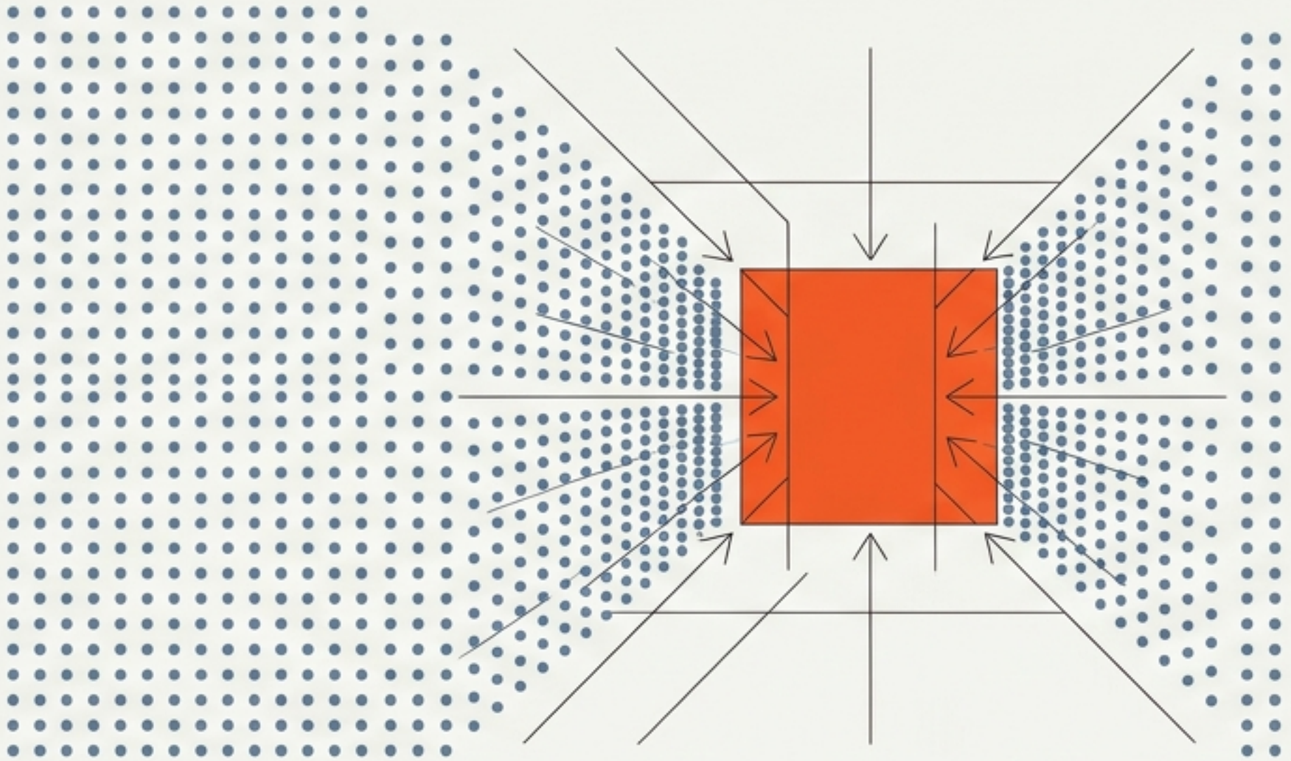
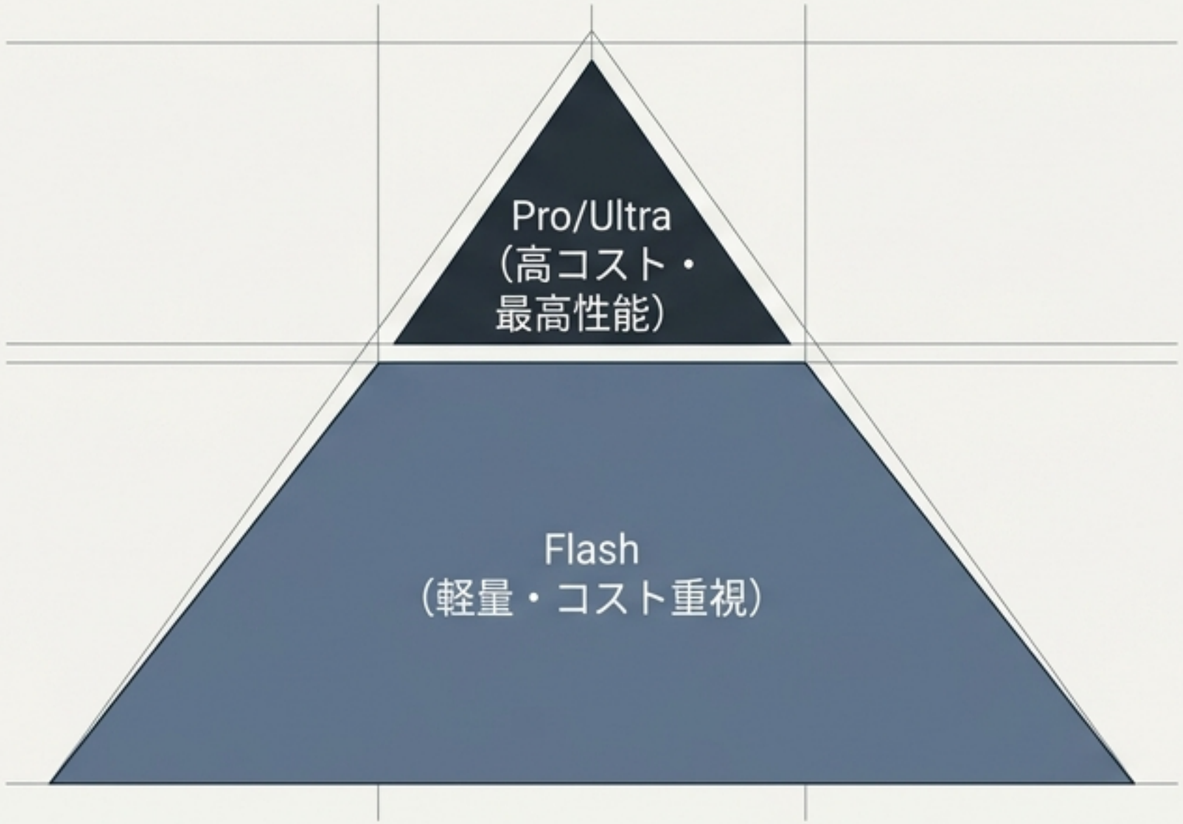
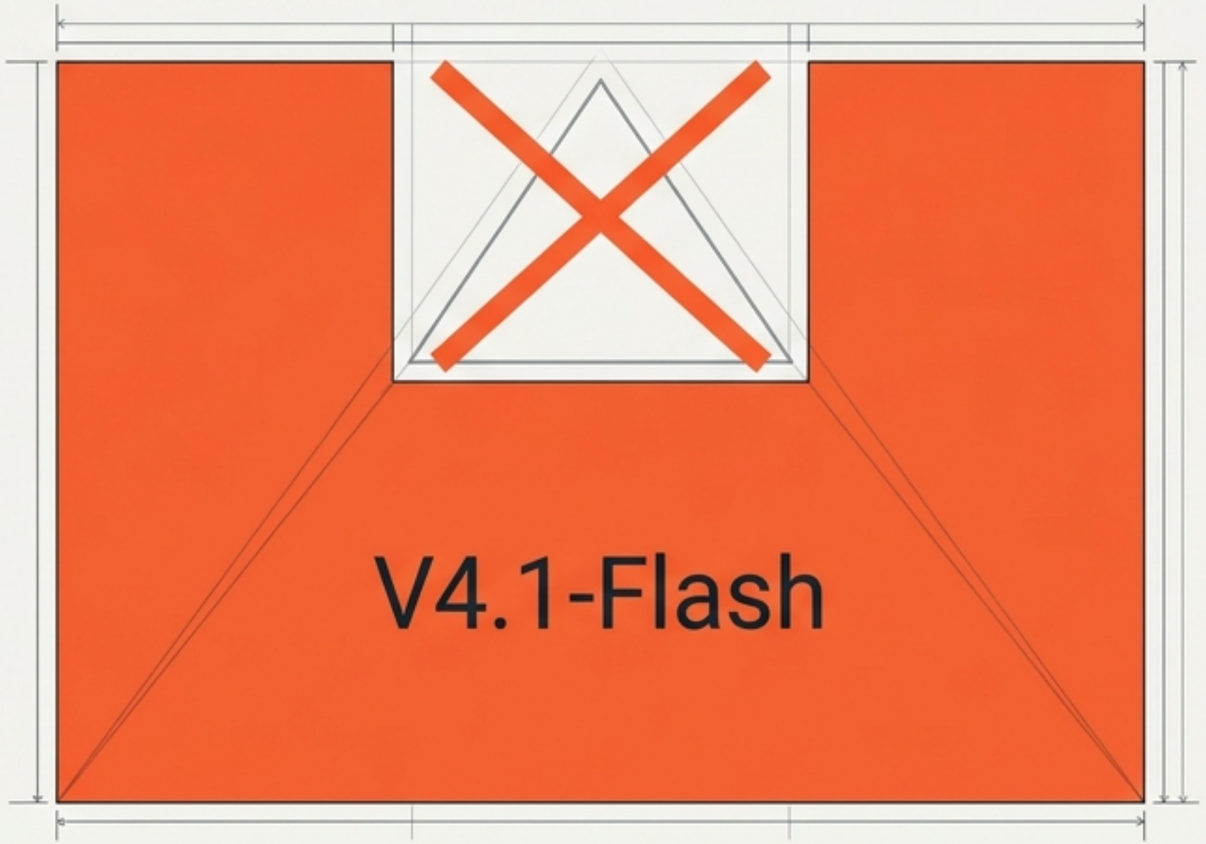


Proモデルの終焉とゼロ限界費用AIの夜明け

DeepSeek-V4.1-FlashによるAI市場構造の変革と「下剋上」の全貌

業界の階層秩序を破壊する構造的な下剋上





旧パラダイム（〜2026.08）

フラッグシップ（Pro）が最高性能を独占し、Flashはコスト重視の軽量タスクに限定される棲み分け。

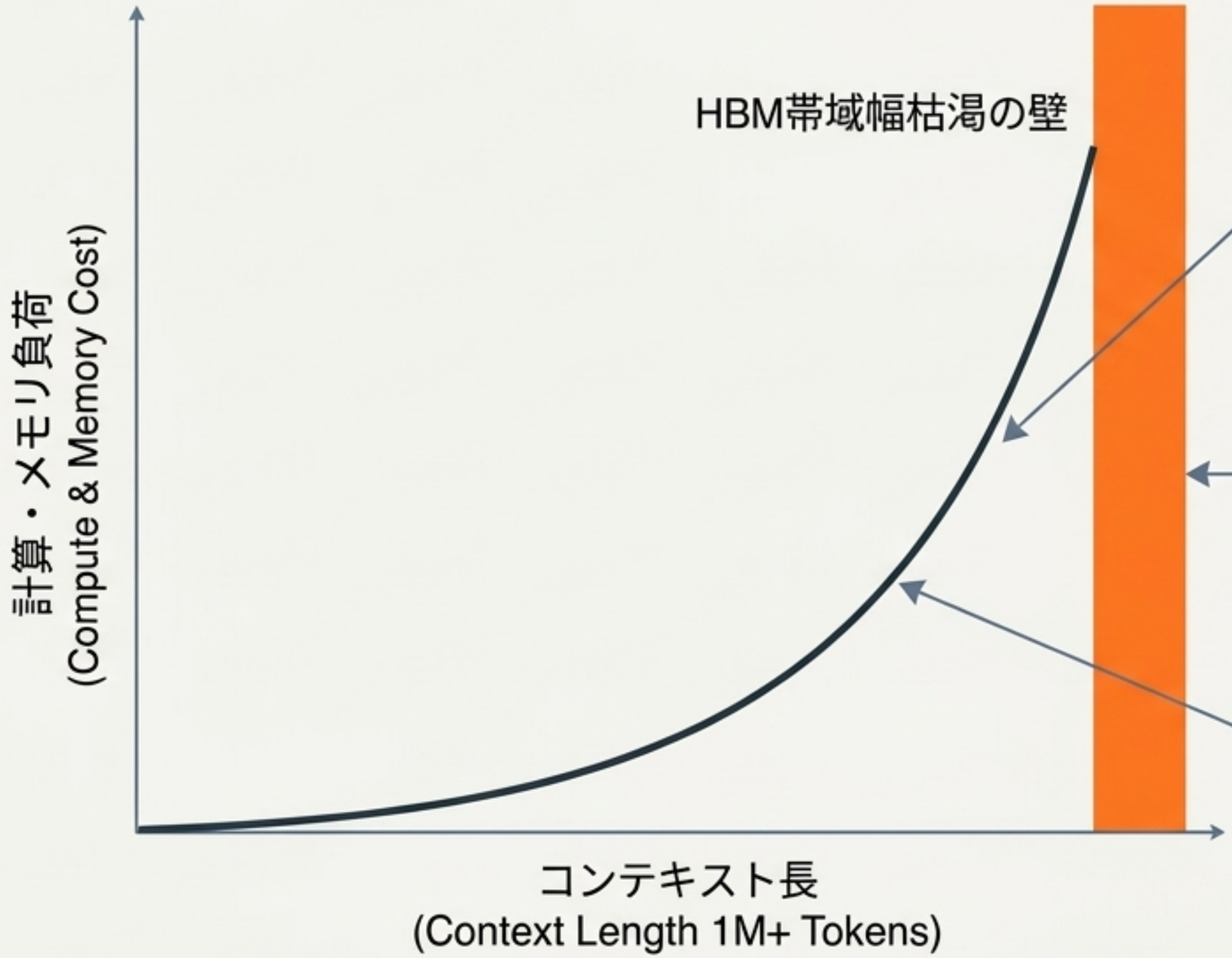
新パラダイム（2026.09〜）

DeepSeek-V4.1-Flashが、同社の最上位モデル「V4-Pro」を総合性能・速度・コストの全方位で凌駕。

即時統廃合

V4-Pro、旧V4-Flash、Vision-Expは廃止。全リクエストはV4.1-Flashへ自動転送。軽モデルがフロンティア級インフラへ昇格。

AIスケーリングを阻む2つの物理的ボトルネック



推論コストの指数関数的増大

Transformerデコーダオンリー構成では、100万（1M）トークン到達時に計算負荷が非現実的な水準へ跳ね上がる。

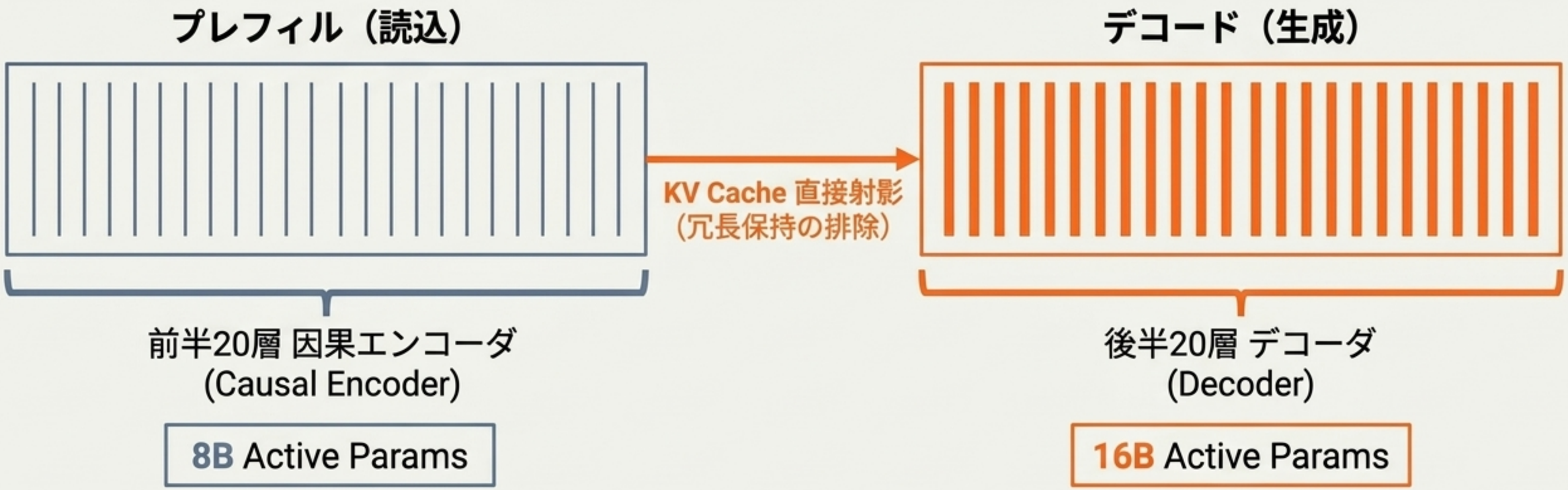
メモリ帯域幅（HBM）の枯渇

過去の文脈を保持するKVキャッシュが物理メモリを占有し、サーバーの並行処理能力を致命的に低下させる。

結論

既存の対称的なアーキテクチャのままでは、長大コンテキストと自律エージェントの商業展開は不可能。

非対称Causal Encoder-Decoder（CED）の導入



プレフィルの極小化

入力文脈の圧縮・読解は前半のエンコーダのみで処理。稼働パラメータをわずか80億（8B）に抑制。

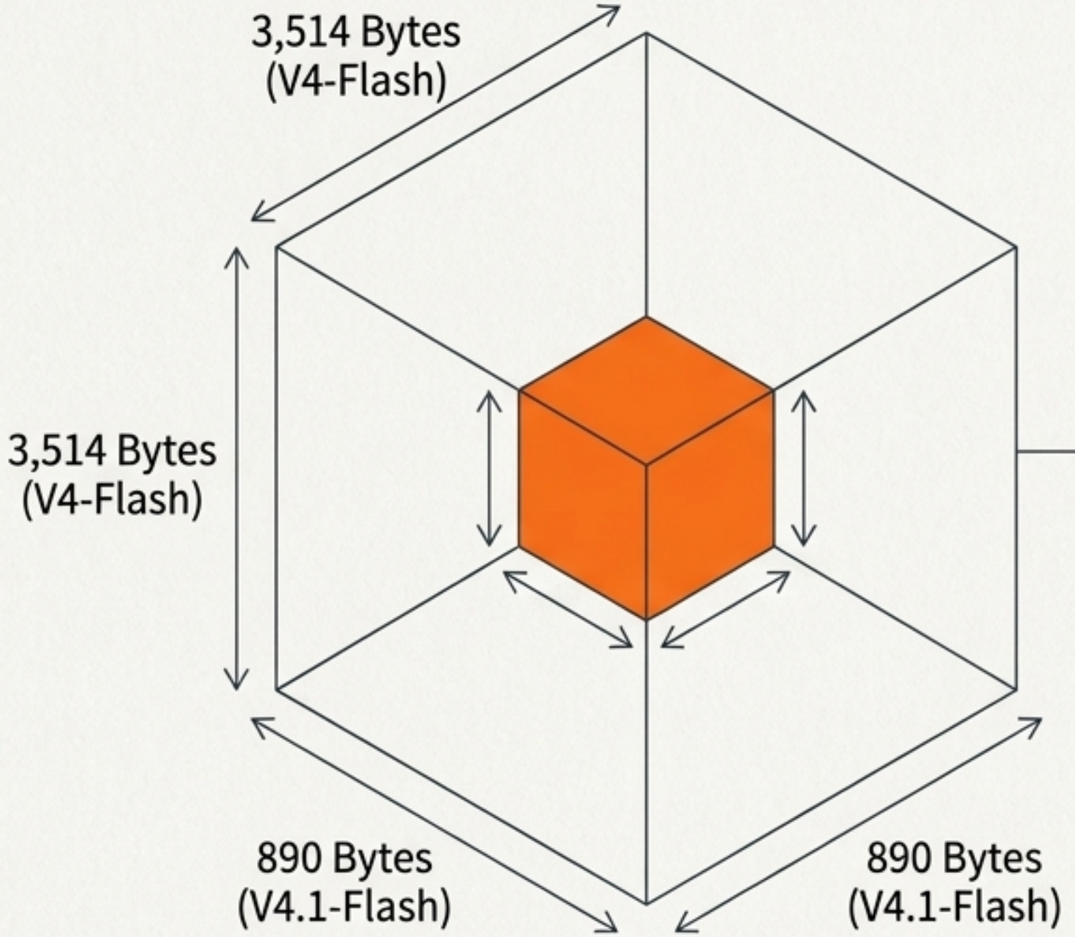
デコードの最適化

出力生成は後半のデコーダ（稼働16B）が担当し、高度な推論力を担保。

冗長性の排除

層間での冗長な状態保持を根本から排除し、計算コストを劇的に削減。

物理メモリ消費を極限まで抑制する超圧縮技術



CSA2（Compressed Sparse Attention 2）

アテンション層を機能分割し、階層型スパースインデクサーで探索空間を制限。

FP4 KV Cache

E2M1フォーマット適用。数値精度を維持したままKVの物理表現を極小化。

SWA Bounded Replay

直近トークンのみを動的再計算。SSD等の低速外部ストレージへのKV退避（オフロード）を完全に排除。

結果: トークンあたりのKV消費量を前世代比で約1/4（初代比1/437）へ圧縮。

DeepSeek-V4.1-Flash システム諸元

総パラメータ数 (MoE)

552B (5,520億)

共有エキスパート1 /
ルーティングエキスパート384

アクティブパラメータ

8B / 16B

非対称スケーリング
(Prefill / Decode)

最大コンテキスト長

1,000,000 Tokens

エージェントループ完全対応

最大出力トークン

256,000+ Tokens

長大コードベースの全体生成

KVキャッシュ / トークン

890 Bytes

HBM要求量 1/4へ激減

提供ライセンス

MIT License

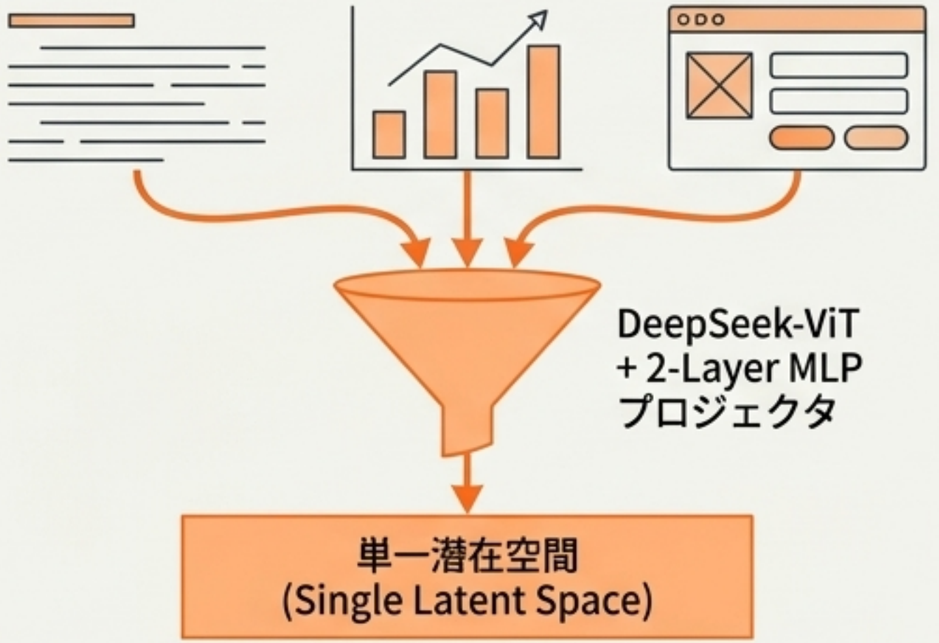
重み・コード完全無償公開

後付けアダプタの廃止とネイティブ・マルチモーダル統合

旧式（ボルトオン型）

非効率な多段階プロセス

ネイティブ統合型（V4.1-Flash）



ゼロからの事前学習

45兆（45T）トークン規模のテキスト・画像混合コーパスを用いて基盤から再設計。

単一潜在空間での処理

高解像度画像、複雑なUI画面、構造化文書、図表をテキストと同一のトークン列として直接自己回帰処理。

統合の成果

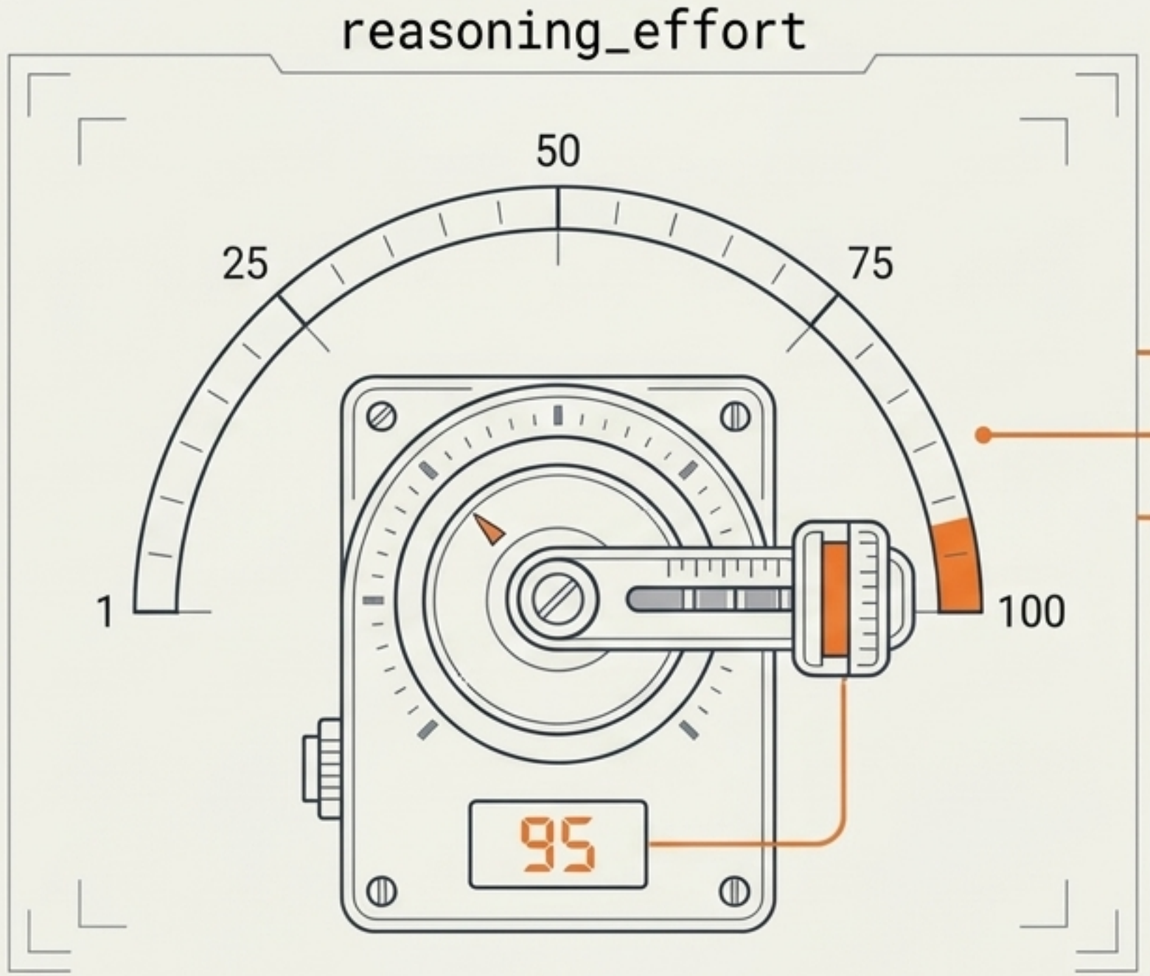
「高速だが視覚が弱い」という軽量モデルの常識を覆し、フロンティア級のマルチモーダル解析を業界最低水準コストで実現。

性能プロファイル：自律エージェント領域での絶対的優位

評価領域（指標）	V4.1-Flash	V4-Pro	Claude Opus 5
実コード修正 (DeepSWE v1.1)	**74.2**	62.7	74.0
業務自動化 (AutomationBench)	**54.8**	43.2	50.3
端末自律操作 (Terminal-Bench 2.1)	**90.6**	-	-
専門科学推論 (GPQA Diamond)	90.9	**92.4**	-
最難関学術試験 (HLE)	36.8	39.1	**56.3**

HLE（Humanity’s Last Exam）などの極限の抽象・学術推論では依然として巨大稠密モデル（Opus 5等）に軍配が上がるが、ソフトウェアエンジニアリングや環境操作の実務タスクではV4.1-Flashが完全な勝利を収めている。

限界を突破する動的推論機構：「reasoning_effort」



課題の背景

8B/16Bという極めてスパースな軽量設計は、手続き型処理に強い反面、多段階の抽象概念を統合する学術推論で不利に働く。

動的ダイヤルによる解決

APIパラメータ`**reasoning_effort**`（1〜100の整数値）により、モデル自身が計算ステップ（Test-Time Compute）を動的に配分。

柔軟なトレードオフ

- Low (1-10): 日常的なエージェントパイプラインを超高速・低コストで処理。
- High (90-100): HLE等の難問に対し、計算量を集中投下して正答率を強制的に引き上げる。

限界費用の破壊：100万トークン 0.003ドル時代

キャッシュヒット時の極限単価

KVキャッシュ圧縮の恩恵を直接還元。入力単価は100万トークンあたり$0.003（約0.5円）へ到達。

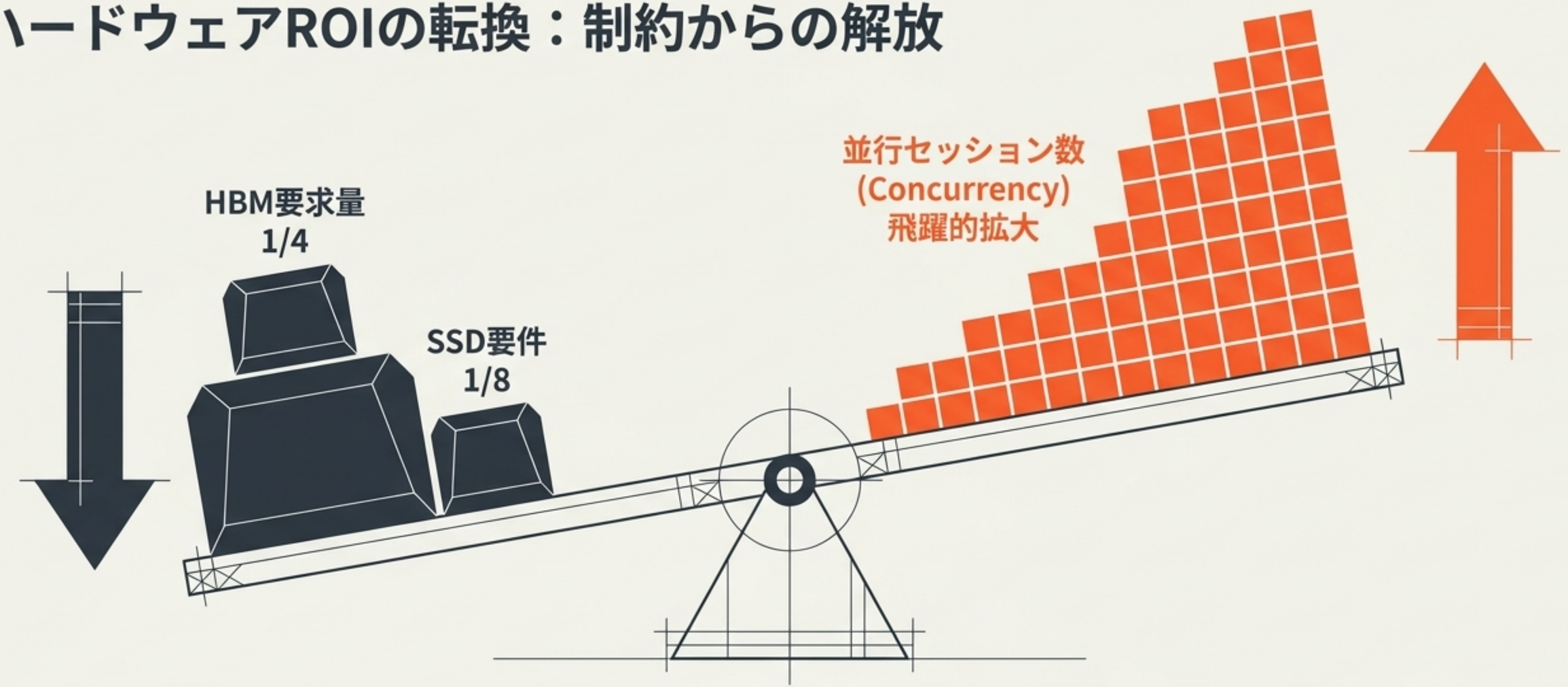
オフピーク戦略

平日の特定時間（ピーク）以外は、すべて半額のオフピーク料金を強制適用し稼働率を平準化。

プロプライエタリ独占崩壊

1Mトークン数十ドルを課金するクローズドモデルの投資対効果（ROI）正当化は事実上不可能に。

ハードウェアROIの転換：制約からの解放



HBM枯渇問題の解消

入力読込が8Bに抑制され、KVキャッシュが物理的に1/4に圧縮されたことで、最先端HBM搭搭載チップへの過度な依存が緩和。

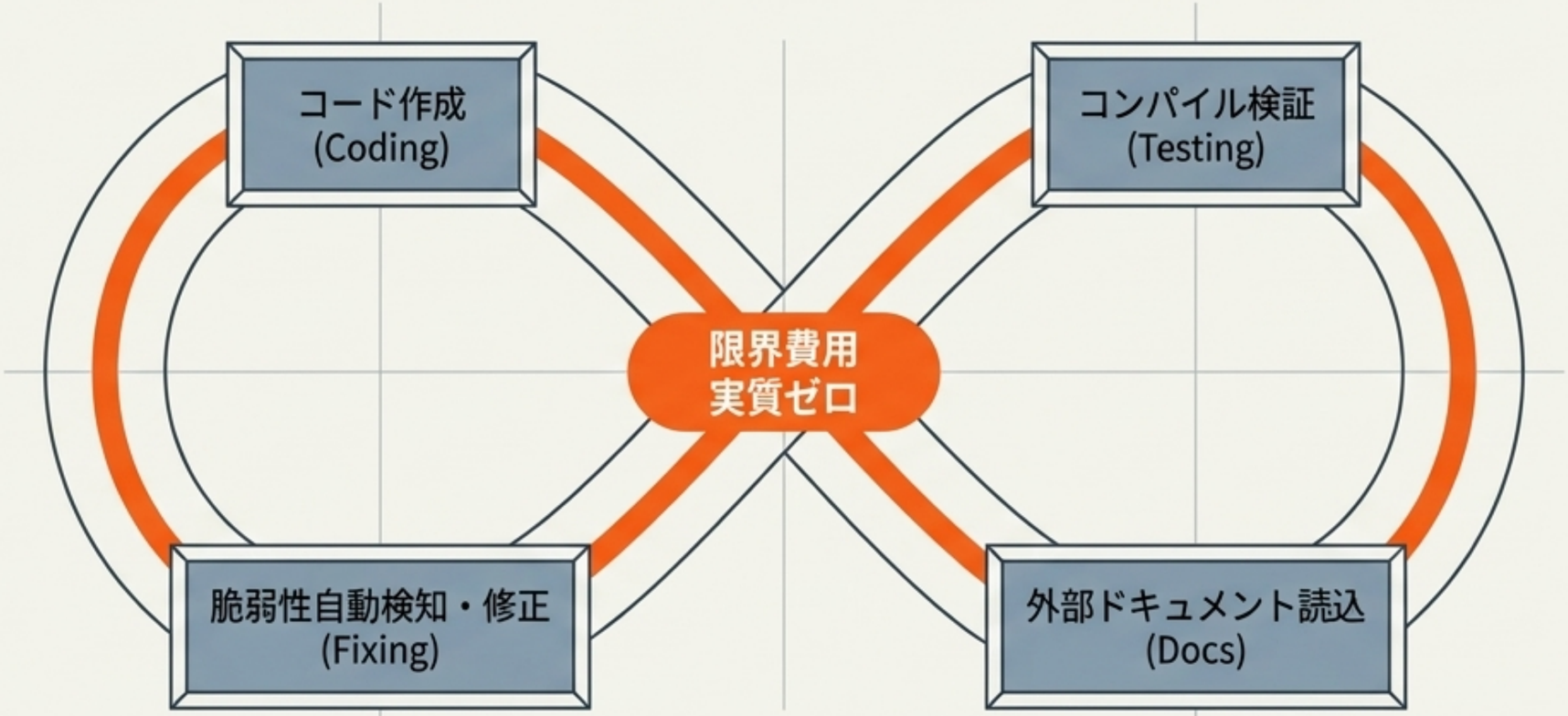
既存インフラの価値最大化

新規の莫大な半導体設備投資を行わずとも、既存の汎用GPUクラスタのまま推論スループット（並行処理数）を数倍に引き上げ可能。

データセンターの経済性

サーバー単位の収益性が劇的に向上。インフラ制約ではなく、ワークフロー設計がボトルネックとなる時代へ。

ゼロ限界費用・自律型エージェント経済の成立



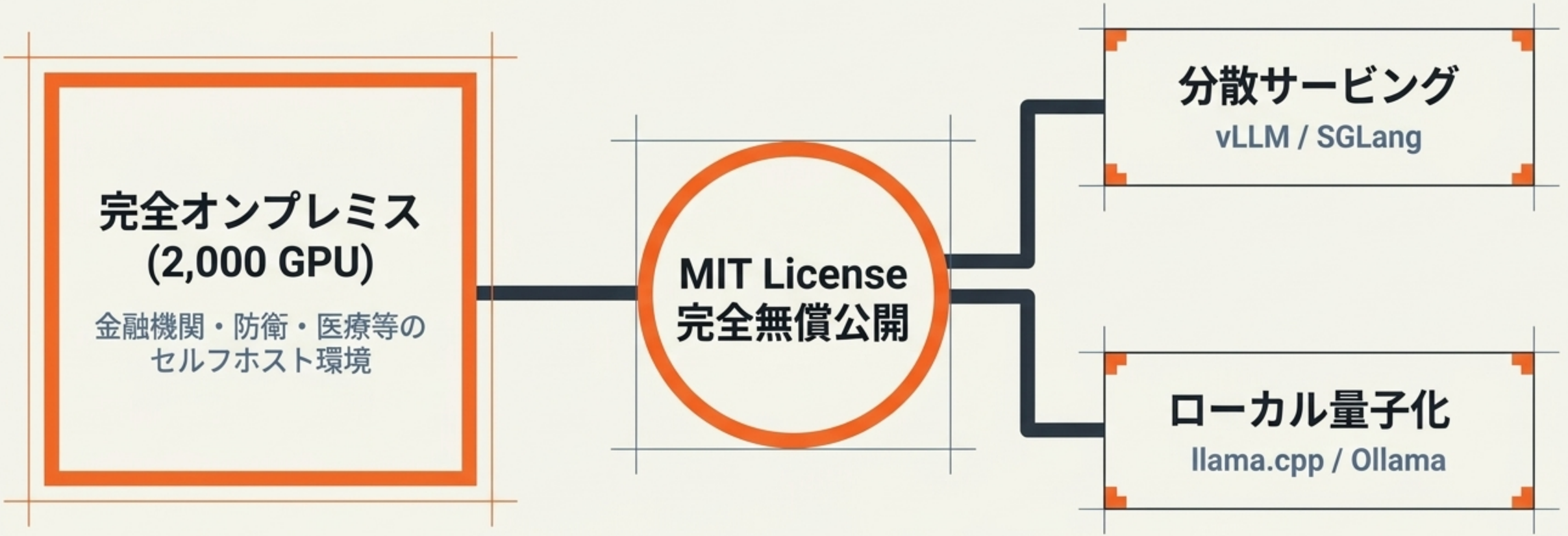
これまでの障壁

エージェントが試行錯誤を繰り返すたびに累積する長大な文脈（KVキャッシュ）維持コストが、商業運用の最大の壁だった。

超高頻度タスクの商業化

- 数千体のエージェントを24時間常時稼働。
- リポジトリ全コードの継続的リファクタリング。
- 結論：「知能の呼び出し」から「労働力の常時自律稼働」への完全なシフト。

オープンエコシステムとオンプレミス主権



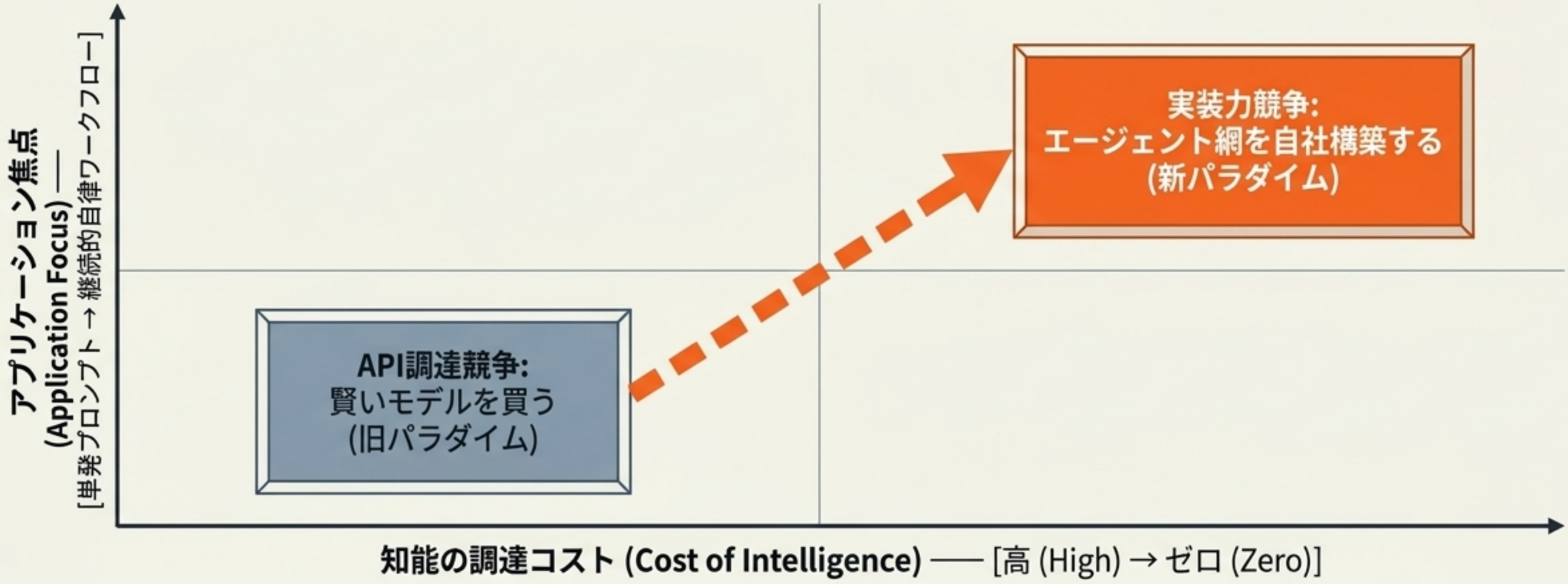
MITライセンスの衝撃

ブラックボックス化を進める西側AI企業に対し、商用利用・改変・再配布が自由なフロンティア級モデルとして強力なオルタナティブを提供。

エンタープライズ統制環境の構築

データ主権や国外移転規制をクリアするため、大規模GPUクラスタを用いた完全オンプレミス運用支援を確立。パブリック APIを利用できないセクターを攻略。

戦略的インサイト：競争優位性の源泉シフト



企業が直面する本質的な問いは、「どの高価なモデルを買うか」から、「遍在する極めて安価な知能（V4.1-Flash）を前提に、いかに洗練された自律エージェントネットワークを自社ワークフローに実装できるか」へと完全に移行した。

DeepSeek-V4.1-Flash : 3つの確定事項

01

アーキテクチャの下剋上

非対称CEDと極限のKVキャッシュ圧縮技術（890 Bytes/token）により、推論効率とフロンティア級性能が両立し、巨大なProモデルは過去のものとなった。

02

限界費用の破壊

100万トークン0.003ドルという驚異的な価格設定は、ソフトウェア自動化における投資対効果（ROI）の前提を根底から覆す。

03

自律エージェント経済の開幕

AIは一部の富裕企業の「希少資源」から、あらゆるシステムに遍在する「コモディティインフラ」へと転換した。常時稼働エージェントの実装力が次代の勝者を決める。