White Paper

Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro – GEN AI

Authors

Ecosystem Edge AI System Architect:

Abhijit Sinha

Timothy Miskell

Yuan Kuok Nee

Ecosystem Enabling Engineer:

Shin Wei Lim

Supermicro: Toby McClean Principal Solutions Architect

Key Contributors

Ecosystem Edge AI System Verification and Qualification Manager:

Edel Curley

1 Introduction

Intel® AI Edge Systems offers a balance between computing and AI acceleration to deliver optimal TCO, scalability, and security. AI Edge systems enable customers to jumpstart development through a hardened system foundation verified by Intel®. AI Edge systems enable the ability to add AI functionality through continuous integration into business applications for better business outcomes and streamlined implementation efforts.

To support the development of these Edge AI systems, Intel® is offering reference design and verified reference blueprints with AI system configurations that are tuned and benchmarked for different AI system types that support Edge AI Workloads. Verified reference blueprints (VRB) include Hardware BOM, Foundation Software configuration (OS, Firmware, Drivers) tested and verified with supported Software stack (software framework, libraries, orchestration management).

This document describes a verified reference blueprint architecture using the 5th Gen Intel® Xeon® Scalable processor family, and Intel® Data Center GPU Flex 140/170 or Intel® Arc™ A380/A750.

Intel® AI Edge System Verified Reference Blueprint helps end users simplify design choices by bundling hardware and software pieces together while making the high performance more predictable. The solution leverages hardened hardware, firmware, and software to allow customers to integrate on top of this known-good foundation.

Intel ® AI Edge Systems Verified Reference Blueprint provides numerous benefits to ensure end users have excellent performance for their Edge AI Inference applications. Some of the key benefits of the blueprint, based on the 5th Generation Intel® Xeon® Scalable Processor Family processor and Intel® Data Center GPU Flex 140/170 or Intel® Arc™ A380/A750:

- High core counts and per-core performance
- Compact, power-efficient system-on-chip platform
- Streamlined path to cloud-native operations
- Accelerated AI inference using Intel® AMX and Intel® DL Boost
- Multiple discrete GPU support to accelerate for AI inference workload
- The X® kernel of Intel® GPUs integrates Extended Vector Engine (XVE) and Extended Matrix Engine (XMX), which accelerate AI workflow and provide powerful and real-time computing power support for AI inference at the edge
- Accelerated encryption and compression
- Platform-level security enhancements

Table of Contents

Figures

Tables

Document Revision History

Doc ID	Revision Number	Description	Date
849098	1.0	Initial release	March 2025

2 Design Compliance Requirements

This chapter focuses on the design requirements for Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS111-AD-WRN and Supermicro SYS-E403-13E-FRN2T.

2.1 Hardware Requirements

The checklists in this chapter are a guide for the platform tested as part of Intel ® AI Edge Systems Verified Reference Blueprint – Computer Vision. The hardware specifications are detailed below.

Ingredient	Supermicro SYS111-AD-WRN	Supermicro SYS-E403-13E-FRN2T
Processor	Intel® 14th Generation Core™ i9-14900E Processor 8 P-Cores, 16 E-Cores, 65 W or equivalent SKU	5th Gen Intel® Xeon® Gold 6538N Processor at 2.1GHz, 32C/64T, 205W or higher number SKU
Total Memory	128 GB DDR5 4800 MT/s	Option 1: DRAM only configuration: 256 GB (16x 16 GB DDR5, 4800 MHz) Option 2: DRAM only configuration: 512 GB (32x 16 GB DDR5, 4800 MHz)
Storage (Boot Drive)	480 GB or equivalent boot drive	480 GB or equivalent boot drive
Storage (Capacity)	1 TB or equivalent boot drive	Minimum 1 TB or equivalent drive
Graphics	Intel® Arc™ A380 Intel® Arc™ A750	Intel® Data Center Flex 170 (Up to 2) Intel® Data Center Flex 140 (Up to 3)
LAN on Motherboard (LOM)	1 Gbps I219-LM for Operation, Administration and Management (OAM)	10 Gbps or 25 Gbps port for video streaming 1/10 Gbps port for Management Network Interface Controller (NIC)
Product Image		

Table 1 Intel ® AI Edge Systems Verified Reference Blueprint –GEN AI Configuration

2.2 BIOS Settings

To meet the performance requirements for an Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro systems, Intel® recommends using the BIOS settings to enable processor p-state and c-state with Intel® Turbo Boost Technology ("turbo mode") enabled. Hyperthreading is recommended to provide higher thread density.

Setting	Value
Hardware Prefetcher	Enabled
Intel® (VMX) Virtualization Technology	Enabled
Hyper-Threading	Enabled
Intel® Speed Shift Technology	Enabled
Turbo Mode	Enabled
C-States	Enabled
Enhanced C-States	Enabled
C-State Auto Demotion	C1
C-State Un-Demotion	C1
MonitorMWait	Enabled
Enforce DDR Memory Frequency POR	POR
Maximum Memory Frequency	Auto
Primary Display	Auto
Internal Graphics	Auto
Graphics Clock Frequency	Max CdClock freq based on Reference Clk
VT-d	Enabled
Re-Size BAR Support	Enabled
SR-IOV Support	Enabled

Table 2 BIOS Settings for Supermicro SYS-111-AD-WRN

Setting	Value
Memory Page Policy	Adaptive
ICCP Pre-Grant Level	AMX
Speed Step (P-States)	Enable
Turbo Mode	Enable
EIST PSD Function	HW_ALL
Hardware P-States	Native Mode with No Legacy Support
Enable Monitor MWAIT	Enable
Enhanced Halt State (C1E)	Enable
Hyperthreading (ALL)	Enable
Hardware Prefetcher	Enable
LLC Prefetch	Enable
Extended APIC	Enable
Intel Virtualization Technology	Enable

Table 3 BIOS Version 2.1 Settings for Supermicro SYS-E403-13E-FRN2T

2.3 Solution Architecture

Figure 1 shows the architecture diagram of Intel ® AI Edge Systems Verified Reference Blueprint –GEN AI. The software stack consists of a single category of AI software: GEN AI. The GEN AI application is containerized using docker.

For the GEN AI use case, we are using the Intel® Extensions for Pytorch (IPEX). The Generative AI benchmark on the 14th Generation Intel® Core and 5th Generation Intel® Xeon® Scalable Processors leverages vLLM, which, as shown in the figure below, performs continuous batching of requests to the LLM. The inference is performed using multiple models including:

- Flan-t5
- TinyLlama
- Phi-3 4K-Instruct
- Llama3 8B
- GPT-NEOX 20B

The Large Language Model (LLM) proxy workload highlights the Generative AI processing capabilities of the Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS111-AD-WRN and Supermicro SYS-E403-13E-FRN2T for GEN AI

configuration directly on either Intel® 14th Generation Core or 5th Generation Intel® Xeon® Scalable Processors or offloaded to Intel® Arc™ GPUs along with Intel® Data Center Flex GPUs.
Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS111-AD-WRN and Supermicro SYS-E403-13E-FRN2T, ensure that the results of the system follow the expected results as shown below to baseline the performance of the platform. The results shown include performance values for the next token latency, the achievable number of tokens per second, along with the time per query.

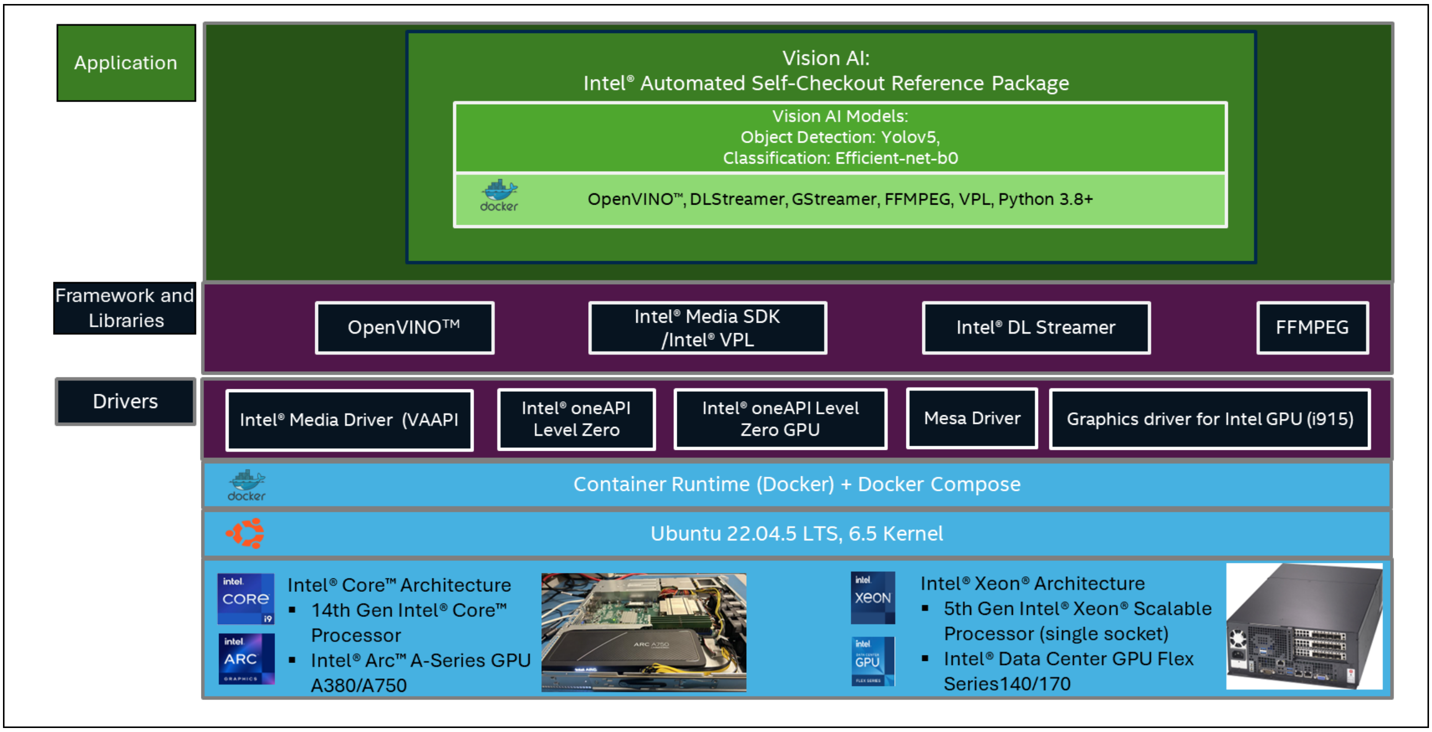

Application
Vision AI:
Intel® Automated Self-Checkout Reference Package
Vision AI Models:
Object Detection: Yolov5,
Classification: Efficient-net-b0
docker
OpenVINO™, DLStreamer, GStreamer, FFMPEG, VPL, Python 3.8+
Framework and Libraries
OpenVINO™
Intel® Media SDK /Intel® VPL
Intel® DL Streamer
FFMPEG
Drivers
Intel® Media Driver (VAAPI
Intel® oneAPI Level Zero
Intel® oneAPI Level Zero GPU
Mesa Driver
Graphics driver for Intel GPU (i915)
docker
Container Runtime (Docker) + Docker Compose
Ubuntu 22.04.5 LTS, 6.5 Kernel
intel CORE i9
intel ARC GRAPHICS
Intel® Core™ Architecture
▪ 14th Gen Intel® Core™ Processor
▪ Intel® Arc™ A-Series GPU A380/A750
intel XEON
intel GPU
Intel® Xeon® Architecture
▪ 5th Gen Intel® Xeon® Scalable Processor (single socket)
▪ Intel® Data Center GPU Flex Series140/170

Figure 1 Architecture of Intel® AI Edge Systems - GEN AI

The Gen AI benchmarks on leverage OpenVINO™, the Intel Extensions for Pytorch (IPEX), along with vLLM. In the case of the vLLM approach, as shown within the figure below, continuous batching of requests are performed to the LLM. In contrast to the OpenVINO™ and IPEX approaches where a fixed input and output token size are utilized, for the vLLM benchmark the input and output token sizes are variable.

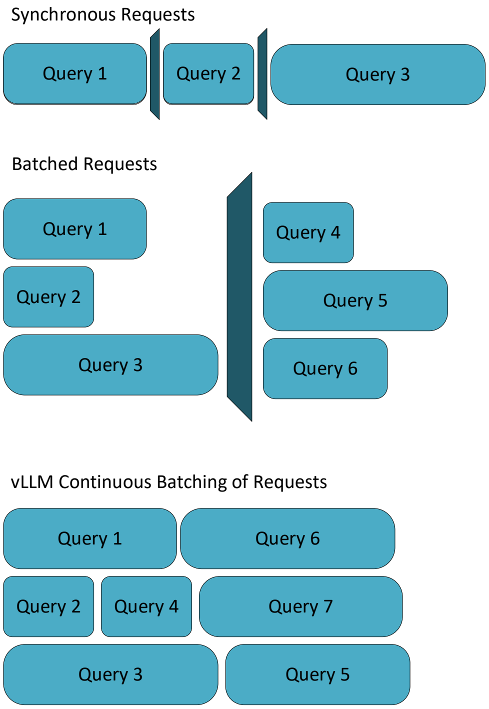


Figure 2 vLLM Continuous Batching

The table below is a guide for assessing the conformance to the software requirements of the Intel ® AI Edge Systems Verified Reference Blueprint Ensure that the platform meets the requirements listed in the table below.

Ingredient	SW Version Details	
OS	Ubuntu 22.04.5 LTS	
Kernel	6.5 (in-tree generic)	
Microcode	Core: 0x123	
	Xeon: 0x21000161	
Docker Engine	27.1.0	
Docker Compose	2.29	
Intel® Level Zero for GPU	1.3.29735.27	
Intel® Graphics Driver for GPU (i915)	24.3.23	
Framework /Toolkit	**CPU:** PyTorch v2.3.100+cpu Deepspeed v0.14.0 Transformers v4.38.1 IPEX-LLM	GPU: PyTorch v2.1.0a0+cxx11.abi Deepspeed @ed8aed5 Transformers v4.37.0 IPEX-LLM
Model Topology	EleutherAI/gpt-neox-20b meta-llama/Llama-3-8b-hf microsoft/Phi-3-mini-4k-instruct TinyLlama/TinyLlama-1.1B-Chat-v1.0	
Libraries	oneDNN v3.4.1 oneCCL v2021.11 torch-ccl v2.3.0+cpu Intel® Neural Compressor v2.4.1	
Quantization methods	weight-only-quantization	
Warmup steps	1	
Number of Iterations	4	
Batch Size	1, 2, 4, 8, 16	

Ingredient	SW Version Details
Beam Width	1 (greedy search)
Input Token Size	32, 256, 1024, 2048
Output Token Size	1024
Compiler	GCC version 12.3.0
Python	3.10+

Table 4 SW Configuration

2.4 Platform Technology Requirements

This section lists the requirements for Intel's advanced platform technologies.

Enterprise AI requires Intel® AVX (Advance Vector Extensions) or AMX (Intel® Advance Matrix Extensions) to be enabled to reap the benefits of hardware-accelerated convolution.

2.5 Platform Security

For Intel® AI System for the Edge, it is recommended that Intel® Boot Guard Technology to be enabled so that the platform firmware is verified suitable during the boot phase.

In addition to protecting against known attacks, all Intel® Accelerated Solutions recommend installing the Trusted Platform Module (TPM). The TPM enables administrators to secure platforms for a trusted (measured) boot with known trustworthy (measured) firmware and OS. This allows local and remote verification by third parties to advertise known safe conditions for these platforms through the implementation of Intel® Trusted Execution Technology (Intel® TXT).

2.6 Side Channel Mitigation

Intel® recommends checking your system's exposure to the "Spectre" and "Meltdown" exploits. This reference implementation has been verified with Spectre and Meltdown exposure using the latest Spectre and Meltdown Mitigation Detection Tool, which confirms the effectiveness of firmware and operating system updates against known attacks

The spectre-meltdown-checker tool is available for download at https://github.com/speed47/spectre-meltdown-checker.

3 Platform Tuning and GPU Driver Setup

3.1 Additional Linux Packages Installation

3.1.1 Install Docker

Follow the instructions at https://docs.docker.com/engine/install/Ubuntu*/ to install Docker Engine on Ubuntu*.

3.1.2 Install Intel® Arc™ GPU Drivers

Refer to the following for instructions on installing the Intel® Client GPU driver: https://dgpu-docs.intel.com/driver/client/overview.html#installing-client-gpus-on-ubuntu-desktop-22-04-lts. Refer to Table 4 for a list of the installed software versions.

3.1.3 Install Intel® Data Center GPU Drivers

In case the end user is installing a server OS, then the following instructions will need to be followed. For a desktop OS refer to section 3.1.2.

Refer to the following for instructions on installing the Intel® Data Center GPU driver: https://dgpu-docs.intel.com/driver/installation.html#ubuntu. Refer to Table 4 for a list of the installed software versions.

4 Performance Verification

This chapter aims to verify the performance metrics for the Intel ® AI Edge Systems Verified Reference Blueprint to ensure that there is no anomaly seen. Refer to the information in this chapter to ensure that the performance baseline for the platform is as expected.

The Supermicro SYS111-AD-WRN and Supermicro SYS-E403-13E-FRN2T solutions were tested on August 06, 2024, with the following hardware and software configurations listed within Table 1 and Table 4.

4.1 Memory Latency Checker (MLC)

The Memory Latency Checker which can be downloaded from https://www.Intel®.com/content/www/us/en/developer/articles/tool/Intel®r-memory-latency-checker.html. Download the latest version, unzip the tarball package, go into the Linux* folder, and execute `./mlc`. Table 5 and Table 6 below should be used as a reference for verifying the validity of the system setup.

Key Performance Metric	Supermicro SYS111-AD-WRN	Supermicro SYS-E403-13E-FRN2T
Idle Latency (ns)	123.8	150.3
Memory Bandwidths between nodes within the system (using read-only traffic type) (MB/s)	53577	260425

Table 5 Memory Latency Checker

Peak Injection Memory Bandwidth (1 MB/sec) using all threads	Supermicro SYS111-AD-WRN	Supermicro SYS-E403-13E-FRN2T
All Reads	52574	255504
3:1 Reads-Writes	50359	211857
2:1 Reads-Writes	50319	202797
1:1 Reads-Writes	50145	186870
STREAM-Triad	50231	206753
Loaded Latencies using Read-only traffic type with Delay=0 (ns)	464.78	183.11
L2-L2 HIT latency (ns)	47.0	73.6
L2-L2 HITM latency (ns)	47.3	74.7

Table 6 Peak Injection Memory Bandwidth (1 MB/sec) Using All Threads

Note: If the latency performance and memory bandwidth performance are outside the range, please verify the validity of the Platform components, BIOS settings, kernel power performance profile used, and other software components.

4.2 GenAI Performance Benchmarks

For the Intel® AI Edge Systems Verified Reference Blueprint with Supermicro SYS111-AD-WRN, the platform should be able to support the Phi-3-mini-4K-instruct, TinyLlama, and the flan-t5 models all while maintaining sub 100 millisecond inference times. For the Intel® AI System for the Edge Verified Reference Blueprint with Supermicro SYS-E403-13E-FRN2T platform should be able to support the GPT-NEOX 20B, Llama3 8B, the Phi-3-mini-4K-instruct models, and the TinyLlama models all while maintaining sub 100 millisecond inference times.

4.2.1 GenAI benchmarks on Supermicro SYS111-AD-WRN System

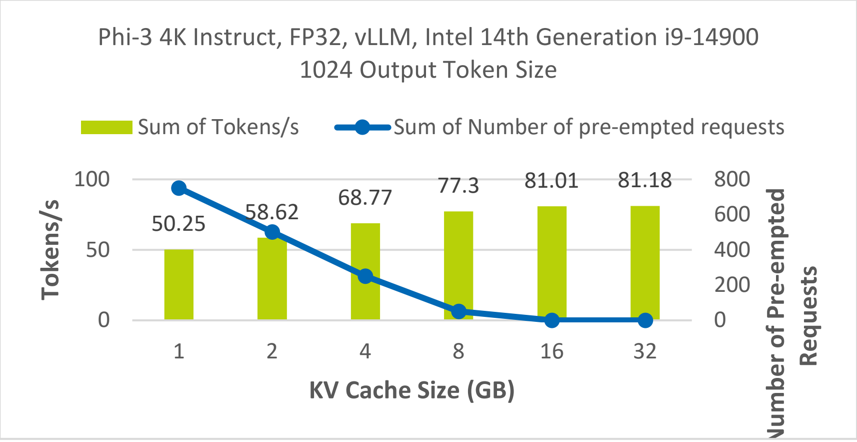


Figure 3 Performance for Phi-3 mini 4K-Instruct model on Intel® 14th Generation i9-14900

For the Phi-3 4K-Instruct model on Intel® 14th Generation 19-14900 is able to achieve up to 81.01 tokens per second using a KV cache size of 16 GB. In this case the number of pre-empted requests remains at zero.

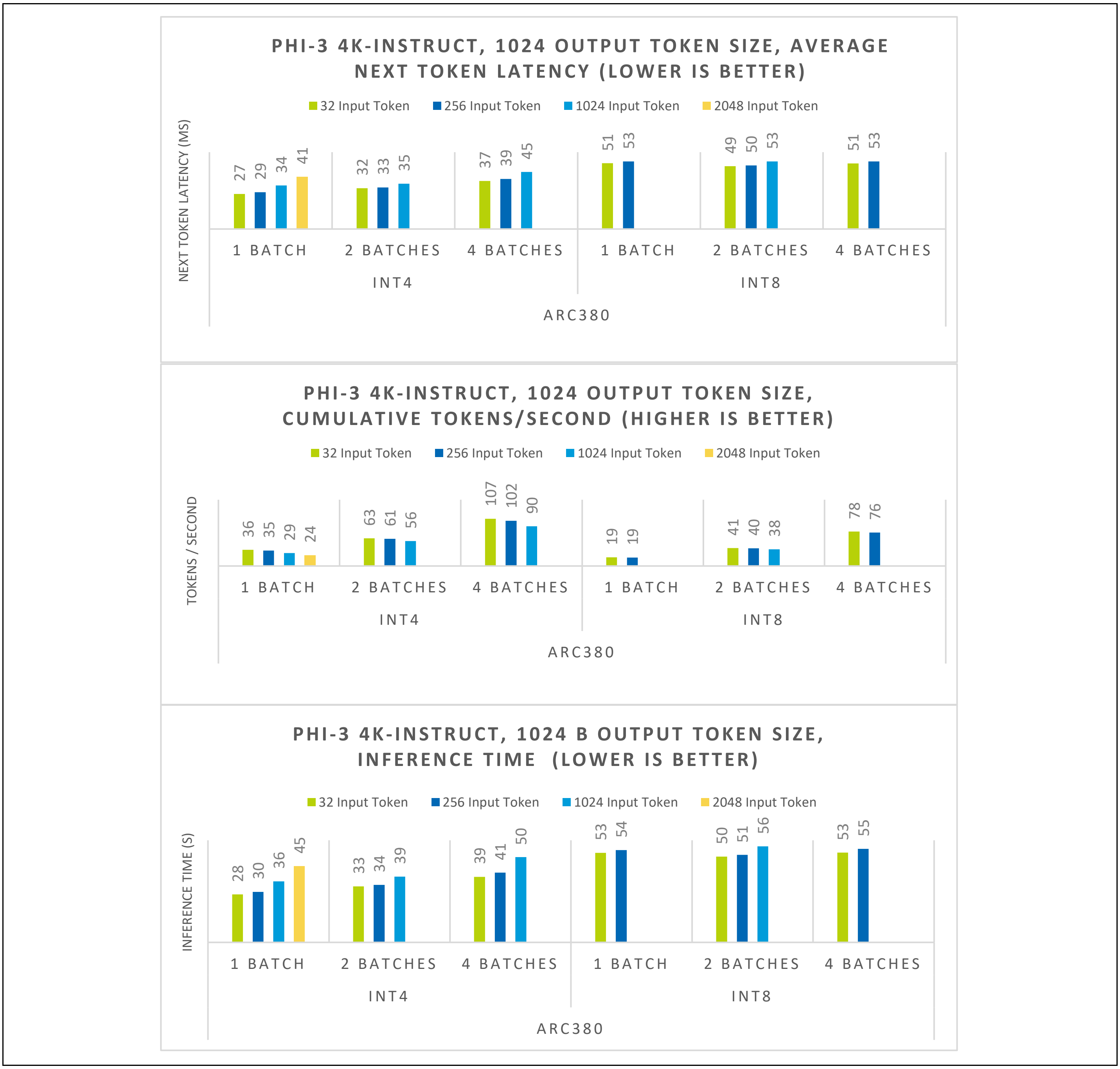


Figure 4 Performance for Phi-3 mini 4K-Instruct model on Intel® Arc A380 GPU

For the Phi-3 4K-Instruct model, a single Intel® Arc™ A380 GPU is able to achieve a next token latency down to 27 ms for a single batch size using an input token size of 1024 with INT4 precision. For the Phi-3 4K-Instruct model, a single Intel® Arc™ A380 GPU can achieve up to 107 tokens per second with a batch size of 4 using an input token size of 32 with INT4 precision. For the Phi3 4K-Instruct model, a single Intel® Arc™ A380 GPU can achieve an inference time down to 28 sec for a single batch size using an input token size of 32 with INT4 precision.

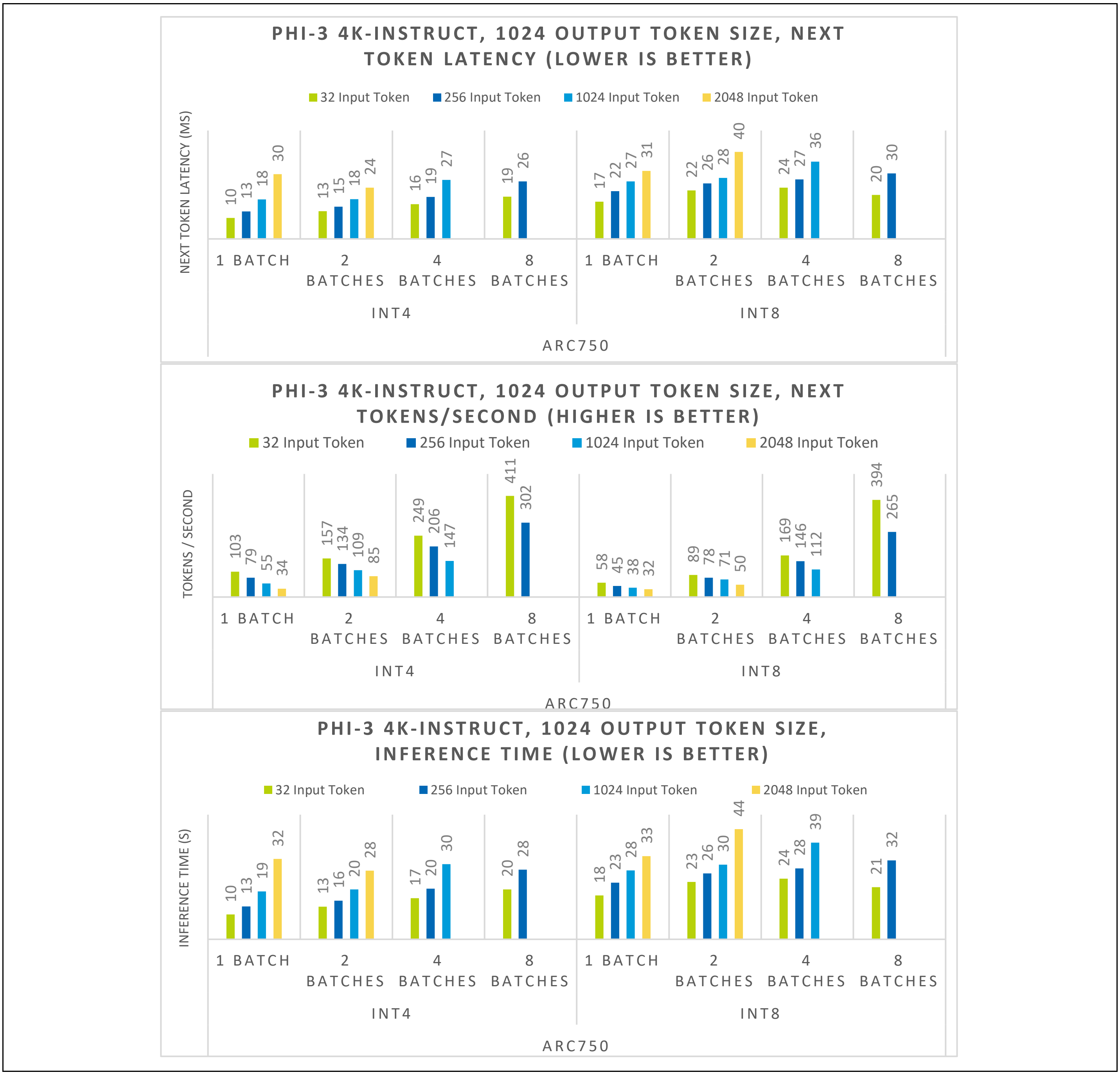


Figure 5 Performance for Phi-3 mini 4K-Instruct model on Intel® Arc A750 GPU

For the Phi-3 4K-Instruct model, a single Intel® Arc™ A750 GPU can achieve a next token latency down to 10 ms for a single batch size using an input token size of 32 with INT4 precision. For the Phi-3 4K-Instruct model, a single Intel® Arc™ A750 GPU can achieve up to 411 tokens per second with a batch size of 8 using an input token size of 32 with INT4 precision. For the Phi-3 4K-Instruct model, a single Intel® Arc™ A750 GPU can achieve an inference time down to 10 sec for a single batch size using an input token size of 32 with INT4 precision.

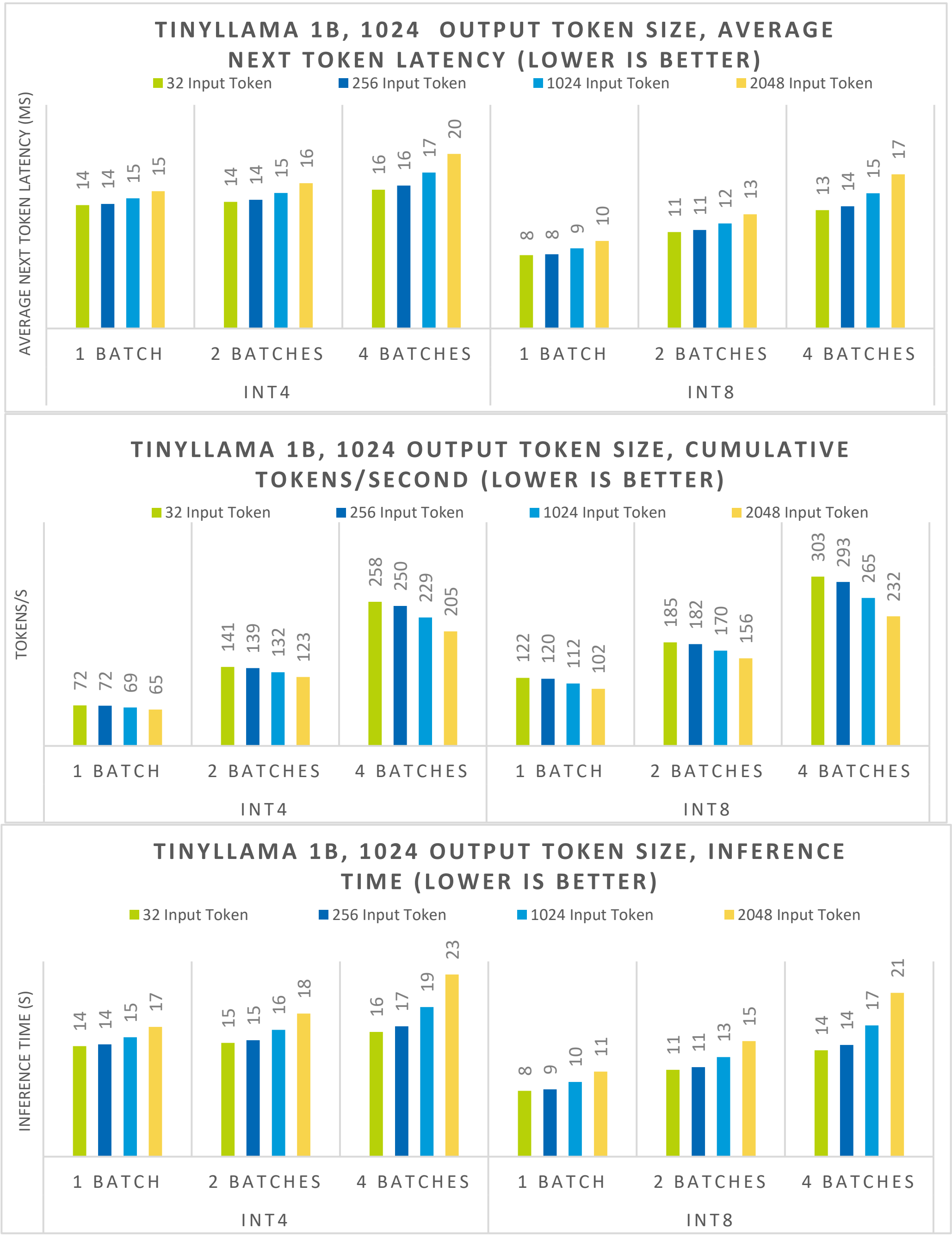
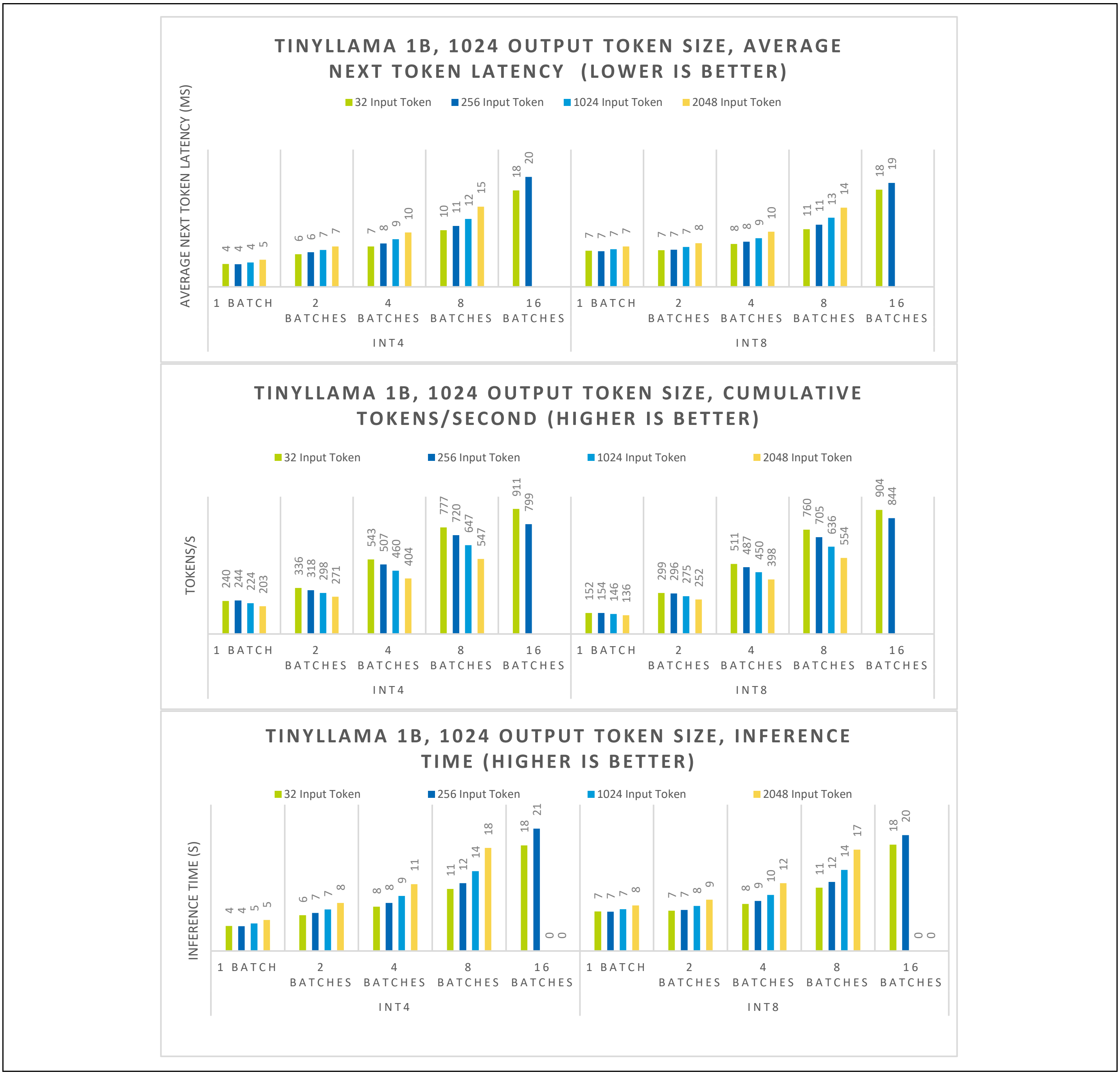


Figure 6 Performance for TinyLlama model on Intel® Arc™ A380 GPU

For the TinyLlama model, a single Intel® Arc™ A™ GPU can achieve a next token latency down to 8 ms for a single batch size using an input token size of 256 with INT8 precision. For the TinyLlama model, a single Intel® Arc™ A380 GPU can achieve up to 303 tokens per second with a batch size of 4 using an input token size of 32 with INT8 precision. For the TinyLlama model, a single Intel® Arc™ A380 GPU can achieve an inference time down to 8 sec for a single batch size using an input token size of 32 with INT8 precision.



TINYLLAMA 1B, 1024 OUTPUT TOKEN SIZE, AVERAGE NEXT TOKEN LATENCY (LOWER IS BETTER)

TINYLLAMA 1B, 1024 OUTPUT TOKEN SIZE, CUMULATIVE TOKENS/SECOND (HIGHER IS BETTER)

TINYLLAMA 1B, 1024 OUTPUT TOKEN SIZE, INFERENCE TIME (HIGHER IS BETTER)

Figure 7 Performance for TinyLlama model on Intel® Arc A750 GPU

For the TinyLlama model, a single Intel® Arc^{TM} A750 GPU can achieve a next token latency down to 4 ms for a single batch size using an input token size of 1024 with INT4 precision. For the TinyLlama model, a single Intel® Arc^{TM} A750 GPU can achieve up to 911 tokens per second with a batch size of 16 using an input token size of 32 with INT4 precision. For the TinyLlama model, a single Intel® Arc^{TM} A750 GPU can achieve an inference time down to 4 sec for a single batch size using an input token size of 256 with INT4 precision.

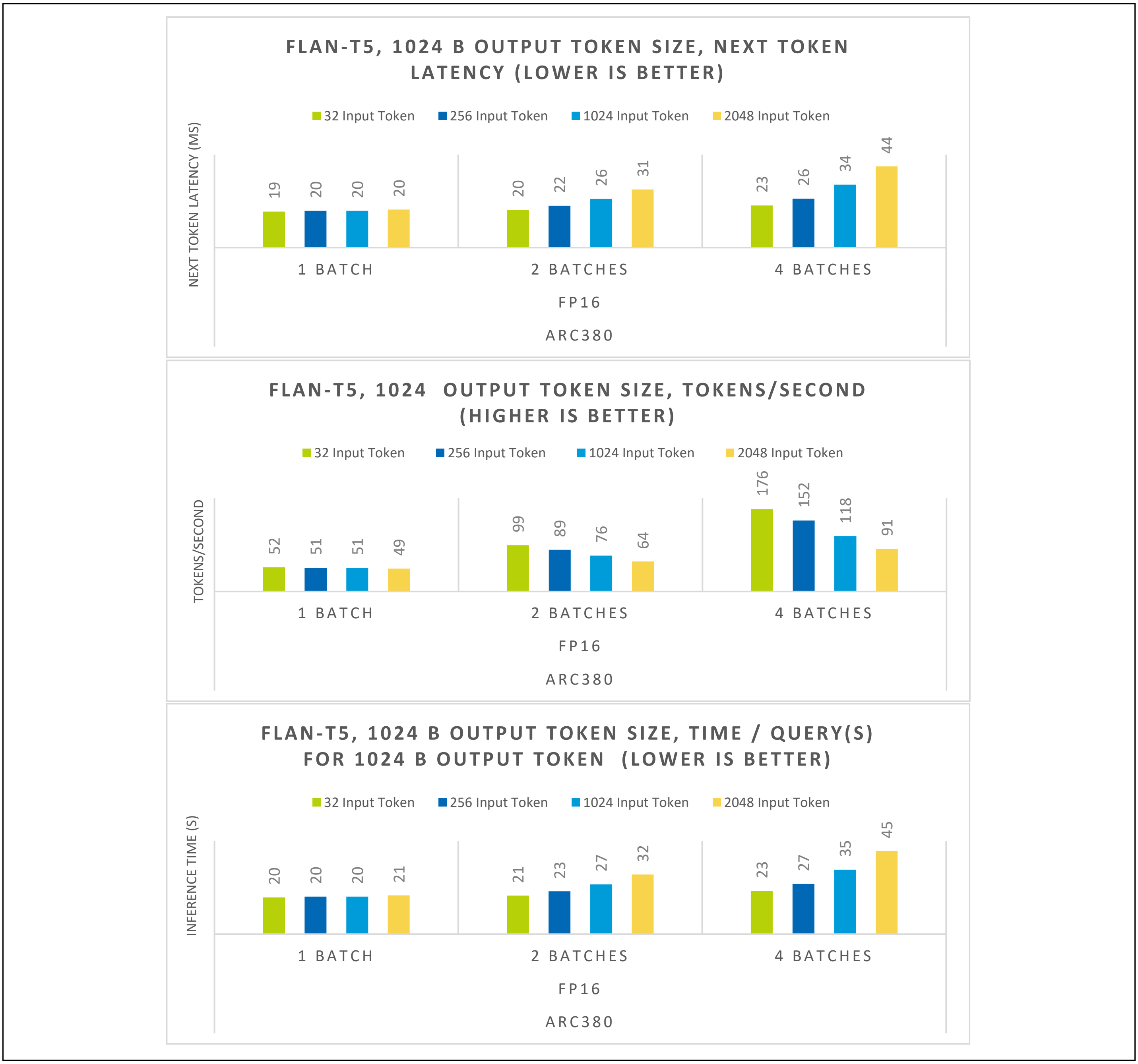


Figure 8 Performance for flan-t5 model on Intel® Arc™ A380 GPU

For the flan-t5 model, a single Intel® Arc™ A380 GPU can achieve a next token latency down to 19 ms for a single batch size using an input token size of 2048 with FP16 precision. For the flan-t5 model, a single Intel® Arc™ A380 GPU can achieve up to 176 tokens per second with a batch size of 4 using an input token size of 32 with FP16 precision. For the flan-t5 model, a single Intel® Arc™ A380 GPU can achieve an inference time down to 20 sec for a single batch size using an input token size of 32 with FP16 precision.

4.2.2 GenAI benchmarks on Supermicro SYS-E403-13E-FRN2T Platform

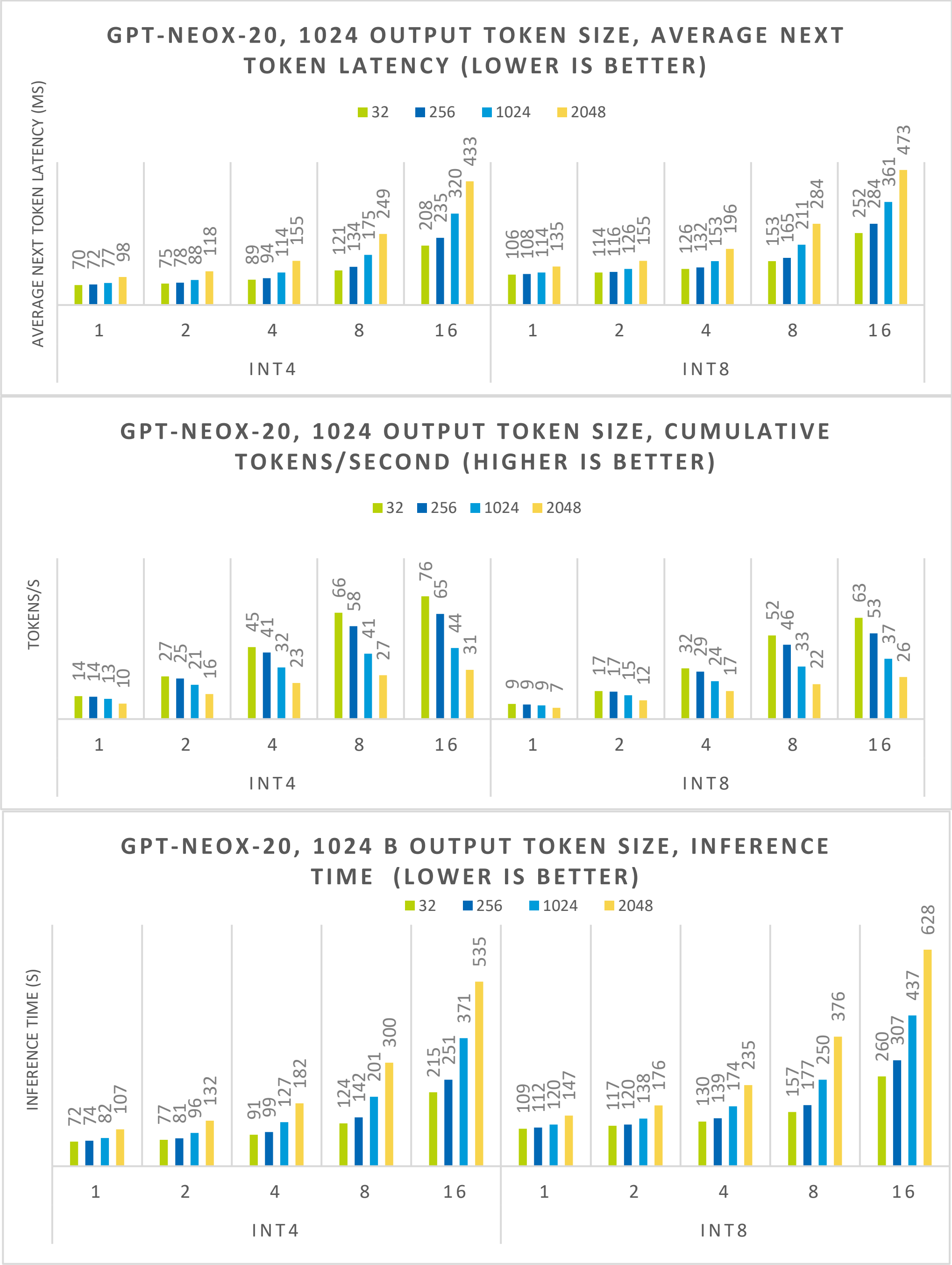


Figure 9 Performance for GPT-NEOX 20B Model on Supermicro E403 (CPU only)

For the GPT-Neox-20 model, a single socket 6538N CPU can achieve a next token latency down to 70 ms for a single batch size using an input token size of 32 with INT4 precision with an inference time of 72 sec.

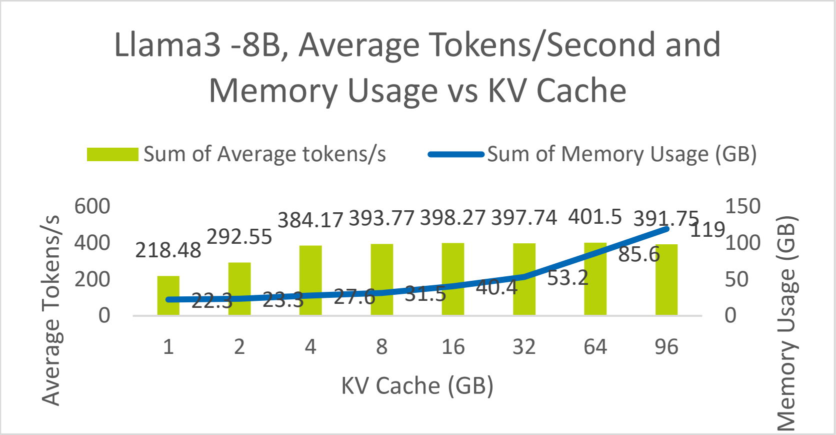


Figure 10 vLLM-IPEX-LLM Performance with Llama 3 8B Model on Supermicro E403 (CPU only)

For the Llama3-8b model with BF16 precision, using vLLM framework, a single socket 6538N CPU can achieve 384 cumulative tokens/s with 4GB of KV Cache.

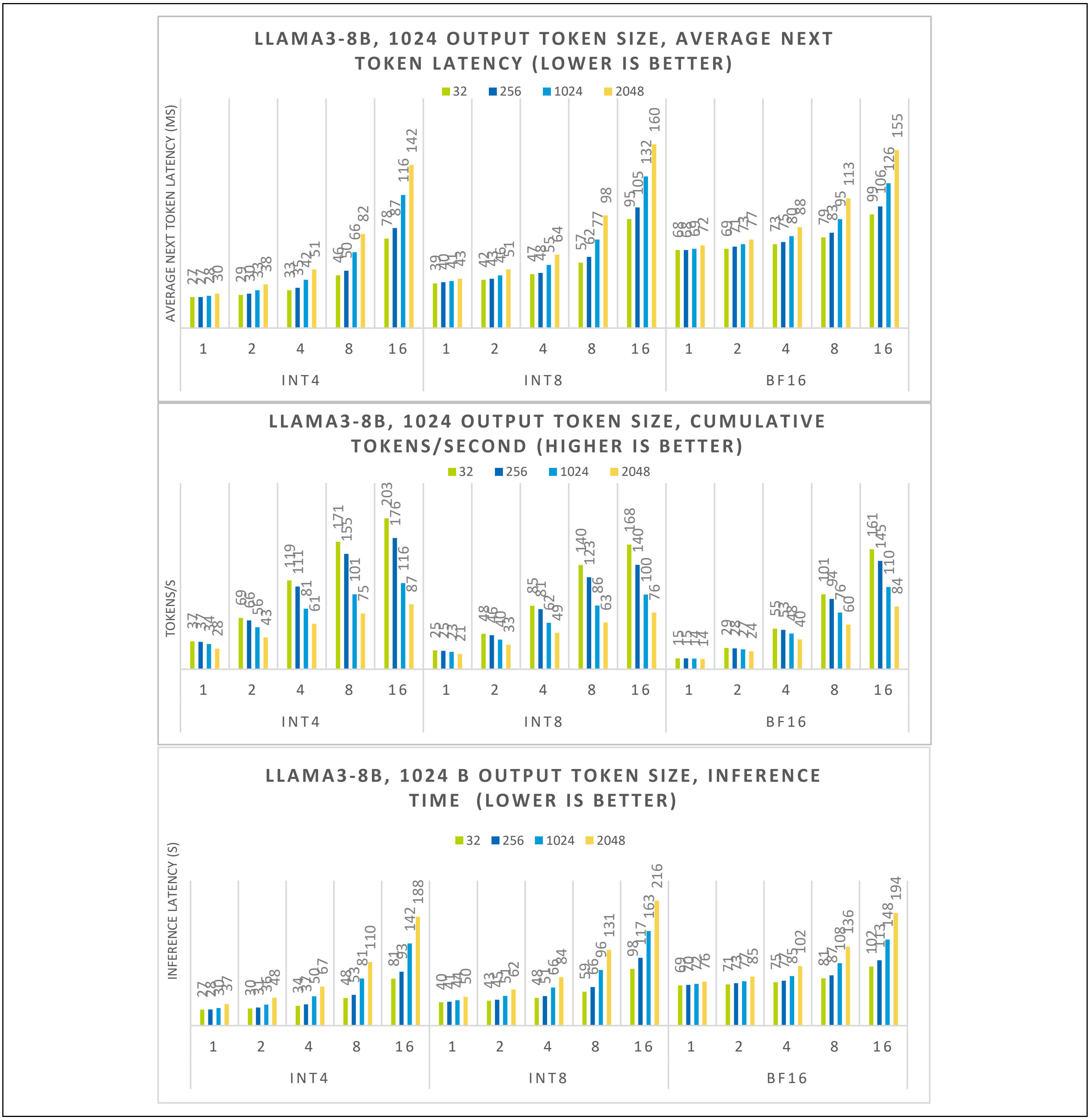


Figure 11 Performance for Llama 3 8B Model on Supermicro E403 (CPU only)

For the Llama3-8b model, a single socket 6538N CPU can achieve a next token latency down to 27 ms for a single batch size using an input token size of 256 with INT4 precision with an inference time of 28 sec.

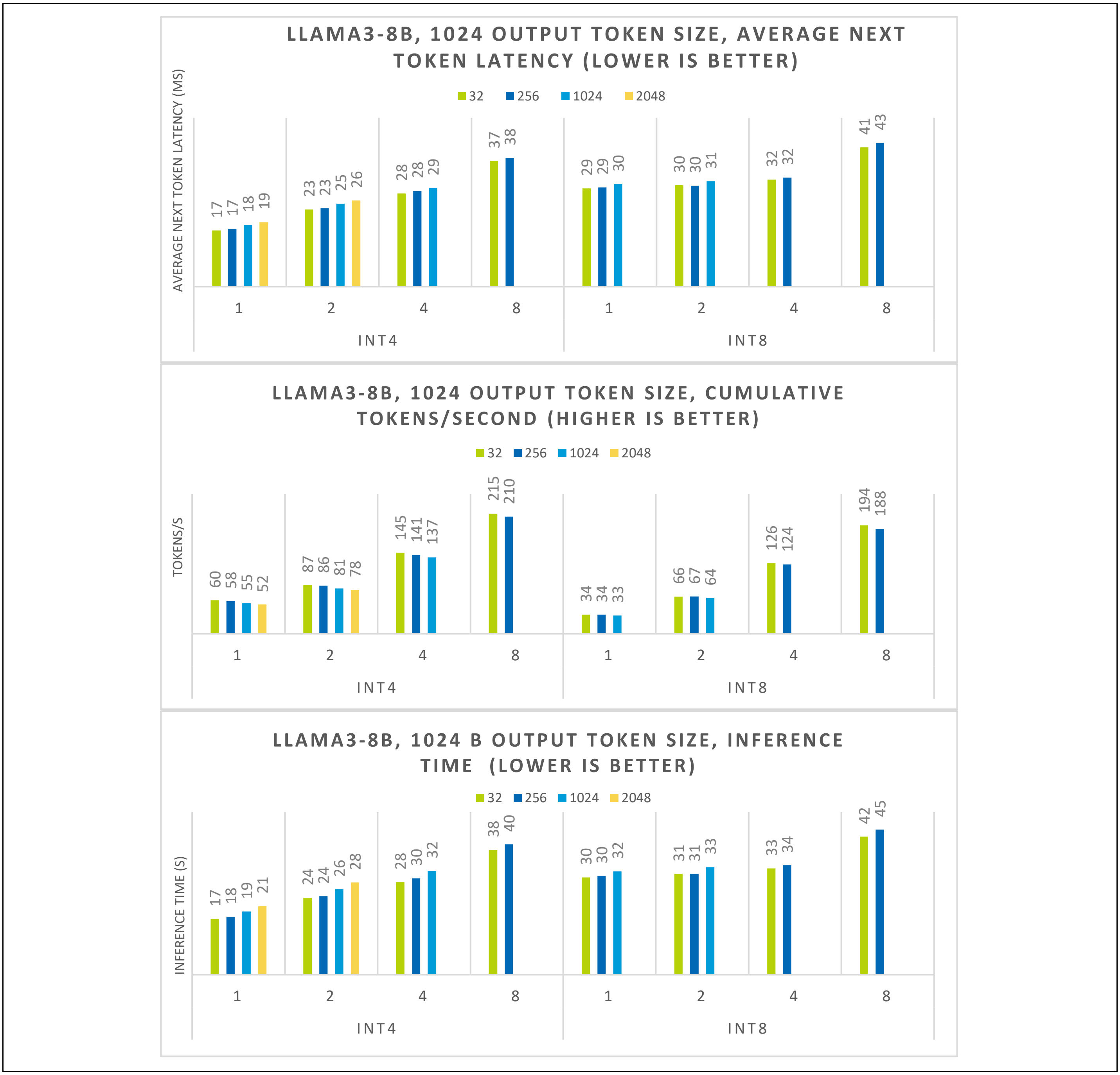


Figure 12 Performance for Llama 3 8B Model on Supermicro E403 (1x Intel® Flex 170)

For the Llama3-8b model, a single Intel® Data Centre GPU Flex 170 can achieve a next token latency down to 17 ms for a single batch size using an input token size of 256 with INT4 precision with an inference time of 18 sec. The FP16 model can't be loaded to a single Flex 170 due to memory restrictions on the GPU.

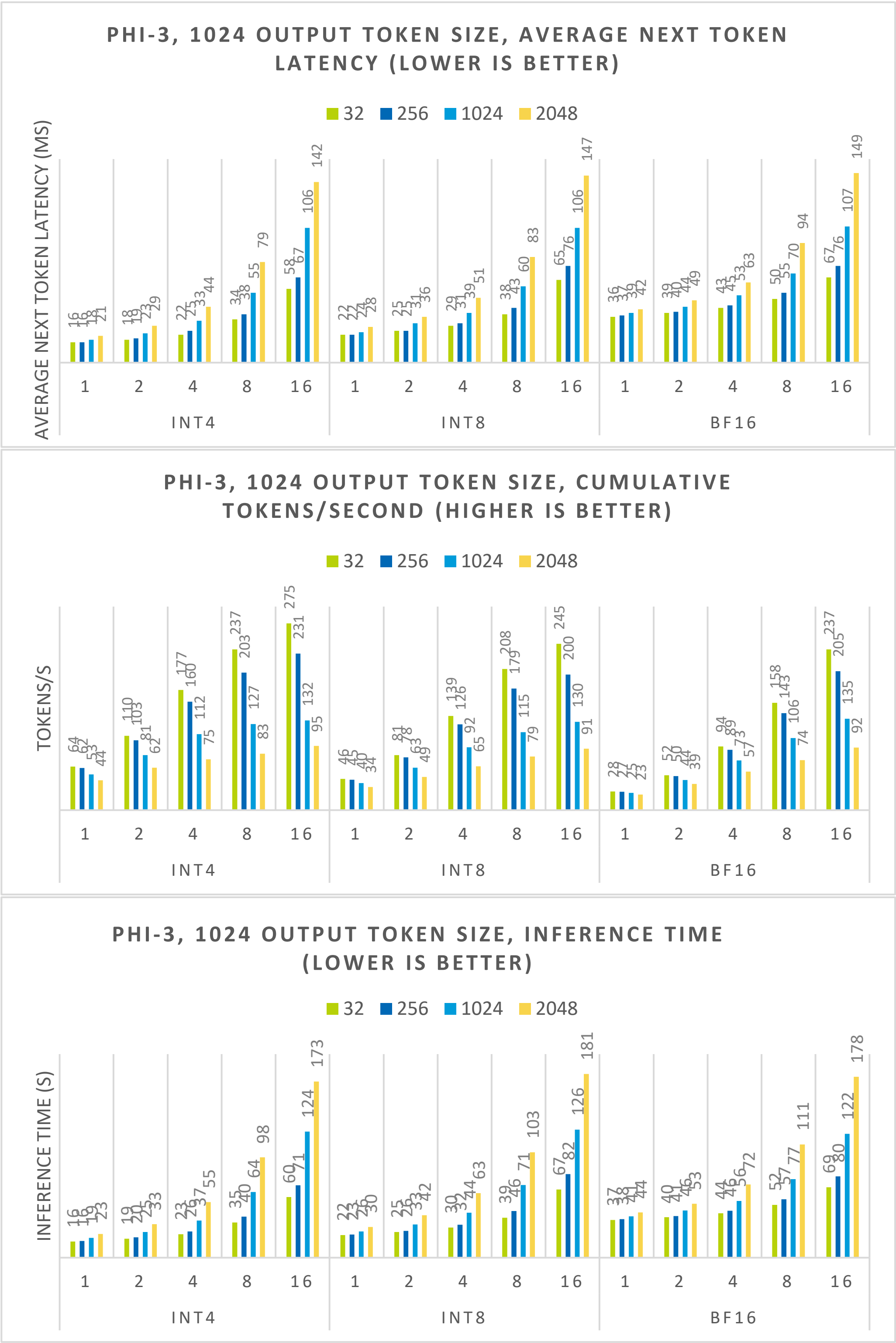


Figure 13 Performance for Phi-3-mini-4k-instruct Model on Supermicro E403 (CPU Only)

For the Phi3 mini 4k instruct model, a single socket 6538N CPU can achieve a next token latency down to 16 ms for a single batch size using an input token size of 32 with INT4 precision with an inference time of 16 sec.

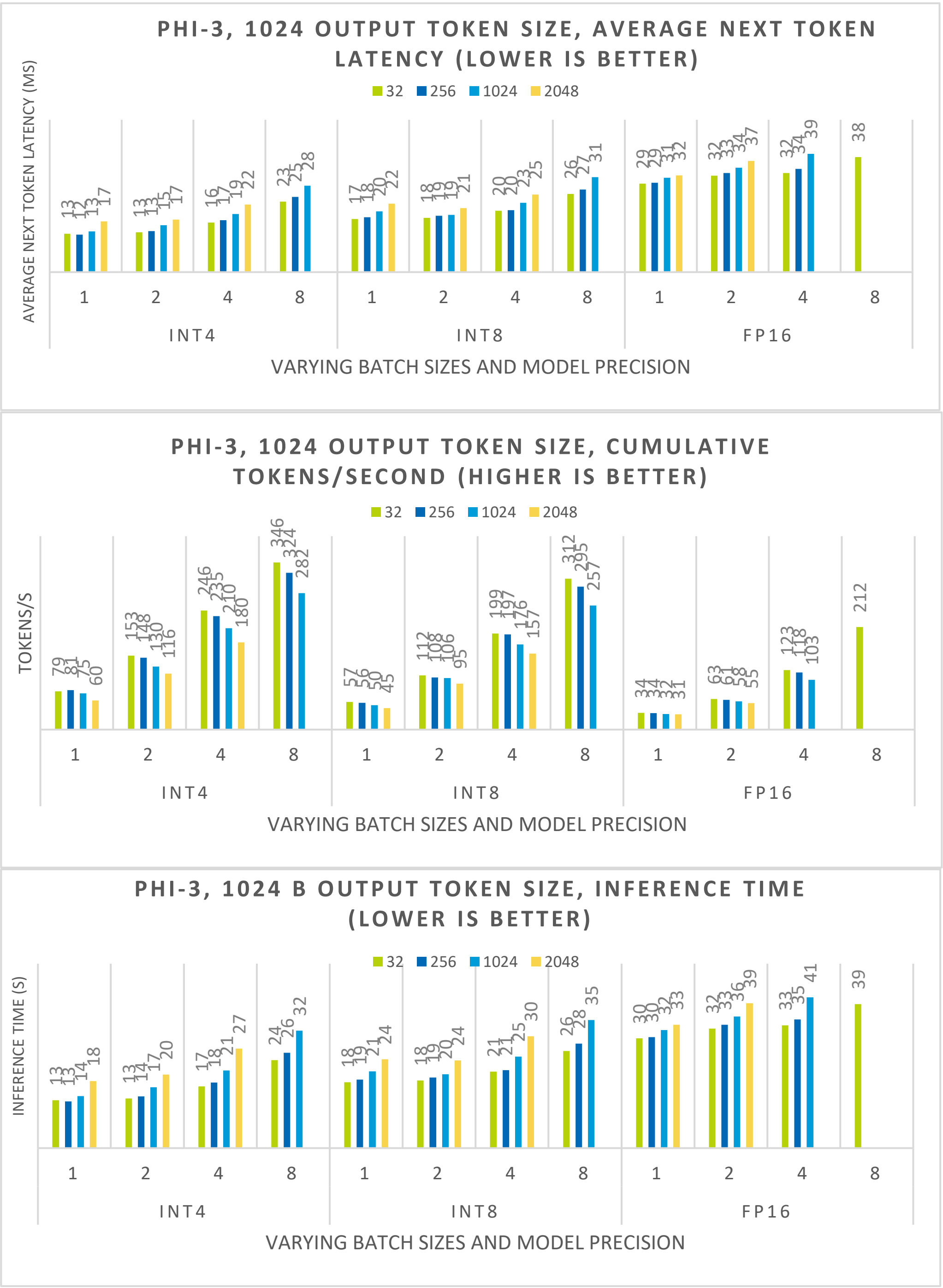


Figure 14 Performance for Phi-3-4k-Mini Model on Supermicro E403 (1x Intel® Flex 170)

For the Phi-3 mini 4k instruct model, a single Intel® Data Centre GPU Flex 170 can achieve a next token latency down to 13 ms for a single batch size using an input token size of 1024 with INT4 precision with an inference time of 14 sec.

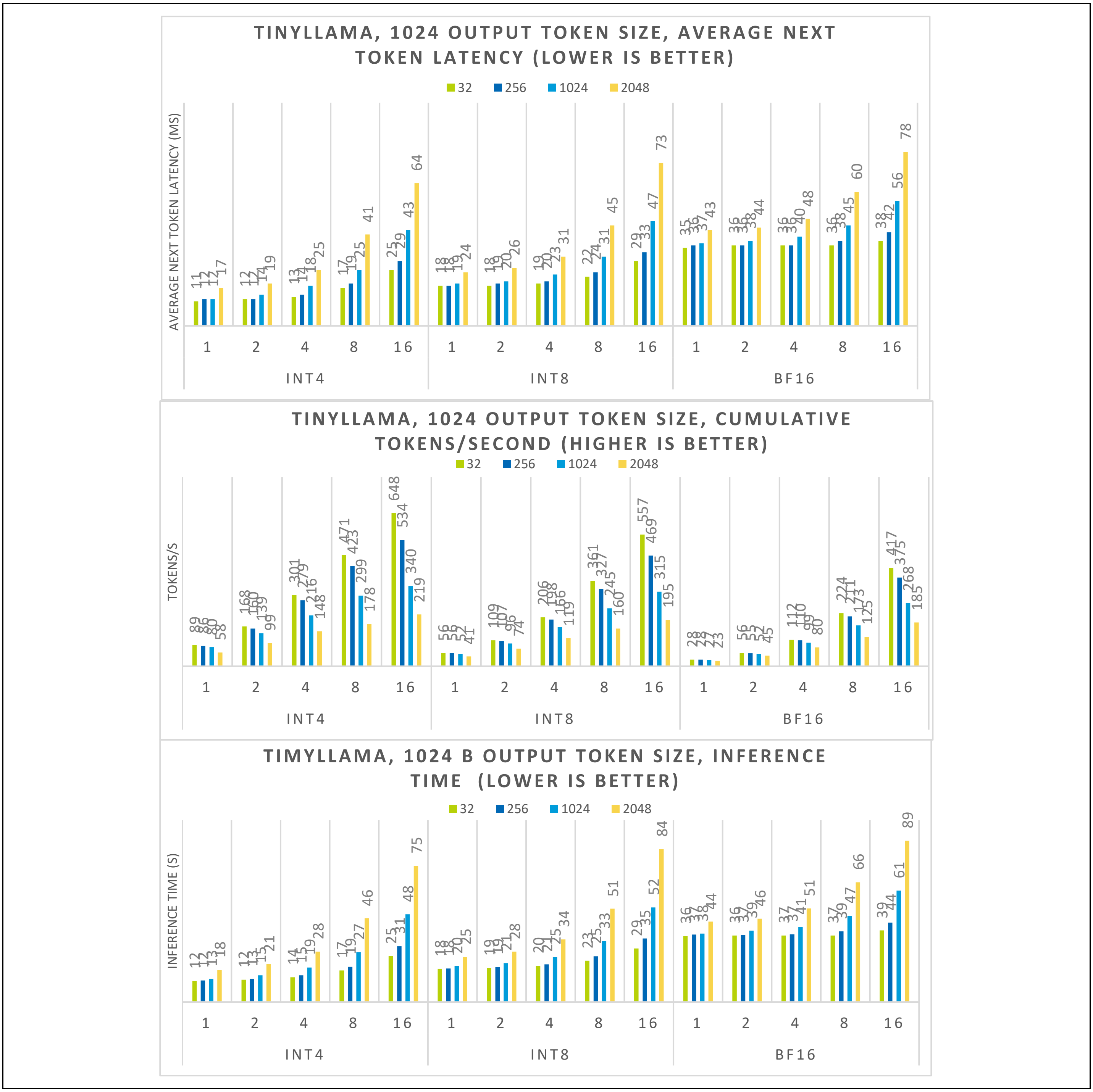


Figure 15 Performance on TinyLlama Model on Supermicro E403 (CPU Only)

For the Tinyllama model, a single socket 6538N CPU can achieve a next token latency down to 11 ms for a single batch size using an input token size of 32 with INT4 precision with an inference time of 12 sec.

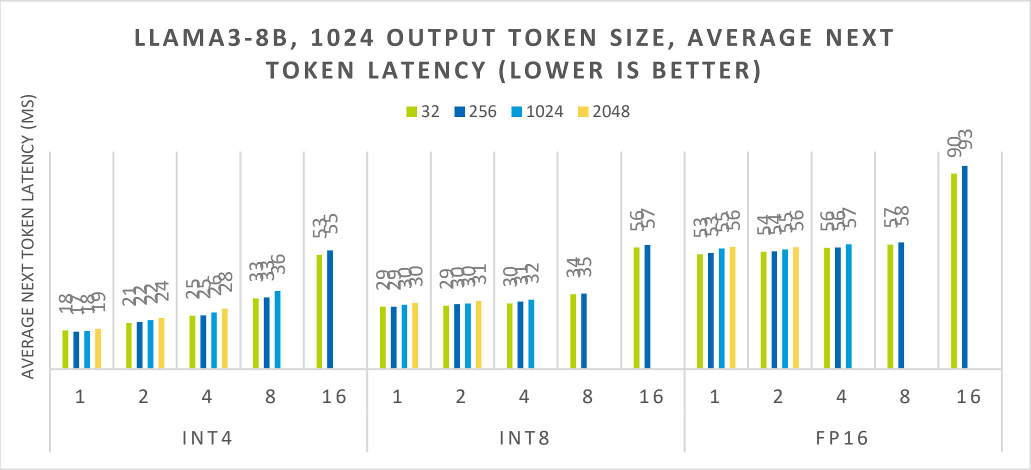
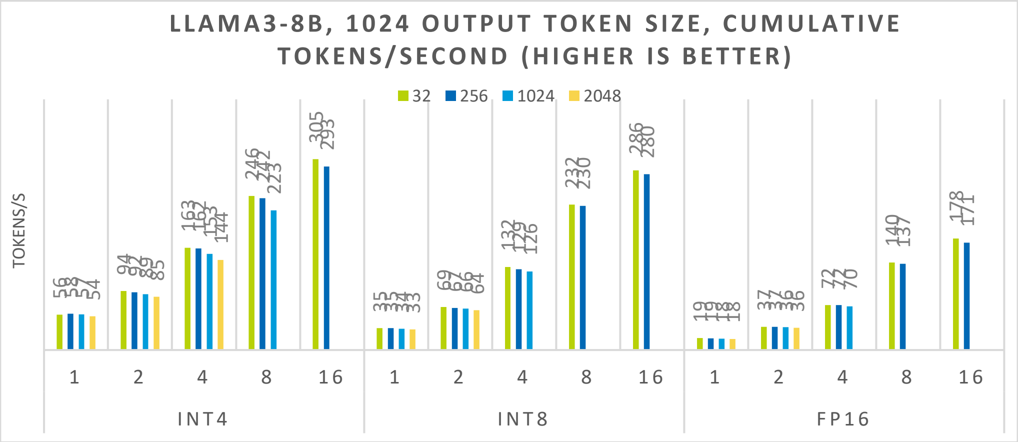
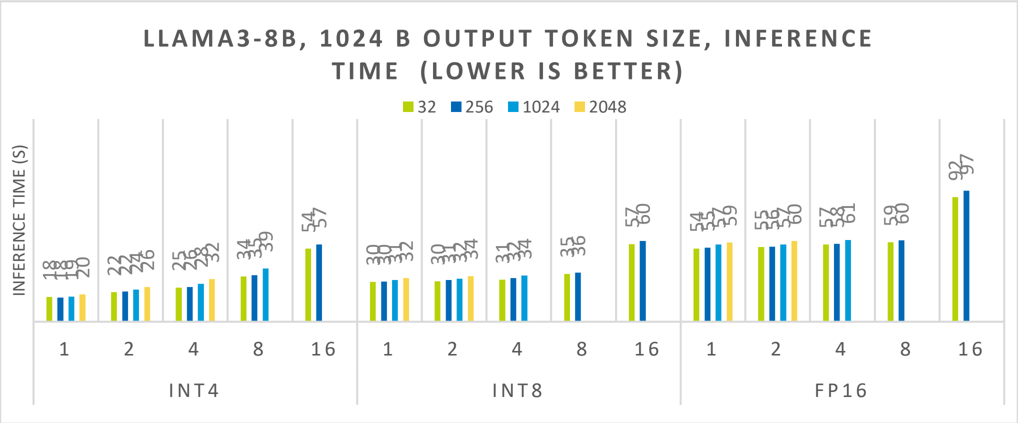


Figure 16 Llama 3 8B Performance on 2 x Flex 170 with Pipeline Parallel Configuration

For the Llama3-8b model, two Data Centre Intel® Data Centre GPU Flex 170 can run the model in FP16 precision. The model can achieve a next token latency of 53 ms for a single batch size using an input token size of 256 with FP16 precision with an inference time of 55 sec.



Note: The performance scaling is not linear while performing distributed inference using pipeline parallelism due to PCIe bandwidth limitations. If the model fits on one GPU, the best performance may be achieved by running 2 instances of the model on individual GPUs rather than distributing it over multiple cards.



5 Summary

The Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS111-AD-WRN and Intel® Arc™ GPUs along with single socket 5th Gen Intel® Xeon® Scalable processors with multiple Intel® Data Center Flex GPUs on Supermicro SYS-E403-13E-FRN2T addresses the capabilities for AI Inference offering the value propositions detailed within the tables below.

For the Intel ® AI Edge Systems Verified Reference Blueprint with Supermicro SYS-111-AD-WRN and Supermicro SYS-E403-13E-FRN2T, the system should be able to deliver results as shown in the tables below a baseline to the expected performance of this solution.

Model	Precision	Configuration	Input-Output Token Size	Batch Size	Number of Tokens / Second
vLLM Phi-3-mini	BF16	Intel® Core i9-14900	Variable	Variable	81.01
Phi-3-min	INT4	1x Arc™ A380	32-1024	4	107
	INT4	1x Arc™ A750	32-1024	8	411
TinyLlama	INT8	1x Arc™ A380	32-1024	4	303
	INT4	1x Arc™ A750	32-1024	16	911
flan-t5	FP16	1x Arc™ A380	32-1024	4	176

Table 7 GEN AI Use Case with Supermicro SYS111-AD-WRN

Model	Precision	Configuration	Input-Output Token Size	Batch Size	Number of Tokens / Second
GPT-Neox-20b	INT4	1x Intel® Xeon® Gold 6538N Processor	32-1024	1	14
vLLM Llama3 8B	BF16	1x Intel® Xeon® Gold 6538N Processor	Variable	Variable	384
Llama3 8B	INT4	1x Intel® Xeon® Gold 6538N Processor	32-1024	8	171
	INT4	1x Intel® Data Centre GPU Flex 170	32-256	8	215
	INT4	2x Intel® Data Centre GPU Flex 170	32-256	8	246
Phi-3-mini	INT4	1x Intel® Xeon® Gold 6538N Processor	32-1024	8	208
	INT4	1x Intel® Data Centre GPU Flex 170	32-256	8	346
TinyLlama	INT4	1x Intel® Xeon® Gold 6538N Processor	32-1024	16	695

Table 8 GEN AI Use Case with Supermicro SYS-E403-13E-FRN2T

This Configuration combined with architectural improvements, feature enhancements, and integrated Accelerators with high memory and IO bandwidth, provides a significant performance and scalability advantage in support for today's AI workload.

The Intel® Core and Intel® Xeon Scalable Processor platforms are optimized for AI intensive workloads coupled with Intel® Arc™ and Intel® Data Center Flex GPUs.

1 Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure. Your costs and results may vary. Intel technologies may require enabled hardware, software or service activation. © Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. *Other names and brands may be claimed as the property of others.

2 Configuration

Test by Intel August 6th 2024
See Hardware Configuration – Table 1