



Put Every Prompt in Its Place

Nicolas Dubé

CVP, DATA CENTER SYSTEMS AND SOLUTIONS, INTEL



Open Communities

Driving the open foundation for scalable AI



Open
hardware



Open
software



Open
interfaces



Open
infrastructure



Open Communities

Driving the open foundation for scalable AI



Open
hardware



Open
software



Open
interfaces



Open
infrastructure

116 Contributions
made to OCP

Intel has been an active contributor and member since its founding in 2011.

33 Sessions at this
Global Summit

Across **13 tracks**

Today
Built for large training clusters

Tomorrow
Built for inference everywhere



Right-Sized Intelligence

Reducing cost

Advancing sustainability

Without giving up quality



Inference cost

One-size routing

~60%

lower cost

Right-sized routing

<1% performance drop



Per-query energy

>80%

lower energy

<https://www.callosum.com/blog/programmable-heterogeneity>

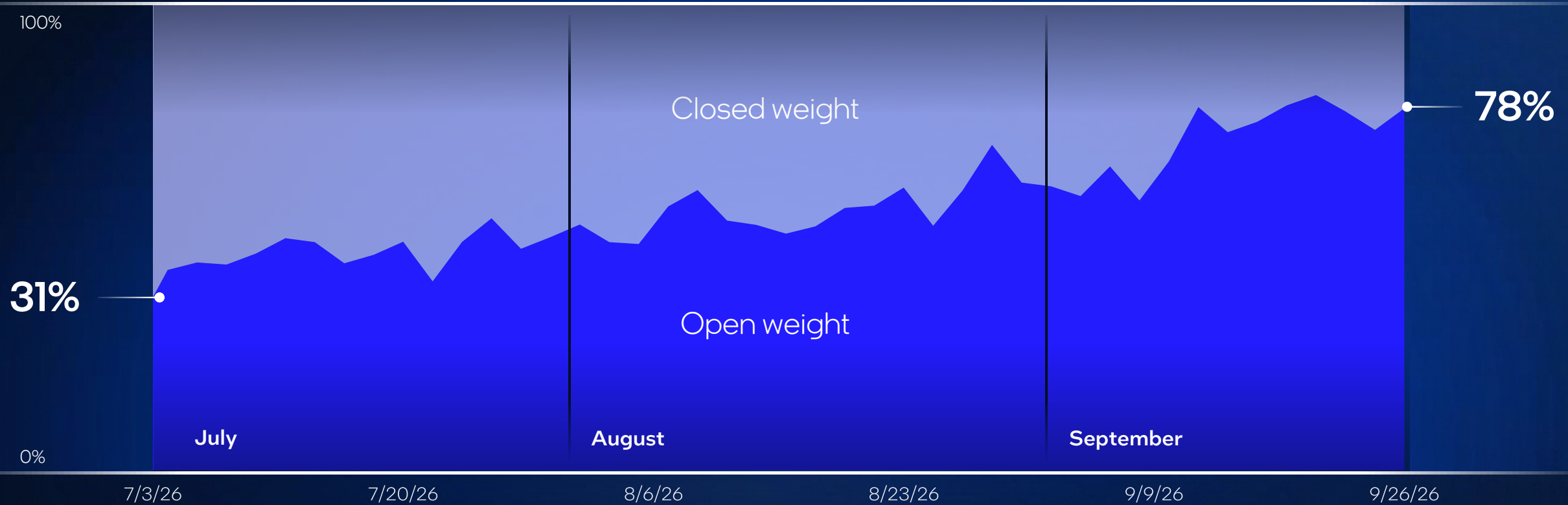
[BEST-Route: Adaptive LLM Routing with Test-Time Optimal Compute \(The University of British Columbia\)](#)

[Right-Sizing the Model to the Question: Cutting AI Chatbot Energy Without Losing Quality \(Frits AI\)](#)

[Energy use of AI inference, efficiency pathways, and test-time scaling \(Microsoft\)](#)

Open Weight Models Growing in Use

% of Tokens Open vs. Closed Model



[Open vs. Closed Token Volume \(Vercel\)](#)

Today
Built for large training clusters

Tomorrow
Built for inference everywhere

Expanding & diversifying workloads

Monolithic
intelligence

Uniform
compute

Large scale
data centers

Right-sized
intelligence

Efficient
compute

Distributed
inference

Compute
For Inference

INTEL® GPU D-SERIES
CODENAMED

Crescent Island

Compute designed for inference

BUILT FOR AGENTIC AI

1311/2622

TFLOPs
FP8/MxFP4

32

Xe Cores
Xe 3P GPU IP

LARGE CONTEXT CAPACITY

1.5TB/s

Memory bandwidth

Up to

480GB*

LPDDR5x

LOW-POWER FAMILIAR FORM

350W

Air-cooled PCIe

OPEN SOFTWARE STACK

Day 0

Functionality

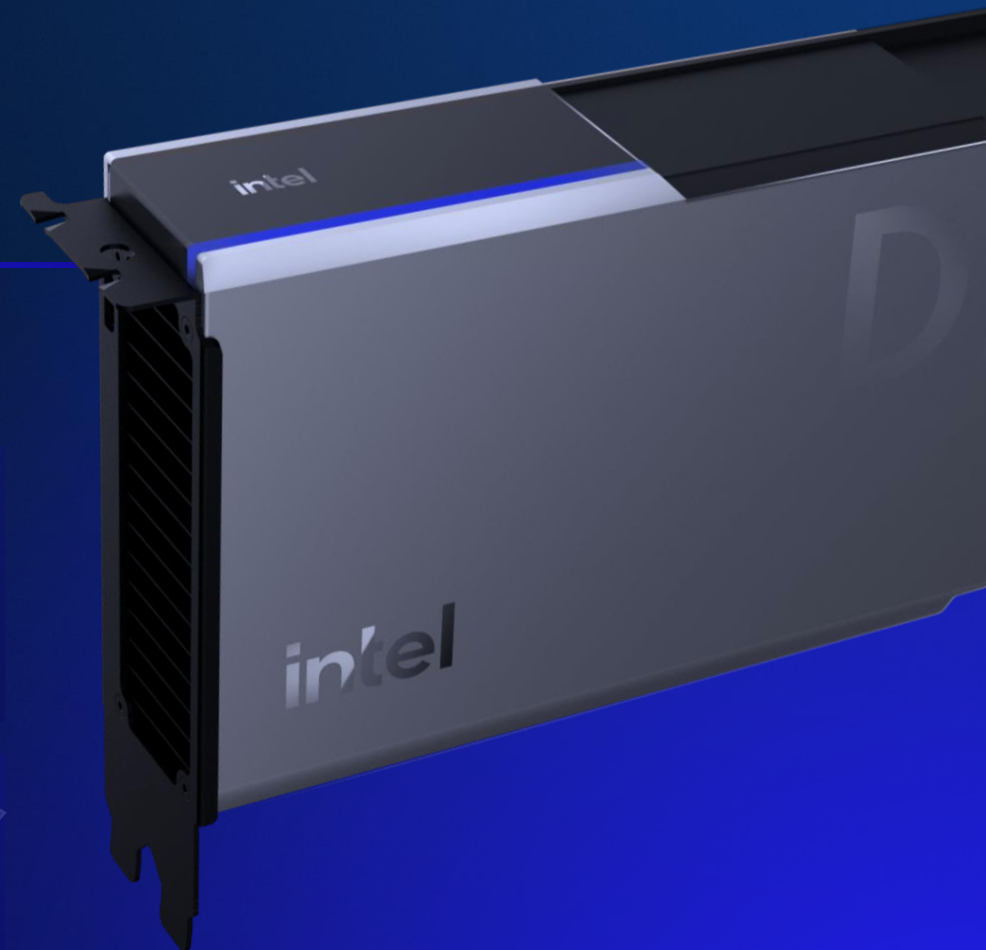
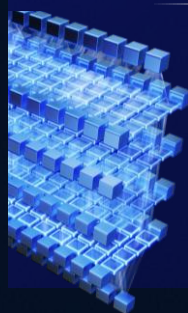
256

XMV engines

16

deep systolic

Xe Matrix Extensions (XMX)



*Intel branded PCIe card has 160GB LPDDR5x; design enables partners to build ODM branded cards with flexible options up to 480GB memory.

INTEL® GPU D-SERIES CODENAMED

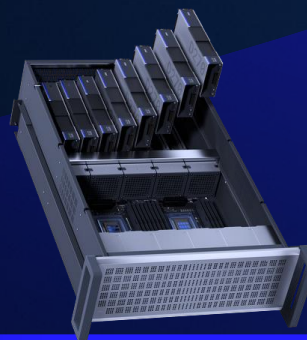
Crescent Island Scalable Performance

Across the full AI model continuum

x1



x8



x80



Throughput

AI Model Continuum

GPT-oss
20B

Kolibri-1
78B

Llama 4
Scout

Qwen3
235B

Deep
Seek
V4-Flash

GLM-4.5
-

Llama
Maverick

Trinity
400B

Deep
Seek
V3-R1

GLM 5.3
753B

Kimi K2
K2, K2.5 & K2.6

Mistral
Le Chonk

Deep
Seek
V4 Pro

Grok 3
2.7T

Kimi K3

Fast Agents

Multimodal Doc Q&A

Agentic Coding

Knowledge Work

Agentic Workflows

Frontier Reasoning

Routing

High Vol. Chat

Multimodal Reasoning

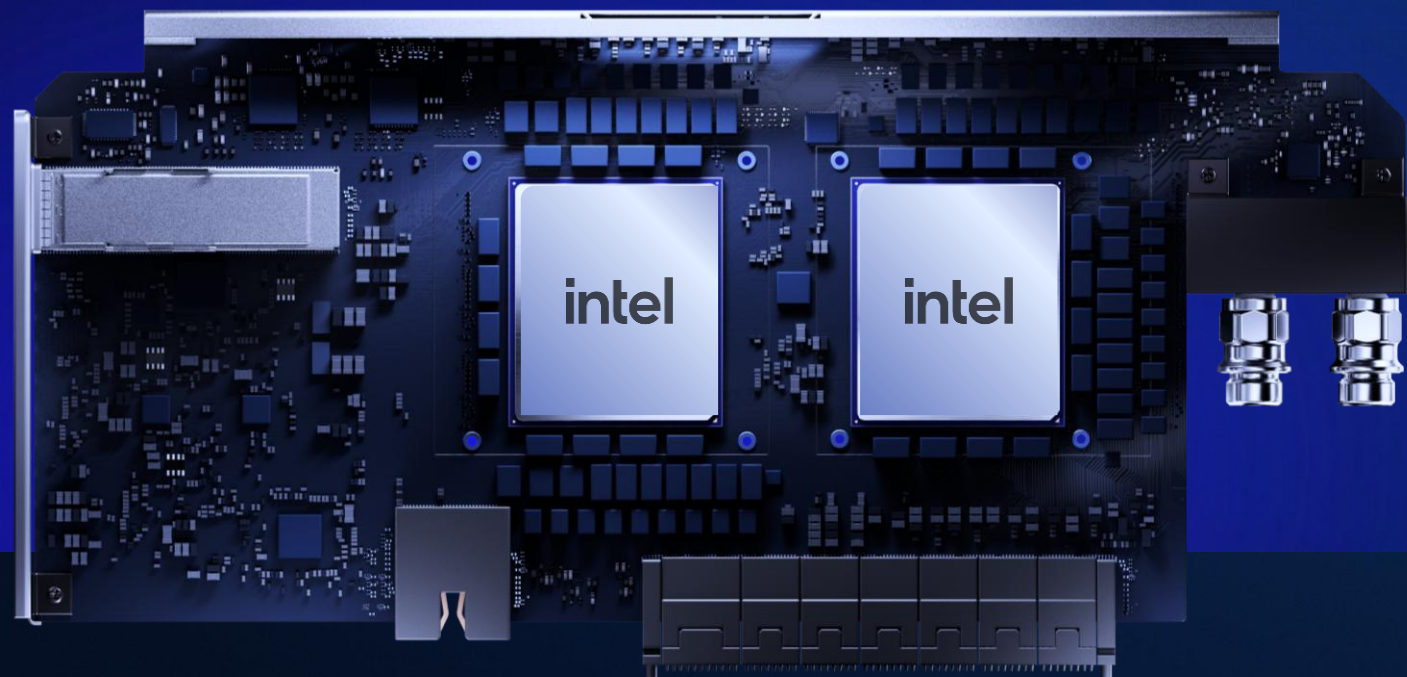
Agent Swarms

Long Horizon Coding



The Open Adapter Card (OAC)

Driving Form Factor Innovation



INTEL® GPU D-SERIES

Growing Ecosystem of Customers and Partners

AI Natives | Appliances | ISV & Software Ecosystem | Neo Clouds | OxMs

Accton	Ample ^{AI}	Anyscale	Arcee ^{AI}	ASRock Rock	AsteraLabs.	ASUS	Canonical	Cirrascale
CISCO	CORNELIS	COMPAL	Mind ^{CONFIDENTIAL}	CORESPAN SYSTEMS	DataKrypto	DELL Technologies	FarmGPU	FOXCONN [®] HON HAI TECHNOLOGY GROUP
Fsas Technologies	GIGA COMPUTING	HPE		vLLM Inferact	inties A ZUCI COMPANY	Inventec	Iterate.ai	INTERNAL [®]
JUSPAY / xync	KAMI WAZA	lablup	LAZARUS	leidos	Lenovo	Ltx	Liquid	MITAC MITAC Computing
MULTIVERSE COMPUTING	na ce	NeoCognition	NetApp	NUTANIX	OpenHands	PEGATRON	QCT [®]	RadixArk
Red Hat	Sail	SUPERMICRO	tensormesh	think [™]	VC2 GIVING COMPUTE DIRECTION	vmware [®] by Broadcom	WEKA [®]	wistron [®]

Includes several OCP members



Today
Built for large training clusters

Tomorrow
Built for inference everywhere

Expanding & diversifying workloads

Monolithic
intelligence

Uniform
compute

Large scale
data centers

Right-sized
intelligence

Efficient
compute

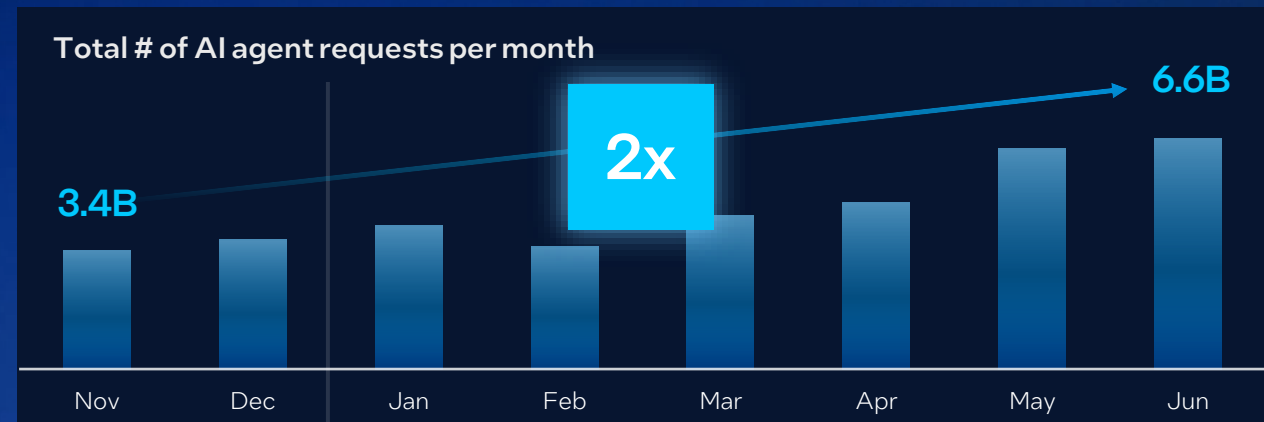
Distributed
inference

Inference
Everywhere

Agents Are on the Rise

Triple the number of agents per enterprise within a year

Doubling of total agent requests per month within half a year



[The AI Traffic Report Q2 2026: Agentic Traffic Surged 45%, With Meta Taking the Lead \(DataDome\)](#)

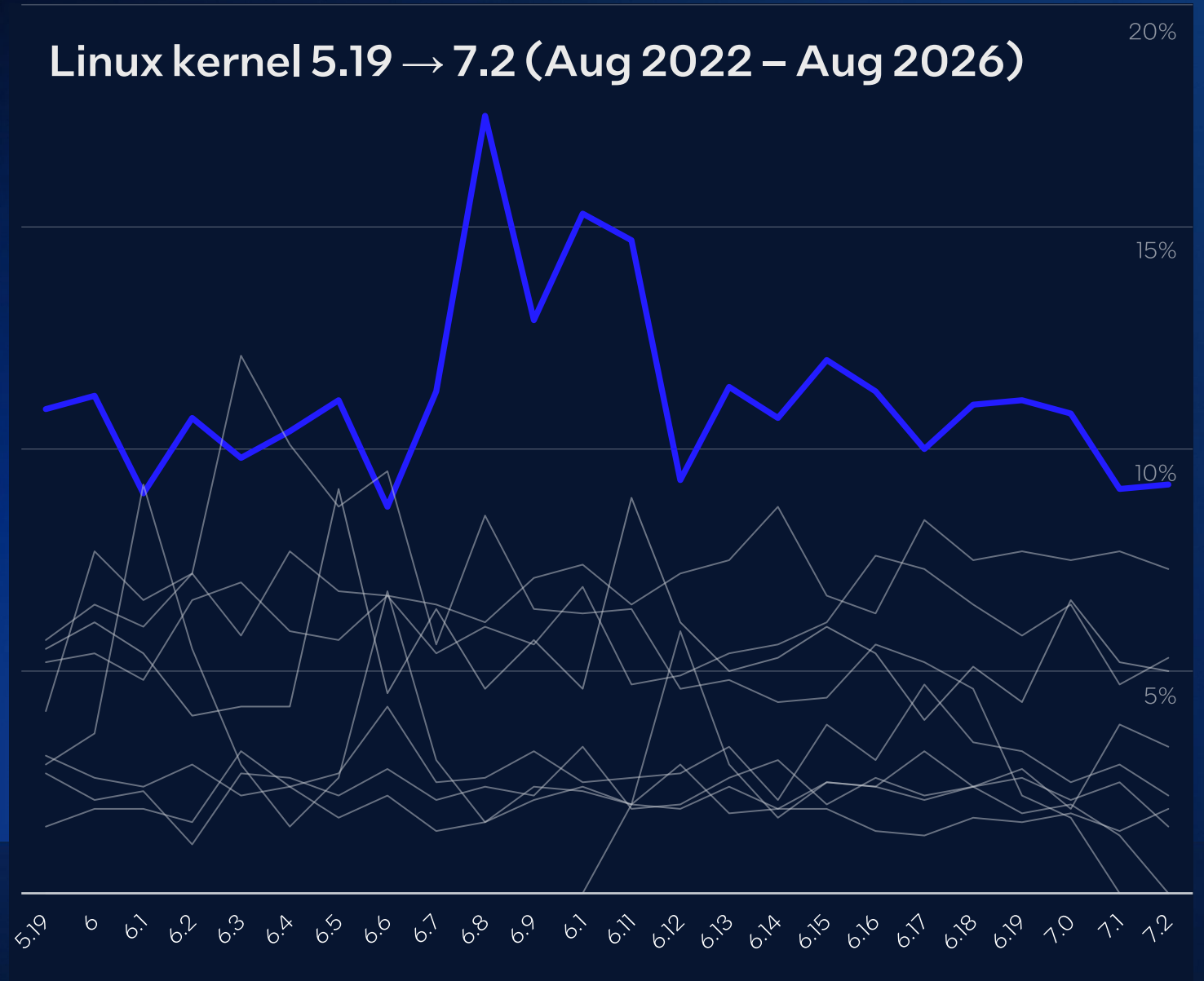
[Salesforce Agentic Enterprise Index Second Edition \(SalesForce\)](#)

Agents That Just Run

Open hardware through OCP

Open software through Linux

Intel a founding member of both

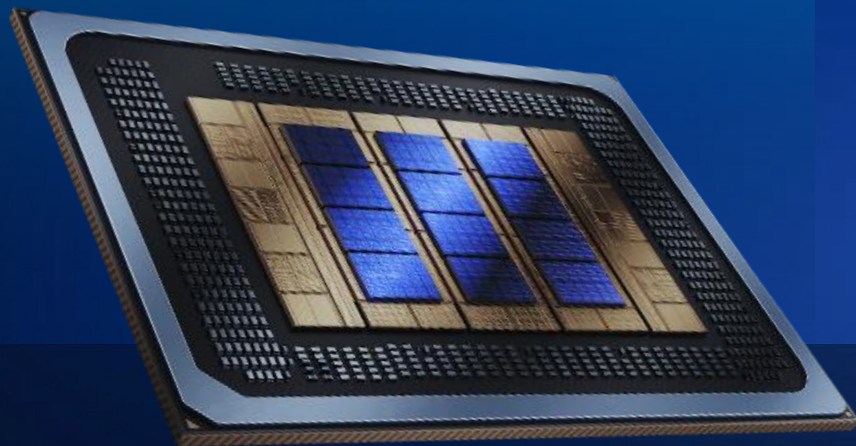


Intel® Xeon® CPUs for Agentic AI

Intel® Xeon® 6+

Dense cores for throughput & efficiency

Up to **288 E-cores**

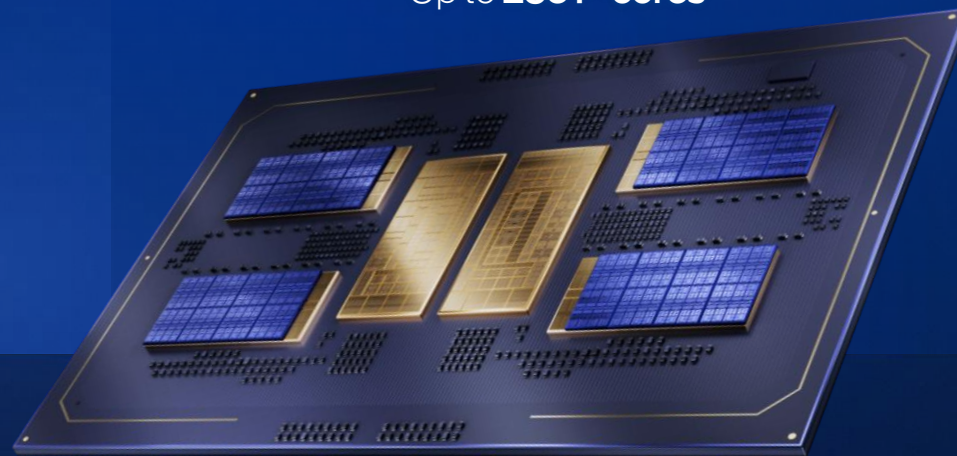


NEXT GEN INTEL® XEON® CODENAMED

Diamond Rapids

Best single-thread performance

Up to **256 P-cores**



Dense Intel® CPU Racks, a Proven Systems Heritage

2008
Sun Blade 6048

192
CPUs / rack

768
cores / rack

X6275
Xeon X5570



2014
HPE Apollo 8000

288
CPUs / rack

5,184
cores / rack

XL730f Gen9
Xeon E5-2699 v3

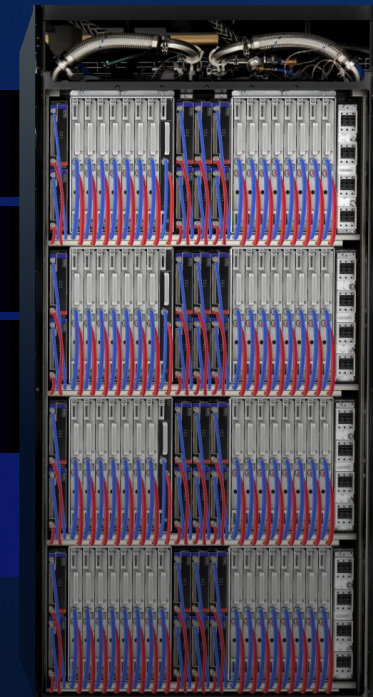


2023
HPE Cray EX

512
CPUs / rack

28,672
cores / rack

EX420 / EX4000
Xeon Max 9480



Intel® Xeon®: Agentic AI Rack Ecosystem

Faster, safer, more predictable, and ready for your workload

ASRock
Rock

ASUS

GIGABYTE™

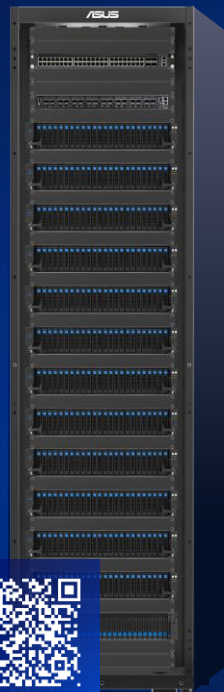
MITAC
MitAC Computing

msi

SUPERMICR



ORv3_{OCP}



ORv3_{OCP}



ORv3_{OCP}

Partners with solutions available today

The Right Tool for the Job

Orchestration Intel® Xeon® CPU



Classify workload

Route requests

Coordinate resources

Batch workload

Token Generation

Scale for throughput

Intel® GPU D-series

Fast token workload

Token Generation

Scale for low latency

Intel® GPU D-series (Prefill)
SambaNova RDU (Decode)



Agent loop

Execute Intel® Xeon® CPU

CPU-native operations

Assemble response

Return to application



One more thing...



Silicon Photonics for Next Gen System Architecture

Partnering with pioneers to accelerate solutions at ecosystem scale

CAPACITY & SCALABILITY

InP on 300mm Silicon

Delivering at High Volume



INTEGRATION & RELIABILITY

Integrated Lasers

Proven at scale, key enabler for NPO



UNIQUELY POSITIONED TO ENABLE

High Bandwidth Design

Necessary for the next generation of AI



Put Every Prompt in Its Place

Crescent Island

Intel® GPU D-Series



Open System Architecture

GPU, CPU and RDU based



Silicon Photonics

for AI at Scale





Building the Open for Enterprise AI

On display at 2026 OCP Global Summit

Dense compute for
agentic AI inference

Agentic AI

High-density inference

Enterprise workloads

Open Accelerator
Card (OAC)

Open ecosystem

Composable

Scalable

AI Computing
Continuum

Cloud to edge

Open infrastructure

Intelligence anywhere

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.

See backup for configuration details. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks

Results have been estimated or simulated.

Intel technologies may require enabled hardware, software or service activation.

Your costs and results may vary.

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

All product plans and roadmaps are subject to change without notice.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

© 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.

