



インテルの AI 製品の最新情報

エクセルソフト株式会社

intel ai



目次

- インテルの AI ポートフォリオ
- ハードウェア
 - インテル® Xeon® プロセッサー
 - インテル® Gaudi® AI アクセラレーター
- インテル® AI ソフトウェアのアップデート

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
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

fiber intel AI クラウド

OpenVINO®

oneAPI

AI PC

エッジ AI

データセンター & クラウド AI

AI ネットワーク

intel CORE ARC intel XEON intel ARC intel XEON intel GAUDI intel eTHERNET

生産性、創造性、セキュリティにおける新しい AI 体験

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

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

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



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

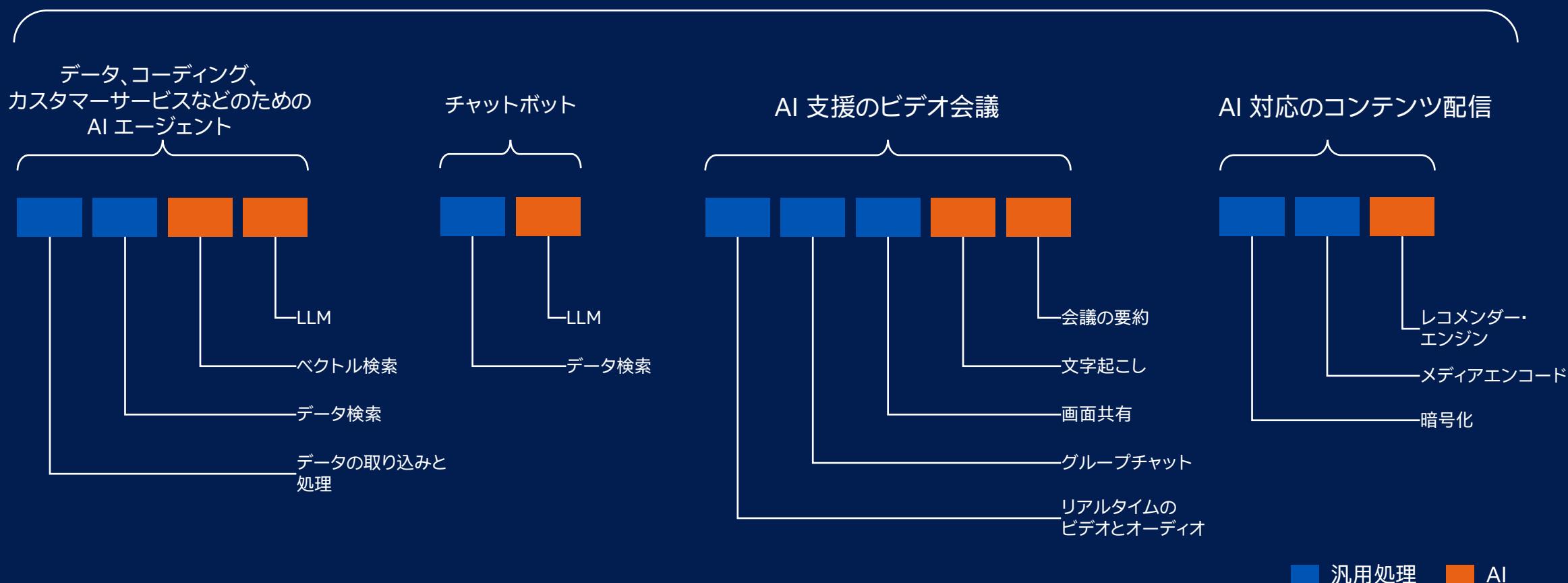


企業規模で CPU 上に構築 & 運用

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

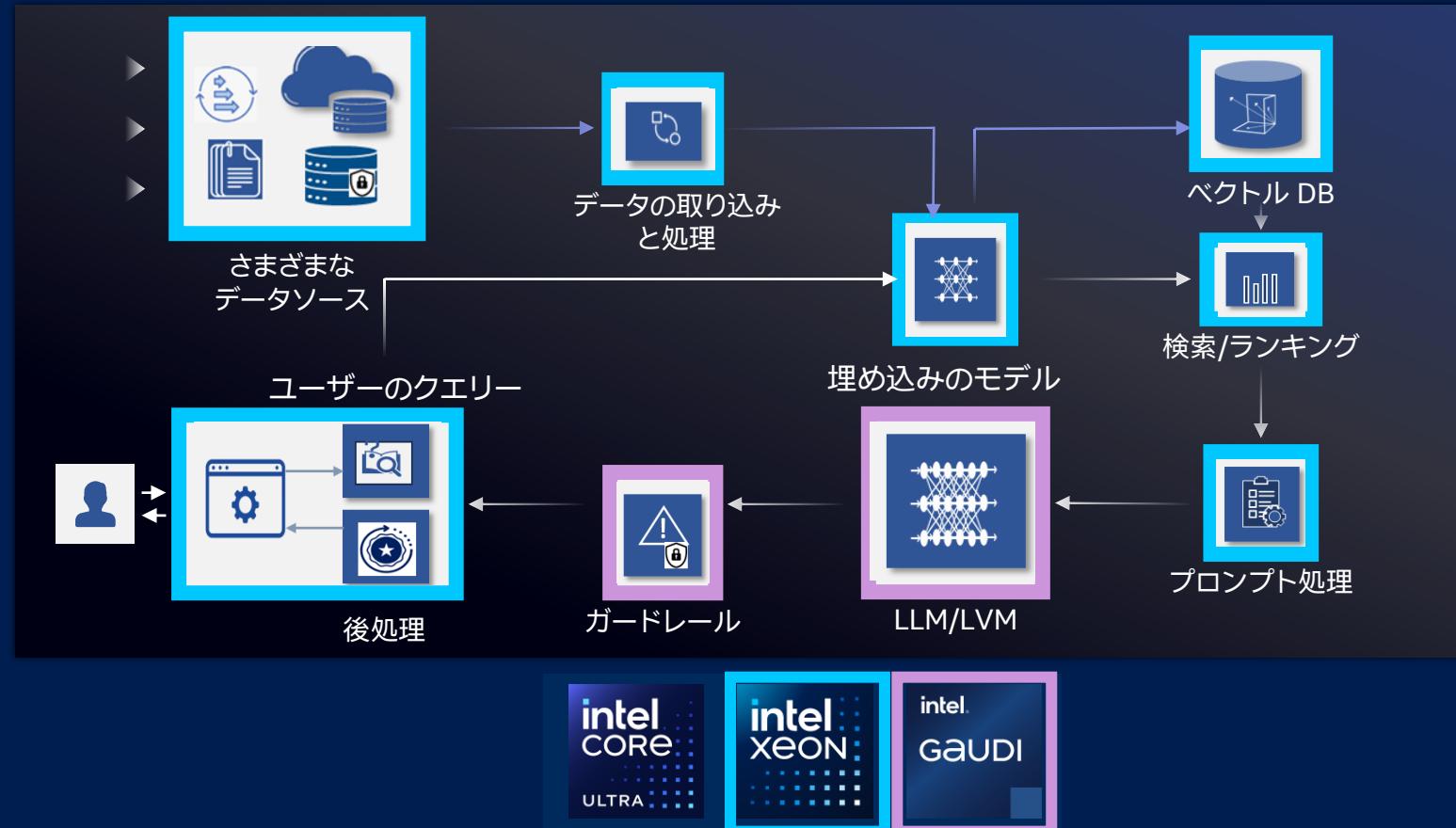
オープン/クローズド

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



Performance-cores (P-cores) 搭載 インテル® Xeon® 6 プロセッサー 6900P シリーズ

CPU あたり最大 128 コア

メモリー帯域幅を増加

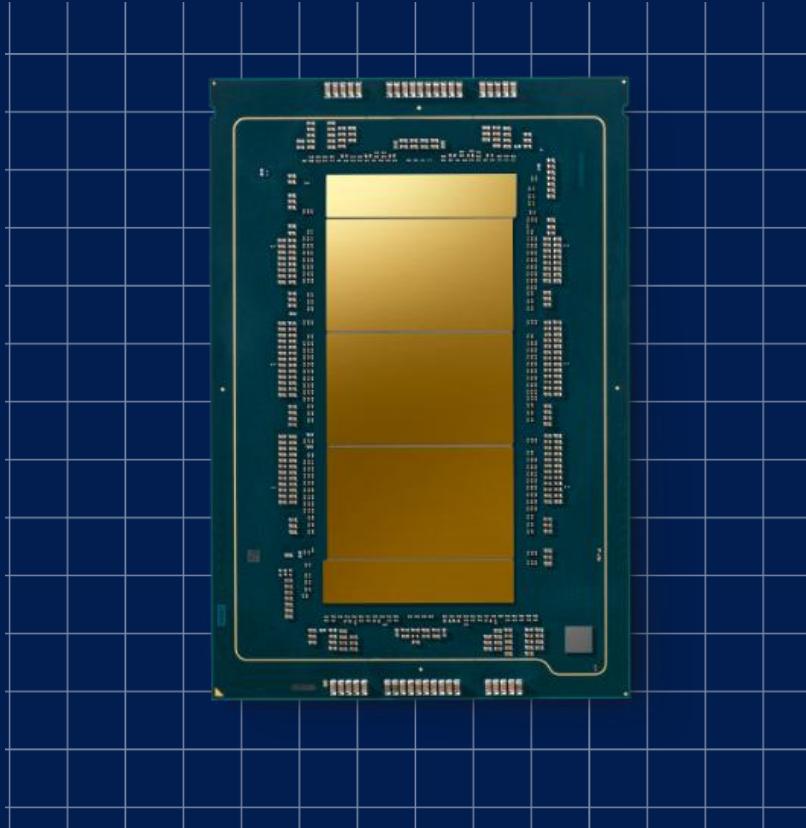
MRDIMM によりメモリー帯域幅を
最大 2.3 倍に
vs 第 5 世代インテル® Xeon® スケーラブル・
プロセッサー¹

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

P-cores 搭載インテル® Xeon® 6 プロセッサーはインテル® 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ーラブルなソリューションを提供

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

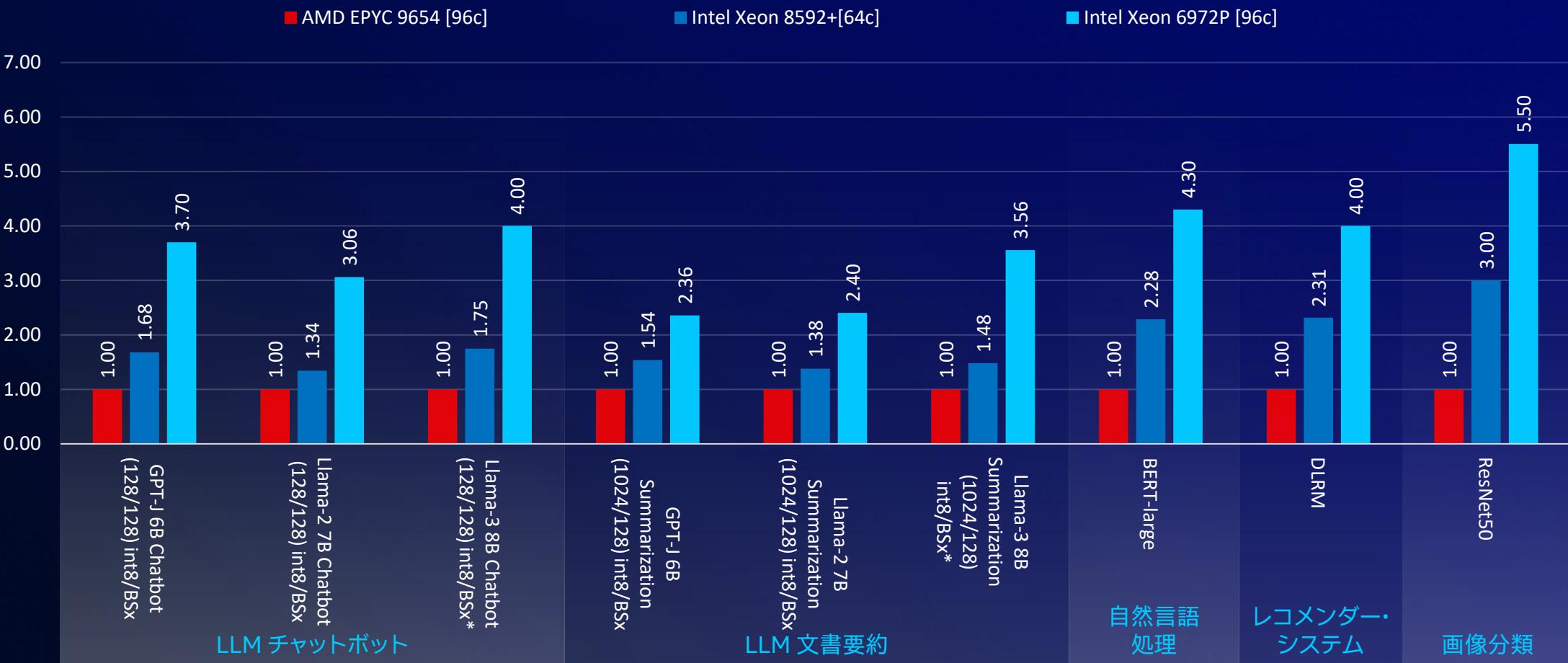
インテル® アドバンスト・ベクトル・エクステ ンション 2 (インテル® AVX2)

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

P-cores 搭載インテル® Xeon® 6 プロセッサー製品における AI 推論性能の向上

AMD EPYC 9654 [192c] 対比

AI 推論性能



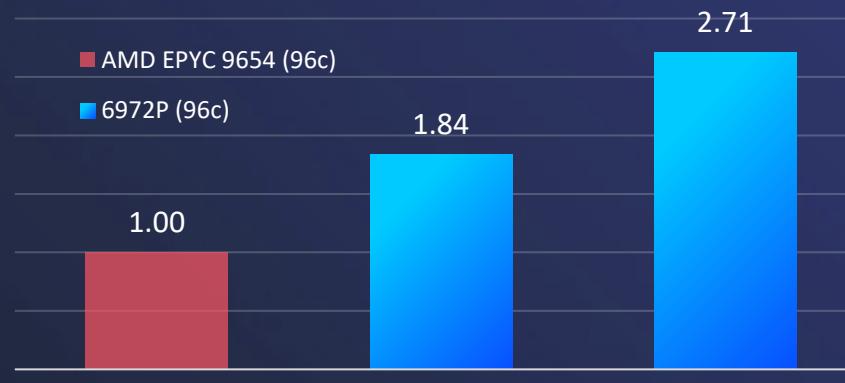
See [18] for workloads and configurations. Results may vary.

* AMD EPYC 9654 with BF16 on Llama3 8B Chatbot; did not meet SLA on INT8

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)

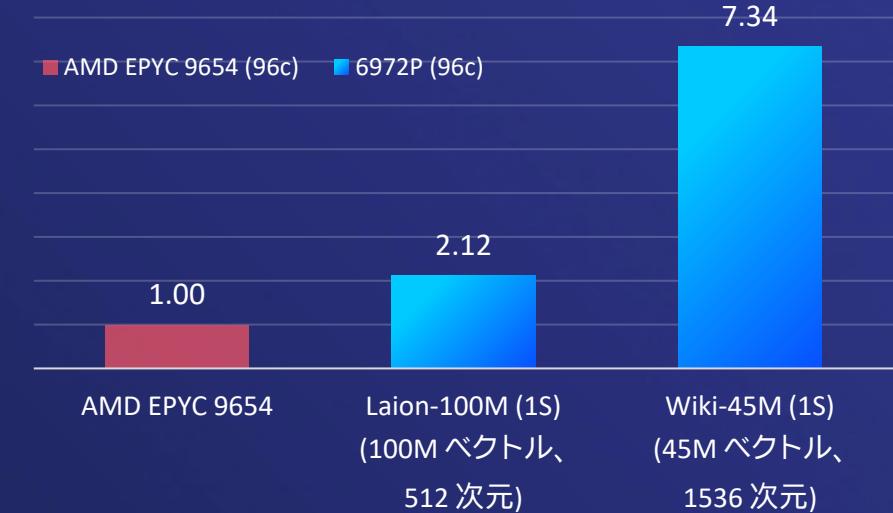
インテル® AMX 命令を活用してベクトル・データベースのインデックス化の性能改善

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)

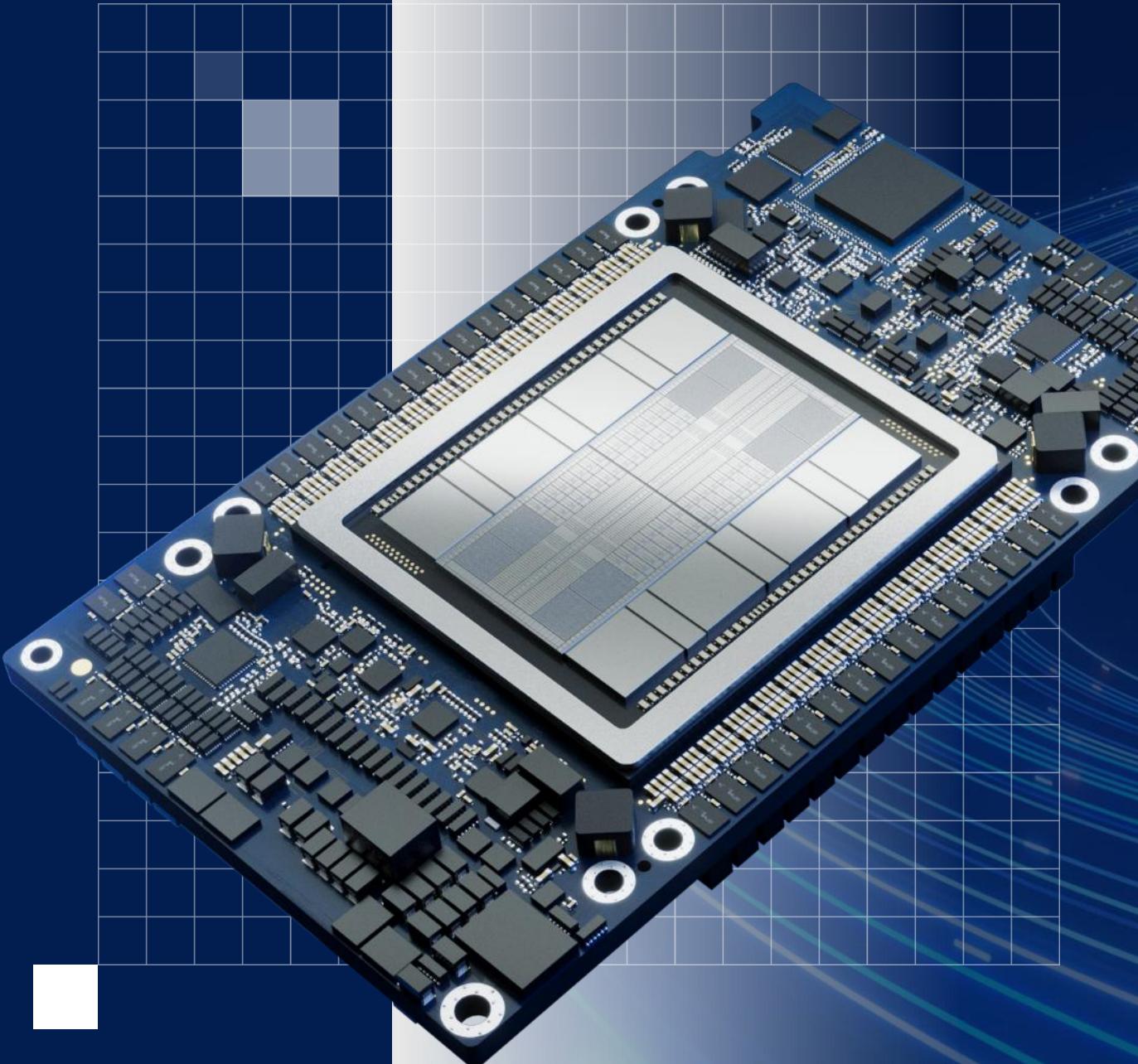


インテル® SVS のベクトル最適化による
ベクトル・データベースの検索の性能改善

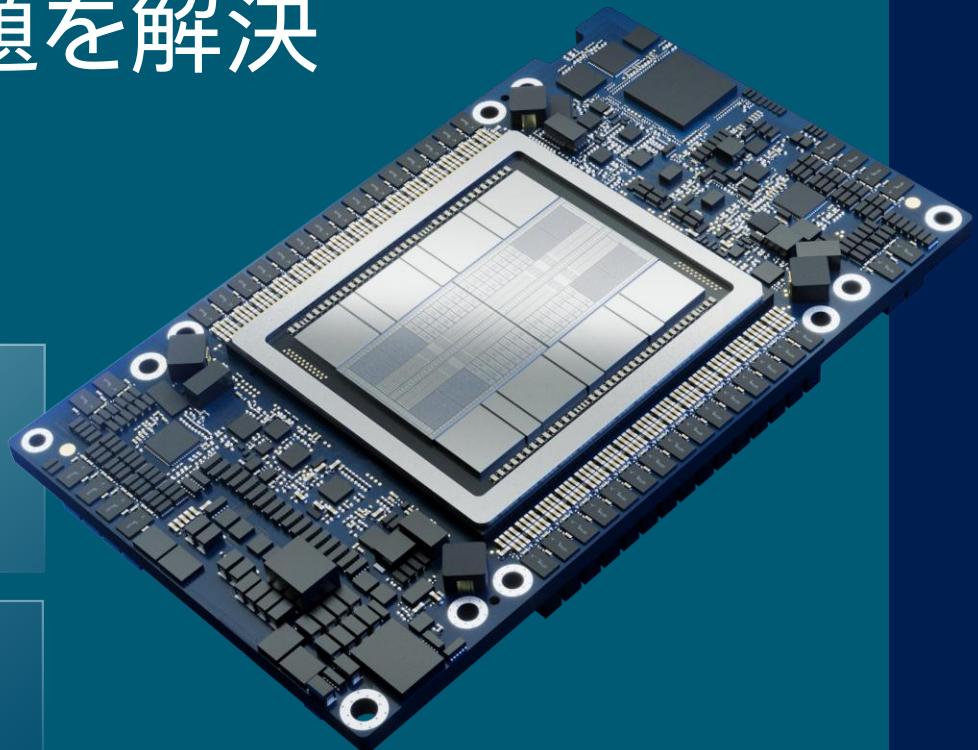
インテル® SVS-Graph による
類似性検索 (値が大きいほど高性能)



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)¹
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)²
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 <https://developer.nvidia.com/deep-learning-performance-training-inference/training> (英語) の Large Language Model の表の 2024/1/29 時点の掲載データに基づく。インテル® Gaudi® 3 アクセラレーターの 2024/3/28 時点の予測値との比較。結果は異なる場合があります。

2: NVIDIAとの比較は [Overview — tensorrt_llm documentation \(nvidia.github.io\)](https://github.com/nvidia/tensorrt_llm/tree/main/doc) (英語) の 2024/5/8 時点の掲載データに基づく。報告されている数値は GPU ごとの値。インテル® Gaudi® 3 アクセラレーターの 2024/3/28 時点の予測値との比較。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 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

コスト効率良好
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: 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

DeepSpeed
への統合

LLM サービング
への統合

量子化ソフトウェア
への統合

量子化ツールキット

PyTorchへの統合

グラフ・コンパイラ

ユーザーの
カスタム
TPC カーネル

最適化された
TPC カーネル・
ライブラリー

行列演算
ライブラリー

集合通信ライブラリー
(CCL)

ユーザー・モード・ドライバー/ランタイム環境

演算器用ドライバー

ネットワーク用ドライバー

プロプライエタリー

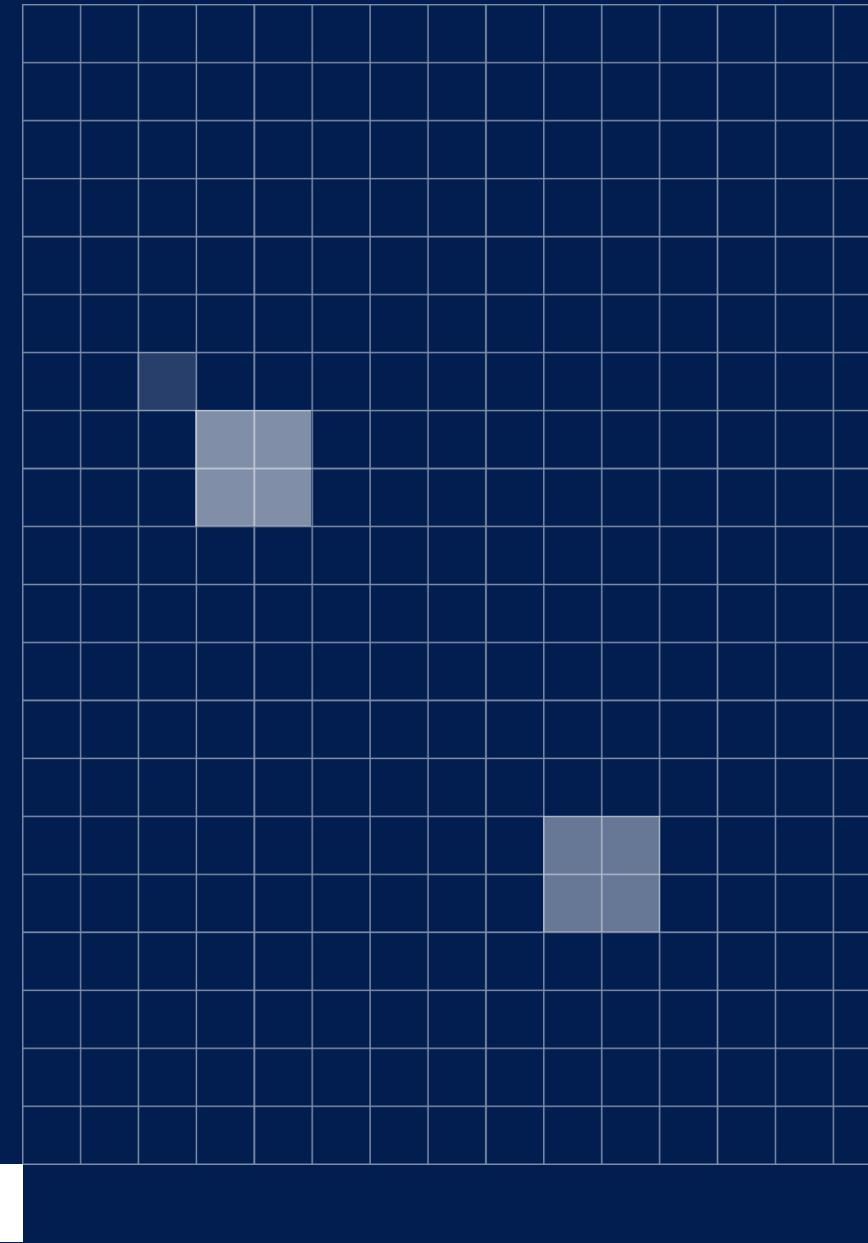
エコシステムへの統合

プラグイン

インテル® Gaudi® ソフトウェア・スイート・アップデート (1.19-1.21)

	1.19.0/1.19.1/1.19.2	1.20.0/1.20.1	1.21.0/1.21.1
General	<ul style="list-style-type: none"> Kubernetes versions 1.30, 1.31, 1.32 Ubuntu 24.04 on Gaudi 2 RHEL 8.6 with Python 3.11 RDMA PerfTest tool 	<ul style="list-style-type: none"> Diagnostic Tool Secure Boot for Intel® Gaudi® AI accelerators on Ubuntu Software backward/forward compatibility for Gaudi 3 	<ul style="list-style-type: none"> Gaudi 3 with the Intel® Tiber™ AI Cloud
PyTorch	<ul style="list-style-type: none"> PyTorch 2.5.1 	<ul style="list-style-type: none"> PyTorch 2.6.0 PyTorch Lightning 2.5.0.post0. 	<ul style="list-style-type: none"> PyTorch Lightning 2.5.1
Inference	<ul style="list-style-type: none"> TGI-gaudi 2.3.1 vLLM-fork v0.6.4.post2 <ul style="list-style-type: none"> Multi-step scheduling HPU with Tensor Parallelism Asynchronous Output Processing Long context with LoRA (up to 128k) Automatic Prefix Caching Repetition penalty Structured Output (guided JSON) FusedMoE Non-invasive model graph splitting 	<ul style="list-style-type: none"> TGI-gaudi 2.3.2 vLLM-fork v0.6.6.post1 <ul style="list-style-type: none"> Multiprocessing backend Speculative decoding 	<ul style="list-style-type: none"> vLLM-fork <ul style="list-style-type: none"> Automatic Prefix Caching Pipeline Parallelism - see Pipeline Parallelism section. Guided Decoding V1 Support - initial support Multimodality Exponential Bucketing - initial support Delayed Sampling - initial support FP16 support - limited models (for further details, see Intel Gaudi vLLM fork) Multi node INT4 support (AWQ/GPTQ) - limited models (for further details, see Intel Gaudi vLLM fork) Support for RedHat Openshift AI Split QKV optimizations for BF16
Training	<ul style="list-style-type: none"> Megatron-LM <ul style="list-style-type: none"> LLaMA 3.1 8B on 8 cards (Gaudi2/3) LLaMA 3.1 70B on 64 cards (Gaudi2/3) Mixtral 8x7B BF16 on 32 cards (Gaudi2) 		

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!
 - PyTorch 向けインテル® エクステンション (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 プロセッサー)
 - 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/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://dev-discuss.pytorch.org/t/intel-gpu-enabling-status-and-feature-plan/2389>

OpenVINO™ ツールキット 2025 アップデート

- 新しいハードウェアへの対応
 - インテル® Core™ Ultra プロセッサー (シリーズ 2 含む)
 - インテル® Arc™ B シリーズ・グラフィックス
 - インテル® Core™ 200V シリーズ (FP16-NF4 精度対応)
- 生成 AI を含むより幅広いモデルへの対応やモデル圧縮技術をサポート
 - 多数の新しい LLM/VLM/画像生成モデルに対応
 - GGUF 形式や LoRA アダプター、Text-to-Speech など新パイプラインを追加
 - INT8/INT4 圧縮や QAT による軽量化と高精度化を実現

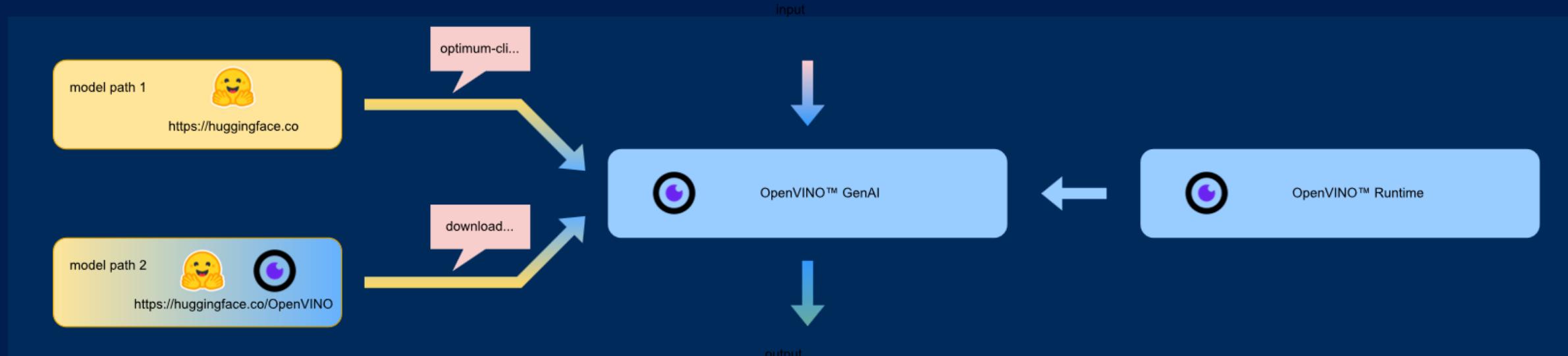
■ OpenVINO™ モデルサーバー

- Windows ネイティブバイナリーの提供 (Python 依存なしの軽量 C++ 版)
- Hugging Face Hub からの自動モデル取得
- Agentic AI 対応 (外部ツール呼び出し)
- 画像生成エンドポイント (OpenAI API 互換)

■ OpenVINO™ ツールキット GenAI.API

- GGUF 形式モデルや Text-to-Speech など新パイプラインに 対応
- KV キャッシュ暗号化や SnapKV でメモリー管理を強化
- JavaScript API も拡張され、LangChain.jsとの連携が可能に

OpenVINO.GenAI



```
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))
```

OpenVINO™ Integration with Optimum

```
import openvino_genai as ov_genai
pipe = ov_genai.LLMPipeline(model_path, "AUTO")
print(pipe.generate("The Sun is yellow because"))
```

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® プロセッサーと、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® GPU 向けの PyTorch などの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