NVIDIA と Google Cloud が加速する AI ファクトリーと新しい産業革命

シニア テクニカル マーケティング マネージャー 澤井 理紀

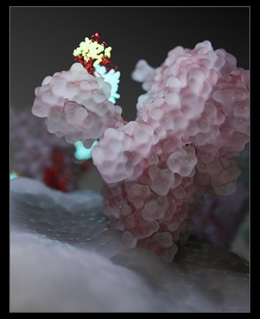
あらゆるところに AI が導入

AI ファクトリーがもたらす新たな産業革命

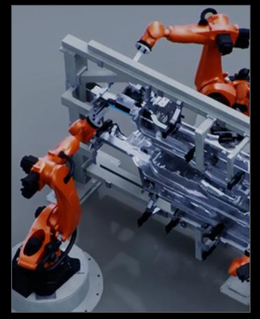
輸送

自然リソース

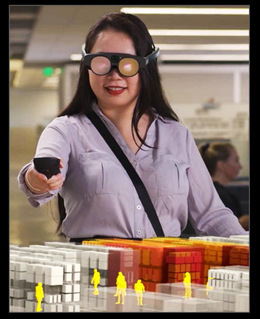
ヘルスケア



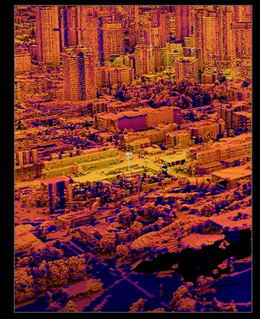
製造



小売り



通信



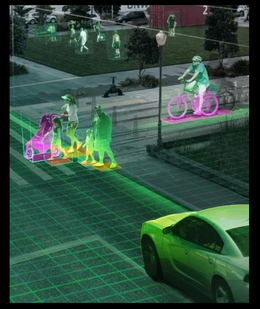
インターネット

金融

HPC

ロボティクス

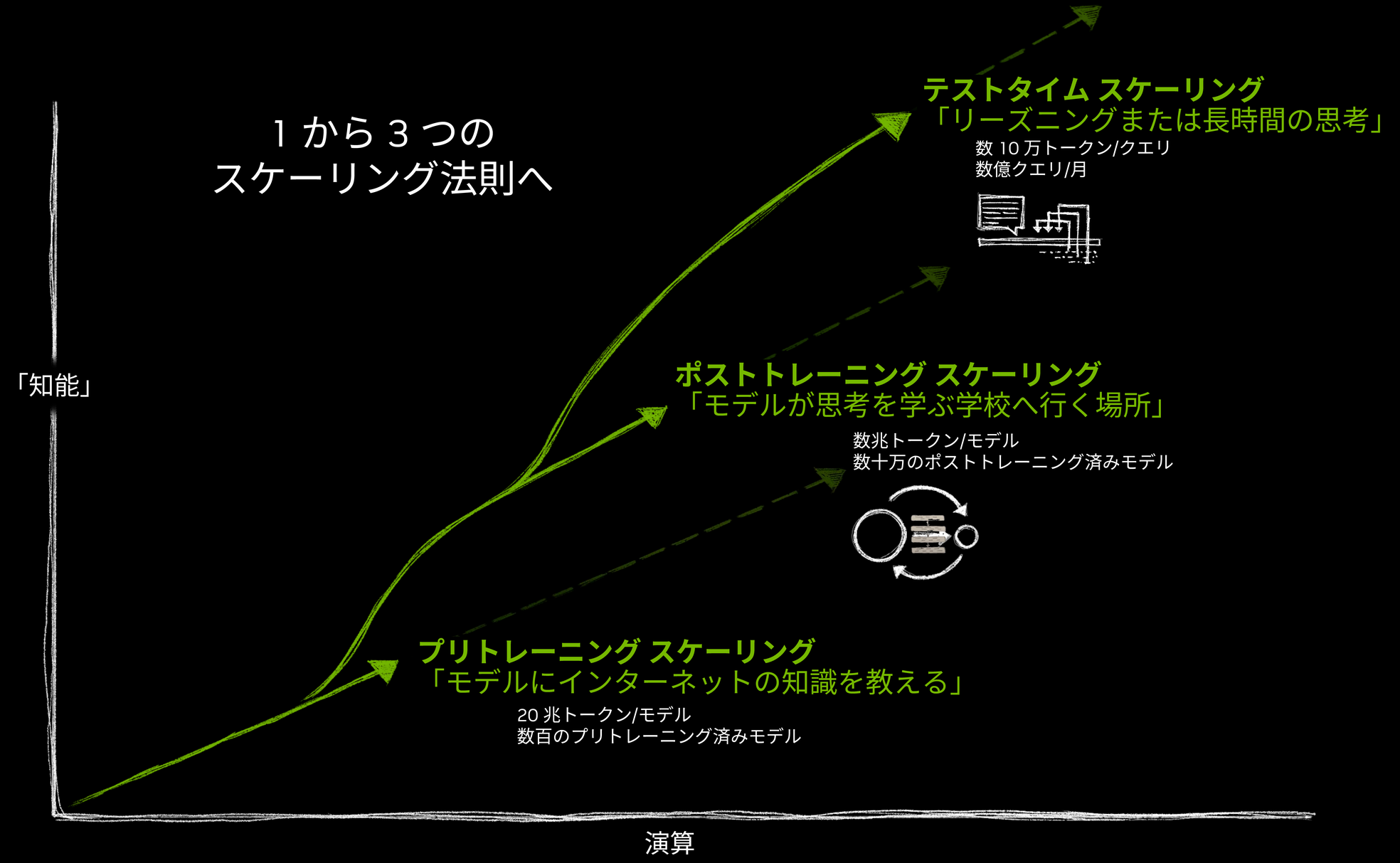
スマート シティ



教育

トークン

AI スケーリング則が演算需要の指数関数的な増加を牽引



エージェント型 AI がすべての人の働き方を変革

10 億以上
ナレッジ
ワーカー

3000 万
ソフトウェア
開発者

1500 万
コール センター
エージェント

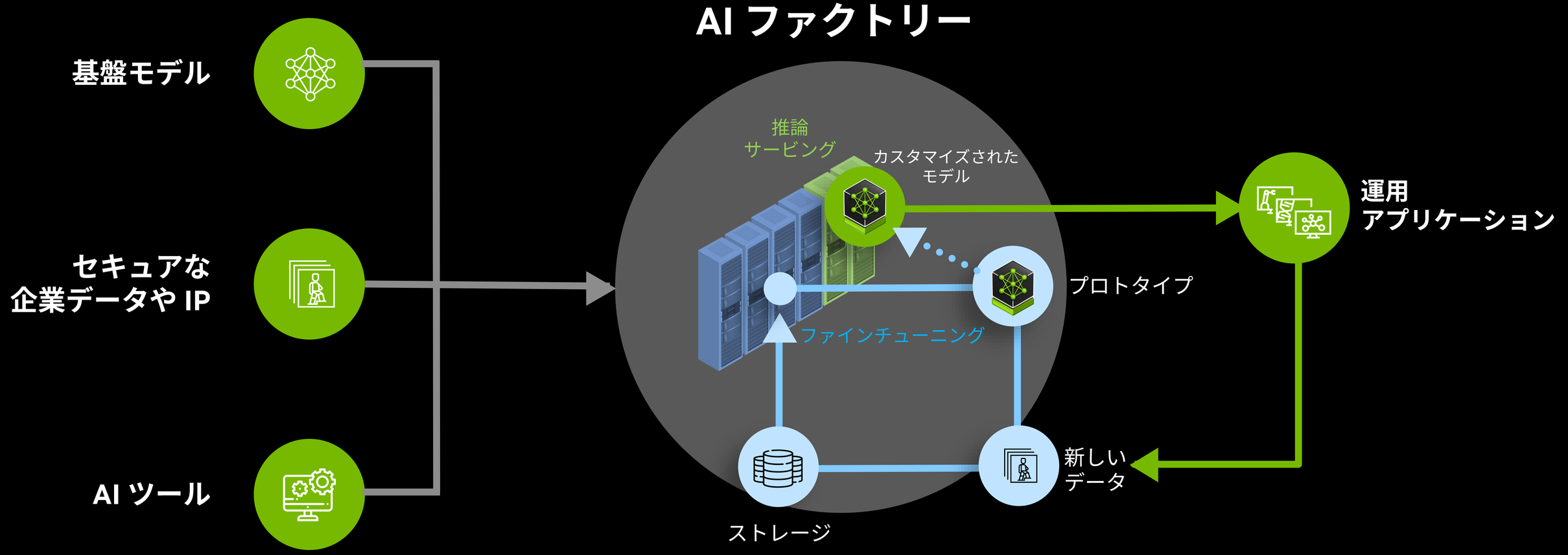
800 万以上
科学者と研究者

5500 万
IT 専門家

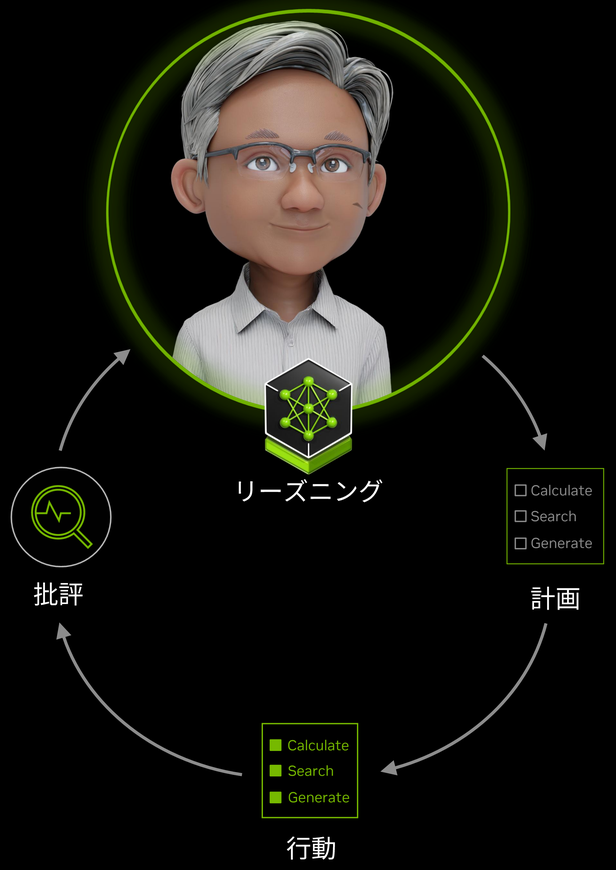
5000 万
世界中の
クリエイター

NVIDIA と Google Cloud が AI ファクトリーを実現

大規模にエンタープライズのインテリジェンスを製造



リーズニングは高度な AI エージェントを構築する鍵



リーズニング モデルはエージェントを支援：

- 文脈の理解
- RAG | 正確な回答の生成
- 曖昧さへの対応 | クエリの書き換え

リーズニングによる計算需要の指数的増加

計算量とともに増加するコスト — トークン単価の削減にはフルスタックのイノベーションが必要

ナレッジ
プリトレーニング済みモデルがインターネットの知識をシミュレート
100 トークン | $0.50 / 百万トークン

What is NVIDIA?

NVIDIA is a multinational technology company that specializes in designing and manufacturing graphics processing units (GPUs), high-performance computing hardware, and artificial intelligence (AI) technologies. The company was founded in 1993 by Jensen Huang, Chris Malachowsky, and Curtis Priem.

NVIDIA is headquartered in Santa Clara, California, and has become one of the leading companies in the field of computer graphics, gaming, and AI.

リーズニング
リーズニング モデルは問題を思考する
~1 万トークン | 100 倍のコスト

I need to figure out where to seat my family of 8 for dinner at a round table. My in-laws don't get along with my parents. My sister needs to sit next my little brother to help him eat. My wife really doesn't want to sit next to my mom. Where should each person be seated?

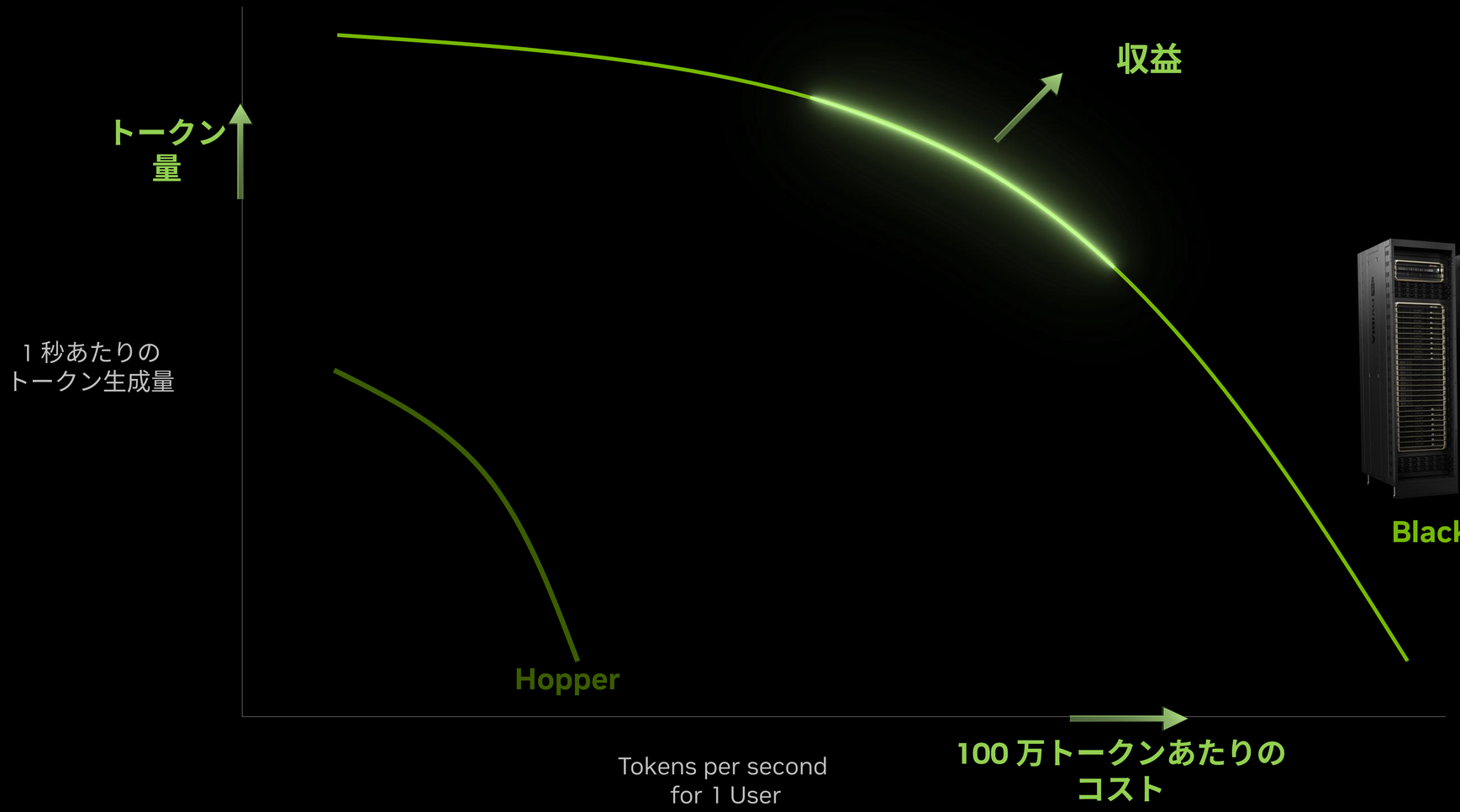
To seat your family of 8 harmoniously at a round table
while respecting all constraints, here's the optimal arrangement:

You Wife Dad Mom Sister Little Brother Mother-in-law Father-in-law

Key Constraints Met:

- In-laws vs. Parents: Separated by your sister, brother, and you.
- Sister & Brother: Adjacent for assistance.
- Wife & Mom: Not next to each other (separated by you and dad).

AI ファクトリーの生産性向上

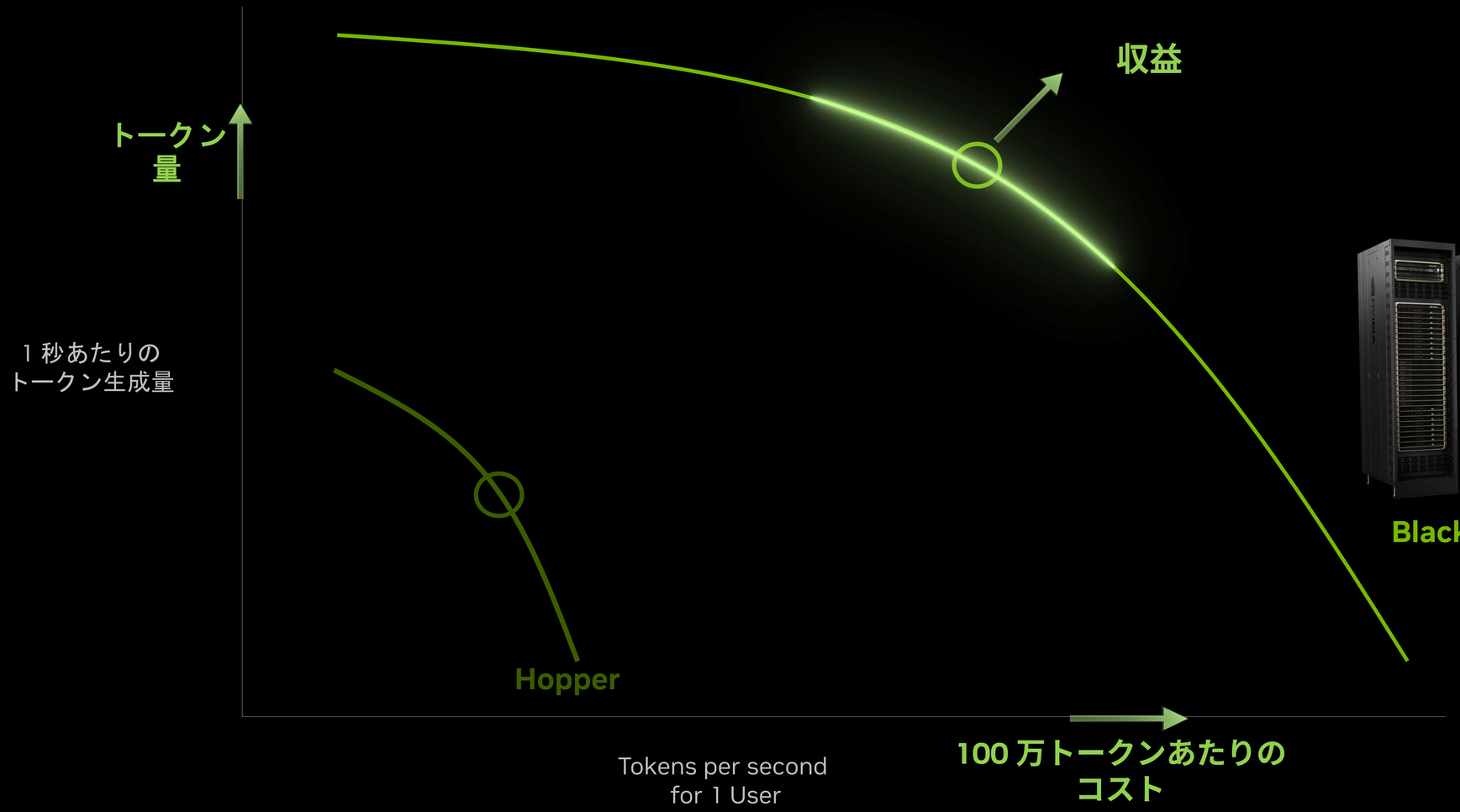


Blackwell

72 Blackwell GPUs
36 Grace CPUs
13.4 TB HBM3e | 576 GB/s
130 TB/s NVLink

AI ファクトリーの生産性向上

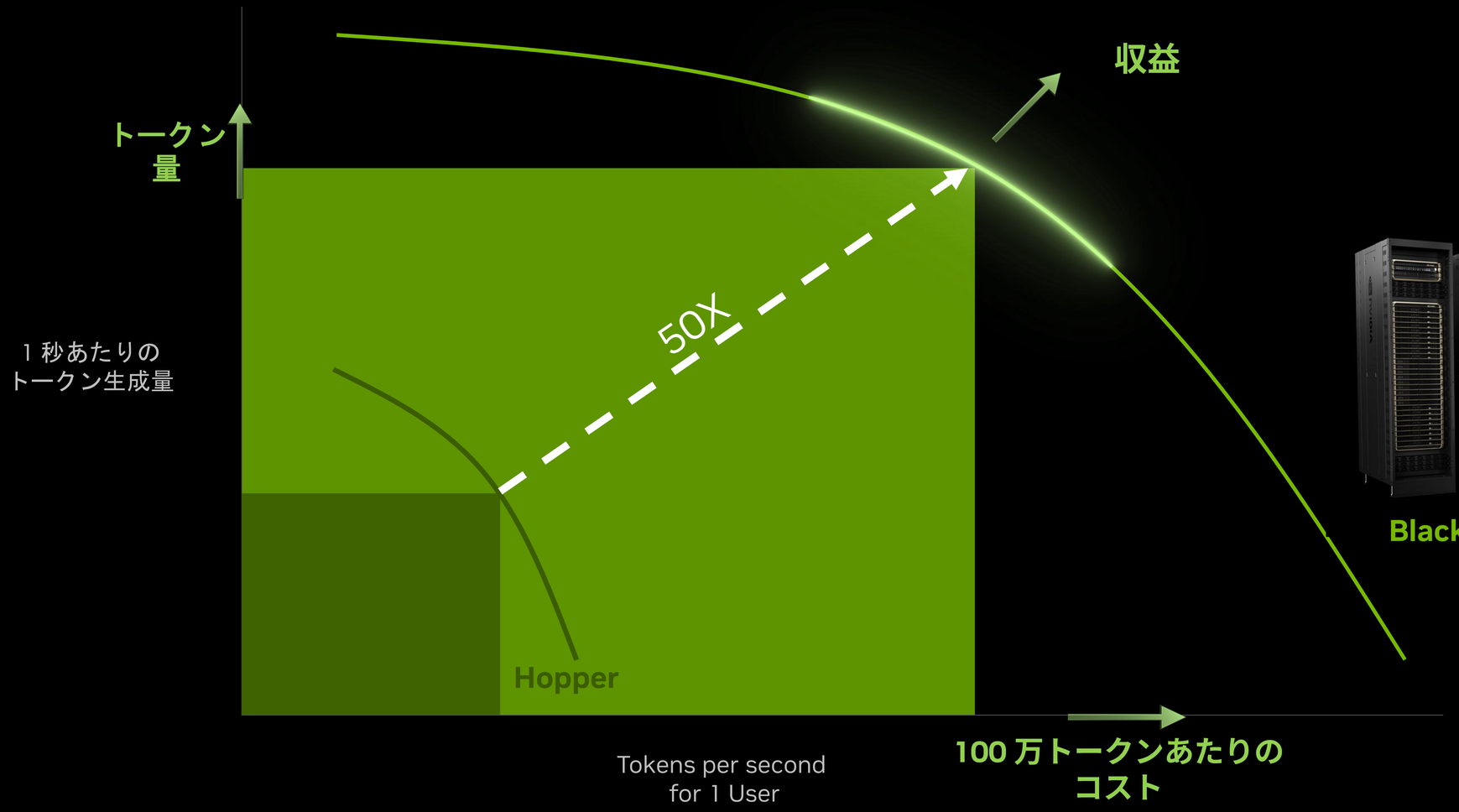
高スループット × 高インタラクティビティ = 総トークン出力



Blackwell

72 Blackwell GPUs
36 Grace CPUs
13.4 TB HBM3e | 576 GB/s
130 TB/s NVLink

AI ファクトリーの出力が収益を牽引



Blackwell

72 Blackwell GPUs
36 Grace CPUs
13.4 TB HBM3e | 576 GB/s
130 TB/s NVLink

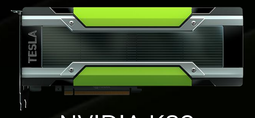
NVIDIA と Google Cloud、
最先端 AI をクラウドに提供

NVIDIA と Google Cloud: フルスタック コラボレーションの歴史

GPU インスタンスからフレームワーク、ソフトウェア、そしてその先へ

2017
TensorFlow 1.0 on NVIDIA GPUs and TensorRT

P100 with NVLink



NVIDIA K80

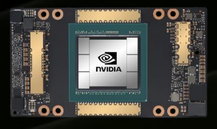
2018
JAX Support

x16 V100 Shape

2019
TensorFlow 2.0

N1 (T4)

2020
Spark, Dataproc and RAPIDS

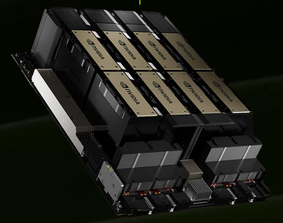


A2 (A100)

2021
Dataflow

2022
OpenXLA

2023
Vertex AI



A3 (H100)

G2 (L4)

2024
Gemma and Cloud Run NIM on GKE

A3 Ultra (H200)

2025
DGX Cloud on GB200

A4X (GB200)

A4 (B200)

NVIDIA Blackwell が Google Cloud から利用可能

A4 VM
NVIDIA HGX B200
提供中

Blackwell GPU x8

14.4 TB/s GPU-to-GPU 帯域幅

1.44 TB 総 GPU メモリ

NVIDIA ConnectX-7 NIC

性能と多用途

Compute Engine

GKE

A4X VM
NVIDIA GB200 NVL72
プレビュー

最大 Grace CPU x36 と Blackwell GPU x72

130 TB/s GPU-to-GPU 帯域幅

13.4 TB 総 GPU メモリ

NVIDIA ConnectX-7 NIC

大規模 AI ファクトリー

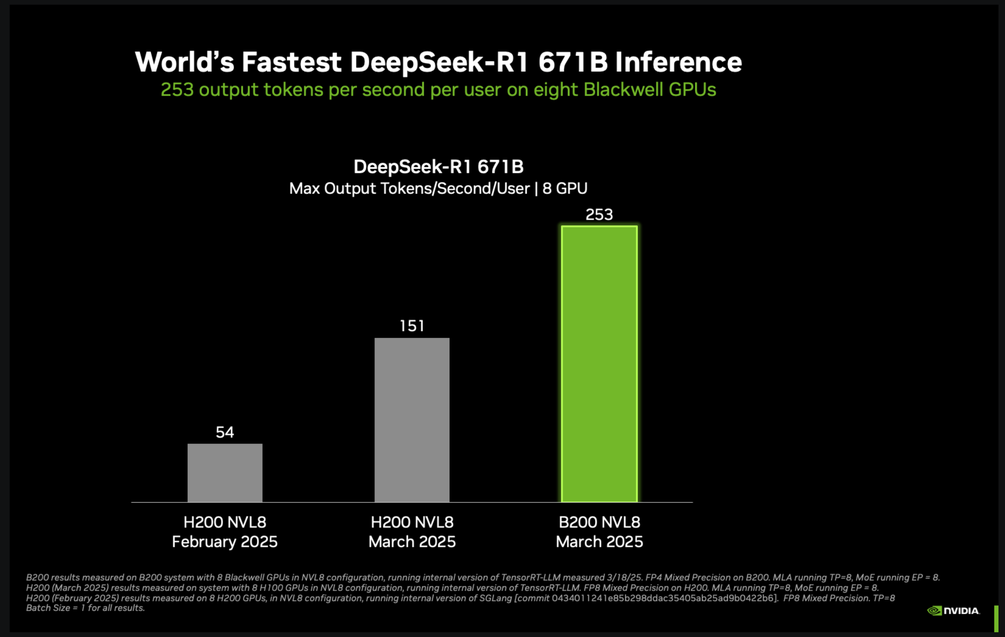
Vertex AI

Dynamic Workload Scheduler

DeepSeek-R1 推論で Blackwell が記録を樹立

わずか 1 ヶ月で約 5 倍の性能向上を達成

- DGX B200 で
  ユーザーあたり 250 TPS/ユーザーの最小レイテンシ、
  30,000 TPS/サーバーの最大スループットを達成
- 推論開発者ツールのオープン エコシステムによって
  実現されたパフォーマンス
- DeepSeek-R1 の FP4 実装により、
  高精度かつ高性能を達成
- Vertex AI Model Garden で利用可能

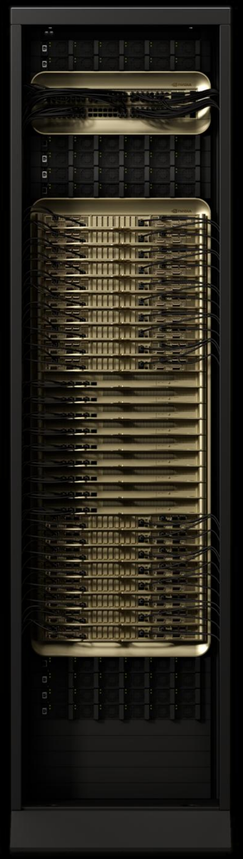


精度を維持しつつ大幅な高速化

	MMLU	GSM8K	AIME 2024	GPQA Diamond	Math-500
DeepSeek R1-FP8	90.8%	96.3%	80%	69.7%	95.4%
DeepSeek R1-FP4	90.7%	96.1%	80%	69.2%	94.2%

GB200 NVL72 が最高のシステム スループットを提供

シングル NVLink ドメインの 72 GPU による Llama 3.1 405B ベンチマーク



Blackwell GPU x72
Grace CPU x36
13.4 TB HBM3e | 576 GB/s
130 TB/s NVLink

Google Cloud 上の NVIDIA Blackwell で新たな開拓を実現

NVIDIA Blackwell が、Google Cloud 上の DGX Cloud で利用可能

DGX Cloud でトレーニングし、あらゆる Google Cloud プラットフォームに展開

NVIDIA DGX Cloud

Google Cloud

Google Cloud と共同開発された、フルマネージド AI トレーニング プラットフォーム

あらゆるレイヤーで最適化

トレーニング期間短縮のための NVIDIA AI およびクラウド エキスパートへのアクセス

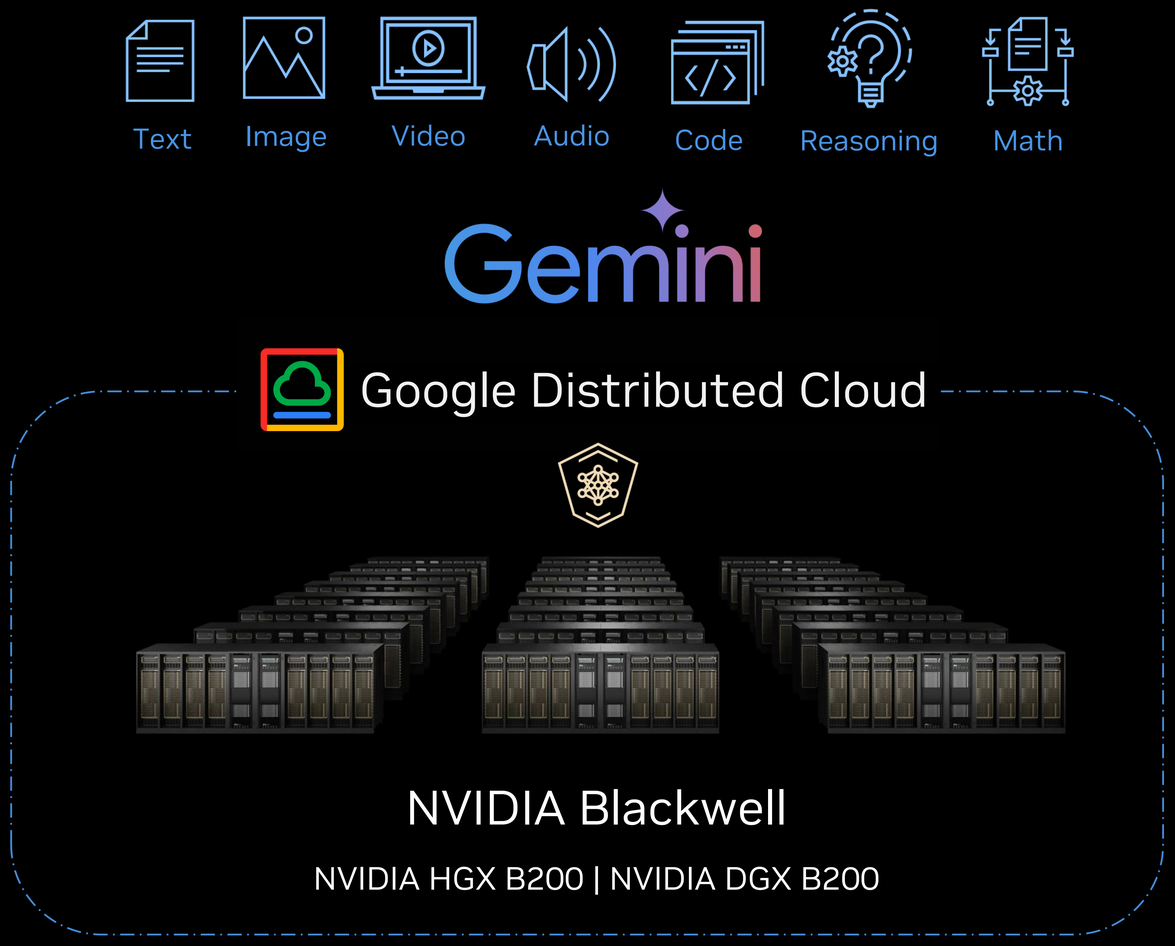
短期的な大規模、連続した高性能クラスター

NVIDIA によるフルマネージド、シングル コンタクト ポイント

NVIDIA GB200 NVL 72 が DGX Cloud で近日提供予定、Google Cloud Marketplace で利用可能

Google Gemini の高度なリーズニング能力をオンプレミスで

NVIDIA Blackwell on Google Distributed Cloud が、エージェント型 AI を規制業界の企業にもたらす



2025 年第 3 四半期にパブリック プレビュー

Google Cloud 上の
NVIDIA 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

A4 VM
NVIDIA B200

柔軟な 1、2、4、8 GPU 構成
GPU メモリ: 最大 640 GB

最大ネットワーク帯域幅: 1,000 Gbps

柔軟で、コスト効率の高い推論、
小さなトレーニング ワークロード

A3 High VM
NVIDIA H100

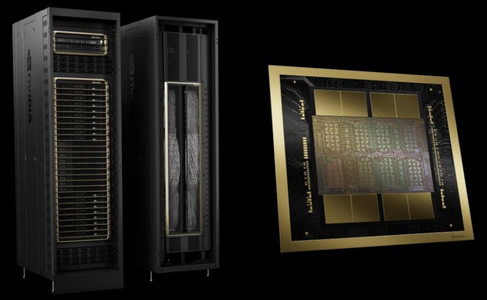
A4X VM
NVIDIA GB200 NVL72

VM あたり 8 GPU
GPU メモリ: 640 GB

最大ネットワーク帯域幅: 1,800 Gbps

分散 AI トレーニングと推論に最適化

A3 Mega VM
NVIDIA H100



GB300
Coming Soon

VM あたり 8 GPU
GPU メモリ: 1128 GB

最大ネットワーク帯域幅: 3,600 Gbps

大規模マルチノード AI トレーニングと推論

A3 Ultra VM
NVIDIA H200 + NVIDIA ConnectX-7 NICs

NVIDIA Hopper GPUs

エージェント AI を可能にする
高度なソフトウェア

NVIDIA はエンタープライズ向けエージェント型 AI の構成要素を提供

NVIDIA AI Blueprint

リサーチ アシスタント エージェント
顧客サービス エージェント
ソフトウェア セキュリティ エージェント
バーチャル ラボ エージェント
ビデオ解析 エージェント

NVIDIA NeMo

Curator
Customizer
Evaluator
Guardrails
Retriever

NVIDIA NIM

理解と リーズニング
情報検索
AI セーフティ
デジタル ヒューマン
ビジュアル コンテンツ生成
デジタル 生物学
フィジカル AI

Dynamo

アクセラレーテッド インフラ

Google Cloud

NVIDIA と Google による AI リーズニング モデルの進化

NVIDIA AI プラットフォーム上での推論に最適化された最先端のリーズニング AI モデル

deepseek

Gemma

Gemini

Llama 4

Llama Nemotron

Mistral AI

Qwen

独自のデータで AI エージェントをオンボード、トレーニング

ガードレールでセキュリティ、安全性、トピックの関連性を確保

NVIDIA NeMo

Curator → Customizer → Evaluator → Guardrails

Google Cloud

- Compute Engine
- GKE
- Cluster Toolkit
- Vertex AI Managed Fine-Tuning (NEW)

NVIDIA

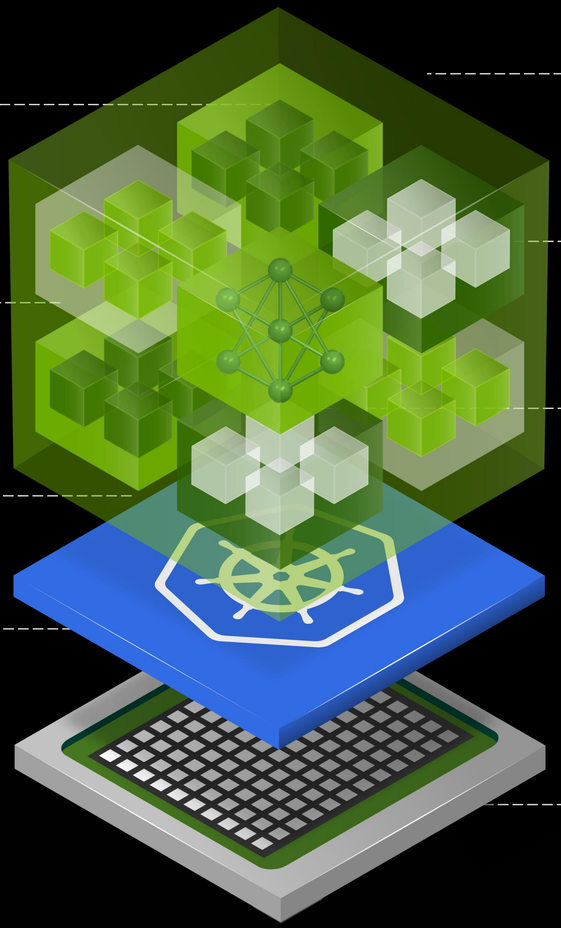
NVIDIA NIM 推論マイクロサービス

標準API
テキスト、音声、画像
ビデオ、3D、生物学

管理と監視
ヘルスチェック、ID、メトリクス
モニタリング、シークレット管理

クラウドネイティブ スタック
GPU Operator、Network Operator

Kubernetes



エンタープライズ サポート
安定した API、品質保証、継続的なアップデート、セキュリティ パッチ、サポート

NVIDIA Dynamo、TensorRT、TensorRT-LLM
cuBLAS、cuDNN、インフライト バッチ処理、メモリ最適化、FP8 量子化

最適化されたモデル
シングル GPU、マルチ GPU、マルチノード

CUDA

Google Cloud

NVIDIA

AI エージェント構築のための NVIDIA Blueprint

高速でスマートなエンタープライズグレード AI エージェントを構築するための簡単な出発点

NVIDIA AI Blueprint

- リファレンス アプリケーション
- サンプル データ
- リファレンス コード
- アーキテクチャ
- カスタマイズ ツール
- オーケストレーション ツール

- PDF to Podcast
- 顧客サービス向けの AI アシスタント
- コンテナ セキュリティのための脆弱性分析
- 創薬のための仮想スクリーニング
- ビデオ検索と要約

専門的な AI エージェント

リサーチ アシスタント エージェント

顧客サービス エージェント

ソフトウェア セキュリティ エージェント

バーチャル ラボ エージェント

ビデオ解析 エージェント

Google Cloud

NVIDIA

NVIDIA Dynamo

リーズニング モデルの大規模サービングのための AI 推論ソフトウェア

分散型および分離型のサービング

30 倍
AI ファクトリーの
スループットと
収益

DeepSeek モデル

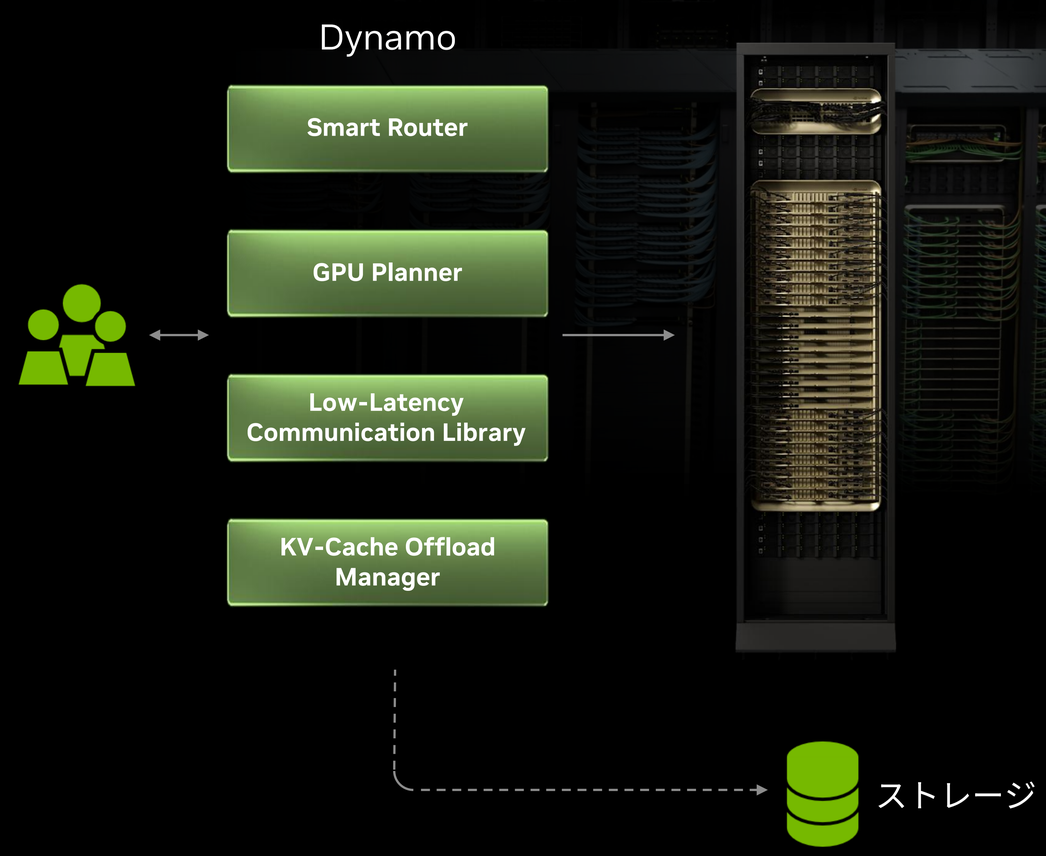
2 倍
AI ファクトリーの
スループットと
収益

Llama モデル

1,000 以上
分散推論のための
GPU スケールアップ

完全なオープン ソースとオープン バックエンド

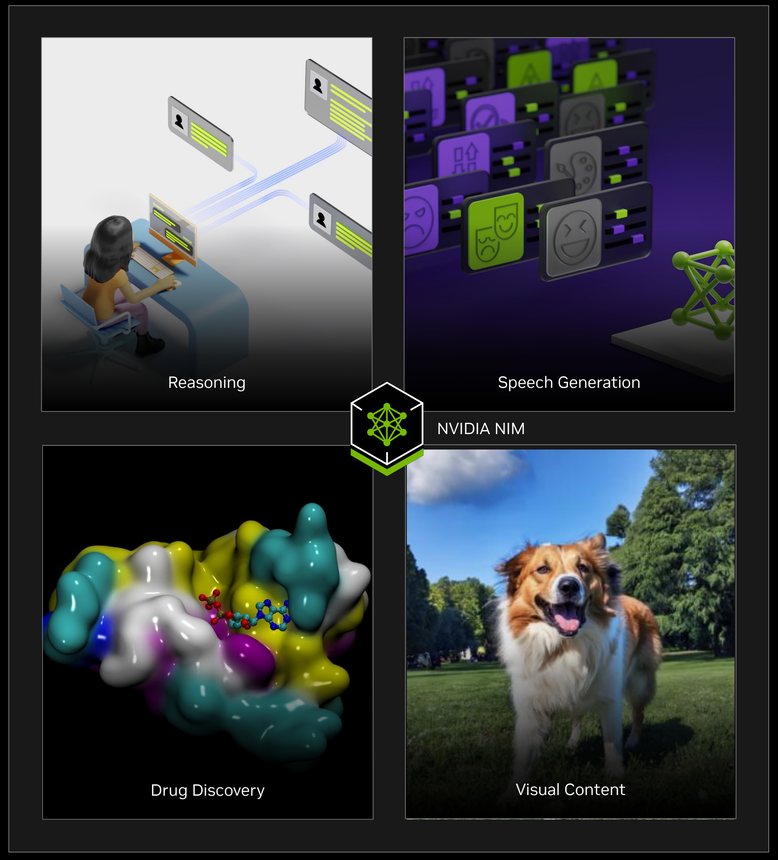
PyTorch　vLLM　SGL　TensorRT



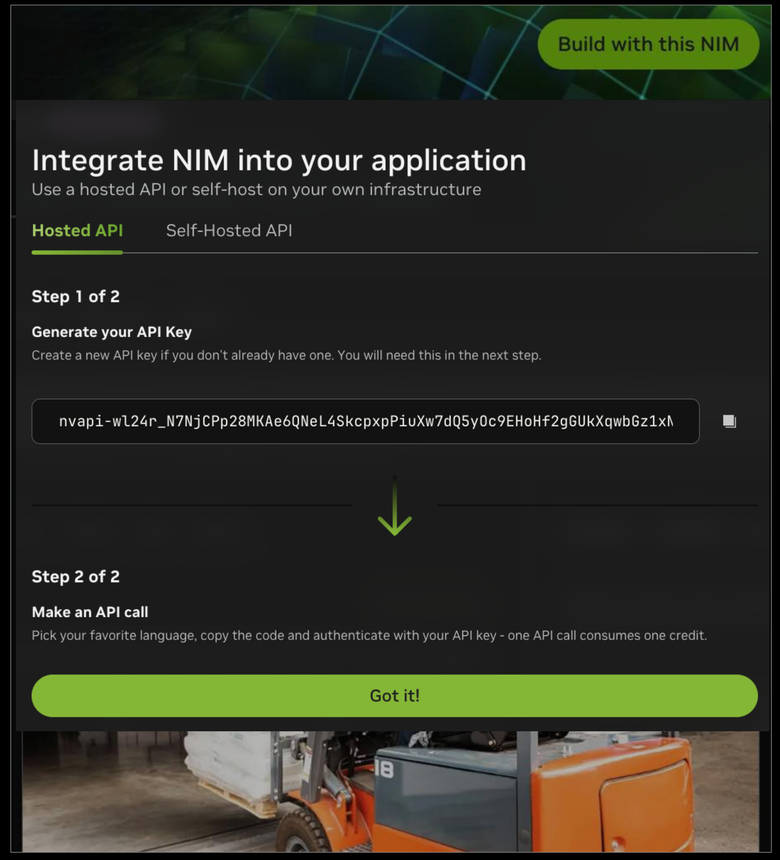
Projected performance subject to change

Google Cloud で NVIDIA NIM を体験、プロトタイプ作成、展開

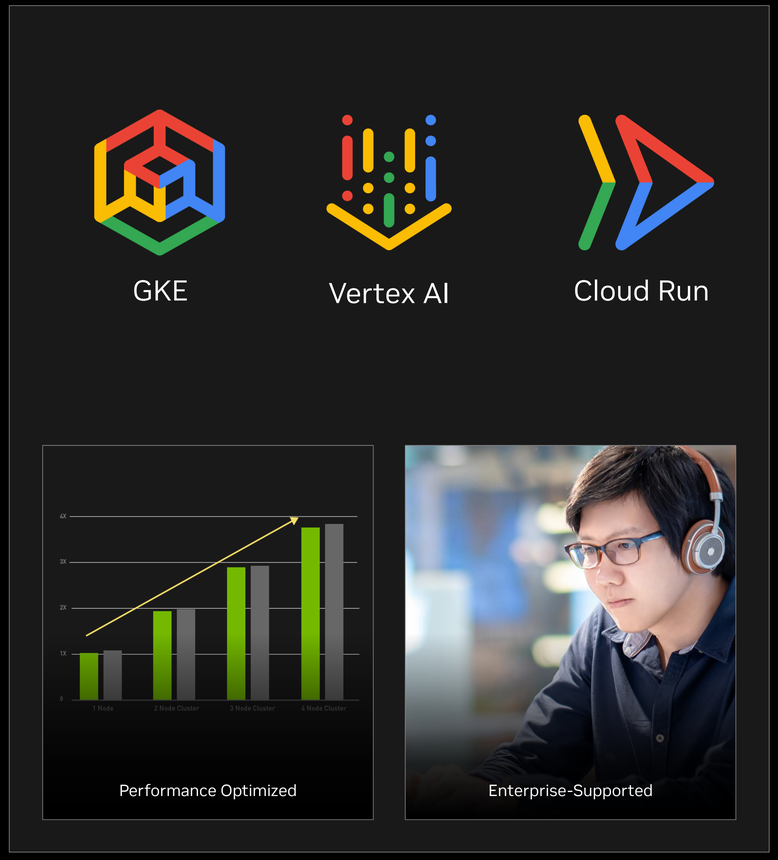
build.nvidia.com から始めましょう



ai.nvidia.com でモデルを体験



build.nvidia.com で API を使って
プロトタイプを作成



Google Cloud に NIM を展開

NVIDIA と Google Cloud で AI データ レイヤーを加速

Dataproc

バッチ データ処理 | ETL

RAPIDS

APACHE Spark™

5倍 高速、**80%** 低コスト

Dataflow

データ パイプラインの効率化 | ML 推論

TensorRT

beam

30倍 高速な
データ パイプライン

AlloyDB

ベクトル インデックス構築の最適化

cuVS

9倍 高速な
HNSW インデックス構築

エージェント型 AI ワークフローのあらゆるステップを加速

Dataproc	Dataflow	Vertex AI	Google Kubernetes Engine	Cloud Run
Dataproc、Dataflow 上の NVIDIA RAPIDS で **データ処理を加速**	GKE 上の NVIDIA NeMo フレームワークで **モデルの構築とカスタマイズ**	GKE または Vertex AI 上の NVIDIA Dynamo と TensorRT-LLM で **推論性能を最適化**	GKE、Vertex AI または Cloud Run 上の NVIDIA NIM で **推論を加速**	GKE 上の NVIDIA NIM と NeMo Retriever で **RAG を実装**
APACHE Spark RAPIDS	NVIDIA NEMO	TensorRT	NVIDIA NIM	NeMo Retriever

NVIDIA AI Enterprise

Google Cloud 上の NVIDIA 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

NVIDIA と Google Cloud:
顧客の成功を実現

Spotify

アプリ内コンテンツ発見機能の強化

Spotify Google Cloud

毎日、何十万もの音楽、ポッドキャストの
エピソード、オーディオブックを処理

数百万の Spotify ユーザーが好みのコンテンツを
簡単に見つけられるよう、60 秒のプレビューを生成

人気番組や日刊ニュースのような時間的制約のある
コンテンツには、ほぼ瞬時のプレビューが必要

NVIDIA と Google Cloud Dataflow のストリーミン
グ パイプラインとして展開

プレビュー生成にかかる時間が
2 時間から 2 分に短縮

30倍 高速化

Toyota

製造最適化の改善

TOYOTA

Google Cloud

トヨタは日本で年間 312 万台の車を製造

欠陥検出や設備監視などの手作業は
労働集約的で、生産を遅らせる

Google Cloud と NVIDIA に展開された
AI プラットフォームにより、
工場労働者が AI モデルを作成可能

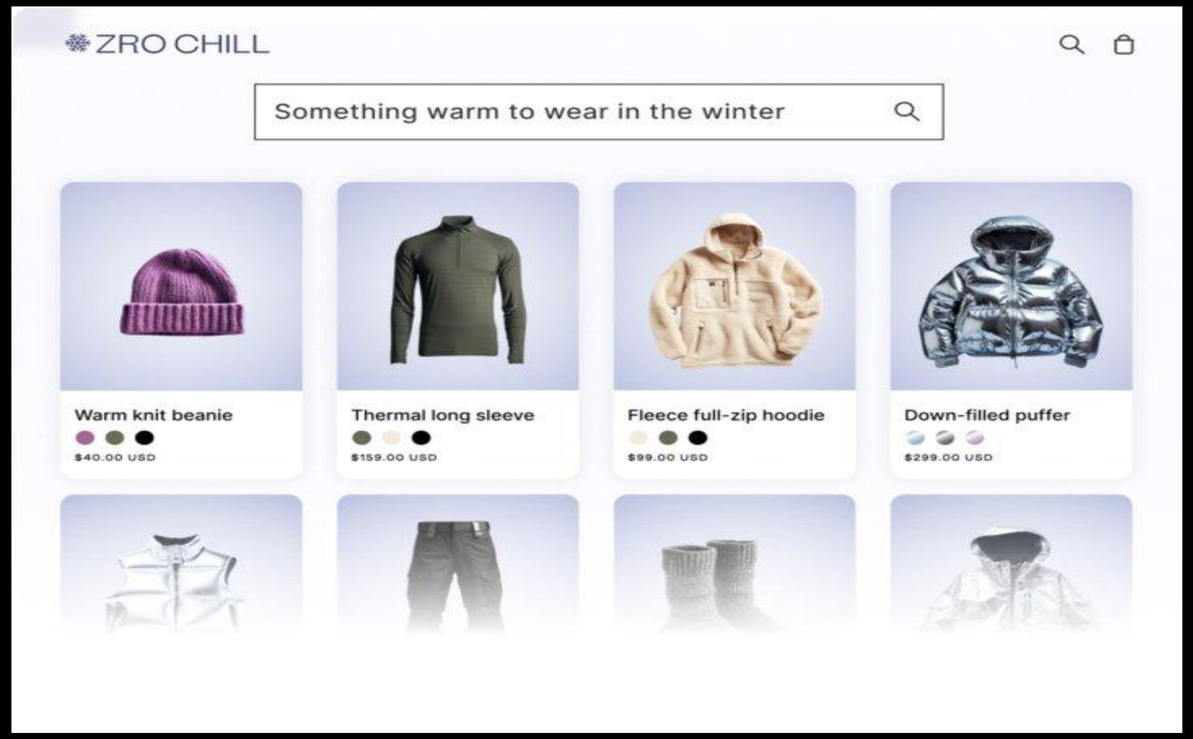
AI は欠陥検出、製造プロセスの監視、
設備故障の予測を支援

10,000+
時間以上を節約

10,000+
AI モデル

Shopify

リアルタイム インテリジェンスで検索を売上につなげる



shopify Google Cloud

175 カ国以上で、スタートアップからグローバルブランドまで数百万のビジネスにサービスを提供

キーワードベースの検索では、
関連性の低い商品を推奨

消費者の意図を理解し、関連性の高い商品を返すために、AIを活用したセマンティック検索を導入

商品リストと関連性の高い検索結果のリアルタイム更新により、加盟店の売上が向上

2 億 1,600 万
1 日あたりの埋め込み


NVIDIA