intel innovation

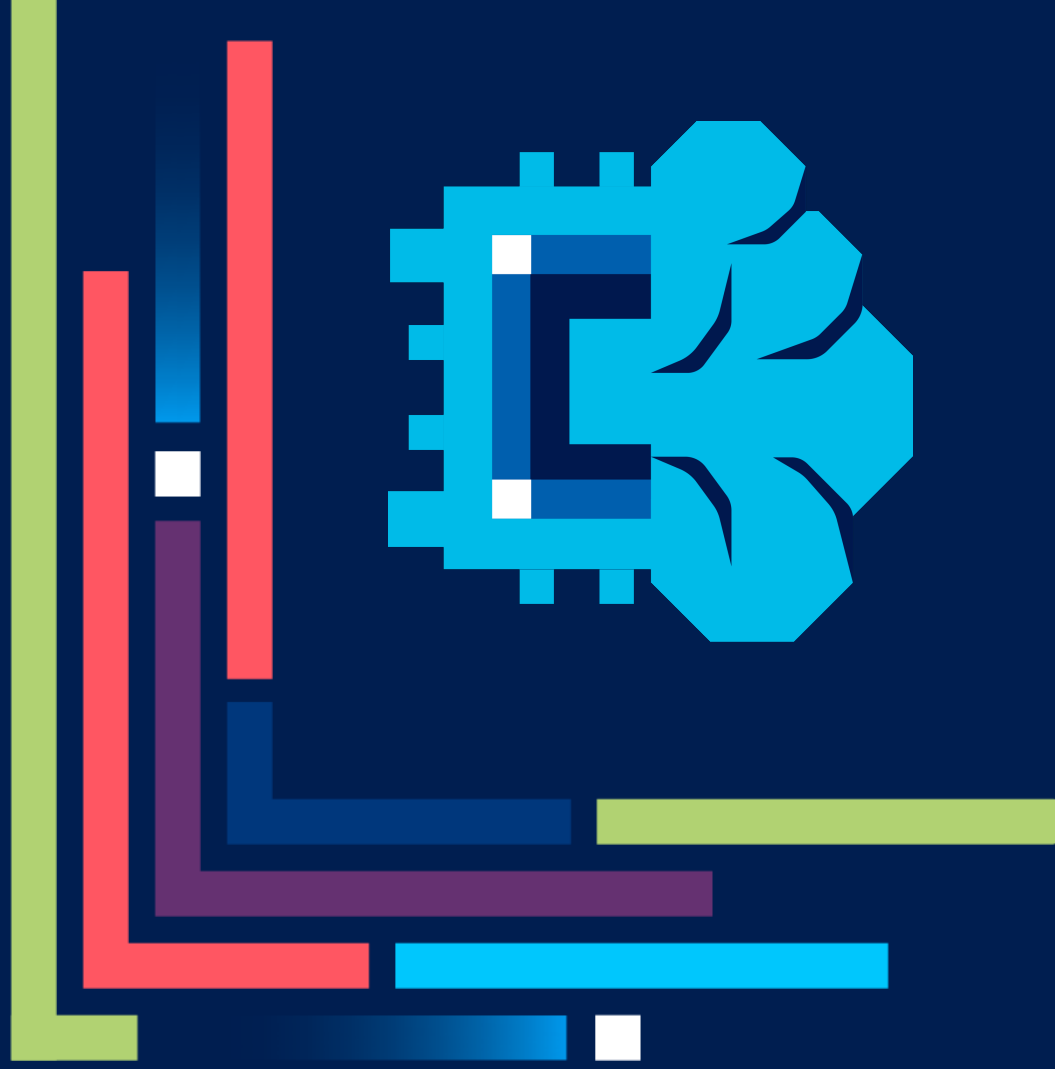
A New Paradigm for AI Productivity & Performance

Seetha . Nookala

Senior Director

Datacenter Graphics, HPC/AI ,APAC & Japan

December 2023



AI Exponentially Increasing In Quality & Size

intel innovation

Source: Amazon Web Services , 2022

Yet Inaccessible For Most – Cost of AI Today

Training Cost

GPT-3

$1.65M

(3,640 petaFLOPS-days) costs if trained on Google TPU v3

GPT-4

$40M

(450,000 petaFLOPS-days) , 7,600 GPUs running for a year

Inferencing Cost

ChatGPT

$40M

to process prompts per month with 100 million active users

Bing AI Chatbot

$4B

Bing AI Chatbot to serve responses to all Bing users

Source: CSET, CNBC

Driving AI Ubiquity Will Unlock New Levels of Innovation

Small AI Models

Smartphones & Mobile Devices	Smart Cameras & Surveillance	Industrial IoT	Autonomous Vehicles	Edge Servers & Gateways
Facial Recognition	Object Detection	Predictive Maintenance	Object Detection	Local Decision-Making
Voice Assistants	Anomaly Detection	Quality Control	Driver Monitoring	Data Filtering
Drones & UAVs	Smart Home Devices	Wearables	Robots	Retail and Customer Analysis
Obstacle Avoidance	Thermostats	Health Monitoring	Navigation	Foot Traffic Analysis
Aerial Data Processing	Home Security	Gesture Control	Object Manipulation	Personalized Marketing
Agriculture Sensors & Equipment	Healthcare	Environmental Monitoring	Smart Grids & Energy Management	Sports and Entertainment
Livestock Monitoring	Remote Patient Monitoring	Air Quality Sensing	Power Distribution	Player and Audience Analytics
Precision Farming	Medical Imaging	Earthquake Detection	Grid Security	AR Headsets

Large

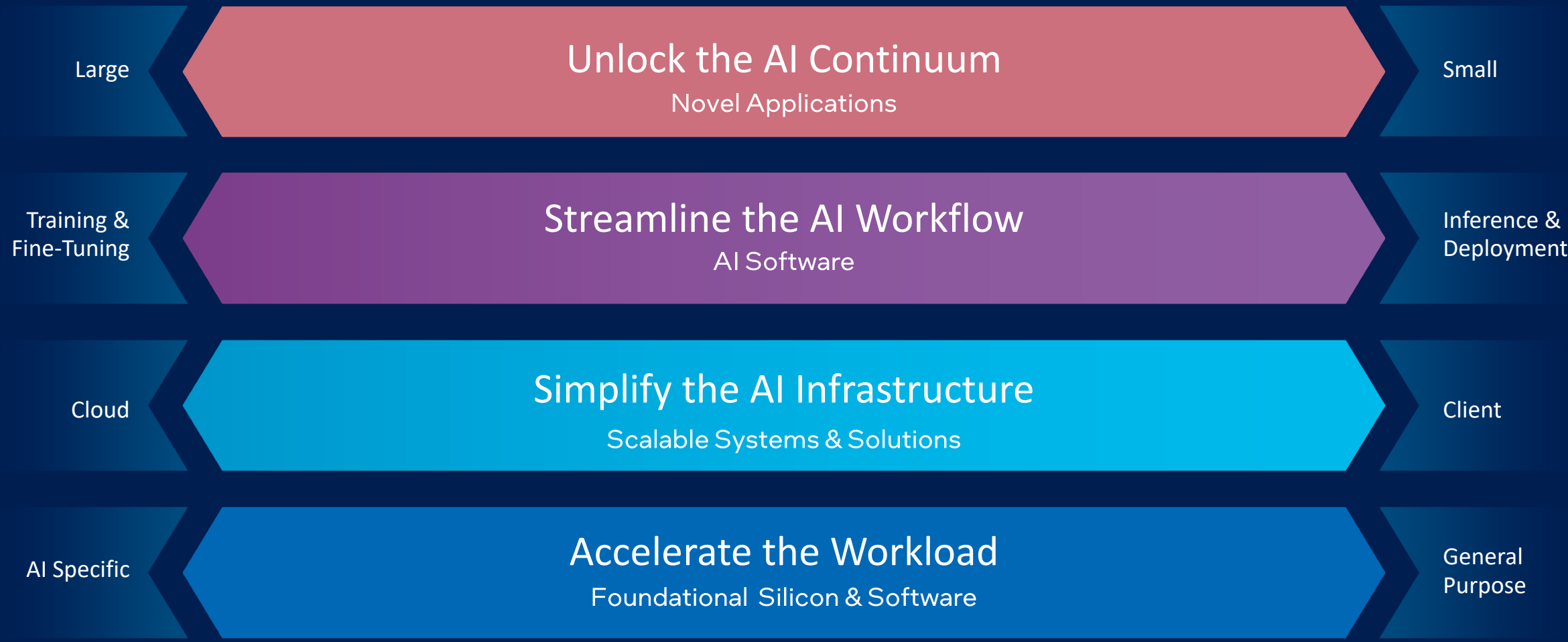
Unlocking the AI Continuum

From the Largest Data Center Scale AI, to the Smallest AI That Fits On Any Device

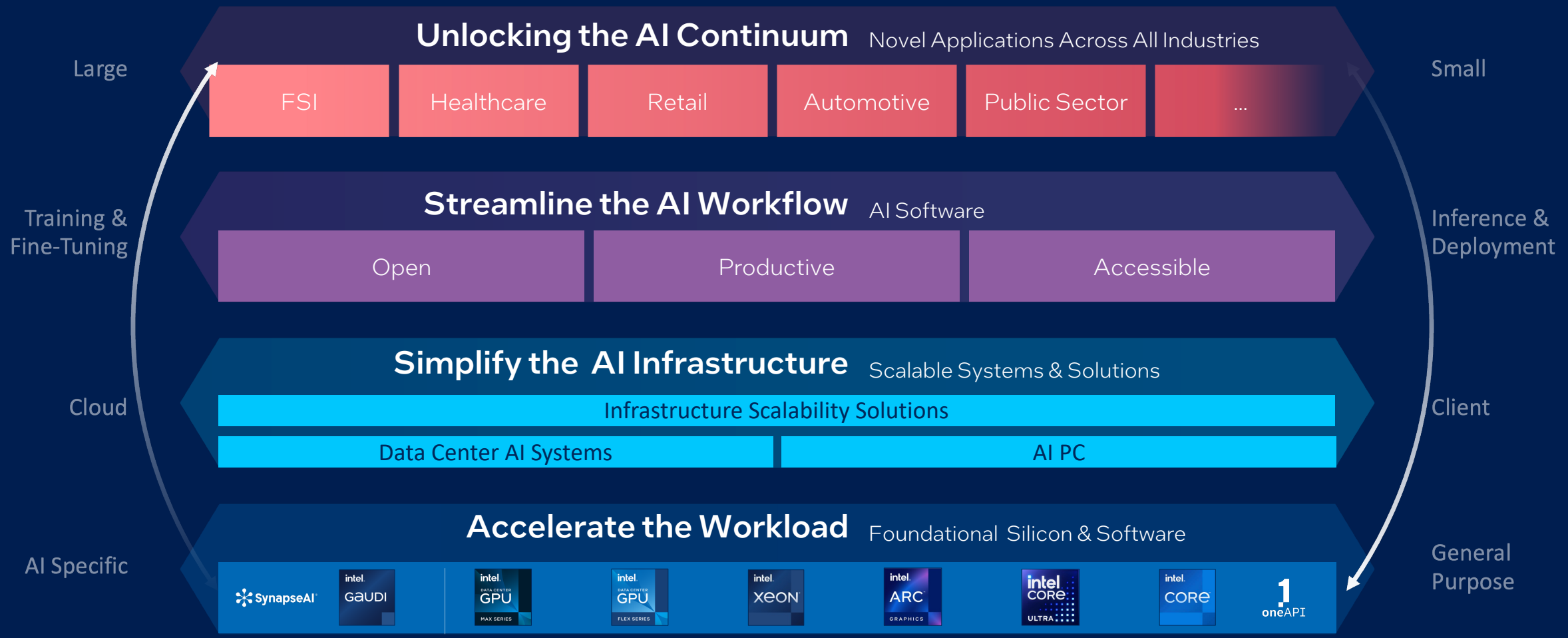
Small

intel innovation

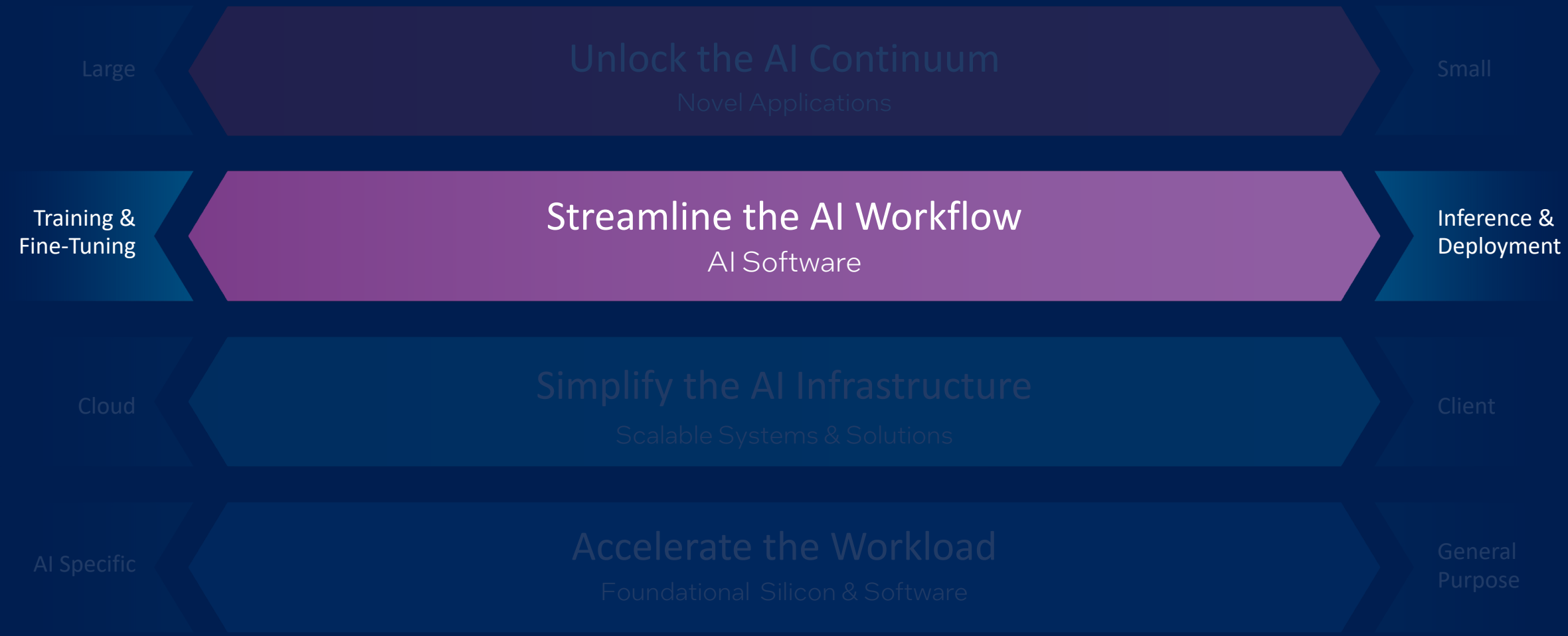
Bringing AI Everywhere



Intel's Approach



Bringing AI Everywhere



AI Workflow Today

Proprietary

Limiting Customization and Hindering Innovation

Fragmented

Challenges Integrating Various AI Frameworks and Libraries

Complex

Barrier For Developers Seeking to Harness the Full Potential of AI

Streamline the AI Workflow

Open

julia · SYCL · Java · OpenCL · OpenXLA · OpenMP · C++ · oneAPI

Productive

Solutions	Pre-configured containers	AI tool Selector
Tooling	Optimized Extensions	OpenVINO
	Modin	CNVRG
References	AI Reference Kits	Hugging Face Collaboration

Accessible

Ecosystem Engagement

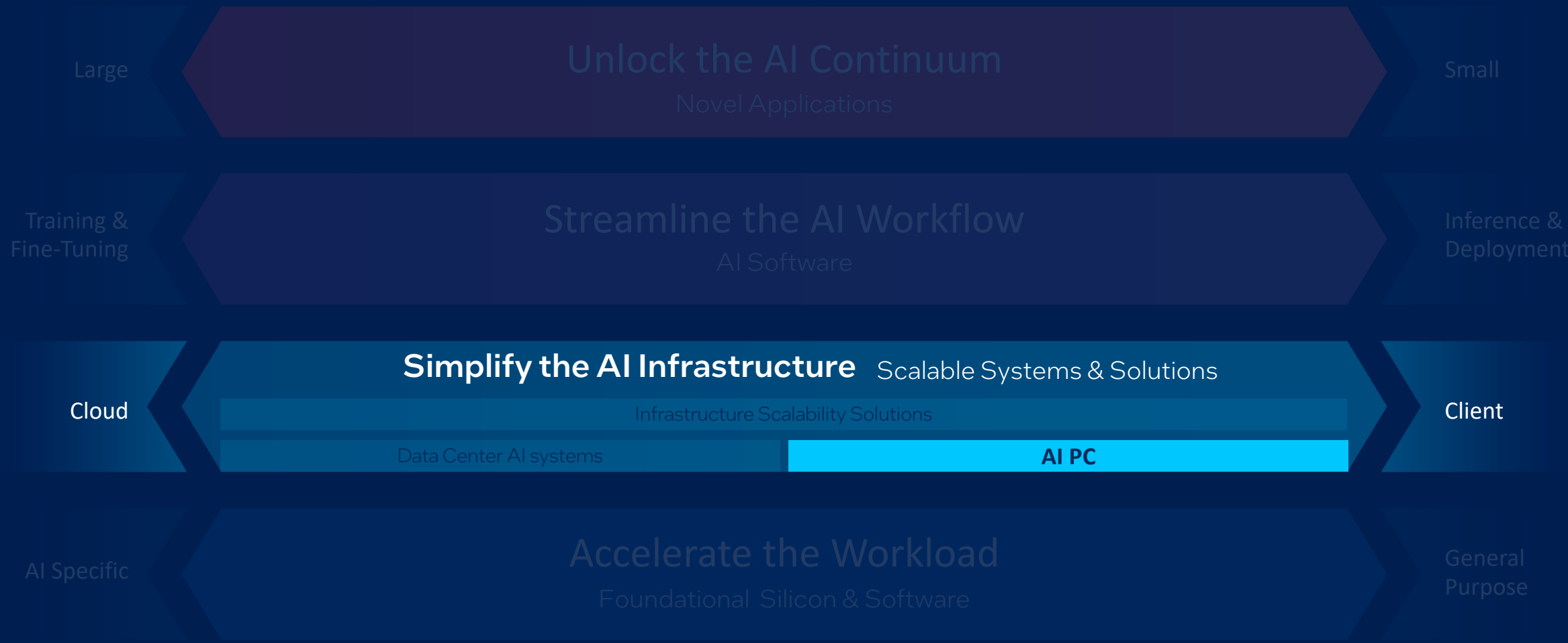
- Industry & Academia
- Solutions Marketplace
- High-Touch Support

Developer Training

- MLOPS training
- Centers of Excellence
- Documentation & Tutorials
- Training Videos
- Summits & Hackathons
- Liftoff Program

intel innovation

Bringing AI Everywhere



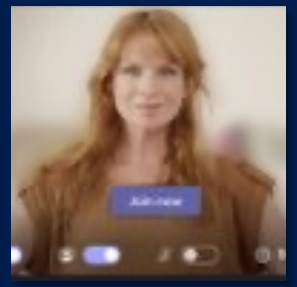
Age of the AI PC

Enabling a World-class, AI-ready PC with High Performance and Broad Compatibility

Enhanced Audio Effects

Elevated Video Collaboration & Streaming

Creator & Gaming Effects



Today – Enhancements

AI Assistants Know Your Daily Context

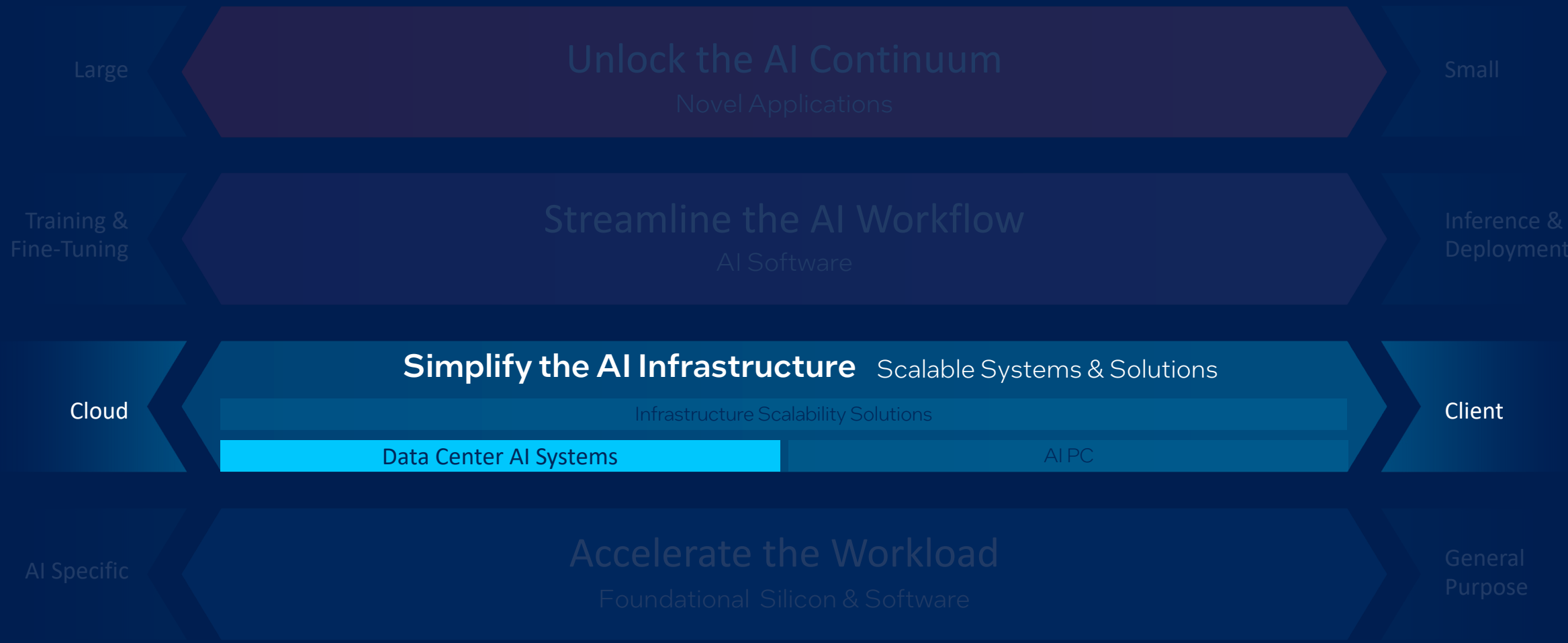
More Creative, Productive, & Collaborative

Across Everything You Do

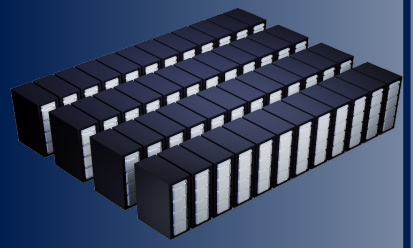
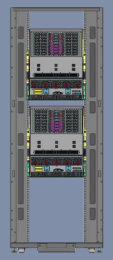
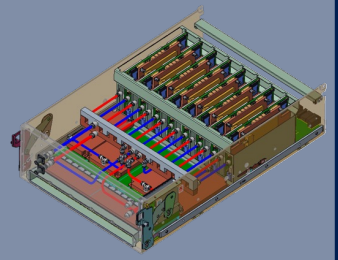
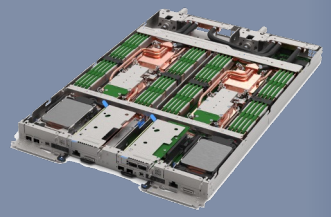
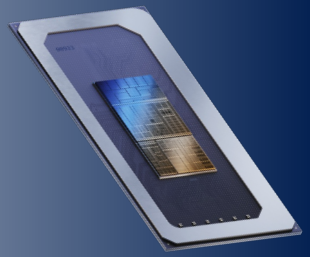
Tomorrow – Everything

intel innovation

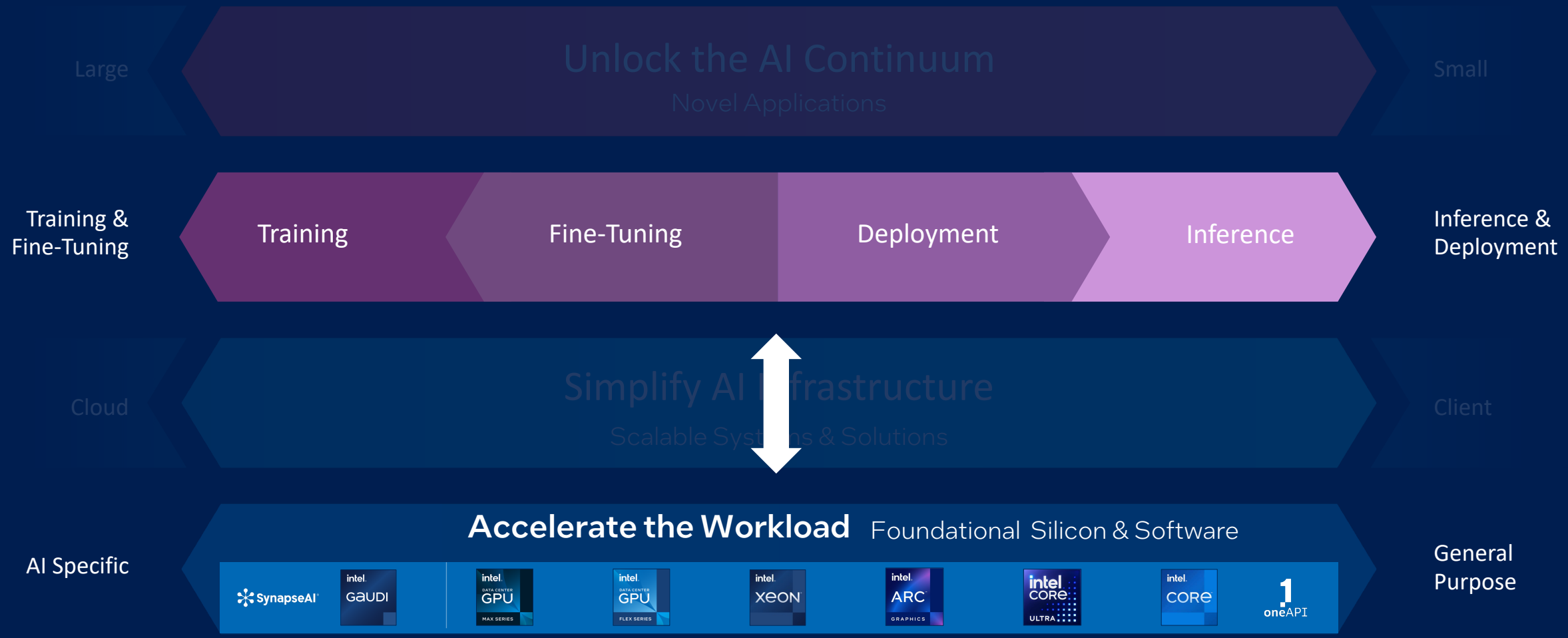
Bringing AI Everywhere



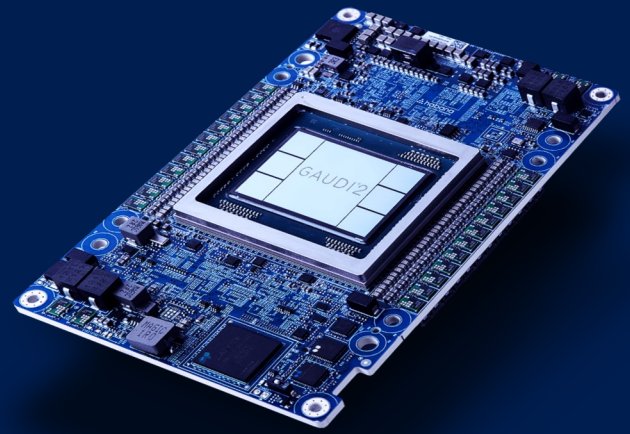
Scalable Systems for Simple AI Infrastructures

Large	State of the Art Model	100B+ Parameters	Up to 100B Parameters	Up to 20B Parameters	<1B Parameters	Small
Training & Fine-Tuning	Training	Peak Inference	Mainstream Inference / Fine-Tuning	Baseline Inference	Endpoint Inference	Inference & Deployment
						
	Cluster & Data Center Scale	Multi-Node Deployment per Rack	Multi-GPU or Multi-Socket CPU	Single- Socket CPU	AI PC	

Bringing AI Everywhere



Intel® Gaudi® 2 AI Accelerator



7nm Process Technology

24 Tensor Processor Cores

96 GB On-Board HBM2

48 MB SRAM

24 Integrated Ethernet ports

Proven Performance

- The ONLY alternative to H100 for training LLMs based on MLPerf
- Trained GPT-3* model in 311 minutes on 384 Intel Gaud2 accelerators

Price Performance

- Intel Gaudi2 accelerators with FP8 estimated to deliver price-performance >H100
- ~2x price-performance to A100

Scalability

- 95% linear scaling on MLPerf GPT-3 training benchmark
- Access large Intel Gaudi2 cluster on the Intel Developer Cloud

Ease of Use

- Software optimized for deep learning training and inference
- PyTorch, Hugging Face, Optimum Library optimizations

intel innovation

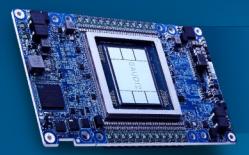
Performance metrics based on MLPerf Training 3.0 benchmark. For configuration details, see the results published by MLPCommons..
GPT-3 model tested on MLPerf Training 3.0 consisted of representative 1% slice of the entire GPT-3 model Performance expectations for Gaudi2 with FP8 based on Intel internal evaluation June 2023
Price-performance claim based on comparable pricing of Intel Gaudi server and Nvidia A100 server and MLPerf Inference 3.1 Results. , Aug 2023. See Supermicro for server pricing. Price-performance claim based on significant pricing differential between Intel Gaudi2 and Nvidia H100 server, MLPerf Training 3.0 Results, May 2023 and internal estimates of performance advancement with FP8. See Supermicro for server pricing. Results

Seamless Code Transitioning

Performant AI Code with Minimal Changes

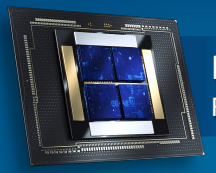
TensorFlow PyTorch deepspeed ...

Across Generations & Architectures



Intel® Gaudi® 2 AI Accelerator

Intel® Gaudi® 3 AI Accelerator



Next Gen GPU (Codenamed Falcon Shores)

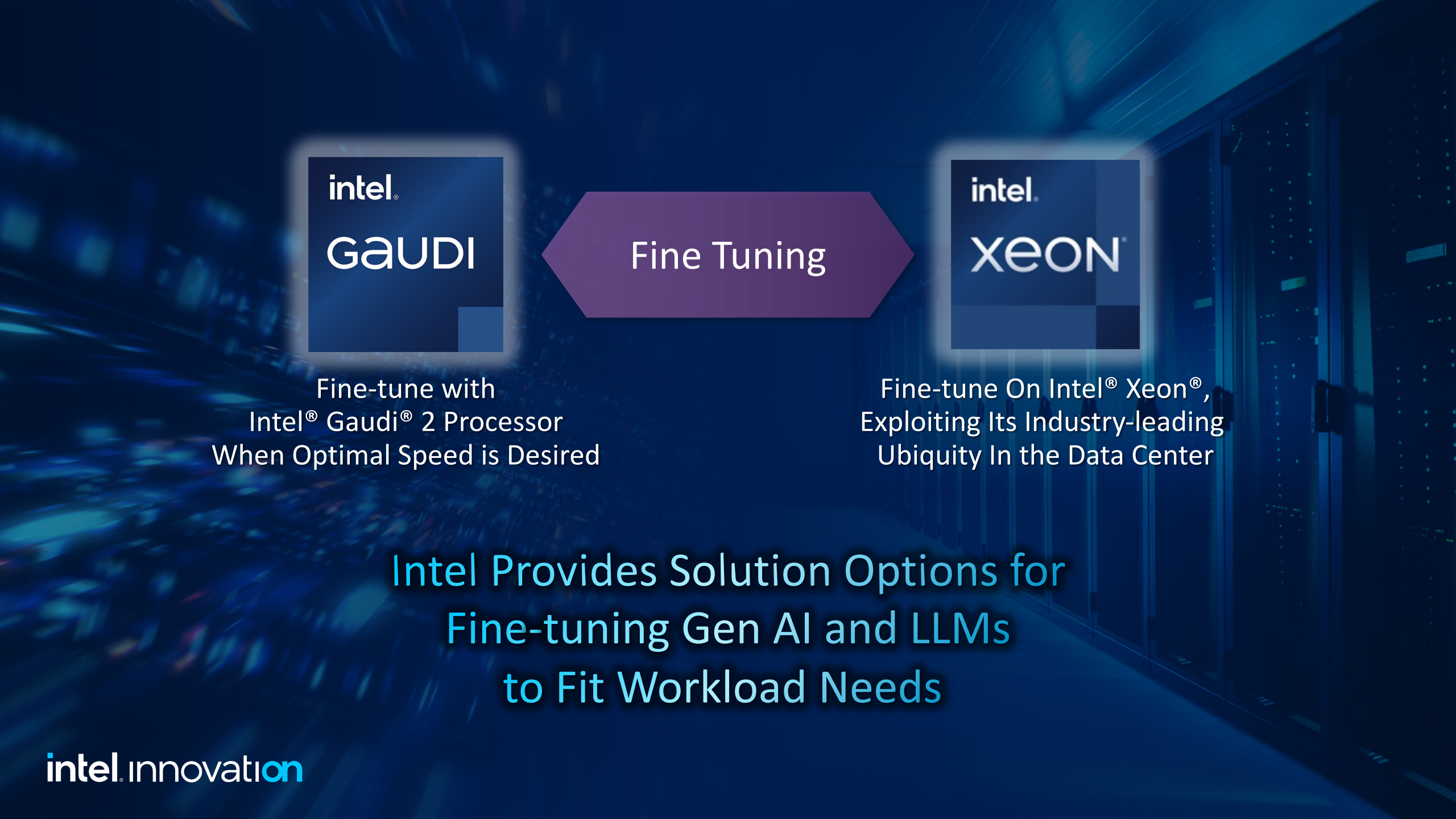
SynapseAI®

Gaudi Software Suite

Powered by Transitioning Into a Single Software Environment

oneAPI

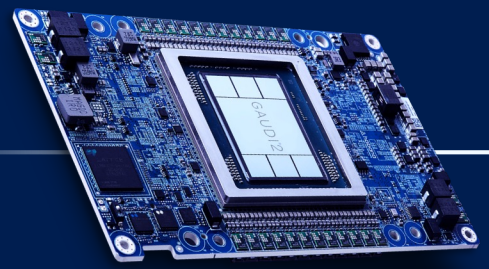
Unified Programming Model

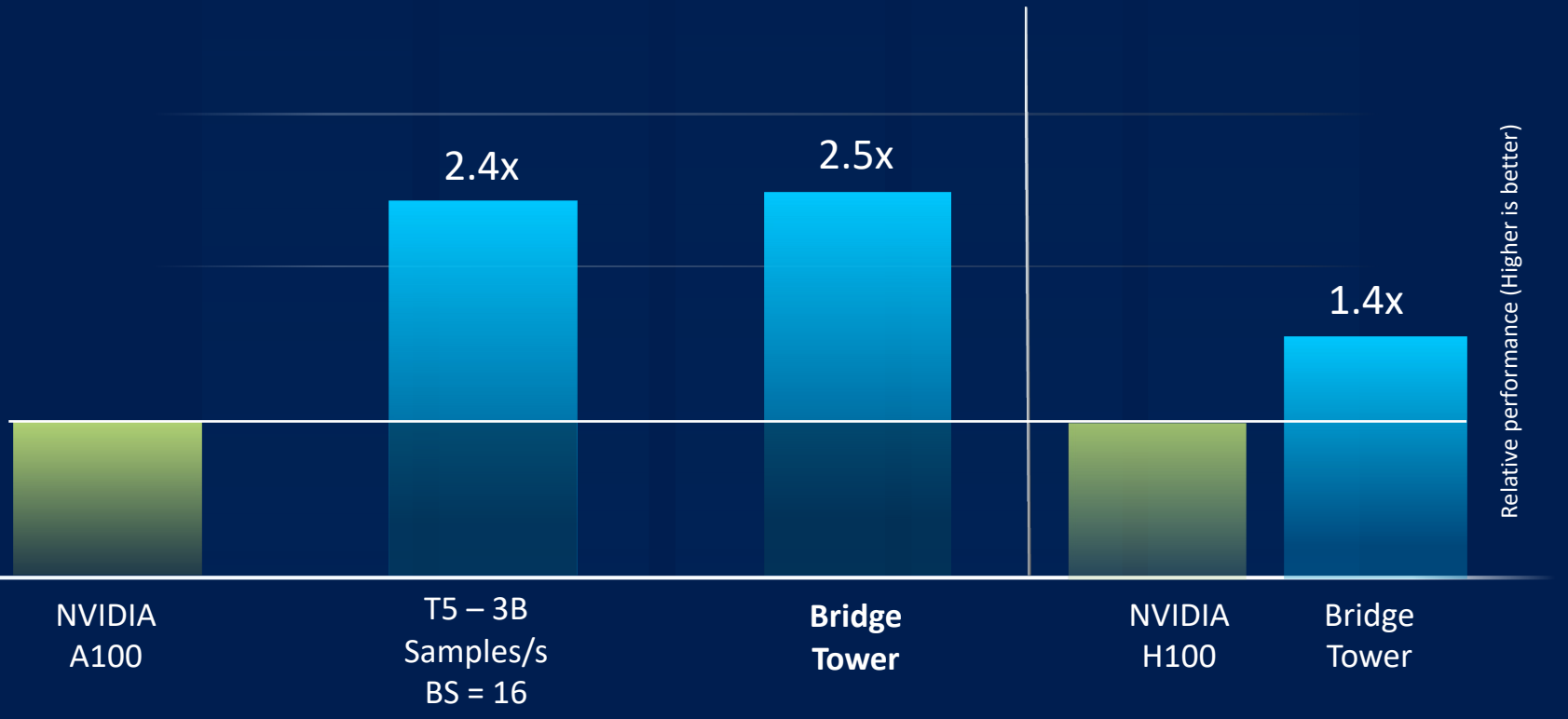

intel
GAUDI
Fine Tuning
intel
XEON
Fine-tune with
Intel® Gaudi® 2 Processor
When Optimal Speed is Desired
Fine-tune On Intel® Xeon®,
Exploiting Its Industry-leading
Ubiquity In the Data Center
Intel Provides Solution Options for
Fine-tuning Gen AI and LLMs
to Fit Workload Needs
intel innovation

intel GAUDI

Fine-tuning Across Numerous LLMs

Hugging Face Evaluations Substantiate Intel® Gaudi® 2 Accelerator LLM Performance vs. Nvidia A100 and H100





intel innovation

Visit https://habana.ai/habana-claims-validation for workloads and configurations. Results may vary.
https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/bridgetower

Out-of-the Box Intel® Xeon® Fine Tuning

Optimized Models & Spaces

Dolly	LLAMA2	MPT	LDM3D	Whisper	*100k's Mode*

Intel Optimized Hugging Face Libraries & Tools

Transformers	Diffusers	Accelerate	PEFT	Optimum
Fine Tuning	Use Cases	Fine Tuning at Scale	Efficient Fine Tuning	Performance Optimization

Foundational Stack

PyTorch + IPEX

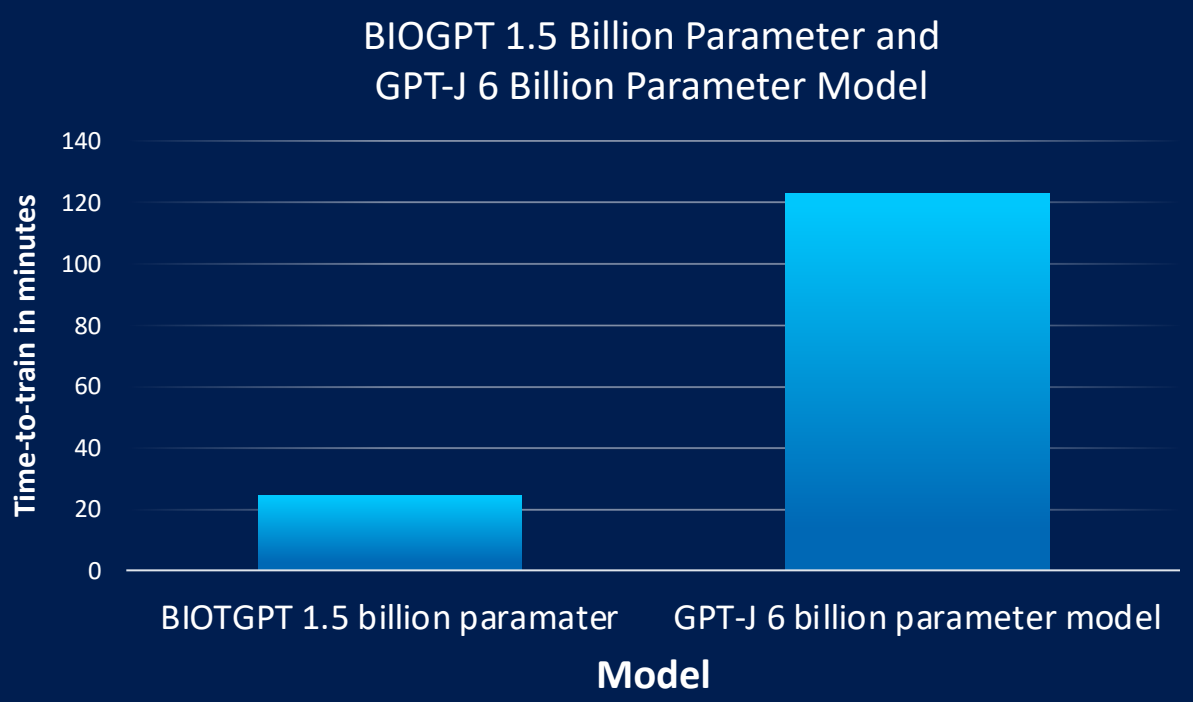
deepspeed + IDEX

ITREX

ONNX RUNTIME

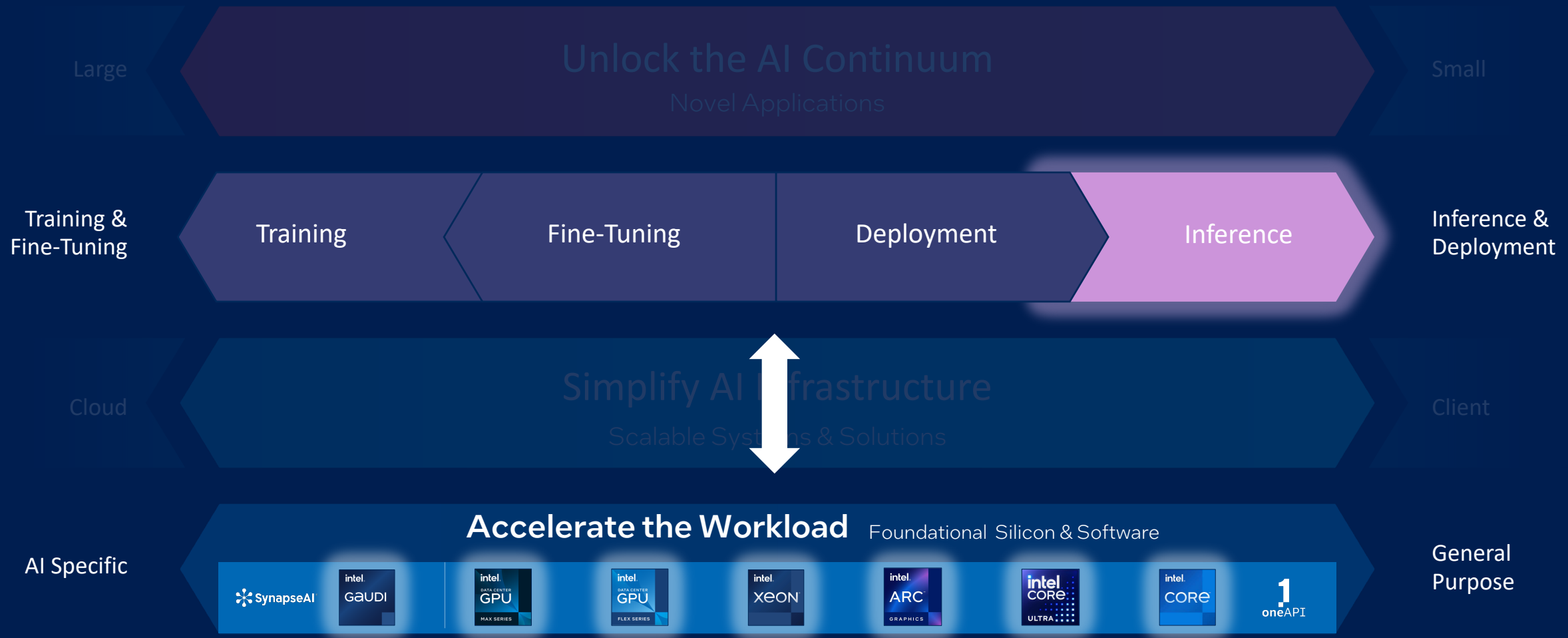
OpenVINO

Multi-node Fine Tuning Open-source Commercial Large Foundational Models In Minutes To Hours



intel innovation

Inference

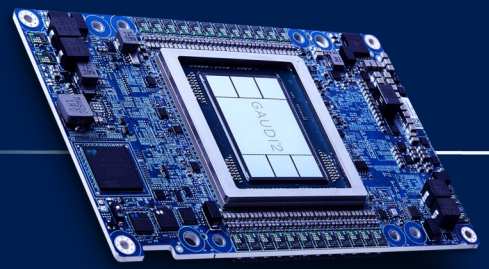

Large
Unlock the AI Continuum
Novel Applications
Small
Training & Fine-Tuning
Training
Fine-Tuning
Deployment
Inference
Inference & Deployment
Cloud
Simplify AI Infrastructure
Scalable Systems & Solutions
Client
AI Specific
Accelerate the Workload
Foundational Silicon & Software
SynapseAI
intel. GAUDI
intel. DATA CENTER GPU MAX SERIES
intel. DATA CENTER GPU FLEX SERIES
intel. XEON
intel. ARC GRAPHICS
intel. CORE ULTRA
intel. CORE
oneAPI
General Purpose

intel GAUDI

Energy Efficiency

Throughput-per-Watt on BLOOMZ 176B Inference is 1.79x better than H100; 1.61x better than A100

Inference Advantage Across Multiple LLM Performance Metrics

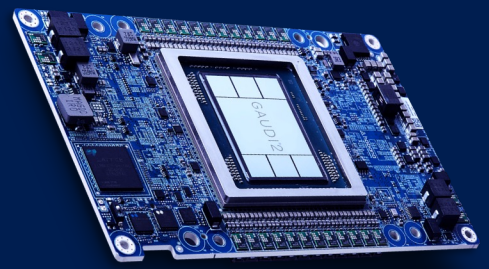


intel innovation

https://huggingface.co/blog/habana-gaudi-2-benchmark
https://huggingface.co/blog/habana-gaudi-2-bloom
Visit https://habana.ai/habana-claims-validation/for workloads and configuration regarding power consumption claims. **Results may vary.**

Intel® Gaudi® 2 AI Accelerator:

Solving LLM Challenges



Inference on GPT-J

Intel Gaudi 2 Accelerator with FP8

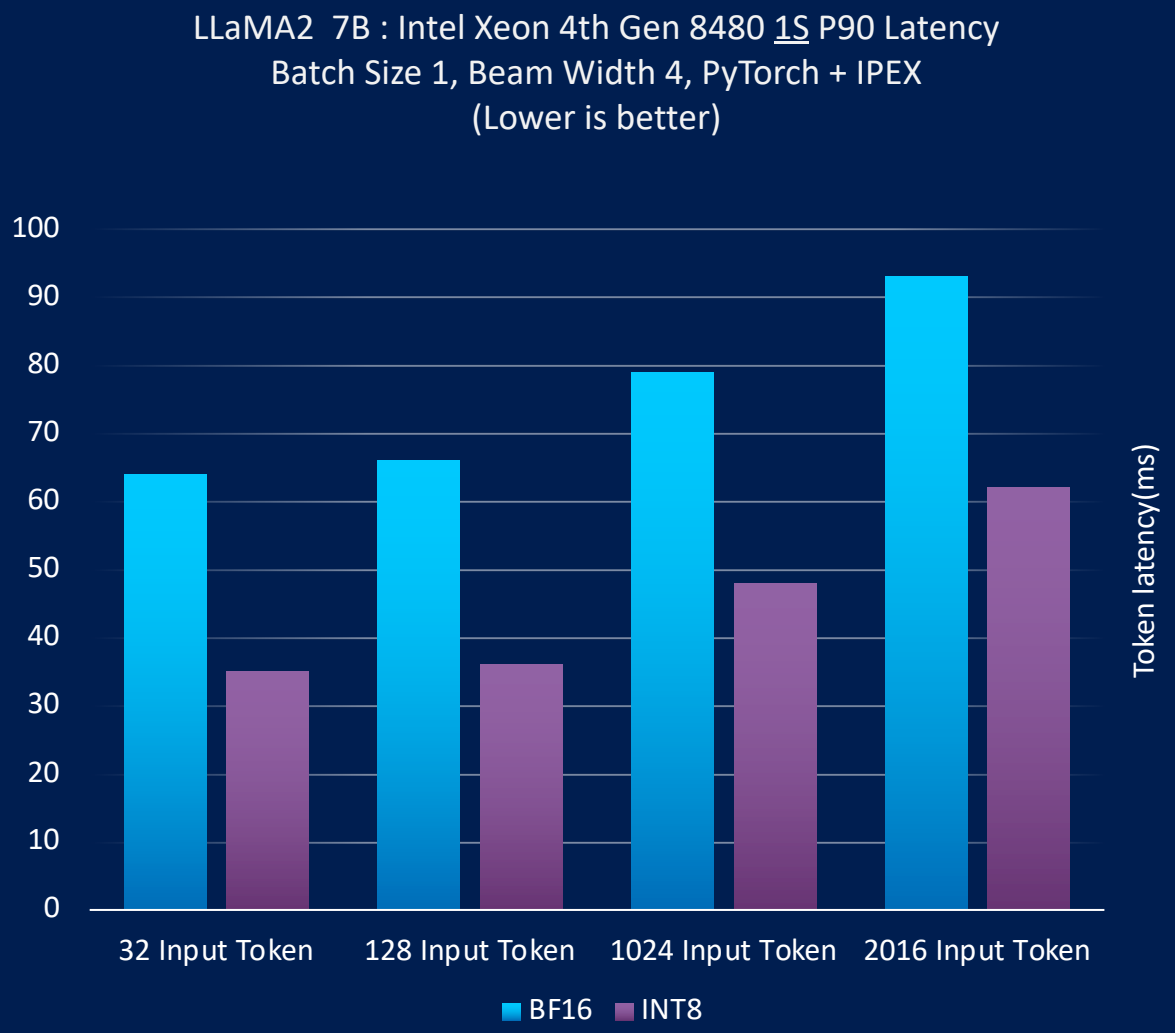
- Near-parity* on GPT-J with H100
- Outperformed A100 by 2.4x (Server) and 2x (Offline)
- Achieved 99.9% accuracy with FP8

GPT-J On MLPerf Inference Benchmark

intel innovation

Source: MLPerf Inference 3.1 Data Center Benchmark Results: https://mlcommons.org/en/inference-datacenter-31/
Intel® Gaudi® 2 AI accelerator on GPT-J Vs H100 with 1.09x (Server) and 1.28 (Offline). Results may vary

LLaMA2 (7B) Inference with 4th Gen Intel® Xeon® Processors



- Use any popular industry standard AI libraries
- Intel AI Platform validated with over 300 inference models
- One socket of 4th Gen Intel® Xeon® processors can run LLaMa2 chatbots in under 100ms 2nd token latency

intel innovation

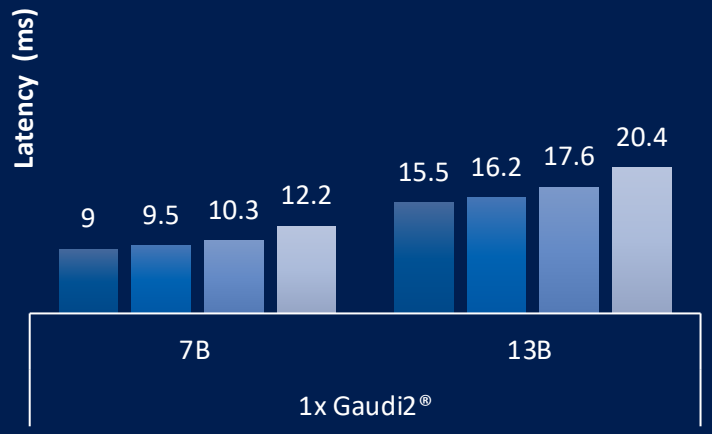
Results shown for bare metal.

Inference Across Multiple Products

LLaMA2 7B & 13B Inference

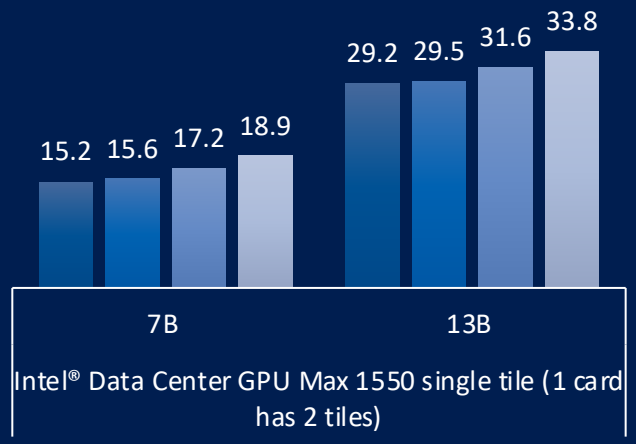
Llama 2 Next Token Latency (Lower is Better)

Greedy Mode, Mixed Precision (bfloat16), BS = 1, 256 Output Tokens



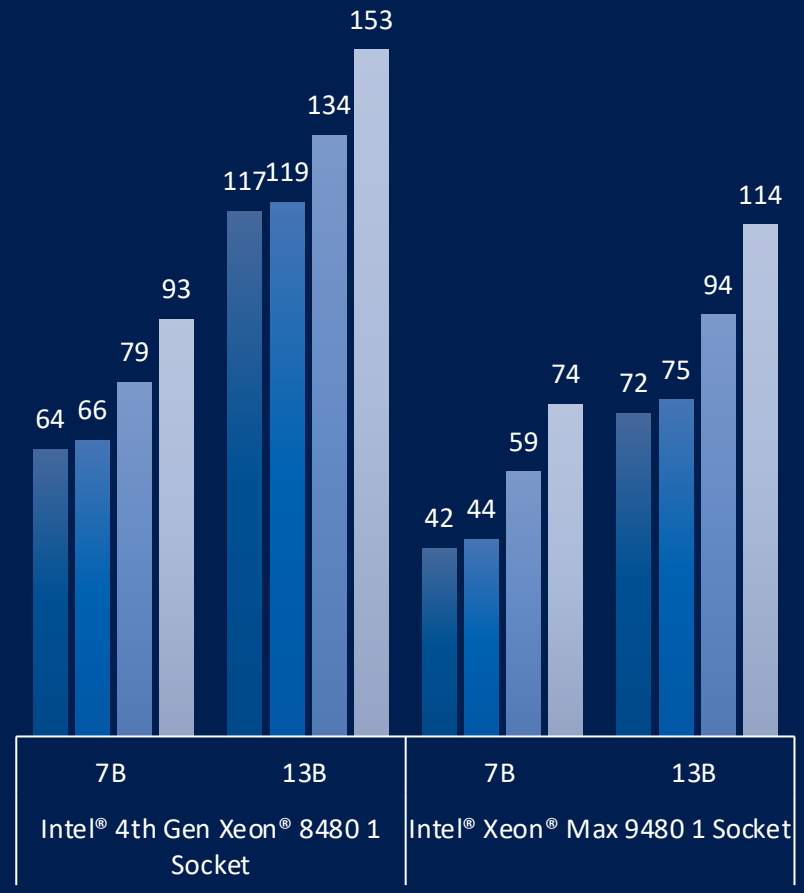
Intel® Gaudi2 ®AI Accelerator

On 1 Tile (out of 2 tiles per card) Intel® Data Center GPU Max 1550



Intel Data Center GPU Max 1550

On 1 Socket Intel® Xeon® Scalable Processor

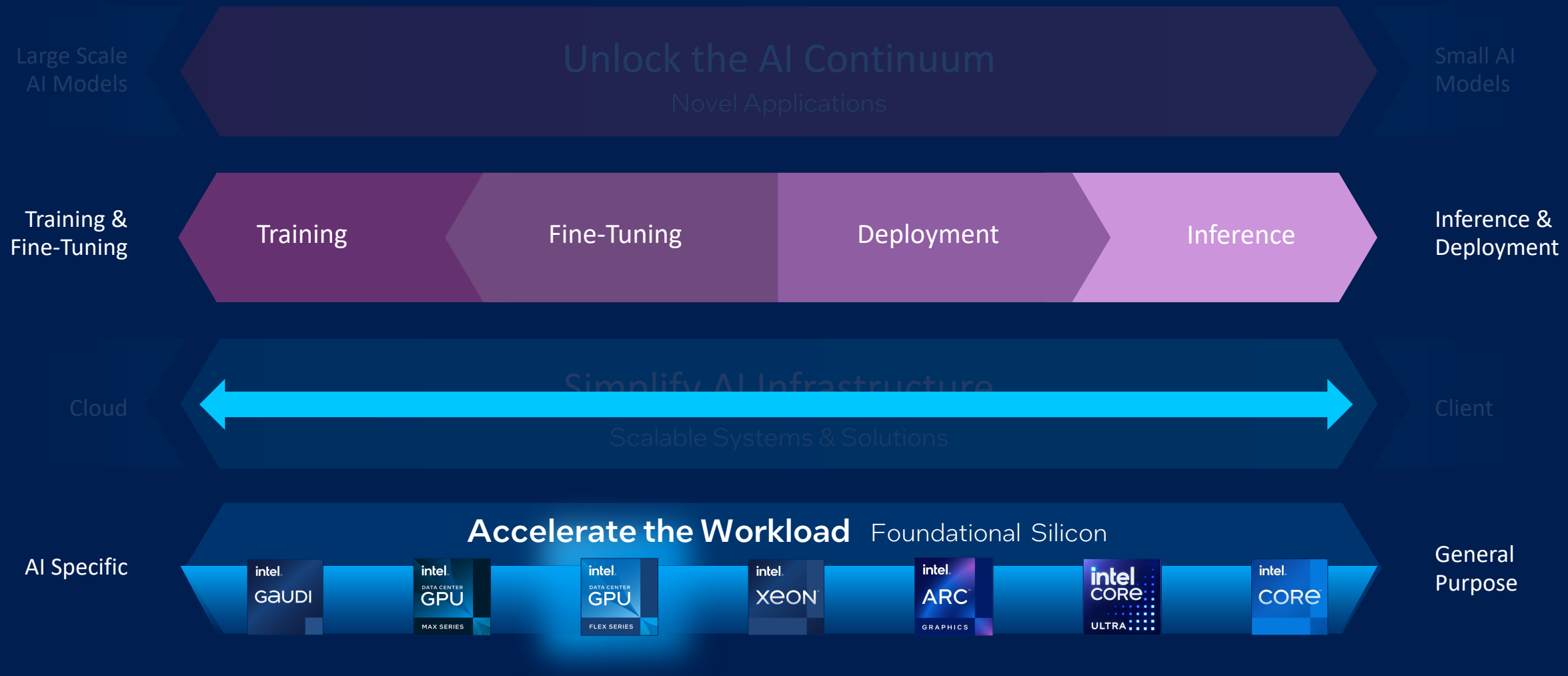


Intel Xeon Scalable Processor

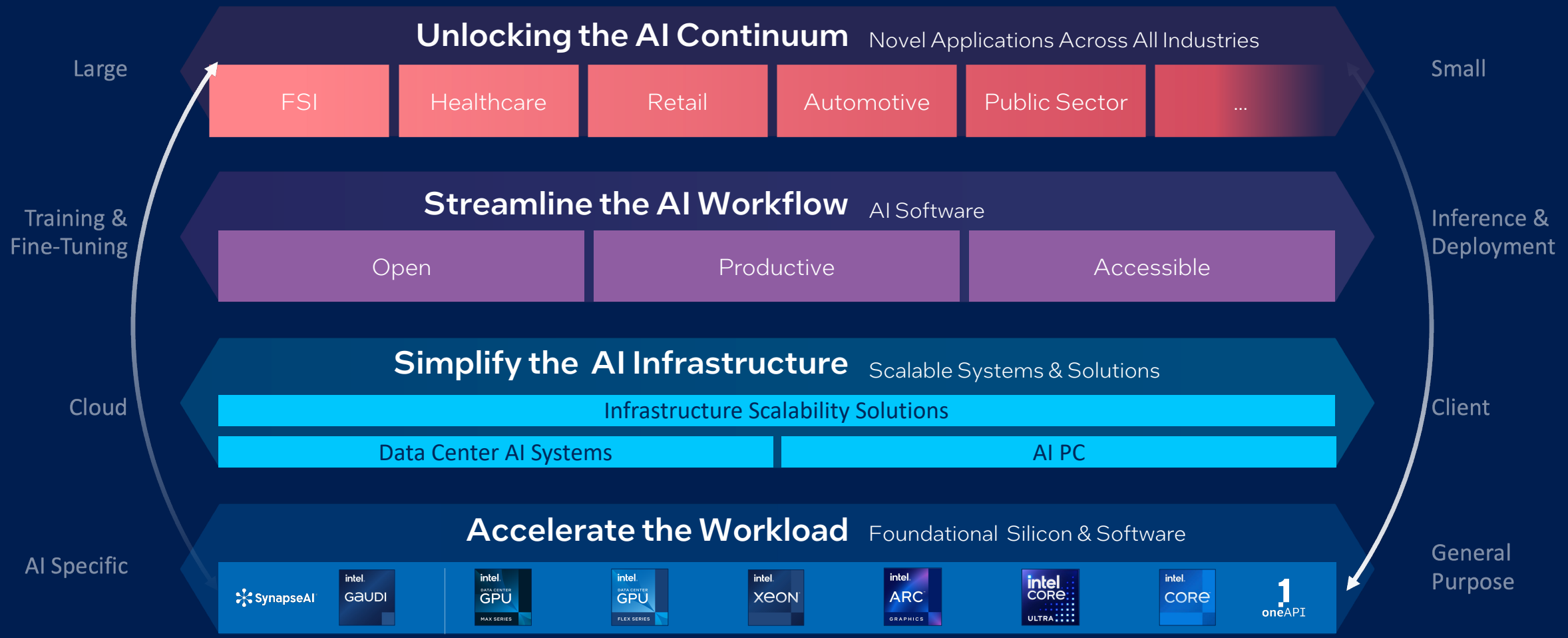
intel innovation

For configuration details see: https://habana.ai/habana-claims-validation/ and www.intel.com/PerformanceIndex

Bringing AI Everywhere



Intel's Approach




intel
innovation
Thank You!

Notices and Disclaimers

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

intel innovation