

# インテルHPC-AI の今

2023年12月7日

矢澤 克巳  
インテル株式会社 HPC事業開発部長

intel<sup>®</sup>

# Siliconomy

## シリコノミー

シリコンの可能性によって実現する経済成長  
現代経済の維持と発展に半導体は不可欠

- テクノロジーは現代の生活に広く浸透しており、コンピューティングの需要が拡大している
- 「Superpowers」(コンピューティング、接続性、インフラ、AI、センシングなど)が相互に結合し、新たな可能性を切り拓いている
- これにより、コンピューティング技術が急速に進化し、その基盤としてシリコンが重要な役割を果たしている。

シリコノミーの世界へようこそ



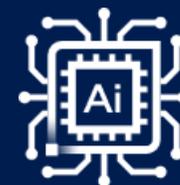
# インテルのフォーカス・エリア



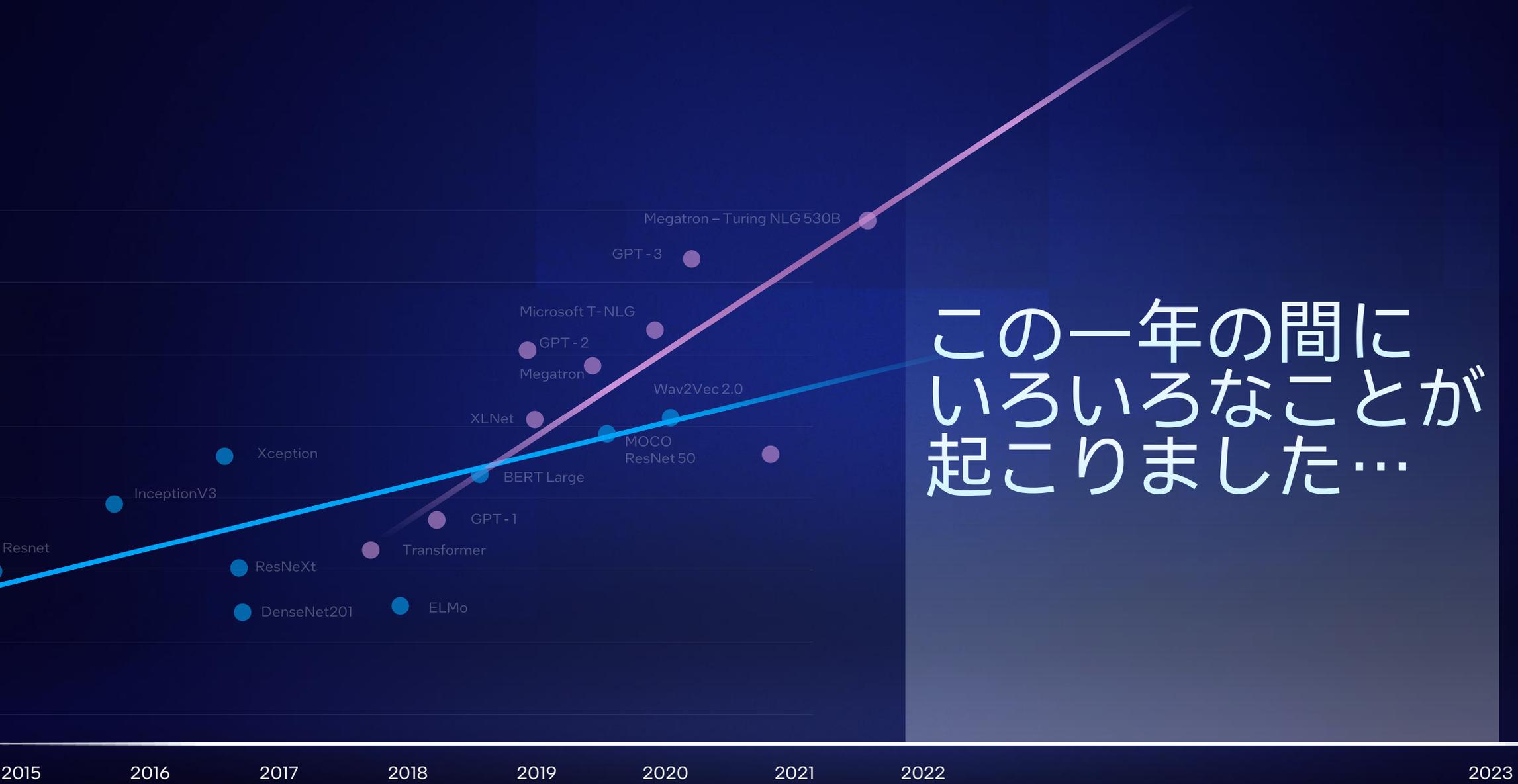
グローバル・  
サプライチェーン  
の強靱化



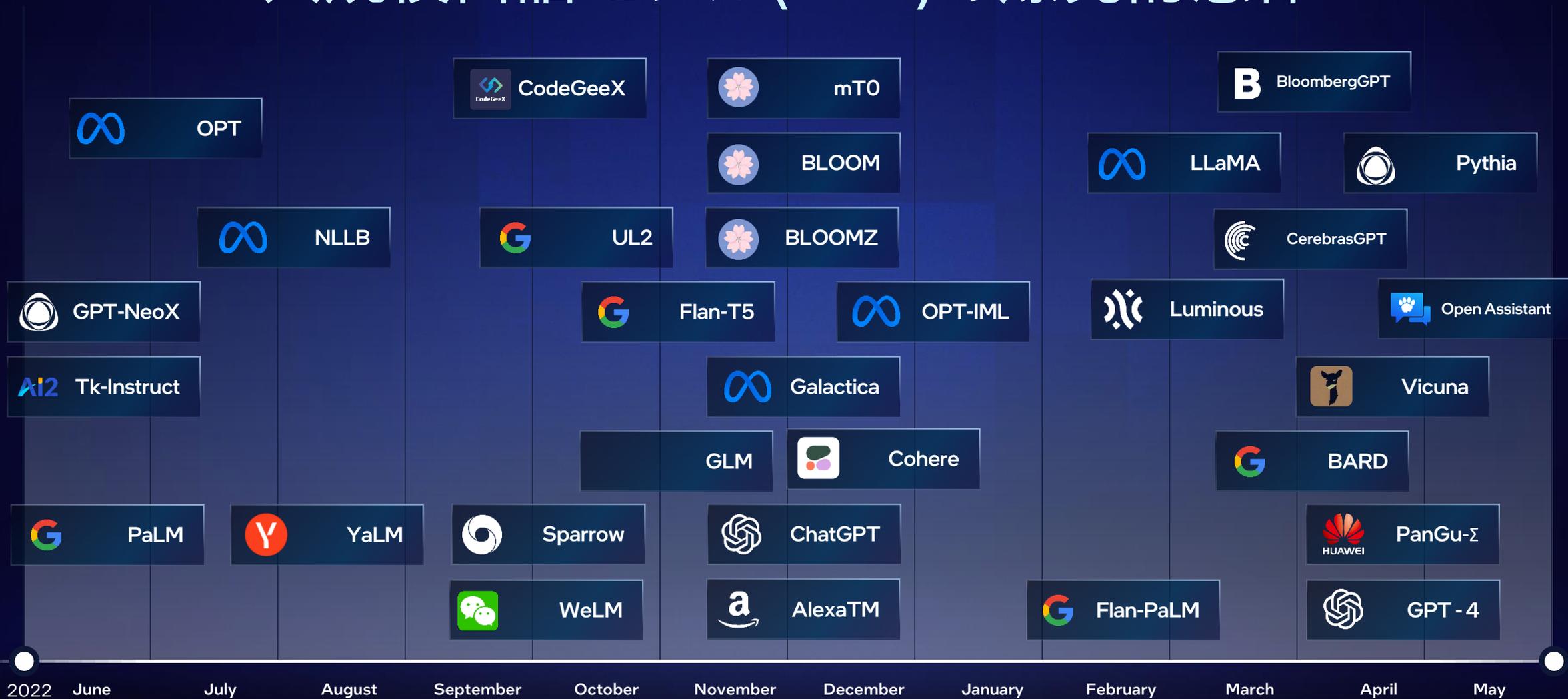
ムーアの法則の  
継続



AI Everywhere



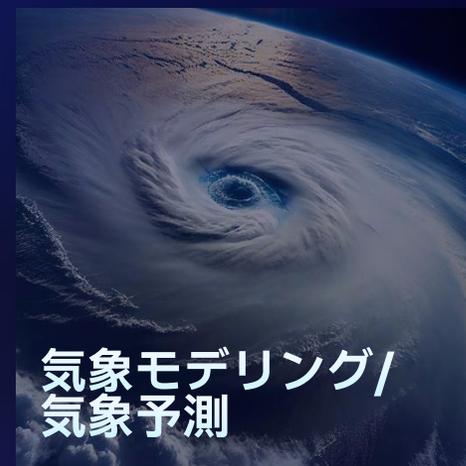
# 大規模言語モデル(LLM)の爆発的急伸



2022 June July August September October November December January February March April May 2023



# 科学分野に AIブームが到来...



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

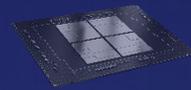
## インテル® Xeon® CPU マックス ・シリーズ

## インテル® データセンター GPUマックス ・シリーズ

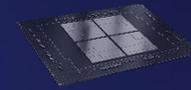


# インテル データセンター ロードマップ

E-Core  
CPU



インテル® Xeon®  
プロセッサ  
コードネーム Sierra Forest

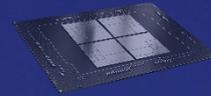


インテル® Xeon®  
プロセッサ  
コードネーム Clearwater Forest

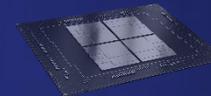
P-Core  
CPU



第四世代 インテル® Xeon®  
スケーラブル・プロセッサ



第五世代 インテル® Xeon®  
プロセッサ  
コードネーム Emerald Rapids



インテル® Xeon®  
プロセッサ  
コードネーム Granite Rapids

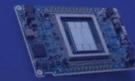


インテル® Xeon®  
CPU マックス・シリーズ

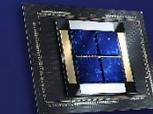
AI  
専用アクセラレータ



Habana®  
Gaudi® 2



Habana®  
Gaudi® 3



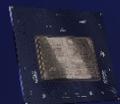
次世代 GPU  
コードネーム Falcon Shores

HPC & AI  
GPU

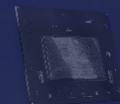


インテル® Data Center GPU マックス・シリーズ

Visual Cloud  
GPU

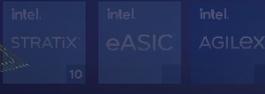
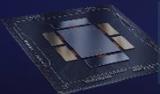


インテル® Data Center GPU Flex Series



インテル® 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+

**Sierra Forest demo**  
2 processors with 288 cores

intel.  
**XEON**

```
processor      : 567  
processor      : 568  
processor      : 569  
processor      : 570  
processor      : 571  
processor      : 572  
processor      : 573  
processor      : 574  
processor      : 575  
demo@srfubuntu22p4-3405:~$
```

**288 cores**  
**Intel® Xeon® with E-core**  
Launch 2024

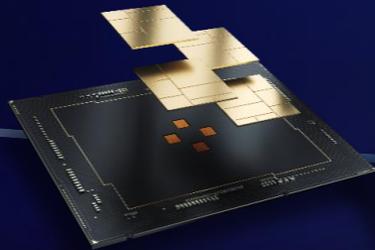
# Intel® Xeon® プロセッサ

第四世代 Intel®  
Xeon® プロセッサ

アクセラレータエンジンを搭載し  
ワークロードに最適化

ML Perf 学習: コスト効率に優れた Intel® イーサネットを使用し、97~100%のマルチノード・スケーリングを実現

提供中

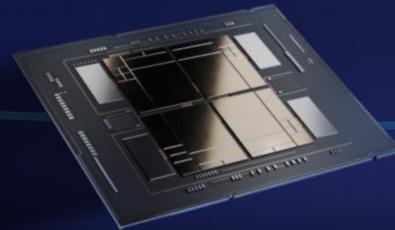


Intel® Xeon® CPU  
マックス・シリーズ

高いメモリバンド幅  
HBM2e

第四世代 Xeon と  
プラットフォーム  
互換

提供中



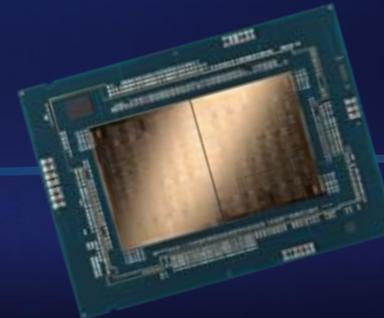
第五世代 Intel®  
Xeon® プロセッサ

Formerly codenamed  
Emerald Rapids

消費電力当たりの  
より高い性能

第四世代 Xeon と  
プラットフォーム  
互換

2023年12月14日  
発表予定



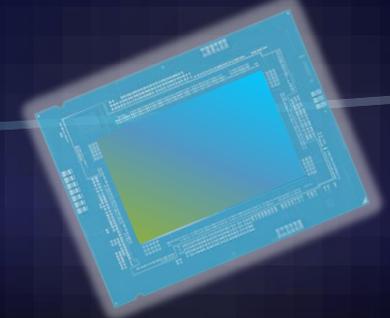
将来の Intel® Xeon®  
プロセッサ

Codenamed Granite Rapids

Intel 3 で作る  
最初の P-Core Xeon

多いコア数,  
高いメモリバンド幅,  
AIアクセラレーション  
の拡張

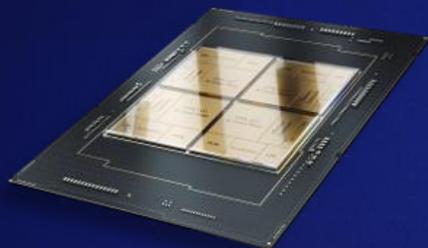
2024年提供に  
向け開発中



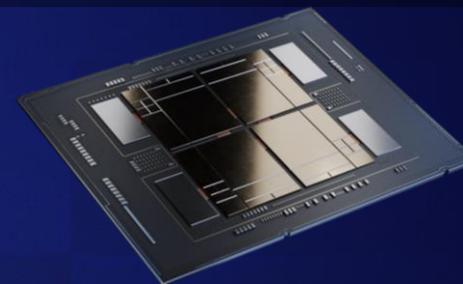
# インテル® Xeon® プロセッサー

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



ブレイクスルー・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

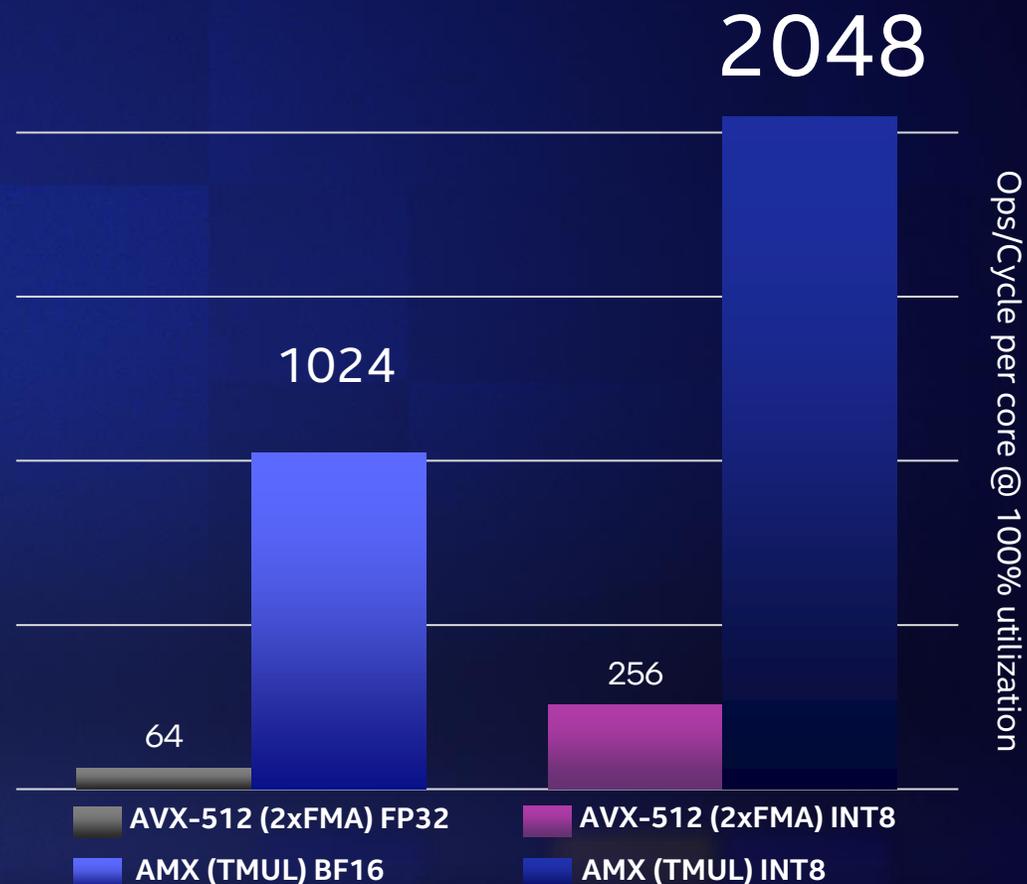
ディープラーニング  
データタイプ

- int8
- Bfloat16

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](http://www.intel.com/ArchDay21claims). Results may vary



標準  
ベンチマーク

地球システム  
モデリング

エネルギー

金融  
サービス

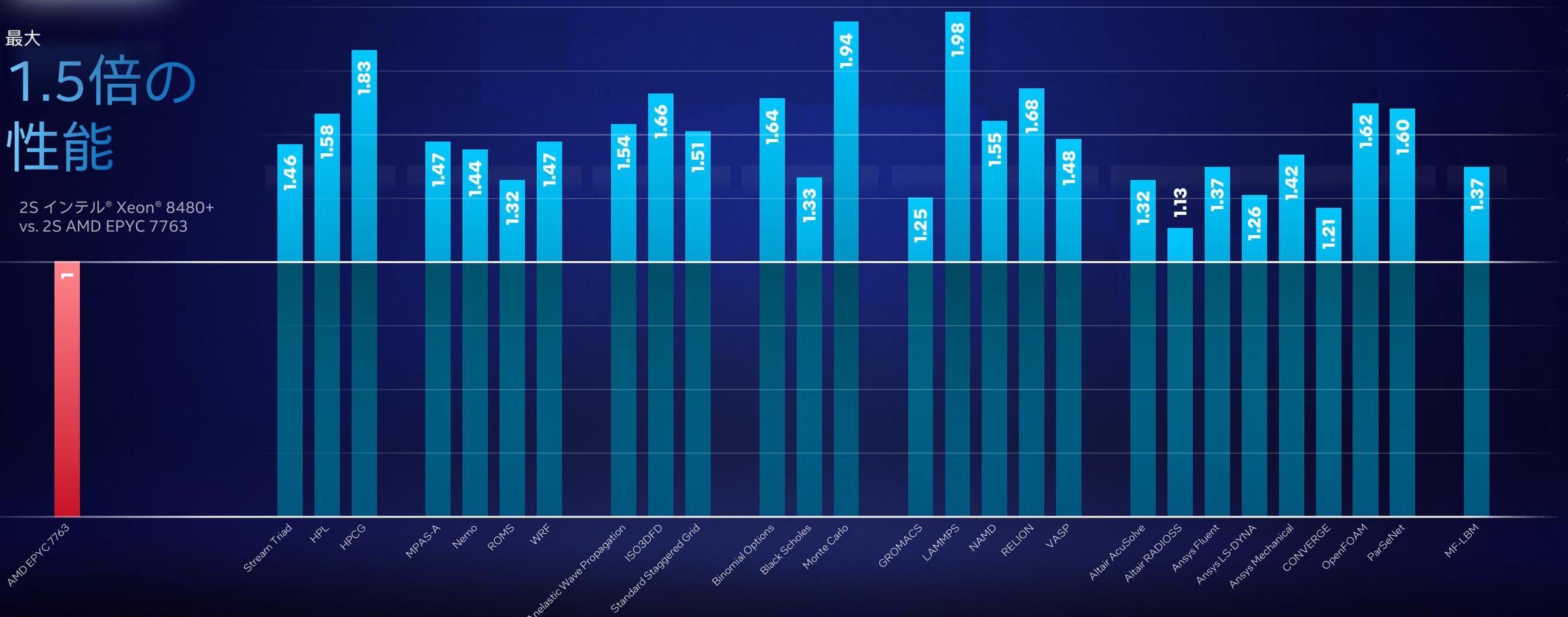
ライフとマテリアル  
サイエンス

製造

物理

# 最大 1.5倍の 性能

2S インテル® Xeon® 8480+  
vs. 2S AMD EPYC 7763



Relative performance (Higher is better)

Visit [www.intel.com/performanceindex](http://www.intel.com/performanceindex) for workloads and configurations. Results may vary

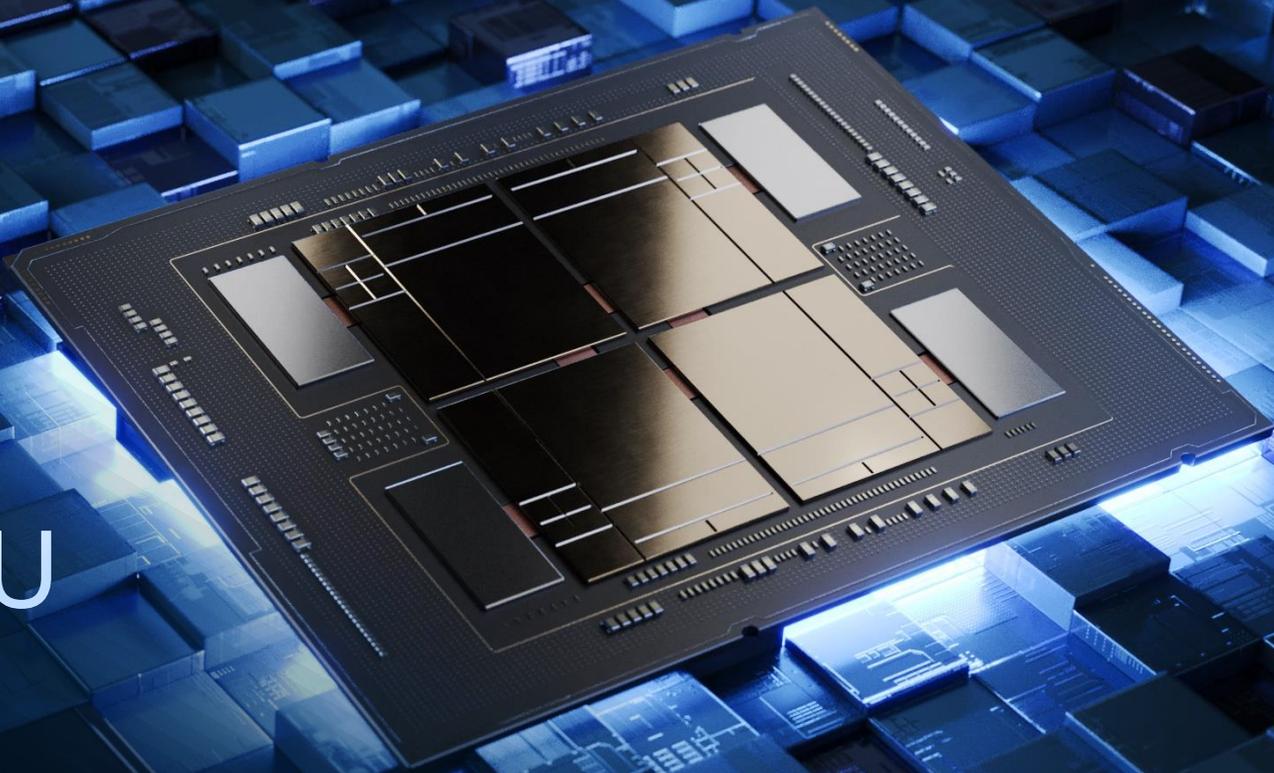
This offering is not approved or endorsed by OpenCFD Limited, producer and distributor of the OpenFOAM software via [www.openfoam.com](http://www.openfoam.com), and owner of the OPENFOAM® and OpenCFD® trademark





# X86初 HBMを持つ唯一のCPU

最適なメモリ選択肢となりうる



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



## 64GB

HBM2e

4 stacks of  
16GB

最大

## 220GF/s

HPCG

最大

## 2GB

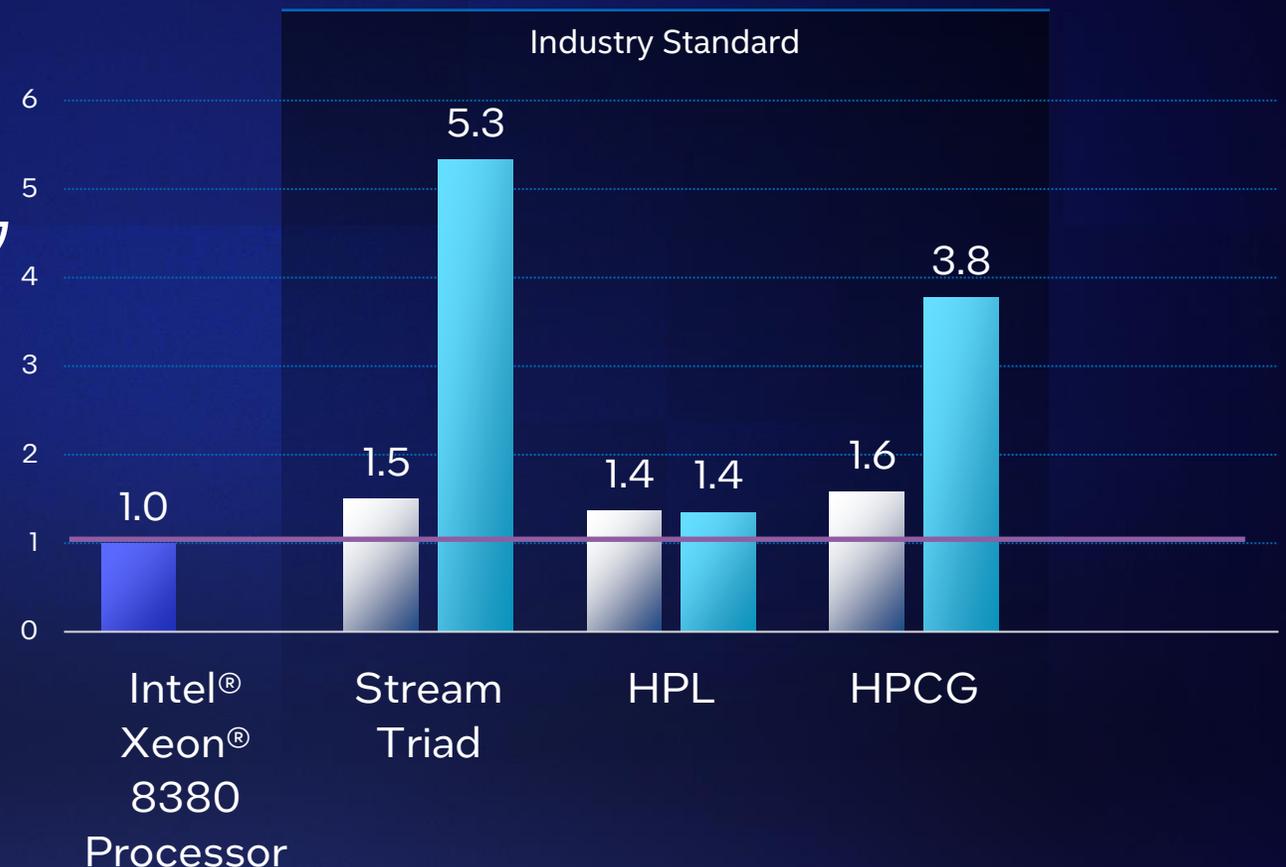
HBM per Core



# メモリ帯域を要求するベンチマーク CPU MAXでは最大 5倍の性能

2S インテル® Xeon® CPU マックス・シリーズ vs.  
2S 第三世代 インテル® Xeon® 8380 プロセッサ

Relative Perf. Higher is better



■ 第四世代 インテル® Xeon® スケーラブル・プロセッサ ■ インテル® Xeon® CPU マックス・シリーズ

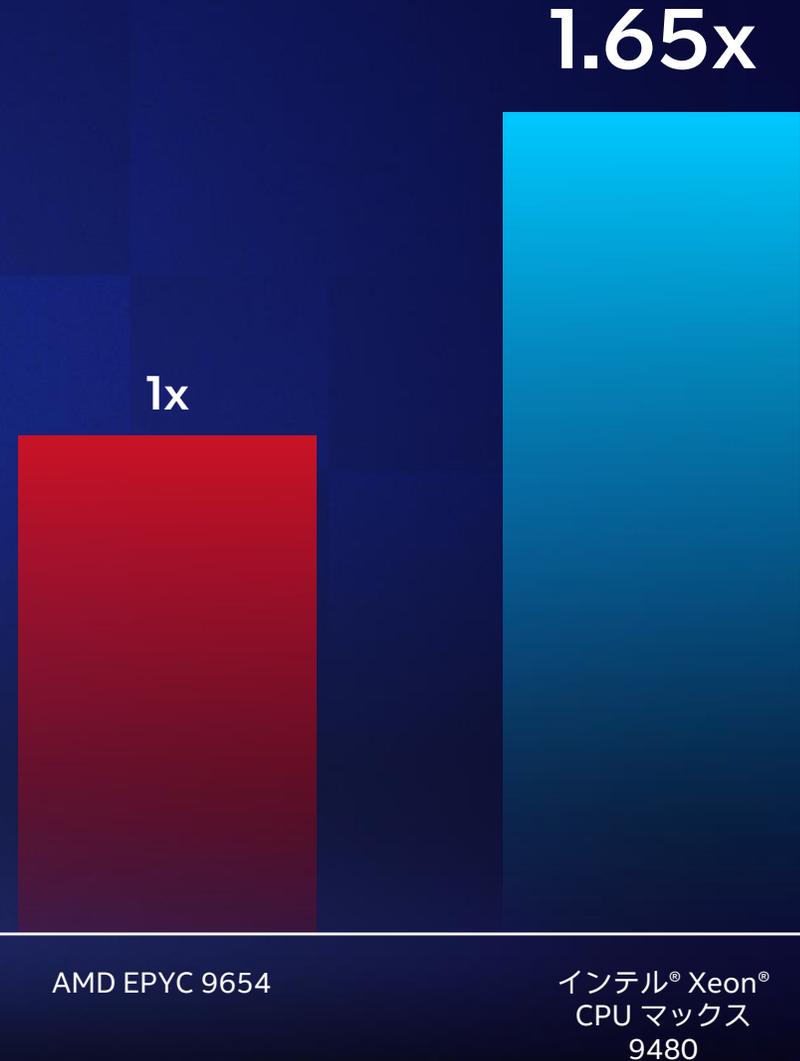
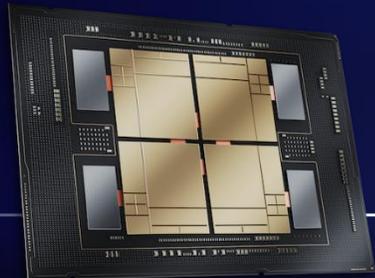
See backup for workloads and configurations. Results may vary.





# コア数差を超えた性能

HPCG 性能



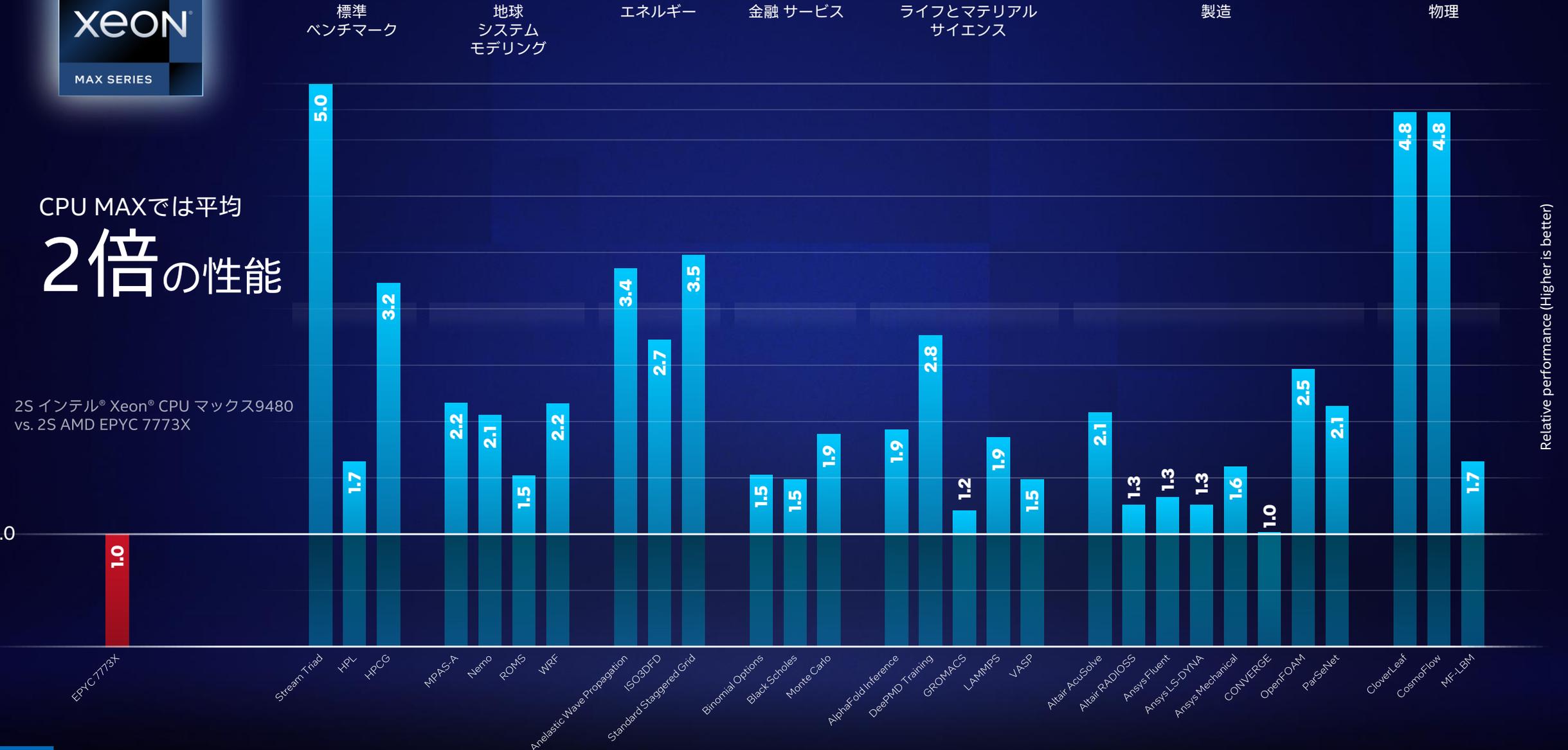
Relative performance (Higher is better)





CPU MAXでは平均  
**2倍**の性能

2S インテル® Xeon® CPU マックス9480  
vs. 2S AMD EPYC 7773X



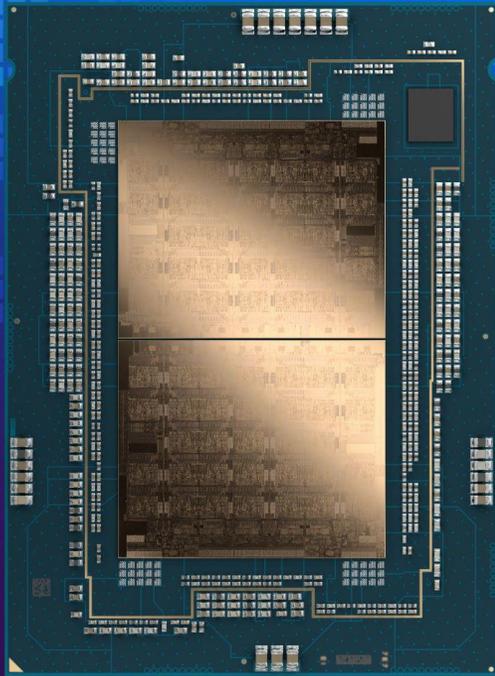
Relative performance (Higher is better)

See backup 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](http://www.openfoam.com), and owner of the OPENFOAM® and OpenCFD® trademark.  
MLPerf™ HPC-AI v0.7 Training ベンチマーク Performance. Result not verified by MLCommons Association. Unverified results have not been through an MLPerf™ review and may use measurement methodologies and/or workload implementations that are inconsistent with the MLPerf™ specification for verified results. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See [www.mlcommons.org](http://www.mlcommons.org) for more information.





intel  
XEON

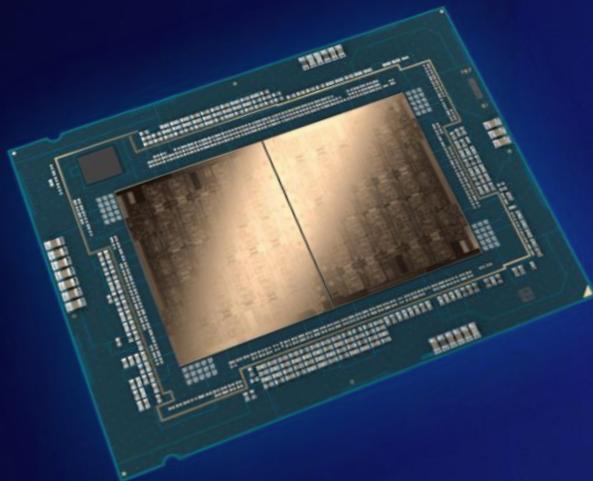


# 第五世代 インテル® Xeon® プロセッサー

Formerly codenamed Emerald Rapids

2023年12月14日

# 第五世代 インテル® Xeon® プロセッサー



ワークロードに最適化された  
電力効率の高いコンピュータ

高速メモリ 8x DDR5, 5600 MT/s

拡張 I/O: CXL タイプ1と2, PCIe Gen 5, 80レーン

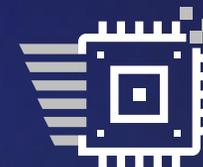


AI  
音声認識

最大

1.4倍

higher throughput  
(rec/sec) on  
Xeon 8592+  
vs  
Xeon 8480+ (BF16)



HPC  
LAMMPS -  
Copper

最大

1.4倍

higher performance  
On  
Xeon 8592+  
vs  
Xeon 8480+



メディア  
Transcode  
(FFMPEG)

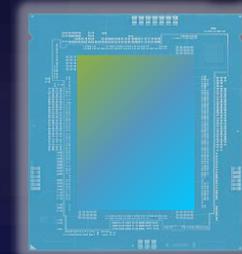
最大

1.2倍

aggregate FPS  
on  
Xeon 8592+  
vs  
Xeon 8480+

# 将来の Intel® Xeon® プロセッサ

codenamed Granite Rapids



## Extends the Intel® Xeon® processor Advantage

# 2-3x

より高いAIワークロード性能<sup>1</sup>

# 2.8x

より高いメモリバンド幅<sup>1</sup>

# 2.9x

DeepMD+LAMMPS AI インファレンス<sup>2</sup>

- より多いコア数, より高い周波数, Intel® Advanced Matrix Extensions (Intel® AMX)
- FP16 を追加
- 12 メモリチャネルおよび MCR サポート (large LLMモデルに有効)
- HPCとAIワークロードに最適なTCOを実現

<sup>1</sup>- Based on architectural projections as of August 21, 2023 vs prior generation platforms. Your results may vary.

<sup>2</sup>- See [intel.com/performanceindex](https://www.intel.com/performanceindex) for workloads and configurations. Your results may vary.

# MCR DIMM

## Memory

Granite Rapids向け2-Rank RDIMMによる  
同容量でより高いメモリバンド幅を持つ

最大

**8800**

MT/s

最大

**83% peak**

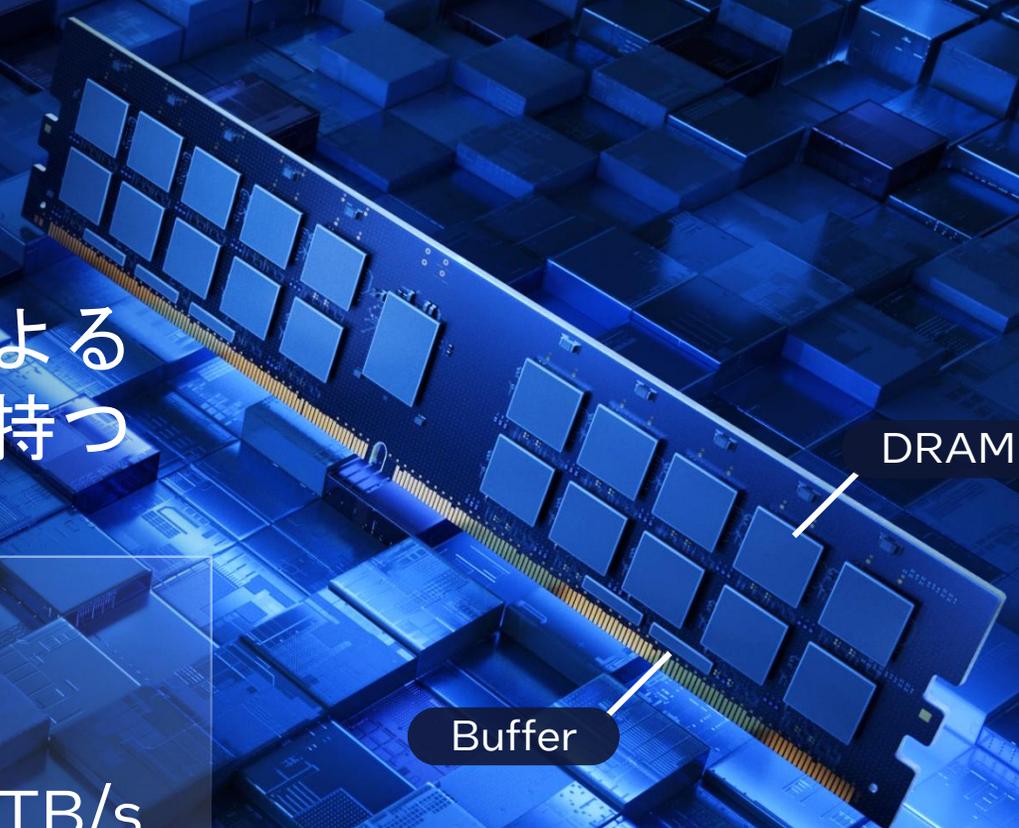
B/W increase

2 socket

**>1.5 TB/s**

DRAM

Buffer





# インテル® Data Center GPU マックス・シリーズ

最大  
**128**  
Xe HPC  
Cores

A diagram of the Xe-core architecture showing a grid of Xe HPC cores. Labels include "Xe-core", "Vector Engine", "Xe HPC", and "Load/Store".

**52TF**  
Peak FP64  
Throughput

**839TF**  
Peak BF16  
Throughput

最大  
**128 MB**  
HBM2e  
Memory

**16**  
Xe Links

A diagram of the Xe Link architecture showing a controller connected to a switch, which is connected to a link processor and a server rack. Labels include "Bridge", "Switch", "Link Processor", "Server Rack", and "Controller".

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



標準ベンチマーク  
 エネルギー  
 地球システムモデル  
 金融サービス  
 ライフとマテリアルサイエンス  
 製造  
 物理

平均  
**1.7倍の性能**

インテル DC GPU マックス・シリーズ 1550 vs. Nvidia A100 80G PCIe

1.0



Relative performance (Higher is better)

See backup for workloads and configurations. Results may vary.





標準  
ベンチマーク

エネルギー

地球  
システム  
モデル

金融 サービス

ライフとマテリアル  
サイエンス

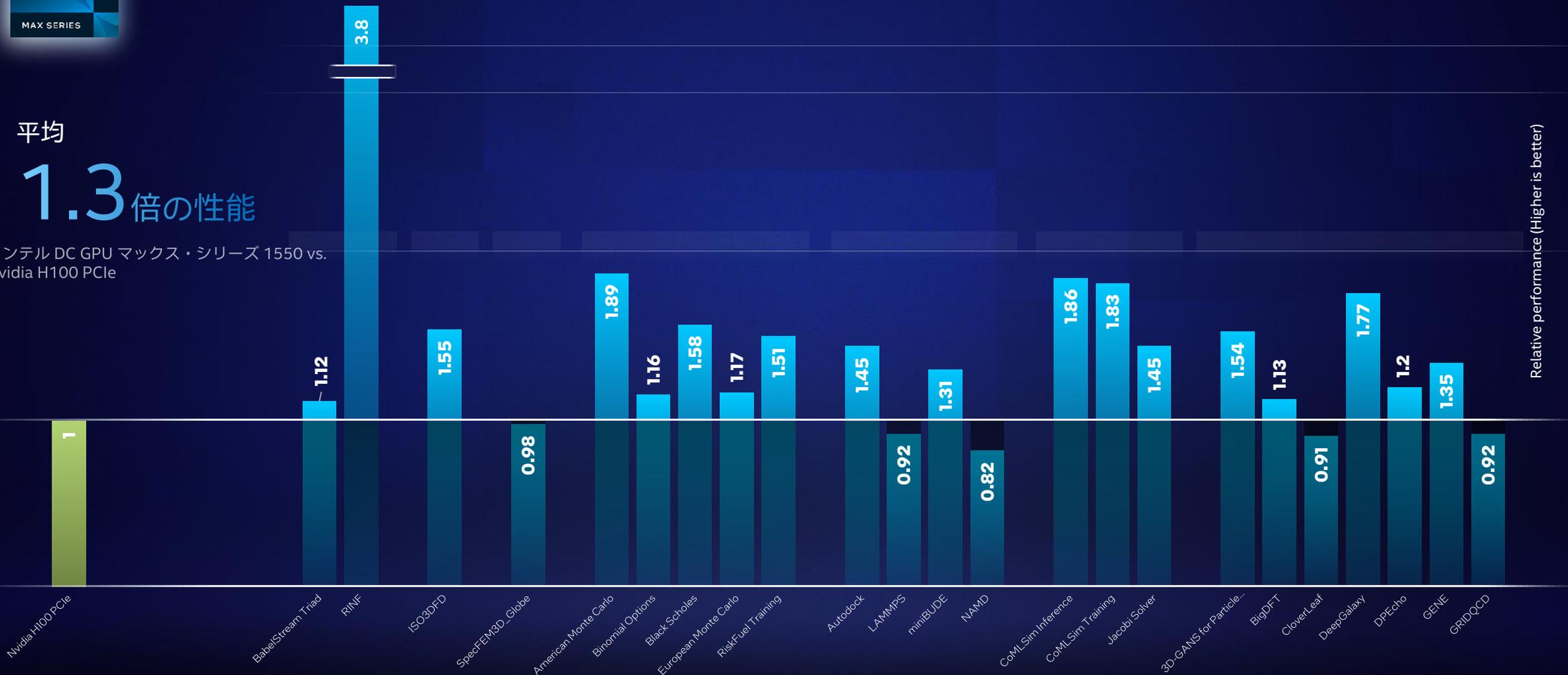
製造

物理

平均

# 1.3倍の性能

インテル DC GPU マックス・シリーズ 1550 vs.  
Nvidia H100 PCIe



Relative performance (Higher is better)



# 100+ HPC Apps Running



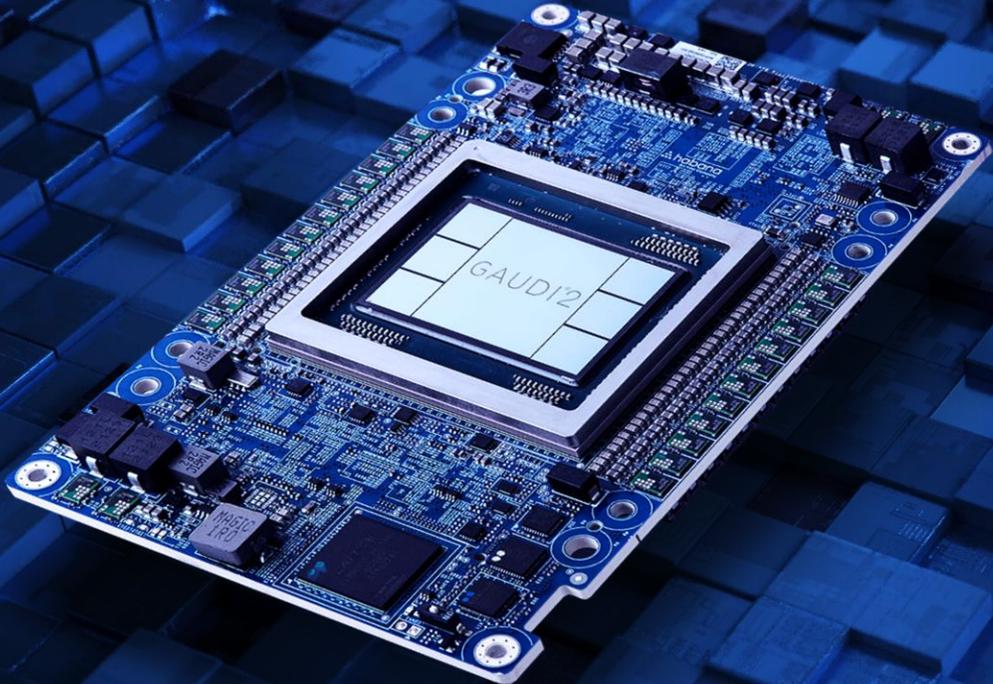
標準ベンチ	D,S,H,I,BF-GEMM	DAOS	Graph500	HPCG	HPL	IO500	MLPerf 2.0
	MLPerf.HPC	OSU	SPEChpc	Stream Triad	MLBench	SPEC ACCEL	SpMV
	RINF	ELPA	Ginkgo	HeFFTe	HYPRE	MFIX (AMReX)	
物理	HACC	DPEcho	GENE	NEKBONE	nekRS	OpenMC	
	XGC	CloverLeaf	Deep Galaxy	Gadget	GRID QCD	MILC	
	QUDA	Chroma	HotQCD	BQCD	CERN 3D GAN		
ライフサイエンス	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					

# 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			

# GAUDI<sup>®</sup>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

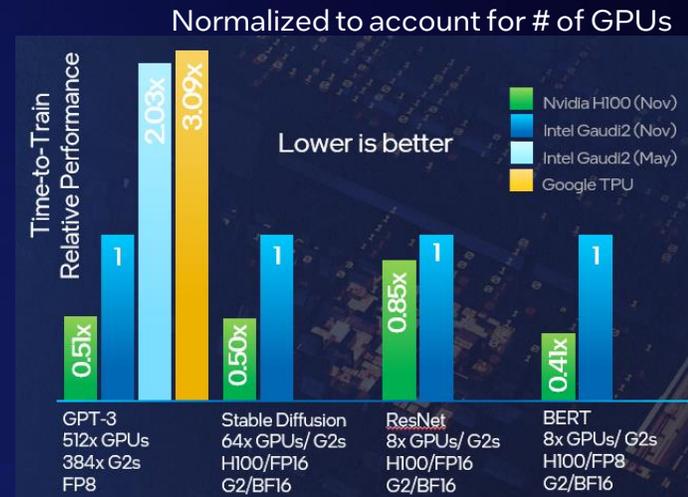
On-Prem

Supermicro Gaudi2 Server

# Intel® Gaudi® 2 AI アクセラレータ

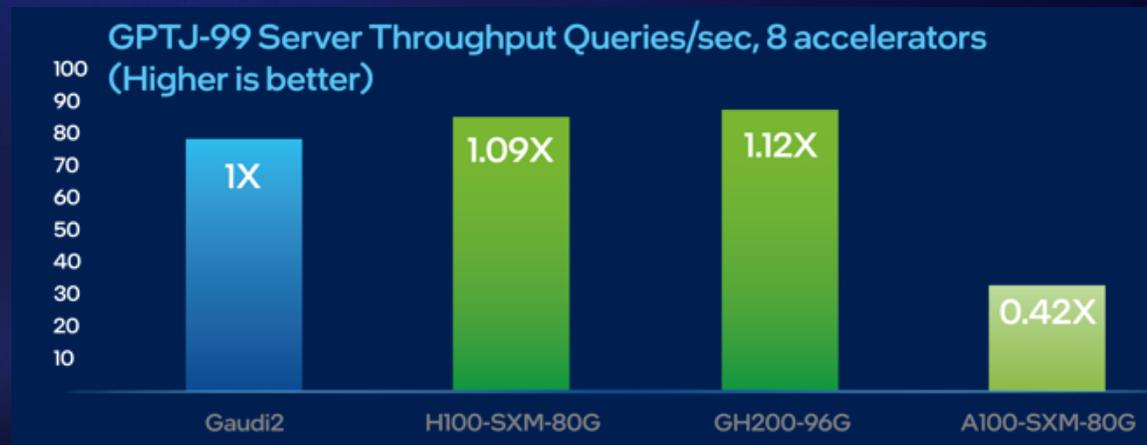
## MLPerf GPT-3 学習ベンチマーク

- MLPerfの結果を提出した唯一のAIアクセラレータ—3社のうちの1社
- FP8により103% GPT-3性能を向上
- 価格性能の優位性



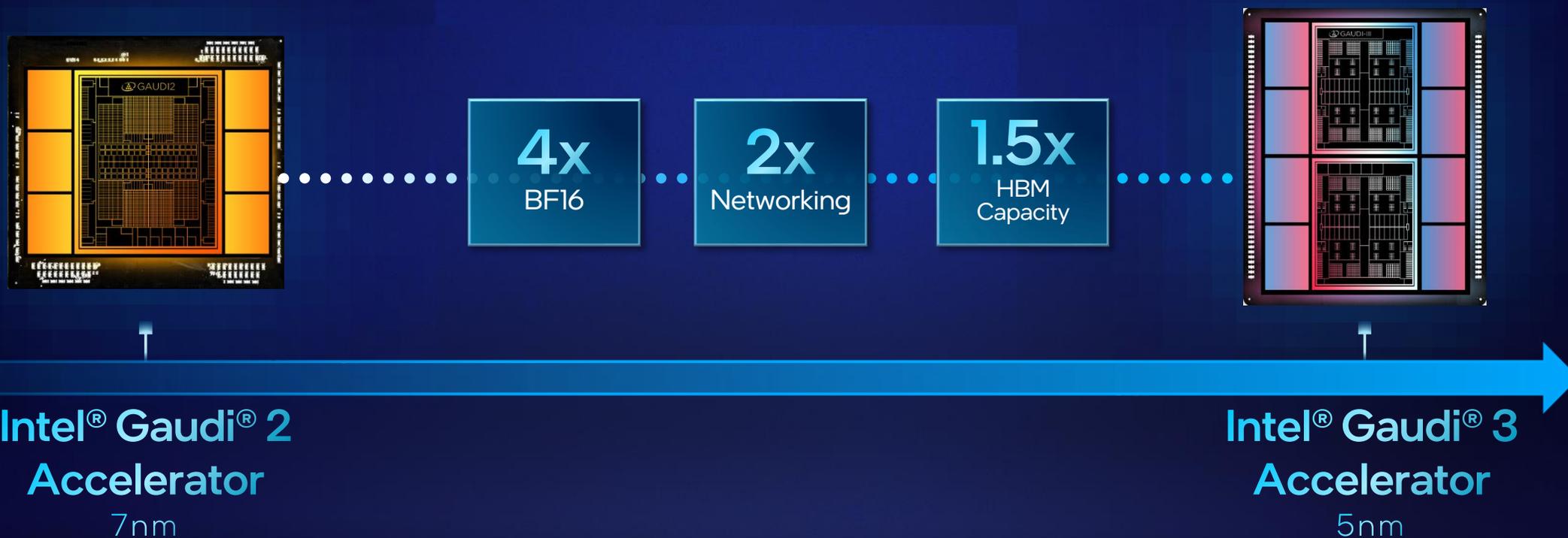
## MLPerf GPT-J 推論ベンチマーク

- H100とほぼ同等のパフォーマンス
- A100と比較して2倍の性能



Visit <https://mlcommons.org/en/> for workloads and configurations. Results may vary.

# Intel® Gaudi® 3 AI アクセラレータ 2024年予定



# アクセラレータと GPU ロードマップ

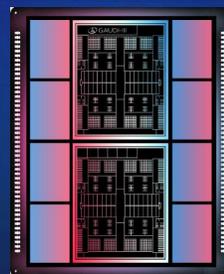
HPC/AI



AI

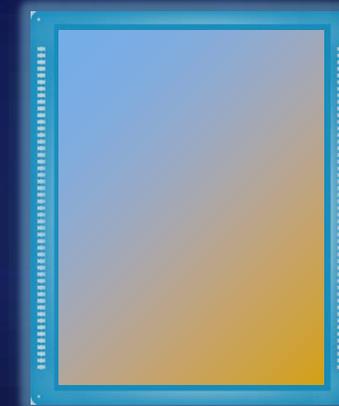


Intel® Data Center  
GPU Max Series



Intel® Gaudi® 2 Accelerator

Intel® Gaudi® 3 Accelerator



Next Generation GPU  
Codenamed  
**Falcon Shores**



# Falcon Shores

GPU

AI & HPC向け  
次世代 GPU

Habana と Xe IPを統合

タイルベースのモジュラー・アーキテクチャー

HBM3とI/Oは拡張できるように設計

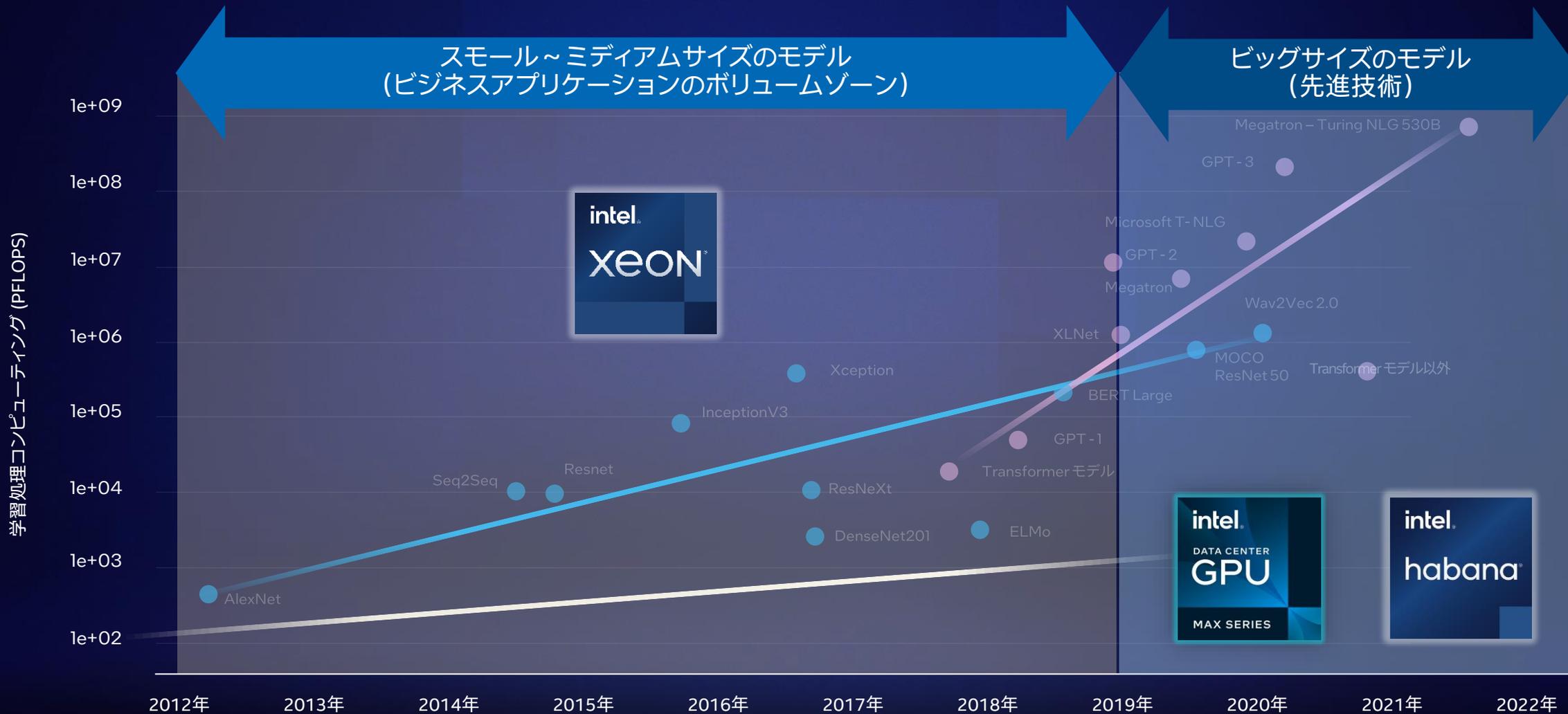
標準イーサネット・スイッチング

柔軟なCPUとGPU比率

単一のGPUプログラミング・インターフェース

CXLプログラミングモデル

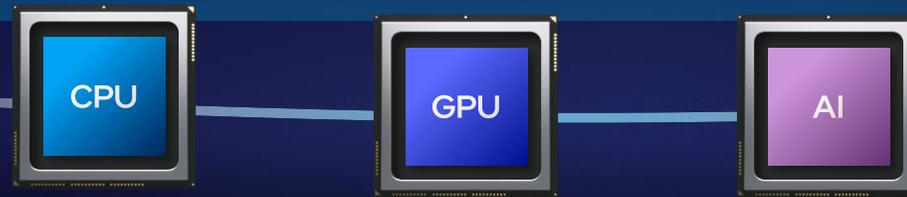
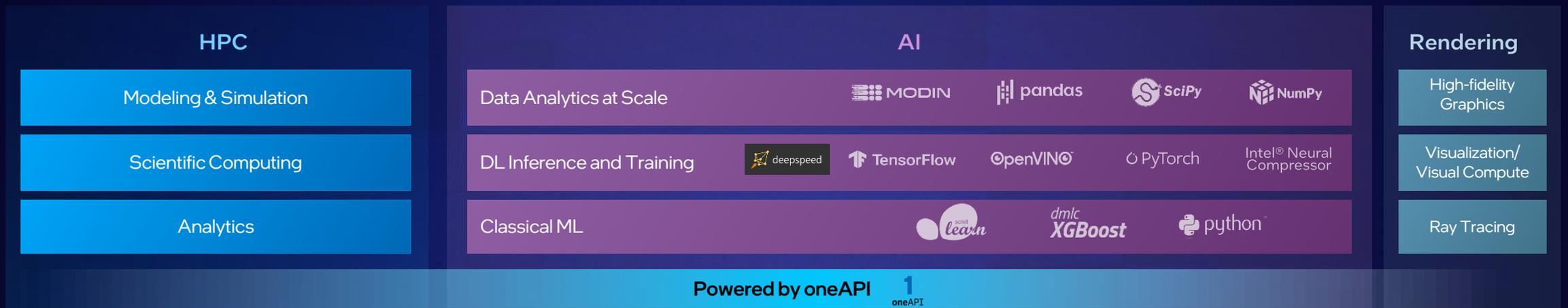
# こういったモデルに適している？



出典: Moore, S. (2022年), 「Nvidia's Next GPU Shows That Transformers Are Transforming AI」、IEEE Spectrum

# 柔軟で包括的なオープンソフトウェアスタック

インテルハードウェアの価値、アプリケーションのパフォーマンス、開発の生産性を最大化



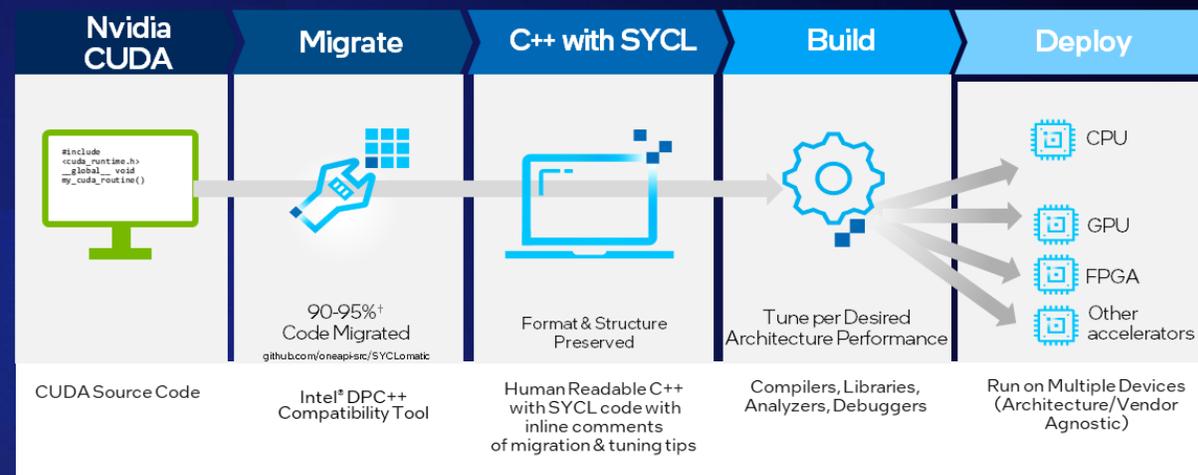
# CUDA から C++/SYCLへの移行

アーキテクチャごとに異なるコードを書いたり保守したりする必要はない



[CUDA to SYCL Migration Portal](#)

- 好みのアクセラレータを選択し、性能と移植性を備えたコードを再利用
- 単一のC++ with SYCLコードベースで、複数のベンダーの複数のアーキテクチャのアクセラレータ上で実行可能
- インテル® DPC++ Compatibility Toolとオープンソース SYCLomaticは、典型的なCUDAアプリケーションの約90~95%を自動的にSYCLに移行
- CUDA から SYCL への移行ポータルでは、チュートリアル、ベストプラクティス、コードサンプル、アプリケーションカタログ、コミュニティサポートを参照可能



## Migration Success Examples:



<sup>1</sup>Intel estimates as of March 2023. Based on measurements on a set of 85 HPC benchmarks and samples, with examples like Rodinia, SHOC, PENNANT. Results may vary.

\*Other names and brands may be claimed as the property of others. SYCL is a trademark of the Khronos Group Inc.

# Intel® Developer Cloud

最新のインテル製ハードウェアとソフトウェアでマルチアーキテクチャアプリケーションとテスト環境を構築

開発者向け

最新のインテル製CPUとGPUと、インテルに最適化されたAIソフトウェアへのアクセスし開発者に提供する

企業向け

インテル製品とテクノロジーの導入と展開を加速し、新しいソフトウェアとサービスを創出する

パートナー向け

パフォーマンスとコストを最適化したインテルAIコンピュート・サービスを顧客に提供する

無償利用のクラウドクレジットあり

## intel® Developer Cloud

Available software tools and optimizations



Available platforms



Visit [Intel® Developer Cloud](https://cloud.intel.com) at [cloud.intel.com](https://cloud.intel.com)

# 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

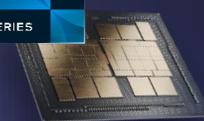
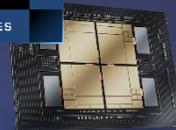
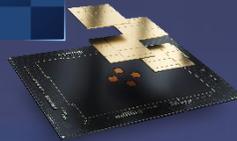
Scale

Open

Trusted

Choice

すべてのHPCとAIニーズに対応する製品群



The Intel logo is centered on a dark blue background. It features the word "intel" in a white, lowercase, sans-serif font. A small, light blue square is positioned above the letter "i". To the right of the word "intel" is a registered trademark symbol (®).

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](http://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 Service 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.