

OCP Global Summit | 2025

Deploying Heterogeneous Hardware in the Era of Agentic AI

Saurabh Kulkarni

VP, AI Technical Product Management

Agentic AI is the Means to Unlocking Enterprise Value in AI

~80% of AI compute cycles in inference by 2028*

Diverse workload requirements

Throughput, low latency, perf/\$/W and system-level tuning required

Physical AI

Real-time inferencing for physical world interaction

Agentic AI

Billions of daily inferences across thousands of models

Frontier AI

Model training - massive, infrequent, hyperscaler-scale workloads

2015

2020

2025

2030

2035

*Source: Gartner, published 08/2024

Agentic AI Future

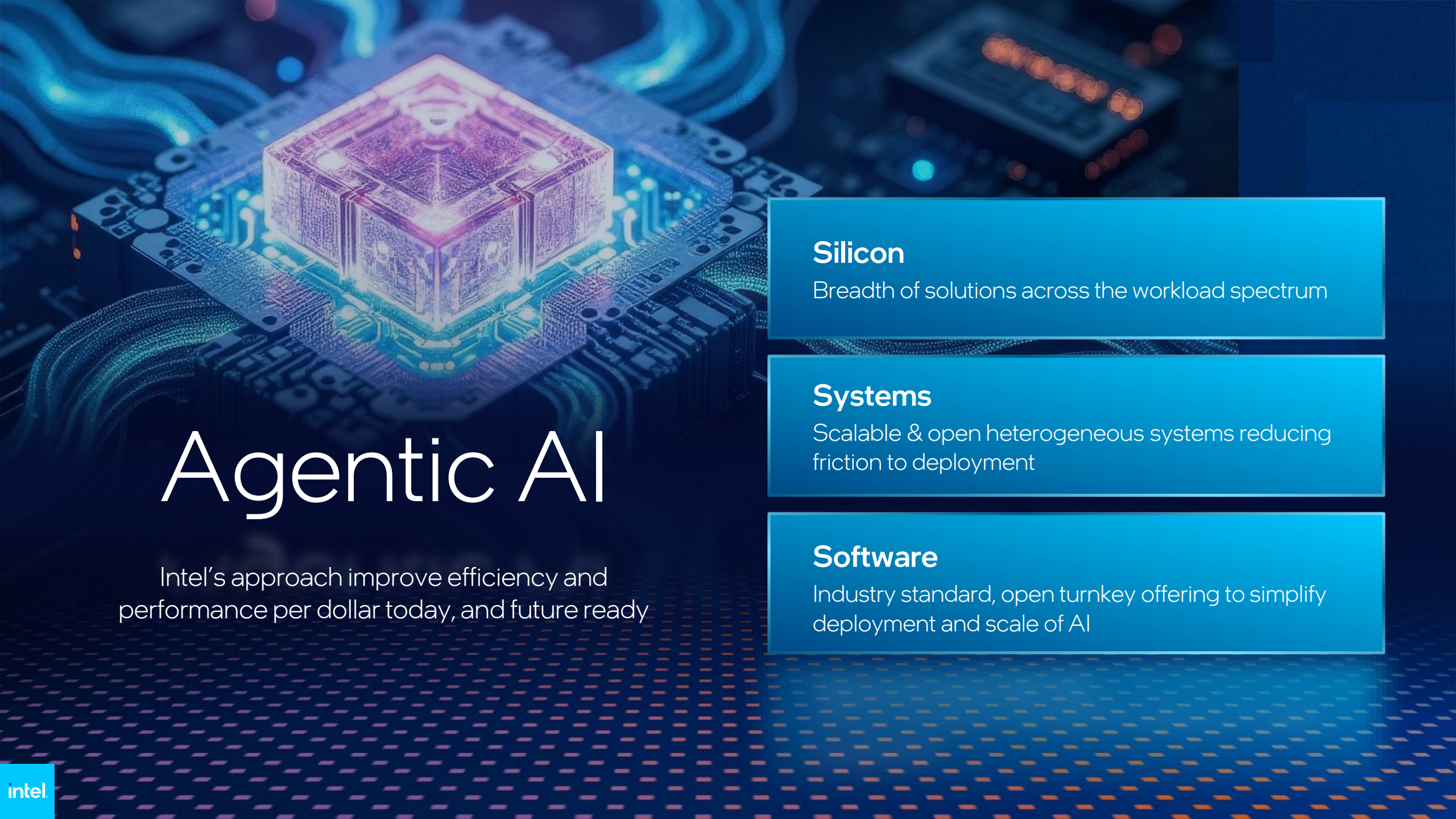
Complex multi-agent workflow

Diverse models & APIs

Different building blocks with diverse hardware req.

Need for hardware diversity and openness is exacerbated





Agentic AI

Intel's approach improve efficiency and performance per dollar today, and future ready

Silicon

Breadth of solutions across the workload spectrum

Systems

Scalable & open heterogeneous systems reducing friction to deployment

Software

Industry standard, open turnkey offering to simplify deployment and scale of AI



Agentic AI

Intel's approach improve efficiency and performance per dollar today, and future ready

Silicon

Breadth of solutions across the workload spectrum

Systems

Scalable & open heterogeneous systems reducing friction to deployment

Software

Industry standard, open turnkey offering to simplify deployment and scale of AI

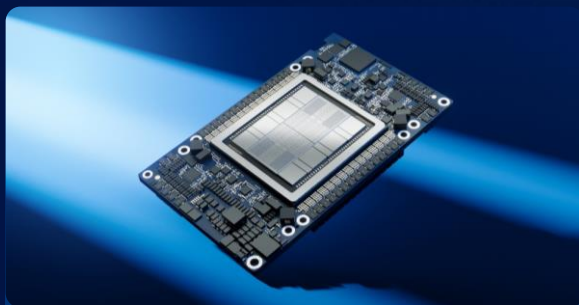
Intel AI Data Center Portfolio



Open heterogeneous strategy to deliver **systems, software & GPUs**

Intel® Gaudi® 3 AI Accelerator

Delivered to Meet Diverse Needs



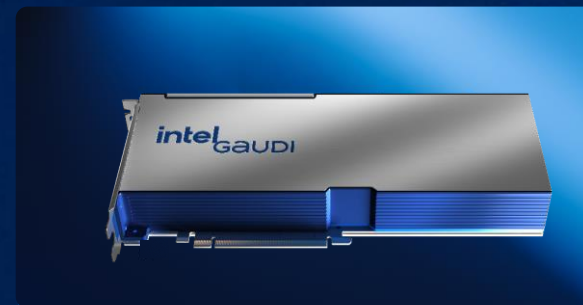
Accelerator Card

OAM-Compliant
(HL-325L)



Universal Baseboard

HLB-325



PCIe CEM

Add-In Card
(HL-338)



Bhavesh Patel
Dell Fellow, CTO office

Next Gen

Intel® Data Center GPU

Designed for tokens/watt built on a reliable open software stack

160GB LPDDR5x

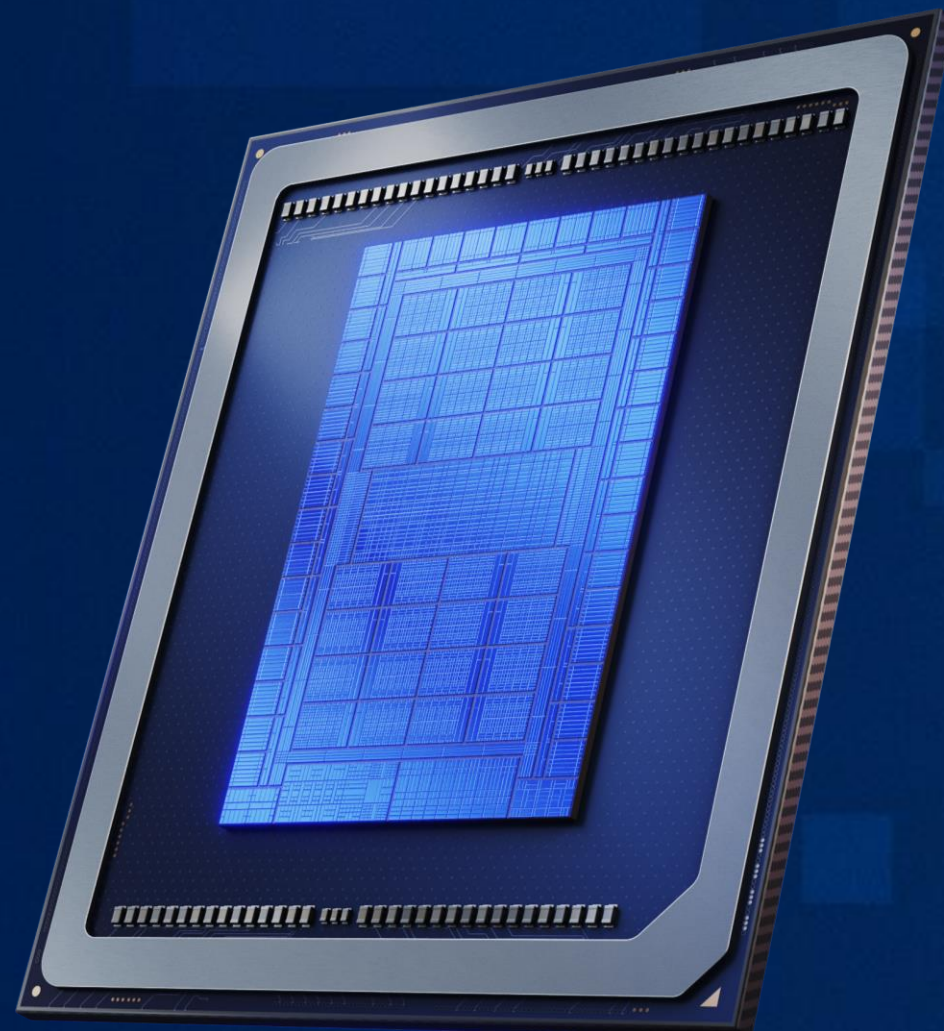
to meet the needs of memory intensive AI workloads

Latest Generation Xe 3 IP

General purpose GPU architecture for competitive perf/watt

Widest Range AI datatypes

Support from FP4/MXP4 to FP32 and FP64



AI Execution Roadmap

Unveiling an annual predictable GPU cadence

Deliver Agentic AI foundations

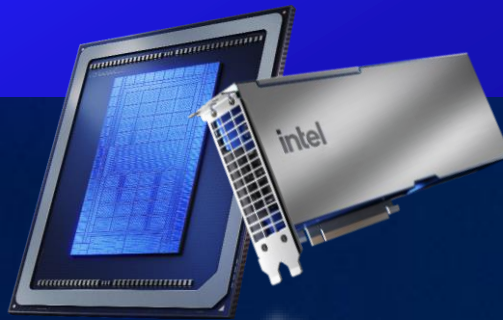
Heterogeneous infrastructure

A differentiated, systems-level approach including Xeon, Gaudi, Arc GPU

Scale Agentic AI solutions

Inference-optimized GPU

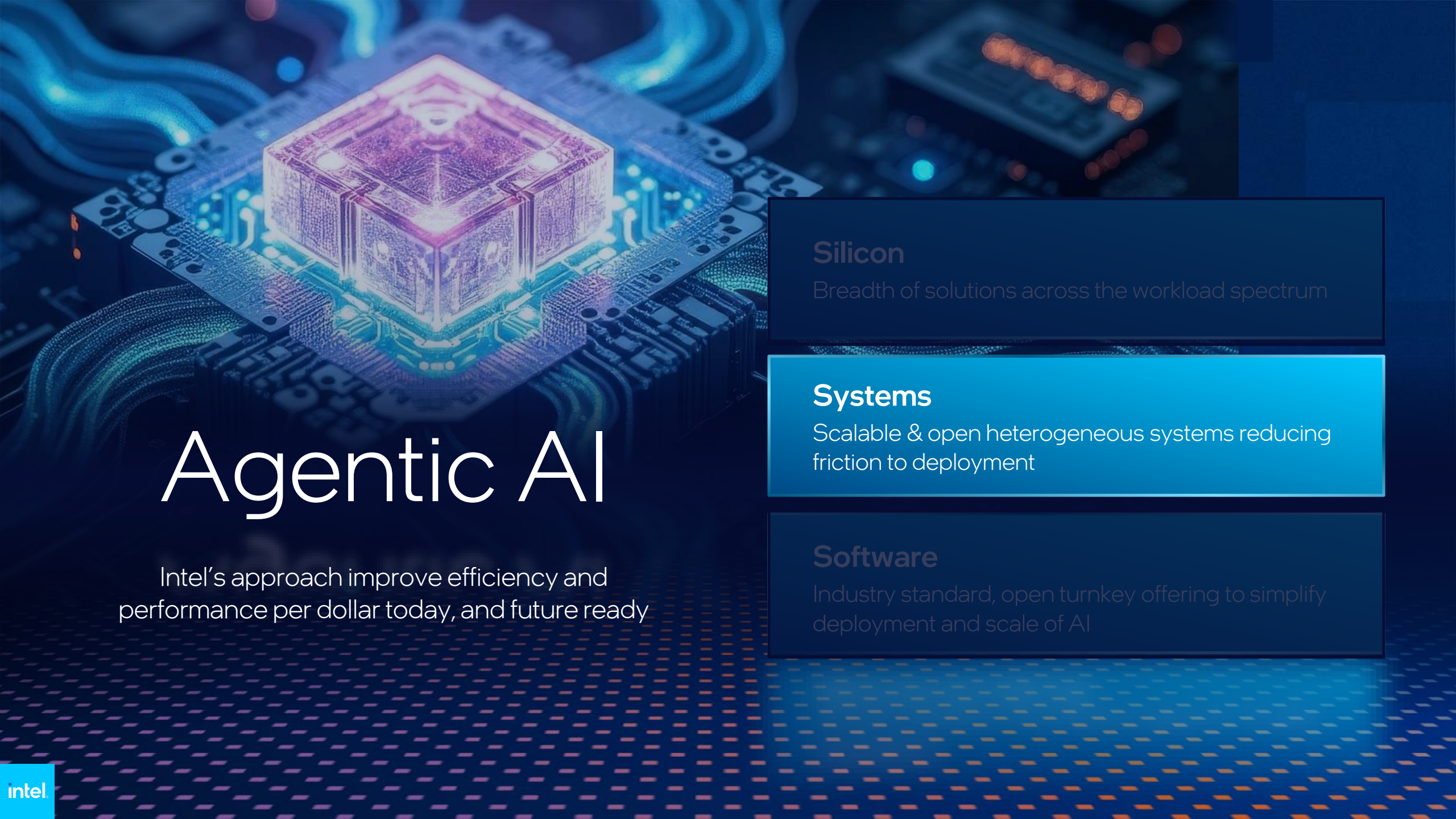
with enhanced memory, bandwidth and enterprise-level inference



Expand Technology & infrastructure

Next-gen inference-optimized GPU & Shores product line

tailored for agentic AI and training workloads



Agentic AI

Intel's approach improve efficiency and performance per dollar today, and future ready

Silicon

Breadth of solutions across the workload spectrum

Systems

Scalable & open heterogeneous systems reducing friction to deployment

Software

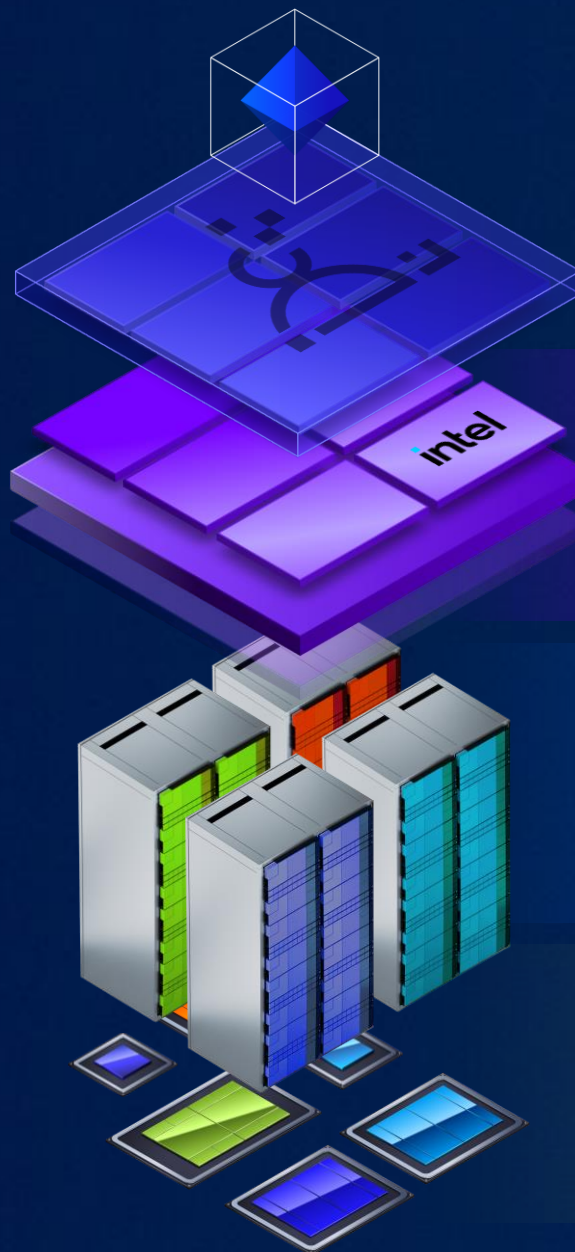
Industry standard, open turnkey offering to simplify deployment and scale of AI



Scalable Heterogenous Systems

Disrupting the current approach

Solving customer needs by
reducing friction to deploy
AI at scale



Open AI software stack

Programmability

Orchestration

Multi-framework/
model/API

System infrastructure

Enterprise-friendly
server form factors

Open ethernet-based
scale-out networking

Heterogenous hardware

Multi-architecture

Multi-vendor

Rack Scale System

Inference & Fine Tuning

Open, Modular &
Scalable Architecture

High Bandwidth
Scale-up Networking



76.8 TB/s
HBD Bandwidth*



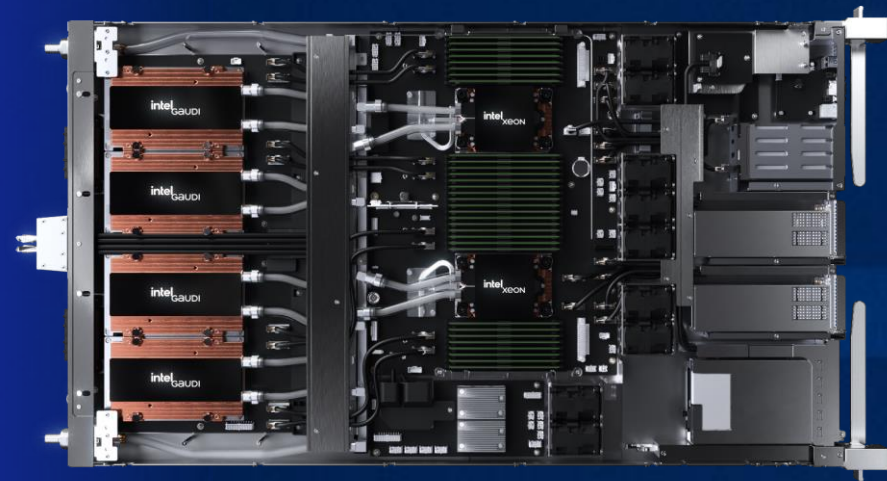
64
Gaudi 3 OAMs



8.2 TB
HBM2e



120 kw
Rack power



* 76.8 TB/S bidirectional scale-up bandwidth.

Open System Development

Foundation for Heterogeneous AI Infrastructure

Open System Design



Open Rack Design
(OCP v2+)



Open Bus Bar
PSU



Open Scale-up
connectivity



Open Scale-out
Connectivity

Open Compute Standards



Chiplet interface



Memory interface



I/O standards

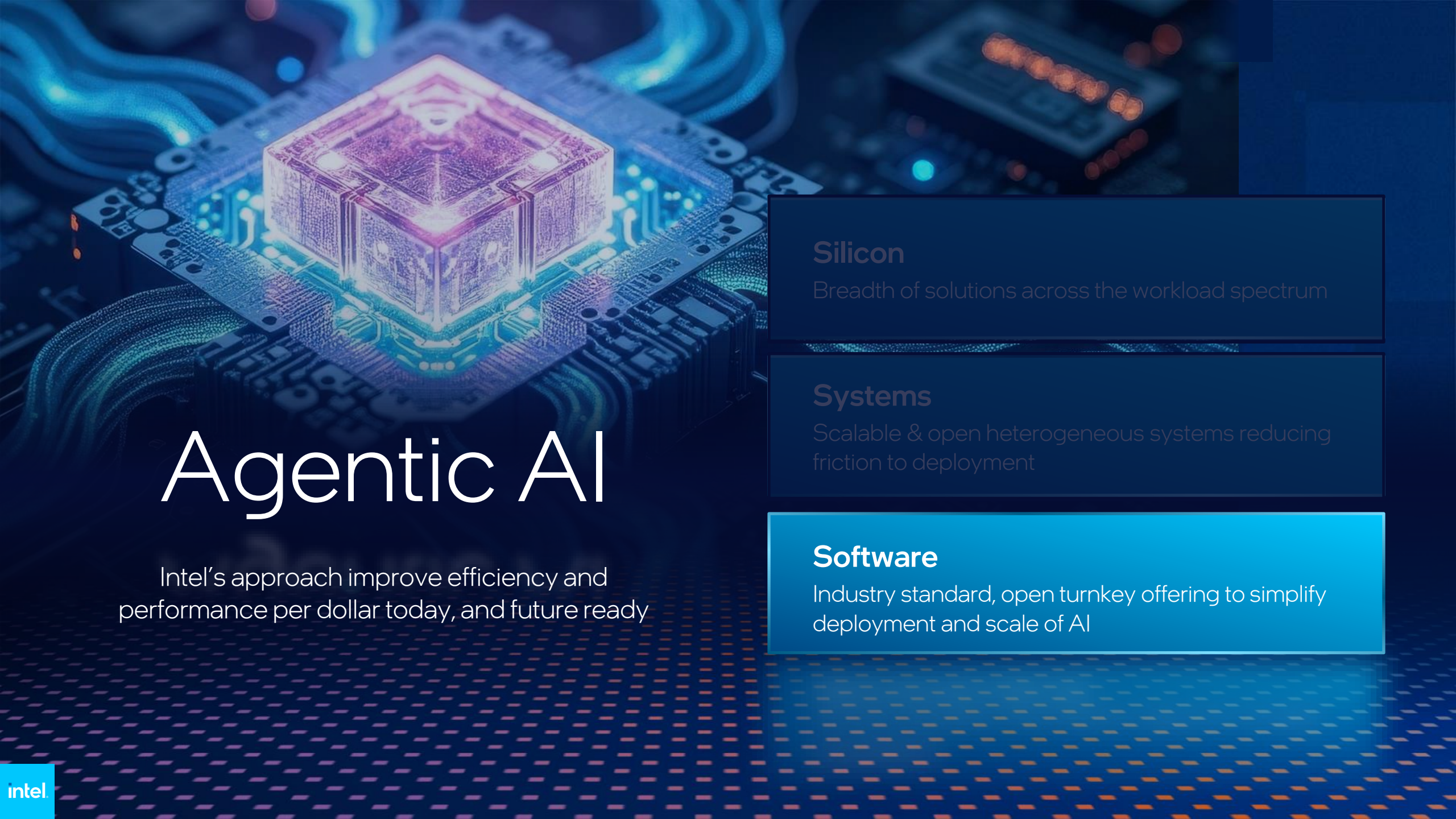


Storage standards



Mike Hendricks

Associate Vice President, Solutions and Ecosystem



Agentic AI

Intel's approach improve efficiency and performance per dollar today, and future ready

Silicon

Breadth of solutions across the workload spectrum

Systems

Scalable & open heterogeneous systems reducing friction to deployment

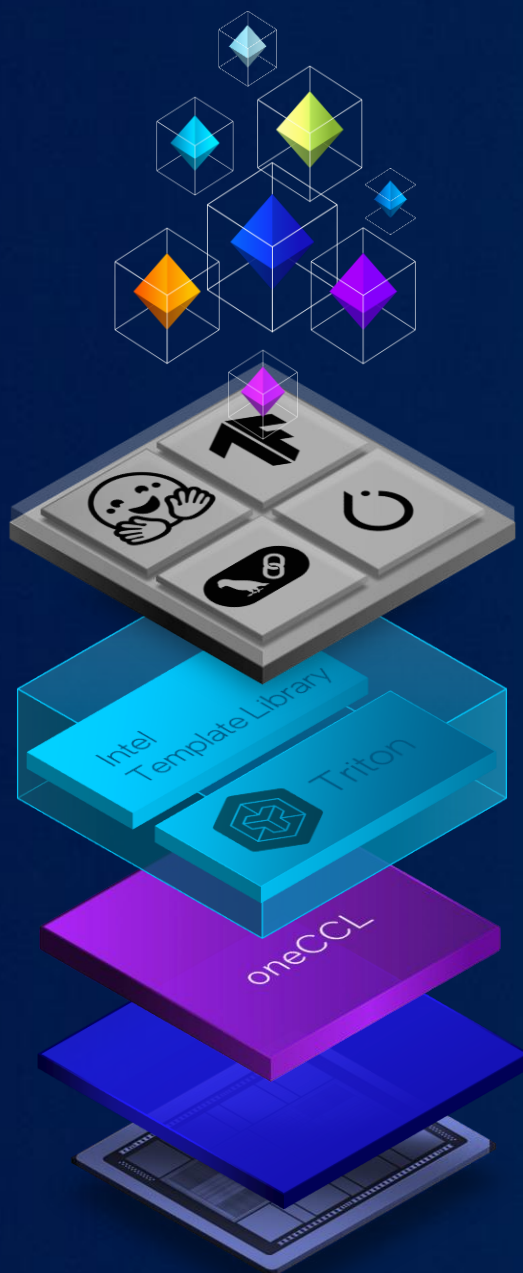
Software

Industry standard, open turnkey offering to simplify deployment and scale of AI

intel

Open AI Software Stack

Industry-standard frameworks
integration for a frictionless
developer experience



Applications

AI agents, enterprise apps

Frameworks

Day 0 functionality and performance
Direct upstream

Compute

Day 0 OpenAI Triton enablement
Intel Template Library (CUTLASS like)

Communication

OneCCL for parallel computing across multiple GPUs
Familiar framework abstractions (NCCL like)

Hardware & Firmware

Fully Validated firmware, drivers and reference systems
Performance analysis toolchain (Vtune based)

OEM Ecosystem

ASRock
— Rack —

ASUS

DELL
Technologies

Fsas
Technologies

GIGABYTE™

Inventec

QCT

SUPERMICR

The Intel logo is centered on a dark blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small, bright blue square is positioned above the dot of the letter "i".

intel