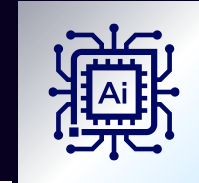




Bringing AI everywhere.

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

Artur Bar





Bringing AI
everywhere



\$

AI is transforming how we work and live everyday

From facial recognition to personalized learning.
AI is here to improve the life of every person on the planet.

Data Encryption

Facial Recognition

Personalized Learning

AI Based Rendering

Video Conference

Digital Assistants

Recommendation

Purchase Recommendation

Code Generation

Robotics Vision

Inventory Management

Recommendation Systems

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 AI²

More than
75%

of enterprise-
managed data
will be created &
processed outside
the data center or
cloud by 2025³

\$300B

worldwide GenAI spending set to
exceed \$300B by 2026¹

50%

of edge deployments
will involve AI by 2026⁵

80%

of PCs to be
AI PCs by 2028⁶

of enterprises will use
Gen AI by 2026⁷

AI as disruptive as the Internet

Generative AI predicted to
add up to \$4.4T of value to
global economy by 2040⁴

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/;](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.

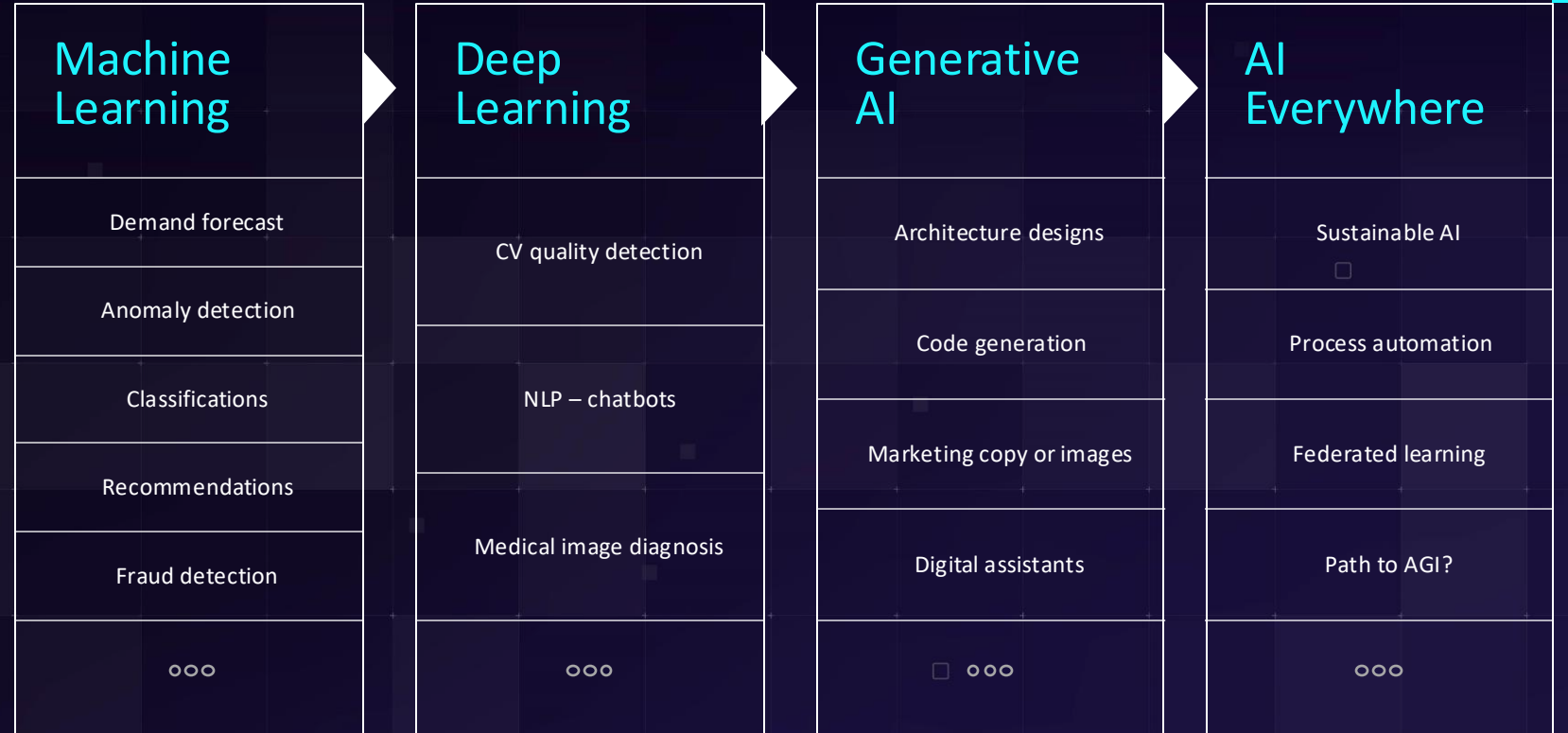
4. Worldwide Artificial Intelligence Spending Guide (IDC)

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 rapid growth of AI



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)

Why is AI challenging?

Complexity

Rapidly growing number of methods, capabilities, data types and sizes, and infrastructure requirements to run AI

Costs

Increasing costs due to increased compute demand as AI becomes more widely adopted and consumed

Operationalizing

Many steps and skill sets required to get AI from proof of concepts through to production in a scalable, sustainable process

Data security and privacy

Activating sensitive or regulated data globally while remaining secure and compliant

Human impact

Ensuring AI technology advances responsibly, ethically and equitably with a comprehensive approach that lowers risks, improves lives and optimizes benefits

□ Business outcomes

Bringing AI everywhere

Enabling AI in every platform with world-changing technology that improves the life of every person on the planet.



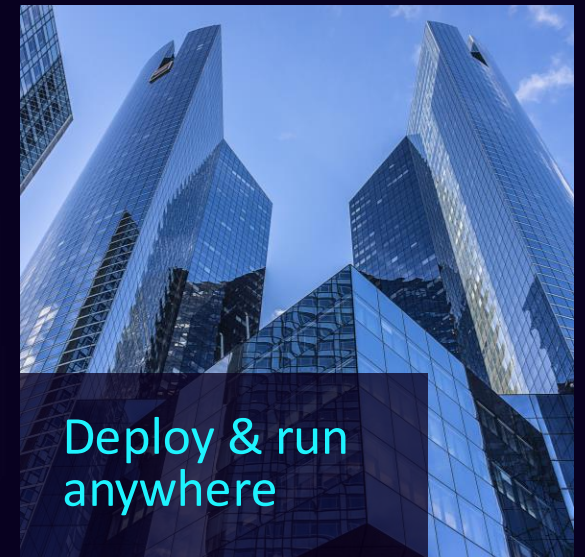
Accelerate innovation



Maximize value



Stay secure & responsible



Deploy & run anywhere

Accelerate innovation

Drive new opportunities and key business outcomes with optimized performance using the modern software tools preferred by AI developers.

Open Programmability

PyTorch

TensorFlow

python™

Hugging Face

UXL
United Acceleration Foundation

SYCL

Choice Compatibility

1
oneAPI

OpenVINO™

SYCLomatic

OpenXLA

OpenAI
Triton

AI

MLIR

Trust AI inference

Secure AI

Secure workloads

Secure models

Secure data in use

Scaled Develop and test

Intel® Developer Cloud

cloud.intel.com

Open, accelerated, connected computing for AI

Multivendor

Multiarchitecture

Hardware / Architecture

intel
XEON

intel
Gaudi

intel
GPU
FLEX SERIES

intel
GPU
MAX SERIES

intel
CORE
ULTRA

intel
ARC
GRAPHICS

intel
ETHERNET

Small, medium and large models

Full system, full cluster

Latest Intel CPUs, accelerators
and software

Maximize value

Choose the hardware and software optimized for all your AI compute needs and available today.

Unlock new and enhanced experiences with

the AI PC: 300+ AI-accelerated ISV features throughout 2024



Accelerate AI with the broadest hardware portfolio that

matches compute and connectivity with your complete AI needs



Create new opportunities from the client and edge to the data center & cloud

with hardware optimized by software **and open standards** for tomorrow's AI



Stay secure & responsible

Protect your AI initiatives and data, confidentiality, responsibly and in compliance with regulations with the built-in security features across Intel platforms.

Protect

Help protect sensitive, private data, models and usage



Compliance

Comply with security and privacy regulations



Trust

Boost protection with Intel® hardware-enabled security



Software reliability



Workload and data protection



Foundational security

[Learn more](#)

Deploy & run anywhere

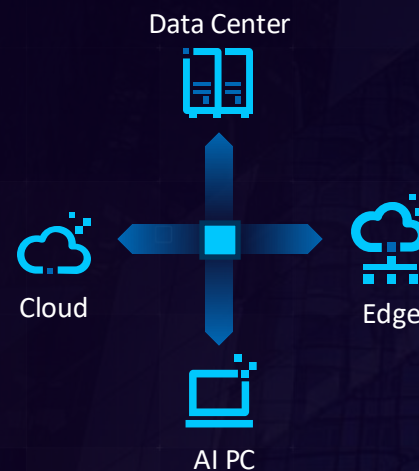
Go from concept to production faster from the client and edge to the data center and cloud

using a common AI runtime
to avoid manual recoding

OpenVINO™



on optimized solutions that
run on any device or architecture

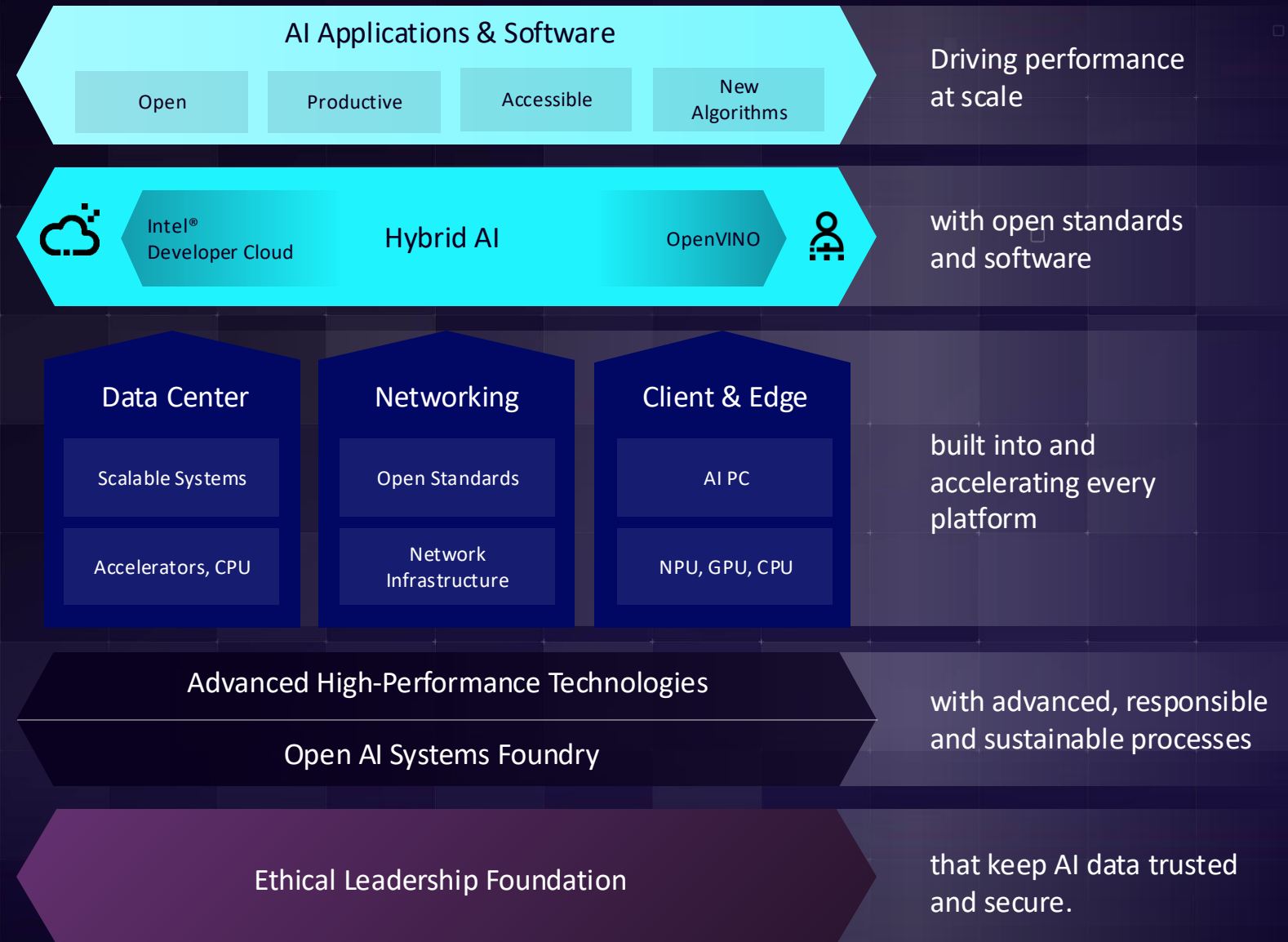


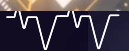
with client-to-cloud security that
protects sensitive data and
models



Intel's AI stack

Bringing AI everywhere





Bringing AI
everywhere



+



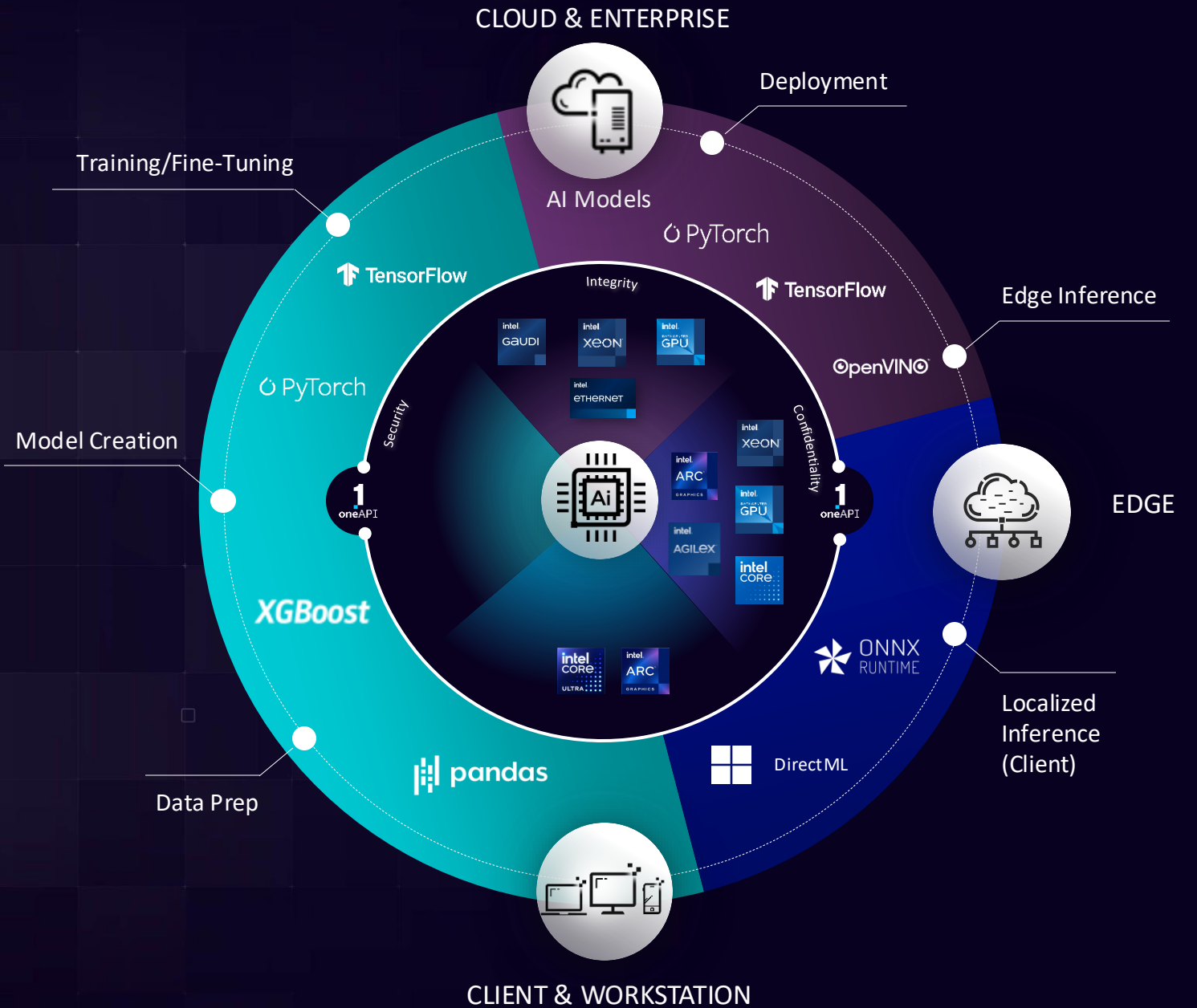
\$

Intel® AI product positioning

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

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.

□ Bringing AI everywhere

Intel AI portfolio

Open Software Environment



Deep Learning Acceleration



Dedicated Deep Learning Training and Inference

General Acceleration



AI Visual Inference, VDI, Media Analytics

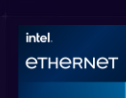


Parallel Compute, HPC, AI for HPC, Data Center

General Purpose



Real-Time, Medium Throughput, Low Latency, and Sparse Inference



Medium to Small Scale Training and Fine Tuning

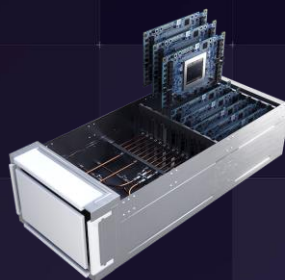


Edge and Network AI Inference



Inference on Client

Building scalable AI systems



Enterprise Training & Deployment of LLMs

Intel® Core™
Processors
Meteor Lake

Xeon Workstations Servers
and racks
4th Gen Intel® Xeon® SP

Training
Parameters

N/A

~ 1B + E2E data
Pipeline

Fine-tuning
Parameters

N/A

<~ 10B
1 - 8+ nodes

Inference

~20 TOPS

< 20B *

Dedicated DL training and inference of LLMs

Gaudi®2 Server
Dual-socket Xeon SP
with 8 Gaudi® 2 devices

1 MegaPOD
8 Gaudi® 2 Servers +
3 400G leaf switches

MegaPOD Cluster
Sized as needed

~ 20B

~ 70B

~ 350B

10Bs

Shared across models of many sizes

10B-100Bs

Shared across models of many sizes

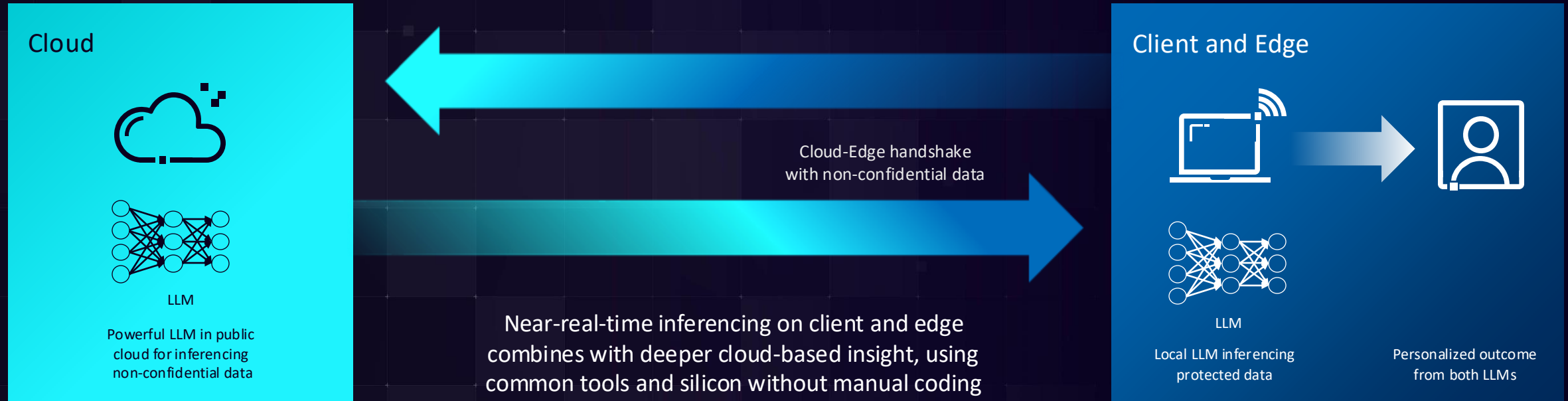
Large Scale
Distributed Training
10Bs - 1T+

Largest foundational models

Distributed Inference
Shared across models of many sizes

* Gaudi2 numbers are estimated based on 16-bit precision. 8-bit precision allows for larger models.

Hybrid AI: Seamless edge-to-cloud coordination



HEALTHCARE

Use Generative AI to automate creation of personalized emails to patients while protecting privacy



RETAIL

Inference video data at the Edge and gain insights from Generative AI without backhauling costs



ENTERPRISE

Use Generative AI for productivity gains without exposing confidential information to public cloud



□ Intel AI portfolio

AI PC powered by Intel® Core™ Ultra processors



3 Powerful
AI Engines

GPU

High Throughput
Ideal for AI-
accelerated digital
content creation

NPU

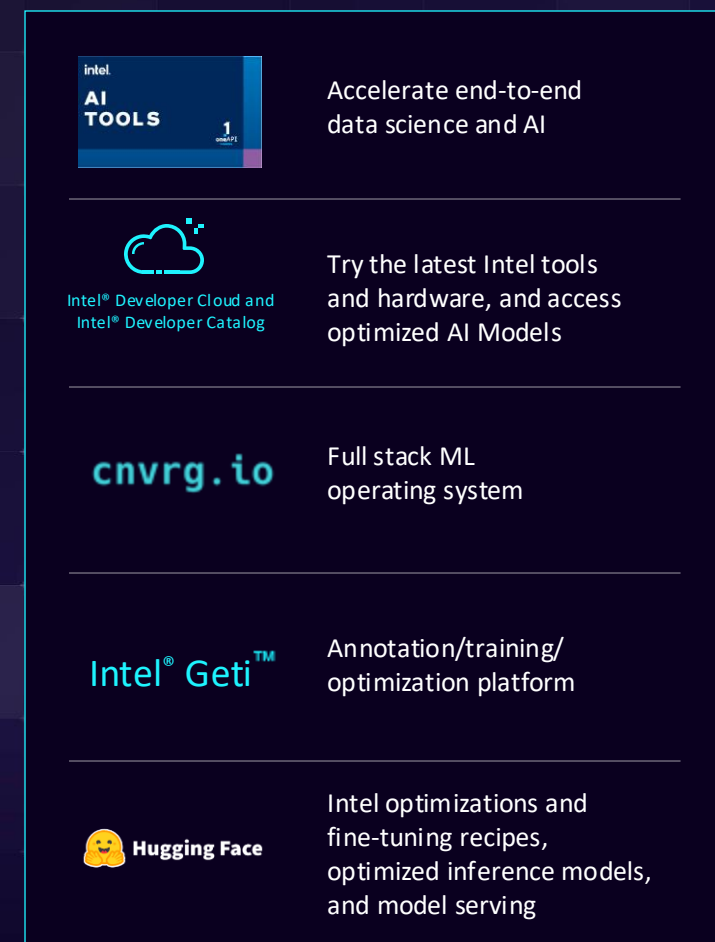
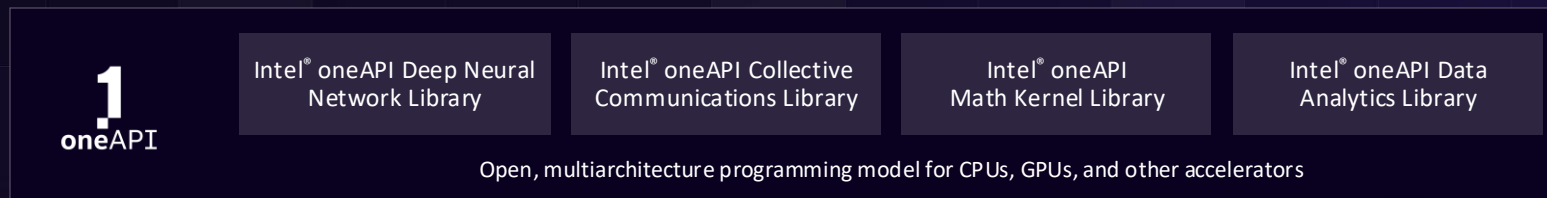
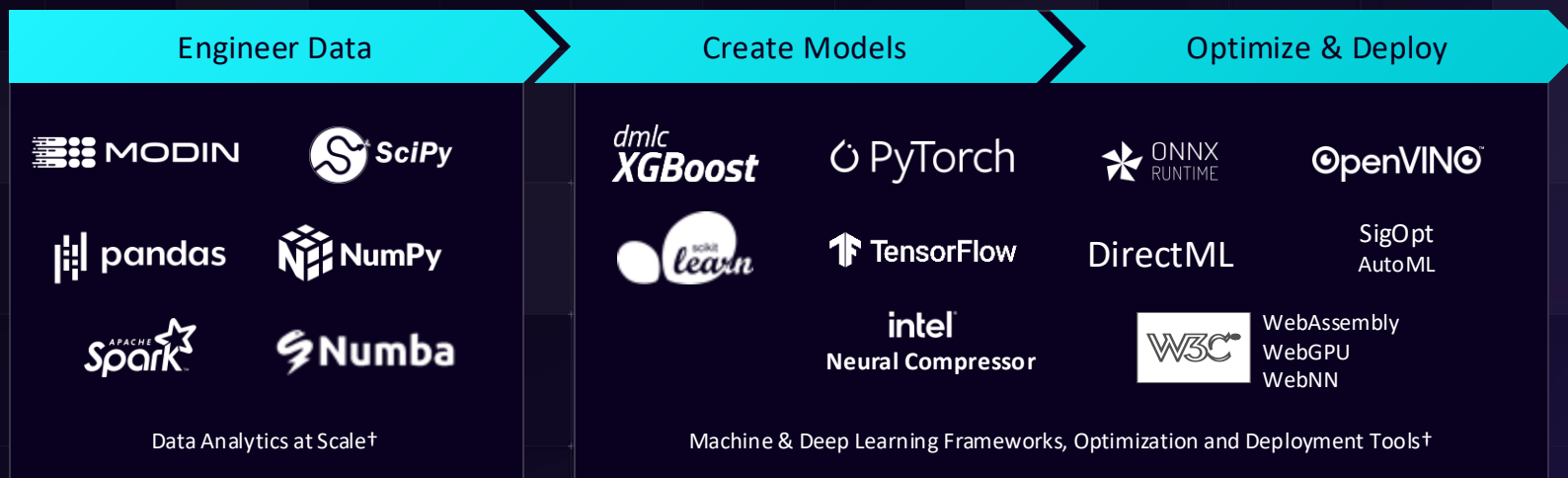
Low Power
Ideal for sustained AI
workloads and AI
offload for battery life

CPU

Fast Response
Ideal for
low-latency AI
workloads



Intel AI software portfolio



Note: components at each layer of the stack are optimized for targeted components at other layers based on expected AI usage models, and not every component is utilized by the solutions in the rightmost column

† This list includes popular open-source frameworks that are optimized for Intel hardware

□ Bringing AI everywhere

intel[®] Developer Cloud

Accelerate AI development using Intel-optimized software on the latest Intel[®] Xeon[®] processors, Intel[®] Data Center GPUs, and Intel[®] Gaudi[®] 2 accelerators.

cloud.intel.com



Get started with Intel

Get hands-on experience with the latest Intel[®] technologies. Empower your AI skills with Intel.



Early technology access

Evaluate pre-release Intel platforms and Intel-optimized software stacks.



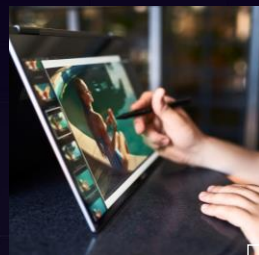
Deploy AI at scale

Speed up AI deployments with the latest tools and libraries on Intel[®] Developer Cloud.

AI use cases

Creator: Photo & video search & editing

Faster, more natural filters, higher quality previews & faster export times with automated, quicker searches.



Mainstream gaming

New AI features for in-game, 3D animation for added realism, transcription & chat translation.



Creator: Text to image

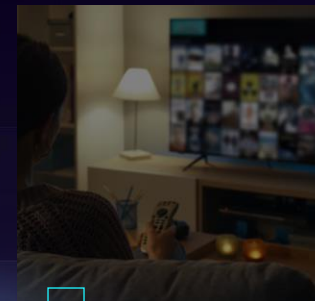
New AI effects & features for creating images with just a few descriptive words – marketing, advertising, design.

AI on the PC

“Unlocking the mundane”

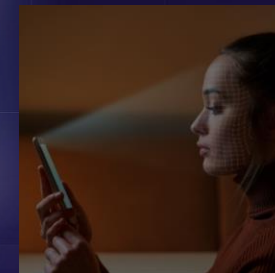
Collaboration/streaming

New AI capabilities for next-gen video conferencing, streaming and collaboration, preserving battery life.



Productivity

AI assistants for writing, creating, coding and offline features, like text & grammar prediction.



Accessibility

AI-assisted audio-visual capabilities for diverse user needs, making it easier to create and be productive on the PC.

With over 100 ISV's and 300 AI features planned for 2024, the Intel® Core™ Ultra offers a comprehensive AI PC experience with improved power & performance vs. prior generation*

AI use cases

Manufacturing

Using edge analytics and machine learning, Audi's Neckarsulm factory implemented a solution to inspect 100 percent of the five million welds they make each day while inferring the results of each weld within 18 msec. Audi was able to reduce labor costs by 30 to 50¹ percent and free employees for more valuable work at the company.



Cities & transportation

Ferrovial began the AIVIA Smart Roads initiative— aspiring to build roadways where conventional and autonomous vehicles can coexist. Along with Intel, Microsoft Azure, Capgemini, and others, Ferrovial is designing a ruggedized 5G- and AI-enabled roadside unit to relay near-real-time information to vehicles, warning of potential collisions, wrong-way driving, traffic conditions, and hazards, including pedestrians.



AI at the Edge

Retail

Nearly 80 percent of shoppers now prefer nontraditional checkout.² Nourish + Bloom is capitalizing on this trend with their autonomous grocery store. Just download the app, scan into the store, load groceries, walk out, and get a credit card charge. Customer journeys are faster and store personnel spend more time with customers and stocking shelves, according to data collected on each customer visit.



Healthcare

Primary ciliary dyskinesia (PCD) is a rare, inherited condition that causes defects in human cilia. It can cause respiratory distress early in life, leading to a lifetime of congestion, coughing, and chronic infections. Detection traditionally requires manual acquisition of cilia cross-sectional images for analysis using electron microscopy. To make it easy for medical professionals to apply their knowledge to an AI-based solution, the Royal Brompton team uses the Intel® Geti™ computer vision platform.



With over 84,000 edge AI deployments and counting, Intel offers a wide range of multipurpose compute and accelerators to deploy edge AI virtually anywhere.

1. "Intel® Helps Audi Achieve Precision Manufacturing & Industrial Automation," Intel. Results may vary.

2. "New Study Finds Self-Service Checkout Options Gaining Favor Across Demographic Groups," PYMNTS, Sept. 23, 2021, <https://www.pymnts.com/news/retail/2021/new-study-finds-self-service-checkout-options-gaining-favor/>

AI outcomes



+

AI innovation across the Data Center



Education

Teacher Assistant

Student Study Buddy

Parent Chat Portal

Health

Drug
Discovery

Doctor
Co-pilot

Patient Family
Chatbot

Finance

Algorithmic Trading

Customer Portfolio
Assistant

Risk / Credit
Assessment

Retail

Product Promotion

Customer Interface
and Sentiment
Tool

Image Shopping Aid

Government

Gov Services Chatbot

Document Search
Summarization

Live Language
Translation

Energy

Energy Consumption
Forecasting

Operational
Performance

Energy Trading
Assistant

Automotive

Autonomous Car
Development

Multi-language in car
aid

Supply Chain
Optimization

Manufacturing

Factory Automation

Predictive
Maintenance

Precision Agriculture

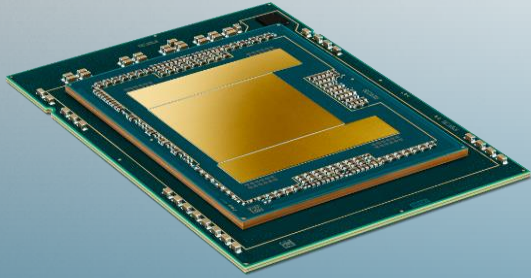
Telco

Personalized
Customer Services

Network Automation

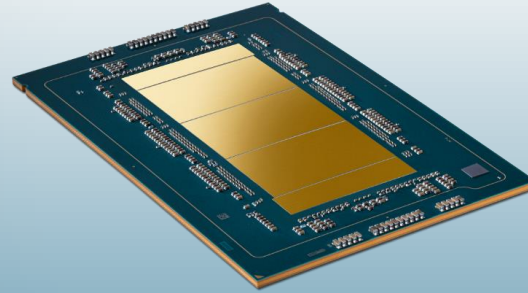
Operational
Performance

Next-Generation Data Center Solutions



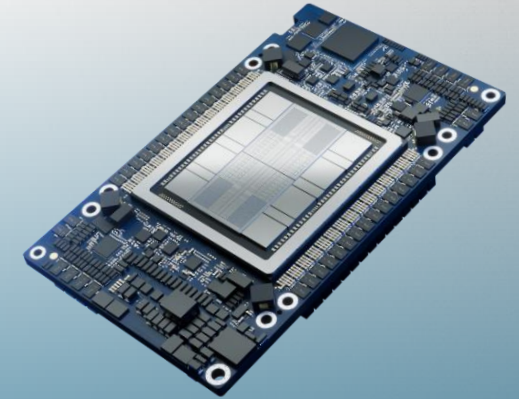
Intel Xeon 6 processor
(Efficient-Core)

Optimized for performance-per-watt in high-density compute and scale-out workloads



Intel Xeon 6 processor
(Performance-Core)

Optimized for per-core-performance in compute-intensive workloads



Intel Gaudi 3 AI Accelerators

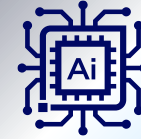
Bringing choice to GenAI with performance, scalability and efficiency

Open Ecosystems • Open Software • Open Platforms



Thank you

Bringing AI everywhere



The Intel logo is centered on a dark blue background with a subtle grid pattern. It features a small cyan square above the first 'i' of the word 'intel' in a white, lowercase, sans-serif font. A registered trademark symbol (®) is located to the right of the word.

intel®

□ Bringing AI everywhere

Legal Notices and Disclaimers

For notices, disclaimers, and details
about performance claims, visit
www.intel.com/PerformanceIndex or
scan the QR code:

```
lower = 100  
upper = 1000  
percent = (value - lower) / (upper - lower) * 100  
for sign in ['negative', 'upper', 'lower']:  
    if percent > 0:  
        for number in range(1, percent + 1):  
            print(sign, number, '%')  
    else:  
        for number in range(1, abs(percent) + 1):  
            print(sign, number, '%')
```



© 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.