



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

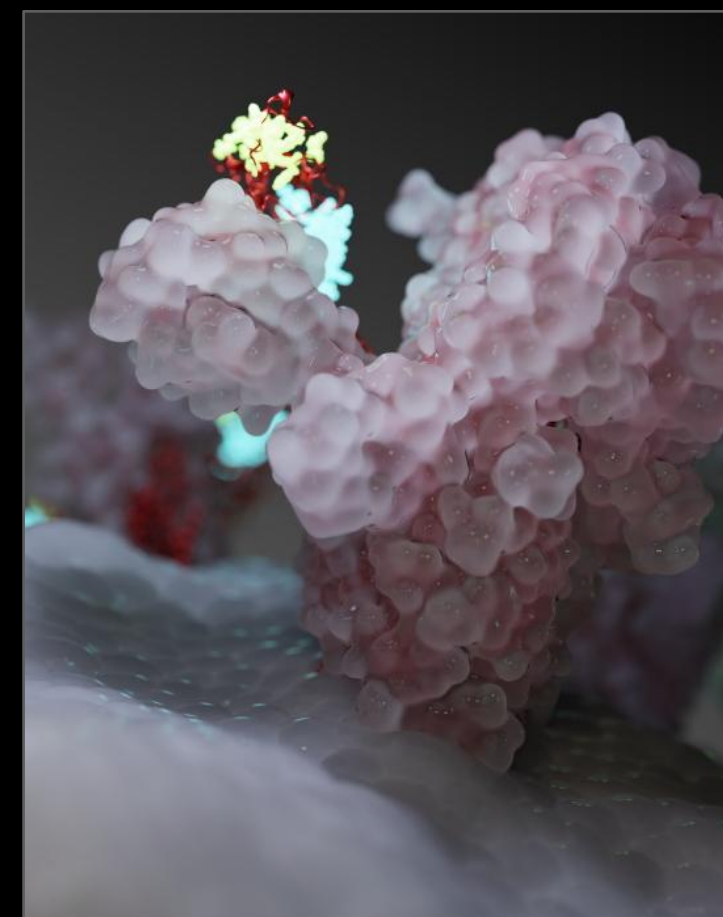
輸送



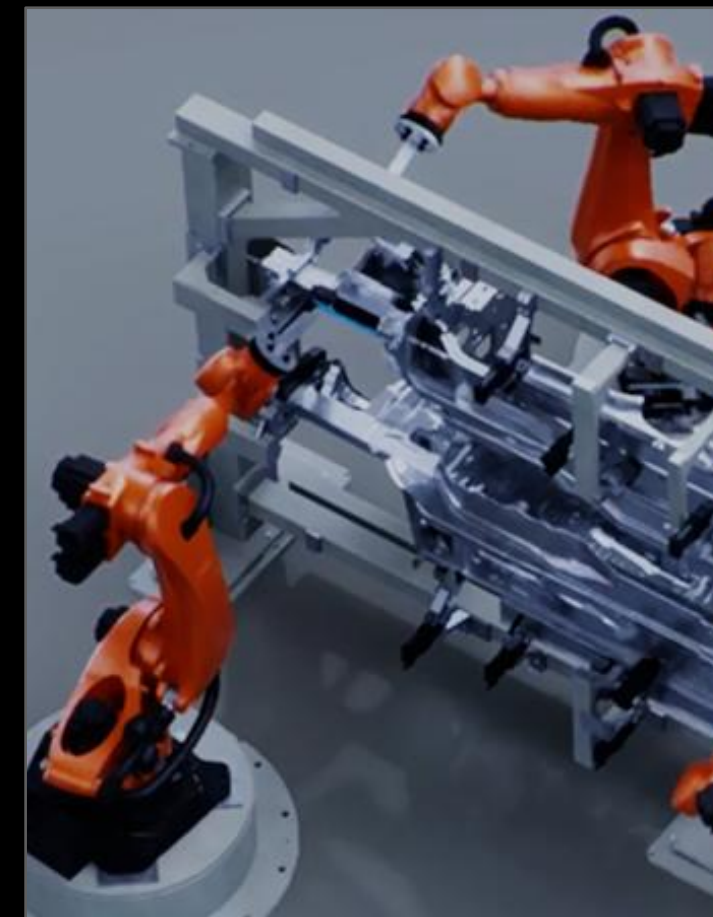
自然リソース



ヘルスケア



製造



小売り



通信



インターネット



金融



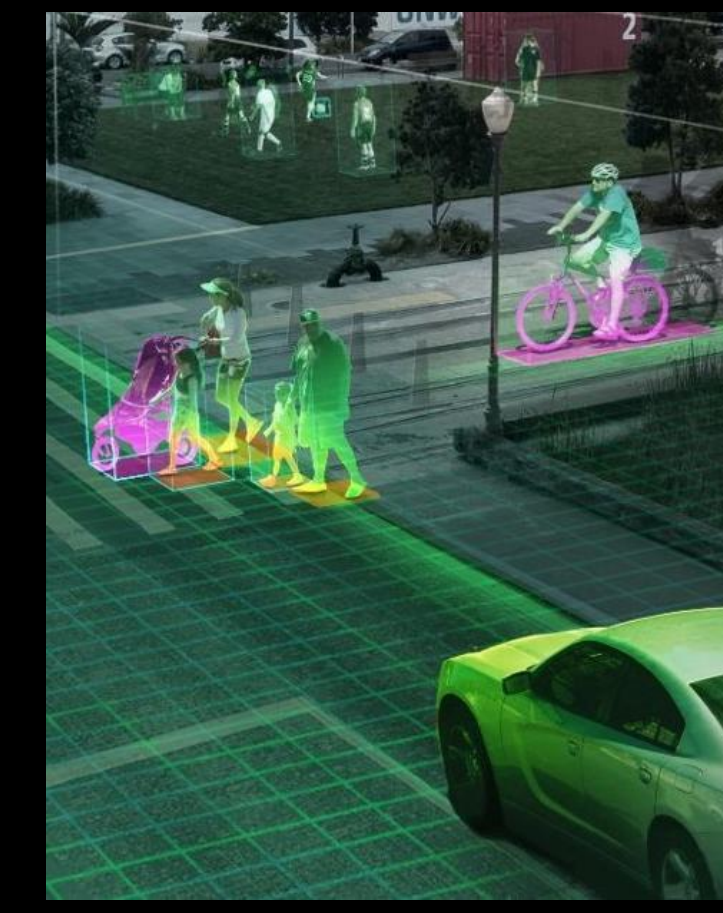
HPC



ロボティクス



スマートシティ



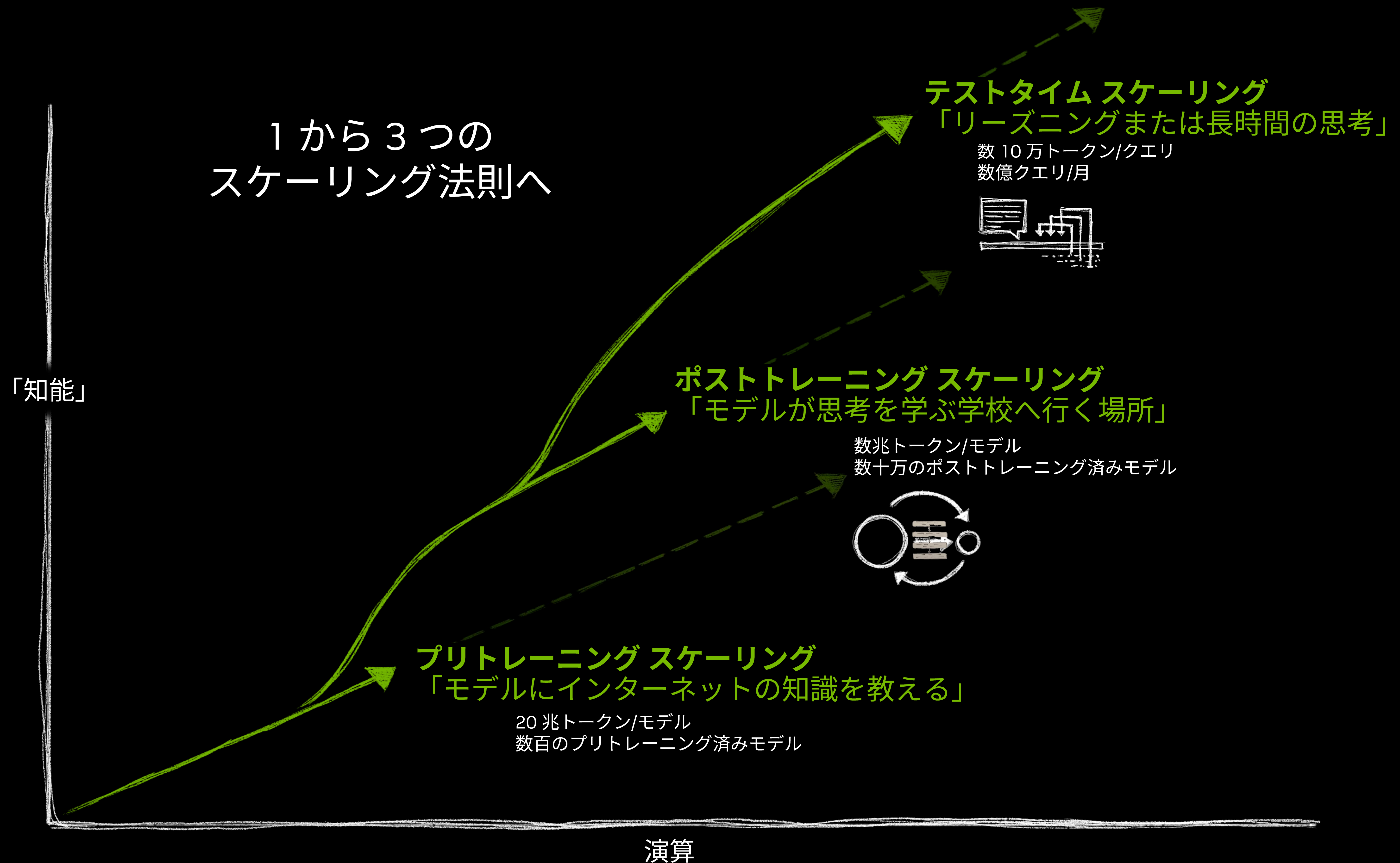
教育



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

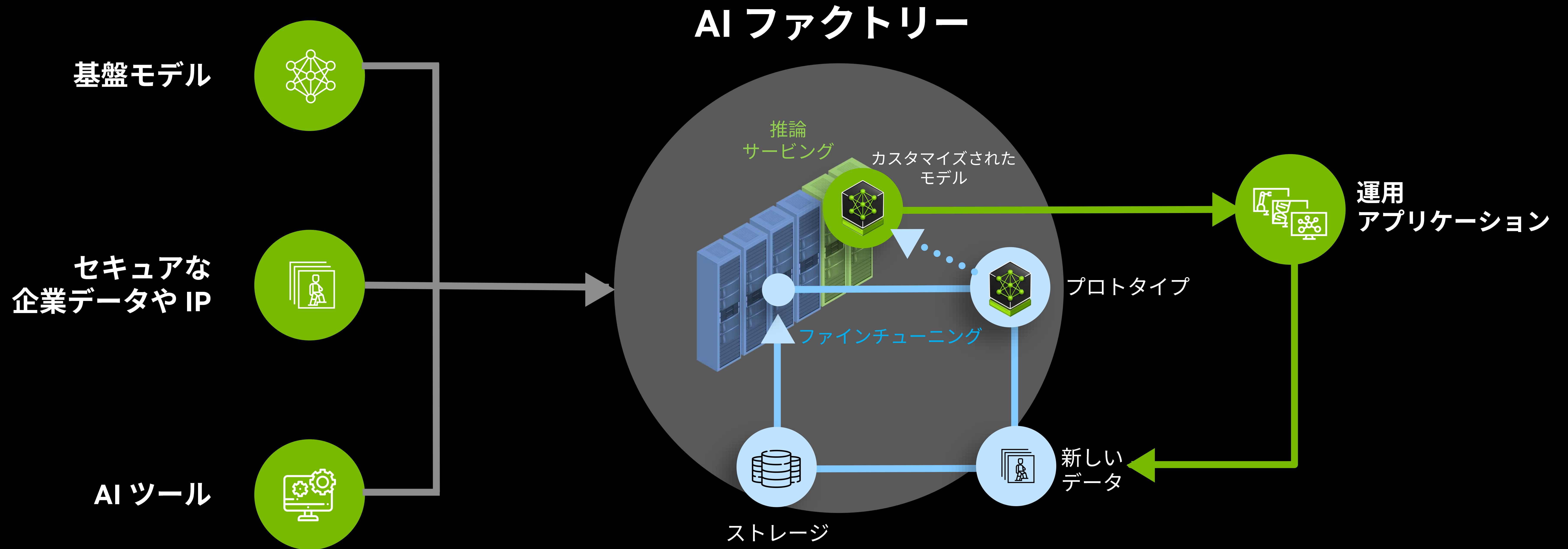
5500 万
IT 専門家

5000 万
世界中の
クリエイター

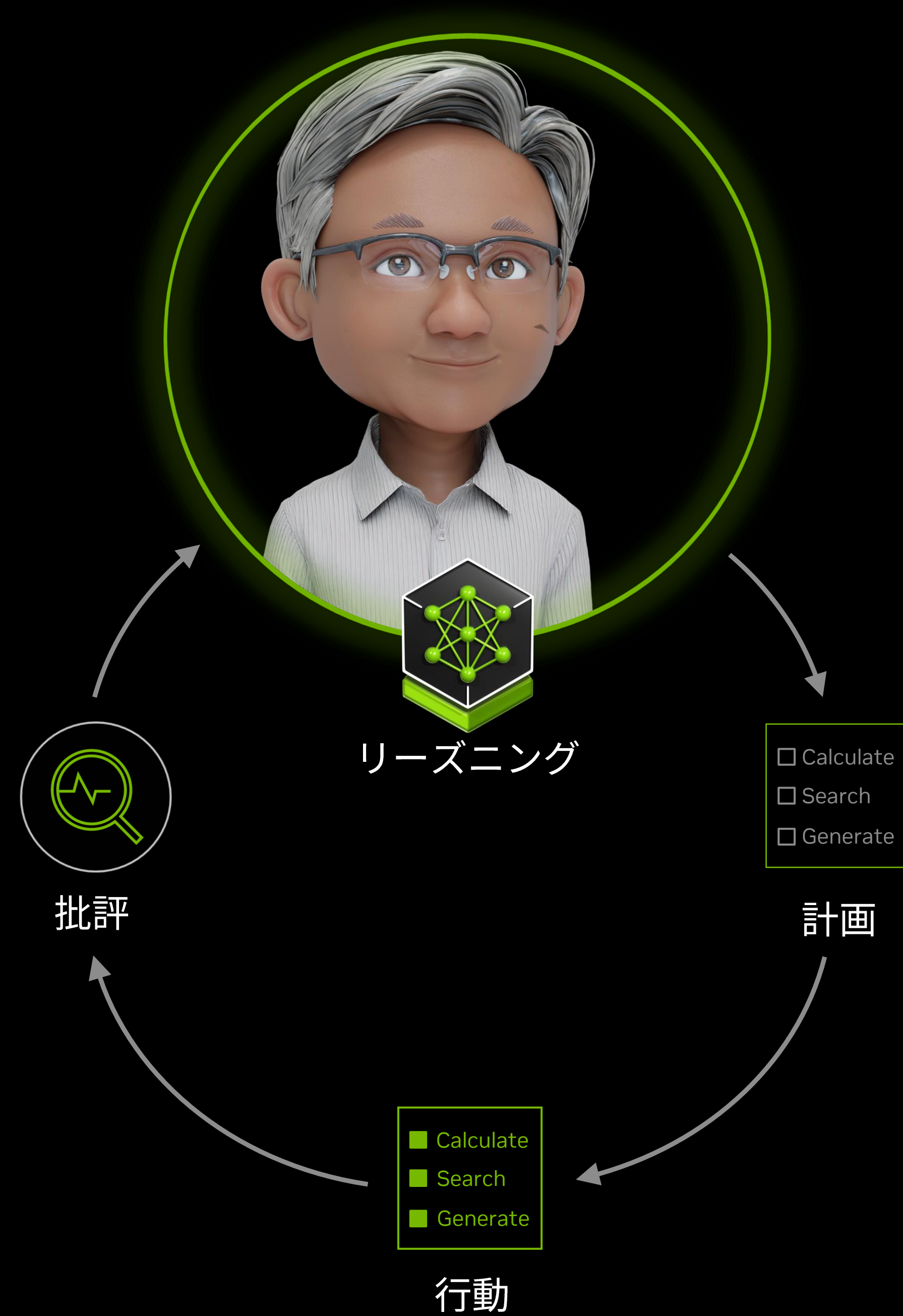


NVIDIA と Google Cloud が AI ファクトリーを実現

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

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

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.

Okay, so I need to figure out seating for a family dinner with 8 people. Let's list out the requirements first to make sure I don't miss anything.

First, the family has 8 members. Let me confirm who they are. The user mentions: in-laws (which I assume are the wife's parents), parents (user's own parents), sister, little brother, wife, and the user. That's 8 people. The user says "my family of 8" which includes the user, their spouse (wife), their parents (mom and dad), their in-laws (wife's parents, mom and dad), sister, and little brother. That's 8: user, wife, user's mom, user's dad, user's sister, and user's little brother.

Now, the constraints:

1. Sister needs to sit next to little brother to help him eat. So sister and little brother must be adjacent.
2. Wife doesn't want to sit next to user's mom. So wife and user's mom should not be adjacent.

Additional considerations: Typically, seating arrangements at a rectangular table have two ends and sides. But sometimes it's a round table. The problem doesn't specify, so maybe I should assume a round table because it's common for family dinners and allows for a circular arrangement. If it's a round table, seating is in a circle, so everyone has two neighbors. If it's rectangular, the ends have only one neighbor, but the problem might be easier with a round table. Let me proceed with a round table unless specified otherwise.

So, the goal is to seat 8 people around a round table with the given constraints.

First, list all the people:

1. User (let's call him U)
2. Wife (W)
3. Mom (M)
4. Dad (D)
5. Sister (S)
6. Little Brother (LB)
7. Mother-in-law (MI)
8. Father-in-law (FI)

Constraints:

- WM and WD (in-laws) should be separated by S, LB, and U.
- S must sit next to LB.
- W should not sit next to M.

Let me start by placing S and LB next to each other. Let's place them in two adjacent positions, say 1 and 2.

Now, the next strict constraint is that the sister and little brother need to be adjacent, which is already satisfied. So let's place S and LB next to each other.

Also, W doesn't want to sit next to M. So I need to ensure W and M are not adjacent.

Another thing to consider is that in-laws (WM and WD) should be separated by S, LB, and U. But the problem doesn't mention that the in-laws can't sit together, only that they should be separated by S, LB, and U. So it's possible that UM and UD would sit together, and WM and WD would sit together, but they would be separated by S, LB, and U. Let's assume that couples can be separated if needed, but perhaps it's better to keep them together unless conflicting with constraints. Let's proceed step by step.

First, place S and LB next to each other. Let's put S at position 1 and LB at position 2.

Now, we need to place U and W. Let's put U at position 3 and W at position 4. This satisfies the constraint that W is not next to M (position 5).

We need to keep WM and WD separated by S, LB, and U. So let's place M at position 5 and D at position 6. This keeps WM and WD separated by S, LB, and U.

Let me try to place the in-laws (MI and FI) in positions 7 and 8. This keeps them separated from the parents (M and D) by U and W.

Alternatively, since in-laws can sit together, let's try placing MI at position 7 and FI at position 8. This keeps them separated from the parents (M and D) by U and W.

Let me try to create two groups: one with UM and UD, and another with WM and WD, separated by other people.

But also, W shouldn't be next to UM. So W needs to be away from UM.

Let me attempt a possible arrangement:

Starting with S at 1, LB at 2. Then moving clockwise, let's try to place U and W in positions that can help separate the conflicting groups.

Let's try:

1: S 2: LB 3: ? 4: ? 5: ? 6: ? 7: ? 8: ?

We need to fill positions 3-8 with U, W, UM, UD, WM, WD.

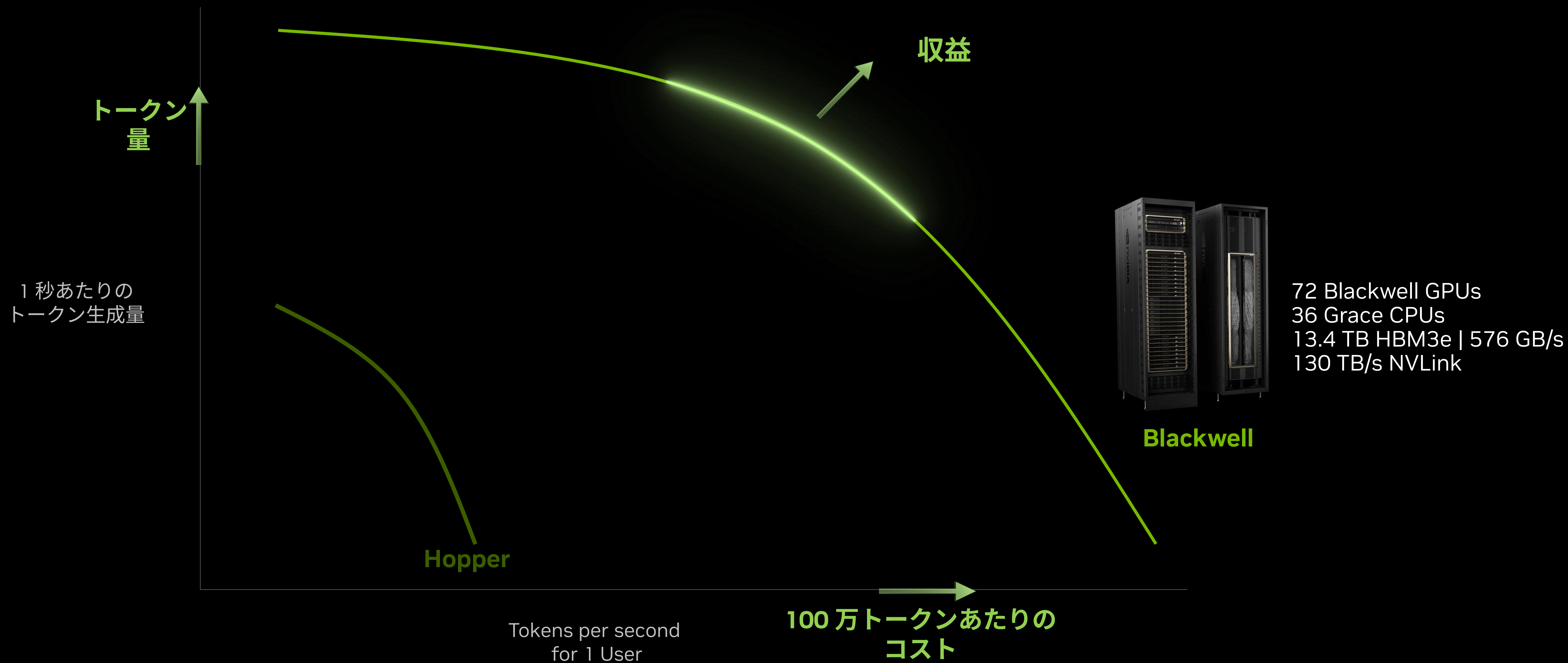
Let me consider placing the user (U) next to LB (position 2) to have a family member next to the little brother. So position 3: U. Then position 8 (next to S) could be someone else.

But then U at 3 would be next to LB at 2 and position 4. Let's see.

Alternatively, maybe place U and W opposite S and LB. Let's see, in an 8-seat round table, opposite of S (position 1) would be position 5. So position 5 could be U or W. Let's try placing U at 5 and W at 6, but need to check constraints.

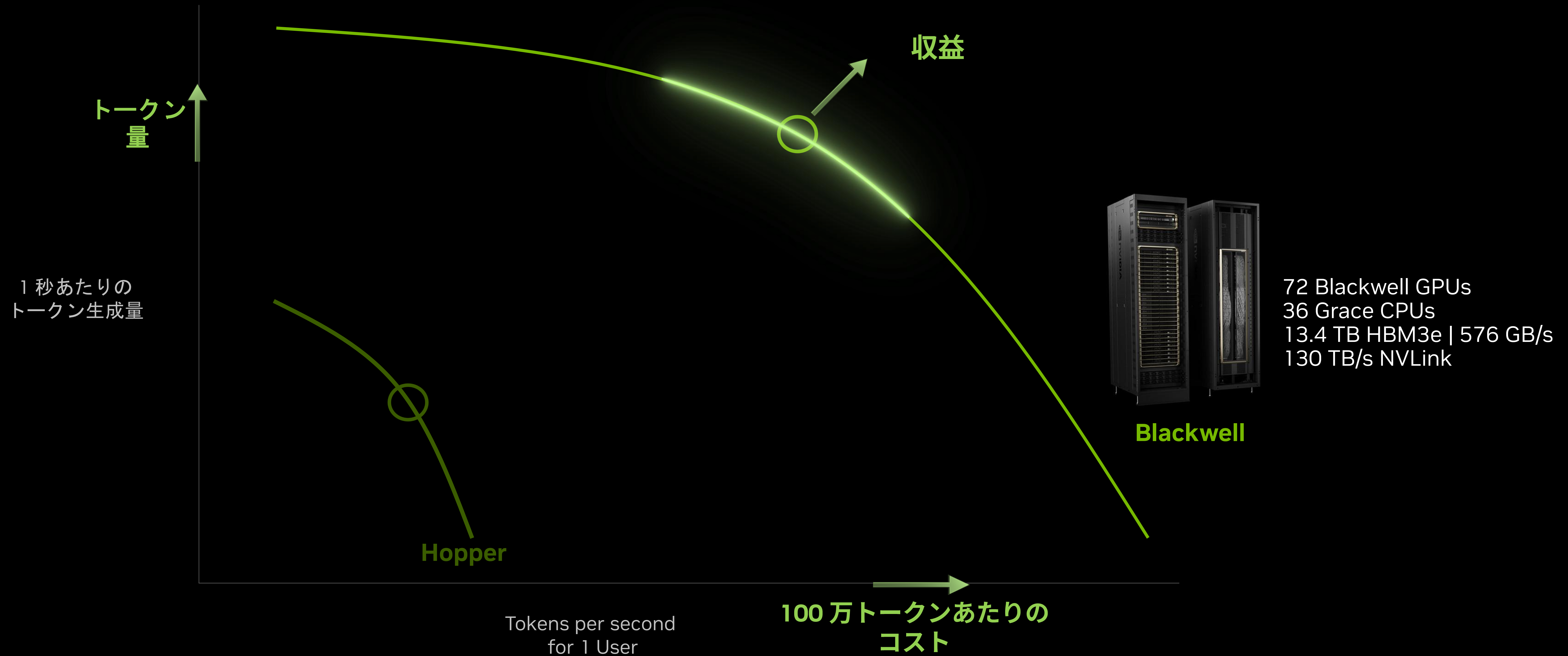
Wait, maybe a better approach is to divide the table into sections. Let's say S and LB are at 1 and 2. Then, to separate the in-laws from parents, we can place the parents (UM and UD) on one side and in-laws (WM and WD) on the other side, with U and W in between.

AI ファクトリーの生産性向上

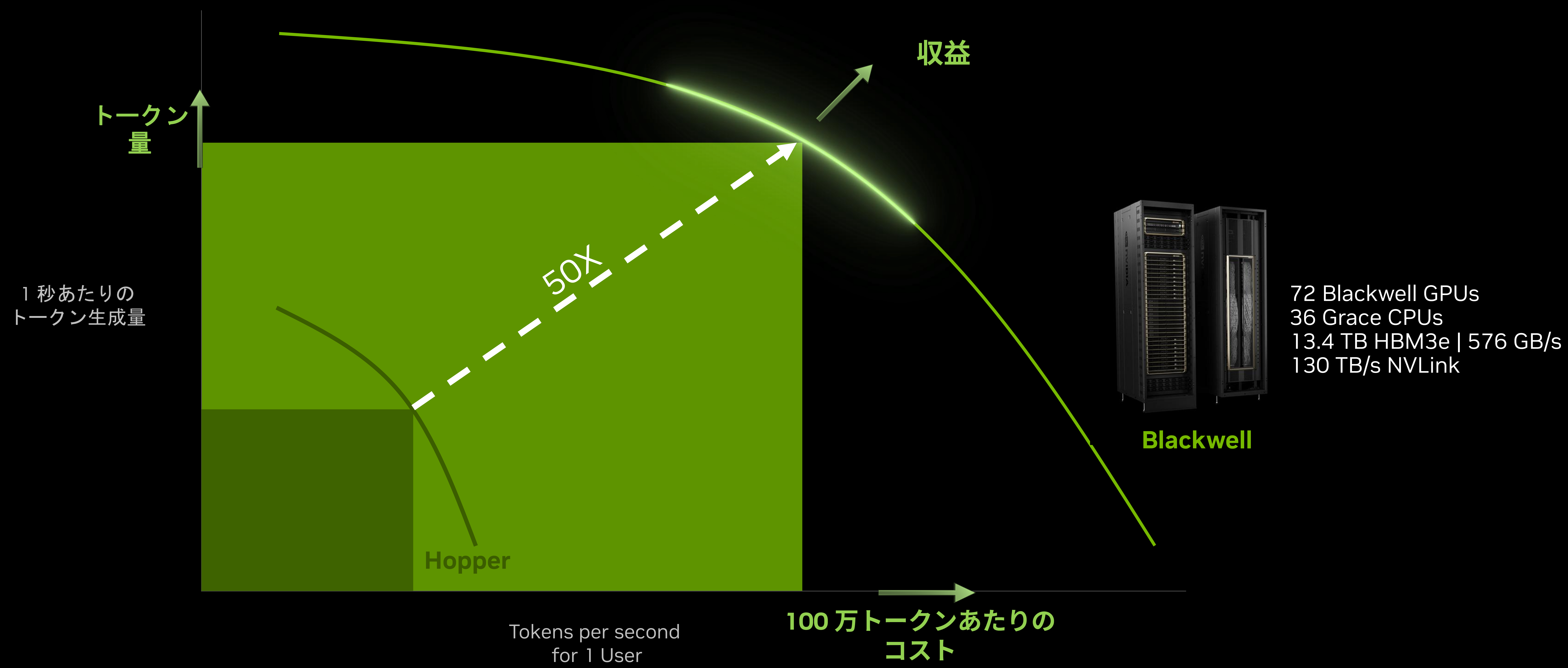


AI ファクトリーの生産性向上

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



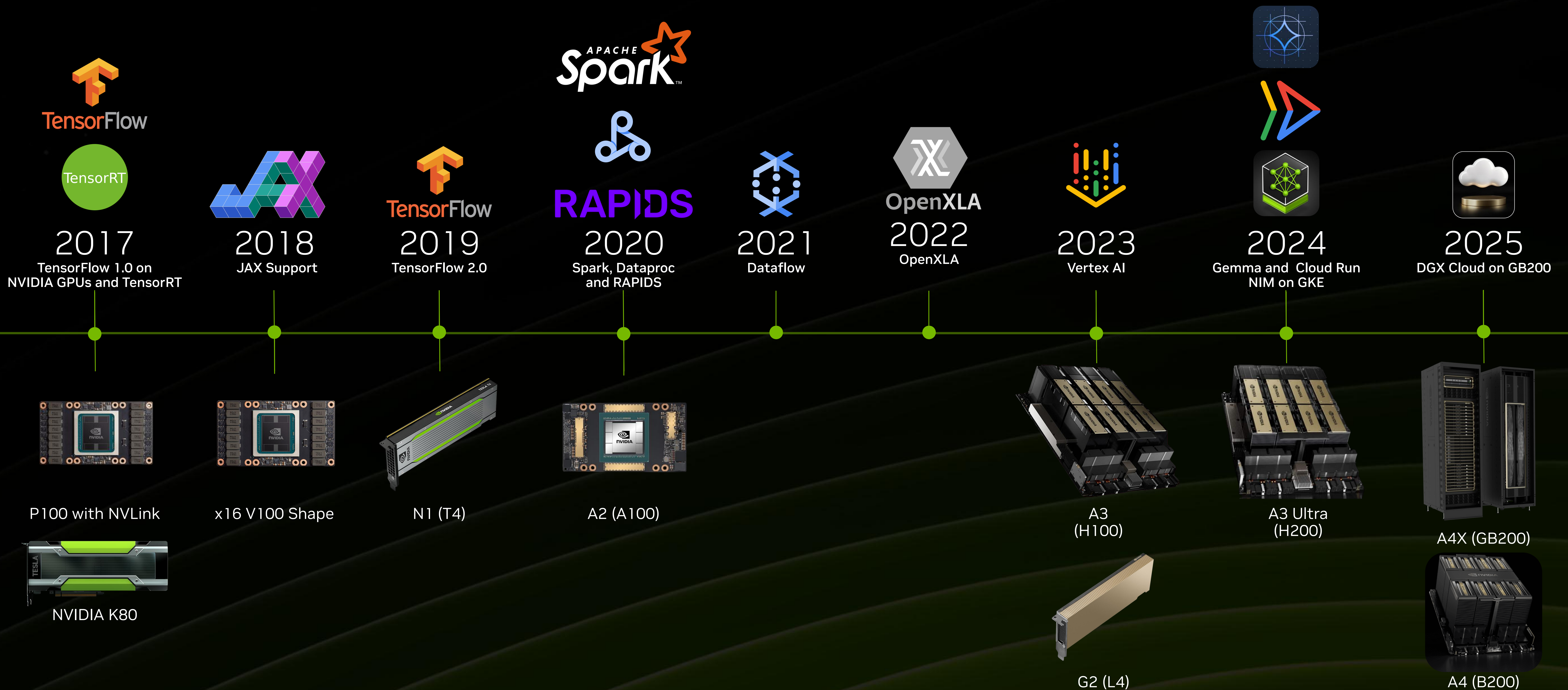
AI ファクトリーの出力が収益を牽引



**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



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



Vertex AI



Dynamic Workload
Scheduler

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

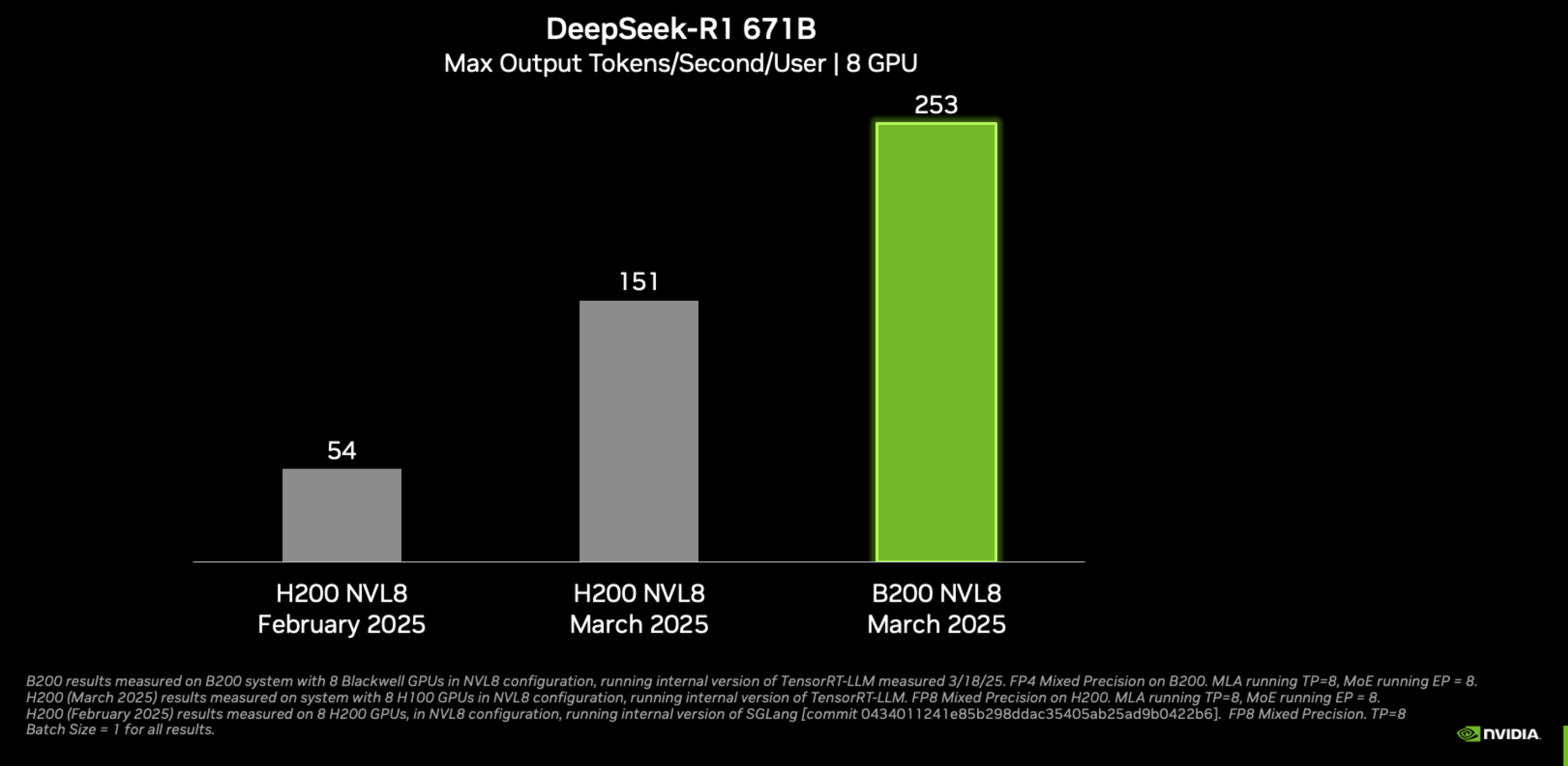
大規模 AI ファクトリー

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

World's Fastest DeepSeek-R1 671B Inference
253 output tokens per second per user on eight Blackwell GPUs



精度を維持しつつ大幅な高速化

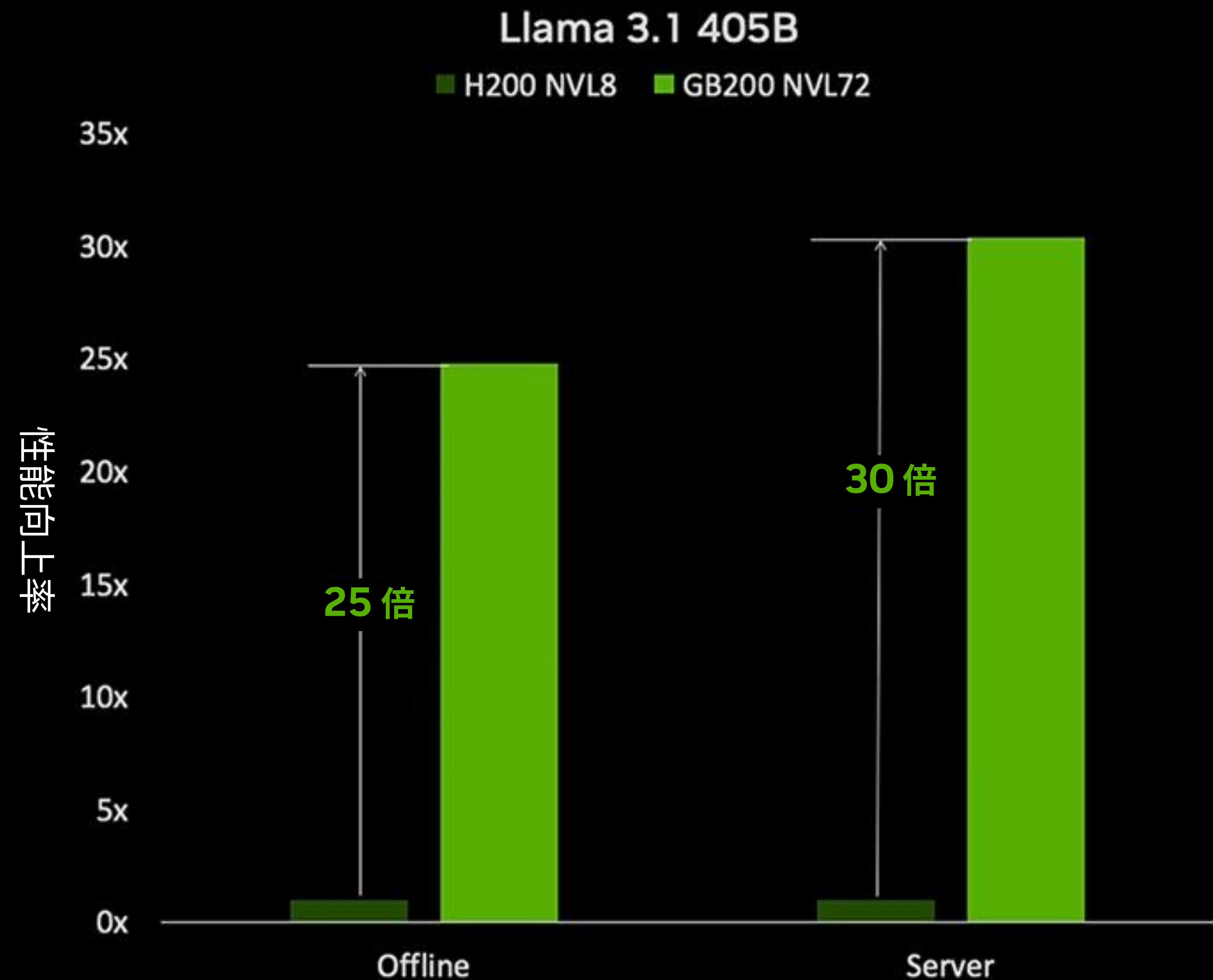
| | MMLU | GSM8K | AIME 2024 | GPQA Diamond | Math-500 |
|-----------------|-------|-------|-----------|--------------|----------|
| DeepSeek R1-FP8 | 90.8% | 96.3% | 80% | 69.7% | 95.4% |
| DeepSeek R1-FP4 | 90.7% | 96.1% | 80% | 69.2% | 94.2% |

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



MLPerf Inference v5.0, Closed, Data Center. Results retrieved from www.mlcommons.org on April 2, 2025.
Results retrieved from the following entries: 5.0-5.0-0058, 5.0-0060. The MLPerf name and logo are registered and unregistered trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information

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 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



Text



Image



Video



Audio



Code



Reasoning



Math

Gemini



Google Distributed Cloud



NVIDIA Blackwell

NVIDIA HGX B200 | NVIDIA DGX B200

2025 年第 3 四半期にパブリック プレビュー

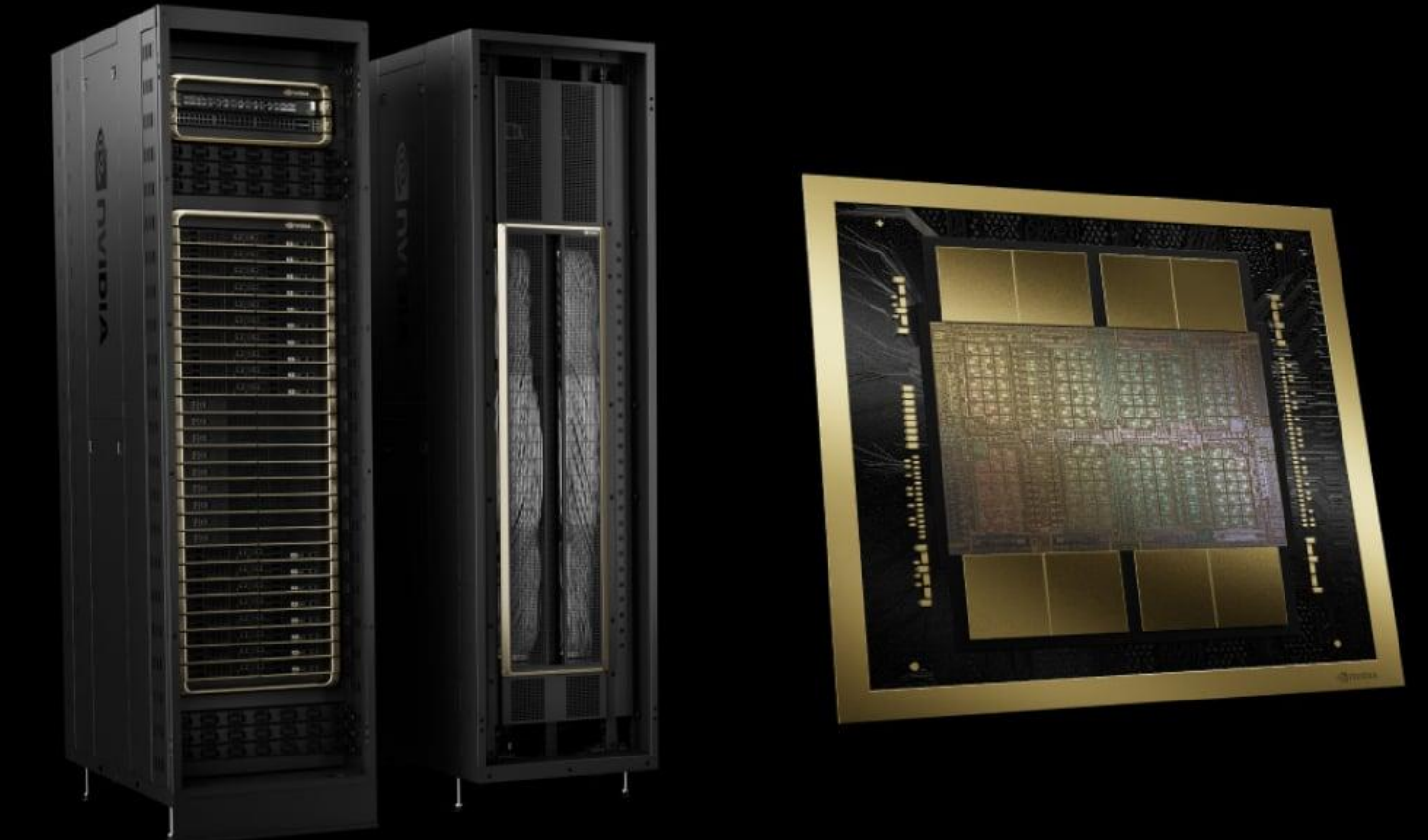
Google Cloud 上の NVIDIA アクセラレーテッド コンピューティング プラットフォーム



A4 VM
NVIDIA B200



A4X VM
NVIDIA GB200 NVL72



GB300
Coming Soon

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

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

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 High VM
NVIDIA H100

A3 Mega VM
NVIDIA H100

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

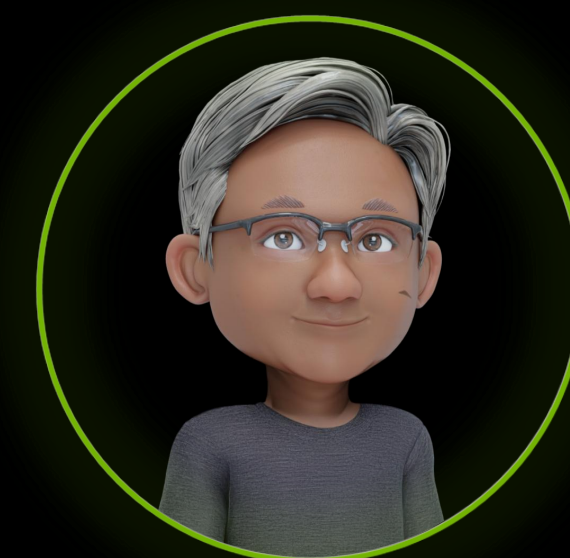
NVIDIA AI Blueprint



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



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

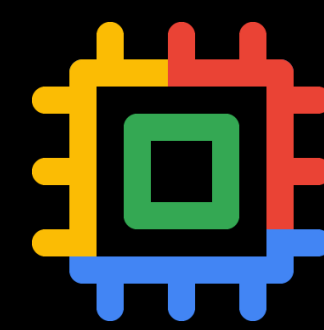
Curator

Customizer

Evaluator

Guardrails

Google Cloud



Compute Engine



GKE

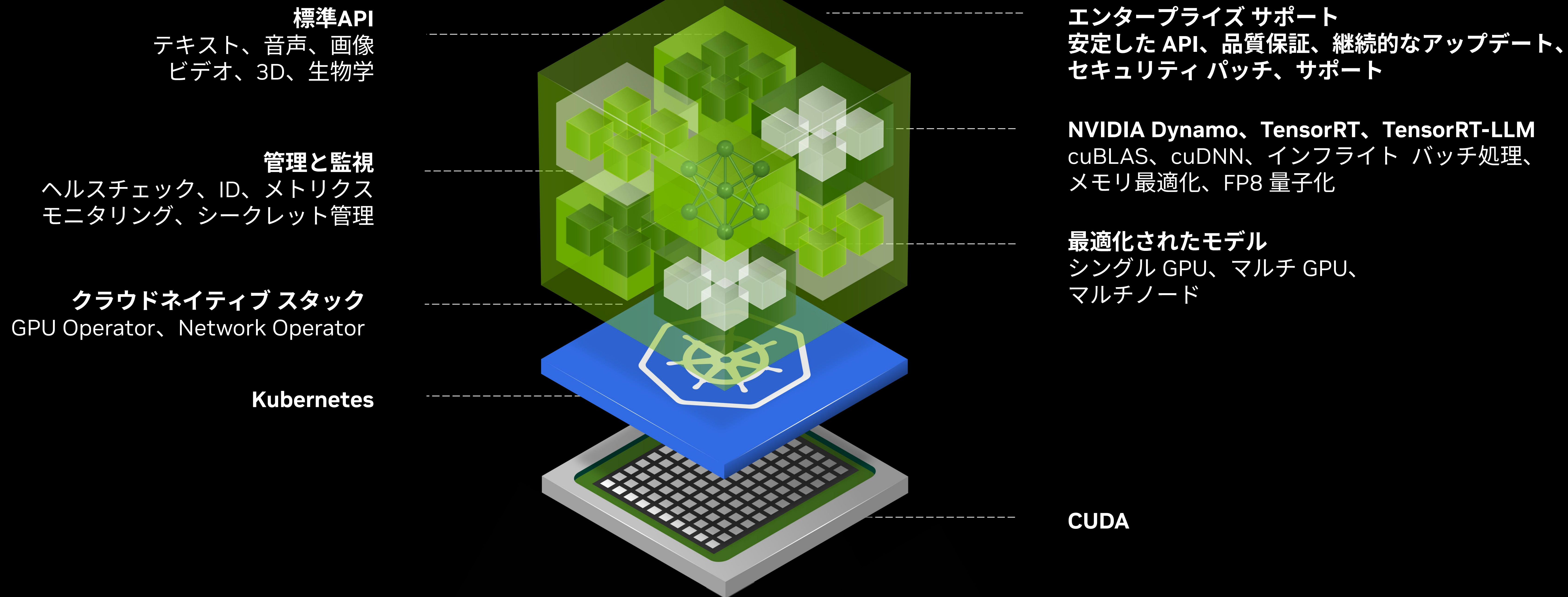
Cluster Toolkit



Vertex AI
Managed Fine-Tuning

NEW

NVIDIA NIM 推論マイクロサービス

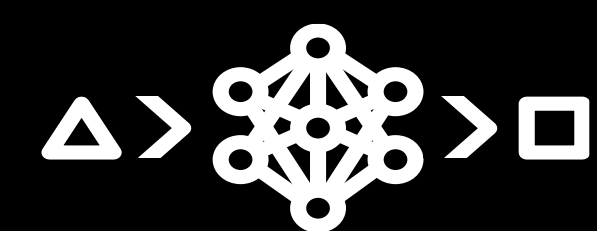


Google Cloud

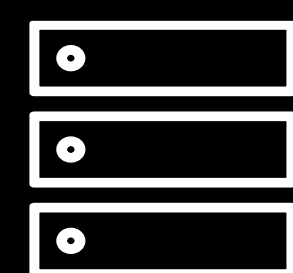
AI エージェント構築のための NVIDIA Blueprint

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

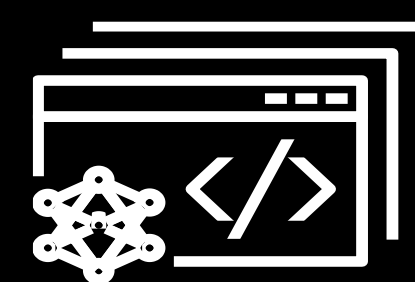
NVIDIA AI
Blueprint



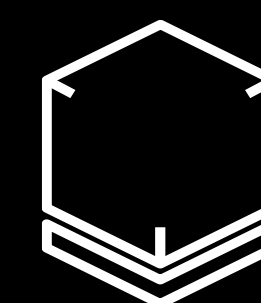
リファレンス
アプリケーション



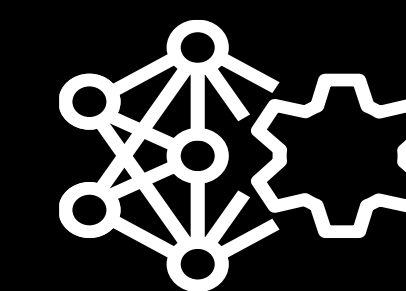
サンプル データ



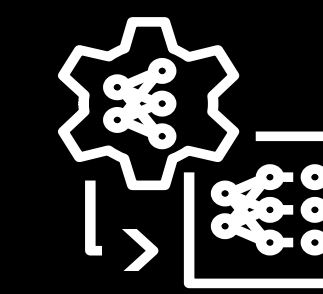
リファレンス コード



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 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

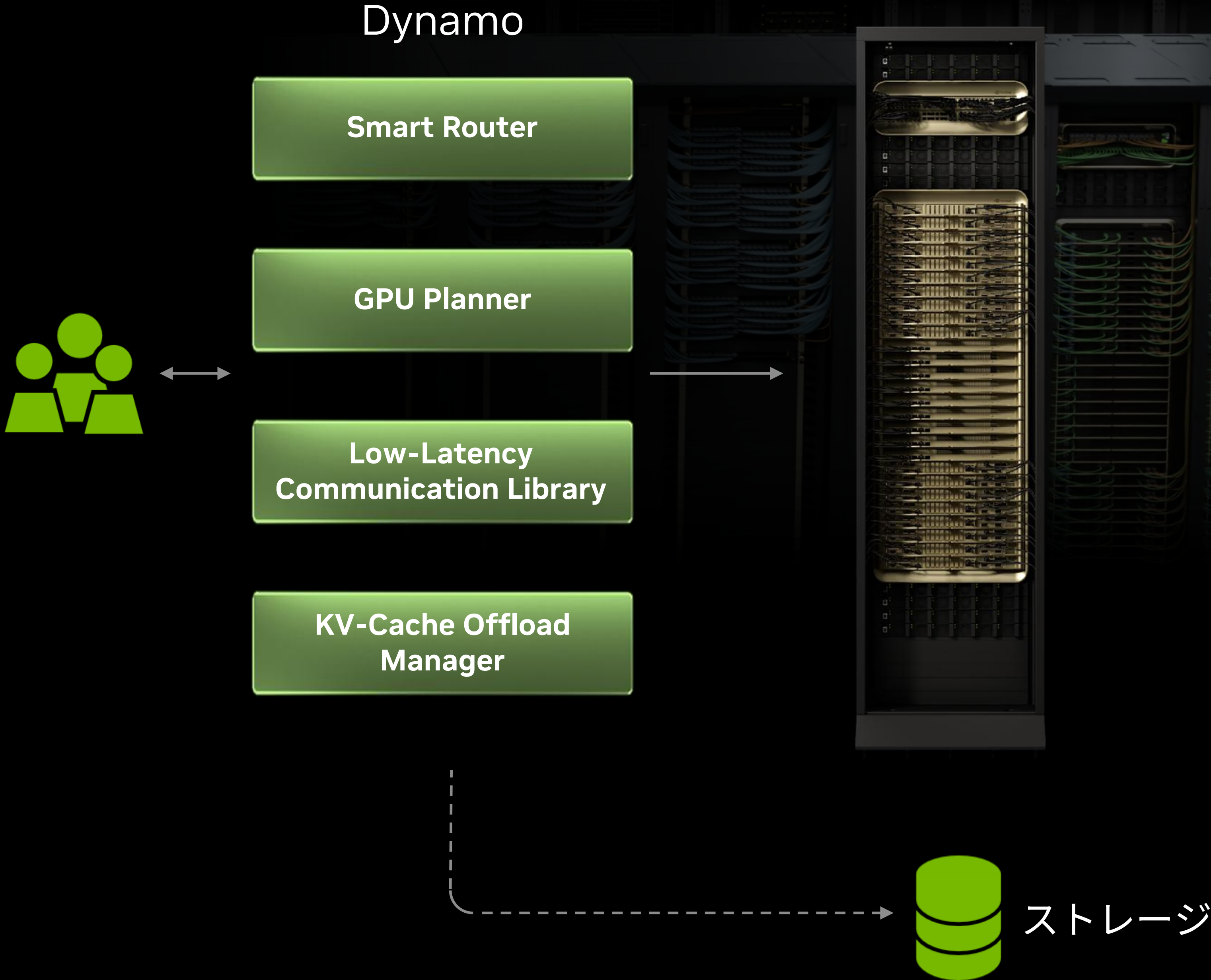
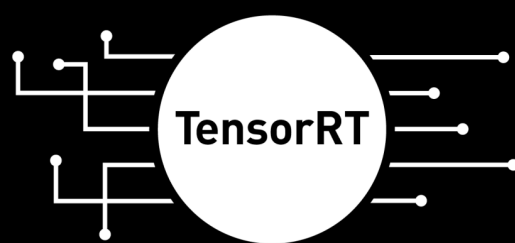
分散推論のための
GPU スケールアップ

完全なオープンソースとオープンバックエンド

PyTorch

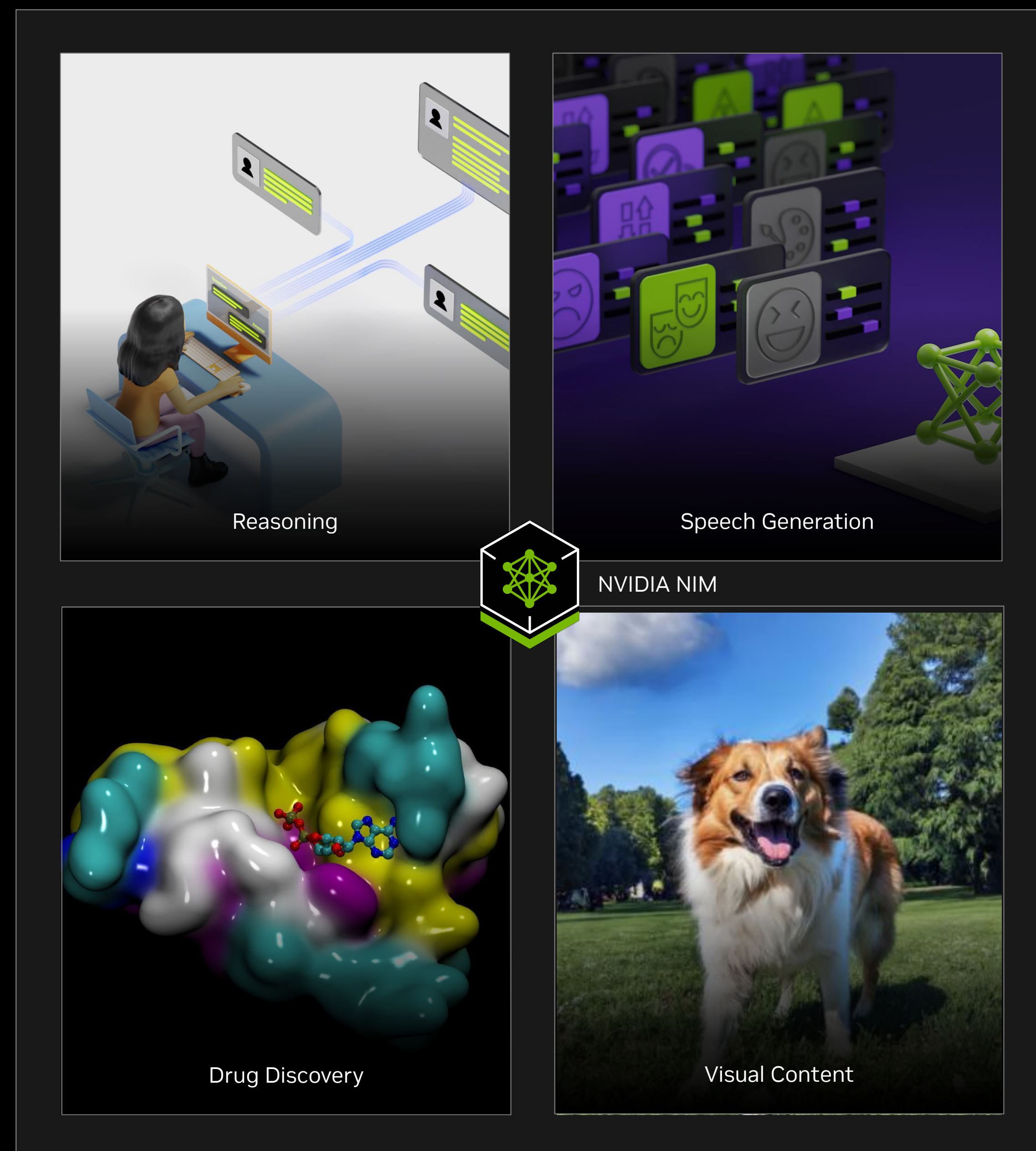
vLLM

SGL

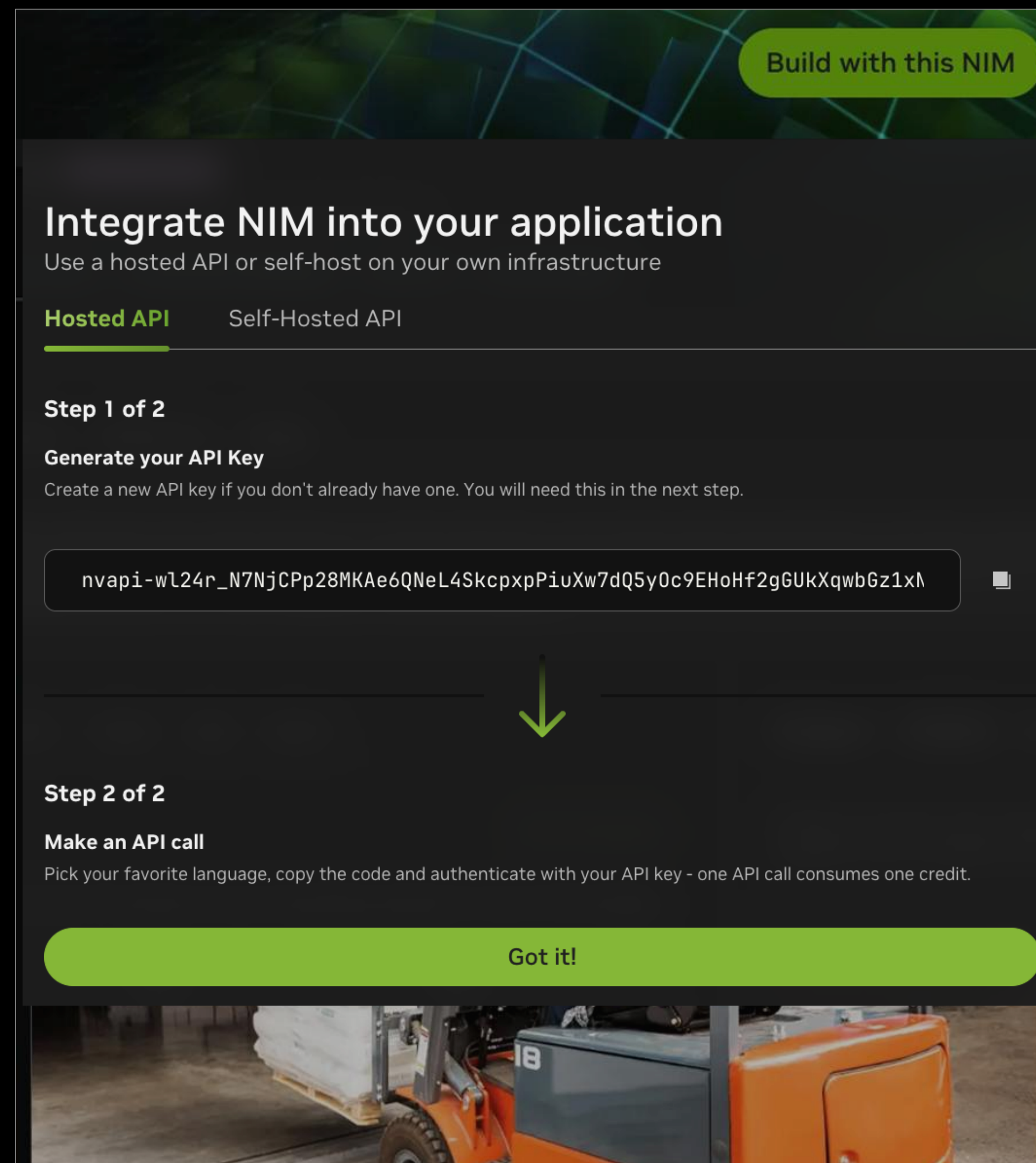


Google Cloud で NVIDIA NIM を体験、プロトタイプ作成、展開

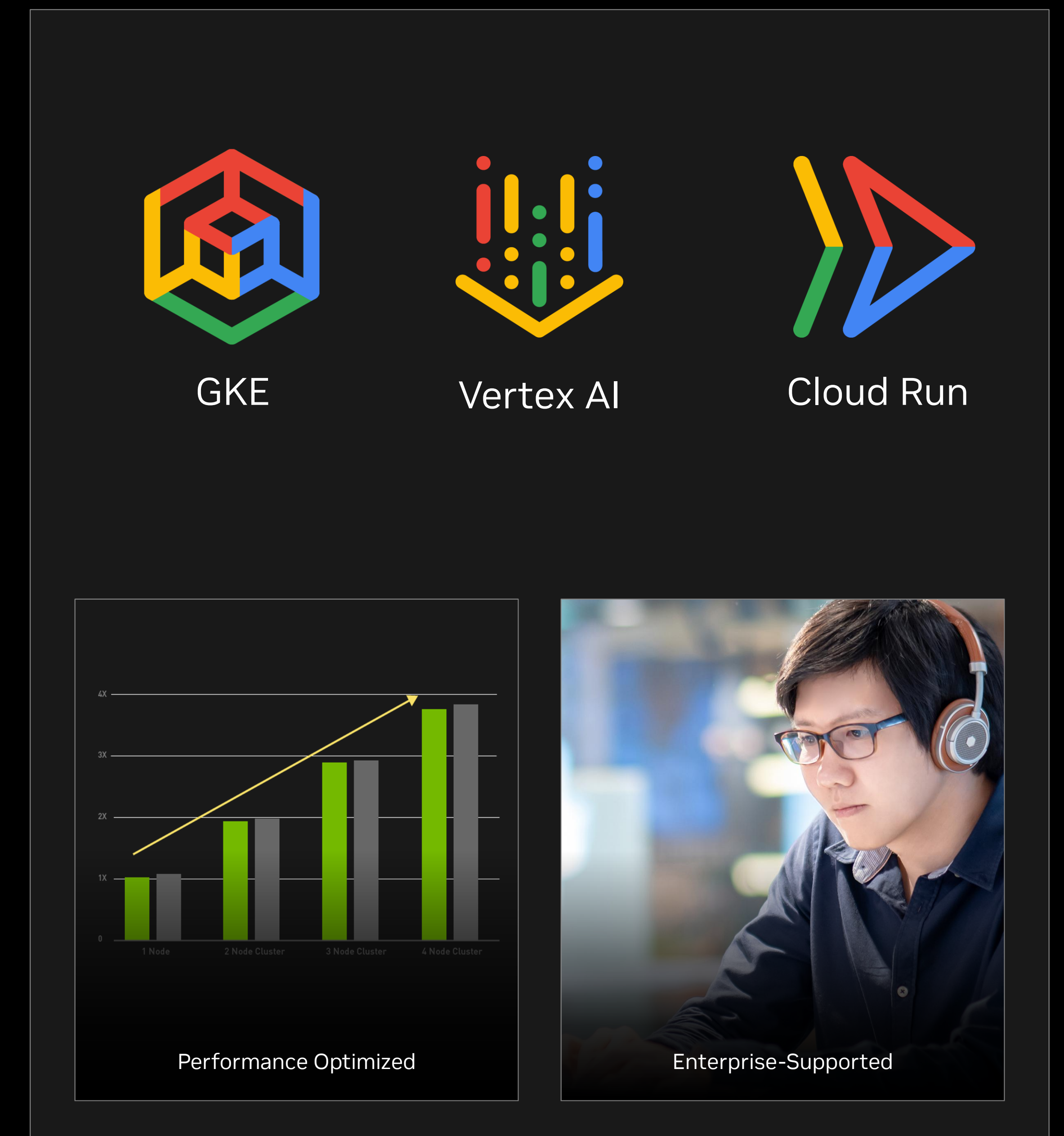
build.nvidia.com から始めましょう



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

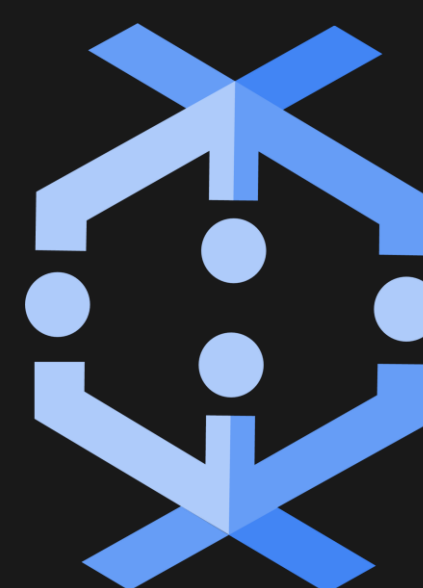


バッチ データ処理 | ETL

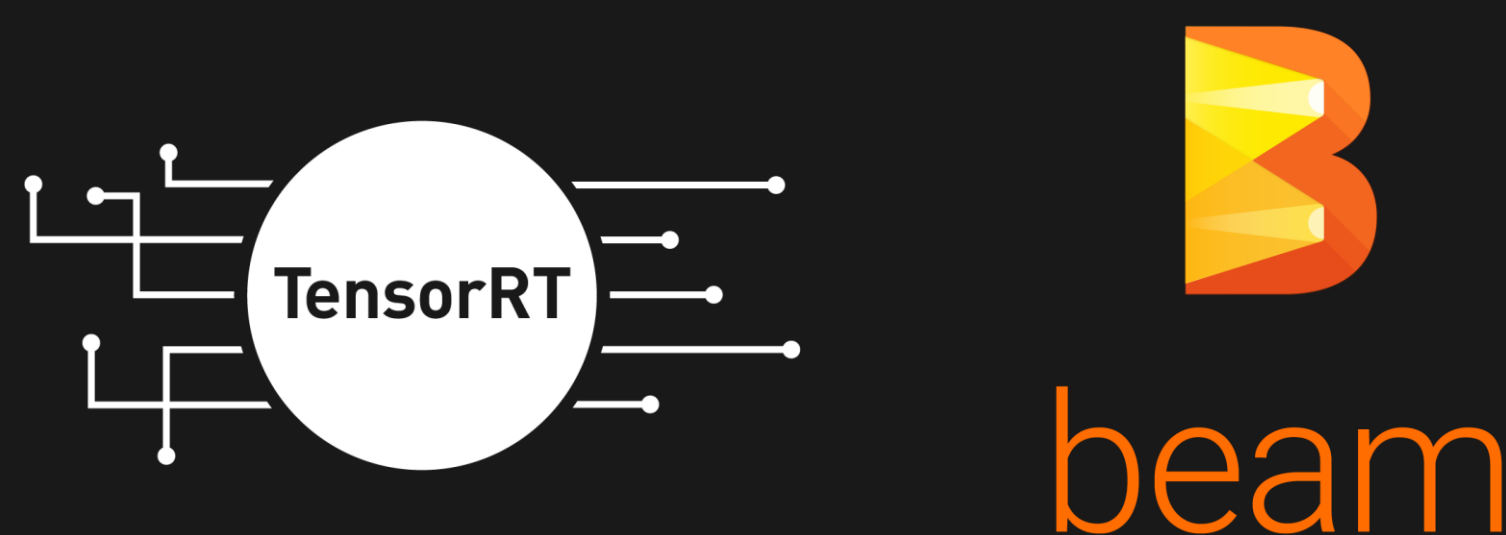


5倍 高速、**80%** 低コスト

Dataflow



データ パイプラインの効率化 | ML 推論



30倍 高速な
データ パイプライン

AlloyDB

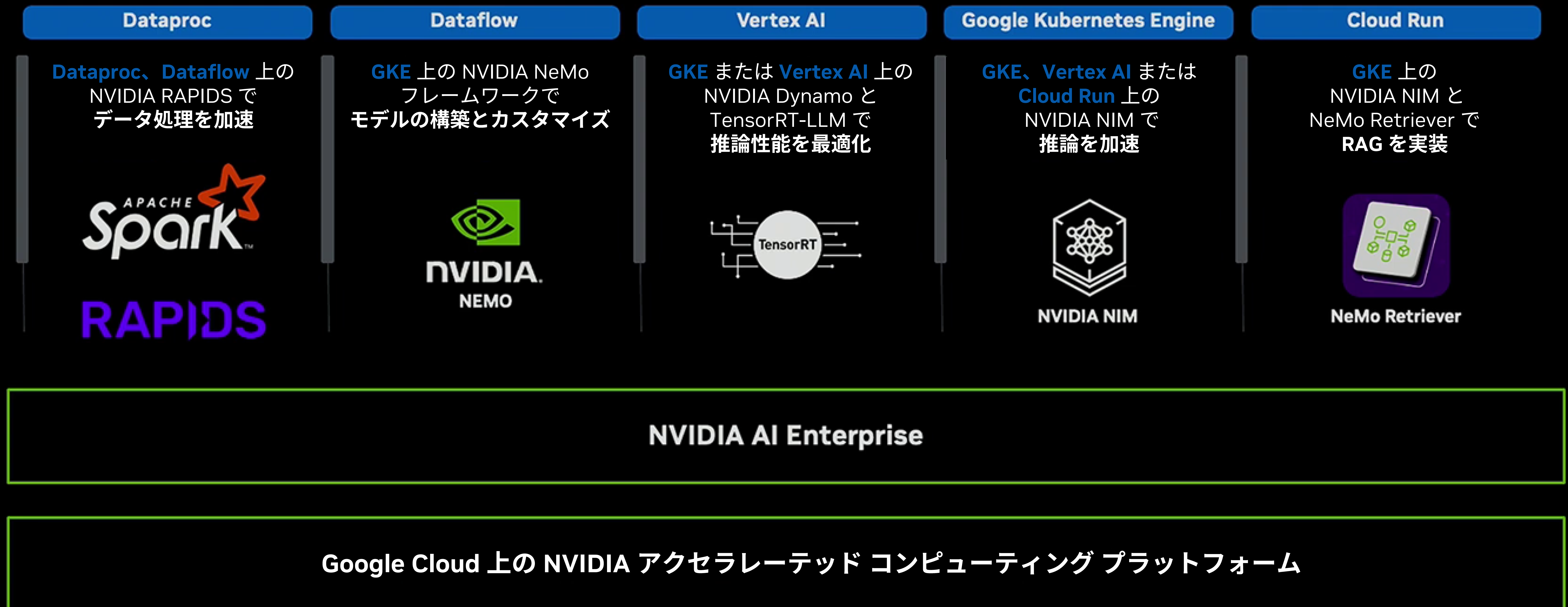


ベクトル インデックス構築の最適化



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



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



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

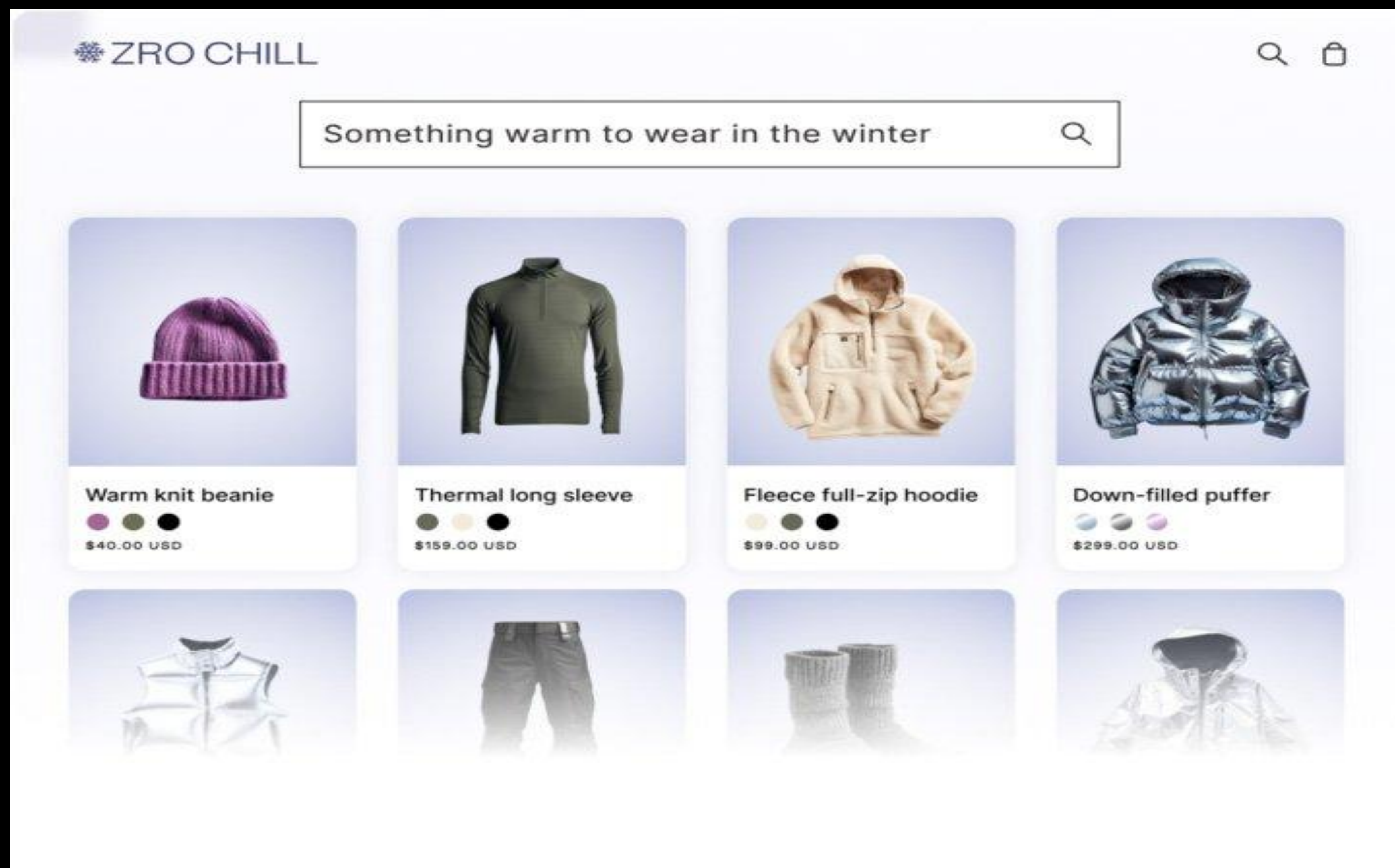
AI は欠陥検出、製造プロセスの監視、
設備故障の予測を支援

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



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

