*AI Engineering*
https://www.sciltp.com/journals/aieng

*Review*

# A Review of AI Computing Hardware for Enabling 6G Wireless Networks: Neuromorphic Chips, NPUs, and TPUs

Jason Lee* and Yu-Dong Yao

Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA
* Correspondence: jlee76@stevens.edu

**How To Cite**: Lee, J.; Yao, Y.-D. A Review of AI Computing Hardware for Enabling 6G Wireless Networks: Neuromorphic Chips, NPUs, and TPUs. *AI Engineering* **2026**, *2*(2), 10. https://doi.org/10.53941/aieng.2026.100010

Received: 30 November 2025
Revised: 29 January 2026
Accepted: 7 June 2026
Published: 16 July 2026

**Abstract:** The introduction of 6G Networks and its deep integration of Artificial Intelligence (AI) necessitates powerful and efficient hardware processors to manage massive data flows, ensure network reliability, and support demanding use cases. To overcome these challenges, the industry is shifting away from conventional processing architectures (CPUs and GPUs) toward specialized hardware accelerators. This paper reviews three critical categories of processing chips: Neuromorphic chips, Neural Processing Units (NPUs), and Tensor Processing Units (TPUs). It also reviews other emerging chips that are still relatively unknown such as Quantum Processing Units (QPUs) and Photonic Integrated Circuits (PICs). Specifically, this paper will review different research tasks and display tables of hardware efficiency and wireless system metrics.

**Keywords:** neural processing unit (NPU); tensor processing unit (TPU); compute-in-memory (CIM); spike neural networks (SNNs); ultra-reliable low-latency communication (URLLC); multi-processor system-on-chip (MPSoC); photonic integrated circuits (PICs); quantum computing; energy efficiency; systolic array; digital twinning; integrated sensing and communication (ISAC)

## 1. Introduction

The escalating complexity of modern wireless communication systems and the introduction of the next-gen, 6G Wireless Networks, demonstrates a massive shift into integrated Artificial Intelligence (AI) natively. AI is now a core functional element for managing massive data flows, ensuring optimal network reliability, and achieving unprecedented performance goals. This necessity stems from 6G's vision of supporting use cases like integrated sensing and communication (ISAC) and ubiquitous computing, which demand pervasive, real-time intelligence across the network [1]. The shift toward Neuromorphic Semantic Communications exemplifies this [2], where event-driven principles are leveraged to reduce communication power and tailor resource consumption directly to the input signal's semantic meaning [3]. This core integration of AI is crucial for supporting demanding network applications such as self-navigating vehicles and metaverse experiences.

The tremendous potential of 6G with AI, however, is very dependent on the hardware required to execute its efficiency. AI Models require billions of operations per second, creating a massive computational load. To sustain this kind of power, the industry has moved away from conventional processing architectures, such as CPUs and GPUs, toward specialized processing units such as Neural Processing Units (NPUs) and Tensor Processing Units (TPUs) that can demonstrate the kind of power needed to utilize the power of AI while maintaining energy efficiency and low latency. This search for non-conventional, highly efficient computing is driven by the fact that even minor performance improvements at the hardware level translate directly into major energy savings at the system level [4].

Meeting the rigorous demands of 6G in AI imposes severe constraints on computer systems. The high complexity of conventional algorithms as well as tight energy efficiency and power consumption constraints calls for solutions outside of the traditional hardware architecture. As the scale of AI deployments grows (e.g., training massive models), it becomes imperative that hardware support not only massive parallelism but also optimize for

compute-in-memory (CIM) technologies to break the memory bottleneck [5]. Concepts like the Compute Continuum Layer (CCL) and Zero-Trust Layer (ZTL) formalize this necessity in 6G architecture, ensuring that resources from the device to the cloud are seamlessly and securely managed, allowing computation to be opportunistically placed where it is most efficient [1]. This shift towards highly specialized hardware is necessary to enable the next and future generations of wireless communication systems.

The competitive advantage in future generation networks lies in superior hardware performance metrics. The AI industry is currently locked in a race to boost processing speed while reducing latency and power consumption. For communication systems, this is crucial for enabling Ultra-Reliable Low-Latency Communication (URLLC) [1] and achieving deterministic performance in highly stochastic wireless environments. Performance gains often come from innovations that streamline data paths, such as In-network Computing (INC) architectures that bypass CPU bottlenecks using shared data planes and specialized accelerators like Data Processing Units (DPUs) and other non-conventional processing units [6].

The remainder of this article is organized as follows. Section 2 goes in-depth on 6G's shift towards AI and networks. Section 3 will give a general overview of all the major hardware accelerators covered in this article. Section 4 discusses Neuromorphic chips in depth and analyze its hardware efficiency and wireless performance. Section 5 discusses NPU and TPU chips in depth and analyzes its hardware efficiency and wireless performance. Section 6 discusses other processors and summarizes a combination of their hardware efficiency and wireless performance. The article concludes with Section 7.

## 2. The Revolutionary Architecture of 6G

6G represents a fundamental shift from simple data connectivity to a unified platform where AI and computing are native to the network fabric. The evolution supports a more goal-oriented communication where the network prioritizes the purpose of the data rather than merely ensuring the accurate transmission of bits. This architecture creates a seamless combination of the physical, digital, and human worlds through a distributed learning framework that spans from the core to the extreme edge [7].

6G aims to achieve ultra high speeds and ultra low latencies, far beyond 4G and 5G, aiming to achieve speeds as high as 1 Tbps. To support these extreme data rates as well as maintain ultra low latencies, this architecture utilizes Terahertz frequencies and optical wireless communications, necessitating advanced hardware innovations like Reconfigurable Intelligent Surfaces (RIS) and Photonic Integrated Circuits (PICs) [8]. 6G can also effectively function as a virtualized extension of a device's own hardware when resource-constrained devices such as robots need to offload computational tasks to the network [1]. Finally, 6G Architecture is also exploring neuromorphic chips and optical computing technologies to achieve high performance, energy efficient processing.

## 3. Overview of all Hardware Accelerators and Processors

Hardware accelerators and advanced processors are critical for 6G networks because they provide the massive parallel computation needed to handle extremely high data rates, real-time signal processing, and AI-intensive network optimization. By offloading intensive tasks, they enable lower latency, higher efficiency, and intelligent wireless systems.

(1) Neuromorphic chips are a new, modern type of hardware designed to mimic the human brain's architecture and function.
(2) The Central Processing Unit (CPU) is a general purpose processor that is good for sequential tasks and system control. However, it is less efficient than specialized processors that require massive parallel-processing capabilities.
(3) The Application-Specific Integrated Circuit (ASIC) is a chip customized for a specific purpose, offering the highest efficiency, performance, and power savings for a certain task.
(4) The Graphics Processing Unit (GPU) is great at parallel computations with thousands of cores, making it perfect for deep learning models.
(5) The Neural Processing Unit (NPU) is specifically built to efficiently execute the matrix and vector arithmetic common in AI and machine learning models.
(6) The Tensor Processing Unit (TPU) is built by Google to speed up the immense matrix calculations that form the core of TensorFlow machine learning workloads.
(7) The Field Programmable Gate Array (FPGA) is a reconfigurable integrated circuit that allows developers to customize the logic gates after manufacturing to create highly tailored, parallel hardware acceleration.
(8) The Quantum Processing Unit (QPU) utilizes quantum phenomena like superposition and entanglement to perform calculations exponentially faster than classical processors for certain complex problems.
(9) The Multi-Processor System-on-Chip (MPSoC) is an integrated circuit that combines multiple processor cores with other components like programmable logic, memory, and I/O interfaces onto a single chip.

## 4. Neuromorphic Focus

### *4.1. Why Neuromorphic Chips?*

Since neuromorphic chips are specifically designed to mimic the human brain's structure, they have been a main focus of optimization and modeling for integrating AI within 6G Networks [9]. Mimicking the brain structure [10] has led to the significant improvements within the 6th gen network architecture while maintaining a strong power-to-energy consumption ratio. Table 1 gives a summary of in-depth metrics of neuromorphic processors.

**Table 1.** In-depth metrics of neuromorphic processors.

| **Neuromorphic Processor Metrics** | | | |
|---|---|---|---|
| ***Research Task*** | ***Hardware Efficiency Metrics*** | ***Wireless System Metrics*** | ***References*** |
| Edge AI & IoT Optimization | **Power Consumption:** Measured in milliwatts (mW) or microwatts ($\mu$W) for always-on operation. **Energy Efficiency:** High TOPS/W compared to traditional CPUs/GPUs. **Wake-Up Latency:** Sub-millisecond wake-up times for event-driven tasks. | **Bandwidth Usage:** Reduced by local processing (disconnected edge computing). **Latency:** Minimized data movement reduces end-to-end latency. | [2,5,10,11] |
| Enhancing Communication Systems | **Computational Power:** 1000$\times$ more power efficient than CPUs/GPUs **Operation Reduction:** >45% reduction in operations for MIMO detection versus conventional methods. | **Bit Error Rate (BER):** Reaches $10^{-4}$ when the antenna-to-stream spatial ratio is $>4$. **Spectral Efficiency:** Optimized spectrum usage through AI-driven management. **Latency:** <1 millisecond latency for 6G applications. | [9,11] |
| Real-Time Autonomous Systems | **Inference Latency:** <50 ms required for safe operation. **Frame Rate:** Processing high-resolution sensor data at 30–60 FPS. **Energy Efficiency:** Optimizing performance per watt to fit vehicle power budgets. | **Ultra-Low Latency:** Essential for V2X communication, collision avoidance, and cooperative driving. **Reliability:** Maintaining robust communication in fast-changing environments. | [12] |
| Hardware Acceleration & Compute-in-Memory | **Energy per Operation:** Extremely low energy (e.g., 0.15 pJ per synaptic event). **Memory Bandwidth:** >1 TB/s of internal bandwidth for feeding compute arrays. | **Throughput:** Supports high-throughput data streams (e.g., 100 MHz bandwidth). | [5,10,11] |

### *4.2. Optimization Methods*

Optimization Methods with Neuromorphic Chips often prioritize on model compression and resource-aware adaptation to gain maximum performance and energy efficiency.

Model compression shrinks the size and computational demands of AI tools on processor chips. Neuromorphic chips achieve this by doing the following:

Quantization reduces the numerical precision of neural network weights and activations from 32-bit floating point to 8-bit, 16-bit, or lower precision formats [12]. This substantially reduces the memory footprint and accelerates inference speed, enabling faster inference and a significant reduction in memory usage, often with minimal impact on accuracy.

Pruning involves removing weights from a neural network that contribute little to the final output, while sparsity techniques encourage networks to have many zero-valued weights, allowing for more efficient computation using sparse matrix operations. By pruning and leveraging sparsity, it is possible to reduce model size without sacrificing performance, speed up inference by skipping unnecessary computations, and improve energy efficiency by minimizing memory accesses [12].

Neural Architecture Search (NAS) automates the design of artificial neural network architectures, helping in optimizing topology, node connections, and operators [12]. NAS is broken down into three main components: Search Space, Search Strategy, and Evaluation Strategy. These three components respectively define the type of

neural network to optimize, defines the approach to explore the search space, and evaluates the performance of a potential design without constructing it.

AI techniques are directly applied to optimize communication protocols and physical layer (PHY) processing, particularly in new wireless domains. Some techniques include:

(1) Multi-user, multi-input multi-output (MU-MIMO) technology has been central to the evolution towards future wireless networks. To enable this kind of technology, Quadratic Unconstrained Binary Optimization (QUBO) [11] is used to utilize quantum-inspired algorithms to optimize MU-MIMO in neuromorphic processors.

### *4.3. Hardware Efficiency*

Hardware efficiency is defined by maximizing performance per watt. While maximum hardware efficiency is dependent on each computing processor, neuromorphic processor's brain-like architecture can result in significantly higher speed while reducing power consumption [13].

Spike Neural Networks (SNNs) are artificial neural networks that mimic natural neural networks. They replicate the function of biological neural cells and are expected to achieve much higher performance with a much smaller energy footprint compared to conventional architectures [2]. Neuromorphic communications deploy spike neural networks to encode and decode semantic information, reducing the communication power by communicating asynchronously via sparse impulse radio (IR) waveforms.

Compute-in-memory (CIM) is a model that places processing capabilities directly inside or alongside memory such as analog RRAM or digital SRAM to perform calculations where the data is stored, rather than transferring to CPU registers first [5]. This significantly improves power usage and performance of moving data between the memory and processor.

### *4.4. Wireless System Metrics*

Neuromorphic processors emphasize minimal power, high throughput, and low latency communication. The core performance metrics that represent a significant leap from previous generation networks include Data Rate/ Throughput, Latency, Energy Efficiency, and Security.

The overall performance is often measured by analyzing the trade-off between competing metrics, which can be tuned by the system hyperparameters $\lambda$. The overall system performance must manage the inherent tension between these factors, as demonstrated by trade-offs, such as optimizing Expected Loss ($L(\lambda)$) versus minimizing Energy Consumption ($E(\lambda)$), which are dynamically controlled by tuning system parameters [2].

## 5. NPU and TPU Focus

### *5.1. Why NPU and TPU are Unique among the Processors*

General NPUs and TPUs stand amongst the multitude of different processing units by specializing in the mathematical background of tensor operations and matrix multiplication that create the basis of deep learning and other AI applications.

NPUs usually use dataflow architectures that are optimized to reflect the structure of neural networks. They manage data dependencies and movement directly in specialized hardware, contrasting with the thread management required by GPUs.

TPUs employ a systolic array architecture where data flows through a large grid of Arithmetic Logic Units (ALUs) without intermediate writes to memory. This minimizes the Von Neumann bottleneck and maximizes data reuse, making them highly efficient for massive matrix multiplications required for Transformer models [14]. Table 2 gives a summary of in-depth metrics of NPU and TPU processors.

### *5.2. Optimization Methods*

These processors primarily utilize their advanced data management techniques to optimize their usage within 6G Networks:

Preemption allows processors to allow high priority tasks such as URLLC to interrupt and automatically allocate resources.

Spatio-temporal sharing [14] combines both spatial and temporal sharing to optimize resource utilization in two dimensions. Tasks are allocated to different resources concurrently (spatial sharing) while also being scheduled to share the same resources at different times (temporal sharing). This hybrid approach improves utilization of modern multi-core and multi-accelerator systems.

Remote Direct Memory Access (RDMA) [6] allows two computers to transfer data directly between their

memory without involving the processor on either end. This significantly increases performance by lowering latency and freeing up resources.

**Table 2.** In-depth metrics of NPU and TPU processors.

| NPU and TPU Processor Metrics | | | |
|---|---|---|---|
| ***Research Task*** | ***Hardware Efficiency Metrics*** | ***Wireless System Metrics*** | ***References*** |
| AI Accelerator Optimization & In-Network Computing | **Resource Utilization:** Maximizing the usage of Processing Elements (PEs) and Matrix Multiply Units (MXUs) through dynamic fine-grained allocation. **Memory Access Reduction:** Using dataflow architectures such as systolic arrays to minimize energy-intensive off-chip memory access. **Pipeline Efficiency:** Bypassing NPU/TPU intervention during data transfer via RDMA for direct accelerator communication. | **System Throughput ($T_p$):** Optimized allocation can increase throughput by up to $15\times$. **End-to-End Latency:** Accelerator-chained processing can reduce service delay by over 28%. **Jitter:** Hardware offloading can improve jitter by up to 95%. | [3,8,15] |
| Energy Efficiency & Intelligent Integrated Circuits | **Sparsity Exploitation:** Reducing dynamic power by processing only non-zero activations (zero-skipping). **Active Duty Cycle:** Minimizing active time of the main processor or radio through wake-up receivers. **Power Consumption ($P_c$):** Reducing energy per task through DVFS and efficient circuit design. | **Energy-Delay Product:** Balancing speed and energy usage. **Wake-Up Latency:** Time required to activate the main radio from deep sleep. | [9,11] |
| Security, Sensing, and Reliability Assurance | **Model Footprint:** Reducing memory and compute requirements for ML-based security tasks. **Simulation Fidelity:** Accuracy of Digital Twin models for threshold tuning. **Sensing Resolution:** Hardware capabilities for high-precision localization and sensing/communication integration. | **Reliability Guarantees:** Statistical tools such as Learn-Then-Test to bound error rates. **Detection Accuracy:** Precision and recall for IDS and sensing tasks. **Spectral Efficiency:** Efficient allocation of spectrum for sensing and communication. | [1,3,8,15] |

### *5.3. Hardware Efficiency*

The efficiency of NPUs and TPUs is largely led by their ability to reduce memory movement and exploit data characteristics within limited power usage.

A major constraint on hardware efficiency is the memory wall because accessing off-chip memory consumes a significant amount of energy. TPUs and NPUs's use of dataflows helps mitigate this by reusing data stored in local, energy efficient memory, such as SRAM [14]. For example, Eyeriss [14], a type of NPU, leverages a row-stationary dataflow to optimize energy efficiency in CNNs by minimizing data movement and maximizing local data reuse on a spatial architecture.

NPUs and TPUs deploy dynamic management strategies to align energy expenditure. Dynamic voltage and frequency scaling (DVFS) [15] is a widely used power management method where the processor lowers its clock speed and corresponding voltage supply during periods of low activity or simple tasks. This scaling approach significantly reduces levels of energy consumption, especially in memory-bound tasks. Dynamic fine-grained allocation [14] is a technique that addresses resource underutilization by allowing multiple machine learning workloads to share the hardware concurrently.

### *5.4. Wireless System Metrics*

Rigorous verification methods are essential for any sensitive applications that must guarantee latency and reliability regardless of dynamic hardware allocation. Digital Twinning Learn the Test (DT-LTT) [3] enhances the spectral efficiency of a direct application of LTT through a digital twin-based pre-selection of candidate thresholds for sensing, detection, and decision making, particularly during the final validation step on the physical hardware.

To achieve the low latency verified by methods like DT-LTT, foundational architectural changes are implemented to accelerate data flow. To avoid bottlenecks with hardware, Remote Direct Memory Access (RDMA) [6] allows a network device to transfer data directly to and from application memory on another system, bypassing a processor's latency. For time-critical data paths, RDMA is a necessity. In addition, at the network layer, simplifying the Radio Access Network (RAN) structure is possible by implementing the Lower Layer Split (LLS) [1]. It streamlines the network hierarchy and is crucial for physically supporting advanced spatial communication techniques like Distributed-MIMO.

These techniques contribute to significant boosts in reducing latency, reducing jitter, and boosting system throughput. For instance, the In-network Service Acceleration Platform (ISAP)'s accelerator [6] chained processing, led by NPUs, reduces latency by 28% and jitter by 95% in data processing, and achieves improvements in overall network delay and jitter by 97% and 99% respectively, compared to a CPU. As mentioned in Hardware Efficiency, dynamic fine-grained allocation also increases system throughput in deep learning scenarios by 15×.

## 6. Other Processing Units Focus

### *6.1. Advanced Processor Architectures Beyond Neuromorphic, NPU, and TPU*

Neuromorphic processors, other NPUs, and TPUs excel at deep learning workloads, other processors also excel in other computational areas to optimize AI within 6G Networks. GPUs still remain a powerhouse staple in in accelerating high-throughput parallel workloads in RANs [16]. MPSoCs are becoming increasingly common in edge-computing scenarios due to its diverse integration of multiple computing units, combining CPUs, GPUs, and Deep Learning Accelerators (DLAs) to balance energy and efficiency accordingly [17].

Beyond silicon-based electronics, light-matter based electronics such as PICs and Optical Neural Networks (ONNs) perform linear operations such as Matrix Vector Multiplication (MVM) at the speed of light [4]. QPUs are also being explored to solve NP-hard (nondeterministic polynomial time) optimization problems used in wireless networks, such as Maximum Likelihood detection and Linear Detection. A QPU's quantum state enables physics-based principles to navigate problem solving more efficiently than classical electronics [18]. Table 3 gives a summary of in-depth metrics of advanced processor architectures beyond neuromorphic, NPU, and TPU.

**Table 3.** In-depth metrics of other processors.

| Other Processor Metrics | | | |
|---|---|---|---|
| ***Research Task*** | ***Hardware Efficiency Metrics*** | ***Wireless System Metrics*** | ***References*** |
| Edge AI & Heterogeneous Computing | **Energy Gains:** Integrating DVFS into the process on edge GPUs can achieve energy gains up to 53%. **Energy Efficiency:** Dynamic mapping frameworks on MPSoCs achieve up to 2.1× greater energy efficiency than GPU-only deployments. | **Training Time:** GPU-accelerated RAN testbeds reduce AI model training time by 32% without sacrificing accuracy. **Latency/Efficiency Gains:** Workload mapping strategies yield approximately 1.57× latency speedup and 3.38× energy efficiency gains for GNNs. | [16,17] |
| Physics-Based Computing | **Computational Speed:** Hybrid hardware architectures integrating digital electronics with analog photonics show throughputs 2.8–14× faster than traditional GPUs. **Energy Reduction:** Photonic architectures can reduce energy consumption by about 25%. **Energy Dissipation:** Reverse quantum computation strategies enable operations with theoretical zero long-term energy dissipation. | **Benchmarks:** Quantum solvers for MIMO detection are benchmarked using TTS and TTB to meet wireless latency requirements. **Response Rate:** MZIs used in photonic computing demonstrate response rates of 10 Gb/s with nonlinear activation functions. | [4,18] |
| 6G System Intelligence Architecture | **Sustainability:** 6G targets energy efficiencies around 1 pJ/bit. **Goal-Oriented Metrics:** Efficiency is shifting toward "Energy per Goal" or "Energy per Inference", reflecting task completion rather than raw throughput. | **Foundational KPIs:** Peak data rates up to 1 Tbps, air-interface latency as low as 0.1 ms, and reliability up to 99.99999%. **Scale and Mobility:** Systems must support up to $10^7$ devices/km$^2$ and mobility up to 1000 km/h, imposing significant throughput and adaptability constraints. | [7,19,20] |

### *6.2. Optimization Methods*

The diverse complexity of processing units calls for an equally diverse set of optimization methods to fully utilize the extent of AI in 6G. Similar to NPUs and TPUs, GPUs can utilize DVFS settings, so it can use Hardware-aware NAS for designing efficient Deep Neural Networks (DNNs). This with defined hardware parameters enable the creation of frameworks such as Hardware-aware Dynamic Neural Architecture Search (HADAS) or Dynamic Clock Frequency Scaling for Hardware-aware NAS that can maximize AI accuracy while minimizing latency and energy consumption [17].

MPSoCs utilizes Deep Learning Accelerators (DLAs) which offer superior energy efficiency for specific inference workloads. Frameworks like "Map-and-Conquer" explicitly target these heterogeneous platforms, using evolutionary algorithms to partition dynamic neural network workloads across DLAs and GPUs based on input complexity. This approach allows for the dynamic mapping of inference stages, ensuring that simpler inputs are processed by the most energy-efficient units (e.g., DLAs), while more complex tasks utilize high-performance units, thereby maximizing the energy-latency trade-off in edge AI applications [17].

Quantum and Photon-based optimizations are relatively unknown compared to some other techniques created for more traditional electronics. Quantum Annealing (QA) and its variants, such as Reverse Annealing (RA), allow for warm-started optimization that can correct initial classical solutions toward global optima, significantly enhancing performance in massive IoT connectivity scenarios. Simulated Annealing (SA) and Parallel Tempering (PT) algorithmic techniques near-optimal detection performance in some extreme MIMO regimes when implemented in CPUs and GPUs [18]. In addition to PICs and ONNs performing MVM, microring resonators (MRRs) enable wavelength-division multiplexing (WDM) which allows for ultra-fast parallel optical convolution operations, enhancing RANs [4].

### *6.3. Hardware Efficiency*

Utilizing GPU acceleration enables virtualized Layer (L1) and Layer (L2) processing, significantly increasing computational efficiency and reducing latency. GPU-accelerated RAN can reduce AI model training time by approximately 32% without sacrificing accuracy, which directly contributes to faster network adaptability. The speed gains result from the faster data transfer and quicker memory operations on the GPU compared to CPUs. Integrating optimization techniques like DVFS into the Hardware-aware Neural Architecture Search (HW-aware NAS) process on edge GPUs, such as the NVIDIA Jetson AGX Xavier, has demonstrated the potential for significant energy gains, reaching up to about 53% [16].

MPSoCs utilizing the "Map-and-Conquer" framework has achieved up to 2.1× greater energy efficiency compared to only using GPUs. Similarly, for Graph Neural Networks (GNNs), mapping-aware frameworks have shown up to 3.38× energy efficiency gains by leveraging the specific strengths of MPSoC components [17]. The definition of hardware efficiency itself is evolving towards "Green" communication paradigms, showing that 6G is targetting energy efficiencies around 1 pJ/bit [19].

Quantum and Photon-based processors offer a fundamental shift in approaching thermal and energy limitations of silicon-based computing. While current quantum annealers require a significant amount of power for cooling, reverse quantum computation strategies enable operations to perform with a theoretical zero energy dissipation [18]. Hybrid hardware architectures integrating digital electronics with analog photonics has indicated throughputs 2.8 to 14 times faster than traditional GPUs while also reducing energy consumption by about 25% [4].

### *6.4. Wireless System Metrics*

Traditional CPU and GPUs are evaluated against strict KPIs, including peak data rates up to 1 Tbps, air-interface latency as low as 0.1 ms, and reliability levels as high as 99.99999% [7]. In AI environments, these KPIs extend to model training efficiency and prediction accuracy. GPU-accelerated testbeds have shown a 32% reduction in Mean Squared Error (MSE) for traffic prediction tasks compared to baseline models [16]. In addition, the ability to support massive connectivity up to $10^7$ devices/km$^2$ and high mobility (up to 1000 km/h) imposes rigorous throughput and adaptability constraints on the underlying processing hardware [20].

MPSoCs have yielded approximately 1.57× in latency speedup and 3.38× energy efficiency gains through efficient workload mapping strategies, relative to single-unit deployments [17].

Quantum and Photonic based applications have introduced unique metrics tailored to their unique physics-based approaches. For quantum annealers solving MIMO detection problems, benchmarks such as "Time-to-Solution" and "Time-to-BER" have been essential in testing speeds for wireless communications [18]. Photonic computing is evaluated on "computational density" (e.g., TOPS/mm$^2$) and extreme energy efficiency (e.g., Joules per Multiply-Accumulate operation) which are vital for overcoming the thermal bottlenecks of electronic chips. A

Mach-Zehnder interferometer (MZI), which is a type of PICs, achieved a response rate of 10 Gb/s through nonlinear activation functions [4].

## 7. Conclusions

This article provides a comprehensive review of the core integration of Artificial Intelligence (AI) and specialized hardware accelerators in enabling the next-generation 6G Wireless Networks. We focused on Neuromorphic chips, NPUs, TPUs, and other advanced processors that address the massive computational loads and significant energy efficiency and low-latency constraints inherent in 6G architectures. It is seen that, Neuromorphic processors, which mimic the human brain's structure, utilize Spike Neural Networks (SNNs) and Compute-in-Memory (CIM) to achieve significantly higher speed while substantially reducing power consumption for edge AI applications. For NPUs and TPUs, specialized dataflow architectures like the systolic array and optimization methods such as Remote Direct Memory Access (RDMA) greatly improve system throughput (up to 15×) and reduce service delay by over 28% and jitter by up to 95%. Advanced processors, including MPSoCs, QPUs, and PICs, will deliver enhanced energy efficiency, with MPSoCs achieving up to 2.1× greater energy efficiency than GPU-only deployments, and PICs demonstrating throughputs 2.8 to 14 times faster than traditional GPUs. Finally, hardware specialization is crucial for implementing foundational 6G requirements like the Compute Continuum Layer (CCL) and supporting Ultra-Reliable Low-Latency Communication (URLLC). With the adoption of these specialized accelerators, future deployments of 6G will produce exceptional performance with aggressive energy efficiencies that redefine intelligent network capability.

As 6G hardware specialization continues to advance, standardized evaluation becomes more and more essential to benchmark diverse sets of chips. The MLPerf Power framework, developed by MLCommons, is an industry standard for measuring the energy efficiency of ML systems. By providing a unified benchmark for measuring power consumption across various scales, MLPerf Power allows for objective approaches for most optimal solutions with minimal energy usage.

## Author Contributions

J.L.: conceptualization, investigation, writing—original draft preparation; Y.-D.Y.: supervision, methodology, writing—review and editing. Both authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

Not applicable.

## Conflicts of Interest

The authors declare no conflict of interest.

## Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Gemini 3.0 Pro to perform grammar and spelling checks. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References

1. Gramaglia, M.; Larrabeiti López, D.; Picazo Martínez, P.; et al. *Towards 6G Architecture: Key Concepts, Challenges, and Building Blocks*; Zenodo: Geneva, Switzerland, 2025.
2. Chen, J.; Park, S.; Popovski, P.; et al. Neuromorphic Semantic Communications with Wake-Up Radios. In Proceedings of

the 2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Lucca, Italy, 10–13 September 2024; pp. 91–95.
3. Chen, J.; Park, S.; Popovski, P.; et al. Neuromorphic split computing with wake-up radios: Architecture and design via digital twinning. *IEEE Trans. Signal Process.* **2024**, *72*, 4635–4650.
4. Song, X.; Fu, Z.; Wang, J.; et al. Optical Neuromorphic Technology Catalyzes the Next-Generation Mobile Communication Technology. *Adv. Intell. Syst.* **2025**, *7*, 2500344.
5. Corradi, F.; Zjajo, A.; Poehls, L.M.B.; et al. Neuromorphic Edge Computing: Challenges, Opportunities, and Current Solutions. In Proceedings of the 2025 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED), Reykjavík, Iceland, 6–8 August 2025.
6. Baba, H.; Hirai, S.; Hayashi, K.; et al. 6G In-Network Computing Architecture for Service Acceleration: Prototype and Evaluation. In Proceedings of the 2025 IEEE Wireless Communications and Networking Conference (WCNC), Milan, Italy, 24–27 March 2025.
7. Chafii, M.; Bariah, L.; Muhaidat, S.; et al. Twelve scientific challenges for 6G: Rethinking the foundations of communications theory. *IEEE Commun. Surv. Tutorials* **2023**, *25*, 868–904.
8. Lambrechts, J.W.; Sinha, S.; Sengupta, K.; et al. Intelligent integrated circuits and systems for 5G/6G telecommunications. *IEEE Access* **2024**, *12*, 21402–21419.
9. Mbakaan, C.; Kpelai, J.; Tyavnande, M. AI-Driven Terahertz 6G Wireless Communication Systems: From Hardware Design to Network Intelligence. *Front. Results Appl. Sci.* **2025**, *1*, 1–18.
10. Ivković, J.; Ivković, J.L. Exploring the potential of new AI-enabled MCU/SOC systems with integrated NPU/GPU accelerators for disconnected Edge computing applications: Towards cognitive SNN Neuromorphic computing. In Proceedings of the EdTech International Scientific Conference, Belgrade, Serbia, 26–27 May 2023; pp. 12–22.
11. Katsaros, G.N.; Ducoing, J.D.L.; Nikitopoulos, K. Towards Neuromorphic processing for next-generation MU-MIMO detection. In Proceedings of the 2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Lucca, Italy, 10–13 September 2024; pp. 581–585.
12. Sophie, E.R. Real-Time Processing of Neural Networks in Autonomous Vehicles. 2025. Available online: https://www.researchgate.net/publication/391150676_Real-Time_Processing_of_Neural_Networks_in_Autonomous_Vehicles (accessed on 29 January 2026).
13. Kumar, P.V.; Kulkarni, A.; Mendhe, D.; et al. AI-Optimized hardware design for Internet of Things (IoT) devices. In Proceedings of the 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 9–10 April 2024; pp. 21–26.
14. Huang, J.; Lin, W.; Wu, W.; et al. On Efficiency, Fairness and Security in AI Accelerator Resource Sharing: A Survey. *ACM Comput. Surv.* **2025**, *57*, 1–35.
15. Hoque, M.M.; Ahmad, I.; Suomalainen, J.; et al. On Resource Consumption of Machine Learning in Communications Security. *TechRxiv* **2024**. https://doi.org/10.36227/techrxiv.173152709.90384402/v.
16. Basaran, O.T.; Zafar, H.; Kasparick, M.; et al. Next-Gen AI-on-RAN: AI-Native, Interoperable, and GPU-Accelerated Testbed Towards 6G Open-RAN. In Proceedings of the ICC 2025–IEEE International Conference on Communications, Montreal, QC, Canada, 8–12 June 2025; pp. 5362–5367.
17. Bouzidi, H. Efficient deployment of deep neural networks on hardware devices for edge ai. Ph.D. Dissertation, Université Polytechnique Hauts-de-France, Valenciennes, France, 2024.
18. Kim, M. Quantum and Quantum-Inspired Computation for MIMO Communications in Wireless Networks. Ph.D. Thesis, Princeton University, Princeton, NJ, USA, 2023.
19. Strinati, E.C.; Belot, D.; Falempin, A.; et al. Toward 6G: From new hardware design to wireless semantic and goal-oriented communication paradigms. In Proceedings of the ESSCIRC 2021–IEEE 47th European Solid State Circuits Conference (ESSCIRC), Grenoble, France, 13–22 September 2021; pp. 275–282.
20. Yang, H.; Alphones, A.; Xiong, Z.; et al. Artificial-intelligence-enabled intelligent 6G networks. *IEEE Netw.* **2020**, *34*, 272–280.