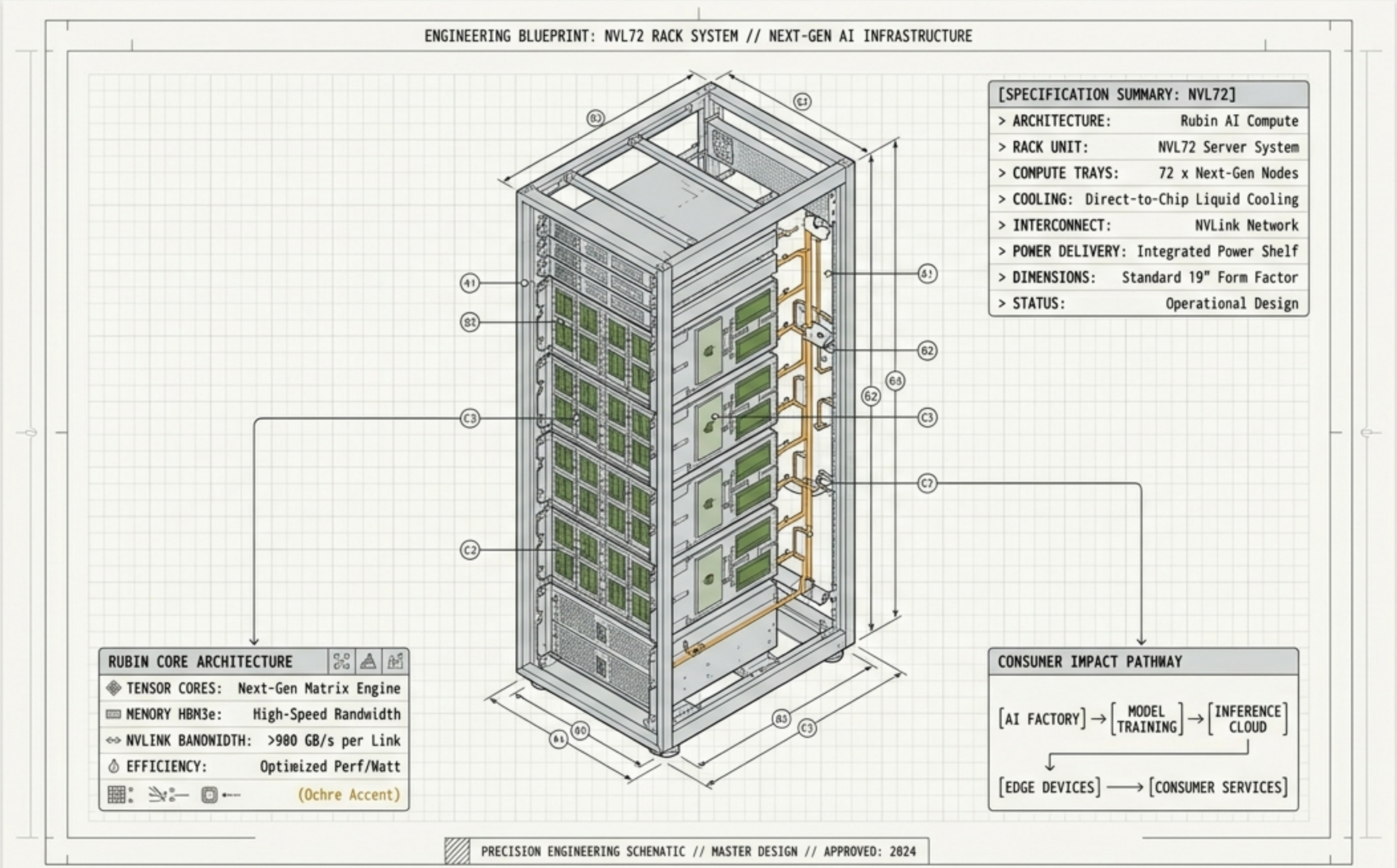
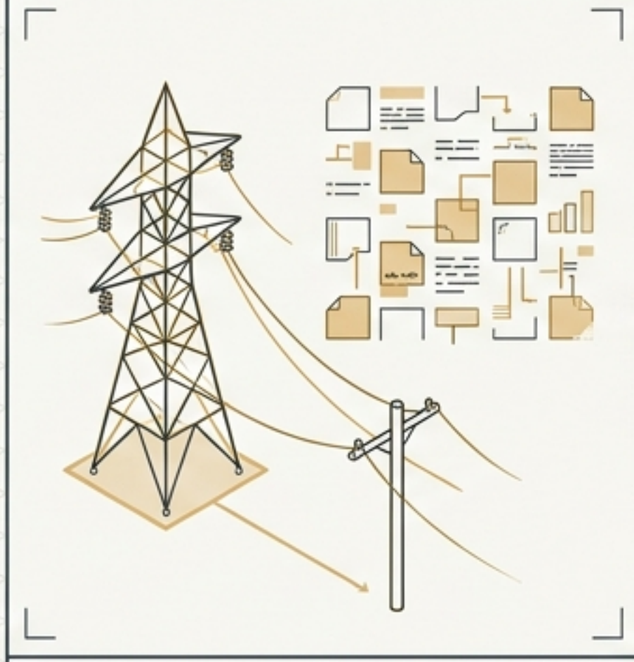
次世代AIファクトリーの設計図：Rubinアーキテクチャからコンシューマーまでの完全解剖



データの保管庫から「トークン生成工場」への進化

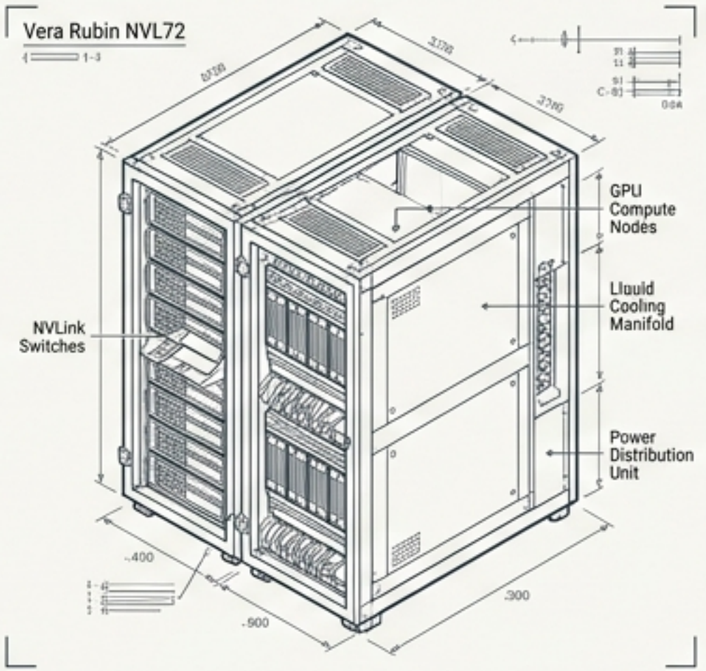
AIファクトリーは、単なるサーバーの集合体ではありません。電力、シリコン、データを継続的に知能（トークン）へと変換する、24時間稼働の生産プラットフォームです。

入力



電力・未構造化データ

処理



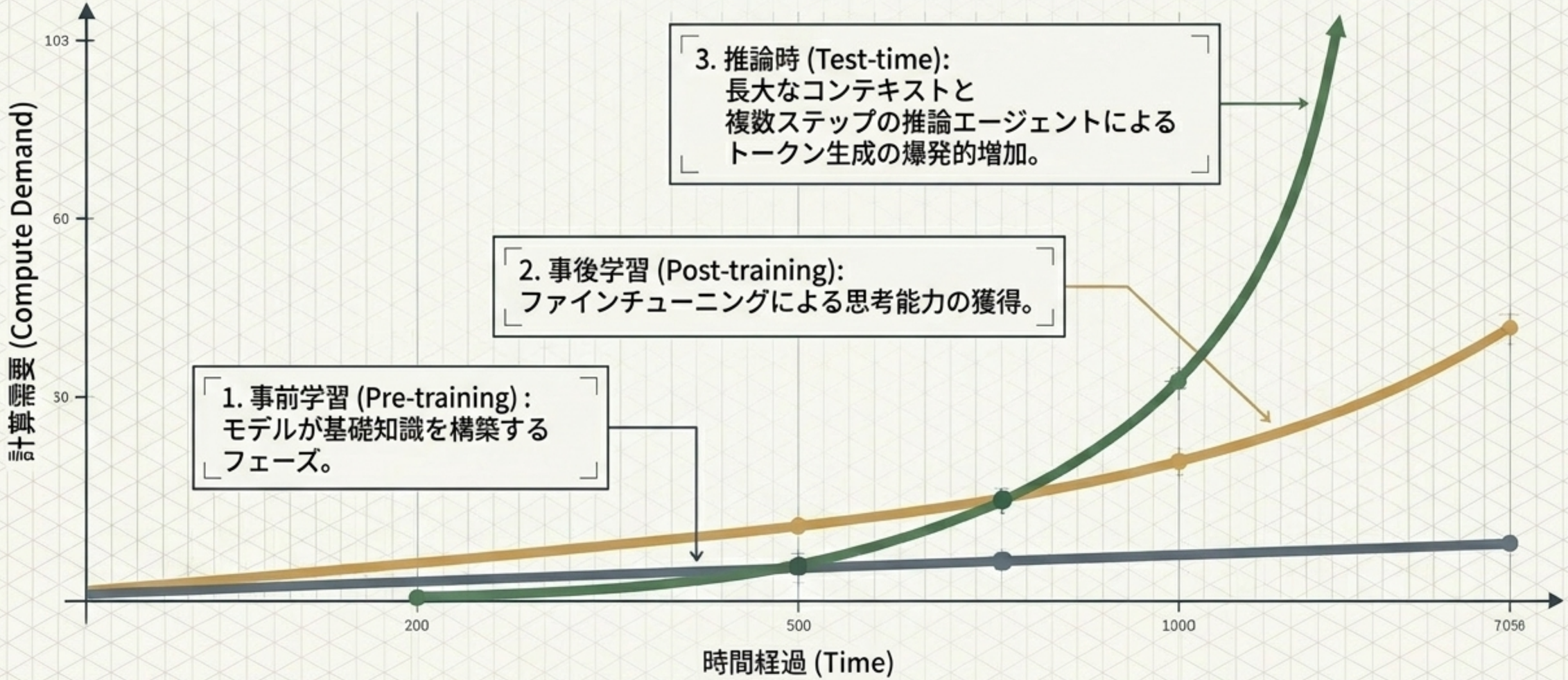
推論・エージェント・MoE（混合専門家モデル）

出力

知能（トークン）

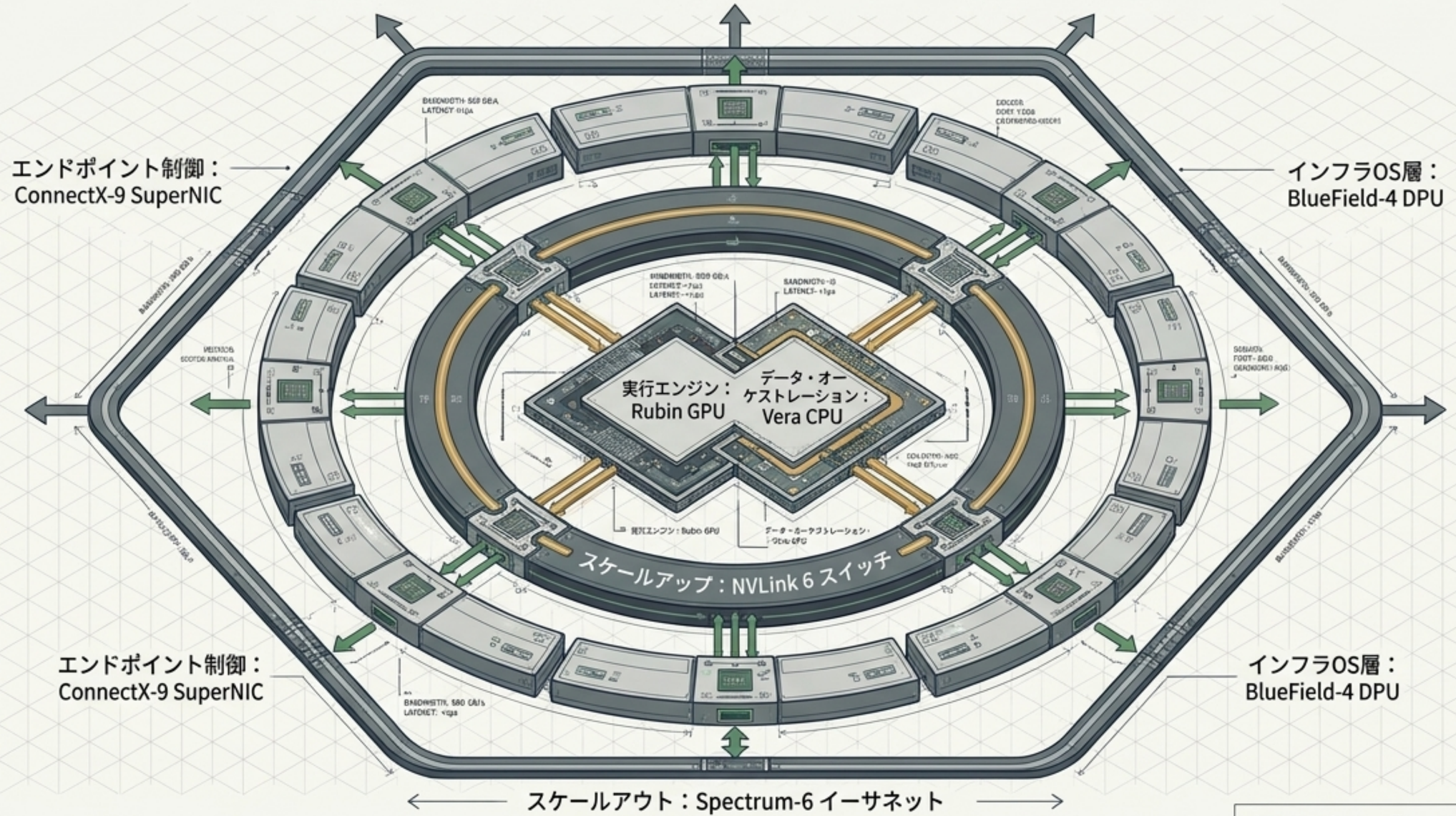
インフラの限界を押し上げる3つのスケーリング則

※推論における「低遅延」と「高スループット」のジレンマが、
新たなハードウェアアーキテクチャを要求しています。



6つのチップが統合された「単一のスーパーコンピュータ」

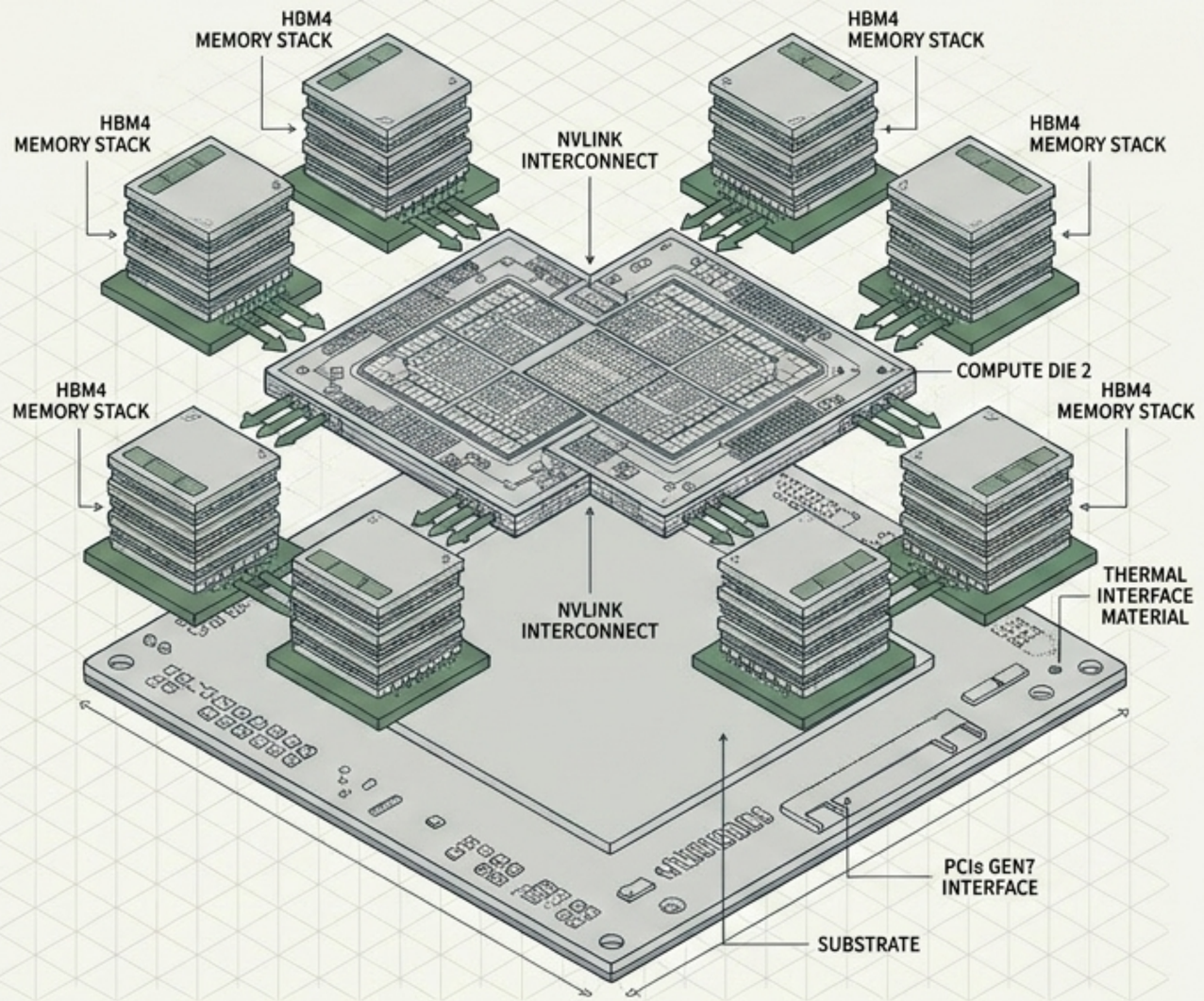
孤立したコンポーネントではなく、システム全体として協調設計（Co-design）されたプラットフォーム。



AI Factory Ecosystem Diagram

実行エンジン：NVIDIA Rubin GPU

トランスフォーマー時代のワークロードに最適化された、3,360億トランジスタを誇るシリコンの頂点。



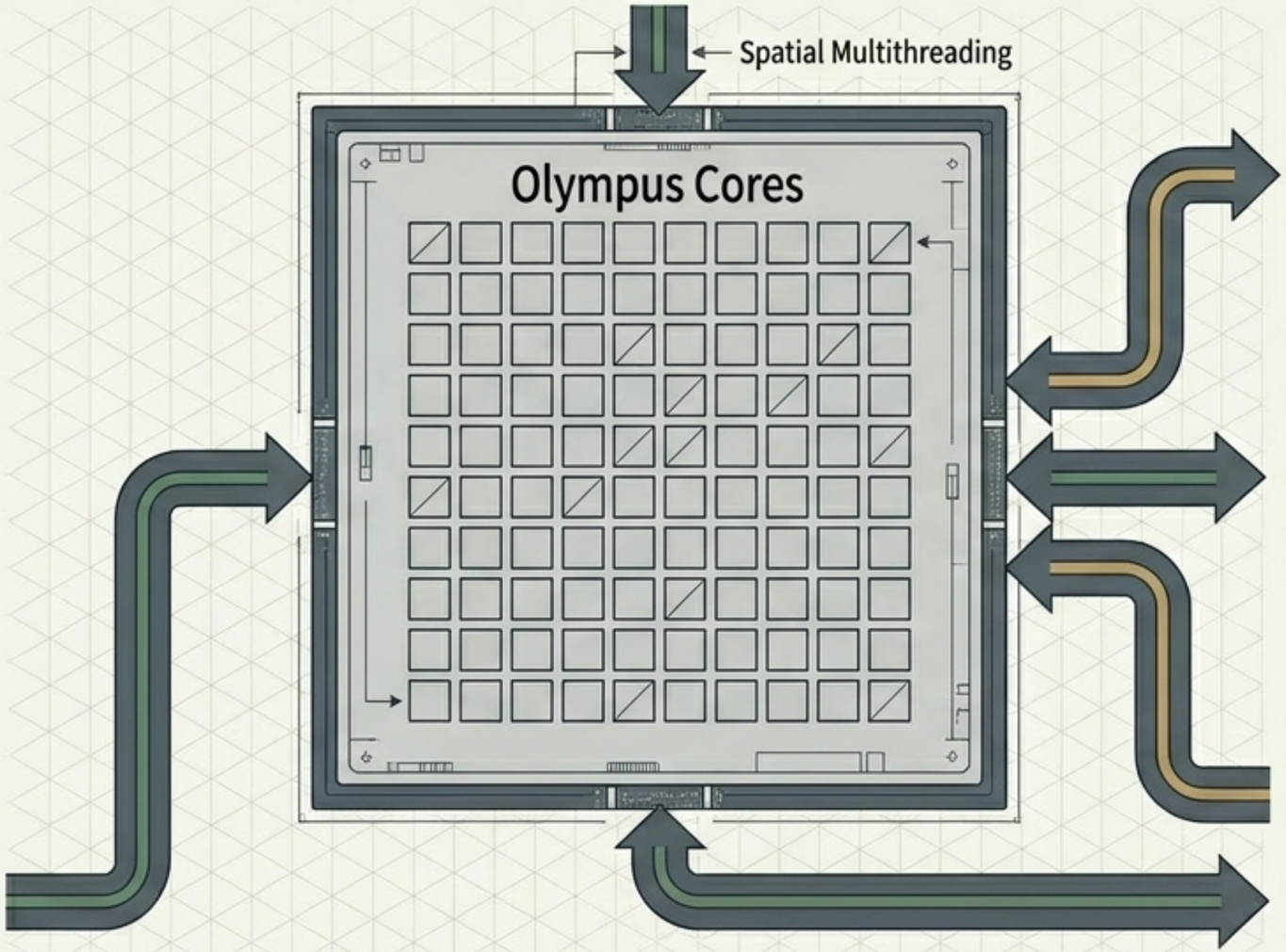
TECHNICAL SPECIFICATIONS / DIAGNOSTIC DATA

- トランジスタ数：
  3,360億
- ストリーミングマルチプロセッサ（SM）：
  224基
- メモリ：
  HBM4（最大288GB / 帯域幅22TB/s）
- 第5世代Tensorコア搭載：
  NVFP4 推論性能 50 PFLOPS / 学習性能 35 PFLOPS
- Softmaxアクセラレーションの大幅な向上

データエンジン：NVIDIA Vera CPU

GPUの稼働率を最大化するための、オーケストレーションとデータ移動の要。

※AIファクトリーのマルチテナント環境に必要なハードウェアレベルの分離と予測可能性を提供。



- コア：88基のNVIDIAカスタムOlympusコア (Arm v9.2) /
  スレッド：176 (リソースを物理的に分割するSpatial Multithreading)

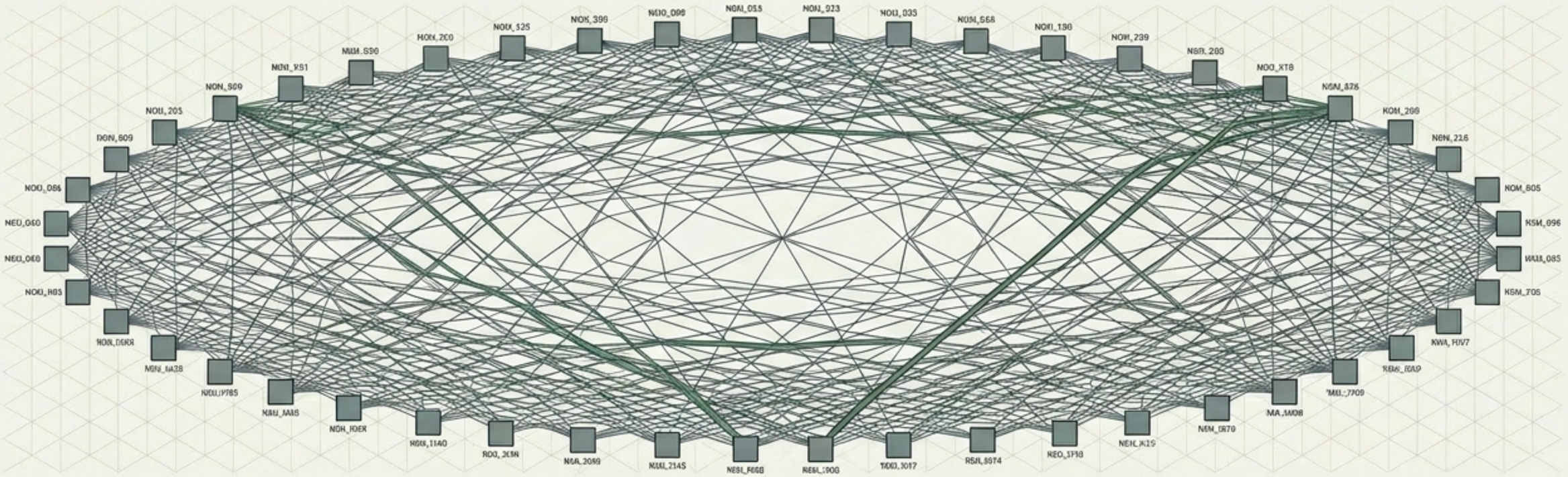
- メモリ帯域幅：最大1.2 TB/s
  (LPDDR5X, 50W未満の超高効率)

- コヒーレント帯域幅：1.8 TB/s
  (NVLink-C2CによるCPU-GPU間の
  統一メモリアドレス空間)

AI Factory Ecosystem Diagram

スケールアップ・ファブリック：NVLink 6

72基のRubin GPUを、単一の巨大なアクセラレータとして結合。
MoEモデル特有の動的なトークンルーティングのボトルネックを解消。



- GPU間双方向帯域幅：3.6 TB/s（Blackwellの2倍）

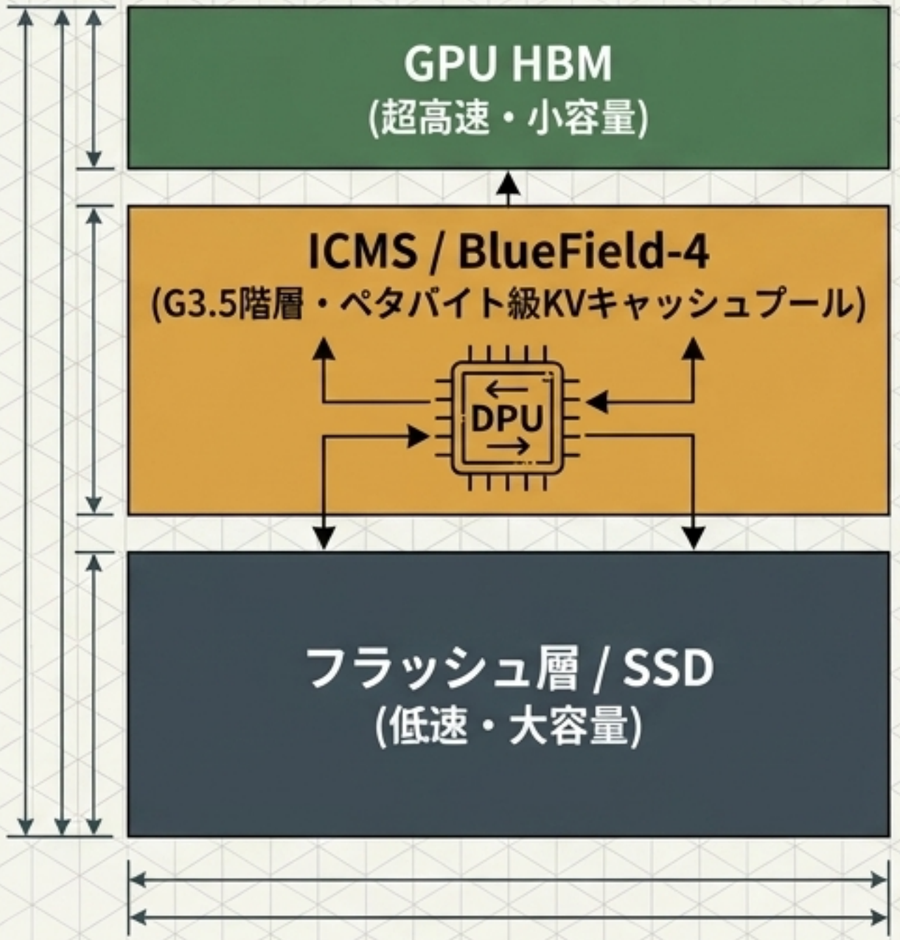
- In-Network Compute（SHARP）：スイッチ内で集合演算（All-Reduce等）を実行し、トラフィックを最大50%削減。1トレイあたり14.4 TFLOPSのFP8演算能力。

- 完全なAll-to-Allトポロジによる均一なレイテンシ。

AI Factory Ecosystem Diagram

AIファクトリーのOS層とICMS（推論コンテキストメモリ）

数百万トークンに及ぶエージェントのコンテキスト（KVキャッシュ）を効率的に管理する新機軸。

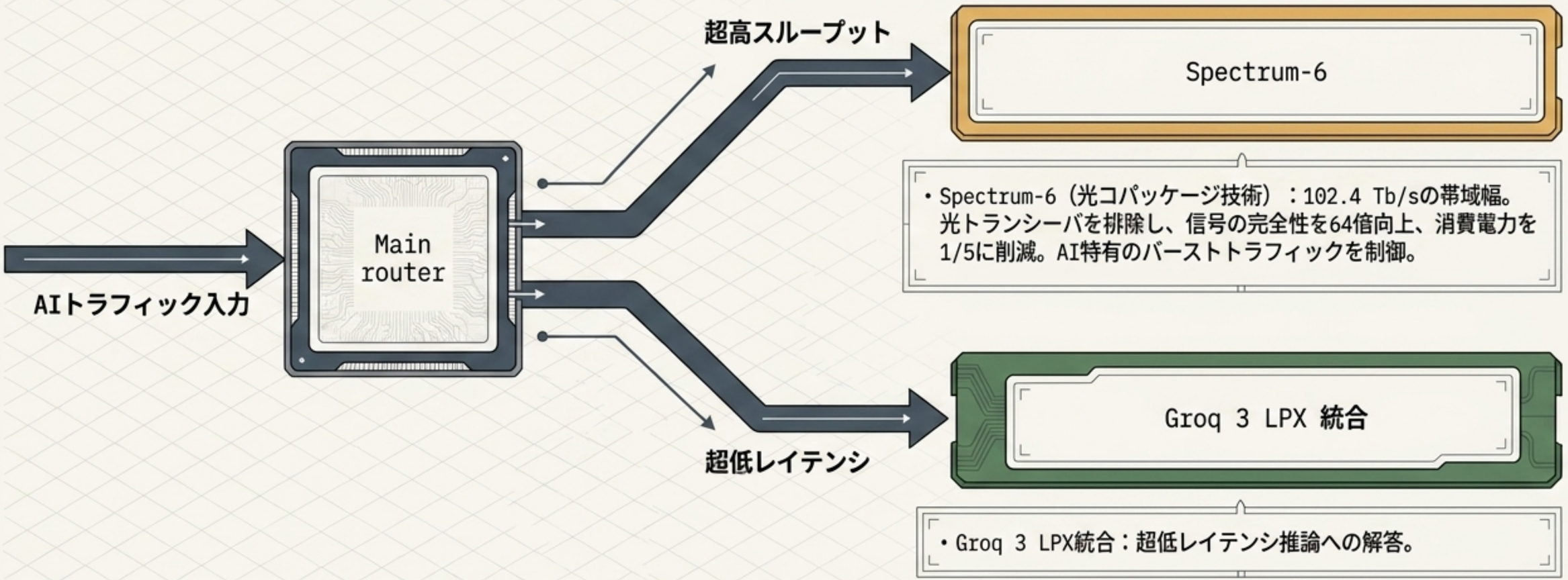


- BlueField-4 DPU：64コアのGrace CPUとConnectX-9を統合。セキュリティとネットワーク制御をホストからオフロード。
- ICMS（Inference Context Memory Storage）：「G3.5」階層として機能するEthernet接続のフラッシュ層。GPUの貴重なHBMを圧迫せず、長大なKVキャッシュを共有リソースとしてプール。トークン生成効率を最大5倍に向上。

AI Factory Ecosystem Diagram

スケールアウトと低遅延の極限：Spectrum-6 & Groq 3 LPX

「高スループット（Rubin）」と「低遅延（Groq）」という相反する要件を、非集約型推論アーキテクチャによって単一のプラットフォーム内で両立。

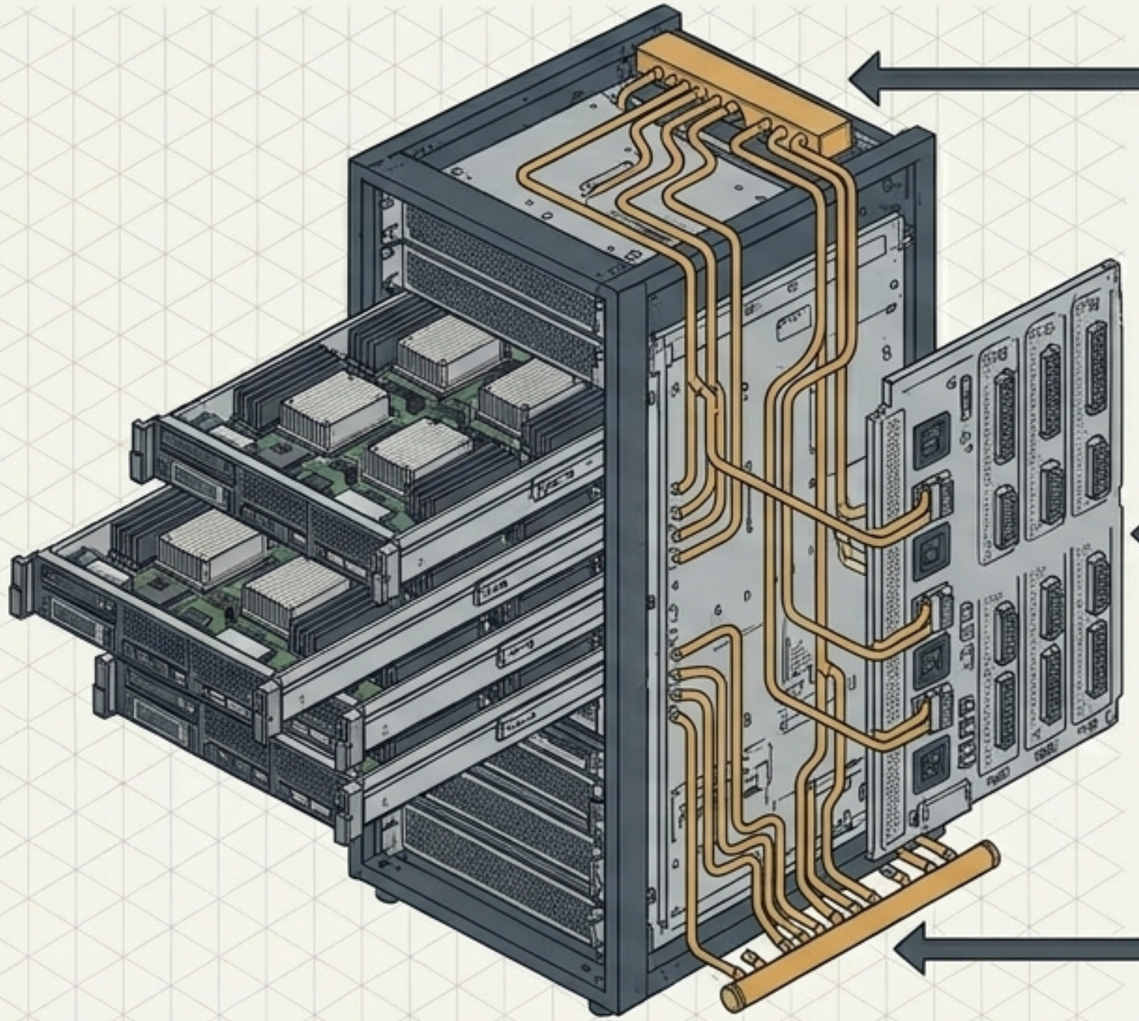


- Spectrum-6（光コパッケージ技術）：102.4 Tb/sの帯域幅。光トランシーバを排除し、信号の完全性を64倍向上、消費電力を1/5に削減。AI特有のバーストトラフィックを制御。

- Groq 3 LPX統合：超低レイテンシ推論への解答。

演算の物理単位：Vera Rubin NVL72 コンピュートトレイ

シリコンの革新を、展開可能で保守性の高い工場ユニットへと変換。



- 構成：36基のVera CPU + 72基のRubin GPU
- モジュール式ケーブルフリー設計：組み立て時間を1.5時間（Blackwell）から約5分へと劇的に短縮（18倍の高速化）。
- 液冷システム：45°Cの温水による単相直接液冷（D2C）システム。
- DGX SuperPOD：このNVL72を8ラック連結し、数万GPU規模の展開の基礎単位とする。

世代間進化マトリクス：エンタープライズ・アーキテクチャ

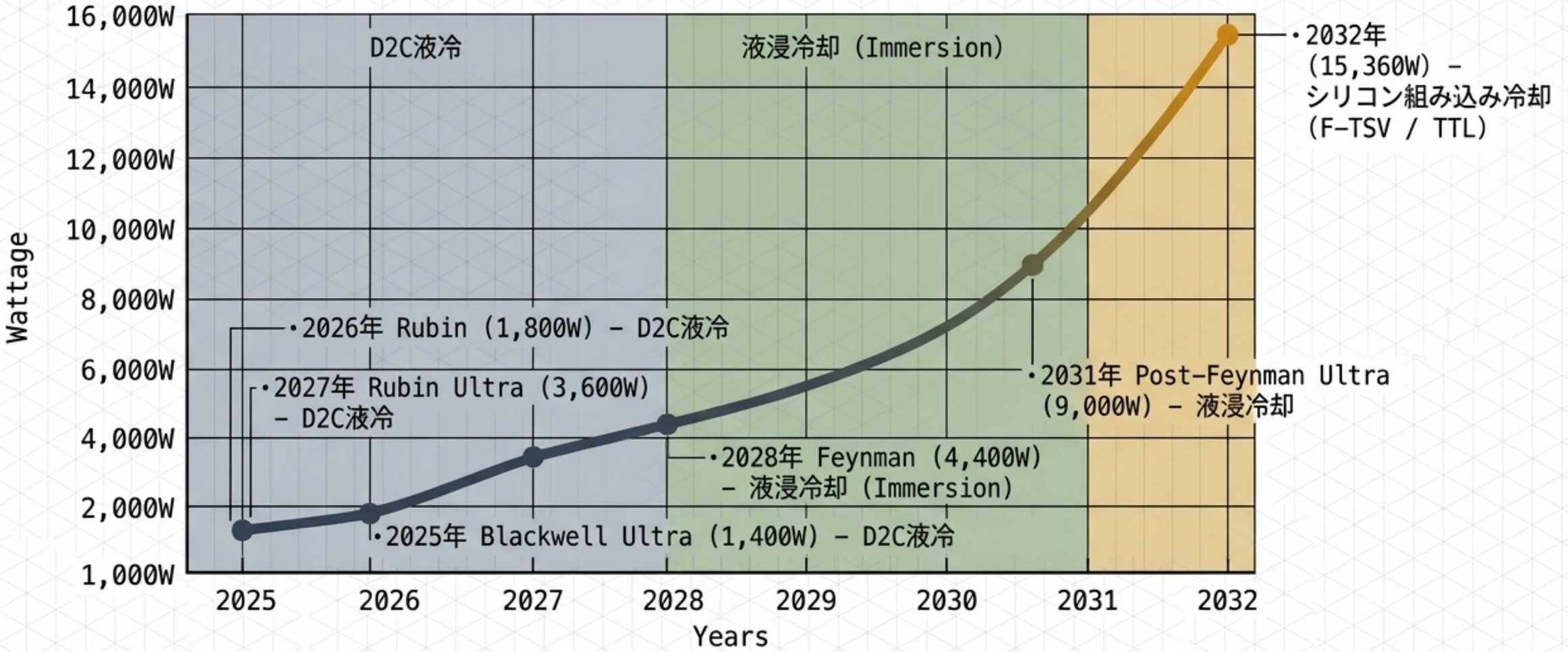
単なる演算性能の向上だけでなく、帯域幅と精度エミュレーションの劇的な飛躍。

※ * はTensorコアによるオザキスキームを用いた高精度エミュレーション時の性能。

仕様	Hopper	Blackwell	Rubin
トランジスタ数	800億	2,080億	3,360億
FP32 ベクター（TFLOPS）	67	80	130
FP64 マトリクス（TFLOPS）	67	150*	200*
NVFP4 推論（PFLOPS）	N/A	10	50
スケールアップ帯域幅	900 GB/s	1.8 TB/s	3.6 TB/s

熱力学の限界点：AIプロセッサのサーマル・ホライズン

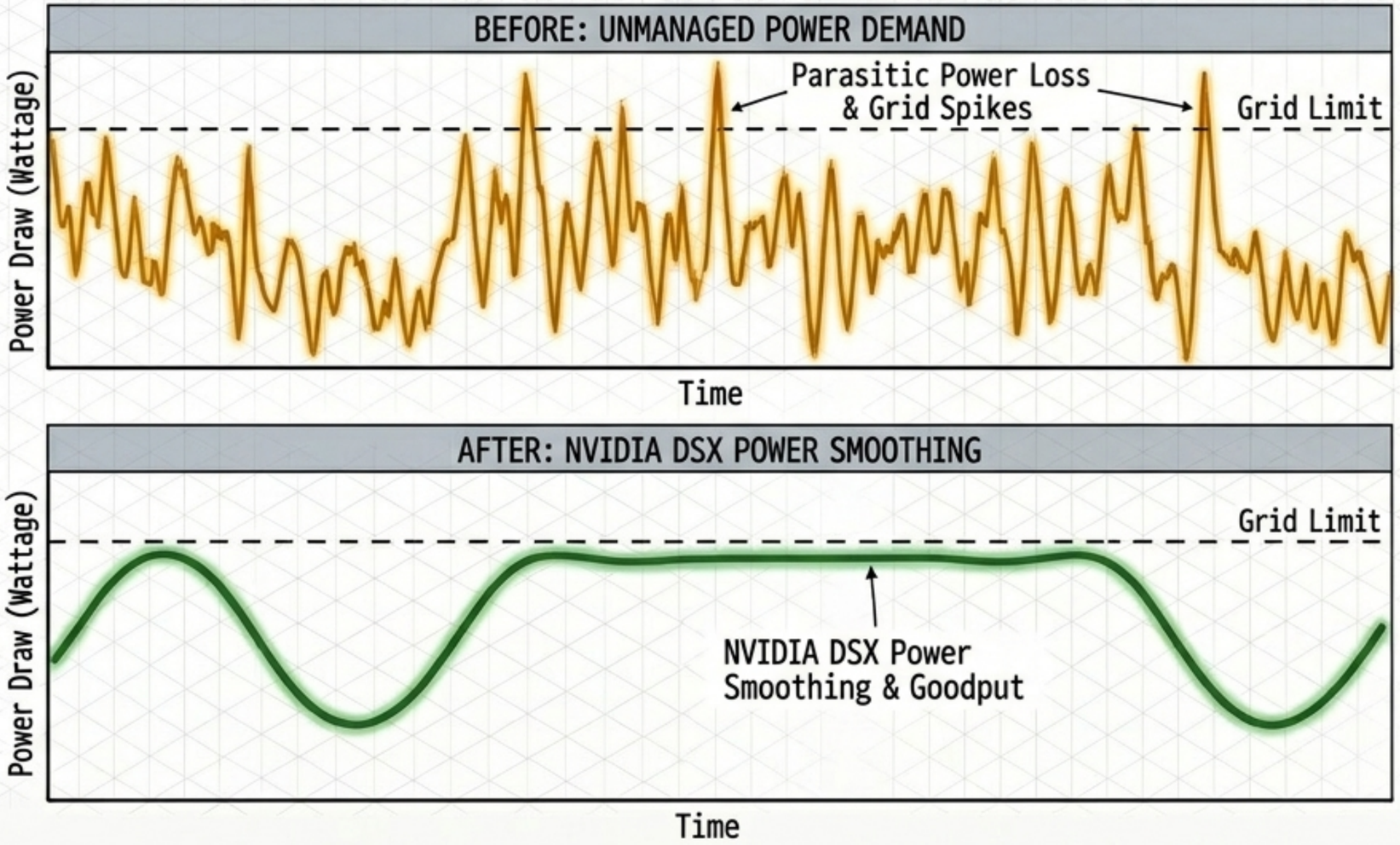
KAISTの予測に基づく、TDP（熱設計電力）の爆発的増加と冷却技術のパラダイムシフト。



AI Factory Ecosystem Diagram

寄生電力の排除と「Goodput」の最大化

同期されたAIワークロードは、電力網に致命的なスパイクを引き起こします。NVIDIAはこれをソフトウェアと局所的エネルギーバッファで平滑化します。

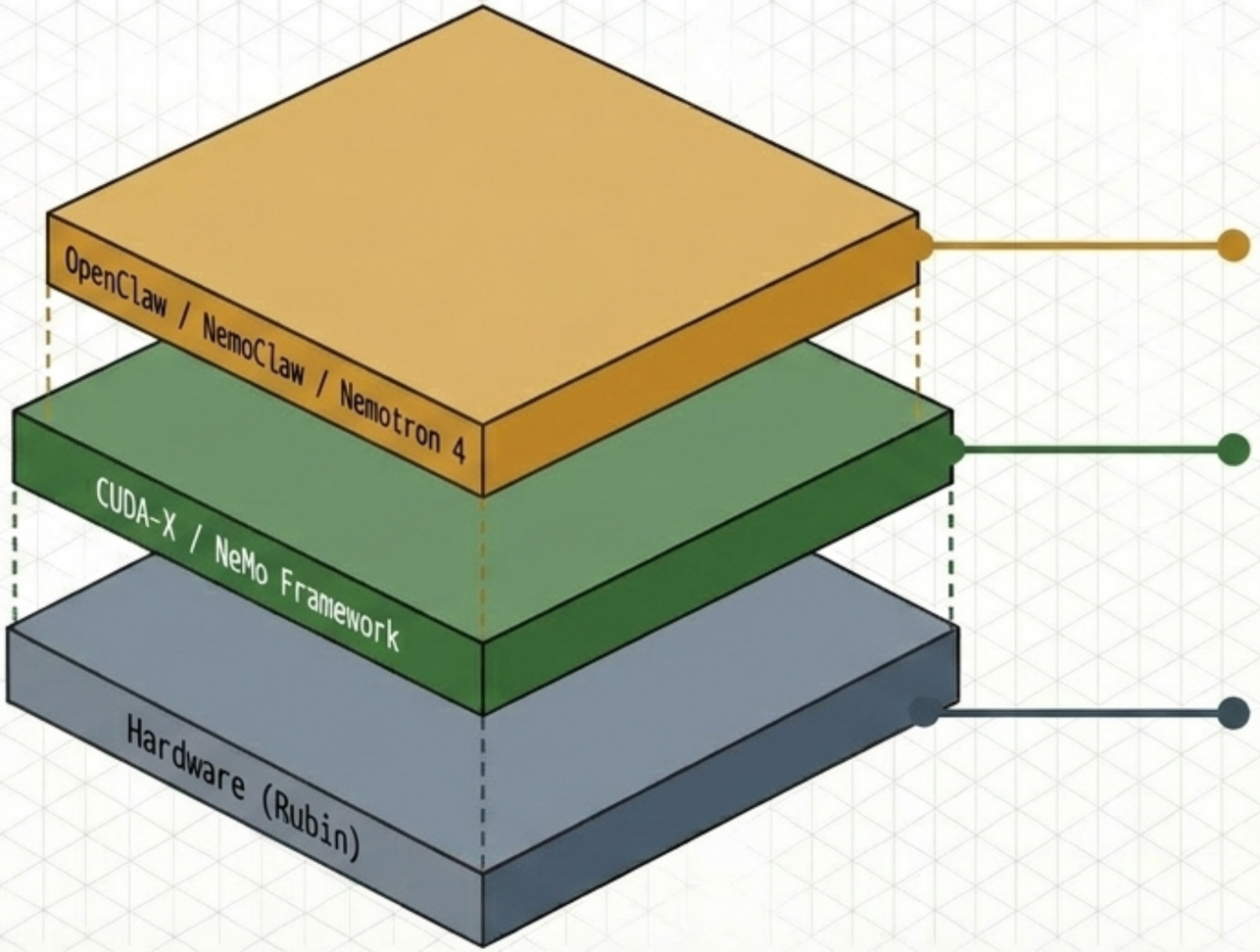


DSX SOLUTIONS & BENEFITS

- DSXソフトウェア（DSX Flex / DSX Boost）：電力網の限界を学習し、クラスタレベルで電力バジェットを動的に配分。
- Goodputの最適化：無駄な待機電力（Parasitic Power）を削減し、同じ電力枠内で最大30%多くのGPUを稼働。グリッドのエネルギーを純粋な「トークン生成」へと変換。

ソフトウェア層：CUDA誕生から20年、エージェントAIの基盤へ

ハードウェアの進化をシームレスに引き出す、NVIDIAの真の「王冠（Crown Jewel）」。

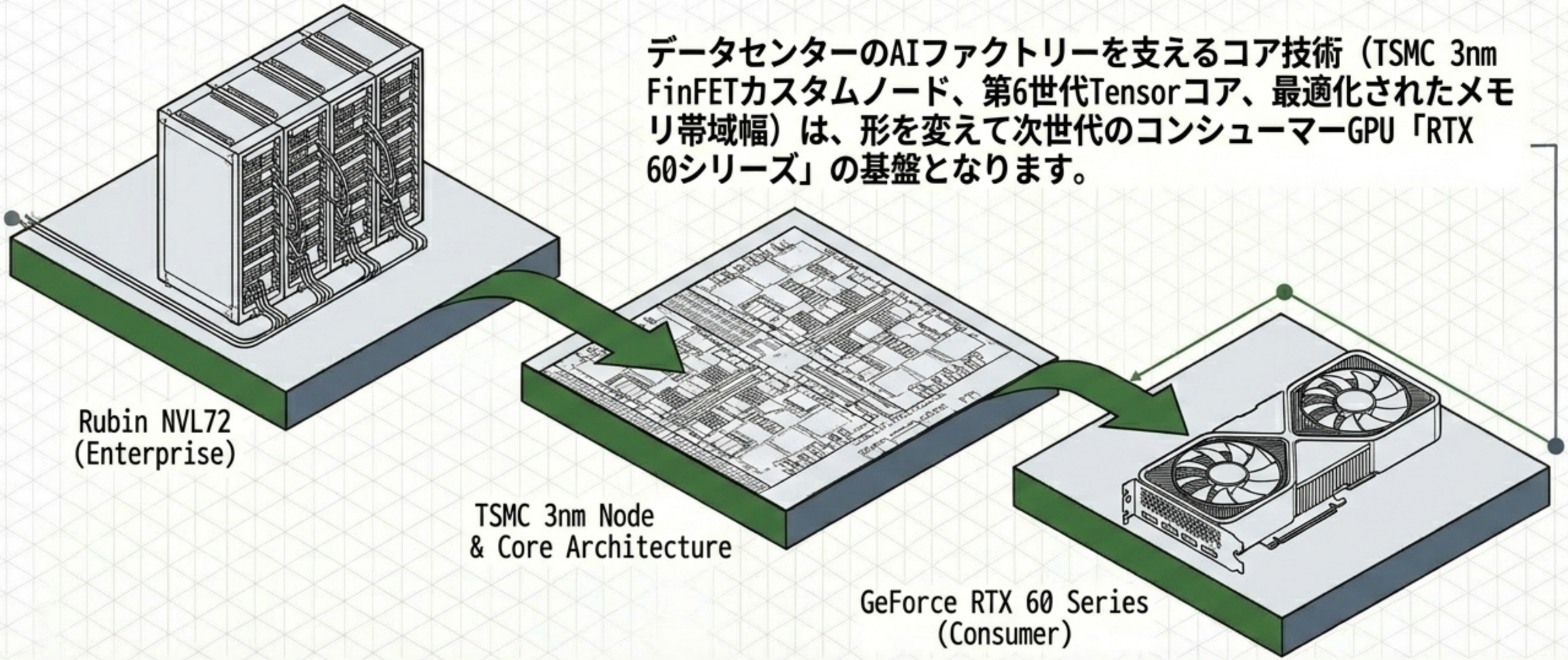


- NemoClaw：オープンソースのAIエージェントOS「OpenClaw」をエンタープライズをエンタープライズのセキュリティ基準で展開。わずか2行のシェルコマンドで自律型エージェントを構築可能。
- NeMo Framework：大規模モデルの学習・カスタマイズの統合ワークフロー。
- CUDA-X ライブラリ：cuDNN等の基盤層。ハードウェアの進化を自動的に性能向上へ変換。

エンタープライズからコンシューマーへの技術波及

「GeForceはNVIDIA最大のマーケティングキャンペーンである」 - Jensen Huang

データセンターのAIファクトリーを支えるコア技術（TSMC 3nm FinFETカスタムノード、第6世代Tensorコア、最適化されたメモリ帯域幅）は、形を変えて次世代のコンシューマーGPU「RTX 60シリーズ」の基盤となります。



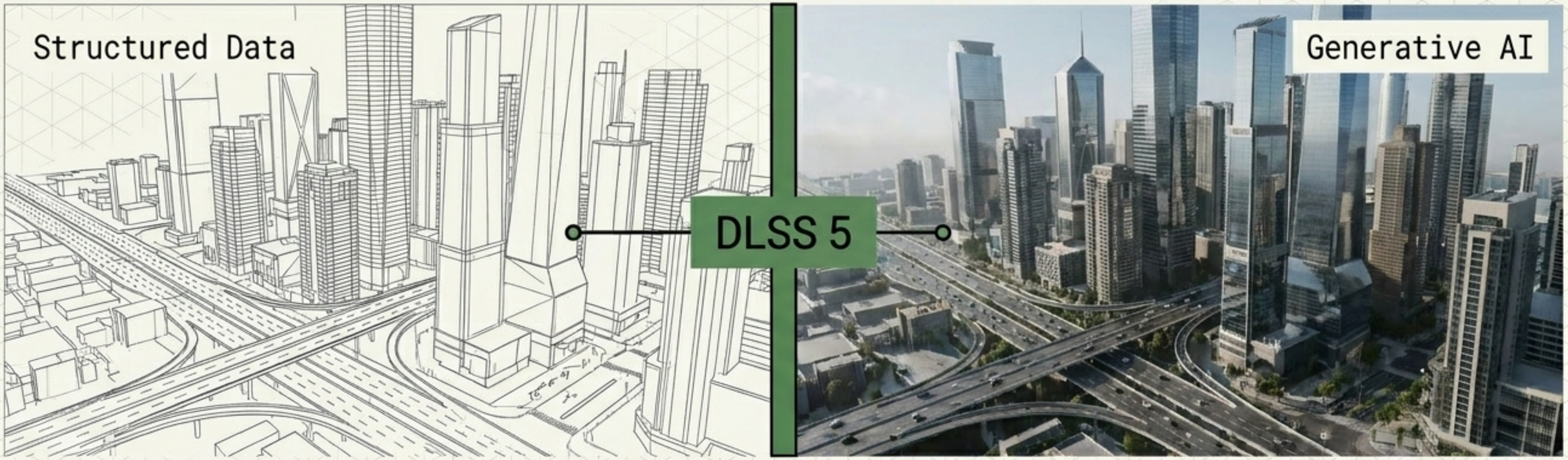
次世代ゲーミングGPU：RTX 60シリーズのリーク仕様と予測

Rubinアーキテクチャの消費者向け展開（GR20Xチップ群）。純粋なラスタライズ性能でRTX 50シリーズ比30〜35%の向上が見込まれる。

モデル	GPUコア	SM数	バス幅	VRAM
RTX 6090	GR202	192	512-bit	32GB GDDR7
RTX 6080	GR203	96	320-bit	20GB GDDR7
RTX 6070	GR205	60	256-bit	16GB GDDR7
RTX 6060	GR206	36	128-bit	12GB GDDR7
RTX 6050	GR207	24	96-bit	9GB GDDR7

DLSS 5：リアルタイムレンダリングの未来

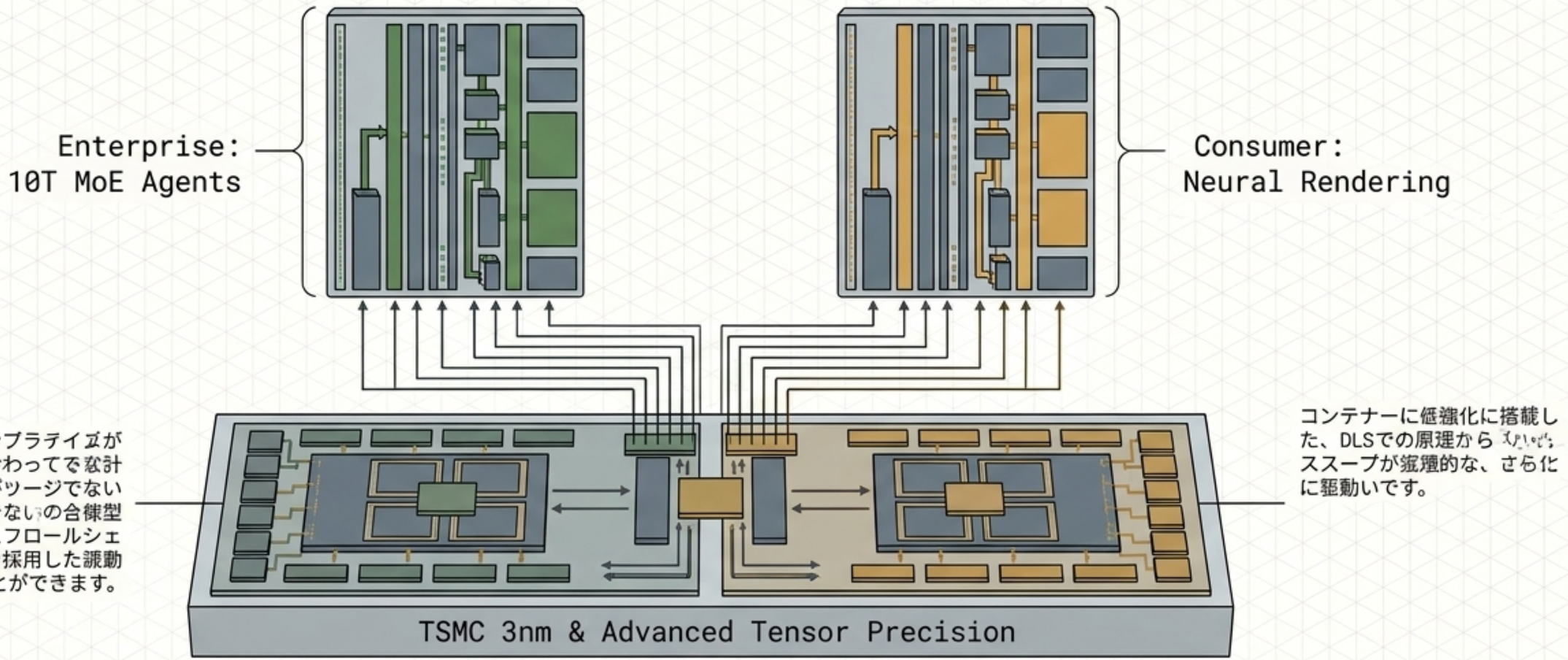
第6世代Tensorコアと第5世代RTコアが実現する、パス・トレーシング性能の2倍の飛躍。



・構造化データと生成AIの融合：	・AIによる完全な描画への移行：
ゲームエンジンからのジオメトリ・テクスチャ（構造化データ）を「プロンプト」として扱い、ニューラルレンダリングによる生成AIで写実的な映像をリアルタイムに構築。	伝統的なアルゴリズムによるシミュレーションから、機械学習ベースの生成モデルへのパラダイムシフト。

The Unified Architecture：単一のイノベーション、多様な出力

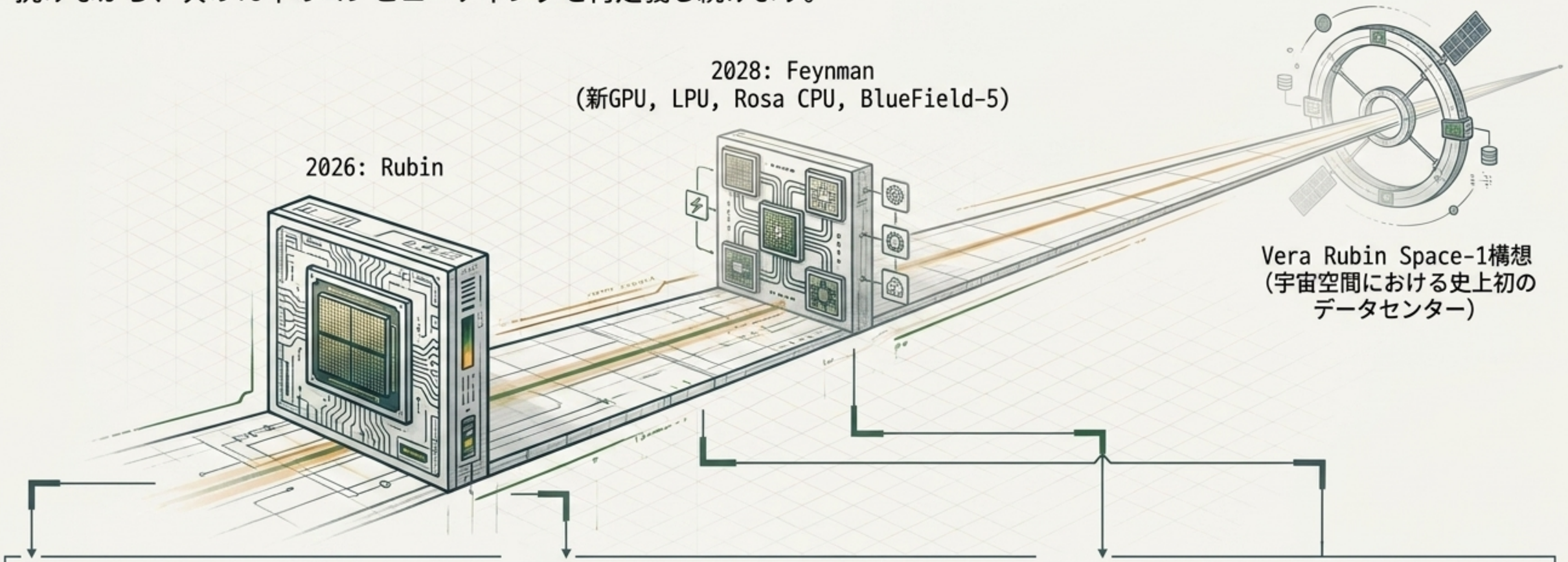
NVIDIAは、エンタープライズとコンシューマーで異なる問題を解いているのではありません。
「計算物理学のボトルネック」という単一の課題を解決し、それを異なるパッケージで提供しています。



TSMCの3nmノードと進化した低精度演算（NVFP4 / FP8）の極限の最適化が、データセンターでは「自律型AIエージェントのトークン生成」を推進し、デスクトップでは「DLSS 5による写実的レンダリング」を駆動するのです。

限界の先へ：Feynmanと宇宙空間データセンター

Rubinは到達点ではなく、新たなプラットフォームシフトの始まりに過ぎません。NVIDIAは、物理法則の限界に挑みながら、次の10年のコンピューティングを再定義し続けます。



Rubin: データセンターの性能を根本から変革する、NVIDIAの次世代アーキテクチャ。

Feynman: 計算物理学の限界を押し広げる新世代。LPU、Rosa CPU、BlueField-5を統合し、新たなコンピューティングの地平を拓く。

Vera Rubin Space-1: 地球を超えて。宇宙空間における史上初のデータセンター構想。無限の可能性への挑戦。