Charpenay Côme Sio 1
14/10/25

Avancée Technologique GPU

Veille Technologique 1

Architecture GPU : évolutions matérielles récentes

Nouvelles générations (2024–2025)

Les architectures NVIDIA Blackwell (B200 / GB200) et AMD Instinct MI300X représentent une rupture par rapport aux générations précédentes :

Fabricant	Architecture	Gravure	Mémoire	Innovations majeures
NVIDIA Blackwell (GB200)	**Dual GPU sur un seul module (NVLink Switch intégré)**	**4 nm TSMC**	**Jusqu'à 192 Go HBM3e**	**Fusion de deux GPU via un switch NVLink 5e gen, bande passante 1,8 To/s, calcul FP8/FP4 optimisé IA**

AMD Instinct MI300X	**APU hybride (CPU + GPU empilés en 3D)**	**5 nm + 6 nm**	**192 Go HBM3**	**Architecture chipset empilée 3D (CPU Zen 4 + GPU PCDNA3), très grande bande passante mémoire**
Intel Ponte Vecchio (Xe-HPC)	**Multi-tile modulaire**	**7 nm + 10 nm**	**Jusqu'à 128 Go HBM2e**	**Conception multi-dies hétérogène (plusieurs types de nœuds), interconnexion EMIB/Foveros**

Innovations thermiques et refroidissement

- **Refroidissement liquide direct (Direct Liquid Cooling, DLC) :**
  Les centres HPC et supercalculateurs (comme *Jupiter* ou *Frontier*) utilisent des plaques froides sur GPU pour évacuer 700–1000 W par carte, impossible à gérer à l'air.
  ➜ Permet des densités extrêmes (jusqu'à 100 kW/rack).

- **Vapor Chamber & microcanaux :**
  Nouvelles chambres à vapeur intégrées directement sur le die GPU ou le substrat interposer pour éviter les "hotspots".
  ➜ Exemple : NVIDIA B200 a une double vapor chamber + cold plate sur HBM.

- **Capteurs thermiques distribués :**
  Chaque HBM, VRM, et cluster SM (Streaming Multiprocessor) a maintenant des capteurs thermiques intégrés.
  ➜ Le firmware ajuste en temps réel la tension et la fréquence par zone (DVFS localisé).

Interconnexions ultra-rapides

NVLink 5.0 (Blackwell) → 1,8 To/s entre GPU (vs 900 Go/s sur Hopper).

Infinity Fabric 3 (AMD) → ~1,5 To/s inter-GPU sur MI300X.

CXL 3.0 (Compute Express Link) → standard émergent pour mémoire partagée entre GPU, CPU et accélérateurs tiers (Micron, Samsung, etc.).

PCIe Gen6 / Gen7 → jusqu'à 128 Go/s bidirectionnel prévu d'ici 2026.

Innovations logicielles et algorithmes

Nouvelles précisions de calcul

FP8 et FP4 :

Nouvelles représentations flottantes introduites pour IA et inférence.

➜ FP8 offre 2× plus de throughput que FP16 à précision similaire pour modèles IA.

➜ FP4 permet 4× plus de performances pour inférence légère (grandes LLM).

Ces formats sont maintenant gérés nativement par Tensor Cores Blackwell et MI300X.

Schedulers matériels intelligents

- **Les GPU modernes utilisent des ordonnanceurs autonomes qui gèrent dynamiquement la répartition des threads et la priorité entre calculs IA / graphisme / physique.**

- **Sur NVIDIA, le Dynamic Work Schedule Dws détecte la charge et redirige les SM en microsecondes.**

- **Objectif : maximiser l'occupation GPU sans intervention CPU.**

Compilation et API unifiées

- **CUDA 13 introduit la compilation unifiée CPU–GPU pour architectures Grace Blackwell.**

- **HIP (AMD) et SYCL (Intel / Khronos) convergent vers une portabilité croissante (code unique pour plusieurs marques).**

- **Vulkan 1.4 / DirectX 12 Ultimate ajoutent la gestion fine de la mémoire vidéo et du Ray Tracing matériel.**

GPU + IA embarquée pour gestion interne

Les pilotes et firmwares (notamment chez NVIDIA) intègrent désormais de l'IA embarquée :

- **Predictive fan control : régulation du refroidissement en fonction de modèles thermiques prédictifs.**
- **Dynamic Voltage Management : IA qui ajuste fréquence et tension selon le type de charge.**
- **Thermal learning loops : apprentissage adaptatif du comportement thermique de chaque puce.**

Capteurs et surveillance intégrée

Les GPU récents comportent plusieurs dizaines de capteurs matériels embarqués, dont :

Type de capteur	Fonction
Thermique (par cluster SM, HBM, VRM)	**Gère throttling local et contrôle du flux liquide/air**
Électrique (tension/courant)	**Mesure en direct la consommation réelle de chaque domaine**
Stress / vieillissement	**Suivi de l'usure des interconnexions et du substrat**
Capteurs optiques internes (sur wafers test)	**Mesurent les micro-variations de fréquence et température en fabrication**
EMI / bruit	**Détectent les interférences électromagnétiques pour stabilité PCIe/NVLink**

Projets futurs (déjà en développement)

NVIDIA Rubin (2026)

Successeur de Blackwell, prévu en 3 nm, avec :

- **NVLink 6.0 (>2 To/s),**
- **mémoire HBM4,**
- **compute FP2 (pour inférence ultra-légère),**
- **refroidissement liquide intégré d'usine (pas d'air cooling par défaut).**

AMD Instinct MI400 (2026–2027)

Basé sur architecture CDNA4 :

- **empilement 3D complet (HBM sur GPU + GPU sur CPU),**
- **256 Go HBM4 possible,**
- **intégration directe dans le socket CPU (pas de PCIe).**

Intel Falcon Shores

Projet hybride CPU+GPU unifié avec mémoire CXL universelle et architecture mixte Xe3 / E-core.
Objectif : remplacer les clusters hétérogènes par un seul type de module “compute universel”.

Mémoire optique / photonique intégrée

Des laboratoires (IBM, MIT, TSMC) testent des GPU avec bus photonique (transmission lumière au lieu d’électrons).
→ Avantage : bande passante >10 To/s, chaleur réduite, aucune interférence EM.
→ Premiers prototypes d’ici 2027.