PRODUCT BRIEF

SUPERMICRO SYS-E403 SERVER EXCELS FOR AI AT THE EDGE

Supermicro Edge System Ideal for AI Inferencing

TABLE OF CONTENTS

Executive Summary

Recently, the demand for real-time AI-Inference applications has rapidly increased. The power of Artificial intelligence (AI) is not only being used in large data centers but also at the Edge. Edge computing differs from a climate-controlled data center, often deployed in environments where the physical conditions and network connectivity may not be stable. Moreover, the Edge AI application usually requests real-time inference with low latencies, for instance, self-driving cars, real-time object detection, and AI in the manufacturing production line. Therefore, it is essential to have a high-capacity system that can adapt to the specific scenario and the application in many industries. Supermicro's Edge system, the SYS-E403-12P-FN2T [1], can handle many AI at-the-Edge applications.

This product brief discusses the AI inference capacity of the Supermicro SYS-E403-12P-FN2T. Supermicro has verified the system's ability by comparing the results when running the MLPerf inference [Table 4] and OpenVino benchmark [Table 8]. In addition, Supermicro investigated the performance of installing different GPUs in the SYS-E403-12P-FN2T. As a result, the system has been certified for the video-analytics test of Intel® Edge Software Device Qualification (Intel® ESDQ) [11]. Furthermore, the SYS-E403-12P-FN2T with a single NVIDIA A30 has been fully certified with NVIDIA's latest NCS 2.6 [3].

Supermicro SuperServer SYS-E403-12P-FN2T

The SuperServer SYS-E403-12P-FN2T is a short-depth wall-mount edge workhorse server with rich storage and networking options and support for accelerator and GPU technologies needed for AI/ML applications. It can be deployed in various environments and can support a comprehensive set of AI services with the appropriate GPU, including Industrial Automation, Retail, and Smart Cities.

System Configuration and AI/ML Software Stack

The three PCI-E slots allow the SYS-E403-12P-FN2T to have up to three computing accelerators. As a result, the system is compatible with several NVIDIA AI computing accelerators. Furthermore, the compact hardware design and AI computation ability allow the SYS-E403-12P-FN2T to adapt to various AI applications and scenarios at the Edge. In this document, the experiments include the test on four different NVIDIA GPUs to evaluate the performance of the SYS-E403-12P-FN2T. Also, the SYS-E403-12P-FN2T with a single NVIDIA A30 has been certified with NVIDIA's latest NCS 2.6. Please check Table 1 below for more detailed information on the exact test system specifications.

Key application areas for the SYS-E403 include:

- AI Inference and DL/ML applications on the Edge
- Multi-Access Edge Computing (MEC)
- Universal Customer Premise Equipment (uCPE)
- Network Function Virtualization (NFV)
- Industrial Automation, Retail, Smart Medical Expert Systems

Figure 1 - AI/ML Software Stack

Configuration Item	Description
System SKU	SYS-E403-12P-FN2T
CPU	Intel® Xeon® Gold 6338 CPU @ 2.00GHz
GPU	3x PCI-E 4.0 x16 slot with GPU/Accelerator Support
Memory	8 DIMM slots
Network	2x 10GbE BaseT port(s)

CUDA	11.6
OS	Ubuntu 20.04.4
Docker Engine	20.10.8
NVIDIA Driver	510.47.03
NVIDIA Docker	2.6.0

Table 1 - Test System Configuration

MLPerf Inference Benchmarks

MLPerf-Inference presents a benchmarking method for evaluating and quantifying ML "inference" systems in diversified conditions and scenarios (e.g., Hardware, Software, Scenario, DL/ML model,…). The benchmarks conducted in this paper were executed using MLPerf Inference version 2.0. In addition, five scenarios and seven benchmarks were defined to enable the representative testing of a wide variety of inference platforms and use cases.

MLPerf Inference engine uses the Edge system's Offline, SingleStream, and MultiStream test scenarios. We tested four different NVIDIA GPU products: the NVIDIA A2, T4, A10, and A30. In addition, Supermicro tested the NVIDIA A2, T4, and A10 GPUs in configurations. For the NVIDIA A30, only a single GPU configuration was tested due to the design of the chassis. Please refer to Table. 2 below for more information.

Index	GPU Type	Quantity
1	NVIDIA A2	2
2	NVIDIA T4	2
3	NVIDIA A10	2
4	NVIDIA A30	1

Table 2 - Experimental GPU Configurations

The benchmark scenarios were tested on the system.

- **Offline** – Represents the system processing all the data in a single query. The metric of Offline is throughput, which is the number of samples that the system can process per second—used in, for example, Photo Categorization.

- **SingleStream** – Represents the system that processes the query as soon as it completes the previous query. The metric is the 90th percentile measured latency. Lower latencies mean the system performs better. They are used in, for example, typing auto-complete and real-time AR.

- **MultiStream** – This represents the application with a stream of queries, but each query comprises multiple inferences, reflecting a variety of industrial automation and remote-sensing tasks [see Figure 2 below].

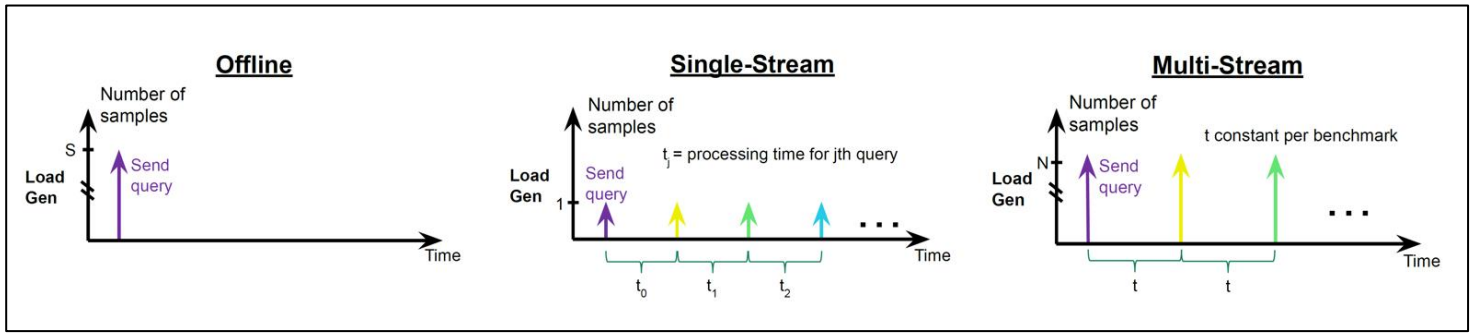

Figure 2 - The Scenarios of an Edge System (2)

Experiment

To investigate the performance of AI inference on SYS-E403-12P-FN2T, we designed experiments that measure the system performance with various GPUs by introducing MLPerf Inference, which provides a standardized and reliable AI evaluation. Furthermore, this method allows us to analyze the efficiency of equipping different GPUs on the SYS-E403-12P-FN2T to better understand the entire system's characteristics.

We picked two representative models as the primary test items to simplify the experiments.

1. ResNet50 [4] is one of the fundamental models in computer vision. Numerous state-of-the-art models use ResNet as their backbone network.

2. BERT [5] is one of the powerful models that can extend to many applications in Natural Language Processing (NLP).

Furthermore, object detection is one of the popular tasks of AI Edge systems. SSD [6] is the essential object detection model that equips various backbone networks to adapt to different scenarios. With ResNet34 as the backbone, SSD-ResNet34 achieves higher accuracy than SSD-MobileNet. MobileNet is a lightweight model that focuses on increasing the throughput. Therefore, it is affordable to deploy MobileNet on Edge systems and embedding devices. Also, the experiment includes RNNT [7], the speech recognition task that processes human speech into a written text format.

To ensure the experimental environment is the same. The tests have been executed inside a Docker container. Except for the GPU, all the software and hardware configurations are identical. The SYS-E403-12P-F2NT is a highly flexible Edge server that supports many different NVIDIA GPUs. Table 3 shows the results of running the MLPerf Inference v2.0 benchmark on various configurations.

MLPerf Inference v2.0														
		ResNet50			BERT		RNNT		SSD-MobileNet			SSD-ResNet34		
		Offline	SingleStream	MultiStream	Offline	SingleStream	Offline	SingleStream	Offline	SingleStream	MultiStream	Offline	SingleStream	MultiStream
GPU Type	1xA2	v	v	v	v	v	v	v	v	v	v	v	v	v
	1xT4	v	v	v	v	v	v	v	v	v	v	v	v	v
	1xA10	v	v	v	v	v	v	v	v	v	v	v	v	v
	1xA30	v	v	v	v	v	v	v	v	v	v	v	v	v
	2xA2	v	v	v	v	v	v	v	v	v	v	v	v	v
	2xT4	v	v	v	v	v	v	v	v	v	v	v	v	v
	2xA10	v	v	v	v	v	v	v	v	v	v	v	v	v

Table 3 - Experiment list

Results

Table 4 lists all of the MLPerf Inferencing 2.0 results. For the **Offline** scenario, the higher the number, the better. On the other hand, for **streaming** tests, the lower the number, the better. One thing that can be easily observed is that – in most cases – the performance of the system increases with a higher-performance GPU. In addition, the performance also increases when the number of GPUs increases. Table 4 below includes benchmark results for all configurations and scenarios tested.

MLPerf Inference v2.0														
		ResNet50			BERT		RNNT		SSD-MobileNet			SSD-ResNet34		
		Offline	SingleStream	MultiStream	Offline	SingleStream	Offline	SingleStream	Offline	SingleStream	MultiStream	Offline	SingleStream	MultiStream
GPU Type	1xA2	3061.6	0.70	3.13	251.5	8.78	1024.6	102.25	4940.9	0.44	2.03	73.4	14.33	112.03
	1xT4	5558.6	0.83	2.12	394.9	10.15	1267.0	72.57	6758.3	0.46	1.43	131.8	8.46	64.69
	1xA10	13039.9	0.45	0.94	1050.8	2.55	4569.8	34.07	19512.0	0.30	0.71	307.8	3.73	26.97
	1xA30	18364.0	0.49	0.86	1701.3	2.22	6711.0	23.49	26001.2	0.30	0.65	484.6	2.78	18.49
	2xA2	6152.7	0.74	3.16	509.3	8.37	2093.5	102.18	9958.2	0.47	2.07	148.0	14.26	110.33
	2xT4	11197.7	0.85	1.88	798.2	8.12	2513.4	71.94	13217.6	0.49	1.34	254.1	6.83	52.65
	2xA10	25904.6	0.48	0.90	2140.3	2.33	9055.4	33.40	37907.1	0.32	0.72	606.6	3.28	23.50

**Metric: Offline: Throughput (sample/second) | SingleStream: Latency (ms) | MultiStream: Latency (ms)*

Table 4 - The result of MLPerf Inference v2.0

Figures 3, 5, and 7 below illustrate the results for the **Offline** scenarios. It's apparently to observe that the GPU with better computation ability gets better performance. The throughput of the dual GPU is two times higher than the single GPU. That shows the impressive linear scaling ability, which meets the expectations.

Figures 4, 6, and 8 illustrate the results for **SingleStream**. We noticed that the SingleStream test doesn't show the same tendency as Offline. The performance of dual-GPU does not scale linearly and could be due to bottlenecks on the SUT or CPU.

We do not recommend adding 2nd GPU for this particular workload until more research can be done to understand and resolve the bottlenecks.

For clarity, **MultiStream** is not included in the charts below. Also, the statistics of RNNT and SSD-ResNet34 are not visualized in this product brief. However, the figures below already provide sufficient information. As can be seen, a dual-processor configuration leads to an increase in performance.

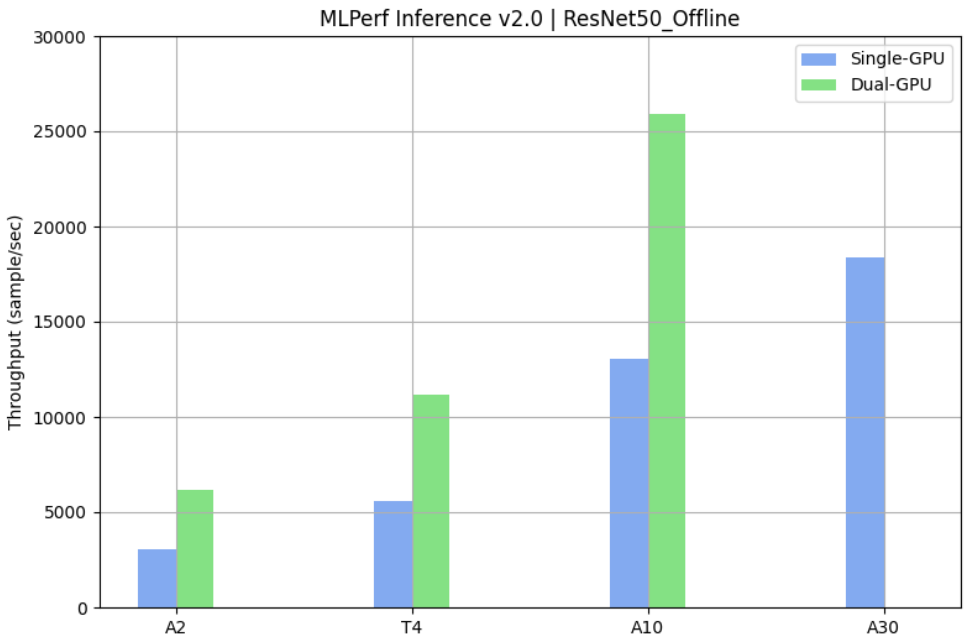

Figure 3 - The result of MLPerf Inference – ResNet50 Offline (Throughput: Higher is Better)

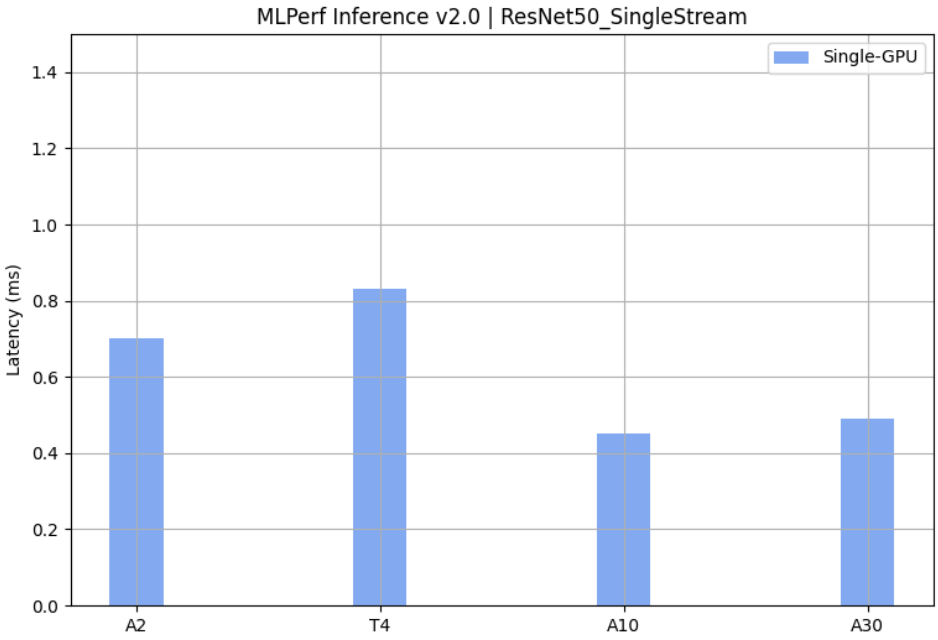

Figure 4 - The result of MLPerf Inference – ResNet50 SingleStream, (Latency: Lower is Better)

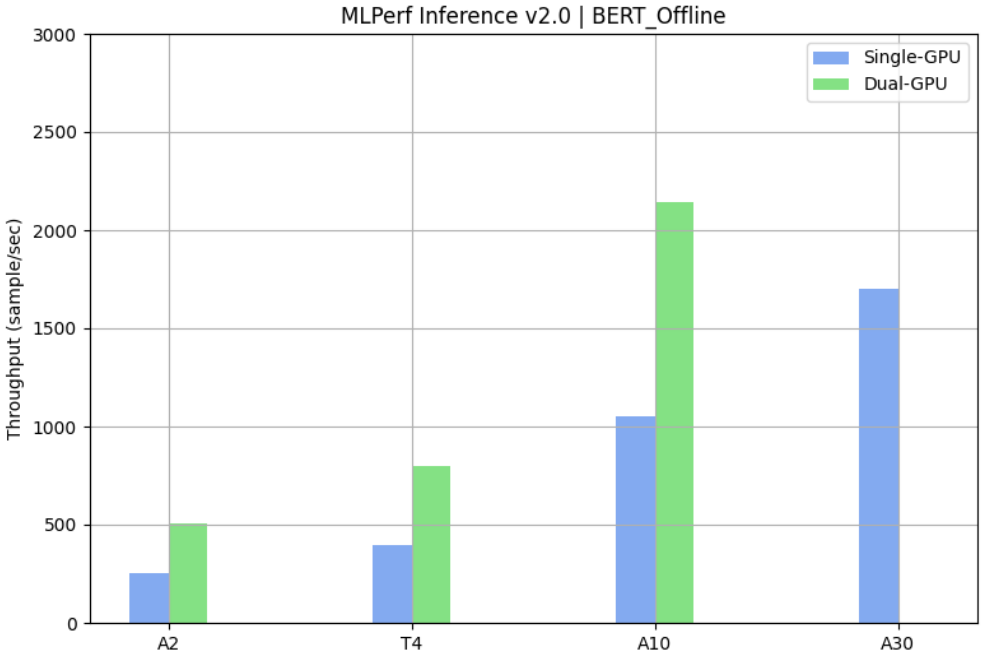

Figure 5 - The result of MLPerf Inference – BERT Offline (Throughput: Higher is Better)

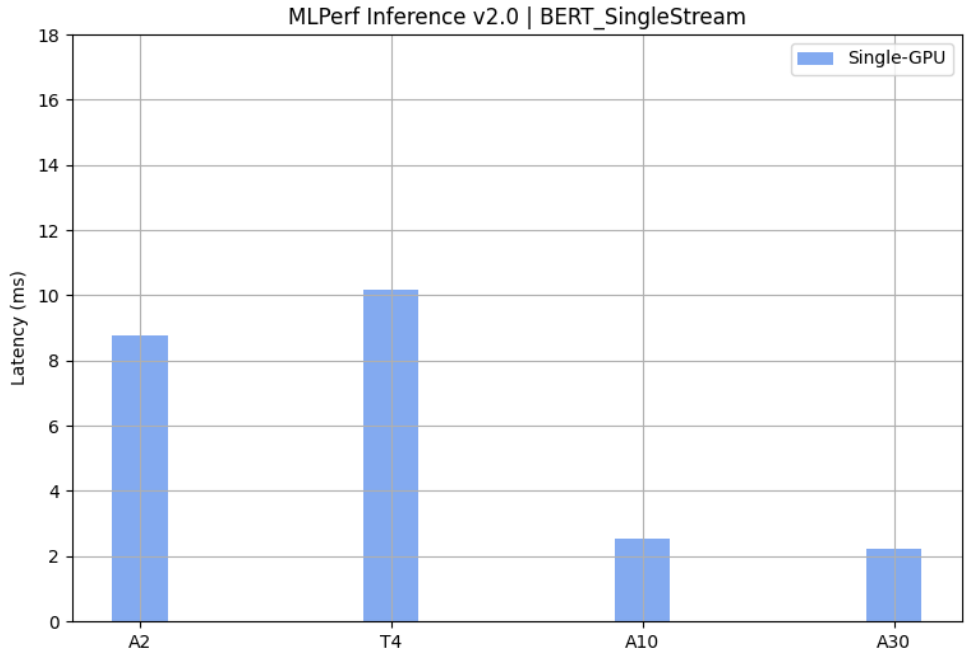

Figure 6 - The result of MLPerf Inference – BERT SingleStream, (Latency: Lower is Better)

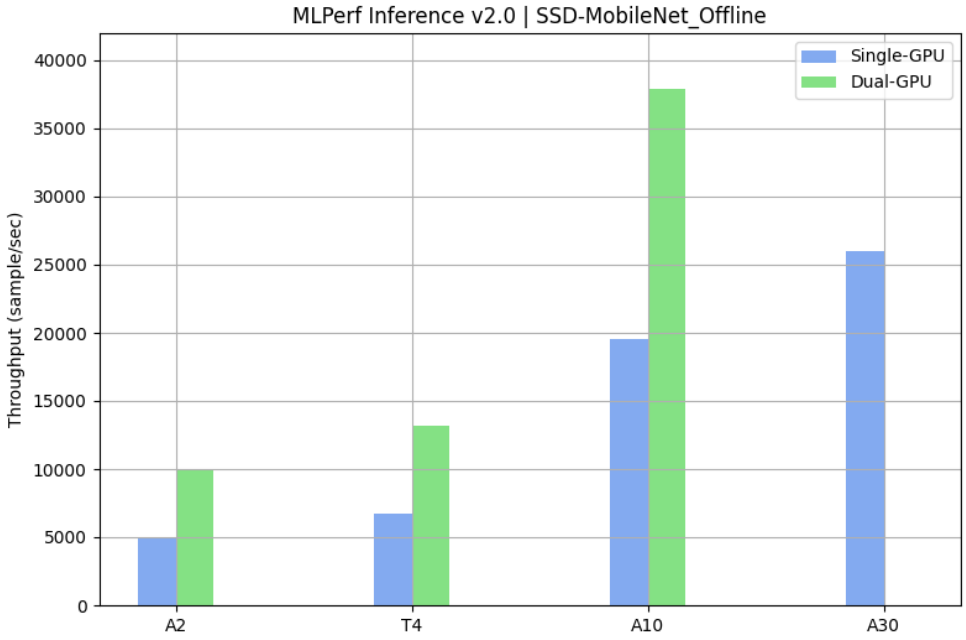

Figure 7 - The result of MLPerf Inference – SSD-MobileNet Offline (Throughput: Higher is Better)

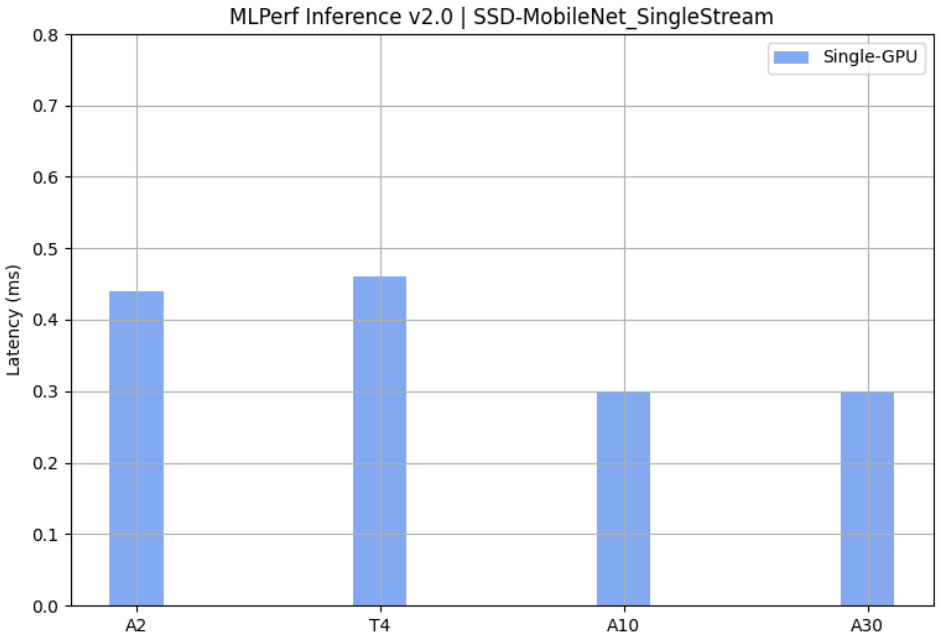

Figure 8 - The result of MLPerf Inference – SSD-MobileNet SingleStream, (Latency: Lower is Better)

OpenVino™ Benchmark

OpenVino benchmark is one of the tools of the Intel® Distribution of OpenVino™ toolkit, which helps accelerate deep learning inference across a variety of Intel® processors and accelerators [10]. The OpenVino benchmark is the tool that can evaluate the CPU ability on AI inference. The experiments have been done on two systems. Please check the system configurations below in Table 5. We can observe that the performance of SYS-220HE-FTNR [12] is roughly 2x faster than E403-12P-FN2T. The reason is that the SYS-220HE-FTNR has dual sockets, and SYS-E403-12P-FN2T is a single-socket system. Therefore, per processor, the SYS-E403-12P-FN2T has a similar performance (see Table 6 for more details).

System configurations		
System SKU	SYS-E403-12P-FN2T	SYS-220HE-FTNR
CPU	Intel® Xeon® Gold 6338 CPU @ 2.00GHz	2x Intel® Xeon® Platinum 8352V CPU @ 2.10GHz
GPU	3x PCI-E 4.0 x16 slot	4x PCI-E 4.0 x16 slot
Memory	8 DIMM slots	32 DIMM slots
Network	2x 10GbE BaseT port(s)	2x AIOM network slots

Table 5 - System configurations for OpenVino benchmark

OpenVino v2022.1.0				
System	**Precision**	**ResNet50-TF**	**ResNet34 PyTorch**	**MobileNet PyTorch**
SYS-220HE-FTNR	INT8	Performance: **2953.99 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **3773.46 fps** Accuracy: **N/A** Batch size: **10** Streams: **16**	Performance: **13955.78 fps** Accuracy: **N/A** Batch size: **10** Streams: **16**
	FP16	Performance: **677.60 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **774.26 fps** Accuracy: **N/A** Batch size: **10** Streams: **16**	Performance: **5291.20 fps** Accuracy: **N/A** Batch size: **10** Streams: **16**
SYS-E403-12P-FN2T	INT8	Performance: **1346.50 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **1780.31 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **7720.95 fps** Accuracy: **95%** Batch size: **10** Streams: **16**
	FP16	Performance: **319.40 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **315.96 fps** Accuracy: **95%** Batch size: **10** Streams: **16**	Performance: **2219.06 fps** Accuracy: **95%** Batch size: **10** Streams: **16**

Table 6 - The result of the OpenVino Benchmark

Conclusion

In this article, Supermicro has investigated the capacity of SYS-E403-12P-FN2T by using MLPerf Inference v2.0. The range of these tests confirms the flexibility and compatibility the SYS-E403-12P-FN2T offers. Furthermore, even without fine-grain parameter tuning, the outcome of the tests shows that the SYS-E302-12P-FN2T delivers an impressive and competitive performance compared to published MLPerf Inference submissions.

Also, SYS-E403-12P-FN2T with a single NVIDIA A30 has been certified with NVIDIA's latest NCS 2.6. As a result, the SYS-E403-12P-FN2T is extremely flexible for all AI/ML application scenarios.

Overall, the SYS-E403-12P-FN2T provides excellent performance and flexibility in AI Edge inferencing applications and is a great fit for deployments at the Intelligent Edge.

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

References

[1] https://www.supermicro.com/zh_tw/products/system/iot/box_pc/sys-e403-12p-fn2t

[2] MLPerf Inference Benchmark (arXiv:1911.02549v2 [cs.LG] 9 May 2020), https://arxiv.org/abs/1911.02549

[3] https://catalog.ngc.nvidia.com/

[4] https://arxiv.org/abs/1512.03385

[5] https://arxiv.org/abs/1810.04805

[6] https://arxiv.org/abs/1512.02325

[7] https://arxiv.org/abs/2103.09935

[8] https://mlcommons.org/en/

[9] https://mlcommons.org/en/inference-edge-20/

[10] https://docs.openvino.ai/latest/openvino_docs_performance_benchmarks.html

[11] https://www.intel.com/content/www/us/en/developer/articles/community/esh-recommended-hardware-program-overview.html

[12] https://www.supermicro.com/zh_tw/products/system/hyper/2u/sys-220he-ftnr