Processeurs Neuromorphiques
Enjeux, limites et architectures hybrides

Vincent CAUQUIL *CPE Lyon TECM 5PSM 2023-2026*
Lyon, France, vincent.cauquil@cpe.fr

***Résumé*—L'explosion de la complexité des modèles d'intelligence artificielle, tels que les Transformers, se heurte aujourd'hui à la fin de la loi de Moore et au « Memory Wall ». Ce rapport examine le paradigme neuromorphique comme alternative durable, en comparant deux approches architecturales majeures : le tout numérique et le signal mixte (In-Memory Computing). En s'appuyant sur un état de l'art rigoureux, nous quantifions le fossé technologique entre la fiabilité des puces numériques (plafonnant autour de 15 TOPS/W « Tera Operations Per Second per Watt ») et le potentiel d'efficacité extrême des technologies émergentes type RRAM/PCM (visant les 100 TOPS/W). L'étude démontre que si les architectures mixtes offrent des densités de calcul supérieures, leur adoption reste freinée par la variabilité stochastique et le coût des conversions de données. Nous concluons que l'avenir de l'IA durable réside non pas dans une opposition, mais dans des architectures hétérogènes 3D combinant la robustesse du numérique et l'efficacité massive de l'analogique.**

***Index Terms*—Calcul neuromorphique, Réseaux de neurones à impulsions (SNN), In-Memory Computing, Efficacité énergétique, Mémoires non-volatiles.**

I. Introduction

L'ESSOR fulgurant de l'intelligence artificielle (IA), catalysé par l'avènement des réseaux de neurones profonds (DNN) et des modèles de langage massifs (LLM) tels que GPT-4, a précipité une crise majeure dans l'architecture des systèmes de calcul. Cette trajectoire de croissance, bien que performante, devient écologiquement et économiquement insoutenable : la demande en puissance de calcul double désormais tous les 3, 4 mois, dépassant largement les gains historiques offerts par la loi de Moore.

Parallèlement, l'industrie des semi-conducteurs se heurte à des limites physiques fondamentales. Le ralentissement de la loi de Moore s'accompagne de la fin de la mise à l'échelle de Dennard [1]. Cette stagnation implique que la simple miniaturisation des transistors ne garantit plus, à densité de puissance constante, les gains d'efficacité énergétique d'autrefois. Ce phénomène exacerbe le problème du *Dark Silicon* [2] aujourd'hui, nous disposons désormais de plus de transistors sur une puce que nous ne pouvons en alimenter simultanément sans provoquer de surchauffe critique.

Cependant, le goulot d'étranglement le plus critique ne se situe plus dans le calcul pur, mais dans le mouvement des données. L'architecture de Von Neumann traditionnelle, caractérisée par la séparation physique entre l'unité de traitement (CPU/GPU) et la mémoire, impose un va-et-vient incessant des données via un bus limité. Cette dichotomie engendre le « Mur de la Mémoire » (*Memory Wall*), illustré par la Fig. 1, où la latence et la bande passante mémoire bride les performances des processeurs. Plus grave encore est l'impact énergétique : comme l'a quantifié Horowitz [2], accéder à une donnée en mémoire DRAM (Dynamic RAM) coûte énergétiquement 100 à 1000 fois plus cher que d'effectuer une opération arithmétique (MAC) sur cette même donnée.

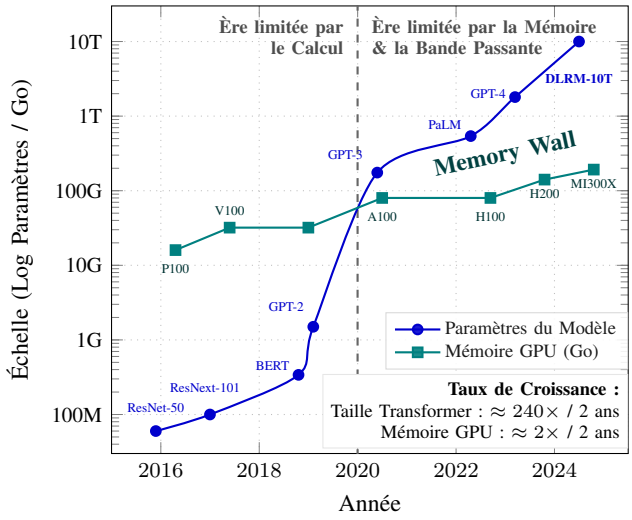

FIGURE 1. Illustration du « Memory Wall » pour l'IA : divergence croissante entre la taille des modèles Transformers (paramètres) et la capacité mémoire des accélérateurs matériels (GPU) [3], [4].

Face à cette impasse architecturale, l'ingénierie neuromorphique propose un changement de paradigme reposant sur une double rupture : algorithmique et matérielle.

Sur le plan **algorithmique**, l'approche s'inspire de l'efficacité biologique. Le cerveau humain, capable de tâches cognitives complexes avec un budget d'environ 20 Watts, surpasse les supercalculateurs actuels grâce au traitement événementiel (*event-driven*) via des Réseaux de Neurones à Impulsions (SNN) [5], [6]. Contrairement aux architectures classiques synchronisées par horloge, les SNN exploitent la parcimonie temporelle : ils ne consomment de l'énergie que lorsqu'un événement (*spike*) survient.

Cependant, la **traduction matérielle** de ces principes divise la communauté scientifique sur la meilleure façon d'implémenter ces réseaux :

1) Les architectures **neuromorphiques numériques** (ex : Intel Loihi 2 [7], IBM TrueNorth [8] ou ReckOn [9]) privilégient la technologie CMOS standard pour simuler la dynamique neuronale. Elles garantissent précision,

déterminisme et facilité de mise à l'échelle, mais restent limitées par la densité des mémoires SRAM.

2) Les architectures à **signal mixte** et le calcul en mémoire (*In-Memory Computing* - IMC). Bien que l'IMC soit également utilisé pour accélérer les DNN classiques, il permet ici d'émuler physiquement les synapses des SNN en exploitant des dispositifs non-volatiles émergents (RRAM, PCM, FeFET) [10]–[12]. Si ces approches promettent des densités et des efficacités records (jusqu'à 18.7 TOPS/W [13]), elles se heurtent aux défis de la variabilité et du bruit analogique.

Ce papier propose une analyse comparative de ces deux voies technologiques. Après avoir exposé les fondements théoriques des SNN (Section II), nous détaillerons l'état de l'art des processeurs numériques (Section III) et des approches mixtes (Section IV), avant de discuter des perspectives d'hybridation.

II. Fondements Théoriques : Les Réseaux de Neurones à Impulsions

Pour saisir les enjeux matériels, il est impératif de comprendre le modèle de calcul que ces processeurs visent à accélérer : le Réseau de Neurones à Impulsions (SNN). Contrairement aux réseaux de neurones artificiels (ANN) standards qui utilisent des fonctions d'activation continues (ex : ReLU « Rectified Linear Unit ») et une propagation synchrone, les SNN opèrent sur des événements binaires discrets, ou « spikes », dans le domaine temporel [14], [15].

A. Comparaison Architecturale

La différence fondamentale entre l'informatique classique et neuromorphique réside dans l'organisation de la mémoire et du calcul, comme illustré par la Fig. 2. L'architecture de Von Neumann (Fig. 2a) souffre du goulot d'étranglement du bus partagé : chaque instruction et chaque donnée doivent transiter par un canal unique, limitant la bande passante et augmentant la latence.

À l'inverse, l'architecture neuromorphique (Fig. 2b) adopte une topologie distribuée. La mémoire (synapses) est co-localisée directement au sein des unités de calcul (neurones). Cette architecture massivement parallèle élimine le besoin de déplacer les données sur de longues distances, réduisant drastiquement la consommation énergétique [6].

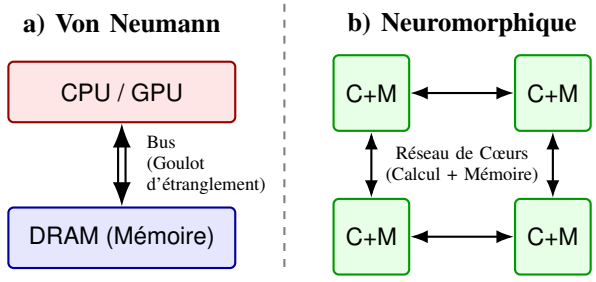

Figure 2. Comparaison conceptuelle. (a) L'architecture de Von Neumann sépare calcul et mémoire. (b) L'architecture neuromorphique co-localise les deux (C+M) pour réduire le mouvement des données.

B. Le Neurone Leaky Integrate-and-Fire (LIF)

L'unité de calcul fondamentale est le neurone *Leaky Integrate-and-Fire* (LIF). Historiquement formalisé par Lapicque [16] et théorisé dans le contexte computationnel moderne par Gerstner [17], ce modèle décrit l'évolution dynamique du potentiel de membrane $V_{mem}(t)$.

Les spikes entrants provenant des neurones pré-synaptiques sont pondérés par les poids synaptiques w_i et intégrés selon l'équation différentielle suivante :

$$\tau_m \frac{dV_{mem}}{dt} = -(V_{mem} - V_{rest}) + R_m \sum_i w_i \delta(t - t_i) \quad (1)$$

Où τ_m est la constante de temps de la membrane, V_{rest} le potentiel de repos, et $\delta(\cdot)$ est la fonction de Dirac modélisant l'arrivée instantanée du spike à l'instant t_i. Lorsque V_{mem} atteint un seuil critique V_{th}, le neurone émet une impulsion (spike) et son potentiel est instantanément réinitialisé. Cette dynamique non-linéaire est cruciale car elle assure la **parcimonie** (*sparsity*) : contrairement aux ANN où tous les neurones sont activés à chaque couche, un neurone SNN reste silencieux et ne consomme de l'énergie que lorsqu'un événement significatif survient [18].

C. Encodage et Communication (AER)

L'efficacité énergétique du système dépend fortement de la stratégie d'encodage de l'information :

— **Rate Coding :** L'information est représentée par la fréquence moyenne des spikes. Bien que robuste au bruit, cette méthode est énergivore car elle nécessite de nombreux spikes pour une précision élevée.

— **Time-to-First-Spike (TTFS) :** L'information réside dans le timing précis du premier spike (latence). Cette méthode est ultra-efficace car elle permet de transmettre de l'information avec un seul événement [19].

Pour le transport des données, la majorité des puces utilisent le protocole **Address-Event Representation (AER)**. Lorsqu'un neurone s'active, il transmet son adresse numérique sur un bus asynchrone, permettant de virtualiser la connectivité massive du cerveau (milliers de synapses par neurone) sur un substrat de silicium limité [20].

D. Apprentissage : De la STDP à l'e-prop

L'apprentissage dans les SNN reste un défi ouvert. Pour l'apprentissage embarqué (*on-chip*), la règle biologique *Spike-Timing-Dependent Plasticity* (STDP) est souvent utilisée. Elle module les poids synaptiques w_{ij} en fonction de la causalité temporelle stricte entre le spike pré-synaptique et le spike post-synaptique [21].

Cependant, la STDP locale peine à optimiser des réseaux profonds pour des tâches complexes. Des algorithmes plus récents comme l'**Eligibility Propagation (e-prop)**, théorisé par Bellec et al. [22] et implémenté matériellement dans la puce ReckOn [9], permettent d'approximer la rétropropagation du gradient (BPTT) en temps réel avec une empreinte mémoire minime, ouvrant la voie à l'apprentissage autonome en périphérie (*edge*).

III. Architectures Neuromorphiques Numériques

Les processeurs neuromorphiques numériques représentent la branche la plus mature et la plus fiable du domaine. Ils exploitent les processus logiques CMOS standard pour simuler la dynamique neuronale de manière numérique. L'avantage principal de cette approche est le **déterminisme** : un neurone numérique garantit une sortie identique pour une séquence d'entrées donnée, indépendamment de la température ou des variations du processus de fabrication [14].

La recherche actuelle continue d'optimiser ces implémentations. Des travaux récents proposent notamment des conceptions de neurones LIF numériques ultra-compacts, optimisés pour réduire la surface de silicium tout en maintenant une précision bit-à-bit, essentielle pour les systèmes basés sur la technologie CTT (*Charge-Trap Transistor*) [23].

A. État de l'art des processeurs numériques

1) Intel Loihi et Loihi 2 (La flexibilité): L'architecture Loihi d'Intel a introduit la plasticité synaptique programmable au cœur du silicium. Sa deuxième itération, Loihi 2, fabriquée avec le procédé Intel 4 (7 nm EUV, « Extreme Ultraviolet »), offre une densité 15 fois supérieure. Une innovation clé est le support des « spikes gradués » (*graded spikes*), qui transportent une charge utile entière (payload) plutôt qu'un simple événement binaire, améliorant la précision [24]. À grande échelle, le système *Hala Point*, intégrant 1 152 puces Loihi 2, démontre une efficacité dépassant 15 TOPS/W en précision 8-bits [25].

2) ReckOn (L'apprentissage à la périphérie): Le processeur ReckOn illustre la spécialisation pour l'apprentissage embarqué à ultra-basse consommation. Fabriqué en 28 nm FD-SOI (Fully Depleted Silicon On Insulator), il implémente un SNN récurrent entraîné via l'algorithme *e-prop*. Son innovation majeure réside dans l'utilisation de traces d'éligibilité pour découpler la mise à jour des poids de l'historique temporel complet, réduisant le surcoût mémoire de l'apprentissage à seulement 0.8 % [9]. Il affiche une consommation de 5.3 pJ par opération synaptique (SOP), soit environ 0.189 TOPS/W, un chiffre qui inclut le coût élevé des mises à jour synaptiques en ligne.

3) SENECA (L'hétérogénéité): Alors que ReckOn optimise une dynamique spécifique, l'architecture SENECA vise la flexibilité via une approche hétérogène. Chaque cœur intègre un contrôleur RISC-V pour gérer les règles de plasticité non-standard et un contrôleur de boucle (*Loop Buffer*) pour accélérer les opérations tensorielles répétitives. En nœud 22 nm, SENECA atteint 2.8 pJ/SOP [26], confirmant la tendance vers des systèmes hybrides « many-core ».

B. Le Goulot d'Étranglement de l'Interconnexion

À mesure que les SNN numériques passent à l'échelle (millions de neurones), le réseau d'interconnexion (NoC) devient le point critique. Les architectures hiérarchiques en arbre se révèlent souvent supérieures aux maillages linéaires pour les implémentations multi-cœurs. L'introduction de l'arbitrage stochastique dans les routeurs permet de réduire la congestion dans les tampons FIFO, améliorant la latence pire-cas par rapport à un arbitrage *round-robin* classique [20].

Néanmoins, une limitation fondamentale persiste, déplacer des paquets numériques, même de manière asynchrone et parcimonieuse, à travers une grande puce consomme une énergie significative (pJ/bit/mm) comparée à l'intégration analogique locale.

C. Défis Physiques de l'Approche Numérique

Malgré leur fiabilité, les architectures numériques se heurtent à trois murs physiques :

1) **Densité SRAM :** Le stockage des poids synaptiques en mémoire SRAM (Static RAM) est coûteux en surface. Une cellule SRAM standard à 6 transistors (6T) occupe une surface bien supérieure aux dispositifs non-volatiles émergents (1T-1R) [10].
2) **Courants de Fuite :** La SRAM étant volatile, elle nécessite une alimentation constante pour retenir les données, créant un plancher de consommation (*leakage floor*) qui empêche d'atteindre un véritable mode veille « zéro watt ».

IV. Signal Mixte et Calcul en Mémoire (IMC)

Si l'architecture numérique élimine le goulot global de Von Neumann, elle reste entravée par la séparation locale entre logique et SRAM, consommant de la puissance dynamique pour déplacer les données. Elle se borne ainsi à *simuler* le neurone.

L'IMC s'affranchit de cette limite en passant à l'*émulation physique*. En exploitant le signal mixte au cœur de la matrice, il supprime le mouvement des données et transforme la mémoire en synapses actives, offrant la densité critique requise pour l'IA embarquée.

A. Principes du Calcul Analogique

L'opération dominante dans les réseaux de neurones (DNN et SNN) est la multiplication matricielle (MVM). L'IMC analogique mappe cette opération directement sur la physique du circuit via une structure *crossbar* :

— Le **poids synaptique** W_{ij} est stocké sous forme de conductance G_{ij} dans un dispositif non-volatile.
— L'**activation d'entrée** V_{in} est appliquée sous forme de tension (ou de largeur d'impulsion).

En vertu de la loi d'Ohm ($I = G \cdot V$) pour la multiplication locale et de la loi des nœuds de Kirchhoff ($\sum I$) pour l'accumulation le long des lignes de bits, le calcul s'effectue dans le domaine analogique avec une complexité temporelle $O(1)$, indépendante de la taille de la matrice [10].

B. Technologies Mémoire Habilitantes (NVM)

Plusieurs technologies de mémoire non-volatile émergentes servent de substrat physique à ces architectures :

— **Resistive RAM (RRAM) :** Stocke l'information en modifiant la résistance d'une couche diélectrique (ex : HfO_2) via la formation de filaments conducteurs. Elle

est hautement scalable ($4F^2$) mais souffre de phénomènes de relaxation de conductance [27].
- **Phase-Change Memory (PCM) :** Repose sur la transition réversible d'un verre de chalcogénure (GST) entre états amorphe (résistif) et cristallin (conducteur). Utilisée dans le cœur HERMES, elle offre un stockage multi-niveaux robuste [11].
- **Ferroelectric FET (FeFET) :** Utilise la polarisation d'un matériau ferroélectrique dans la grille d'un transistor. Des études récentes démontrent une capacité de 2 bits/cellule avec une endurance élevée et une consommation inférieure à la Flash [12].

C. État de l'art des processeurs mixtes

1) NeuRRAM (La flexibilité avant tout): NeuRRAM représente une avancée majeure dans l'IMC basé sur RRAM. Gravée en 130 nm, cette puce intègre 48 cœurs et 3 millions de dispositifs RRAM. Elle introduit l'architecture TNSA (*Transposable Neurosynaptic Array*) qui permet d'entrelacer physiquement les circuits neuronaux dans la matrice, autorisant des flux de données bidirectionnels. De plus, NeuRRAM utilise une lecture en **mode tension** (mesure du taux de décharge) plutôt qu'en mode courant, évitant la saturation et permettant d'activer les 256 lignes simultanément pour une efficacité énergétique doublée par rapport aux références numériques [10].

2) Cœur HERMES (La précision analogique): Le cœur HERMES (14 nm CMOS + PCM) innove par sa méthode de conversion Analogique-Numérique. Au lieu d'utiliser des ADC (*Analog-to-Digital Converters*) classiques coûteux en surface et énergie, il emploie des Oscillateurs Contrôlés par Courant (CCO). Le courant de sortie module la fréquence de l'oscillateur, qui est numérisée par un simple comptage d'impulsions, atteignant une efficacité de 10.5 TOPS/W [11].

3) Macro RRAM 22 nm (Le compromis précision-énergie): Une macro-cellule RRAM de 4 Mb gravée en 22 nm illustre parfaitement le compromis inhérent à l'analogique. Pour des opérations binaires (1-bit), elle atteint l'efficacité spectaculaire de **195.7 TOPS/W**. Cependant, lorsqu'elle est configurée pour une précision de 8 bits, l'efficacité chute à 11.9 TOPS/W [27]. Cela souligne que l'« avantage analogique » est maximal pour les faibles précisions, typiques des réseaux de neurones quantifiés ou impulsionnels.

V. Défis de Conception et Limitations

Si les architectures mixtes promettent une efficacité énergétique théoriquement illimitée, leur déploiement à grande échelle exige un co-design algorithme-matériel pour mapper des topologies complexes sur des crossbars physiquement rigides malgré les non-idéalités analogiques intrinsèques

A. Variabilité et Stochasticité

Contrairement aux bits numériques stables, les états de conductance analogiques sont intrinsèquement bruités. Cette variabilité se manifeste à plusieurs échelles temporelles et spatiales :

- **D2D (Device-to-Device) :** Les imperfections lithographiques lors de la fabrication entraînent des disparités statiques entre les dispositifs. Deux memristors programmés avec la même tension n'auront jamais exactement la même conductance.
- **C2C (Cycle-to-Cycle) :** Le processus de formation et de rupture des filaments conducteurs (dans les RRAM) est stochastique, rendant l'écriture probabiliste plutôt que déterministe.
- **Dérive Temporelle (Drift) :** La conductance des dispositifs (notamment PCM) tend à se relaxer ou à dériver après la programmation, dégradant la précision de l'inférence au fil du temps [18].
- **Chute IR (IR Drop) :** Dans les grandes matrices *crossbar*, la résistance parasite des lignes métalliques provoque une chute de tension non-négligeable, introduisant des erreurs spatiales dans le calcul de la somme pondérée.

B. La Variabilité comme Atout (Beneficial Variability)

Paradoxalement, ce bruit n'est pas toujours nuisible. Dans le contexte des SNN entraînés avec des règles locales non-supervisées, une certaine dose de stochasticité peut agir comme un régularisateur naturel. Des études montrent qu'une variabilité de conductance modérée ($\sigma \approx 5\%$) permet aux réseaux d'échapper aux minima locaux lors de l'apprentissage et d'améliorer la généralisation par rapport à des dispositifs « parfaits », mimant ainsi le bruit synaptique biologique [28].

C. Le Goulot d'Étranglement des Convertisseurs (ADC)

L'interface analogique-numérique constitue le principal verrou énergétique, représentant souvent plus de 60% de la consommation d'un cœur IMC. Trois architectures s'opposent selon les contraintes ciblées :

- **SAR (Successive Approximation Register) ADC :** Standard de l'industrie pour l'efficacité à faible précision (4-8 bits), il est cependant limité par sa surface qui croît exponentiellement avec la résolution, le réservant aux réseaux quantifiés.
- **Sigma-Delta ($\Sigma\Delta$) :** Indispensable pour la fiabilité. Par suréchantillonnage, il lisse le bruit stochastique inhérent aux RRAM (bruit télégraphique), échangeant la latence contre une robustesse accrue [10].
- **Conversion Temporelle :** Adoptée par l'architecture HERMES [11], cette approche transforme le courant en intervalles de temps. Elle élimine les références de tension coûteuses en surface et tire profit de la vitesse de commutation des transistors numériques modernes.

D. Intégration 3D : La Conquête du Volume

Face à la saturation inéluctable de la surface du silicium, une question fondamentale s'impose : *pourquoi rester confiné aux limites du plan 2D quand l'expansion vers la troisième dimension offre une rupture de densité ?* Pour maximiser la connectivité synaptique et minimiser les délais d'interconnexion, l'industrie adopte désormais une approche stratifiée :

1) Collage Hybride (Hybrid Bonding): Cette technologie permet d'augmenter la densité d'interconnexion verticale au-delà de 12 000 contacts/mm^2. Elle propulse l'état de l'art des accélérateurs numériques comme l'*AMD Instinct MI300X* qui, en empilant matrices de calcul et mémoire HBM3, atteint une bande passante de 5.3 To/s et une efficacité de **7 TOPS/W** (FP8) [29]. Ce chiffre sert de référence « plancher » pour les architectures neuromorphiques mixtes visant les 100 TOPS/W.

2) Intégration Monolithique (3D-SoC): Pour franchir le mur de la densité, l'intégration monolithique (ex : CoolCube) grave séquentiellement les couches de transistors les unes sur les autres. Cette approche élimine les contraintes d'alignement mécanique, offrant une densité de vias **10 000 fois supérieure** aux technologies classiques. Cela autorise une connectivité verticale à la granularité de la synapse individuelle, sous réserve de maîtriser les budgets thermiques de fabrication (< 500°C) [30].

VI. Domaines d'Application : De l'Edge au Cloud

La diversité des architectures neuromorphiques (numériques vs mixtes) permet de cibler un spectre applicatif extrêmement large, allant de l'implant médical sous contrainte thermique stricte jusqu'à l'accélération massive dans les centres de données.

A. Implants Médicaux Éco-efficients

L'une des frontières les plus prometteuses pour le matériel neuromorphique concerne les dispositifs médicaux implantables (prothèses neuronales, pacemakers intelligents). Ces systèmes opèrent sous des contraintes critiques : une autonomie sur batterie de plusieurs années et une dissipation thermique minimale (< 1°C d'échauffement tissulaire) pour éviter la nécrose. L'approche SNN à signal mixte est ici idéale. Elle permet de traiter les signaux biologiques bruts (ECG « Électrocardiogramme », EMG « Électromyogramme », LFP « Local Field Potential ») localement, en temps réel, sans le coût énergétique prohibitif de la transmission sans fil des données brutes vers le cloud. Des prototypes gravés en 65 nm ont démontré la capacité d'apprendre et de classifier des gestes via EMG avec une consommation de l'ordre du microwatt [31]. Pour ce domaine, la métrique pertinente n'est plus le TOPS/W, mais l'« Information par Joule ».

B. Fusion de Capteurs et Calcul In-Sensor

Au-delà du traitement classique, la technologie RRAM permet de fusionner le capteur et le processeur (*In-Sensor Computing*) :

- **Localisation d'Objets (pMUT) :** L'intégration de transducteurs ultrasonores piézoélectriques (pMUT) directement sur une matrice RRAM permet de traiter les échos analogiques à la source, éliminant les ADC frontaux énergivores pour des tâches de localisation 3D.
- **Routage TCAM :** Les mémoires RRAM sont utilisées pour créer des mémoires adressables par contenu ternaires (TCAM). Cette approche permet de découpler le chemin de recherche du chemin de programmation, autorisant une reconfiguration dynamique de la topologie du réseau sans interrompre l'inférence [10].

C. Centres de Données Durables

Si l'*Edge AI* vise le milliwatt, les centres de données affrontent une crise du mégawatt. L'intégration d'accélérateurs neuromorphiques numériques (type Intel Loihi [24]) dans les racks de serveurs offre une voie de délestage pour des charges de travail spécifiques. Contrairement aux GPU optimisés pour le calcul matriciel dense, les processeurs neuromorphiques excellent dans les problèmes d'optimisation combinatoire (ex : problème du voyageur de commerce, routage de paquets) et l'analyse de graphes clairsemés. Leur asynchronisme permet de réduire drastiquement l'empreinte carbone globale de l'infrastructure pour ces tâches non-euclidiennes [25].

VII. Analyse Comparative

Le choix entre une architecture numérique et une architecture à signal mixte ne se résume pas à une simple course à la performance ; il implique un compromis complexe entre l'efficacité énergétique, la précision de calcul et la flexibilité de programmation. Le Tableau I synthétise les différences fondamentales identifiées dans notre étude.

Caractéristique	**Numérique (ex : Loihi 2)**	**Mixte / IMC (ex : NeuRRAM)**
Principe de Calcul	Logique Booléenne	Physique (Analogique)
Efficacité	≈ 15 TOPS/W (Hala Point)	$10 - 195$ TOPS/W (selon précision)
Densité Mémoire	Faible (SRAM)	Haute (RRAM/PCM)
Précision	Déterministe (INT8)	Stochastique (1 à 8 bits)
Immunité au Bruit	Totale	Faible (compensée par l'entraînement)
Maturité Matérielle	Industrielle (Procédé 7 nm)	Prototype Recherche (Labo / 130 nm)
Maturité Logicielle	**Élevée** (SDKs existants)	**Faible** (Co-Design requis)

TABLE I
COMPARAISON DES ARCHITECTURES NUMÉRIQUES VS MIXTES BASÉE SUR L'EFFICACITÉ ET LA DENSITÉ.

A. Le Fossé de l'Efficacité Réelle

Théoriquement, le calcul analogique offre un avantage d'efficacité de deux ordres de grandeur ($100\times$). Cependant, en pratique, cet écart se réduit à un facteur de $2\times$ à $10\times$ à cause du coût énergétique des périphériques (ADC/DAC) et des circuits de contrôle numérique nécessaires.

B. L'Avantage Décisif de la Non-Volatilité

Néanmoins, pour les applications de veille permanente (*always-on sensing*), l'architecture mixte possède un atout majeur : la non-volatilité. Contrairement à la SRAM numérique qui consomme du courant de fuite en permanence pour retenir les données, une mémoire RRAM ou PCM peut être totalement éteinte sans perte d'information, permettant une consommation de veille quasi-nulle [1].

VIII. Conclusion et Perspectives

Le domaine neuromorphique se trouve à une jonction critique. L'analyse comparative de ce rapport souligne une bifurcation technologique majeure pour l'avenir de l'IA durable.

A. Le Dilemme : Fiabilité vs Efficacité

Les **architectures numériques** (Loihi [24], ReckOn [9]) offrent une fiabilité industrielle et des outils matures, mais se heurtent à un « plafond d'efficacité » imposé par la densité de la SRAM. À l'inverse, les **technologies mixtes** (NeuRRAM [10], HERMES [11]) promettent de briser ce plafond en visant les 100 TOPS/W, mais leur adoption reste freinée par la variabilité des composants et le coût des conversions analogique-numérique.

B. Vers une Symbiose Hybride

L'avenir ne réside pas dans un choix binaire, mais dans une architecture hétérogène convergente :

1) **Calibration Numérique :** L'intégration de cœurs de contrôle flexibles et open-source (RISC-V) devient indispensable pour orchestrer les accélérateurs analogiques. Ils permettent de calibrer dynamiquement la stochasticité et d'apporter la flexibilité algorithmique manquante aux puces purement IMC [26].
2) **Intégration 3D Dense :** Pour maximiser la bande passante entre ce contrôle numérique et les matrices mémoires, l'empilement vertical (via *Hybrid Bonding*, *CoolCube* [30]) est la clé de voûte des futures performances.
3) **Co-Design Algorithme-Matériel :** Le matériel seul ne suffit pas. Comme l'indique la *Roadmap* 2022 [1], il est urgent de développer des réseaux de neurones capables non seulement de tolérer le bruit du matériel, mais de l'exploiter.

En acceptant une précision stochastique au profit d'une efficacité extrême, l'approche hybride forge le paradigme nécessaire pour intégrer l'intelligence dans le monde physique.

Références

[1] D. V. Christensen, R. Dittmann, B. Linares-Barranco *et al.*, "2022 roadmap on neuromorphic computing and engineering," *Neuromorphic Computing and Engineering*, vol. 2, no. 2, p. 022501, 2022.

[2] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," *ISSCC Digest of Technical Papers*, pp. 10–14, 2014.

[3] I. O'Connor and A. Bosio, "Introduction," Séminaire : Sustainable and Trustworthy Hardware Architectures for AI, 2025, PSM-ESE.

[4] Celestial AI, "Technology architecture : The memory wall," https://www.celestial.ai/technology-1, 2024.

[5] J. Yang, L. Wang, and Y. Chen, "Biologically-inspired neuromorphic computing," *Signal Transduction and Targeted Therapy*, vol. 8, no. 1, p. 319, 2023.

[6] C. D. Schuman, T. E. Potok *et al.*, "A survey on neuromorphic architectures for running artificial intelligence algorithms," *arXiv preprint arXiv :1705.06963*, 2017.

[7] Intel Corporation, "A look at Loihi 2 - Open Neuromorphic," https://open-neuromorphic.org, 2025.

[8] F. Akopyan *et al.*, "TrueNorth : Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, 2015.

[9] C. Frenkel and G. Indiveri, "ReckOn : A 28nm sub-mm^2 task-agnostic spiking recurrent neural network processor enabling on-chip learning over second-long timescales," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[10] W. Wan, R. Kubendran, C. Schaefer *et al.*, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, pp. 504–512, 2022.

[11] R. Khaddam-Aljameh, M. Stanisavljevic, J. Mas *et al.*, "HERMES core – a 14nm CMOS and PCM-based in-memory compute core using an array of 300ps/LSB linearized CCO-based ADCs," in *2021 Symposium on VLSI Circuits*, 2021, pp. 1–2.

[12] S. De, F. Müller, H.-H. Le *et al.*, "READ-optimized 28nm HKMG multi-bit FeFET synapses for inference-engine applications," *IEEE Journal of the Electron Devices Society*, vol. 10, 2022.

[13] J.-S. Kim, L. Park, M. Kwon *et al.*, "A 18.7 TOPS/w mixed-signal spiking neural network processor with 8-bit synaptic weight on-chip learning," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 8, pp. 2362–2376, 2022.

[14] C. Frenkel *et al.*, "A 0.086-mm^2 12.7-pj/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 1, pp. 145–158, 2019.

[15] W. Maass, "Networks of spiking neurons : the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.

[16] L. Lapicque, "Recherches quantitatives sur l'excitation électrique des nerfs traitée comme une polarisation," *J. Physiol. Pathol. Gen*, vol. 9, pp. 620–635, 1907.

[17] W. Gerstner and W. M. Kistler, *Spiking neuron models : Single neurons, populations, plasticity*. Cambridge University Press, 2002.

[18] F. Moro, "Memristive analog computing and innovative sensors for neuromorphic systems," Ph.D. dissertation, Université Grenoble Alpes, 2023.

[19] S. Thorpe, A. Delorme, and R. Van Rullen, "Spike-based strategies for rapid processing," *Neural networks*, vol. 14, no. 6-7, pp. 715–725, 2001.

[20] K. Boahen, "Point-to-point connectivity between neuromorphic chips using address events," *IEEE Transactions on Circuits and Systems II*, vol. 47, no. 5, pp. 416–434, 2000.

[21] G.-q. Bi and M.-m. Poo, "Synaptic modification in cultured hippocampal neurons : dependence on spike timing, synaptic strength, and postsynaptic cell type," *Journal of neuroscience*, vol. 18, no. 24, pp. 10 464–10 472, 1998.

[22] G. Bellec, F. Scherr, A. Subramoney, E. Hajek, D. Salaj, R. Legenstein, and W. Maass, "A solution to the learning dilemma for recurrent networks of spiking neurons," *Nature communications*, vol. 11, no. 1, p. 3625, 2020.

[23] O. T. Gumus, M. Karimi, and B. Vaisband, "Digital LIF neuron for CTT-based neuromorphic systems," in *Proceedings of the Great Lakes Symposium on VLSI 2023*, 2023, pp. 561–566.

[24] M. Davies, N. Srinivasa *et al.*, "Loihi : A neuromorphic manycore processor with on-chip learning," in *IEEE Micro*, vol. 38, no. 1, 2018, pp. 82–99.

[25] Intel Corporation, "Intel builds world's largest neuromorphic system to enable more sustainable AI," www.intc.com, 2024, accessed : 2025-12-20.

[26] S. Gazanagha *et al.*, "SENECA : building a fully digital neuromorphic processor, design trade-offs and challenges," *Frontiers in Neuroscience*, vol. 17, p. 1187252, 2023.

[27] C.-X. Xue, J.-M. Hung, H.-Y. Kao *et al.*, "16.1 a 22nm 4Mb 8b-precision ReRAM computing-in-memory macro with 11.91 to 195.7TOPS/w for tiny AI edge devices," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, 2021, pp. 245–247.

[28] M. Suri and V. Parmar, "Bio-inspired stochastic computing using resistive RAM," *Frontiers in Neuroscience*, vol. 14, p. 636, 2020.

[29] J. Wuu, R. Agarwal, M. Ciraula *et al.*, "3d V-Cache : The implementation of a hybrid-bonded 64MB stacked cache for a 7nm x86-64 CPU," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 428–429.

[30] M. M. Shulaker, G. Hills, R. S. Park *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature*, vol. 547, pp. 74–78, 2017.

[31] E. Donati, M. Payvand, N. Risi, R. Krause *et al.*, "A compact online-learning spiking neuromorphic biosignal processor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 6, pp. 1127–1138, 2022.