de calcul" ultra-rapide des GPU. InfiniBand connecte les nœuds pour les tâches HPC traditionnelles. Mais c'est CXL qui agit comme le tissu conjonctif, le "système nerveux" interne du nœud hétérogène. Sans CXL, l'intégration de huit processeurs serait un cauchemar de transferts de données explicites, gérés par logiciel et passant par le CPU. CXL transforme ce problème de transfert de données en un problème de gestion de mémoire hiérarchique. Il permet à un FPGA ou un NPU de traiter des données résidant dans la mémoire principale du CPU comme si elles étaient locales, grâce à la cohérence de cache matérielle.[105] Il permet à un pool de mémoire externe d'être partagé de manière cohérente entre le CPU et plusieurs accélérateurs. C'est cette capacité à créer un espace mémoire cohérent et partagé qui transforme une collection de puces disparates en un système de calcul unifié et programmable.

4.2 Hiérarchie et Cohérence de la Mémoire

L'architecture mémoire sera organisée en une hiérarchie à plusieurs niveaux pour équilibrer performance, capacité et coût :

- **Niveau 0 (L0) - Mémoire sur puce :** La mémoire la plus rapide, HBM (High Bandwidth Memory) intégrée directement sur les puces GPU et TPU. Elle offre une bande passante de plusieurs téraoctets par seconde mais une capacité limitée (de l'ordre de 100-200 Go).
- **Niveau 1 (L1) - Mémoire Locale :** La mémoire DDR5 directement attachée aux sockets CPU. Elle offre une grande capacité (plusieurs téraoctets par nœud) avec une latence faible, mais une bande passante inférieure à celle de la HBM.
- **Niveau 2 (L2) - Mémoire Poolée CXL :** Des tiroirs d'extension mémoire connectés via CXL. Cette mémoire, basée sur des DIMM DDR5, forme un vaste pool partageable entre plusieurs nœuds. Sa latence est plus élevée que celle de la mémoire locale, mais elle offre une capacité et une flexibilité de configuration quasi illimitées, permettant de construire des "nœuds à mémoire géante" à la demande.[101]

La cohérence de cache à travers cette hiérarchie est la clé de sa performance. Le protocole CXL.cache garantit que les données mises en cache par un accélérateur (par exemple, un FPGA) à partir de la mémoire du CPU sont maintenues cohérentes avec le cache du CPU lui-même, en utilisant des protocoles de cohérence basés sur des états comme MESI.[100] Pour la communication CPU-GPU, des solutions émergentes comme l'interconnexion NVLink-C2C (Chip-to-Chip), présente dans les superchips NVIDIA Grace Hopper, montrent la voie vers une mémoire entièrement unifiée et cohérente à très haute bande passante entre CPU et GPU, un modèle que CXL vise à standardiser pour l'ensemble de l'écosystème.[75]

4.3 Comparaison des Technologies d'Interconnexion

Le tableau ci-dessous résume les rôles complémentaires des principales technologies d'interconnexion au sein d'un nœud.

Caractéristique	**PCIe 5.0**	**NVLink 4.0**	**CXL 3.0**
Bande Passante (x16, bidirectionnelle)	128 Go/s	900 Go/s (par GPU)	256 Go/s
Latence Typique	Plusieurs microsecondes	Dizaines de nanosecondes	~200 ns (au-dessus de la latence DRAM)
Support de Cohérence de Cache	Non (transferts DMA)	Oui (GPU-GPU via Unified Memory)	**Oui (CPU-Périphérique, Périphérique-Périphérique)**
Cas d'Usage Principal	Connexion de périphériques I/O (stockage, réseau)	Interconnexion GPU-GPU à très haute vitesse	**Expansion mémoire cohérente, pooling, accélération**

Ce tableau illustre clairement que ces technologies ne sont pas concurrentes mais synergiques. PCIe sert de couche physique de base, NVLink crée un cluster de calcul GPU ultra-rapide, et CXL fournit la couche de cohérence mémoire essentielle pour faire fonctionner l'ensemble du système hétérogène comme un tout unifié.

5. Écosystème Logiciel : Orchestration Intelligente et Programmation

Le matériel le plus puissant est inutile sans une pile logicielle capable de l'exploiter de manière efficace et productive. Pour une architecture d'une hétérogénéité sans précédent, l'écosystème logiciel est le composant le plus critique, responsable de l'abstraction de la complexité, de l'ordonnancement intelligent des tâches et de la gestion des flux de données.

5.1 Modèles de Programmation Unifiés

Le défi majeur pour les développeurs est d'éviter d'avoir à maîtriser huit paradigmes de programmation distincts et à réécrire leur code pour chaque type de processeur. L'objectif est d'atteindre la **portabilité du code** (écrire une fois, exécuter partout) et, idéalement, la **portabilité de la performance** (obtenir de bonnes performances sans réécriture majeure).[18]

- **Approches Standards :** Des standards ouverts comme **SYCL**, promu par le Khronos Group, et son implémentation de référence **oneAPI** par Intel, visent à fournir un modèle de programmation unique basé sur le C++ moderne pour cibler de manière transparente les CPU, les GPU et les FPGA à partir d'une seule base de code.[18] Cette approche sera privilégiée pour les composants supportés.
- **Approches Spécifiques :** Pour les accélérateurs hautement spécialisés, une abstraction complète est souvent contre-productive car elle masque les caractéristiques qui les rendent performants. Par conséquent, l'accès à ces processeurs se fera via leurs SDK natifs, encapsulés dans des bibliothèques appelables depuis le code principal :
  - **DPU :** Le framework **NVIDIA DOCA** pour la programmation des fonctions de réseau et de stockage.
  - **QPU :** Des bibliothèques comme **Qiskit** (pour IBM) ou des intégrations de fournisseurs (pour IonQ) pour définir et soumettre des circuits quantiques.
  - **TSU :** Des bibliothèques spécifiques (par exemple, thrml d'Extropic) pour définir les modèles énergétiques et lancer les processus d'échantillonnage.

5.2 Orchestration des Tâches et des Données (Workflows)

La gestion de l'exécution d'applications complexes sur cette architecture est la tâche de l'orchestrateur de workflows.

- **Modélisation des Workflows :** Les applications scientifiques et d'IA complexes sont naturellement modélisées sous forme de **Graphes Orientés Acycliques (DAGs)**. Dans ce modèle, les nœuds représentent les tâches de calcul (par exemple, une étape de simulation, une inférence de modèle, un transfert de données) et les arêtes représentent les dépendances de données et de contrôle entre ces tâches.[19]
- **Ordonnancement (Scheduling) :** L'ordonnanceur de bas niveau (par exemple, **Slurm** ou **Kubernetes**) est responsable de l'allocation des ressources physiques (nœuds, processeurs, mémoire) aux tâches. Pour les environnements hétérogènes, des heuristiques classiques comme **HEFT (Heterogeneous Earliest Finish Time)** peuvent être utilisées comme point de départ. HEFT priorise les tâches en fonction de leur chemin critique et les assigne au processeur qui permet le temps de fin d'exécution le plus précoce, en tenant compte des coûts de calcul et de communication.[20]
- **Orchestration Intelligente Assistée par IA :** Un ordonnanceur basé sur des heuristiques statiques est insuffisant pour gérer la dynamique et la complexité de notre système. Nous proposons la mise en place d'une couche d'orchestration

de plus haut niveau, s'appuyant sur des solutions comme **Lenovo LiCO** ou **Adaptive Computing Moab**.[16] Cette couche implémentera une **orchestration intelligente**, utilisant des techniques d'IA pour optimiser le placement des tâches. Un **Sélecteur d'Heuristique Dynamique (DHS)**, par exemple, pourrait analyser les caractéristiques d'un workflow (taille des données, type de calculs, dépendances) et choisir dynamiquement entre plusieurs métaheuristiques d'optimisation comme les **Algorithmes Génétiques (GA)** ou l'**Optimisation par Essaim Particulaire (PSO)** pour trouver le meilleur mappage tâches-processeurs. Ces algorithmes explorent un vaste espace de solutions pour optimiser des objectifs multiples tels que la réduction du temps total d'exécution (makespan), la minimisation des coûts ou la réduction de la consommation d'énergie.[21]

5.3 Pile Logicielle par Domaine

La gestion de la diversité logicielle sera assurée par la conteneurisation, qui permet d'isoler les environnements d'exécution et leurs dépendances complexes. Des technologies comme **Singularity** (privilégiée en HPC pour sa sécurité) ou Docker seront utilisées pour encapsuler les piles logicielles spécifiques à chaque domaine.[109]

- **Communication : MPI** (et ses implémentations comme Open MPI, MPICH) reste le standard de facto pour la communication inter-nœuds dans les applications scientifiques, en particulier pour les charges fortement couplées.[110]
- **Accélérateurs IA et Graphiques :**
  - **GPU :** La pile **NVIDIA CUDA**, incluant les bibliothèques cuDNN, cuBLAS, NCCL, et les frameworks d'apprentissage profond comme PyTorch et TensorFlow.
  - **TPU :** La pile logicielle de Google, centrée sur **JAX** et **TensorFlow**, utilisant le compilateur XLA.
  - **NPU :** Le **Qualcomm AI Stack** ou les bibliothèques **OpenVINO** d'Intel pour l'optimisation et le déploiement de modèles d'inférence.
- **Accélérateurs Spécifiques :**
  - **FPGA :** Les environnements de développement de haut niveau comme **AMD Vitis** ou **Intel oneAPI** pour la programmation en C++/OpenCL, en plus des langages de description matérielle traditionnels (VHDL/Verilog).
  - **DPU :** Le SDK **NVIDIA DOCA** pour programmer les moteurs d'accélération et les cœurs Arm du DPU.
  - **QPU :** Le SDK **Qiskit** pour l'interaction avec les processeurs IBM, et les API des fournisseurs pour d'autres plateformes.

L'orchestrateur intelligent sera responsable de déployer le bon conteneur sur le nœud approprié et de gérer les flux de données entre ces conteneurs, assurant ainsi une exécution fluide des workflows hétérogènes.

6. Modèles de Workflows Applicatifs Hybrides

Cette section illustre, à travers des exemples concrets, comment les différents types de processeurs collaborent au sein de workflows complexes pour résoudre des problèmes scientifiques et technologiques de pointe. Ces scénarios démontrent la synergie architecturale et la justification de l'intégration de chaque composant spécialisé.

6.1 Simulation Scientifique avec Analyse IA en Temps Réel

Ce workflow combine la simulation numérique traditionnelle avec l'analyse de données par IA pour un pilotage intelligent ("computational steering").

- **Scénario :** Simulation de la combustion dans un moteur-fusée. L'objectif est de détecter en temps réel l'apparition d'instabilités de combustion, qui peuvent être des phénomènes rares et difficiles à prédire avec les seuls modèles

physiques.

- **Déroulement du Workflow :**
  1. **Simulation Principale (GPU) :** Le code de simulation de dynamique des fluides (CFD), massivement parallèle et exigeant en calculs FP64, s'exécute sur les **GPU**. Il résout les équations de Navier-Stokes à chaque pas de temps pour modéliser l'écoulement des gaz et les réactions chimiques.[111]
  2. **Extraction et Envoi des Données (CPU/DPU) :** À intervalles réguliers (par exemple, tous les 100 pas de temps), le **CPU** orchestre l'extraction d'un "snapshot" des données de simulation (champs de pression, température, vitesse) depuis la mémoire du GPU. Le **DPU** peut être utilisé pour compresser ces données à la volée afin de réduire la charge sur le réseau interne.
  3. **Détection d'Anomalies (NPU) :** Les données extraites sont envoyées au **NPU**. Celui-ci exécute un modèle d'IA (par exemple, un auto-encodeur variationnel) pré-entraîné pour reconnaître les signatures de données correspondant à des régimes de combustion stables. Toute déviation significative par rapport à la reconstruction du modèle est identifiée comme une anomalie potentielle, signalant le début d'une instabilité.[111] Le NPU est idéal pour cette tâche de surveillance continue en raison de sa faible consommation d'énergie et de sa faible latence.
  4. **Boucle de Contrôle (CPU) :** Si le NPU déclenche une alerte, il en informe le **CPU**. Le CPU peut alors prendre une décision en temps réel : soit enregistrer un état détaillé de la simulation pour une analyse ultérieure, soit modifier dynamiquement les paramètres de la simulation en cours sur le GPU (par exemple, en affinant le maillage dans la zone d'instabilité) pour mieux capturer le phénomène.

6.2 Pipeline d'Inférence IA Accéléré par le Matériel

Ce workflow illustre un pipeline de traitement de données de bout en bout, optimisée pour l'inférence IA à très haut débit, typique des applications en périphérie de réseau ou dans le cloud.

- **Scénario :** Analyse en temps réel d'un flux vidéo pour la détection d'objets dans une ville intelligente.
- **Déroulement du Workflow :**
  1. **Ingestion et Sécurité (DPU) :** Le flux vidéo arrive sous forme de paquets réseau sur l'un des ports du **DPU**. Le DPU utilise ses accélérateurs matériels pour décharger le CPU des tâches de bas niveau : terminaison TLS/IPsec, inspection des paquets pour un pare-feu de niveau 4, et équilibrage de charge initial.[66]
  2. **Pré-traitement des Données (DPU) :** Les cœurs Arm et les accélérateurs du DPU effectuent ensuite le pré-traitement des images directement sur les données du flux : décodage vidéo, redimensionnement des images à la taille attendue par le modèle d'inférence, et normalisation des pixels (par exemple, conversion en tenseurs).[113]
  3. **Transfert Direct de Données (DPU → TPU/GPU) :** Grâce à des technologies comme GPUDirect (pour les GPU) ou des transferts DMA optimisés via CXL, les tenseurs d'images pré-traitées sont envoyés directement de la mémoire du DPU vers la mémoire HBM du **TPU** (ou du GPU), en contournant la mémoire système du CPU. Cela minimise la latence et libère les cycles du CPU.[114]
  4. **Inférence (TPU/GPU) :** Le **TPU**, avec son architecture systolic array, exécute le modèle de détection d'objets (par exemple, YOLO ou un Vision Transformer) à très haut débit, en traitant les images par lots.[116]
  5. **Post-traitement et Action (CPU) :** Les résultats de l'inférence (coordonnées des boîtes englobantes et classes des objets) sont renvoyés au **CPU**. Le CPU effectue un post-traitement léger (par exemple, suppression des détections non maximales) et déclenche les actions appropriées (par exemple, envoyer une alerte, stocker les métadonnées).

6.3 Workflow de Chimie Quantique (VQE)

Ce workflow illustre l'interaction complexe et à faible latence requise pour les algorithmes quantiques-classiques variationnels.

- **Scénario :** Calcul de l'énergie de l'état fondamental d'une petite molécule pour la découverte de médicaments.
- **Déroulement du Workflow :**
  1. **Boucle d'Optimisation Classique (CPU/GPU) :** Le **CPU** exécute la boucle principale de l'algorithme d'optimisation (par exemple, SPSA, L-BFGS). Il maintient l'état de l'optimiseur et, à chaque itération, génère un nouvel ensemble de paramètres (angles de rotation des portes quantiques) pour le circuit d'essai ("ansatz").[117] Le **GPU** peut être utilisé pour accélérer certaines parties de l'optimiseur classique si elles sont parallélisables.
  2. **Compilation et Contrôle en Temps Réel (FPGA) :** Les paramètres du circuit sont envoyés au **FPGA**. Le FPGA agit comme un contrôleur matériel en temps réel ultra-précis. Il traduit les portes quantiques et les paramètres abstraits en une séquence de signaux analogiques (impulsions micro-ondes ou laser) avec une synchronisation à la nanoseconde. Ces signaux sont envoyés au **QPU** pour manipuler l'état des qubits.[59]
  3. **Exécution Quantique (QPU) :** Le **QPU** exécute le circuit paramétré. L'état des qubits évolue selon les principes de la mécanique quantique. À la fin du circuit, une mesure est effectuée, projetant chaque qubit dans un état classique (0 ou 1). Ce processus est répété des milliers de fois ("shots") pour construire une distribution de probabilités des états de sortie.
  4. **Lecture et Agrégation (FPGA/CPU) :** Les résultats des mesures sont lus par l'électronique de contrôle, souvent gérée par le **FPGA** pour une lecture rapide. Les statistiques agrégées sont renvoyées au **CPU**.
  5. **Calcul de la Fonction de Coût (CPU) :** Le **CPU** utilise les statistiques de mesure pour calculer la valeur attendue de l'hamiltonien de la molécule, qui correspond à son énergie pour les paramètres donnés. Cette valeur est utilisée par l'optimiseur pour décider du prochain jeu de paramètres, et la boucle recommence.

6.4 IA Générative Efficace (Modèle de Diffusion)

Ce workflow démontre comment une tâche d'IA peut être décomposée et distribuée entre des processeurs déterministes et probabilistes pour une efficacité énergétique maximale.

- **Scénario :** Génération d'une image à haute résolution à partir d'un texte (prompt) en utilisant un modèle de diffusion latente.
- **Déroulement du Workflow :**
  1. **Encodage du Prompt (CPU/GPU) :** Le **CPU** reçoit le prompt textuel et le transmet à un encodeur de texte (souvent un modèle Transformer comme CLIP) qui s'exécute sur le **GPU** pour le convertir en un vecteur d'enchâssement numérique.
  2. **Processus de Dénaturation Itératif (Orchestration CPU) :** Le **CPU** orchestre le processus de dénaturation, qui consiste en une série d'étapes itératives. À chaque étape, un modèle (généralement un U-Net) prédit le bruit à retirer d'une image latente bruitée.
  3. **Étape de Prédiction (GPU) :** La partie déterministe de chaque étape, c'est-à-dire l'exécution du réseau de neurones U-Net pour prédire le bruit, est une tâche de calcul intensive qui est effectuée sur le **GPU**.[118]
  4. **Étape d'Échantillonnage (TSU) :** La partie stochastique de chaque étape, qui consiste à ajouter une petite quantité de bruit gaussien ou à échantillonner à partir d'une distribution de probabilités conditionnelle, est déchargée sur le **TSU**. Le TSU est conçu pour effectuer cette opération d'échantillonnage de manière native et avec une consommation d'énergie bien moindre qu'un GPU qui doit utiliser des générateurs de nombres pseudo-aléatoires.[89]
  5. **Alternance et Décodage Final (CPU/GPU) :** Le **CPU** gère l'alternance entre le GPU (pour la prédiction) et le TSU

(pour l'échantillonnage) à chaque étape du processus de dénaturation. Une fois le processus terminé, l'image latente "propre" finale est envoyée à un décodeur (VAE) s'exécutant sur le **GPU** pour la convertir en l'image finale en pixels.

7. Considérations Opérationnelles et de Sécurité

La conception d'un supercalculateur de cette envergure ne se limite pas à la sélection et à l'intégration de composants matériels et logiciels. Sa viabilité à long terme dépend de la prise en compte rigoureuse des contraintes opérationnelles telles que la consommation d'énergie, la fiabilité et la sécurité.

7.1 Gestion de l'Énergie et du Refroidissement

- **Le Défi du "Mur de l'Énergie" :** L'un des défis les plus importants pour les systèmes à l'échelle exaflopique et au-delà est de maintenir la consommation électrique totale dans une enveloppe soutenable, généralement fixée entre 20 et 30 MW.[8] Dépasser cette limite entraîne des coûts d'exploitation prohibitifs et des exigences d'infrastructure de centre de données irréalisables. Notre architecture, avec sa densité de calcul extrême, doit être conçue dès le départ avec l'efficacité énergétique comme priorité absolue.
- **Stratégies de Mitigation :**
  1. **Refroidissement Liquide Direct :** Tous les composants générant une forte densité de chaleur, notamment les CPU, GPU et TPU, seront équipés de systèmes de refroidissement liquide direct ("direct-to-chip liquid cooling"). Cette technologie est beaucoup plus efficace que le refroidissement par air pour évacuer la chaleur, ce qui permet aux processeurs de fonctionner à des fréquences plus élevées et de manière plus stable, tout en réduisant la consommation d'énergie globale du système de refroidissement.[31]
  2. **Efficacité Intrinsèque des ASICs :** Un pilier fondamental de notre stratégie énergétique est l'utilisation intensive de circuits intégrés spécifiques à une application (ASIC). Les NPU, TPU et TSU sont conçus pour exécuter leurs tâches respectives avec une efficacité énergétique (opérations par watt) bien supérieure à celle des processeurs généralistes (CPU) ou même des GPU.[15] En déchargeant les tâches sur ces accélérateurs spécialisés, la consommation globale du système pour un workflow donné est considérablement réduite.
  3. **Gestion Dynamique de l'Énergie :** La pile logicielle intègrera des outils avancés de gestion de l'énergie, tels que l'**Energy Aware Runtime (EAR)**, qui peut être couplé avec des ordonnanceurs comme Slurm.[16] Ces outils surveillent en temps réel la charge des applications et ajustent dynamiquement la fréquence et la tension des processeurs (DVFS - Dynamic Voltage and Frequency Scaling) pour minimiser la consommation d'énergie sans impacter la performance des chemins critiques.

7.2 Fiabilité et Tolérance aux Pannes

- **Le Défi de l'Échelle :** Un système composé de dizaines de milliers de nœuds et de millions de cœurs de calcul est confronté à une probabilité de défaillance matérielle non négligeable, voire quasi certaine sur une période d'exécution donnée. La défaillance d'un seul composant ne doit pas entraîner l'échec de l'ensemble d'une simulation ou d'un entraînement qui pourrait durer plusieurs jours ou semaines.[122]
- **Stratégies de Résilience :**
  1. **Checkpoint/Restart :** Des mécanismes de "checkpointing" réguliers seront implémentés au niveau du système et des applications. L'état complet d'un calcul sera périodiquement sauvegardé sur un système de fichiers parallèle robuste. En cas de défaillance d'un nœud, le calcul peut être redémarré à partir du dernier point de contrôle valide, en perdant seulement une fraction du temps de calcul.
  2. **Redondance des Composants :** L'infrastructure réseau (tissus InfiniBand et Ethernet) sera conçue avec une

topologie redondante (par exemple, Fat-Tree ou Dragonfly+) pour qu'il n'y ait pas de point de défaillance unique. La défaillance d'un commutateur ou d'un lien n'isolera pas de larges portions du système.

3. **Algorithmes Tolérants aux Pannes :** Encourager et supporter le développement d'algorithmes qui sont intrinsèquement résilients à la perte de nœuds de calcul. C'est particulièrement pertinent pour les charges de travail faiblement couplées et certains algorithmes d'IA (par exemple, l'entraînement de données asynchrone) où la perte d'un "worker" peut être gérée sans arrêter l'ensemble du processus.
4. **Surveillance Prédictive des Pannes :** Les flux de données de monitoring (températures, tensions, taux d'erreur) provenant de l'ensemble du système seront analysés en temps réel par des modèles d'IA s'exécutant sur les NPU. L'objectif est de prédire les pannes imminentes de composants (disques, modules de mémoire, alimentations) avant qu'elles ne se produisent, permettant une maintenance proactive.

7.3 Sécurité de l'Infrastructure

- **Le Défi des Environnements Ouverts :** Les supercalculateurs destinés à la recherche scientifique fonctionnent souvent dans des environnements ouverts, où la collaboration et le partage de données sont encouragés. Ce modèle est en conflit direct avec les approches de sécurité traditionnelles basées sur un périmètre fortifié ("castle-and-moat").[124] Dans ce contexte, la priorité n'est pas seulement la confidentialité, mais surtout l'**intégrité du calcul** (garantir que les résultats ne sont pas altérés, accidentellement ou malicieusement) et la **provenance des données** (pouvoir tracer l'origine et les transformations de chaque donnée).[124]
- **Stratégies de Sécurité :**
  1. **Architecture "Zero Trust" :** Aucun composant, utilisateur ou flux de données n'est considéré comme fiable par défaut, même s'il se trouve à l'intérieur du périmètre du centre de données. Chaque accès à une ressource doit être authentifié, autorisé et chiffré.
  2. **Sécurité Accélérée par le Matériel (DPU) :** Les DPU joueront un rôle central dans la mise en œuvre de cette architecture. Ils utiliseront leurs capacités matérielles pour :
     - **Isoler les Workloads :** En gérant la virtualisation du réseau, les DPU peuvent créer des micro-segments pour chaque application ou utilisateur, empêchant tout mouvement latéral dans le réseau en cas de compromission d'un nœud.
     - **Chiffrement à la Volée :** Le trafic réseau peut être chiffré (IPsec/TLS) par les accélérateurs cryptographiques du DPU à la vitesse de la ligne, sans impacter les performances du CPU.
     - **Racine de Confiance Matérielle :** Les DPU fourniront des fonctionnalités de "secure boot" et d'attestation à distance, garantissant que seuls des logiciels validés et non altérés s'exécutent sur les nœuds de calcul.[65]
  3. **Traçabilité et Provenance :** La pile logicielle d'orchestration intègrera des mécanismes pour enregistrer la provenance de toutes les données et de tous les calculs. Chaque résultat produit sera accompagné de métadonnées décrivant précisément le code, les données d'entrée, la version des logiciels et la configuration matérielle utilisés pour le générer, garantissant ainsi la reproductibilité et la vérifiabilité des résultats scientifiques.

8. Conclusion et Recommandations Stratégiques

8.1 Synthèse de l'Architecture Proposée

Ce Dossier Système d'Information a présenté la conception d'un supercalculateur HPC hétérogène de nouvelle génération, une architecture conçue pour répondre aux défis computationnels les plus complexes à l'intersection de la simulation scientifique, de l'intelligence artificielle et du calcul quantique. La philosophie fondamentale de cette conception est la **spécialisation hétérogène**, une réponse directe aux contraintes physiques du "mur de l'énergie" et du goulot

d'étranglement des données.

En intégrant huit types de processeurs spécialisés — CPU, GPU, FPGA, DPU, NPU, TPU, TSU et QPU — cette architecture ne se contente pas d'agréger de la puissance de calcul. Elle crée un écosystème co-conçu où chaque composant matériel est sélectionné pour son efficacité optimale sur une classe de tâches spécifique. Le CPU agit en chef d'orchestre, le GPU en moteur de calcul parallèle, le DPU en gestionnaire d'infrastructure, et les autres accélérateurs (FPGA, NPU, TPU, TSU, QPU) fournissent des capacités sur mesure pour des charges de travail allant de la logique reconfigurable à l'inférence IA basse consommation, en passant par le calcul probabiliste et quantique.

Le succès de cette intégration matérielle repose sur un tissu d'interconnexion multicouche (NVLink, CXL, InfiniBand) et une pile logicielle sophistiquée. CXL, en particulier, joue un rôle de "système nerveux" en permettant une mémoire cohérente et partageable entre les composants hétérogènes. Au sommet, un orchestrateur de workflows intelligent, assisté par l'IA, est le cerveau du système, responsable du mappage dynamique des tâches sur les ressources les plus appropriées.

En somme, cette architecture n'est pas simplement une collection de matériel de pointe, mais un système holistique où le matériel, le logiciel et les modèles de workflows sont pensés ensemble pour maximiser la performance, l'efficacité énergétique et la flexibilité scientifique.

8.2 Feuille de Route pour le Déploiement

Le déploiement d'un système d'une telle complexité doit être abordé de manière pragmatique et phasée, afin de gérer les risques technologiques et de permettre une montée en puissance progressive des capacités. La feuille de route suivante est recommandée :

- **Phase 1 : Construction du Socle HPC et IA (Mois 1-18)**
  - **Objectif :** Déployer un système de production robuste pour les charges de travail HPC et IA conventionnelles.
  - **Composants :** Mettre en place les nœuds de calcul avec le socle de base : **CPU** (AMD EPYC), **GPU** (NVIDIA H200), **DPU** (NVIDIA BlueField-3) et **FPGA** (AMD Versal).
  - **Infrastructure :** Déployer le réseau InfiniBand/Ethernet, le système de fichiers parallèle, et l'infrastructure de refroidissement liquide.
  - **Logiciel :** Mettre en place la première version de l'orchestrateur de workflows, basé sur Slurm avec des heuristiques de type HEFT, et supporter les piles logicielles matures (CUDA, MPI, Vitis).
- **Phase 2 : Intégration des Accélérateurs IA Spécialisés (Mois 19-30)**
  - **Objectif :** Étendre les capacités du système pour l'inférence à haute efficacité et les workflows optimisés pour les écosystèmes logiciels spécifiques.
  - **Composants :** Intégrer les nœuds équipés de **NPU** et de **TPU**.
  - **Logiciel :** Faire évoluer l'orchestrateur pour gérer la coexistence de multiples écosystèmes logiciels (CUDA, JAX/TensorFlow) via la conteneurisation. Développer les premières versions des politiques d'ordonnancement assistées par IA.
- **Phase 3 : Déploiement des Technologies de Rupture (Mois 31-48)**
  - **Objectif :** Positionner le supercalculateur à l'avant-garde de la recherche en informatique en offrant un accès à des paradigmes de calcul émergents.
  - **Composants :** Intégrer les **TSU** et les **QPU** en tant que ressources expérimentales. L'accès à ces composants se fera via des files d'attente dédiées et des API spécifiques gérées par l'orchestrateur.
  - **Logiciel :** Développer des workflows hybrides avancés qui combinent ces nouvelles ressources avec les

composants classiques, en se concentrant sur les cas d'usage décrits dans la section 6.

8.3 Axes de Recherche et Développement Futurs

Le déploiement de cette architecture n'est pas une fin en soi, mais le début d'un programme de recherche et de développement continu, essentiel pour exploiter pleinement son potentiel. Les axes prioritaires sont :

1. **Orchestration de Workflows Assistée par IA :** La recherche doit se concentrer sur le développement d'algorithmes d'ordonnancement qui apprennent automatiquement les caractéristiques des applications et les performances des différents processeurs. L'utilisation de l'apprentissage par renforcement pour optimiser dynamiquement les décisions de placement en fonction d'objectifs en temps réel (performance, énergie, coût) est une voie particulièrement prometteuse.
2. **Modèles de Programmation Unifiés pour une Hétérogénéité Radicale :** Alors que SYCL/oneAPI offre une solution pour les CPU/GPU/FPGA, il n'existe pas de paradigme unifié pour intégrer des processeurs non déterministes comme les TSU et les QPU. La recherche sur des langages et des compilateurs de plus haut niveau, capables d'abstraire ces différentes sémantiques de calcul, sera cruciale pour la productivité des développeurs.
3. **Co-conception d'Algorithmes Hybrides :** Le plein potentiel de cette architecture ne sera atteint que lorsque les scientifiques et les développeurs commenceront à concevoir de nouveaux algorithmes qui tirent parti nativement de la synergie entre les différents types de processeurs. Il est essentiel de favoriser la recherche sur des algorithmes qui décomposent les problèmes en sous-tâches déterministes, probabilistes, logiques et quantiques, et qui exploitent les chemins de données à faible latence entre les différents accélérateurs.

En investissant dans cette architecture visionnaire et en poursuivant activement ces axes de R&D, l'institution se positionnera non seulement comme un leader mondial du calcul haute performance, mais aussi comme un pionnier dans la définition de la prochaine ère du calcul scientifique.

Ouvrages cités

1. Qu'est-ce que le calcul haute performance (HPC) ? – Explication du ..., dernier accès : octobre 31, 2025, https://aws.amazon.com/fr/what-is/hpc/
2. Faire évoluer les charges de travail d'IA dans un environnement HPC - Intel, dernier accès : octobre 31, 2025, https://www.intel.fr/content/www/fr/fr/high-performance-computing/hpc-artificial-intelligence.html
3. High Performance Computing (HPC) and AI - IBM, dernier accès : octobre 31, 2025, https://www.ibm.com/think/topics/hpc-ai
4. Solutions pour l'informatique Exascale - Altair, dernier accès : octobre 31, 2025, https://altairengineering.fr/exascale
5. Overview of the ECP - Exascale Computing Project, dernier accès : octobre 31, 2025, https://www.exascaleproject.org/about/
6. How to Scale AI Workloads within an HPC Environment - Intel, dernier accès : octobre 31, 2025, https://www.intel.com/content/www/us/en/high-performance-computing/hpc-artificial-intelligence.html
7. Supercalculateur exaflopique - Wikipédia, dernier accès : octobre 31, 2025, https://fr.wikipedia.org/wiki/Supercalculateur_exaflopique
8. The Conversation : "Calcul haute performance et ordinateurs ..., dernier accès : octobre 31, 2025, https://www.univ-grenoble-alpes.fr/actualites/the-conversation/sciences/the-conversation-calcul-haute-performance-et-ordinateurs-superpuissants-la-course-a-l-nbsp-exascale-nbsp--1178749.kjsp
9. Data Locality in High Performance Computing, Big Data, and Converged Systems: An Analysis of the

Cutting Edge and a Future System Architecture - MDPI, dernier accès : octobre 31, 2025, https://www.mdpi.com/2079-9292/12/1/53
10. Qu'est-ce que le calcul haute performance (HPC) et comment il fonctionne | Hivenet, dernier accès : octobre 31, 2025, https://compute.hivenet.com/fr/post/understanding-the-impact-of-high-performance-computing-hpc
11. Heterogeneous computing - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/Heterogeneous_computing
12. Heterogeneous Computing and Architecture | Multi-Processors ..., dernier accès : octobre 31, 2025, https://www.electronicsforu.com/technology-trends/heterogeneous-computing-architecture
13. What is heterogeneous compute? – Arm®, dernier accès : octobre 31, 2025, https://www.arm.com/glossary/heterogeneous-compute
14. Heterogeneous Computing: Integrating CPUs, GPUs, TPUs ... - IJFMR, dernier accès : octobre 31, 2025, https://www.ijfmr.com/papers/2019/3/48991.pdf
15. What is a Neural Processing Unit (NPU)? - IBM, dernier accès : octobre 31, 2025, https://www.ibm.com/think/topics/neural-processing-unit
16. Lenovo Intelligent Computing Orchestration (LiCO) - Lenovo Press, dernier accès : octobre 31, 2025, https://lenovopress.lenovo.com/lp0858.pdf
17. MOAB HPC SUITE - Adaptive Computing, dernier accès : octobre 31, 2025, https://adaptivecomputing.com/moab-hpc-suite/
18. Unified Programming Models for Heterogeneous High-Performance ..., dernier accès : octobre 31, 2025, https://www.researchgate.net/publication/369807373_Unified_Programming_Models_for_Heterogeneous_High-Performance_Computers
19. Workflow Models for Heterogeneous Distributed Systems - CEUR-WS.org, dernier accès : octobre 31, 2025, https://ceur-ws.org/Vol-3606/invited77.pdf
20. Workflow Simulation and Multi-Threading Aware Task ... - PEARL, dernier accès : octobre 31, 2025, https://pearl.plymouth.ac.uk/cgi/viewcontent.cgi?article=2296&context=secam-research
21. AI-Enhanced Hybrid Scheduling Framework for Scientific Workflows ..., dernier accès : octobre 31, 2025, https://www.researchgate.net/publication/396865202_AI-Enhanced_Hybrid_Scheduling_Framework_for_Scientific_Workflows_in_Intelligent_Cloud-Connected_Devices
22. AI- Enhanced Hybrid Scheduling Framework for ... - IGI Global, dernier accès : octobre 31, 2025, https://www.igi-global.com/ViewTitle.aspx?TitleId=392280&isxn=9798337375038
23. La saga des supercalculateurs - CEA DAM, dernier accès : octobre 31, 2025, https://www-dam.cea.fr/dam/wp-content/uploads/2021/11/CEA-TERA-210521-BD.pdf
24. A scientific approach to workload-aware computing on AWS | AWS ..., dernier accès : octobre 31, 2025, https://aws.amazon.com/blogs/hpc/a-scientific-approach-to-workload-aware-computing-on-aws/
25. Qu'est-ce que le calcul haute performance (HPC) - Oracle, dernier accès : octobre 31, 2025, https://www.oracle.com/africa-fr/cloud/hpc/what-is-hpc/
26. High Performance Computing (HPC) - Amazon AWS, dernier accès : octobre 31, 2025, https://aws.amazon.com/hpc/
27. What Is High-Performance Computing (HPC)? - IBM, dernier accès : octobre 31, 2025, https://www.ibm.com/think/topics/hpc
28. What is high performance computing (HPC) | Google Cloud, dernier accès : octobre 31, 2025, https://cloud.google.com/discover/what-is-high-performance-computing
29. The Role of GPUs in Training Models | Core Scientific, dernier accès : octobre 31, 2025, https://corescientific.com/resources/blog/unlocking-ai-potential-with-hpc-role-of-gpus-in-training-models/

30. Supermicro Expands Collaboration with NVIDIA, dernier accès : octobre 31, 2025, https://insidehpc.com/2025/10/supermicro-expands-collaboration-with-nvidia/
31. ASUS Rolls Out NVIDIA GB300 NVL72 Rack Solution to Usher in the Next Era of Datacenters, dernier accès : octobre 31, 2025, https://press.asus.com/news/press-releases/asus-nvidia-gb300-nvl72-rack-solution/
32. How HPC Is Powering AI, Big Data, and the Future of Innovation | Core Scientific, dernier accès : octobre 31, 2025, https://corescientific.com/resources/blog/high-performance-computing-powering-ai-big-data-and-future-of-innovation/
33. Train With Mixed Precision - NVIDIA Docs, dernier accès : octobre 31, 2025, https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html
34. Multi-GPU Benchmark: B200 vs H200 vs H100 vs MI300X - Research AIMultiple, dernier accès : octobre 31, 2025, https://research.aimultiple.com/multi-gpu/
35. High-Performance Computing in Big Data Analytics: Architectures, Scalability, and Optimization Strategies, dernier accès : octobre 31, 2025, https://ijaidsml.org/index.php/ijaidsml/article/download/26/24
36. What is High Performance Data Analytics?, dernier accès : octobre 31, 2025, https://www.polestarllp.com/glossary/high-performance-data-analytics
37. HPDA (What Is It & Why Is It Important?) - WEKA, dernier accès : octobre 31, 2025, https://www.weka.io/learn/glossary/ai-ml/hpda/
38. High Performance Computing (HPC) Server Solutions - Supermicro, dernier accès : octobre 31, 2025, https://www.supermicro.com/en/solutions/high-performance-computing
39. Strategic Insights: High Performance Computing As A Service, dernier accès : octobre 31, 2025, https://www.openpr.com/news/4245602/strategic-insights-high-performance-computing-as-a-service
40. What is Hybrid Quantum Computing - QMware, dernier accès : octobre 31, 2025, https://www.qmware.com/blog/hybrid-quantum-computing/what-is-hybrid-quantum-computing/
41. What are hybrid quantum-classical algorithms? - Milvus, dernier accès : octobre 31, 2025, https://milvus.io/ai-quick-reference/what-are-hybrid-quantumclassical-algorithms
42. Hybrid Quantum-Classical Algorithms: The Future of Computing ..., dernier accès : octobre 31, 2025, https://www.spinquanta.com/news-detail/hybrid-quantum-classical-algorithms-the-future-of-computing20250123075527
43. From GPUs to FPGAs – An Introduction to High-Performance Computing - Fotis I. Giasemis, dernier accès : octobre 31, 2025, https://fotisgiasemis.com/blog/hpc-gpu-fpga-intro/
44. AMD EPYC™ 9754 AI/ML Performance, dernier accès : octobre 31, 2025, https://www.amd.com/content/dam/amd/en/documents/epyc-business-docs/performance-briefs/amd-epyc-9754-pb-aiml.pdf
45. AMD EPYC 128 CORE Processor 9754 2.25GHZ Base / 3.1GHZ Max 256MB L3 Cache TDP 360W SP5 Socket (Bergamo) (4TH Gen) - ITCreations.com, dernier accès : octobre 31, 2025, https://www.itcreations.com/product/140945
46. AMD EPYC 9754 2.25GHz 128-core 360W Processor for HPE Data sheet, dernier accès : octobre 31, 2025, https://www.hpe.com/psnow/doc/PSN1014752656CHEN
47. Intel Xeon-Platinum 8490H 1.9GHz 60-core 350W Processor for HPE Data sheet, dernier accès : octobre 31, 2025, https://www.hpe.com/psnow/doc/PSN1014739132VNEN
48. Intel® Xeon® Platinum 8490H 1.9GHz 60 Core Processor, 60C ..., dernier accès : octobre 31, 2025, https://www.dell.com/en-sg/shop/intel-xeon-platinum-8490h-19ghz-60-core-processor-60c-120t-16gt-s-113m-cache-turbo-ht-350w-ddr5-4800-customer-install/apd/338-clrh/processors
49. Intel PK8071305074901 Intel Xeon Platinum 8490H Processor Tray - PROVANTAGE, dernier accès : octobre 31, 2025, https://www.provantage.com/intel-pk8071305074901~7ITEP8KY.htm

50. AMD EPYC Processors for Computer-Aided Engineering (CAE ..., dernier accès : octobre 31, 2025, https://www.amd.com/content/dam/amd/en/documents/partner-hub/epyc/amd-epyc-9004-series-vs-intel-platinum-8490h-media-entertainment-workloads-solution-brief-competitive.pdf
51. AMD EPYC 9754 vs Intel Xeon Platinum 8470 [cpubenchmark.net] by PassMark Software, dernier accès : octobre 31, 2025, https://www.cpubenchmark.net/compare/5752vs5693/AMD-EPYC-9754-vs-Intel-Xeon-Platinum-8470
52. Accelerating The Cloud with Heterogeneous Computing - USENIX, dernier accès : octobre 31, 2025, https://www.usenix.org/event/hotcloud11/tech/final_files/Suneja.pdf
53. Cornell Virtual Workshop > Understanding GPU Architecture > GPU ..., dernier accès : octobre 31, 2025, https://cvw.cac.cornell.edu/gpu-architecture/gpu-characteristics/applications
54. Independent Performance Analysis of Leading GPUs: AMD MI300X, NVIDIA H100, NVIDIA H200, dernier accès : octobre 31, 2025, https://artificialanalysis.ai/articles/independent-analysis-of-leading-gpus-amd-nvidia
55. MI300X vs H200: Future of Exascale Computing, dernier accès : octobre 31, 2025, https://blog.neevcloud.com/mi300x-vs-h200-the-future-of-supercomputing
56. What are the Elements of High Performance Computing (HPC) Nowadays?, dernier accès : octobre 31, 2025, https://altimetrikpoland.medium.com/what-are-the-elements-of-high-performance-computing-hpc-nowadays-137e64c241af
57. Role of FPGAs in High-Performance Computing [2024], dernier accès : octobre 31, 2025, https://www.logic-fruit.com/blog/fpga/role-of-fpgas-in-high-performance-computing/
58. Using FPGAs for High-Performance Computing: Challenges and Opportunities, dernier accès : octobre 31, 2025, https://runtimerec.com/using-fpgas-for-high-performance-computing-challenges-and-opportunities/
59. Forthcoming IBM Paper Expected to Show Quantum Algorithm Running on Inexpensive AMD Chips, dernier accès : octobre 31, 2025, https://thequantuminsider.com/2025/10/24/forthcoming-ibm-paper-expected-to-show-quantum-algorithm-running-on-inexpensive-amd-chips/
60. IBM Runs Quantum Computing Algorithm on AMD FPGA Chips ..., dernier accès : octobre 31, 2025, https://www.techpowerup.com/342217/ibm-runs-quantum-computing-algorithm-on-amd-fpga-chips
61. VERSAL™ PREMIUM VP1902 ADAPTIVE SOC - AMD, dernier accès : octobre 31, 2025, https://www.xilinx.com/content/dam/xilinx/publications/product-briefs/2118851-versal-premium-vp1902-product-brief.pdf
62. AMD Versal™ Premium VP1902 Adaptive SoC, dernier accès : octobre 31, 2025, https://www.amd.com/en/products/adaptive-socs-and-fpgas/versal/premium-series/vp1902.html
63. Intel Stratix 10 GX/SX. Product Table. Datasheet. Specifications, dernier accès : octobre 31, 2025, https://www.skyblue.de/uploads/Datasheets/intel_pb_stratix-10-product-table.pdf
64. A Survey on Heterogeneous Computing Using SmartNICs and Emerging Data Processing Units - arXiv, dernier accès : octobre 31, 2025, https://arxiv.org/html/2504.03653v2
65. What is DPU(Data Processing Units)? - GIGABYTE Global, dernier accès : octobre 31, 2025, https://www.gigabyte.com/Glossary/dpu
66. What Is a DPU? | NVIDIA Blog, dernier accès : octobre 31, 2025, https://blogs.nvidia.com/blog/whats-a-dpu-data-processing-unit/
67. What is a DPU (Data Processing Unit)? - Premio Inc, dernier accès : octobre 31, 2025, https://premioinc.com/blogs/blog/what-is-a-dpu-data-processing-unit
68. ThinkSystem NVIDIA BlueField-3 QSFP112 Adapters - Lenovo Press, dernier accès : octobre 31, 2025, https://lenovopress.lenovo.com/lp1809.pdf
69. /nvidia-bluefield-dpu-3 - PNY Technologies, dernier accès : octobre 31, 2025, https://www.pny.com/en-eu/nvidia-bluefield-dpu-3

70. AMD Pensando Salina 400 DPU Spotted - ServeTheHome, dernier accès : octobre 31, 2025, https://www.servethehome.com/amd-pensando-salina-400-dpu-arm-neoverse/
71. AMD Pensando Salina 400 DPU: New Features and Insights Unveiled - ColoCrossing, dernier accès : octobre 31, 2025, https://www.colocrossing.com/blog/amd-pensando-salina-400-dpu-new-features/
72. What is an NPU? And why is it key to unlocking on-device ..., dernier accès : octobre 31, 2025, https://www.qualcomm.com/news/onq/2024/02/what-is-an-npu-and-why-is-it-key-to-unlocking-on-device-generative-ai
73. NPU (Neural Processing Units) | Samsung Semiconductor Global, dernier accès : octobre 31, 2025, https://semiconductor.samsung.com/support/tools-resources/dictionary/the-neural-processing-unit-npu-a-brainy-next-generation-semiconductor/
74. TOPS of the Heap: Qualcomm Unveils Snapdragon X2 Elite Extreme ..., dernier accès : octobre 31, 2025, https://www.pcmag.com/news/snapdragon-summit-qualcomm-unveils-snapdragon-x2-elite-extreme-cpu
75. NVIDIA Grace Hopper Superchip Architecture Whitepaper, dernier accès : octobre 31, 2025, https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper
76. Apple M5 vs Snapdragon X2 Elite Extreme benchmarks: The early verdict is in, and it's a surprise | Tom's Guide, dernier accès : octobre 31, 2025, https://www.tomsguide.com/computing/cpus/apple-m5-vs-snapdragon-x2-elite-extreme-benchmarks-the-early-verdict-is-in-and-its-a-surprise
77. Snapdragon X Elite | Best Laptop Performance - Qualcomm, dernier accès : octobre 31, 2025, https://www.qualcomm.com/laptops/products/snapdragon-x-elite
78. Arrow Lake (microprocessor) - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/Arrow_Lake_(microprocessor)
79. Intel Core Ultra 5 245K Review - AI Performance with NPU ..., dernier accès : octobre 31, 2025, https://www.techpowerup.com/review/intel-core-ultra-5-245k/26.html
80. TPU vs GPU: A Comprehensive Technical Comparison - Wevolver, dernier accès : octobre 31, 2025, https://www.wevolver.com/article/tpu-vs-gpu-a-comprehensive-technical-comparison
81. TPU vs GPU: Choosing the Right Hardware for Your AI Projects | DigitalOcean, dernier accès : octobre 31, 2025, https://www.digitalocean.com/resources/articles/tpu-vs-gpu
82. AI Hypercomputer TPUs - TPU options | Google Cloud Skills Boost, dernier accès : octobre 31, 2025, https://www.cloudskillsboost.google/paths/2806/course_templates/1405/documents/568755
83. TPU v5p | Google Cloud Documentation, dernier accès : octobre 31, 2025, https://docs.cloud.google.com/tpu/docs/v5p
84. TPU architecture - Google Cloud Documentation, dernier accès : octobre 31, 2025, https://docs.cloud.google.com/tpu/docs/system-architecture-tpu-vm
85. TPU vs GPU: What's the real difference? - Telnyx, dernier accès : octobre 31, 2025, https://telnyx.com/learn-ai/tpu-vs-gpu
86. Anthropic expands Google Cloud TPU use to boost AI research ..., dernier accès : octobre 31, 2025, https://www.eenewseurope.com/en/anthropic-expands-google-cloud-tpu-use-to-boost-ai-research/
87. Thermodynamic Computing: From Zero to One | Extropic, dernier accès : octobre 31, 2025, https://extropic.ai/writing/thermodynamic-computing-from-zero-to-one
88. TSU 101: An Entirely New Type of Computing Hardware | Extropic, dernier accès : octobre 31, 2025, https://extropic.ai/writing/tsu-101-an-entirely-new-type-of-computing-hardware
89. Extropic Claims 10,000x Energy Savings With New Probabilistic AI ..., dernier accès : octobre 31, 2025, https://www.vktr.com/ai-news/extropic-claims-10000x-energy-savings-with-new-probabilistic-ai-chip/
90. Inside X0 and XTR-0 | Extropic, dernier accès : octobre 31, 2025, https://extropic.ai/writing/inside-x0-and-xtr-0
91. What is Hybrid Quantum Computing? - IonQ, dernier accès : octobre 31, 2025,

https://ionq.com/resources/what-is-hybrid-quantum-computing
92. IBM Heron - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/IBM_Heron
93. Processor types | IBM Quantum Documentation, dernier accès : octobre 31, 2025, https://quantum.cloud.ibm.com/docs/guides/processor-types
94. List of quantum processors - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/List_of_quantum_processors
95. IonQ Forte, dernier accès : octobre 31, 2025, https://ionq.com/quantum-systems/forte
96. Unlocking Ultra-Fast GPU Communication with NVIDIA NVLink ..., dernier accès : octobre 31, 2025, https://uvation.com/articles/unlocking-ultra-fast-gpu-communication-with-nvidia-nvlink-nvlink-switch
97. NVLink - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/NVLink
98. Understanding NVIDIA NVLink | Scaleway Documentation, dernier accès : octobre 31, 2025, https://www.scaleway.com/en/docs/gpu/reference-content/understanding-nvidia-nvlink/
99. NVLink & NVSwitch: Fastest HPC Data Center Platform | NVIDIA, dernier accès : octobre 31, 2025, https://www.nvidia.com/en-us/data-center/nvlink/
100. Architectural and System Implications of CXL-enabled Tiered Memory - arXiv, dernier accès : octobre 31, 2025, https://arxiv.org/html/2503.17864v1
101. From GPUs to Memory Pools: Why AI Needs Compute Express Link ..., dernier accès : octobre 31, 2025, https://www.eetimes.com/from-gpus-to-memory-pools-why-ai-needs-compute-express-link-cxl/
102. CXL: ENABLING A HETEROGENOUS, COMPOSABLE, NEXT-GENERATION DATA CENTER - Moor Insights & Strategy, dernier accès : octobre 31, 2025, https://moorinsightsstrategy.com/wp-content/uploads/2022/08/CXL-Enabling-A-Heterogeneous-Composable-Nxt-Gen-DC-By-Moor-Insights-And-Strategy.pdf
103. Compute Express Link - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/Compute_Express_Link
104. GPU Graph Processing on CXL™-Based Microsecond-Latency ..., dernier accès : octobre 31, 2025, https://www.kioxia.com/en-jp/rd/technology/topics/topics-63.html
105. Compute Express Link (CXL) 3.0: All You Need To Know | Servermall Blog, dernier accès : octobre 31, 2025, https://servermall.com/blog/compute-express-link-cxl-3-0-all-you-need-to-know/
106. Purging CXL cache coherency dilemmas - Siemens Digital Industries Software, dernier accès : octobre 31, 2025, https://resources.sw.siemens.com/en-US/white-paper-purging-cxl-cache-coherency-dilemmas/
107. Heterogeneous vs. Homogeneous Computing Environments - Intel, dernier accès : octobre 31, 2025, https://www.intel.com/content/www/us/en/docs/sycl/introduction/latest/01-homogeneous-vs-heterogeneous.html
108. HPC Orchestration for the Post-exascale Era (HOPE) - Italy for Artificial Intelligence, dernier accès : octobre 31, 2025, https://it4lia-aifactory.eu/services/hpc-orchestration-for-the-post-exascale-era-hope/
109. Lenovo Intelligent Computing Orchestration (LiCO) Product Guide, dernier accès : octobre 31, 2025, https://lenovopress.lenovo.com/lp0858-lenovo-intelligent-computing-orchestration-lico
110. Message Passing Interface - Wikipedia, dernier accès : octobre 31, 2025, https://en.wikipedia.org/wiki/Message_Passing_Interface
111. NPU vs GPU: What's the Difference? - IBM, dernier accès : octobre 31, 2025, https://www.ibm.com/think/topics/npu-vs-gpu
112. NPU 101: A Guide to Next-Generation AI Performance – Rugged ..., dernier accès : octobre 31, 2025, https://dtresearch.com/blog/2025/08/28/npu-101-a-guide-to-next-generation-ai-performance/
113. Programmable packet-optical network security and monitoring using ..., dernier accès : octobre 31, 2025, https://opg.optica.org/jocn/abstract.cfm?uri=jocn-17-2-A178
114. Deploy a model to Cloud TPU VMs | Vertex AI - Google Cloud Documentation, dernier accès : octobre

31, 2025, https://docs.cloud.google.com/vertex-ai/docs/predictions/use-tpu
115. Improving inference time in multi-TPU systems with profiled model segmentation - arXiv, dernier accès : octobre 31, 2025, https://arxiv.org/html/2503.01025v1
116. Balanced segmentation of CNNs for multi-TPU inference - arXiv, dernier accès : octobre 31, 2025, https://arxiv.org/html/2503.01035v1
117. Scaling Hybrid Quantum–HPC Applications with the Quantum ..., dernier accès : octobre 31, 2025, https://arcb.csc.ncsu.edu/~mueller/ftp/pub/mueller/papers/sfwm25.pdf
118. Unlocking AI innovation: GPU-as-a-Service with Red Hat, dernier accès : octobre 31, 2025, https://www.redhat.com/en/blog/unlocking-ai-innovation-gpu-service-red-hat
119. How To Sample From a Generative AI Model During Training ..., dernier accès : octobre 31, 2025, https://wandb.ai/capecape/stacking_tables/reports/How-To-Sample-From-a-Generative-AI-Model-During-Training--Vmlldzo1MzQyMDYw
120. Qualcomm unveils new AI engine and enters the rack-scale business - Computing UK, dernier accès : octobre 31, 2025, https://www.computing.co.uk/news/2025/chips-and-components/qualcomm-unveils-new-ai-engine-and-enters-the-rack-scale-business
121. How heterogeneous computing optimizes deep learning workloads | ADLINK Blog, dernier accès : octobre 31, 2025, https://www.adlinktech.com/en/heterogeneous-computing
122. Qu'est-ce que le calcul hautes performances (HPC)? | Google Cloud, dernier accès : octobre 31, 2025, https://cloud.google.com/discover/what-is-high-performance-computing?hl=fr
123. DOE Explains...Exascale Computing - Department of Energy, dernier accès : octobre 31, 2025, https://www.energy.gov/science/doe-explainsexascale-computing
124. ASCR Cybersecurity for Scientific Computing Integrity - ENERGY, dernier accès : octobre 31, 2025, https://science.osti.gov/-/media/ascr/pdf/programdocuments/docs/ASCR_Cybersecurity_For_Scientific_Computing_Integrity_Report_2015.pdf