HPC-AI Convergence: Intel is Bringing AI Everywhere

Dr. Jean-Laurent Philippe
Intel EMEA CTO

intel®

Our Transformation Journey

CPU Company	**XPU Company** **Multi-architecture** (CPU, GPU, FPGA, purpose-built ASICs)
Silicon Focused	**Platform Focused** **Solutions** including silicon, software and services
Traditional IDM	**Modern IDM 2.0** Disaggregated **design**, unmatched manufacturing scale, innovating across all areas key to product leadership

intel ai

AI performance, available right now

Bringing AI everywhere

Scalable AI computing platforms

AI Software & Services

Fast development with open source tools and workflows that offer choice and flexibility

AI PC

Workforce productivity: 300+ AI-enabled features on the AI PC

Edge AI

Efficiency at the edge: Performance designed for space and power constraints

Data Center & Cloud AI

AI acceleration with performance per dollar advantages

AI Networking

High speed connectivity: Standards-based connectivity with excellent scalability and cost advantages

intel ai

Intel Data Center & AI Priorities

Investing to return to sustained leadership

Re-invigorating

x86 leadership

x86 ecosystem advisory group

Intel® Xeon® roadmap execution

Innovating

in efficient AI systems

Power-efficient accelerators

Integrated rack designs

Open-source software

Diversifying

our portfolio to enable our customers to differentiate

Intel® Xeon® customization

x86 chiplets

Bringing AI everywhere

Intel helps you put AI to work fast while maximizing the value of your hardware and software investments

Accelerate Innovation & Maximize Value

intel ai

Intel helps businesses and organizations...

Fast development in an open ecosystem

Open source tools and workflows offer choice and flexibility

AI Software & Services

intel ai Software

Accelerate on an open ecosystem

Open frameworks and libraries to accelerate performance—includes PyTorch optimizations and GenAI models on Hugging Face

OpenVINO

Develop once, deploy everywhere

Deploy across diverse hardware with minimal code modification

intel tiber Edge Platform

Bring AI to the edge

Scalable edge and AI solutions on standard hardware with cloud-like ease

intel tiber Developer Cloud

Build in the cloud

Open, optimized AI models, frameworks, and libraries

oneAPI

No vendor lock-in

Optimized libraries for all major AI frameworks, with tools to migrate code from CUDA

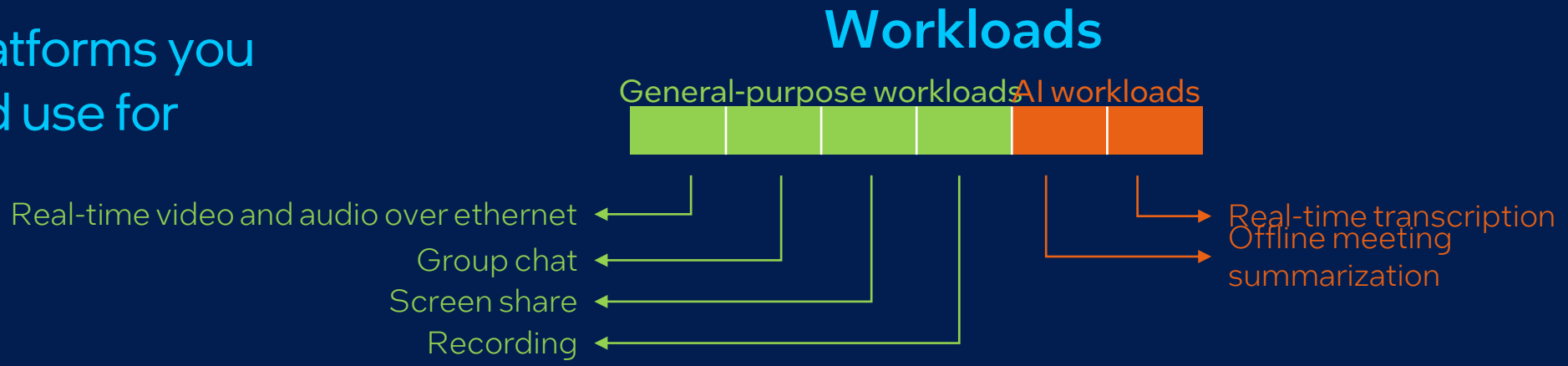
intel ai

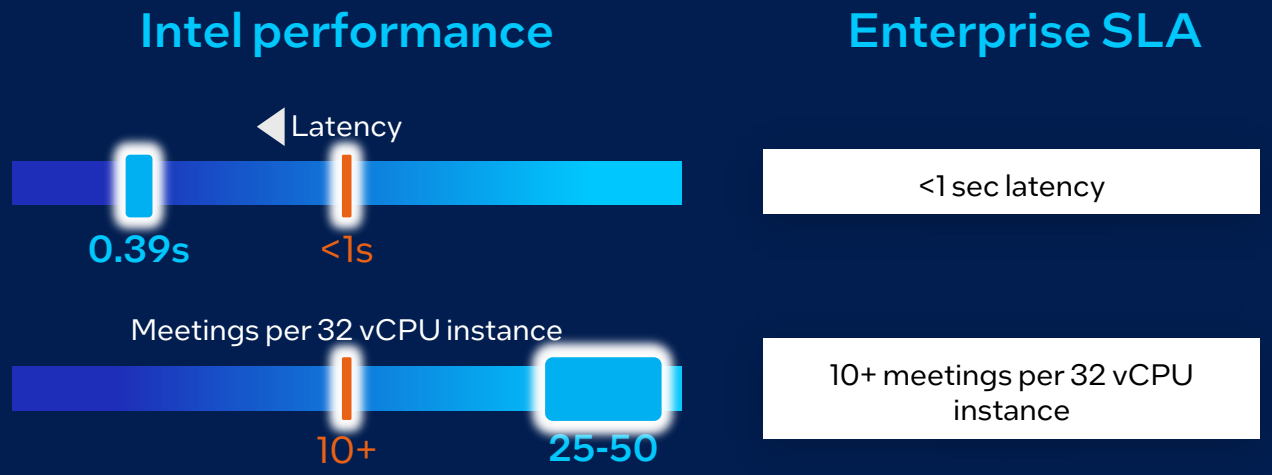
Higher resource utilization

intel XEON

Run AI on the platforms you already have and use for other workloads

Case study: Video conferencing service





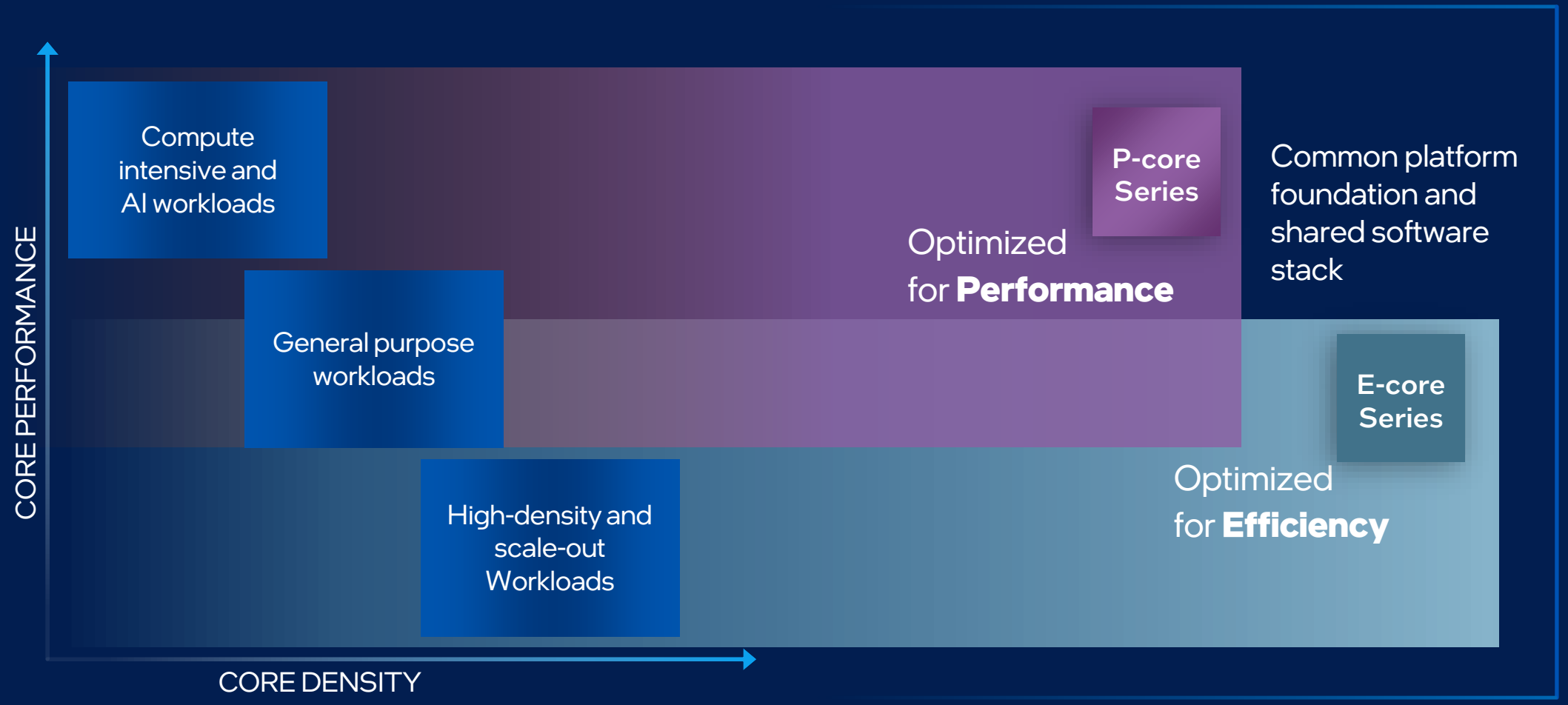
Intel® Xeon® Processor

The CPU designed for AI

A strong foundation for the entire AI pipeline—from data prep to inference—and as the host processor for AI accelerators

Intel® Xeon® 6 Processors

The best processors to meet diverse performance and efficiency requirements



Intel Xeon 6 Processors

P-core & E-core SKUs
Up to **288** cores
per processor

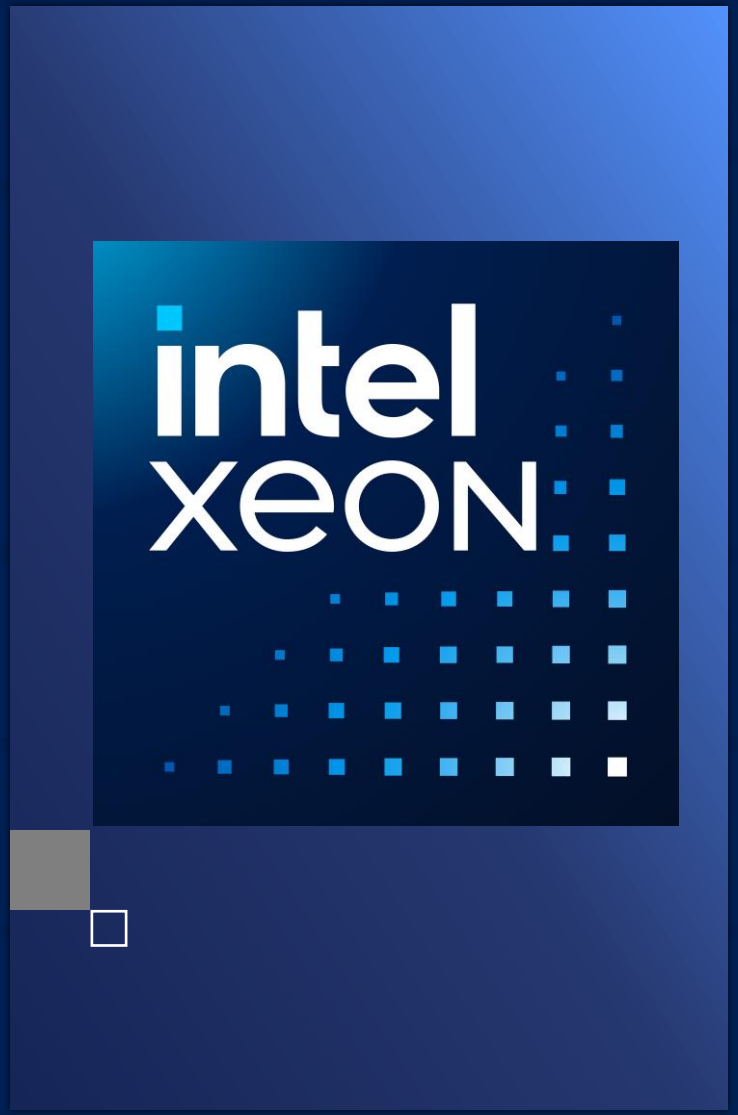
Increased Memory Bandwidth
Up to **1.7x (DDR5)**
vs. 5th Gen Intel® Xeon® CPU

Up to **2.3x (MRDIMM)****
vs. 5th Gen Intel Xeon CPU

Increased I/O Bandwidth
1.2x (PCIe 5)
vs. 5th Gen Intel Xeon CPU

Increased Inter-Socket Bandwidth
Up to **1.8x (UPI 2.0)**
vs. 5th Gen Intel Xeon CPU

Compute Express Link* (CXL*) 2.0
Type 1, Type 2 & Type 3



Socket Scalability
1 to 8S

Increased Last Level Cache
Up to 504MB total

Integrated IP Accelerators
Intel® QuickAssist Technology (Intel® QAT)
Intel® In-Memory Analytics Accelerator (Intel® IAA)
Intel® Data Streaming Accelerator (Intel® DSA)
Intel® Dynamic Load Balancer (Intel® DLB)

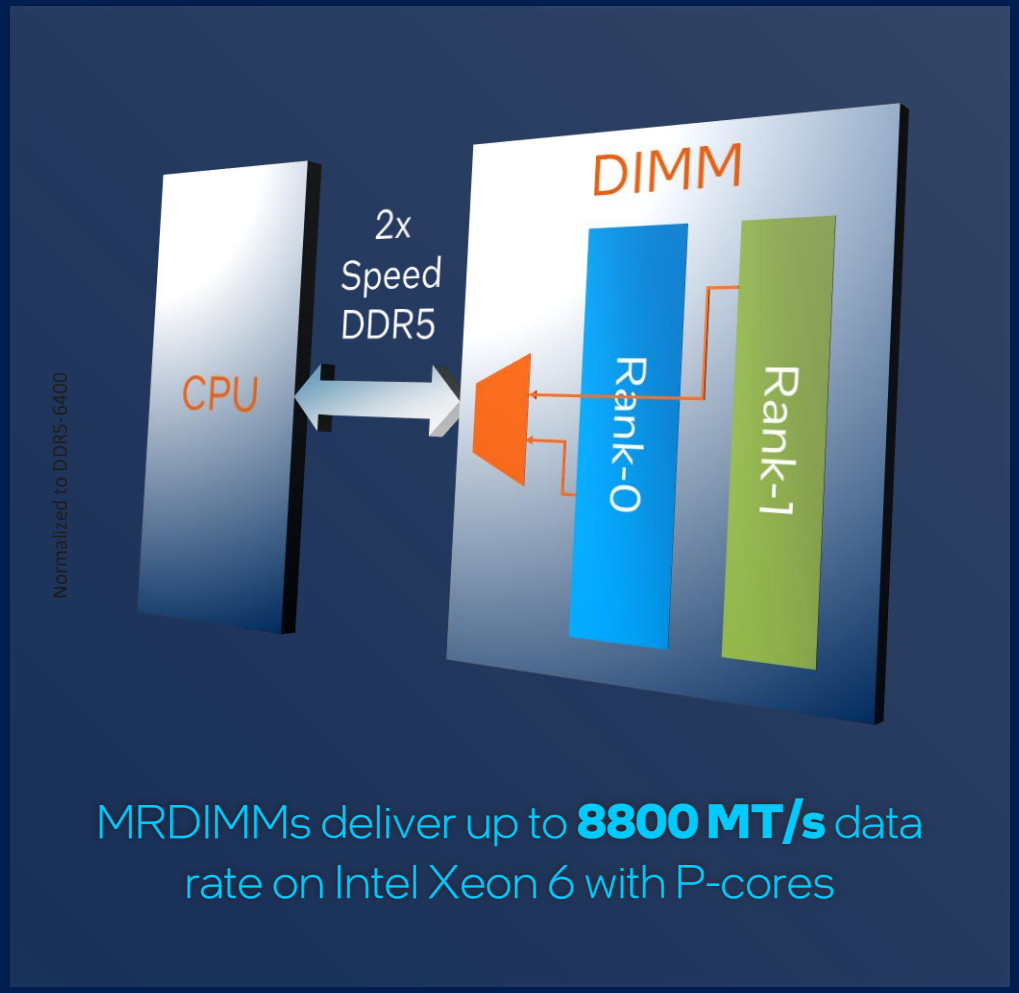
AI Accelerator built in
Intel® Advanced Matrix Extensions (Intel® AMX)**

Hardware Enhanced Security
Intel® Trust Domain Extensions (Intel® TDX)
Intel® Software Guard Extensions (Intel® SGX)

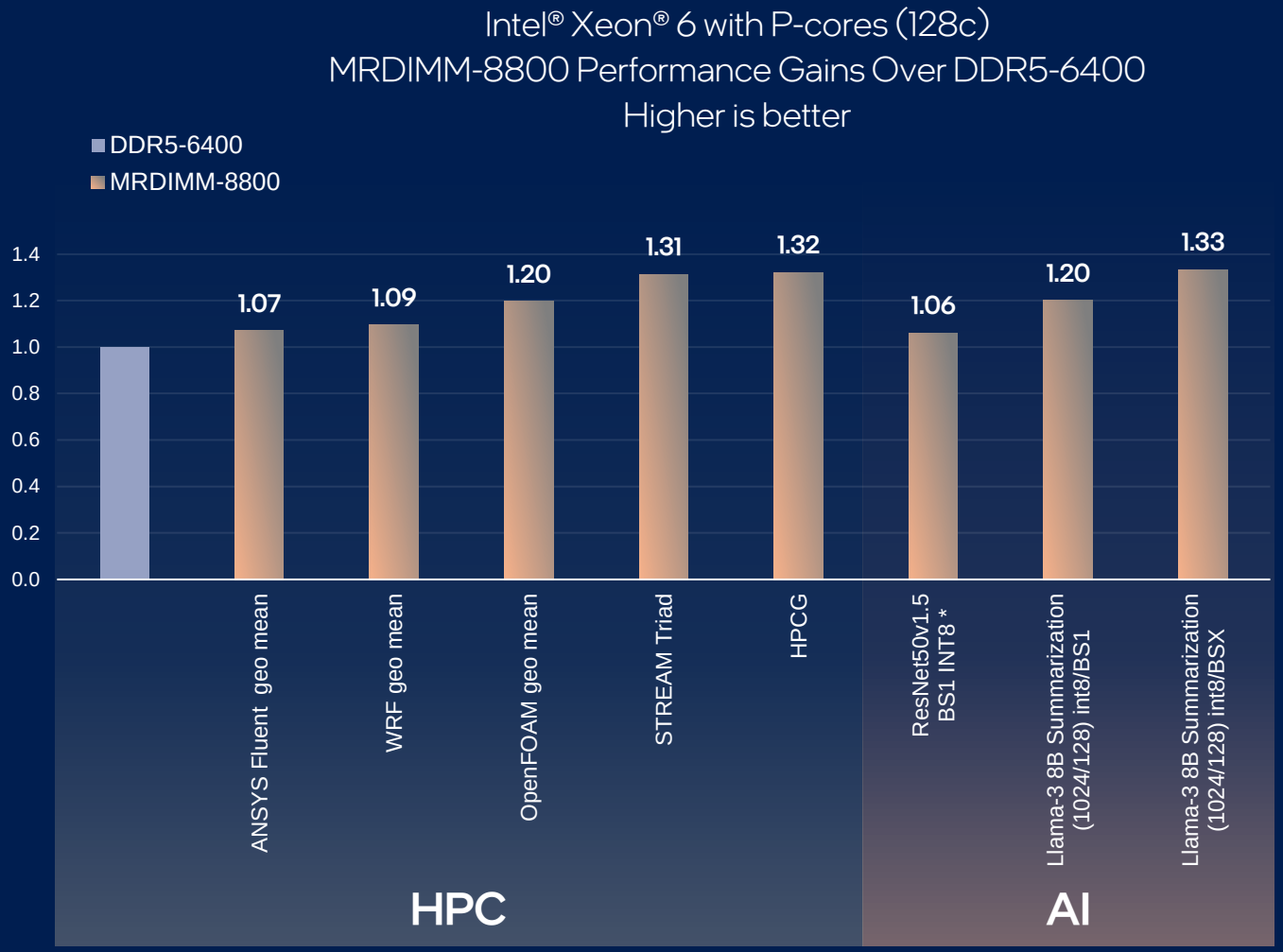
Common OS and Firmware
Simplify development and deployment

Multiplexed Rank DIMMs

First to market on Intel Xeon 6 processors with P-core



MRDIMMs deliver up to **8800 MT/s** data rate on Intel Xeon 6 with P-cores



See intel.com/processorclaims: Intel Xeon 6. Results may vary. * 6972P (96c) used.
This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark

Intel® Xeon® 6 with Performance-cores (P-cores)

Embrace and quickly scale AI everywhere with world's best CPU for AI

Run models up to

70B

parameters

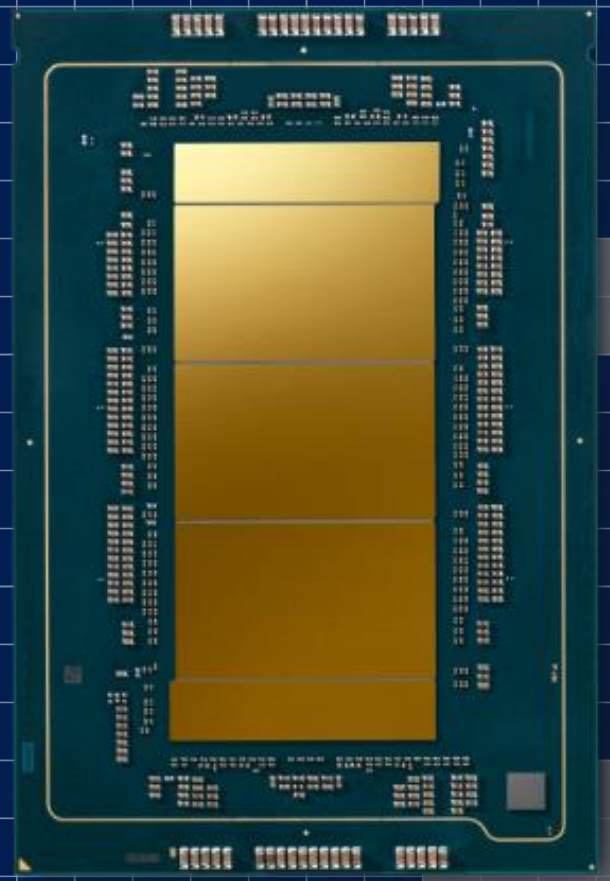
On Llama-2 models. See Vision 2024 section of intel.com/performanceindex for workloads and configurations. Your results may vary.

Up to

3.7x

higher AI performance vs. alternative competitive processor

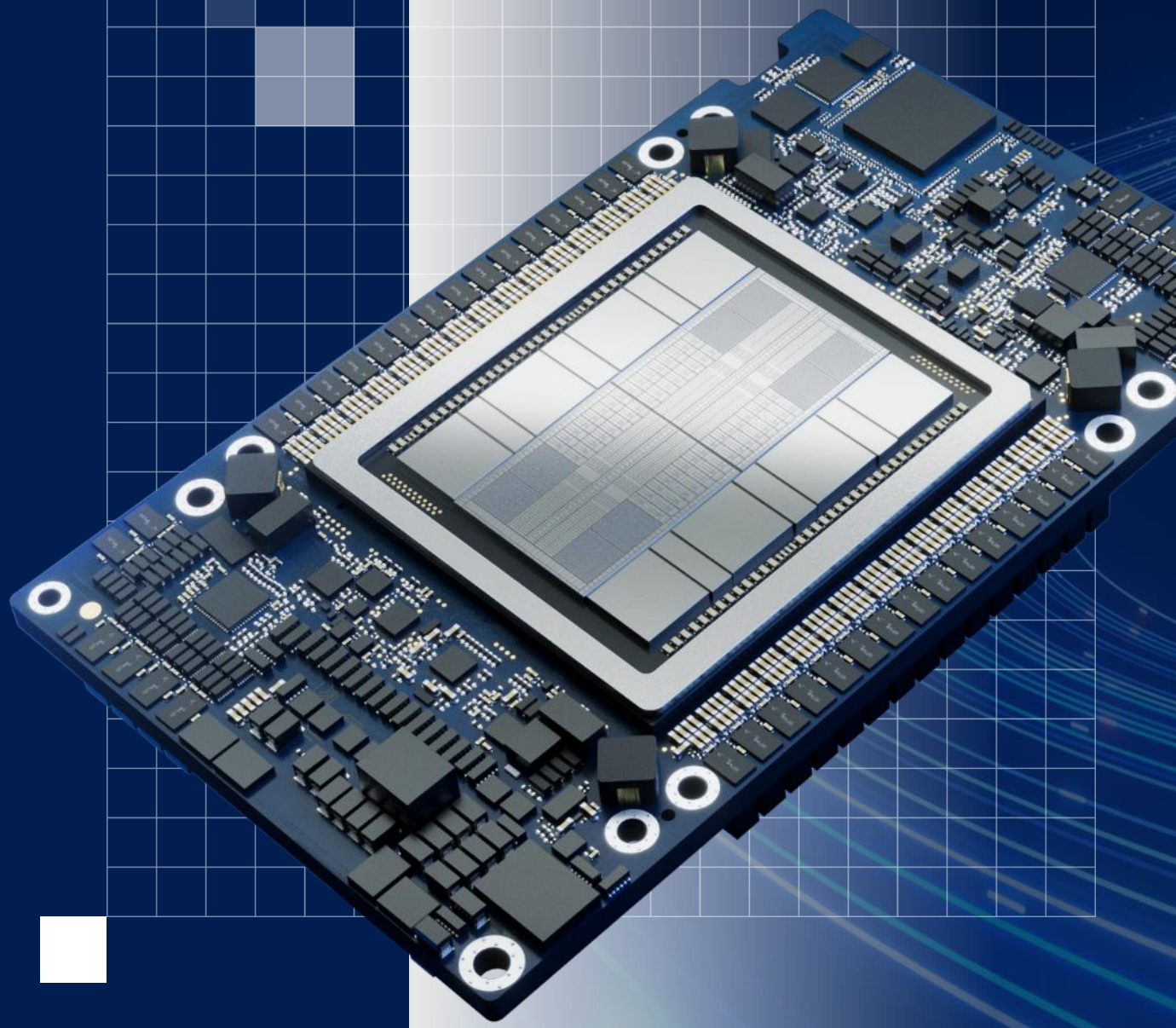
See [9A210] at intel.com/processorclaims: Intel® Xeon® 6. Results may vary.



Intel® Gaudi® AI Accelerator

Bringing choice to GenAI

Enterprise-ready performance, scalability, and efficiency that unlocks more GenAI solutions for more customers



GAUDI 3 - Delivering Price Performance

Advantage

1.19x

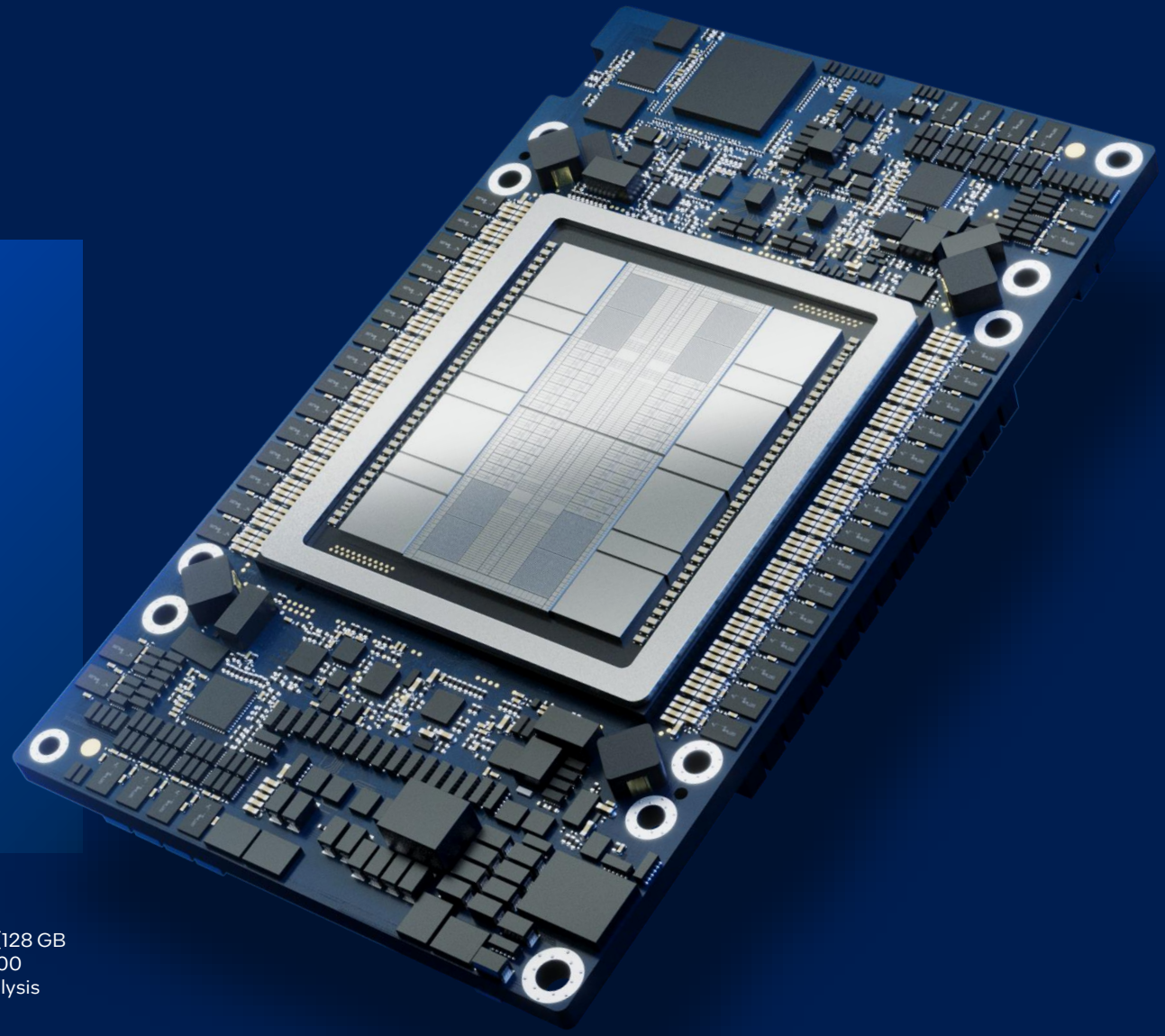
Inference Throughput
LLaMA 2 70B

Intel® Gaudi® 3 AI accelerator
Vs H100

~2x perf/$

Inference Throughput
LLaMA 2 70B

Intel® Gaudi® 3 AI accelerator
Vs H100



Source Intel measured results vs H100 data sources: https://github.com/NVIDIA/TensorRT-LLM/blob/main/docs/source/performance/perf-overview.md input-output sequences: 128-2048tps on 2 accelerators/GPUs. Intel results obtained on September 9th 2024. Hardware: Two Intel Gaudi 3 AI Accelerators (128 GB HBM) vs two Nvidia H100 GPU (80 GB HBM). Software: Intel Gaudi software release 1.18.0. See Nvidia link for H100 software details. Results may vary. Pricing estimates based on publicly available information and Intel internal analysis

intel CORE ULTRA

AI PC Momentum

40M

AI PCs shipped through 2024

intel ai

Creator: Photo & video search & editing

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

Collaboration/streaming

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

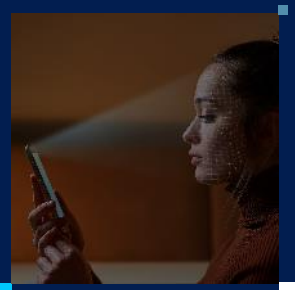
Mainstream gaming

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

AI on the PC

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.

Creator: Text to image

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

"Eliminating the mundane"

With over 100 ISVs and 300 AI features, the Intel® Core™ Ultra offers a comprehensive AI PC experience with improved power & performance vs. prior generation[1]

1. Intel Dec. 14, 2023, AI Everywhere event news release. AI features may require software purchase, subscription or enablement by a software or platform provider, or may have specific configuration or compatibility requirements. Details at intel.com/AIPC. Results may vary.

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® Tiber™ AI Cloud
Intel® Tiber™ Developer Cloud

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

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 performance, available right now

Bringing AI everywhere

Scalable AI computing platforms

AI Software & Services

Fast development with open source tools and workflows that offer choice and flexibility

AI PC

Workforce productivity: 300+ AI-enabled features on the AI PC

Edge AI

Efficiency at the edge: Performance designed for space and power constraints

Data Center & Cloud AI

AI acceleration with performance per dollar advantages

AI Networking

High speed connectivity: Standards-based connectivity with excellent scalability and cost advantages

intel

Thank you

intel


intel ai