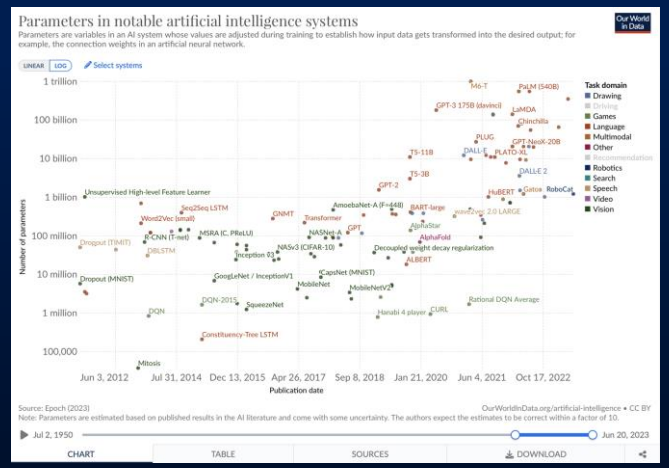
intel ai summit
英特爾 AI 科技論壇

Bringing AI Everywhere

Accelerating GenAI for Enterprise

Sean Kuo
AI Center of Excellence APJ Sales Lead
March 27th, 2024

AI Exponentially Increasing In Quality & Size



Model Size

15,000x
Increase In 5 Years

GPT4+
>1 Trillion

Megatron-Turing
530B

GPT-3
175B

Transforme r
65M

BERT
340M

GPT-2
1.5B

GPT-2 8B
8.3B

TS
11B

Turing-NLG
17B

Mid 2017 | 2018 | 2019 | Mid 2019 | Late 2019 | 2020 | Mid 2020 | Late 2021 | 2022

intel AI summit

Source: Amazon Web Services , 2022

Amazing Innovation

But Huge Models are Inaccessible For Most

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

intel Ai summit

Source: CSET, CNBC

Amazing Innovation

But Huge Models are Inaccessible For Most

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

intel Ai summit

Source: CSET, CNBC

Specialized AI Models

The answer for the "masses"

Large Foundational Model

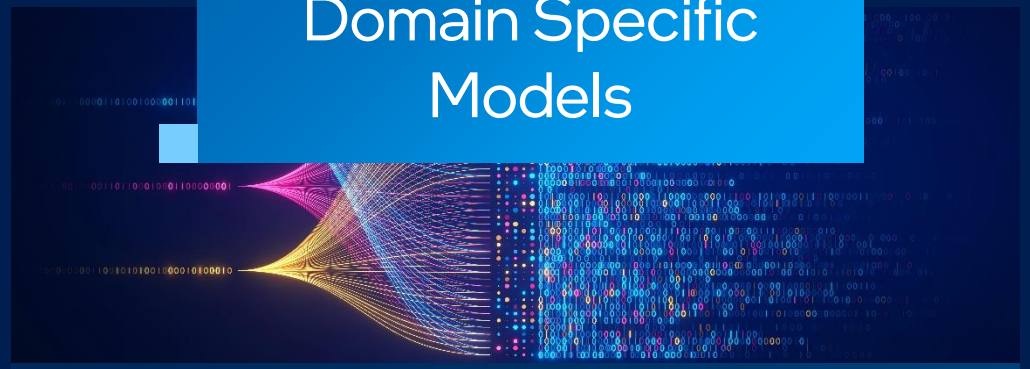
Advantages

+ Incredible all-in-one, out-of-the-box versatility: text, programming, continual natural language conversation and plain summarization
+ Surprisingly, compelling outcomes

Challenges

- Big (>100B parameters), expensive- $4m+ to train, $3m per month for inferencing
- Hallucinations; lack of explainability, intellectual property issues
- Frozen in time (sampling)

Domain Specific Models



Advantages

+ 10-100x smaller models while maintaining/improving accuracy
+ Economical on general-purpose compute
+ Correctness; Source attribution; Explainability
+ Utilizing private/enterprise data
+ Continuously updated information

Challenges

- Reduced range of tasks
- Requires few-shot fine-tuning and indexing

Specialized AI Models

The answer for the "masses"

Large Foundational Model

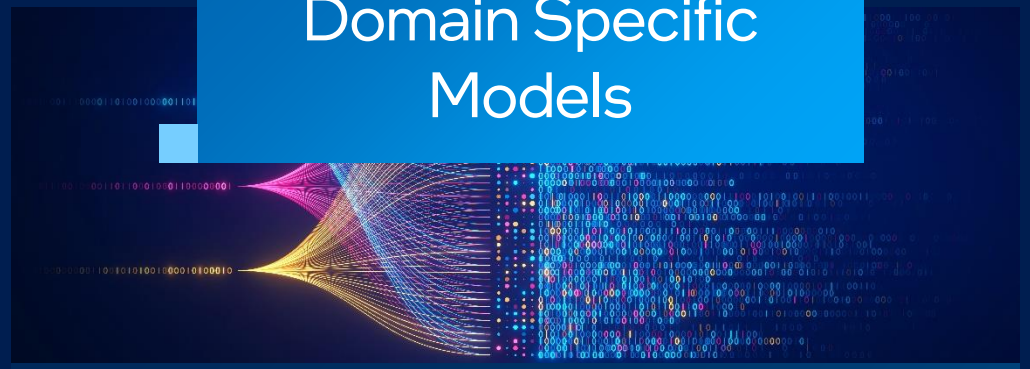
Advantages

+ Incredible all-in-one, out-of-the-box versatility: text, programming, continual natural language conversation and plain summarization
+ Surprisingly, compelling outcomes

Challenges

- Big (>100B parameters), expensive- $4m+ to train, $3m per month for inferencing
- Hallucinations; lack of explainability, intellectual property issues
- Frozen in time (sampling)

Domain Specific Models



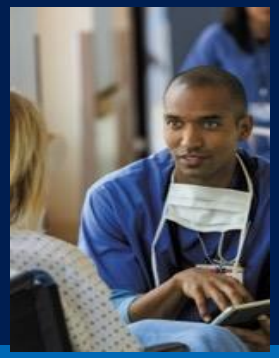
Advantages

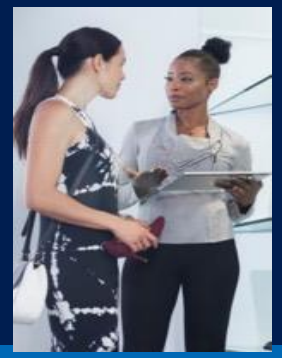
+ 10-100x smaller models while maintaining/improving accuracy
+ Economical on general-purpose compute
+ Correctness; Source attribution; Explainability
+ Utilizing private/enterprise data
+ Continuously updated information

Challenges

- Reduced range of tasks
- Requires few-shot fine-tuning and indexing

Specialized Models Enable Scale






Education	Health	Finance	Retail	Government	Energy	Automotive	Manufacturing	Telco
Teacher Assistant	Drug Discovery	Algorithmic Trading	Product Promotion	Gov Services Assistant	Energy Consumption Forecasting	Autonomous Car Development	Factory Automation	Personalized Customer Services
Student Study Buddy	Doctor Assistant	Customer Portfolio Assistant	Customer Interface and Sentiment Tool	Document Search Summarization	Operational Performance	Multi-language in car aid	Predictive Maintenance	Network Automation
Parent Chat Portal	Patient Family chatbot	Risk / Credit Assessment	Image Shopping Aid	Live Language Translation	Energy Trading Assistant	Supply Chain Optimization	Precision Agriculture	Operational Performance

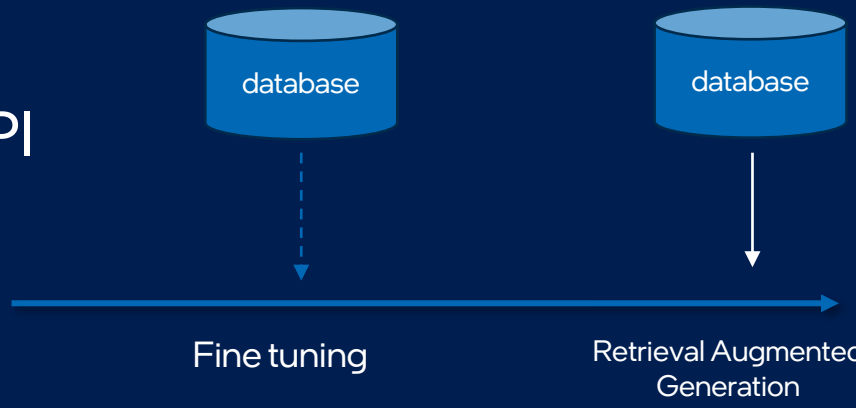
intel Ai summit

Enterprise options to build specialized GenAI

1.

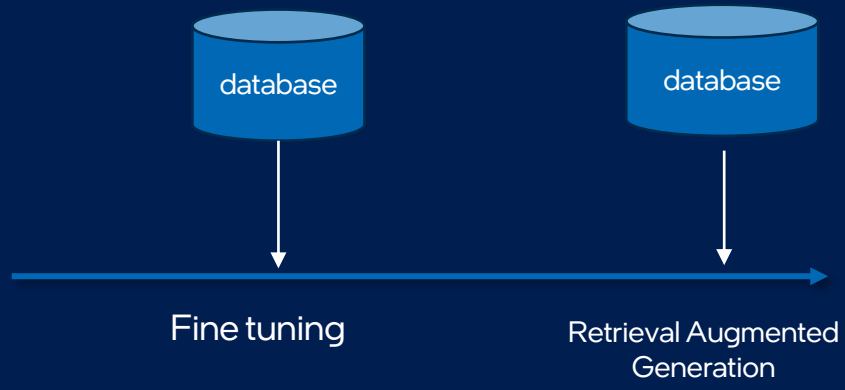
API

Large Foundation Models



2.

General Open Models



Domain Specific Models

Finance ChatBot

Personalized Marketing Content

Code Generator

Clinical Notes Transcription

Text-to-Speech Generator

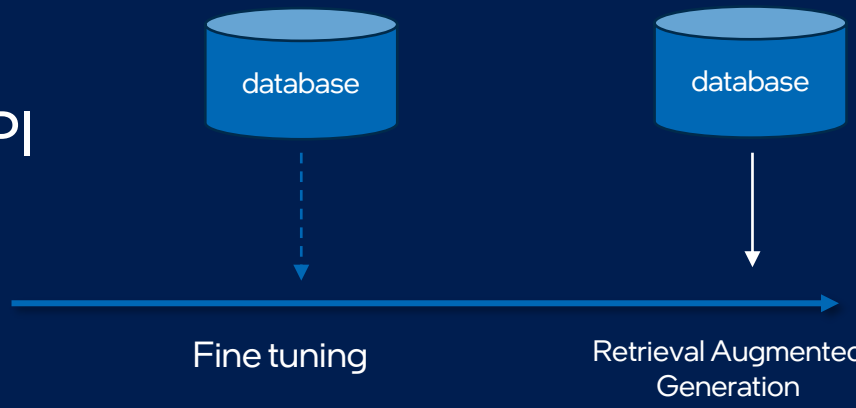
intel AI summit

Enterprise options to build specialized GenAI

1.

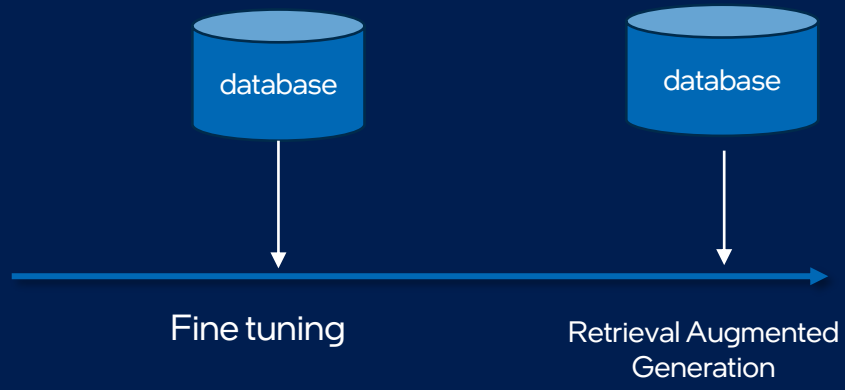
API

Large Foundation Models



2.

General Open Models



Domain Specific Models

Finance ChatBot

Personalized Marketing Content

Code Generator

Clinical Notes Transcription

Text-to-Speech Generator

intel AI summit

Intel has the software tools developers use to scale AI Everywhere

Upstream

Integrated acceleration to popular open-source software

PyTorch, TensorFlow, ONNX RT, more ...

Intel Extension

Easily pluggable extensions to open-source software

Intel Extension for PyTorch, Intel Extension for TensorFlow, more ...

Intel Tools

Tools / Kits which improve productivity and performance on Intel hardware

OpenVINO™, oneContainer Portal, Intel Neural Compressor, SynapseAI®, Developer Cloud

Across major software channels (PyPI, Anaconda, Intel, Apt, Yum, Docker) and ecosystems, (Optimum Intel through Hugging Face)

intel Ai summit

Bringing AI Everywhere

Large

Unlock the AI Continuum

Novel Applications

Small

Training & Fine-Tuning

Streamline the AI Workflow

AI Software

Inference & Deployment

Cloud

Simplify the AI Infrastructure

Scalable Systems & Solutions

Client

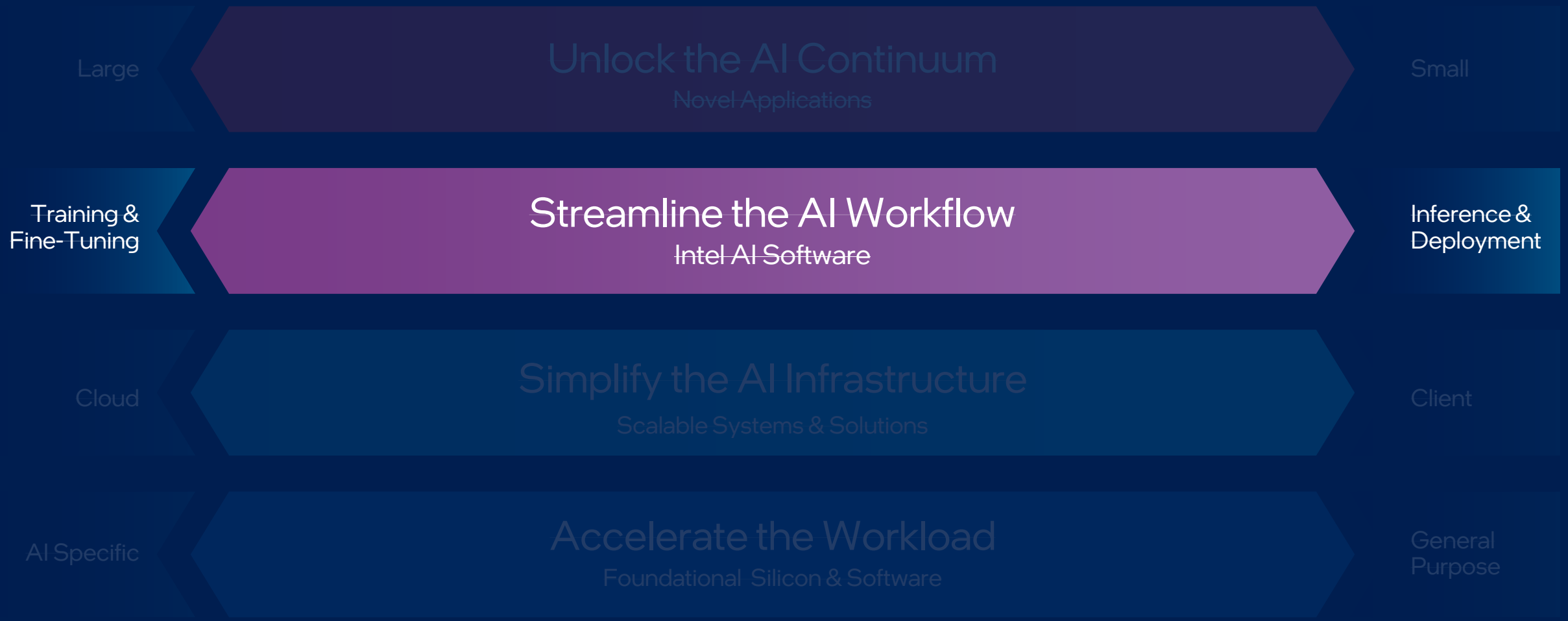
AI Specific

Accelerate the Workload

General Purpose

intel AI summit

Bringing AI Everywhere



intel AI summit

Streamline the AI Workflow

Training | Fine-Tuning | Deployment | Inference

MODIN, SciPy, pandas, NumPy, Spark, Numba, dmlc XGBoost, PyTorch, ONNX RUNTIME, OpenVINO, Hugging Face, scikit learn, TensorFlow, DirectML, cnvrg.io, SigOpt AutoML, intel Neural Compressor

Open

julia, SYCL, Java, OpenCL, OpenXLA, OpenMP, C, 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 AI summit

Streamline the AI Workflow

Training | Fine-Tuning | Deployment | Inference

MODIN, SciPy, pandas, NumPy, Spark, Numba, dmlc XGBoost, PyTorch, ONNX RUNTIME, OpenVINO, Hugging Face, scikit learn, TensorFlow, DirectML, cnvrg.io, SigOpt AutoML, intel Neural Compressor

Open

julia, SYCL, Java, OpenCL, OpenXLA, OpenMP, C, 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

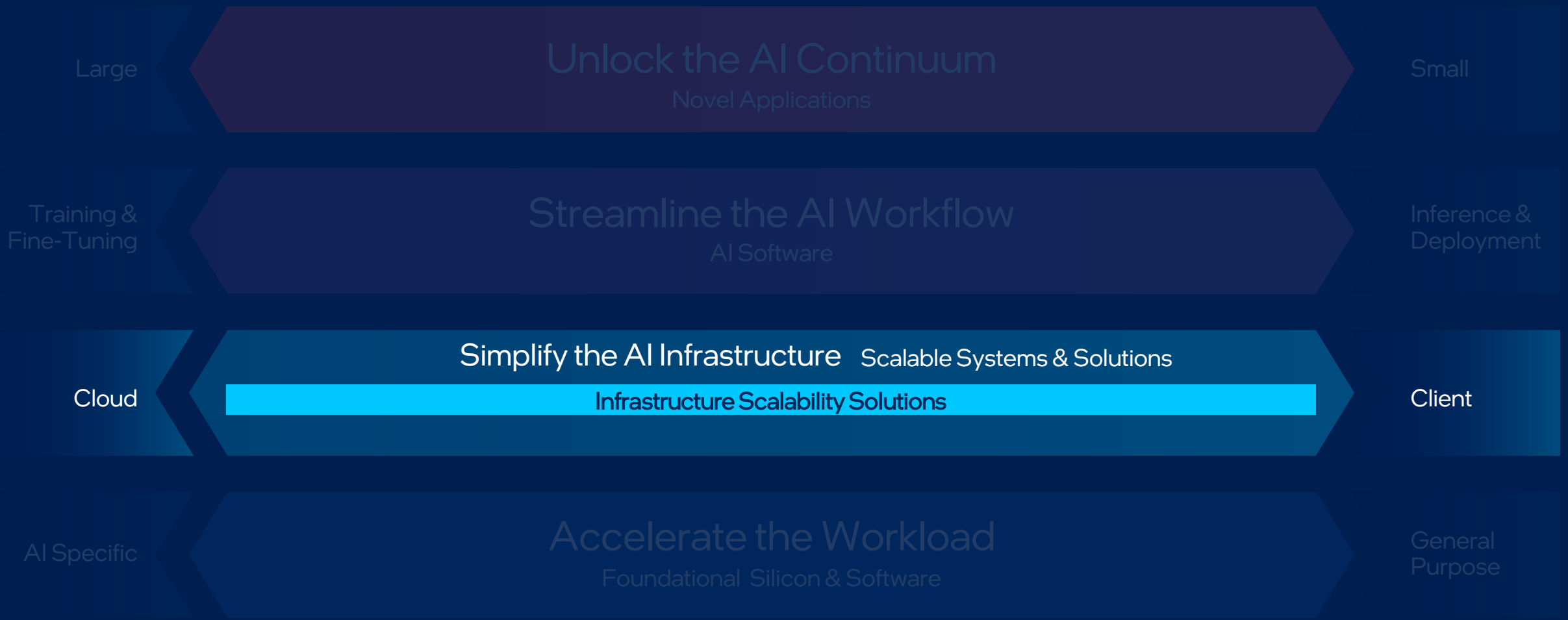
- Industry & Academia
- Solutions Marketplace
- High-Touch Support

Developer Training

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

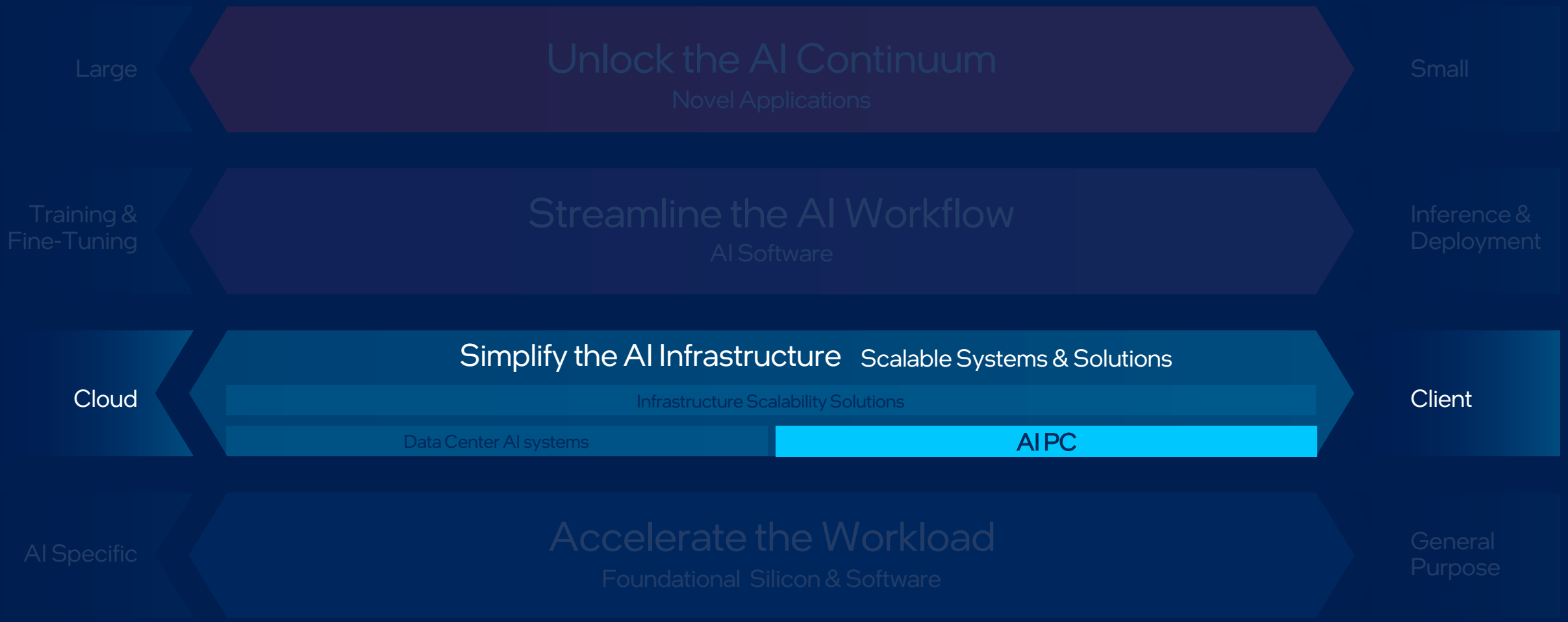
intel AI summit

Bringing AI Everywhere



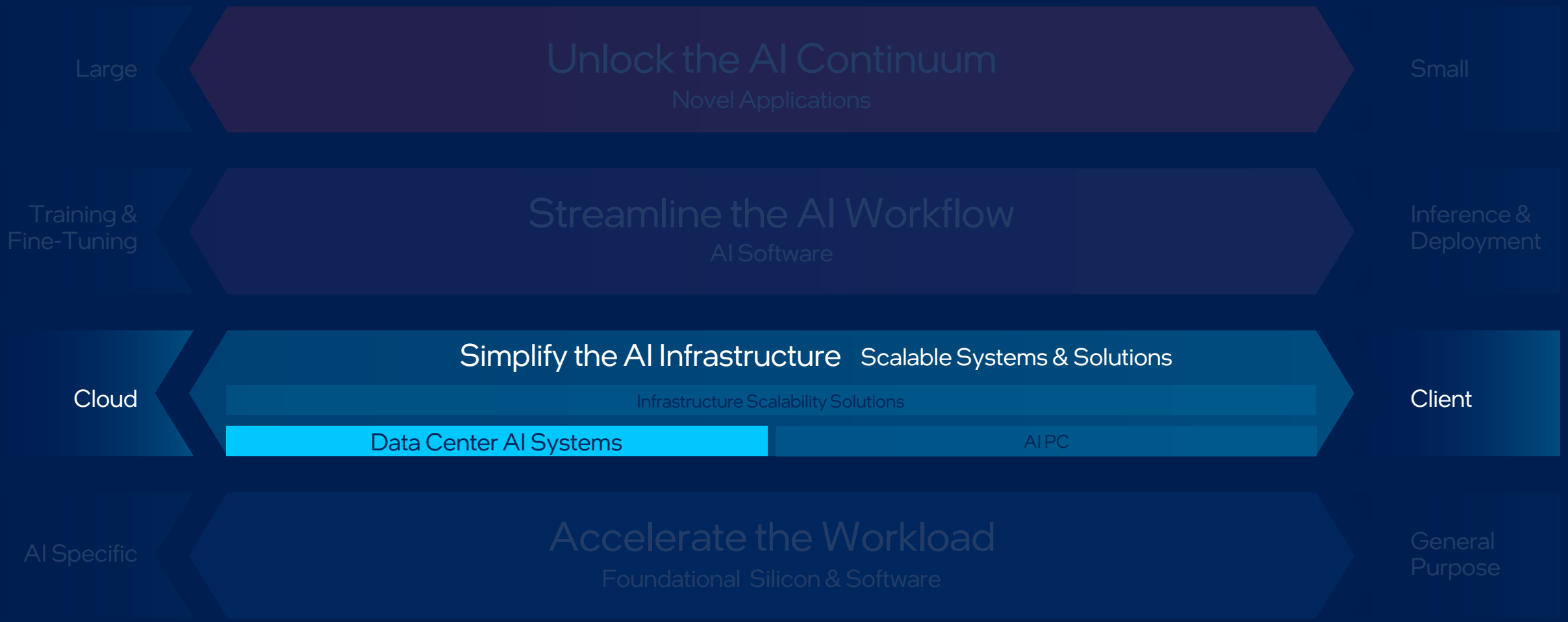
intel Ai summit

Bringing AI Everywhere



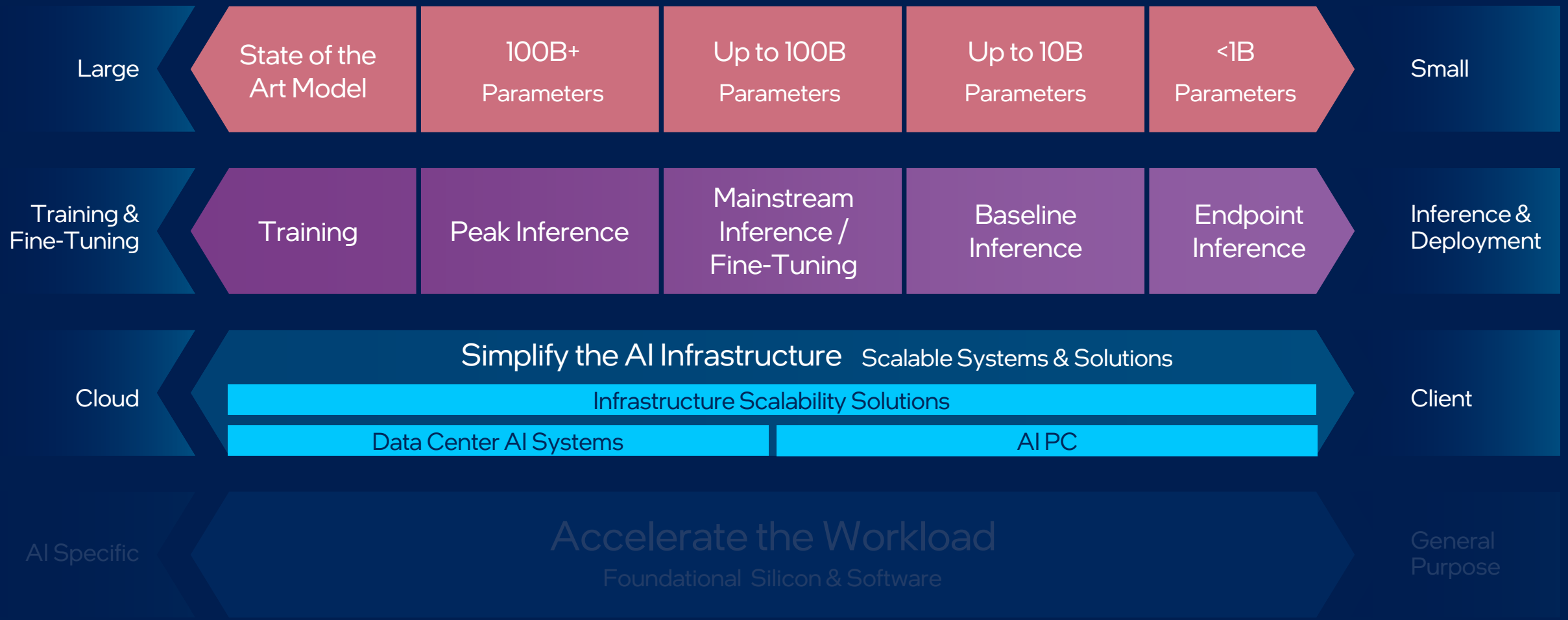
intel Ai summit

Bringing AI Everywhere



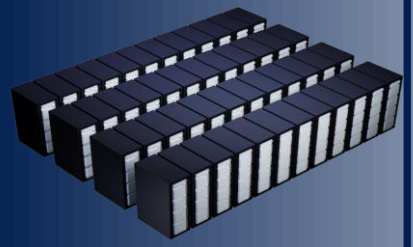
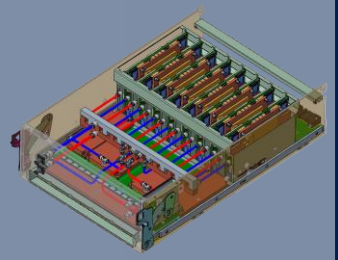
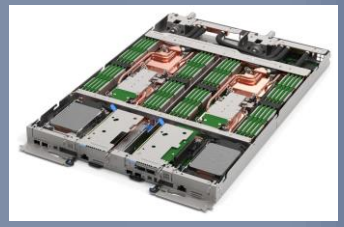
intel Ai summit

Scalable AI Workloads From Cloud to Client



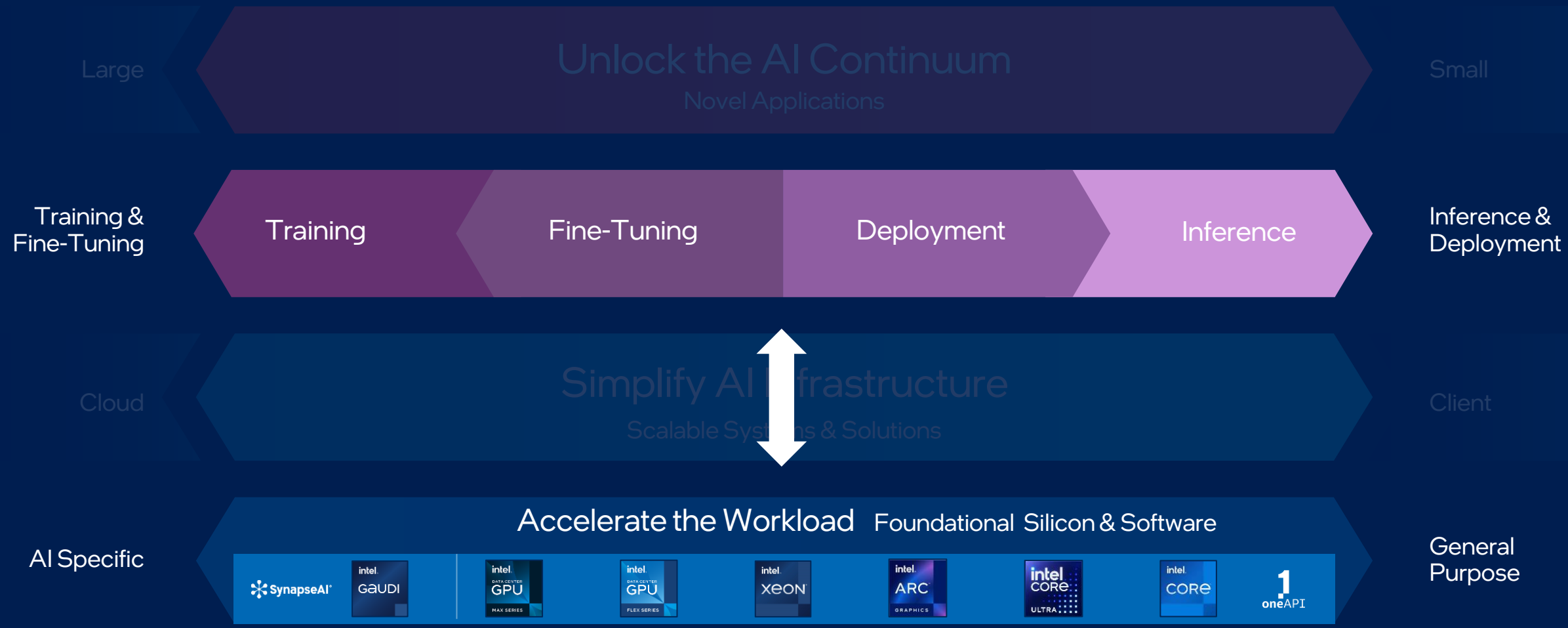
intel AI summit

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	

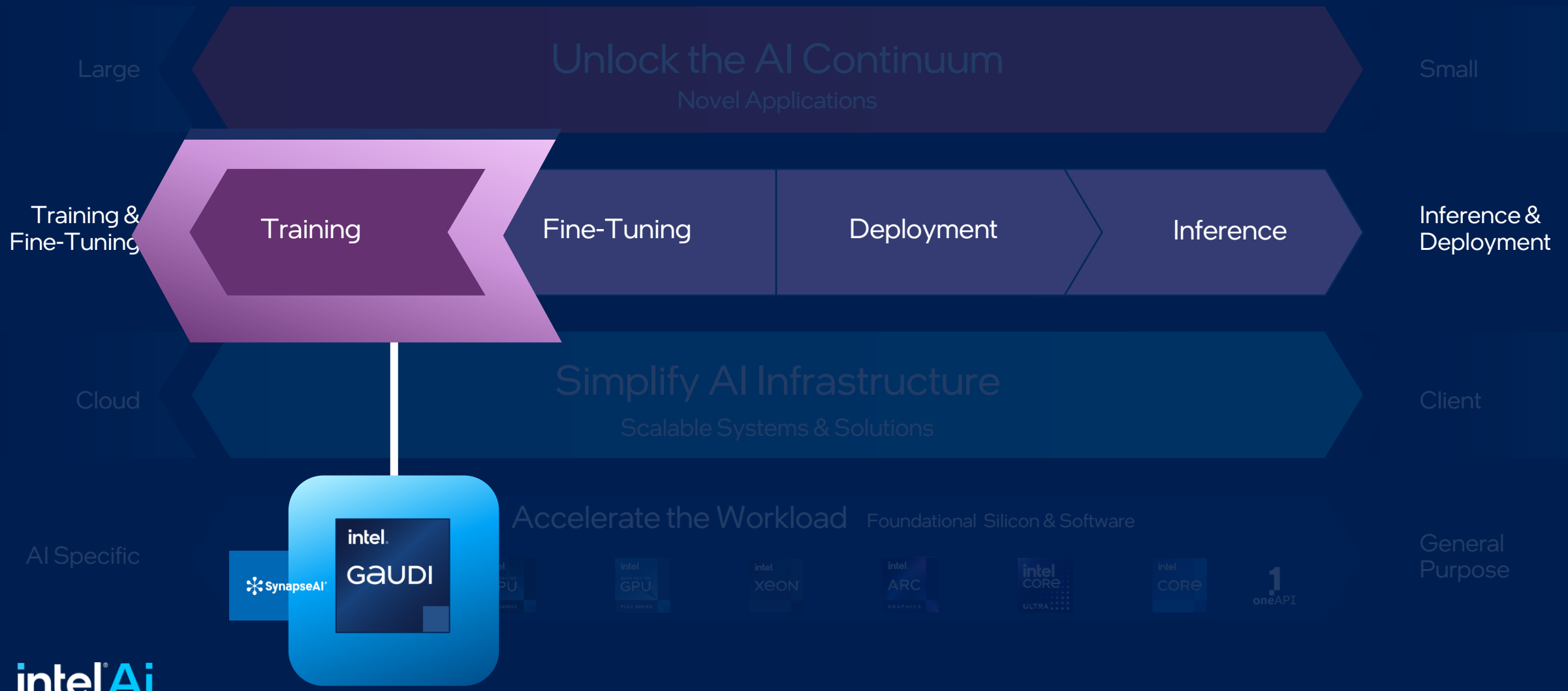
intel Ai summit

Bringing AI Everywhere



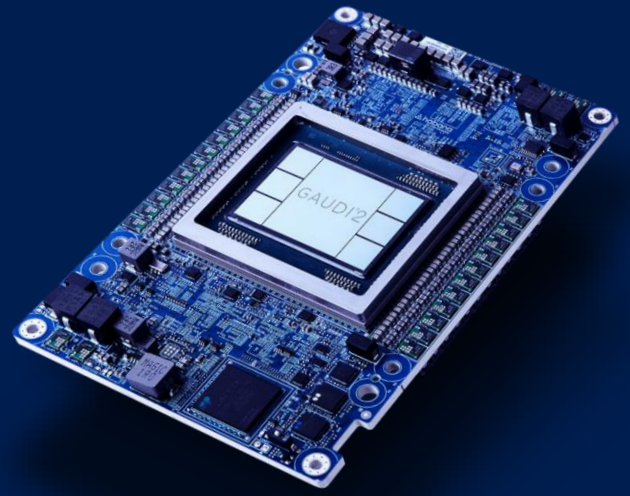
intel AI summit

Bringing AI Everywhere



intel AI summit

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 TTT doubled in 2023 from 311 minutes in Jun to 153 min in Nov'23 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 Ai summit

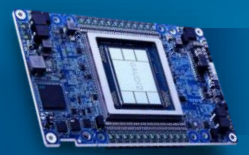
Performance metrics based on MLPerf Training 3.0 benchmark. For configuration details, see the results published by MLPCommons..
GPT-3 model tested on MLPerf Training 3.0 consisted of representative 1% slice of the entire GPT-3 model Performance expectations for Gaudi2 with FP8 based on Intel internal 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

Seamless Code Transitioning

Performant AI Code with Minimal Changes

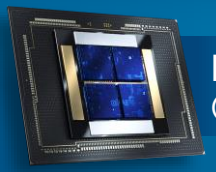
TensorFlow PyTorch deepspeed ...

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 Ai summit

Stability.AI

Performance Evaluation of Gaudi

- Multimodal Diffusion Transformer model (MMDiT) on Stable Diffusion 3 (in preview now)
- MMDiT: 50% faster training than H100-80GB & 3x faster than A100-80GB
- Stable Beluga 2.5, 70B language model-- 28% faster inference speed vs. the A100.
- Better Performance, TCO and took less than one day to port the code base

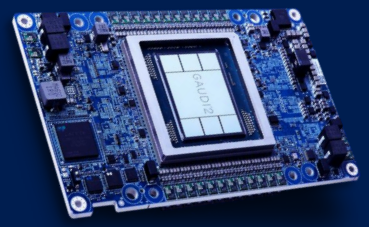
stability.ai

See blog for workloads and configurations. Results may vary
https://stability.ai/news/putting-the-ai-supercomputer-to-work

intel

intel Ai summit

Intel® Gaudi® 2 AI Accelerator Up To 55% Faster Than NVIDIA H100 In Stable Diffusion, 3x Faster Than A100 In AI Benchmark Showdown



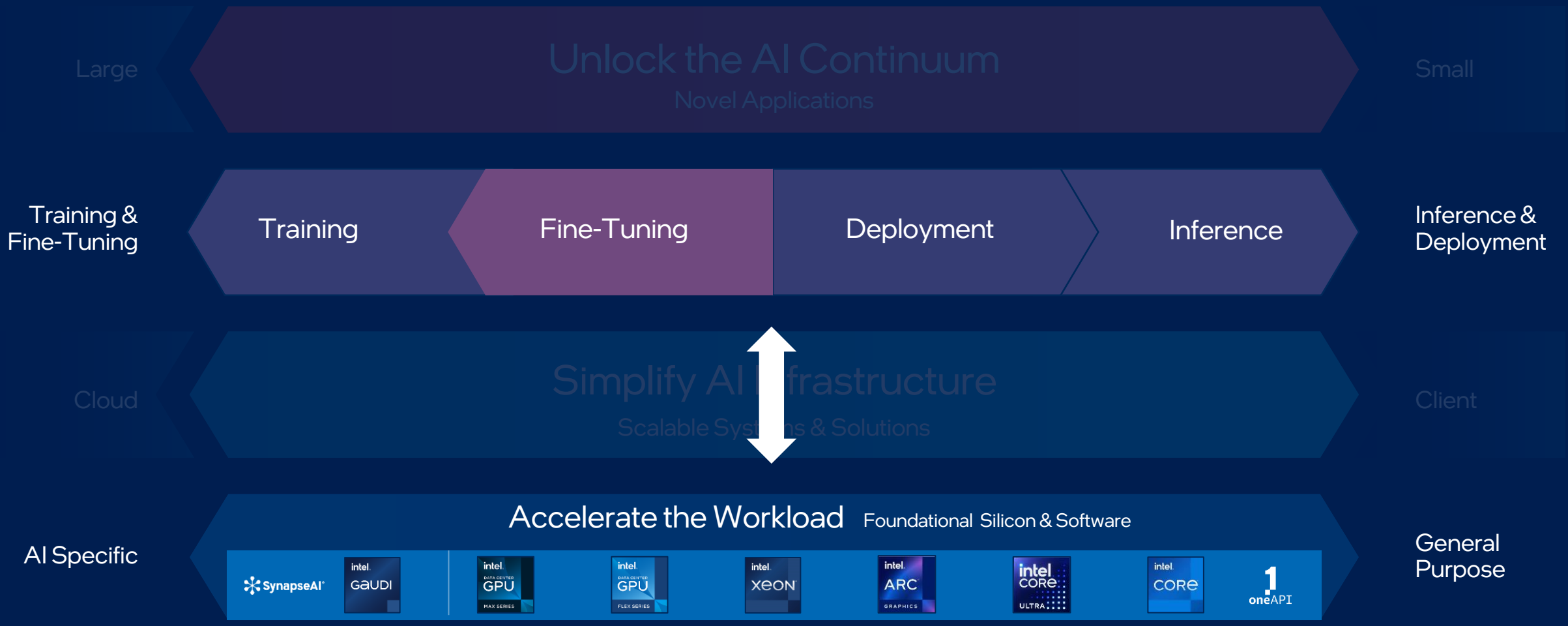
Device	Attention	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec (100-MA)
Gaudi2	FusedSDPA	2	16	32	512	1,254
Gaudi2	FusedSDPA	2	16	16	256	927
H100-80GB	xFormers	2	16	16	256	595
A100-80GB	xFormers	2	16	16	256	381

Device	# Nodes	# Accelerators (total)	Batch Size per Accelerator	Total Batch Size	Images / sec	Images / sec / device
Gaudi2	32	256	16	4096	12,654	49.4
A100-80GB	32	256	16	4096	3,992	15.6

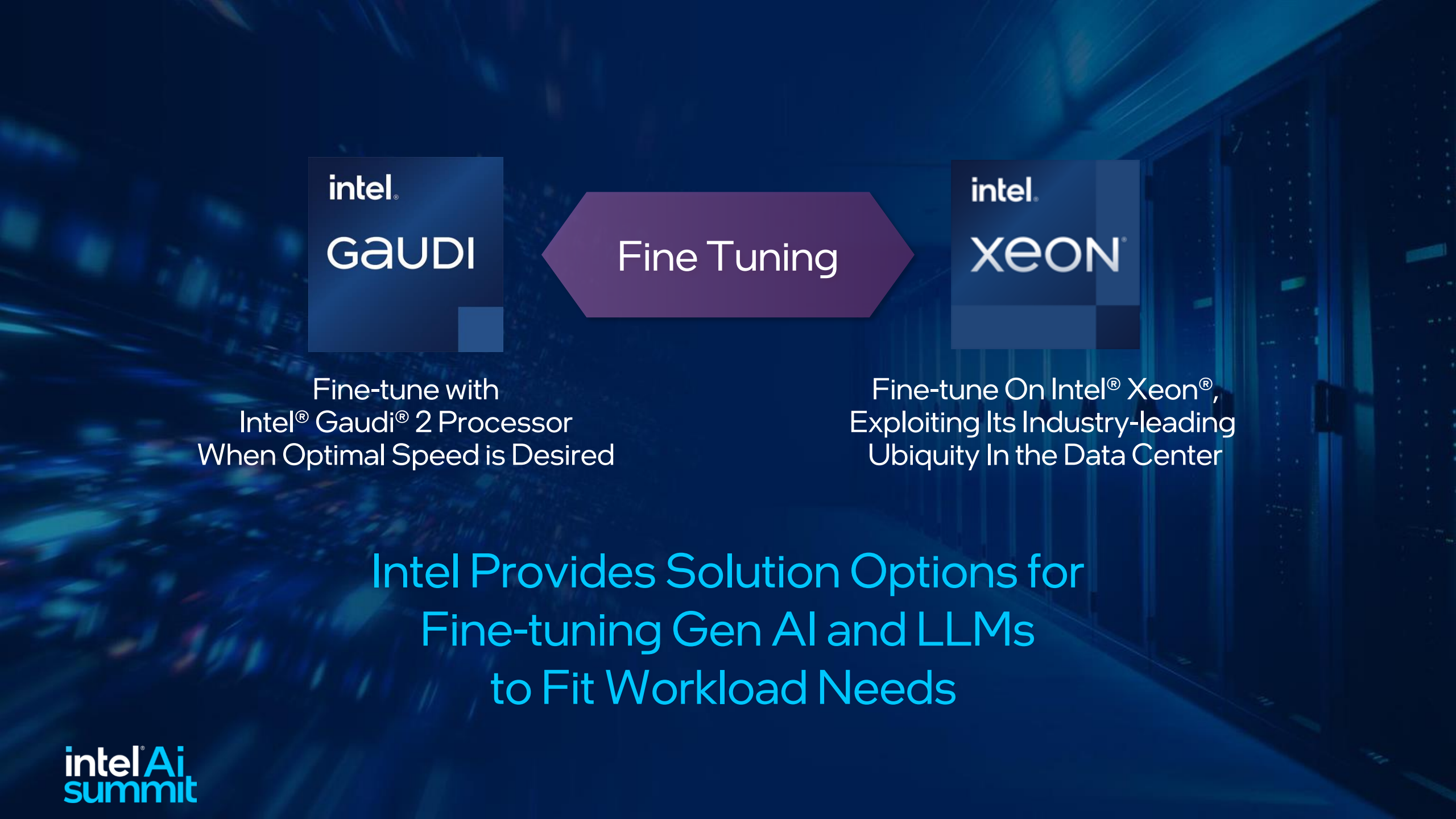
intel AI summit

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

Fine Tuning

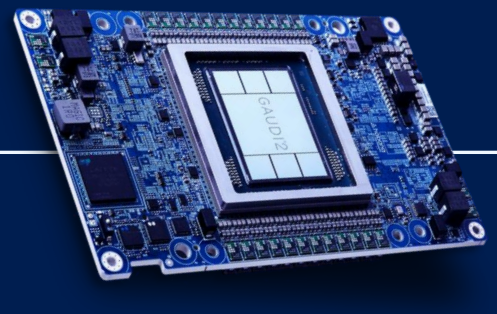


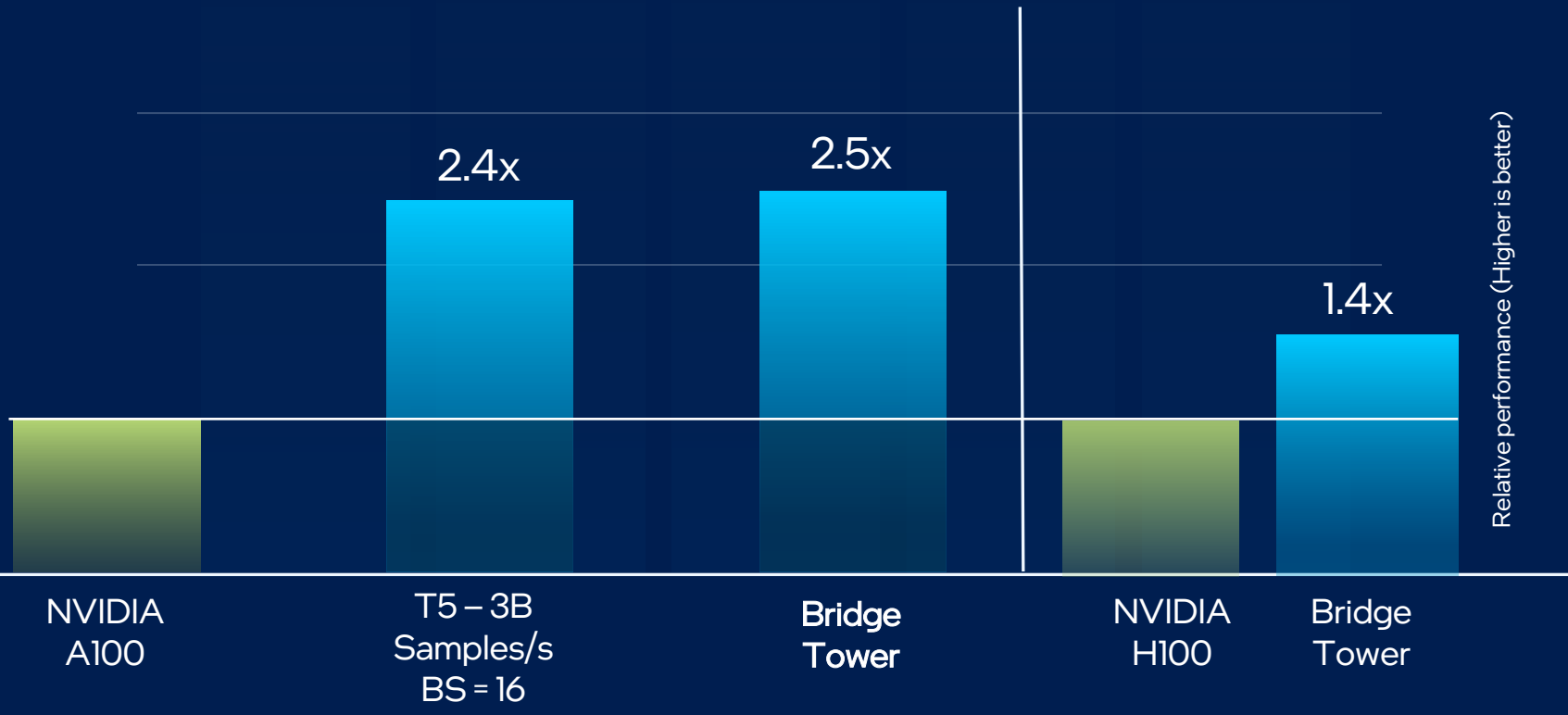
intel AI summit


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 Ai summit

Fine-tuning Across Numerous LLMs

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





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

5th Gen Intel® Xeon® Fine Tuning

Optimized Models & Spaces

Dolly	LLAMA 2	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

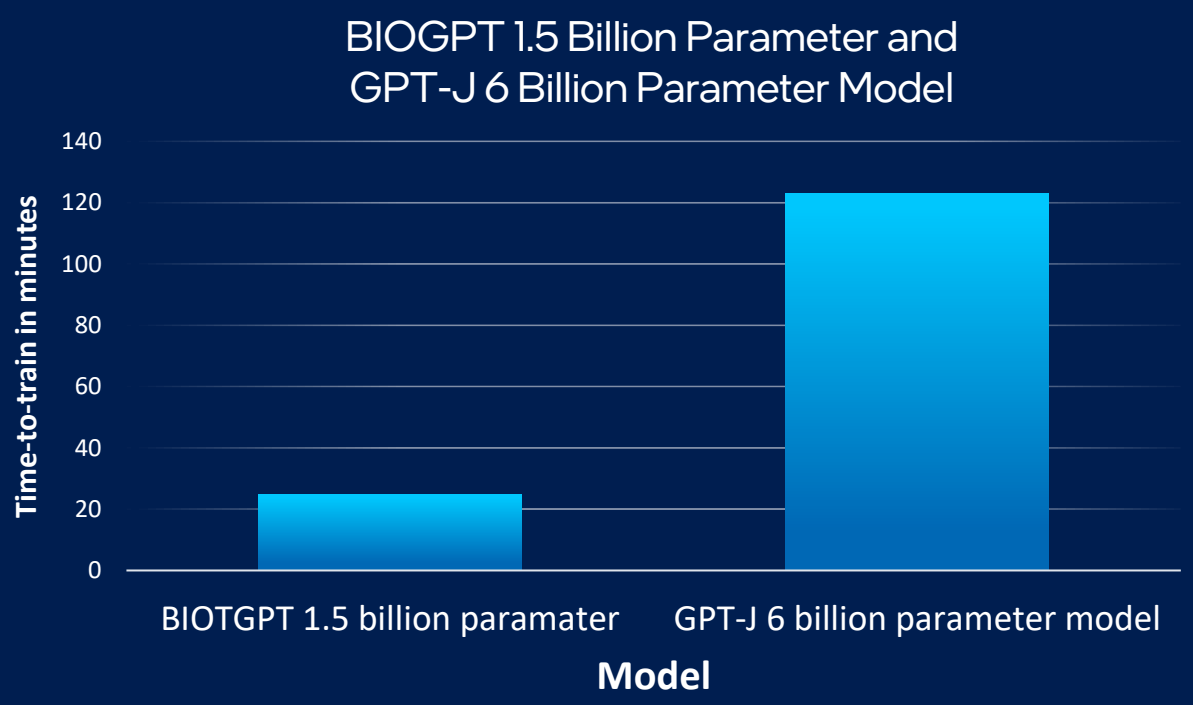
deepspeed + IDEX

ITREX

ONNX RUNTIME

OpenVINO

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



intel Ai summit

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

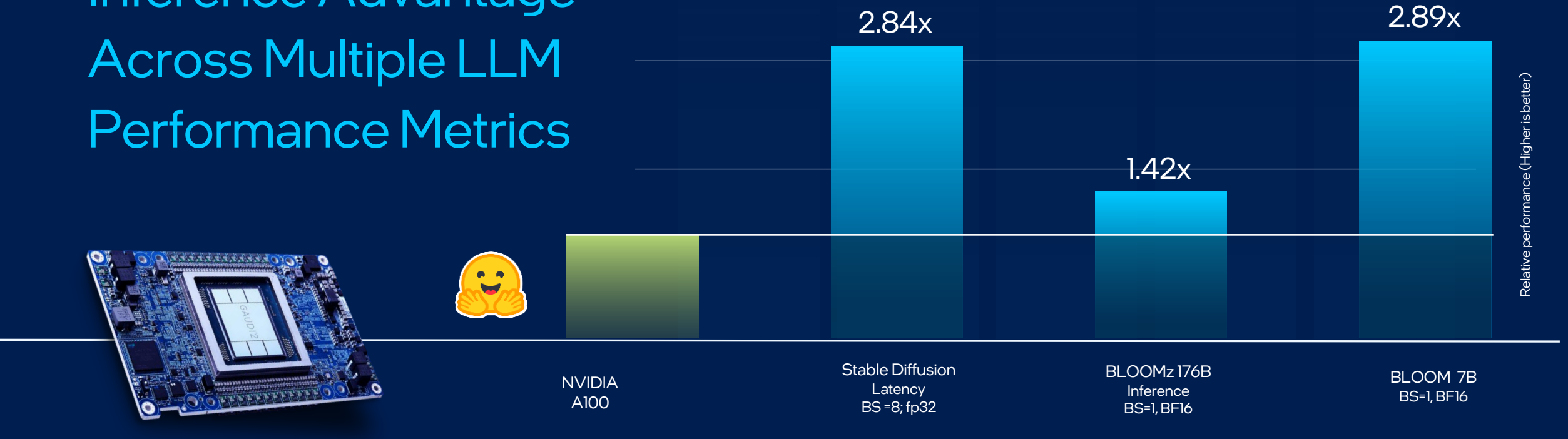
intel Ai summit

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

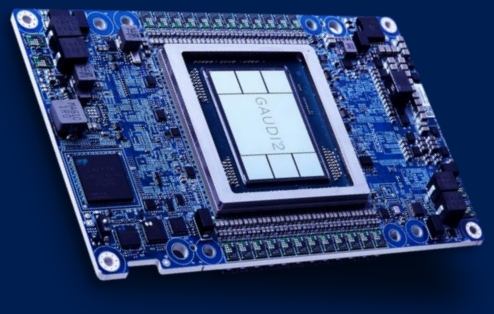


intel Ai summit

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

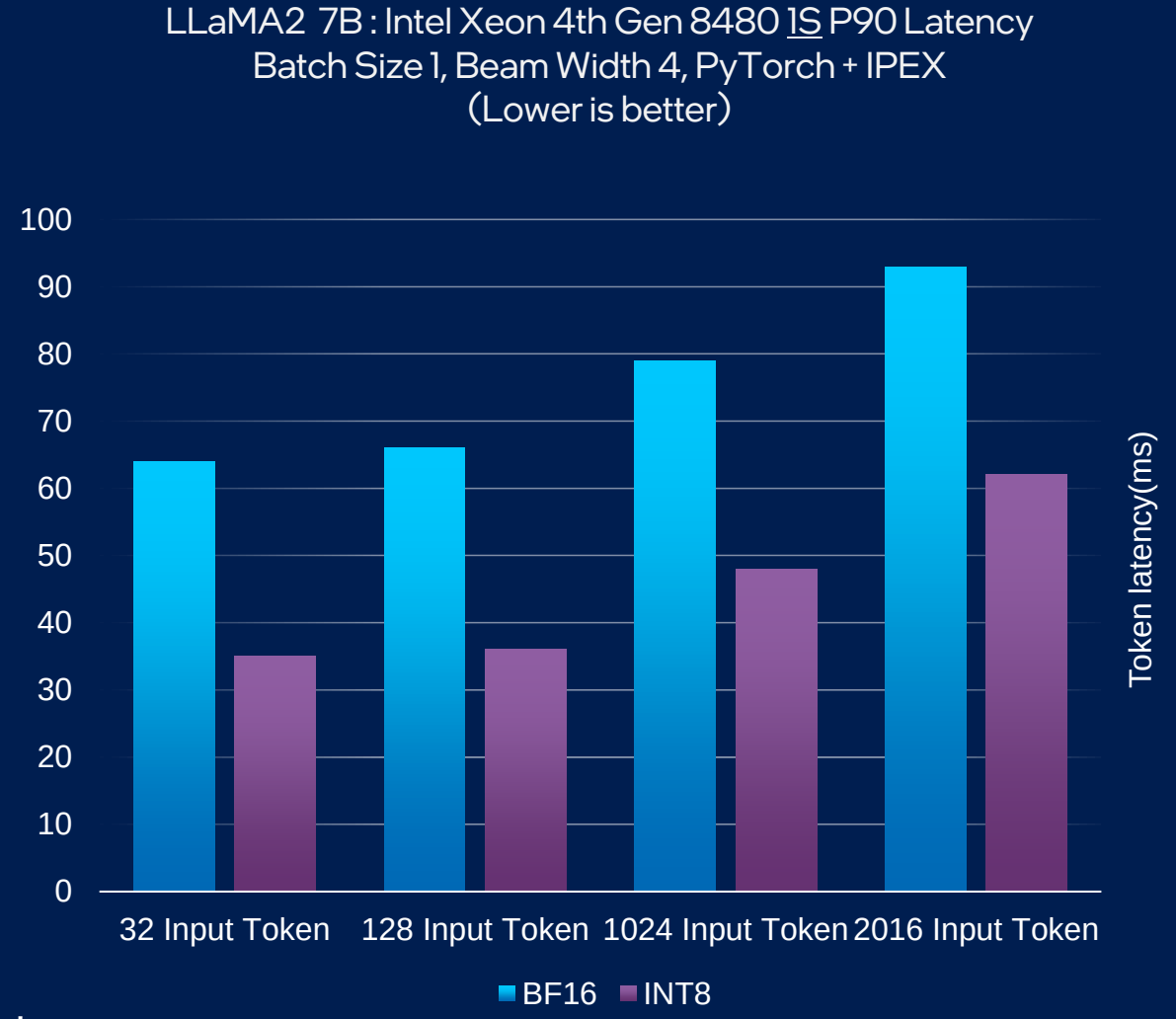
- Near-parity* on GPT-J with H100
- Outperformed A100 by 2.4x (Server) and 2x (Offline)
- Achieved 99.9% accuracy with FP8

GPT-J On MLPerf Inference Benchmark

intel Ai summit

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

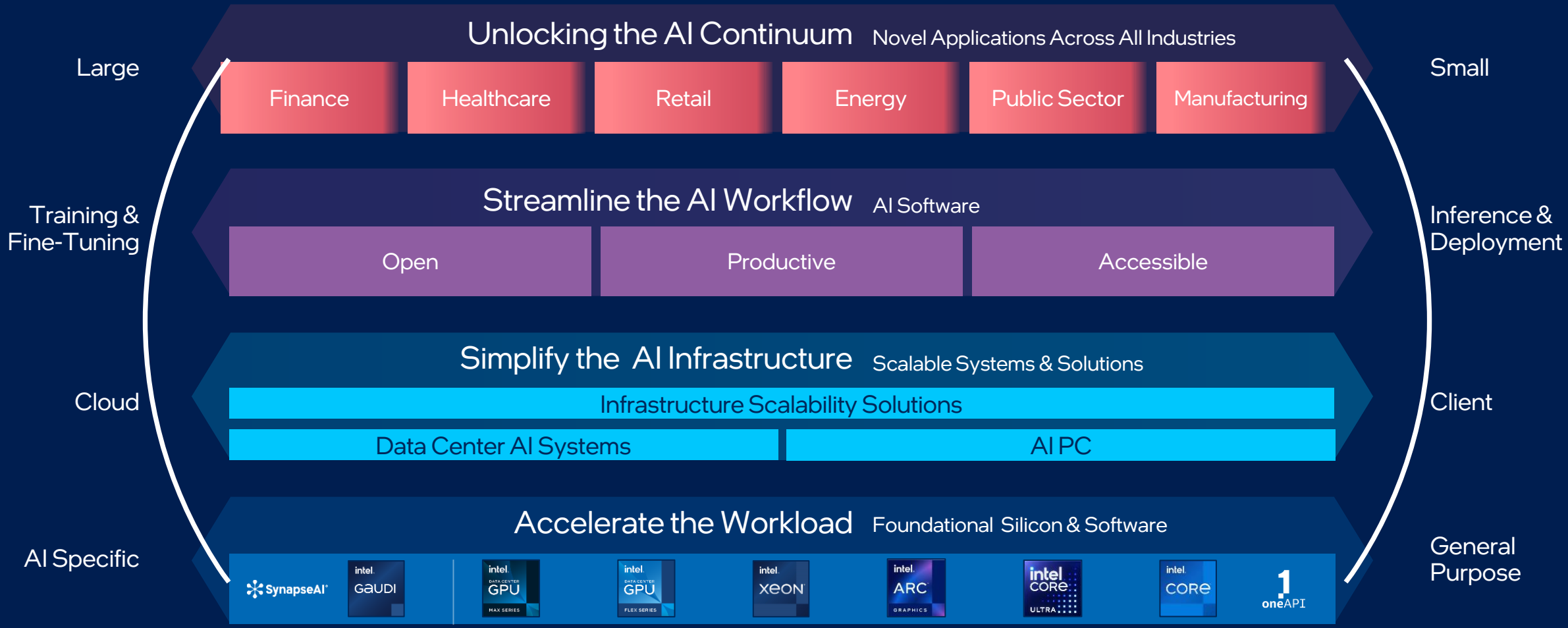
LLaMA2 (7B) Inference with 4th Gen Intel® Xeon® Processors



- 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

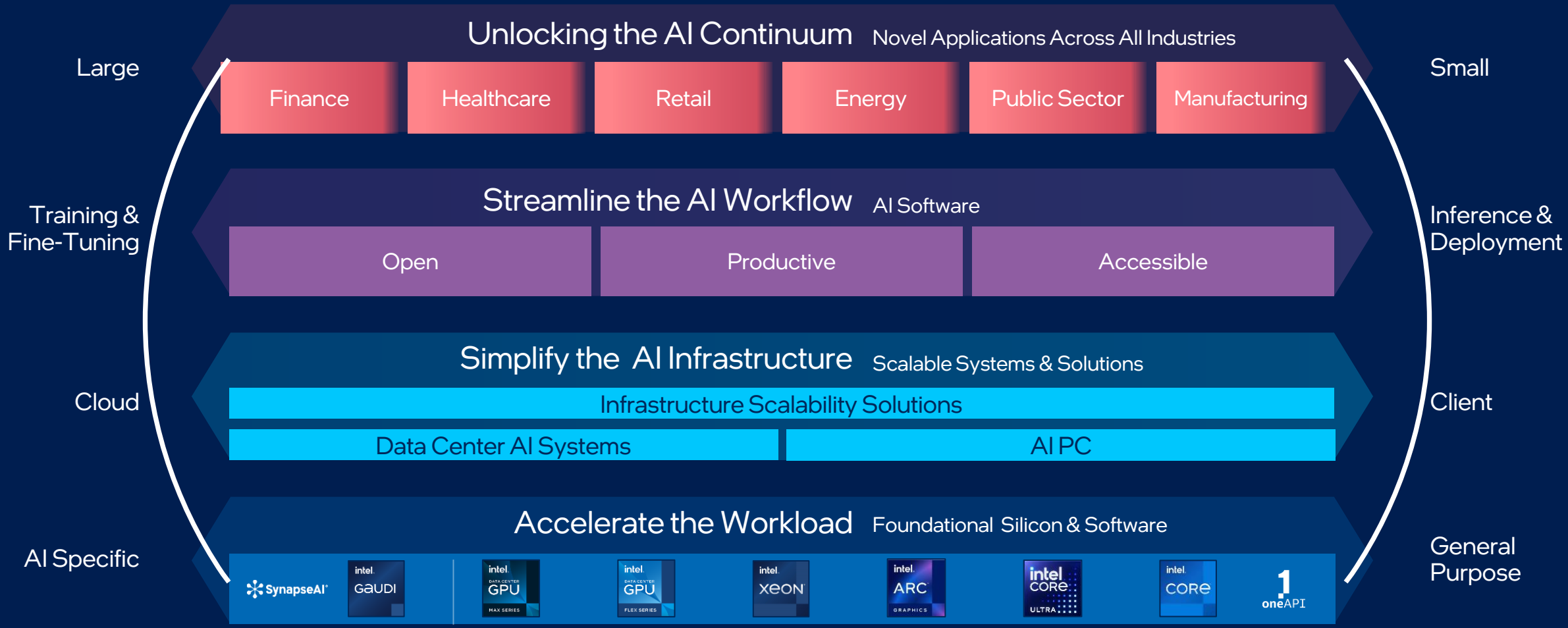
Results shown for bare metal.

Intel's Approach



intel AI summit

Intel's Approach



intel AI summit

您是否有興趣申請 Intel® Developer Cloud 優惠券？

掃描右側 QR Code
將能獲得一組使用優惠代碼
搶先體驗 Intel® 的硬體和軟體雲端 AI 服務

價值美金 $250，有效期限至 2024 年 7 月 30 日

Notices and Disclaimers

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

Thank You!