ハイパフォーマンス・ソフトウェア・カンファレンス 2023

インテル AI の「いま」
～アクセラレーターからソフトウェアまで～

インテル株式会社
AIセンター・オブ・エクセレンス
AIテクニカル・ソリューション・スペシャリスト
大内山　浩

intel®

注意事項および免責条項

性能は、使用状況、構成、その他の要因によって異なります。詳細については、http://www.Intel.com/PerformanceIndex/ (英語) を参照してください。性能の測定結果は、システム構成に記載された日付時点のテストに基づいています。また、現在公開中のすべての更新プログラムが適用されているとは限りません。構成の詳細については、http://www.Intel.com/InnovationEventClaims/ (英語) を参照してください。絶対的なセキュリティーを提供できる製品やコンポーネントはありません。

すべての製品計画およびロードマップは、予告なく変更されることがあります。 製品版出荷前のシステムやコンポーネントで測定された結果は、インテル・リファレンス・プラットフォーム (新しいシステムの社内サンプルモデル)、インテル社内の分析、アーキテクチャー・シミュレーション、モデリングでの推定 / シミュレーション結果を含め、情報提供のみを目的としています。システム、コンポーネント、仕様、構成に対する今後の変更によって、結果は異なる場合があります。インテルのテクノロジーを使用するには、対応したハードウェア、ソフトウェア、またはサービスの有効化が必要となる場合があります。

本資料は、明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず、いかなる知的財産権のライセンスも許諾するものではありません。 開発コード名は、一般向けに発表または出荷されていない製品やテクノロジー、サービスを識別するためにインテルによって使用されているものです。いずれも「商用」の名称ではなく、商標としての機能を前提としたものではありません。

インテルは、Principled Technologies が統括する BenchmarkXPRT 開発コミュニティーを含め、さまざまなベンチマーク制定団体 / 機関へのスポンサーとしての参画、また技術的サポートの提供によって、ベンチマークの開発に貢献しています。

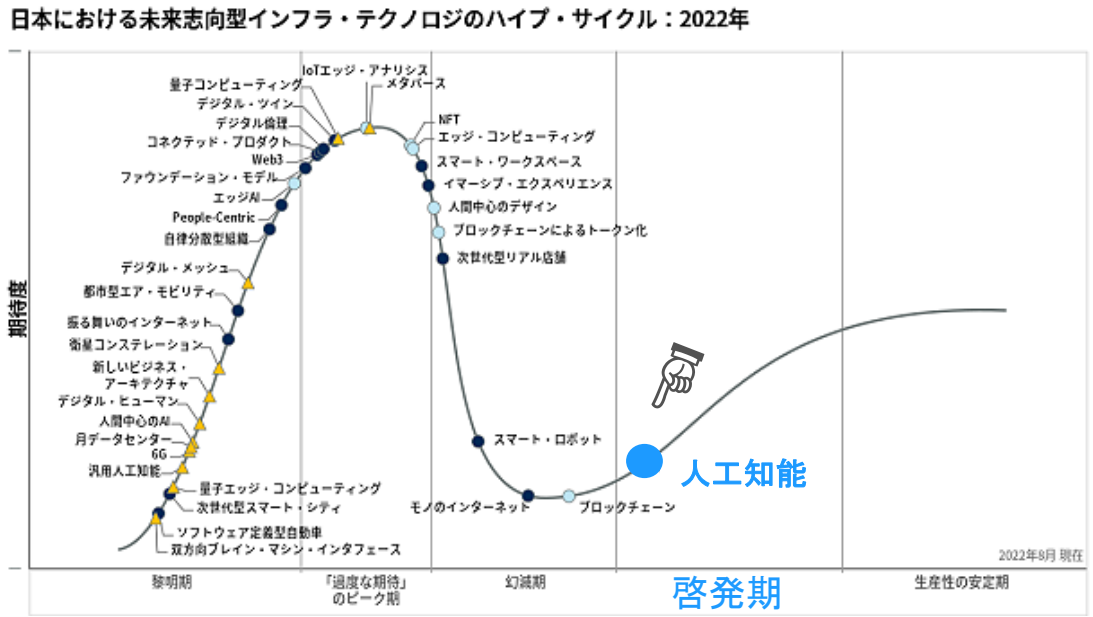
将来的な計画や予測について言及している本プレゼンテーション資料内の記述は、多数のリスクや不確定要素を伴う将来の見通しです。「想定される」、「見込まれる」、「意図する」、「目標とする」、「計画する」、「考えられる」、「求める」、「推定する」、「継続する」、「可能性がある」、「予定である」、「期待する」、「はずである」、「仮定する」などの表現やその変化形および類似表現は、いずれも将来の見通しであることを示しています。推定、予想、予測、不確定な事象、または仮定について言及している、またはこれらに基づいている記述も、今後リリースされる製品やテクノロジー、こうした製品やテクノロジーに期待される利用可能性とメリット、市場機会、インテルの事業および関連市場に見込まれるトレンドに関する記述を含め、将来の見通しであることを示しています。経営陣による現在の予測に基づくものであり、多数のリスクや不確定要素を伴う将来の見通しです。これらの要因によって、実際の結果はこれらの予測的記述に明示的または黙示的に示された結果と著しく異なる可能性があります。実際の業績をインテルの予測と大きく異ならせる重要な要因には、インテルの投資家向け IR サイト (https://www.intc.com/) または証券取引委員会 (SEC) のウェブサイト (https://www.sec.gov/) で入手可能な Form 10-K および Form 10-Q に関するインテルの最新の報告書を含め、インテルが SEC に提出 / 登録した報告書に記載されています。インテルは、法律で開示が義務付けられている場合を除き、新しい情報、新規開発、その他の結果にかかわらず、本プレゼンテーション資料に記載されたいかなる記述も、更新する義務を一切負わないことを明示的に表明します。

インテルは、サードパーティーのデータについて管理や監査を行っていません。ほかの情報も参考にしてデータの正確さを評価してください。

AIは「作る」から「使う」時代へ

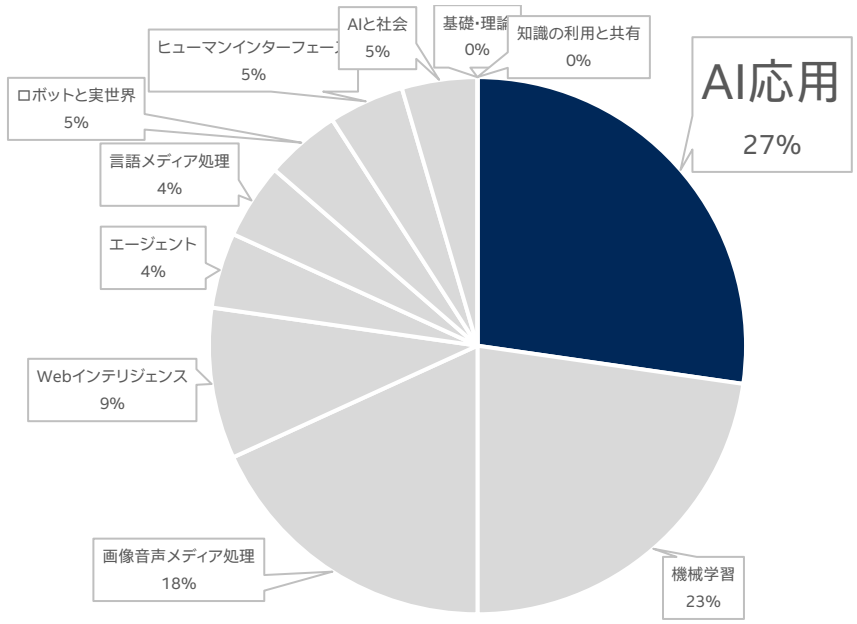
本格的な普及時代へと突入したAI

ガートナー･ハイプ･サイクル(2022年)



「日本における未来志向型インフラ･テクノロジのハイプ･サイクル：2022年」
(ガートナー ジャパン)

2022年 AI博士論文研究分野



人工知能学会誌 vol37 (2022年)

Democratizing AI for Everyone

オープン

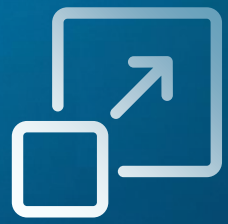
エコシステム

選択

互換性

信頼

ワークロード



スケール

デリバリー＆デプロイ

AIインフラに選択肢を

要件に応じた最適なH/Wを選択しTCOの最小化へ

CPU

GPU

Coming soon

ASIC

インテル® Xeon®プロセッサー製品のAI対応の歴史

第3次AIブーム開始
⇓

2012 ・・・ 2017 2018 2019 2020 2021 2022 2023~

New!

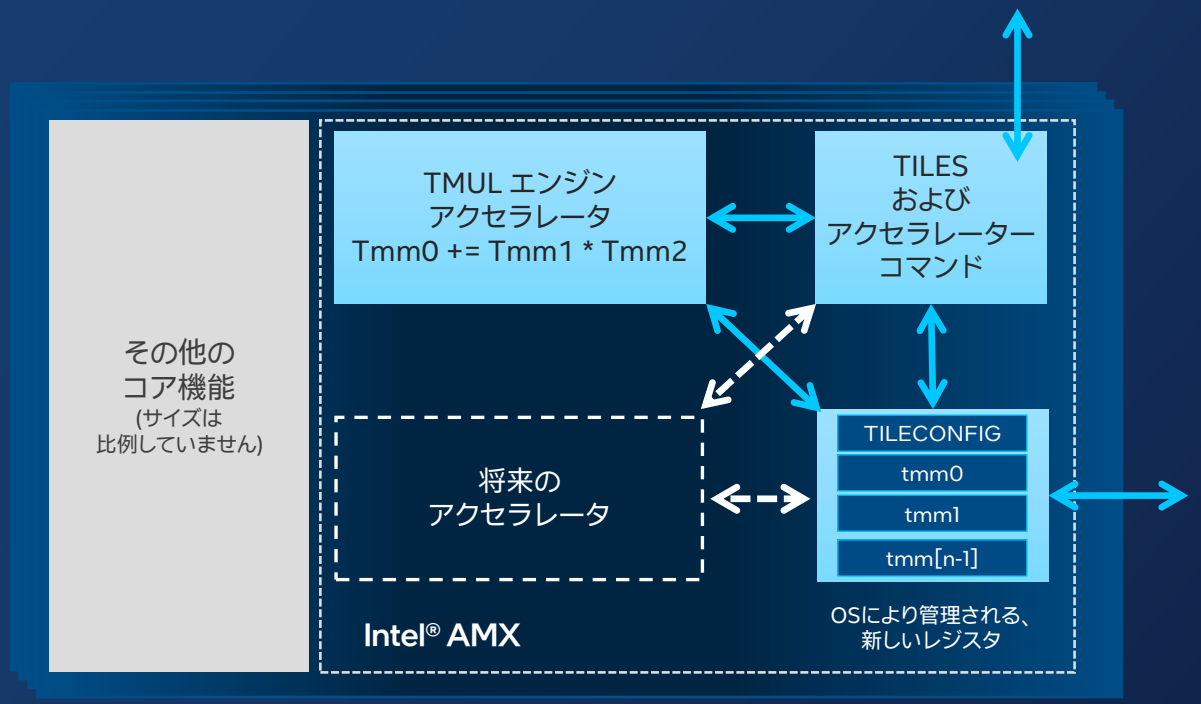
	2017	2019	2021	2023~
	第1世代インテル® Xeon® スケーラブル・プロセッサー（コードネーム：Skylake）	第2世代インテル® Xeon® スケーラブル・プロセッサー（コードネーム：Cascade Lake）	第3世代インテル® Xeon® スケーラブル・プロセッサー（コードネーム：Ice Lake）	第4世代インテル® Xeon® スケーラブル・プロセッサー（コードネーム：Sapphire Rapids）
	AVX-512 **(ベクトル演算高速化)**	**DL Boost** **(量子化による推論高速化)**	**前世代より1.5倍高速化**	**AMX** **(行列演算高速化)**
対象AIﾜｰｸﾛｰﾄﾞ	推論	推論	推論	学習&推論

インテル® アドバンスド マトリックス エクステンションズ (Intel® AMX)

第4世代 インテル® スケーラブル・プロセッサー各コアに内蔵されたDeep Learningアクセラレータ

データを
大きな塊で保存

大きな行列を**一つの処理**で計算可能な命令



サポートされるデータタイプ

BF16: 学習と推論

INT8: 推論

intel XEON Accelerate with Xeon

インテル® AMXを使うには特別なソフトウェアが必要?

いいえ、必要ありません。
インテルが最適化した下記OSSフレームワークをご使用ください。

フレームワーク	学習	推論
TensorFlow (V2.9~) + Intel® Extension for TensorFlow*	○	○
PyTorch (v2.0~) + Intel® Extension for PyTorch*	○	○
OpenVINO™ (v2022.3~)	–	○

PyTorchでResnet50の推論をする場合の例

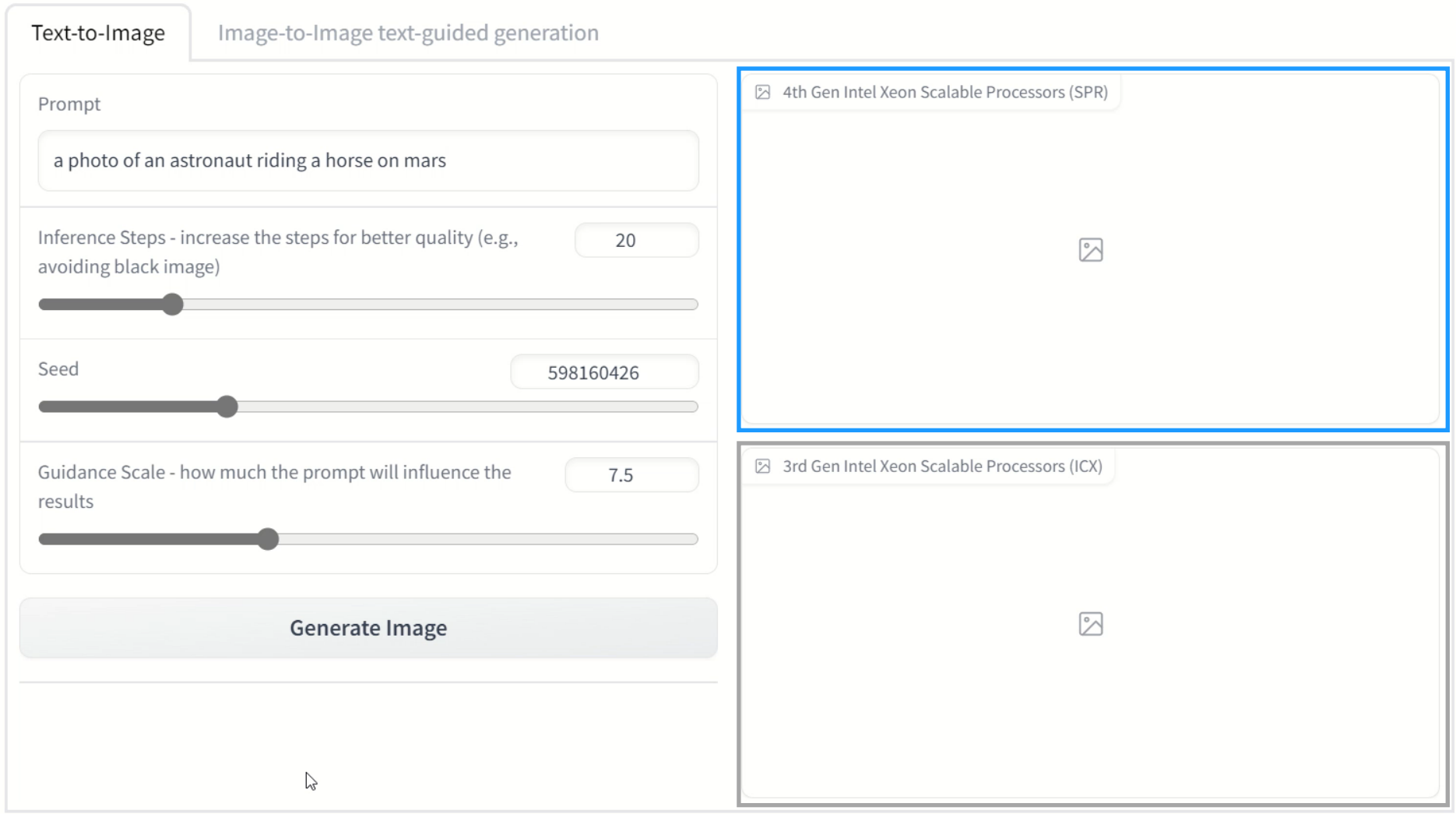
AMX無しで実行する場合（FP32のみ）

```
import torch
import torchvision.models as models

model = models.resnet50(weights='ResNet50_Weights.DEFAULT')
model.eval()
data = torch.rand(1, 3, 224, 224)

with torch.no_grad():
  model(data)
```

AMX有りで実行する場合（FP32+BF16）

```
import torch
import torchvision.models as models

model = models.resnet50(weights='ResNet50_Weights.DEFAULT')
model.eval()
data = torch.rand(1, 3, 224, 224)

with torch.no_grad():
  with torch.cpu.amp.autocast(dtype=torch.bfloat16):
    model(data)
```

Stable Diffusion を用いた画像生成デモ



第4世代Xeon
AMX＋AVX-512利用
(FP32+BF16)

第3世代Xeon
AVX-512のみ利用
(FP32)

https://huggingface.co/spaces/Intel/Stable-Diffusion-Side-by-Side

PyTorchでResnet50の学習をする場合の例

AMX無しで実行する場合（FP32のみ）

```
import torch
import torchvision

LR = 0.001
DOWNLOAD = True
DATA = 'datasets/cifar10/'

transform = torchvision.transforms.Compose([
  torchvision.transforms.Resize((224, 224)),
  torchvision.transforms.ToTensor(),
  torchvision.transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
train_dataset = torchvision.datasets.CIFAR10(
  root=DATA, train=True, transform=transform, download=DOWNLOAD,
)
train_loader = torch.utils.data.DataLoader(
  dataset=train_dataset, batch_size=128
)

model = torchvision.models.resnet50()
criterion = torch.nn.CrossEntropyLoss()
optimizer = torch.optim.SGD(model.parameters(), lr = LR, momentum=0.9)
model.train()

for batch_idx, (data, target) in enumerate(train_loader):
  optimizer.zero_grad()

  output = model(data)
  loss = criterion(output, target)
  loss.backward()
  optimizer.step()
torch.save({
   'model_state_dict': model.state_dict(),
   'optimizer_state_dict': optimizer.state_dict(),
   }, 'checkpoint.pth')
```

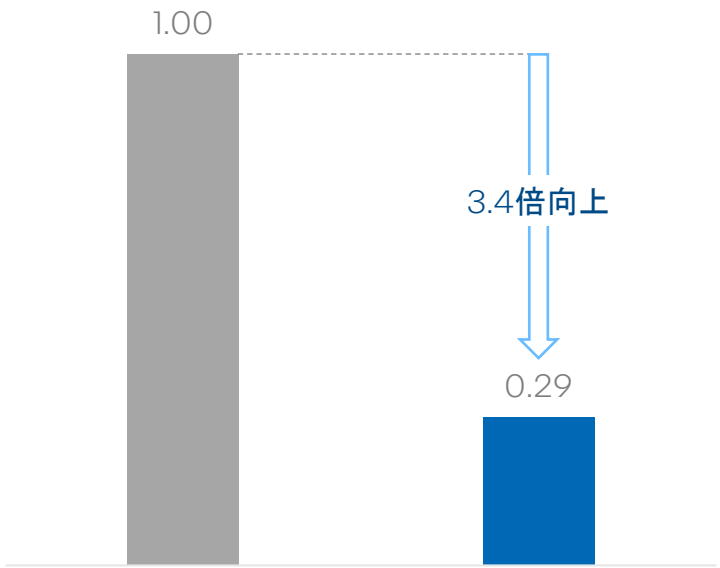
AMX有りで実行する場合（FP32+BF16）

```
import torch
import torchvision

LR = 0.001
DOWNLOAD = True
DATA = 'datasets/cifar10/'

transform = torchvision.transforms.Compose([
  torchvision.transforms.Resize((224, 224)),
  torchvision.transforms.ToTensor(),
  torchvision.transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
train_dataset = torchvision.datasets.CIFAR10(
  root=DATA, train=True, transform=transform, download=DOWNLOAD,
)
train_loader = torch.utils.data.DataLoader(
  dataset=train_dataset, batch_size=128
)

model = torchvision.models.resnet50()
criterion = torch.nn.CrossEntropyLoss()
optimizer = torch.optim.SGD(model.parameters(), lr = LR, momentum=0.9)
model.train()

for batch_idx, (data, target) in enumerate(train_loader):
  optimizer.zero_grad()
  with torch.cpu.amp.autocast(dtype=torch.bfloat16):
    output = model(data)
    loss = criterion(output, target)
    loss.backward()
  optimizer.step()
torch.save({
  'model_state_dict': model.state_dict(),
  'optimizer_state_dict': optimizer.state_dict(),
  }, 'checkpoint.pth')
```

第4世代インテル® Xeon® スケーラブル・プロセッサー モデル 学習性能

ResNet50 学習時間

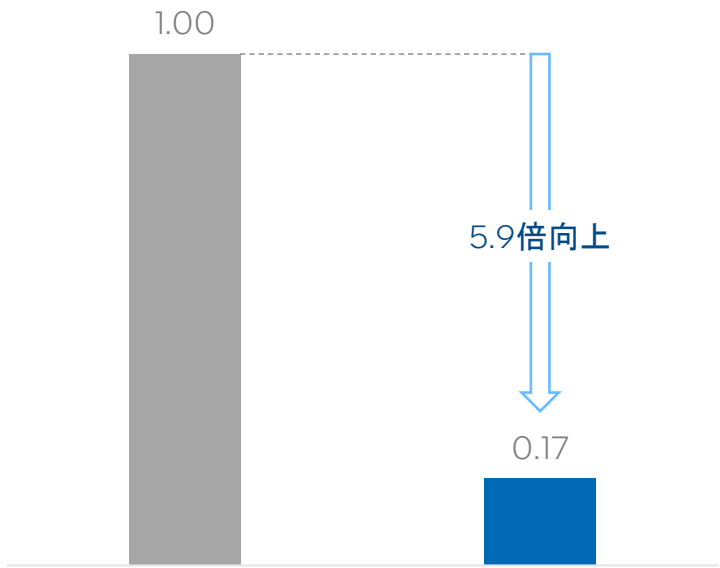
Lower is better



ソースコード：
https://intel.github.io/intel-extension-for-pytorch/latest/tutorials/examples.html

Mask R-CNN 学習時間

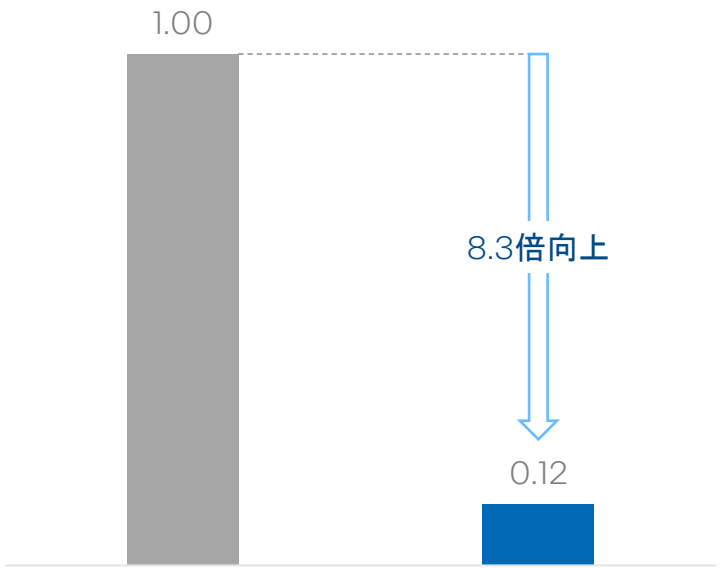
Lower is better



ソースコード：
https://github.com/IntelAI/models/blob/master/quickstart/object_detection/pytorch/maskrcnn/training/cpu

DistilBERT FT時間

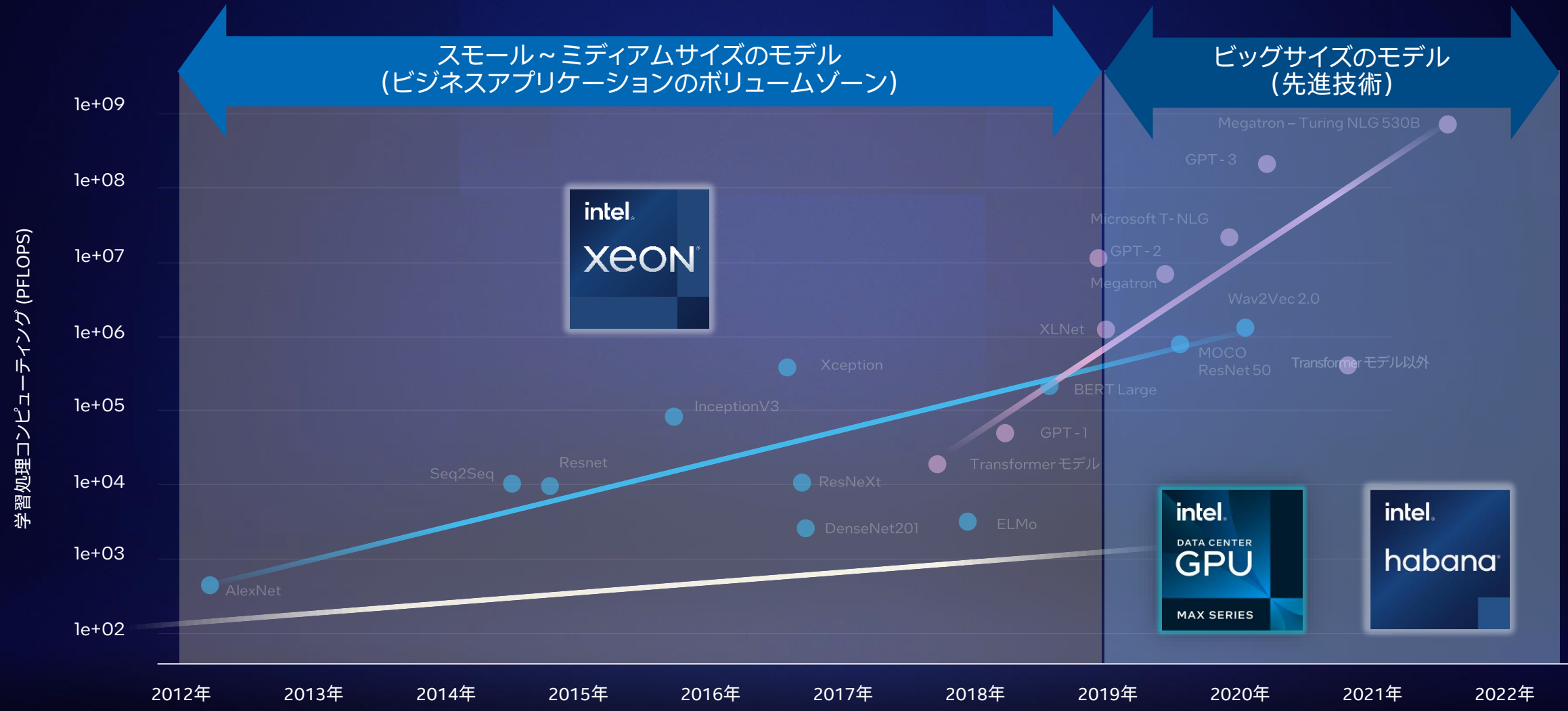
Lower is better



参照元：
https://huggingface.co/blog/intel-sapphire-rapids

注：本ベンチマークは個人によるものであり、インテルの公式結果ではありません。

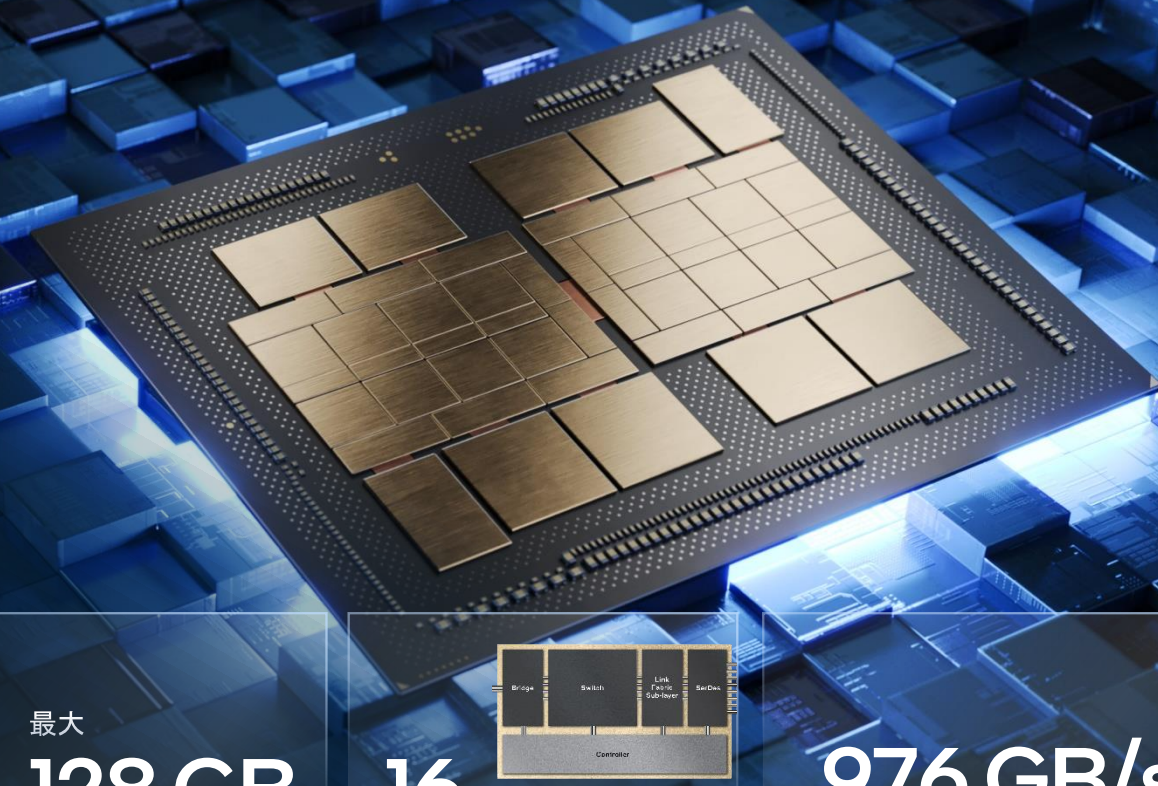
どういったモデルに適している?



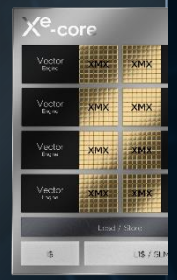
出典: Moore, S. (2022年)、「Nvidia's Next GPU Shows That Transformers Are Transforming AI」、IEEE Spectrum

intel
DATA CENTER
GPU
MAX SERIES

インテル® データセンター GPU マックス・シリーズ



最大
128
Xe HPC コア



52TF
ピーク FP64 スループット

839TF
ピーク BF16 スループット

最大
128 GB
HBM2e メモリー

16
Xe Links

976 GB/s
Xe Links を経由する GPUからGPUの 通信

intel
ISC High Performance
The HPC Event.

AI 対応モデル

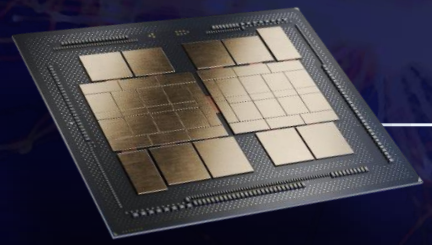
コンピュータービジョン 画像分類	ResNet-50 v15	ResNeXt-101	ResNet-101	EfficientNet-B7	SE-ResNeXt50	TSM	
	Adorym	CosmoFlow	RegNetY-32Y	ResNeXt-101	Candle Uno	Swin Transformer	
画像セグメンテーション	Cosmic Tagger	Mask R-CNN	DenseNet169	FFN	3D-Unet		
	PointNet	DeepCAM	DRN-D-54	ResNeXt3D-101			
物体検出	SSD-ResNet34	SSD-ResNet50	EfficientDet	ShuffleNet	YOLO-v3	YOLO-v4	RetinaNet-ResNet50
	Deep Fusion	CascadeRCNN-	MobileNet v3	SSD-MobileNet	MMA	ResNet101-FPN	
NLP 言語モデル	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			
音声生成	FastSpeech2	Tacotron-2 with LPCNet					
レコメンデーション	DLRM	DSSM	ESSM	Wide & Deep	DeepFM		
	DIN	AttRec	DIEN	MMOE			

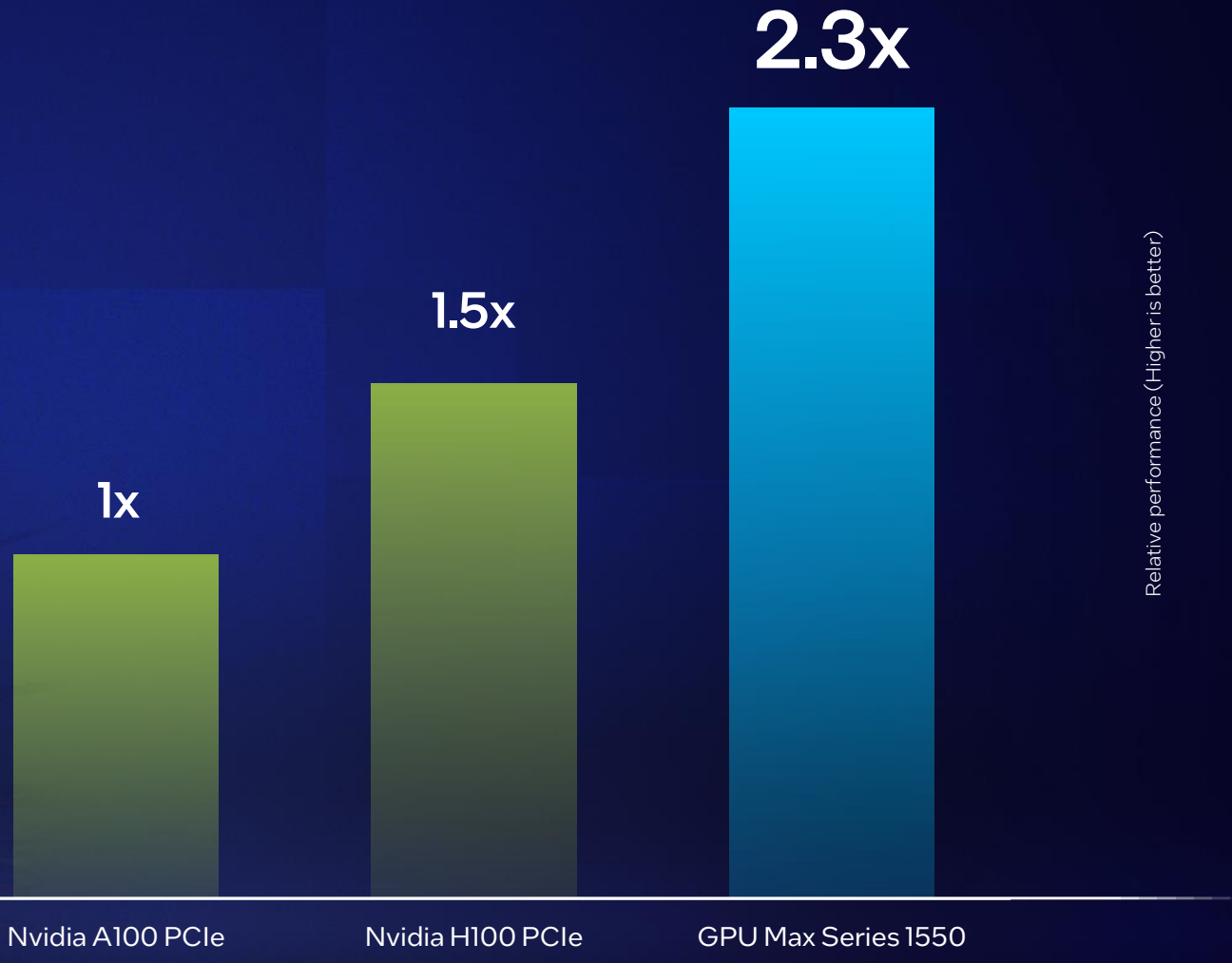
intel

intel DATA CENTER GPU MAX SERIES

“粒子の通り道”の解明をより早く

CERN 3D-GAN (BS= 256)





intel ISC High Performance The HPC Event.

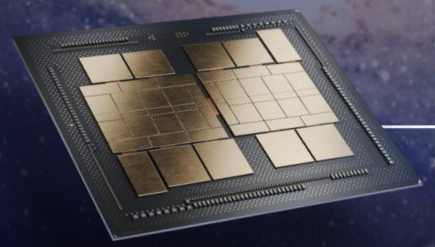
Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

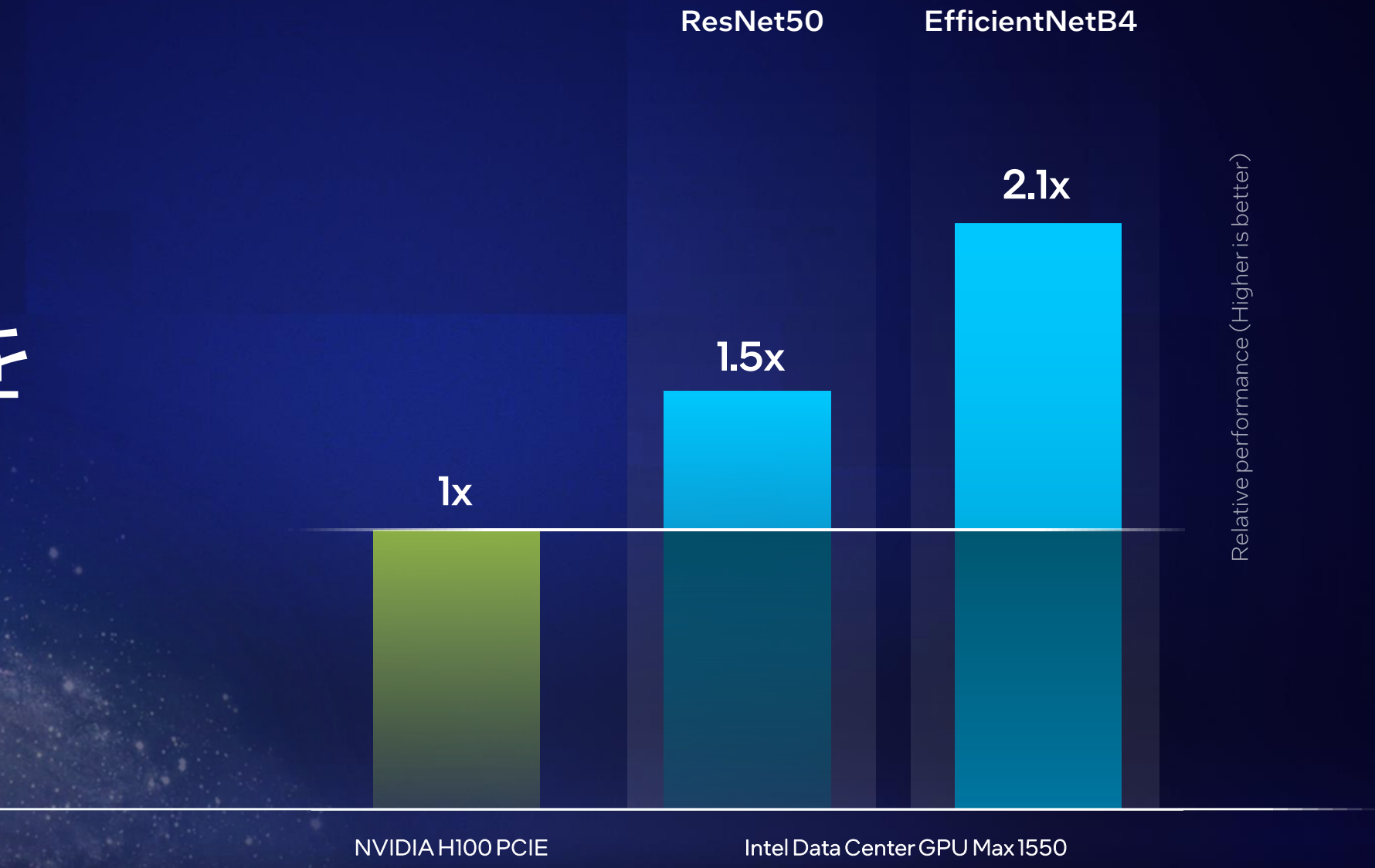
Get story details from Maxwell

intel DATA CENTER GPU MAX SERIES

"光速"で銀河の種類を分類

on DeepGalaxy





intel

ISC High Performance The HPC Event.

See backup for workloads and configurations. Results may vary.

AIインフラに選択肢を

要件に応じた最適なH/Wを選択しTCOの最小化へ

CPU

GPU

Coming soon

ASIC

多様なハードウェアをシームレスに行き来する

オープン・ソース・ソフトウェアエコシステムの拡充

インテル® oneAPI ベース・ツールキット

SynapseAI® Software Suite

CPU

GPU

ASIC

Coming soon

TensorFlow*／PyTorch*用のインテル® 拡張ライブラリー

インテル® チップの性能をより引き出すために

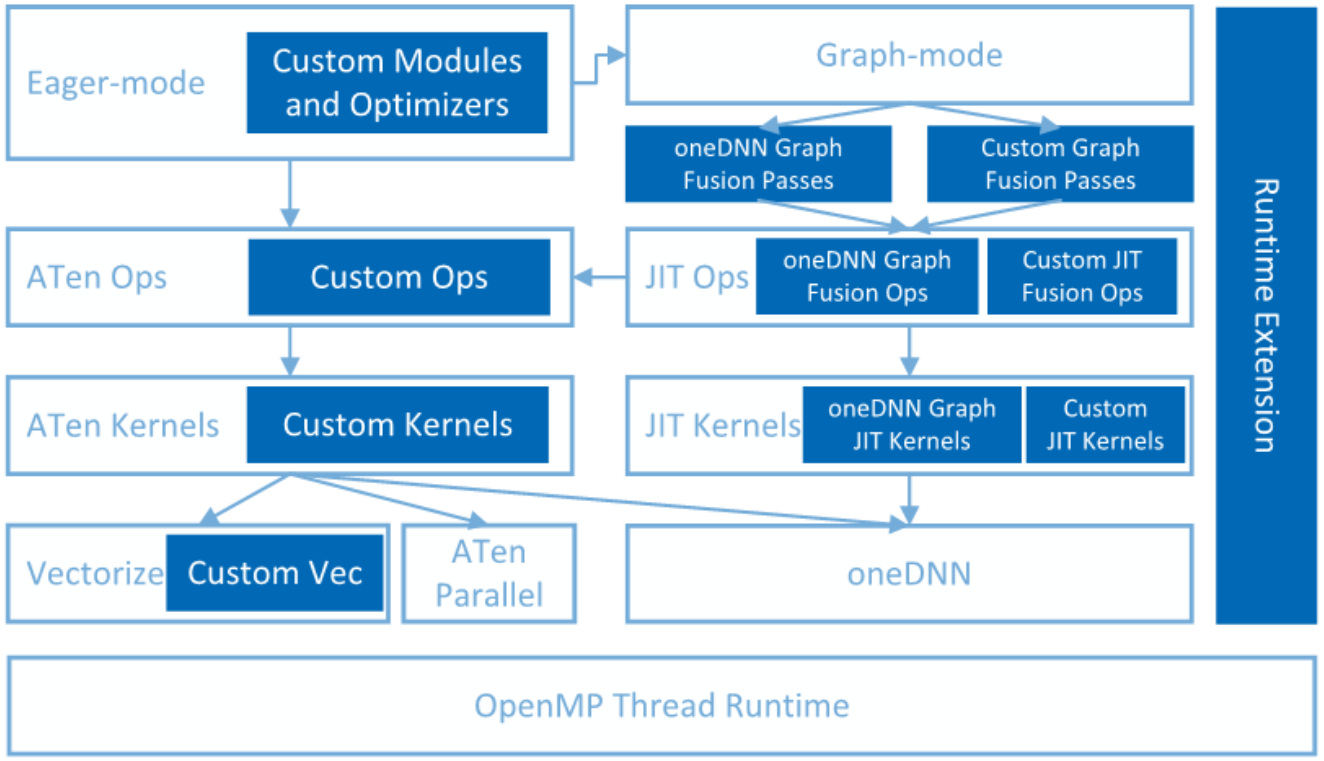
TensorFlow

Intel® Extension for TensorFlow*
(ITEX)

PyTorch

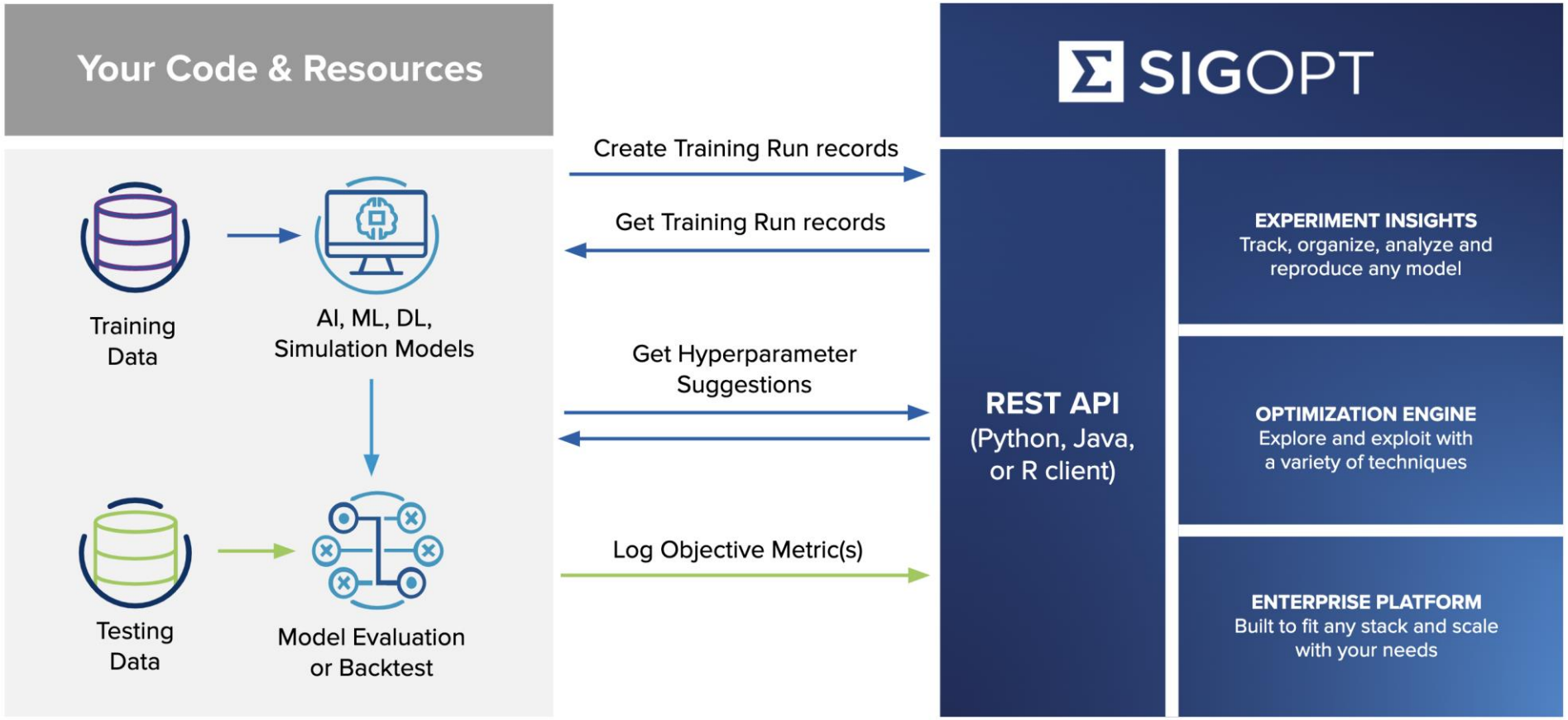
Intel® Extension for PyTorch*
(IPEX)

参考：Intel® Extension for PyTorch*のアーキテクチャー



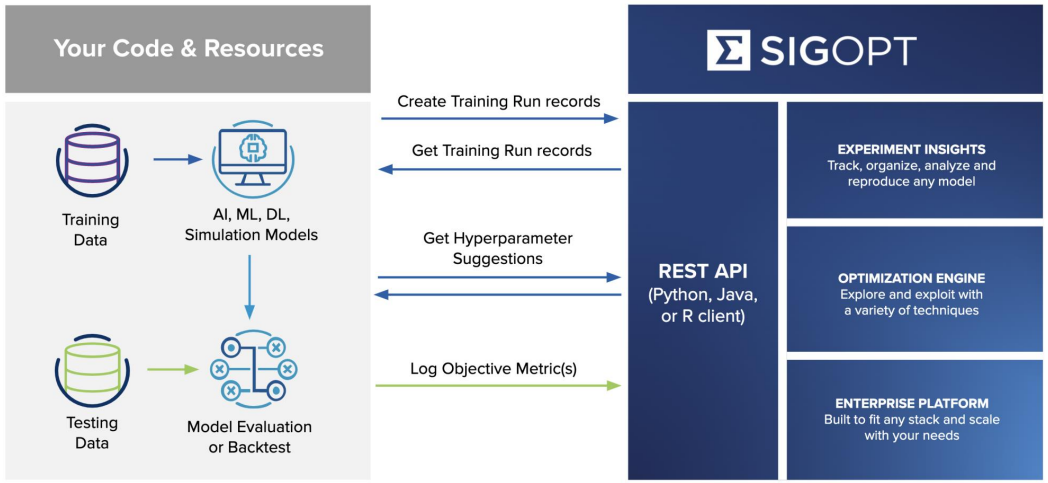
SigOpt

SaaS型ハイパーパラメータチューニングエンジン



SigOpt

SaaS型ハイパーパラメータチューニングエンジン



オープンソース化

SigOpt Github 検索

多様なハードウェアをシームレスに行き来する

オープン・ソース・ソフトウェアエコシステムの拡充

インテル® oneAPI ベース・ツールキット

SynapseAI® Software Suite

CPU

GPU

ASIC

Coming soon

実践的なAI知見のオープン化

インテル® AI リファレンス・キット

Transformers

DeepSpeed

TensorFlow

OpenVINO™

PyTorch

oneAPI

インテル® oneAPI ベース・ツールキット

SynapseAI® Software Suite

CPU

GPU

ASIC

intel Xeon

intel Data Center GPU Max Series

Coming soon

intel habana

インテル® AI リファレンス・キット

Intel AI ref kit | 検索

アクセンチュアと共同で開発した、
多様なAIユースケースにてより迅速にAIを開発するための実践手引き

金融・保険
- クレーム文書の自動化
- クレジットカード取引における不正検知
- デフォルトリスク予測
- 災害時の鑑定プロセス

プロセスオートメーション
- ビジュアル・プロセス・ディスカバリー
- インテリジェント・ドキュメント・インデクシング
- インボイス・トゥ・キャッシュの自動化
- 過去の資産の文書処理

製造・ユーティリティ
- ドローンナビセグメンテーション
- 電力線故障検出
- 予測的資産分析
- 外観品質検査
- デジタルツイン
- エンジニアリング・デザイン
- 交通カメラの物体検出
- 空力 / 流体プロファイリング

ヘルス&ライフサイエンス
- 医用画像診断
- 疾患予測
- セラピスト向けAIトランスクライブ

カスタマーケア
- 購買予測
- 顧客セグメンテーション
- 顧客解約予測
- カスタマーケア・チャットボット
- 需要予測
- 商品リコメンデーション
- 受注から納品までの予測

技術・セキュリティ
- バーティカル・サーチ・エンジン
- ネットワーク侵入検知
- データ保護
- IoT(データストリーミング異常検知)

合成データ
- AI合成データ(構造化)
- AI合成データ(非構造化-テキスト)
- AI合成データ(非構造化-画像)
- AI合成データ(非構造化-音声)

1. More info at **https://www.intel.com/aireferencekit**
2. Downloads available from GitHub at **https://github.com/oneapi-src**

AIの民主化を促進するインテルAI製品

インテル® AI リファレンス・キット

Transformers

DeepSpeed

TensorFlow

OpenVINO™

PyTorch

oneAPI

インテル® oneAPI ベース・ツールキット

SynapseAI® Software Suite

CPU

GPU

ASIC

intel XEON

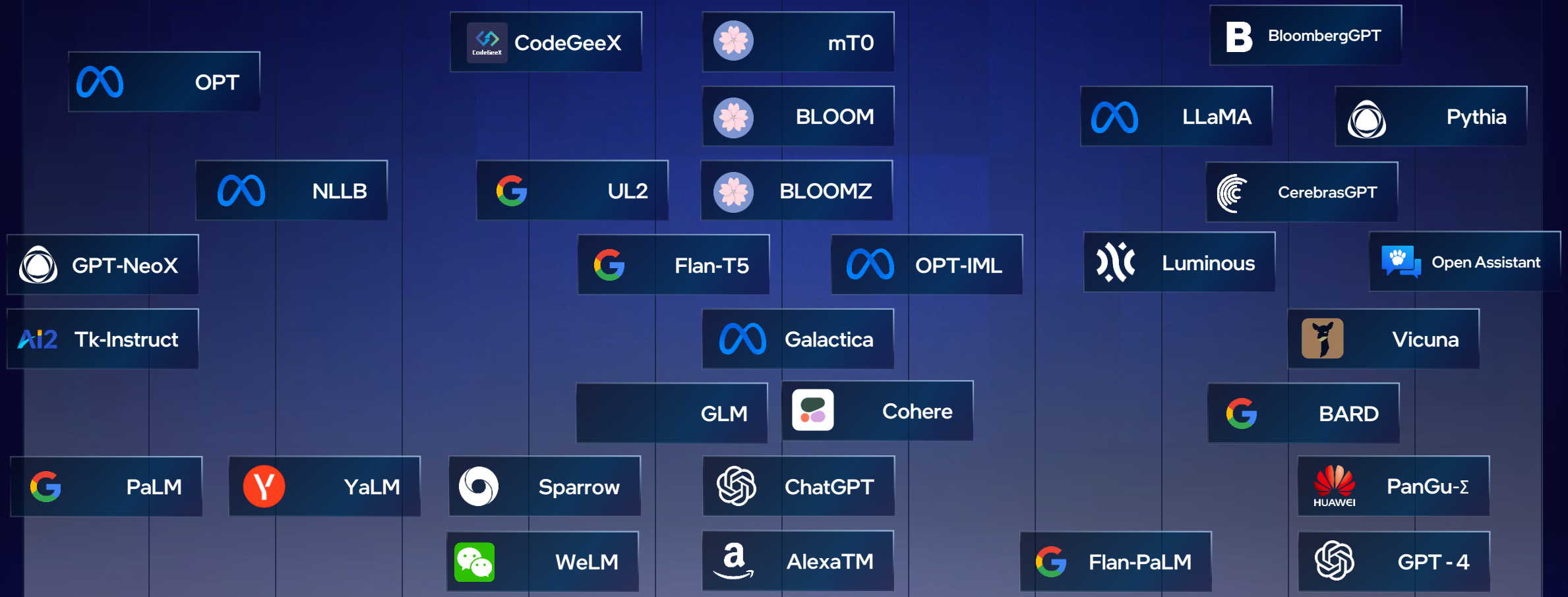
intel DATA CENTER GPU MAX SERIES

Coming soon

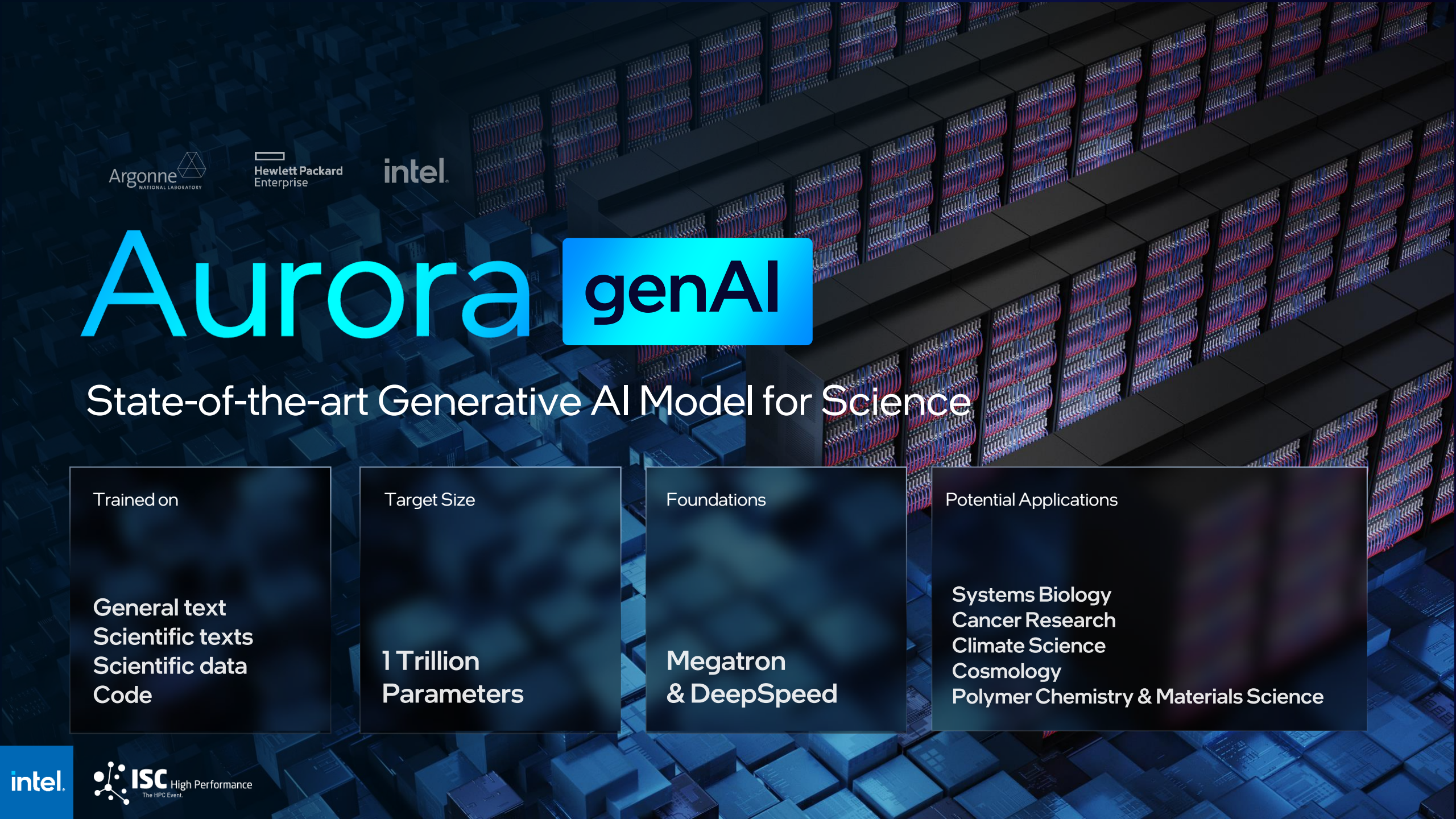
intel habana

大規模言語モデル（LLM）の台頭

大規模言語モデル（LLM）の台頭

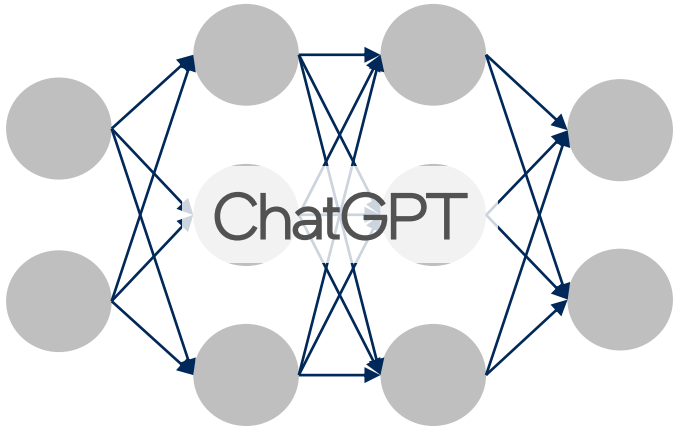


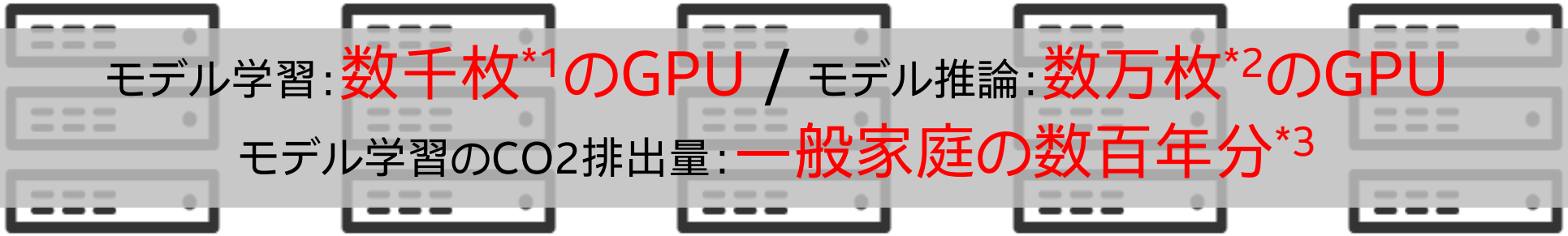
Source: Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.


Argonne NATIONAL LABORATORY
Hewlett Packard Enterprise
intel
Aurora genAI
State-of-the-art Generative AI Model for Science
Trained on
General text
Scientific texts
Scientific data
Code
Target Size
1 Trillion Parameters
Foundations
Megatron & DeepSpeed
Potential Applications
Systems Biology
Cancer Research
Climate Science
Cosmology
Polymer Chemistry & Materials Science
intel
ISC High Performance
The HPC Event.

LLM～インフラ＆カーボンフットプリント～

例：ChatGPT





モデル学習：数千枚*1のGPU / モデル推論：数万枚*2のGPU

モデル学習のCO2排出量：一般家庭の数百年分*3

*1 https://www.youtube.com/watch?v=4q9-yf1eU8c&t=952s
*2 https://www.semianalysis.com/p/the-inference-cost-of-search-disruption
*3 https://aiindex.stanford.edu/report/

LLMのインフラをめぐる動き

テクノロジー　2023年4月19日 / 1:26 午前 / 1ヶ月前更新

米マイクロソフト、独自のＡＩチップ開発へ＝報道

ロイター編集　　1 分で読む

米マイクロソフトが、「チャットＧＰＴ」のような人工知能（ＡＩ）を利用したチャットボット（自動応答システム）向けの独自のＡＩチップ「Ａｔｈｅｎａ」を開発している。テクノロジー情報サイトのジ・インフォメーションが１８日、関係者の話として報じた。２０２１年７月撮影（２０２３年　ロイター/Dado Ruvic/Illustration/File Photo/File Photo）

*1

TECH DRIVERS

Google reveals its newest A.I. supercomputer, says it beats Nvidia

PUBLISHED WED, APR 5 2023•1:12 PM EDT | UPDATED WED, APR 5 2023•2:55 PM EDT

Kif Leswing
@KIFLESWING

WATCH LIVE

KEY POINTS

- Google published details about its AI supercomputer on Wednesday, saying it is faster and more efficient than competing Nvidia systems.
- While Nvidia dominates the market for AI model training and deployment, with over 90%, Google has been designing and deploying a chip, called Tensor Processing Unit, for artificial intelligence since 2016, partially for internal use.

In this article

NVDA +0.76 (+0.20%)　GOOGL +0.21 (+0.17%)

Follow your favorite stocks
CREATE FREE ACCOUNT

Google headquarters is seen in Mountain View, California, United States on September 26, 2022.
Tayfun Coskun / Anadolu Agency / Getty Images

*2

SambaNovaが生成AI機能を企業向けに提供

新しいSambaNova Suiteにより最も正確な事前学習済みの生成AIモデルとオープンソースモデルを提供

SambaNova Systems Japan合同会社
2023年3月1日 09時00分

2023年2月28日、カリフォルニア州パロアルト – 生成AIを支える事前学習済みファンデーション（基盤）モデルを初めて市場に投入したSambaNova Systemsは、オンプレミスまたはクラウドで展開される、企業向けの最も精度の高い生成AIモデルを集めた新しいSambaNova Suiteを発表します。

「どの分野であれ、企業が競争力を持ち続けるためには、生成AI機能を備えた独自の基盤モデルが必要になります。新しいSambaNova Suiteは、オンプレミスまたはクラウドで、既存のワークフローやアプリケーションに安全かつオープンに組み込まれ、企業全体で生成AIを迅速に導入することができます」と、SambaNova Systemsの製品担当上級副社長のMarshall Choy氏は述べています。

SambaNova Suiteは、最先端のオープンソースモデルと、SambaNovaによって事前学習されたSambaNova GPTのようなモデルの両方を提供し、いずれも企業のデータを使ってファインチューニングすることでより高い精度を得ることができます。SambaNova SuiteはシンプルなAPIで簡単にアクセスでき、既存のビジネスプロセスと統合することですぐに価値を提供します。

「生成AIは、世界の注目、興奮、想像力を集めています。いまSambaNovaは企業向けに精度の高い生成AIモデルを最適化し、企業が独自のデータをモデルに適用し、クラウドやデータセンターなどあらゆる場所で展開することを可能にしています。これにより生成AIは単なる誇大宣伝や興奮から、企業に真の価値を提供するものへと進化することができるのです」とIDCで人工知能・自動化マーケットリサーチを担当するグループ副社長のRitu Jyoti氏は述べています。

Samba Suiteは以下の機能を提供します：
企業向けに最適化された、**最も精度の高い事前学習済み生成AIモデル**

- 競合他社に先駆けて最新モデルの恩恵を受けることができる**最新のオープンソースモデル**
- お客様の環境、お客様自身のデータでファインチューニングされ、最高レベルのセキュリティ、柔軟性、機密保持を実現する、**お客様によるモデルのオーナーシップ**
- 他のクラウドベースの代替製品よりも柔軟なドメイン適応とデータガバナンスを実現する、モデル学習のための**専用モデルバックボーン**
- 拡張性があり、AI導入のさまざまな段階にある企業のニーズに対応する、**エンタープライズクラスのセキ**

*3

*1 https://jp.reuters.com/article/microsoft-ai-idJPKBN2WF1DH
*2 https://www.cnbc.com/2023/04/05/google-reveals-its-newest-ai-supercomputer-claims-it-beats-nvidia-.html
*3 https://prtimes.jp/main/html/rd/p/000000005.000106960.html

LLMインフラにも選択肢を

要件に応じた最適なH/Wを選択しTCOの最小化へ

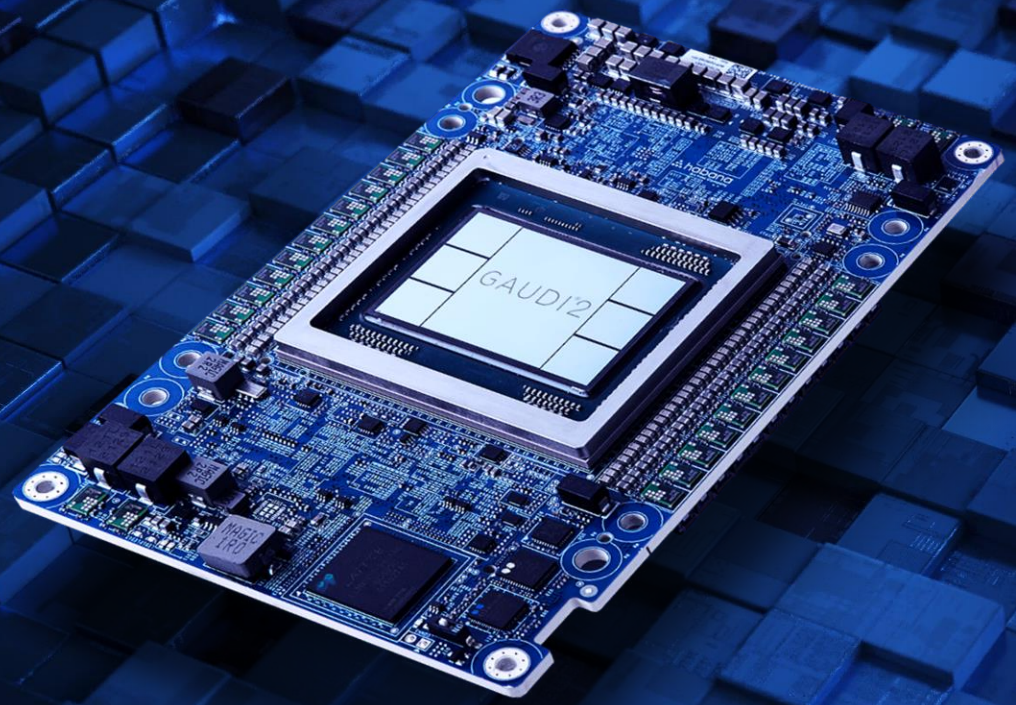
CPU

GPU

Coming soon

ASIC

GAUDI®2



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

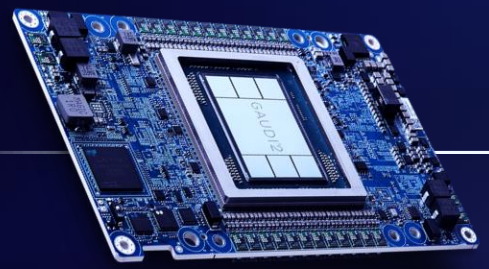
Availability

クラウド	Intel Developer Cloud
オンプレミス	Supermicro Gaudi2 Server

intel

ISC High Performance
The HPC Event.

Advantage across numerous performance metrics

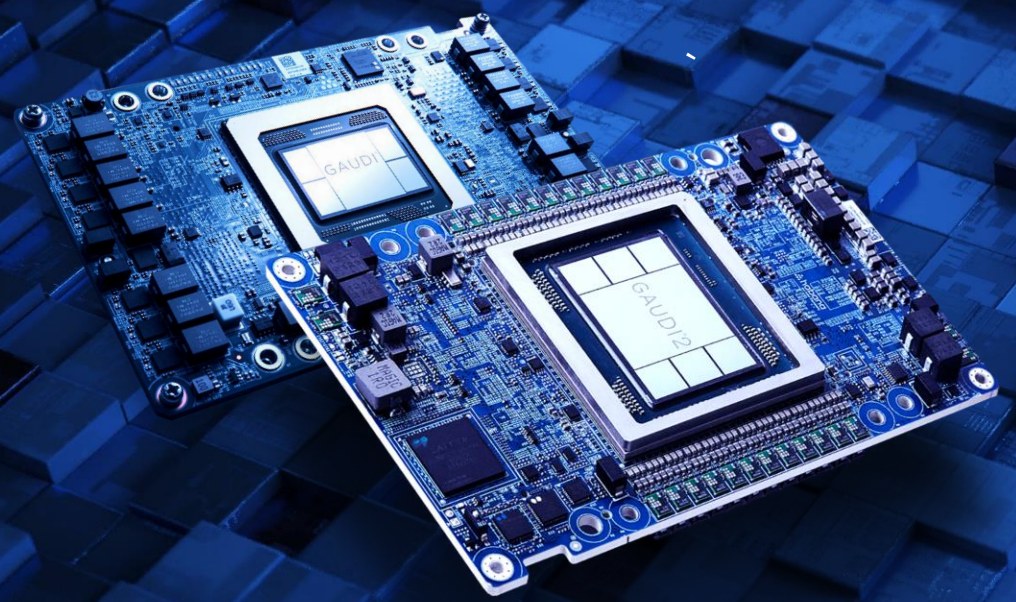


Visit www.intel.com/performanceindex for workloads and configurations. Results may vary

https://huggingface.co/blog/habana-gaudi-2-benchmark , https://habana.ai/habana-claims-validation

intel habana

生成AIの次なる波を後押し



大規模スケール
インテルAIスーパーコンピューティングクラスターでの学習

高速な推論スピード
1760億パラメーターのBLOOM LLMにおけるHabana Gaudi2

リアルタイムマルチモーダルセマンティック検索
Intel Developer Cloud上のGaudi 2での実行

120億パラメータのDolly 2.0がGPT-3並みの性能を発揮
Habana Gaudi 2で実行中

テキストを没入感のある3Dに変換する新型モデル
インテルAIスーパーコンピューターで学習

intel ISC High Performance The HPC Event.

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

LLMの民主化を体現するための共同事例

BCGとインテルにて共同開発

- 生成AIの学習
  - BCGの約50年分の業務データを使用
  - Habana Gaudi2を学習インフラで使用
- デプロイメント
  - BCG社員向けに情報検索サービスとして展開
- 結果
  - 社員満足度：41% UP
  - 作業効率：39% UP

https://www.bcg.com/ja-jp/press/10may2023-intel-bcg-announce-collaboration-enterprise-grade-secure-generative-ai


intel
Developer Cloud
cloud.intel.com
GAUDI®2
intel
XEON
intel
XEON
MAX SERIES
intel
DATA CENTER
GPU
MAX SERIES

Democratizing AI for Everyone

オープン

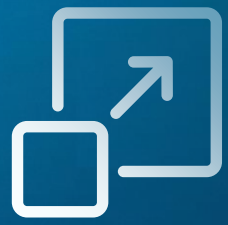
エコシステム

選択

互換性

信頼

ワークロード



スケール

デリバリー＆デプロイ

ご清聴ありがとうございました

intel


intel