AI everywhere:
Revolutionizing AI in the hybrid cloud

Presentation by Stefan Englet, Alliance Manager for Red Hat in EMEA

AI is evolving rapidly

AI is as disruptive as the Internet

$300B
worldwide GenAI spending set to exceed $300B by 2026[1]

Generative AI predicted to add up to $4.4T of value to global economy by 2040[4]

Growth of **large model sizes** (1T+ parameter models) and **smaller, nimbler models** (~10B parameters)

80% of PCs to be AI PCs by 2028[6]

of enterprises will use Gen AI by 2026[7]

58% of CEOs from leading public companies actively investing in AI[2]

More than **75%** of enterprise-managed data will be created & processed outside the data center or cloud by 2025[3]

AI inferencing driving up compute costs; exceeding the pace of Moore's Law

50% of edge deployments will involve AI by 2026[5]

1. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights
2. https://chiefexecutive.net/the-rise-of-the-ai-ceo /;
3. Gartner©, Hyperscalers Stretching to the Digital Edge, July 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All right reserved.
4. Worldwide Artificial Intelligence Spending Guide (IDC)
5. Gartner®, Building an Edge Computing Strategy, Thomas Bittman, 12 April 2023. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All right reserved
6. Source: Boston Consulting Group
7. Gartner news release – Oct. 10, 2023.

The AI Pipeline runs on Intel

Data

Data Collection
Data Validation
Data Preparation
Data Ingestion
Data Labeling
Data Transformation
Data Analysis
Data Cleanup

Majority of data prep already happens on Intel® Xeon® processors and that continues today

Train

Feature Creation
Model Selection
Evaluation
Model Training
Validation

Traditional Machine Learning
Intel Xeon's large memory and fast cores excel running ML workloads

Deep Learning
GPUs & Gaudi accelerator are well-suited, Intel Xeon's built-in AI capabilities deliver compelling TCO

Deploy

Model Optimization
Model Packaging
Performance Monitoring
Endpoint Config
Model Serving

2/3rd of data center inferencing is run on **Intel Xeon Processors**[1]

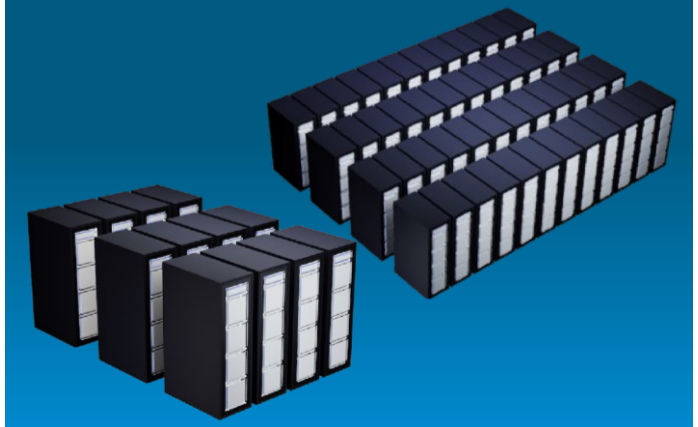
Database, Data Warehouse, Data Lake, Streaming Data, Feature Store, Model Registry

1 Based on Intel market modeling of the worldwide installed base of data center servers running AI Inference workloads as of December 2022.

AI PC Node
AI Developer Productivity & Light Inference

Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inference



Super Cluster
Training, Tuning, Peak Inference

Mega Cluster
Large Scale Training & Inference

AI PC
Broadest AI SW Ecosystem

ENTERPRISE AI & EDGE AI
Open Standard, "Ready to Use"

DATA CENTER AI
AI Open, Scalable Systems & Reference Arch

intel core ULTRA | intel ARC GRAPHICS | intel core ULTRA | intel xeon | intel gaudi | intel xeon | intel gaudi | intel IPU

ANNOUNCING Intel enterprise AI for Red Hat® OpenShift® AI and Red Hat® Enterprise Linux

Customer and Ecosystem Managed Software

Intel® Tiber™ Edge Platform

PyTorch

OpenVINO™

Intel® Tiber™ AI Studio

Red Hat OpenShift AI

Red Hat Ansible Automation Platform

Red Hat OpenShift Container Platform

Red Hat Enterprise Linux

intel CORE ULTRA

intel. ARC GRAPHICS

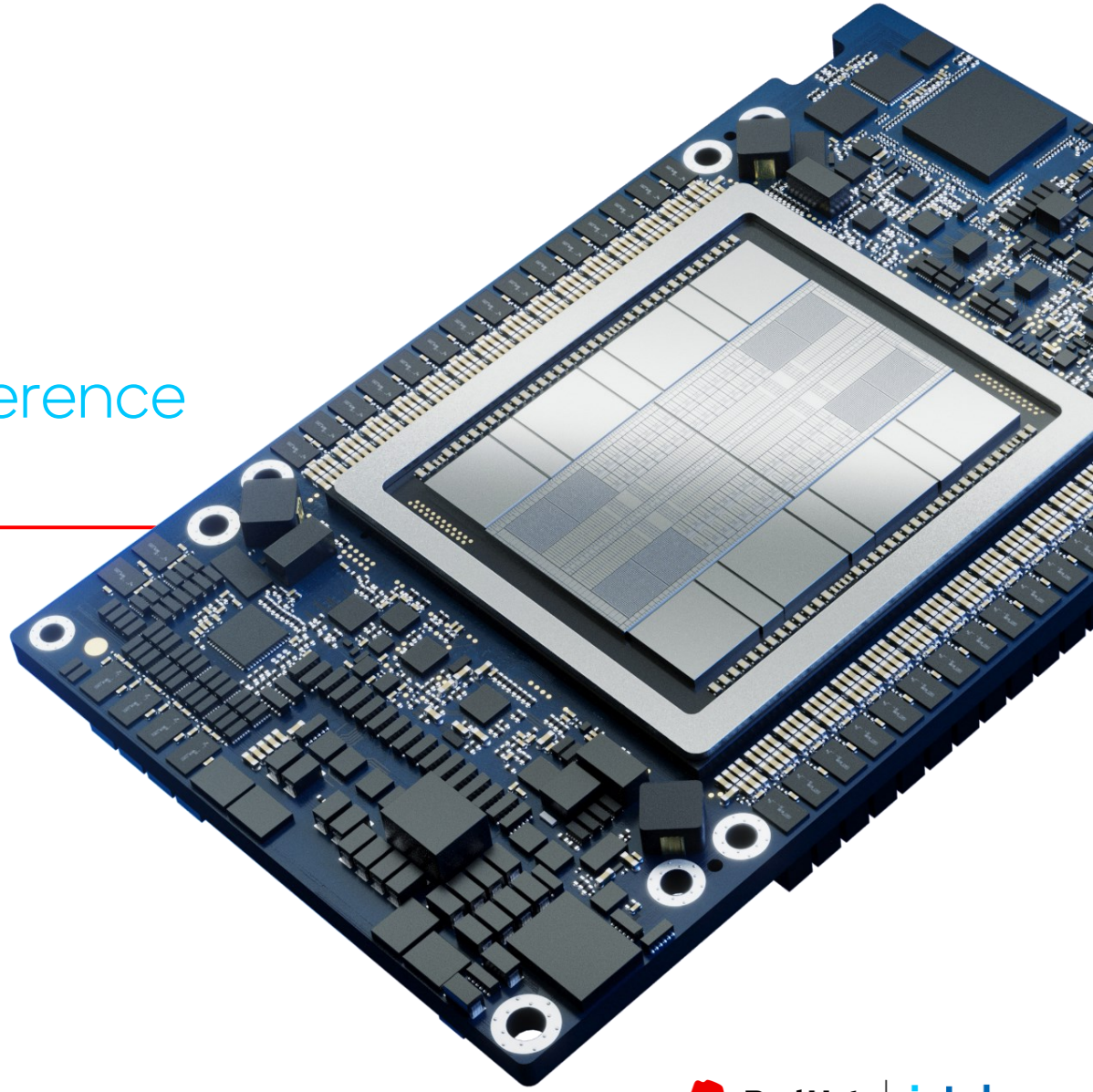
intel GAUDI

intel XEON

intel. IPU

Intel® Gaudi® 3 AI Accelerator

Giant Leap in Performance and Productivity for AI Training & Inference



SUPERMICRO

DELL Technologies

Lenovo

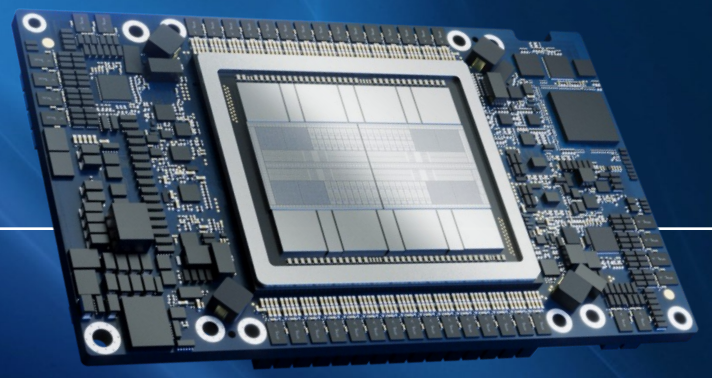
Hewlett Packard Enterprise

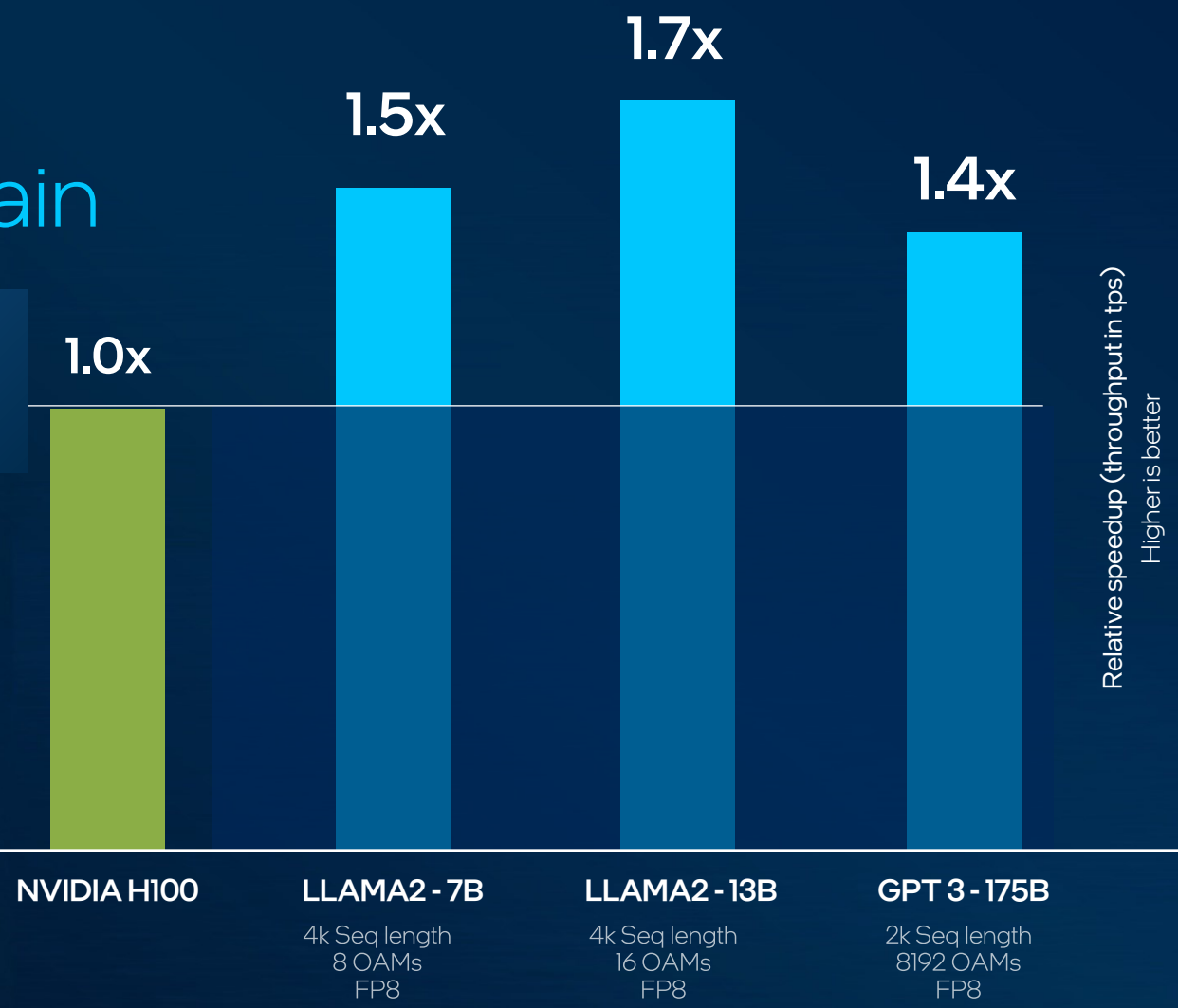
Red Hat | intel.

intel GAUDI

1.5x faster time-to-train

Average projection for Intel® Gaudi® 3 Accelerator vs. Nvidia H100, running common Large Language Models*





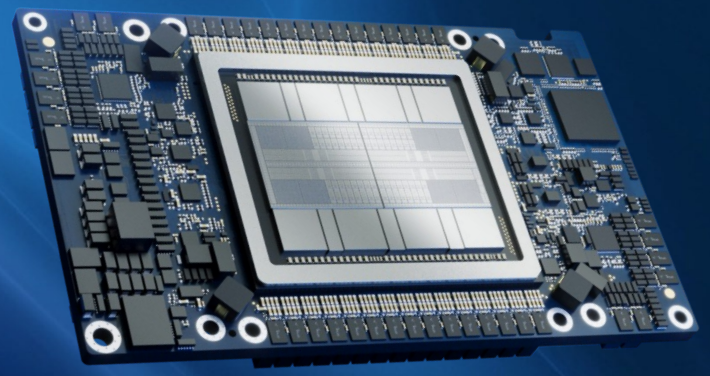
Link to data

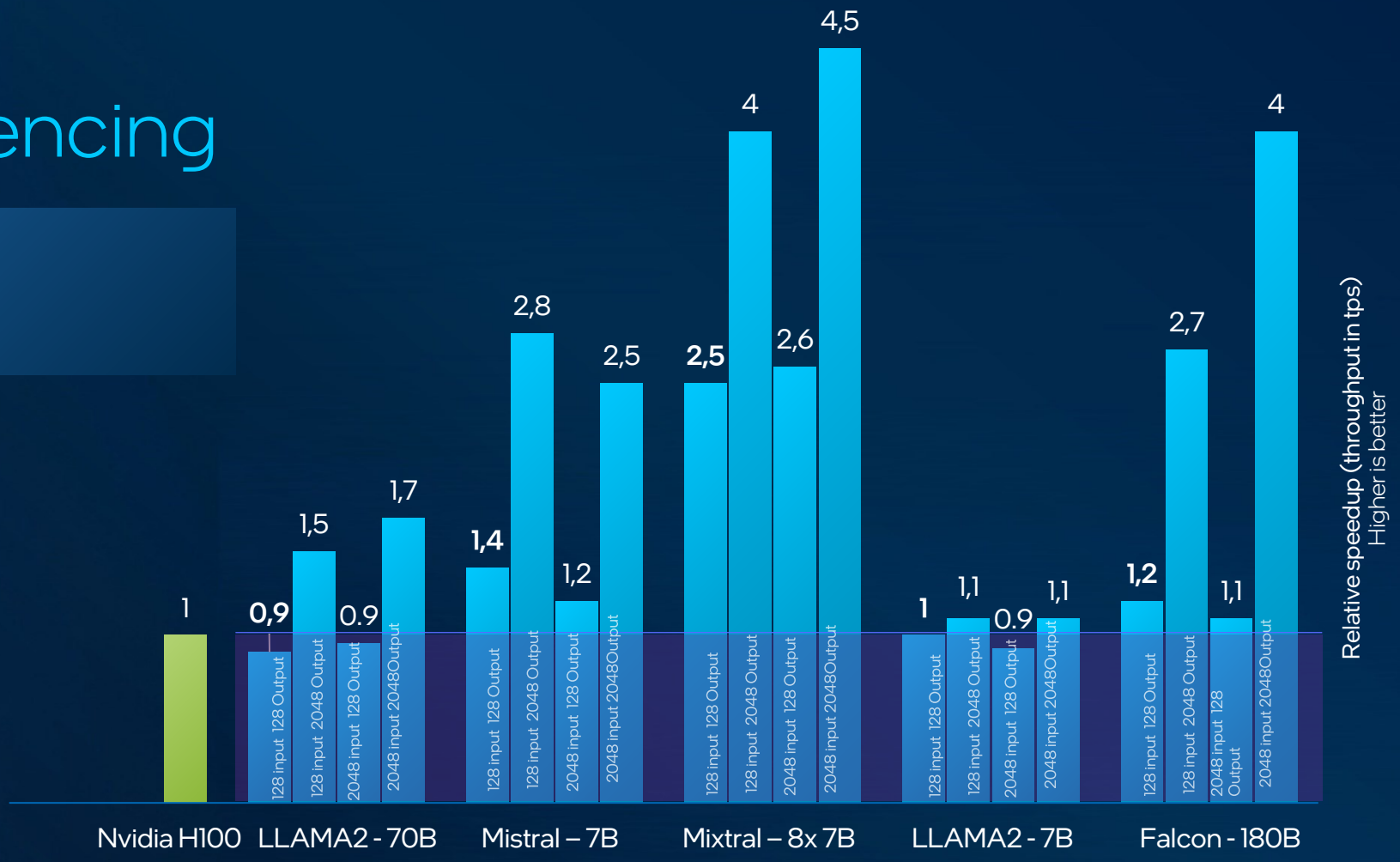
* NV H100 comparison based on : https://nvidia.github.io/TensorRT-LLM/performance/perf-overview.html, May 28, 2024 → "Large Language Model" tab
Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-13B & GPT3-175B as of 3/28/2024. Results may vary

intel GAUDI

2x faster inferencing

Average projection for Intel® Gaudi® 3 Accelerator vs. Nvidia H100, running common Large Language Models*

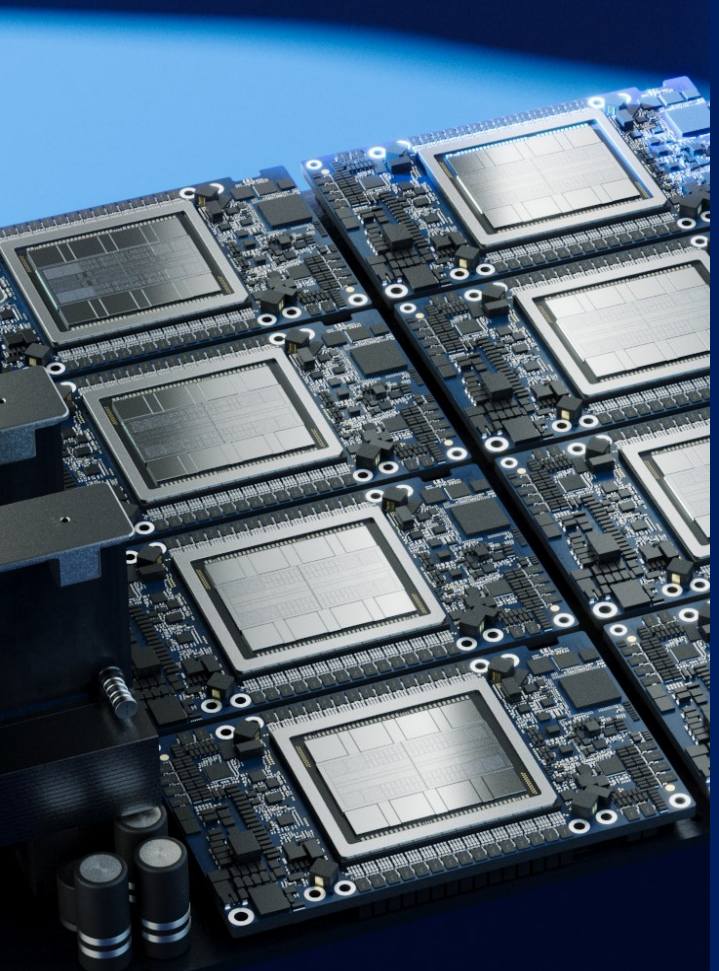




Source for Nvidia performance: Overview — tensorrt_llm documentation (nvidia.github.io), May, 2024. Reported numbers are per GPU.
Intel Gaudi 3 projections by Habana Labs, Apr 2024; Results may vary

intel GAUDI

Making Gen AI More Accessible



Addressing Cost Barriers

Gaudi 3 AI Accelerator kit

USD 125K

8X Gaudi 3 AI Accelerators+ Universal Baseboard (UBB) (List Price)

Gaudi 2 AI Accelerator kit

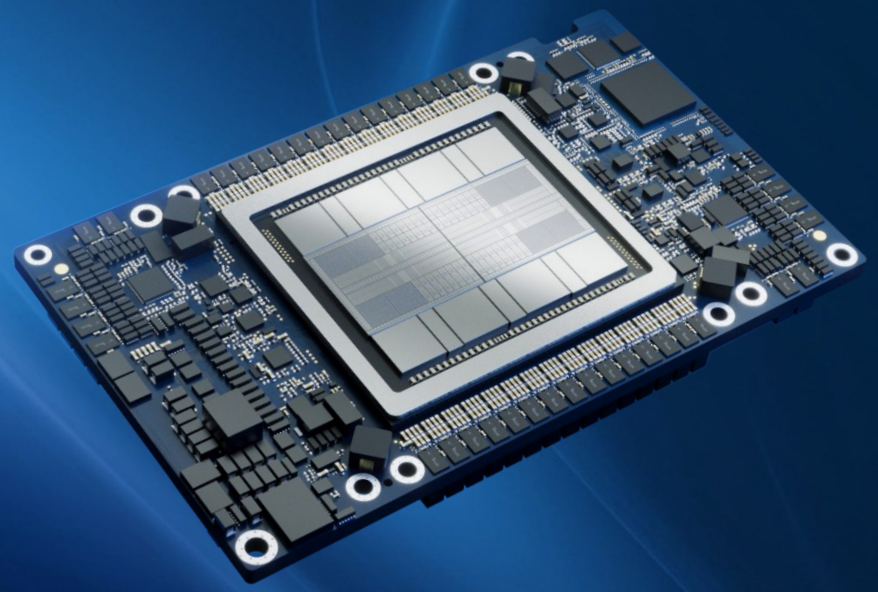
USD 65K

8X Gaudi 2 AI Accelerators+ Universal Baseboard (UBB) (List Price)

Pricing guidance for cards and systems is for modeling purposes only. Please consult your original equipment manufacturer (OEM) of choice for final pricing. Results may vary based upon volumes and lead times.

intel GAUDI

Delivering Price Performance Advantage



2.3x Perf/$

Inference Throughput

Gaudi 3 AI Accelerator
Vs H100

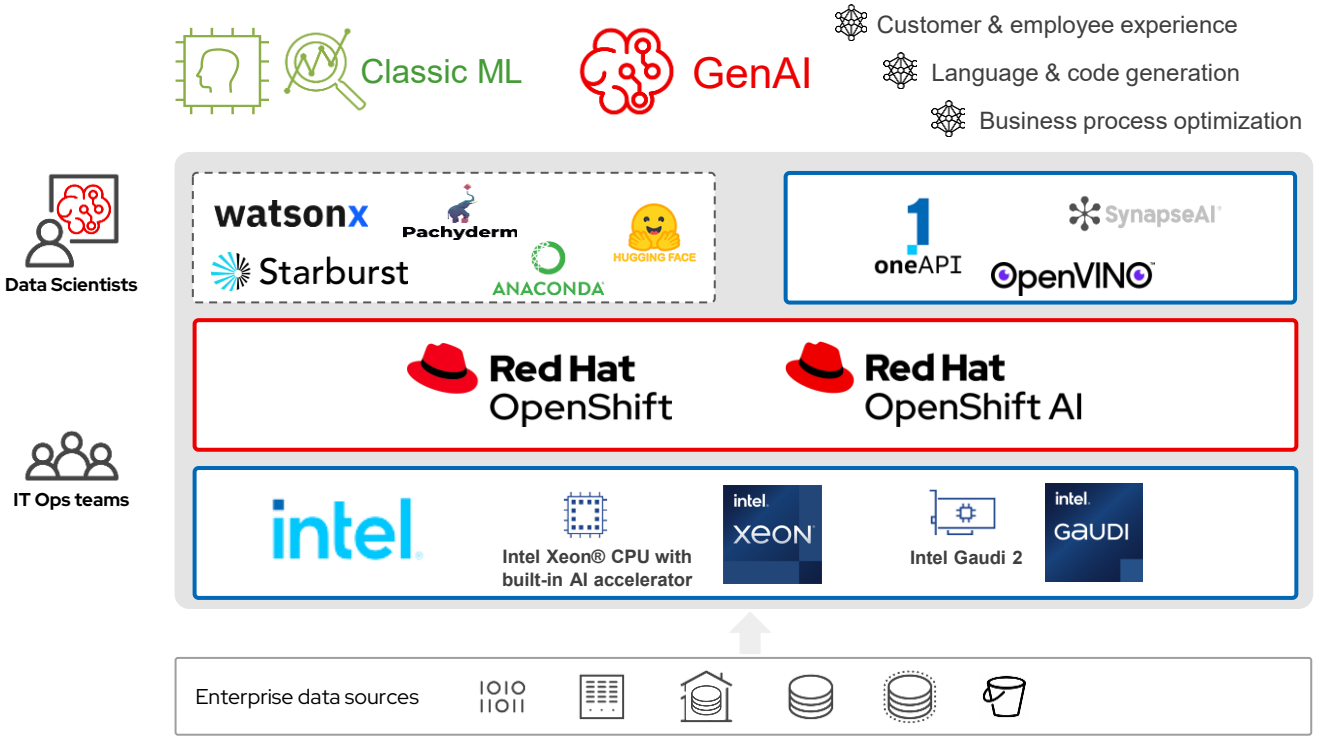
1.9x Perf/$

Training Throughput

Gaudi 3 AI Accelerator
Vs H100

Source Intel projected results vs H100 data sources: https://developer.nvidia.com/deep-learning-performance-training-inference/ai-inference and https://developer.nvidia.com/deep-learning-performance-training-inference/training
Intel results obtained in April 2024. Results may vary. Pricing estimates based on publicly available information and Intel internal analysis

Example: Building a GenAI stack

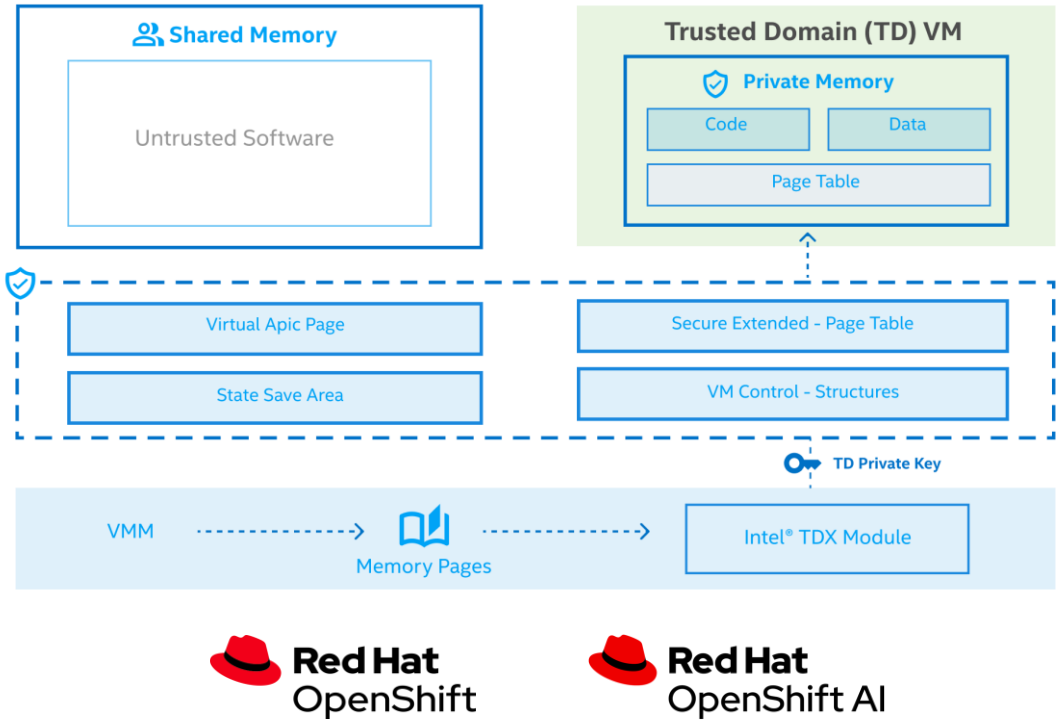


Partnering to deliver
AI solutions **at the**
Intel 5g Innovation Center
with AI Sweden consortium

- Red Hat and Intel partnered with AI Sweden, a national center for applied AI, on NLP applications
- Deeper, product collaboration focused on customer enablement with OpenShift AI, Intel Xeon, Gaudi 2 and the Intel AI Suite
- Enable testing, validation and proof of concepts with partners and customers
- Receive support for building AI applications

Confidential Containers for AI workload security in the public cloud

With Intel Trusted Domain Extensions (TDX)

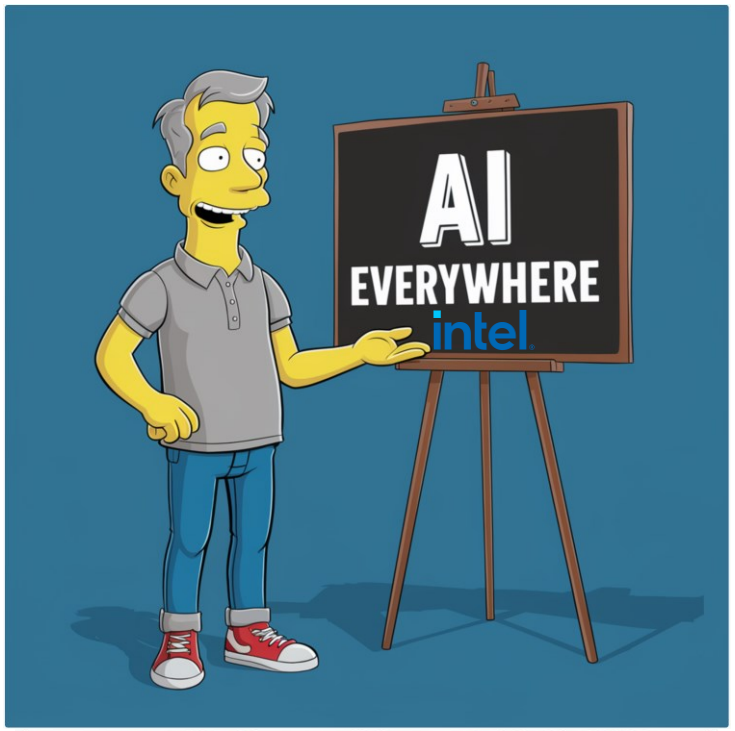


- High value IP such as AI models require additional security
- Confidential Containers allow to safely process sensitive data in the cloud
- Intel TDX providing hardware-isolated VMs to protect containers from unauthorized access
- On Azure DCesv5 and ECesv5-series or on-prem with 5th Gen Intel Xeon CPUs

More information - https://www.redhat.com/en/blog/enhancing-ai-workload-security-in-the-public-cloud

Summary

AI is evolving rapidly but Intel has you covered!



Picture by Ideogram.ai

- Intel has a rich portfolio of AI-optimized hardware for your entire AI pipeline
- Well integrated in OpenShift and OpenShift AI, delivering application performance and developer productivity
- To learn more, join the Intel and Red Hat AI developer program
- You can immediately action on AI with
  - Dell will launch the Intel Gaudi 3 AI Accelerator on the Dell PowerEdge XE9680 server
  - Intel Tiber Developer Cloud
  - Dell APEX Cloud Platform for Red Hat OpenShift

Red Hat + intel.

Over **25** Years of Partnership

Muchas gracias