



# インテル®の AI 製品の最新情報

インテル株式会社 戸谷 大介



intel ai

# インテルのAI ポートフォリオ

スケーラブルな  
AI コンピューティン  
グ・プラットフォーム



## AI ソフトウェア & サービス

intel ai  
ソフトウェア

intel  
Tiber AI クラウド

OpenVINO™

1  
oneAPI

オープンソース & エンド・ツー・エンドの AI ワークフロー エンタープライズ・ソフトウェア製品



AI PC



生産性、創造性、セキュリ  
ティにおける新しい AI 体験



エッジ AI



リアルタイムなデータ処理やアクショ  
ンのための高性能なコンピューティン  
グ



データセンター  
& クラウド AI



オープンで入手性の高い計算スタック  
で構築されたスケーラブルなシステム



## AI ネットワーク

intel  
ETHERNET

シームレスな AI 体験のためのスケーラブルなクラウド・ツー・エッ  
ジ・ツー・クライアント・ネットワーク

# AI ワークロードごとに最適な計算機システム

## 大規模で AI に特化

AI が **主要な** ワークロード

汎用処理のサイクル

AI 処理のサイクル



AI に特化したアプリケーションやサービス

GPU や AI アクセラレーター  
によるクラスター

## “General-Purpose (汎用な)” AI

AI は多数のワークロードの1つ



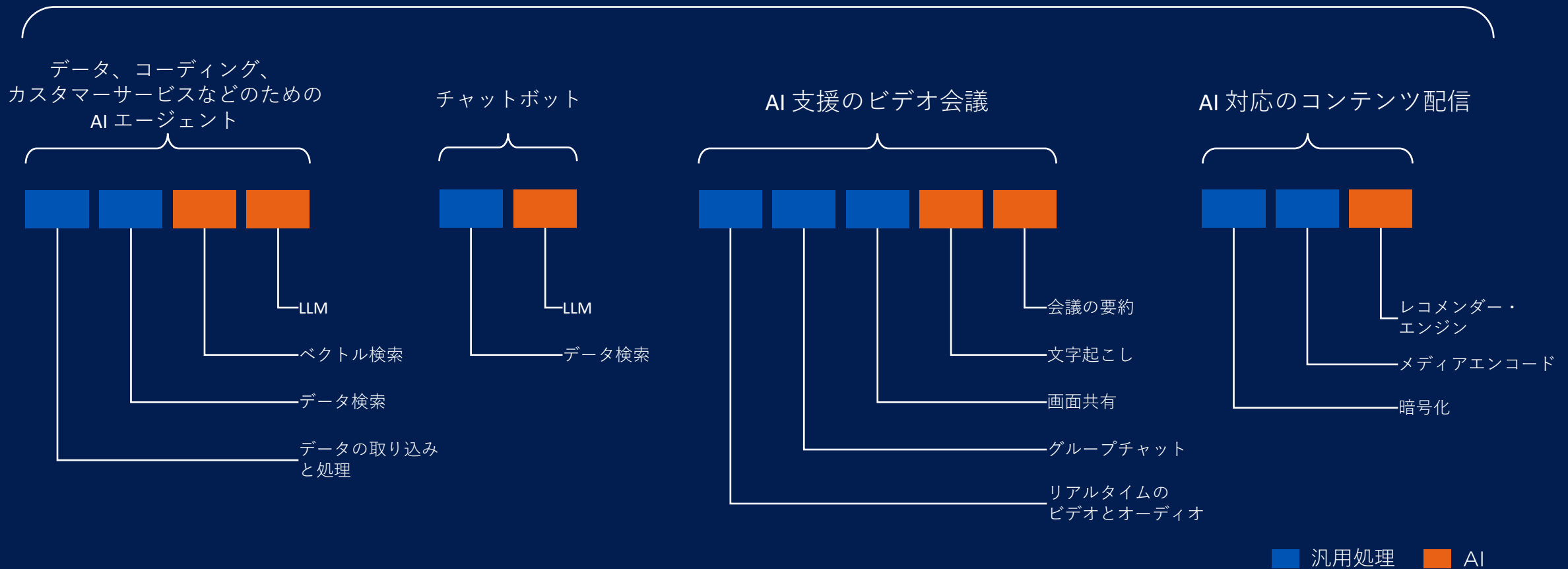
同じインフラ上で実行される複数のワークロード (AI を含む)

企業規模で CPU 上に構築 & 運用

# AI ワークロードごとに最適な計算機システム

AI を同じインフラ上で実行されるワークロードの1つとして扱いシステムの複雑化を回避

既存の **x86** インフラストラクチャーを活用し、AI および非 AI のエンタープライズ・ワークロードの**両方**を同じ基盤上で実行



# RAG (検索拡張生成) により企業のデータを活用

企業内のデータ

セキュリティー  
と機密性

データの  
近くでの処理

成熟し  
予測可能

CPU ベース

現在

## 企業内の データ

成熟性が高く、より  
安全で機密性が高い

x86 CPU ベース



## RAG

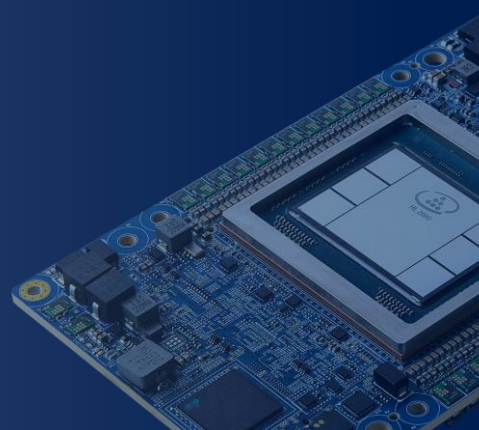
埋め込み & ベクトル化

将来

## AI モデル

急速に進化し、  
非常に破壊的

アクセラレーター・ベース



AI モデル

現在パブリックな  
データに基づく

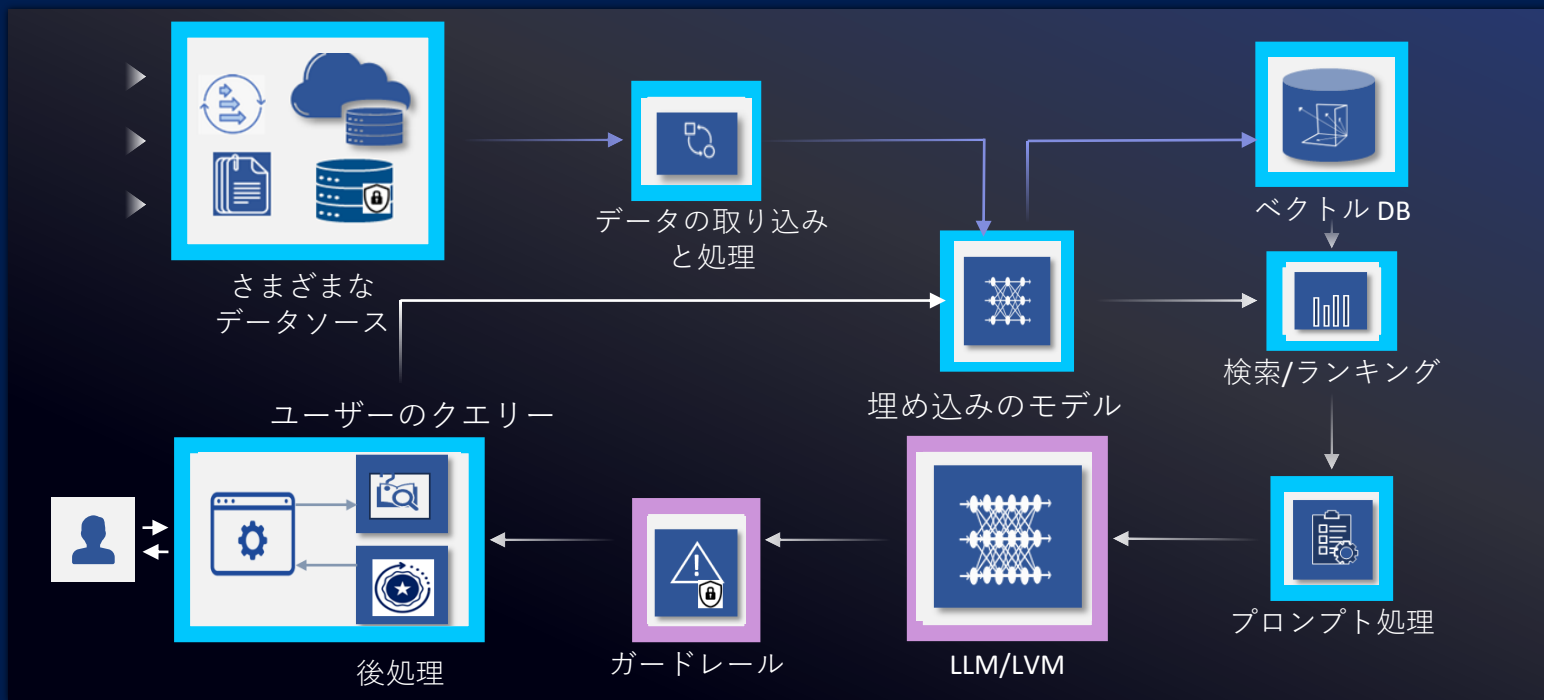
オープン/クローズド

急速な変化

アクセラレー  
ター・ベース

# Open Platform for Enterprise AI OPEA

- 生成 AI のシステムに必要なコンポーネントのスタック構造とエンドツーエンドのワークフローのアーキテクチャー設計図
- 断片化されたエコシステムの複雑さを軽減し、ソリューションの生産規模を拡大
- Linux Foundation とのパートナーシップにより、業界リーダー間のコラボレーションと貢献を促進



# インテル® Xeon® プロセッサー

AI 向けに設計された CPU



# インテル® Xeon® 6 プロセッサ Performance-core (P-core) 製品 6900P シリーズ

CPU 当たり 最大 128 コア

メモリー帯域幅を増加

MRDIMM によりメモリー帯域幅を  
最大 2.3 倍に

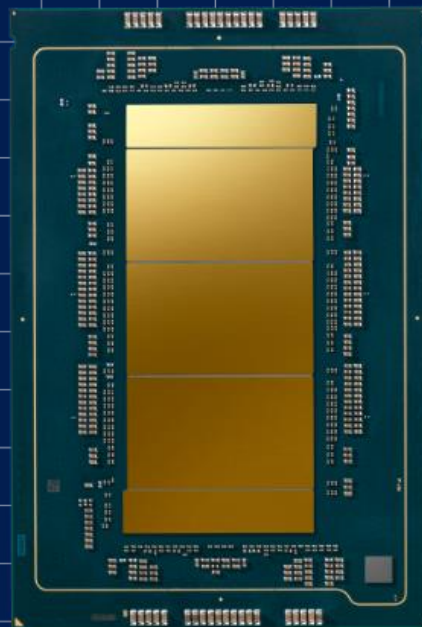
vs 第 5 世代 インテル® Xeon® スケーラブル・プロセッサ<sup>1</sup>

LLCを増加

L3 キャッシュは最大 504MB で、  
大きな L3 アクセスサイズでも非常に  
低いレイテンシーを実現

インテル® AI ソフトウェア

GenAI、エッジ・デプロイメント、  
機械学習にわたる AI 開発のための  
多様なツールが利用可能



インテル® アドバンスト・マトリクス・  
エクステンション (インテル® AMX)

FP16 ベースのモデルで AI パフォーマンス  
を向上

インテル® アドバンスト・ベクトル・  
エクステンション 512 (インテル® AVX-512)

固有の命令セットとコアごとに 2 つの  
512 ビット FMA ユニットを搭載

インテル® アドバンスト・ベクトル・エ  
クステンション 2 (インテル® AVX2)

新しい VNNI 命令と BF16 および FP16 の高  
速なアップ/ダウン変換

インテル® Xeon® 6 プロセッサ + インテル® Gaudi® アクセラレーター

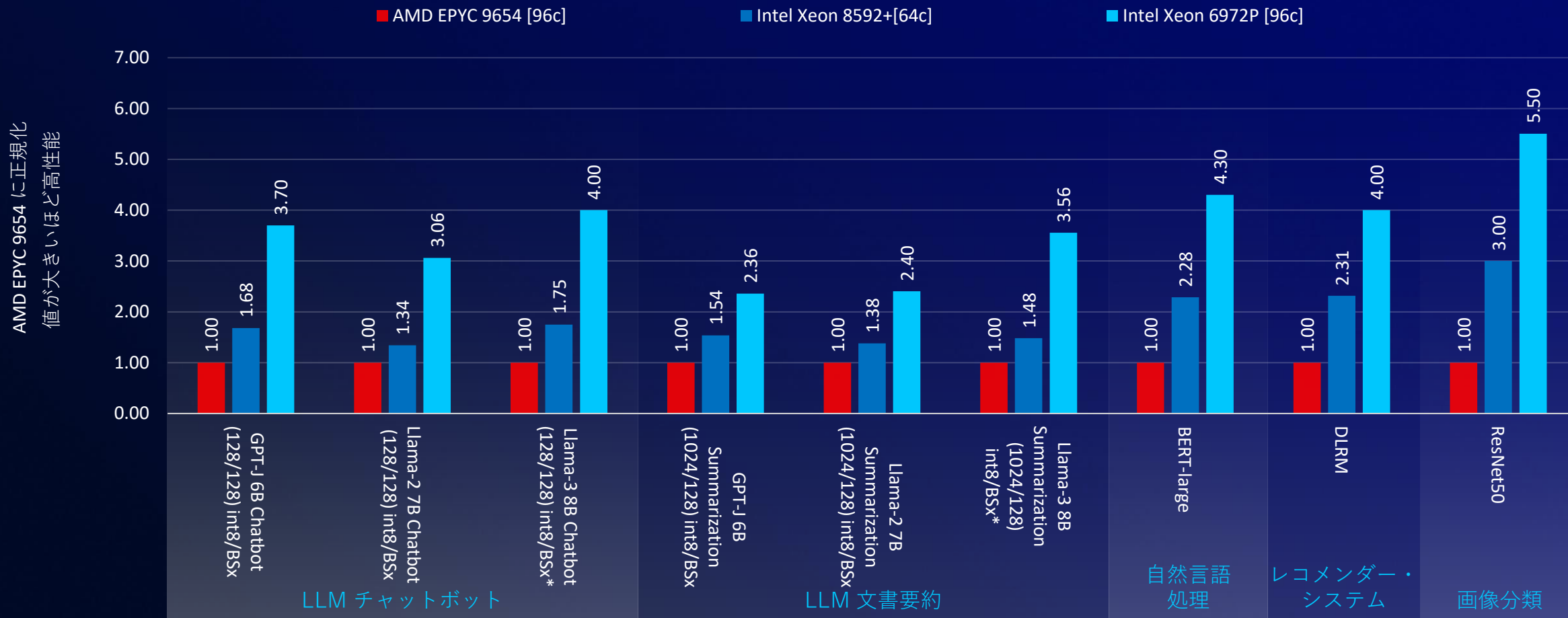
P-core を搭載したインテル® Xeon® 6 プロセッサはインテル® Gaudi® アクセラレーターと完全に検証  
次世代の AI トレーニングのための容易にスケラブルなソリューションを提供

<sup>1</sup>ワークロードと構成については、[intel.com/performanceindex](https://www.intel.com/performanceindex) の ISC 2024 セクションをご覧ください。結果は異なる場合があります。インテルの技術には、対応するハードウェア、ソフトウェア、またはサービスの有効化が必要な場合があります。



# インテル® Xeon® 6 プロセッサ (P-cores 採用) 製品における AI 推論性能の向上

AI 推論性能

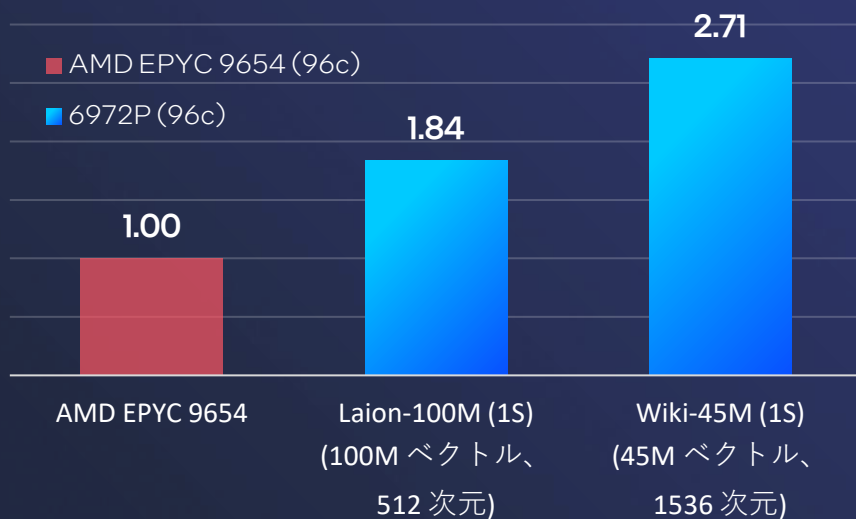


# ベクトル・データベース向けインテル® SVS のベクトル最適化

## インテル® スケーラブル・ベクトル・サーチ (インテル® SVS)

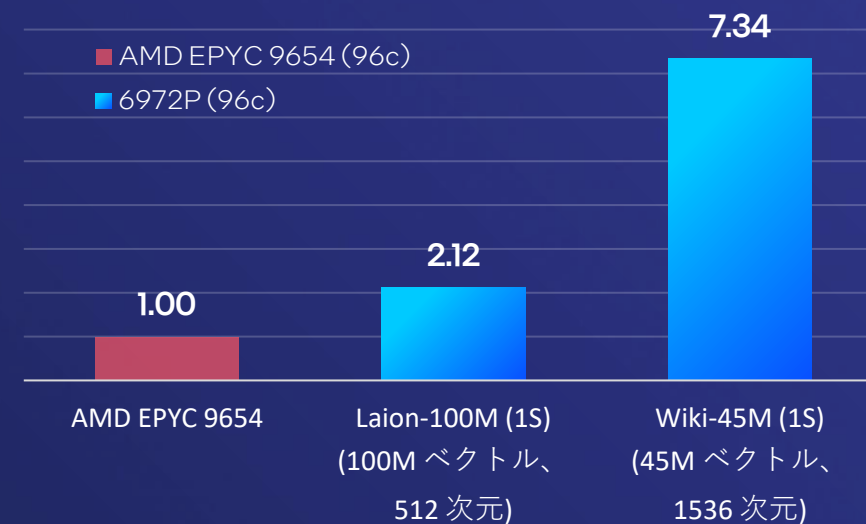
インテル® AMX 命令 を活用してベクトル・データベースのインデックス化の性能改善

インテル® SVS-Inverted File Index (IVF)  
構築 速度向上 (値が大きいほど高性能)

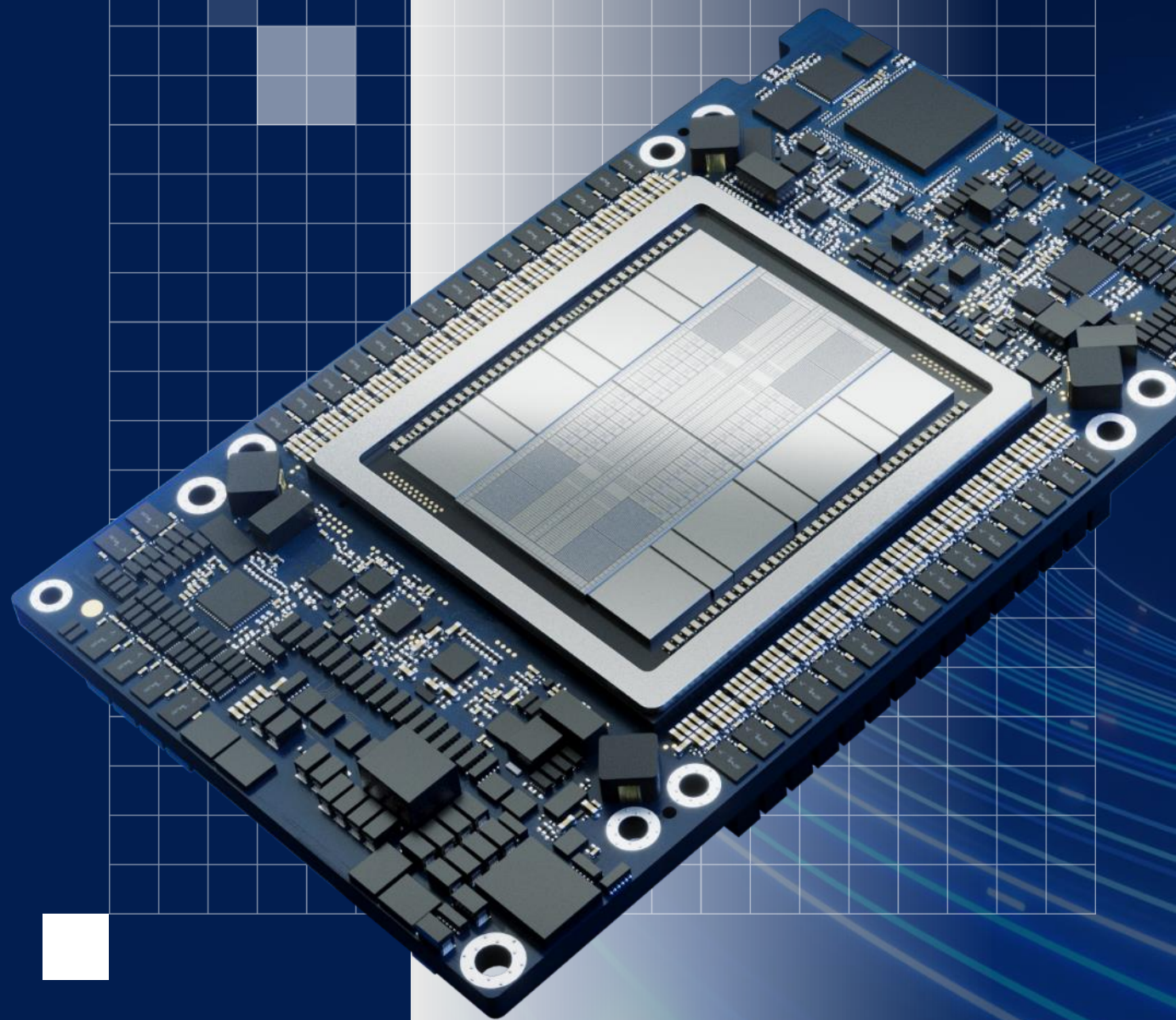


インテル® SVS のベクトル最適化 による  
ベクトル・データベースの検索の性能改善

インテル® SVS- Graph による  
類似性検索 (値が大きいほど高性能)



インテル® Gaudi®  
AI アクセラレーター  
生成 AI に新しい選択肢を



## インテル® Gaudi® 3 AI アクセラレーターで 生成 AI の課題を解決

### 40% 短縮

学習処理時間

(NVIDIA H100 との比較)<sup>1</sup>

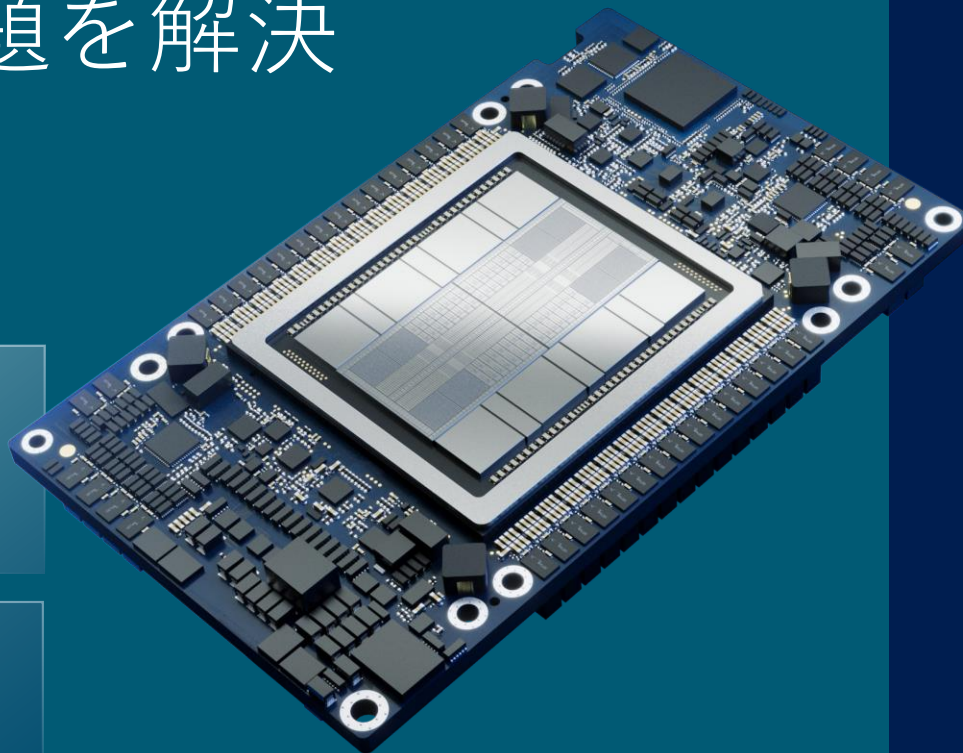
代表的な大規模言語モデルでの平均(予測値)

### 50% 高速

推論速度

(NVIDIA H100 との比較)<sup>2</sup>

代表的な大規模言語モデルでの平均(予測値)



### 生成 AI に より多くの選択 肢

生成 AI の実運用に向け  
て供給の安定性と価格上  
昇への懸念を解消

### 優れた コスト性能比

実サービスに向けて ROI  
を改善

8 カード搭載ベースボード  
が USD 125K\*\* (標準価  
格)

### オープンな規格 によるデバイス 間接続

業界標準の RoCE 対応  
イーサネットによる接続  
オンチップ・ネットワー  
クで直接デバイス間を接  
続

### オープンな ソフトウェア・ プラットフォーム

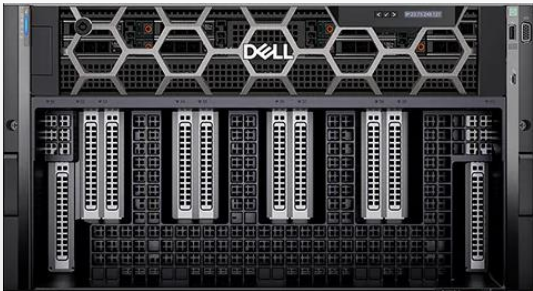
PyTorch や HuggingFace  
など代表的なエコシステ  
ムのソフトウェアに対応  
わずか数行のコード修正  
で移行可能

1: NVIDIA との比較はこちらのデータに基づく: <https://developer.nvidia.com/deep-learning-performance-training-inference/training>, 2024年1月29日 → “Large Language Model” の表とインテル® Gaudi® 3 アクセラレーターの 3/28/2024 の時点の予測値の比較。結果は異なる場合があります。

2: 出典: NVIDIA との比較は 2024/5/8 の時点でこちらに掲載されているデータに基づく [Overview — tensorrt\\_llm documentation \(nvidia.github.io\)](https://github.com/nvidia/tensorrt-llm/blob/main/docs/Overview.md) 報告されている数値は GPU ごとの値。3/28/2024 時点でのインテル® Gaudi® 3 アクセラレーターの予測値との比較。インテル® Gaudi® ソフトウェア 1.14.0 を使用。結果は異なる場合があります。

# インテル® Gaudi® 3搭載サーバー製品

**DELL**Technologies



## Dell PowerEdge XE9680

Air-cooled  
Dell AI Factory



**Hewlett Packard  
Enterprise**



## HPE Cray XD680

Air-cooled & Liquid-cooled



## Supermicro X14

Air-cooled & Liquid-cooled  
Equipped with Intel® Xeon® 6  
processors

intel



# インテル® Gaudi® 3 AI アクセラレーターを IBM Cloud に搭載し エンタープライズ AI の 効率化を推進

そして IBM の watsonx AI & データ・プラットフォームでインテル® Gaudi® 3 AI アクセラレーターをサポート

## コスト効率良く

生成 AI をスケールし、TCO を下げる

## ハイブリッド・クラウド環境で AI を安全に拡張

## 開発を容易にする

オープンなコミュニティー・ベースの  
プラットフォーム

# インテル® Gaudi® ソフトウェア・スイート

現在主に使われている生成 AI のフレームワークへ統合

FP16 と BF16 のサポートに加えて FP8 量子化のサポート

## 主なプロプライエタリーなソフトウェア・レイヤー

グラフ・コンパイラー: すべてのエンジンの依存性とスケジューリングのロジック

行列演算ライブラリー: MME を構成

TPC カーネル: 行列演算以外のすべての演算

集合通信ライブラリー (CCL)

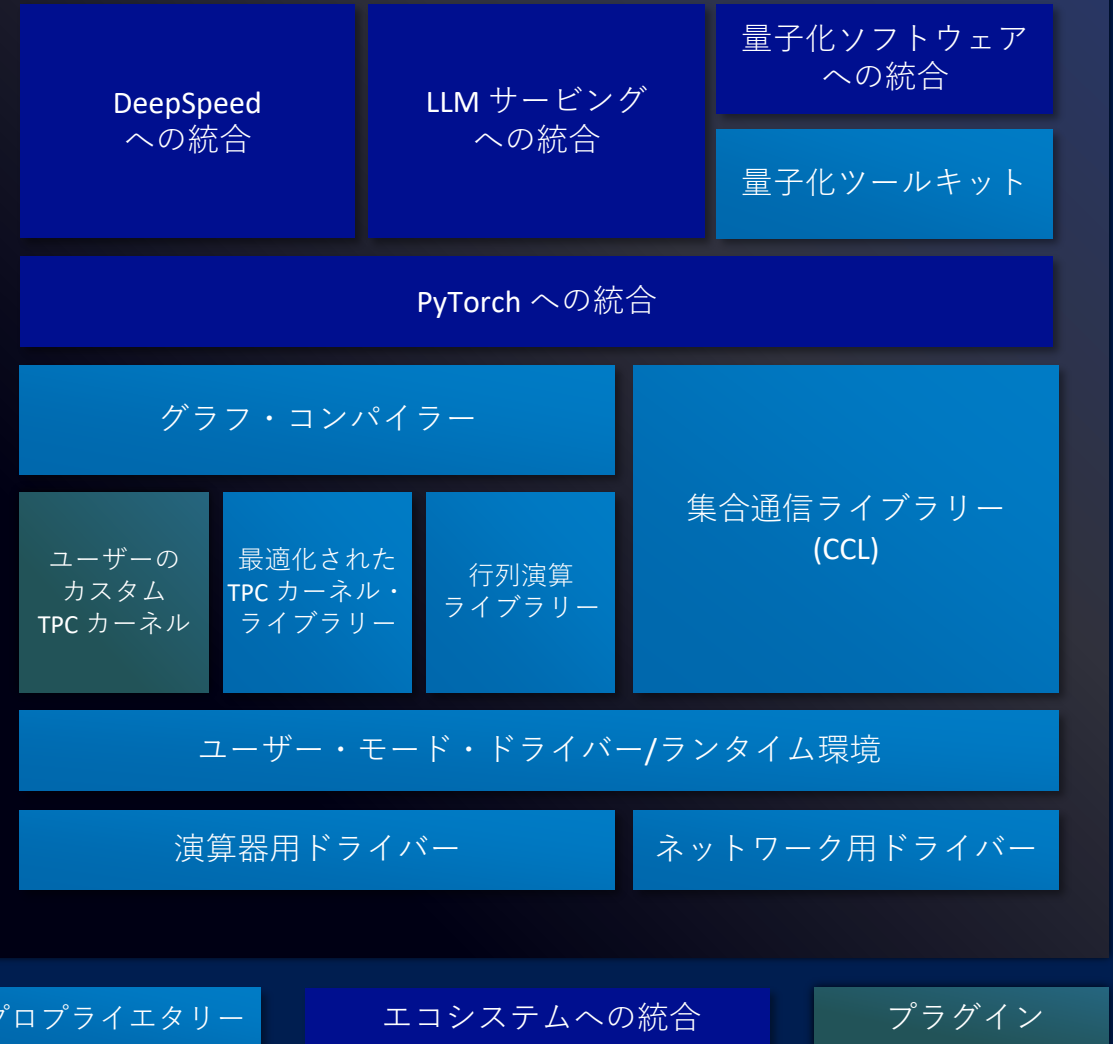
## いくつかの TPC カーネルのソース

インテル® Gaudi® アクセラレーターに最適化された  
TPC カーネル・ライブラリー

カスタムなユーザーカーネル

MLIR ベースの融合カーネル: グラフのコンパイル中に生成

## インテル® Gaudi® ソフトウェア・スイート



# Intel Gaudi Software Suite アップデート(1.17-1.19)

	1.17/1.17.1(Aug`24)	1.18 (Oct`24)	1.19/1.19.1(Dec`24)
General	<ul style="list-style-type: none"> <li>Intel Gaudi Base Operator for Kubernetes</li> <li>Offline Trace Parser tool</li> <li>Hugging Face Optimum 1.12.1</li> <li>INC (Intel Neural Compressor) replace HQT for FP8 Quantize</li> </ul>	<ul style="list-style-type: none"> <li>Ubuntu 24.04 for Gaudi3</li> <li>H.264 support in MediaPipe</li> <li>Hugging Face Optimum 1.13.1</li> <li>DeepSpeed 0.14.4</li> <li>Performance improvement for FusedSDPA</li> </ul>	<ul style="list-style-type: none"> <li>Kubernetes versions 1.30, 1.31, 1.32</li> <li>Ubuntu 24.04 on Gaudi 2</li> <li>RHEL 8.6 with Python 3.11</li> <li>RDMA PerfTest tool</li> </ul>
PyTorch	<ul style="list-style-type: none"> <li>PyTorch 2.3.1</li> <li>PyTorch Lightning 2.3.3</li> <li>Distributed Tensor (DTensor)</li> <li>TorchServe</li> </ul>	<ul style="list-style-type: none"> <li>PyTorch 2.4.0</li> <li>INC UINT4 Quantize</li> </ul>	<ul style="list-style-type: none"> <li>PyTorch 2.5.1</li> </ul>
Inference	<ul style="list-style-type: none"> <li>TGI-gaudi 2.0.1</li> <li>Hugging Face Optimum Inference                             <ul style="list-style-type: none"> <li>LLaMA 2 7B on 1 card</li> <li>LLaMA 2 70B up to 2 cards</li> <li>LLaMA 3 8B on 1 card</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>TGI-gaudi 2.0.4</li> <li>vLLM-fork                             <ul style="list-style-type: none"> <li>Attention with Linear Biases (ALiBi)</li> <li>Quantization with Intel Neural Compressor (INC)</li> <li>LoRA adapters</li> <li>Long context (up to 32k)</li> <li>Initial support for torch.compile</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>TGI-gaudi 2.3.1</li> <li>vLLM-fork v0.6.4.post2                             <ul style="list-style-type: none"> <li>Multi-step scheduling HPU with Tensor Parallelism</li> <li>Asynchronous Output Processing</li> <li>Long context with LoRA (up to 128k)</li> <li>Automatic Prefix Caching</li> <li>Repetition penalty</li> <li>Structured Output (guided JSON)</li> <li>FusedMoE</li> <li>Non-invasive model graph splitting</li> </ul> </li> </ul>
Training		<ul style="list-style-type: none"> <li>Megatron-DeepSpeed                             <ul style="list-style-type: none"> <li>LLaMA 3.1 8B FP8/BF16 on 8 cards (Gaudi2/3)</li> <li>LLaMA 3.1 70B FP8/BF16 on 64 cards (Gaudi2/3)</li> <li>Megatron-LM Preview</li> </ul> </li> </ul>	<ul style="list-style-type: none"> <li>Megatron-LM                             <ul style="list-style-type: none"> <li>LLaMA 3.1 8B on 8 cards (Gaudi2/3)</li> <li>LLaMA 3.1 70B on 64 cards (Gaudi2/3)</li> <li>Mixtral 8x7B BF16 on 32 cards (Gaudi2)</li> </ul> </li> </ul>





# インテル® AIソフトウェア



# PyTorch – インテル® GPU をネイティブサポート



- PyTorch 2.5 からインテル® GPU をサポート！
  - Intel® Extension for PyTorch (IPEX) が不要に
  - インテルのクライアント向け GPU (インテル® Arc™ グラフィックス、インテル® Core™ Ultra プロセッサ)
  - インテル® データセンター GPU マックス・シリーズ
- torch.compile と eager モードに対応
- Linux と Windows に対応
- FP32/BF16/FP16、AMPに対応
- インテル® Tiber™ AI クラウドの無料アカウントでインテル® データセンター GPU マックス・シリーズ上で “PyTorch on Intel GPU” などの学習用ノートブックを利用可能

```
$ pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/test/xpu
```

```
# CUDA CODE
tensor = torch.tensor([1.0, 2.0]).to("cuda")

# CODE for Intel GPU
tensor = torch.tensor([1.0, 2.0]).to("xpu")
```

[https://pytorch.org/docs/stable/notes/get\\_start\\_xpu.html](https://pytorch.org/docs/stable/notes/get_start_xpu.html)

<https://dev-discuss.pytorch.org/t/intel-gpu-enabling-status-and-feature-plan/2389>

# OpenVINO™ ツールキット 2024.4～2024.6 アップデート

## ■ 新しいハードウェアへの対応

- 2024.4: インテル® Core™ Ultra プロセッサ (シリーズ 2) (開発コード名: Lunar Lake)
- 2024.5: インテル® Xeon® 6 プロセッサ (P-cores 採用) (開発コード名: Granite Rapids)、インテル® Core™ Ultra プロセッサ 200V シリーズ (開発コード名: Allow Lake-S)
- 2024.6: インテル® Arc™ B シリーズ・グラフィックス (開発コード名: Battlemage)

## ■ 生成 AI を含むより幅広いモデルへの対応や最適化、モデル圧縮技術をサポート

- GLM-4-9B Chat、MiniCPM-1B、Llama 3 and 3.1、Phi-3-Mini、Phi-3-Medium、Llama 3.2 (1B & 3B)、Gemma 2 (2B & 9B)、YOLO11

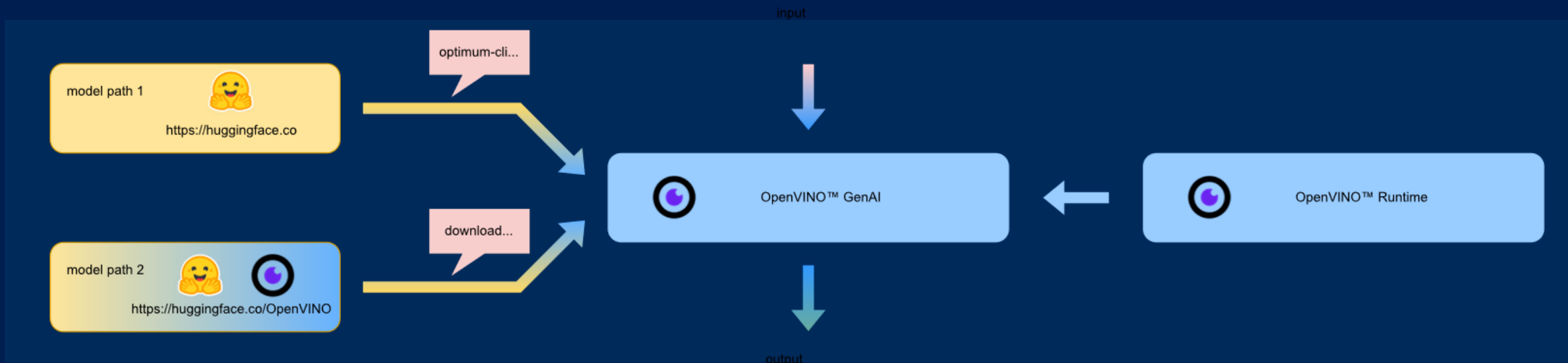
## ■ 2024.5: JAX/Flax のサポート (Preview)

## ■ OpenVINO™ モデル・サーバー

- 2024.4: OpenAI API によるテキスト生成の正式サポートと性能向上のための機能追加 (プレフィクス・キャッシュ、KV キャッシュ圧縮など)
- 2024.5: OpenAI API でのテキスト埋め込みのエンドポイントや Cohere API でランキングのエンドポイント (RAG での利用)
- 2024.5: Windows 用バイナリーパッケージのサポート (Preview)

## ■ OpenVINO™ ツールキット GenAI API

- 2024.5: マルチモーダル・パイプライン (Preview)、投機的デコード、LoRA アダプター (Preview) NPU での LLM のサポート



```
from optimum.intel import OVModelForCausalLM
from transformers import AutoConfig, AutoTokenizer

model_dir = r"llama-3-8b-instruct\INT4_compressed_weights"

ov_config = {'PERFORMANCE_HINT': 'LATENCY', 'NUM_STREAMS': '1', "CACHE_DIR": ""}
chat_model = OVModelForCausalLM.from_pretrained(model_dir, device="AUTO",
                                                config=AutoConfig.from_pretrained(model_dir),
                                                ov_config=ov_config)
chat_tokenizer = AutoTokenizer.from_pretrained(model_dir)

prompt="The Sun is yellow because"
inputs = chat_tokenizer(prompt, return_tensors="pt").to(chat_model.device)
input_length = inputs.input_ids.shape[1]
outputs = chat_model.generate(**inputs, max_new_tokens=256,
                              do_sample=True, temperature=0.6, top_p=0.9, top_k=50)
tokens = outputs[0, input_length:]
print(chat_tokenizer.decode(tokens, skip_special_tokens=True))
```

import openvino\_genai as ov\_genai  
pipe = ov\_genai.LLMPipeline(model\_path, "AUTO")  
print(pipe.generate("The Sun is yellow because"))

**OpenVINO™ Integration with Optimum**

**OpenVINO™ GenAI API**

## まとめ

- AI を継続的にビジネスで活用していくためには、ワークロードに合わせて最適なハードウェアを用いてインフラストラクチャーを構築することが重要
- RAGなどの企業におけるAIの用途には、継続して AI 機能を改善しているインテル® Xeon® プロセッサと、生成 AI 向けに性能と高コスト効率に優れたインテル® Gaudi® 3 AI アクセラレーターを適材適所で利用
- OpenVINO™ ツールキットやインテル®のGPU向けのPyTorchなどソフトウェアも積極的に開発中

ご清聴ありがとうございました！

# 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、<https://www.Intel.com/PerformanceIndex/> (英語) を参照してください。

性能の測定結果は、構成に示されている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アップデート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、補足資料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またはコンポーネントはありません。

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。

各アクセラレーターの利用可否は **SKU** ごとに異なります。詳細については、担当のインテル販売代理店までお問い合わせください。

実際のコストや結果は異なる場合があります。

©2025 Intel Corporation. Intel、インテル、Intel ロゴ、その他のインテルの名称やロゴは、Intel Corporation またはその子会社の商標です。その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

intel ai