Embedded AI performances of Nvidia's Jetson Orin SoC series

Agathe Archet, Nicolas Gac, François Orieux, Nicolas Ventroux

Embedded AI performances of Nvidia's Jetson Orin SoC series

Agathe Archet *† Nicolas Gac † François Orieux † Nicolas Ventroux *

* Thales Research and Technology, 1 Avenue Augustin Fresnel, 91120 Palaiseau, France

† Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes
3 rue Joliot Curie, 91190 Gif-Sur-Yvette, France

***Abstract*—Energy efficiency is key in many embedded systems that must achieve best performances for a given power budget. Additionally, new neural network-based applications combine multiple processing needs. For such applications, heterogeneous system-on-chips, such as the Nvidia Jetson Orin series, include different computing capabilities to propose new interesting latency and power consumption trade-offs. But, choosing the suitable Jetson module for a given application's need can be confusing since these modules have many operating ranges and several accelerators. In this paper, we evaluate through emulation the embedded performances of popular neural networks to provide a first hands-on insight of all Jetson Orin modules.**

***Index Terms*—Heterogeneous computing, Convolution Neural Networks, Embedded AI, Jetson Orin modules, DLA, Computer vision**

I. Introdution

To face more complex Convolution Neural Networks (CNNs) and energy-intensive inferences on GPUs, new AI-based accelerators are used to improve embedded systems' overall SWaP (Size, Weight, and Power) constraints. Among these solutions, Nvidia's Jetson Orin series are heterogeneous System-on-Chips (SoC) composed of various computing capabilities to answer the different application needs and better comply with embedded constraints. Together, each Orin module's accelerators propose new interesting latency and power consumption trade-offs for the inference of CNNs. This paper aims at identifying the potential performance variations between the Jetson Orin modules and their accelerators for different CNN inferences.

II. The Orin series' SoC and accelerators

The Jetson Orin series is composed of three SoC subfamilies, with two SoC/modules each: the AGX Orin for high performance, the Jetson NX Orin for average performance and power, and the Jetson Orin Nano for low-power computing. Each module has up to three kinds of units: a CPU for general-purpose processing, an Ampere GPU for intensive data parallelism, and optionally one or two Deep Learning Accelerators (DLA) dedicated to energy-efficient neural network processing, as shown in Table I. During a CNN inference, when combining all the units, the Jetson Orin series cover a throughput from 20 to 275 sparse TOP/s (Tera Operations per Second) for a power consumption between 5 and 60 W [1].

TABLE I: Jetson Orin series characteristics [2][3]

Jetson	**Nano**		**NX**		**AGX**	
Orin module	4 GB	8 GB	8 GB	16 GB	32 GB	64 GB
CPU cores	6	6	6	8	8	12
CUDA cores (GPU)	512	1024	1024	1024	1792	2048
Tensor cores (GPU)	16	32	32	32	56	64
DLA cores	–	–	1	2	2	2
TOP/s	20	40	70	100	200	275
Power (W)	5-10	7-15	10-20	10-25	15-40	15-60

Aside from the module families, the accelerators themselves allow a variety of inferences inside a single Orin module. In addition to their different hardware specifications, as seen in Table II, the GPU and the DLA accelerators also experience distinct software optimizations during CNN inferences on complexity reduction and memory optimization. Such specificities result in inferences with variable latencies and power consumption.

TABLE II: AI Accelerators on Jetson AGX Orin 64GB [2][1]

Accelerator	**GPU**	**2 × DLA 2.0**
Type	GP-GPU	fixed-funtion accelerator
Computing units	2048 CUDA cores + 64 Tensor Cores	6 IP-based modules
AI performances*	170 TOPs	52.5 TOPs per DLA
Power consumption	High	2.5 more effcient than the GPU (TOPs/Watt)

* sparse INT8

III. CNN deployment workflow and benchmark

To identify the enabled latency and power consumption scopes, a benchmark of common CNNs is implemented (Table III).

TABLE III: CNN descriptions

Name	**Task**	**Input size**	**GMAC**
ResNet50	classification	3x224x224	4.2
MobileNetV2	classification	3x224x224	0.39
ResNet34-SSD	object detection	3x1200x1200	216.8
MobileNet-SSD	object detection	3x300x300	1.3
UNet	semantic segmentation	3x128x128	10.6

TABLE IV: Latency and energy efficiency benchmark of various CNN accross the Jetson Orin series SoCs * +

CNN name	Target	Orin configuration (board + power mode)													
		AGX				NX				Nano					
		64 GB		32 GB		16 GB		8 GB		8 GB		4 GB	
		MAXN	15W	MAXN	15W	MAXN	10W	MAXN	10W	7W		5W	
ResNet50	GPU	0.79 | 151	3.76 | 106	1.19 | 144	3.58 | 99.2	1.39 | 149	2.96 | 117	1.64 | 142	2.94 | 118	4.27 | 93.7		4.16 | 105	
	DLA	1.64 | 138	2.05 | 188	1.69 | 141	2.11 | 162	1.75 | 150	6.37 | 61.1	1.75 | 158	6.38 | 62	-		-	
MobileNetV2	GPU	0.49 | 30.4	1.83 | 21.9	0.71 | 28.3	1.78 | 20.3	0.86 | 29.1	1.52 | 23.6	0.97 | 28.9	1.53 | 23.5	1.13 | 18.2		1.55 | 20.1	
	DLA	1.04 | 22.1	1.87 | 22.5	1.07 | 22.5	1.86 | 20.5	1.46 | 20.6	4.36 | 9.40	1.46 | 21.3	4.36 | 9.43	-		-	
ResNet34-SSD	GPU	8.20 | 485	59.5 | 328	12.7 | 508	59.5 | 302	20.8 | 442	58.4 | 299	24.8 | 432	58.4 | 299	30.4 | 232		59.4 | 251	
	DLA	29.4 | 374	59.8 | 369	37.6 | 316	69.8 | 280	64.0 | 243	233 | 92.4	63.0 | 256	231 | 93.6	-		-	
MobileNet-SSD	GPU	0.52 | 88.5	2.04 | 65.8	0.76 | 85.8	1.99 | 60.4	0.87 | 84.7	1.78 | 67.3	1.00 | 80.8	1.78 | 66.7	1.27 | 51.6		1.94 | 57.9	
	DLA	1.11 | 71.8	2.02 | 68.4	1.15 | 70.6	1.98 | 63.5	1.59 | 62.8	5.11 | 26.1	1.59 | 64.8	5.10 | 26.1	-		-	
UNet	GPU	0.63 | 379	4.34 | 229	1.13 | 350	4.37 | 208	1.49 | 328	4.01 | 221	1.76 | 316	3.98 | 220	2.16 | 169		3.92 | 190	
	DLA	3.08 | 196	4.49 | 224	4.12 | 148	5.39 | 165	5.32 | 139	16.2 | 63.4	5.31 | 145	16.2 | 63.7	-		-	

* Cell content : latency (ms) | Energy efficiency (GMAC/s/W) + Row content : worst value(s) / best value(s)

First, CNNs of various applications are defined or reused from Nvidia documentation. Then, the networks are deployed according to the usual workflow for Nvidia embedded SoC (Figure 1).

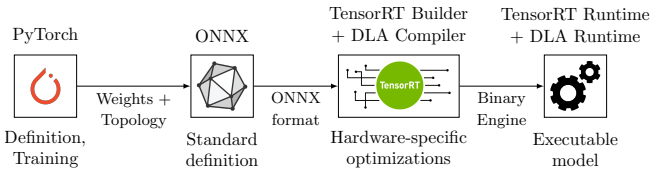

Fig. 1: Inference workflow steps to the Jetson AGX Orin.

Nvidia's TensorRT inference framework applies further optimizations [3] and provides a binary for the target. The six possible Orin hardware targets are emulated with the Jetson Orin Development Kit and Jetpack 5.1 to get accurate performances. Finally, each inference performance is evaluated with mean latency and power consumption measurements. Latency is retrieved from TensorRT's log outputs, and power consumption is estimated with the method described in [4], with an error below 3.5% on Orin AGX 64GB.

IV. Experimentation and Analysis

The CNNs benchmark is analyzed over 500 inferences on the GPU or on one DLA, in INT8 format, by varying the following parameters: the *power mode* (the highest and lowest configurations, and the lowest one for the Nano modules), and the hardware target. The results are presented in Table IV and Figure 2 shows an example of plotted performances.

As seen in Figure 2, a gap in power estimation remains for the Jetson Orin Nano boards despites the name of the power mode (e.g., 8.9 W measured in 5W mode). This difference could come from the emulation itself or the measurement method's approximations.

Under emulated conditons, the Orin AGX series proposes the fastest inferences, but it also has the most energy-efficient inferences. From the CNN benchmark, the latency for a model on the GPU can vary by a factor of at least 4.2 (MobileNetV2), and for its energy-efficiency by a factor of 1.7 (MobileNet-SSD). For the DLA, these factors are at least 3.5 (ResNet50) and 2.4 (MobileNet-V2). Latency and energy-efficiency differences between each board are not constant factors. They can vary depending on the inferred CNN. Between Orin boards, GPU performances differ more on latency and less on energy than the DLA performances.

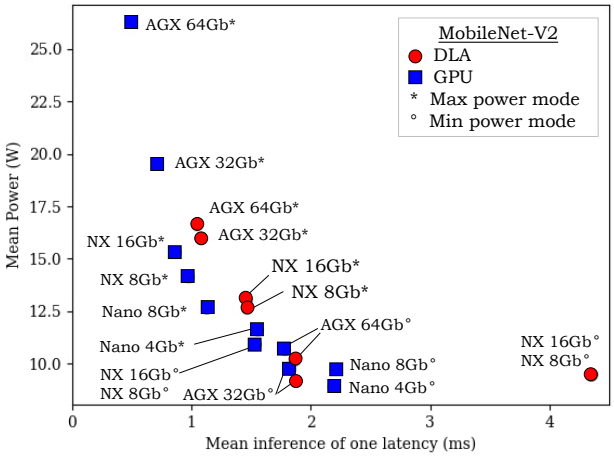

Fig. 2: Jetson Series' performance on MobileNetV2

V. Conclusion

By using emulation on Nvidia's Orin DevKit, it is possible to obtain an overview of the Jetson Orin series performance for AI. With a benchmark of 5 different CNNs, the latency and the energy-efficiency variations among each Orin module's accelerators depends on the inferred CNN and the used accelerator (GPU or DLA). Yet, with the DLA accelerator, this new heterogenous SoC family provides a even more complete performance spectrum within each module to meet embedded needs. These results and conclusion need to be confirmed with real platforms, especially for the energy consumption.

References

[1] L. S. Karumbunathan, "NVIDIA Jetson AGX Orin Series Technical Brief v1.2," Nvidia, Tech. Rep., 2022.

[2] R. Cherukuri, "Leveraging deep learning accelerators on nvidia agx platforms," 2022, nvidia GTC Digital Spring.

[3] Nvidia, "Tensorrt 8.5 developer guide," 2023.

[4] C. F. Rodrigues et al., "Fine-grained energy profiling for deep convolutional neural networks on the jetson tx1," in *IISWC*, 2017.