第390回 NRIメディアフォーラム
「ITロードマップ 2025年版」

# AI向け次世代コンピューティング

チーフリサーチャー　藤吉　栄二

株式会社　野村総合研究所
DX基盤事業本部
IT基盤技術戦略室

2025年3月25日

NRI

Envision the value,
Empower the change

# 目次

# AIにおけるコンピューティング

AIにおけるコンピューティング

# 生成AIの利活用を支える、「学習」と「推論」

- 学習と推論では処理内容が異なる
- 質の高いAI基盤モデルを生み出すことに世界的な関心が集まるなか、**学習環境の選定**が話題に

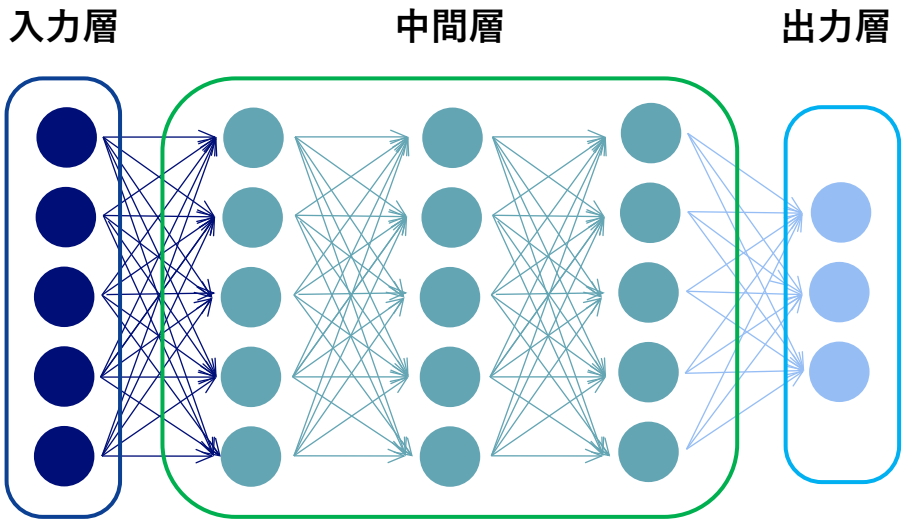
| | 学習（Training） | 推論（Inference） |
|---|---|---|
| 仕組み | 大量のデータをニューラルネットワークに入力。<br>重みづけを調整した特徴データの<br>組み合わせパターンを作成する | 学習されたAI基盤モデルに新たな事象を入力。<br>データを分析し、予測や分類を生成 |
| 計算リソースへの要求 | **高速な計算能力**<br>（大規模なデータセットを処理し、<br>モデルパラメータを反復的に調整） | **低レイテンシ**<br>（リアルタイムデータを効率的に処理） |
| | **大容量ストレージ**<br>（学習するデータセットを保存） | **スケーラビリティ**<br>（さまざまな推論要求を処理） |
| | **高速相互接続**<br>（学習時間を短縮） | **信頼性とサポート**<br>（ダウンタイムを最小化） |

出所）https://aivres.com/blog/ai-training-vs-inferencing-infrastructure-comparison
出所）https://developer.nvidia.com/blog/nvidia-deep-learning-inference-platform-performance-study/

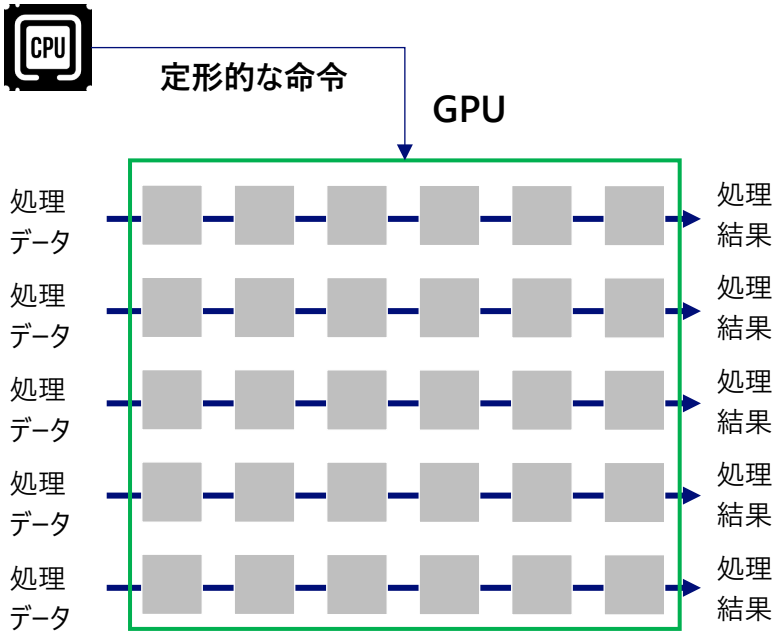
# GPUはAIの学習におけるアクセラレーターに

- 生成AIのAI基盤モデル生成（学習）の基本は、**「ニューラルネットワーク」による深層学習**
- CG（コンピュータ・グラフィクス）や科学計算の並列計算機として利用されてきた**GPUが、生成AIの「学習」におけるアクセラレーターとして活用**されるように

## 深層学習におけるニューラルネットワーク




## GPUは並列計算



# AI基盤モデルの学習では大量の計算リソースが必要

■「スケーリング則」によると、大量のGPU、データを用意すれば、性能が良いAI基盤モデルを開発できる

■ 従って、「最新の高性能GPUを大量に持つ企業がAI基盤モデル開発競争で勝利する」考えが一般化

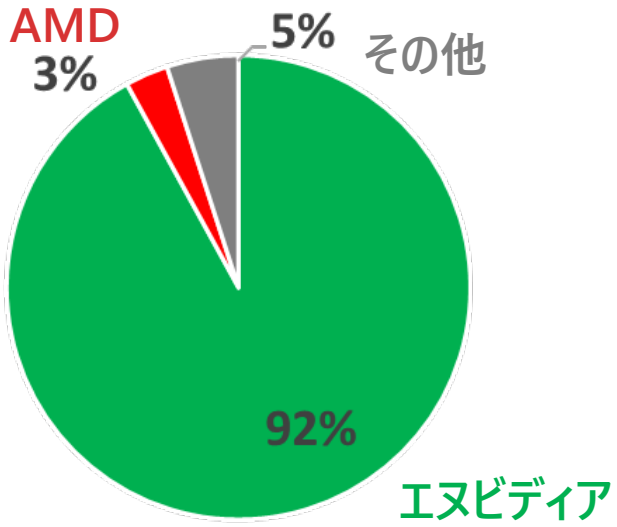
| AI基盤モデル | | 学習終了 | 利用GPU | GPU個数 |
|---|---|---|---|---|
| OpenAI | GPT-3 | Apr/2020 | V100 | 10,000 |
| Meta | Llama 1 | Jan/2023 | A100 | 2,048 |
| Meta | Llama 2 | Jun/2023 | A100 | 2,048 |
| Amazon | Titan | Apr/2023 | A100 | 13,760 |
| OpenAI | GPT-4 | Aug/2022 | A100 | 25,000 |
| Google | Gemini | Nov/2023 | TPUv4 | 57,000 |
| Meta | Llama 3 70B | Apr/2024 | H100 | 24,576 |
| Meta | Llama 3 405B | Apr/2024 | H100 | 24,576 |
| xAI | Grok 2 | Jun/2024 | H100 | 20,000 |
| xAI | Grok 3 | Dec/2024 | H100 | 100,000 |

出所）LifeArchitect.ai
https://lifearchitect.ai/

# データセンター向けAIコンピューティングで、エヌビディアは独走

- エヌビディア製GPUはデータセンター向けAIプロセッサとしても利用。**シェアは9割以上**
- **CUDA (Compute Unified Device Architecture)は、エヌビディア製GPU専用の並列計算プラットフォーム。**
  ゲーム開発者の囲い込み手段として提供されてきたが、2000年代にGPUが科学技術計算用途として利用されるようになると、**AI研究者を囲い込むためのツール**としても重要な役割を果たすようになった

## データセンター向けGPUのベンダーシェア（IoT Analytics調べ。2023年）




出所）IoT　Analytics
https://iot-analytics.com/wp/wp-content/uploads/2023/12/Generative-AI-market-share-of-leading-vendors-2023.png

## 2006年のCUDA発表以降、エヌビディアはHPCの進展にあわせてCUDAを進化

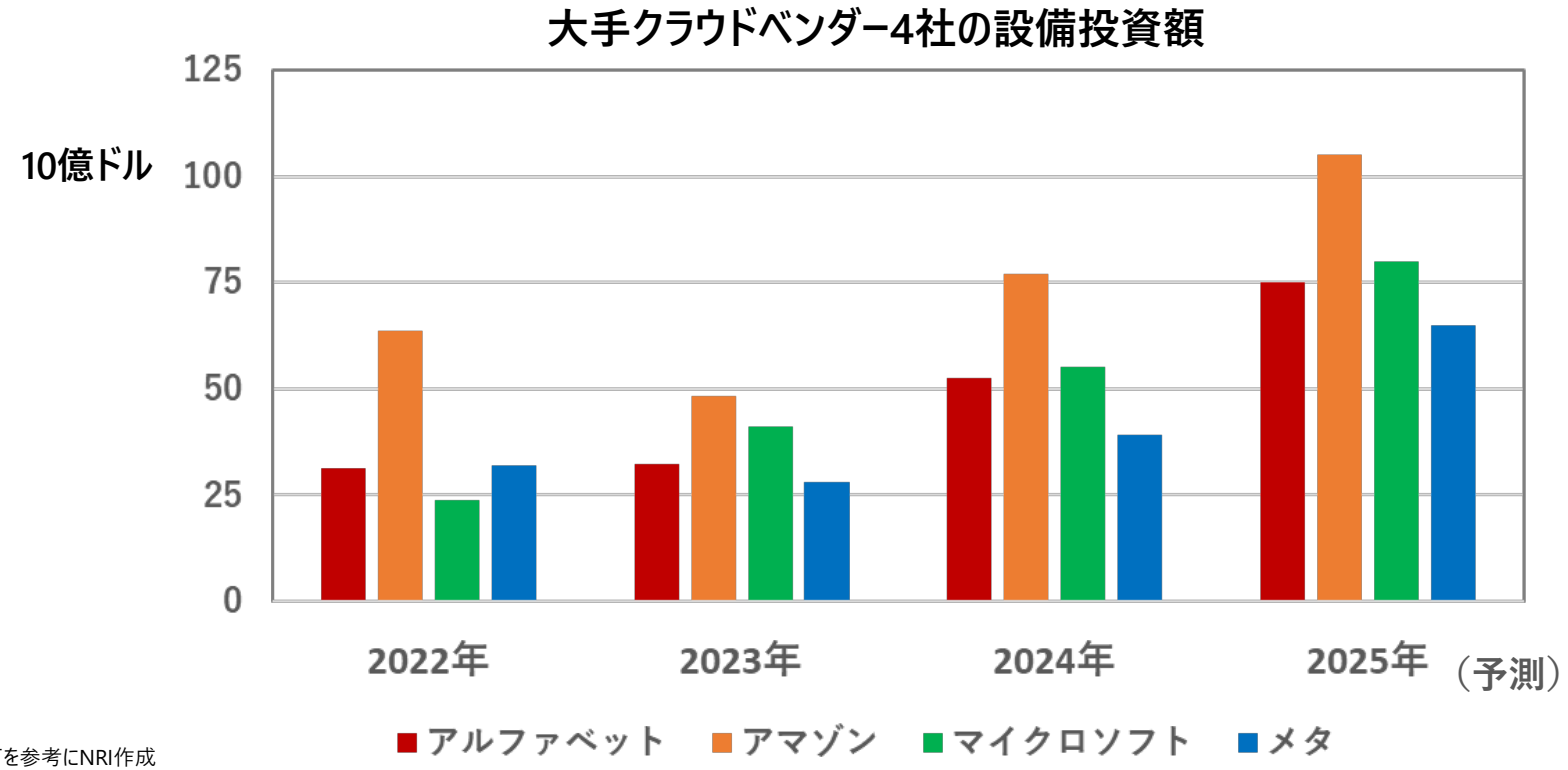
CUDA Toolkit 12.3（2023年10月20日リリース）では、6G の研究、量子コンピューティング、ゲノミクス、創薬、ロジスティクスの最適化などハイパフォーマンス コンピューティング （HPC）の分野、そしてロボティクス、サイバーセキュリティ、データ分析などの作業を支援する60以上のアップデートを追加

**2008 年の導入以来 3,300 万回以上ダウンロード（2022年3月時点）**

出所）Nvidia
https://www.nvidia.com/ja-jp/about-nvidia/press-releases/2022/nvidia-introduces-60-updates-to-cuda-x-libraries-opening-new-science-and-industries-to-accelerated-computing/

# 大手クラウドベンダーは、エヌビディア製GPUを大量調達。AI向けの設備投資を強化

- 2024年、マイクロソフトは48.5万個のエヌビディアHopperチップを購入、メタは22.4万個購入（※）
- 2025年の大手各社の設備投資は、前年比20％以上の増加の見込み

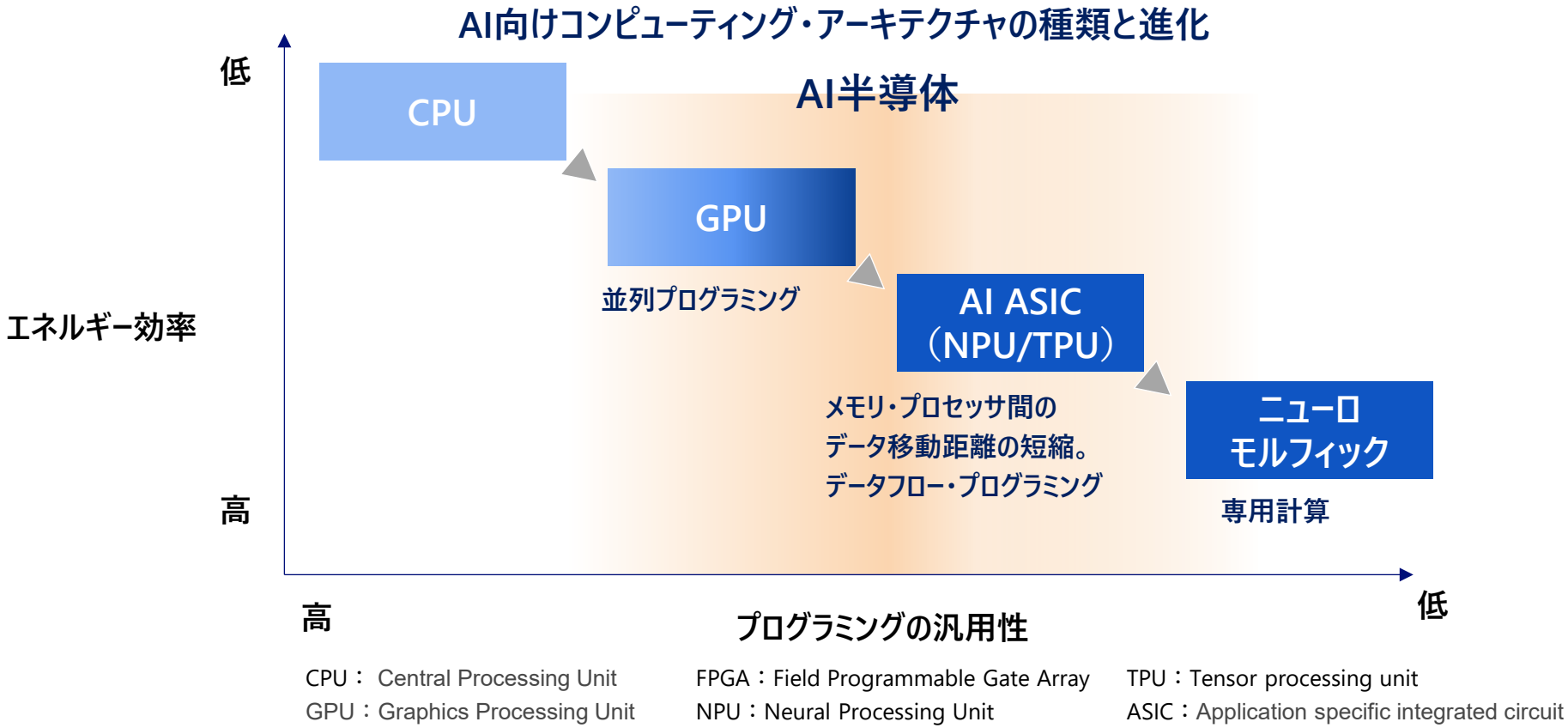



出所）以下を参考にNRI作成
https://www.businessinsider.com/big-tech-ai-capex-spend-meta-google-amazon-microsoft-earnings-2025-2
https://finance.yahoo.com/news/big-tech-set-to-invest-325-billion-this-year-as-hefty-ai-bills-come-under-scrutiny-182329236.html
https://www.wsj.com/tech/ai/the-ai-spending-race-is-still-on-as-google-antes-up-6671401d
（※）https://www.windowscentral.com/microsoft/microsoft-reportedly-acquired-the-most-nvidia-gpus-compared-to-its-rivals-including-google-and-meta-for-its-ai-projects-translating-to-485-000-chips-and-usd31-billion-in-expenditure

# AI半導体の動向

# 「AI半導体」としてのGPU、AI ASIC

- **AI半導体はアンブレラターム。**「効率的にAIの演算処理を行うことに特化した半導体デバイス」を指す
- CPUやGPU、FPGAといったプロセッサの一般的な分類の枠組みで整理することは難しいが、**一般的にはAI用で利用されるGPUやASICを示す**ことが多い




CPU： Central Processing Unit
GPU：Graphics Processing Unit
FPGA：Field Programmable Gate Array
NPU：Neural Processing Unit
TPU：Tensor processing unit
ASIC：Application specific integrated circuit

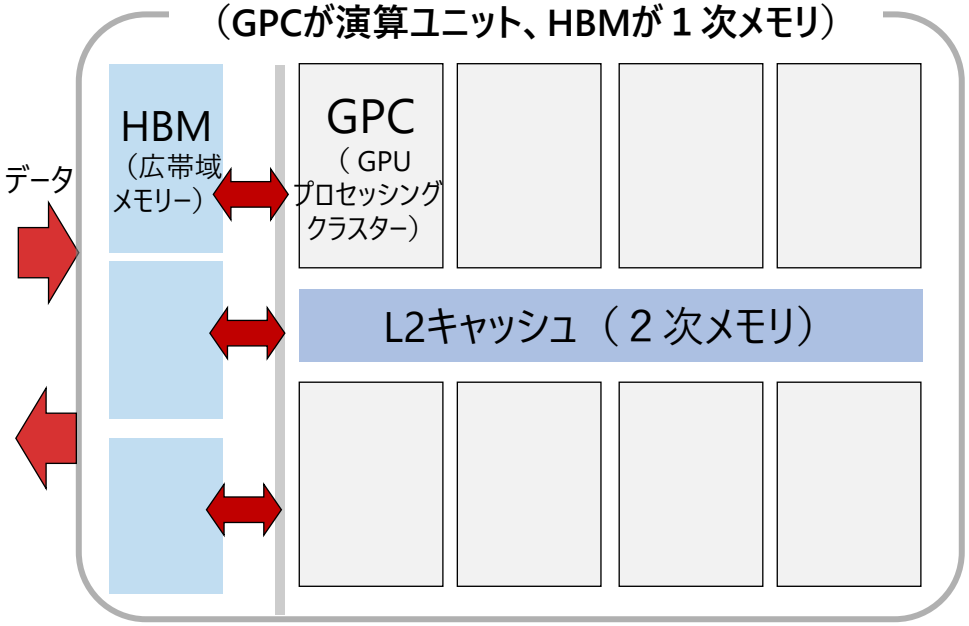
出所）NEDO「人工知能を支えるハードウェア分野の 技術戦略策定に向けて」（2018年10月31日）https://www.nedo.go.jp/content/100884652.pdf
出所）NAIST「Interview　創発的先端人材育成フェローシップ採択者　押尾 怜穏」 http://isw3.naist.jp/IS/is-fellowship/pr/interview/2023/oshio/
ほかを参考にした。

# GPUは最新HBM（広帯域メモリー）の採用により、パフォーマンスを向上

- GPUでは、CPUの指示を受けてGPU内の1次メモリであるHBM（広帯域メモリー）にデータ転送され、演算ユニットで並列計算が実行される。HBMのボトルネック化を回避させるため、GPUベンダーは**大容量HBMの採用とデータ転送速度の高速化でパフォーマンスを向上**させている
- データ転送は**エネルギー消費が大きく、消費電力に影響**。さらに、**汎用計算ならではのオーバーヘッド**がある

## GPUでは、HBMのメモリ容量、HBMと演算ユニットの間の通信帯域が、GPUの性能に影響を与える

エヌビディアH100 概念図
（GPCが演算ユニット、HBMが１次メモリ）




## エヌビディアは、2028年までのGPUロードマップを発表。GPUに搭載されるHBMは大容量化

| | コードネーム | 製品名 | HBMタイプ、搭載個数 | HBM総容量 | 消費電力（TDP） |
|---|---|---|---|---|---|
| 2028年 | Feynman | – | （次世代HBM） | – | – |
| 2027年 | Rubin Ultra | – | HBM4e x16個 | 1 T | – |
| 2026年 | Rubin | – | HBM4 x8個 | 288GB | – |
| 2025年 | Blackwell Ultra | B300 | HBM3e x8個 | 288GB | – |
| 2024年 | Blackwell | B200 | HBM3e x6個 | 192GB | 1200W |
| 2023年 | Hopper | H200 | HBM3e x6個 | 144GB | 1000W |
| 2022年 | Hopper | H100 | HBM3 x6個 | 80GB | 700W |
| 2020年 | Ampere | A100 | HBM2 x6個 | 40GB | 400W |

出所）以下を参考にNRI作成
https://developer.nvidia.com/ja-jp/blog/nvidia-hopper-architecture-in-depth/

# クラウドベンダーやスタートアップがAI専用ASICの開発に着手

■ クラウドベンダーの狙いは、**GPU調達コスト対策、パートナーおよび自社のAI基盤モデルの運用最適化**

■ **AI時代の本格化を見据え、スタートアップがAI専用ASIC市場に参入**

| ベンダー | AI ASIC | 概要 |
|---|---|---|
| グーグル | Trillium TPU | 自社製LLM「Gemini」に加え、派生のオープンモデル「Gemma」をサポート。第7世代TPUを開発中 |
| AWS | Trainium2 | 最大数兆のパラメータを持つAI基盤モデルおよび学習用 |
| | Inferentia2 | AI基盤モデルの推論に加え、小規模のAI基盤モデル（SLM）のファインチューニングも可能 |
| マイクロソフト | Maia 100 | Azureに展開される大規模なAIワークロード向けに特別に設計 |
| メタ | MTIA ver2 | 推論用アクセラレーター。オープンソースのRISC-Vベースの演算ユニットで構成 |
| テスラ | D1 | テスラのスパコン「Dojo」向けとして開発 |

| ベンダー | AI ASIC | 概要 |
|---|---|---|
| Graphcore | IPU | 10EFLOPSの「Good Computer」を発表。2024年7月、ソフトバンクグループが買収 |
| Groq | TSP（LPU） | グーグルでTPUを開発したエンジニアが起業。KDDIが出資 |
| Cerebras | WSE | 21.5㎠の大型半導体「ウェハースケールエンジン」を開発。東京エレクトロンデバイスが取り扱い |
| SambaNova Systems | RDU | サンマイクロシステムズやスタンフォード大出身者が起業。日本にデータセンター構築 |
| Tenstorrent | Wormhole | アップルやAMDのCPU開発を担当したジム・ケラー氏がCEO。オープンソースのRISC-Vを採用 |
| Preferred Networks | MN-Core | 神戸大学との共同研究を経て、MN-Core 2を開発。推論専用「L100」を開発中 |
| EdgeCortix | SAKURA- II | 8Wの消費電力で60TOPSを実現する推論専用のAIアクセラレーター |

# 参考）AI ASICは、データフロー型アーキテクチャ

■ AI ASICでは、乗算が行われるたびに、その結果が次の乗算ユニットに渡される。データとパラメータの乗算結果をすべて合計したものが出力。**行列乗算処理中に外部メモリにアクセスしない**ため、**データの移動距離を短縮させて、処理を高速化かつエネルギー消費を抑えることが可能**

グーグル　TPU（Tensor Flow Processor）の動作概念図

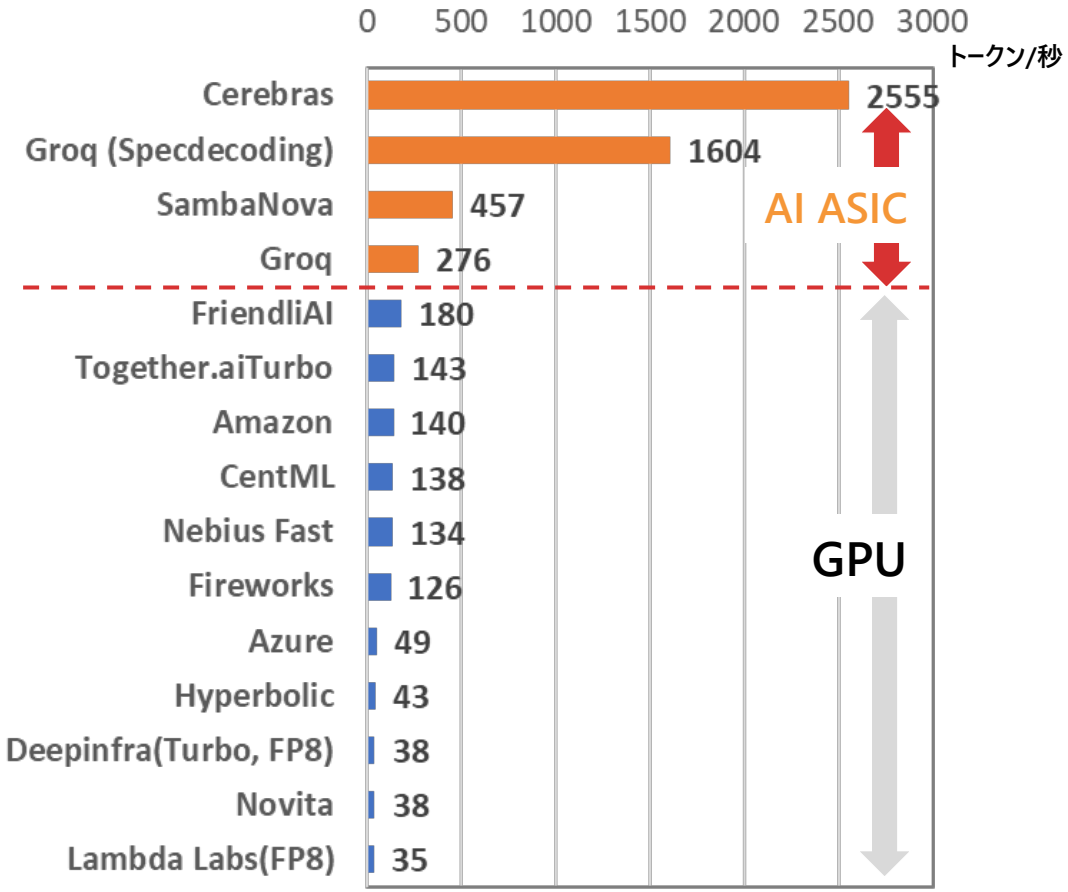
まず、HBMからパラメータを Matrix Multiplication Unit（MXU）に読み込み

次にHBMからデータを読み込み。乗算のたびに、その結果が次のユニット（乗算アキュムレータ）に渡される。データとパラメータの乗算結果をすべて合計したものが出力


出所）グーグル
https://cloud.google.com/tpu/docs/intro-to-tpu?hl=ja

# AI ASICであれば、推論の高速処理、大規模モデルの運用が容易

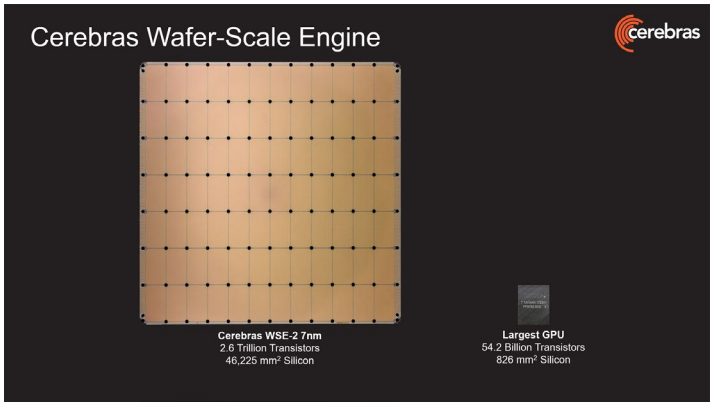
## Output Speed (tokens per second)
## Llama3.3 Instruct 70Bを用いて、1秒あたり返信可能なトークン数




出所）Artificial Analysis
https://artificialanalysis.ai/models/llama-3-3-instruct-70b/providers

※2025年3月20日時点

## Cerebras WSE-3は、コア数90万個
## オンチップメモリ44GB搭載（エヌビディアH100の880倍）




出所）Cerebras
https://cerebras.ai/company/press-kit/

## SambaNova Systems SN40Lシステムは、ラック1本でDeepSeek−R1のAI基盤モデルを実装可能


出所） SambaNova Systems
https://sambanova.ai/blog/sambanova-cloud-launches-the-fastest-deepseek-r1-671b

# 大手クラウドベンダーは、他社GPUと自社AI ASICの二刀流。ユーザーの需要を様子見

- AWS、グーグルは、クラウドユーザー向けに**自社製AI ASICのインスタンスを提供**
- マイクロソフトのAI ASIC「Maia 100」は、第一世代。インスタンス未提供（2025年3月時点）
  - マイクロソフト製AI基盤モデル「MAI（Microsoft Artificial Intelligence）」開発に利用している可能性を指摘される（※）など、まずはマイクロソフト内での利用を通じて、技術成熟を図る見込み

| | AI半導体ベンダー | 提供インスタンス（カッコ内は利用可能なAI半導体） |
|---|---|---|
| マイクロソフト | エヌビディア | ND-family（A100、H100、H200） |
| | AMD | ND-family（MI300X） |
| AWS | エヌビディア | P5e（H200）、P5（H100）、P4（A100）、P3（V100）<br>G3（M60）、G4dn（T4）、G5（A10G）、G6（L4）、G6e（L40S） |
| | インテル | DL1（Gaudi） |
| | クアルコム | DL2q（AI 100） |
| | AWS | G5g（Graviton2）、Trn2（Trainium2）、Trn1（Trainium）、<br>Inf2（Inferentia2）、Inf1（Inferentia） |
| グーグル | エヌビディア | A3 Ultra（H200）、A3 （H100）、A2（A100）、G2（L40）、N1 （T4）<br>N1+P100（P100）、N1+V100 （V100）、 |
| | グーグル | Trillium、TPU v5p、TPU v5e、TPU v4 pod<br>TPU v3 pod、TPU v3 device、TPU v2 pod、TPU v2 device |

※2025年3月10日時点

※出所）SiliconANGLE、‘Microsoft reportedly develops LLM series that can rival OpenAI, Anthropic models’,2025年3月7日
https://siliconangle.com/2025/03/07/microsoft-reportedly-develops-llm-series-can-rival-openai-anthropic-models/

# AI半導体における、GPUかAI ASICかの議論の行方は？

■ 経済学のジェヴォンズ・パラドックスに従うと、「AIがより効率的で利用しやすくなるにつれて利用が増加し、いくらあっても足りないほどAIはコモディティー化する」→**AI半導体全体の需要が増加**

- ジェヴォンズ・パラドックス＝「資源効率の向上は長期的な資源消費の減少ではなく増加」

## Groqのジョナサン・ロスCEOへのインタビュー（AIM Americas）

"Training should be done on GPUs（**学習はGPUで行われるべき**）," Ross said. "I think NVIDIA will sell every single GPU they make for training."

Ross added that if Groq were to deploy high volumes of lower-cost inference chips, the demand for training would increase. "The more inference you have, the more training you need, and vice versa（**推論が多ければ多いほど、必要なトレーニングも増えるし、その逆もまた同様**）," he said.

## AI ASICベンダーの取り組み（2025年）

| | 取組み | 発表日 |
|---|---|---|
| Groq | サウジアラビアと**15億ドル（2250億円）の契約**を締結 | 2025/2/10 |
| Cerebras | 北米と欧州に**計6か所の推論データセンター新設**を発表。Mistral、Perplexityに加え、HuggingFaceとAlphaSenseが採用を発表 | 2025/3/11 |
| SambaNova Systems | ソフトバンクと協業し、**アジア太平洋地域全体で高速AI推論を提供** | 2025/3/5 |

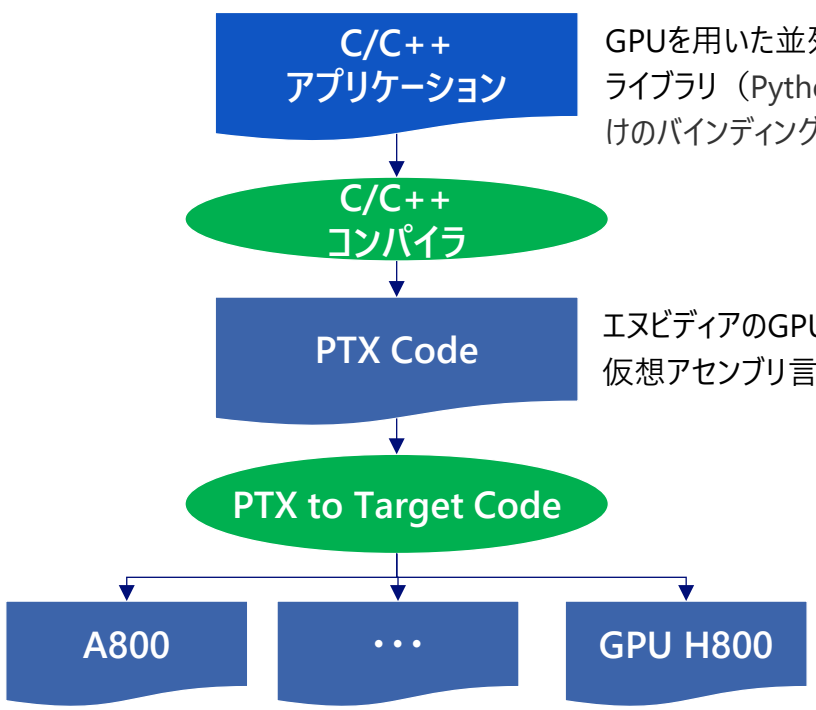
出所）AIM Americas,' Groq Aims to Provide At Least Half of the World's AI Compute, Says CEO',2025年2月25日
https://analyticsindiamag.com/global-tech/groq-aims-to-provide-at-least-half-of-the-worlds-ai-compute-says-ceo/

# DeepSeekショック以降のAIコンピューティング

# 「DeepSeek R-1」のインパクトの一つは、CUDAを利用せず、かつ機能に制約があるGPU「H800」を約2000個使って、AI基盤モデルを開発できたこと

- DeepSeekは、NVIDIAが提供する中間コードの一種であるPTX（Parallel Thread Execution）を直接記述。**CUDAのAPI（C言語やC++拡張ライブラリ）の利用を回避して計算効率を向上**
  - さらに、**学習におけるGPUのメモリー消費を減らすために、演算精度を落とす**（16ビット浮動小数点演算から8ビット浮動小数点演算へ）ことで、GPUの利用個数を削減

## CUDAを用いたプログラミングの流れ




GPUを用いた並列計算用のライブラリ（PythonやJava向けのバインディングもある）

エヌビディアのGPU向け仮想アセンブリ言語

**一般的なエヌビディアユーザー**

並列処理に関わる最適化やGPUの世代の違い、エラー処理はCUDAが実施。アプリケーションの保守性が高い

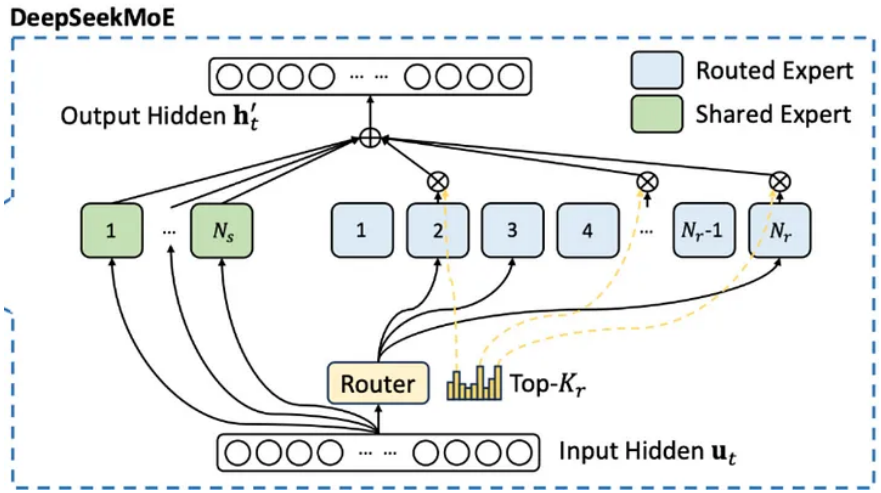
**DeepSeek**

アセンブリ言語に近く、開発が難しく、保守性が悪いが、計算のオーバーヘッドを減らし、処理を高速化できる

# さらにMoE（Mixture of Experts）、強化学習の拡張などを実施。卓越したエンジニアリング力によって推論を強化

- 複数の専門家AIモデルを採用。**必要なモデルのみを選択して動かす**ため、計算リソースを節約
  - 小さなステップに問題を分解して出力を検証するCoT（Chain-of-Thought）を採用し、精度を向上
  - さらに、強化学習（Reinforcement Learning）で学習するため、コスト圧縮に貢献
- MoEやCoTは従来から提案されていたが、**DeepSeekはアセンブラレベルでのGPU制御とアルゴリズムの工夫で、モデル開発におけるコスト問題と実行時の性能問題に対応**

## MoE（Mixture of Experts）概念図




DeepSeekでは、256個の専門家AIモデル（Routed Expert）と一般知識を持つAIモデル（Shared Expert）を準備

出所）DeepSeek
https://deepseekai.jp/different-models-of-deepseek/

## 強化学習における工夫

| 特長 | 関連キーワード | 概要 |
|---|---|---|
| **強化学習前の準備** | **コールドスタート教師あり微調整（SFT）** | DeepSeek-v3をベースに、約1,000 個の高品質な Chain-of-Thought (CoT) のサンプルが手動でキュレーション済み。これにより、データの精度が向上 |
| **強化学習の採用** | **グループ相対ポリシー最適化(GRPO)の採用** | 複数の生成結果をグループ化し、モデルの複数の異なる出力を比較し、そのグループ内での相対的な良さに基づいて学習されるため、計算効率が向上 |
| **小さなモデルへの蒸留** | **小さなモデル（15億～700億パラメータ）に凝縮** | 学習済みモデルから80万個のサンプルが生成。サンプルには600Kの推論タスクと200Kの一般タスクが含まれる。蒸留中は強化学習が行われないため、計算効率が向上 |

出所）以下を参考にした
https://www.mygreatlearning.com/blog/deepseek-r1-features-use-cases/
https://dev.to/prathameshdevadiga/deepseek-r1-internals-made-easy-16ia

# Inference（推論）とReasoning（論理的推論）

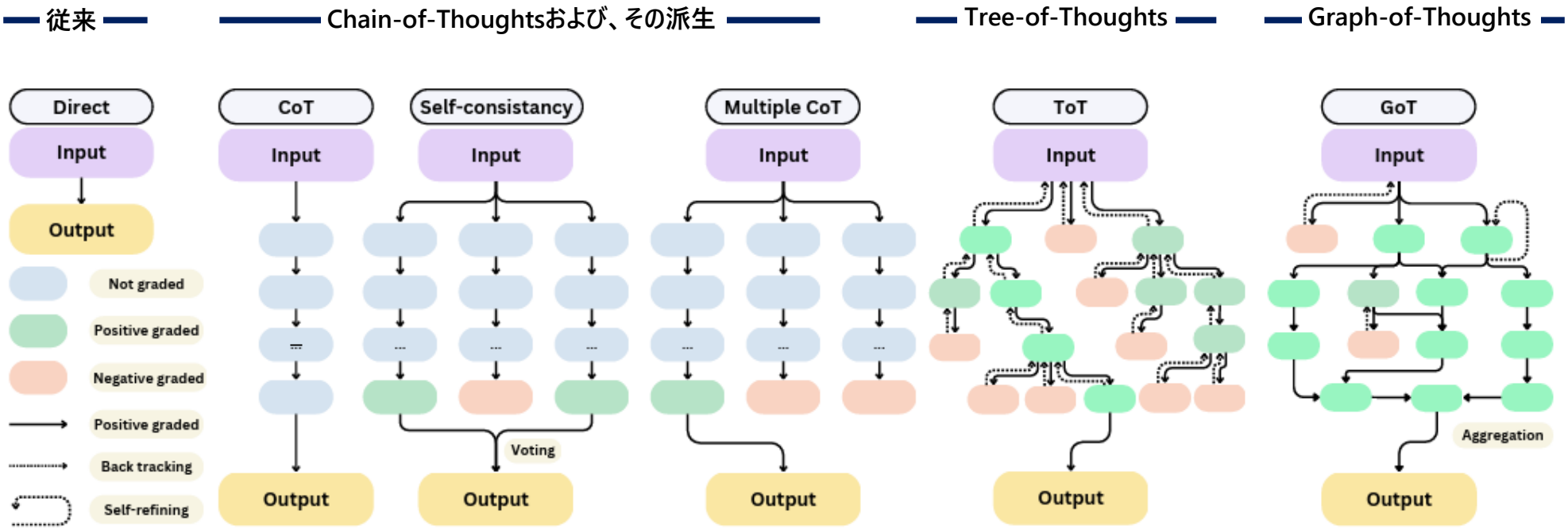
■ OpenAIは、2025年2月に発表した**GPT-4.5を強化学習を利用しない巨大な知識ベース**と位置づけ

- 「GPT-4.5 のようなモデルが、事前学習を通じてより高度な知識と知性を獲得することで、**将来的にはReasoningやツール活用能力を持つエージェントの、より強固な基盤となる**ことが期待されます。」（OpenAIホームページ）
- DeepSeek-R1は、DeepSeek-V3をベースとして、教師あり学習と強化学習を繰り返して初期モデルを作成した

| | Inference（推論） | Reasoning（論理的推論） |
|---|---|---|
| 実行 | AI基盤モデルの実行<br>（通常の推論） | Chain of Thoughtなどの手法を<br>利用して多段的に推論を行う |
| 用途 | 情報検索、コンテンツ作成など<br>日常的なタスク | 数学的問題など<br>複雑な問題解決 |
| 利用する<br>AI基盤モデル | DeepSeek-V3<br>OpenAI GPT-4.5<br>Llama | DeepSeek-R1<br>OpenAI o1、o3-mini、<br>Gemini 2.0 Flash Thinking |
| AI基盤モデルの役割 | 知識 | 論理あるい推理 |

# Reasoning（論理的推論）では、さまざまな手法が提案されている

- Tree-of-Thoughts（ToT）やGraph-of-Thoughts（GoT）は、AI基盤モデルの推論過程をツリーやグラフ構造で表現する手法
  - GoTでは、事前に定義された枝やルートに縛られることなく、新しい考えや情報を追加できるため、Chain-of-ThoughtsやTree-of-Thoughtsと比較すると、思考の範囲が限定されることが少なくなるメリットがある

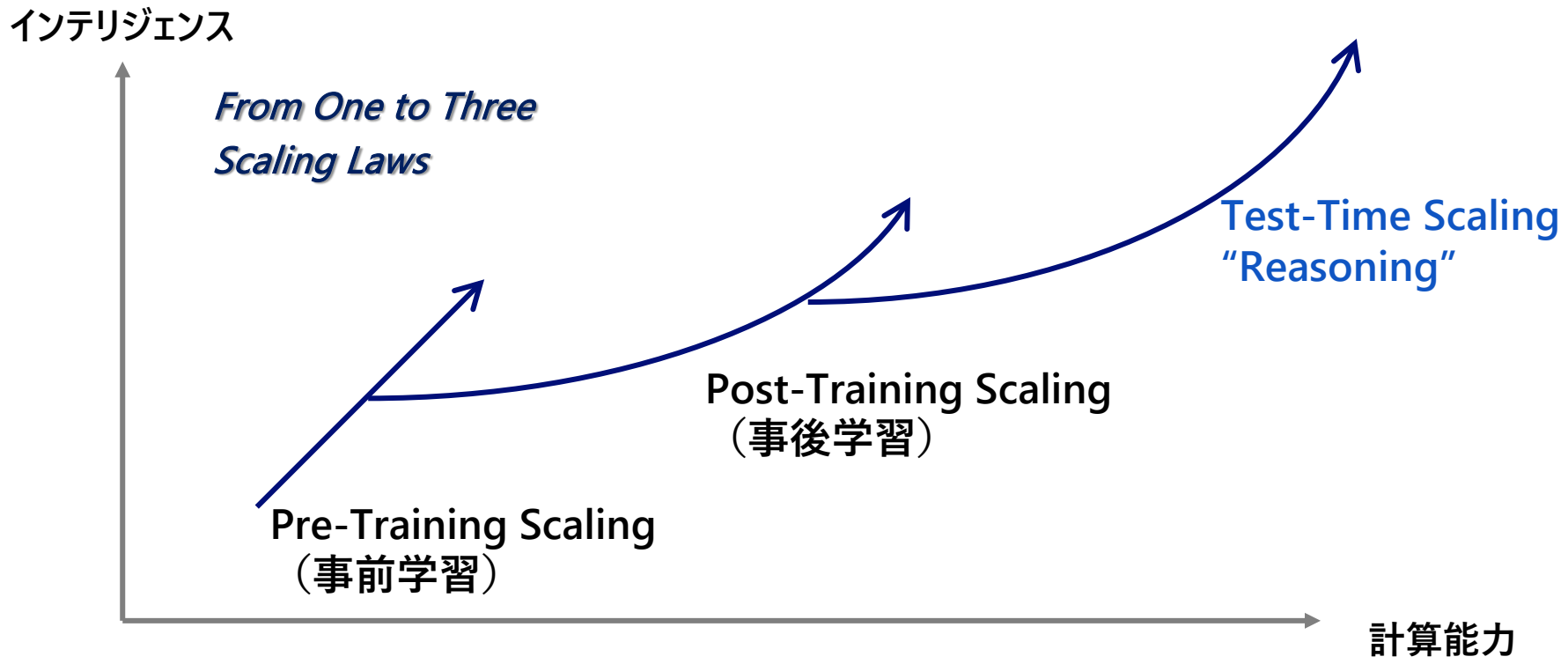



出所）Komal Kumarほか,'LLM Post-Training: A Deep Dive into Reasoning Large Language Models',2025年2月28日
https://arxiv.org/pdf/2502.21321v1

# Reasoning （論理的推論）による、「スケール則」の進化

- 従来のInference（推論）と異なり、**Reasoningでは計算が複雑。計算能力が必要**
- エヌビディアは自社のGPUが貢献する場面が増えるとして、Reasoningの台頭を歓迎

CES 2025の基調講演でエヌビディアCEOが示したスライド（イメージ）



# 今後の展望

# 今後の注目ポイント

1. サービスとしてのGPU/AI ASIC
2. エッジにおけるAI活用（エッジAI）
3. ソブリンAI（AI主権）への対応

今後の展望

# サービスとしてのGPU、AI ASIC

■ GPU、AI ASICのサービス提供の拡大は、**AIに取り組みたい企業にとっては好機**

■ **選択肢が豊富な大手クラウドか？料金で訴求する国内GPUサービス、AI ASICサービスか？**

| | 事業者 | サービス名 | GPU | | | AI ASIC | 参考料金（一部料金は、オンデマンド利用かつ24時間x30日、1ドル150円を前提に計算） |
|---|---|---|---|---|---|---|---|
| | | | H200 | H100 | A100 | | |
| 大手クラウドベンダー | マイクロソフト | NDファミリ | ○ | ○ | ○ | | 約1077万円/月　ND96isr H100 v5：H100 8基選択時 |
| | AWS | EC2 G4/5/6、P3/4/5 | ○ | ○ | ○ | | 約1145万円/月　p5en.48xlarge ：H200 8基選択時 |
| | | Trainium、Inferentia | | | | Trainium、Inferentia | 約267万円/月　trn1n.32xlarge：Trainium 16基選択時 |
| | グーグル | Compute Engine A3、A2、G2、N1 | | ○ | ○ | | 約969万円/月　a3-highgpu-8g：H100 8基選択時<br>約312万円/月　a2-highgpu-8g：A100 8基選択時 |
| | | Cloud TPU | | | | Trillium、TPU v2~v5 | 約35万円/月（チップ費用のみ。システム費用は除く）（東京リージョン、asia-northeast1選択時） |
| 国内クラウド・GPUサービスベンダー | さくらインターネット | 高火力DOK | | ○ | ○※ | | 約75万円/月　H100 1基　通常料金適用時 |
| | GMOインターネット | GPUクラウド | ○ | ○※ | | | 専用プラン：380万円/月<br>共有プラン：GPU−100円/分、CPU-20円/分<br>H200 8基　エヌビディア推奨構成を選択時 |
| | NTT-PC | WebARENAGPU | | | ○ | | 約23万円/月　t80-1-a-standard：A100 1基選択時 |
| | ハイレゾ | GPUSOROBAN | ○ | | ○ | | 約273万円/月　1ノード-スタンドアロン：H200 8基構成時 |
| AI ASICベンダー | Preferred Networks | Preferred Computing Platform | | | | MN-Core2 | 約170万円/月　1ノード（MN-Core2 8基構成）選択時 |
| | SambaNova Systems | SambaNova Cloud | | | | SN40L | API：LLM毎に変動、専有利用：要相談 |

※別サービスブランド名で提供中

今後の展望

# AI ASICスタートアップは利用可能なモデルを充実化し、ユーザーに訴求

## Groq Cloud

最新のAI基盤モデル、
音声認識、画像に対応

（言語）
- DeepSeek R1 Llama 70B
- DeepSeek R1 Distill Qwen 32B 128k
- Qwen 2.5 32B Instruct 128k
- Qwen 2.5 Coder 32B Instruct 128k
- Qwen QwQ 32B
- Mistral Saba 24B
- Llama 3.2 1B (Preview) 8k
- Llama 3.2 3B (Preview) 8k
- Llama 3.3 70B Versatile 128k
- Llama 3.1 8B Instant 128k
- Llama 3 70B 8k
- Llama 3 8B 8k
- Mixtral 8x7B Instruct 32k
- Gemma 2 9B 8k
- Llama Guard 3 8B 8k
- Llama 3.3 70B SpecDec 8k

（音声）
- Whisper V3
- Whisper Large v3 Turbo
- Distil-Whisper

（画像）
- Llama 3.2 11B Vision 8k (Preview)
- Llama 3.2 90B Vision 8k (Preview)

灰色はプレビュー版

## Cerebras Cloud

Cerebrasで学習したAI基盤モデルを
利用できる

（言語）
- Llama 3.3
- Mistral
- JAIS
- FALCON
- T5
- STARCODER
- GIGAGPT
- CRYSTALCODER
- BTLM-CHAT
- CEREBRAS-GPT
- Med42

（画像）
- DIFFSION　Transformer

Cerebras学習モデル

モデルの詳細記述が不明なものがあるため、
ホームページ記載のモデル表記に従った

## SambaNova Cloud

最新の超大型AI基盤モデル、
日本発のLLM（Swallow）を利用できる

（言語）
- DeepSeek-R1 671B
- DeepSeek-R1-Distill-Llama-70B
- Ai2 Llama-3.1-Tulu-3-405B
- Llama-3.1-Swallow-70B-Instruct-v0.3
- Llama-3.1-Swallow-8B-Instruct-v0.3
- Meta-Llama-3.1-405B-Instruct
- Meta-Llama-3.1-70B-Instruct
- Meta-Llama-3.1-8B-Instruct
- Meta-Llama-3.2-1B-Instruct
- Meta-Llama-3.2-3B-Instruct
- Meta-Llama-3.3-70B-Instruct
- Qwen2.5-72B-Instruct
- Qwen2.5-Coder-32B-Instruct
- QwQ-32B-Preview

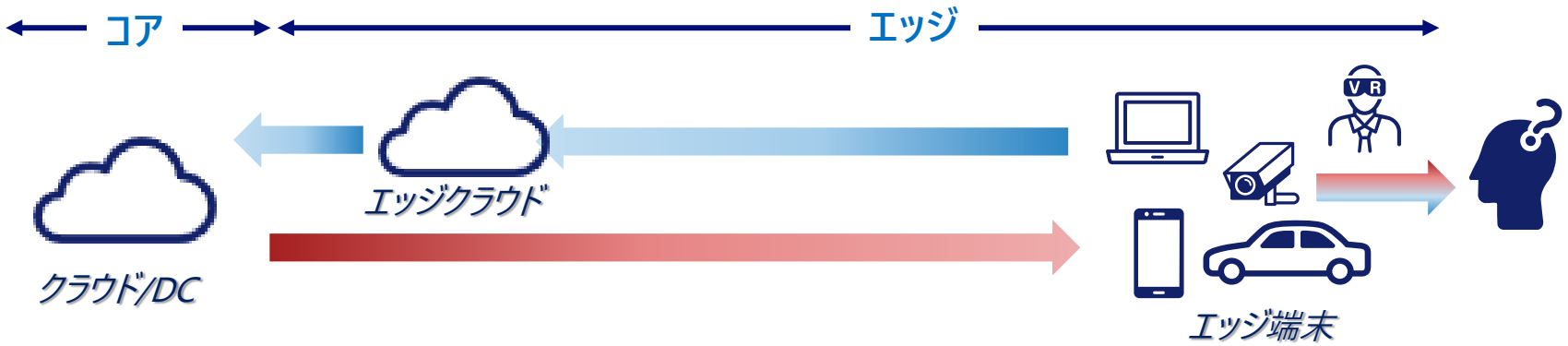
（画像）
- Llama-3.2-11B-Vision-Instruct
- Llama-3.2-90B-Vision-Instruct

※各社とも2025年3月10日時点の提供内容

今後の展望

# AI半導体の次の主戦場は、エッジへ

- **エッジAIとは、エッジコンピューティングとAIの融合**。ネットワークの端の端末でAI処理を実施
- AI PCから自動車、ロボット、センサーなど、エッジで利用されるデバイスの種類はさまざま
  - 「エヌビディアは新たにASIC部門を設立、推論AIチップの市場拡大を想定し、1000人のエンジニアを台湾で採用」(台湾メディアのCommercial Times)




| AIコンピューティングへの要求 | | |
|---|---|---|
| | • AI基盤モデル開発のための学習<br>• 継続追加学習、ファインチューニング<br>• 大型AI基盤モデルを用いた推論 | • 特定タスクの実行（低消費電力対応）<br>• 機密性の高いデータを扱った推論<br>• 小型AI基盤モデル、小型端末への実装 |
| | GPUが有利 ⟷ | AI ASICが有利 |

今後の展望

# 新たなAI半導体ビッグプレイヤー登場への期待

■ エッジを含む、次世代のAIコンピューティングインフラでの市場獲得に挑むプレイヤーが続々登場

| ベンダ | 設立国 | 設立年 | 事業概要 | 調達額（百万米ドル。2025年3月10日時点） | |
|---|---|---|---|---|---|
| SambaNova Systems | 米国 | 2017 | 再構成可能なデータフロー ユニット (RDU) を開発 | 1,100 | |
| Tenstorrent | カナダ | 2016 | Tensix コアとRISC-Vを組み合わせたAIチップを提供。ラピダスと提携 | 1,000 | |
| Groq | 米国 | 2016 | AI アプリケーションに超高速を実現する言語処理ユニット (LPU) を開発 | 1,000 | |
| Lightmatter | 米国 | 2017 | 推論向け photonic chipsを開発 | 822 | |
| Cerebras Systems | 米国 | 2016 | データセンター向けAIシステムを開発・提供 | 715 | NASDAQ上場申請中 |
| GraphCore | 英国 | 2016 | AIコンピューティング用のビルディングブロックとリファレンスデザインを提供 | 682 | ソフトバンクが買収 |
| Hailo | イスラエル | 2017 | エッジAIプロセッサを開発 | 344 | |
| Celestial AI | 米国 | 2020 | AIプロセッシング向け光インターコネクトの技術開発を実施 | 339 | |
| NUVIA | 米国 | 2019 | データセンター向けシリコンチップを開発 | 293 | |
| SiMa.ai | 米国 | 2018 | エッジAI向け超低電力ソフトウェアとチップを開発 | 270 | |
| Blaize | 米国 | 2010 | 自動車、スマートビジョン向けAIプラットフォーム開発 | 242 | NASDAQ上場（BZAI） |
| Enfabrica | 米国 | 2019 | GPUネットワークインターフェースコントローラーチップを開発 | 240 | |
| Rebellions | 韓国 | 2020 | シリコンアーキテクチャと深層学習のギャップを埋めるAIアクセラレータ開発 | 224 | |
| Kneron | 米国 | 2015 | 特定用途向け集積回路とソフトウェアを開発 | 212 | |
| Cambricon Technologies | 中国 | 2016 | AIチップ開発 | 200 | |
| Mythic | 米国 | 2012 | アナログ マトリックス プロセッサ (AMP) を開発 | 165 | |
| EnCharge AI | 米国 | 2022 | analog in-memory-computing AI chipsを開発 | 163 | |
| Untether AI | カナダ | 2018 | AI推論ワークロードを加速するように設計された高性能AIチップの開発 | 152 | |
| Axelera | オランダ | 2021 | エッジコンピューティング向けAIチップ開発 | 137 | |
| EdgeQ | 米国 | 2018 | 5GとAIに特化したチップを開発 | 126 | |
| Etched.ai | 米国 | 2022 | transformer modelsの動作に特化したAIチップ開発 | 125 | |
| Esperanto Technologies | 米国 | 2014 | 機械学習を目的とした新しいRISC-Vベースのチップを開発 | 124 | |
| Luminous Computing | 米国 | 2018 | AI向けphotonics chipを開発 | 115 | |
| Prophesee | フランス | 2014 | neuromorphic vision systemsを開発 | 110 | |
| MatX | 米国 | 2018 | 大規模言語モデル用AIチップを開発 | 105 | |

出所）CrunchBase、AI Sartupsほかを参考に作成

今後の展望

# AI ASICの先。究極はニューロモルフィックコンピューティングへ

- ニューロモルフィックは、人間の脳の神経回路を模倣したもの。IBMとインテルが研究で先行
- 試作品が登場しているが基礎研究段階。将来、**低消費電力で動作できる特長を生かし、エッジAIの適用先を広げることができる**
  - 人間の脳はおよそ20Wで動作。現在使われているデジタルAI計算の１万分の１の電力

## ニューロモーフィック・チップ開発ベンダー

| ベンダー名 | 製品名 |
|---|---|
| ABR | Nengo Brain Board |
| BrainChip | Akida neural processor |
| GrAI Matter Labs | GrAI One |
| General Vision | NeuroMem |
| IBM | TrueNorth |
| Intel | Loihi |
| Nepes | NM500 |
| Rain Neuromorphics | （製品開発中） |
| SynSense | DYNAP |

## MercedesのコンセプトカーVision EQXX

**音声認識（キーワード検出）システムに、BrainChipを利用。「音声制御よりも5～10倍効率的」**

出所）Mercedes-Benzグループ
https://mercedes-benz-media.co.uk/models/electric-concept-vision-eqxx/photos?page=2

今後の展望

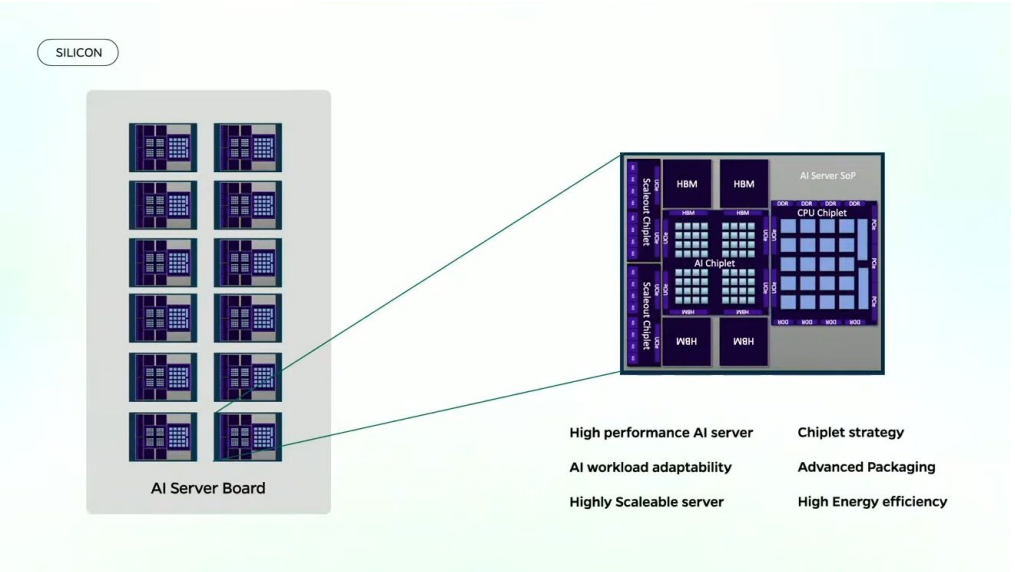
# ソブリンAI（主権AI）は、AI半導体市場に影響を及ぼす可能性がある

- ソブリンAIとは「各国が自国のインフラ、データ、人材を活用して独自にAIを開発・運用する能力」のこと
- 各国政府がAI戦略を掲げ、自国のAI産業育成、AIの実用化に取り組むなか、**自国での適用を想定し、AI半導体の開発を行う企業が登場**

**中国では自治体や国営企業がDeepSeekの採用を開始。大手メーカーは、エヌビディア以外のGPUを搭載したAIサーバーを提供**

| メーカー | 動向 |
|---|---|
| Inspur | ・オールインワンDeepSeek R1ベースのAIサーバーを提供。**エヌビディアH20 と中国製AIプロセッサ**から選択可能 |
| Huawei | ・Atlas 800I A2（DeepSeek AIサーバー）を2025年2月に発表。**HuaweiのAscend 910B GPUを搭載**（DeepSeekは、推論モデルの実行にAscend 910C GPUを使用している） |

**AI企業のOla Krutrimは、インド初のAIチップを開発。2025年に製品化。自国のスパコンへの適用とインドの多種多様な言語に対応予定**




出所） Ola Krutrim
https://www.youtube.com/watch?v=EP1x_9LMp50&t=4098s

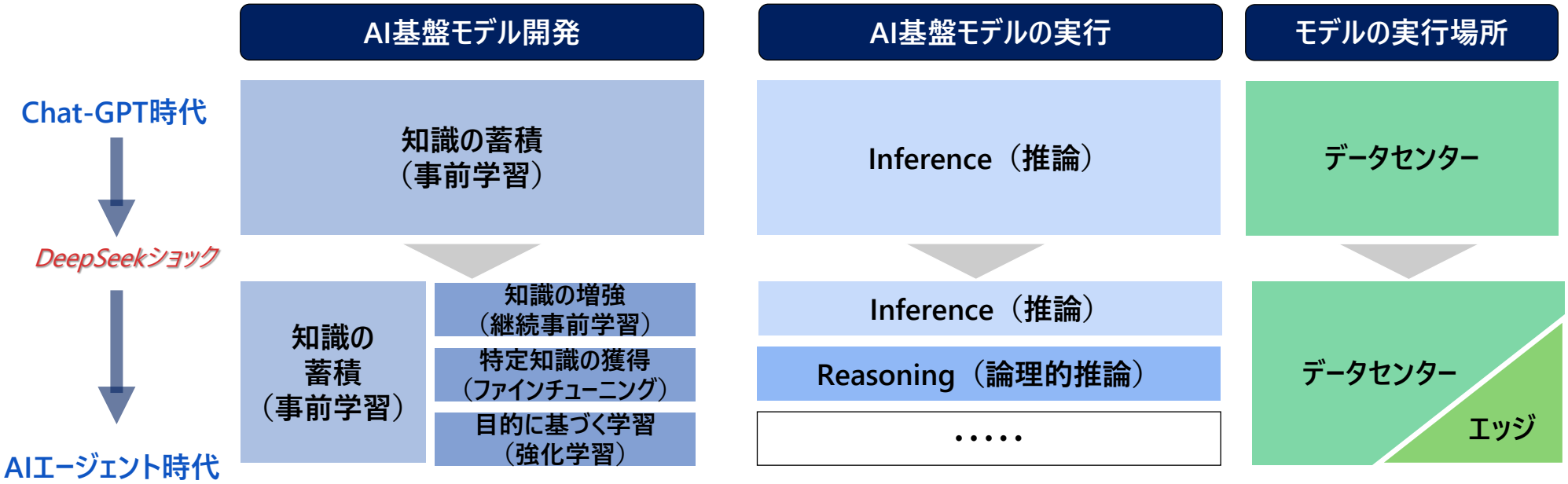
今後の展望

# これからのAIコンピューティング

■ **AIは、モデルの実行で真価を発揮**

■ アルゴリズム研究の進化は著しく、新たな学習方式、推進方式提案の可能性もある。「コスト効率が高いコンピューティング環境をどう準備すればよいか」の問いに対する回答は益々難しく

- "Ross said Groq contemplates selling its **LPUs as a "nitro boost to GPUs**". 「GroqのLPU（AI ASIC）は、GPUのニトロブーストとして販売することを検討している。」（※）




※出所）AIMAmerica
https://analyticsindiamag.com/global-tech/groq-aims-to-provide-at-least-half-of-the-worlds-ai-compute-says-ceo/

# まとめ

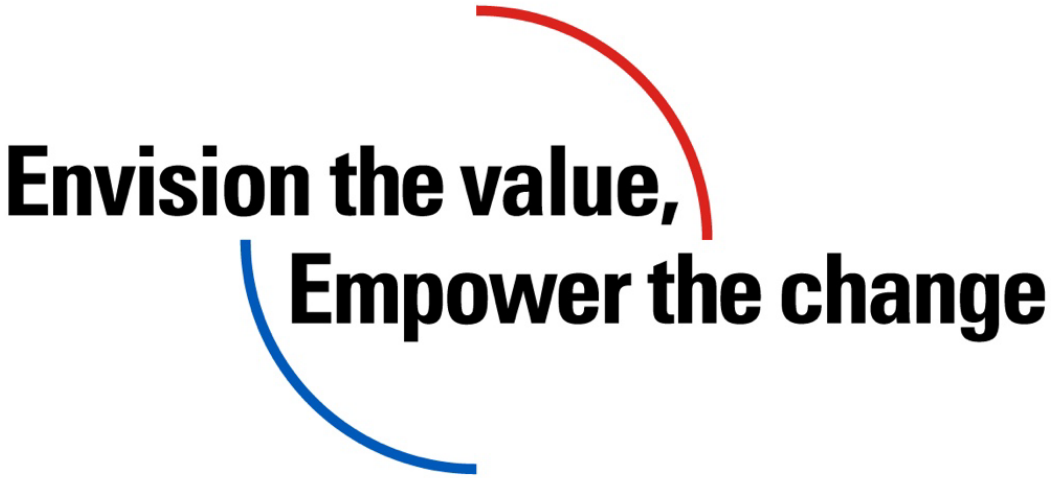
■AI半導体への高い期待
- エヌビディアのGPUがデータセンター向けAIアクセラレーターとして席巻
- 大手クラウドベンダーは、AI向け設備投資を継続

■AI半導体の新たな選択肢、AI ASIC
- AI ASIC開発ベンダーは、クラウドサービスを提供開始
- 大手スタートアップは、最新のオープンソースAI基盤モデルを準備し、ユーザーに訴求

■DeepSeekは、ハードとソフトのエンジニア力で大手AIベンダーに対抗できることを示した
- ハード：CUDAに依存しない、アセンブリレベルでGPUを制御
- ソフト：MoEとCoT、GRPOなどアルゴリズムを工夫、精度の高いReasoning（論理的推論）を実現

**今後の注目ポイント**

- サービスとしてのGPU/AI ASIC：AIへに取り組む企業の選択肢
- エッジにおけるAI活用（エッジAI）：AIの本格活用、AI ASIC活躍の場
- ソブリンAI（AI主権）への対応：市場獲得に向けたAI半導体の新たな提案


Envision the value,
Empower the change