

CHESAPEAKE LARGE-SCALE ANALYTICS CONFERENCE

BENCHMARK OF REAL-WORLD AI INFERRING WORKLOADS ON THE INTEL GAUDI 2 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

CHESAPEAKE LARGE-SCALE ANALYTICS CONFERENCE
NOVEMBER 5TH 2024

FEDDATA COMPANY PROPRIETARY

FEDDATA

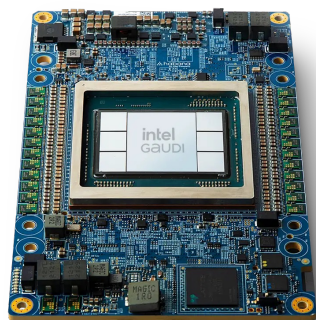
Overview

- Intel Gaudi 2 Accelerators
 - Present and discuss the results on AI inferencing workloads for the Gaudi 2 and Nvidia accelerators
 - We will focus on “real-world” applications: serve multiple concurrent user queries on a RAG AI platform (Retrieval Augmented Generative AI)
 - The goal is to present a holistic comparison (temperature profiles, power draw, etc...)
 - Latency analysis on token generation
 - Serving multiple concurrent requests
- Demo of OnPremGPT, an AI multimodal application using audio conversations
- A quick dive on the topic: Software Bill of Materials (SBOM)

This presentation does not represent an endorsement of the technologies presented herein.

Hardware Platform - Intel Gaudi 2 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

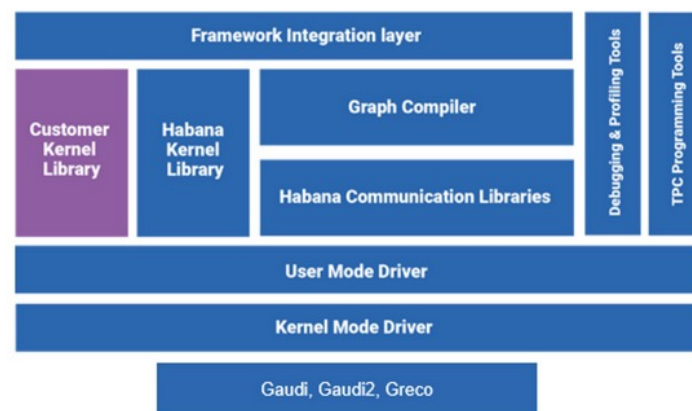
MMEs can be loosely described the Intel equivalent of Nvidia Tensor Cores

Supports all popular data types required for deep learning, including FP32, TF32, BF16, FP16, and FP8

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



Development Ecosystem



FEDDATA

Software Stack - AI Inferencing Engine



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

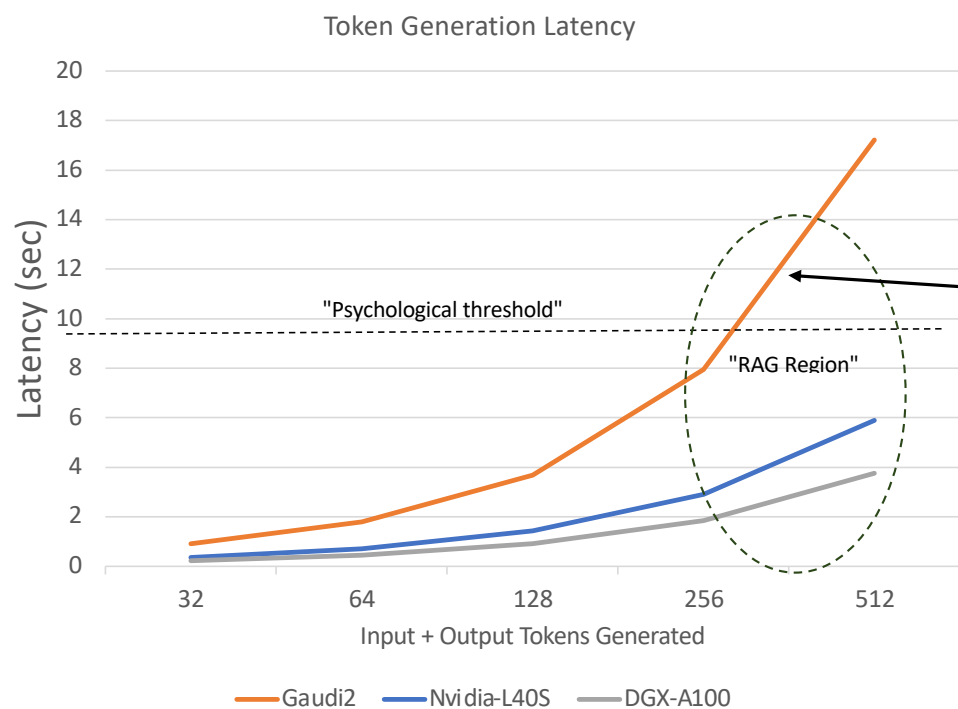
- State-of-the-art serving throughput
- Efficient management of attention key and value memory with **PagedAttention**
- Continuous batching of incoming requests
- Fast model execution with CUDA/HIP graph
- Quantization: GPTQ, AWQ, SqueezeLLM, FP8 KV Cache
- Optimized CUDA kernels
- Ported to Intel Gaudi 2 architecture
- Seamless integration with popular HuggingFace models
- Support NVIDIA GPUs and AMD GPUs, as well as Intel Gaudi HPUs (Habana Processing Units)

Inferencing Benchmarks

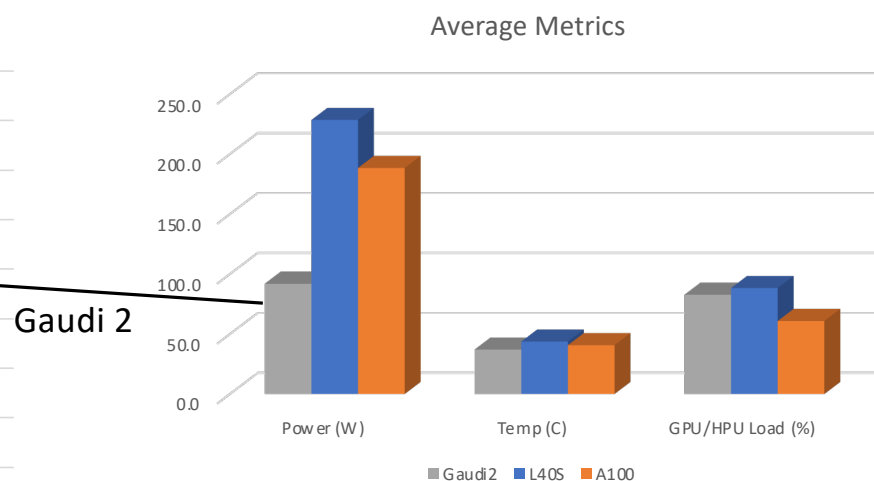
- Token Generation Latency
 - Benchmark the latency in seconds of processing a single batch of requests
 - Requests are defined as a user's request to the LLM to generate coherent text
- Online Serving Throughput
 - Benchmark a real-world scenario where the RAG AI application is served to multiple concurrent users, i.e., 1000, with 1000-1500 tokens (uniform distribution) generated per query
- Intel Gaudi 2 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

Benchmark: Token Generation Latency



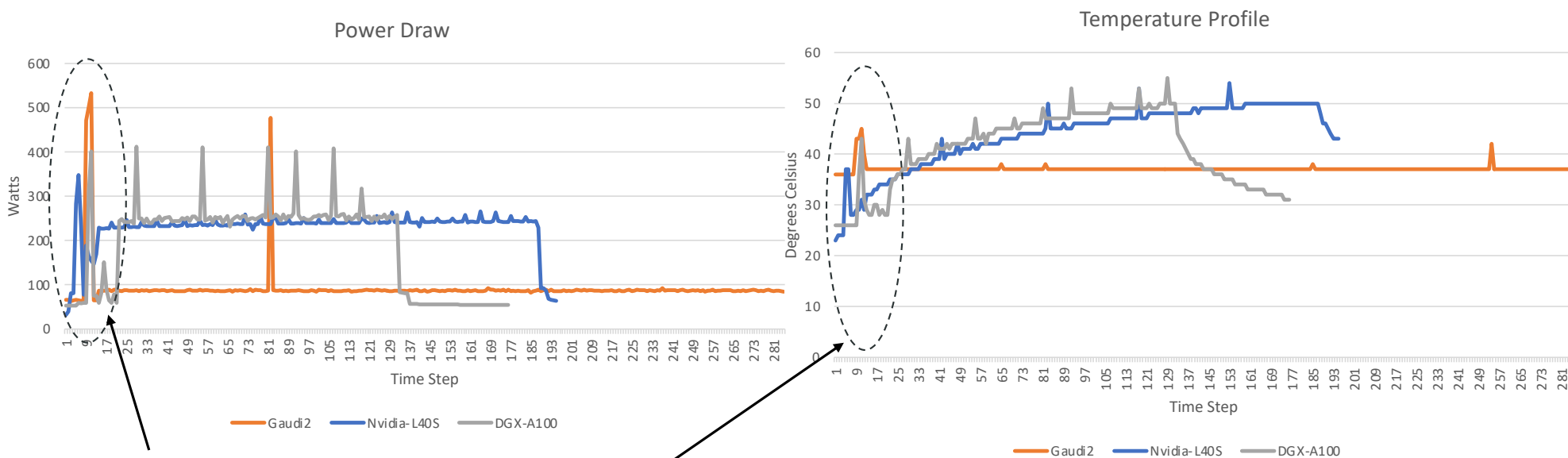
Lower is better...



Nvidia accelerators -> low latency, better performance

Intel Gaudi 2 -> higher latency, could be "acceptable" for RAG applications; however, the accelerator presents much lower power consumption and cooling requirements for comparable loads

Benchmark: Token Generation Latency (cont'd)



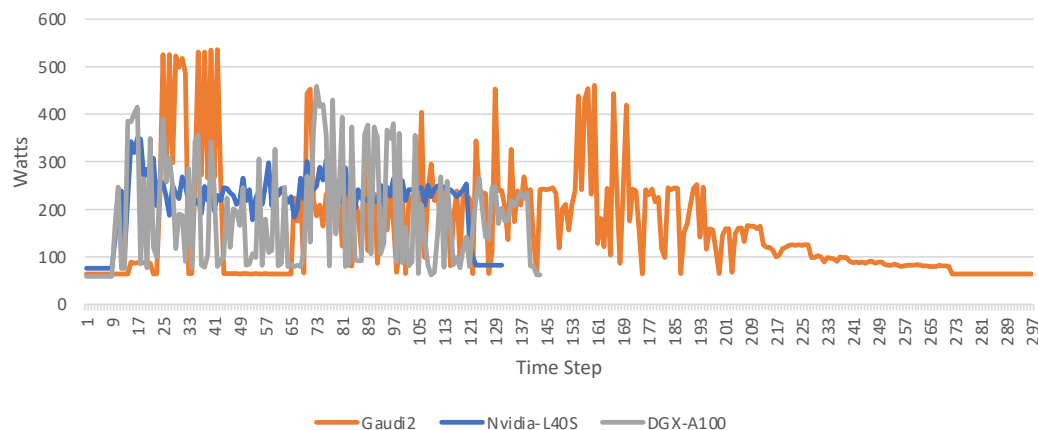
Large Language Model (LLM) initialization,
weights are loaded on the accelerator memory

Benchmark: Serving Multiple Requests (users)

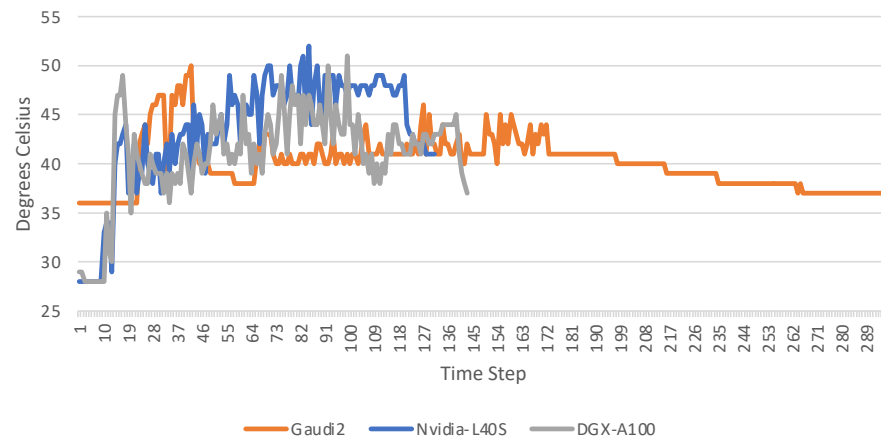
	Total tokens	Throughput			Latency	
		Requests/s	Input tok/s	Output tok/s	Mean Latency (ms)	P99 (ms)
Intel Gaudi 2	441550	4	963.49	801.19	534.93	2023.28
Nvidia-L40S	441003	9.14	2203.45	1827.27	238.1	434.84
DGX-A100	440897	7.71	1859.85	1541.52	288.24	472.86

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

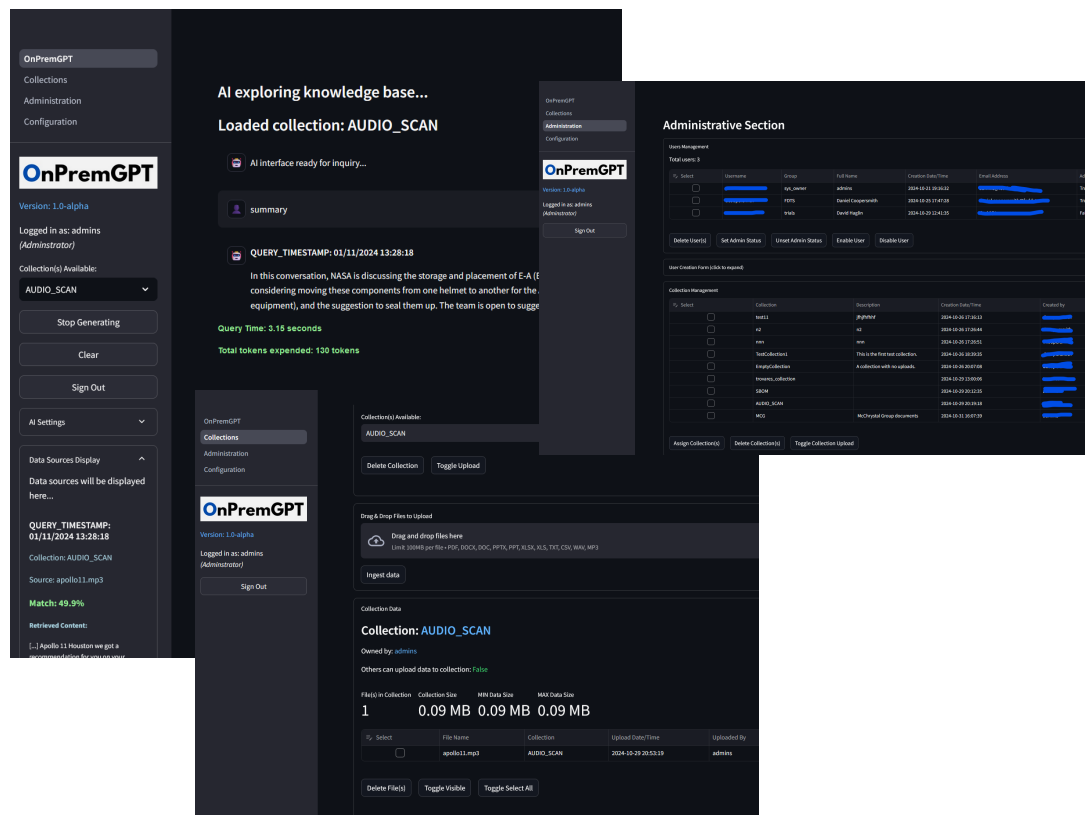
Power Draw



Temperature Profile



OnPremGPT - a Multimodal RAG AI



RAG: Retrieval Augmented Generative AI

OnPremGPT is an enterprise-level RAG AI designed for multi-user environments

Total control over the hardware, the data content and configuration, and security

Highly flexible and configurable

Deployable at the edge

“Audio RAG” capabilities using a modified version of OpenAI Whisper (local model)

APIs available for custom applications

FEDDATA

FEDDATA COMPANY PROPRIETARY

OnPremGPT as an “Audio RAG”

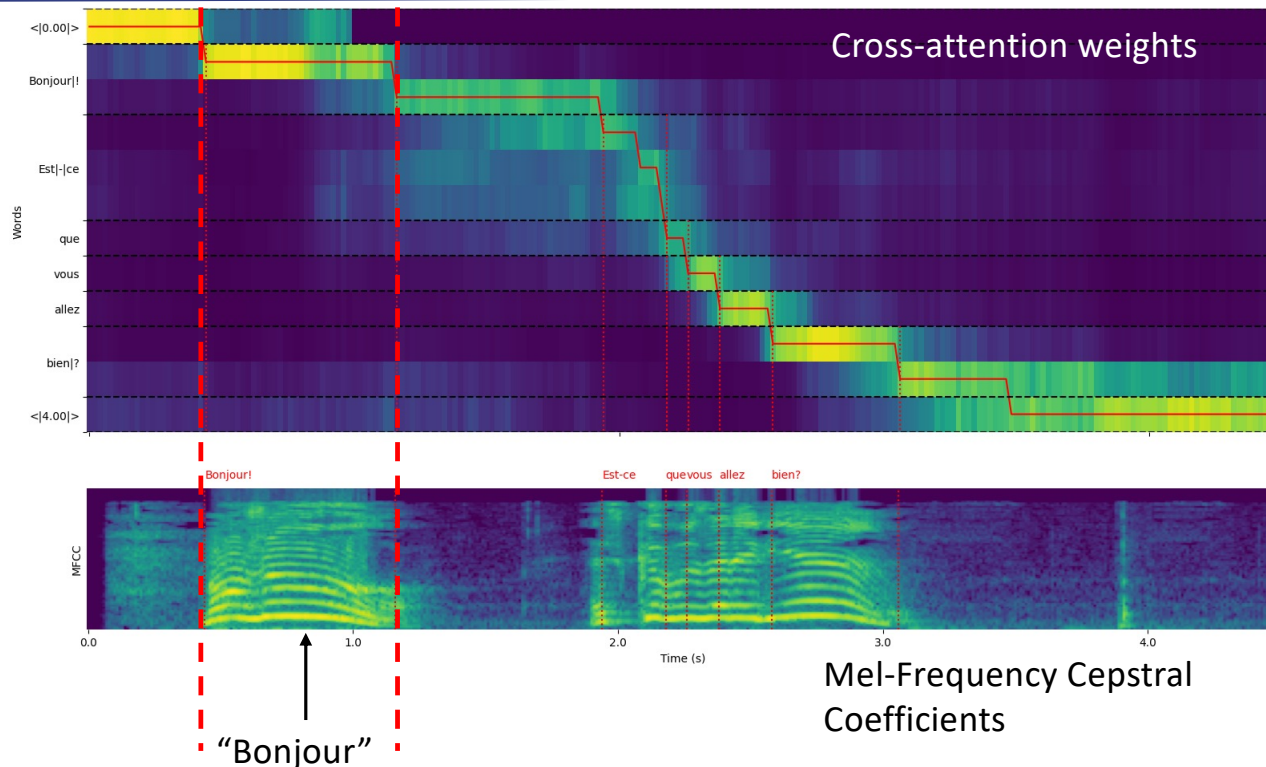
- Idea is to turn conversations into data to be ingested by the RAG to enable contextual search within the conversation
- Capability to process audio data (.wav, .mp3, etc..)
- Multilingual speech recognition, speech translation, and language identification

Work in progress:

- Improve on prediction of word timestamps
- Detect specific waveforms (sounds) in a conversation, such as an idling car, environmental sounds

Based on work from J. Lourador, whisper-timestamped),

<https://github.com/linto-ai/whisper-timestamped>)



Establishing a Software Bill of Materials (SBOM)

- Q) What is a SBOM?
- A) A SBOM is a formal, machine-readable inventory of software components and dependencies, information about those components, and their relationships
- Q) Why do we need a SBOM?
- A)
 - Increase visibility and transparency on software supply chain
 - Identify SW components for risk analysis (from a cybersecurity standpoint)
 - Facilitates the identification of suspicious or counterfeit software Components
 - Improves accountability of secure software development practices

Framing Software Component Transparency: Establishing a Common Software Bill of Materials (SBOM)

Third Edition

Tooling and Implementation Working Group hosted by the Cybersecurity and Infrastructure Security Agency (CISA)

September 3, 2024



Photo by Luke van Zyl on Unsplash

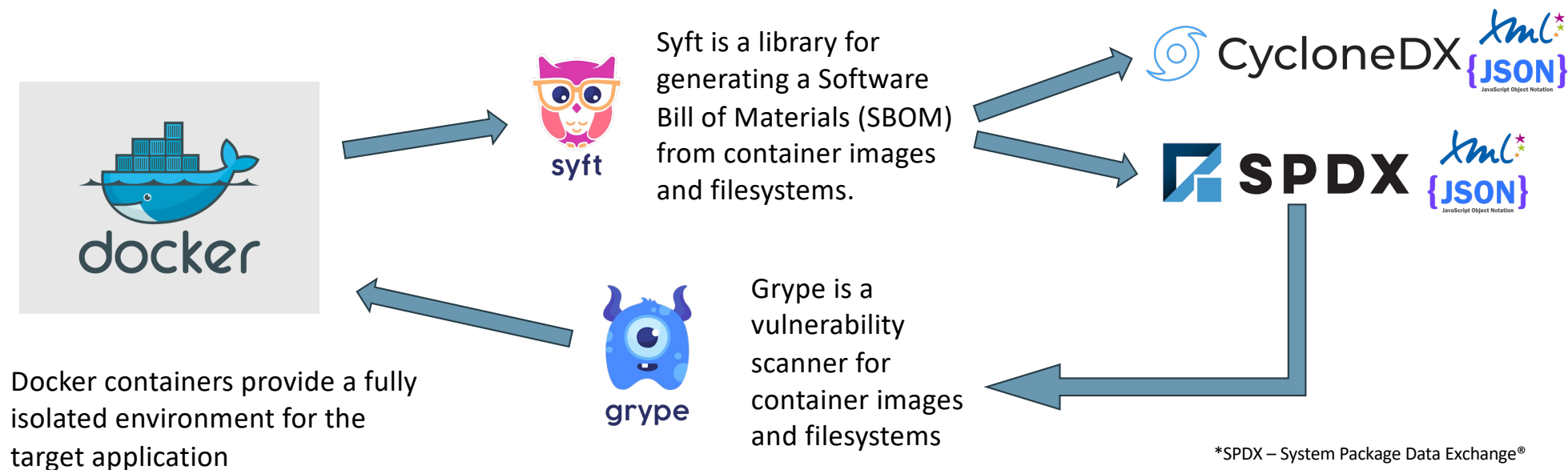
<https://www.cisa.gov/sites/default/files/2024-10/SBOM%20Framing%20Software%20Component%20Transparency%202024.pdf>

FEDDATA

FEDDATA COMPANY PROPRIETARY

SBOM Generation and Lifecycle

- Plethora of tools out there, both open-source and proprietary...
- This is an industry on its own, and several companies offer this service



FEDDATA

FEDDATA COMPANY PROPRIETARY

Example of OnPremGPT SBOM – CycloneDX and SPDX

```
<bom serialNumber="urn:uuid:acf3990a-e447-46c2-bfab-e4d732ff5a35" version="1">
  <metadata>
    <timestamp>2024-10-28T19:28:43Z</timestamp>
  </metadata>
  <tools>
    <components>
      <component type="application">
        <author>anchore</author>
        <name>syft</name>
        <version>1.14.2</version>
      </component>
    </components>
  </tools>
  <component bom-ref="a721df5bd956634" type="container">
    <name>onprem-gpt</name>
    <version>v1_0_alpha</version>
  </component>
  <properties>
    <property name="syft:image.labels:com.docker.compose.project">onprem-gpt</property>
    <property name="syft:image.labels:com.docker.compose.service">onprem-gpt</property>
    <property name="syft:image.labels:com.docker.compose.version">2.29.7</property>
  </properties>
</bom>
```

CycloneDX Specification 1.6 XML

```
JSON Raw Data Headers
Save Copy Collapse All Expand All (slow) Filter JSON
{
  "spdxVersion": "SPDX-2.3",
  "dataLicense": "CC0-1.0",
  "SPDXID": "SPDXRef-DOCUMENT",
  "name": "onprem-gpt",
  "documentNamespace": "https://anchore.com/syft/image/onprem-gpt-df37f41e-9f49-4488-90c7-950aaf8c07bc",
  "creationInfo": {
    "licenseListVersion": "3.25",
    "creators": [],
    "created": "2024-10-28T19:34:40Z"
  },
  "packages": [],
  "files": {
    "0": {
      "fileName": "/etc/ImageMagick-6/coder.xml",
      "SPDXID": "SPDXRef-File-etc-ImageMa_er.xml-fd6d60ea617ac461",
      "fileTypes": [],
      "checksums": [],
      "licenseConcluded": "NOASSERTION",
      "licenseInfoInFiles": [],
      "copyrightText": "",
      "comment": "layerID: sha256:d23b5e61_29f5af29864847008211613"
    }
  }
}
```

SPDX Specification 2.3 JSON

Example of a SBOM Security Scan – Grype

NAME	INSTALLED	FIXED-IN	TYPE	VULNERABILITY	SEVERITY
cryptography	42.0.8	43.0.1	python	GHSA-h4gh-qq45-vh27	Medium
gnutls	3.8.3-4.el9_4	10:3.8.3-4.el9_4_fips	rpm	ELSA-2024-12364	Medium
gnutls	3.8.3-4.el9_4	10:3.7.6-23.el9_3.4_fips	rpm	ELSA-2024-12336	Medium
openssl	1:3.0.7-28.0.1.el9_4	10:3.0.7-28.0.1.el9_4_fips	rpm	ELSA-2024-12675	Medium
openssl-libs	1:3.0.7-28.0.1.el9_4	10:3.0.7-28.0.1.el9_4_fips	rpm	ELSA-2024-12675	Medium
python-unversioned-command	3.9.18-3.el9_4.5	0:3.9.18-3.el9_4.6	rpm	ELSA-2024-8446	Medium
python3	3.9.18-3.el9_4.5	0:3.9.18-3.el9_4.6	rpm	ELSA-2024-8446	Medium
python3-libs	3.9.18-3.el9_4.5	0:3.9.18-3.el9_4.6	rpm	ELSA-2024-8446	Medium
stdlib	go1.18.2	1.21.11, 1.22.4	go-module	CVE-2024-24790	Critical
stdlib	go1.18.2	1.19.10, 1.20.5	go-module	CVE-2023-29405	Critical
stdlib	go1.18.2	1.19.10, 1.20.5	go-module	CVE-2023-29404	Critical
stdlib	go1.18.2	1.19.10, 1.20.5	go-module	CVE-2023-29402	Critical
stdlib	go1.18.2	1.19.9, 1.20.4	go-module	CVE-2023-24540	Critical
stdlib	go1.18.2	1.19.8, 1.20.3	go-module	CVE-2023-24538	Critical
stdlib	go1.18.2	1.21.0-0	go-module	CVE-2023-24531	Critical
stdlib	go1.18.2	1.22.7, 1.23.1	go-module	CVE-2024-34158	High
stdlib	go1.18.2	1.22.7, 1.23.1	go-module	CVE-2024-34156	High
stdlib	go1.18.2	1.21.12, 1.22.5	go-module	CVE-2024-24791	High
stdlib	go1.18.2	1.21.8, 1.22.1	go-module	CVE-2024-24784	High
stdlib	go1.18.2	1.21.9, 1.22.2	go-module	CVE-2023-45288	High
stdlib	go1.18.2	1.20.0	go-module	CVE-2023-45287	High
stdlib	go1.18.2	1.20.12, 1.21.5	go-module	CVE-2023-45285	High

Identification of critical vulnerabilities including CVE identifier, severity, etc...

Excellent way to harden the application...however it exposes attack surfaces similarly to what

OpenVAS or Nessus scans can do on a web application.

Morale: we need to be careful in how the SBOM is shared/maintained.

FEDDATA

FEDDATA COMPANY PROPRIETARY

Conclusions

- Intel Gaudi 2 Accelerators
 - Nvidia accelerators provide higher throughput (tokens/sec) with lower latency
 - However...Intel Gaudi 2 provides acceptable performance for RAG AI type of applications with lower power and cooling requirements, and especially total cost of ownership (TCO)
 - Positive experience in terms of using Habana libraries, PyTorch Lightning, porting of applications to Intel Gaudi 2 seems to be well addressed
- OnPremGPT and Multimodal AI
 - Good results obtained in analyzing audio conversations
 - Very powerful contextual search capabilities, **not just word pattern recognition**
 - AI multimodality will take us closer to Artificial General Intelligence (AGI)
- SBOMs are necessary but need to be approached with the right mindset

Future Work

- Benchmark the **Intel Gaudi 3** 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 the development of multimodal AI capabilities