NEW YORK CITY | JULY 10, 2024

aws SUMMIT

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AIM206-S

SPONSORED BY INTEL

The hidden potential of CPUs in AI

Mackenzie Lamon
Global GTM and Business Development Lead
Intel Corporation

Colin Best
Data and AI Lead
Accenture

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- AI market review
- Using a CPU for AI/generative AI
- Intel and Accenture – ecosystem partnership at scale

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

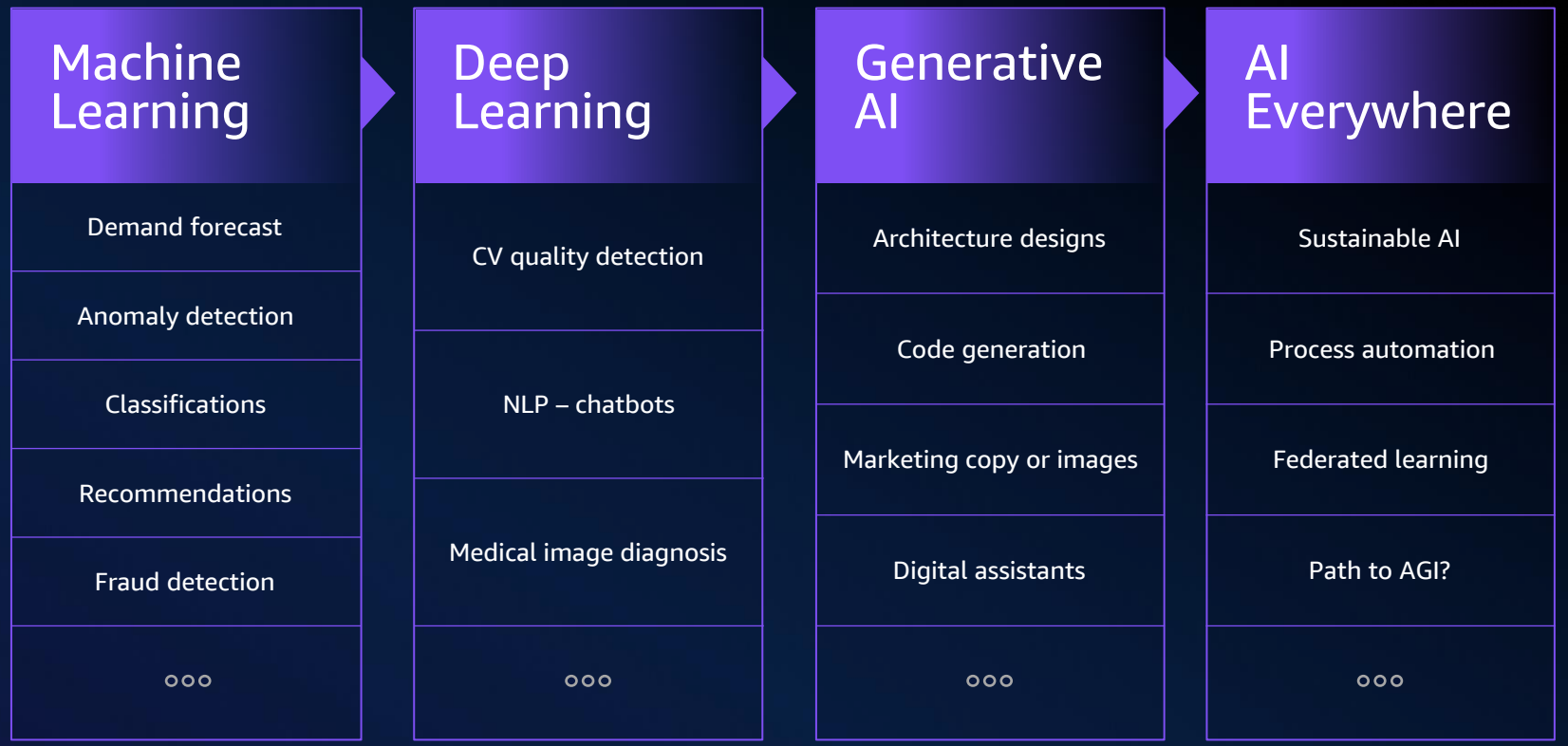
AI market review

The answer isn’t always a GPU

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Market view

AI is evolving rapidly



What AI needs

- Data, compute, networking, memory and algorithms
- Speed training and fine tuning on large and nimble compute clusters (from weeks to days to hours)
- Responsibly deploy and inference anywhere on all devices (milliseconds)

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.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

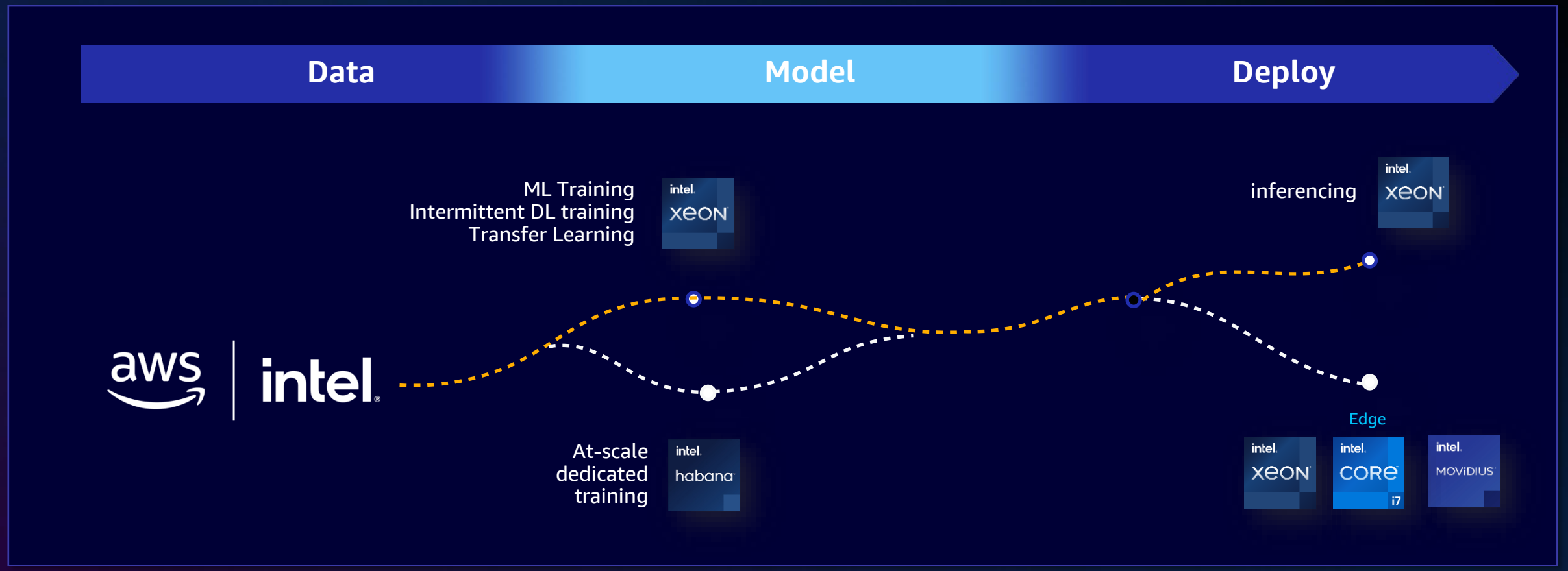
Why a CPU?

- Availability: CPUs widely deployed across availability
- Performance: CPUs can meet customer SLAs (including generative AI)
- Open platform: Open SW ecosystem with broad model/tool support
- Sustainability: Model optimizations, software optimizations

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

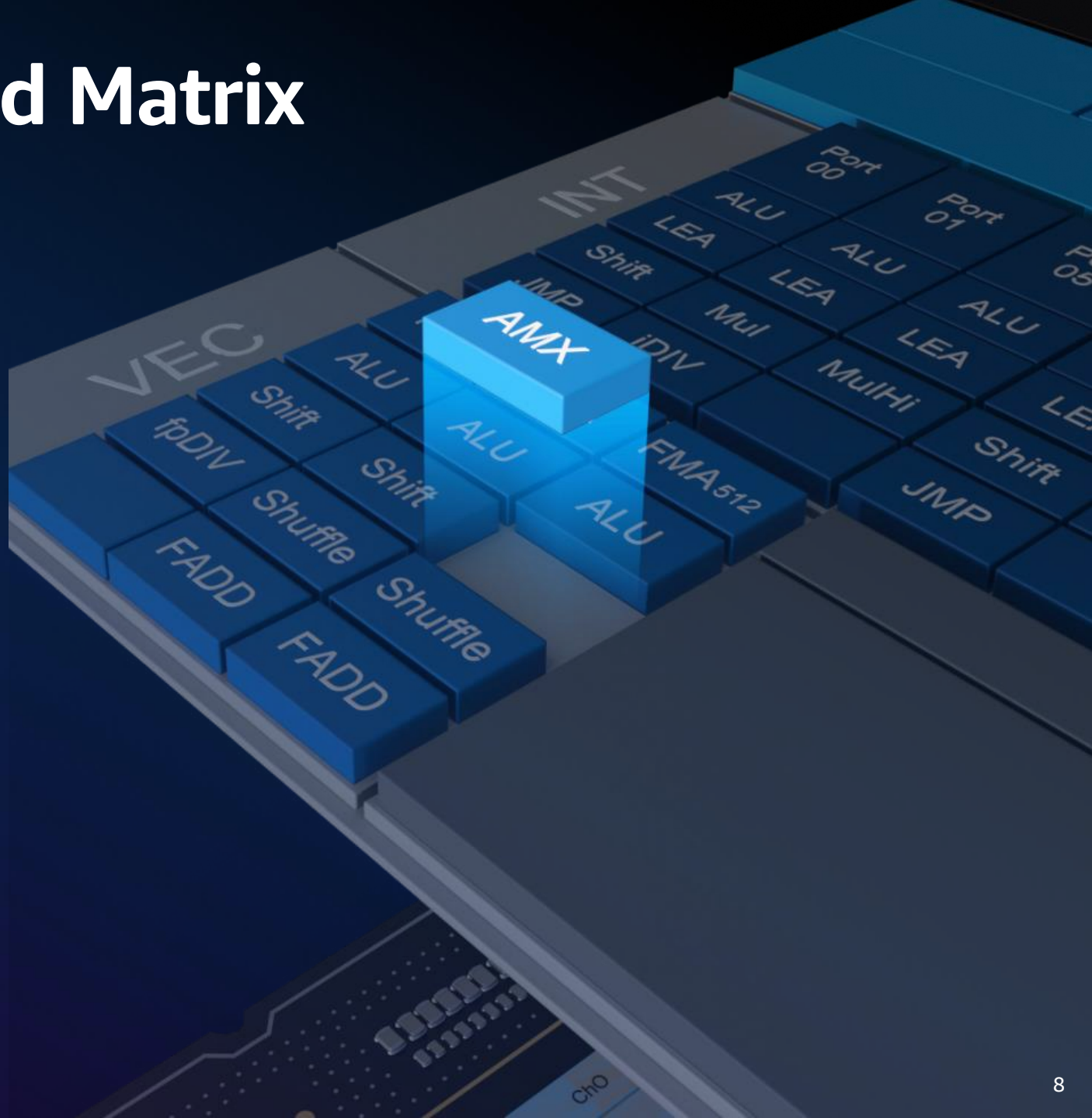
AWS Intel AI strategy



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is Intel® Advanced Matrix Extensions (AMX)?

Dedicated silicon on each Sapphire Rapids core that helps accelerate **deep learning** workloads



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intel® AMX accelerates deep learning use cases

Intel® Advanced Matrix Extensions (AMX)

BF16 and INT8 data types



Recommender Systems



Natural Language Processing



Image Recognition
Object Detection

Intel® Advanced Vector Extensions (AVX-512)

FP32 and FP64 data types



Data Analytics



Classical Machine Learning

FUN FACT:

Many DL workloads are “mixed precision” and 4th Gen Xeon can seamlessly transition between AMX and AVX-512 as needed

aws

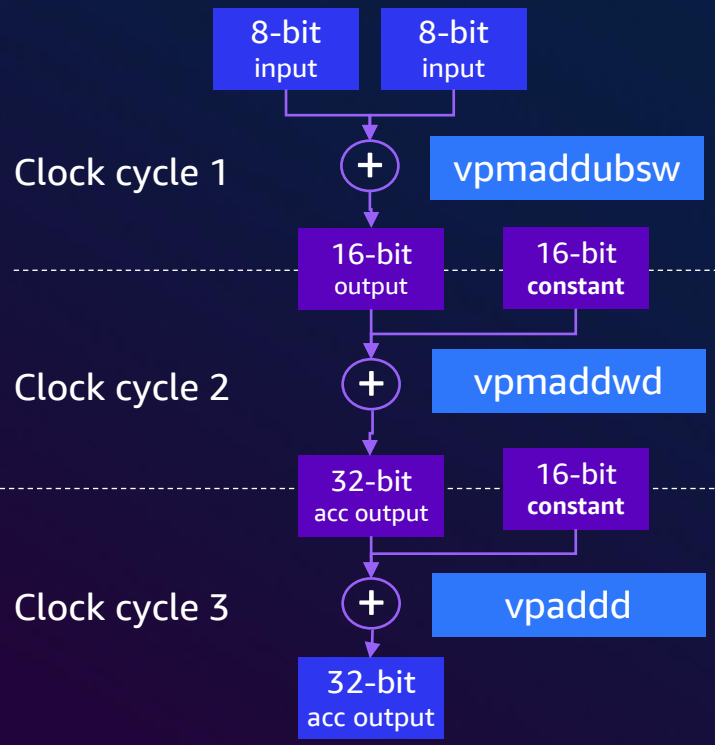
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

M7i & C7i have 8x* higher peak FLOPs / TOPs compared to M6i & C6i instances

New to 7th gen instances

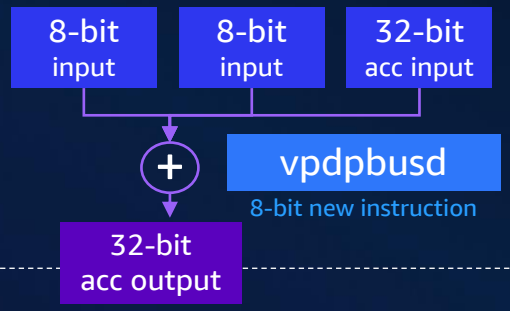
Intel® AVX-512

85 int8 ops/cycle/core
with 2 FMAs



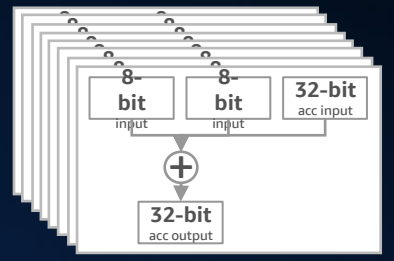
Intel® AVX-512 VNNI

256 int8 ops/cycle/core
with 2 FMAs



Intel® AMX

2048 INT8 ops/cycle/core
Multi-fold MACs in one instruction



CPU scalar and vector + matrix extensions now available in a single chip

4th Gen Intel Xeon can seamlessly transition between AMX and AVX-512 as needed

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intel AI portfolio

Intel AI hardware portfolio

aws | intel

400+ Intel instances

15 years of collaboration

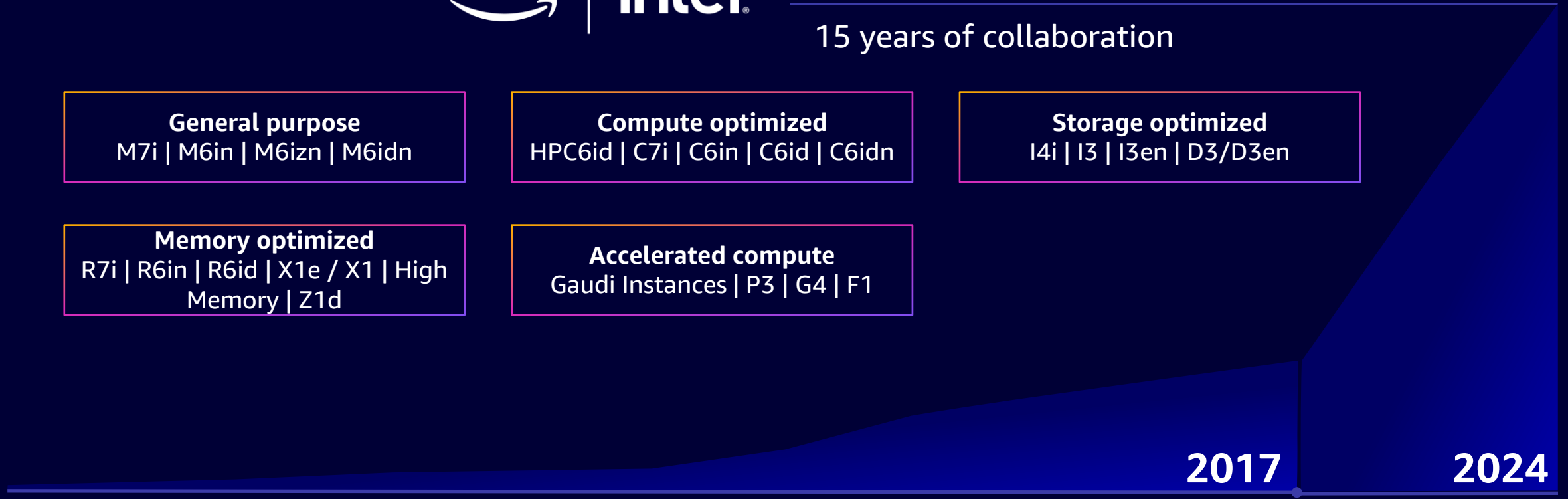
General purpose
M7i | M6in | M6izn | M6idn

Compute optimized
HPC6id | C7i | C6in | C6id | C6idn

Storage optimized
I4i | I3 | I3en | D3/D3en

Memory optimized
R7i | R6in | R6id | X1e / X1 | High Memory | Z1d

Accelerated compute
Gaudi Instances | P3 | G4 | F1



aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Deploy with SageMaker

Amazon SageMaker now supports M7i, C7i, and R7i instances for ML Inference

Posted On: Jan 11, 2024

All 7i instances come with AMX acceleration



Deploy inference endpoints for embedding models and large language models

M7i/C7i instances in SageMaker provide up to 5x higher throughput* for embedding models.

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intel AI portfolio

Intel AI software portfolio



Engineer Data

MODIN · SciPy · pandas · NumPy · Spark · Numba

Data Analytics at Scale†

Create Models

dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO · scikit learn · TensorFlow · DirectML · SigOpt AutoML · intel Neural Compressor · W3C WebAssembly WebGPU WebNN

Machine & Deep Learning Frameworks, Optimization and Deployment Tools†

Optimize & Deploy

1 oneAPI

- Intel® oneAPI Deep Neural Network Library
- Intel® oneAPI Collective Communications Library
- Intel® oneAPI Math Kernel Library
- Intel® oneAPI Data Analytics Library

Open, multiarchitecture programming model for CPUs, GPUs, and other accelerators

CLOUD & ENTERPRISE | **CLIENT & WORKSTATION** | **EDGE**

---

intel AI TOOLS — Accelerate end-to-end data science and AI

Intel® Developer Cloud and Intel® Developer Catalog — Try the latest Intel tools and hardware, and access optimized AI Models

intel Tiber — Full stack ML operating system

Intel® Geti™ — Annotation/training/optimization platform

Hugging Face — Intel optimizations and fine-tuning recipes, optimized inference models, and model serving

aws

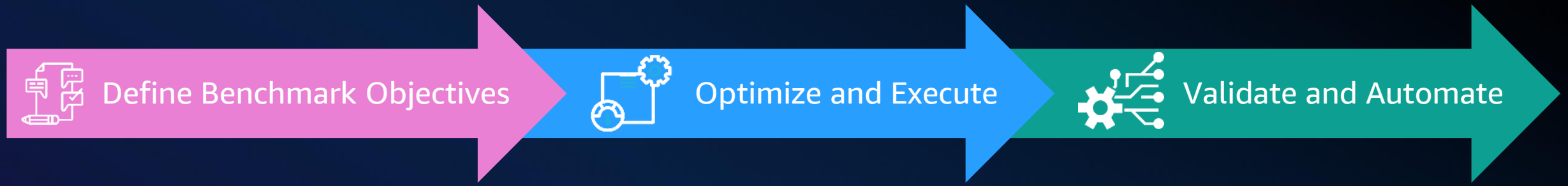
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Using a CPU for AI/generative AI

Do the easy things first

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI Benchmarking services



Identify AI workload parameters @ AWS

- Identify if current benchmarks are available to customize
- Customize existing benchmark or build new benchmark to customer specific application environment
- Identify resources from customer and Intel for project and develop a timeline

Software optimizations for AI on Intel technology

- Tuning parameters for buffering and caching
- Activating Hardware Accelerators
- Simple Code Changes
- Kernel parameter optimizations
- Validated config files

Before and after performance results and automate

- Internal Validation with 3 teams
- Terraform, Cloud Init, and Ansible scripts for automated customizations
- Upload into CSP Marketplace (optional)

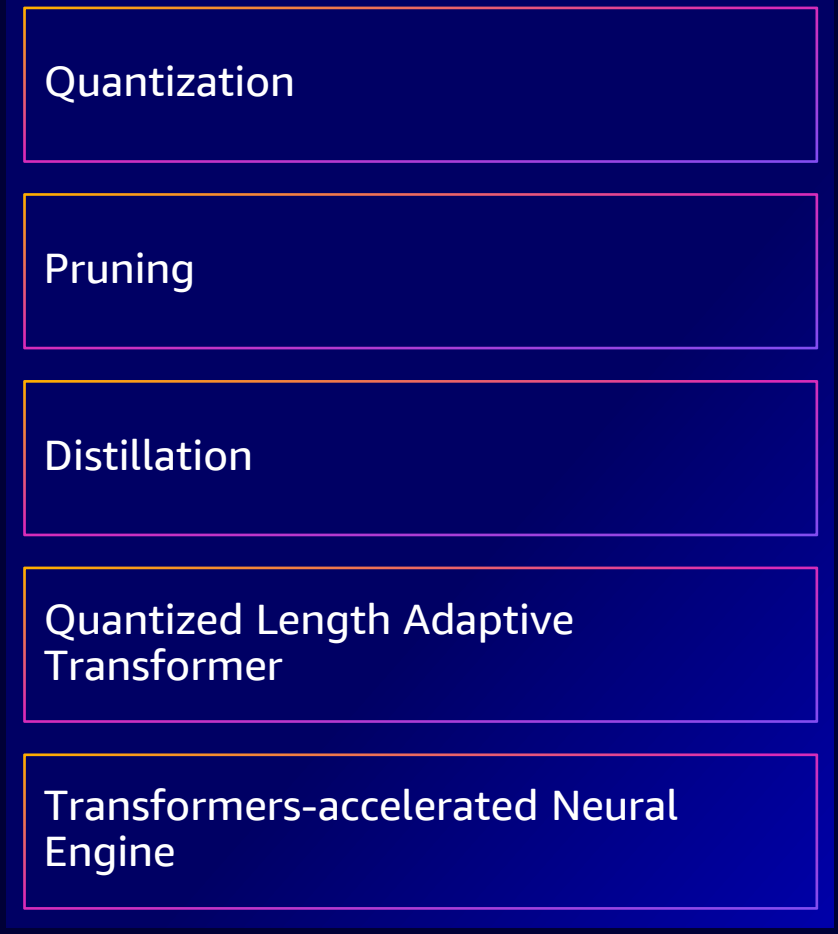
aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimizing large language models

- **SW optimizations:**
  - Optimizing Hugging Face Transformers: Intel® Extension for Transformers.
  - Framework optimizations: Intel® Extension for PyTorch, Intel® Extension for Tensorflow, Intel® Neural Compressor.

- **AWS Gaudi instances:**
  - DL1 (Gaudi 1)

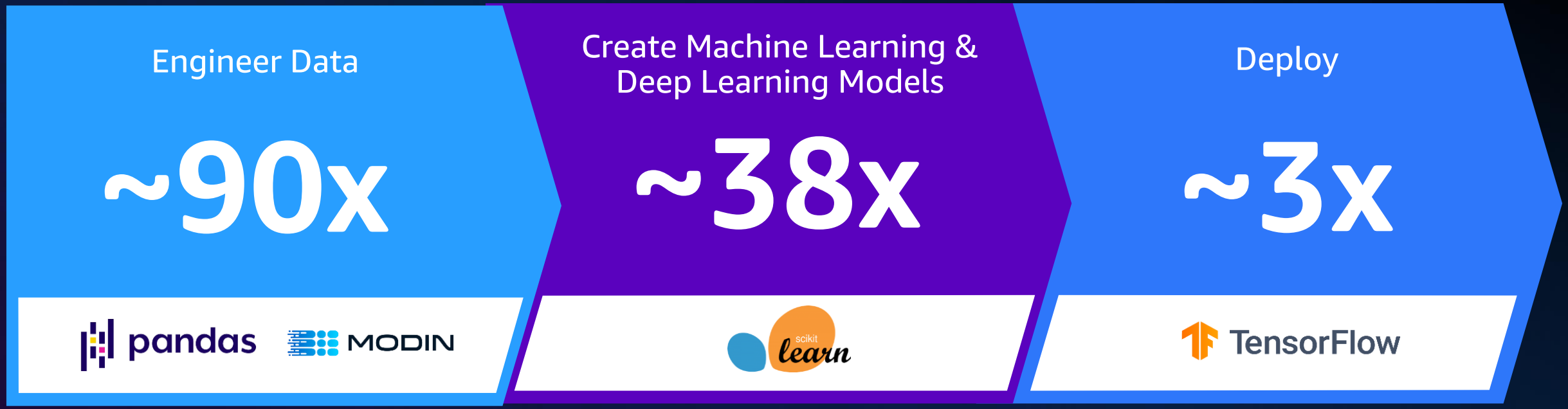
- **Intel CPU instances supporting AMX:**
  - **M7i, C7i, R7i** (4th Generation Intel® Xeon® Scalable Processors).



Intel® Extension for Transformers features

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

One-line code change = Huge performance gains



import modin.pandas as pd

from sklearnex import patch_sklearn
patch_sklearn()

TF_ENABLE_ONEDNN_OPTS=1

Click here for workloads and configurations. Results may vary

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intel and Accenture

Ecosystem partnership at scale

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

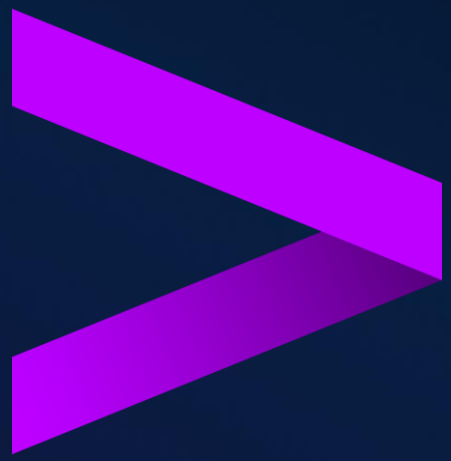
Realize client transformation through our distinct partnership

accenture

With unrivaled market and industry insights, **Accenture** accelerates market adoption of new technology through joint solution development and early evaluation of use cases and real-life client scenarios.

intel

As one of the largest R&D organizations in the world, **Intel** offers Accenture and its clients early access to new technology with tight integration between software, hardware and cloud applications – and corresponding SMEs.



Accelerated co-innovation

We marry cutting-edge Intel technology with Accenture's industry expertise through joint investment into our new and existing offerings.

Capabilities alignment

We maximize impact by facilitating initial client engagements, serving the right experts, and enabling an ongoing Accenture sales motion.

Consistent client outcomes

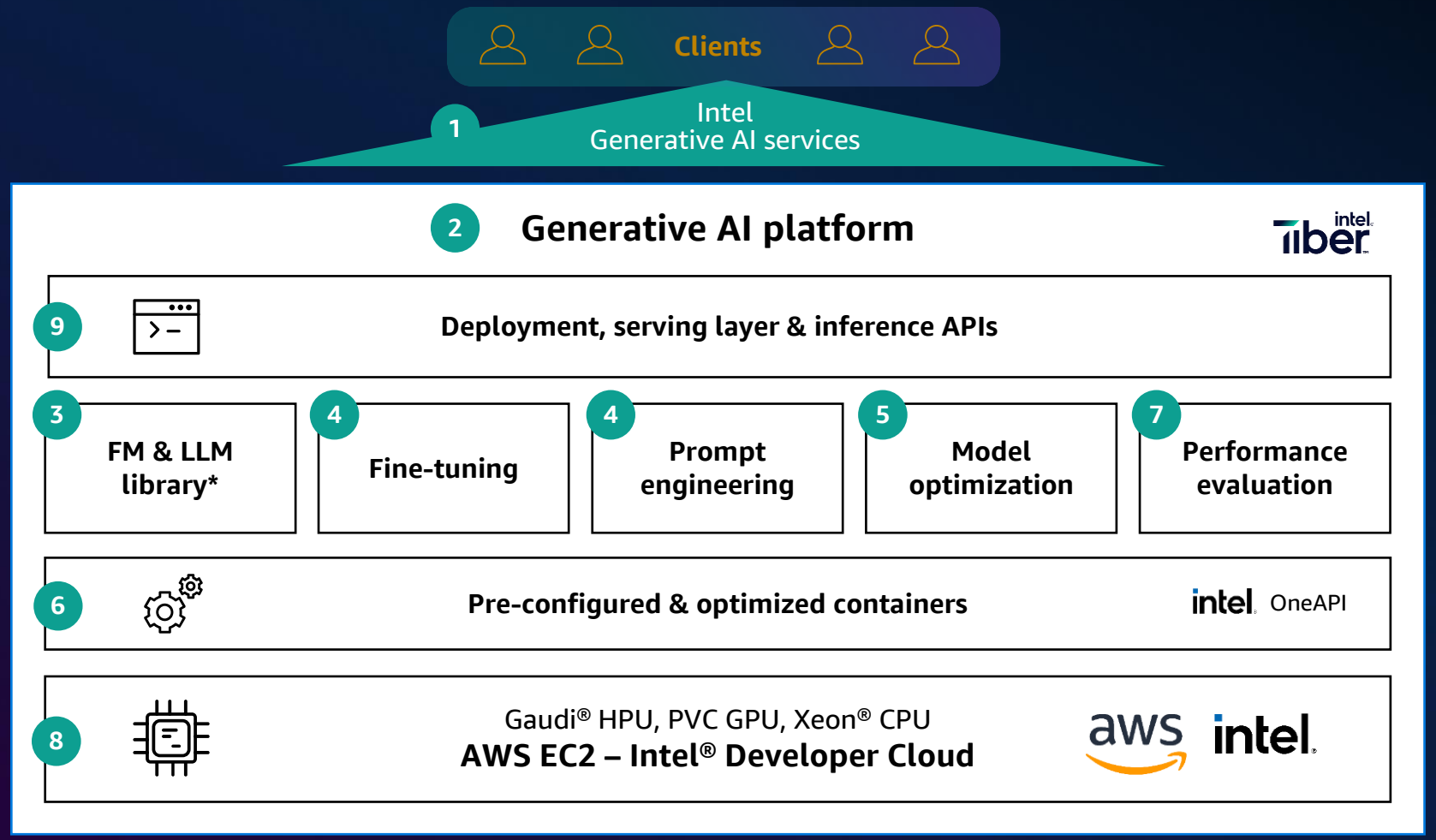
We facilitate consistent client outcomes and evangelize success for the benefit of other clients.

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI and LLM architecture, powered by Intel:

A step-by-step process



* MPT, Falcon, GPT-J, Llama 2 and other open source commercially available models or client BYOM

1. Define use case and create solution architecture.
2. Set-up Intel's MLOps tool, cnvrg.io, as collaboration environment with preconfigured features and access to compute on cloud/IDC.
3. Choose the model that meet performance & accuracy or BYOM.
4. Customize the model by fine-tuning or Intel's fastRAG tool to improve relevancy of responses.
5. Optimize the model using Intel tools like INC or OpenVINO™ to improve performance.
6. Configure models and optimize on selected hardware using Intel® oneAPI and AI/analytics toolkits.
7. Evaluate model/architecture choices and find the most optimized and performant alternative.
8. Set-up infrastructure for development and production workloads on public cloud or Intel® Developer Cloud.
9. Deploy models on scalable end points and set-up APIs for consumption.

aws

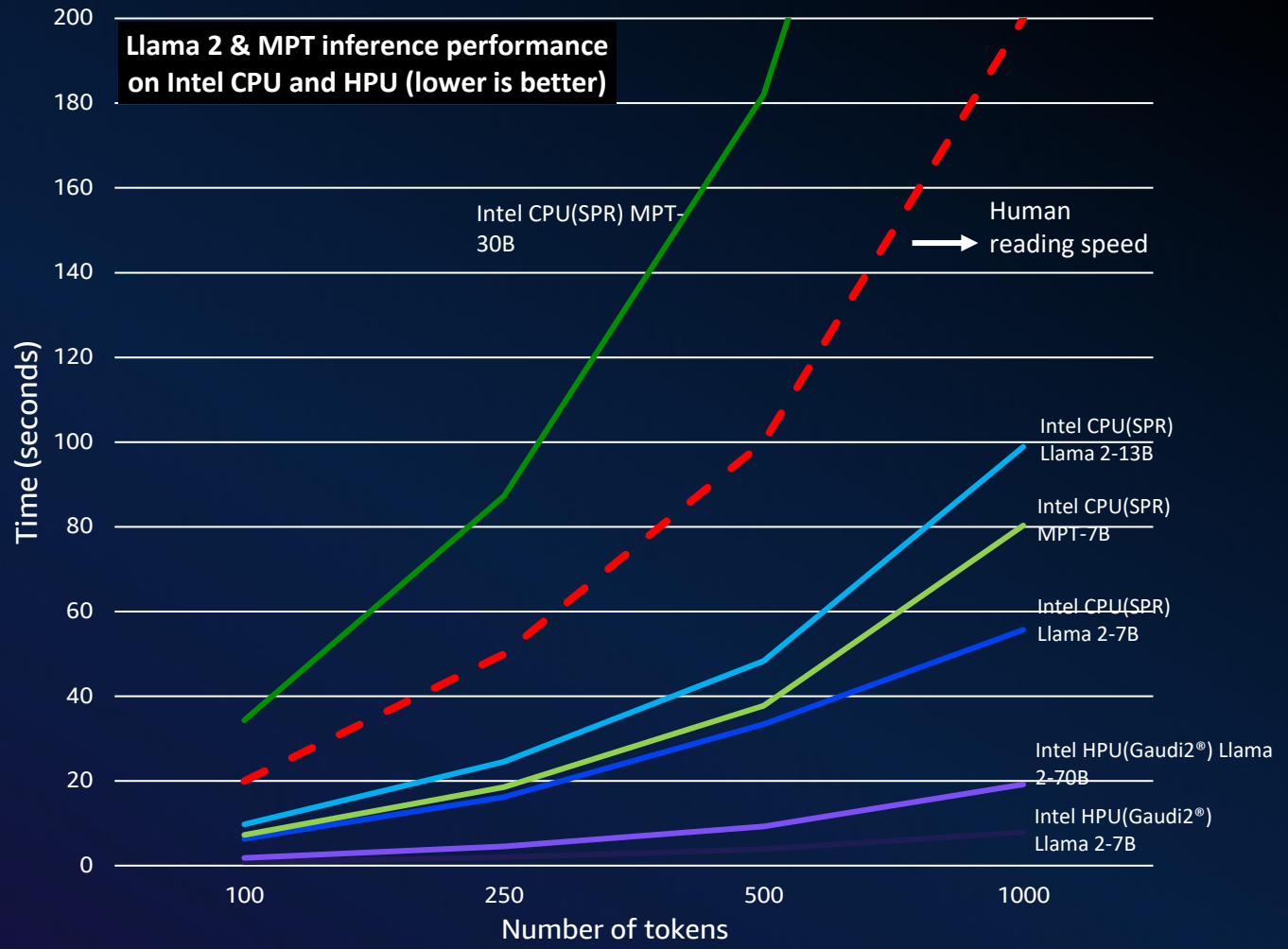
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Generative AI performance on Intel architecture

Intel's CPUs and HPUs provide cost-saving and performance optimization opportunities for generative AI inference and fine-tuning workloads

Key findings:

- Intel CPUs (Sapphire Rapids/SPR) available on AWS and Intel® Cloud provide improved price to performance for small to medium LLMs (sub 15B) such as MPT-7B and Llama 2-7B (see chart)
- Intel HPUs (Gaudi®) available on AWS and Intel® Cloud offer better inference price to performance for bigger LLMs such as Falcon 40B and Llama 2-70B
- Intel HPUs (Gaudi2®) on Intel® Cloud deliver better fine-tuning performance with less than one third of the price of alternative GPU instances on AWS
- Broad availability of Intel CPUs on the cloud makes them a great choice for LLM inference workloads at scale
- Confidential computing feature (Intel® SGX and Intel® TDX) embedded in Intel CPUs ensures security of data at the time of processing and is a differentiating factor



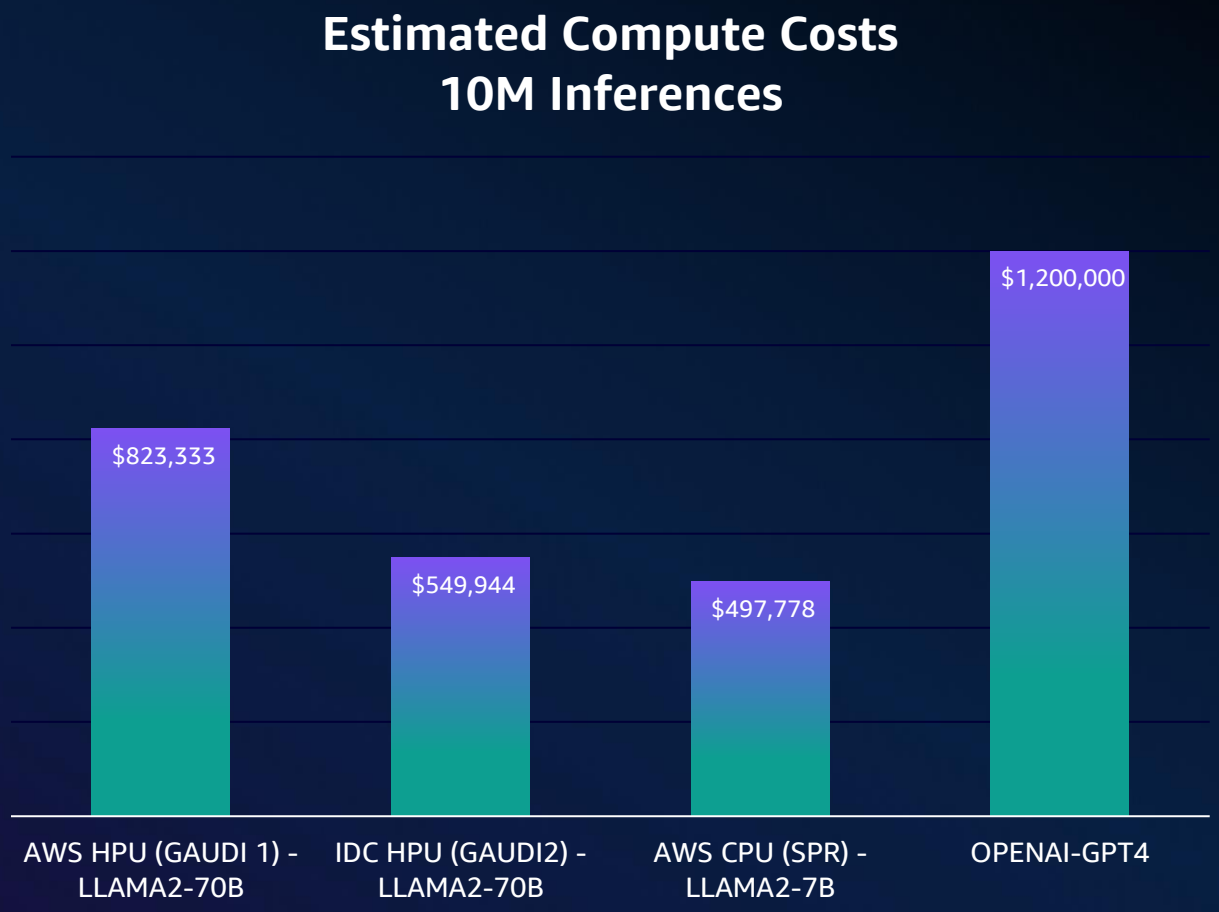
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cost analysis shows great opportunities for smaller LLMs on CPU and larger ones on HPU

- Cost calculations based on AWS and IDC on demand pricing and average 1k tokens per inference call

Findings/Assumptions

- Llama 2-70B model used here is arguably as factually accurate as Open AI (GPT4) and can be run on Intel's HPUs on AWS and IDC
- Llama 2-7B is the smaller version and has great performance on Intel's CPUs. When less output creativity and modality is needed or for edge use cases, this is the cheapest and most feasible option
- Running inference on Intel cloud HPU provides 50%+ cost savings
- Overhead costs for building and maintaining a scalable API like OpenAI are not included in cost calculations



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-tuning vs. RAG – When to use which technique

	Fine-tuning	RAG
I want my LLM to…	• Perform a new complex task • Answer in a new tone or format • Produce content in my style	• Answer factually • Have working knowledge of my companies documents • Use external or evolving information
Examples	• Custom Document Classifiers • Stylized Assistants • Domain-specific content generation • Guardrails	• Business-aware QA Chatbots • Service Ticket Handling • Knowledge based instruction following • Domain specific content generation
Challenges	• Need plenty of example training data • Not guaranteed to be factual • Evaluation is pivotal	• Accurate models require excellent retrieval • Adds another Vector DB architectural component

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

MLOps enablement and the use of Intel Gaudi2 & Xeon to deliver client POCs

1- AllerGen for a Retail Client



Developed generative-AI enabled AllerGen app for retail customer to provide grocery shoppers with real-time & personalized allergy alerts, analysis as well as alternate product recommendations.

2- Process Automation for Automotive Client



Developing a generative AI & RAG-enabled process automation capability to speed up RFP generation in Automotive industry.

3- Customer care chatbot for telco client

Fine-tuned Llama2 70B to interact with customers and extract data from internal systems to provide improved responses.

4- Multi-modal knowledge retrieval for utility client

Used generative AI to “read” and query millions of asset documents and use that gained knowledge to accelerate field management decisions.

Consumption layer: Endpoints, APIs, OpenAI Wrapper

Sustainability

MLOps & LLMOps – Intel Tiber

Multimodal - LangChain

Responsible AI

Inference at Scale

CPU - SPR

HPU – Gaudi2®

Fine-tuning PEFT-RLHF

HPU – Gaudi2®

Vector store – RAG

CPU - SPR

Model evaluation & metrics

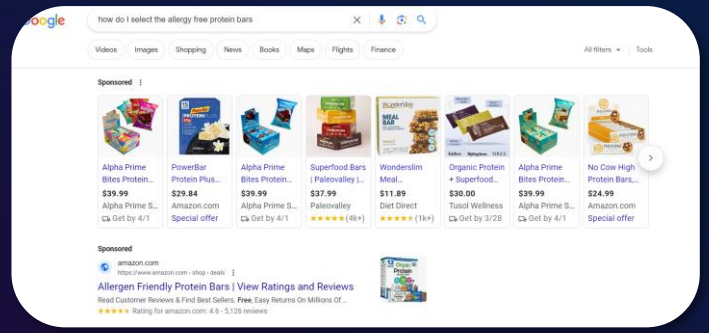
aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Copyright © 2024 Accenture. All rights reserved.

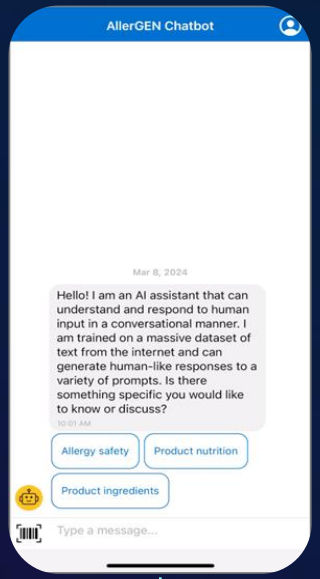
A paradigm shift is emerging within the shopping journey ecosystem (digital to start)

Search becomes the start and end of the purchase journey



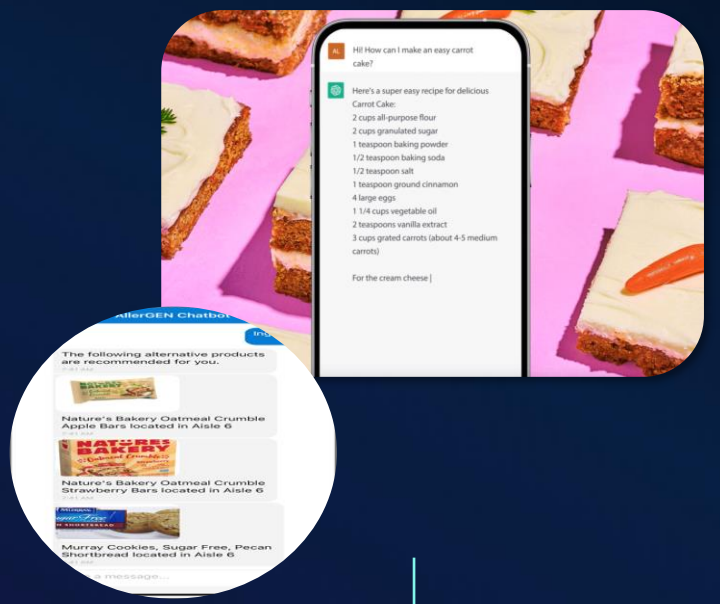
Google – Bing –You.com - Perplexity

LLMs become influencers

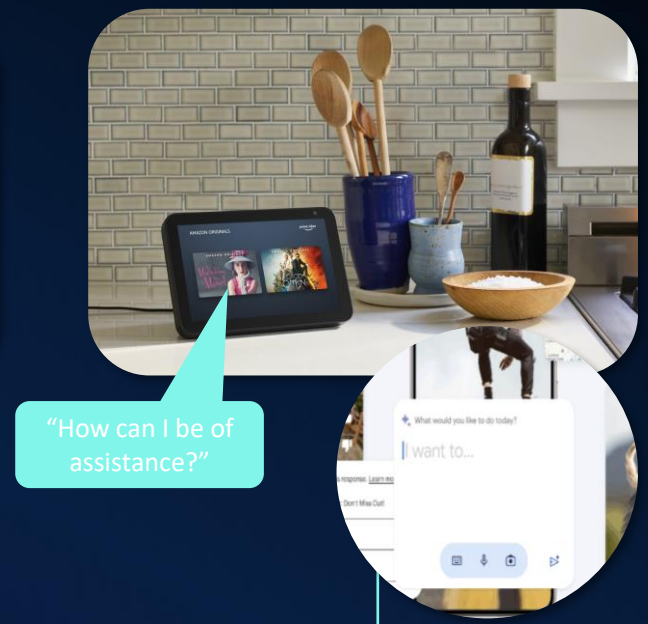


Allergen Chatbot

<Shopping> Assistants become useful



Chatbot Assistants



Alexa LLM, Google Assistant with Gemini

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customer use case: AllerGEN

Business challenge

Develop an Advanced Product Allergen Scanner, Analysis and Recommendation System that provides users with personalized allergy profiles, personalized analysis and alerts as well as alternate recommendations. The objective is to empower users to make informed choices about allergen-free products and enhance their overall well-being.

Solution

- **Advanced Allergen Scanner:** Develop a barcode scanner system that accurately recognize complex ingredient information and potential allergens even in cases where the variations of ingredients are not explicitly listed.
- **User Allergy Profiling:** Personalized warnings on scanned products containing allergens that match the user profiles.
- **Product Recommendations with Allergy Insights:** Leverage a database of products categorized by allergen content, ingredient analysis, and user preferences to provide suggestions on similar allergen-free products. Provide insights into emerging allergens, cross-reactivity patterns, or trends in allergic reactions.

Benefits:

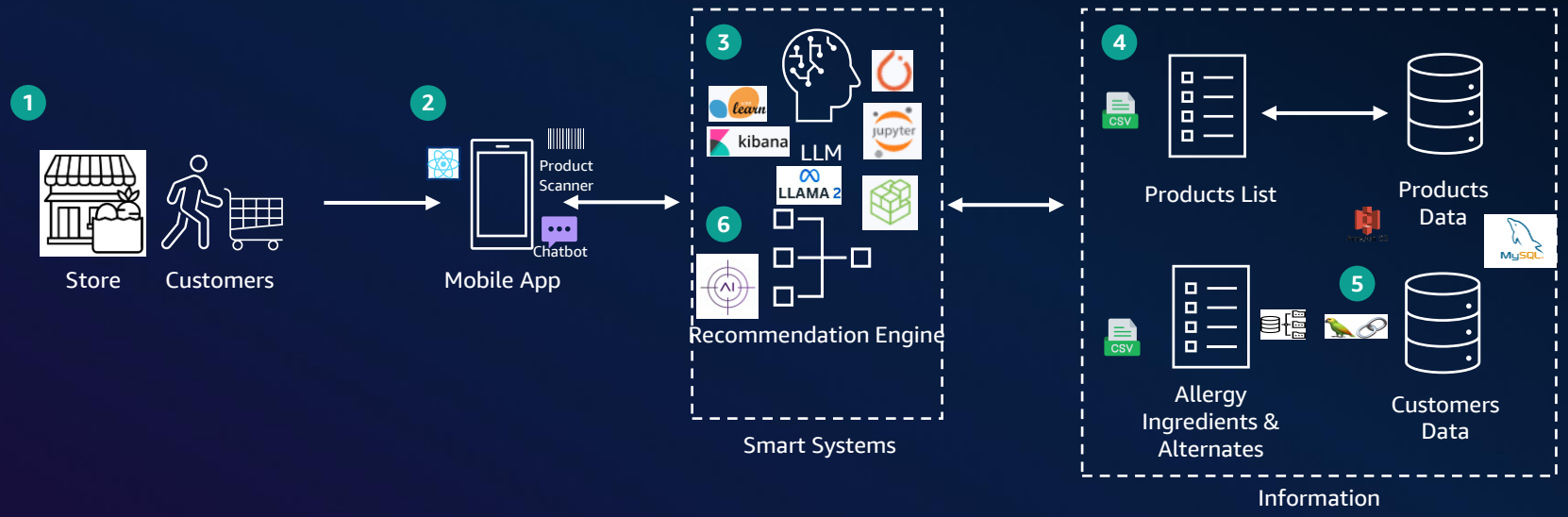
- Social Impact and Inclusivity
- Customer Engagement
- Customer Safety and Health
- Market Differentiation



aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Copyright © 2024 Accenture. All rights reserved.

AllerGEN use case flow

Generative AI enables customers to make informed decisions about allergen-free products while also improving their overall wellbeing by providing personalized allergy profiles, allergy warnings, and alternative recommendations





1. Customers enter a physical store to shop
2. Customers access through secure technologies via their mobile app
3. A generative AI LLM enhances the customer shopping experience by responding to questions asked in the AllerGEN chat or product barcodes scanned.
4. Products and customer data are analyzed using data analytics and generative AI insights to make real-time adjustments to product offerings and customer interactions.
5. AI-powered technologies identify allergen-free products based on customer allergy profiles.
6. After it warns customers about an allergy, the app recommends alternate products tailored to the specific customer needs via the recommendation engine.
7. In summary, the customer receives a one-of-a-kind, personalized experience through real-time tracking of the product list, allergy ingredients, and alternates, as well as a chat assistant that understands and fulfills their needs.

- ✓ Entirely new in-store & online shopping experience
- ✓ Support the customer journey with intelligent bots
- ✓ Elevate the customer experience
- ✓ Contextual, hyper-personalized in product interactions

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Copyright © 2024 Accenture. All rights reserved.

Key takeaways

1. Intel hardware is an effective way to run many AI/generative AI workloads

2. Intel software extends the benefits of Intel hardware

3. Engaging with Intel’s partner ecosystem can help you scale your solution or solve your use case

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Notices & disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

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 does not control or audit third-party data. You should consult other sources to evaluate accuracy.

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.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



Please complete the session survey in the mobile app

Mack Lamon
mackenzie.lamon@intel.com

Colin Best
colin.best@accenture.com

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Portfolio focus on Infrastructure and AI Application & SW

	Infrastructure Software SW-driven simplicity to abstract HW complexity	AI Application & Software Cloud-to-edge Hybrid AI platform for edge AI applications
Definition	Edge Node Stack Fleet Management (Day 0/1/2) Life Cycle Management Network Security & Unified Control Wired & Wireless Connectivity Ops & Services at Edge	Hybrid AI AI Model Life Cycle Management Privacy Preserving Data Analytics Application Orchestration Secure Software Supply Chain Ops & Services at Edge
Portfolio *Not exhaustive*	Intel® Edge Controls for Industrial ipdk Infrastructure Programmer Development Kit DPDK Data Plane Development Kit Intel® FIDO Device Onboard Intel® Infrastructure Power Manager	OpenVINO™ intel GeTi™ intel EDGE INSIGHTS FOR INDUSTRIAL Edge AI Box
Value Prop	Reduce TCO Secure Onboarding Network Security IT OT Sprawl Edge Native	Ease of Use Deployment Scale Cloud to Edge Quick & Secure Start Management Scale

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI solutions available in the

AWS Marketplace

Real Time Continuous Optimization

By Granulate

1 external review

Granulate employs real-time continuous performance optimization to allow organizations to handle compute workloads with 60% fewer servers while improving performance by 40%, with no code changes required.

Available as a SaaS with multiple pricing options

With more coming to the Marketplace soon!

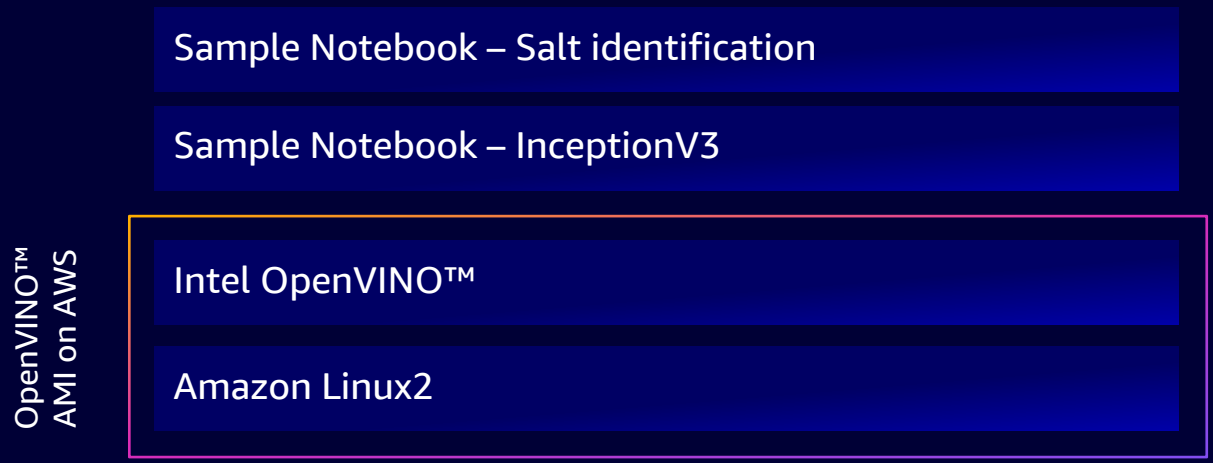
Intel® Distribution of OpenVINO™ Toolkit

By Intel | Ver 2023.2.0

★★★★☆ 3 AWS reviews

OpenVINO™ is an open-source toolkit for optimizing and deploying AI inference. OpenVINO™ toolkit is a comprehensive toolkit for quickly developing applications and solutions that solve a variety of tasks including emulation of human vision, Generative AI, LLMs, automatic speech recognition, natural...

Architecture diagram showing OpenVINO™ AMI on EC2 and its samples



Source: AWS Marketplace: Intel® Distribution of OpenVINO™ Toolkit (amazon.com)
AWS Marketplace: Granulate (amazon.com)

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI is evolving rapidly

Underlying data technologies….



Graph Databases



Data Lakehouse



Data Fabrics



Synthetic Data

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]

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

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

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

80% of enterprises will use Gen AI by 2026[7]

AI as disruptive as the Internet

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

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

Growth of **large model** sizes (1T+ parameter models)

Growth of **smaller, nimbler models** (~10B parameters)

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.

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

August 22, 2023

Gartner Forecasts Worldwide AI Chips Revenue to Reach $53 Billion in 2023

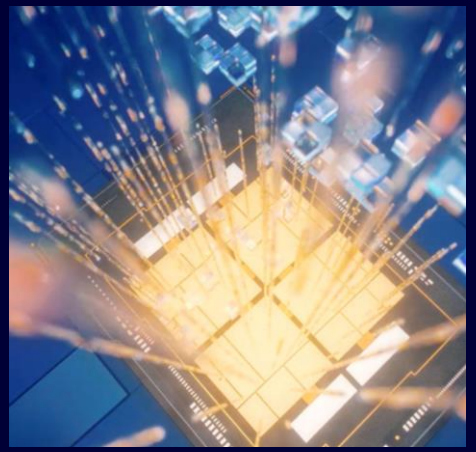
By 2027, AI Chips Revenue Is Forecast to More Than Double

Semiconductors designed to execute artificial intelligence (AI) workloads will represent a $53.4 billion revenue opportunity for the semiconductor industry in 2023, an increase of 20.9% from 2022, according to the latest forecast from Gartner, Inc.

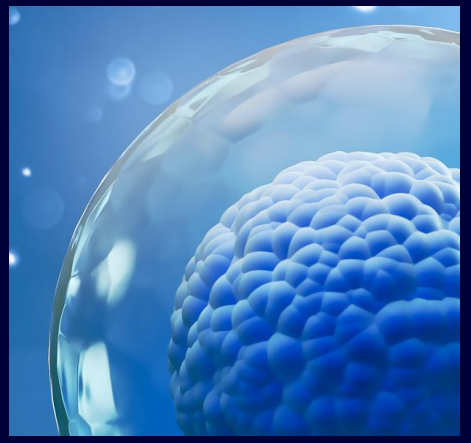
"The developments in generative AI and the increasing use of a wide range of AI-based applications in data centers, edge infrastructure, and endpoint devices require the deployment of high-performance graphics processing units (GPUs) and optimized semiconductor devices. This is driving the production and deployment of AI chips," said Alan Priestley, VP Analyst at Gartner.
AI semiconductor revenue will continue to experience double-digit growth through the forecast period, increasing 25.6% in 2024 to $67.1 billion (see Table 1). By 2027, AI chips revenue is expected to be more than double the size of the market in 2023, reaching $119.4 billion.

aws
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Bringing AI everywhere









Intel® AI product positioning

Enabling AI in every platform…from client and edge to data center and cloud.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

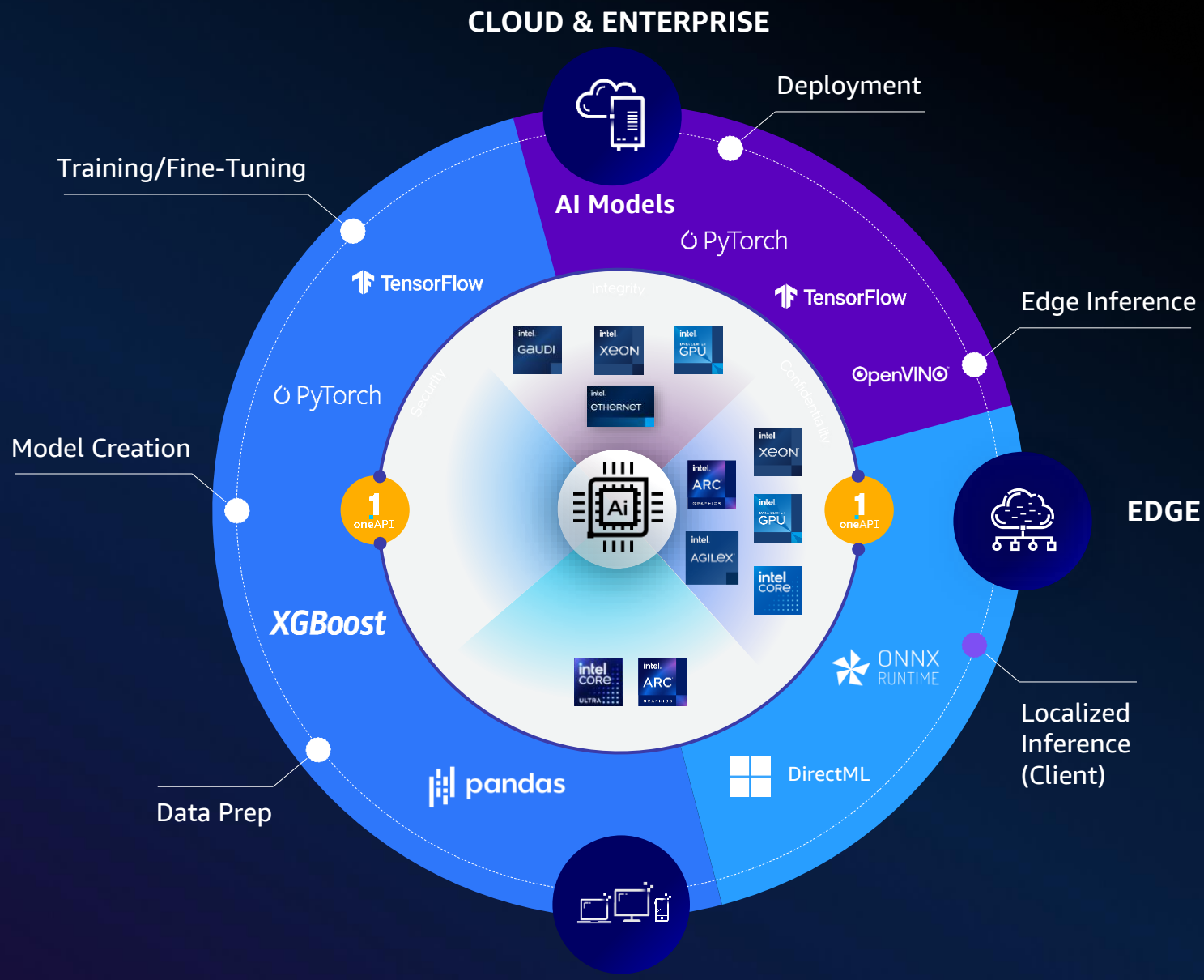
Core customer needs

1. Deploying scalable AI in less time with more choice
    1. Cost consideration, in your environment, BYOM (model that works best for you)
2. Deliver AI architecture that scales economically
3. Keeping data secure

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AI Continuum

Bringing AI everywhere



Note: Intel® Core™ Ultra processors integrate NPU low power inference engine from 15th Gen processors onwards.

© 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.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key objectives to touch on

- Intel products can play a key role in your AI architecture
- Accenture’s platform a leading-edge example of delivering peak value in GenAI
  - Low-code, no code, immediately deploy scalable AI endpoints
- Intel/Accenture partnership extends the impact of our ecosystem

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Structure

1. Intel and AI everywhere
2. Low hanging fruit to extend the impact of your AI efforts
    1. Hardware / Software
    2. Partnership with Intel
3. Partnership in action : Accenture
4. Call to action

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.