Anatomie des unités de calcul

CPU, GPU, TPU, NPU, DPU, QPU, LPU

Stéphane FOSSE
fosse.fr

03 mai 2026

Résumé

Le sigle *xPU* est devenu un tic de langage du secteur. Derrière cette inflation de lettres se trouve pourtant une logique architecturale cohérente : depuis que la loi de Moore ne suffit plus à garantir des gains de performance, la spécialisation du silicium est devenue le principal levier. Cet article propose un tour d'horizon des grandes familles d'unités de calcul — CPU, GPU, TPU, NPU, DPU, QPU, APU, LPU — de leur histoire, de leurs principes architecturaux et de leurs cas d'usage.

Quelle est la différence entre RISC et CISC ?

Tout commence au début des années 1970, chez IBM. Pour comprendre le RISC, il faut d'abord comprendre ce contre quoi il réagit. En 1964, IBM lance le System/360, une gamme unifiée de mainframes partageant un même jeu d'instructions (ISA). L'idée, portée notamment par Fred Brooks, est brillante commercialement : un seul logiciel tourne sur toutes les machines de la gamme, du petit modèle à 8 bits au grand à 64 bits. Pour y parvenir, le System/360 s'appuie sur la microprogrammation : chaque instruction complexe est décomposée en micro-instructions stockées dans une mémoire de contrôle. Les machines bon marché exécutent les mêmes instructions en plus de cycles ; les machines haut de gamme les exécutent directement en câblé. IBM gagne son pari : les descendants du System/360, devenus Z, génèrent encore 10 milliards de dollars de revenus annuels.

Mais cette logique porte en elle ses propres limites. À mesure que les circuits intégrés progressent, les jeux d'instructions CISC (Complex Instruction Set Computer) grossissent. Le VAX-11/780 de Digital Equipment Corporation, lancé en 1977, dispose d'un store de contrôle de 5 120 mots par 96 bits, contre 256 mots par 56 bits pour son prédécesseur. Des traces de programmes réels révèlent alors quelque chose d'embarrassant : 30 % des instructions consistent à déplacer des données entre registres et mémoire, et les branchements absorbent jusqu'à un tiers du temps d'exécution. La complexité du jeu d'instructions ne sert pas l'essentiel du travail.

C'est à partir de ce constat que John Cocke, ingénieur au laboratoire Thomas J. Watson d'IBM, développe à partir de 1974 ce qui deviendra le projet IBM 801. À l'origine, la machine est conçue pour gérer un réseau de commutation téléphonique traitant 300 appels par seconde, soit environ 20 000 instructions par appel : une application temps réel, intensive mais répétitive. Cocke observe que des instructions simples, de longueur uniforme, exécutables en un seul cycle d'horloge avec pipeline, offrent un débit supérieur à celui de machines CISC pourtant plus sophistiquées sur le papier. En 1980, le prototype de l'IBM 801 tourne à 15 MHz pour une puissance de 15 MIPS, alors remarquable [5].

Ni Cocke ni son équipe n'inventent le terme « Reduced Instruction Set Computer ». C'est David Patterson, professeur à l'Université de Californie Berkeley, qui le forge — en travaillant en parallèle sur le projet RISC de Berkeley financé par la DARPA, entre 1980 et 1983. John Hennessy mène un projet concurrent à Stanford, qui donnera naissance au processeur MIPS. Ces trois lignées — IBM 801, Berkeley RISC, Stanford MIPS — posent les fondations de toute l'architecture processeur moderne [2].

L'argument RISC repose sur trois principes : des instructions simples et de longueur fixe que le compilateur peut facilement ordonnancer, une architecture load/store qui sépare strictement les accès mémoire du traitement arithmétique, et un grand nombre de registres généraux pour minimiser les allers-retours en mémoire. Ces principes favorisent le pipeline et l'exécution superscalaire. Ils ont guidé ARM, SPARC, PowerPC et, depuis 2010, RISC-V.

La réalité des processeurs actuels est plus nuancée. Les puces x86 d'Intel et AMD, dont l'ISA CISC remonte au 8086 de 1978, décomposent en interne leurs instructions complexes en micro-opérations de type RISC. Inversement, les cœurs ARM intègrent depuis longtemps des extensions SIMD (NEON), de l'exécution spéculative et des pipelines out-of-order qui les éloignent du minimalisme d'origine. L'Apple M1, sorti en 2020, réfute définitivement l'idée que CISC serait l'avenir des ordinateurs de bureau : basé sur ARM, il égale ou dépasse les Intel Xeon contemporains en réduisant la consommation d'environ 90 %. Comme le formulent

John Hennessy et David Patterson dans leur conférence Turing de 2018 : « le marché tranche finalement les débats d'architecture » [9].

Pourquoi la fin de la loi de Moore a-t-elle relancé l'innovation architecturale ?

Pendant quarante ans, Gordon Moore a eu raison. Sa loi empirique de 1965 — la densité des transistors double environ tous les deux ans — a permis à chaque génération de logiciels de bénéficier automatiquement de performances supérieures sans modifier son code. Le modèle économique de l'informatique reposait entièrement là-dessus. Or, depuis le milieu des années 2010, la miniaturisation ralentit. La loi de Dennard, qui prédisait que la densité de puissance des transistors resterait constante à mesure qu'ils rétrécissaient, a cessé de s'appliquer vers 2005. Les processeurs multicœurs ont masqué le problème pendant une décennie, mais la parallélisation a ses propres limites.

Dans leur article fondateur publié dans les *Communications of the ACM* en 2019, Hennessy et Patterson diagnostiquent la rupture et annoncent paradoxalement un nouvel âge d'or pour l'architecture : puisque le silicium généraliste ne progresse plus suffisamment vite, la spécialisation devient le seul vecteur de gains significatifs. Un processeur généraliste consomme entre 10 et 4 000 fois plus d'énergie à simplement récupérer et décoder une instruction qu'à effectuer l'opération arithmétique correspondante : cette inefficacité était acceptable quand Moore compensait ; elle ne l'est plus. William Dally, directeur de la recherche chez NVIDIA, et ses collègues de Stanford le formulent en 2020 dans les *Communications of the ACM* : les accélérateurs dédiés à un domaine (Domain-Specific Accelerators, DSA) sont désormais l'un des rares leviers restants pour continuer à progresser [7].

C'est dans ce contexte que prolifèrent les xPU. Chacun répond à un goulot d'étranglement précis : parallélisme massif pour le graphisme et l'IA (GPU), opérations matricielles entières pour l'inférence de réseaux de neurones (TPU), traitement efficace à faible consommation pour l'edge (NPU), déchargement des tâches réseau et sécurité dans les datacenters (DPU), génération de tokens LLM à latence déterministe (LPU), calcul quantique pour les problèmes d'une complexité exponentiellement difficile pour l'ordinateur classique (QPU).

Qu'est-ce qu'un GPU et comment est-il devenu central pour l'IA ?

Le GPU (Graphics Processing Unit) naît dans les années 1990 de la convergence de plusieurs lignées de recherche en infographie parallèle : les workstations Silicon Graphics et leur API OpenGL, les systèmes Pixel Planes de l'Université de Caroline du Nord, et les travaux de la NASA sur le traitement massivement parallèle d'images. NVIDIA introduit le terme « GPU » en 1999 avec le GeForce 256, premier processeur graphique capable de gérer l'intégralité du pipeline OpenGL en silicium.

Le GPU est fondamentalement une machine SIMD (Single Instruction, Multiple Data) : un même flux d'instructions s'applique simultanément à des milliers de données différentes. Cette architecture, parfaitement adaptée au rendu 3D qui consiste à appliquer les mêmes transformations géométriques à des millions de pixels en parallèle, se révèle idéale pour les opérations matricielles au cœur de l'apprentissage profond. Le passage au GPGPU (General-Purpose computing on GPU) s'opère en 2006 lorsque NVIDIA introduit CUDA (Compute Unified Device Architecture), une extension du langage C permettant de programmer les GPU pour des calculs non graphiques. Brook et Sh, deux langages académiques de 2002-2004, avaient préparé le terrain.

Le résultat de trente ans de recherche financée en grande partie par des fonds publics américains (DARPA, NSF) : NVIDIA est aujourd'hui l'entreprise la plus valorisée au monde, portée par la demande en GPU pour l'entraînement de modèles de langage. Un H100, sorti en 2022, contient des dizaines de milliards de transistors et intègre des Tensor Cores spécialisés dans les multiplications matricielles à précision réduite (FP16, BF16, INT8). Cette spécialisation progressive du GPU le rapproche paradoxalement des accélérateurs dédiés qu'il contribue à concurrencer [13].

Qu'est-ce qu'un TPU et pourquoi Google l'a-t-il conçu en 15 mois ?

En 2013, les équipes de Google réalisent que si la croissance des calculs d'IA liée aux réseaux de neurones se poursuit, il faudra doubler le nombre de datacenters en opération : une perspective économiquement intenable. La solution retenue est un circuit intégré entièrement dédié à l'inférence de réseaux de neurones, conçu et déployé en 15 mois — un délai extraordinairement court pour un ASIC de cette envergure. Norm Jouppi, architecte principal du projet et co-concepteur du processeur MIPS, décrit lui-même le rythme comme « une conception de chip très rapide, assez remarquable » [11].

Le cœur du TPU v1, déployé dans les datacenters de Google à partir de 2015, est un réseau systolique de $256 \times 256 = 65\,536$ unités multiply-accumulate (MAC) 8 bits, offrant un débit de crête de 92 téraopérations par seconde. Il s'accompagne d'une mémoire on-chip de 28 mébioctets gérée explicitement par le logiciel, sans cache matériel. Ce choix délibéré tranche avec l'architecture GPU : pas de spéculation, pas de cache hiérarchique, pas d'exécution hors-ordre. Le modèle d'exécution est entièrement déterministe, ce qui garantit une latence de réponse au 99^{e} percentile bien maîtrisée, essentielle pour les services en temps réel comme Google Search, Street View, Photos ou Translate.

La comparaison avec les contemporains est saisissante : le TPU v1 s'avère 15 à 30 fois plus rapide qu'un GPU Nvidia K80 ou qu'un CPU Intel Haswell pour les charges d'inférence de Google, avec un ratio performance/watt 30 à 80 fois supérieur. Il fonctionne à 700 MHz sur un procédé 28 nm et consomme 40 watts en charge [10]. Sa conception suit par ailleurs un jeu d'instructions CISC : des instructions complexes qui coordonnent l'ensemble de la matrice systolique réduisent le nombre d'instructions à décoder, ce qui minimise la surface de silicium consacrée au contrôle.

Depuis, Google a publié les versions TPU v2, v3 et v4, cette dernière atteignant 275 téraflops en BF16. Un Edge TPU, lancé en 2019, décline l'architecture pour les cas d'usage embarqués, avec 4 TOPS sous 2 watts. Le concept de réseau systolique que le TPU exploite remonte pourtant à 1979, bien avant la vague actuelle de l'IA générative.

Qu'est-ce qu'un NPU et pourquoi trouve-t-on désormais de l'IA dans les smartphones ?

Le NPU (Neural Processing Unit) est la déclinaison du TPU pour les appareils edge : smartphones, laptops, voitures autonomes, objets connectés. Sa contrainte principale n'est pas le débit brut mais l'efficacité énergétique et la latence de traitement sans connexion réseau. Faire tourner un modèle d'IA directement sur le dispositif plutôt que dans le cloud présente deux avantages : la vie privée (les données ne quittent pas l'appareil) et la disponibilité hors ligne.

Apple intègre son premier Neural Engine dans l'A11 Bionic de l'iPhone X en 2017, avec une capacité de 600 milliards d'opérations par seconde. En 2020, l'A14 Bionic porte ce chiffre à 11 billions d'opérations par seconde avec 16 cœurs dédiés. En 2021, l'A15 Bionic atteint 15,8 billions d'opérations par seconde, soit 26 fois la performance du premier Neural Engine en quatre ans. Sur le segment PC, Intel intègre son premier NPU dans les processeurs Meteor Lake (Core Ultra Série 1) en décembre 2023. AMD suit avec les Ryzen 7040 en 2023 (10 TOPS) et les Ryzen AI 300 en 2024 (50 TOPS, architecture XDNA 2). Qualcomm frappe fort avec le Snapdragon X Elite : 45 TOPS, conçu pour satisfaire les critères Copilot+ PC de Microsoft [13].

Les NPU sont architecturalement conçus pour l'arithmétique à faible précision (INT4, INT8, FP8, FP16) et les opérations matricielles denses. Ils misent sur le parallélisme et la minimisation des transferts de données entre mémoire et unités de calcul. Selon IDC, les smartphones embarquant un NPU de 30 TOPS ou plus capables de faire tourner des modèles génératifs localement représentaient 234 millions d'unités en 2024, une croissance de 364 % en un an.

Qu'est-ce qu'un DPU et quel rôle joue-t-il dans la sécurité des datacenters ?

Le DPU (Data Processing Unit) répond à un problème spécifique de l'infrastructure des grands datacenters. Amazon a constaté qu'environ 30 % des cœurs de ses serveurs étaient consacrés à des tâches d'infrastructure : traitement des paquets réseau, chiffrement TLS, gestion du stockage distribué, pare-feu, virtualisation réseau. Des cœurs CPU coûteux et généralistes occupés à faire du routage, c'est un gaspillage économique autant qu'une surface d'attaque potentielle.

Le DPU est essentiellement une SmartNIC avancée : une carte réseau capable de faire tourner un système d'exploitation complet avec ses propres cœurs ARM, sa propre mémoire DRAM, et son propre moteur de chiffrement matériel. NVIDIA a popularisé l'appellation avec la gamme BlueField, issue de l'acquisition de Mellanox en 2020. Le BlueField-2 offre 200 Gbits/s de débit réseau et embarque seize cœurs Armv8, des accélérateurs RDMA (Remote Direct Memory Access) et des moteurs de compression/décompression matériels. Le BlueField-3, présenté en 2022, monte à 400 Gbits/s [12].

Jensen Huang, PDG de NVIDIA, a présenté le DPU en 2020 comme le troisième pilier des datacenters modernes, aux côtés du CPU et du GPU. L'argument est architectural : en isolant les tâches d'infrastructure sur le DPU, on libère les CPU applicatifs, on renforce la ségrégation entre plans de données et plans de contrôle, et on améliore la sécurité. Intel propose une alternative sous le nom d'IPU (Infrastructure Processing Unit). AMD a acquis Pensando en 2022 pour entrer sur ce marché. La nomenclature reste instable : le terme « DPU » s'applique parfois aussi aux unités de traitement en mémoire (Processing-in-Memory) ou aux accélérateurs d'apprentissage profond, ce qui brouille les comparaisons.

Qu'est-ce qu'un APU et pourquoi AMD a-t-il fusionné CPU et GPU sur une même puce ?

L'APU (Accelerated Processing Unit) est une puce qui intègre un CPU et un GPU sur la même puce éléctronique, partageant la même mémoire physique. AMD a lancé la gamme sous le nom de code « Fusion » en 2011 avec les premières puces Bobcat (E-Series) et Llano (A-Series). L'argument initial est économique : pour les segments entrée de gamme et nomade, un GPU discret représente un coût et un encombrement injustifiés.

Mais la vraie rupture architecturale intervient en 2013 avec la puce Kaveri et l'Heterogeneous System Architecture (HSA). L'HSA supprime la frontière mémoire entre CPU et GPU : les deux partagent un espace d'adressage virtuel unifié (hUMA, heterogeneous Unified Memory Architecture), ce qui élimine les copies mémoire coûteuses entre les deux zones. Une application peut envoyer directement un pointeur au GPU sans copier les données. L'impact sur la latence et la consommation est mesurable pour les charges mixtes calcul/rendu.

L'évolution des APU illustre comment les frontières entre unités de calcul s'effacent. Les SoC Apple (M1, M2, M3) appliquent le même principe à une échelle plus ambitieuse, en unifiant CPU, GPU, Neural Engine et décodeurs vidéo dans une mémoire unifiée de haute bande passante. Les APU Ryzen AI 300 d'AMD (2024) intègrent désormais un NPU en plus du CPU et du GPU, faisant de la puce une plateforme hétérogène à trois domaines de calcul distincts.

Qu'est-ce qu'un LPU et en quoi diffère-t-il d'un GPU pour l'inférence de grands modèles de langage ?

Le LPU (Language Processing Unit) est la proposition architecturale de la société Groq, fondée en 2016 par Jonathan Ross, l'un des concepteurs du TPU original chez Google. Le constat de départ est simple : un GPU est conçu pour le parallélisme indépendant de tâches graphiques ; l'inférence d'un LLM est fondamentalement séquentielle, token par token, avec des opérations d'algèbre linéaire très régulières. Un GPU exécute cette charge de manière sous-optimale parce qu'il mobilise une complexité architecturale — caches hiérarchiques, ordonnanceurs dynamiques, DRAM/HBM externe — qui ne sert pas ce cas d'usage.

Le LPU repose sur quatre principes. Le premier est une architecture « logiciel-d'abord » : le compilateur est conçu avant le silicium, de sorte que le hardware reflète exactement ce que le compilateur peut ordonnancer statiquement. Le deuxième est une architecture de chaîne de montage programmable : les données circulent sur des « tapis roulants » entre les unités SIMD, sans besoin de routeurs ou de contrôleurs, y compris entre chips. Le troisième est le déterminisme strict : chaque étape d'exécution est programmée au cycle d'horloge près. Le quatrième est la mémoire on-chip massive : le LPU intègre sa SRAM directement en silicium, avec une bande passante de 80 téraoctets par seconde, contre environ 8 téraoctets par seconde pour la HBM externe des GPU — un avantage d'un facteur dix [8].

En mars 2026, NVIDIA a acquis Groq et présenté le NVIDIA Groq 3 LPX, un système rack-scale à 256 puces LPU interconnectées, offrant 315 pétaflops d'inférence, 128 Go de SRAM totale et 40 pétaoctets par seconde de bande passante SRAM. Ce système est co-déployé avec les GPU Vera Rubin NVL72 pour former une architecture hétérogène : les GPU traitent la phase de prefill (ingestion du contexte) et l'attention, les LPU gèrent la phase de décodage token par token [1], là où la latence est critique. C'est une reconnaissance de facto que le GPU seul ne suffit plus pour les systèmes d'inférence agentique à grande échelle.

Qu'est-ce qu'un QPU et à quel horizon peut-on en attendre des résultats pratiques ?

Le QPU (Quantum Processing Unit) manipule l'information sous la forme de qubits, des états quantiques qui peuvent exister en superposition de 0 et de 1 simultanément. Un registre de n qubits peut représenter 2^n états en même temps, ce qui ouvre théoriquement la voie à des accélérations exponentielles pour certaines classes de problèmes : factorisation d'entiers (algorithme de Shor), recherche dans des bases non structurées (algorithme de Grover), simulation de molécules en chimie ou en physique des hautes énergies.

Keith Britt et Travis Humble, du Oak Ridge National Laboratory, identifient deux modèles d'intégration : une intégration « lâche » où le QPU est un accélérateur externe accessible via le réseau avec des tolérances de latence de plusieurs millisecondes, et une intégration « serrée » où le QPU est couplé directement à des ressources HPC classiques avec des latences de l'ordre de la dizaine de microsecondes pour les opérations de correction d'erreurs quantiques en temps réel [3].

IBM propose des QPU supraconducteurs refroidis à quelques millikelvin au-dessus du zéro absolu, accessibles via le cloud. Google revendique avoir atteint la « suprématie quantique » en 2019 avec son processeur Sycamore à 53 qubits, affirmation contestée par IBM qui a montré que le même calcul pouvait être effectué en 2,5 jours sur son supercalculateur Summit. Les QPU restent aujourd'hui des machines de

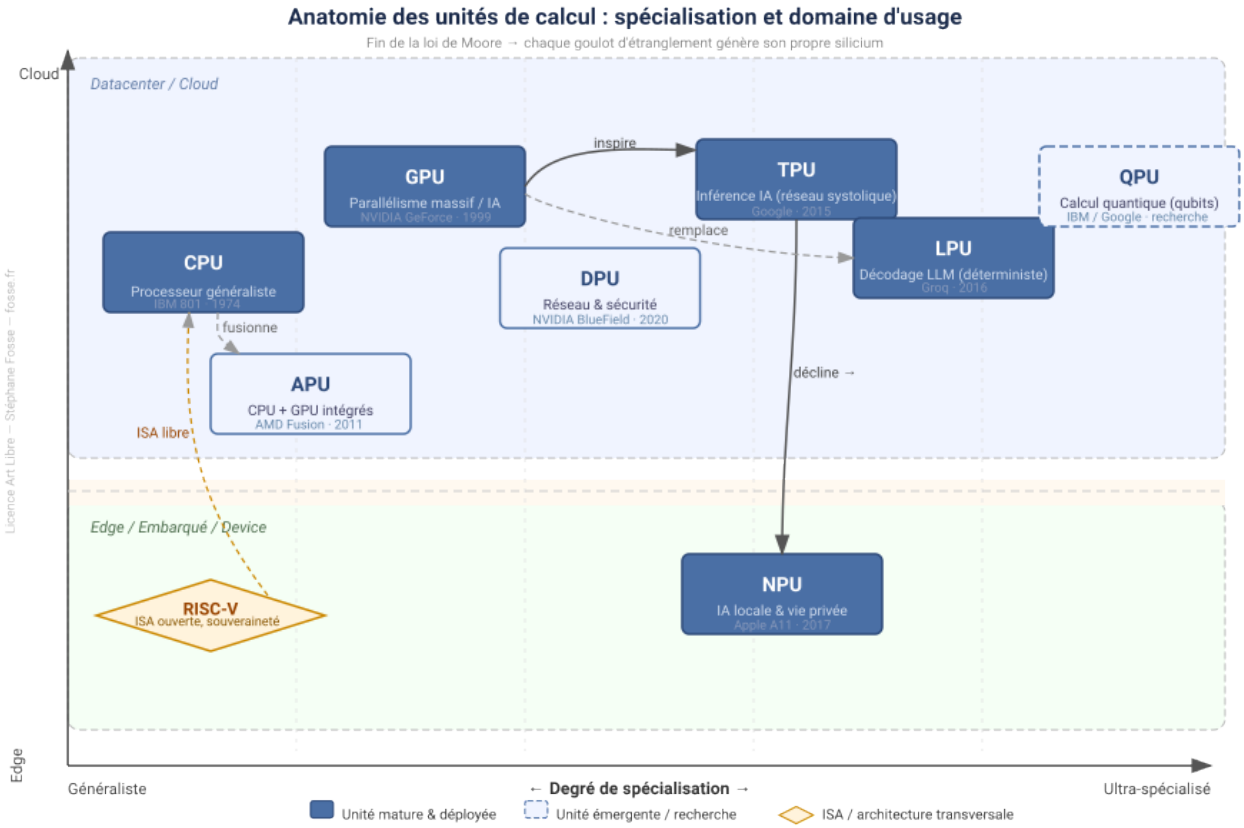

FIGURE 1 – Anatomie des unités de calcul : spécialisation et domaine d'usage

recherche, bruyantes et difficiles à corriger. Le calcul quantique tolérant aux fautes à grande échelle est encore à plusieurs années, voire une décennie, de maturité industrielle.

RISC-V : pourquoi un jeu d'instructions libre est-il un enjeu de souveraineté ?

RISC-V est une ISA ouverte, libre de droits et librement implémentable, née en 2010 dans le groupe de recherche de David Patterson à l'Université de Californie Berkeley. RISC-V se caractérise par une base modulaire : un jeu d'instructions entier minimal auquel s'ajoutent des extensions optionnelles standardisées (M pour la multiplication, A pour l'atomicité, F/D pour le flottant, V pour les vecteurs, H pour la virtualisation).

Ce qui était un projet académique est devenu un enjeu géopolitique. NVIDIA, Western Digital et Alibaba l'utilisent pour leurs designs internes ; l'European Processor Initiative (EPI) s'en empare pour développer des puces HPC souveraines en Europe sans dépendance aux licences américaines ; la Chine a créé la China RISC-V Alliance pour coordonner son industrie nationale de semi-conducteurs. On estimait à 60 milliards le nombre de cœurs RISC-V déployés en 2025 [6]. RISC-V entre dans l'espace avec une sonde NASA. Il arrive dans les laptops (SiFive), les serveurs (Milk-V) et potentiellement les iPhones.

Pour un architecte IT sensible à la question de la souveraineté numérique, RISC-V représente quelque chose d'inédit : la possibilité de construire une filière matérielle complète sans payer de licence à une entreprise américaine ou britannique. C'est le logiciel libre appliqué au silicium. Les risques existent — fragmentation des extensions propriétaires, absence de puces RISC-V hautes performances véritablement compétitives avec x86 ou ARM dans les datacenters en 2026 — mais la trajectoire est claire.

Conclusion : un écosystème hétérogène, pas une guerre de standards

Cinquante ans après le projet IBM 801, la question n'est plus RISC contre CISC : elle a été réglée par le marché, dans un sens qui contredit les catégories initiales. La vraie rupture est celle que Hennessy et Patterson annonçaient en 2019 : la fin du dividende Moore comme mécanisme automatique de progression force chaque domaine applicatif à concevoir son propre silicium.

Le paysage qui en résulte est hétérogène par nature. Dans un datacenter moderne, un serveur d'inférence combine un CPU pour l'orchestration, un GPU ou un LPU pour la génération de tokens, un DPU pour la couche réseau et de sécurité, et demain peut-être un QPU accessible en accélérateur distant. Chaque unité est spécialisée pour sa fonction ; aucune n'est substituable par les autres. IBM Research décrit dès 2026 cette architecture hybride CPU+GPU+QPU comme la prochaine étape du supercalcul.

Ce qui se joue en dessous est une question de souveraineté et de concentration du pouvoir industriel. Le fait que NVIDIA ait absorbé Groq en 2026, qu'AMD ait racheté Pensando et Xilinx, que Google possède ses propres TPU et Amazon ses propres Trainium/Inferentia : ces mouvements signalent que le silicium spécialisé est devenu stratégique. RISC-V est pour l'instant la seule réponse architecturale qui ne passe pas par une licence propriétaire. Ce n'est pas un détail technique : c'est une option de souveraineté.

Références

[1] Kyle Aubrey. Inside NVIDIA Groq 3 LPX: The Low-Latency Inference Accelerator for the NVIDIA Vera Rubin Platform. Anglais. Mars 2026.

[2] Babbage. First RISC: John Cocke and the IBM 801. Anglais. Oct. 2022.

[3] Keith A. Britt et Travis S. Humble. High-Performance Computing with Quantum Processing Units. Anglais. In : *arXiv preprint arXiv:1511.04386* (2015).

[4] Frank Carrubba et al. IBM 801 Microprocessor Oral History Panel. Anglais. Oct. 2014.

[5] John Cocke et V. Markstein. The evolution of RISC technology at IBM. Anglais. In : *IBM Journal of Research and Development* 34.1 (1990), p. 4-11.

[6] Enfang Cui, Tianzheng Li et Qian Wei. RISC-V Instruction Set Architecture Extensions: A Survey. Anglais. In : *IEEE Access* (2023).

[7] William J. Dally, Yatish Turakhia et Song Han. Domain-Specific Hardware Accelerators. Anglais. In : *Communications of the ACM* 63.7 (2020), p. 48-57.

[8] Groq. What is a Language Processing Unit? Anglais. 2024.

[9] John L. Hennessy et David A. Patterson. A New Golden Age for Computer Architecture. Anglais. In : *Communications of the ACM* 62.2 (2019), p. 48-60.

[10] Norman P. Jouppi et al. In-Datacenter Performance Analysis of a Tensor Processing Unit. Anglais. In : *Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17)*. Toronto, ON, Canada : ACM, 2017, p. 1-12.

[11] Kaz Sato, Cliff Young et David Patterson. An in-depth look at Google's first Tensor Processing Unit (TPU). Anglais. Juin 2018.

[12] Nathan Tibbetts, Sifat Ibtisum et Satish Puri. A Survey on Heterogeneous Computing Using SmartNICs and Emerging Data Processing Units. Anglais. In : *arXiv preprint arXiv:2504.03653* (2025).

[13] Bin Xu, Ayan Banerjee et Sandeep Gupta. Hardware Acceleration for Neural Networks: A Comprehensive Survey. Anglais. In : *arXiv preprint arXiv:2512.23914* (2026).