The Agentic Engine

自律型AIを駆動するソフトウェアとインフラストラクチャの未来

Human-in-the-loopのUX設計と、それを支える極限のコンピューティング基盤

なぜ今、自律型AIの「スタック全体」を再考すべきなのか

SOFTWARE & UX

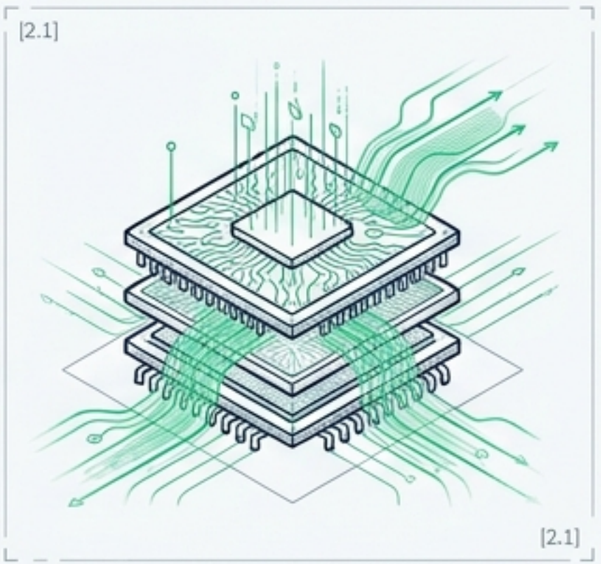
Human-in-the-loop UX

LLMエージェントの自律性が高まる中、安全性と複雑なタスク解決を両立させるため、人間をシステムの一部として組み込むUX（Magentic-UI等）が不可欠となっている。

ブレイクスルーの交差点

この2つの進化が交差するポイントに、次世代エンタープライズAIの真の価値が存在する。

INFRASTRUCTURE



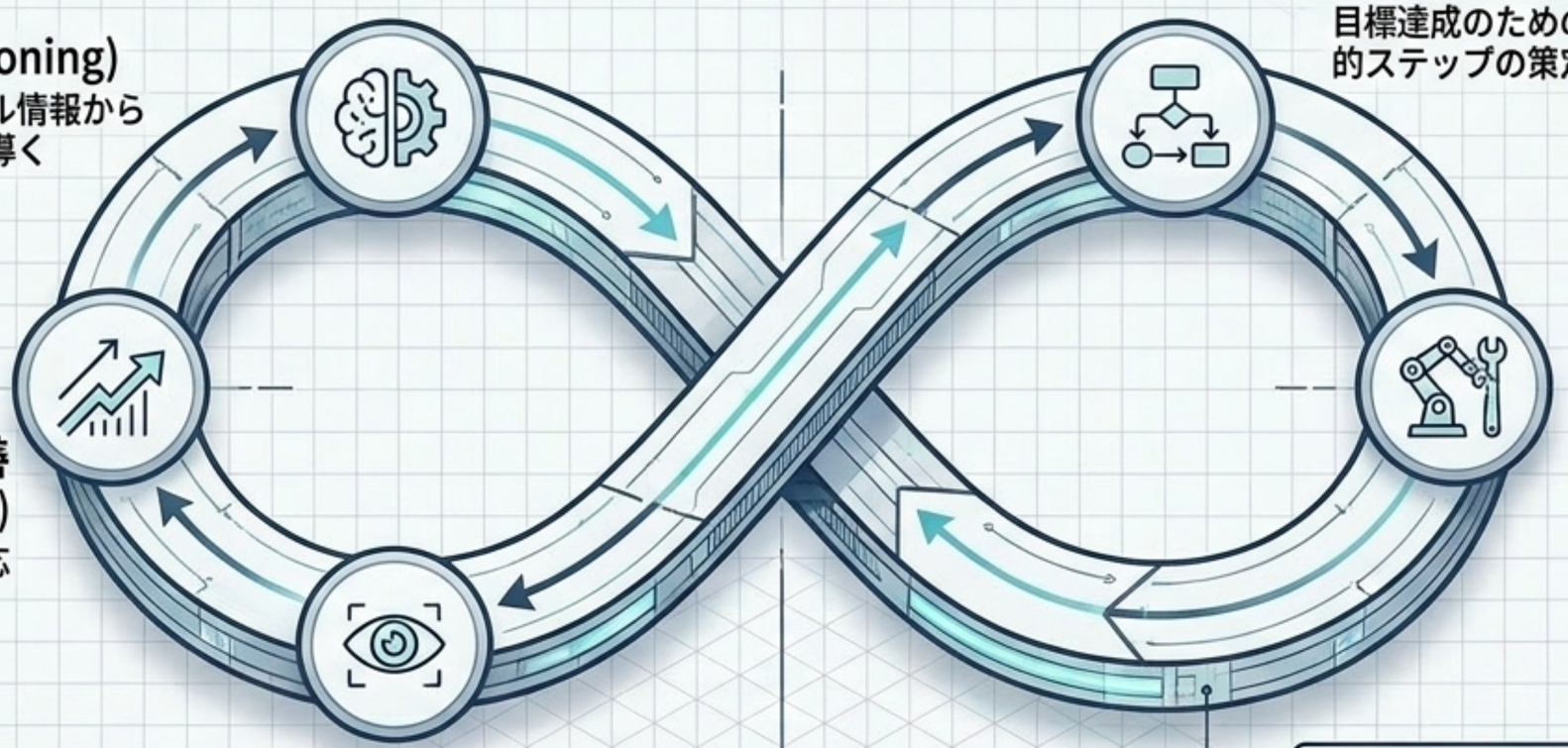
メモリ帯域幅の壁の突破

エージェントのマルチタスクと巨大コンテキストの維持により、既存のVRAMは枯渇。GDDR7やNvidia Rubinのような圧倒的なメモリ帯域幅が必須条件となる。

AIエージェントの中核機能：受動から自律的行動への進化

計画 (Planning)
目標達成のための戦略的ステップの策定

推論 (Reasoning)
マルチモーダル情報から論理的結論を導く



行動 (Acting)
ツール、API、ブラウザへの直接介入

自己改善 (Self-Improvement)
経験からの学習と適応

観察 (Observation)
環境からのフィードバックとコンテキストの収集

生成AIとマルチモーダル基盤モデルにより、システムは意思決定と適応のレベルを備えた『自律性』を獲得した。
(Source: Google Cloud)

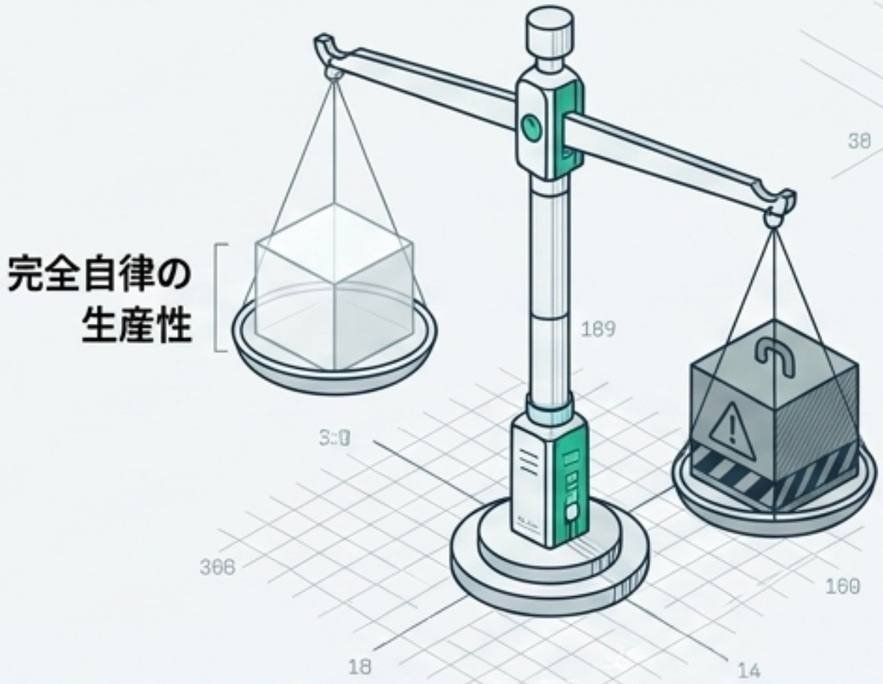
AIインタラクションのパラダイムシフト

受動的（Passive）　予防的・目標指向（Proactive）

予防的・目標指向（Proactive）

Bot	AI Assistant	AI Agent
目的 単純タスクの自動化	**目的** ユーザー作業のサポート	**目的** 自律的かつプロアクティブなタスク実行
機能 ルールベース、限定的学習	**機能** リクエストへの応答、アクションの推奨（最終決定権は人間）	**機能** 複雑な複数ステップの実行、独立した意思決定、環境への適応
インタラクション トリガーやコマンドにのみ応答	**インタラクション** ユーザーの指示に依存するリアクション型	**インタラクション** 人間に代わって目標を追求する予防型

The HITL Imperative: 人間参加型の必然性



倫理的・社会的ダイナミクスの壁

人間の微妙な感情、非言語的合図、道徳的判断を要する状況におけるAIの文脈理解の限界。

環境の不確実性とリスク

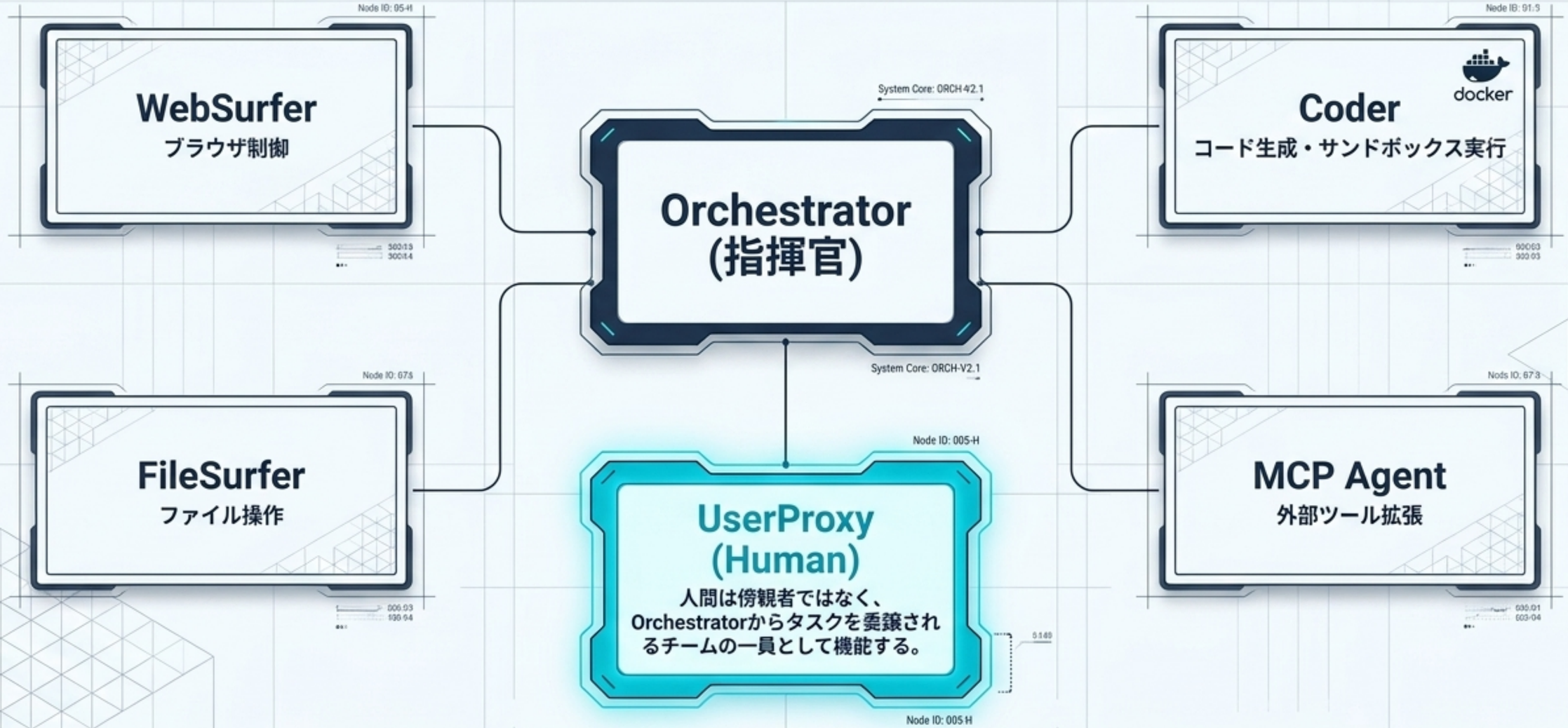
予測不可能な外部ツール操作において、不可逆な変更や意図せぬ購買など致命的エラーを回避するフェイルセーフ。

曖昧さの解消と事前知識

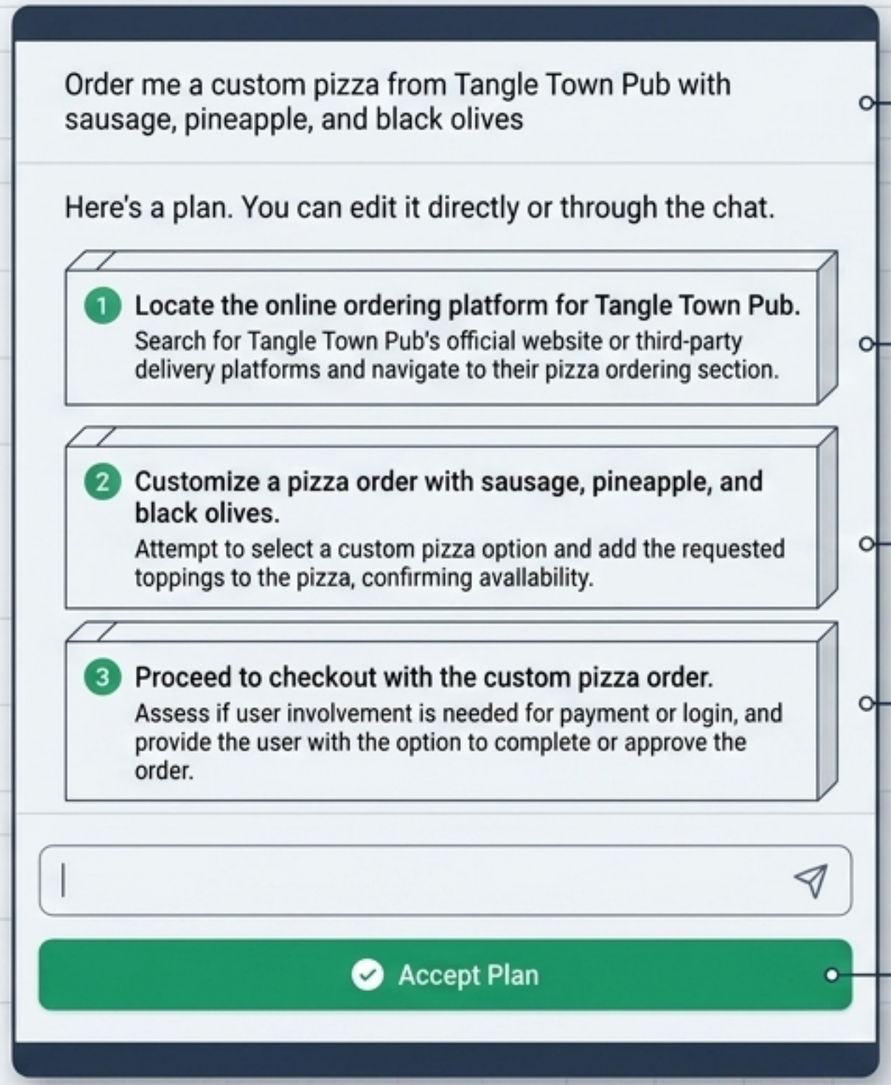
ユーザーの曖昧な指示を解釈し、タスク実行前に人間が持つ「暗黙の制約」をシステムに注入する必要性。

Magentic-UI Architecture:
人間をエージェントチームに組み込む

Magentic UI



Interaction 1: Co-planning（協調計画による目標の共有）



曖昧さの排除

タスクが不明確な場合、Orchestratorは実行前にユーザーに明確化の質問を投げる。

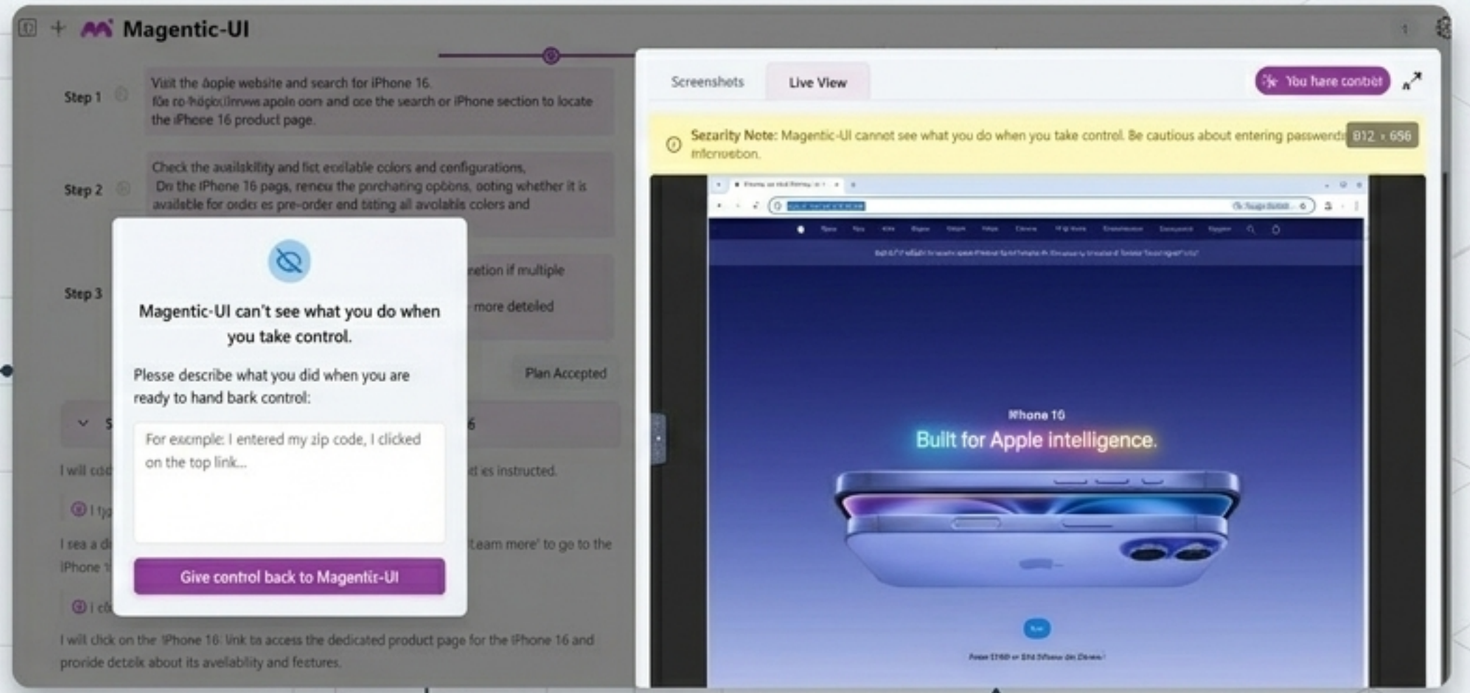
人間の事前知識の注入

ユーザーはこのUI上で直接ステップを編集し、暗黙の制約を事前に計画へ組み込む。

透明性の確保

内部表現ではなく、人間が理解できる自然言語で計画を合意してから実行フェーズへ移行する。

Interaction 2: Co-tasking（シームレスな制御のハンドオフ）



1. AI Execution

エージェントが自律的にブラウザを操作し、計画を進行。

2. Human Interruption

CAPCHAや決済情報入力のため、ユーザーが任意のタイミングで介入。

3. Dynamic Handoff

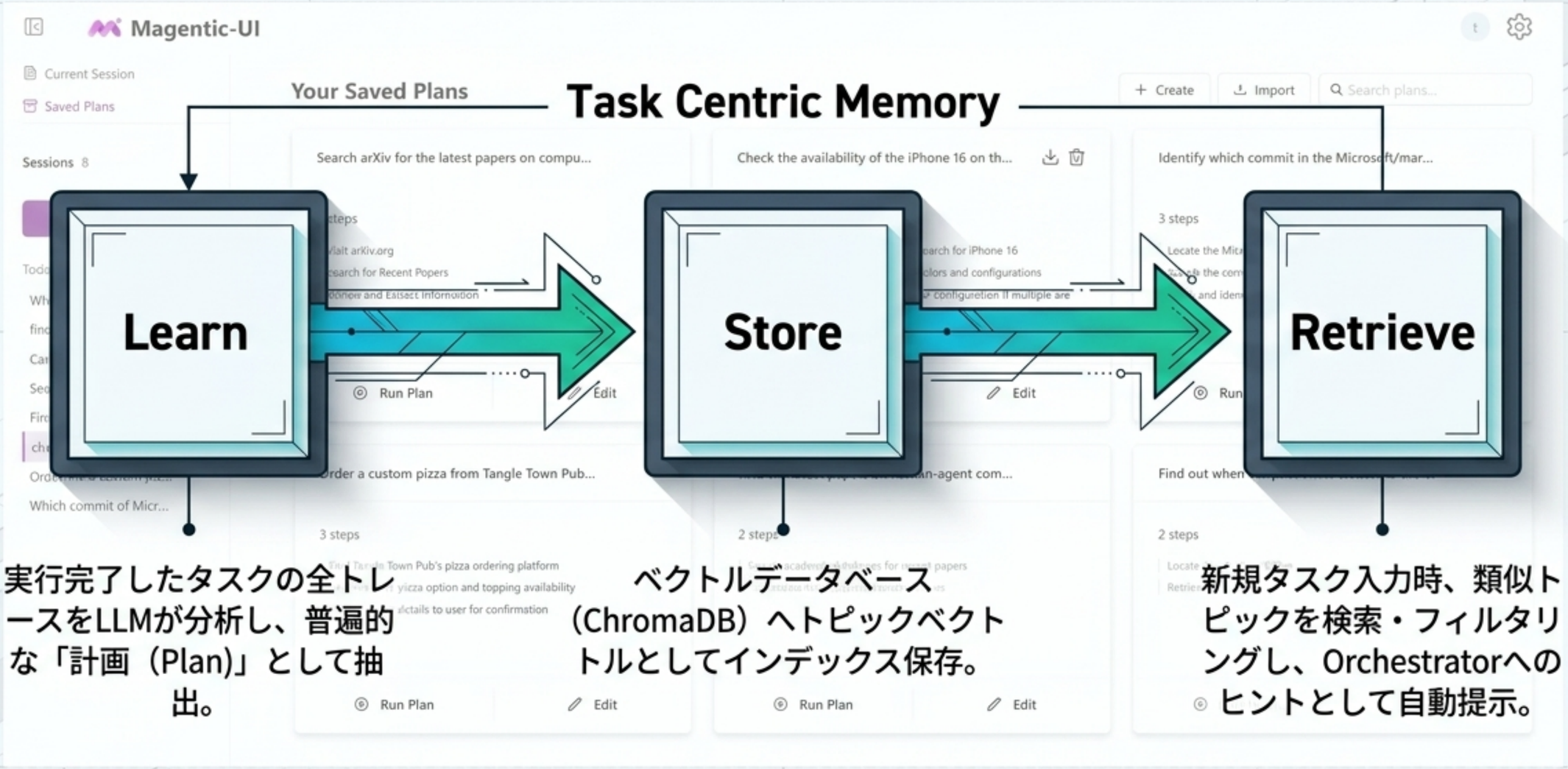
ブラウザの制御権が即座に人間に移行（Dockerサンドボックス内のライブブラウザを直接操作）。

4. Resumption

完了したアクションをチャットでエージェントに伝え、自律実行を再開。

Architectural Blueprint of Autonomy

Interaction 3: Agent Memory（経験の学習と再利用）



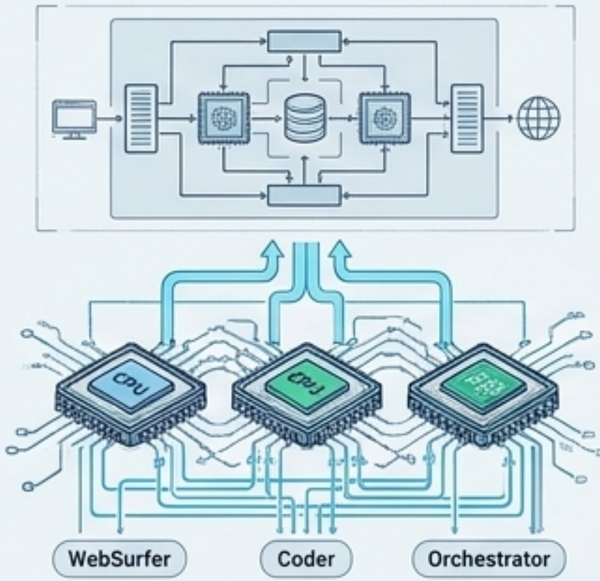
実行完了したタスクの全トレースをLLMが分析し、普遍的な「計画（Plan)」として抽出。

ベクトルデータベース（ChromaDB）へトピックベクトルとしてインデックス保存。

新規タスク入力時、類似トピックを検索・フィルタリングし、Orchestratorへのヒントとして自動提示。

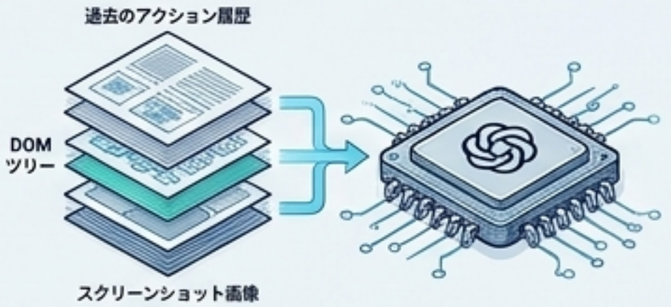
The Hardware Wall: エージェントが直面するインフラの限界

Concurrent Execution



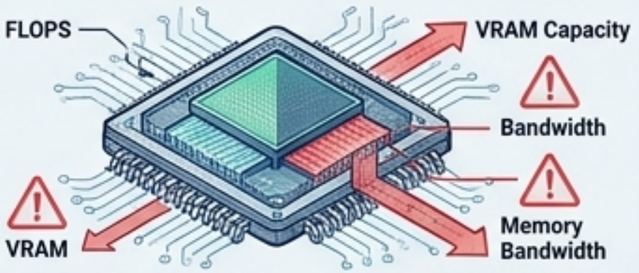
複数の特化型モデル（WebSurfer, Coder, Orchestrator）の同時稼働による演算リソースの占有。

Massive Context Windows

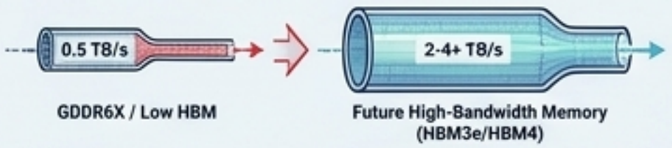


過去のアクション履歴、DOMツリー、スクリーンショット画像を毎ターンLLMに入力し続ける負荷。

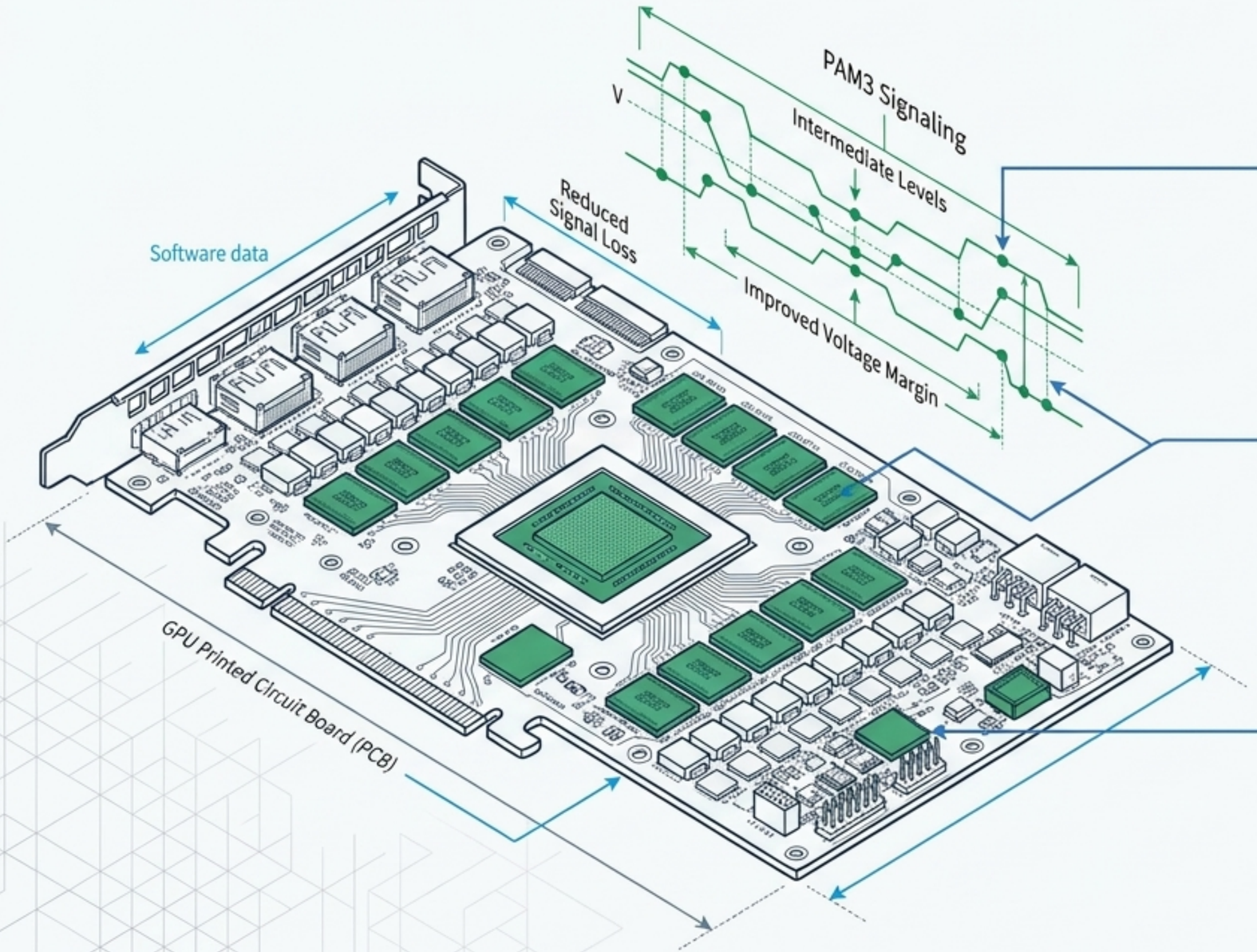
The Bandwidth Crisis



これらの処理は計算能力（FLOPS）以上に「狂気的なメモリ帯域幅」と「VRAM容量」を消費する。従来のGDDR6Xや少量のHBMでは、推論時のデータ転送がボトルネックとなり、エージェントのリアルタイム応答性が完全に損なわれる。



Workstation Breakthrough: GDDR7による帯域幅の解放



PAM3 Signaling

NRZとPAM4の中間方式を採用。信号損失を低減しつつ、高周波数帯域における電圧マージンを劇的に改善。

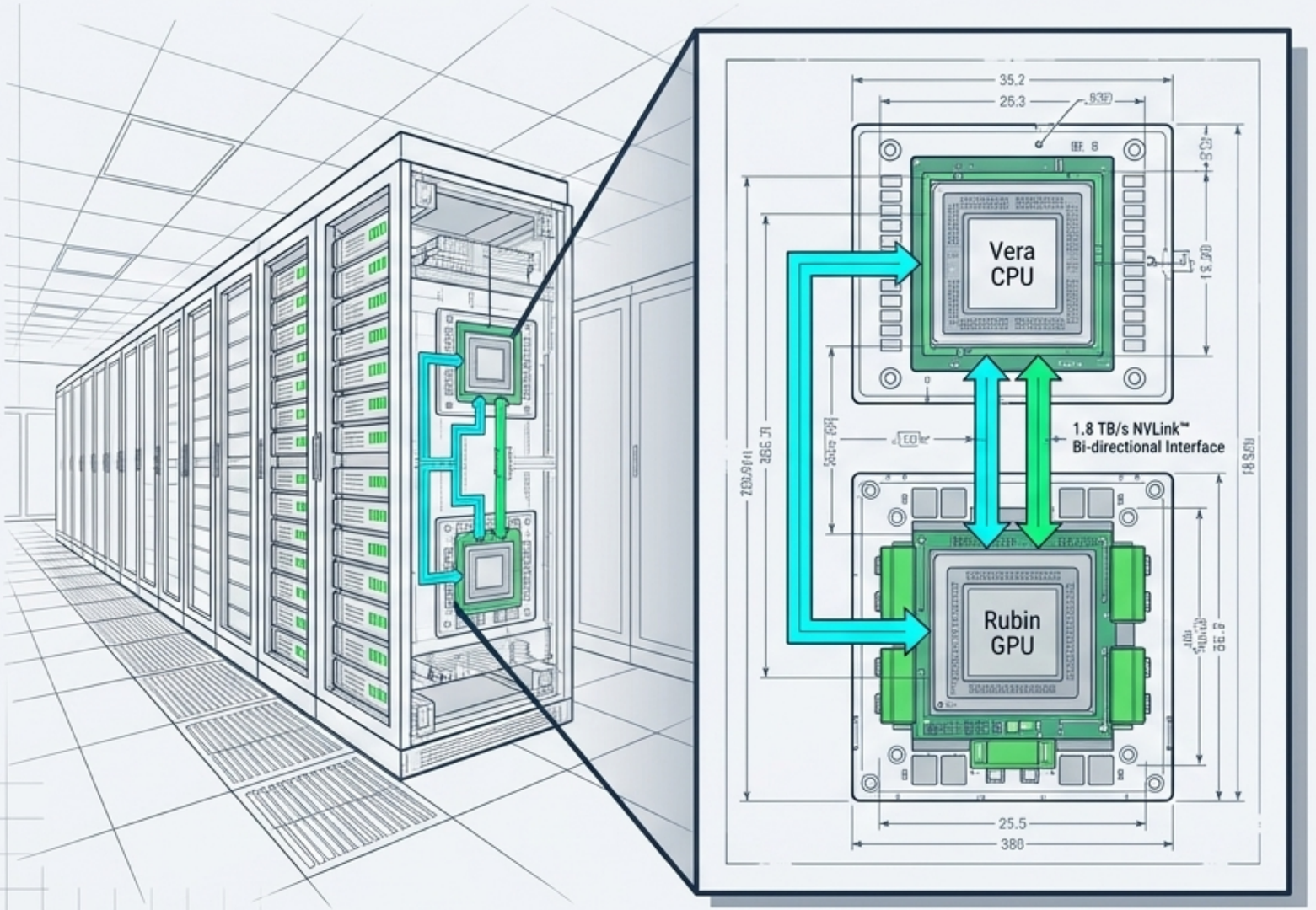
Insane Bandwidth

ピンあたり32〜48 Gb/sに到達。システム全体で 1.5 TB/sの帯域幅を超え、前世代比約60%〜70%のパフォーマンス向上。

Capacity Jump

24Gbモジュールの標準化により、メインストリームで24GB、ハイエンドで36〜48GBのVRAM搭載が可能に。エージェントの巨大コンテキストウィンドウ増大に直接対応。

Data Center Breakthrough: Nvidia Rubin Architecture



Vera CPU

88コアのカスタムARMと176スレッド。GPUとの間に1.8 TB/sの双方向NVLink接続を確保し、CPU-GPU間のボトルネックを解消。

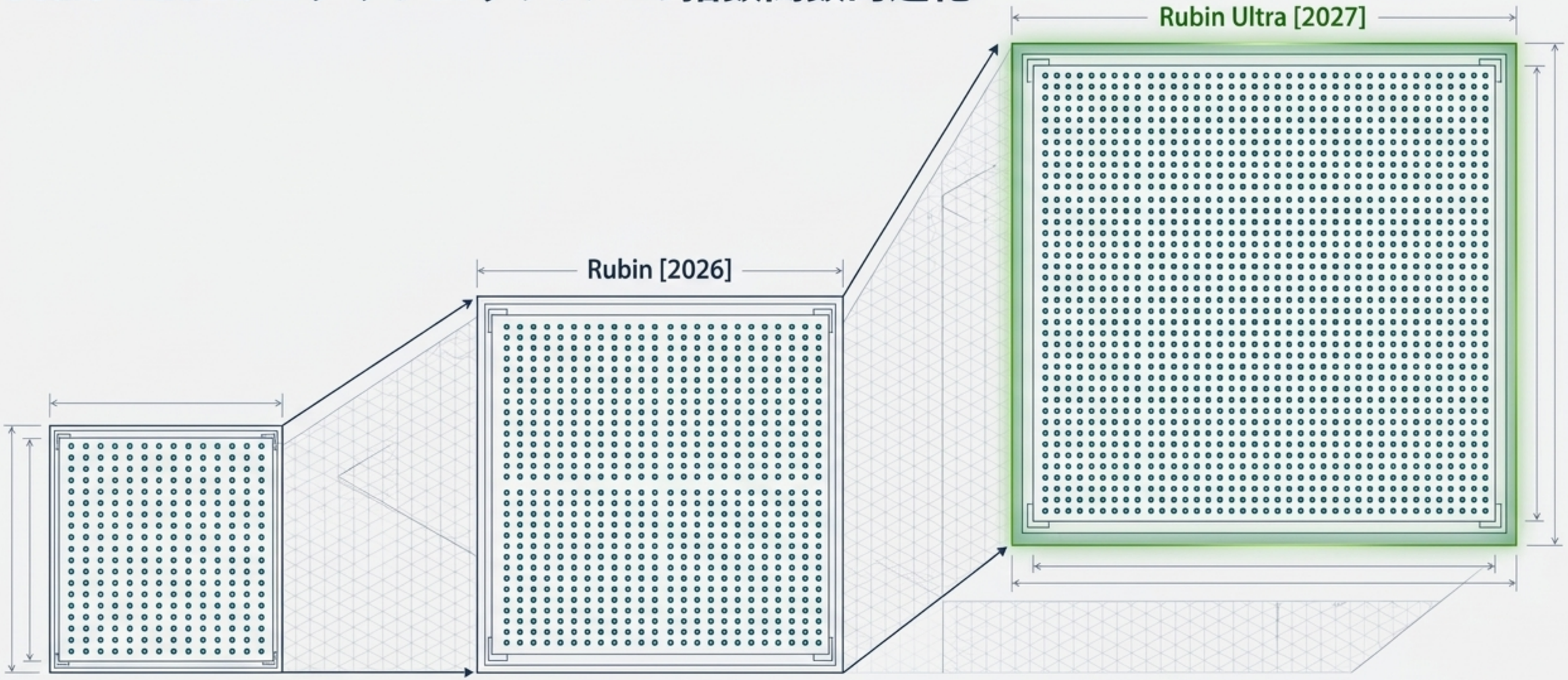
HBM4 / HBM4e Transition

B300のHBM3e（8 TB/s）から進化。RubinではHBM4、次世代のRubin UltraではHBM4eを採用し、各GPUで13 TB/sの驚異的なスループットを実現。

FP4 Dense Compute

大規模な推論タスクやマルチエージェントの並列処理に最適化されたFP4演算能力を大幅に強化し、データセンターのワットパフォーマンスを再定義。

Scale Matrix: エクサフロップスへの指数関数的進化

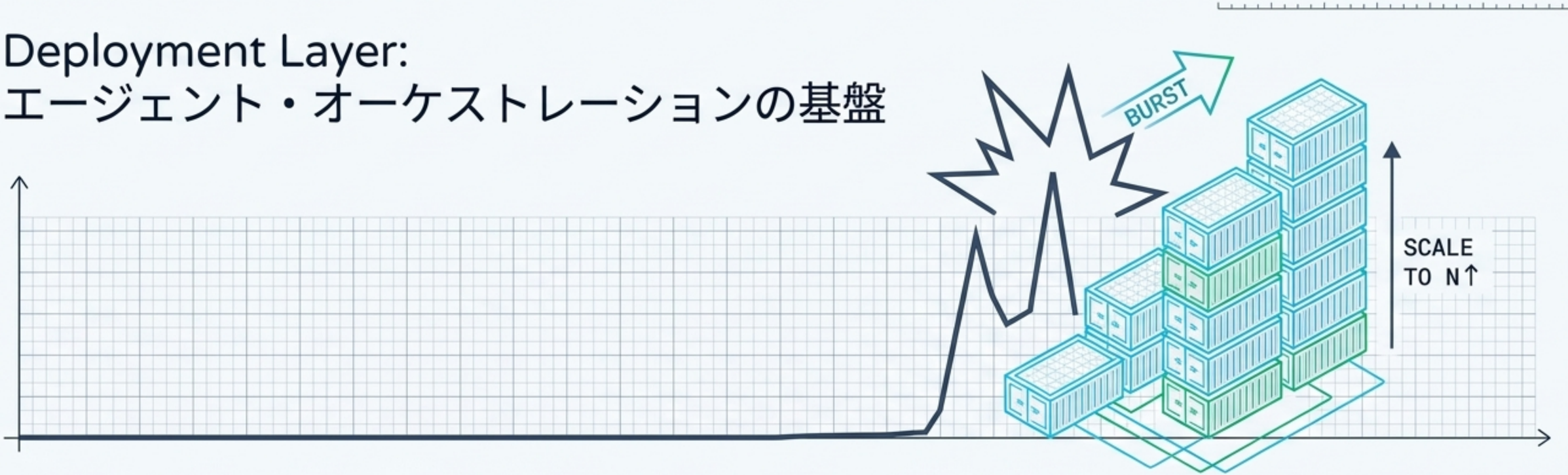


Blackwell Ultra (B300) [2025]
Layout: NVL72 (72 GPUs/Rack)
Compute: 1.1 EFLOPS (Dense FP4)

Rubin [2026]
Layout: NVL144 (144 GPUs/Rack - 2 die per GPU)
Compute: 3.6 EFLOPS (B300の約3.3倍)

Rubin Ultra [2027]
Layout: NVL576 (576 GPUs/Rack - 4 die per package)
Compute: 15 EFLOPS / 365TB Rack Memory
1.5 PB/s NVLink7 throughput

Deployment Layer:
エージェント・オーケストレーションの基盤

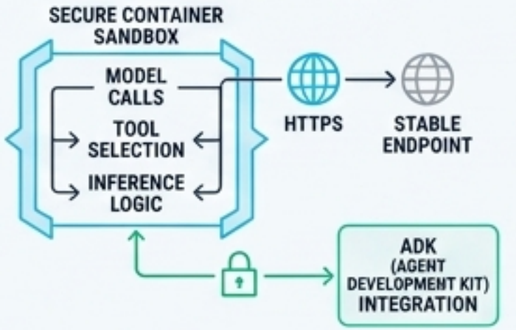


Serverless Efficiency (ゼロスケールダウン)

目標指向のエージェントは推論・計画時に突発的で重いトラフィックを発生させる。Cloud Runはアイドル時にゼロにスケールダウンし、実行中のリソースにのみ課金されるため、コスト効率が極めて高い。

HTTPS Endpoints & Sandboxing

モデル呼び出し、ツール選択、推論プロセスを管理するコアロジックをセキュアなコンテナに封じ込め、安定したエンドポイントを提供。
Agent Development Kit (ADK) とシームレスに統合。



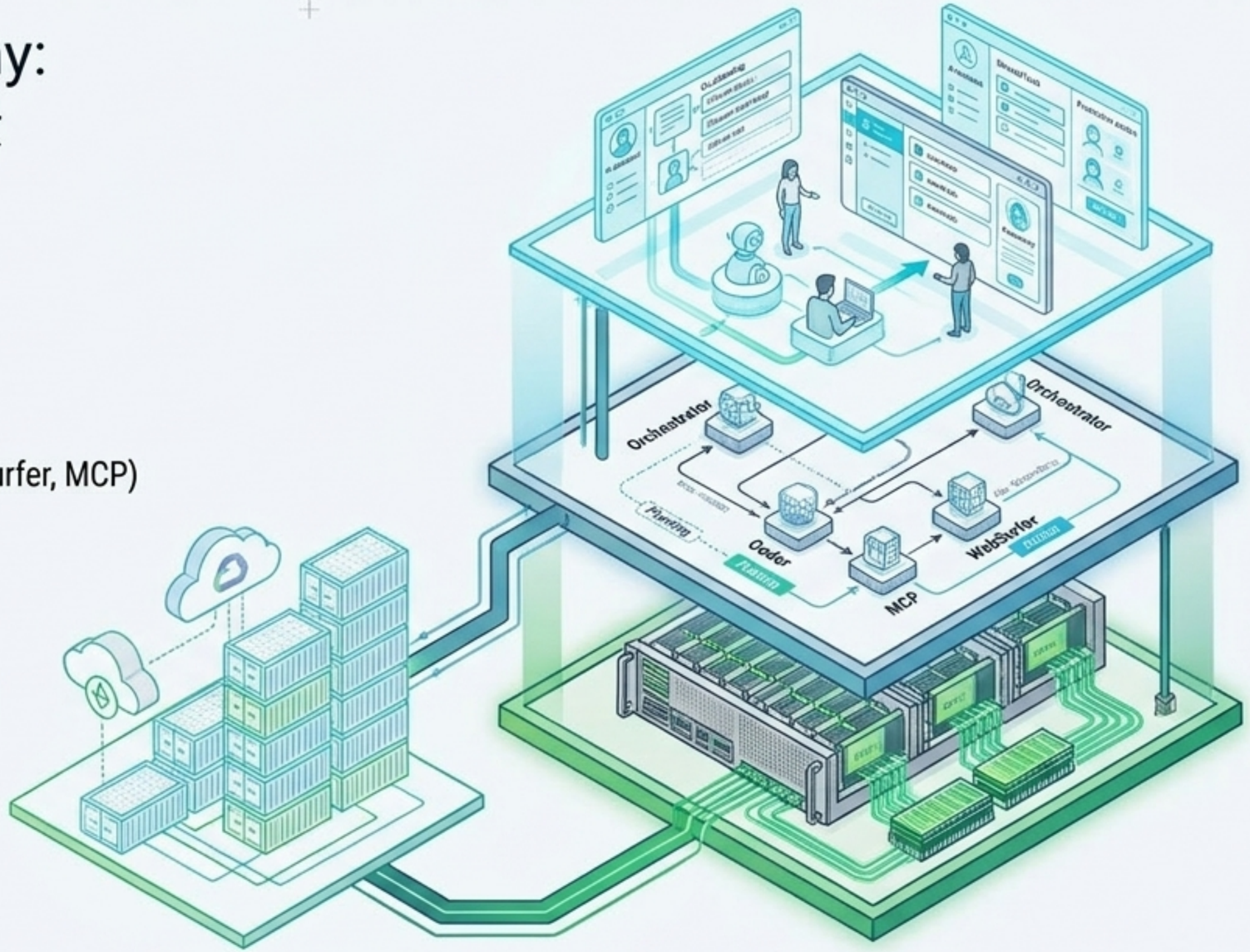
The Blueprint of Autonomy:
自律型AIフルスタック全景

Layer 4: Interaction & UX
Magentic-UI [Co-planning, Co-tasking,
Action Guardsによる人間参加型制御]

Layer 3: Agent Orchestration
Multi-Agent Logic (Orchestrator, Coder, WebSurfer, MCP)
[自律的な推論・計画・行動ループ]

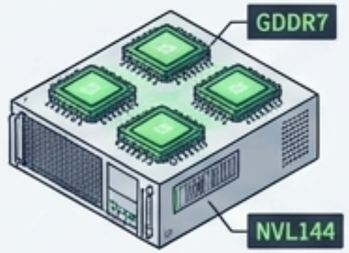
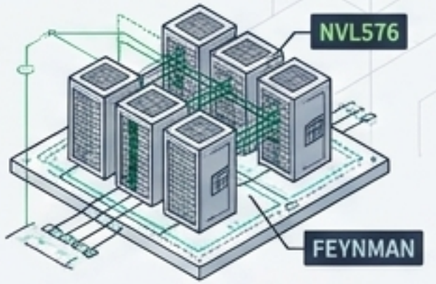
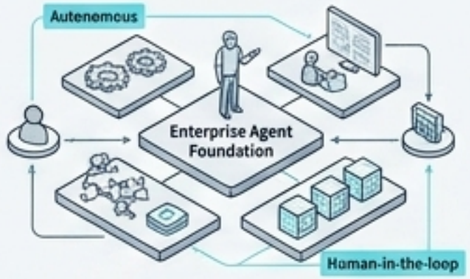
Layer 2: Deployment Infrastructure
Google Cloud Run / Docker Sandboxes
[セキュアでスケーラブルなコンテナ実行環境]

Layer 1: Hardware & Memory
Nvidia Rubin (NVL576) & GDDR7
[圧倒的なコンピュートと帯域幅]



次世代AIの成功は、単一のモデル性能ではなく、最下層のメモリ帯域から最上層の
Human-in-the-loop UXに至る、スタック全体の最適化によって決定づけられる。

Strategic Roadmap (2025-2027): 今、組織が取り組むべきこと

	2025	2026	2027+
Infrastructure (ハードウェア進化)	 Blackwell B300導入	 GDDR7搭載WS普及 / Nvidia Rubin (NVL144) 登場	 Rubin Ultra (NVL576) 稼働 / Feynmanアーキテクチャへ
Software & UX (組織の対応)	 Magentic-UI型プロトタイプの検証開始。 業務特化型MCPサーバーの開発。	 帯域幅の解放に伴うマルチエージェントの 実業務デプロイとCo-tasking UXの標準化。	 完全自律と人間監視がシームレスに統合された エンタープライズ・エージェント基盤の完成。

インフラの限界が突破される2026年に向け、
今すぐ『人間とAIの協調アーキテクチャ』の設計図を描き始めよ。