

DeepSeek-V4.1-Flash : AI経済を塗り替える「超圧縮・高効率」の衝撃

従来の「大規模＝高コスト」の壁

メモリの壁と計算負荷



非効率なリソース消費

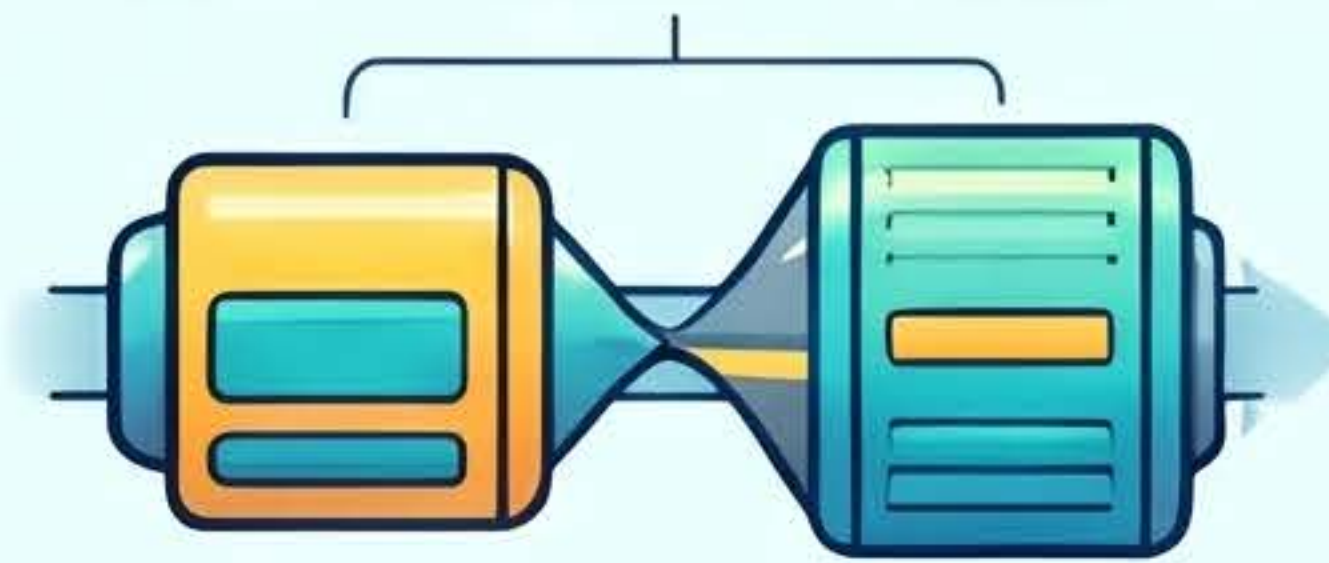
HBM & ストレージ
大量消費



DeepSeek-V4.1-Flash : 非対称・超圧縮アーキテクチャ

工学的革新：メモリの壁を打破する新構造

非対称CED構造による効率化

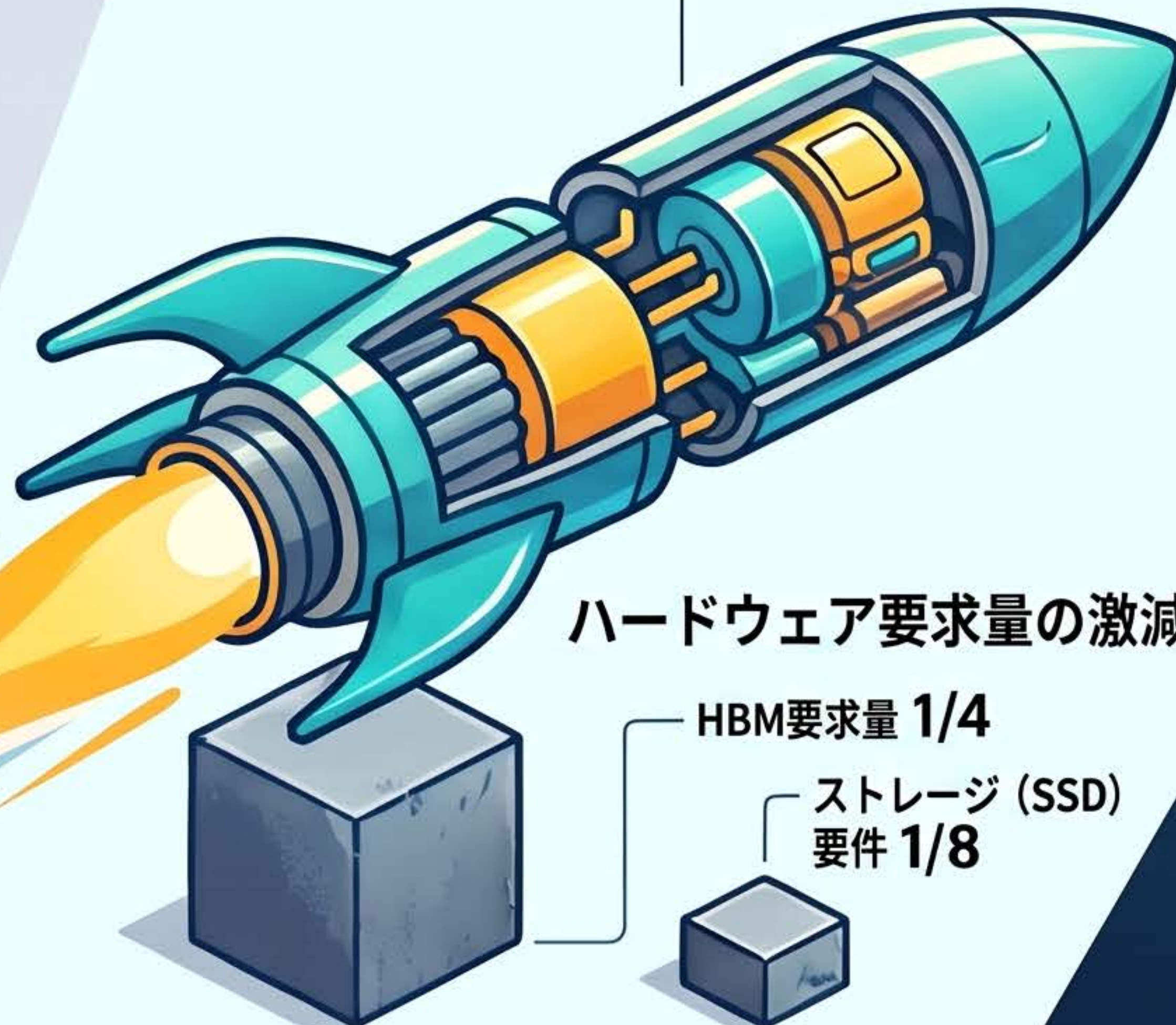


プレフィル (8B) デコード (16B)
計算負荷を最適化

KVキャッシュ容量を
約1/4に圧縮



1トークンあたり890バイト
独自のCSA2技術等



ハードウェア要求量の激減

HBM要求量 1/4

ストレージ (SSD)
要件 1/8

従来HBM/ストレージ ハードウェア要求量の激減

経済的インパクト：エージェント経済の夜明け

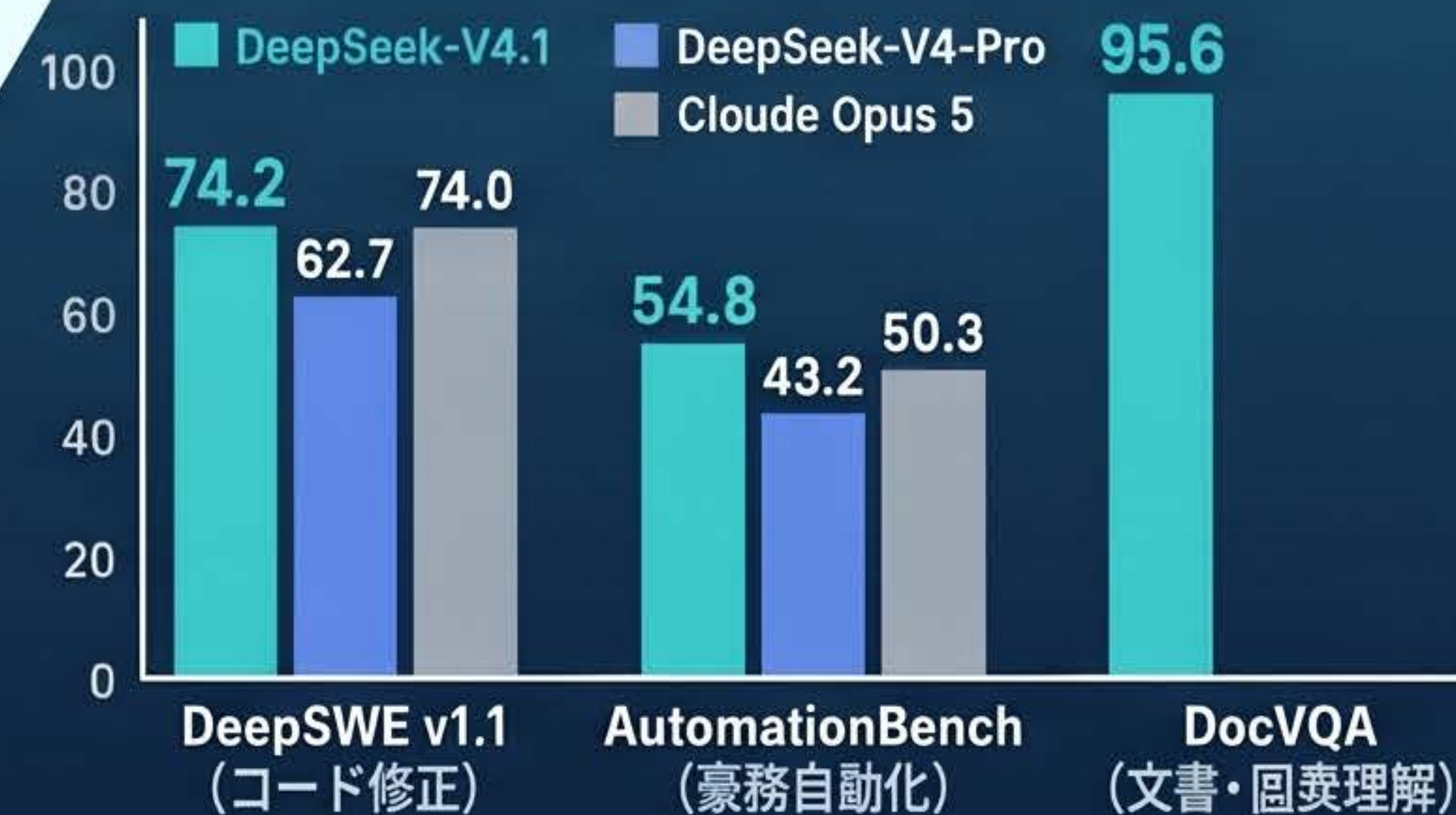


キャッシュヒット単価
\$0.003

100万トークン入力単価、運用コスト実質無効化

フラッグシップ級を凌駕する実務性能

主要エージェント・実務ベンチマーク性能比較



MITライセンスでの
完全無償公開

重みとコードを公開、商用利用・
オンプレミス自律展開を強力支援