SOLUTION BRIEF

SUPERMICRO X13 HYPER EMPOWERS ENTERPRISE AI WORKLOADS ON THE VMWARE PLATFORM

Computational AI workload Use Cases: Large Language Model (LLM) and AI Image Recognition – ResNet50 on Intel® Data Center Flex 170 GPU

Supermicro Hyper Dual Socket Server

TABLE OF CONTENTS

Executive Summary 1
Reference Architecture 2
System Configuration. 4
Summary 5
References 5

Executive Summary

This document presents the verified generative AI and image recognition solutions deployed on the Supermicro X13 Hyper system, powered by 6th Gen Intel® Xeon Scalable processors and the Intel Data Center GPU Flex 170. The critical integrated use cases include a knowledge base utilizing the Large Language Model (LLM) and an AI image recognition powered by ResNet50, both running on virtualized VMware infrastructure. By leveraging Single-Root Input/Output Virtualization (SR-IOV), PCIe resources are directly allocated to one or more Virtual Machines (VMs), enabling the offloading of AI computational workloads from the CPU to the GPU. This solution demonstrates successful virtualized integration, significantly reduces latency, and delivers optimal performance for AI workloads.

Reference Architecture

In designing a high-performance AI solution for enterprise applications on Supermicro X13, SYS-221H-TNR, the goal was to select models that provide cutting-edge capabilities and ensure optimal efficiency across different workloads. The combination of IPEX LLM Meta-Llama-3-8B and ResNet50 was strategically chosen for their complementary strengths in handling diverse AI applications, from language understanding to image recognition.

Both models are integrated within a virtualized VMware infrastructure, taking advantage of Single-Root Input/Output Virtualization (SR-IOV) to allocate PCIe resources efficiently across Virtual Machines (VMs). By offloading AI workloads from the CPU to the GPU, this architecture ensures that computationally intensive tasks, such as inference and LLM processing, are executed with minimal latency and maximum resource utilization.

This holistic approach to AI workload management, combined with the inherent strengths of Meta-Llama-3-8B and ResNet50, enables a flexible, scalable, and high-performance solution.

Furthermore, performance analysis across various configurations revealed that Intel© Data Center Flex 170 GPU handles such workloads exceptionally well, especially when Error Correction Code (ECC) was disabled to maximize the available device memory capacity. This optimization further enhances the model's ability to manage large-scale AI applications efficiently, ensuring optimal resource utilization for enterprise environments.

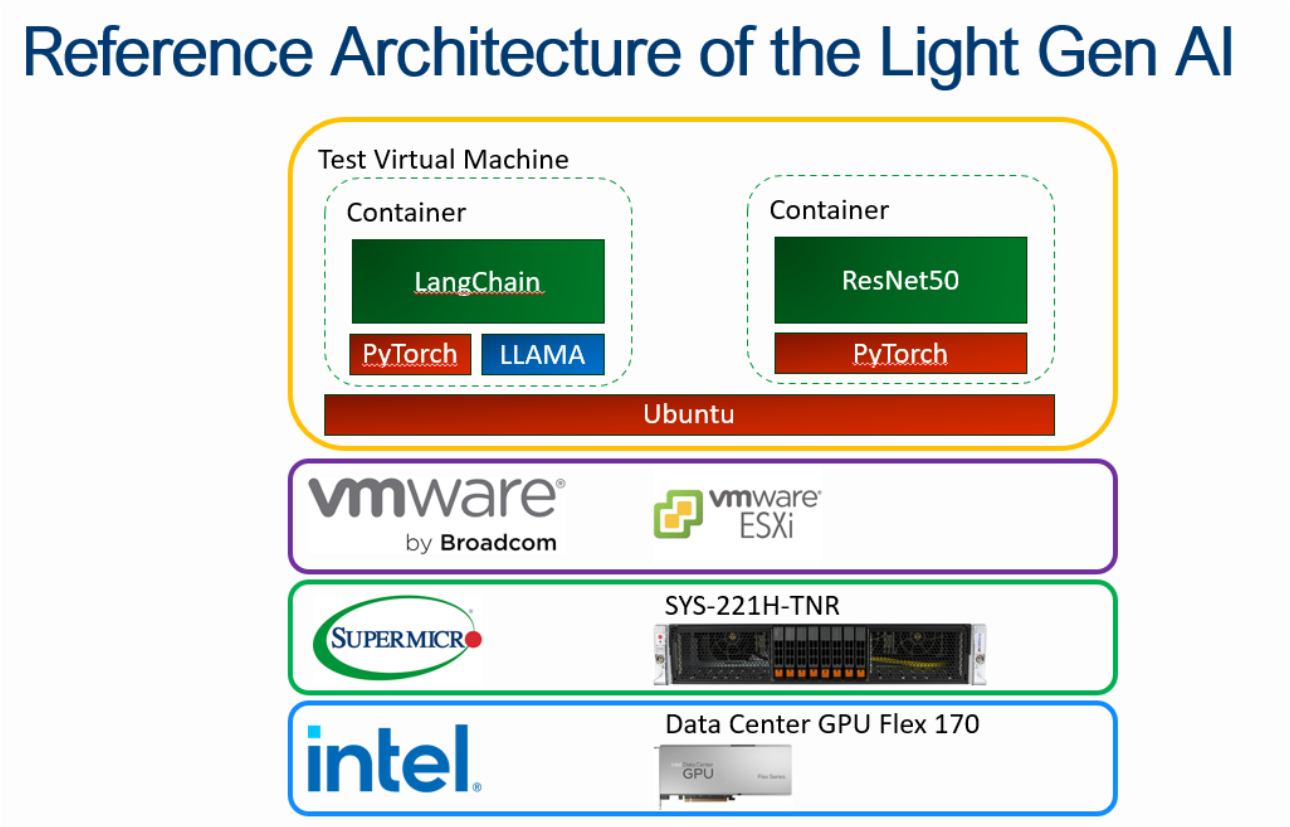

Figure 1 - Solution Reference Architecture

- **IPEX LLM Meta-Llama-3-8B**

IPEX is a PyTorch library for running LLM Meta-Llama-3-8B. The software was selected for its powerful large language model (LLM) architecture, which excels in processing and generating human-like text in the knowledge base (LangChain). This model's first and rest token latency is highly optimized. It is ideal for use cases requiring fast, real-time responses, such as knowledge bases and conversational AI systems. This low-latency performance ensures that even as data complexity increases, the model maintains responsiveness. Additionally, Meta-Llama-3-8B supports various data types, providing the flexibility needed for different text-based AI applications. Whether processing structured data or unstructured text, the model adapts seamlessly to the input type, making it versatile for multiple industries, including healthcare, finance, and customer service.

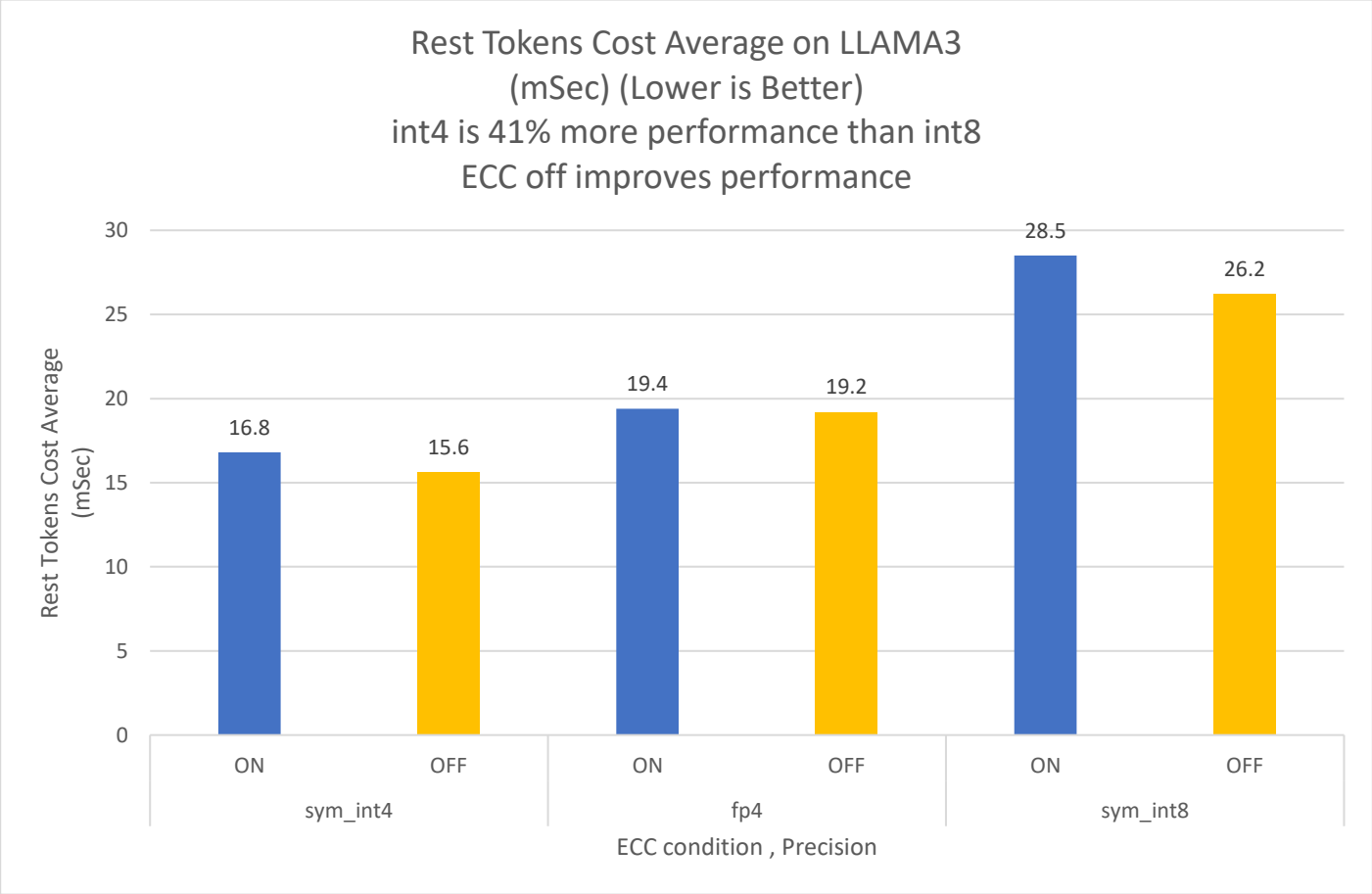

Figure 2 - Performance of LLM - LLAMA 3 with Different Data Types and ECC turned ON and OFF

- **ResNet50**

For image recognition tasks, ResNet50 was chosen due to its proven performance in inference workloads, especially in terms of throughput and latency. In AI-driven environments where rapid and accurate image analysis is critical, such as medical imaging or industrial quality control, ResNet50's deep learning architecture delivers high-precision results while keeping latency low.

Using INT8 and FP16 data formats in ResNet50 further enhances its performance. Running inference with these precision levels, particularly at a batch size of 1024, maximizes throughput without compromising accuracy. This precision flexibility allows the model to strike the right balance between speed and computational efficiency, making it ideal for applications requiring high-speed image processing.

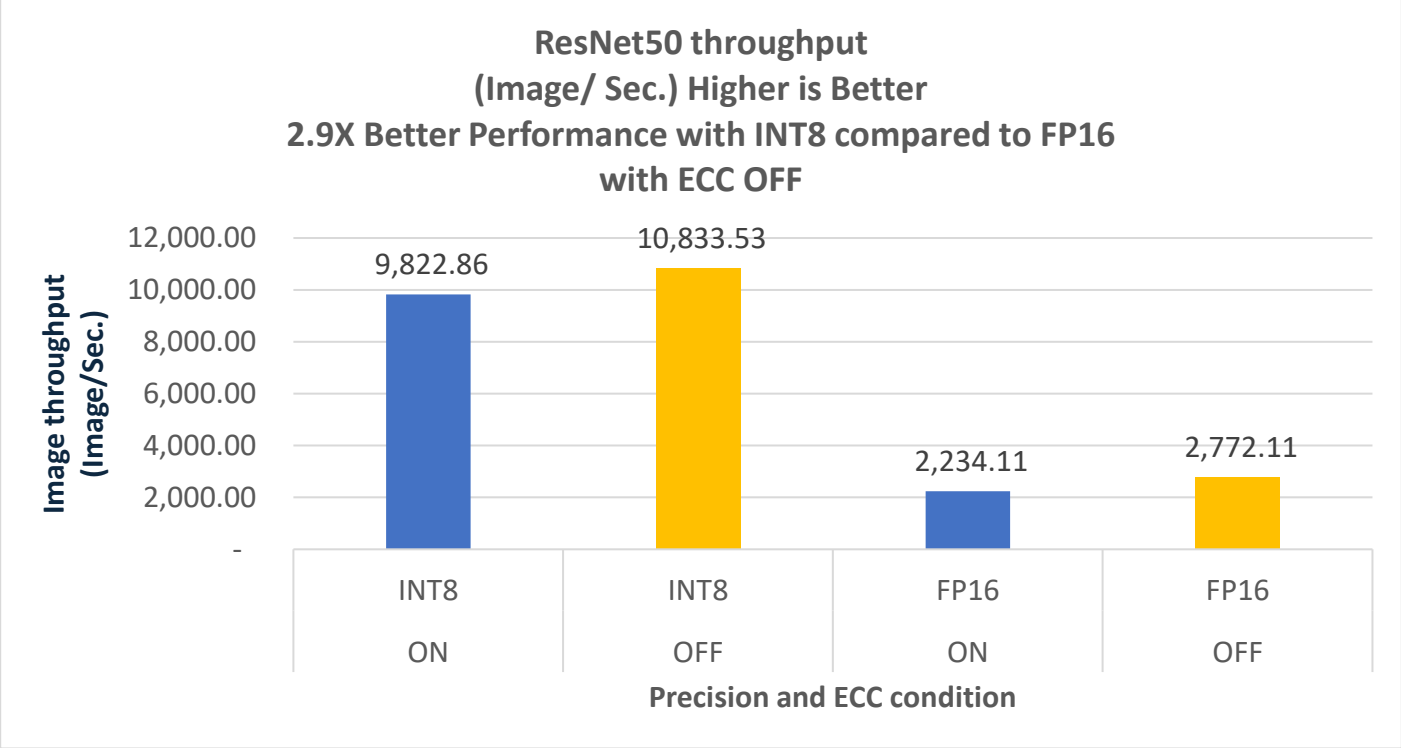

Figure 3 - ResNet50 Performance Comparisons Between INT8 and FP16 Data Formats and ECC ON and OFF

System Configuration

Type	Configuration
System	SYS-221H-TNR
CPU	2x Intel® Xeon® 6548Y (32 Cores, 2.5 GHz)
MEMORY	1TB DDR5 4800 (64GB x 16)
Storage Drive	10 x 6.4TB PCIe Gen 4 x 4
NIC	2x 100GbE PCIe Gen 4 x16
GPU	Intel Data Center Flex 170 x 1

Summary

This AI solution, built on the VMware-certified Supermicro X13 Hyper "SYS-221H-TNR," offers enterprise customers a powerful and scalable platform for AI workloads, bringing the following key advantages:

- VMware Certified Ready Node: Fully supports VMware solutions, ensuring seamless integration and compatibility.
- Balanced PCIe Distribution: Optimized across a dual-processor (DP) architecture, enhancing performance and resource management.
- Flexible Storage Options: Supports up to 24 storage drive bays with NVMe, SAS, and SATA configurations to meet diverse storage capacity needs.
- High-Speed PCIe 5.0: Delivers up to 400Gbps throughput, ensuring high-performance data transfer for demanding applications.
- AI-Optimized GPU Support: Accommodates up to four GPUs in full-height, 10.5” slots, enabling efficient execution of AI workloads.

Combined with 5th Gen Intel Xeon Scalable CPUs, Data Center GPU Flex 170, and SR-IOV for efficient resource allocation, these capabilities ensure a flexible, scalable, high-performance infrastructure tailored for modern enterprise AI applications.

References

Supermicro X13 SYS-221H-TNR: https://www.supermicro.com/en/products/system/hyper/2u/sys-221h-tnr

VMware certification of Supermicro X13 SYS-221H-TNR: https://www.vmware.com/resources/compatibility/detail.php?deviceCategory=server&productid=59936&deviceCategory=server&details=1&partner=105&cpuSeries=165&page=2&display_interval=10&sortColumn=Partner&sortOrder=Asc

Intel Data Center GPU Flex 170: https://www.intel.com/content/www/us/en/products/sku/230019/intel-data-center-gpu-flex-170/specifications.html

Meta LLAMA 3: https://ai.meta.com/blog/meta-llama-3/

ResNet50: https://pytorch.org/vision/main/models/generated/torchvision.models.resnet50.html

LangChain, a knowledge base platform powered by Meta LLAMA 3 LLM and, provides a human like text: https://www.langchain.com/

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.

For more information: www.supermicro.com

INTEL

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, visit www.intel.com