

Benchmarking Ultra-Low-Power μ NPUs

Josh Millar, Yushan Huang,
Sarab Sethi, Hamed Haddadi
Imperial College London

Anil Madhavapeddy
University of Cambridge

Abstract

Efficient on-device neural network (NN) inference has various advantages over cloud-based processing, including predictable latency, enhanced privacy, greater reliability, and reduced operating costs for vendors. This has sparked the recent rapid development of microcontroller-scale NN accelerators, often referred to as neural processing units (μ NPUs), designed specifically for ultra-low-power applications.

In this paper, we present the first comparative evaluation of a number of commercially-available μ NPUs, as well as the first independent benchmarks for several of these platforms. We develop and open-source a model compilation framework to enable consistent benchmarking of quantized models across diverse μ NPU hardware. Our benchmark targets end-to-end performance and includes model inference latency, power consumption, and memory overhead, alongside other factors. The resulting analysis uncovers both expected performance trends as well as surprising disparities between hardware specifications and actual performance, including μ NPUs exhibiting unexpected scaling behaviors with increasing model complexity. Our framework provides a foundation for further evaluation of μ NPU platforms alongside valuable insights for both hardware designers and software developers in this rapidly evolving space.

1 INTRODUCTION

Performing neural network (NN) inference on constrained devices has applications across numerous domains, including wearable health monitoring [1], smart agriculture [2], real-time audio processing [3], and predictive maintenance [4]. On-device inference offers various advantages over cloud-based alternatives: improved latency for time-critical applications, enhanced privacy, as well as reduced operating costs for vendors, by eliminating the need to transmit sensitive data, and improved reliability by removing dependence on network connectivity. Given their unique form factor and low power consumption, microcontrollers (MCUs) are widely used in resource-constrained environments. However, their performance is often constrained by limitations in memory capacity, throughput, and compute.

The computational demands of modern neural networks (NNs) have catalyzed the development of specialized hardware accelerators across the computing spectrum, from high-performance data centers to ultra-low-power and embedded

devices. At the resource-constrained end of the spectrum, microcontroller-scale neural processing units (μ NPUs) have recently emerged, designed to operate within extremely tight power envelopes — in the milliwatt or sub-milliwatt range — while still providing low latency for real-time inference. These devices represent a new class of accelerator, combining the power efficiency of MCUs with the cognitive capabilities previously exclusive to more powerful computing platforms. The core advantage of μ NPUs stems from their ability to exploit the inherent parallelism of neural networks with dedicated multiply-accumulate (MAC) arrays alongside specialized memory structures for weight storage. Such architectural specialization enables μ NPUs to achieve orders of magnitude improvement in latency compared to general-purpose MCUs executing equivalent workloads.

Despite the growing availability of μ NPU platforms, the field lacks a standardized evaluation or comprehensive benchmark suite. Existing benchmarks focus solely on Analog Devices’ MAX78000 [5–7], lacking any side-by-side comparison with other platforms. Hardware vendors provide performance metrics based on proprietary evaluation frameworks, often using disparate NN models, quantization strategies, and other varying optimizations. This heterogeneity across evaluation methods, and absence of independent verification of vendor-provided performance claims, creates uncertainty for hardware designers and embedded software developers in selecting the most suitable μ NPU platform for their application’s constraints. The lack of standardized benchmarking also hampers research by obscuring the relationship between architectural design and real-world performance. Given the rapid pace of development and increasing diversity of available μ NPU platforms, establishing reliable comparative benchmarks has become an urgent need for the field. To this end, we make the following contributions:

- **Side-by-Side Benchmark of μ NPU Platforms:** We conduct the first comparative evaluation of commercially-available μ NPU platforms, enabling direct performance comparisons across diverse hardware architectures under consistent workloads and measurement conditions.
- **Independent Benchmarks:** We also provide the first fine-grained and independent performance benchmarks for several μ NPU platforms that have not previously been subject to third-party evaluation, offering unbiased verification of vendor performance claims.

- **Open-Source Model Compilation Framework:** We develop and release¹ an open-source framework that enables consistent and simplified transplanting of NN models across diverse μ NPU platforms, reducing the engineering overhead associated with cross-platform evaluation.
- **Developer Recommendations:** Informed by our benchmark results, we provide actionable recommendations to developers regarding platform selection, key focus areas for model optimization, and trade-offs for various application scenarios and constraints.

In developing a unified compilation and benchmarking framework, we standardize model representations across the various μ NPU platforms, enabling direct comparison of latency, memory, and energy performance. Our evaluation also includes fine-grained analysis of the various model execution stages, from NPU initialization and memory input/output overheads to CPU pre/post-processing – aspects that can significantly impact end-to-end performance but are often not addressed in technical evaluations. The resulting analysis uncovers both expected performance trends as well as surprising disparities between hardware specifications and actual performance, including μ NPUs exhibiting unexpected scaling behaviors with increasing model complexity. We hope our findings provide valuable insights to both developers and hardware architects.

2 BACKGROUND & MOTIVATION

2.1 Resource-Constrained Neural Computing

The shift from cloud-based to on-device neural computing has numerous advantages for real-time data processing, especially with increasing concerns regarding data privacy and security [8]. Unlike cloud-based solutions, local inference mitigates security risks by processing sensitive data locally, which is particularly advantageous in domains such as medical diagnostics and surveillance [9, 10]. Additionally, local processing reduces end-to-end latency alongside operating costs for model vendors. However, traditional NN accelerators, such as GPUs and TPUs, are ill-suited to resource-constrained environments given their power consumption and large form factors [11, 12].

MCUs are compact, low-power computing platforms, often reliant on a single CPU and shared memory bus [13]. While MCUs are commonly adopted for resource-constrained IoT applications [14–16], they generally lack the computational resources for efficient NN inference. Specifically, the computational capability of typical MCUs is often limited to a few million MAC operations per second, far below the tens of billions MACs/s required for real-time NN inference. Their

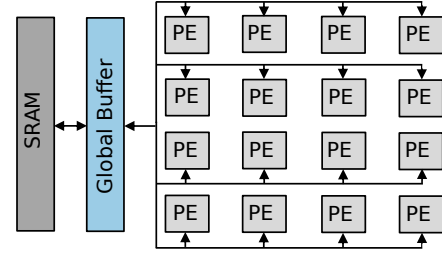


Figure 1: typical μ NPU hardware architecture

absence of dedicated hardware acceleration results in large latency overheads and elevated power consumption during NN processing. Limited SRAM and flash memory also often poses challenges for efficiently managing the large weight matrices required for NN models.

Given the various shortcomings of traditional MCUs, microcontroller-scale μ NPUs are emerging as a response. These specialized NN accelerators offer dedicated neural processing hardware, providing higher throughput for NN workloads, meeting the stringent requirements of real-time NN inference [17–19] while maintaining low-power operation. Collectively, μ NPUs position themselves as a key solution for efficient, real-time NN processing in resource-constrained environments.

2.2 μ NPU Hardware Design

μ NPU hardware design is optimized for efficient tensor operations via specialized MAC units and parallelizable memory hierarchies [20, 21]. Fig. 1 illustrates the architecture of a typical μ NPU, composed of a systolic array of processing elements (PEs). Notably, each PE contains its own MAC units and, importantly, its own weight memory space to avoid memory contention and maximize parallelization. The array of PEs is linked by an inter-PE communication grid, which connects to a large global buffer and SRAM/DRAM via an on-chip network [22]. Efficient memory hierarchy optimization is achieved by partitioning available RAM, along with implementing high-bandwidth memory interfaces and data prefetching mechanisms, addressing the memory bottlenecks faced by traditional MCUs when handling large NN model weights. μ NPUs mainly vary by their number of PEs, PE layout and clustering, memory hierarchy layout, and the availability/amount of storage/MAC units in each PE.

These architectural advantages, coupled with low-power optimization techniques such as power gating, enable μ NPU platforms to deliver low-power, high-throughput performance for real-time NN inference.

2.3 Benchmarking μ NPU Platforms

Adoption of μ NPU Platforms: The increasing demand for on-device neural computing has accelerated the development and commercialization of μ NPUs. This is evidenced by the

¹Upon publication

growing number of vendors, including Arm [23], who have released μ NPU platforms to market.

Need for Comprehensive Benchmarking: Existing work on μ NPU platforms mainly focuses on practical applications and/or model optimizations [24–26], lacking fine-grained performance analysis from a systems perspective. In evaluating memory usage, latency, power, and throughput, across μ NPU platforms, we aim to uncover critical performance bottlenecks, guiding researchers towards more efficient software and NN model design.

Limitations of Existing Benchmarks: Existing benchmarks of μ NPUs focus on a single platform, lacking horizontal comparisons across the now wide variety of available platforms [6, 7, 27]. This narrow perspective limits understanding of the variations in performance and task-based applicability across different μ NPUs. Existing standalone benchmarks also have significant shortcomings. Chiefly, most focus solely on the model’s inference forward pass, overlooking other adjacent operations within the end-to-end model inference or application pipeline(s), such as NPU initialization, memory input/output (I/O), and CPU pre/post-processing. While often neglected, these factors can significantly impact overall performance and efficiency.

3 INFRASTRUCTURE & METHODOLOGY

We begin by detailing our benchmark hardware and models, then provide a comprehensive overview of our benchmarking framework and model inference pipeline.

3.1 Hardware

To provide a comprehensive benchmark, we evaluate a diverse range of widely-used, commercially-available μ NPU platforms, from ultra-low-power μ NPUs to high-performance NPU-equipped system-on-chip (SoC) architectures. These are evaluated alongside MCUs without dedicated neural hardware for comparison. Our selection covers a wide range of computational capabilities (<5 to >500 GOPs), memory configurations (128 KB to 2 MB RAM), and bit-width support (1-bit quantized to 32-bit floating-point operations). Fig. 2 provides a visualization of peak GOPs (Giga Operations Per Second) vs. peak power for the various μ NPU platforms included in our benchmark (on a log scale). Table 3.1 details our set of benchmark μ NPUs, and we provide more detail on each platform below.

The **MAX78000** (or **MAX78K**) [5] from US-based Analog Devices features a Cortex-M4F with a RISC-V coprocessor, each capable of acting as the primary processor, along with a proprietary 30-GOPS CNN accelerator. The latter has a dedicated 512 KB SRAM for input data, 442 KB for weights, and 2 KB for biases, and supports quantized operations at 1, 2, 4, and 8-bit precision. The same fine-grained bit-width

quantization is not yet widely supported on other μ NPU platforms, or indeed in common software libraries designed for ML on resource-constrained devices; TFLite/LiteRT [28], for example, only supports 8-bit integer and 16-bit float weight quantization. The MAX78000 also has 512 KB of flash and 128 KB of CPU-only SRAM. This platform is among the best-documented commercially-available μ NPUs; previous work has benchmarked its CNN accelerator under various configurations [6, 7, 27], alongside exploring optimal model and data loading strategies for its 2D memory layout [29].

The **GAP8** [30], part of GreenWaves Technologies’ GreenWaves Application Processor series, features an 8-core RISC-V cluster and 22.65-GOPS hardware convolution engine for neural network acceleration at 8 or 16-bit precision. The platform has 512 KB of L2 RAM, up to 8 MB of L3 SRAM, and 20MB flash storage, enabling it to store and run larger, more complex models or mixture-of-experts (MoE) architectures. The GAP series of μ NPUs have also been the subject of several recent works, again mainly centered on model optimization [26, 31, 32]; no platform benchmark exists yet.

The **Himax HX6538 WE2** (or HX-WE2) [33] is a more powerful μ NPU platform from Taiwan-based semiconductor manufacturer, Himax Technologies. This platform features a Corstone-300 set up, with Cortex M55 CPU and Ethos U55 NPU, delivering up to 512 GOPs. The platform also features 512KB TCM, 2MB SRAM, and 16MB flash, suitable for large or more complex models, but at an increased power draw.

NXP’s **MCXN947** [34] is part of the MCX N94x line of MCUs, featuring dual Cortex-M33 CPUs and NXP’s eiQ Neutron NPU. The MCXN947 is designed for lower-power applications, with 8-bit neural acceleration of only 4.8 GOPs. The platform features 512 KB RAM and 2 MB flash storage.

Our benchmark also includes MCUs without neural hardware for comparison, to quantify any efficiencies gained from specialized NPU architectures.

The **STM32H7A3ZI** [35] is a high-performance MCU based on the Cortex M7, with 2 MB of flash and 1.4 MB of SRAM. Manufactured by Swiss-based ST Microelectronics, it is frequently used with on-board NNs [16, 36].

The **ESP32s3** MCU [37] features dual-core Tensilica LX6 processors, 512 KB of SRAM, 2MB PSRAM, and 8MB flash. Notably, whilst primarily a low-power MCU, it advertises NN acceleration capabilities with “support for vector instructions ... providing acceleration for neural network computing”. This is achieved via an extended instruction set, which includes 128-bit vector operations, *e.g.*, complex multiplication, addition, subtraction, shifting, and comparison.

We also include the **MILK-V Duo** [38], a RISC-V SoC built around the CVITEK CV1800B processor. Unlike the previous MCUs/ μ NPUs, it runs a Linux OS, supporting more flexible NN workloads at a much-increased power budget.

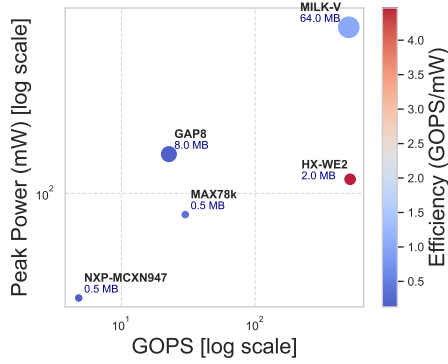


Figure 2: a visualization of the various μ NPUs used in our benchmark, and how they compare in terms of GOPS, peak power draw observed, and *theoretical* efficiency (GOPS/mW).

This platform represents the upper bound of our evaluation in terms of computational power and software flexibility.

3.1.1 Note on CPU Frequency We configure the various μ NPU platforms to operate at a uniform CPU frequency. While this permits direct comparison of architectural efficiency, it should be noted that many of the platforms are capable of operating at higher frequencies than evaluated – approaching the GHz range in some cases. Our method intentionally isolates architectural efficiency, but further experimentation could explore the impact of varying CPU frequencies on end-to-end latency and power consumption.

Other hardware parameters are largely standardized by default. Further work could investigate their impact on performance; the number of active PEs, memory layout configuration, and other hardware-specific optimizations are variably configurable across the μ NPU platforms and can influence overall efficiency.

3.2 Models

Table 2 details the various CNN-based models used in our benchmark, covering image classification, object recognition, and signal reconstruction applications. We provide more detail on each model below.

CIFAR10-NAS: the optimal CNN model generated by neural architecture search (NAS) for the CIFAR-10 dataset, combining diverse convolutional units, pooling layers, and unique connectivity patterns. The model was generated using the Once-for-All (OFA) NAS framework, a weight-sharing-based framework that decouples search and training by constructing a *supernet* model from which various hardware-specific *subnet* models can be derived [39]. This is our largest model, with 74.3 Million MACs (MMACs) and 36.4 Million FLOPs (MFLOPs). Trained on the CIFAR-10 dataset, with 3x32x32 input size and 10-class output.

ResidualNet: a CNN framework built around residual functions, helping to mitigate gradient vanishing. ResidualNet has 37.7 MMACs and 18.5 MFLOPs. Trained on the CIFAR-100 dataset, with 3x32x32 input size and 100-class output.

SimpleNet: a simpler CNN framework composed of a basic stack of convolutional and pooling layers, without complex branches or residual functions. Despite its basic architecture, SimpleNet often outperforms more complex models including certain versions of ResidualNet [40]. SimpleNet has 38.0 MMACs and 18.5 MFLOPs. Trained on the CIFAR-100 dataset, with 3x32x32 input size and 100-class output.

Autoencoder: a symmetric encoder-decoder model. This model is our simplest, with just 0.5 MMACs and 0.2 MFLOPs. Trained on a machine fault detection dataset, generated using the SpectraQuest Machinery Fault Simulator [41]. The input/output size is 3x256.

YoloV1: a single-stage object detection CNN that uses multi-scale output feature maps to predict bounding boxes and class probabilities. YoloV1 has 43.83 MMACs and 21.2 MFLOPs. Trained for person-only detection on the COCO dataset [42], with input size 3x96x96 and 3 output layers, which result from pruning its final layers for cross-platform uniformity (more details below). The non-max suppression (NMS) post-processing step is performed on CPU.

3.2.1 Ensuring Model Uniformity We encountered substantial variability in operator support across the benchmark platforms. The NXP-MCXN947’s eIQ Neutron NPU lacks native support for softmax operations, for example, necessitating its implementation as a CPU post-processing step for relevant models. Similarly, operations associated with non-maximum suppression (NMS) in the YoloV1 model were inconsistently supported across platforms, requiring us to also move the entire NMS operation to CPU post-processing. This explains the unusual multi-component output shape of our YoloV1 model (see Table 2). The benchmark platforms also outline varying levels of support for operator compatibility. The MAX78000, for example, only supports 1D convolution with kernel sizes 1 to 9 and 2D convolution with kernel sizes of 1 by 1 or 3 by 3. Unsupported operations will fall back to CPU execution and incur latency penalties.

By identifying and constructing models using a core subset of operators that are universally supported across all μ NPUs, we aim to ensure that any measured performance differences stem from fundamental architectural discrepancies rather than variations in model compilation and optimization.

3.2.2 Quantization We quantize all benchmark models to INT8 precision, as it is supported by all evaluated NPUs. However, it’s important to note that while this enables a more direct architectural comparison, it may not reflect the optimal accuracy-performance tradeoff on each platform; some NPUs, such as the MAX78000, support lower bit-widths

Table 1: the various μ NPU platforms used in our benchmark.

MCU	CPU(s)	NPU	Flash	RAM	GOPs	Bit Cap.
MAX78000	Cortex-M4 RISC-V	MAXIM-own	512KB	512 KB NPU 128 KB CPU	30	1, 2, 4, 8
HX-WE2 (Corstone-300)	Cortex-M55	Ethos-U55	16 MB	2 MB SRAM 512 KB TCM	512	8,16,32
NXP-MCXN947	Cortex-M33 x2	eIQ Neutron	2 MB	512 KB	4.8	8
GAP8	RISC-V	GAP-own	20 MB L3	512 KB L2 8MB L3	22.65	8,16
STM32H74A3ZI	Cortex-M7	-	2 MB	1.4 MB	-	8, 16, 32
ESP32	Tensilica LX6	-	4 MB	520 KB	-	8, 16, 32
MILK-V	RISC-V XuanTie C906 x2	CV1800B	-	64 MB	500	8, 16, 32

**Table 2: the various models used in our benchmark.
Note: MMACs/MFLOPs are forward-only.**

Model	Input Shape	Output Shape	MMACs	MFLOPs
CIFAR10-NAS	3x32x32	1x10	74.2512	36.3776
ResidualNet	3x32x32	1x100	37.7812	18.4612
SimpleNet	3x32x32	1x100	38.0006	18.4612
Autoencoder	3x256	3x256	0.5455	0.2020
YoloV1	3x96x96	1x12x12x12	43.8294	21.2244
		1x12x12x2		
		1x12x12x10		

(*e.g.*, 1, 2, 4-bit), and others, like the HX-WE2 support floating-point acceleration (*e.g.*, FLOAT16 and 32-bit).

We perform post-training quantization (PTQ) on all models/platforms. While platforms like the MAX78000 support quantization-aware training (QAT) and fused operators, such optimizations produce platform-specific models incompatible with other NPUs. PTQ enables us to maintain structural consistency across all platforms. Moreover, since our primary metrics of interest are latency and power consumption, rather than inference accuracy, PTQ provides a sufficiently representative model for performance evaluation. PTQ was performed using a representative calibration dataset appropriate to each model’s domain. We did not apply per-channel quantization for weights, instead using per-tensor quantization to ensure compatibility across all platforms.

3.2.3 Compilation The various μ NPUs support a wide range of model formats, from platform-optimized versions of common model formats (*e.g.*, TFLite) to platform-specific formats (*e.g.*, CVITEK’s CVIMODEL). To facilitate cross-platform deployment, we developed a custom model compilation workflow for converting our base models into optimized formats for each target NPU. Our workflow ingests Torch (or ONNX/TFLM) base models along with various compiler flags (*i.e.*, target NPU platform, model input dimensions, bit-precision requirements, representative PTQ calibration data,

etc.), producing platform-specific optimized models with accompanying inference code.

The compilation process varies significantly by platform. For example, models targeting the ARM Ethos-U55 (on the HX-WE2) are compiled using the ARM Vela compiler, which ingests TFLiteMicro (TFLM) models and produces binaries optimized for the Ethos-U architecture. Vela applies platform-specific optimizations, including memory reduction. We evaluate both the *Size* optimization strategy, HX-WE2 (S), which minimizes SRAM usage, and the *Performance* strategy, HX-WE2 (P), which prioritizes execution speed using available arena cache if specified.

For other platforms, we utilize their respective toolchains (*e.g.*, the MAX78k’s SDK or the NXP eIQ portal tools). In each case, we configured such tools to maintain model structure equivalence while applying platform-appropriate optimizations. Note that in our model compilation workflow, template inference code is often generated along with a compiled model. However, this doesn’t include model-specific pre/post-processing steps, which should be implemented manually by the developer, who can update the template code as needed.

Fig 3 below details our model compilation framework for converting a base (Torch/ONNX/TFLM) model into various platform-specific formats. Our framework will be released as open-source, and we hope its use can ease the process of cross-platform model compilation and benchmarking.

3.3 Evaluation Metrics

We measure latency, power, energy-efficiency – in terms of number of inference operations per mJ – and memory usage across each benchmark platform and model. The impacts of various platform-specific model optimizations or compilation workflows on model accuracy are out of scope for our study. Latency can be considered proportional to throughput, since batching and other amortization techniques are not practical on μ NPU platforms due to memory constraints.

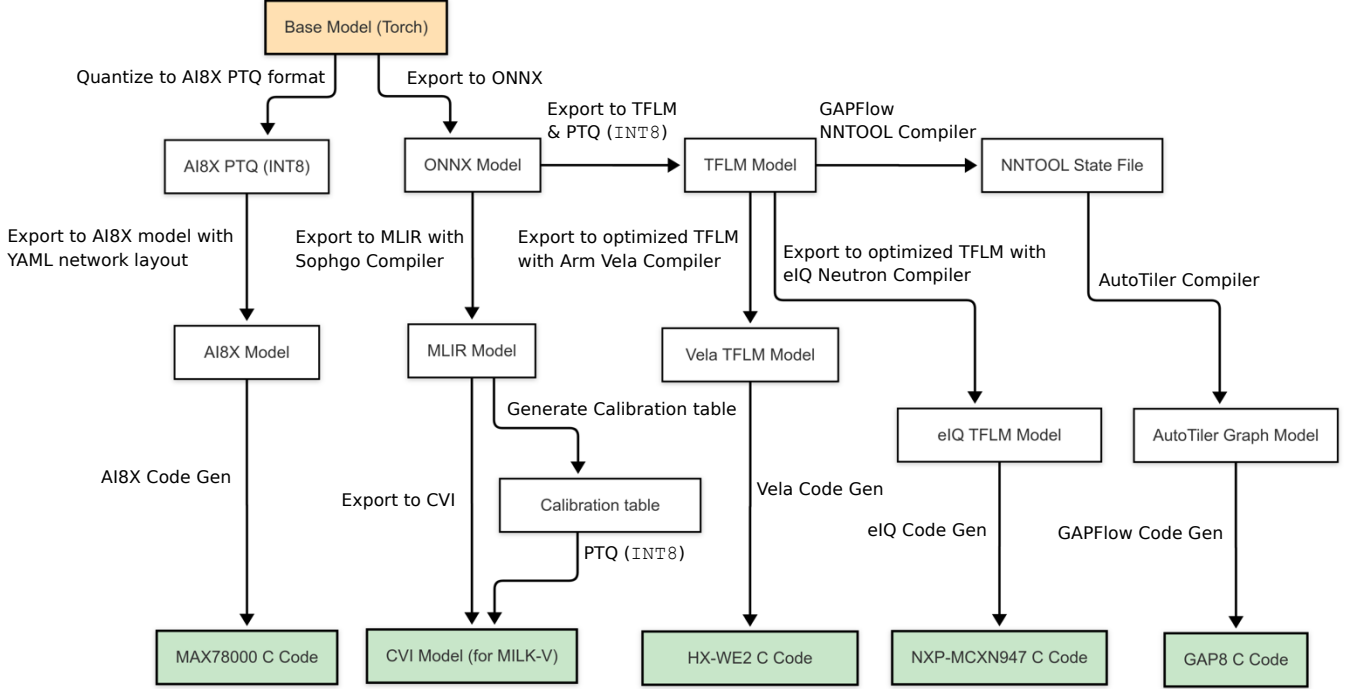


Figure 3: an overview of our model compilation workflow.

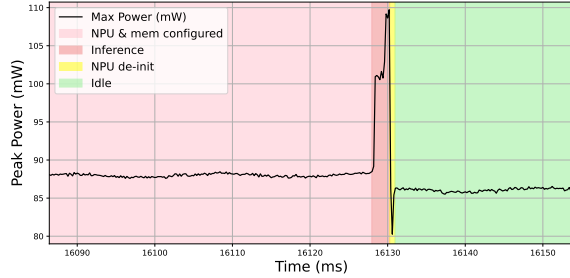


Figure 4: power trace of YoloV1 inference on HX-WE2.

Latency: Latency is measured using each platform’s internal timer. Notably, all MCUs, bar the MILK-V, are configured to run at 100 MHz. The MILK-V does not support manual frequency scaling, only DVFS. However, latency is inversely proportional to CPU frequency, as described by $T = N/f$ (T denotes latency, N the number of cycles required for a task, and f the operating frequency). Accordingly, we normalized MILK-V’s latency to be comparable to performance under uniform frequency conditions. Each model was run for 10 consecutive inferences. We report both the mean latency and standard deviation to account for any run-to-run variability.

Power and Energy: We compute power and energy using the Monsoon High Voltage Power Monitor [43] at a sampling rate of 50 Hz. The input voltage (U) is set to 3.3 V, capturing inference duration (t) and average power (P), from which the average energy consumption ($E = Pt$) is derived. To ensure stable measurements, we analyze only the data recorded

after a 1-minute operation period. Power measurements are gathered over 10 repeated inferences, and we report both mean and standard deviation. Fig. 4 details the power profile of YOLOv1 inference on the HX-WE2’s Ethos-U55 μ NPU.

Inferences per mJ: To quantify energy efficiency, we introduce ‘inferences per mJ’, I_{mJ} , capturing the number of end-to-end inferences (*i.e.*, memory transfer, CPU pre/post-processing, and optionally NPU initialization) performed for each millijoule of energy consumed.

Memory Usage: Memory usage is assessed by analyzing the linker (*.map*) file generated by the compilation toolchain. This file provides a detailed breakdown of memory allocation, including code (*.text*), initialized data (*.data*), and uninitialized data (*.bss*) segments. Flash memory usage is calculated as the sum of the code and initialized data segments (*.text* + *.data*), while RAM usage includes both the initialized and uninitialized data segments (*.data* + *.bss*). For the MAX78000, with its dedicated NPU-only memory, the RAM usage is computed separately for CPU and NPU.

3.4 Performance Breakdown

We break down each stage of model execution and measure per-stage latency and power consumption. This granular analysis helps identify specific bottlenecks in the inference pipeline, alongside measuring overall end-to-end

²MAX78k (C) denotes use of its Cortex-M CPU, and (R) its RISC-V CPU. HX-WE2 (S) denotes model compilation with the Vela *Size* optimization flag, and (P) with the Vela *Performance* flag.

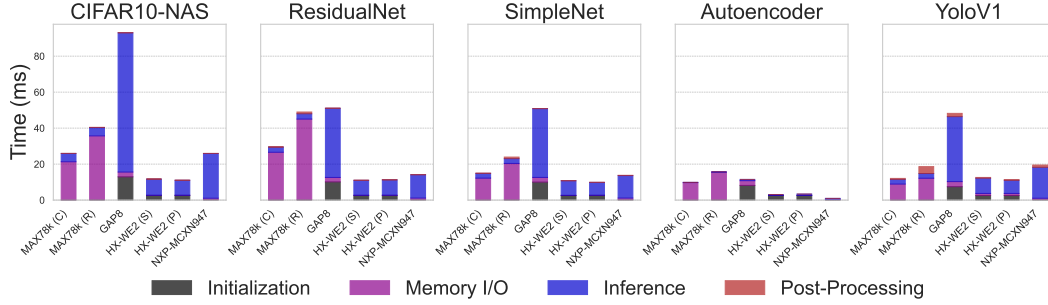


Figure 5: stacked latency for each stage, model, and platform.²

performance. We also measure idle power consumption (*i.e.*, each platform’s base power draw in the absence of active computation). We provide more detail on each stage below:

(NPU) Initialization: This covers any NPU setup overhead, including memory buffer allocation and kernel configuration.

Memory I/O: The cost/overhead of model and input data loading, including movement of input tensors and model weights from flash to NPU DRAM, and vice versa (*i.e.*, output tensors from NPU to CPU SRAM).

Inference: Executing the model’s forward pass on the NPU.

Post-Processing: Any additional operations required to be performed on the CPU. This includes computing softmax outputs for ResidualNet, SimpleNet, and CIFAR10-NAS models. YoloV1 post-processing includes NMS alongside output class softmax. The Autoencoder model does not require post-processing, since it produces direct reconstruction outputs.

Idle: The base power consumption of the various platforms, when not actively performing computation.

For MCUs without neural hardware (*i.e.*, the STM32H7A3ZI and ESP32), Initialization and Memory I/O are combined.

4 RESULTS & DISCUSSION

Table 6, which can found in the supplementary material, details our full latency and power measurements across each stage, model, and platform.

4.1 Power and Efficiency Breakdown

Our results reveal significant variation in efficiency across the benchmark platforms, as shown in Tables 3 and 4.

The MAX78000 (C) with Cortex-M4 CPU active demonstrates the best *overall* efficiency across evaluated models when NPU initialization overhead is considered, with consistent <30ms end-to-end latency. The MAX78000 (R) with RISC-V CPU lags slightly behind. This aligns with previous standalone benchmarks [6].

The NXP-MCXN947 also achieves consistent sub-30ms latency, with its fast initialization and memory I/O offsetting the impact of (moderately) slower inference latency, delivering comparable (and in some cases, improved) efficiency despite its lower-throughput accelerator.

Notably, the power-hungry but low-latency HX-WE2 platform, with Arm Corstone-300 (Cortex-M55 & Ethos-U55 NPU), consistently beats the MAX78000 (C/R) in terms of end-to-end latency across the various models, due to the latter’s large memory I/O overhead. The HX-WE2 (S/P) demonstrates average $\sim 1.93\times$ and $\sim 3.07\times$ speedup in end-to-end latency over the MAX78000 (C) and (R) respectively, but $\sim 3.13\times$ and $\sim 3.33\times$ increase in average power consumption. We find the Vela *Performance*-optimized models, for the HX-WE2, generally achieve slightly lower latency than the *Size*-optimized models. However, their efficiency gain diminishes with model complexity – efficiency on *Performance*-optimized YoloV1 is lower than on its *Size*-optimized variant.

The general-purpose MCUs without dedicated neural hardware – the STM32H7A3ZI and ESP32s3 – demonstrate significantly lower efficiency across all models. This result empirically validates the advantage neural hardware provides for performing on-device inference in constrained environments, with up to 2 orders of magnitude improvement in end-to-end latency in some cases. However, the STM32H7A3ZI’s power consumption during inference (54.91 - 56.11 mW) is comparable to or lower than MAX78000 (C/R) for some models. This is particularly evident for the Autoencoder model, where the STM32H7A3ZI achieves a surprisingly competitive $3.483 I_{mJ}$ – comparable to the best-performing platforms in our suite. This anomaly is likely attributable to the STM32H7A3ZI’s relatively efficient Cortex-M7 core when operating on the Autoencoder’s simple computational structure (0.5455 MMACs). In contrast, the ESP32 consistently exhibits high inference power consumption (129.74 - 157.17 mW) and latency (7.11 - 536.22 ms), despite its advertised support for CPU-accelerated tensor operations. Altogether, while general-purpose MCUs can achieve reasonable efficiency for simple models, they quickly become impractical for more complex NNs.

The MILK-V, our RISC-V SoC, also demonstrates low efficiency across all models, due to its NPU initialization overhead. We observe a different story, however, if initialization overhead is removed from consideration (*i.e.*, for continuous

Table 3: inferences per mJ (I_{mJ}) for evaluated models and platforms, including NPU initialization. The largest I_{mJ} for each model is underlined and bolded, while the second largest is in bold.

	MAX78k (C)	MAX78k (R)	GAP8	NXP-MCXN947	HX-WE2 (S)	HX-WE2 (P)	MILK-V	STM32H7A3ZI	ESP32s3
NAS	<u>1.10</u> ±0.002	0.85±0.001	0.10±0.002	<u>1.07</u> ±0.002	0.79±0.007	0.83±0.006	0.01±0.001	0.03±0.001	0.01±0.001
ResNet	<u>1.24</u> ±0.003	0.85±0.002	0.17±0.002	<u>1.97</u> ±0.003	0.85±0.006	0.84±0.019	0.01±0.001	0.06±0.001	0.02±0.001
SimpleNet	<u>2.29</u> ±0.006	1.65±0.003	0.16±0.005	<u>2.10</u> ±0.004	0.89±0.006	0.99±0.006	0.01±0.001	0.07±0.001	0.02±0.001
Autoenc	<u>3.92</u> ±0.014	2.75±0.008	1.12±0.028	<u>36.95</u> ±0.002	3.57±0.035	3.06±0.038	0.01±0.001	3.48±0.082	0.32±0.001
YoloV1_small	<u>2.27</u> ±0.004	1.76±0.003	0.20±0.005	<u>1.83</u> ±0.006	0.73±0.009	0.81±0.008	0.01±0.001	0.05±0.001	0.01±0.001

Table 4: inferences per mJ (I_{mJ}) for evaluated models and platforms, not including NPU initialization. The largest I_{mJ} for each model is underlined and bolded, while the second largest is in bold.

	MAX78k (C)	MAX78k (R)	GAP8	NXP-MCXN947	HX-WE2 (S)	HX-WE2 (P)	MILK-V
CIFAR10-NAS	<u>1.11</u> ±0.002	0.85±0.001	0.11±0.002	1.09±0.002	0.98±0.009	1.04±0.008	<u>2.75</u> ±0.281
ResNet	1.24±0.003	0.85±0.002	0.22±0.001	<u>2.01</u> ±0.002	1.08±0.008	1.05±0.024	<u>9.31</u> ±2.567
SimpleNet	<u>2.30</u> ±0.006	1.66±0.003	0.21±0.007	2.13±0.004	1.13±0.008	1.29±0.010	<u>4.13</u> ±0.459
Autoenc	3.94±0.014	2.78±0.008	6.25±0.203	<u>47.06</u> ±1.956	<u>22.45</u> ±0.392	12.97±0.232	13.74±1.953
YoloV1_small	<u>2.27</u> ±0.004	1.76±0.003	0.23±0.005	1.86±0.007	0.91±0.013	1.03±0.011	<u>5.75</u> ±0.770

Table 5: flash and RAM use (KB) for evaluated models and platforms. The model with highest flash/RAM for each platform is bolded. Note: MAX78k’s RAM is split into CPU-only and NPU-only.

	MAX78k (C)		MAX78k (R)		GAP8		NXP-MCXN947		HX-WE2 (S)		HX-WE2 (P)		STM32H7A3		ESP32s3	
	Flash	RAM	Flash	RAM	Flash	RAM	Flash	RAM	Flash	RAM	Flash	RAM	Flash	RAM	Flash	RAM
NAS	347.67	4.96+295.51	364.39	6.16+295.51	358.46	534.56	569.94	371.70	127.75	551.87	127.75	538.59	423.61	93.75	674.57	268.86
ResNet	425.38	4.98+372.84	446.92	6.91+372.84	258.32	372.49	471.52	381.89	127.75	618.11	127.75	694.33	456.07	70.97	694.44	268.78
SimpleNet	214.61	5.00+162.55	233.04	6.87+162.55	258.26	351.21	471.08	381.90	127.75	553.18	127.73	566.67	451.86	53.48	698.06	268.77
Autoenc	184.15	6.46+133.59	193.74	6.09+133.59	143.31	196.20	261.36	381.27	125.44	336.06	125.44	336.35	203.57	21.35	445.59	271.89
Yolo	130.43	6.93+41.75	147.96	8.38+41.75	43.29	159.46	287.70	410.83	152.32	263.81	152.32	319.10	119.28	167.52	355.19	268.77

operation). Without initialization, the MILK-V ranks highest for efficiency across almost all benchmark models. Notably, despite a large idle power draw, it achieves blazingly fast inference times (0.17 - 0.61 ms). Thus, for applications where power consumption isn’t a major constraint and workloads don’t require frequent NPU initialization/deinitialization, but low-latency and a compact form factor are key, the MILK-V could prove effective.

Fig. 7 details the power consumption breakdown across all evaluated platforms; among these, the MAX78000 (10.87–80.41 mW) and NXP-MCXN947 (22.91–36.69 mW) exhibit the lowest power draw across the benchmark models, with the NXP showing the lowest variance in peak power across the execution stages, enabling more reliable energy budgeting.

Beyond peak power, idle power consumption is another key consideration for low-power deployments, particularly if workloads run infrequently – idle power also varies significantly across our benchmark platforms. The MAX78000 demonstrates the lowest idle power of the various μ NPU platforms (10.87 mW with RISC-V and 13.21 mW under Cortex-M4). The HX-WE2 platform ranks highest (89.09 mW), raising concerns about its applicability in extremely power-constrained scenarios (such as ones in which long idle durations dominate overall energy usage).

4.2 Latency and Memory I/O Breakdown

4.2.1 NPU initialization NPU initialization times vary significantly across the benchmark platforms, from as low as 0.07 ms on the MAX78000 to 12.94 ms on the GAP8.

However, the actual initialization overhead, with respect to end-to-end latency, is almost negligible on most μ NPU platforms except the GAP8 (7.46 ms to 12.92 ms initialization latency across the various benchmark models). Such overhead could again be problematic for duty-cycled applications, where models must be frequently loaded/unloaded.

4.2.2 Memory I/O Table 5 details flash and RAM usage across our various benchmark platforms and models.

The significant memory I/O latency across all models on the MAX78000 forms an obvious inference bottleneck, with an average of 6.10x and 9.80x (Cortex-M4 and RISC-V) longer spent on memory I/O than actual inference (*e.g.*, 44.89 ms vs. 2.96 ms for ResidualNet with the MAX78k (R), meaning over 90% of end-to-end inference time is dedicated to memory operations rather than computation). This implies the MAX78000’s performance is largely memory-bound, and aligns with previous standalone benchmarks [6]. Notably, memory I/O operations are more efficient on the MAX78000’s Cortex-M4 CPU than its RISC-V one. In contrast, memory I/O

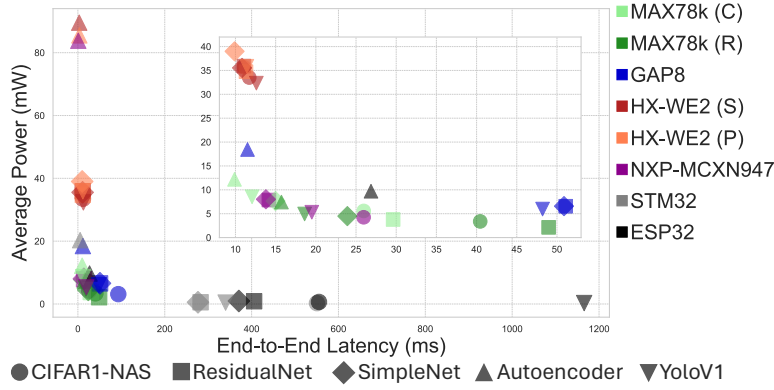


Figure 6: a visualization of average end-to-end latency vs power draw for evaluated models and platforms. The inset graph provides a magnified view of μ NPU platforms with lower end-to-end latency.

operations introduce negligible overhead across the other benchmark platforms with shared SRAM.

Differing from CPUs and GPUs, which rely on a 1D contiguous memory space, μ NPU hardware adopts a 2D memory layout; in this layout, one axis maps to parallel compute cores and the other organizes the logical address space. As shown in Fig. 1, each PE is equipped with its own weight memory space to avoid memory contention and maximize parallelization. This results in a hierarchical architecture with both channel-wise and weight-wise parallelism, though with the constraint that weights must use the same offset.

Recent work [29] has explored optimizing weight loading strategies for such 2D memory layouts to shrink I/O latency when switching models on a single device, including:

- virtualizing weight memory within the accelerator to reduce fragmentation,
- optimizing dynamic weight allocation to minimize loading/unloading overhead,
- and weight preloading, where the next model’s weights are loaded by the idle CPU into unused memory regions before execution.

Further work should include automating memory management, alongside reducing I/O latency for single-model execution, using techniques like just-in-time prefetching, dynamic quantization, or input-adaptive pruning.

4.2.3 Inference Another unexpected finding is the superior inference latency of the MAX78000 compared to the HX-WE2. The MAX78000 (C), for example, demonstrates an average $\sim 2.48\times$ latency improvement of the HX-WE2 (P), despite having significantly lower theoretical compute capacity (30 GOPs vs. 512 GOPs on the HX-WE2). This could be attributed to more optimized weight-stationary dataflow patterns for CNN workloads compared to the Arm Ethos-U55. However, the HX-WE2 still wins in terms of end-to-end latency with much reduced memory I/O latency. The relatively consistent

inference times across different models on the HX platforms also suggest its architecture is optimized for larger models than those in our benchmark suite. The MAX78000 demonstrates more variability in inference latency (ranging from 0.14 ms to 4.63 ms), suggesting greater scalability across differing model complexities.

The GAP8 demonstrates the highest end-to-end latency across all models - averaging $17\times$ slower than the MAX78000, despite having similar compute capacity (22.65 GOPs vs. 30 GOPs on the MAX78000). However, again, the GAP8’s large flash and RAM size make it more suitable for deploying large models or MoE architectures

4.2.4 CPU Post-Processing Post-processing operations, while often overlooked in benchmarking studies, can contribute to end-to-end latency and overall efficiency. We find CPU processing overhead is generally low across most of the evaluated platforms, in comparison to other execution stages, but is non-negligible for YoloV1’s NMS on certain platforms. For instance, the MAX78000 with RISC-V CPU active takes 3.82 ms in post-processing for YoloV1, compared to 2.62 ms spent in actual inference. This outlines the importance of minimizing CPU-dependent post-processing, and highlights a key design consideration with our benchmark; by ensuring all models are fully NPU-compatible across the various platforms, we aim to enable a fair comparison of end-to-end latency, avoiding bottlenecks or penalties caused by unsupported operators falling back to CPU execution. However, in real-world use, developers would build models that are optimized for a given target platform, making it necessary to consider the range of supported operators (which is quite limited on certain NPUs), and accuracy or performance trade-offs that might arise from using other, more compute-capable platforms, with more complex or unmodified models.

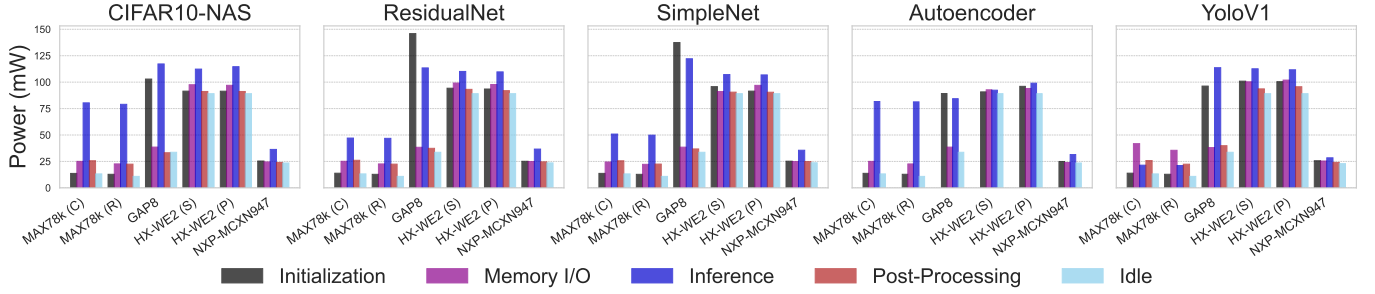


Figure 7: power consumption for each stage, model, and μ NPU.

4.3 Task-Specific Considerations

Memory Constraints and Model Complexity Memory capacity significantly influences the feasible model complexity for each platform. The GAP8’s expansive memory (8MB RAM, 20MB flash) enables deployment of substantially larger models than possible on the MAX78000 (512KB NPU memory, 128KB CPU memory), for example. This difference becomes critical for applications requiring more complex models, such as multi-class object detection or audio classification with large vocabulary sets.

The detailed memory I/O timing data provides additional insights into how different platforms handle model loading. The MAX78000’s long memory I/O times (8.84 - 26.53 ms) are more suitable for persistent model deployment. In contrast the HX-WE2’s comparatively large flash memory and low-latency memory I/O (0.03 - 1.11 ms), but longer initialization times (2.56 - 2.60 ms), are ideal for continuous inference or dynamic model switching.

Operational Modes and Power Profiles The ability to support different operational modes significantly impacts a platform’s suitability for specific applications. The MAX78000 displays high power variation between idle (10.87 - 13.21 mW) and inference (21.13 - 81.67 mW) states; hence, power gating mechanisms could be effectively leveraged in duty-cycled applications.

Further study of the various low-power modes available on our benchmark platforms is needed, including wake-up times, power gating mechanisms, and DVFS. Moreover, dual-CPU platforms with asymmetric co-processing capabilities could improve task distribution between cores — or enable hierarchical task-based wake-up of CPU cores — leading to further power-saving advantages. For instance, MAX78000’s combination of low-power RISC-V and compute-capable Cortex-M4 cores, when used in tandem alongside early-exit strategies for dynamic, low-power inference, could further optimize energy usage during model deployment.

Precision Requirements and Quantization Support The bit-width support of each platform represents another important consideration for application-specific deployment.

The MAX78000’s support for 1, 2, 4, and 8-bit operations enables highly optimized model deployment for applications where lower precision is acceptable, or for models amenable to aggressive quantization.

Conversely, applications requiring higher numerical precision may benefit from platforms like the HX-WE2, which supports floating-point acceleration up to 32-bit precision.

4.4 Summary of Results

We measured power consumption and latency for various model architectures across commercially-available μ NPU platforms. We find GOPS isn’t a reliable predictor for estimating end-to-end latency, and memory bandwidth enormously impacts performance.

The MAX78000 μ NPU, with its Cortex-M4 CPU active, offers the best *overall* efficiency, with NPU initialization considered, delivering consistent sub-30ms end-to-end latency across all models. However, its performance is primarily memory-bound, spending up to 90% of execution time on memory I/O operations. The HX-WE2 platform achieves an average end-to-end $\sim 1.93\times$ speedup over MAX78000 but with $\sim 3.13\times$ higher power consumption. The NXP-MCXN947 also offers relatively comparable (<30 ms) end-to-end latency, with fast initialization and memory I/O; despite its lower computational throughput, it exhibits high efficiency on our lower-complexity, memory-light benchmark models.

General-purpose MCUs demonstrate significantly lower efficiency, empirically validating the advantage of having dedicated neural hardware.

Excluding initialization overhead (*i.e.*, for applications requiring continuous operation), the MILK-V ranks overall highest in terms efficiency, with its large idle power draw outweighed by fast end-to-end inference latency. Notably, the NXP-MCXN947 ranks highest, both with and without NPU initialization, for the Autoencoder model, delivering over $2\times$ efficiency gains on its nearest competitor.

Key platform differentiators include memory capacity (affecting model complexity), power profiles (*e.g.*, MAX78000 shows significant variation between idle and inference states,

beneficial for duty-cycled applications), and precision support (*e.g.*, MAX78000 supports 1-8 bit operations while HX-WE2 supports up to 32-bit).

4.5 Future Directions

Advancing Hardware Architectures: Developing next-generation μ NPU architectures with larger on-chip cache and improved memory throughput is an obvious priority. This would (1) reduce the significant memory I/O overheads observed in certain platforms (*e.g.*, MAX78000) and (2) enable deployment of larger, more capable models or MoE architectures for context-aware inference.

Optimizing Model Weight-Loading: Together with hardware advancements, improved optimization of model architectures and loading strategies to maximize data reuse is also essential. The substantial memory I/O bottlenecks observed across certain platforms underscore the need for μ NPU-specialized weight virtualization, dynamic allocation optimization, and prefetching strategies.

Expanding Operator Support: Currently, most μ NPU platforms exhibit hugely limited operator support, focusing on CNN-based operators. Future designs should incorporate more expansive operator sets, towards supporting more diverse model architectures, such as transformers.

Improving Quantization and Model Compression: Fine-grained bit-width quantization and other non-standard model optimizations also remain inadequately supported across μ NPU platforms. This includes both a hardware and a software aspect, with existing software libraries designed for NN models on resource-constrained devices also generally lacking flexibility; TFLite/LiteRT, for example, only supports 8-bit integer and 16-bit float weight quantization.

Enabling On-Device Training: Current μ NPU platforms exclusively support NN inference, with no support for on-device training. However, model training on-device would enable personalization, continual learning, and adaptation to dynamic distribution shifts, without relying on cloud-based processing – vital for data privacy and remote deployments. Future μ NPU designs should aim to support quantized on-device training, requiring both memory-efficient training algorithms alongside hardware support for backpropagation.

Standardizing Model Formats: The heterogeneity in supported model formats across our various benchmark platforms is another issue. Vendors should aim to move towards unified model formats to reduce cross-platform compilation and deployment overheads.

Developing Accurate Simulators: Finally, reliable software simulators and predictive models for inference latency, power consumption, and memory utilization are notably absent for μ NPUs (and MCUs in general). Such tools would enable developers to optimize deployments without physical hardware, accelerating the end-to-end development cycle.

4.6 Practical Recommendations

We offer the following practical recommendations for embedded developers and hardware designers:

For Energy-Efficiency: The MAX78000 largely outperforms other μ NPU platforms in terms of energy-efficiency (when including NPU initialization), making it particularly well-suited for battery-powered applications. For extended battery life, consider leveraging its ability to dynamically power-gate portions of the system during idle periods.

For Latency-Critical Applications: The HX-WE2 platform offers low-latency with fast NPU initialization, memory I/O, and inference itself, making it best suited for applications requiring responsive model switching, real-time adaptation to changing conditions, or intermittent/duty-cycled operation. The NXP-MCXN947 also achieves low end-to-end inference latency at a significantly lower power budget, making it ideal for power-constrained workloads. Meanwhile, for latency-critical yet space-constrained applications, where power consumption isn't a major constraint and workloads don't require frequent NPU initialization/deinitialization, more powerful SoC architectures, like the MILK-V, could also be suitable. Further work could explore the performance of other SoC platforms [44, 45].

For Large Models: The GAP8's expansive memory makes it uniquely suitable for deploying larger, more complex models or implementing model-switching approaches where multiple specialized networks are employed based on operating conditions, despite its longer initialization times and inference latency. However, again, if power consumption isn't a major concern, the MILK-V's low inference latency and large memory capacity, with SD card support, could also make it a strong alternative.

For Security with Efficiency: Being Arm-based, the NXP-MCXN947 CPU includes Arm's TrustZone [46], enabling hardware isolation between secure and non-secure enclaves. With its competitive end-to-end inference latency (0.96-25.95 ms) and low power draw (22.91–36.69 mW), it may be suitable for security-centric applications without extreme constraints in any one dimension. Future work could explore extending its secure execution environment to integrate μ NPU acceleration via I/O passthrough, enabling protected yet efficient NN inference.

For Simple Models: For sufficiently simple models, general-purpose MCUs like the STM32H7 can achieve competitive efficiency without dedicated neural acceleration, obviating the need for specialized hardware.

4.7 Limitations

Several limitations should be considered when interpreting our benchmarking results:

Frequency Standardization: While enforcing a uniform CPU frequency across all platforms enables direct comparison of architectural efficiencies, it fails to showcase each platform’s peak performance – many of the benchmark platforms can operate at higher frequencies than evaluated.

Fixed Quantization Bit-Width: Our standardized INT8 quantization approach, while enabling fair comparison, does not leverage the full capabilities of platforms supporting lower bit-width operations (*e.g.*, MAX78000’s 1/2/4-bit support) or higher precision computation (*e.g.*, HX-WE2’s FLOAT16/FLOAT32 support). We also only focus on quantization as a means of reducing model size, excluding other optimization methods.

CPU Configuration: We also enforced uniform CPU divider settings across experiments; however, many platforms support variable divider configurations, which could potentially impact overall efficiency profiles.

Model Adaptation Constraints: The requirement to maintain structural consistency across all platforms necessitated compromises in model optimization. Platform-specific optimizations might yield slightly different efficiency profiles than our standardized approach.

Operator Support: Similarly, by ensuring all models are fully NPU-compatible across the various evaluated platforms, we negate the impact of unsupported NN operators. Further work should examine performance scaling across platforms with different sets of supported operators, using more complex or unmodified models, alongside precision-optimized models for each platform, and the impact of platform-specific architectural optimizations.

Development Toolchain Maturity We focus solely on performance metrics in this study. However, another often overlooked factor when selecting a target platform, deserving attention in future studies, is the relative maturity of its development ecosystem and model optimization tools.

5 RELATED WORK

Benchmarking NN Models on Constrained Hardware: A growing body of literature has explored NN benchmarking on constrained and mobile computing platforms. Japana et al.’s MLPerf benchmark introduced the first industry-standard open-source framework for performance evaluation of NNs on mobile devices equipped with diverse NN accelerators and software stacks [47]. Laskaridis et al. recently investigated the efficiency of large language models (LLMs) on various SOTA mobile platforms, including Android, iOS and Nvidia Jetson devices [48]. Reuther et al. explored the performance and power characteristics of a wide range of NN accelerators, spanning cellular GPUs, FPGA accelerators, up to data center hardware [49]. However, previous work on μ NPU platforms has been limited to application-level performance assessments [18, 19] or single-platform

standalone benchmarks [6, 27]. Furthermore, existing single-platform benchmarks often overlook certain operations in the end-to-end model pipeline [7]. Hence, to the best of our knowledge, our work details the first side-by-side and fine-grained benchmarking study of NN models across a number of commercially-available μ NPU platforms.

NN Accelerators for MCUs: NN accelerators offer vast potential in mitigating the computational and memory bottlenecks of traditional MCUs for NN inference. Beyond commercial accelerators (*e.g.*, Arm Ethos-U55), recent work has introduced new, more efficient custom designs. For instance, Venkataramani et al. designed RaPiD, an accelerator tailored for ultra-low-power INT4 inference, achieving an energy efficiency of 3-13.5 TOPS/W (average 7 TOPS/W) [50]. Conti et al. developed the XNOR Neural Engine, a digital, configurable hardware accelerator IP for binary neural networks, integrated into an MCU with an autonomous I/O subsystem and hybrid SRAM/standard cell memory [51].

Efficient On-Device Inference: Deploying NNs on MCUs is constrained by the underlying hardware’s memory capacity and throughput, with power consumption also often emerging as a bottleneck [52, 53]. Numerous works have explored model compression [12, 54, 55], the design of more efficient NN operators and architectures for lower resource usage [56–58], and adaptive NN inference based on input complexity and workload [59–61]. Various hardware-based optimizations have also been studied, such as parallel dataflow processing [21]. Our work aims to further advance efficient NN deployment across μ NPU platforms by identifying the current SOTA alongside existing bottlenecks.

6 CONCLUSION

Our comprehensive evaluation of various NN models across commercially-available μ NPUs reveals both expected trends as well as unexpected findings, contributing to the growing body of knowledge on embedded neural computation.

The significant performance advantages of dedicated neural acceleration are clearly demonstrated, with specialized platforms achieving up to two orders of magnitude higher energy-efficiency compared to general-purpose MCUs. We also highlight that theoretical computational capacity (GOPs) alone is an insufficient predictor of real-world performance. The stage-by-stage breakdown of model inference reveals critical bottlenecks on certain platforms – particularly in memory I/O operations – alongside key insights for future work in hardware and model design. We urge developers to consider trade-offs in latency, energy-efficiency, model complexity, and operational flexibility to achieve optimal performance in real-world deployments. We open-source our benchmarking framework and hope its use can streamline cross-platform model compilation and evaluation.

7 ACKNOWLEDGMENTS

This research was supported in part by the UKRI Open Plus Fellowship (EP/W005271/1: Securing the Next Billion Consumer Devices on the Edge), as well as funding from the Grantham Institute, Imperial College London.

References

- [1] Pietro Mercati and Ganapati Bhat. Self-Sustainable Wearable and Internet of Things (IoT) Devices for Health Monitoring: Opportunities and Challenges. *IEEE Design and Test*, 42(2):35–60, 2025.
- [2] Sarah Condran, Michael Bewong, Md Zahidul Islam, Lancelot Maphosa, and Lihong Zheng. Machine Learning in Precision Agriculture: A Survey on Trends, Applications and Evaluations Over Two Decades. *IEEE Access*, 10:73786–73803, 2022.
- [3] Chanwoo Kim, Dhananjaya Gowda, Dongsoo Lee, Jiyeon Kim, Ankur Kumar, Sungsoo Kim, Abhinav Garg, and Changwoo Han. A Review of On-Device Fully Neural End-to-End Automatic Speech Recognition Algorithms. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 277–283, 2020.
- [4] Emil Njor, Mohammad Amin Hasanpour, Jan Madsen, and Xenofon Fafoutis. A Holistic Review of the TinyML Stack for Predictive Maintenance. *IEEE Access*, 12:184861–184882, 2024.
- [5] Maxim Integrated. MAX78000. 2025. <https://www.analog.com/en>.
- [6] Arthur Moss, Hyunjong Lee, Lei Xun, Chulhong Min, Fahim Kawsar, and Alessandro Montanari. Ultra-Low-Power DNN Accelerators for IOT: Resource Characterization of the MAX78000. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 934–940, 2022.
- [7] Mitchell Clay, Christos Grecos, Mukul Shirvaikar, and Blake Richey. Benchmarking the MAX78000 artificial intelligence microcontroller for deep learning applications. In *Real-Time Image Processing and Deep Learning 2022*, volume 12102, pages 47–52. SPIE, 2022.
- [8] Linghe Kong, Jinlin Tan, Junqin Huang, Guihai Chen, Shuaitian Wang, Xi Jin, Peng Zeng, Muhammad Khan, and Sajal K Das. Edge-computing-driven Internet of Things: A survey. *ACM Computing Surveys*, 55(8):1–41, 2022.
- [9] Ruijin Wang, Jinshan Lai, Zhiyang Zhang, Xiong Li, Pandi Vijayakumar, and Marimuthu Karuppiah. Privacy-preserving Federated Learning for Internet of Medical Things under Edge Computing. *IEEE journal of biomedical and health informatics*, 27(2):854–865, 2022.
- [10] Cheng Wang, Zenghui Yuan, Pan Zhou, Zichuan Xu, Ruixuan Li, and Dapeng Oliver Wu. The security and privacy of mobile-edge computing: An artificial intelligence perspective. *IEEE Internet of Things Journal*, 10(24):22008–22032, 2023.
- [11] Jinhyuk Kim and Shiho Kim. Hardware accelerators in embedded systems. In *Artificial Intelligence and Hardware Accelerators*, pages 167–181. Springer, 2023.
- [12] Ji Lin, Wei-Ming Chen, Yujun Lin, Chuang Gan, Song Han, et al. MCUNet: Tiny deep learning on iot devices. *Advances in neural information processing systems*, 33:11711–11722, 2020.
- [13] Swapnil Sayan Saha, Sandeep Singh Sandha, and Mani Srivastava. Machine learning for microcontroller-class hardware: A review. *IEEE Sensors Journal*, 22(22):21362–21390, 2022.
- [14] Ji Lin, Ligeng Zhu, Wei-Ming Chen, Wei-Chen Wang, Chuang Gan, and Song Han. On-device training under 256kb memory. *Advances in Neural Information Processing Systems*, 35:22941–22954, 2022.
- [15] Young D Kwon, Rui Li, Stylianos I Venieris, Jagmohan Chauhan, Nicholas D Lane, and Cecilia Mascolo. TinyTrain: resource-aware task-adaptive sparse training of DNNs at the data-scarce edge. *arXiv preprint arXiv:2307.09988*, 2023.
- [16] Yushan Huang, Ranya Aloufi, Xavier Cadet, Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Low-Energy On-Device Personalization for MCUs. In *2024 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 45–58. IEEE, 2024.
- [17] Erez Manor and Shlomo Greenberg. Custom Hardware Inference Accelerator for Tensorflow Lite for Microcontrollers. *IEEE Access*, 10:73484–73493, 2022.
- [18] Guanchu Wang, Zaid Pervaiz Bhat, Zhimeng Jiang, Yi-Wei Chen, Daochen Zha, Alfredo Costilla Reyes, Afshin Niktash, Gorkem Ulkar, Erman Okman, Xuanting Cai, et al. Bed: A Real-Time Object Detection System for Edge Devices. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4994–4998, 2022.
- [19] Weining Song, Stefanos Kaxiras, Thiemo Voigt, Yuan Yao, and Luca Mottola. TaDA: Task Decoupling Architecture for the Battery-less Internet of Things. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pages 409–421, 2024.
- [20] Luca Caronti, Khakim Akhunov, Matteo Nardello, Kasim Sinan Yildirim, and Davide Brunelli. Fine-grained hardware acceleration for efficient batteryless intermittent inference on the edge. *ACM Transactions on Embedded Computing Systems*, 22(5):1–19, 2023.
- [21] Taesik Gong, Fahim Kawsar, and Chulhong Min. DEX: Data Channel Extension for Efficient CNN Inference on Tiny AI Accelerators. *Advances in Neural Information Processing Systems*, 37:43925–43951, 2025.
- [22] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. How to Evaluate Deep Neural Network Processors: TOPS/W (Alone) Considered Harmful. *IEEE Solid-State Circuits Magazine*, 12(3):28–41, 2020.
- [23] ARM. ARM Ethos-U Processor Series Brief, 2022. Accessed: 2025-03-12.
- [24] Marco Giordano and Michele Magno. A Battery-Free Long-Range Wireless Smart Camera for Face Recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 594–595, 2021.
- [25] Bakar, Abu and Goel, Rishabh and De Winkel, Jasper and Huang, Jason and Ahmed, Saad and Islam, Bashima and Pawelczak, Przemyslaw and Yildirim, Kasim Sinan and Hester, Josiah. Protean: An energy-efficient and heterogeneous platform for adaptive and hardware-accelerated battery-free computing. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 207–221, 2022.
- [26] Edward Humes, Mozhgan Navardi, and Tinoosh Mohsenin. Squeezed Edge YOLO: Onboard Object Detection on Edge Devices, 2023.
- [27] Yushan Huang, Taesik Gong, SiYoung Jang, Fahim Kawsar, and Chulhong Min. Energy Characterization of Tiny AI Accelerator-Equipped Microcontrollers. In *Proceedings of the 2nd International Workshop on Human-Centered Sensing, Networking, and Multi-Device Systems*, pages 1–6, 2024.
- [28] TensorFlow.org. TensorFlow Lite for Microcontrollers. <https://www.tensorflow.org/lite/microcontrollers>, 2022. Accessed: 2025-03-13.
- [29] Changmin Jeon, Taesik Gong, Juheon Yi, Fahim Kawsar, and Chulhong Min. TinyMem: Boosting Multi-DNN Inference on Tiny AI Accelerators with Weight Memory Virtualization. In *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*, HotMobile '25, page 1–6, New York, NY, USA, 2025. Association for Computing Machinery.
- [30] GreenWaves Technologies. GAP8 Product Brief, 2021. Accessed: 2025-03-12.
- [31] Cristian Ramirez, Adrián Castelló, Héctor Martínez, and Enrique S. Quintana-Orti. Communication-Avoiding Fusion of GEMM-Based Convolutions for Deep Learning in the RISC-V GAP8 MCU. *IEEE Internet of Things Journal*, 11(21):35640–35653, 2024.

- [32] Julian Moosmann, Hanna Müller, Nicky Zimmerman, Georg Rutishauser, Luca Benini, and Michele Magno. Flexible and Fully Quantized Lightweight TinyissimoYOLO for Ultra-Low-Power Edge Systems. *IEEE Access*, 12:75093–75107, 2024.
- [33] Himax Technologies. WiseEye2 AI Processor, 2025. Accessed: 2025-03-12.
- [34] NXP Semiconductors. FRDM-MCXN947 Development Board, 2025. Accessed: 2025-03-12.
- [35] STMicroelectronics. STM32H7A3ZI Microcontroller, 2025. Accessed: 2025-03-12.
- [36] Tommaso Addabbo, Ada Fort, Marco Mugnaini, Valerio Vignoli, Matteo Intravaia, Marco Tani, Monica Bianchini, Franco Scarselli, and Barbara Toniella Corradini. Gravimetric system for enhanced security of accesses to public places embedding a mobilenet neural network classifier. *IEEE Transactions on Instrumentation and Measurement*, 71:1–10, 2022.
- [37] Espressif Systems. ESP32-S3 Datasheet, 2025. Accessed: 2025-03-12.
- [38] Milk-V. Milk-V Duo, 2025. Accessed: 2025-03-12.
- [39] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once-for-All: Train One Network and Specialize it for Efficient Deployment, 2020.
- [40] Seyyed Hossein Hasanpour, Mohammad Rouhani, Mohsen Fayyaz, and Mohammad Sabokrou. Lets keep it simple: Using simple architectures to outperform deeper and more complex architectures, 2023.
- [41] Inc. Analog Devices. Motor Fault Sample Dataset. <https://github.com/analogdevicesinc/CbM-Datasets/tree/main>, 2024. Accessed: 2024-04-01.
- [42] Tsung-Yi Lin, Peizhao Ma, Serge Belongie, and Fei-Fei Li. Microsoft COCO: Common Objects in Context, 2014. Accessed: 2025-03-12.
- [43] Monsoon Solutions Inc. Monsoon High voltage power monitor. 2024. <https://www.msoon.com/>.
- [44] Luckfox. Luckfox Pico. <https://www.luckfox.com/Luckfox-Pico>. Accessed: 2025-03-17.
- [45] Canaan. K230. https://developer.canaan-creative.com/k230/zh/dev/00_hardware/K230_datasheet.html. Accessed: 2025-03-17.
- [46] ARM. ARM TrustZone, 2025. Accessed: 2025-03-17.
- [47] Vijay Janapa Reddi, David Kanter, Peter Mattson, Jared Duke, Thai Nguyen, Ramesh Chukka, Ken Shiring, Koan-Sin Tan, Mark Charlebois, William Chou, et al. MLPerf mobile inference benchmark: An industry-standard open-source machine learning benchmark for on-device AI. *Proceedings of Machine Learning and Systems*, 4:352–369, 2022.
- [48] Stefanos Laskaridis, Kleomenis Katevas, Lorenzo Minto, and Hamed Haddadi. Melting point: Mobile evaluation of language transformers. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 890–907, 2024.
- [49] Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadepally, Sidharth Samsi, and Jeremy Kepner. Survey and benchmarking of machine learning accelerators. In *2019 IEEE high performance extreme computing conference (HPEC)*, pages 1–9. IEEE, 2019.
- [50] Swagath Venkataramani, Vijayalakshmi Srinivasan, Wei Wang, Sanchari Sen, Jintao Zhang, Ankur Agrawal, Monodeep Kar, Shubham Jain, Alberto Mannari, Hoang Tran, et al. RaPiD: AI accelerator for ultra-low precision training and inference. In *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, pages 153–166. IEEE, 2021.
- [51] Francesco Conti, Pasquale Davide Schiavone, and Luca Benini. Xnor neural engine: A hardware accelerator ip for 21.6-fj/op binary neural network inference. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2940–2951, 2018.
- [52] Sayed Saad Afzal, Waleed Akbar, Osvy Rodriguez, Mario Doumet, Unsoo Ha, Reza Ghaffarivardavagh, and Fadel Adib. Battery-free wireless imaging of underwater environments. *Nature communications*, 13(1):5546, 2022.
- [53] Yuchen Zhao, Sayed Saad Afzal, Waleed Akbar, Osvy Rodriguez, Fan Mo, David Boyle, Fadel Adib, and Hamed Haddadi. Towards battery-free machine learning and inference in underwater environments. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications*, pages 29–34, 2022.
- [54] Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155, 2020.
- [55] Muhammad Zawish, Steven Davy, and Lizy Abraham. Complexity-driven model compression for resource-constrained deep learning on edge. *IEEE Transactions on Artificial Intelligence*, 5(8):3886–3901, 2024.
- [56] Yu Pan, Ye Yuan, Yichun Yin, Zenglin Xu, Lifeng Shang, Xin Jiang, and Qun Liu. Reusing pretrained models by multi-linear operators for efficient training. *Advances in Neural Information Processing Systems*, 36:3248–3262, 2023.
- [57] Jakub M Tarnawski, Amar Phanishayee, Nikhil Devanur, Divya Mahajan, and Fanny Nina Paravecino. Efficient algorithms for device placement of dnn graph operators. *Advances in Neural Information Processing Systems*, 33:15451–15463, 2020.
- [58] Lingda Li, Robel Geda, Ari B Hayes, Yanhao Chen, Pranav Chaudhari, Eddy Z Zhang, and Mario Szegedy. A simple yet effective balanced edge partition model for parallel computing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(1):1–21, 2017.
- [59] Stefanos Laskaridis, Alexandros Kouris, and Nicholas D. Lane. Adaptive Inference through Early-Exit Networks: Design, Challenges and Directions. In *Proceedings of the 5th International Workshop on Embedded and Mobile Deep Learning*, EMDL’21, page 1–6, New York, NY, USA, 2021. Association for Computing Machinery.
- [60] Bitu Darvish Rouhani, Azalia Mirhoseini, and Farinaz Koushanfar. Delight: Adding energy dimension to deep neural networks. In *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*, ISLPED ’16, page 112–117, New York, NY, USA, 2016. Association for Computing Machinery.
- [61] Noam Shazeer, Kayvon Fatahalian, William R. Mark, and Ravi Teja Mullapudi. HydraNets: Specialized Dynamic Architectures for Efficient Inference. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2018.

Supplementary Material

Table 6: complete latency (ms) and power (mW) measurements across each stage, model, and platform.

Model	Stage	MAX78k(C)		MAX78k(R)		GAP8		NXP-MCXX947		HX-WE2(S)		HX-WE2(P)		STM32H7A3ZI		ESP32	
		Time	Power	Time	Power	Time	Power	Time	Power	Time	Power	Time	Power	Time	Power	Time	Power
NAS	Initialization	0.07±0.001	13.67±0.02	0.17±0.002	12.84±0.03	12.94±0.237	102.93±1.11	0.24±0.001	25.40±3.39	2.60±0.002	91.50±0.22	2.60±0.001	91.44±0.19	0.43±0.005	47.38±0.07	86.73±0.001	105.93±0.09
	Memory I/O	21.24±0.004	25.05±0.05	35.53±0.008	22.66±0.03	2.66±0.007	38.58±0.95	0.83±0.001	24.52±0.01	0.12±0.001	97.55±1.25	0.12±0.001	97.01±0.71	550.01±0.012	54.91±0.09	468.38±0.001	151.99±0.31
	Inference	4.63±0.002	80.41±0.08	4.64±0.002	79.03±0.05	77.40±0.470	117.22±1.47	24.86±0.002	36.38±0.65	8.99±0.022	112.35±0.76	8.32±0.008	114.62±0.76	0.02±0.001	46.01±0.03	0.02±0.001	101.93±1.86
	Post-Processing	0.02±0.001	25.66±0.04	0.11±0.001	22.43±0.06	0.08±0.001	33.23±1.28	0.02±0.001	24.03±1.19	0.01±0.001	91.19±1.31	0.01±0.001	91.17±1.39	-	-	-	77.73±0.07
ResNet	Idle	-	13.21±0.02	-	10.87±0.01	-	33.67±0.35	-	23.53±0.14	-	89.09±1.30	-	89.09±1.30	-	38.13±0.03	-	77.73±0.07
	Initialization	0.07±0.001	13.87±0.05	0.17±0.001	12.82±0.07	10.13±0.01	145.90±3.59	0.24±0.001	25.14±1.99	2.59±0.001	94.27±0.23	2.60±0.001	95.51±0.01	0.43±0.002	48.32±0.07	87.49±0.09	106.41±0.07
	Memory I/O	26.53±0.002	25.14±0.05	44.87±0.008	22.64±0.04	2.47±0.130	38.39±0.31	0.91±0.001	24.89±0.99	0.12±0.001	99.15±0.45	0.14±0.001	97.74±1.37	282.00±0.030	54.98±0.08	318.15±0.001	144.91±0.27
	Inference	2.89±0.001	47.12±0.09	2.96±0.001	46.86±0.07	38.45±0.01	113.47±0.40	12.97±0.02	36.69±0.52	8.30±0.001	110.13±0.79	8.54±0.015	91.95±1.27	0.08±0.002	50.47±0.09	0.05±0.001	102.79±0.71
SimpleNet	Post-Processing	0.12±0.001	26.10±0.06	0.96±0.002	22.45±0.05	0.04±0.001	37.41±1.35	0.04±0.001	24.60±1.35	0.04±0.001	93.19±0.15	0.04±0.001	89.09±1.30	-	-	-	77.73±0.07
	Idle	-	13.21±0.02	-	10.87±0.01	-	33.67±0.35	-	23.53±0.14	-	89.09±1.30	-	89.09±1.30	-	38.13±0.03	-	77.73±0.07
	Initialization	0.07±0.001	13.75±0.06	0.17±0.001	12.86±0.05	8.17±0.09	89.22±1.28	0.23±0.001	24.95±0.20	2.60±0.002	90.79±0.79	2.60±0.001	96.04±1.30	-	-	-	77.73±0.07
	Memory I/O	9.66±0.001	25.04±0.08	15.40±0.008	22.63±0.06	2.63±0.042	38.55±0.56	0.22±0.001	24.01±1.31	0.03±0.001	92.75±1.83	0.03±0.001	94.03±0.62	0.12±0.002	48.17±0.06	19.76±0.002	103.51±0.59
Autoenc	Inference	0.14±0.001	81.67±0.09	0.16±0.001	81.34±0.07	0.67±0.015	84.35±1.25	0.51±0.001	31.49±0.44	0.45±0.001	92.40±1.19	0.75±0.001	99.02±1.56	5.01±0.110	56.11±0.09	7.11±0.001	157.17±0.17
	Post-Processing	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Idle	-	13.21±0.02	-	10.87±0.01	-	33.67±0.35	-	23.53±0.14	-	89.09±1.30	-	89.09±1.30	-	38.13±0.03	-	77.73±0.07
	Initialization	0.07±0.001	13.90±0.03	0.17±0.001	12.87±0.02	7.46±0.001	96.28±2.90	0.28±0.001	25.66±0.39	2.60±0.003	100.94±1.01	2.60±0.001	100.45±0.59	3.64±0.002	49.04±0.03	628.97±0.002	106.49±0.63
YoloV1	Memory I/O	8.84±0.005	41.82±0.04	11.99±0.002	35.55±0.04	2.83±0.037	38.10±0.11	0.77±0.001	25.41±1.61	1.11±0.001	100.31±3.99	1.11±0.001	101.95±1.90	-	-	-	-
	Inference	2.61±0.002	21.47±0.04	2.62±0.001	21.13±0.05	36.11±0.047	113.67±2.39	17.18±0.001	28.43±1.66	8.48±0.001	112.63±1.30	7.30±0.001	111.69±0.97	336.00±0.013	54.99±0.08	536.22±0.001	129.74±0.07
	Post-Processing	0.54±0.001	25.87±0.06	3.82±0.001	22.41±0.04	1.82±0.002	39.89±0.62	1.27±0.001	24.08±0.15	0.42±0.001	93.63±1.45	0.42±0.001	95.65±0.79	0.48±0.001	45.71±0.04	0.53±0.001	109.57±0.08
	Idle	-	13.21±0.02	-	10.87±0.01	-	33.67±0.35	-	22.91±0.14	-	89.09±1.30	-	89.09±1.30	-	38.13±0.03	-	77.74±0.07

Notes:

- For MCUs without neural hardware, STM32H7A3ZI and ESP32, Initialization and Memory I/O are combined.
- The post-processing for ResidualNet, SimpleNet, and NAS models is composed of a softmax operation.
- The post-processing for YOLOv1 is a NMS (non-max suppression) operation, also with softmax.