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

Edge AI Hardware (Processor) Architecture Study of NVIDIA Jetson Modules

Diya Joshi[1], Mayuri Manchekar[2], Saniya Patil[3], Shruti Dalavi[4]

[1.2.3.4]*Student, Bachelor of Computer Applications, DCITM, Maharashtra, India*

DOI: 10.5281/zenodo.20628098

ABSTRACT

The proliferation of the Internet of Things (IoT) has intensified the demand for specialized edge hardware capable of supporting high-performance artificial intelligence (AI) workloads with low latency and minimal power consumption. This research provides an architectural study of the NVIDIA Jetson module family, specifically examining its heterogeneous system-on-chip (SoC) design as a high-performance edge accelerator. Unlike traditional discrete GPUs, NVIDIA's Jetson architecture integrates an ARM-based CPU and an NVIDIA GPU that share a unified memory pool, which can reach up to 32 GB.
This study evaluates the effectiveness of these modules across diverse machine learning pipelines, including training, inference, and data pre-processing. Performance characterization using benchmarks like TPCx-AI reveals that while Jetson devices offer significant power efficiency and reduced data movement compared to desktop hardware, available memory often serves as the primary bottleneck for scaling complex workloads. Furthermore, the research investigates specialized architectural differences, such as the Compute-to-Cache (C2C) ratio, noting that edge-optimized GPUs require distinct L2 cache configurations compared to traditional variants to maintain efficiency. By analysing the trade-offs between computational throughput, memory constraints, and power usage (typically 5W–10W for modules like the Nano), this work establishes a technical foundation for deploying robust, real-time AI applications in resource-constrained edge environments.

Keyword: ***NVIDIA, Artificial Intelligence, Internet of Things,Jetson AGX Xavier***

1.INTRODUCTION

The rapid growth of the Internet of Things (IoT) and the subsequent demand for real-time data processing have driven the adoption of edge computing. To meet the high computational requirements of modern artificial intelligence (AI) while adhering to strict power constraints, specialized hardware accelerators have become essential. Among these, the NVIDIA Jetson series has emerged as a leading platform, offering a unique heterogeneous system-on-chip (SoC) architecture that bridges the gap between traditional embedded systems and high-performance server hardware.

NVIDIA Jetson modules are characterized by an integrated design where an ARM-based CPU and an NVIDIA GPU share a unified memory pool, which can reach up to 32 GB in high-end modules. This shared memory architecture is a critical feature, as it facilitates efficient data movement between processing units, though memory availability often remains a primary bottleneck for scaling complex machine learning pipelines. These devices are engineered for extreme power efficiency, typically operating within a 5W to 10W thermal design power (TDP) range while supporting parallelized workloads across training, inference, and data pre-processing. By leveraging specialized GPU cores for tensor operations, Jetson modules enable the deployment of sophisticated deep learning models closer to the data source. This study investigates the architectural nuances and performance trade-offs of the Jetson family, establishing a roadmap for optimizing resource-constrained AI deployments at the network edge.

1.1 Unified Memory Architecture (Uma)

A defining characteristic of the Jetson architecture is the shared physical memory between the ARM CPU and the NVIDIA GPU. Unlike desktop systems where data must be transferred over a high-latency PCIe bus, Jetson modules utilize a zero-copy mechanism. This significantly reduces data movement overhead, though it creates a shared resource bottleneck where high-resolution sensors and complex AI models must compete for the same memory bandwidth.

1.2 Heterogeneous Computing Engines

Beyond the CPU and GPU, modern Jetson modules (like the Xavier and Orin series) feature specialized hardware blocks:

1.2.1. Deep Learning Accelerator (DLA): Fixed-function engines designed specifically for inferencing common neural network layers with extreme energy efficiency.

1.2.2. Programmable Vision Accelerator (PVA): A specialized processor for computer vision algorithms (e.g., optical flow, stereo matching) that offloads tasks from the GPU to save power.

1.3 Precision-Based Performance Scaling

The Jetson architecture supports various levels of mathematical precision, including FP32, FP16, and INT8. Utilizing the **Tensor Cores** for mixed-precision arithmetic allows the hardware to perform significantly more operations per clock cycle. However, research shows that while GPU utilization might appear high, Tensor Core utilization often remains low (under 30%) due to scheduling inefficiencies on the CPU side.

1.4 Dynamic Power Management

Jetson modules utilize "MAX-N" and various "Clock-Gating" modes. These architectural presents allow the hardware to dynamically adjust the number of active CPU cores and GPU clock frequencies. This research point is crucial because it demonstrates the trade-off between peak computational throughput and the thermal constraints of a fan less edge enclosure.

1.5 Multi-Tenant Workload Challenges

While Jetson hardware is powerful, it often lacks robust hardware-level isolation for concurrent applications. When multiple AI models run simultaneously (e.g., face detection and voice recognition), they often suffer from "resource contention" at the L2 cache and DRAM levels. This necessitates advanced software schedulers like HaX-CoNN or CP-CNN to manage hardware partitioning effectively.

2. BACKGROUND ON EDGE AI HARDWARE

The shift toward edge AI is driven by the need for localized processing to achieve real-time response, data privacy, and reduced network bandwidth usage. Unlike cloud-based systems, edge AI hardware must operate within stringent resource constraints, primarily concerning power consumption, thermal dissipation, and physical size. Historically, edge devices relied on general-purpose CPUs, which often struggled with the massive parallelization required for deep learning workloads

NVIDIA Jetson modules represent a paradigm shift in this domain by integrating high-performance graphics processing units (GPUs) into a low-power System-on-Chip (SoC) architecture. These modules feature a heterogeneous design where an ARM-based CPU and an NVIDIA GPU—ranging from Maxwell to Orin architectures—share a unified memory pool. This shared memory structure is critical as it eliminates the high-latency data transfers typically found in discrete GPU systems. Furthermore, contemporary Jetson modules incorporate dedicated hardware accelerators, such as Deep Learning Accelerators (DLA) and Tensor Cores, specifically designed to accelerate matrix multiplications and neural network inference. By providing a scalable software stack through the JetPack SDK, these modules enable developers to deploy complex AI models at the edge while maintaining energy efficiency, often operating between 5W and 60W. This architectural maturity has made Jetson a cornerstone for research into autonomous robotics, industrial inspection, and smart city infrastructure.

3.NVIDIA JETSON MODULE ARCHITECTURE

The architectural design of NVIDIA Jetson modules represents a sophisticated heterogeneous System-on-Chip (SoC) tailored specifically for edge AI deployment. At its core, the architecture integrates a high-performance ARM-based CPU with an NVIDIA GPU—ranging from the Maxwell to the latest Orin generations—on a single die. A defining characteristic is the **Unified Memory Architecture (UMA)**, where the CPU and GPU share a common pool of physical RAM, which can scale up to 32 GB in high-end modules like the AGX Xavier. This shared memory eliminates the high-latency overhead of data transfers across a PCIe bus, though it can create bandwidth contention during complex workloads.

Beyond general-purpose cores, the architecture features specialized hardware accelerators designed to offload specific AI tasks. These include:

3.1. Tensor Cores: Dedicated units within the GPU that accelerate large-scale matrix-multiply-accumulate operations essential for deep learning.

3.2. Deep Learning Accelerator (DLA): A fixed-function engine optimized for highly efficient neural network inference, freeing the main GPU for other tasks.

3.3. Vision and Image Processing Engines: Specialized blocks such as the Programmable Vision Accelerator (PVA) and Video Image Compositor (VIC) handle pre-processing and computer vision algorithms at low power. The GPU architecture itself is optimized with a specific **Compute-to-Cache (C2C) ratio** and L2 cache configurations that differ from traditional desktop GPUs to maintain high efficiency within a 5W–60W thermal envelope.

4.PERFORMANCE AND POWER METRICS

4.1Performance Metrics

4.1.1. Computational Throughput: This is measured by GFLOPS (Giga-Floating Point Operations Per Second) or TOPS (Tera-Operations Per Second), representing the theoretical peak performance.

4.1.2. Streaming Latency: Critical for real-time edge applications, this measures the time taken to process a single frame or data packet from input to output.

4.1.3. Inference Speed: Often quantified in Frames Per Second (FPS), this indicates the module's capability to execute deep learning models like YOLO or ResNet-50.

4.1.4. Resource Utilization: Metrics include percentage usage of the CPU, GPU, and specialized Tensor Cores. Research indicates that while GPU utilization may be high, Tensor Core usage often ranges only between 15% and 30%.

4.1.5. Memory Bandwidth: Since Jetson uses a Unified Memory Architecture (UMA), bandwidth and latency of the shared RAM are primary constraints for large-scale AI workloads.

4.2Power Metrics

4.2.1. Thermal Design Power (TDP): Jetson modules typically operate within configurable power envelopes, ranging from 5W (Jetson Nano) to 60W (AGX Orin).

4.2.2. Energy Efficiency: This is expressed as Performance-per-Watt (e.g., FPS/W or GFLOPS/W), serving as a benchmark for battery-operated or fanless deployment.

4.2.3.Power Mode Scalability: Metrics track how hardware performance scales across different presets (e.g., 5W, 10W, or Max-N modes) to balance thermal output and speed.

5.SOFTWARE STACK AND ECOSYSTEM

The software ecosystem for NVIDIA Jetson modules is centred around a unified and comprehensive stack that enables seamless deployment of AI models from the cloud to the edge. At the core of this ecosystem is the **NVIDIA JetPack SDK**, which provides a complete development environment including the Linux for Tegra (L4T) operating system and essential acceleration libraries. This stack is designed to leverage the integrated GPU and specialized hardware engines such as Deep Learning Accelerators (DLA).

Key components of the software stack include:

5.1. CUDA and cuDNN: These fundamental libraries enable high-performance parallel computing and primitive operations for deep neural networks.

5.2. TensorRT: A critical deep learning inference optimizer and runtime that delivers low latency and high throughput by optimizing models specifically for the Jetson hardware architecture.

5.3. Framework Support: The ecosystem provides robust support for popular machine learning frameworks such as TensorFlow, PyTorch, and Caffe, allowing researchers to use standard tools for training and deployment.

5.4. Containerization and Orchestration: Modern edge deployments increasingly utilize Docker and Kubernetes to manage microservices, facilitating consistent software environments across diverse edge devices.

5.5. Specialized SDKs: NVIDIA provides application-specific tools like DeepStream for multi-sensor video analytics and Isaac for robotics, which further extend the capabilities of the hardware for industry-specific use cases.

This integrated software approach ensures that the architectural advantages of the Jetson modules, such as unified memory and heterogeneous cores, are fully utilized for efficient AI execution.

6.COMPARISON JETSON WAS OTHER EDGES PLATFORMS

In the landscape of edge AI hardware, NVIDIA Jetson modules occupy a unique position as high-performance accelerators characterized by their integrated GPU-based architecture. Unlike many specialized edge platforms, the Jetson family—including the Nano, TX2, and AGX Xavier—utilizes a heterogeneous System-on-Chip (SoC) design where an ARM CPU and a powerful GPU share a unified memory pool. This architectural choice provides a distinct advantage in versatility, allowing it to support a wider range of machine learning frameworks and complex data pre-processing tasks compared to fixed-function accelerators.

In contrast, platforms like the Google Coral Edge TPU and Intel Movidius Myriad VPUs are designed primarily for high-efficiency inference on specific model types, such as quantized 8-bit integers. While these ASIC-based competitors often achieve superior power efficiency in narrow tasks, they lack the computational flexibility of Jetson's CUDA cores, which can handle diverse parallel workloads, including localized model training. Furthermore, compared to Field-Programmable Gate Arrays (FPGAs) like the Xilinx Kria series, which offer extreme customizability and deterministic low latency, Jetson modules provide a more accessible software ecosystem through the JetPack SDK. However, research highlights that memory remains a primary bottleneck

for Jetson; while it can support up to 32GB of RAM, data movement and memory availability often limit the scaling of large-scale machine learning pipelines relative to desktop-class hardware. Ultimately, Jetson modules excel as balanced platforms for complex, multi-functional edge applications where software compatibility and raw throughput are prioritized.

7.DEPLOYEMENT CHALLENGES

Deploying AI applications on NVIDIA Jetson architectures involves navigating several hardware-specific hurdles. Based on the architectural studies provided, the following are the primary deployment challenges:

7.1. Memory Bottlenecks: The shared memory architecture (Unified Memory) is often the most significant constraint. High-end modules like the AGX Xavier are limited by available RAM (up to 32GB), which restricts the scaling of complex machine learning pipelines and end-to-end data processing.

7.2. CPU-Side Inefficiencies: Even when GPU resources are available, performance is often hindered by CPU-side overhead. Inefficiencies in thread scheduling and low-level resource management can lead to significant underutilization of specialized components like Tensor Cores.

7.3. Resource Contention in Multi-Tenancy: Jetson hardware frequently lacks robust isolation for concurrent vision workloads. Running multiple models simultaneously leads to contention for the L2 cache and DRAM, causing performance degradation and unpredictable latency.

7.4. Thermal and Power Trade-offs: While designed for low power (5W–60W), the hardware requires careful tuning of "MAX-N" and clock-gating modes. Developers must balance peak computational throughput against strict thermal limits, especially in fanless or enclosed edge environments.

7.5. Architectural Simulation Gaps: A major challenge for researchers is the lack of accurate cycle-level simulators. Existing tools primarily target discrete GPUs, leading to simulation cycle errors of up to 29% when applied to the integrated GPU designs found in Jetson modules.

7.6. Data Movement vs. Pre-processing: While the architecture reduces data, movement compared to cloud setups, the overhead of data pre-processing at the edge remains high unless specifically parallelized using the available compute cores.

8. FUTURE TRENDS IN EDGE AI HARDWARE ARCHITECTURE

8.1. Integration of Specialized Simulation Tools: There is a growing need for cycle-accurate simulators, such as Accel-Sim-J, specifically designed for integrated GPUs in Jetson modules to facilitate the architectural exploration of next-generation edge devices.

8.2. Expansion of On-Chip Memory and Bandwidth: As machine learning pipelines grow in complexity, increasing available RAM and memory bandwidth is critical, as memory is currently the primary limitation for scaling end-to-end AI workloads on edge devices.

8.3. Heterogeneous Resource Orchestration: Future architectures will move toward more sophisticated System-on-Chip (SoC) designs that better balance workloads across diverse compute units, including ARM CPUs, GPUs, and specialized accelerators like Deep Learning Accelerators (DLA).

8.4. Energy-Aware Hardware Design: Optimization will focus on "Performance-per-Watt," with a trend toward hardware that can handle not only model inference but also data pre-processing and model training locally to reduce the energy cost of data movement to the cloud.

8.5. Hardware-Software Co-Design for Miniaturization: There is a continued push toward "tinier" AI, where hardware is architected specifically to support compressed models and tinyML frameworks, enabling intelligence in even more resource-constrained embedded sensing schemes.

8.6. Support for Full-Pipeline Parallelism: Future modules are expected to offer enhanced parallelism that extends beyond AI inference to benefit the entire data lifecycle, from acquisition to pre-processing, achieving near order-of-magnitude improvements in processing time.

9. CONCLUSION

This research highlights that NVIDIA Jetson modules represent a significant advancement in edge AI hardware by successfully integrating high-performance GPU capabilities into a low-power, heterogeneous System-on-Chip (SoC) architecture. A primary conclusion is that while these modules offer substantial energy efficiency and reduced data movement compared to traditional desktop hardware, available memory serves as a critical constraint for scaling complex end-to-end machine learning pipelines.

The architectural study reveals that Jetson's unified memory—where the CPU and GPU share the same RAM—enables sophisticated computations close to the data source. However, to fully exploit this potential, specialized software optimizations like TensorRT and parallelization are necessary to improve throughput across data pre-processing, training, and inference phases. Furthermore, accurately modeling these integrated GPUs requires refined simulation tools, as traditional simulators designed for discrete GPUs often fail to capture the unique performance characteristics of edge-based integrated architectures.

In summary, the NVIDIA Jetson family provides a robust foundation for mobile and embedded AI applications, effectively balancing computational power with portability. While memory capacity and resource management remain challenges for larger models, the platform's ability to deliver high performance-per-watt makes it a cornerstone for future developments in autonomous systems and real-time edge analytics. Continuous improvements in hardware-aware simulation and workload scheduling will be essential to further maximize the efficiency of these integrated edge architectures.

10. REFERENCE

[1]. Ul Mehmood, M., Ulasyar, A., Ali, W., Zeb, K., Zad, H.S., Uddin, W., Kim, H.J. (2023). A new cloud-based IoT solution for soiling ratio measurement of PV systems using artificial neural network. Energies, 16(2): 996. https://doi.org/10.3390/en16020996
[2]. Santosh Pashikanti, Edge Computing and AI: Extending Cloud Capabilities with NVIDIA and Kubernetes, North American Journal of Engineering and Research Est. 2020
[3]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, and A. Konwinski, “A View of Cloud Computing,” Communications of the ACM, vol. 53, no. 4, pp. 50–58, Apr. 2010.
[4]. T. Guo, “Cloud-Edge Collaborative Inference Systems for Real-Time Applications,” IEEE Cloud Computing, vol. 7, no. 2, pp. 61–71, Mar./Apr. 2020.
[5]. NVIDIA Edge Computing Solutions. Available: https://www.nvidia.com/en-us/edge-computing/
[6]. P. SK, S. A. Kesanapalli, and Y. Simmhan, “Characterizing the per formance of accelerated jetson edge devices for training deep learning models,” Proceedings of the ACM on Measurement and Analysis of Computing Systems, vol. 6, no. 3, pp. 1–26, 2022.
[7] X. Hou, C. Xu, C. Li, J. Liu, X. Tang, K.-t. Cheng, and M. Guo, “Improving efficiency in multi-modal autonomous embedded systems through adaptive gating,” in IEEE Transactions on Computers, 2024.
[8] X. Hou, X. Tang, J. Liu, C. Li, L. Liang, and K.-t. Cheng, “Wasp: Efficient power management enabling workload-aware, self-powered aiot devices,” in IEEE Transactions on Parallel and Distributed Systems, 2024.
[9] X. Hou, P. Tang, C. Li, J. Liu, C. Xu, K.-T. Cheng, and M. Guo, “Smg: A system-level modality gating facility for fast and energy-efficient multimodal computing,” in IEEE Real-Time Systems Symposium, 2023.
[10] X. Hou, J. Liu, X. Tang, C. Li, J. Chen, L. Liang, K.-T. Cheng, and M. Guo, “Architecting efficient multi-modal aiot systems,” in Proceedings of the 50th Annual International Symposium on Computer Architecture, 2023.