International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **03** | **2026**

Role of Hexagon Processor in On-Device AI

Nishtha Kambli[1], Chandana Maral[2], Diksha Mestri[3], Adnya Acharekar[4]

[1,2,3,4]*Student, Bachelor of Computer Application, DCITM, Maharashtra, India*

DOI: 10.5281/zenodo.20629641

ABSTRACT

Establishing the foundational role of specialized hardware in mobile ecosystems, this abstract examines the Qualcomm Hexagon Digital Signal Processor (DSP) as a critical pillar for on-device AI acceleration. As mobile deep learning demands have evolved from simple tasks to complex large language model (LLM) inference, the Hexagon processor has transformed from a traditional signal handler into a sophisticated Neural Processing Unit (NPU).

The Hexagon architecture, particularly starting with the V6 68x series, introduced specialized extensions like Hexagon Vector Extensions (HVX) and dedicated Tensor Accelerators to optimize AI inference. These advancements allow for high-throughput, low-latency execution of quantized neural networks, significantly reducing the computational burden on mobile CPUs and improving power efficiency. Research highlights that Hexagon-based NPUs now deliver performance comparable to desktop-class GPUs in specific AI benchmarks, making real-time, privacy-preserving local processing feasible. Furthermore, recent systems like HeteroInfer demonstrate the potential of integrating the Hexagon NPU with mobile GPUs through unified memory architectures to further accelerate heterogeneous AI workloads.

This study underscores that the Hexagon processor's evolution is essential for overcoming the resource constraints of portable devices, enabling the next generation of context-aware, self-learning mobile applications. By balancing high-speed tensor operations with extreme energy optimization, the Hexagon DSP remains a primary enabler for sustainable on-device AI.

Comparative benchmarks reveal that latest-generation Hexagon processors achieve performance levels approaching desktop-class GPUs in specific AI workloads, facilitating real-time, privacy-preserving local processing. This study concludes that the Hexagon processor's evolution is vital for overcoming the inherent resource limitations of mobile ecosystems, serving as a primary enabler for sustainable, context-aware, and self-learning edge devices. By optimizing tensor operations and reducing CPU overhead, the Hexagon DSP remains central to the future of on-device intelligence.

1. INTRODUCTION

The rapid evolution of mobile System-on-Chips (SoCs) has transitioned smartphones from simple communication tools into powerful hubs for artificial intelligence (AI). Initially, mobile devices relied on general-purpose CPUs for tasks like voice recognition and predictive text, but the escalating complexity of deep learning models necessitated specialized hardware acceleration. Central to this shift is the Qualcomm Hexagon Digital Signal Processor (DSP), which has evolved from a dedicated audio and voice handler into a sophisticated engine for on-device AI.

The Hexagon architecture's significance lies in its ability to handle the high-throughput, parallel computational demands of neural networks while maintaining extreme energy efficiency. By utilizing features such as Hexagon Vector Extensions (HVX) and dedicated Neural Processing Units (NPUs), the processor excels at executing quantized deep learning workloads. This specialization allows for real-time AI inference—such as image enhancement and natural language processing—directly on the device, reducing reliance on cloud computing and enhancing user privacy.

Furthermore, as mobile AI enters the era of Large Language Models (LLMs), the Hexagon processor plays a vital role in heterogeneous computing environments, working alongside GPUs to distribute heavy workloads efficiently. This research explores how the Hexagon DSP overcomes the inherent resource constraints of mobile electronics, serving as a primary enabler for the next generation of self-learning, context-aware smart devices.

Building on the previous sections, here are additional points regarding the role of the Hexagon processor in on-device AI, focusing on architectural advantages, performance benchmarks, and its emerging role in Large Language Model (LLM) acceleration:

1.1 Evolution of Functionality:
While Digital Signal Processors (DSPs) were originally designed for voice coding and radio signal processing, they have not been displaced by CPUs or GPUs in mobile devices because they offer superior performance at significantly lower power consumption.

1.2 Hardware Acceleration for Quantized Models:
The Hexagon processor is specifically optimized for hardware acceleration of quantized neural networks, which allows it to run models like MobileNet in under 25ms—considerably faster than a mobile CPU, which typically takes 60-65ms for the same task.

1.3Performance Scaling:
The computational power of mobile AI accelerators, including the Hexagon DSP, has been evolving rapidly, nearly doubling with each new generation of System-on-Chip (SoC).

1.4Approaching Desktop-Class Performance:
Modern mobile Neural Processing Units (NPUs), of which the Hexagon is a primary example, are reaching performance levels comparable to CUDA-compatible Nvidia graphics cards from recent years, enabling the execution of complex, deep AI models directly on portable devices.

1.5Heterogeneous LLM Inference:
Recent advancements in frameworks like HeteroInfer and HeteroLLM leverage the Hexagon NPU alongside mobile GPUs to distribute LLM inference workloads.

1.6 Efficient Tensor Partitioning: Heterogeneous inference engines can now solve optimal tensor partitioning strategies for LLMs by analyzing model structures to determine where workloads—such as attention projections and feed-forward network computations—can be split between the NPU and GPU to maximize parallel execution.

1.7 Resource Constraint Mitigation: By offloading data-intensive tasks from the CPU to the Hexagon processor, mobile devices can handle advanced computer vision and natural language processing tasks without the thermal or battery drain associated with general-purpose processors.

2. ARCHITECTURE EVOLUTION FROM DSP TO NPU

The evolution of the Qualcomm Hexagon architecture represents a strategic transition from a specialized signal processor to a high-performance Neural Processing Unit (NPU) tailored for edge intelligence. Originally designed in the 1990s for audio and radio signal handling, the Hexagon Digital Signal Processor (DSP) maintained a presence in mobile SoCs due to its superior power efficiency compared to general-purpose CPUs.
The architectural shift began with the introduction of Hexagon Vector Extensions (HVX), which enabled the processor to handle massive 1024-bit vector operations, making it suitable for image processing and early deep learning tasks. As AI workloads intensified, the architecture integrated dedicated hardware for tensor acceleration, effectively transforming the DSP into a modern NPU. This evolution allowed for the high-throughput execution of quantized (INT8) neural networks, which are critical for balancing computational accuracy with mobile thermal constraints.
By the fourth generation of mobile accelerators, this evolved architecture began approaching the performance levels of CUDA-compatible desktop GPUs. Modern frameworks now leverage this NPU to handle complex sub-graphs of Large Language Models (LLMs), using heterogeneous partitioning to distribute data-intensive projections to the Hexagon unit while reserving the GPU for memory-bound decoding phases. This transformation ensures that the Hexagon processor remains the primary enabler for sustainable, low-latency, on-device AI.

3. POWER EFFICIENCY AND ALWAYS ON AI PERFORMANCE

The Qualcomm Hexagon processor serves as a primary driver for power-efficient, "always-on" AI performance in mobile ecosystems. Its fundamental advantage lies in a specialized architecture that achieves higher throughput with a fraction of the power required by mobile CPUs or GPUs. Unlike general-purpose processors that consume significant energy handling diverse tasks, the Hexagon Digital Signal Processor (DSP) is optimized for quantized, repetitive mathematical operations typical of deep learning.
Data from the provided literature highlights that the Hexagon processor can execute quantized neural networks like MobileNet in under 25ms, compared to over 60ms on a CPU. This speed-to-power ratio is critical for maintaining "always-on" functionality—such as voice trigger detection, environmental sensing, and background scene

recognition—without depleting the battery. By offloading these persistent tasks from the primary application processor to the low-power Hexagon unit, the system prevents thermal throttling and minimizes power leakage.
Furthermore, the introduction of Hexagon Vector Extensions (HVX) and dedicated Tensor Accelerators allows for massive parallelization of 8-bit integer (INT8) workloads. This specialization ensures that context-aware features remain active even when the device is in a low-power state. This research concludes that the Hexagon's ability to balance high-speed tensor execution with extreme energy optimization is the cornerstone of sustainable on-device AI, enabling devices to be more responsive and intelligent while maximizing battery longevity.

4. HETEROGENEOUS COMPUTING INTEGRATION

Heterogeneous computing integration is a sophisticated strategy that distributes AI workloads across a device's diverse processing cores—primarily the Hexagon NPU (Neural Processing Unit), GPU, and CPU—to maximize throughput and energy efficiency. Rather than relying on a single processor, modern frameworks like HeteroLLM and HeteroInfer treat the mobile SoC as a unified computational fabric.
The core of this integration lies in task-specific partitioning. The Hexagon processor is specifically engineered for high-throughput, quantized tensor operations (INT8), making it the ideal candidate for the "prefill phase" of Large Language Models (LLMs) and matrix-heavy attention projections. Conversely, mobile GPUs, which offer higher memory bandwidth, are better suited for the "decoding phase," which is often memory-bound rather than compute-bound. Strategic integration involves:

4.1Operator Partitioning: Splitting deep learning sub-graphs so that each layer runs on the hardware best suited for its specific tensor shape.

4.2Unified Memory Access: Leveraging shared memory architectures to minimize data-copying overhead between the Hexagon NPU and the GPU.

4.3Dynamic Load Balancing: Utilizing offline profiling to characterize the "order-sensitive" performance of the NPU, ensuring that execution sequences are optimized to prevent thermal throttling.
This synergy allows on-device AI to transcend the limitations of individual cores, enabling complex generative models to run within the strict power and thermal envelopes of portable electronics.

5. HARDWARE ACCELERATION FOR NEURAL NETWORK

Hardware acceleration for neural networks refers to the use of specialized silicon logic to perform the massive parallel computations required by deep learning more efficiently than a general-purpose CPU. In the context of the Qualcomm Hexagon processor, this acceleration has transitioned from simple signal processing to a sophisticated Neural Processing Unit (NPU) architecture designed for high-throughput tensor operations.
The Hexagon's primary strength in hardware acceleration lies in its ability to handle quantized inference. While CPUs are designed for complex branching and floating-point accuracy, the Hexagon is optimized for 8-bit integer (INT8) math. This specialization allows it to execute models like MobileNet in under 25ms, whereas a mobile CPU might take over 60ms. By utilizing Hexagon Vector Extensions (HVX) and dedicated Tensor Accelerators, the processor can perform thousands of multiply-accumulate (MAC) operations in a single clock cycle, which is the fundamental mathematical requirement for neural network layers.
Furthermore, hardware acceleration on the Hexagon is tightly integrated with the Android Neural Networks API (NNAPI). This allows the system to offload specific sub-graphs of an AI model—such as the compute-heavy attention projections in Large Language Models (LLMs)—directly to the Hexagon NPU. This offloading not only reduces latency but also prevents thermal throttling of the main CPU, ensuring that on-device AI remains responsive and sustainable within the power-constrained environment of a mobile device.

6. SOFTWARE ABSTRACTION AND DEVELOPER ACCESS

Software abstraction and developer access are critical for leveraging the Qualcomm Hexagon processor, transitioning it from a closed signal handler to an accessible AI engine. At the core of this abstraction is the Android Neural Networks API (NNAPI), which serves as an intermediary layer between high-level machine learning frameworks and device-specific hardware.
Developers typically access the Hexagon processor through TFLite delegates or specialized vendor SDKs, such as the Qualcomm AI Stack. These tools abstract the underlying complexity of Hexagon's Very Long Instruction Word (VLIW) architecture, allowing developers to offload computational sub-graphs—like convolution and matrix multiplication—directly to the Neural Processing Unit (NPU) or DSP. This abstraction ensures that models

developed in standard frameworks like TensorFlow or PyTorch can be automatically optimized for Hexagon's specialized INT8 and FP16 capabilities without requiring low-level assembly programming.

Recent advancements in heterogeneous frameworks, such as HeteroLLM, further simplify developer access by using offline profilers to characterize hardware performance. These systems determine the optimal partitioning of tensors across the Hexagon NPU and mobile GPU, automating the deployment of complex Large Language Models (LLMs). This layered software ecosystem is essential for overcoming deployment complexity, ensuring that sophisticated AI applications can run efficiently across various hardware generations while maintaining consistent performance.

7. METHODOLOGY

The methodology for evaluating the role of the Hexagon processor in on-device AI involves a multi-layered approach centered on benchmarking, architectural profiling, and heterogeneous system analysis.

7.1 Benchmarking Frameworks: Research utilizes standardized tools like AI Benchmark to measure real-time performance across diverse tasks, including image classification, super-resolution, and natural language processing. These benchmarks assess execution time and memory consumption under various hardware configurations.

7.2 Performance Characterization: The methodology distinguishes between floating-point and quantized inference. Specialized attention is given to the Hexagon's efficiency in running quantized neural networks (e.g., MobileNet) via Android NNAPI drivers or TFLite delegates, comparing these speeds directly against mobile CPU results.

7.3 Architectural Profiling: Modern approaches, such as those used in the HeteroInfer and HeteroLLM frameworks, involve an offline profiling stage where the computational power of the NPU (Hexagon) is characterized for specific tensor shapes and sequence lengths.

7.4 Heterogeneous Partitioning: To optimize Large Language Model (LLM) inference, a solver analyzes the model structure to identify partitionable operators. This allows for the development of parallelization strategies that distribute workloads—such as attention projections—between the Hexagon NPU and the mobile GPU.

7.5 Cross-Platform Comparison: The methodology includes a comparative analysis of Hexagon-equipped Snapdragon chipsets against competitors from HiSilicon, Samsung, and MediaTek to establish industry performance baselines. Building on the previous sections, here are additional points regarding the architectural impact, performance benchmarks, and heterogeneous system integration of the Hexagon processor in on-device AI:

7.6 Historical Integration and Evolution: DSPs have been integral to mobile phones since the early 1990s, initially handling voice coding and radio signal processing. Their evolution into AI-centric engines like the Hexagon series stems from their inherent power efficiency compared to general-purpose CPUs.

7.7 Substantial Speed Advantages: Performance data shows that the Hexagon DSP (such as the version 685) can execute quantized neural networks like MobileNet in under 25ms, which is significantly faster than the **60–65ms** required by a mobile CPU.

7.8 Rapid Generation-over-Generation Growth: The computational power of mobile AI accelerators, including the Hexagon series, has historically evolved rapidly, nearly doubling with each new generation of System-on-Chip (SoC).

7.9 Approaching Desktop Performance: By the 4th generation of mobile NPUs, these mobile-side accelerators began approaching the performance levels of CUDA-compatible Nvidia graphics cards from previous years, enabling complex model execution on portable hardware.

7.10 Heterogeneous Synergy with GPUs: Modern frameworks like HeteroInfer and HeteroLLM leverage the Hexagon NPU alongside mobile GPUs to optimize Large Language Model (LLM) inference. This allows for parallel execution where tasks are split between the NPU and GPU based on which processor is more efficient for a specific operator.

7.11 Quantized Precision Focus: The Hexagon processor is highly optimized for INT8 (8-bit integer) operations, which provide a high performance-to-power ratio. While information on 16-bit floating-point (FP16) power is often undisclosed by vendors, researchers often estimate it to be roughly half of the INT8 capability.

7.12 Efficient Profiling and Deployment: Advanced inference engines use offline profiling—often completed in under 20 minutes—to characterize the Hexagon NPU's performance across various tensor shapes. This data allows a solver to determine the optimal partitioning strategy for real-time AI tasks.

8. CONCLUSION

In summary, the Qualcomm Hexagon processor has established itself as an indispensable component for on-device AI, moving beyond its origins as a simple signal handler to become a sophisticated neural engine. Research across

various generations of Snapdragon SoCs demonstrates that the Hexagon DSP, particularly when utilizing specialized extensions like Hexagon Vector Extensions (HVX), consistently outperforms standard mobile CPUs in deep learning tasks. For instance, benchmarks show that Hexagon-based acceleration can execute quantized models like MobileNet significantly faster than general-purpose processors, often reducing inference times from over 60ms to under 25ms.

Furthermore, the emergence of heterogeneous computing frameworks such as HeteroInfer and HeteroLLM highlights a new phase in the processor's role. By partitioning complex AI sub-graphs between the Hexagon NPU and mobile GPU, these systems maximize throughput while adhering to the strict thermal and power limits of mobile hardware. This synergy is particularly vital for the next generation of Large Language Models (LLMs), where the Hexagon's high-efficiency tensor acceleration handles data-intensive projections that would otherwise overwhelm the device.

Ultimately, the Hexagon processor serves as a primary enabler for sustainable, privacy-preserving edge intelligence. Its ability to provide desktop-class AI performance within a mobile power envelope ensures that sophisticated, real-time AI applications remain feasible on portable electronics. As mobile AI continues to evolve toward more complex generative models, the continued advancement of specialized architectures like the Hexagon will be essential for maintaining the balance between performance and battery longevity.

In addition to the summary provided, here are further critical points regarding the Role of the Hexagon Processor in On-Device AI based on the technical analysis of the provided research:

8.1 Sustainability and Thermal Management: Unlike mobile CPUs and GPUs that often suffer from thermal throttling during prolonged AI workloads, the Hexagon processor is designed for high sustained performance. By offloading continuous tasks—such as real-time video bokeh or live translation—to the DSP, the system maintains a stable temperature, preventing the "performance drops" common in general-purpose processors.

8.2 Privacy-First AI Architecture: The Hexagon processor enables a "zero-cloud" approach. Because it can process complex models locally with minimal latency, sensitive data (such as biometric patterns, voice recordings, or personal documents) never needs to leave the device, fulfilling the core security requirements of modern Edge Intelligence.

8.3 Vectorization via HVX: A key technical differentiator is Hexagon Vector Extensions (HVX). This allows the processor to handle massive 1024-bit vector operations in a single clock cycle. This architectural feature is what enables high-resolution image processing and "super-resolution" tasks to run in real-time without draining the battery.

8.4 Unified Memory Access: In heterogeneous frameworks like HeteroLLM, the Hexagon NPU utilizes a shared memory architecture with the GPU and CPU. This minimizes "data copying" overhead, which is often the primary bottleneck in mobile AI. The ability to directly access shared tensors allows the Hexagon to jump into the computational pipeline seamlessly.

8.5 Support for Always-On Contextual Awareness: Due to its ultra-low power "sensing hub" capabilities, the Hexagon processor is the primary driver for always-on AI features. It can remain active to detect voice triggers or environmental changes while the rest of the SoC remains in a sleep state, providing a balance of "smart" functionality and extreme energy conservation.

8.6 Developer Ecosystem Maturity: The integration of Hexagon through the Qualcomm AI Stack and Android NNAPI ensures that developers do not need to write assembly code to access its power. The existence of "delegates" in TFLite allows for automated optimization, making the Hexagon's advanced architecture accessible for mainstream app development.

9. REFERENCES

[1]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)

[2]. Alzantot, M., Wang, Y., Ren, Z., Srivastava, M.B.: Rstensorflow: Gpu enabled tensorflow for deep learning on commodity android devices. In: Proceedings of the 1st International Workshop on Deep Learning for Mobile Systems and Applications. pp. 7–12. ACM (2017)

[3]. ArmNN: https://github.com/arm-software/armnn. Retrieved on: 30.09.2018

[4]. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

[5] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once for all: Train one network and specialize it for efficient deployment. In ICLR, 2020. 1, 2

[6] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. CoRR, abs/1812.00332, 2018. 3
[7] Yu-Hsin Chen, Joel S. Emer, and Vivienne Sze. Eyeriss v2: A flexible and high-performance accelerator for emerging deep neural networks. CoRR, abs/1807.07928, 2018.