PRODUCT BRIEF

SUPERMICRO X13 SERVERS, POWERED BY 5TH GEN INTEL® XEON® PROCESSORS DELIVER REMARKABLE PERFORMANCE ENHANCEMENTS WHEN USING INTEL® AMX

New Intel CPUs incorporate Advanced Matrix Extensions (Intel® AMX), an integrated AI accelerator, which significantly improves AI inference workloads such as language modeling, object detection, and image recognition.

Supermicro CloudDC System

TABLE OF CONTENTS

Executive Summary

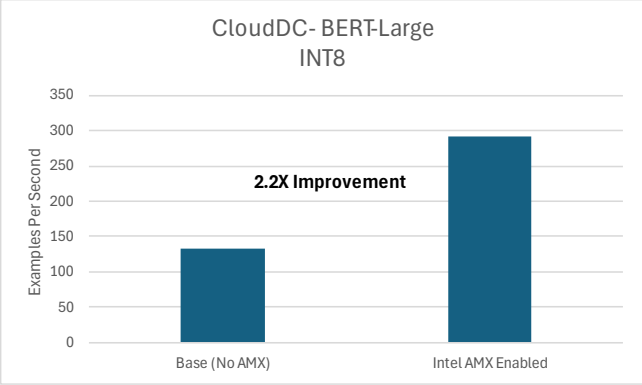
The evaluation of performance improvement resulting from Intel's Advanced Matrix Extensions (Intel AMX) AI accelerator, integrated into the 5th Gen Intel Xeon Processors, involved conducting tests across two Supermicro X13 SuperServer systems. These tests encompassed three distinct AI inference workloads: language modeling (BERT-large), image recognition (ResNet50 v1.5), and object detection (SSD-ResNet34). Initially, these workloads were executed on the Supermicro X13 CloudDC SuperServer systems without utilizing Intel AMX. Subsequently, the evaluation extended to leveraging Intel integrated AI accelerators, with testing performed using two different precision modes: INT8 and bfloat16.

What are Data Types?

As AI is becoming so ubiquitous, a new field that refers to the data type used is now visible, and customers need to understand the differences. The benchmarks below all refer to data types. Briefly, the two data types used in these benchmarks refer to the amount of precision in the data. INT8 refers to an integer number with 8 bits of range in these tests. This data type can be used when more precision is not necessary and is faster than higher precision data types. The bfloat16 data type is a 16-bit data type designed specifically for machine learning applications. It balances the higher precision of a 32-bit floating point (float32) and the lower memory footprint of a 16-bit half-precision (float16).

Inference on Supermicro CloudDC SuperServer

- AI inference on the Supermicro X13 CloudDC SuperServer system observed a throughput of 132.7 Examples per Second during inference on BERT-Large using INT8 data type.
- When the same workload was executed on the same Supermicro X13 CloudDC SuperServer utilizing Intel built-in Advanced Matrix Extensions (AMX), the system's performance achieved a much higher throughput of 292.6 Examples per Second.
- Enabling Intel AMX AI built-in accelerator on 5th Gen Xeon Processors resulted in a substantial 2.2x performance gain while running BERT-Large AI inference (INT8).
- The performance improvements were even more compelling when the workload was run using bfloat16 precision. With the bfloat16 data type, Supermicro X13 CloudDC SuperServer AI inference achieved 30.8 Examples per Second on BERT-Large. After enabling Intel AMX in the processor, this performance soared to 130.4 Examples per Second, representing an astonishing 4.2x improvement when utilizing Intel AMX.



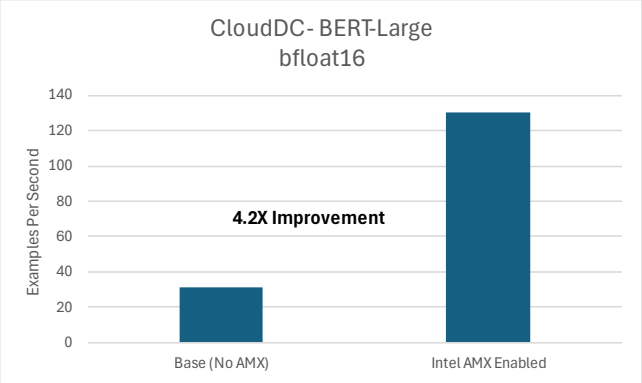
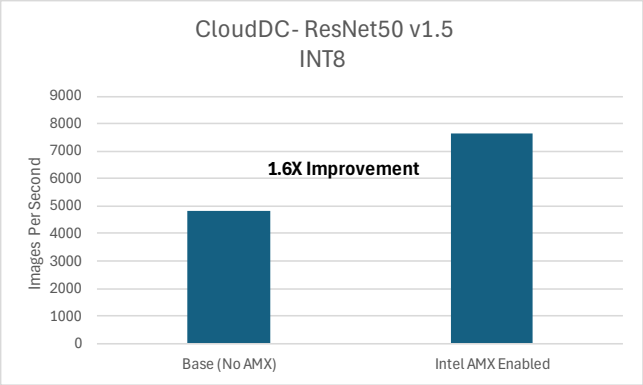
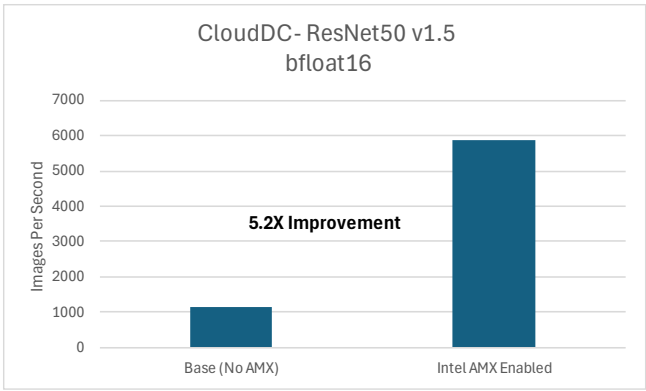


Image Recognition AI Inference on Supermicro CloudDC SuperServer

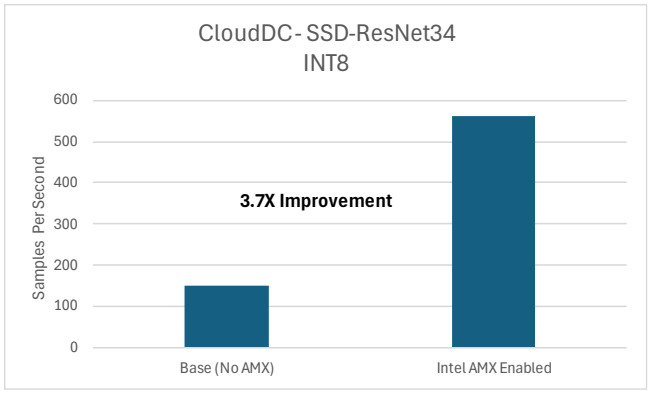
- AI image recognition inference on the Supermicro X13 CloudDC SuperServer achieved a performance of 4815.8 Images per Second during real-time inference using INT8 data type with ResNet50 v1.5.
- When the same workload was run using Intel AMX on the same system, performance improved to 7642.8 Images per Second.
- Leveraging Intel Advanced Matrix Extensions led to a 59% performance boost.
- When the bfloat16 data type was utilized, Supermicro X13 CloudDC SuperServer AI inference achieved 1136.3 Images per Second on ResNet50 v1.5 real-time inference. After enabling Intel AMX, this performance climbed to 5864.1 Images per Second, an impressive 5.2x performance improvement.

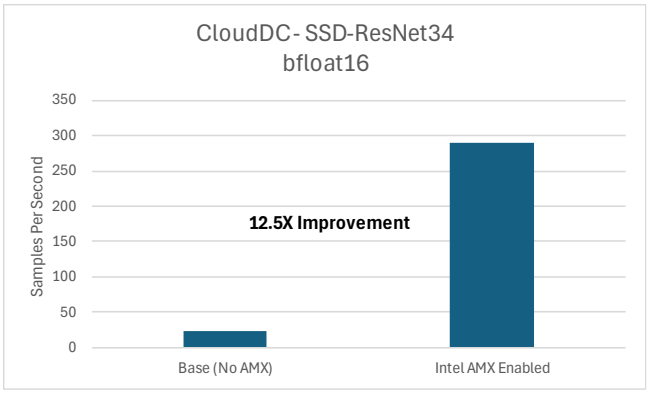




Object Detection AI Inference on Supermicro CloudDC SuperServer

- The improvements were even more significant in object detection AI inference tasks. The Supermicro X13 CloudDC SuperServer demonstrated a performance of 150.8 Samples per Second during real-time inference using INT8 precision on SSD-ResNet34.
- When the same workload was completed utilizing Intel AMX on 5th Gen Xeon, the system's performance achieved a remarkable throughput of 561.0 Samples per Second.
- Enabling Intel AMX resulted in a substantial 3.7x performance boost, leveraging Intel 5th Gen Xeon processors on Supermicro X13 CloudDC SuperServer.
- Running the same workload on the same system with bfloat16 data type led to 23.3 Samples per Second. The performance was improved to 290.2 Samples per Second after enabling Intel AMX, representing an astonishing 12.5x throughput improvement.





The use of Intel AMX results in faster processing times, increased throughput, and enhanced performance, making it an ideal choice for businesses looking to optimize their AI workloads.

Conclusion

Several AI inference workloads benchmark results were collected on the Supermicro X13 CloudDC featuring 5th Gen Intel Xeon Processors, these workloads include language modeling (BERT-Large), image recognition (ResNet50 v1.5), and object detection

(SSD-ResNet34) using both INT8 and bfloat16 data types. The workloads were first run without leveraging Intel's AI built-in accelerator in 5th Gen Xeon CPUs, named Intel Advanced Matrix Extensions (AMX). Then, the same workloads run on the same system while utilizing Intel AMX. The results across all three AI inference workloads experienced a significant boost in throughput after enabling Intel Xeon AI integrated accelerator, up to 12.5x performance improvement with Intel AMX. This shows that although large AI training models benefit from Supermicro SuperServer systems featuring accelerators like the Intel Gaudi2 or the Intel Data Center GPU Max series, various AI inference workloads can efficiently run on Intel 5th Gen Xeon processors.

Appendix

Testing Methodology

- Hardware Configuration
  - Supermicro X13 CloudDC SuperServer SYS-621C-TN12R
  - CPU: 2x Intel Xeon Platinum 8592+ (64cc, 1.9 GHz, 350 W)
  - Memory: 16x MEM-DR564L-SL01-ER56
  - Drives: 1x HDS-SMP-SSDPF2KX038T1
  - AIOM: 1x AOC-ATG-i2TM

Software

- Operating System and Kernel
- Ubuntu 22.04.03 LTS, Kernel: 5.15.0-89-generic

Process

- AVX-512 Vector Neural Network Instructions (VNNI) is used when AMX instruction is off.
- The AI inference workload results were obtained and can be replicated with documentation found on the latest Intel AI Reference Model webpage.
- ResNet50 v1.5
- Real Dataset Output uses the tf_records dataset from ImageNet.

Further Information:

Supermicro X13 Systems – www.supermicro.com/x13

Supermicro X13 CloudDC System Information: https://www.supermicro.com/en/products/clouddc

Intel AMX Description: https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html

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. See www.supermicro.com