intel VISION venture on

Solving the world's complex computational challenges with Intel's data center GPU, codenamed Ponte Vecchio

AMIL011

Ogi Brkic , GM & Vice President

Radhika Rao, Sr. Director

Meeting the compute needs of the future

SOLVE THE WORLD MOST CHALLENGING PROBLEMS BY PROVIDING ZETTA-SCALE COMPUTING CAPABILITIES BEFORE THE END OF THE DECADE

HPC & Data Center GPU Segments

Super Compute

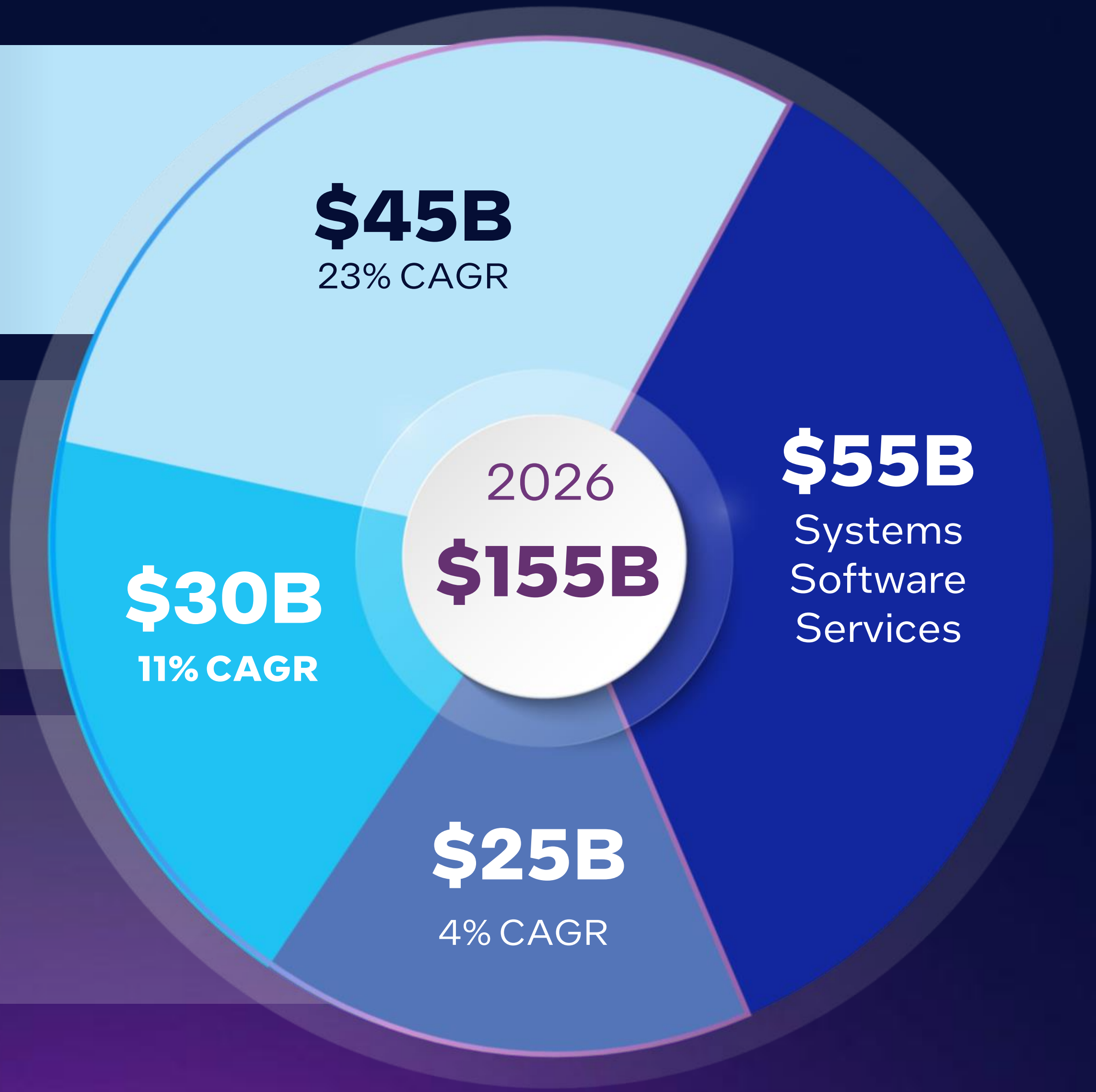
HPC - AI
Media & Visual Cloud

Custom Compute

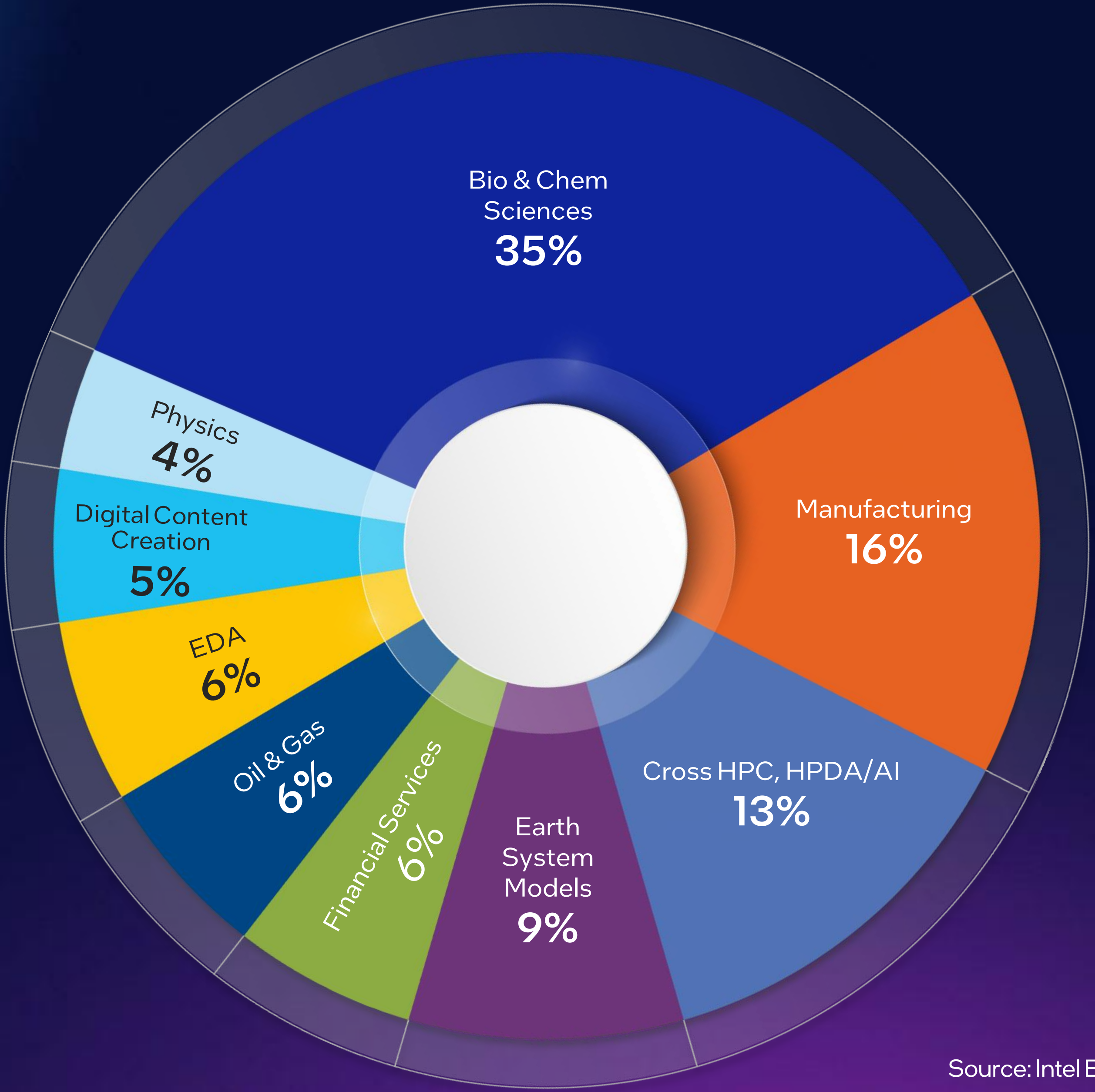
Blockchain
Supercomputing at the Edge

Visual Compute

Gaming
Content Creation
Metaverse



Top HPC Applications by Vertical

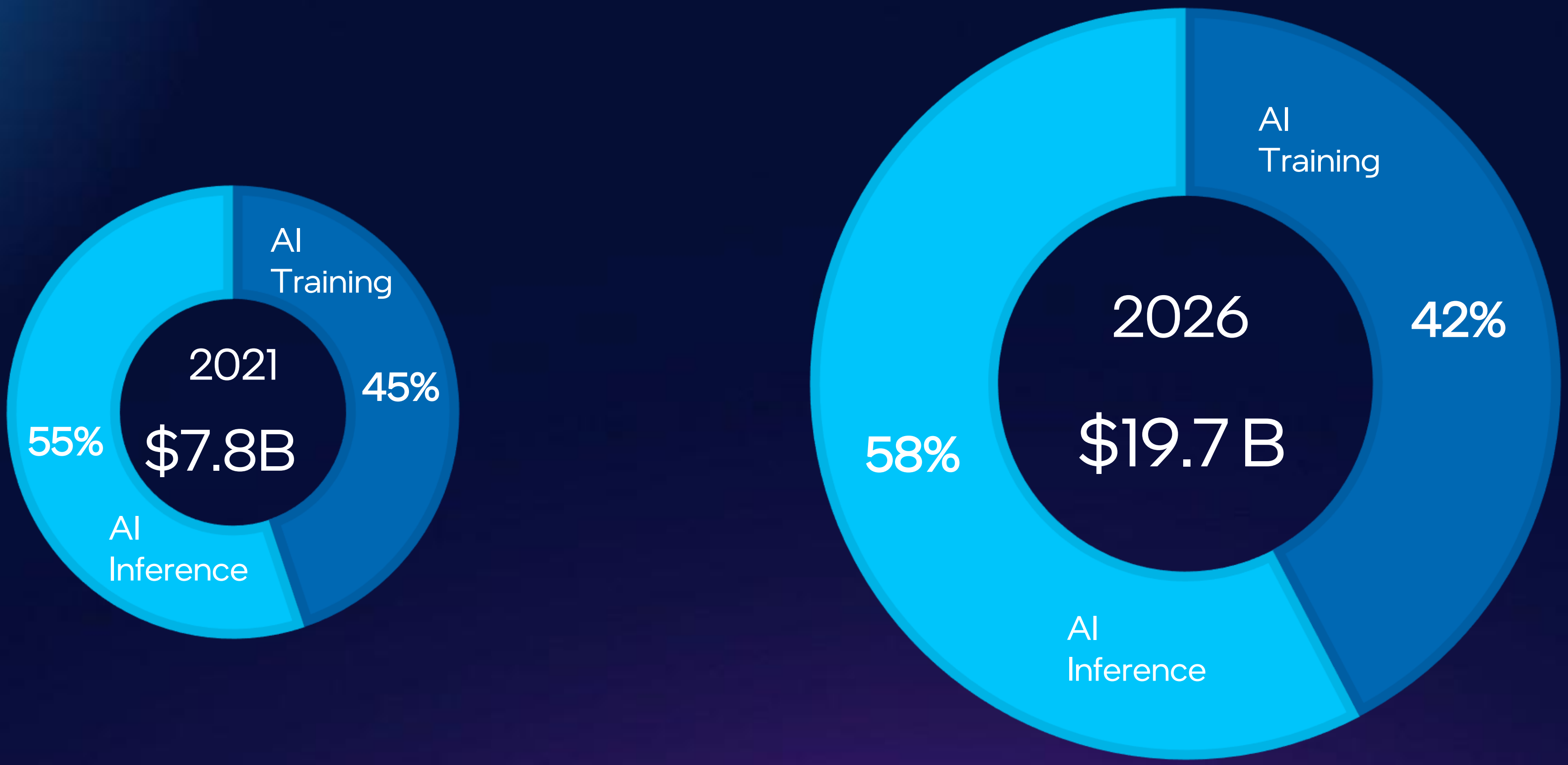


intel VISION

Source: Intel Estimates including Hyperion and other 3rd party data

AI Training and Inference Opportunity

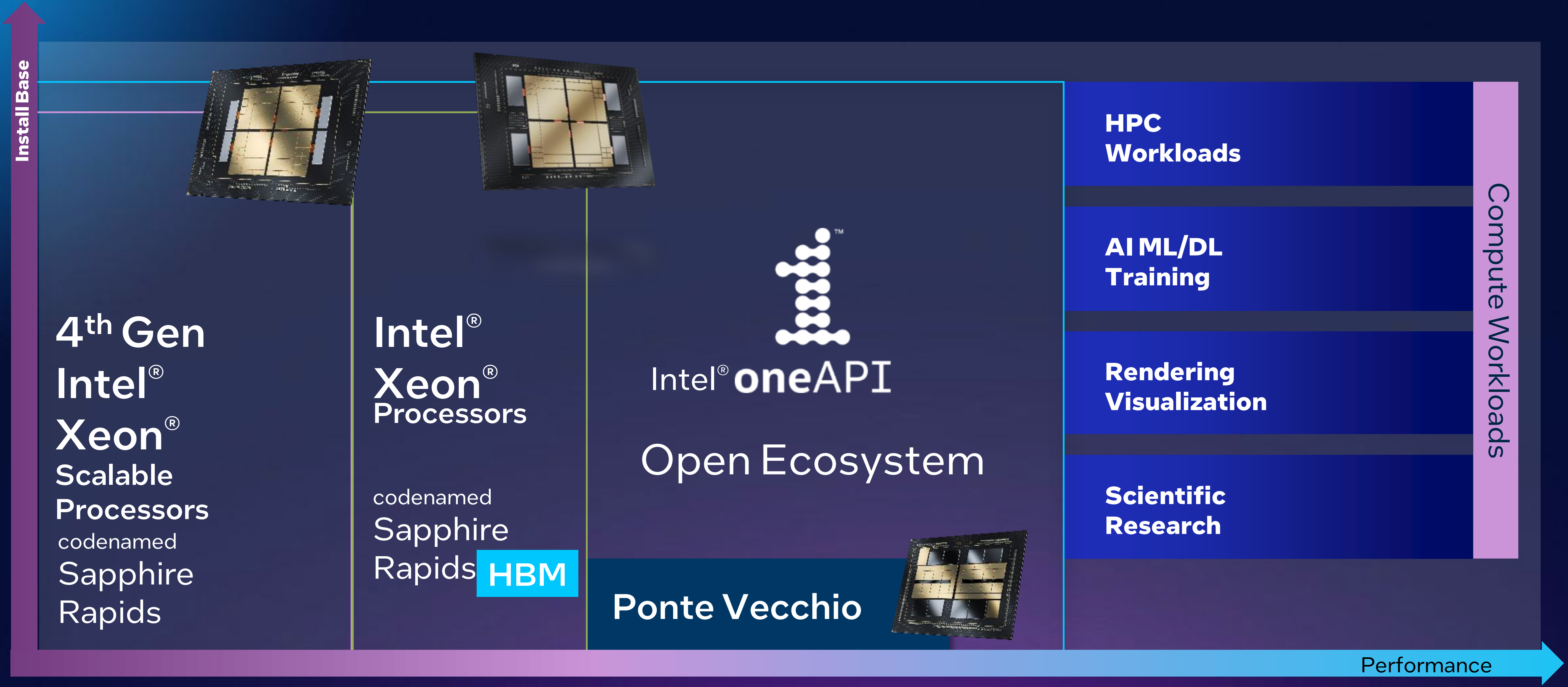
CPU and GPU



intel
VISIOn

Source: Intel Estimates

HPC - AI Super Compute Strategy



Install Base

4th Gen Intel® Xeon® Scalable Processors codenamed Sapphire Rapids

Intel® Xeon® Processors codenamed Sapphire Rapids HBM

Intel® oneAPI™

Open Ecosystem

Ponte Vecchio

HPC Workloads

AI ML/DL Training

Rendering Visualization

Scientific Research

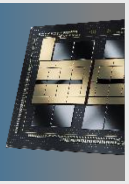
Compute Workloads

Performance

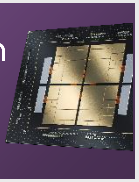
Annual Refresh Cycle

oneAPI

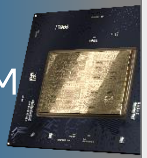
Ponte Vecchio



4th Gen Xeon HBM



Arctic Sound-M

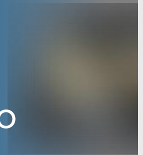


4th Gen Intel Xeon® processors

2022

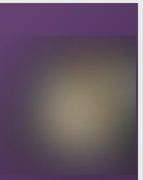
oneAPI

Ponte Vecchio Next

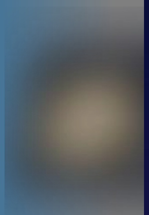


Xeon HBM

Next



Arctic Sound Next



Xeon Emerald Rapids

2023

oneAPI

Falcon Shores

XPU

>5x

Memory Capacity & B/W
Compute density in x86 socket
Performance/Watt

New Tile Based Flexible & Scalable Architecture

Scalable Architecture for all Super Computing Workloads

+

2024

https://edc.intel.com/content/www/us/en/products/performance/benchmarks/investor-day-2022/

Compute

Up to
128 Ray
Tracing Units

**Highest
Compute Density**
socket & node

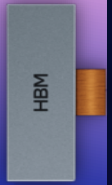
Up to
**128 X^e
Cores**

Memory

Up to
64MB
L1 cache

Up to
408MB
L2 Cache

Up to
**128GB
HBM2e**



I/O

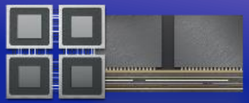
Up to **8
Fully Connected
GPUs**

PCIe
Gen 5

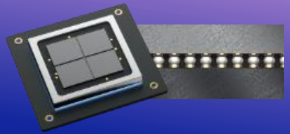
X^e Link
High-Speed
Coherent
Unified GPU Fabric

Xe-Link
Xe-Link
Xe-Link
Xe-Link
Xe-Link
Xe-Link
Xe

Technology



EMIB



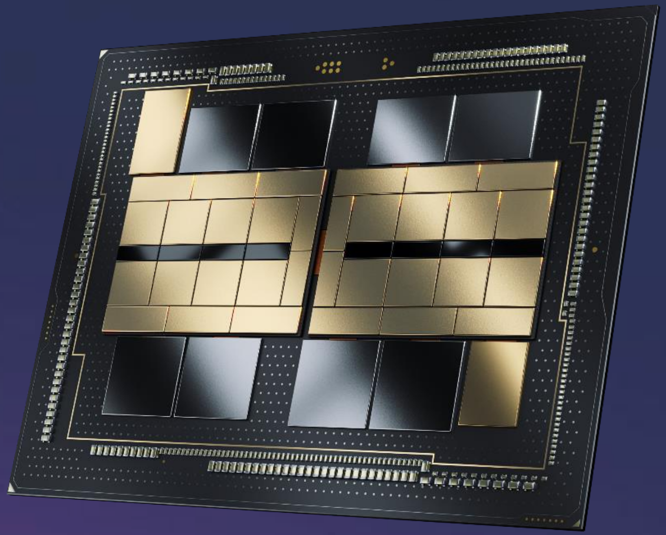
Foveros

**Intel 7
TSMC N5
TSMC N7**

Ponte Vecchio

X^e HPC based GPU

Up to 2.6x[1] Perf
over best in market today



On Track
for Aurora 2 Exaflop Supercomputer[2]

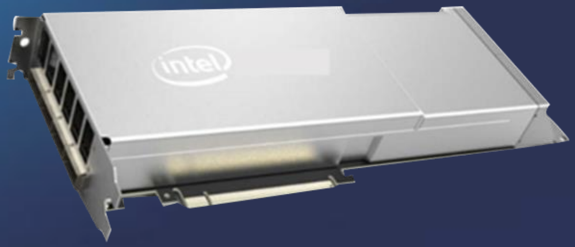
[1] Based on pre-production measurements vs A100.
Learn more at www.intel.com/PerformanceIndex. Results may vary.

[2] >2 exaflop peak performance

intel
VISION

Ponte Vecchio GPU Boards & Systems

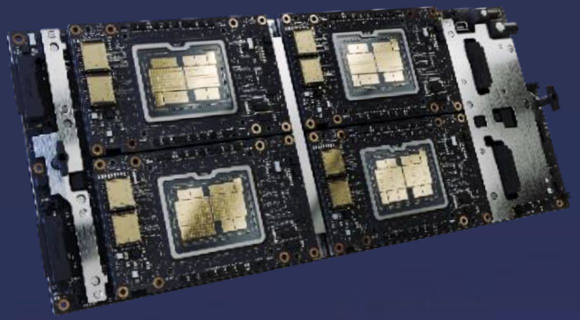
GPU BOARDS
Intel Branded



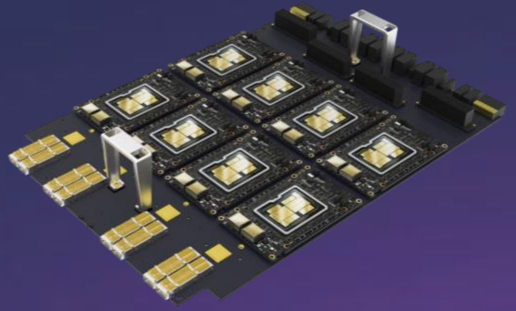
PCIe Add In Cards

OAM Modules

OAM SUBSYSTEMS

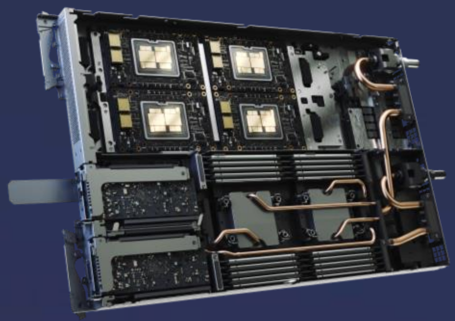


x4 GPU Subsystem



x8 GPU Subsystem

FULL SYSTEMS



1U 4 GPU Server

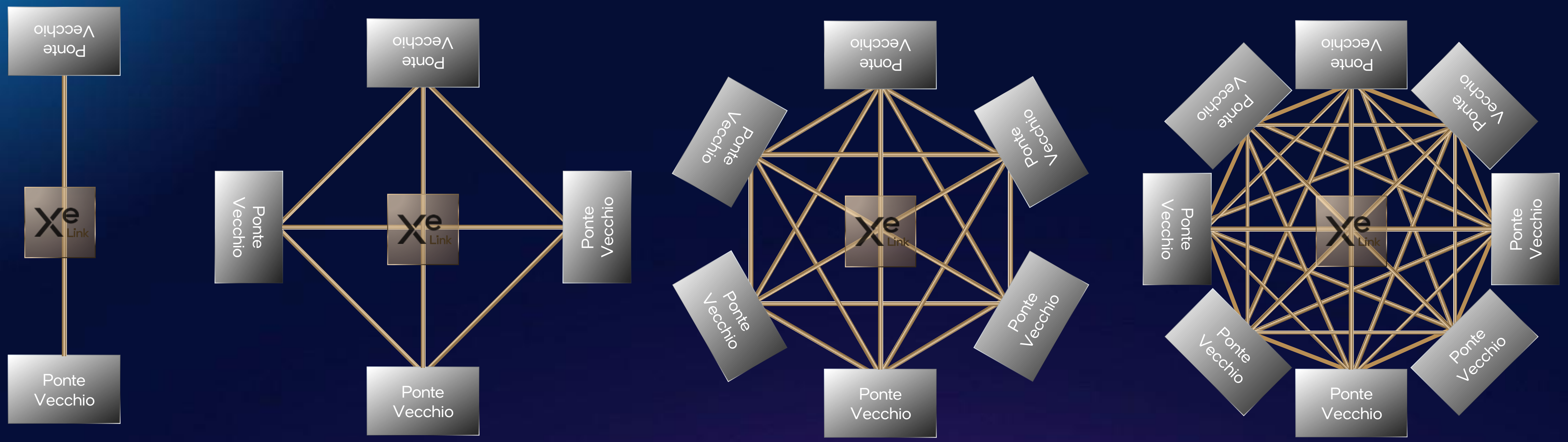
4U-7U GPU Server

Compute Accelerator Market Segmentation by GPU

	Segment	Number of GPUs
	Exascale	> 10K
	Exascale Follow On	Up to 10K
Google Cloud amazon web services Azure	Hyperscalers CSP	Up to 10K
	Large Enterprise and Next Wave CSPs	Up to 1K
SME	Enterprise AI and HPC	Up to 100

Xe Link for Scalability

Enabling a high number of coherent and unified accelerators



Flexibility for Scale Up and Scale Out across GPUs and Nodes

Note:
Xe Links connect gluelessly between PVCs
The Xe Link Logo above is not representing an additional System Device

intel. VISION

oneAPI

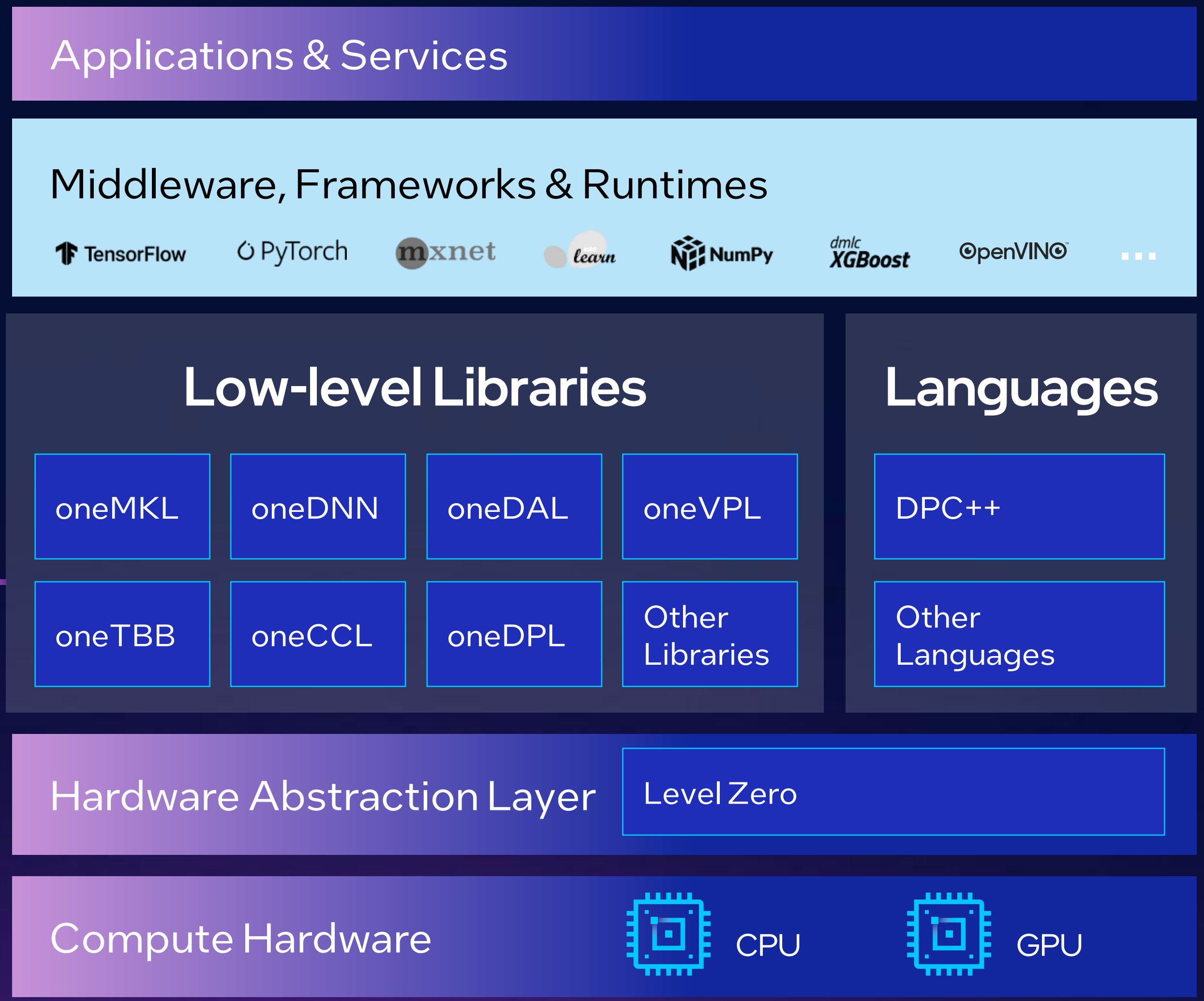
Open, Standards-Based Unified Software Stack

Freedom from proprietary programming models

Full performance from the hardware

Piece of mind for developers

CPU & XPU - Optimized Stack



Applications & Services

Middleware, Frameworks & Runtimes

TensorFlow | PyTorch | mxnet | learn | NumPy | dmlc XGBoost | OpenVINO | ...

Low-level Libraries

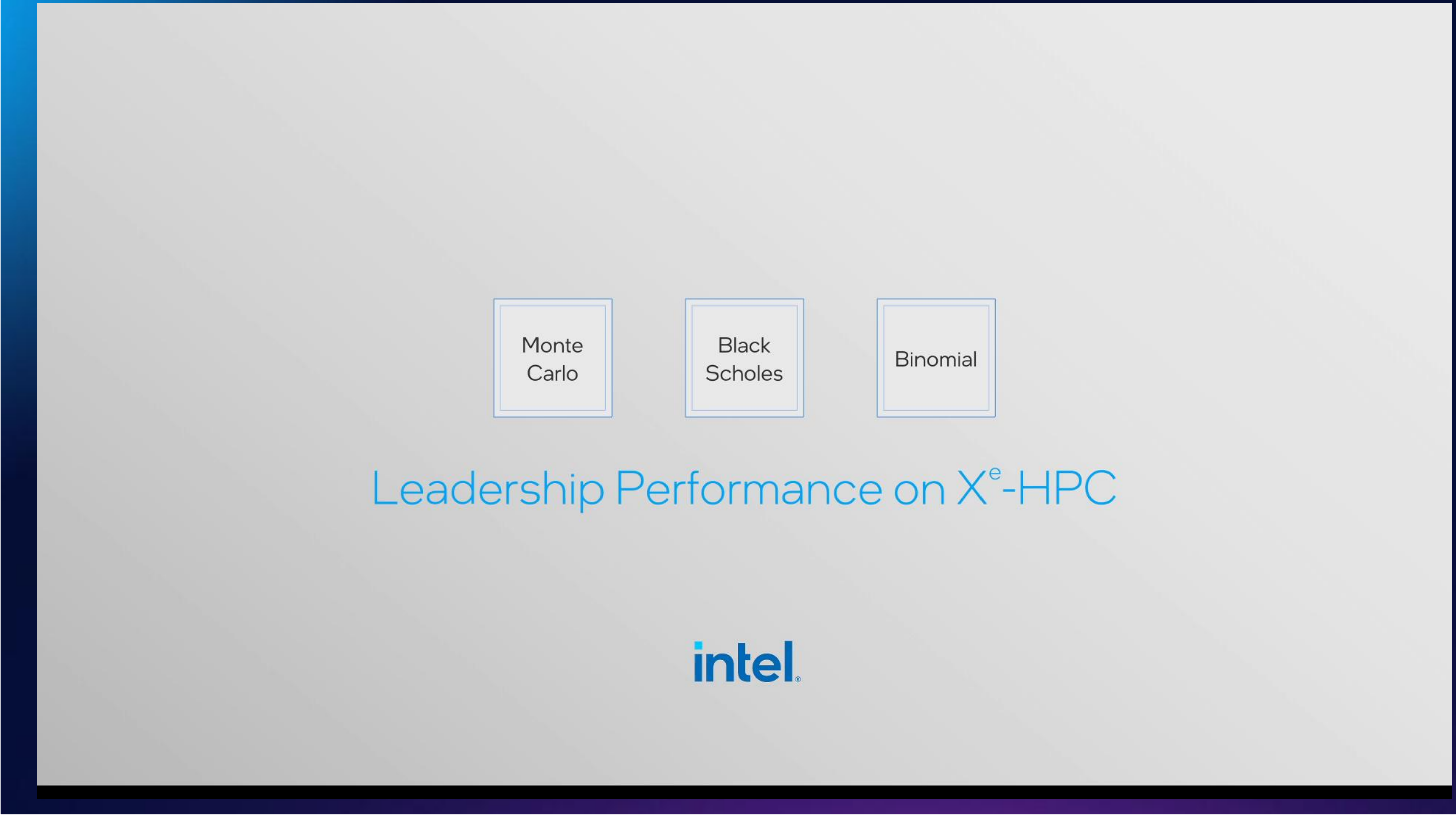
oneMKL	oneDNN	oneDAL	oneVPL
oneTBB	oneCCL	oneDPL	Other Libraries

Languages

- DPC++
- Other Languages

Hardware Abstraction Layer: Level Zero

Compute Hardware: CPU, GPU



Monte Carlo

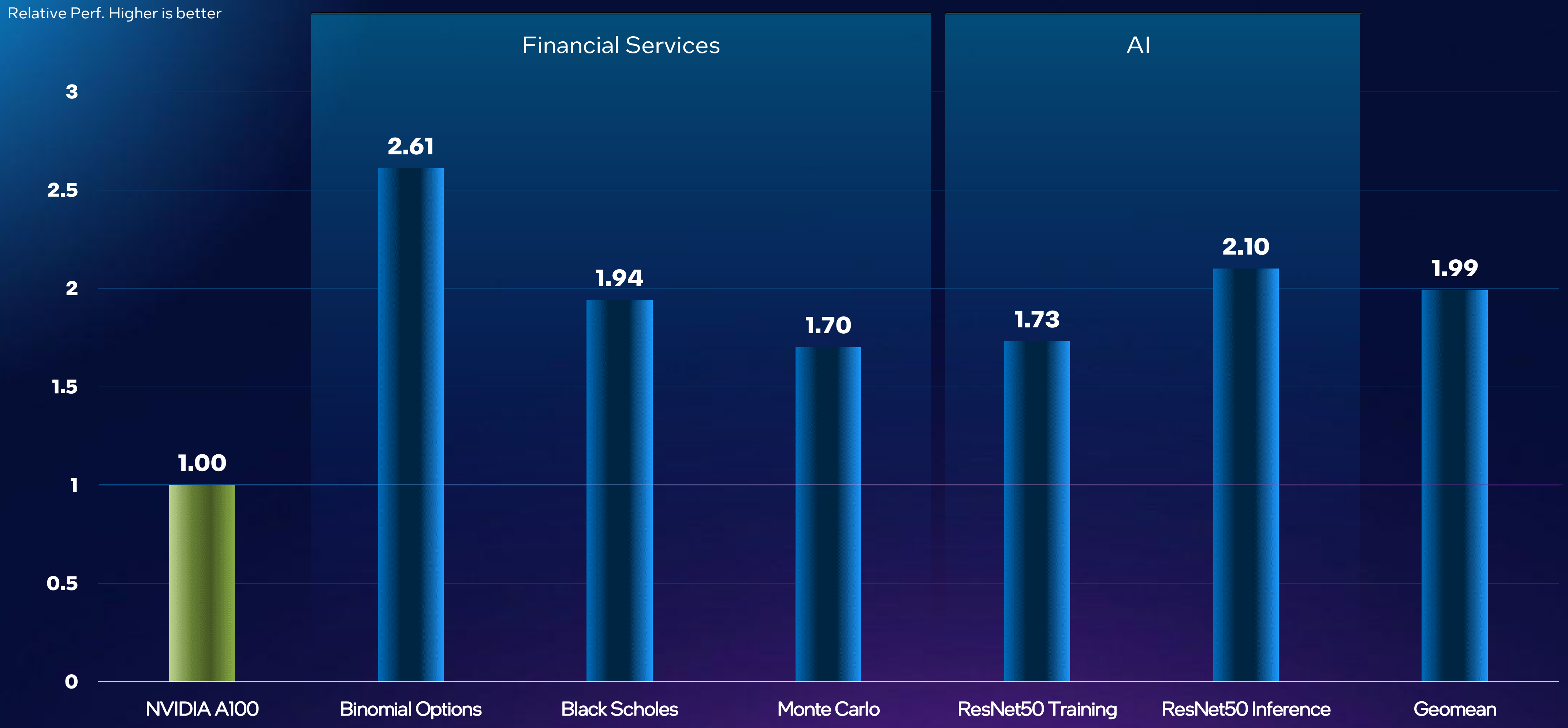
Black Scholes

Binomial

Leadership Performance on X^{e}-HPC

See the Vision section of intel.com/performanceindex for workloads and configurations. Results may vary.

Ponte Vecchio Performance



See the Vision section of intel.com/performanceindex for workloads and configurations. Results may vary.

intel VISION

Research Scientists

Students

SW Developers

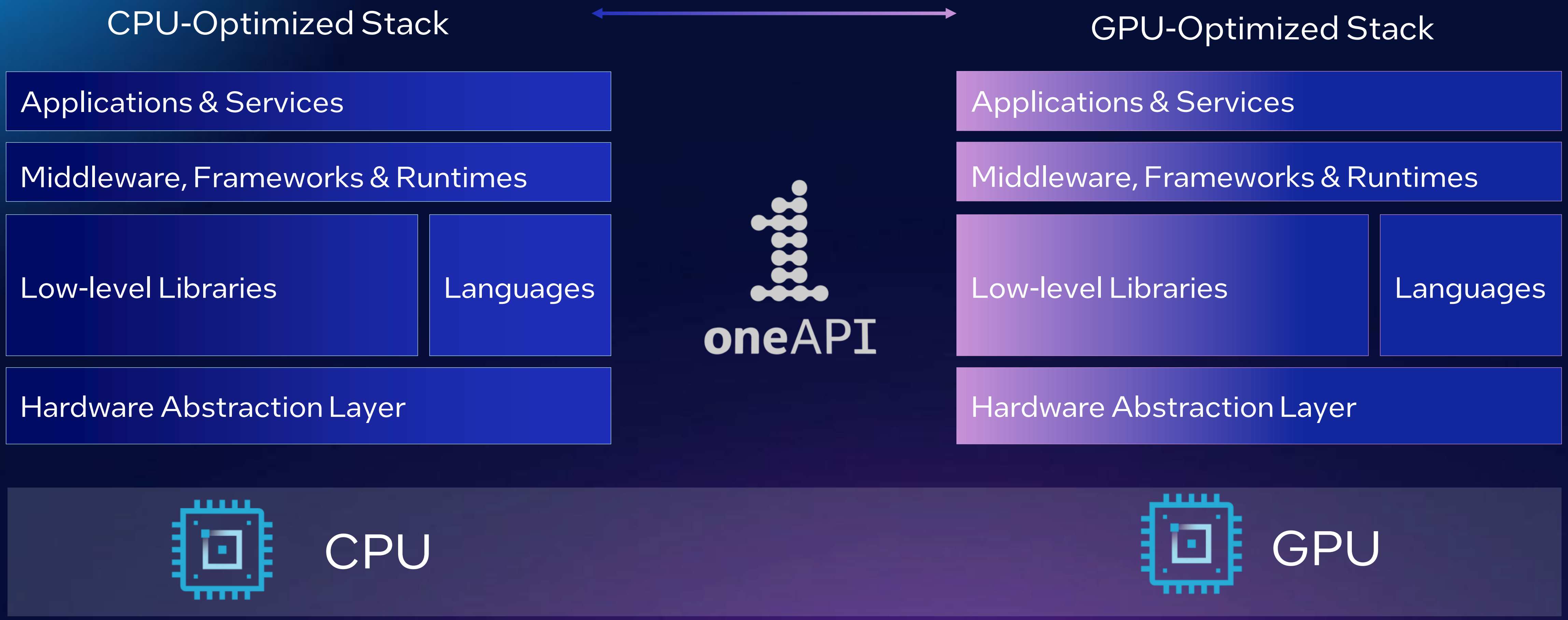
Data Engineers

Win Developers

for HPC/AI Everywhere

Agriculture	Energy	Education	Government
Finance	Telecom	Transport	Smart Home
Media	Health	Industrial	Retail
Academia	Cloud/DC	Edge	Client

Overcoming Separate CPU and GPU Software Stacks with oneAPI



CPU-Optimized Stack

Applications & Services

Middleware, Frameworks & Runtimes

Low-level Libraries

Languages

Hardware Abstraction Layer

oneAPI

GPU-Optimized Stack

Applications & Services

Middleware, Frameworks & Runtimes

Low-level Libraries

Languages

Hardware Abstraction Layer

CPU

GPU

intel®

VISION

venture **on**

Thank you