独立行政法人 経済産業研究所（RIETI）

RIETI
Research Institute of Economy, Trade & Industry, IAA

BBLセミナー
プレゼンテーション資料

2024年11月27日

「アクセラレーテッド・コンピューティング・プラットフォーム・カンパニーNVIDIAとは」

井﨑 武士

https://www.rieti.go.jp/jp/index.html

アクセラレーテッド・コンピューティング・プラットフォーム・カンパニーNVIDIAとは

エンタープライズ事業本部　事業本部長　井﨑 武士

Accelerated Computing Platform Company

- 1993 年創業
- 創業者及び CEO　ジェンスン フアン
- 従業員 29,600 人
- 2024会計年度売上高 609億ドル
- 時価総額　3.6兆ドル

NVIDIA の革新の歴史

1993
ゲームテクノロジー

2000
プロフェッショナル
ビジュアライゼーション

2006
HPC & 科学
コンピューティング

2012
Deep Learning & AI

2018
RTX

2020
Omniverse

2022+
生成 AI

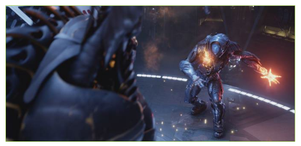
NVIDIA の事業領域

グラフィックス

ゲーミング

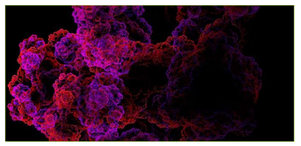
デザイン

レンダリング



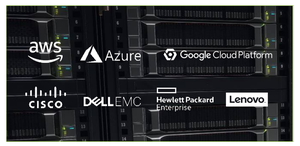
HPC

スーパー
コンピューター

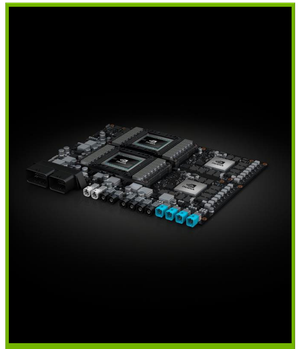


AI

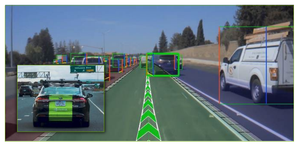
AI 学習



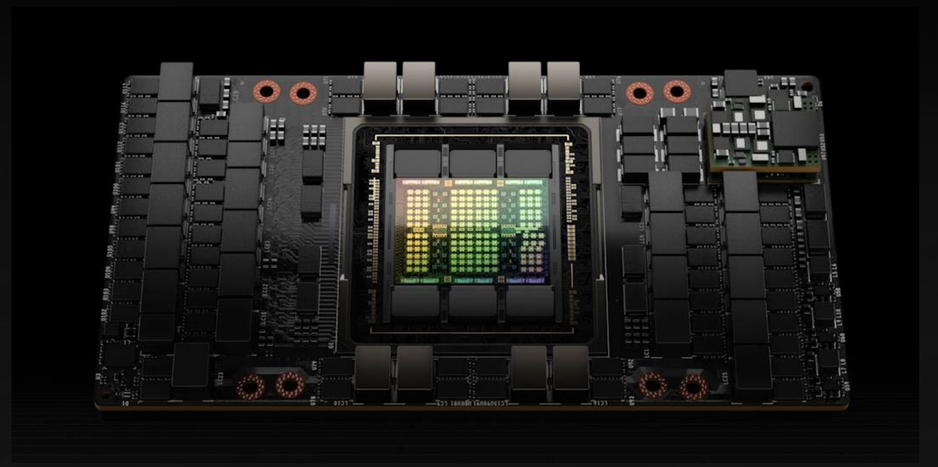
AI 推論



ロボティクス



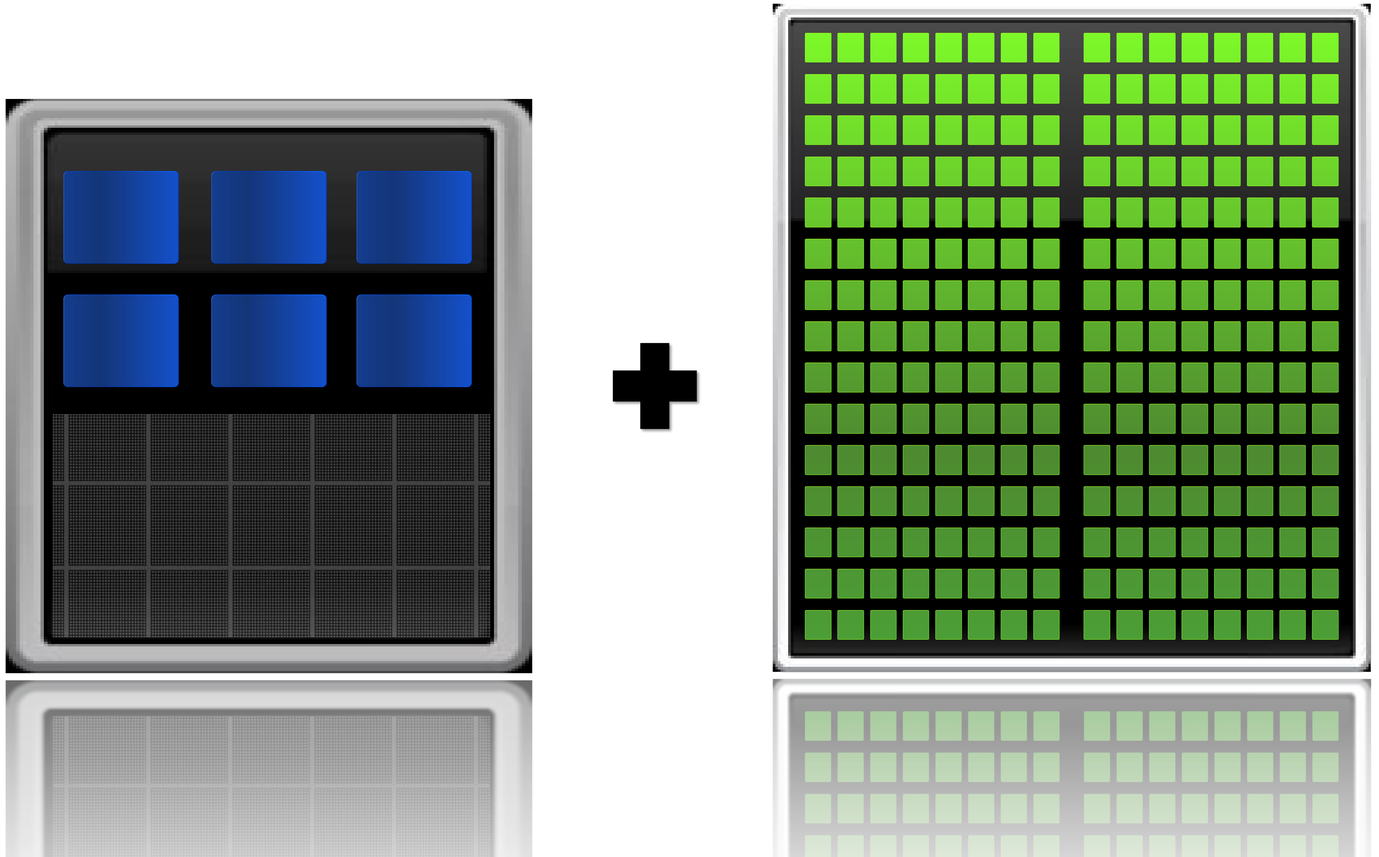
GPU



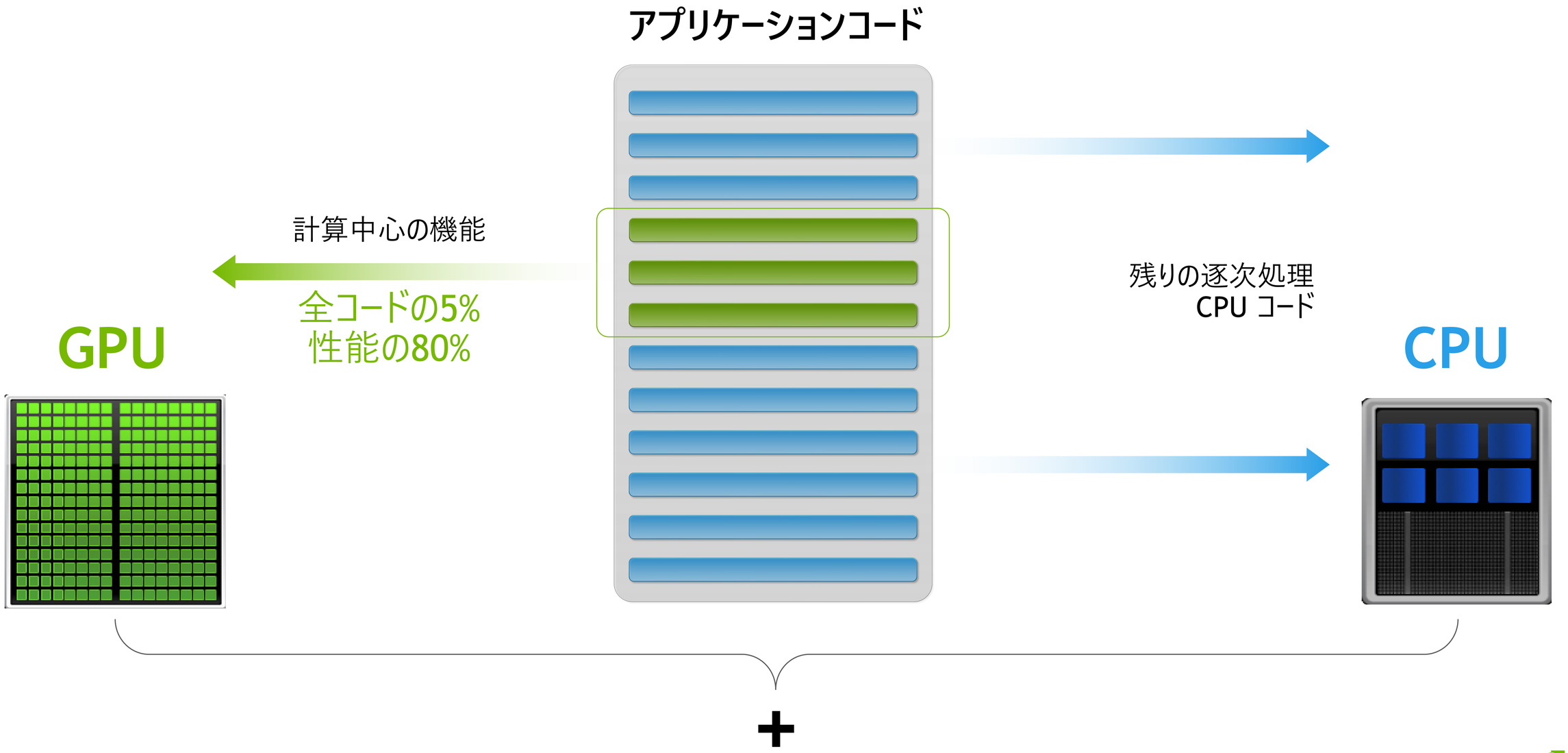
CPUとGPU

CPU

GPU アクセラレータ



GPUアクセラレーションの仕組み




16.7 Million Particles
AOKI Lab.
Tokyo Institute of Technology

NVIDIA のプラットフォーム戦略

NVIDIA アプリケーション フレームワーク

NVIDIA プラットフォーム　ソフトウェア　ライブラリ

NVIDIA HPC

NVIDIA AI

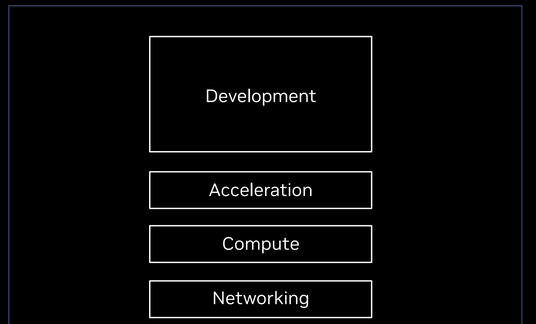
NVIDIA Omniverse

NVIDIA アクセラレーテッドコンピューティング インフラ

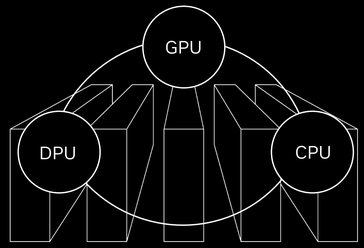
GPU　CPU　DPU

4 つのコアコンピタンス

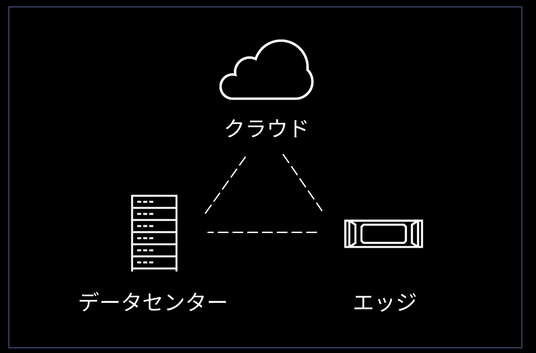
フルスタック



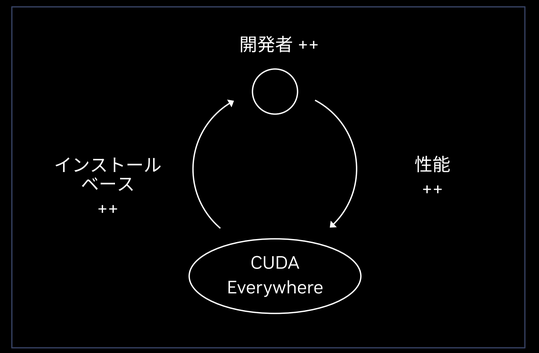
データセンタースケール コンピューティング



エンド・ツー・エンド



単一アーキテクチャ



開発者数: 400万人 / 利用企業: 4万社 / アプリケーション数: 3500超/ スタートアップ: 23,000社

“

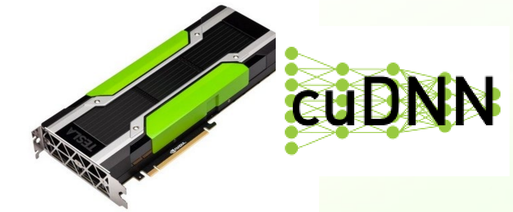
Know the workload

”

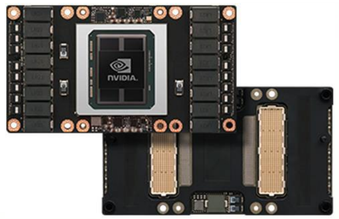
NVIDIA の技術進展

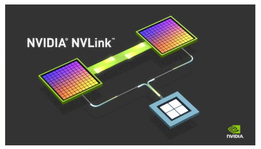
2012
Kepler発表

2015
Maxwell発表
cuDNN



2016
Pascal発表
NVLINK/NCCL
DGX-1

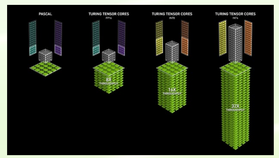




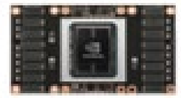
2017
Volta発表
DGX-Staion

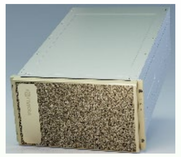
2018
DGX- 2
NV-Switch
NGC/TensorRT/RAPIDS

2019
Turing発表/Omniverse
MAGNUM-IO
DGX/HGX/EGX/AGX
ISACC/CLARA



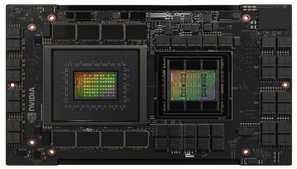
2020
Ampere発表
DGX A100/DGX SPD
JAVIS/MARLIN/AERIAL
Mellanox買収





2021
Triton
RIVA/Nemo/MAXIN/
Metropolis/

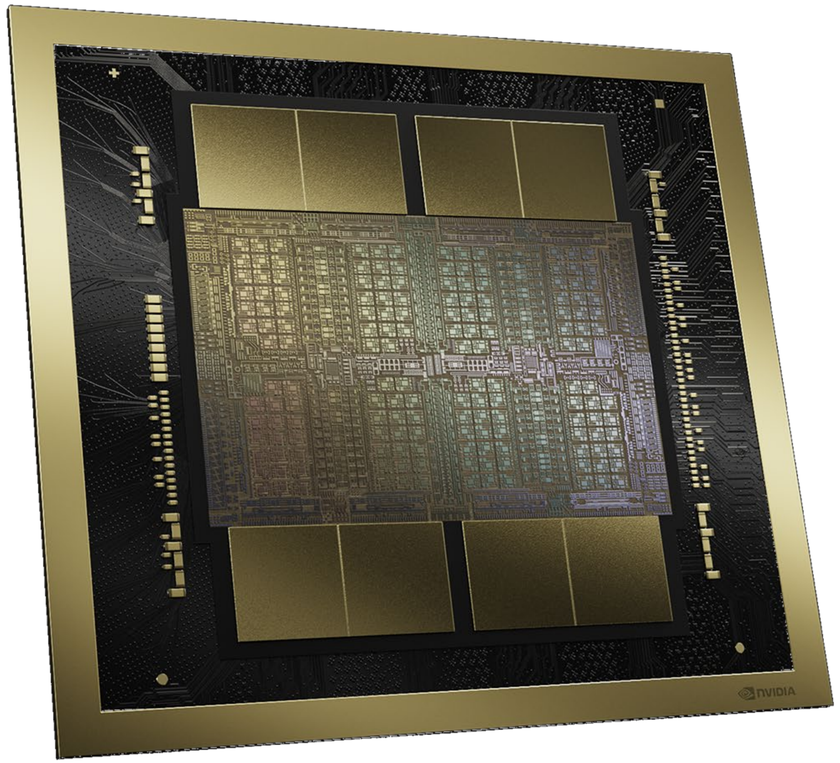
2022
Hopper/Ada Lovelace/Grace発表
Application Framework
Modulus(Earth2) /ACE
Clara Discovery/cuQuantum



2023
DGX Cloud発表
AI Foundation

NVIDIA Blackwell

新たな産業革命のためのエンジン



単一GPUで20 PetaFLOPSのAI性能

従来製品比　学習4倍 | 推論30倍 | 電力＆コスト効率25倍

10 PetaFLOPS FP8 | 20 PetaFLOPS FP4
192GB HBM3e | 8 TB/sec HBM Bandwidth | 1.8TB/s NVLink

AI SUPERCHIP
208B Transistors

2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core

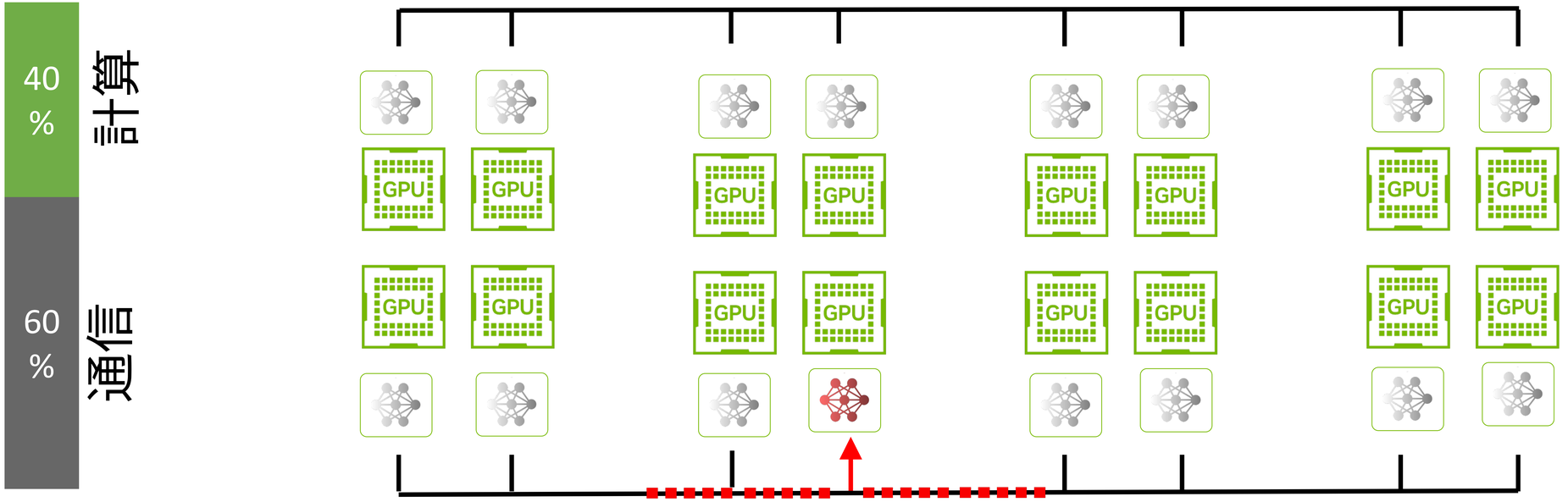
5th GENERATION NVLINK
Scales to 576 GPUs

RAS ENGINE
100% In-System
Self-Test

SECURE AI
Full Performance
Encryption & TEE

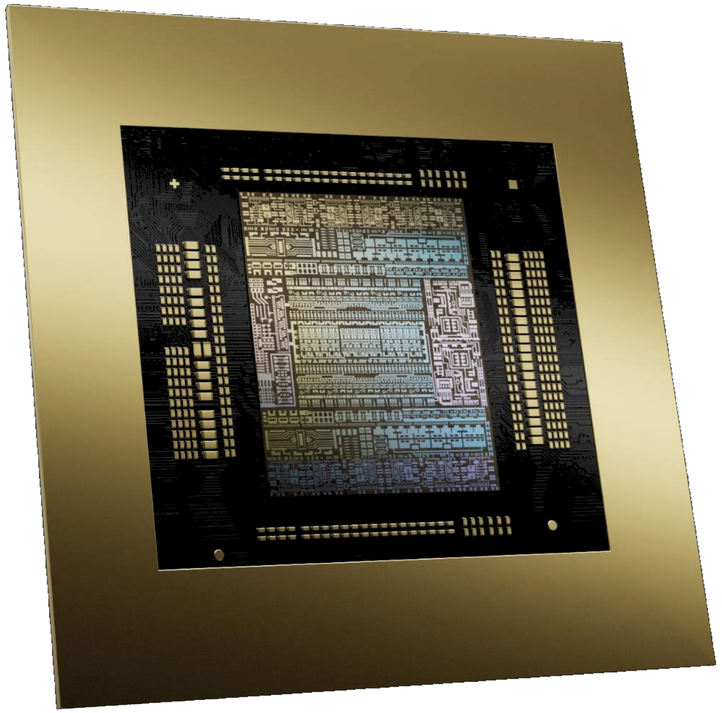
DECOMPRESSION ENGINE
800 GB/s

次世代モデルの通信ボトルネック



第5世代NVLink および NVLink スイッチ

数兆パラメータモデルの効率的なスケーリング



7.2 TB/s　完全な全対全双方向帯域幅

Sharp v4 plus FP8

3.6 TF インネットワークコンピュート

NVLinkを最大576 GPU NVLinkドメインまで拡張

現在のマルチノード相互接続より18倍高速

GB200 NVL72

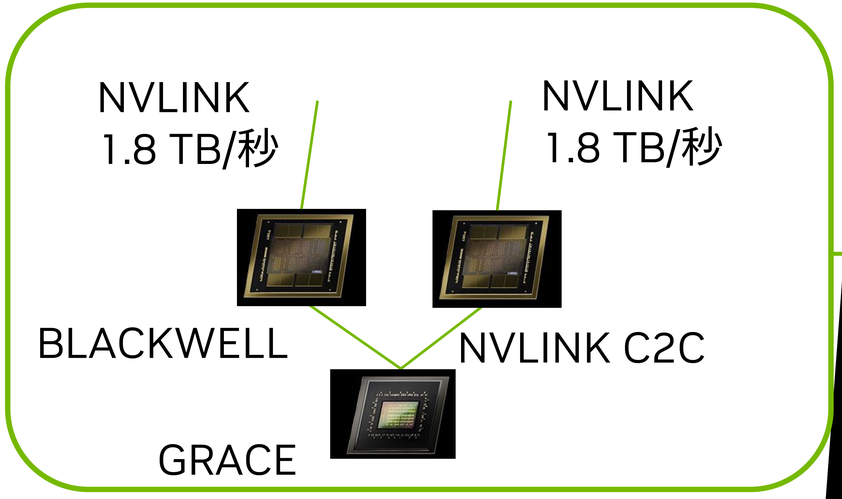
新しいコンピューティング ユニットを提供

GB200 NVL72

- 36 GRACE CPU
- 72 BLACKWELL GPU
- 全結合NVLink Switch ラック

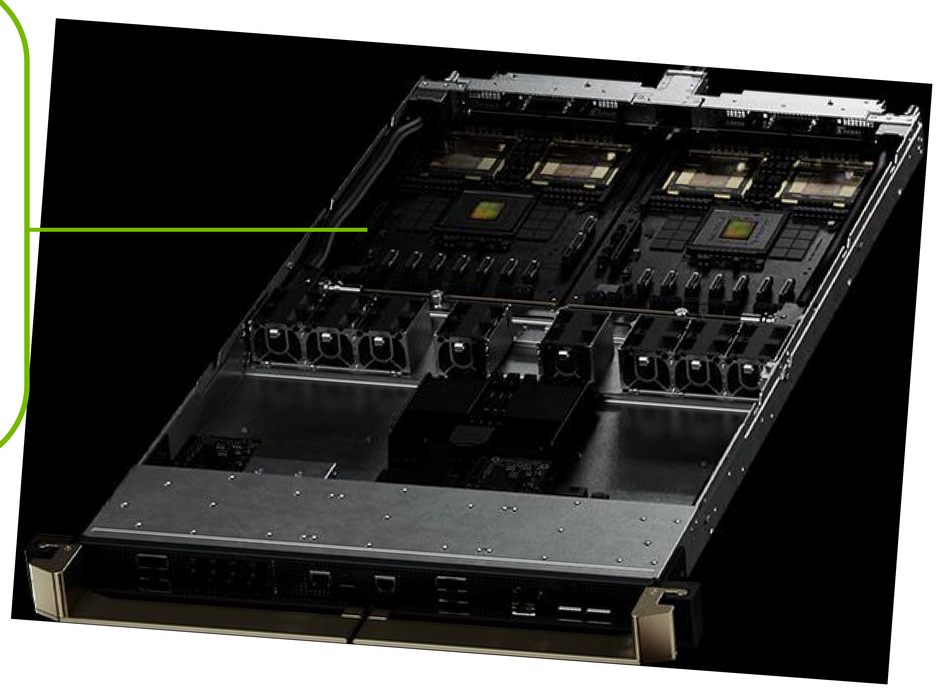
学習 FP8	720 PFLOPs
推論 FP4	1,440 PFLOPs
NVL モデルサイズ	27T params
マルチノードAll to All	130 TB/s
マルチノードAll-Reduce	260 TB/s

GB200 NVL72 コンピュート ノードとインターコネクト ノード



GB200 Superchip

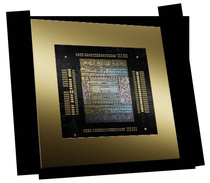
40 PETAFLOPS FP4 AI 推論
20 PETAFLOPS FP8 AI トレーニング
864GB 高速メモリ



GB200 SUPERCHIP コンピュート トレイ

GB200 x2
80 PETAFLOPS FP4 AI 推論
40 PETAFLOPS FP8 AI トレーニング
1728 GB 高速メモリ
1U 水冷
ラックあたり 18 ノード

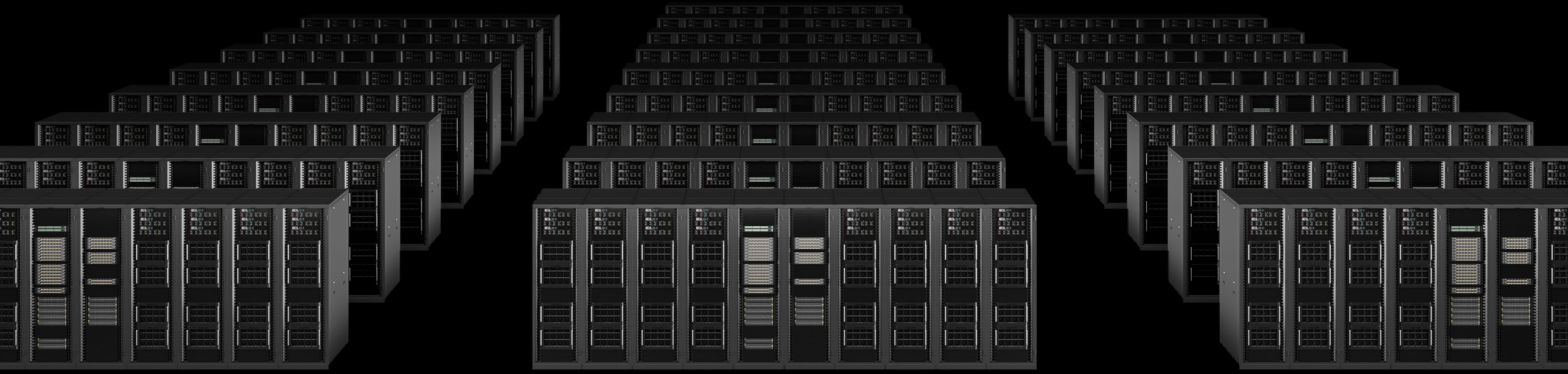
NVLINK SWITCH トレイ



NVLINK SWITCH x2
14.4 TB/秒 総帯域幅
SHARPv4 FP64/32/16/8
1U 水冷
ラックあたり 9 ノード

GPT-MoE-1.8T を90日以内で学習する場合

Hopper
8000 GPUs | 15MW



GPT-MoE-1.8T を90日以内で学習する場合

Blackwell GB200 NVL72
2000 GPUs | 4MW

1/4th the Power

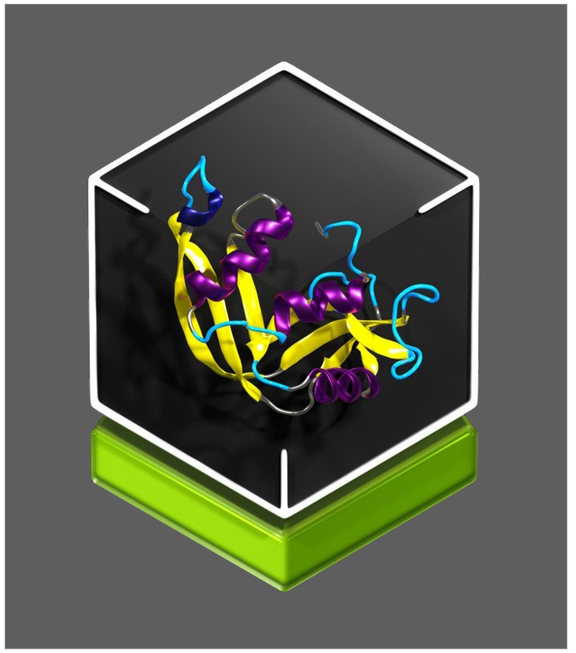
生成AIの新時代

かつてないレベルの生産性を実現

カスタマー・エクスペリエンス
顧客セルフサービスのエージェント体験

コンテンツ制作
パーソナライゼーション
ドメイン別要約

ソフトウェア工学
コーディング・アシスタント



製品研究開発
デザインの強化
シミュレーションとテスト

“... 生成AIは、産業全体で2兆6000億ドルから4兆4000億ドルの価値を生み出す可能性がある”

— McKinsey Digital, “The Economic Potential of Generative AI: The Next Productivity Frontier” 2023

生成AIをどのように使えばいいのか？

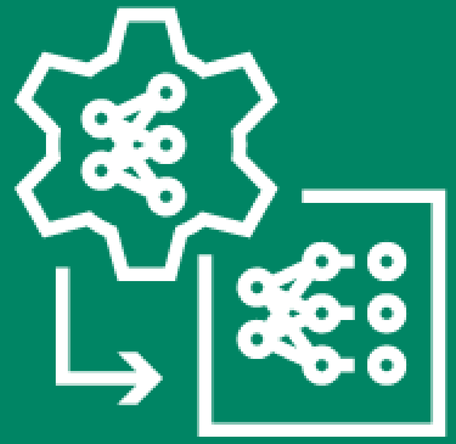
利用者数

最小限のカスタム化

生成AIサービスを利用する - ChatGPT, Google Bard, Amazon Bedrockなどの既存サービス
サービス利用に応じたコンサンプションモデル
早期の市場投入が可能

中間のカスタム化

事前学習モデルのファインチューニング
インフラやリソースに数億の投資が必要
数週間から数ヵ月の開発期間



広範囲なカスタム化

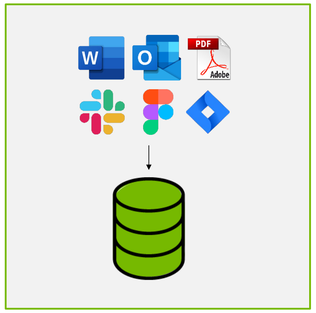
独自の基盤モデル構築もしくは広範囲なファイン
チューニング
インフラやリソースに数十億の投資が必要
半年以上の開発期間

生成AIのカスタム化

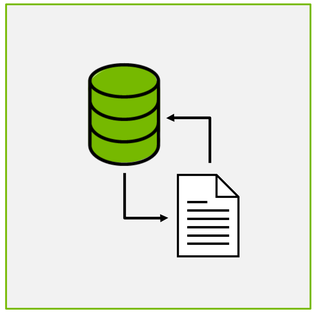
エンタープライズ向けの生成 AI アプリケーションの構築

生成 AI モデルを構築、カスタマイズ、展開するためのエンドツーエンドのクラウドネイティブ フレームワーク

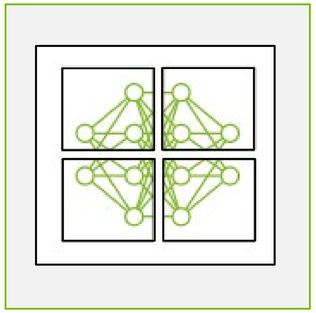
データ
収集



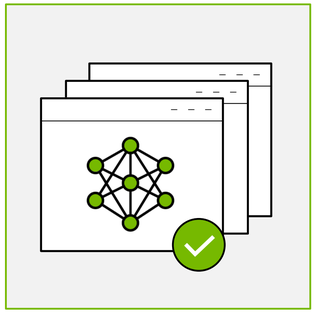
データ
キュレーション



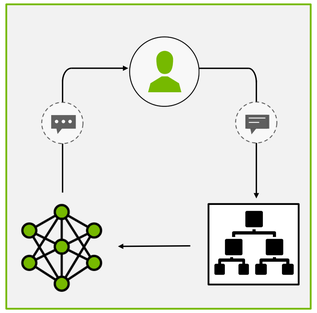
事前学習



モデルの
カスタマイズ



情報検索



推論

ガードレール

データの準備

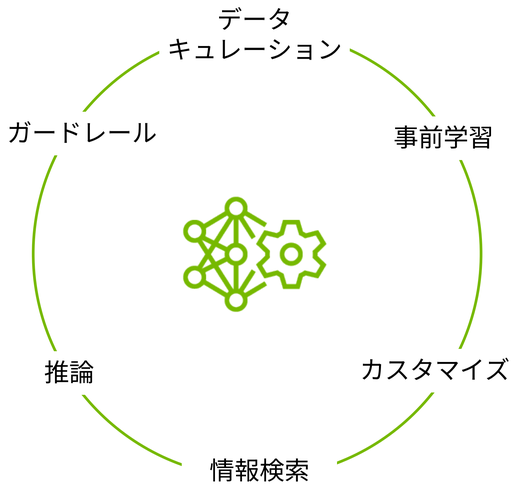
トレーニングとカスタマイズ

展開

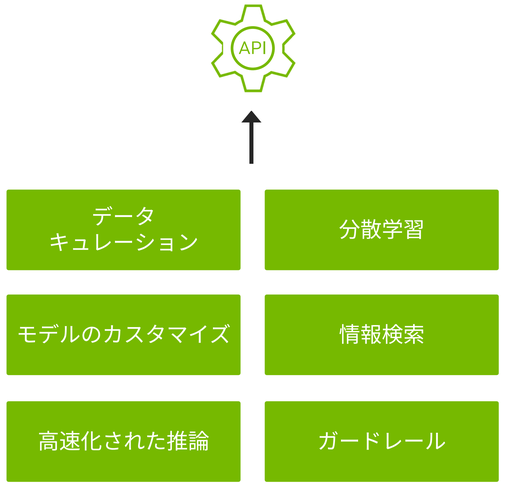
NVIDIA NeMo

AI ファウンドリーを実行してカスタム LLM アプリケーションを作成するためのソフトウェアと API

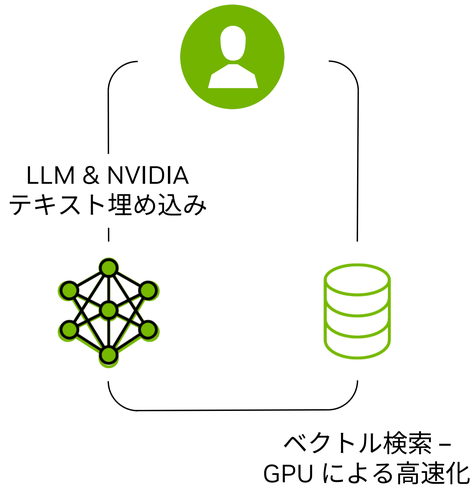
カスタマイズ、最適化、展開のためのフレームワーク



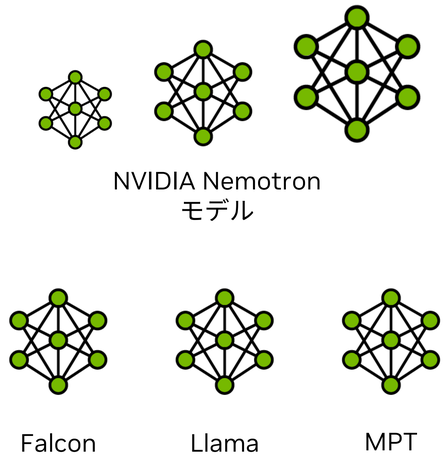
どこでも実行できるマイクロサービスと API



次世代の検索およびチャットボットのための 検索拡張生成 (RAG)

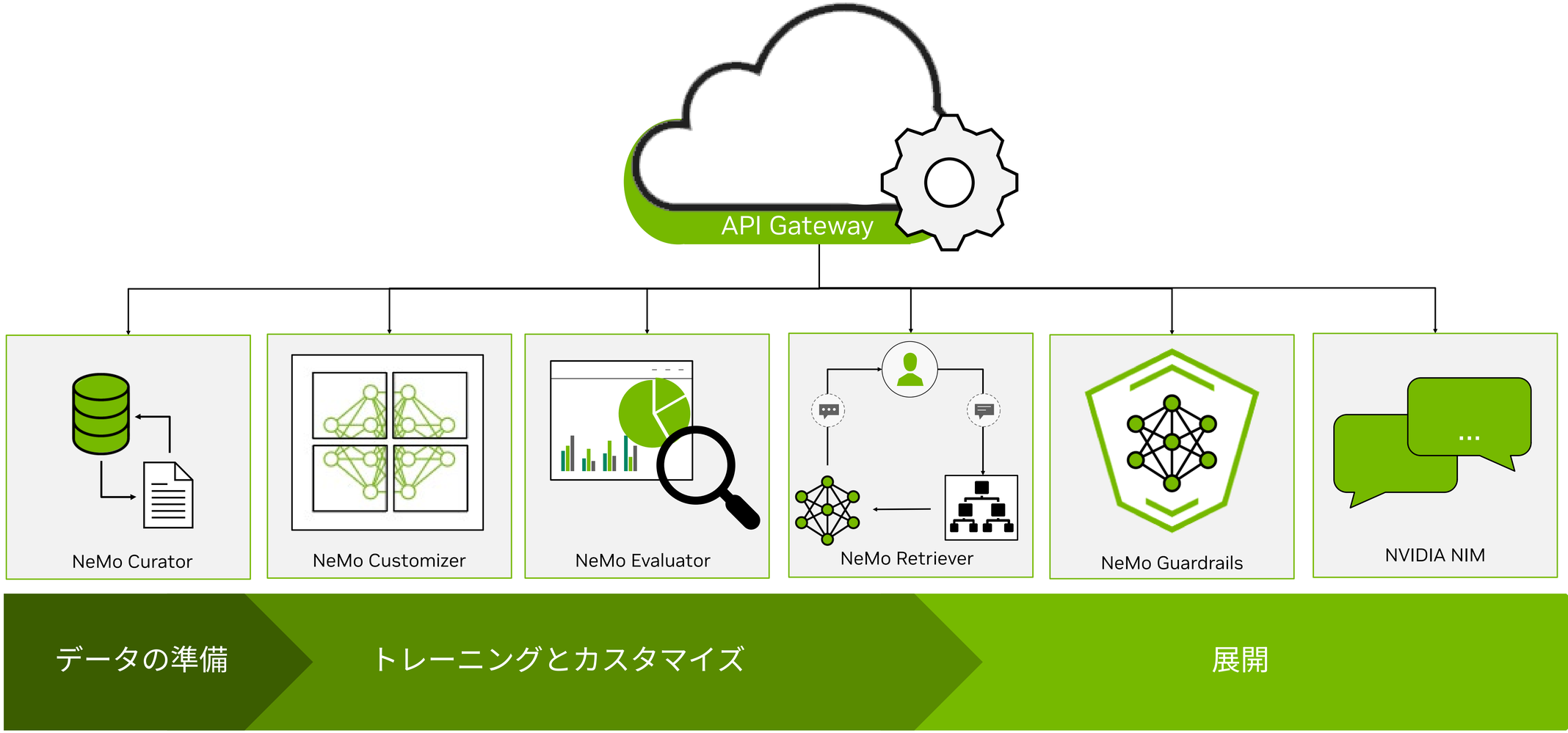


最先端の事前学習済みモデル



エンタープライズ向け 生成 AI アプリケーションの構築

NVIDIA NeMo を使用して生成 AI モデルを構築、カスタマイズ、展開

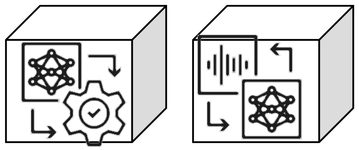


Microsoft Azure　aws　Google Cloud　ORACLE　NVIDIA DGX Cloud　DELL Technologies　Hewlett Packard Enterprise　Lenovo　SUPERMICRO　NVIDIA

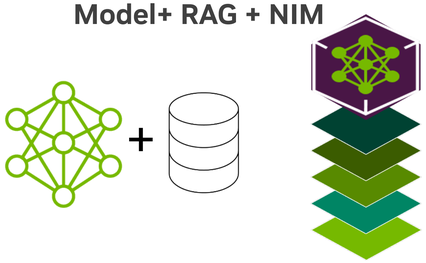
チップ設計のためのLLM アシスタント - ChipNeMo

ドメイン適応型事前学習＋RAG

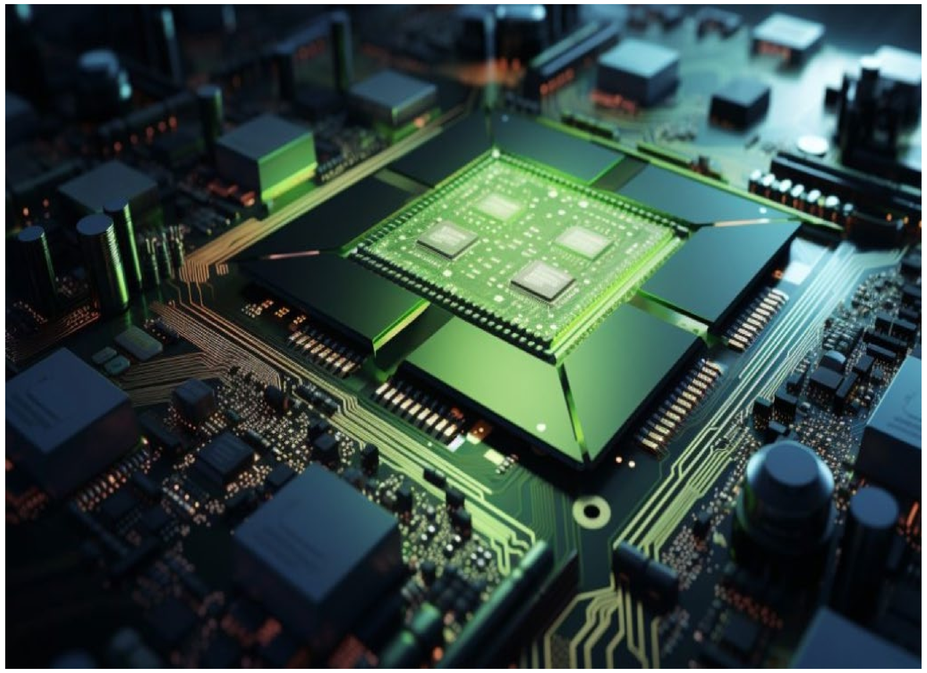
NeMo FW



Model+ RAG + NIM



Aligner + Guardrails



GPU ASIC アーキテクチャのQ&A

バグ分析
& レポート

VLSI ツール用のコード生成

最も複雑なエンジニアリングの1つである半導体設計を支援するため、NVIDIA研究所より構築されたAI**コパイロット**

GPUのアーキテクチャや設計に関する質問に答えながら、技術者が早い段階で技術文書を見つけられるようサポートします
また、チップ設計者が使用する2つの専門言語で**約10〜20行のソフトウェアのスニペットを作成**し、新たなコード開発を容易にします

独自のデータを使って基盤モデルをカスタマイズすることで、研究者たちは、**より小規模な13Bのパラメータモデルが、より大規模な汎用LLMを凌駕できること**がわかりました

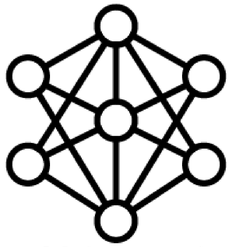
ドメインに適応させたChipNeMoカスタマイズ・ワークフロー

コード生成、要約、設計支援ChatBotの3つのユースケースを実現するドメイン適応型事前トレーニング + SFT

事前学習

1兆トークン
インターネットデータ

10^5-10^6 GPU 時間

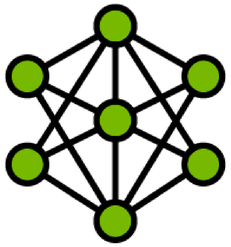


基盤モデル
Llama 2
(70B, 13B, 7B)

ドメイン適応
事前学習

240億トークン
チップ設計
文書・コード

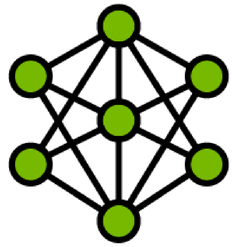
~5000 GPU 時間



ChipNeMo
基盤モデル
(13B, 7B)

教師ありファイン
チューニング
(SFT)
12万8千
チャットデータ
+ 1100
タスクデータ

~100 GPU 時間



ChipNeMo
チャットモデル
(13B, 7B)

モデルの事前学習 | 学習とカスタマイズ | デプロイ

https://arxiv.org/abs/2311.00176

NVIDIA

生成AIに対する企業の動向

認知

ChatGPTは2022年後半に発表され、わずか2ヶ月で1億人以上のユーザーを獲得。あらゆるレベルのユーザーがAIを体験し、そのメリットを肌で感じることができた

検証

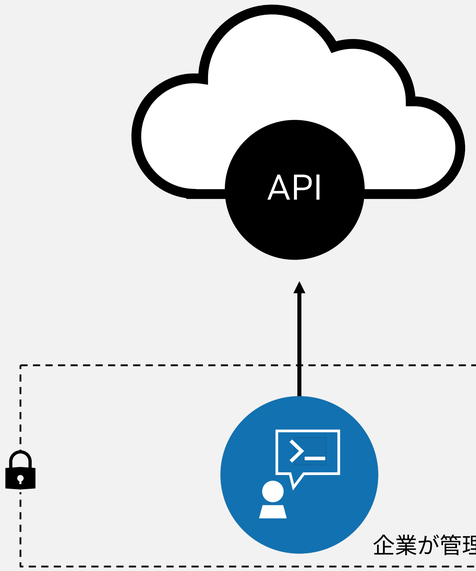
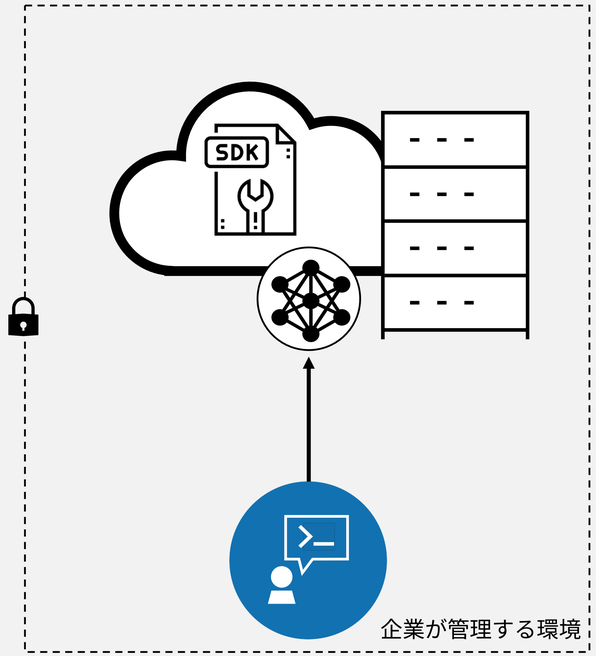
Llama 2、Mistral、NVIDIAなどを含むAPIサービスとオープンモデルで、企業アプリケーション開発者が生成AIアプリケーションのPOCを開始

活用

組織は予算を確保し、本番で生成AIをサポートするための加速インフラを構築する取り組みを強化

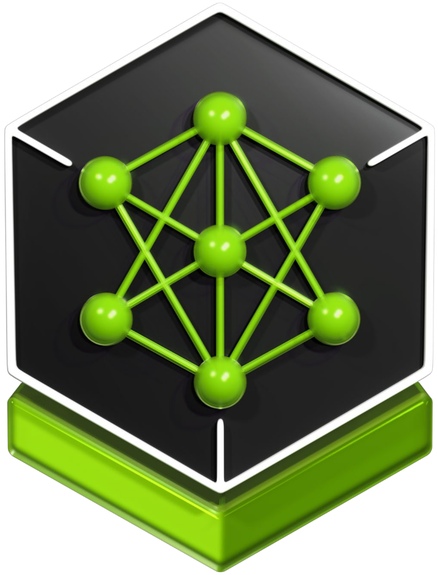
生成 AI の展開 オプション

企業による生成 AI アプリケーションの検証

	マネージド型生成 AI サービス	オープンソース型の展開	
使いやすい 開発用 API			データ センターとクラウドの**どこでも実行可能**
AI を使い始めるための**近道**			セルフ ホスト環境でデータを**安全に管理**
マネージド環境に**制限されたインフラストラクチャー**			各種インフラストラクチャーに合わせた**チューニングが必要**
データとプロンプトは**外部と共有**			API やファインチューニングされたモデル用の**カスタム コード**
限定的な制御を生成 AI 戦略全体に行う			**継続的なメンテナンス**とアップデート

NVIDIA NIM:最大5倍の実行速度を発揮する最適化されたAIモデル

コミュニティモデル - パートナーモデル - NVIDIAモデル



NVIDIA INFERENCE MICROSERVICE

学習済みAIモデル
パッケージ化され最適化されている
CUDAインストールベース

スピーチ

デジタルヒューマン

コンピュータ・ビジョン

バイオロジー

シミュレーション

言語

地域言語

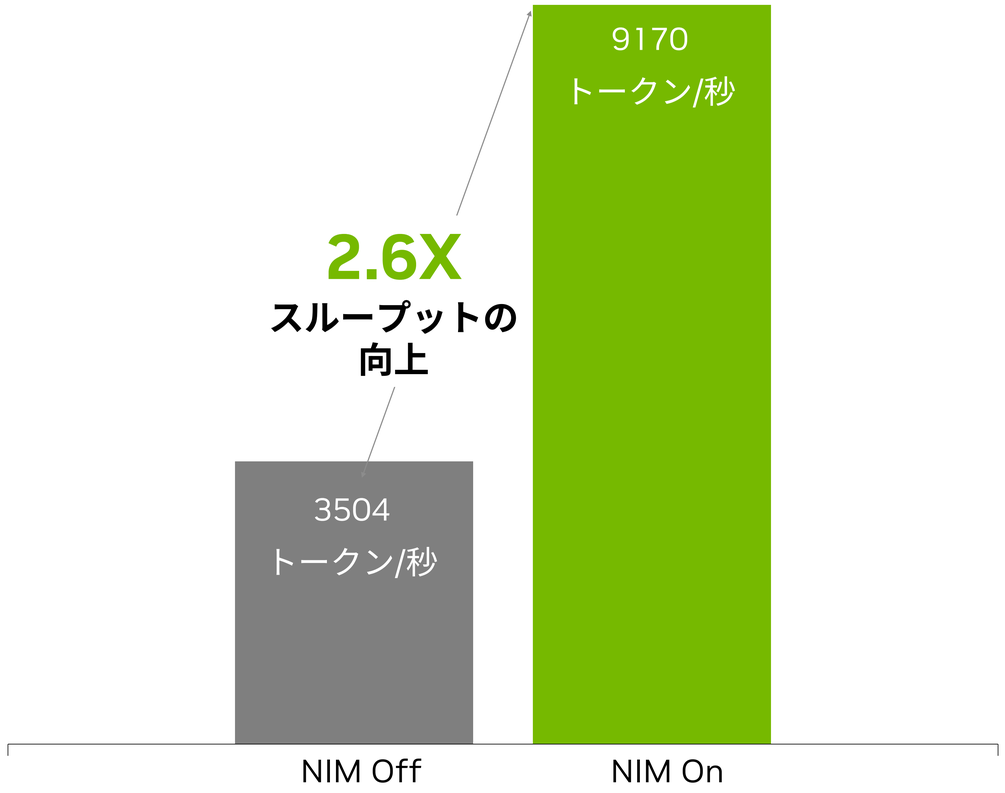
ビジョン・ランゲージ

RAG

ADEPT gettyimages Google Meta MIT MISTRAL AI_ NVIDIA shutterstock snowflake

簡単に効率を上げられる

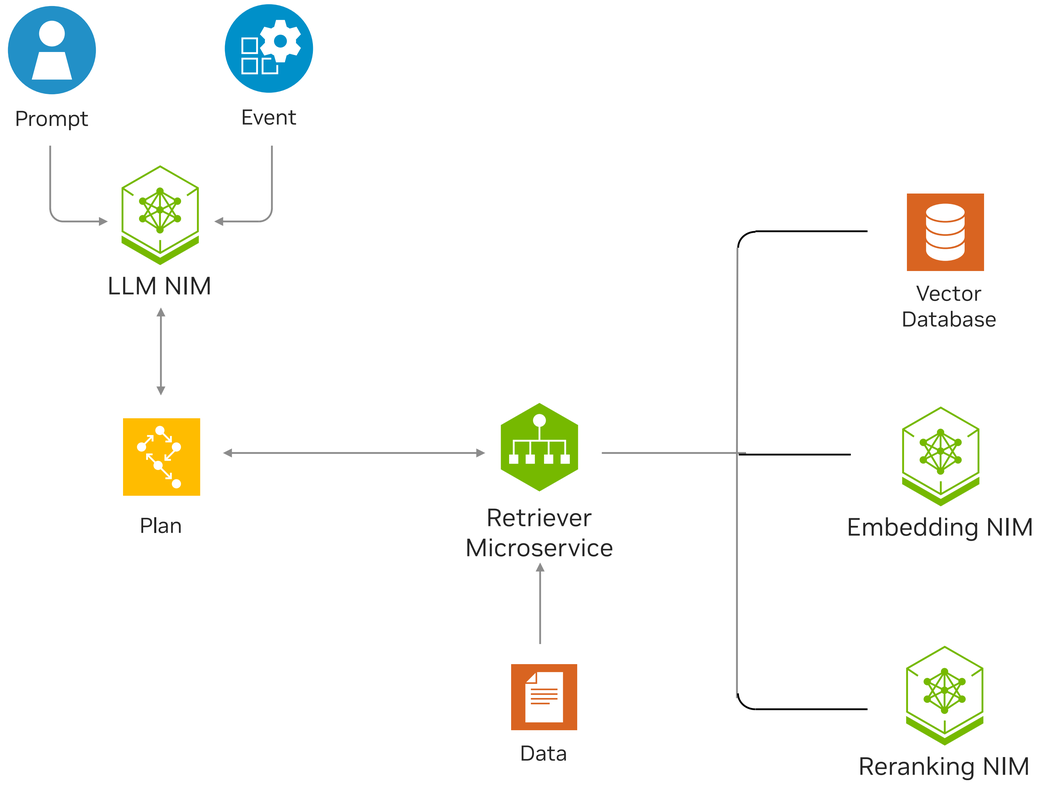
Llama3.1-8B-instruct, 1 x H100SXM



Llama3.1-8B-instruct, 1 x H100SXM; input 1000 tokens, output 1000 tokens. Concurrent requests: 250. NIM OFF : FP8. throughput 3,504 tokens/s, TTFT 1.7s, ITL: 68ms. NIM ON : FP8. throughput 9,170 tokens/s, TTFT 3.4s, ITL: 24ms.

NeMo RetrieverでRAGアプリケーションを強化

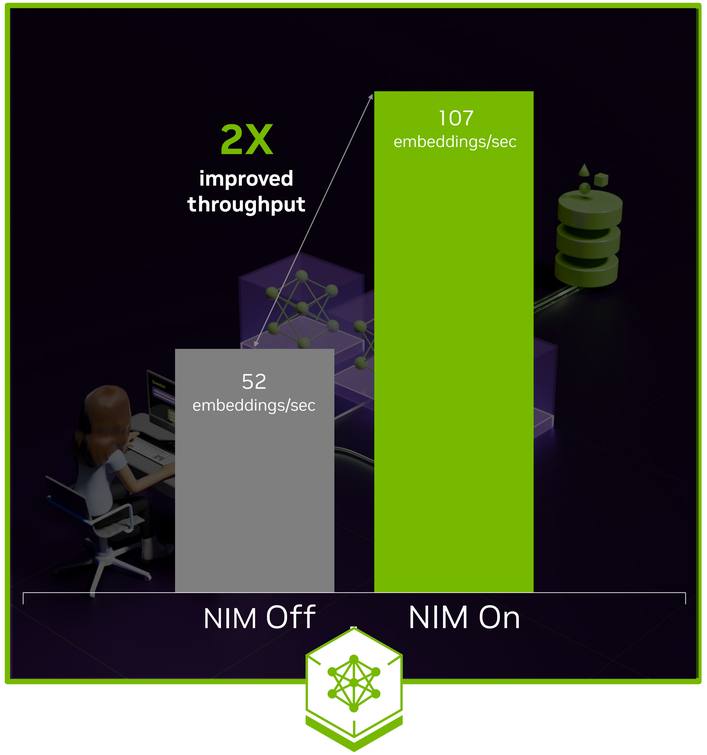
世界最高のオープンな商用テキストQ&A検索パイプライン



- 最適化された推論エンジン
- ワールドクラスのモデルとコミュニティ・モデルのサポート
- 柔軟でモジュール化された展開
- カスタマイズ可能なモデルとパイプライン
- プロダクション・レディ

NVIDIA NeMo Retriever NIM

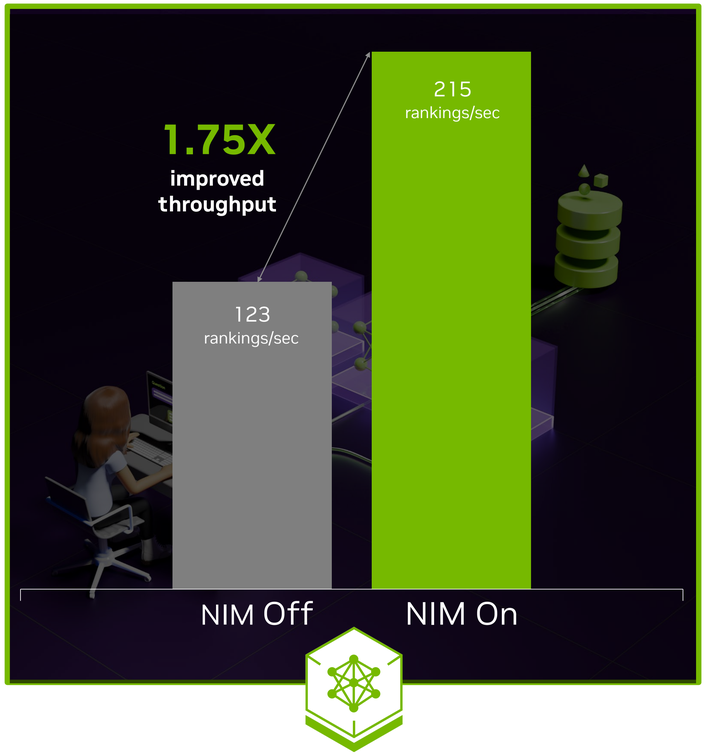
ai.nvidia.comにてダウンロード可能



NV-EmbedQA-Mistral7B-v2

Multilingual text embedding model

NV-EmbedQA-Mistral7B-v2, 1xH100 SXM; passage token length: 512, batch size: 64, concurrent client requests: 3; NIM Off: FP16, P90 latency: ~3.8x; NIM On: FP8, P90 latency: ~1.8s.



NV-RerankQA-Mistral4B-v3

Text reranking for high accuracy question answering

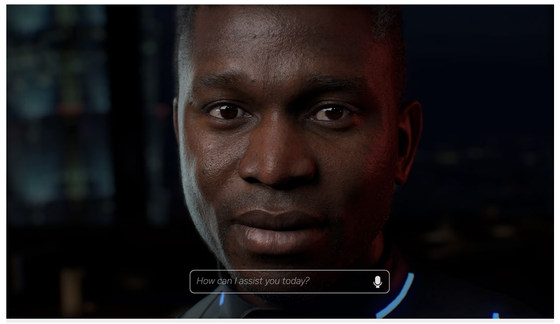
NV-RerankQA-Mistral4B-v3, 1xH100 SXM; query token length: 20, passage token length: 512, batch size: 40, concurrent client requests: 3; NIM Off: FP16, P90 latency: ~1s; NIM On: FP8, P90 latency: ~0.56s

NVIDIA

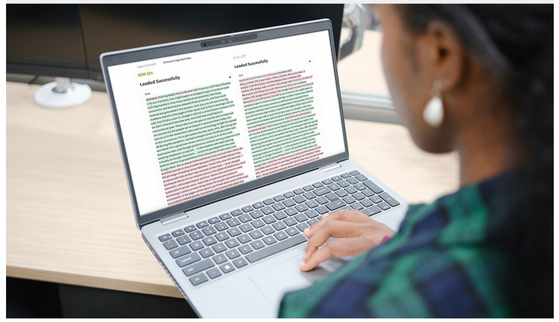
NVIDIA NIM Agent Blueprints

build.nvidia.comで入手可能

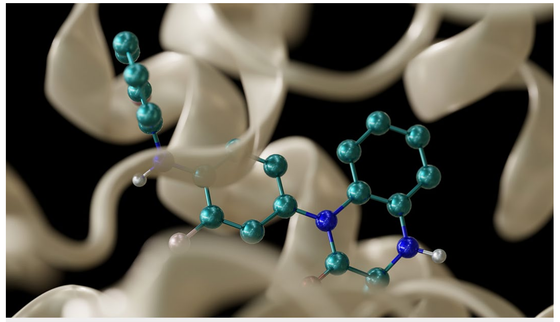
顧客サービスのための
デジタルヒューマン



エンタープライズRAGのための
マルチモーダルPDFデータ抽出



創薬のためのジェネレーティブ・
バーチャル・スクリーニング



■ ■ ■

月次リリース

NVIDIA NIM Agent Blueprint

アプリケーション例

簡単に再現できるインタラクティブな体験

サンプルデータ

ワークフロー・テストのための公開データ

リファレンス・コード

実績のある事前学習済みモデルを活用

アーキテクチャ

API定義、NIMなどを含むリファレンス・アーキテクチャ

カスタマイズ・ツール

モデルのカスタマイズと評価

オーケストレーション・ツール

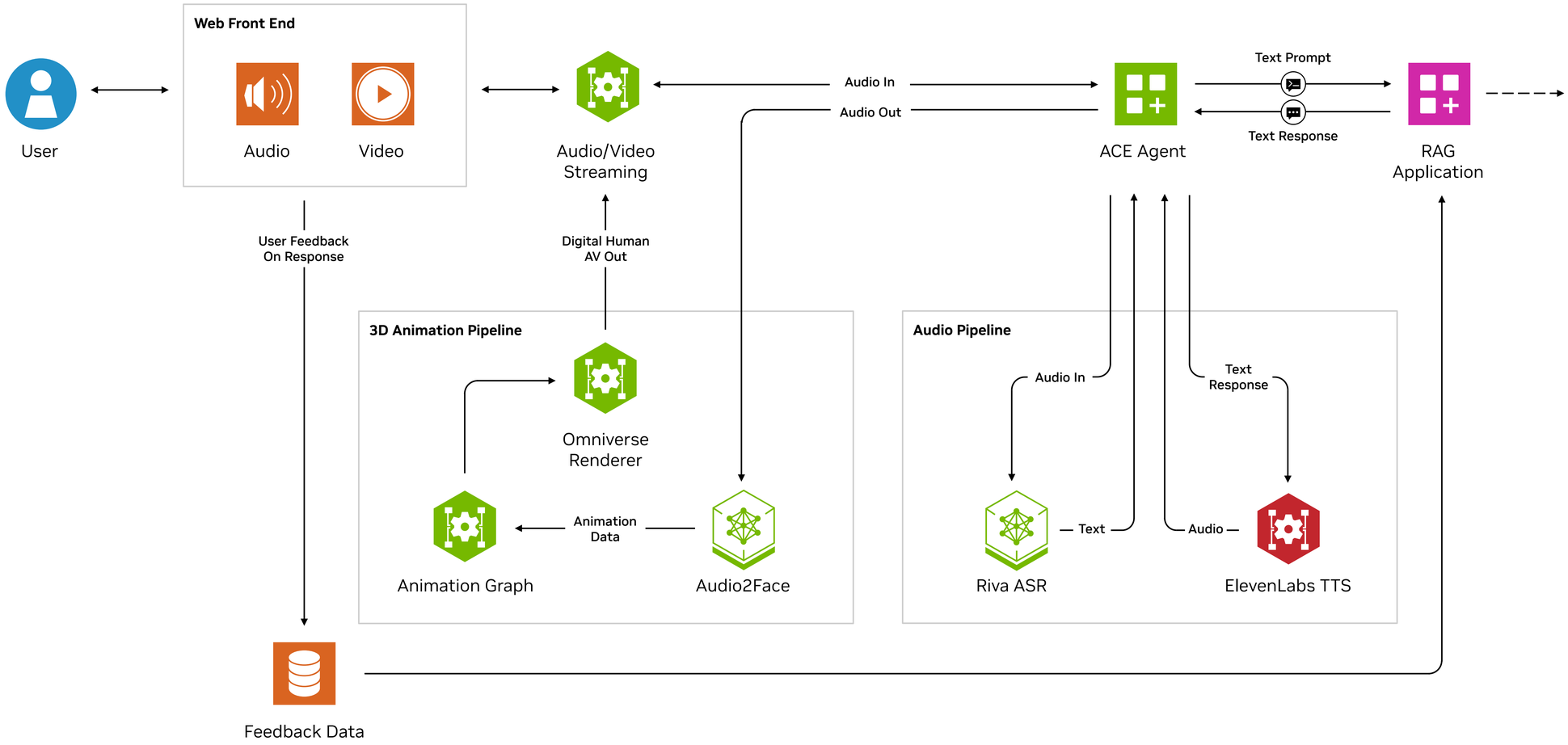
ワークフローマイクロサービスのデプロイと管理

顧客サービスのためのデジタルヒューマン

デジタル・ヒューマン・エコノミー市場は2035年までに125億ドル規模と推定される

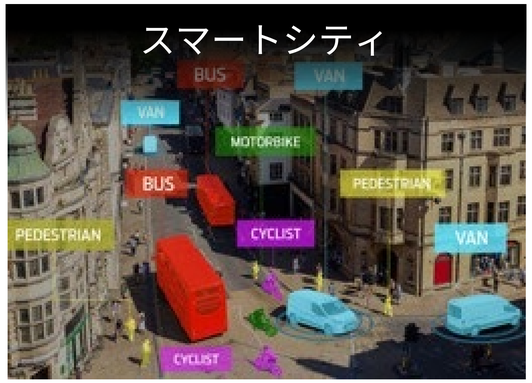
恩恵

- ユーザー向けアプリケーションのエンゲージメントと満足度の向上
- 正確な肌、髪、アニメーション、発話を持つ、本物そっくりの3Dデジタル人間を作成します。
- 企業のアプリケーションやデータとの自然な会話を可能にする



マルチモーダル生成AI－Vision/Language

大量の動画データから重要な洞察を得る



マルチモーダル (VLM)：LLaVA

（分かりやすくするために日本語翻訳しています）

問：この画像のおかしな点はなんですか？

回答：画像は、走行中の車両、特に黄色いタクシーの荷台に置かれたアイロン台の上でアイロンをかけている人の姿である。

このような場所でアイロンがけをすることは、現実的でないだけでなく、アイロンがけをする人に大きな危険をもたらすため、これはこれは異常で危険な状況である。

アイロンがけをしている人は、風雨にさらされ、転倒の危険や車の走行による事故の可能性もある。

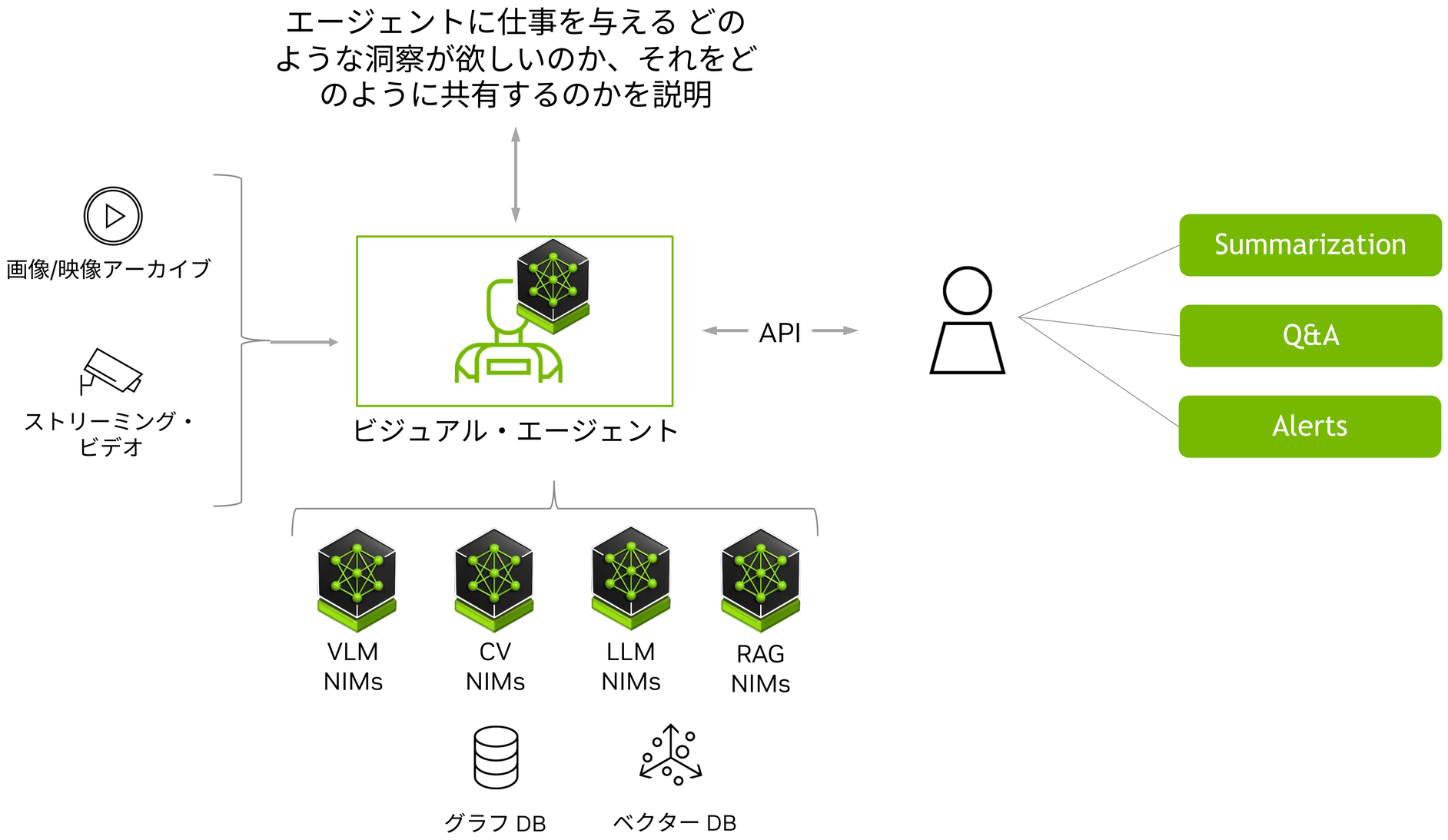
さらに、アイロン台とその人の位置がドライバーの視界を妨げたり、他の道路利用者に危険を及ぼす可能性もある。

ユーモラスでシュールな光景だが、常識や安全対策を無視したものであるため、真に受けることはないだろう。

Demoサイト　https://llava.hliu.cc/ での実行結果より

ビデオ検索と要約 のためのNVIDIA AI Blueprint

Video Search and Summarization



STOP STOP

POLICE

Eureka

大規模言語モデルのコーディングによる人間レベルの報酬設計

Environment Code

```
class ShadowHandPenSpin(VecTask):
  def compute_observations(self):
    self.obj_pose = ...
    self.obj_pos = ...
    self.obj_rot = ...
    self.obj_linvel = ...
    self.obj_angvel = ...

    self.tgt_pose = ...
    self.tgt_pos = ...
    self.tgt_rot = ...

    self.fingertip_state = ...
    self.fingertip_pos = ...

    self.compute_full_state()

  def compute_full_state(self):
    ...
```

Task Description

```
To make the shadow hand spin the pen
to a target orientation
```

Coding LLM (GPT 4)

Reward Candidate Sampling

```
def compute_reward(
  obj_rot, obj_angvel, ...
):
  ...
  # Angular velocity penalty
  av_norm = torch.norm(obj_angvel)
  av_penalty = torch.where(
    av_norm > 2.0,
    torch.exp(av_norm - 2.0)
  )
  ...
```

GPU-Accelerated RL

Reward Reflection

```
We trained a RL policy using the
provided reward function code...
av_penalty: ['0.02', '0.05',
'0.05', '0.04', '0.03', ...]
success_rate: ['0.00', '0.38',
'1.57', '3.01', '3.95', ...]
Please carefully analyze the policy
feedback and provide a new, improved
reward function...
```

Query with Feedback

Eureka

Isaac Manipulator

基盤モデル | モジュラー型 GPU アクセラレーテッドライブラリー

ソリューションインテグレーター

OEM

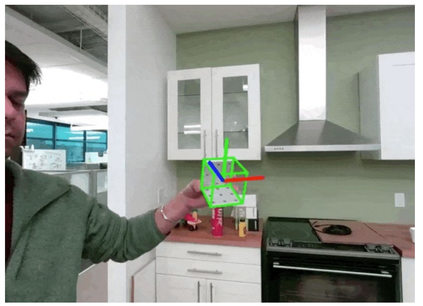
ISV / SDP

開発者プラットフォーム

UR | MiR TERADYNE Robotics　YASKAWA　PICKNIK

FRANKA ROBOTICS　SOLOMON

ISAAC MANIPULATOR



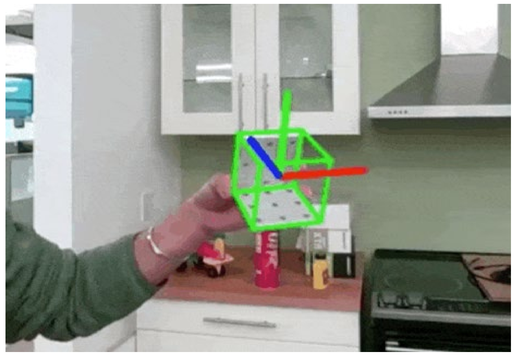
NVIDIA

基盤モデル

合成データを用いた学習済みモデル | 高精度かつ高性能

SyntheticaDETR

インダストリアル向け屋内アセット検出



FoundationPose

未知の物体の 6D 姿勢推定 およびトラッキング

FoundationGrasp

把持ポイントの特定およびアノテーション

cuMotion

GPU-アクセラレーテッド軌道計画

AIの次の波に対応する3つのコンピューター

OMNIVERSE

AI

ROBOT