85TH HPC USER FORUM

BENCHMARK OF REAL-WORLD AI INFERENCING WORKLOADS ON THE INTEL GAUDI2 AI ACCELERATOR AND AN EXAMPLE OF MULTI-MODAL AI

GIANLUCA LONGONI, PH.D., P.E.
EXECUTIVE TECHNICAL DIRECTOR AI/HPC
FEDDATA TECHNOLOGY SOLUTIONS
GIANLUCA.LONGONI@FEDDATA.COM

ARGONNE NATIONAL LABORATORY
SEPTEMBER 5TH 2024

FEDDATA

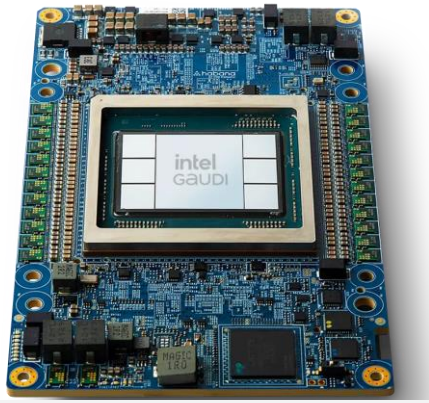
Overview

- Present and discuss the results on AI inferencing workloads for the Gaudi2 and Nvidia accelerators
- “Real-world” applications refer to AI inferencing workloads to serve multiple concurrent users on RAG AI-like applications (RAG = Retrieval Augmented Generative AI)
- Goal is to present an holistic comparison (temperature profiles, power draw, etc…)
- Two benchmarks will be analyzed for the Intel Gaudi2 HPU, Nvidia A100, and L40S GPUs
  - Latency analysis
  - Serving concurrent requests
- Short demo on multi-modal AI
- This presentation does not represent an endorsement of the technologies presented herein

FEDDATA

Hardware Platform - Intel Gaudi2 Accelerator

Intel Gaudi 2 AI

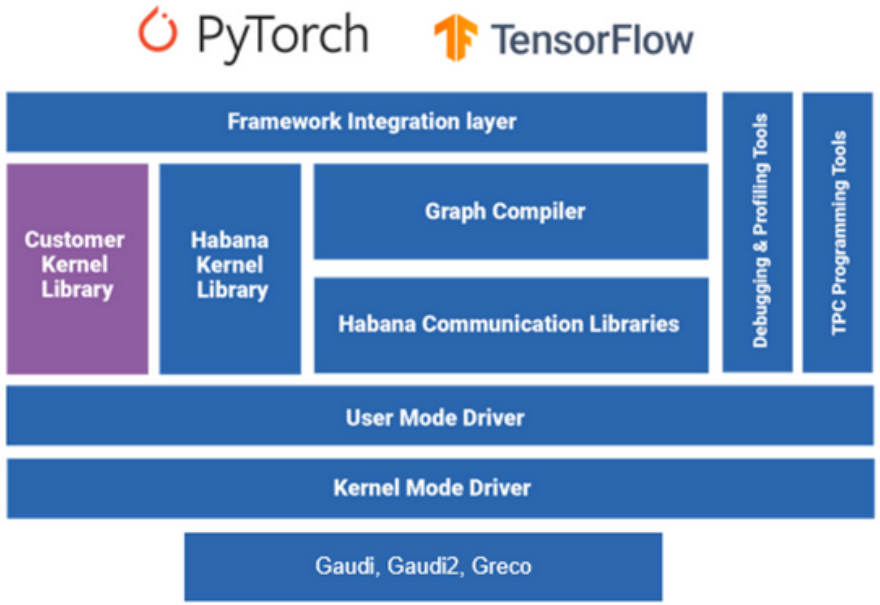


- 7nm process technology
- Heterogeneous compute
- 24 Tensor Processor Cores
- Dual matrix multiplication engines
- 24 100 Gigabit Ethernet integrated on chip
- 96 GB HBM2E memory on board
- 48 MB SRAM
- Integrated Media Control

- Large onboard memory footprint
- Ease of adoption - Ethernet network backbone

SynapseAI®

Development Ecosystem



FEDDATA

Software Stack - AI Inferencing Engine

vLLM is a fast and easy-to-use library for LLM inference and serving

- State-of-the-art serving throughput
- Efficient management of attention key and value memory with **PagedAttention**
- Continuous batching of incoming requests
- Fast model execution with CUDA/HIP graph
- Quantization: GPTQ, AWQ, SqueezeLLM, FP8 KV Cache
- Optimized CUDA kernels
- Ported to Intel Gaudi2 architecture
- Seamless integration with popular HuggingFace models
- Support NVIDIA GPUs and AMD GPUs, as well as Intel Gaudi2

FEDDATA

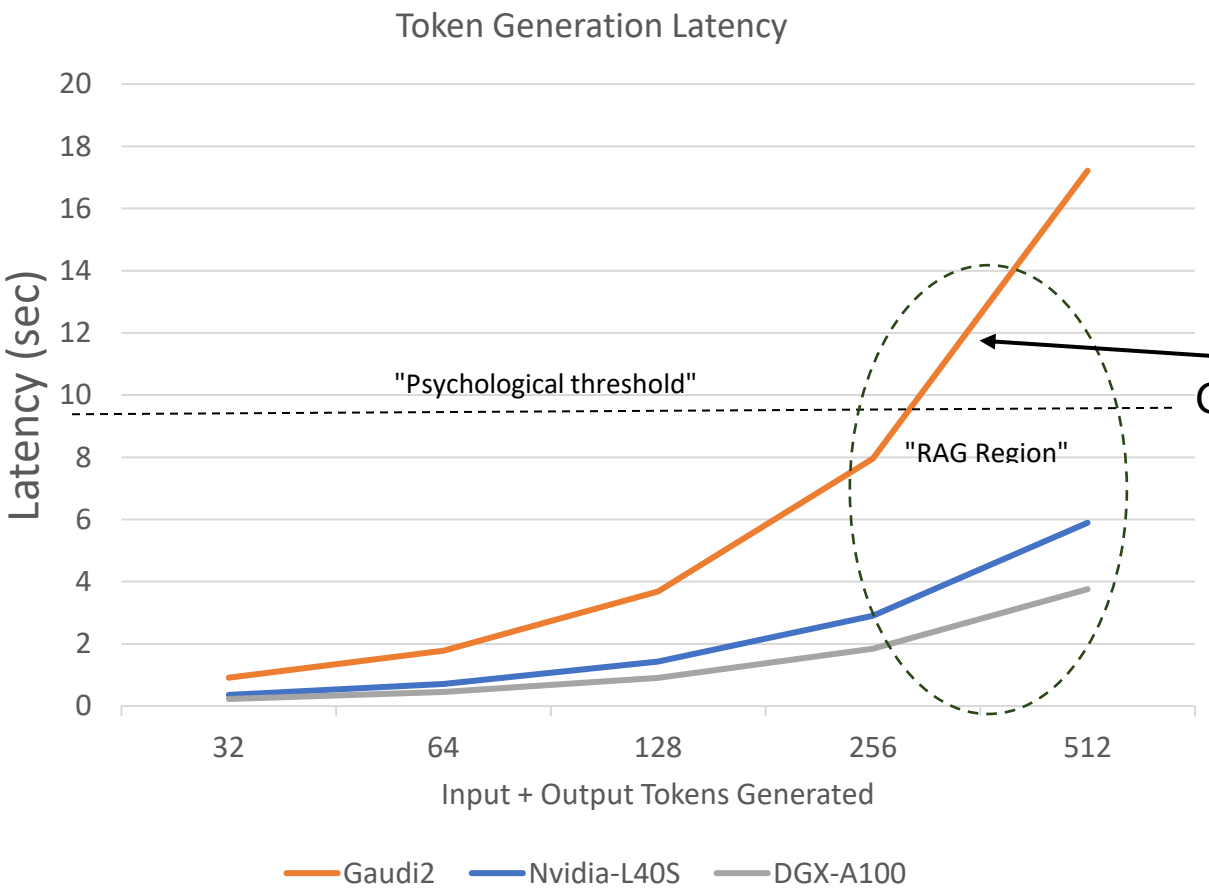
Inferencing Benchmarks

- Token Generation Latency
  - Benchmark the latency in seconds of processing a single batch of requests
  - Requests are defined as a user's request to the LLM to generate coherent text
- Online Serving Throughput
  - Benchmark a real-world scenario where the AI application is served to multiple concurrent users, i.e., 1000.
- Intel Gaudi2 environment
  - vLLM version 0.5.3 + Gaudi 1.16 Habana libraries*
- Nvidia Ennvironment
  - vLLM version 0.5.3 + CUDA 12.4 libraries*

*vLLM was compiled from source on target architectures

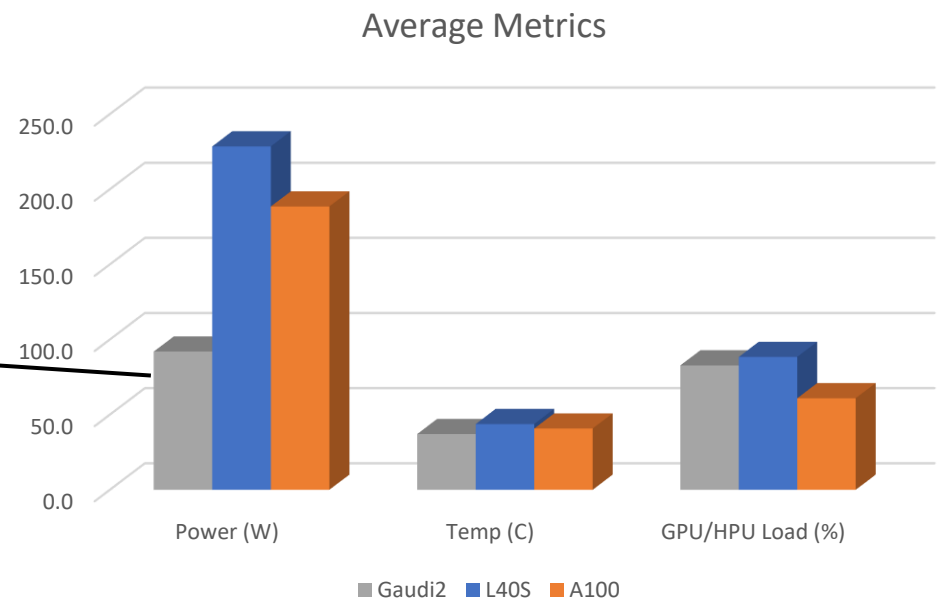
FEDDATA

Benchmark: Token Generation Latency



Lower is better…

Gaudi2

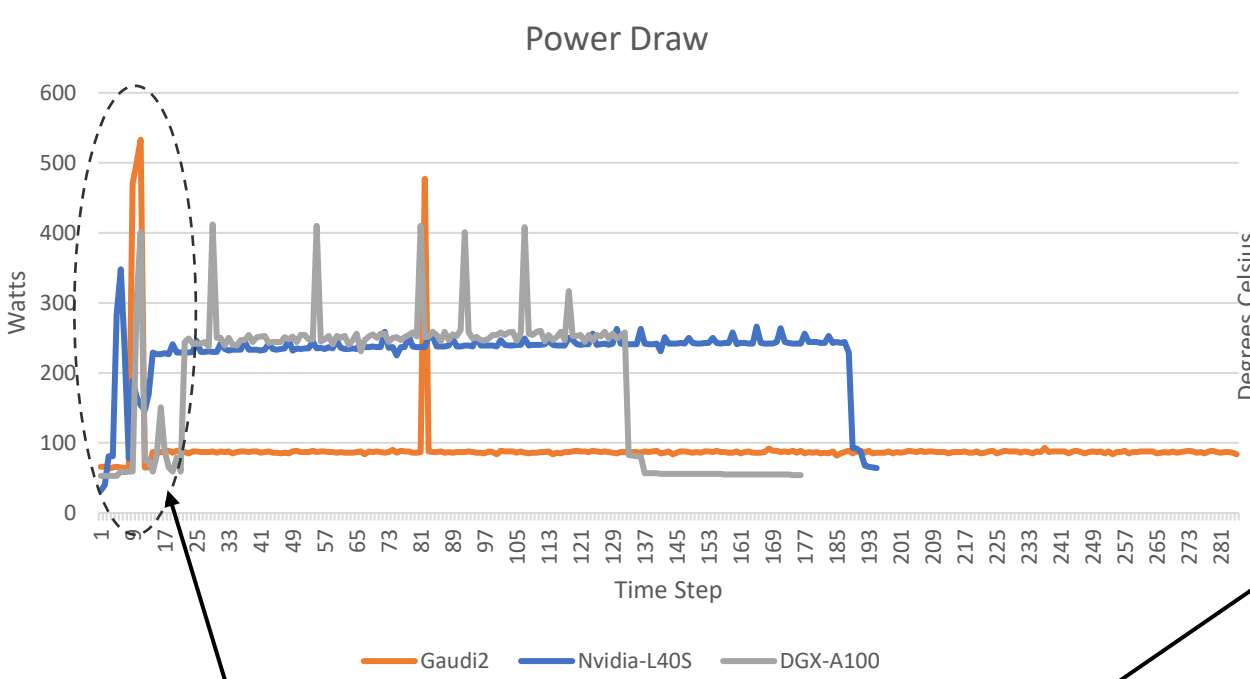


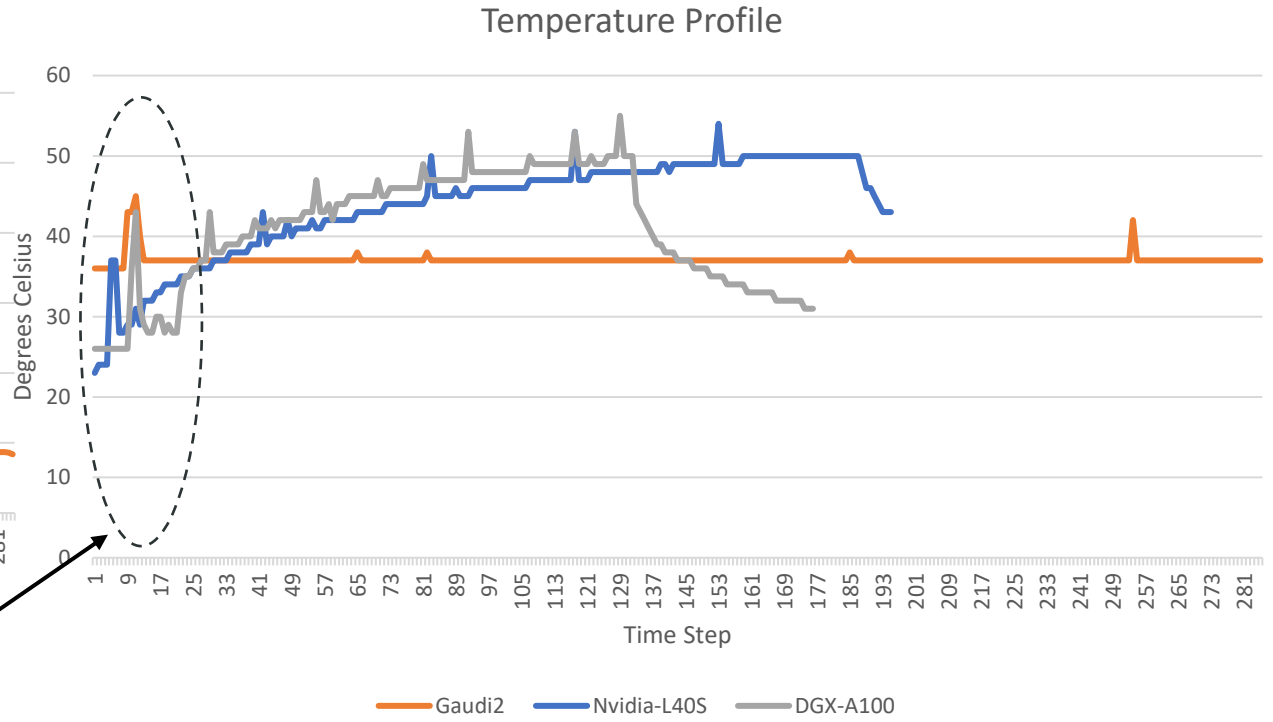
Nvidia accelerators -> low latency, better performance

Intel Gaudi2 -> higher latency, could be “acceptable” for RAG applications; however, the accelerator presents much lower power consumption and cooling requirements for comparable loads

FEDDATA

Benchmark: Token Generation Latency (cont'd)



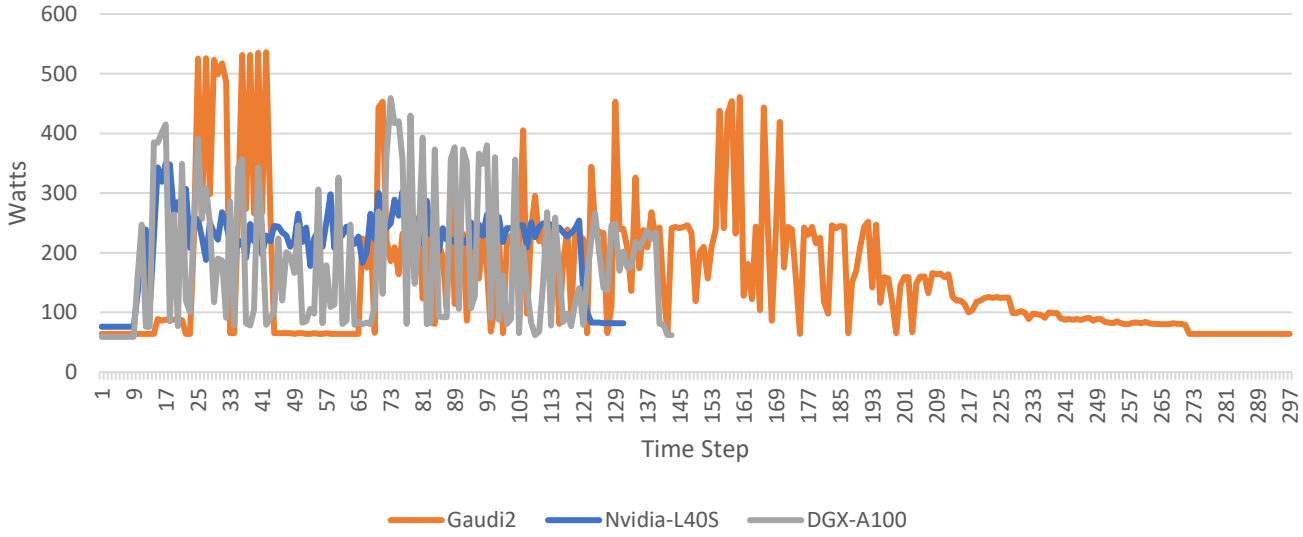


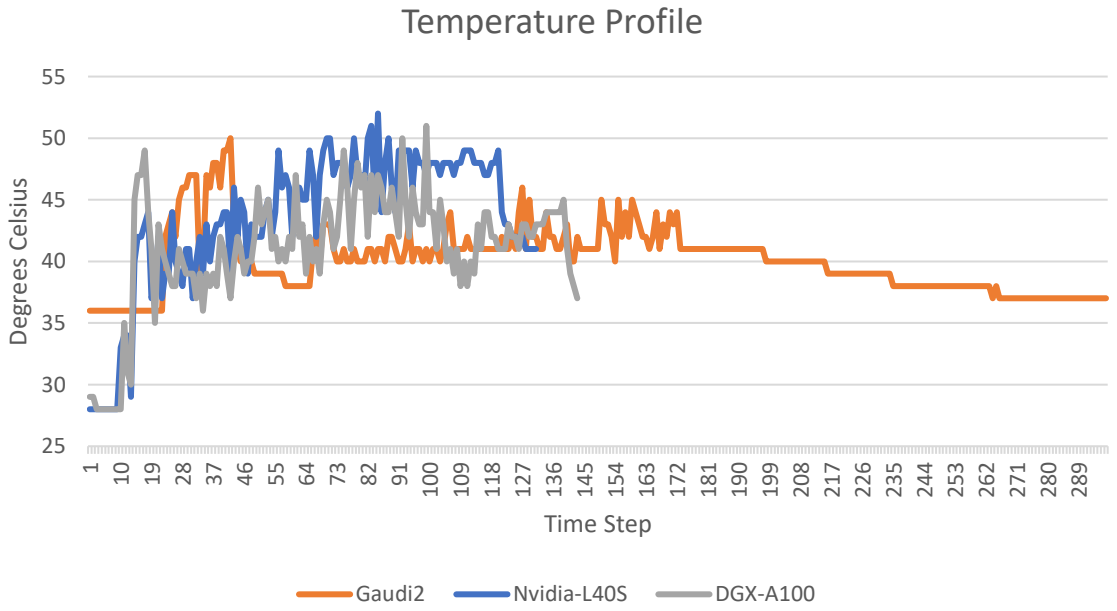
Large Language Model (LLM) initialization, weights are loaded in RAM for each accelerator

Benchmark: Serving Multiple Requests (users)

	Total tokens	Throughput			Mean Latency (ms)	Latency P99 (ms)
		Requests/s	Input tok/s	Output tok/s		
Gaudi2	441550	4	963.49	801.19	534.93	2023.28
Nvidia-L40S	441003	9.14	2203.45	1827.27	238.1	434.84
DGX-A100	440897	7.71	1859.85	1541.52	288.24	472.86

P99: 99th percentile latency. 99% of requests were processed faster than this value, and only 1% took longer





Conclusions

- Nvidia accelerators provide higher throughput (tokens/sec) with lower latency
- However…Intel Gaudi2 provides acceptable performance for RAG AI type of applications with lower power consumption and cooling requirements -> “AI at the edge”
- Positive experience in terms of using Habana libraries, PyTorch Lightning, porting of applications to Gaudi2 seems to be well addressed
- Multi-modal generative AI is a major topic of interest in this “second phase” (post 2023) of the generative AI explosion

FEDDATA

Future Work

- Benchmark the Intel Gaudi3 platform when released and compare to additional GPU accelerators
- Integrate benchmarks for AI inferencing and training in “Keystone”, an industrialized benchmark framework for High Performance Computing (HPC) systems
- Continue on the multi-modal AI path…

FEDDATA

Short video presentation on an “audio-RAG”, a step towards Multi-Modal AI applications

FEDDATA