Application Note

Performance and efficiency benchmarking with TDA4 Edge AI processors

Texas Instruments

ABSTRACT

AI is revolutionizing our lifestyles with constant innovation in deep learning and machine learning driving new use cases across home, retail, and factories. AI at the edge is instrumental for continued success of AI delivering low latency, privacy, and better user experience. The key AI function that happens in an embedded edge device is inference. This is where Texas Instruments (TI) is innovating with TDA4x processor family specially designed to make greener, smarter, and safer edgeAI devices possible.

With industry-leading vision and AI accelerators, TDA4x processors achieve more than 60% higher deep learning performance and energy efficiency compared to leading GPU based architectures. With process and technology leadership, developers can achieve more than six times better deep learning performance compared to leading FPGA based architectures that exist today.

This application note uses the industry standard performance and power benchmarking used to compare the TDA4x system-on-chip (SoC) with other architectures. TDA4x processor family also comes with easy to use, no-cost to low-cost development platforms making it easier for developers to innovate with AI even without any prior experience.

ADVANCE INFORMATION

Table of Contents

List of Figures

1 Introduction

The world population today is 7.8 billion and is on the constant rise with an estimate of 10 billion by 2050 [1]. The growing population needs necessities such as food, clothing and ever-increasing comforts and tools - safely and securely. There is constant technology innovation across all markets - consumer, industrial and automotive to meet these needs. New technologies that we all got used to have made the data generation more affordable and more fun. Think about the number of pictures taken with smart phones and the amount of data being generated from various sensors and edge devices in buildings and factories. All this data is propelling end-to-end automation in factories and buildings to drive productivity to produce more goods and services. This exponentially increases the data that has to be managed - processed, analyzed to take corrective actions. For example, a smart factory could have more than 50,000 sensors and generate several petabytes a day. Even a standard office building will generate hundreds of gigabytes of data. Most of this data will be stored, managed, analyzed and kept right where it was produced, at the edge driven by security, real-time performance and reliability

1.1 Vision Analytics

There are three types of data produced in the edge devices - video, audio and other sensor data. Video-based analytics tend to be more complex as each video is a collection of images per second and the image itself will have multiple channels - Red, Green and Blue. With advances in cameras, vision-based analytics are gaining momentum across many applications - smart video doorbells, video surveillance, drones, robots, autonomous vehicles and last mile delivery. Fundamentally, there are three functions one can implement with vision-based analytics as shown in Figure 1-1: Classification, Detection and Segmentation. You can see from the same image, three different functions that can be implemented based on vision-analytics in an edgeAI system - starting from classifying the image to pixel level analysis of the entire scene.

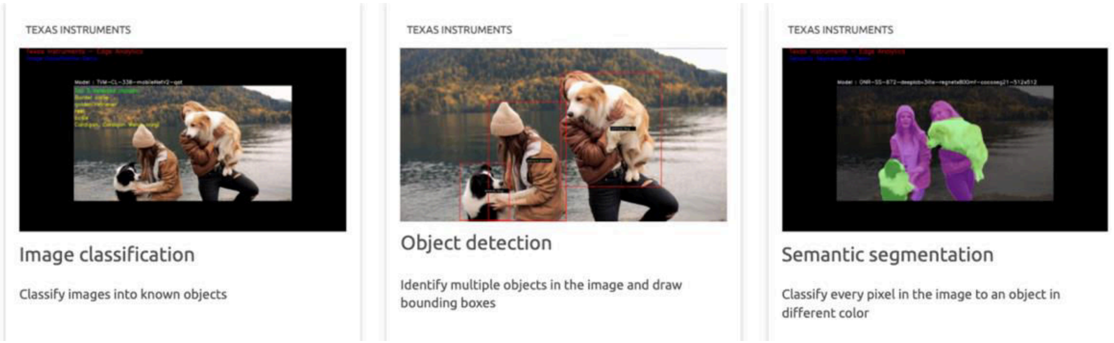

TEXAS INSTRUMENTS
Image classification
Classify images into known objects
TEXAS INSTRUMENTS
Object detection
Identify multiple objects in the image and draw bounding boxes
TEXAS INSTRUMENTS
Semantic segmentation
Classify every pixel in the image to an object in different color

Figure 1-1. Top three vision AI functions

1.2 End Equipments

The AI-enabled vision market is an exciting area with rapid growth expected in the next few years. Vision based analytics have broad use cases across many different markets and end equipment as shown in Figure 1-2.

Figure 1-2. Example Vision AI applications

For example, a robot in factory or warehouse setting or a last-mile robot can use vision-based analytics to do the below functions.

1. Obstacle detection
2. Pose estimation

A surveillance camera can be smarter with edgeAI functionality by analyzing the objects to do extended functions such as:

1. Automatic object detection
2. Intrusion and hazard detection
3. Safety monitoring

A smart shopping cart can use vision-based analytics to completely automate shopping experience for better efficiency and lower costs to consumers by adding functions such as:

1. Automatic checkout
2. Real-time pricing and nutrition information display
3. Retail Analytics

1.3 Deep learning: State-of-the-art

Deep learning model development is a hot research area in the deep learning and AI community. Driven by smartphones and the amount of image data that is being generated, a special focus has been given for image or vision deep learning functions to be able to identify faces, scenes, moods and other information in pictures. A specific type of neural network, Convolution Neural network (CNN), is an enabler for the latest advancements in computer vision. Convolution is a cool technique to detect different features in the input image. Convolution process uses a kernel, also called a filer, to sweep across an image to detect patterns in the image. A kernel is a very small matrix (usually 3x3 or 5x5) with a set of weights corresponding to its size and typically detects one feature in the image such as eyes, nose or a specific expression.

The seminal AlexNet [3] paper in 2012, showed researchers and industry that deep learning was an extremely effective algorithmic processing technique in solving computer vision tasks like classification, object detection and semantic segmentation. This triggered a series of new innovations continuing to improve the inference performance and efficiency targeting myriad of applications from robots and smart retail carts to last mile delivery autonomous delivery systems.

Figure 1-3 below shows popular models used in the industry today starting with AlexNet [4]. In general, there is a clear trade-off between the accuracy of the model and the number of operations used by the model shown in Giga operations (G-ops) [4].

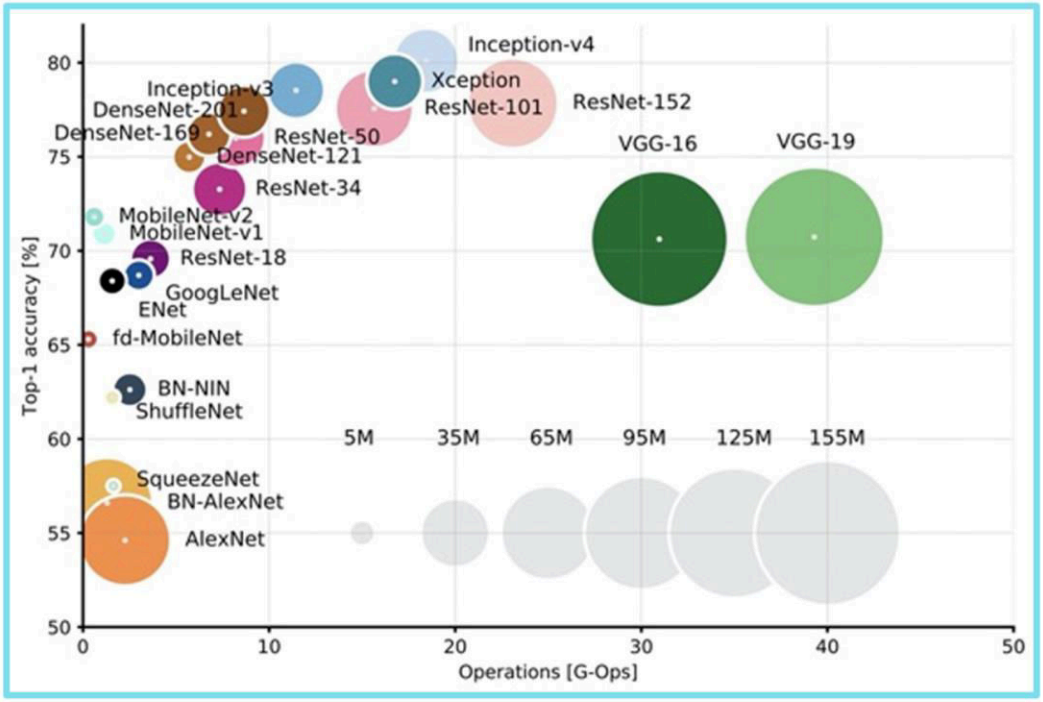


Figure 1-3. Popular deep learning models

This trade-off highlights the need for efficient SoC architectures to be able to run large computations efficiently. Deep learning community is pretty vibrant and constant innovation is happening to improve the model performance and efficiency. TI is constantly looking into new SoC technology innovations to offer this community best-in-class AI capabilities.

It is common practice to use architectures of deep learning networks published in literature. There are open-source implementations of several models. TI is making the process even easier with its own Model Zoo. TI's ModelZoo is a large collection of models that are optimized for inference speed and low power consumption. The models used in this benchmarking app note are examples of such open source models.

2 Embedded edge AI system: Design considerations

There are two aspects required of any neural network or a deep learning model: training and inference. The training function involves using a set of training data to train the model. Once the training is done, the model can be used for inference on a new set of input data. Typically, training is done once or a few times for a given product. Inference, on the other hand, happens all the time for a given edgeAI system.

Figure 2-1 shows the difference between the two steps. **This is the reason why the inference process needs to be optimized for high performance and high energy efficiency for any embedded edge AI device.**

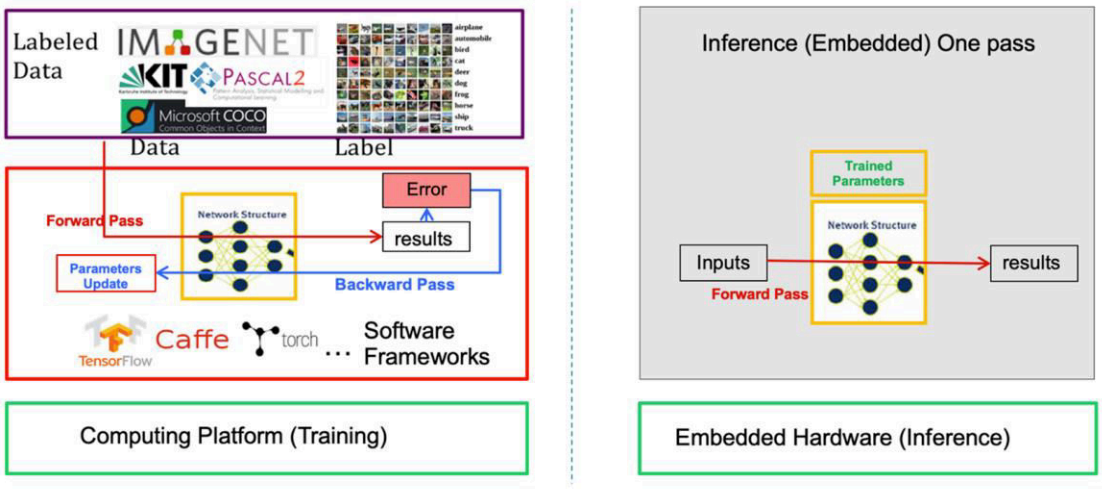
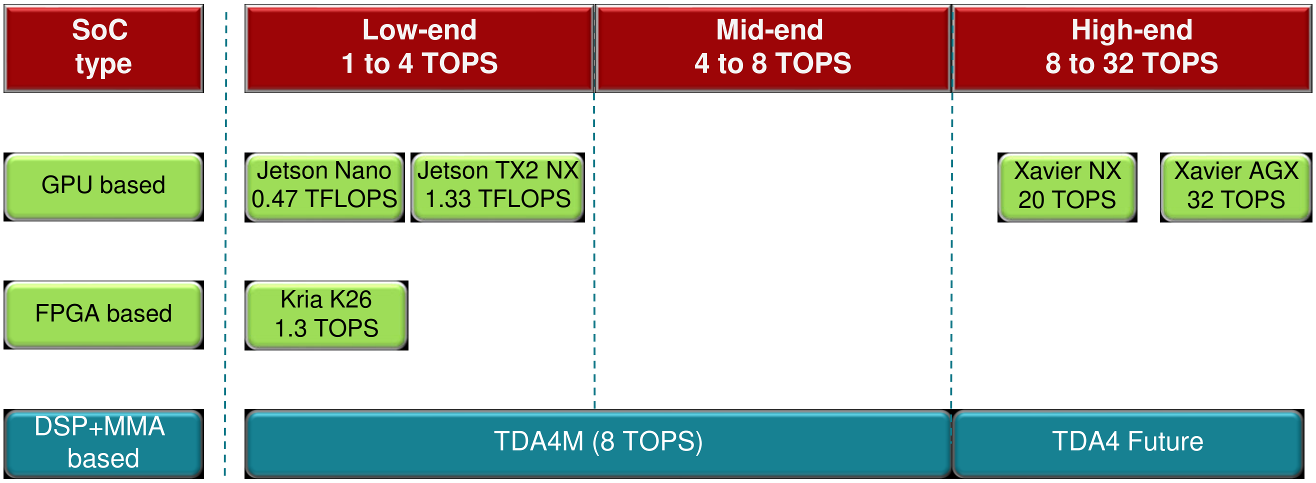


Figure 2-1. Training vs Inference Process

Training a deep learning model typically requires very high TOPS processing engine but for most low- to mid-end edge AI inference applications, TOPS required are in the range of 1 to 32 TOPS. This is the segment that TDA4x processor family is targeting.

2.1 Processors for edge AI: Technology landscape

There are several architectures available for embedded edge AI applications. Figure 2-2 below shows some of the SoC devices based on different architectures – GPU (graphical processing unit), FPGA (field programmable gate array) and other embedded processors [6] [7] [8]. There are many other technologies offering higher than 32 TOPS of deep learning performance but such high performance is typically needed for data center and infrastructure applications. This application note focuses on the edge AI segment so primarily SoC architectures with performance up to 32 TOPS are considered.



SoC type	Low-end 1 to 4 TOPS	Mid-end 4 to 8 TOPS	High-end 8 to 32 TOPS
GPU based	Jetson Nano 0.47 TFLOPS Jetson TX2 NX 1.33 TFLOPS		Xavier NX 20 TOPS Xavier AGX 32 TOPS
FPGA based	Kria K26 1.3 TOPS		
DSP+MMA based	TDA4M (8 TOPS)		TDA4 Future

Figure 2-2. Edge AI SoC landscape in 1 to 32 TOPS range

To measure AI inference performance, TOPS (tera operations per second) can be used as the first indicator. However, actual system performance depends on many other factors influenced by the processor interconnect and data flow architecture. This is because a typical deep learning model has a lot of data movement, and it is not just the computations that affect the inference time but also how the data is handled efficiently. This, in turn, also affects power consumption of the processor and the overall system. So, in this application note, we will look at additional metrics to design an efficient edge AI system.

2.2 Edge AI with TI: Energy-efficient and Practical AI

2.2.1 TDA4VM processor architecture

Using the matrix multiplication accelerator (MMA) as the acceleration for AI functions, the overall TDA4x block diagram is shown in the below Figure 2-3. Based on heterogenous architecture, the TDA4x System on Chip (SoC) optimizes entire platform around easy programming on the multi-core Cortex-A72 microprocessor unit (MPUs) while offloading compute intensive tasks such as deep learning, imaging, vision, video, and graphics processing to the specialized hardware accelerators and programable cores. High throughput and high energy efficiency are enabled by holistic system level integration of these cores using high bandwidth interconnect and smart memory architecture. An optimized system BOM is achieved by advanced integration of the system components.

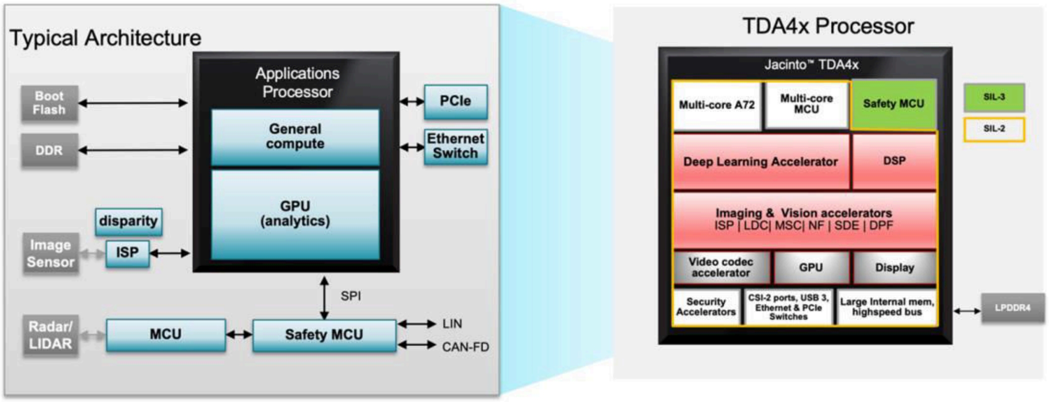


Figure 2-3. Block Diagram

As we discussed in the previous section, TOPS (tera operations per second) are used to measure the deep learning performance comparison. However, actual inference time depends on the efficiency of the system architecture making use of optimum data flow in the system. So, a better performance benchmarking is inference time for a given model at a given input image resolution. If the inference time is lesser, more images can be processed resulting in higher frames per second (FPS). So, FPS divided by TOPS (FPS/TOPS) indicates the deep learning architecture efficiency. Similarly, FPS divided by Watts (FPS/Watt) is a good benchmark for energy efficiency of an embedded processor.

2.2.1.1 Development platform

The TI Edge AI starter kit is a $249 tool based on the TDA4VM SoC offering 8 TOPS of low power, accelerated deep learning capabilities. With comprehensive software using open-source industry standard software, the kit enables developers prototyping smart cameras, edge AI boxes, autonomous machines, and robots easily and quickly.

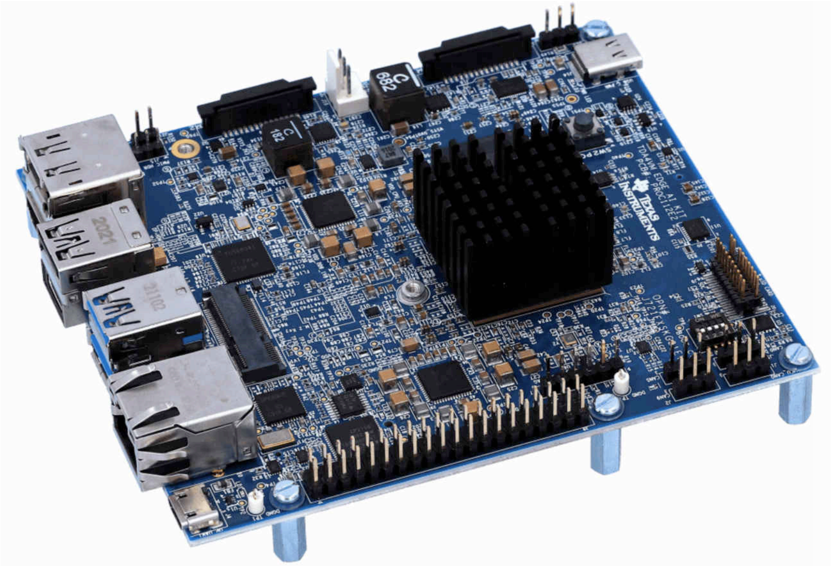


Figure 2-4. $249 Edge AI Starter Kit offering 8 TOPS performance

Figure 2-5 below shows the block diagram of the starter kit. It has many sensor inputs and communications peripherals making it ideal for sensor fusion and vision applications such as image classification, object detection and semantic segmentation.

The board features 2x CSI2 ports compatible with Raspberry PI cameras and production-grade cameras as well as TI's high-speed CSI2 connection allowing connections up to 8 cameras using the 8-port fusion application add-on card.

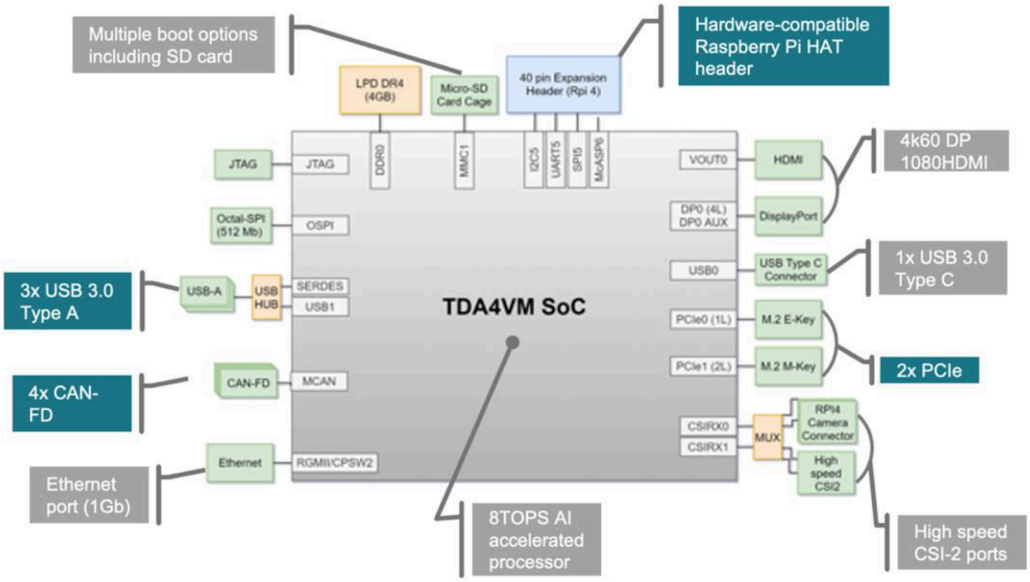


Figure 2-5. Edge AI Starter Kit Block Diagram

Additionally, the board includes 2x USB 3.0 type A ports, 1 Gigabit Ethernet port, a 40 pin Raspberry Pi GPIO header to support expansion boards and for video output, a 4K capable DisplayPort as well as an HDMI port. Raspberry Pi header enables developers to tap into a large eco-system of peripheral boards readily. The board also is designed in small form factor to be useful for prototyping different applications.

2.3 Software programming

TI's Edge AI starter kit come with a comprehensive edge AI software architecture that enables developers to do application development completely in Python or C++ language. There is no need to learn any special language to take advantage of the performance and energy efficiency of the deep learning accelerator with the TDA4x SoC processor.

Figure 2-6 below shows the comprehensive software offering from TI for the edgeAI applications. With strong and robust Linux foundation, developers can program the device with popular open-source frameworks such as, Tensorflow lite, ONNX, and TVM. Figure 2-6 shows the software architecture used for edge AI application development.

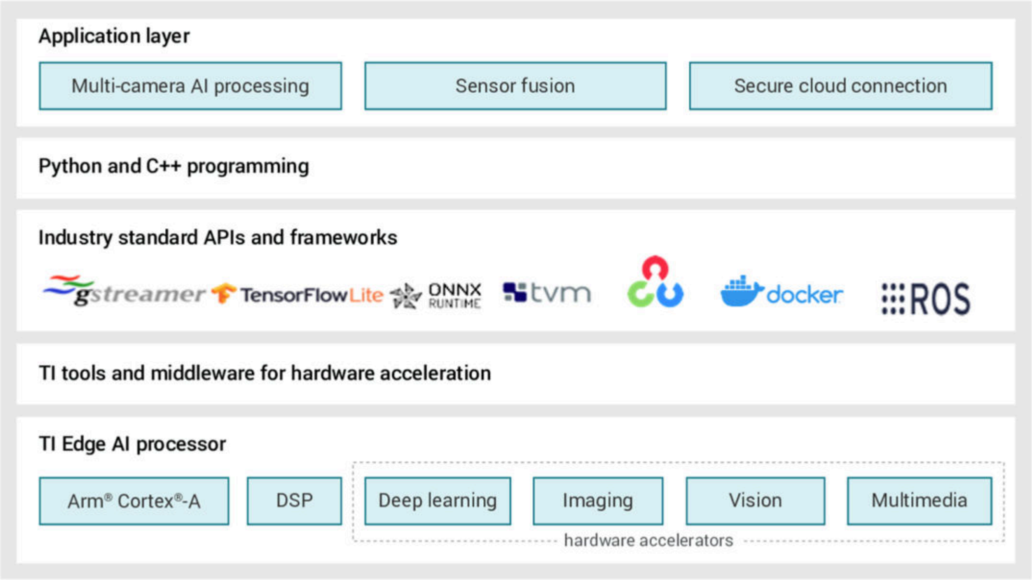


Figure 2-6. Easy to use complete software for easy AI application development

3 Industry standard performance and power benchmarking

While different SoC vendors use different metrics to benchmark deep learning performance and power across multiple devices, there is an initiative in the industry to standardize benchmarking for apples-to-apples comparison. Driven by more than 30 organizations, MLPerf Inference [9] prescribes a set of rules and best practices to ensure comparability across systems with wildly differing architectures. We will primarily use these guidelines for performance and power benchmarking.

3.1 MLPerf models

MLPerf Inference is a benchmark suite for measuring how fast systems can run models in a variety of deployment scenarios. Quoting from the website mlperf.org, the benchmark aims to do: "Fair and useful benchmarks for measuring training and inference performance of ML hardware, software, and services". A major contribution of MLPerf is selection of representative models that permit reproducible measurements. Based on industry consensus, MLPerf Inference comprises models that are mature and have earned community support. MLPerf models are also open source and the software and models for the benchmarks are provided in github repositories in and . Benchmarks are provided for both training and inference of models. Inference includes both cloud scenario as well as edge scenario. In this application note, as we discussed before, we focus on inference of models for edge and mobile benchmarks.

Image classification: As we saw in the introduction, image classification is a commonly used deep learning function for applications that include photo searches, text extraction, and industrial automation, such as object sorting and defect detection. MLPerf uses the ImageNet 2012 data set [10], crop the images to 224x224 in preprocessing, and measure Top-1 accuracy. MLPerf suggests two models: a computationally heavyweight model that is more accurate and a computationally lightweight model that is faster but less accurate. The heavyweight model, ResNet-50 v1.5 [16] is used in this benchmarking and comparison.

Object detection. Object detection is a vision task that determines the coordinates of bounding boxes around objects in an image and then classifies those boxes. Implementations typically use a pretrained image-classifier network as a backbone or feature extractor, then perform regression for localization and bounding-box selection. Object detection is crucial for a multitude of tasks in automotive and robotics, such as detecting hazards and analyzing traffic, and for mobile-retail tasks, such as identifying items in a picture. MLPerf suggests two models: a lightweight model using 300x300 image and a heavyweight model using 1200x1200 image with the COCO data set [11].

Based on this, the two models used in the app note are shown in Table 3-1below.

Note

TI has not officially submitted the results to MLcommons.org yet. These models are used because they represent practical use cases and these models are used by other edge AI SoC vendors.

Table 3-1. TDA4VM performance and power measurements

DL Model	Function	Image size	Data set	Compute requirements per input
ResNet-50	Image Classification	224x224	IMAGENET	8.2 GOPS 25.6 million parameters
SSD MobileNet-V1	Object Detection	300x300	COCO	2.47 GOPS 6.91 million parameters

MLPerf inference standard also defines different scenarios for benchmarking - single-stream, multi-stream, server, and offline. For real-time embedded edge AI systems such as smart cameras, machine vision and robotics, the most relevant scenarios are single-stream and multi-stream use cases involving image and video processing from single and multiple cameras simultaneously. We will be using single-stream use case in the benchmarking.

3.2 Performance and efficiency benchmarking

Performance measurements

In order to run MLPerf Deep Learning Models on TDA4VM starter kit hardware, they need to be converted into a format that is understood by deep learning accelerators in the DAv4M device. TI already has done all the work of converting, optimizing and exporting several models from the original training frameworks in PyTorch, Tensorflow and MxNet into these MMA friendly formats. All these pre-compiled and optimized models are hosted in TI's GitHub repository [22].

The TI Edge AI software development kit (SDK) comes packaged a few pre-imported models as part of the SD card image. For quick evaluation, these are good to run the demos out of box [19].

The starter kit is used for performance benchmarking as it is much smaller and lower cost.

All the benchmarking is done using the Mlperf models and the procedure described in the MLcommons specification. Below table shows FPS, FPS/TOPS, Watts and FPS/Watts measured on the TDA4x EVM.

Table 3-2. TDA4VM performance and efficiency measurements

Model	FPS	FPS/TOPS
ResNet-50	162	20.29
SSD MobileNet-V1	385	48.08

The performance numbers are published on ti.com/edgeai. We can use these numbers to compare with other leading GPU and FPGA based architectures.

3.3 Comparison against other SoC Architectures

MLcommons page [9] has several published results for these models by different vendors for SoCs using three kinds of processors – GPU based, FPGA based and other embedded processors. We will compare TDA4VM results against these other architectures.

3.3.1 Benchmarking against GPU-based architectures

Figure 3-1 plots the above TDA4VM results against a GPU based SoC numbers from the MLcommons page [19] for the single-stream mode use case. As we discussed before, FPS/TOPS is a better indicator of energy efficiency.

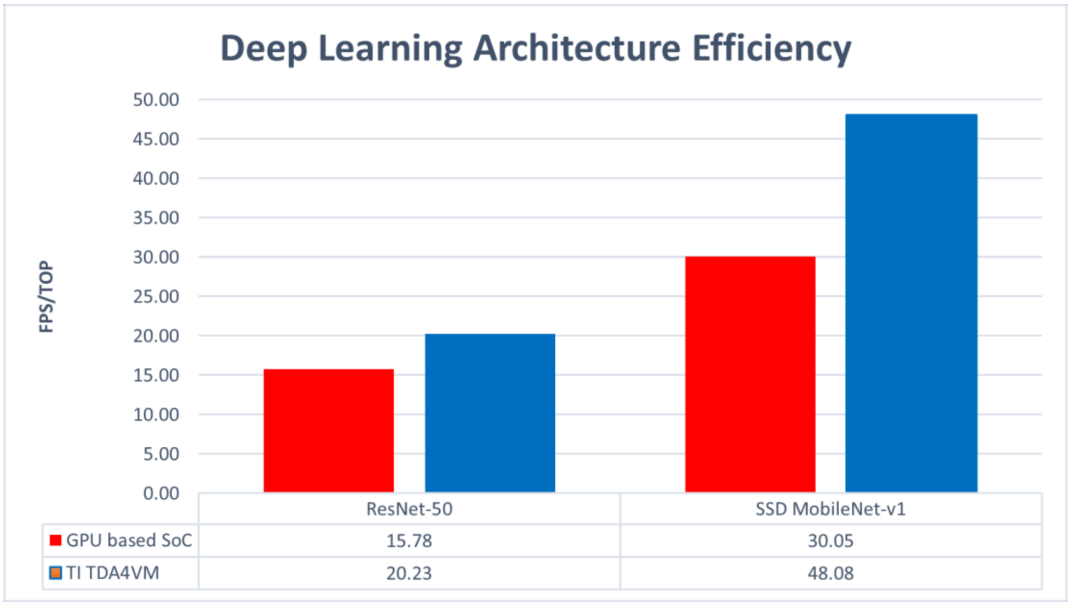


Figure 3-1. GPU based SoC vs TDA4VM: FPS/TOPS comparison

We can see from the comparison that TDA4VM is up to 60% better in terms of FPS/TOPS efficiency. What this means is that 60% less TOPS are needed to run equivalent deep learning functions.

3.3.2 Benchmarking against FPGA based SoCs

The Xilinx Kria K26, is based on the Zynq® UltraScale+™ MPSoC architecture with various deep learning processing unit (DPU) configurations. Xilinx has an application note [8] publishing the FPS/Watt numbers for the MLPerf models chosen in our study. However, these results are not yet published on Mlperf website. So, just comparing the 1.2 TOPS vs 8 TOPS of performance that TDA4VM platform brings, developers will have more than six times the performance boost to design more sophisticated AI vision tasks.

3.3.3 Summary of competitive benchmarking

For deep learning performance and power benchmarking, it is essential to perform apples-to-apples comparison using industry standard machine learning benchmarks. MLPerf Inference [9] is an industry recognized standard for this purpose. Using these guidelines, we can illustrate the benefits of TDA4M edge AI processor resulting in greener and more efficient edge AI systems.

Different metrics used are:

1. TOPS (Tera operations per second)
2. FPS (Frames per second)
3. FPS/TOPS: Frames per second normalized to 1 TOPS

TDA4VM platform's advantage is summarized both from performance and power aspects as below:

- GPU based architectures: TDA4VM offers up to 60% better performance efficiency measured as FPS/TOPS and 60% better power efficiency measured as FPS/Watts.
- FPGA based architectures: TDA4VM offers more than six times performance boost in TOPS metric.

4 Conclusion

AI inferencing technology is a key enabler for broader deployment of edge AI devices across from home to factories with the potential to impact every aspect of our lives. Performance and power benchmarking different edge AI SoC processors is a complex task and having an apples-to-apples comparison is critical for developers to pick up the right device for applications that are cost, size, and power sensitive. In this application note, we discussed industry standard performance and power benchmarking used to compare the TDA4x architecture with GPU-based and FPGA-based architectures.

TI Edge AI tools are bringing state-of-the-art process and technology leadership with the TDA4VM SoC devices. Developers can now achieve **more than 60% better energy efficiency** compared to GPU based devices resulting in greener edge devices. Edge AI devices can also include greater levels of sophistication with more than **six times higher performance** compared to FPGA based solutions.

TDA4x processor family also comes with easy to use, no-cost to low-cost development platforms making it easier for developers to innovate with AI without any prior experience. Developers can also take advantage of an extensive collection of embedded AI projects from TI and also from our third-party ecosystem for faster time to market [20].

Revision History

NOTE: Page numbers for previous revisions may differ from page numbers in the current version.

DATE	REVISION	NOTES
August 2022	*	Initial Release

5 References

1. United Nations: Population Division
2. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way
3. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.
4. Alfredo Canziani, Thomas Molnar, Lukasz Burzawa, Dawood Sheik, Abhishek Chaurasia, Eugenio Culurciello ,"Analysis of deep neural networks".
5. K. Lee, V. Rao, and W. C. Arnold, "Accelerating facebook's infrastructure with application-specific hardware," Facebook, 3 2019.
6. TDAVM: Dual Arm® Cortex®-A72, C7x DSP, and deep learning, vision and multimedia accelerators
7. Jetson Modules
8. K26 SOM: Ideal platform for Vision AT at the edge https://www.xilinx.com/support/documentation/white_papers/wp529-som-benchmarks.pdf
9. MLcommons benchmarking
10. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.
11. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan,P. Doll´ar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014.
12. TensorFlow
13. Onnx Runtime
14. Apache TVM
15. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016.
16. MLPerf, "ResNet in TensorFlow," 2019. M. Naumov, D. Mudigere, H. M. Shi, J. Huang, N. Sundaraman
17. Google mobilenet on embedded systems
18. MLPerf Machine Learning Benchmark Suite - Benchmark Results
19. TDA4VM processor starter kit for Edge AI vision systems
20. Enabling optimized edge AI inference performance, system power and cost
21. MLCommons Github
22. TI's collection of optimized deep learning models
23. Development tools for deep learning runtime
24. TI edgeAI benchmarking repository
25. Linux SDK for edge AI applications on TDA4VM Jacinto processors

IMPORTANT NOTICE AND DISCLAIMER

TI PROVIDES TECHNICAL AND RELIABILITY DATA (INCLUDING DATA SHEETS), DESIGN RESOURCES (INCLUDING REFERENCE DESIGNS), APPLICATION OR OTHER DESIGN ADVICE, WEB TOOLS, SAFETY INFORMATION, AND OTHER RESOURCES “AS IS” AND WITH ALL FAULTS, AND DISCLAIMS ALL WARRANTIES, EXPRESS AND IMPLIED, INCLUDING WITHOUT LIMITATION ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT OF THIRD PARTY INTELLECTUAL PROPERTY RIGHTS.

These resources are intended for skilled developers designing with TI products. You are solely responsible for (1) selecting the appropriate TI products for your application, (2) designing, validating and testing your application, and (3) ensuring your application meets applicable standards, and any other safety, security, regulatory or other requirements.

These resources are subject to change without notice. TI grants you permission to use these resources only for development of an application that uses the TI products described in the resource. Other reproduction and display of these resources is prohibited. No license is granted to any other TI intellectual property right or to any third party intellectual property right. TI disclaims responsibility for, and you will fully indemnify TI and its representatives against, any claims, damages, costs, losses, and liabilities arising out of your use of these resources.

TI’s products are provided subject to TI’s Terms of Sale or other applicable terms available either on ti.com or provided in conjunction with such TI products. TI’s provision of these resources does not expand or otherwise alter TI’s applicable warranties or warranty disclaimers for TI products.

TI objects to and rejects any additional or different terms you may have proposed.

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265