Benchmarking AI Inference at the Edge:

Measuring Performance and Efficiency for Real-World Deployments

brainchip
Essential AI

Contents

Introduction

Current industry benchmarks measuring edge AI inference performance tend to overemphasize isolated TOPS metrics that don’t effectively quantify results for real-world, power-conscious deployments. To benchmark AI inference more accurately for specific edge use cases, we recommend holistically focusing on performance and efficiency.

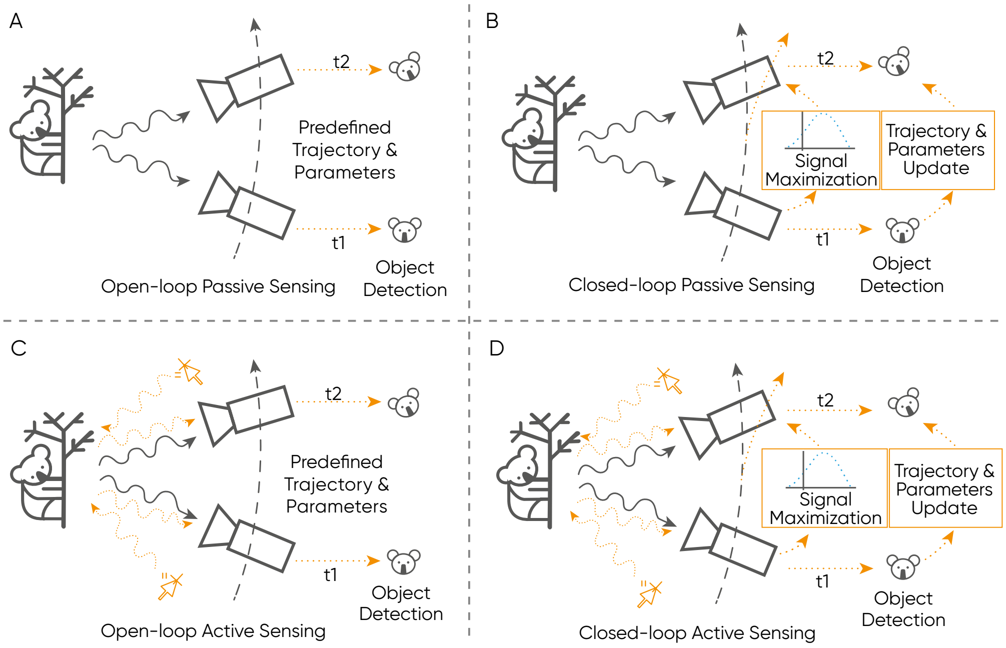
In this paper, we highlight the importance of balancing these important criteria while mapping tangible benefits to three primary verticals: automotive, smart homes, and Industry 4.0. We also reference key BrainChip AI inference performance and efficiency benchmark results – and explain how these numbers were achieved with neuromorphic techniques that significantly reduce latency and power while increasing throughput.

In addition, we discuss potentially formulating a new set of benchmarks that focus on latency, power, and (on-chip) in-memory computation. When fully defined, these benchmarks will allow system designers to further optimize low-power AI inference for complex, multi-modal edge environments.

Chapter 1:
The limits of conventional AI performance benchmarks

As artificial intelligence (AI) hardware and machine learning (ML) algorithms evolved, the semiconductor industry developed a new generation of standardized benchmarks such as MLPerf™ to measure the performance of AI-specific workloads and inference capabilities. These include IBM DVS128 Gesture Dataset, ImageNet, and GLUE. Although some benchmarking organizations continue to introduce new fields and subcategories to measure AI inference at the edge, these additions are frequently limited by an overemphasis on isolated TOPS and do not effectively quantify results for real-world use cases where power consumption is a primary concern.

As a recent NeurIPS paper notes, a small collection of influential benchmarks is currently "valorized" across different AI subfields. However, these benchmarks do not accurately capture edge AI inference subsets or effectively gauge the efficiency of certain neuromorphic techniques.


A
t2
Predefined Trajectory & Parameters
t1
Open-loop Passive Sensing
Object Detection
B
t2
Signal Maximization
Trajectory & Parameters Update
t1
Closed-loop Passive Sensing
Object Detection
C
t2
Predefined Trajectory & Parameters
t1
Open-loop Active Sensing
Object Detection
D
t2
Signal Maximization
Trajectory & Parameters Update
t1
Closed-loop Active Sensing
Object Detection

Open-loop versus closed-loop sensing (Neuromorphic Engineering Needs Closed-Loop Benchmarks)

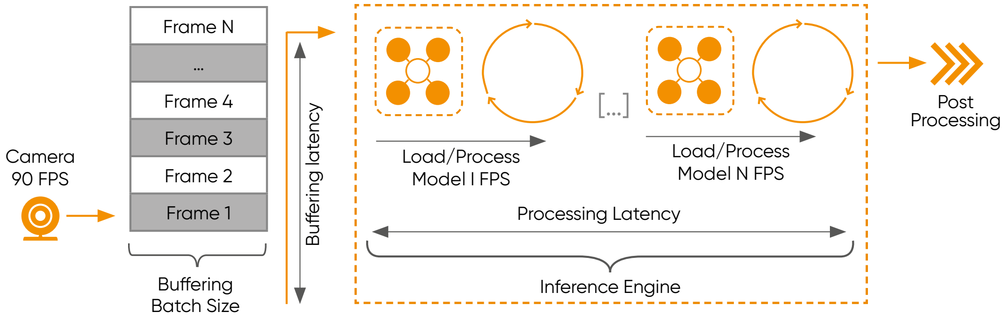
From our perspective, edge AI benchmarks need to move beyond their desktop and data center origins by adopting new criteria that address the unique, power-conscious requirements of cloud-free, local AI inference. As well, edge AI inference benchmarks should be application-based, with dedicated fields emulating multiple sensor inputs to reflect closed-loop, real-world use cases. This comprehensive approach to benchmarking offers a more top-down, system-level view of performance and efficiency.

Chapter 2:
Balancing performance and power at the edge

To accurately measure AI inference capabilities for multi-sensor, edge-specific use cases in a power and thermally constrained environment, benchmarks should holistically assess performance and efficiency. These benchmarks – which should include open-loop and closed-loop datasets – will enable system designers to effectively measure raw performance metrics such as throughput and power consumption while gauging efficacy for real-world tasks. With this data, companies can more precisely calibrate AI inference performance and efficiency for edge-specific verticals such as automotive, smart homes, and Industry 4.0.

Automotive

Consumers expect new vehicles to feature advanced assisted driver assistance systems (ADAS) enabled by LiDAR, radar, and computer vision; as well as highly personalized and responsive in-cabin systems that respond to voice commands, gestures, and facial expressions.



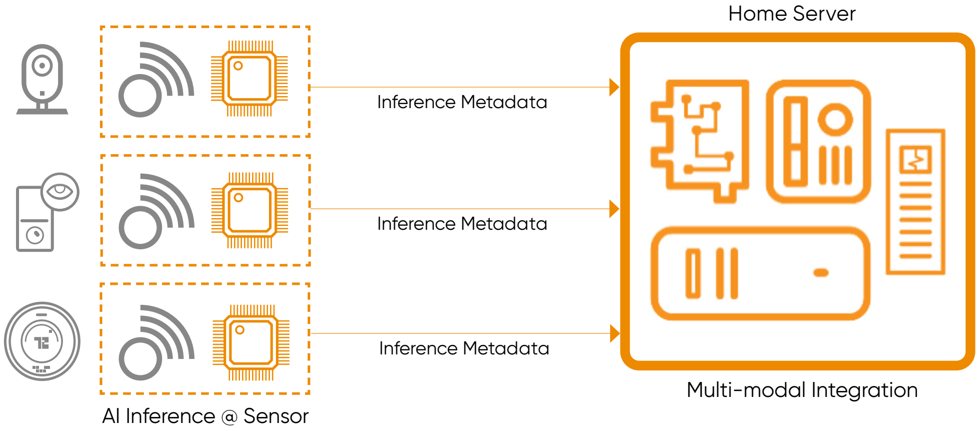
Pipeline considerations: AI inference at the automotive edge

Automotive manufacturers must therefore design edge AI systems that support massive amounts of data throughput from multiple sensors, ensure inference accuracy while minimizing latency, and keep power consumption within a reasonable envelope. The increasing popularity of electric vehicles with finite battery limitations only reinforces the importance of energy-efficient computation.

For automotive use cases, balanced inference benchmarks can offer an especially accurate – and comprehensive – assessment of AI performance in complex, dynamic environments by holistically measuring efficiency and power draw for applications such as keyword spotting and image detection. These benchmarks will help automotive manufacturers implement more responsive in-cabin systems, as well as computer vision and LiDAR systems that detect vehicles, pedestrians, bicyclists, street signs, and objects with incredibly high levels of precision.

Smart Homes

Smart home devices such as personal assistants, video doorbells, and thermostats typically require a minimal physical footprint. These devices also need to quickly respond to voice commands, detect anomalous behavior, and determine next steps by analyzing and interpreting data from multiple sensors.



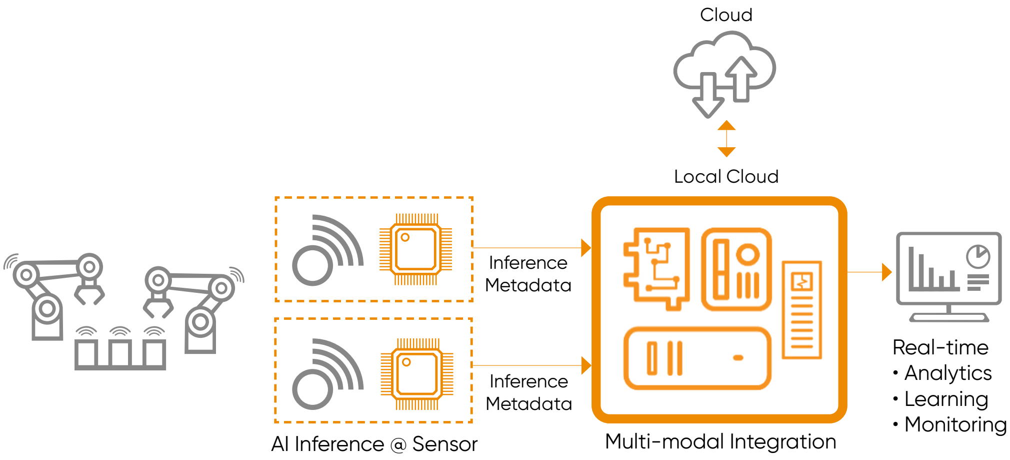
Designing smarter edge consumer devices for multi-modal applications

To enable a new generation of smarter, more proactive consumer devices, manufacturers must design AI subsystems that bolster cloud-free inference capabilities while minimizing processor die size. Enabling AI inference workloads to run efficiently at the edge will help smart home manufacturers reduce overall system footprint, price, and power consumption.

For smart home devices, inference benchmarks should focus on measuring performance, accuracy, and efficiency for tasks such as keyword spotting, object detection, and visual wake words.

Industry 4.0

Factories, warehouses, and loading docks increasingly rely on advanced AI-powered robots to complete challenging physical tasks with minimal human intervention. Equipped with real-time learning capabilities, these robots often leverage sophisticated sensors to see, hear, smell, touch, and even taste.



Balancing efficiency and power at the edge for Industry 4.0

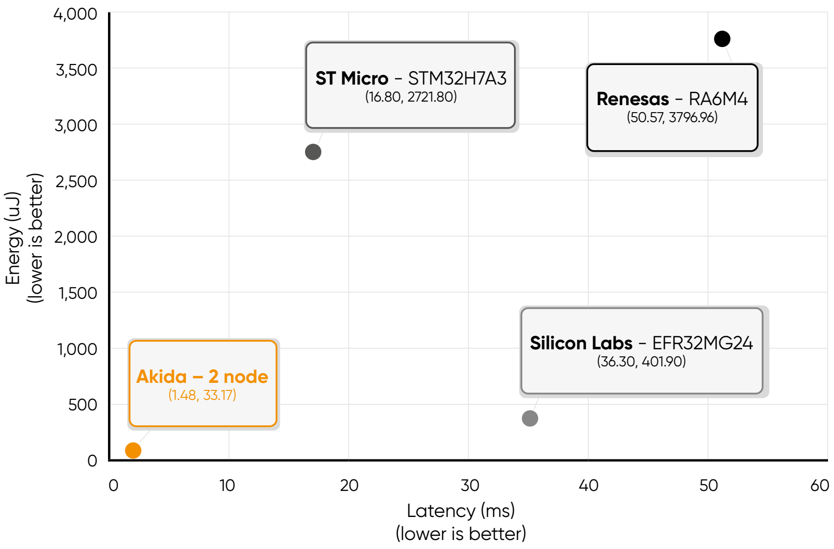
Targeted Industry 4.0 inference benchmarks focused on balancing efficiency and power will enable system designers to architect a new generation of energy-efficient robots that optimally process data-heavy input from multiple sensors. These robots will intelligently respond to dynamic variables, conditions, and instructions by immediately detecting and recognizing new objects.

Chapter 3:
Comparing performance and energy efficiency with tinyML

As highlighted in previous chapters, current AI inference performance benchmarks do not accurately gauge efficacy for real-world edge deployments. However, these benchmarks do effectively demonstrate relative performance. With that, let us look at BrainChip's Akida performance relative to MLPerf published benchmark data. The Akida event-based AI processor is primarily available as IP to be integrated into SoC designs, but the following benchmarks have been run on the AKD1000 silicon platform, which is commercially available for development today.

The benchmark data below is divided into two sets of graphs, each comparing BrainChip's Akida processor using three representative applications. These applications – keyword spotting, object detection, and anomaly detection – play critical roles in enabling the edge-specific verticals highlighted in the previous chapter. The first set of graphs are based on publicly available MLPerf datasets from MLCommons that benchmark Akida and leading MCU vendor offerings.*

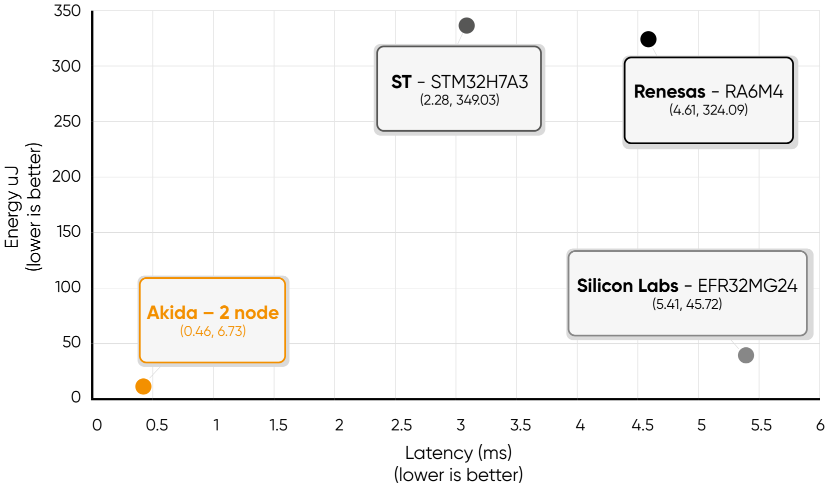
As the first chart illustrates, Akida-based performance delivers extremely low latency using just a fraction of the energy consumed by conventional MCUs.



Latency versus energy for keyword spotting benchmark (DS_CNN @49x10x1)*

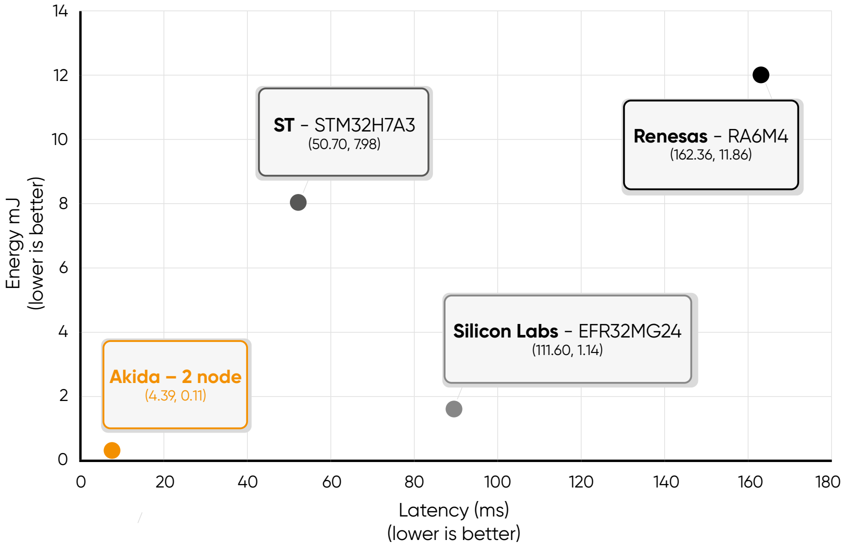
*All data for the MCUs (ST, Renesas, Silicon Labs), DLA (NVIDIA), and TPU (Google) have been taken from published data at the MLCommons tinyML site (https://mlcommons.org/en/inference-tiny-07/). Note: AKD1000 scores are unverified results, have not been through an MLPerf review, and may use measurement methodologies and/or workload implementations that are inconsistent with the MLPerf specification for verified results.

The contrast is even more pronounced when benchmarking Akida against conventional MCUs for visual wake words – a popular edge application that demands complementary levels of efficiency and performance in power-conscious and thermally constrained environments.



Latency versus energy for anomaly detection benchmark in MCU-class devices*

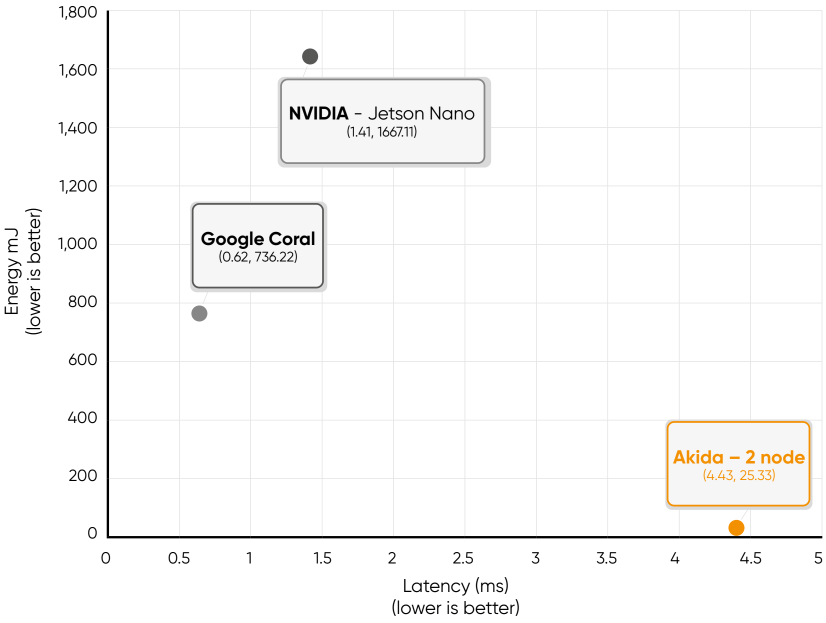
Similarly, the anomaly detection benchmark shows a substantial gain in latency and energy versus the MCU solutions.



Latency versus energy for person/no-person visual wake word benchmark in MCU-class devices*

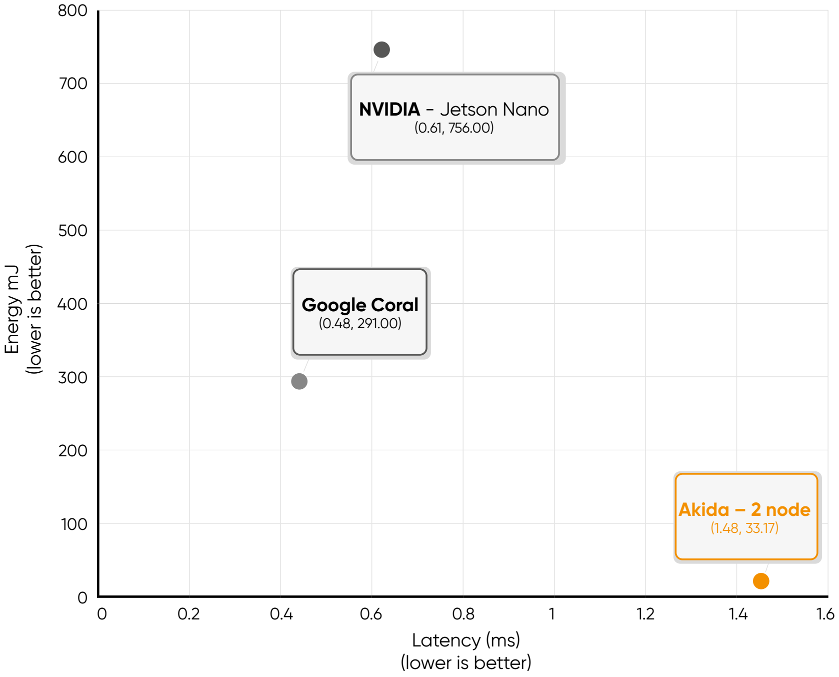
The second set of graphs are also based on publicly available MLPerf datasets from MLCommons that benchmark Akida and leading DLA and TPU vendor offerings. These compare Akida with two higher end edge AI processors: NVIDIA Jetson Nano (a deep learning accelerator) and Google Coral (a tensor processing unit). Even at a significantly lower frequency, Akida matches or outperforms both processors while consuming significantly less power.

This performance and efficiency data is based on a two-node configuration – which we compare here to ensure consistency with earlier MCU comparisons. It should be noted that higher node configurations further reduce latency – and are potentially more efficient as they enable faster compute. However, even with a two-node configuration, Akida's latency fits well within the 5ms target, and the energy consumed per frame by Akida is only a fraction of the power drawn by larger processors.



Latency versus energy for person/no-person visual wake word benchmark*

Similarly, Akida offers notable efficiency benefits for keyword spotting applications, with latency measuring within the 1.5 ms target.



Latency versus energy for keyword spotting benchmark*

Chapter 4:
Maximizing edge AI inference performance and efficiency with Akida

To put these benefits into context, we have consolidated results from the previous section of TinyML benchmarks (a two-node Akida configuration versus published MCUs) into a single, energy efficiency view. To do this, we use the measured data on latency and energy per inference on each device to calculate the following:

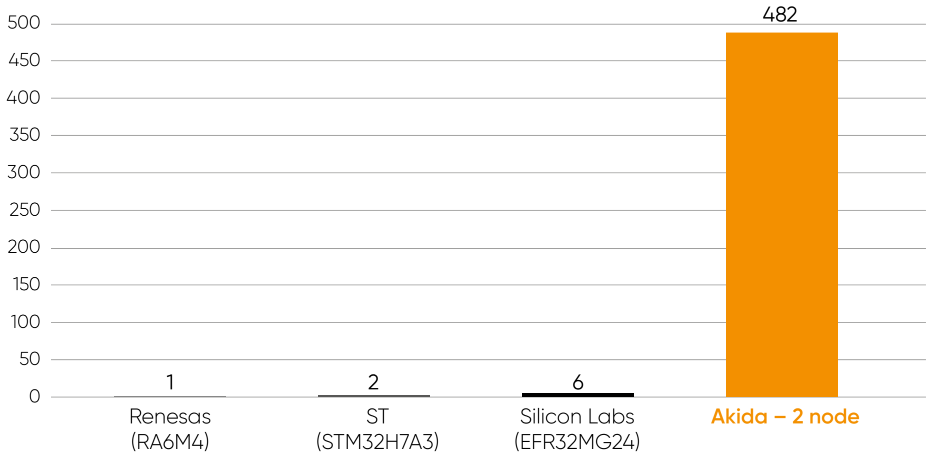
$$\text{Performance}_{\text{device}} = \frac{1}{\text{Latency}_{\text{device}}}$$

$$\text{Efficiency}_{\text{device}} = \frac{\text{Performance}_{\text{device}}}{\text{Energy per inference}_{\text{device}}}$$

Although this equation is based on the latency and energy for tinyML benchmarks, there are additional factors such as model size and load times that could provide a more precise efficiency metric. To compare these results against other published datapoints, we use a relative performance efficiency metric which is normalized against the lowest efficiency device.

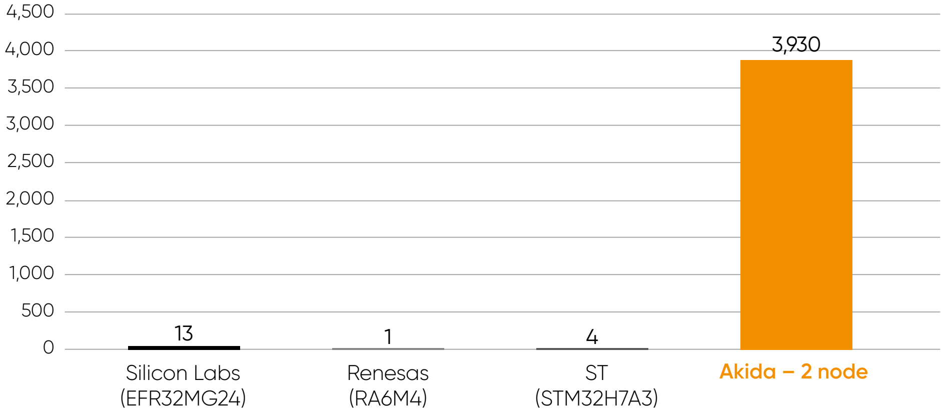
$$\text{Relative performance efficiency} = \frac{\text{Efficiency}_{\text{device}}}{\text{Efficiency}_{\text{lowest efficiency device}}}$$

This view highlights the work done per unit energy by each device.



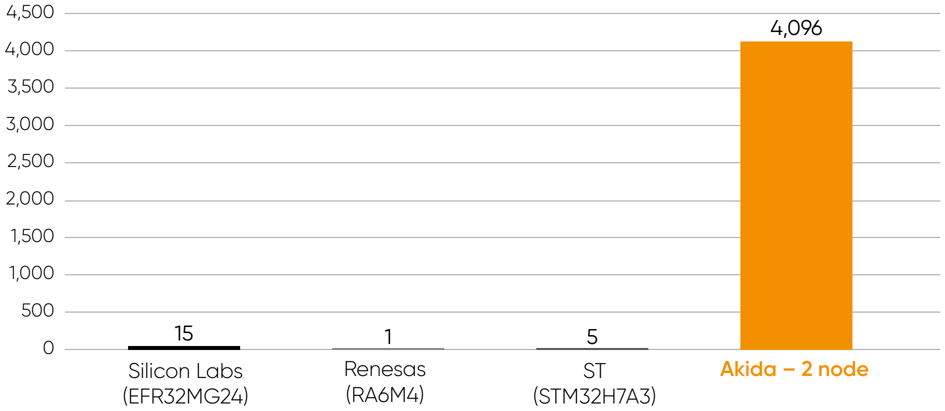
Relative performance efficiency on anomaly detection benchmark

The relative efficiency of the anomaly detection benchmark highlights the benefits of an efficient AI engine over traditional MCUs even for small workloads.



Relative performance efficiency on keyword spotting benchmark

With a slightly larger workload, this efficiency gap significantly increases.



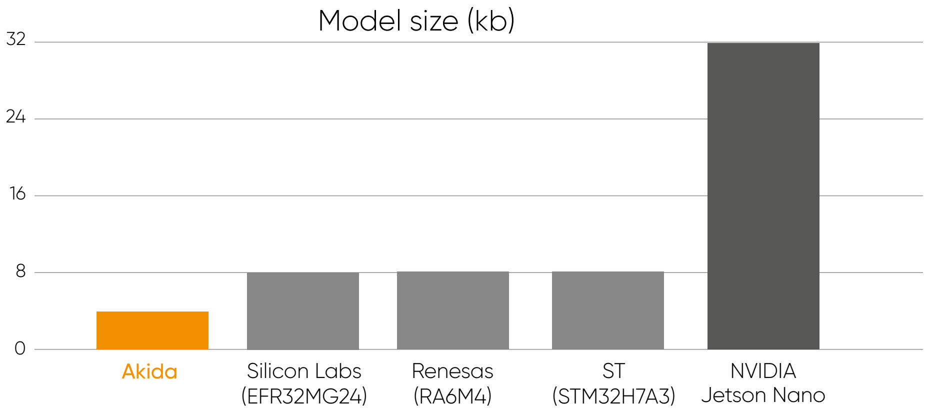
Relative performance efficiency on visual wake word benchmark

Additionally, as we discuss in the next chapter, future devices will not only be running larger AI workloads but likely run multiple networks simultaneously. This is where the efficiency advantage truly unlocks more intelligence in low power devices.

Chapter 5:
Optimizing load times with lighter AI inference models

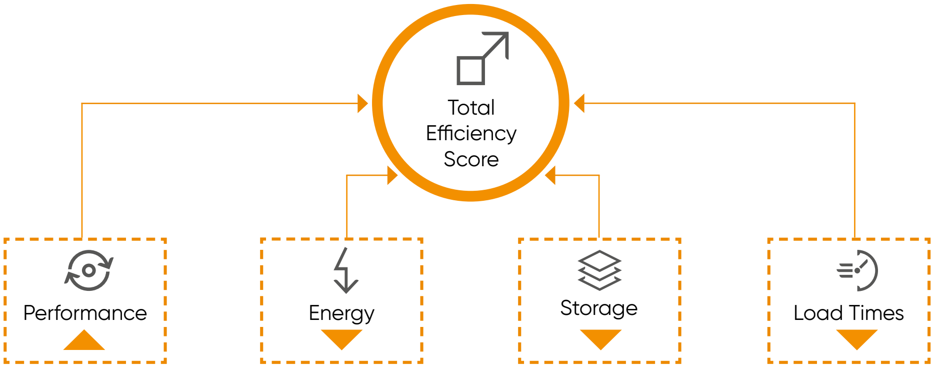
While frames per second (FPS) and energy consumption are some of the most measured metrics, model size is another crucial factor that affects overall efficiency. By leveraging lighter AI inference models, chips with limited on-die memory can run larger workloads. Moreover, decreasing the size of AI inference models helps reduce data movement, allowing systems to efficiently perform parallel computations. Although lowering system bandwidth requirements definitively improves system efficiency and yields increased FPS, future studies analyzing further impacts will likely provide more granular benchmarks.

The graph below illustrates how quantization of weights and activation results in lighter models and significantly improves efficiency.



Lighter models = faster load time & performance

Lastly, it should be noted that reducing model size and implementing advanced optimization techniques significantly improve load time. This is particularly important for neural processors running multiple networks.



Holistically assessing performance and efficiency

Indeed, load time differences are often quite pronounced – ranging from milliseconds to microseconds. Put simply, load times play a major role in supporting efficient multi-sensor, multi-network real-world AI inference deployments.

Conclusion

To enable real-world, power-conscious deployments, benchmarks measuring AI inference capabilities should focus on holistically gauging performance and efficiency for multi-sensor, edge-specific use cases. In this paper, we have proposed a few suggestions to measure efficiency gains more comprehensively and highlighted additional criteria for consideration. However, there is more to be done. Ultimately, we foresee a new set of inference performance benchmarks that will measure efficiency with precision by focusing on three primary criteria: latency, power, and (on-chip) in-memory computation. This triumvirate forms the foundation of AI processing at the edge.

Specific strategies for formulating and implementing the next generation of benchmarks will undoubtedly vary. That is why it is important to collaboratively explore, establish, and promote new methods of measuring inference performance and efficiency for conventional and neuromorphic silicon.

*All data for the MCUs (ST, Renesas, Silicon Labs), DLA (NVIDIA), and TPU (Google) have been taken from published data at the MLCommons tinyML site (https://mlcommons.org/en/inference-tiny-07/). Note: AKD1000 scores are unverified results, have not been through an MLPerf review, and may use measurement methodologies and/or workload implementations that are inconsistent with the MLPerf specification for verified results.