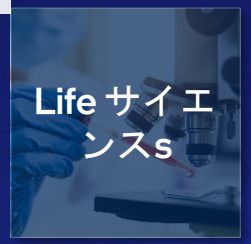
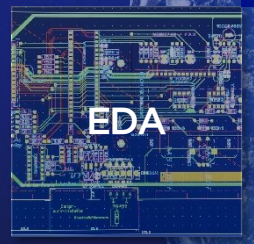
世界の難問を紐解く鍵となる

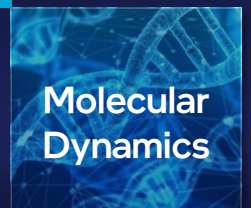
AIで加速するHPC

矢澤　克巳
インテル株式会社　インダストリー事業本部

シミュレーションへの飽くなき要求



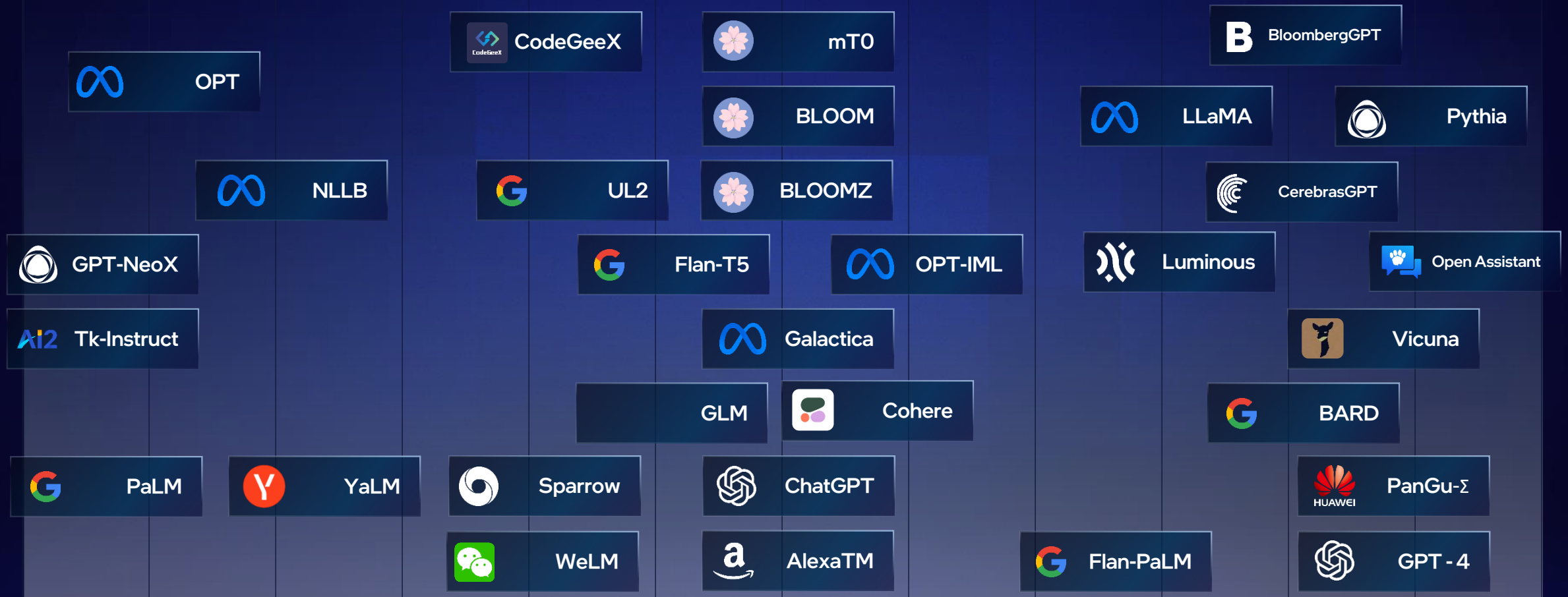




この一年の間にいろいろなことが起こりました…

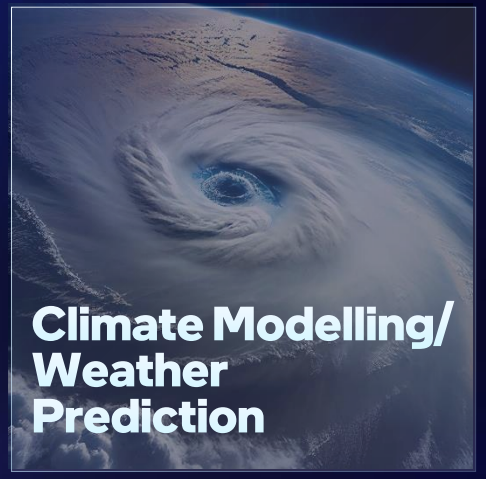
Source: Moore, S. (2022), “Nvidia’s Next GPU Shows That Transformers Are Transforming AI”, IEEE Spectrum.

大規模言語モデル(LLM)の爆発的急伸



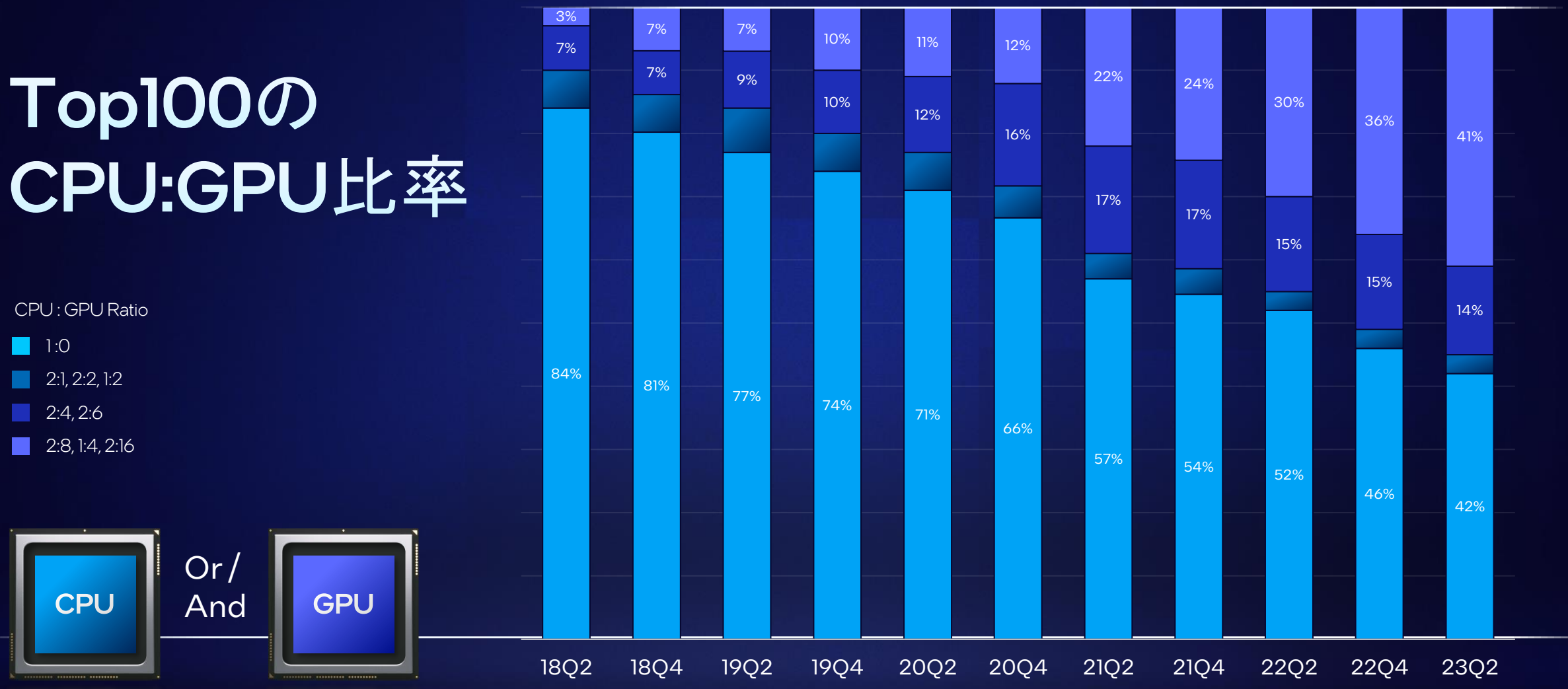
Source: Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

科学分野に
AIブームが到来...



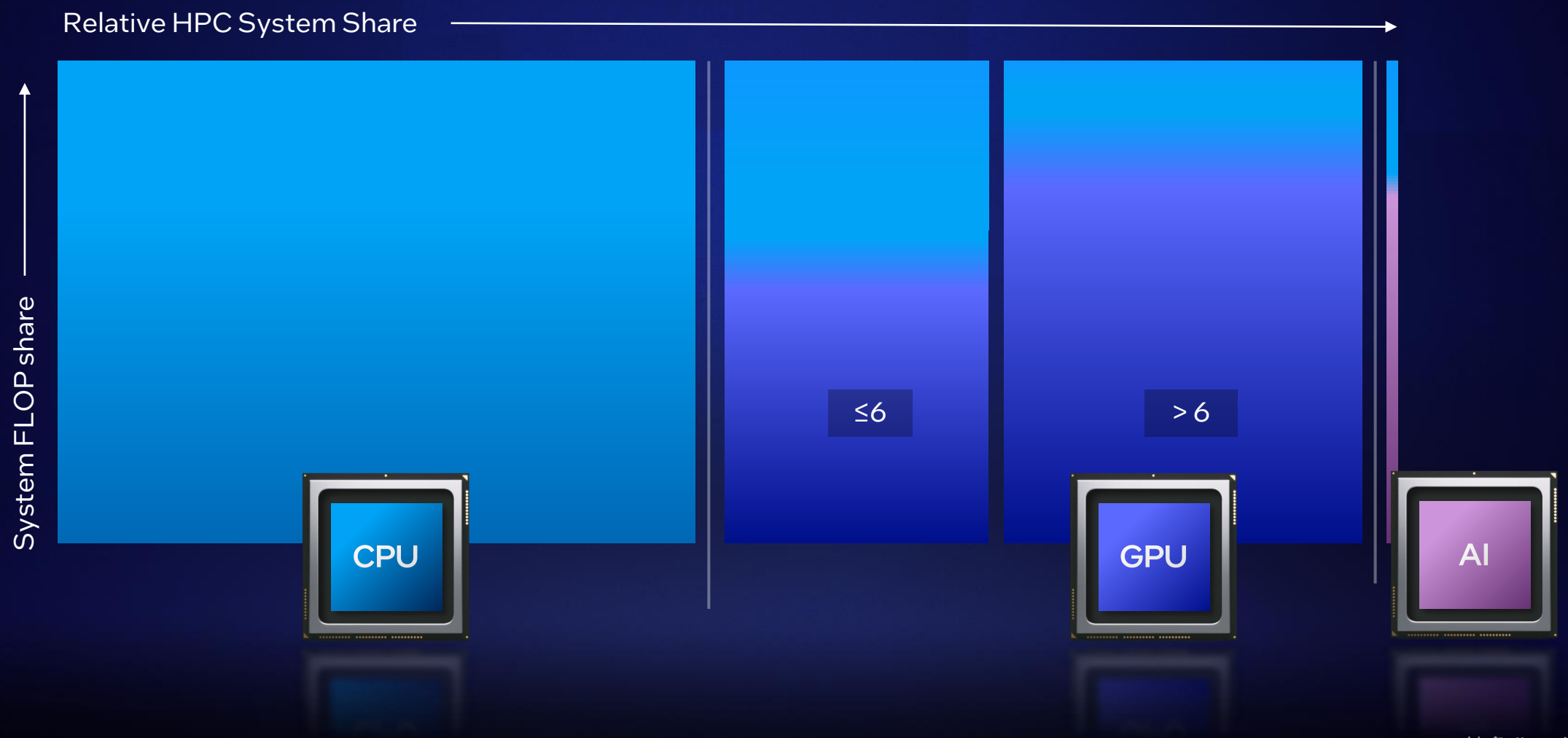
intel

Top100の CPU:GPU比率



Source: Intel analysis based on https://top500.org/lists/top500

HPC利用でのCPU:アクセラレータ比率*

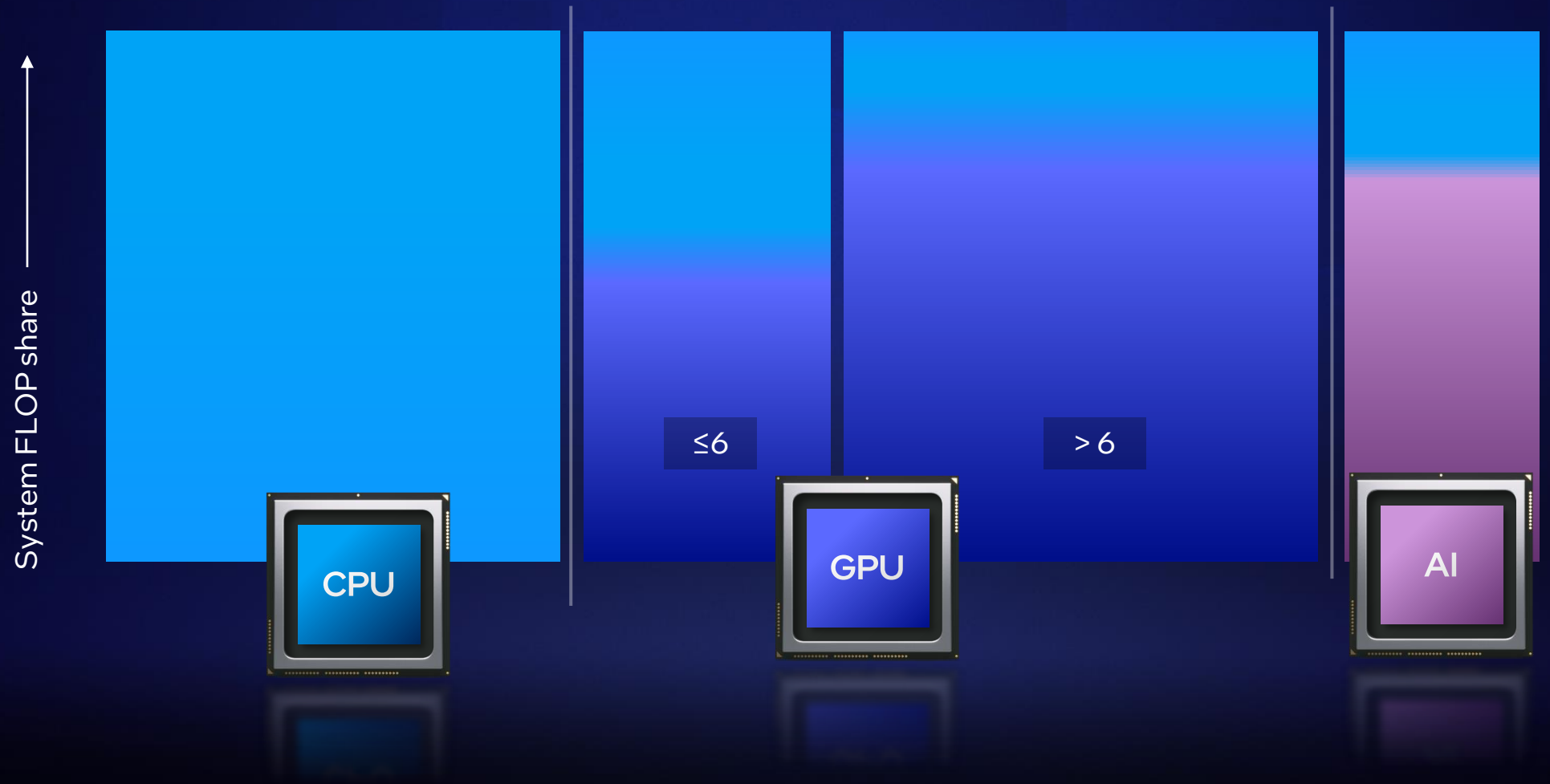


* 抽象化した図

大規模AI利用でのCPU:アクセラレータ比率**

**市場動向から見たコンセプト

HPCとAIの融合をどう実現させるか?



世界でもっとも困難な技術的課題の解決に向け

業界と協力して提供する
サステナブルなHPC

アーキテクチャやベンダーを問わず
オープンかつ統一されたソフトウェア

AIで加速するHPCの未来を支える
計算エンジン

intel

2023年1月10日（日本時間11日）正式発表

第4世代 インテル® Xeon®
スケーラブル・プロセッサー

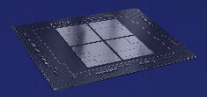
インテル® Xeon® CPU
マックス・シリーズ

インテル® データセンター GPU マックス・シリーズ

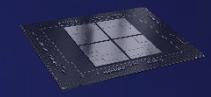
intel

インテル　データセンター　ロードマップ

E-Core CPU

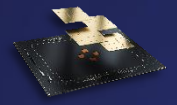


Intel® Xeon® Processor
コードネーム Sierra Forest



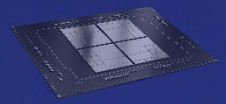
Intel® Xeon® Processor
コードネーム Clearwater Forest

P-Core CPU

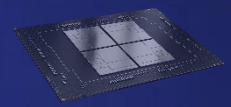


第四世代 Intel® Xeon® Scalable processor

Intel® Xeon® CPU Max Series

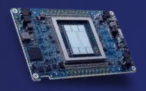


第五世代 Intel® Xeon® Processors
コードネーム Emerald Rapids

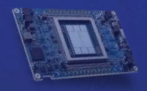


Intel® Xeon® Processors
コードネーム Granite Rapids

AI 専用アクセラレータ

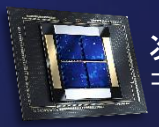


Habana® Gaudi® 2

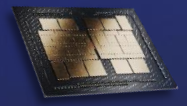


Habana® Gaudi® 3

HPC & AI GPU

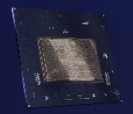


次世代 GPU
コードネーム Falcon Shores

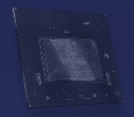


Intel® Data Center GPU Max Series

Visual Cloud GPU

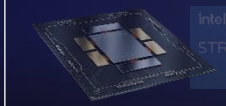


Intel® Data Center GPU Flex Series

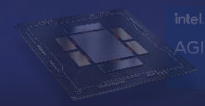


Intel® Data Center GPU Flex Series
コードネーム Melville Sound

FPGA



15+ new FPGAs on schedule to PRQ in 2023



Next Gen FPGAs

2023 — 2025+

intel

ISC High Performance
The HPC Event.

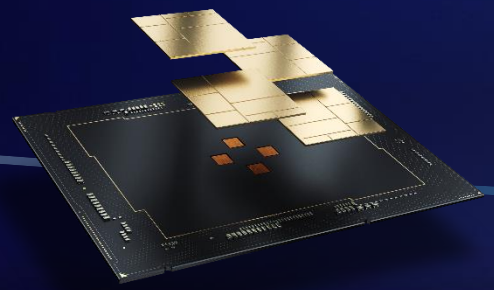
第四世代 Intel® Xeon® Processors

400+ Design Wins
The most ever for any Xeon family

Top 10 Global CSPs* deploying now and throughout 2023

* Cloud service providers

提供中

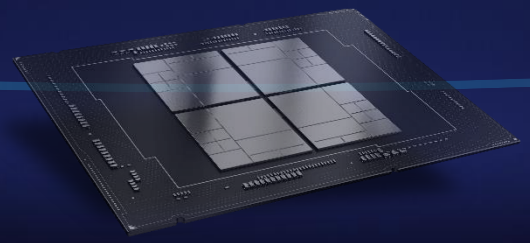


第五世代 Intel® Xeon® Processors
コードネーム Emerald Rapids

より高い消費電力あたりの性能

Same platform as 第四世代 Xeon

Q4 2023予定

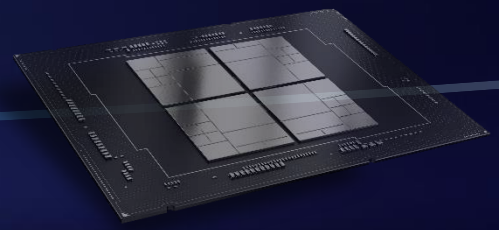


Intel® Xeon® Processors
コードネーム Granite Rapids

Intel 3 で作る初のP-Core Xeon

Increased core density, memory & I/O innovations

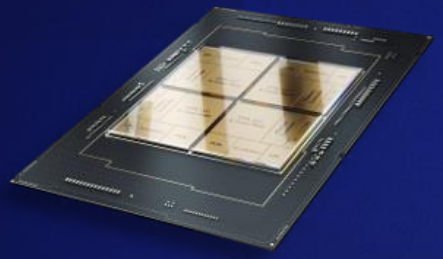
2024予定



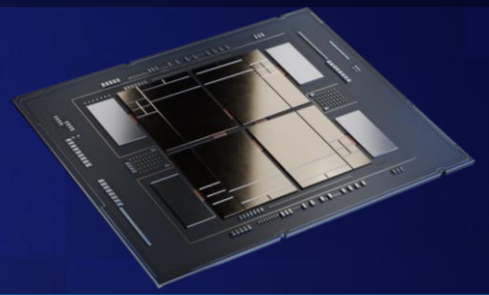
インテル® Xeon® プロセッサー

HPC とAI アクセラレーションに最適化された特徴的な新機能

第4世代インテル® Xeon® スケーラブル・プロセッサー



インテル® Xeon® CPU マックス



ブレークスルー・テクノロジー

DDR5
強化されたメモリバンド幅

PCIE 5
高いスループット

CXL 1.1
次世代 IO

内蔵AI アクセラレーション

インテル® Advanced Matrix Extensions (AMX)
ディープラーニングの推論および学習処理性能を向上

広帯域メモリ

HBM2e
広帯域を必要とするワークロードの性能を飛躍的に向上

インテル® Advanced Matrix Extensions (AMX)

ディープラーニングの推論および学習処理性能を向上

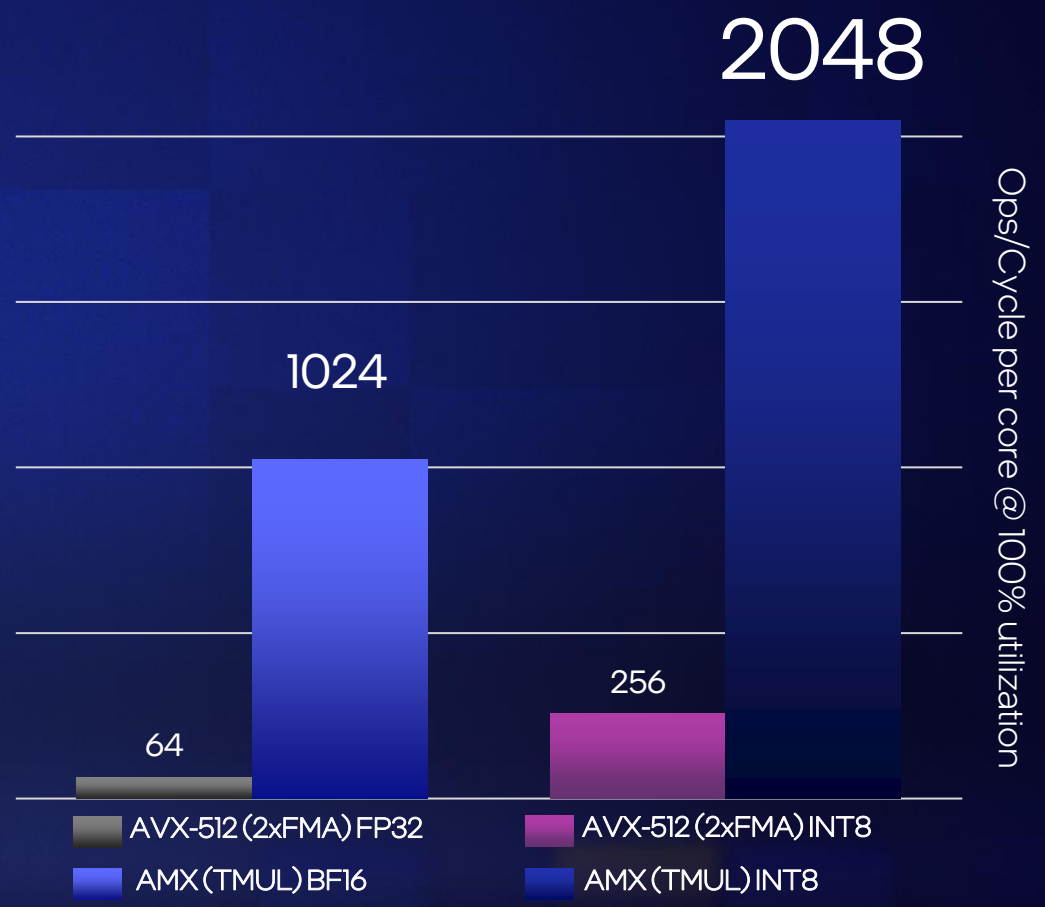
ディープラーニング データータイプ

- int8 with int32 accumulation
- Bfloat16 with IEEE SP accumulation

ISAレベルでの加速

- 完全なインテルアーキテクチャのプログラマビリティ
- 低遅延

業界関連のフレームワークおよびライブラリが利用可能



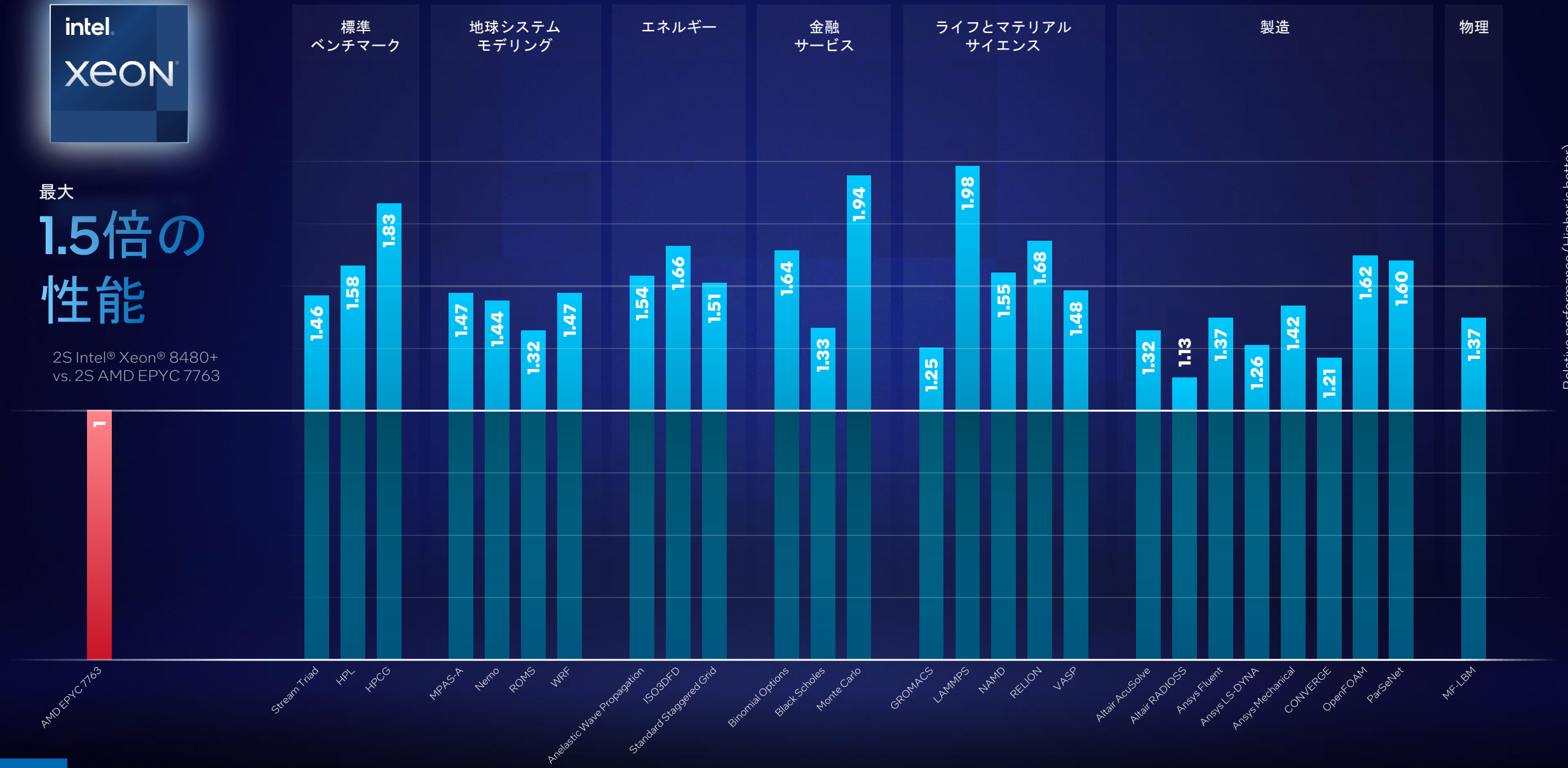
Results have been simulated. For workloads and configurations visit www.intel.com/ArchDay21claims. Results may vary

intel XEON

最大

1.5倍の性能

2S Intel® Xeon® 8480+
vs. 2S AMD EPYC 7763



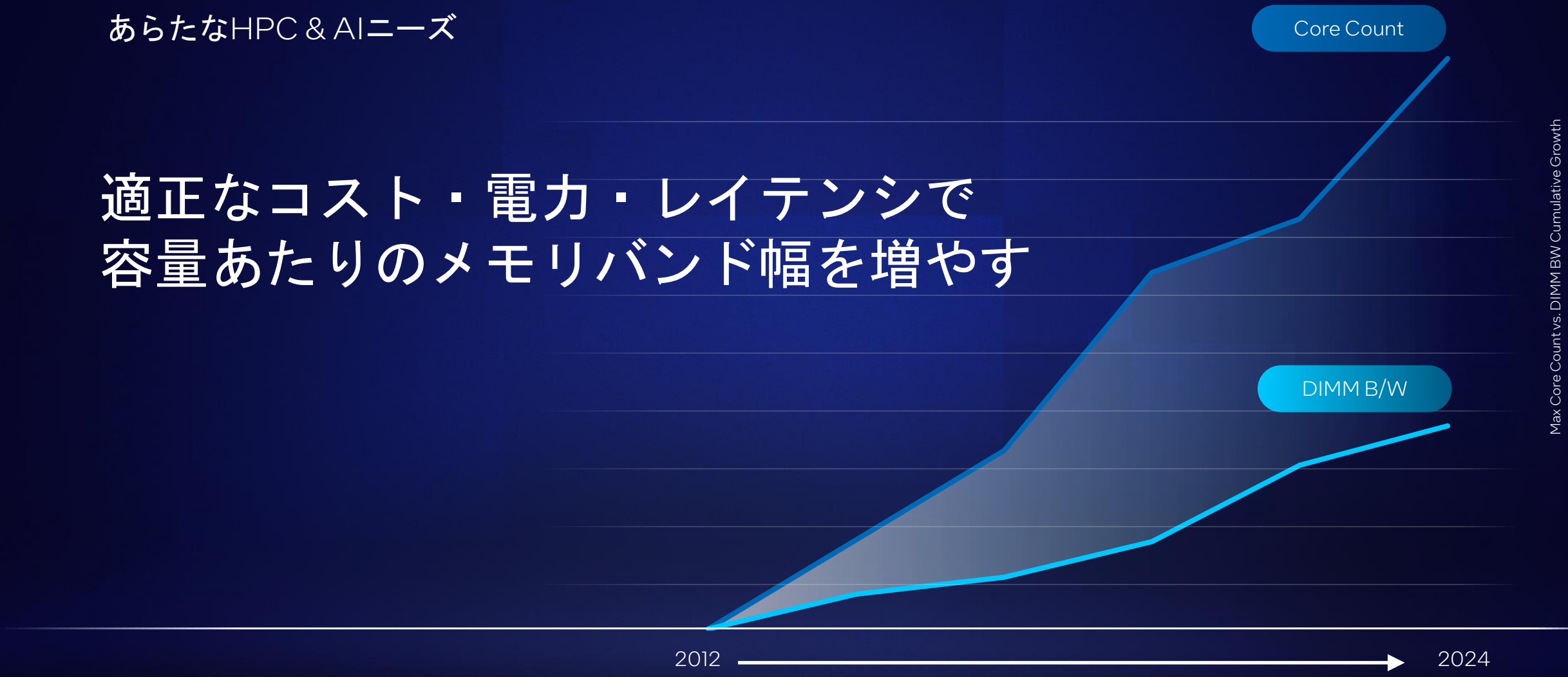
Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via www.openfoam.com, and owner of the OPENFOAM® and OpenCFD® trademark

intel

あらたなHPC & AIニーズ

適正なコスト・電力・レイテンシで
容量あたりのメモリバンド幅を増やす

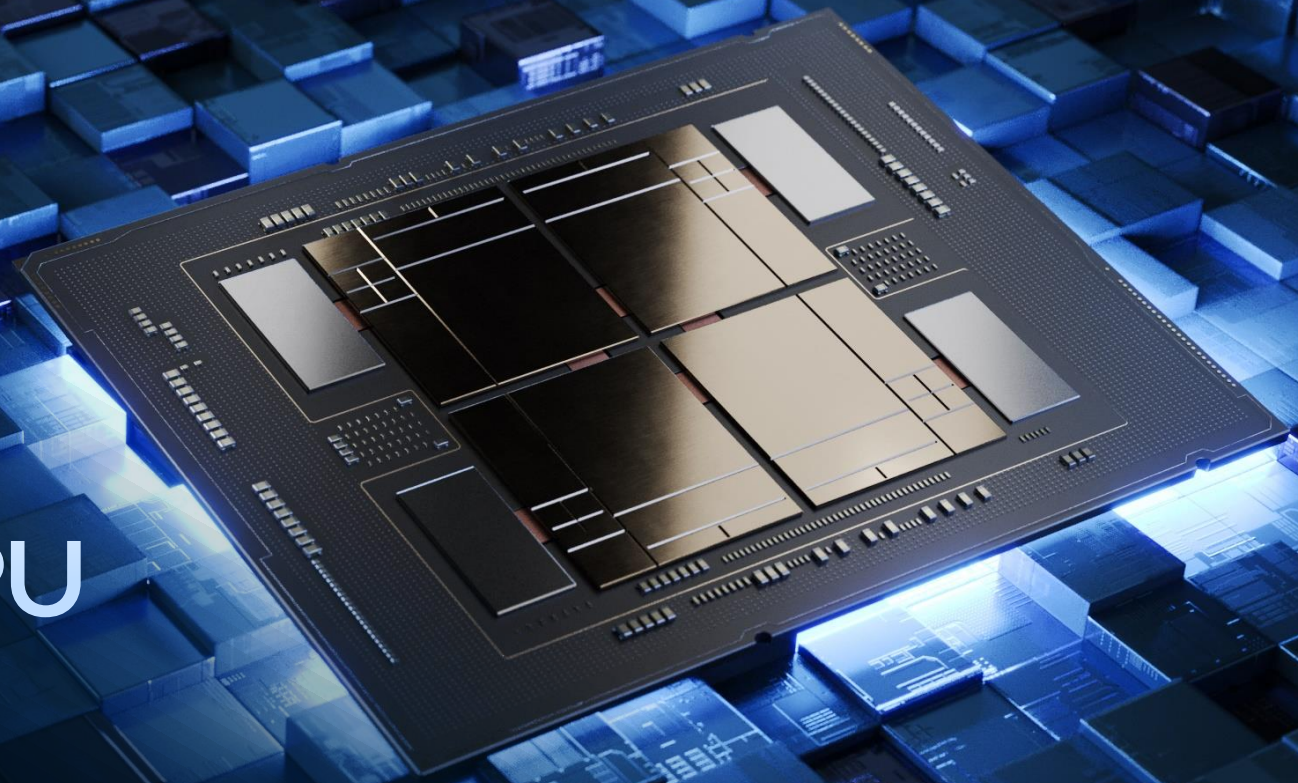


intel

intel XEON MAX SERIES

X86初
HBMを持つ唯一のCPU

最適なメモリ選択肢となりうる



64GB
HBM2e
4 stacks of 16GB

最大
220GF/s
HPCG

最大
2GB
HBM per Core

Memory Modes

HBM Only
Bootable from HBM
No code change

HBM Flat
2 Memory Regions
SW Optimization Needed

HBM Caching
HBM as cache for DDR
No code change

intel

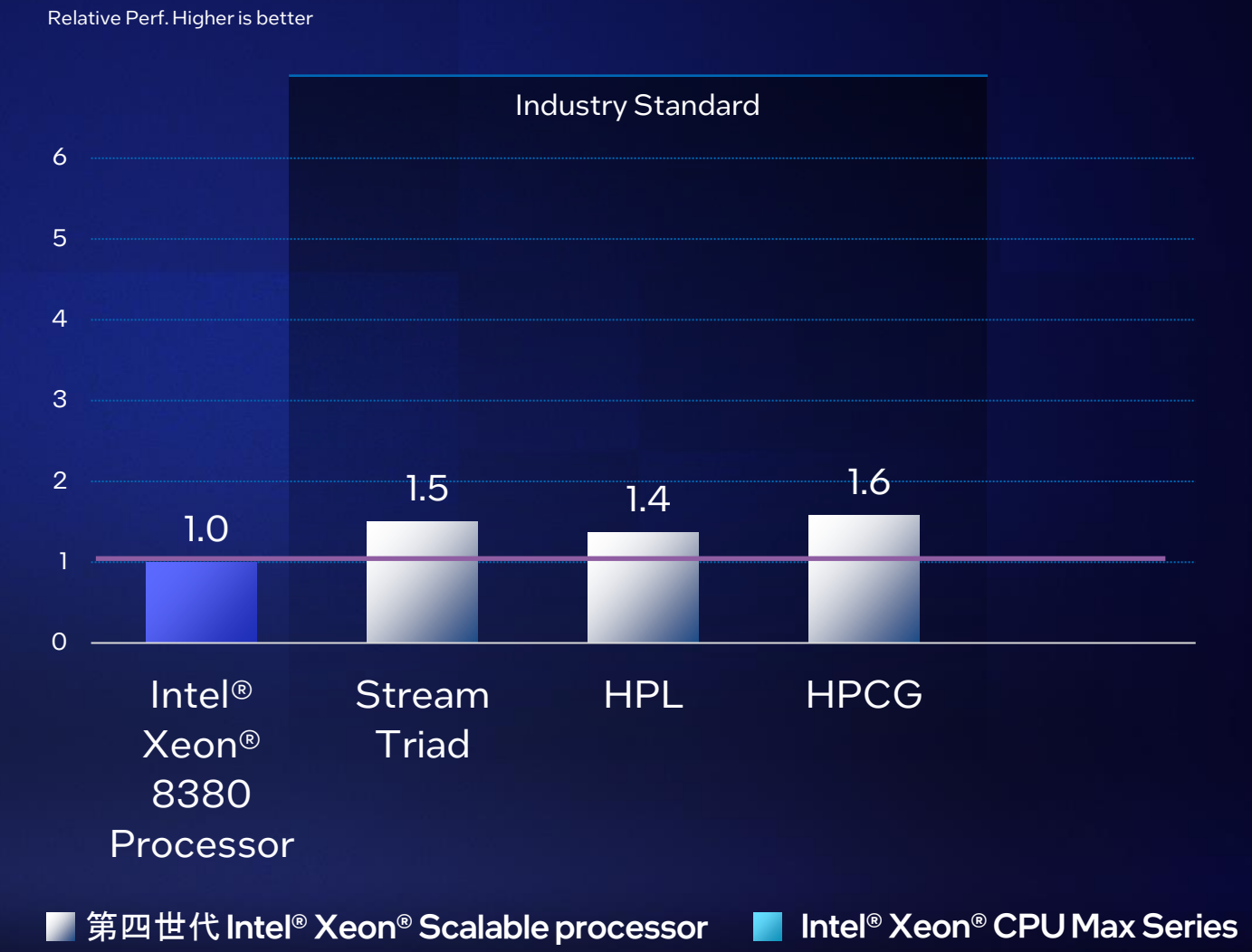
See backup for workloads and configurations. Results may vary.

従来のXeonでは最大

1.6 倍

Performance Industry Standard Benchmarks

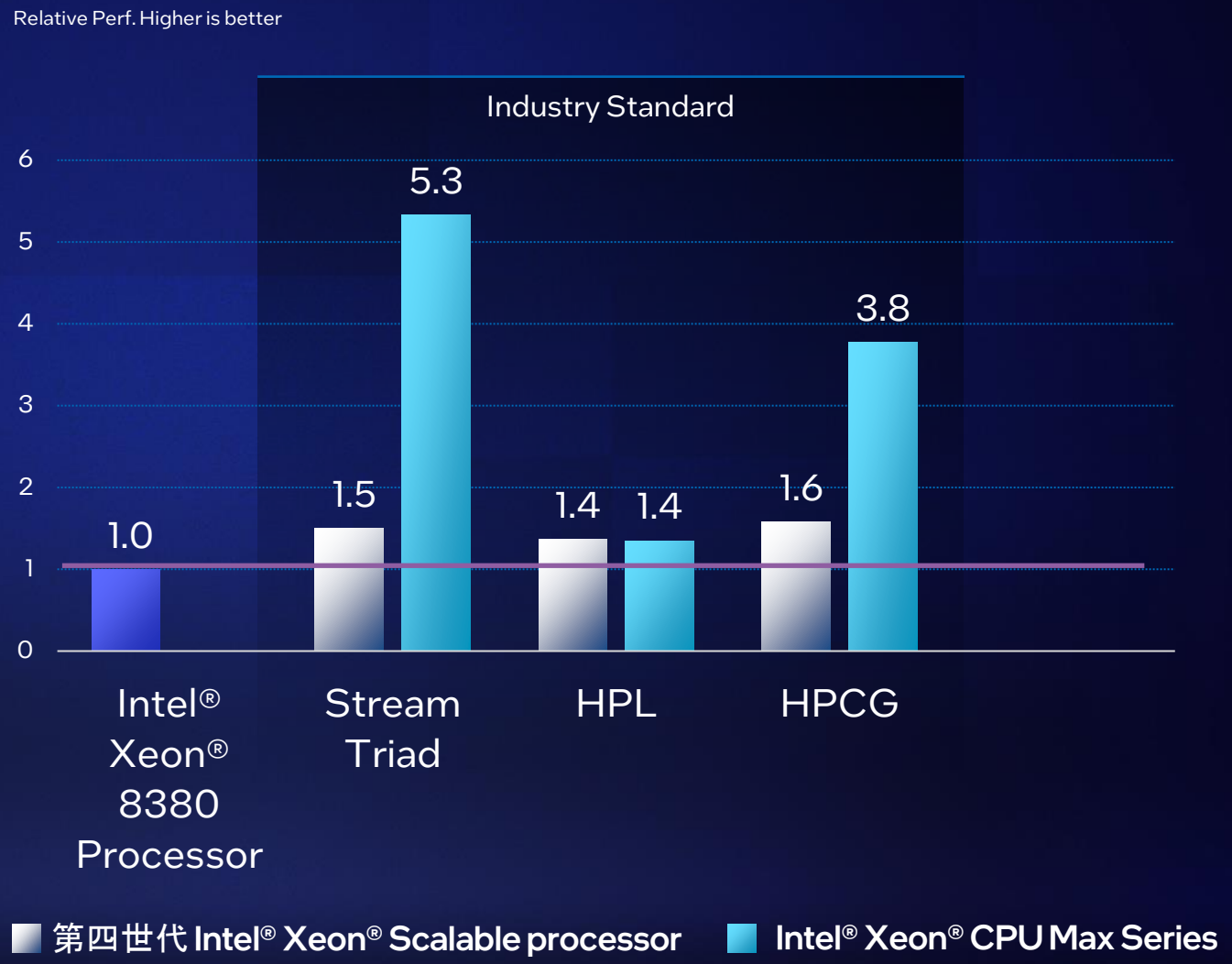
2S 4th Gen Intel® Xeon® processor vs.
2S 第三世代 Intel® Xeon® 8380 processor



See backup for workloads and configurations. Results may vary.

CPU MAXでは最大
5倍
Better Performance for Memory Bandwidth

2S Intel® Xeon® CPU Max Series vs.
2S 第三世代 Intel® Xeon® 8380 processor

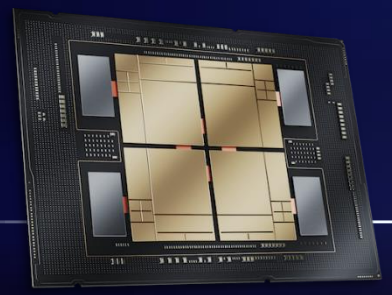


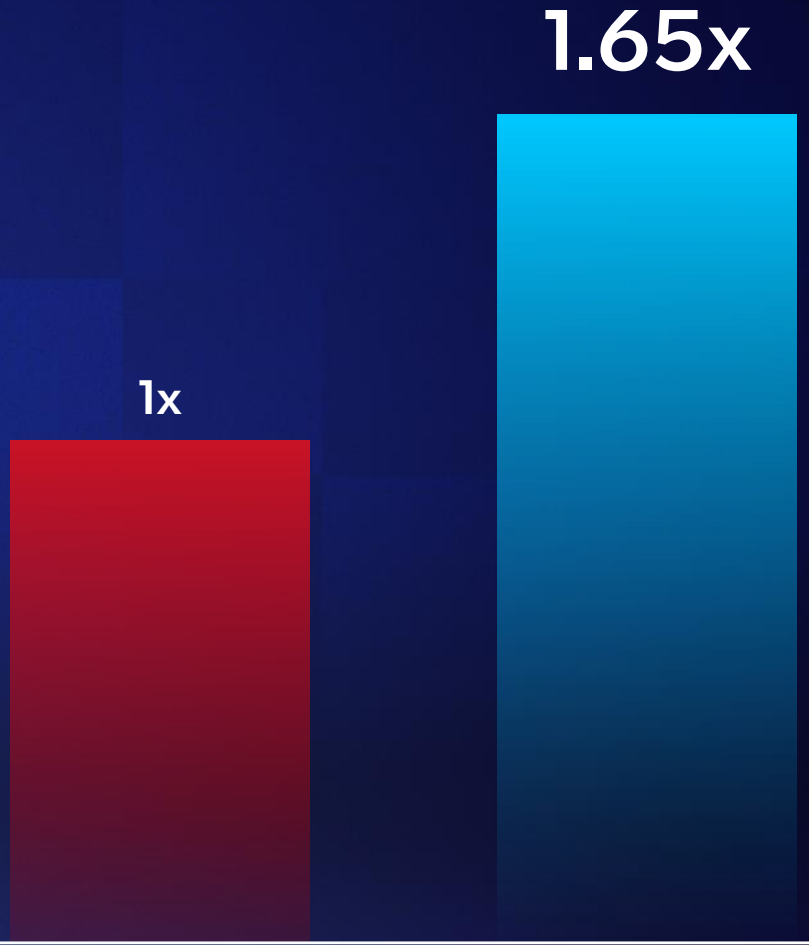
See backup for workloads and configurations. Results may vary.

intel XEON MAX SERIES

コア数差を超えた性能

HPCG Performance





Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

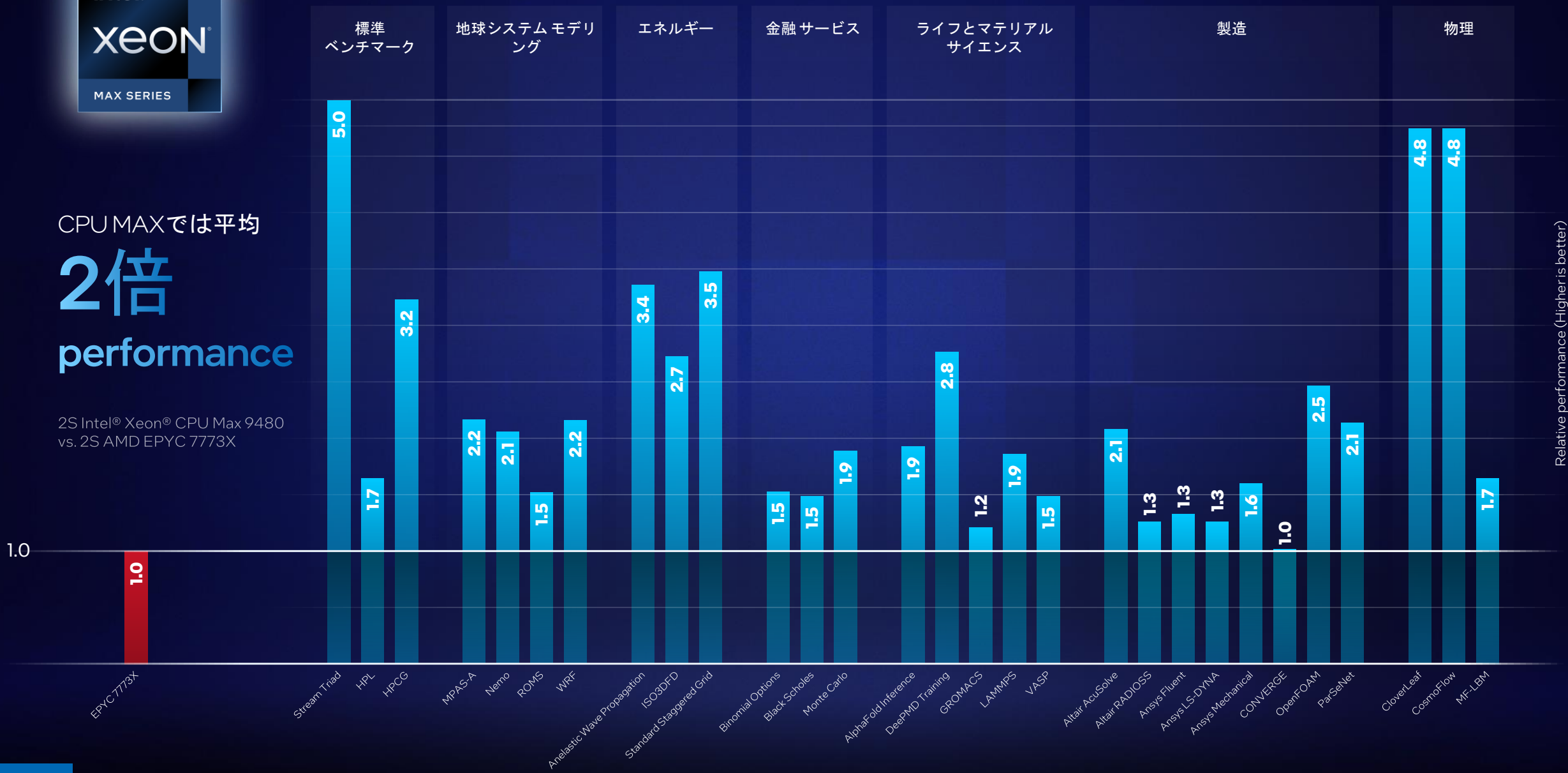
intel

intel XEON MAX SERIES

CPU MAXでは平均

2倍 performance

2S Intel® Xeon® CPU Max 9480
vs. 2S AMD EPYC 7773X



intel

少ないノード数と電力で同等の性能を実現

ALTAIR AcuSolve™

	第三世代 Intel Xeon Scalable Processor		Intel Xeon CPU Max Series
# Node	4	-75%	1
External Memory DIMMS	64		0
Power	1640W	-57%	700W

Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

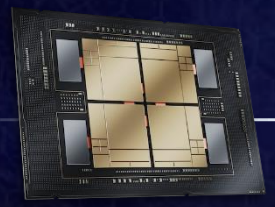
intel XEON MAX SERIES

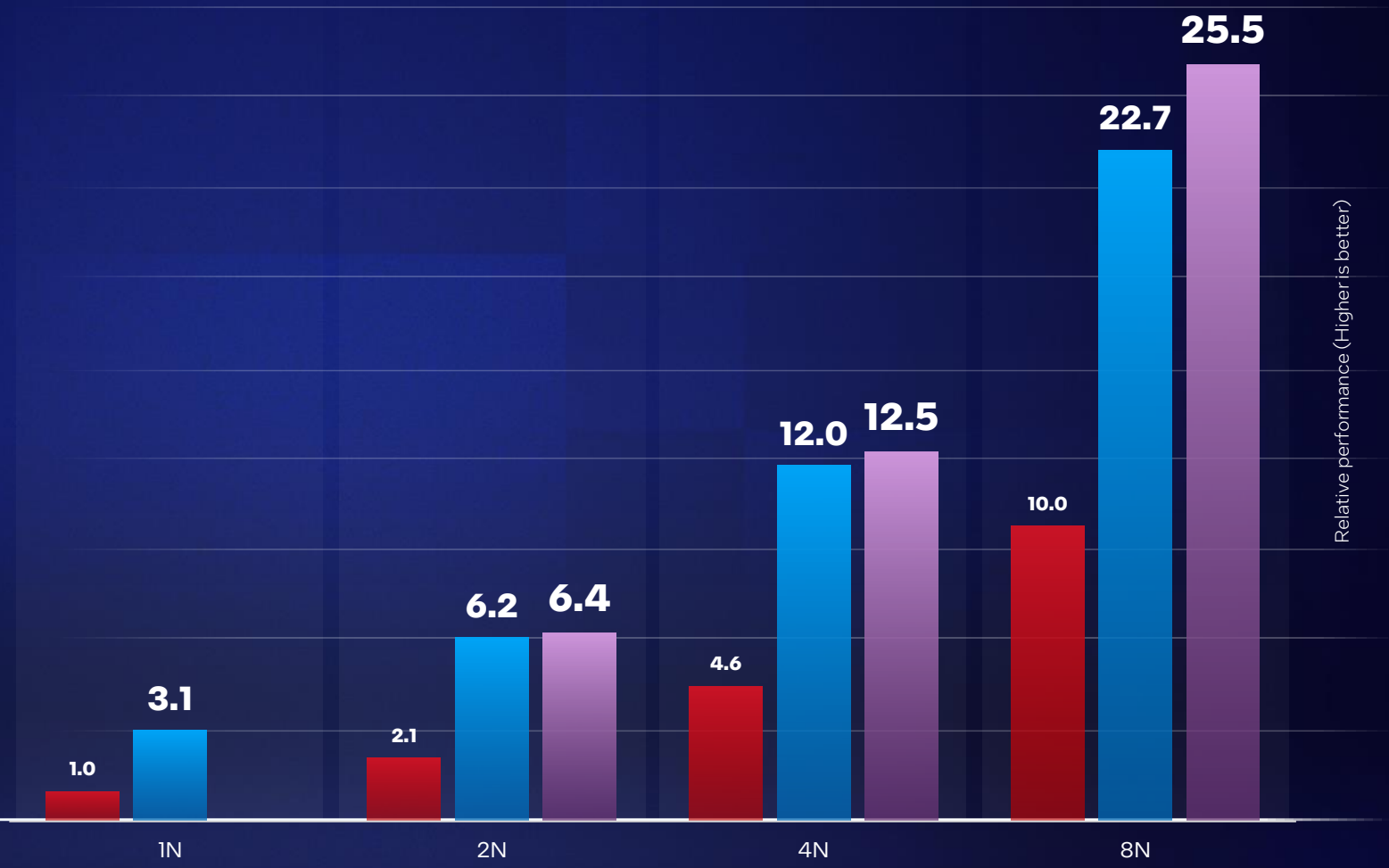
地球システムモデリング

MPAS-A – dyncore 30-km

91% Scaling Efficiency in cache mode

- AMD EPYC 7763
- Intel® Xeon® CPU Max 9480 (Cache Mode)
- Intel® Xeon® CPU Max 9480 (HBM Only)





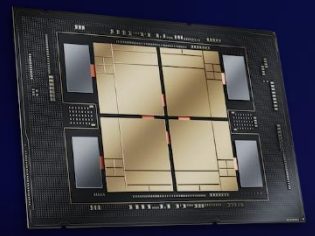
intel ISC High Performance The HPC Event.

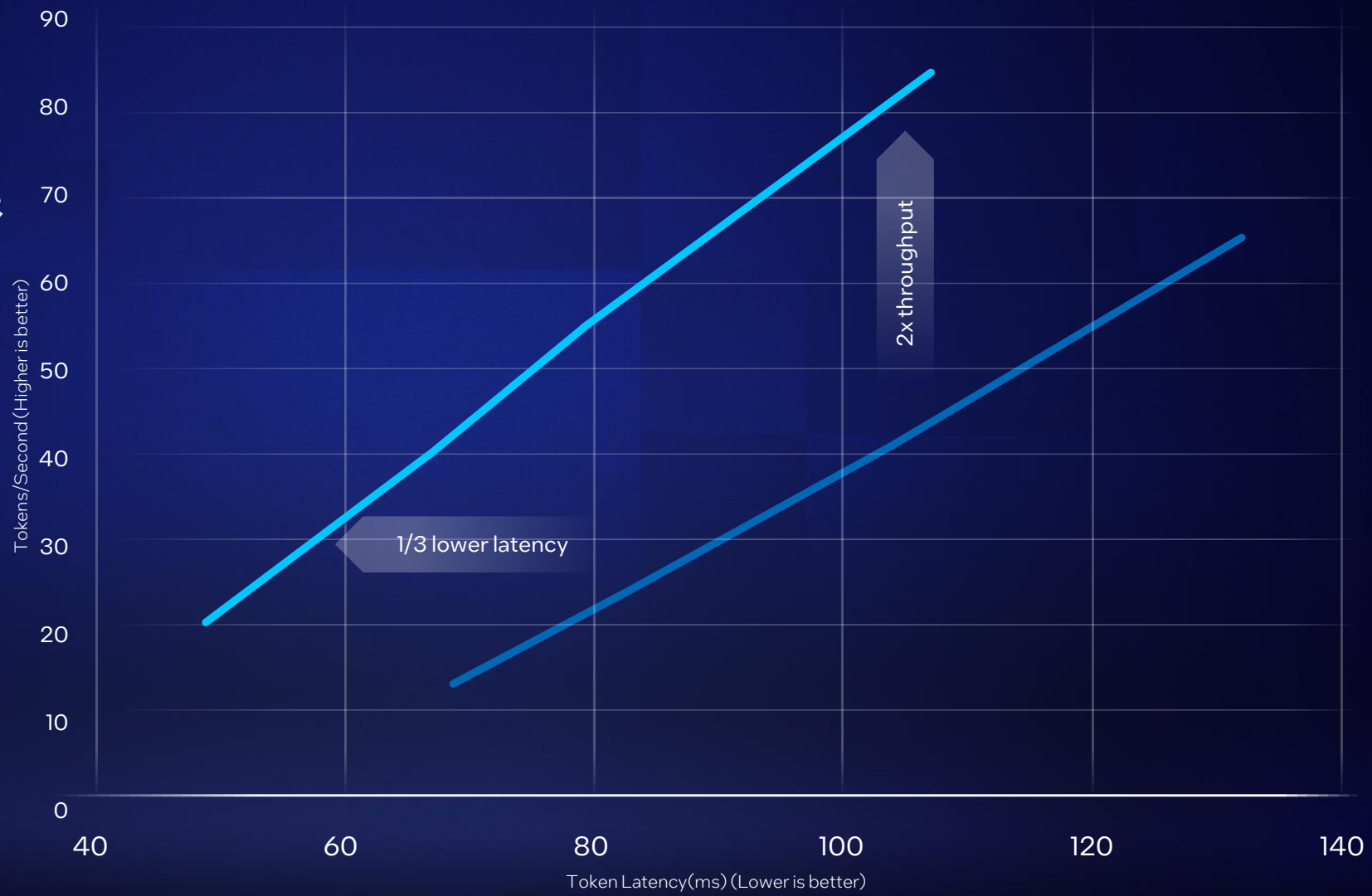
Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

1/3のレイテンシで 2倍の推論 スループット

GPT-J 6B BF16 Inference Perf
(Batch Sizes 1,2,4,8)

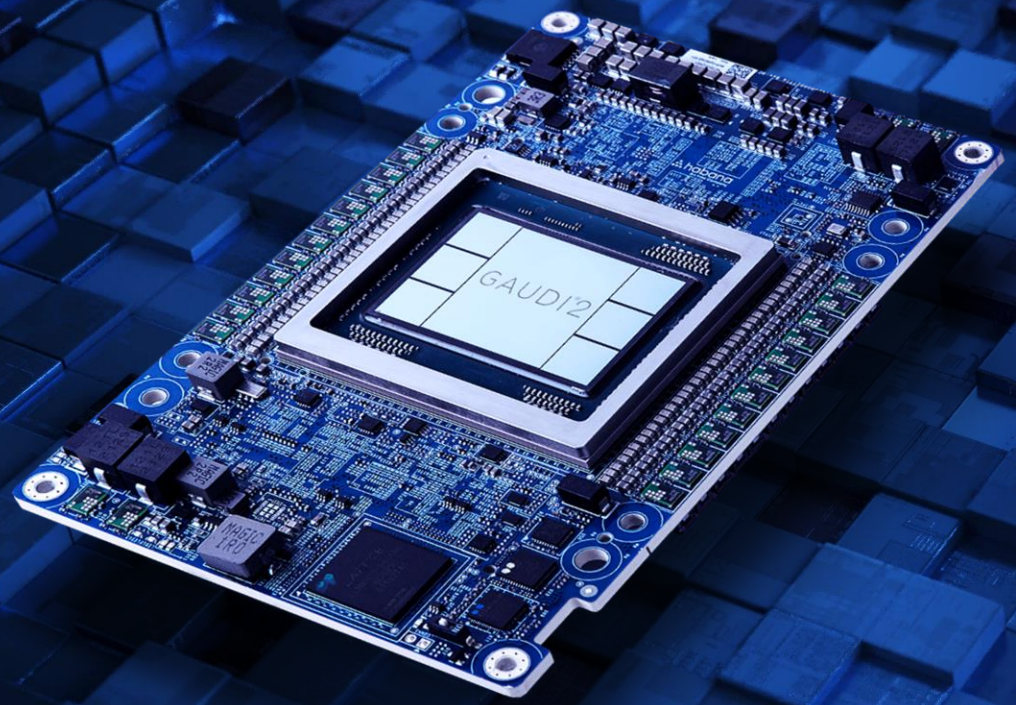
- 1S Intel® Xeon® CPU Max 9480
- 1S Intel® Xeon® CPU 8480





Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

GAUDI®2



7nm Process Technology	**24** Tensor Processor Cores
96 GB On-Board HBM2	**48 MB** SRAM
24 Integrated Ethernet ports	

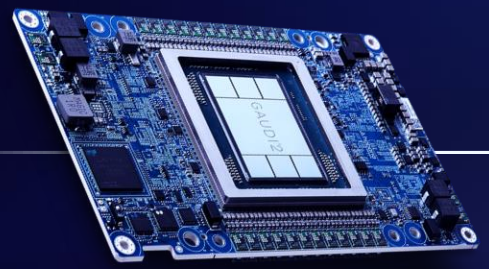
Availability

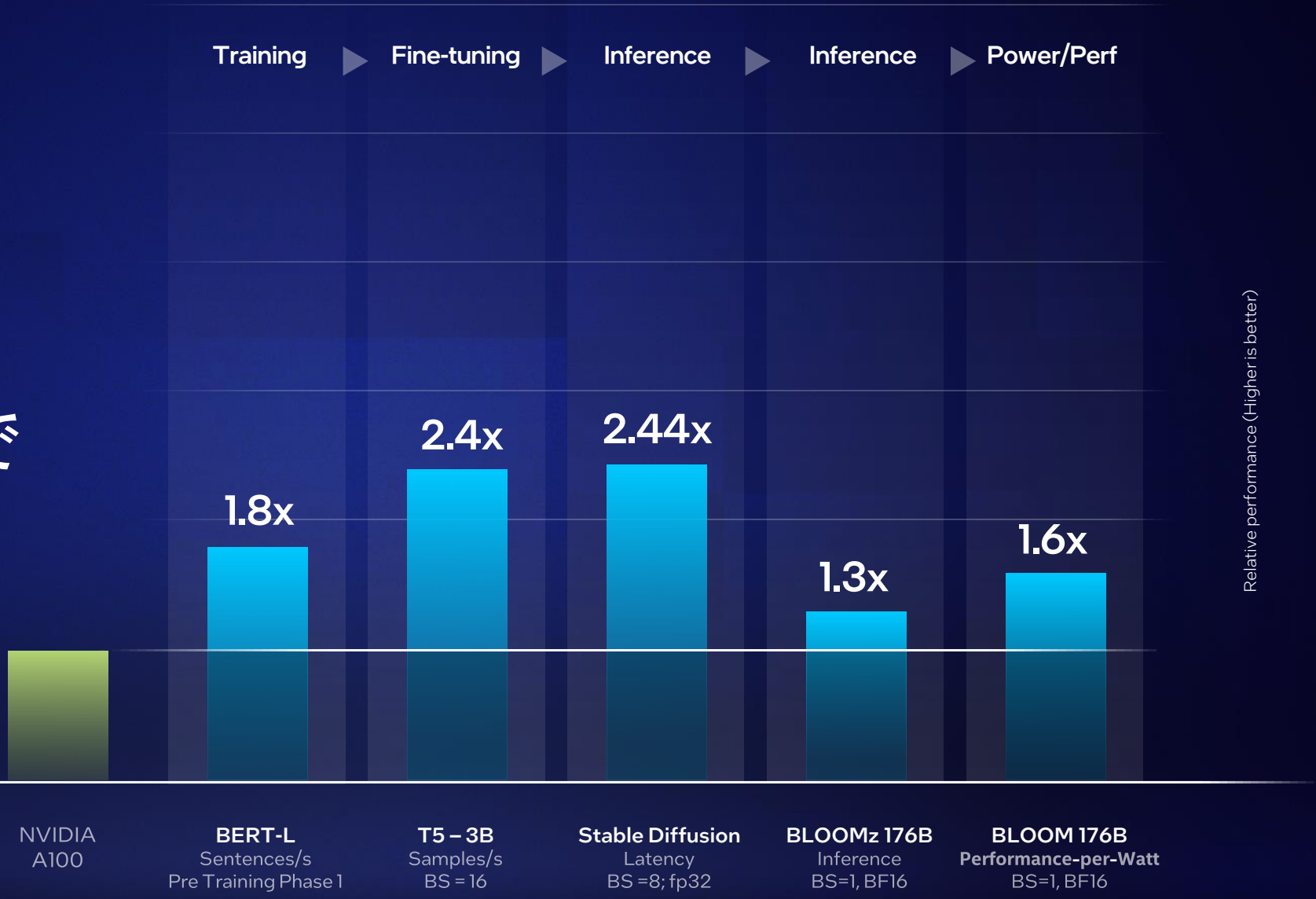
Cloud	Intel Developer Cloud
On-Prem	Supermicro Gaudi2 Server

intel

ISC High Performance
The HPC Event.

数々の性能指標で優位性を示す



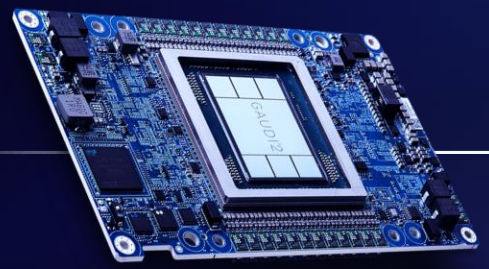


Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

https://huggingface.co/blog/habana-gaudi-2-ベンチマーク , https://habana.ai/habana-claims-validation

大規模で柔軟な統合されたネットワークによるニアなスケーリング

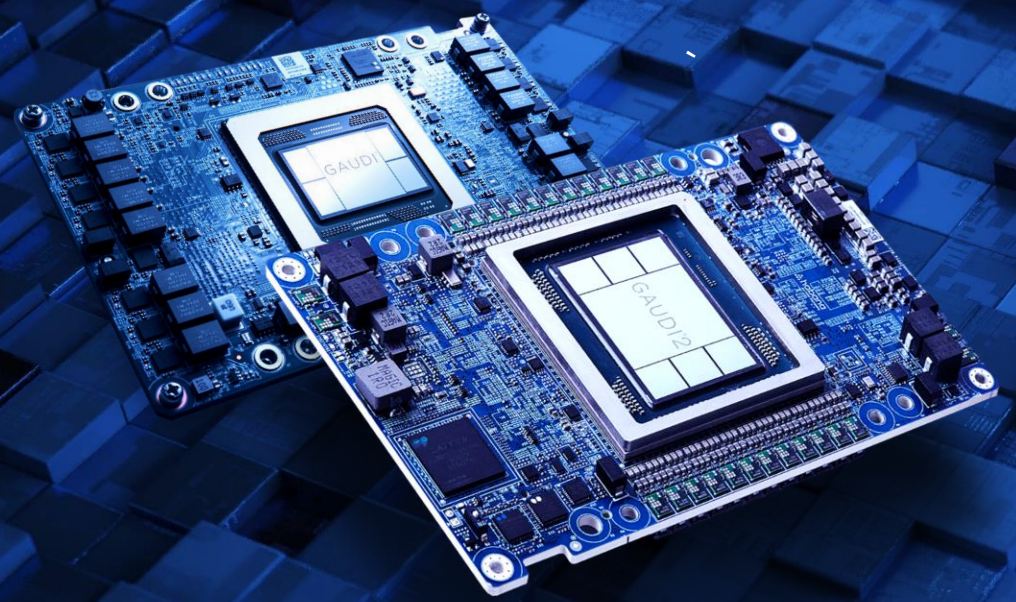
最大64枚のカードまでリニアなスケーリング効率
>99%



Model source: https://github.com/HabanaAI/Model-References/tree/master/PyTorch/generative_models/stable-diffusion-training, Dataset laion2B-en.
Training with BF16, batch size = 16, global batch size = 1024, for 1K iterations. Image size 256x256. Measurements using SynapseAI 1.9.0

intel habana

生成AIに最適



Large-scale
Training on Intel AI supercomputing cluster

Fast inference Speed
Habana Gaudi2 on 176B parameter BLOOM LLM

Real-time multimodal semantic search
Running in the Intel Developer Cloud on Gaudi 2

12B Parameter Dolly 2.0 with GPT-3 like performance
Running on Habana Gaudi 2

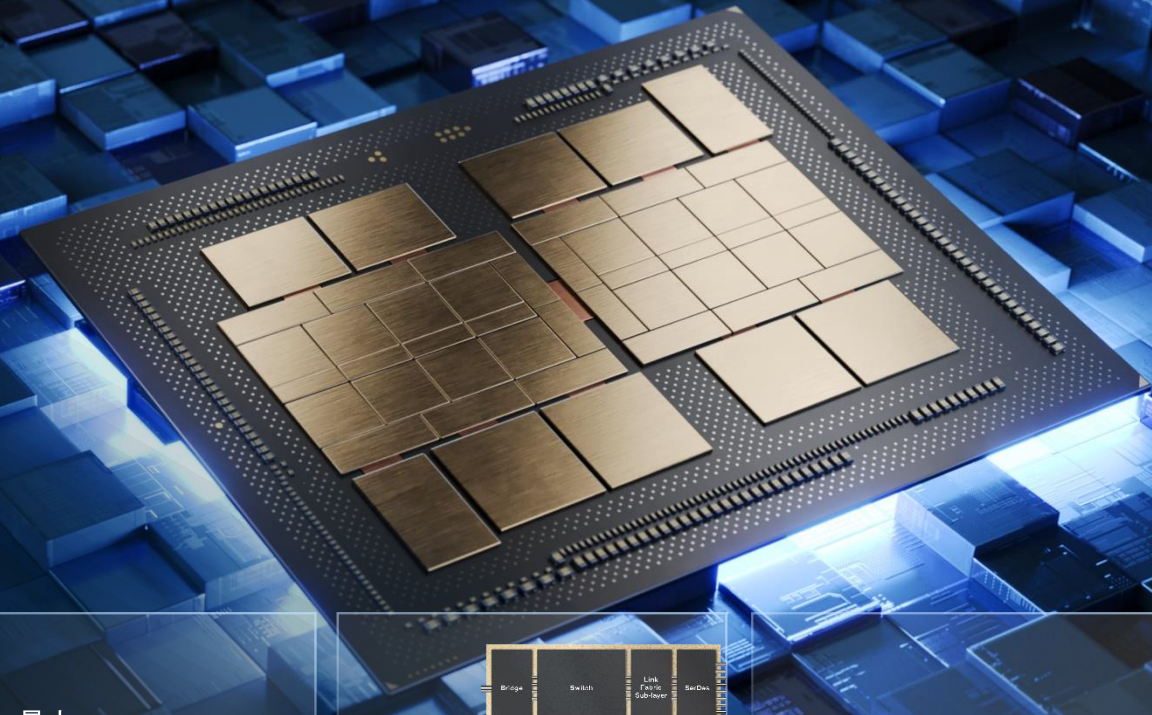
New Model transforms text into immersive 3D
Trained on Intel AI Supercomputer

intel ISC High Performance The HPC Event.

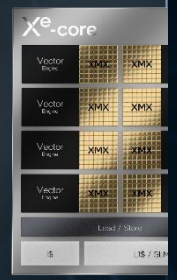
Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

intel
DATA CENTER
GPU
MAX SERIES

Intel® Data Center GPU Max Series



最大
128
Xe HPC Cores



52TF
Peak FP64 Throughput

839TF
Peak BF16 Throughput

最大
128 MB
HBM2e Memory

16
Xe Links

976 GB/s
GPU-to-GPU comms via Xe Links

intel

intel DATA CENTER GPU MAX SERIES

100+ HPC Apps Running

標準ベンチ	D,S,H,I,BF-GEMM	DAOS	Graph500	HPCG	HPL	IO500	MLPerf 2.0
	MLPerf.HPC	OSU	SPEChpc	Stream Triad	MLBench	SPEC ACCEL	SpMV
	RINF	ELPA	Ginkgo	HeFFTe	HYPRE	MFIX (AMReX)	
物理	HACC	DPEcho	GENE	NEKBONE	nekRS	OpenMC	
	XGC	CloverLeaf	Deep Galaxy	Gadget	GRID QCD	MILC	
	QUDA	Chroma	HotQCD	BQCD	CERN 3D GAN		
ライフサイエンス	Autodock-GPU	miniBUDE	AMBER	GROMACS	LAMMPS	NAMD	OpenMM
	Relion	Quantum Espresso	BerkeleyGW	CP2K	NWChem	QMCPACK	DeePMD
製造	ANSYS CoMLSim	Jacobi Solver	Commercial EDA ISV	Commercial CFD ISV			
	ANSYS ParSeNet (SplineNet)	Commercial Multi 物理 ISV	Commercial CFD ISV	Proprietary CFD code			
金融	Binomial Options	Black-Scholes	European Monte Carlo	American Monte Carlo	Riskfuel Risk Calculations	STAC-A2	
地球システム	SPECFEM3D GLOBE	ES3M/MMF	SeisSol				
エネルギー	RTM Stencil Kernel	ISO3DFD					

intel

AI モデルサポート

Computer Vision Image Classification	ResNet-50 v15	ResNeXt-101	ResNet-101	EfficientNet-B7	SE-ResNeXt50	TSM	
	Adorym	CosmoFlow	RegNetY-32Y	ResNeXt-101	Candle Uno	Swin Transformer	
画像セグメンテーション	Cosmic Tagger	Mask R-CNN	DenseNet169	FFN	3D-Unet		
	PointNet	DeepCAM	DRN-D-54	ResNeXt3D-101			
物体検知	SSD-ResNet34	SSD-ResNet50	EfficientDet	ShuffleNet	YOLO-v3	YOLO-v4	RetinaNet-ResNet50
	Deep Fusion	CascadeRCNN-	MobileNet v3	SSD-MobileNet	MMA	ResNet101-FPN	
NLP 言語モデリング	BERT-Large	Stable Diffusion	ALBERT	FastFormers	Transformer-LT	Big Bird	Faster Transformer
	BERT-base	GP-J	BLOOM	DistilBERT	RoBERTa	XLNet	
音声認識	RNN-T	LAS - Listen Attend & Smell	Wave2Vec	QuartzNet			
音声合成	FastSpeech2	Tacotron-2 with LPCNet					
リコメンデーション	DLRM	DSSM	ESSM	Wide & Deep	DeepFM		
	DIN	AttRec	DIEN	MMOE			

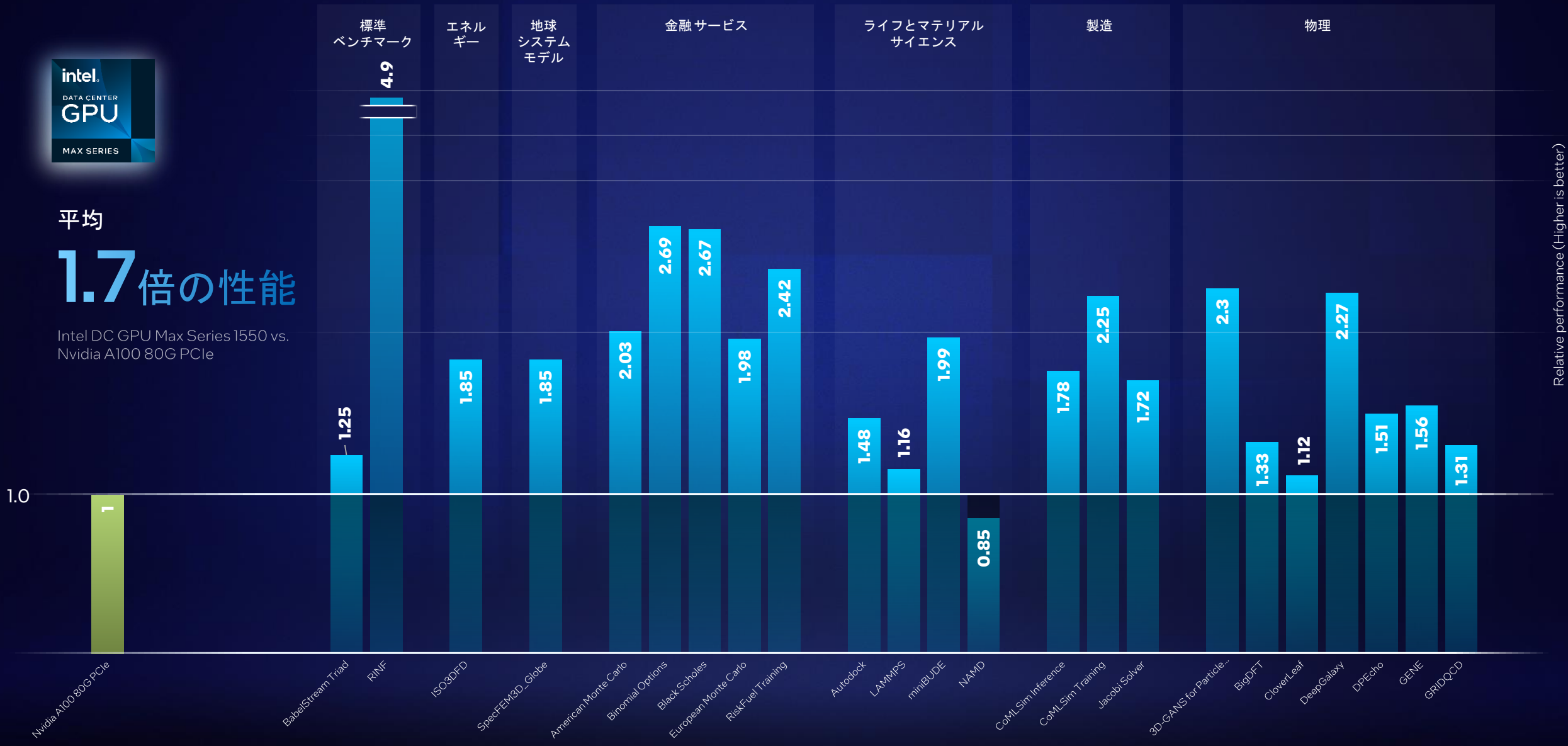
intel

intel DATA CENTER GPU MAX SERIES

平均

1.7倍の性能

Intel DC GPU Max Series 1550 vs. Nvidia A100 80G PCIe



See backup for workloads and configurations. Results may vary.

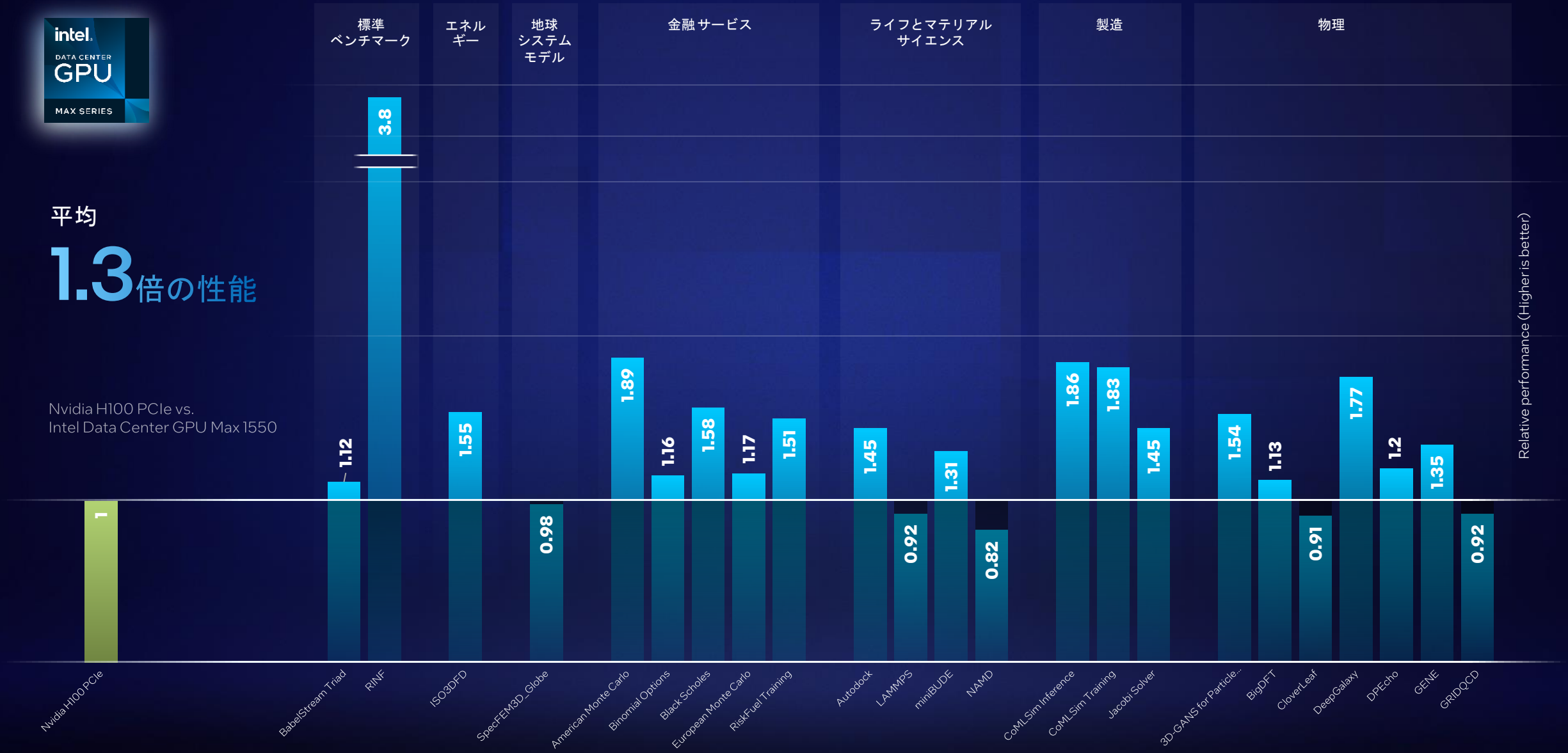
intel

intel DATA CENTER GPU MAX SERIES

平均

1.3倍の性能

Nvidia H100 PCIe vs.
Intel Data Center GPU Max 1550



intel

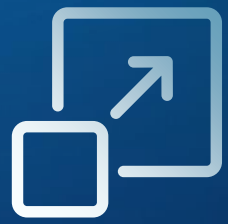
Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

最小限の努力で
最適なソリューションを選択可能にする

Open
Ecosystem

Choice
Compatibility

Trust
Workloads



Scale
Delivery & Deployment

intel.

oneAPI

オープン、
マルチアーキテクチャ、
マルチベンダープログラミング

- オープンな業界仕様
- ハードウェア選択の自由度
- 性能、生産性、ポータビリティ
- コミュニティ主導の標準化

oneAPI

ツールセットとしてインテルが実装

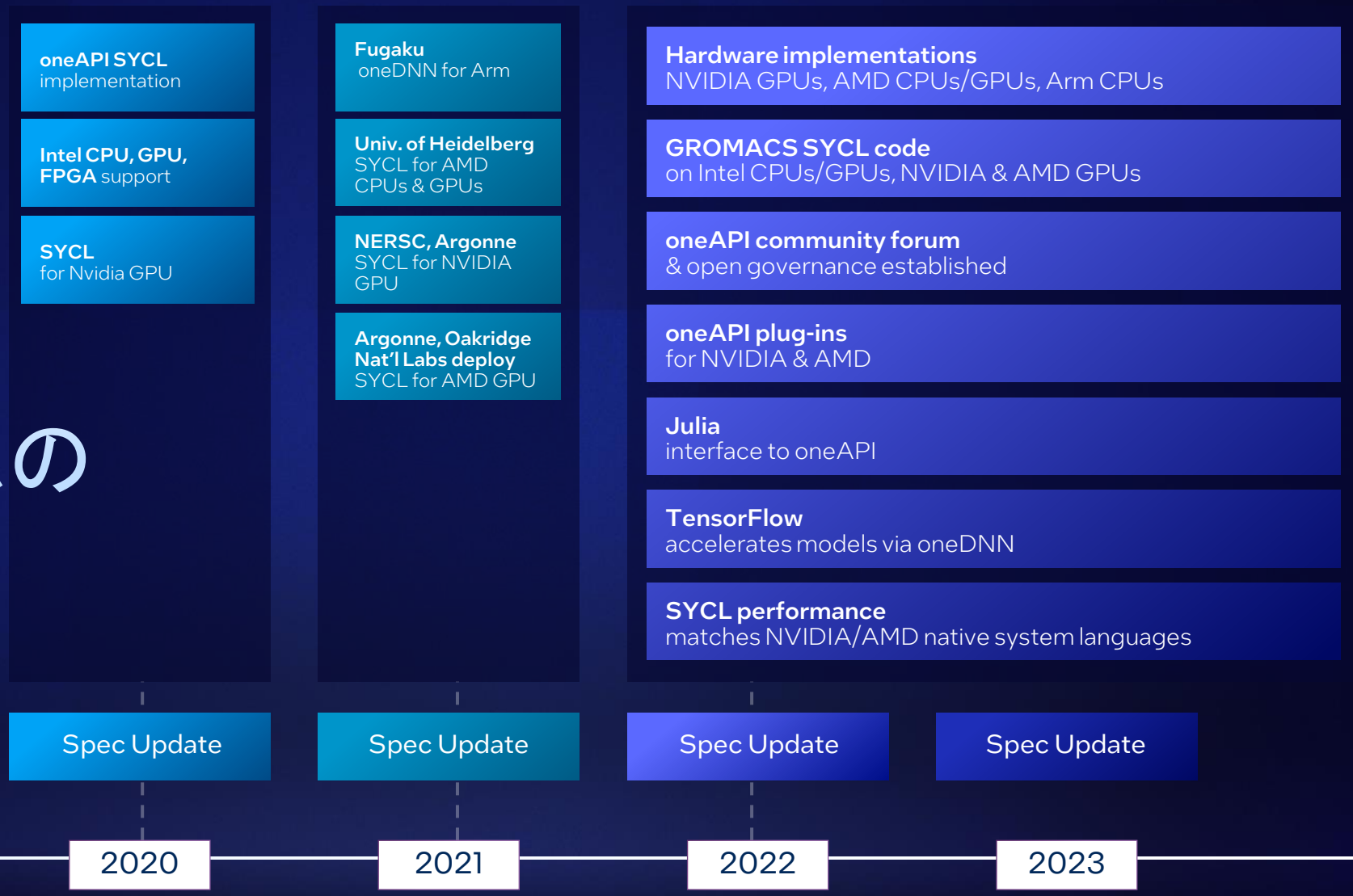
- インテルハードウェアに最適化
- 確かな機能と性能
- Fortran, Python, OpenMP, MPI... サポート
- ダウンロードは無償、商用サポートあり

intel

oneAPI

オープン エコシステムの 進展

採用が堅調に進む



Intel oneAPI および AI ツール
インテルハードウェアの価値を最大化

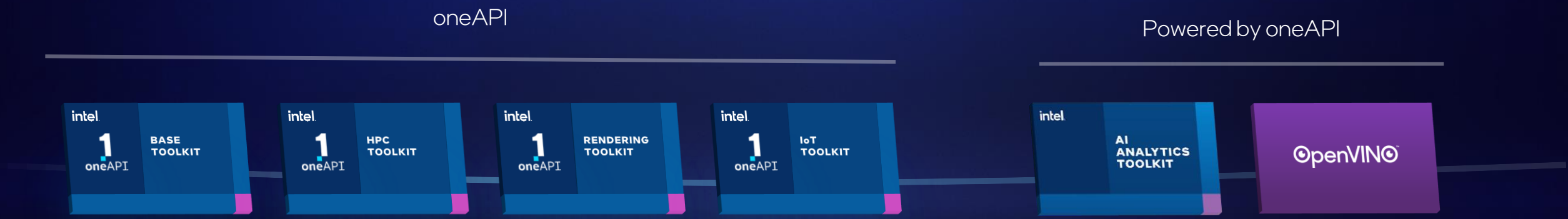
新機能

Accelerated performance & AI
by enabling 第四世代 Intel® Xeon® CPUs & Max Series CPUs & GPUs

Speed up HPC applications
with OpenMP GPU offload & extended support for OpenMP & Fortran
Base & HPC toolkit

Real-time ray tracing
with hardware acceleration on Intel® GPUs & AI-based denoising in milliseconds*

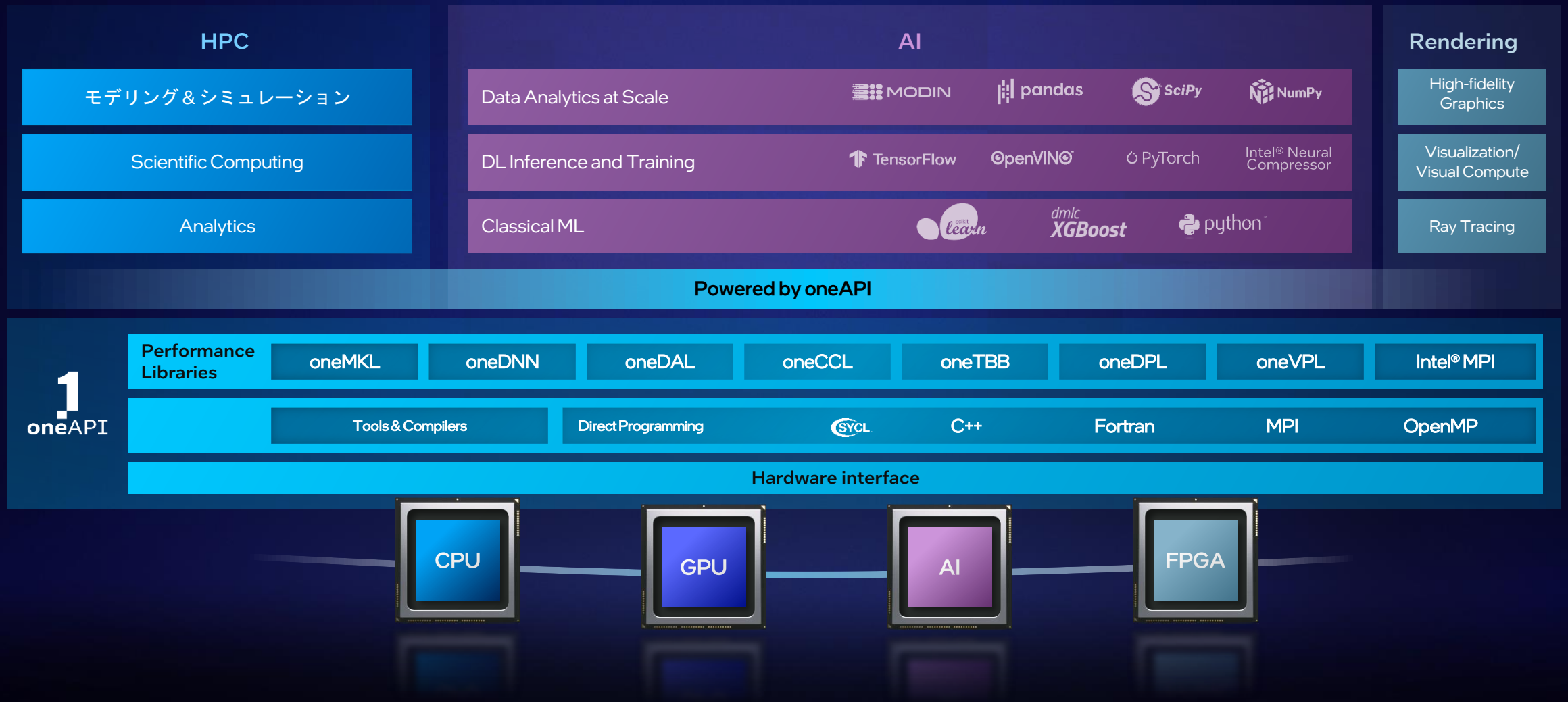
Easier CUDA*-to SYCL* migration
with improved tools
Base Kit



*avail. in open source Intel® Embree & Intel® Open Image Denoise, they will be in the Render Kit's next release

柔軟かつ包括的なオープンソフトウェアスタック

アプリケーション性能と生産性の高い開発環境によりインテルハードウェアの価値を最大化する



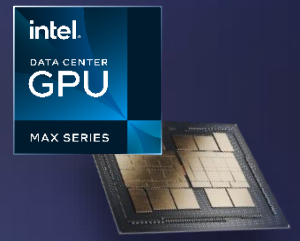
この一年の間に
いろいろなことが起こりました…

AIとHPCアプリケーション性能をリード

強化された統一ソフトウェア層

Scale | Open | Trusted | Choice

すべてのHPCとAIニーズに対応する製品群



*Conceptual Representation

intel


intel

Notices and Disclaimers

Statements in this document that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.

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

Performance varies by use, configuration and other factors. Learn more on the Performance Index site. Intel technologies may require enabled hardware, software or service activation.

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.

Your costs and results may vary.

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

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