AIインフラ

AI Infrastructure

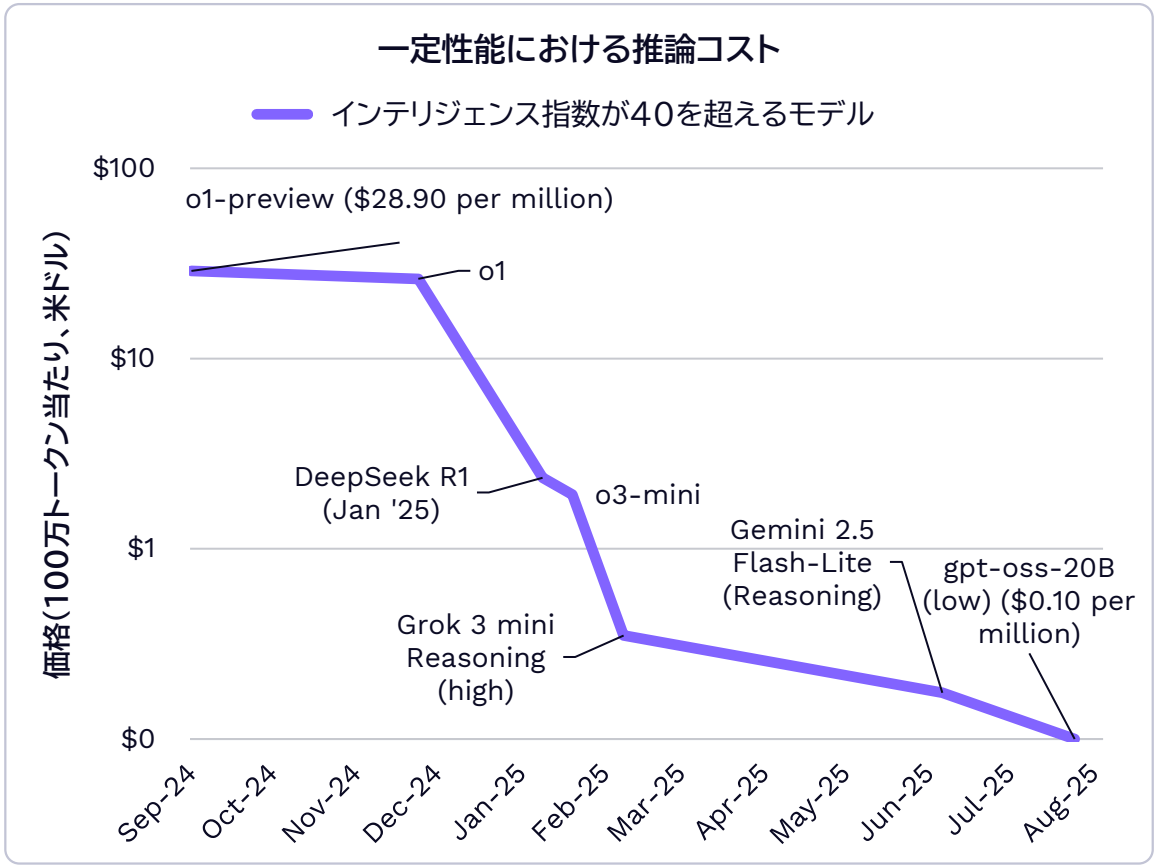
次世代クラウドを形づくるもの

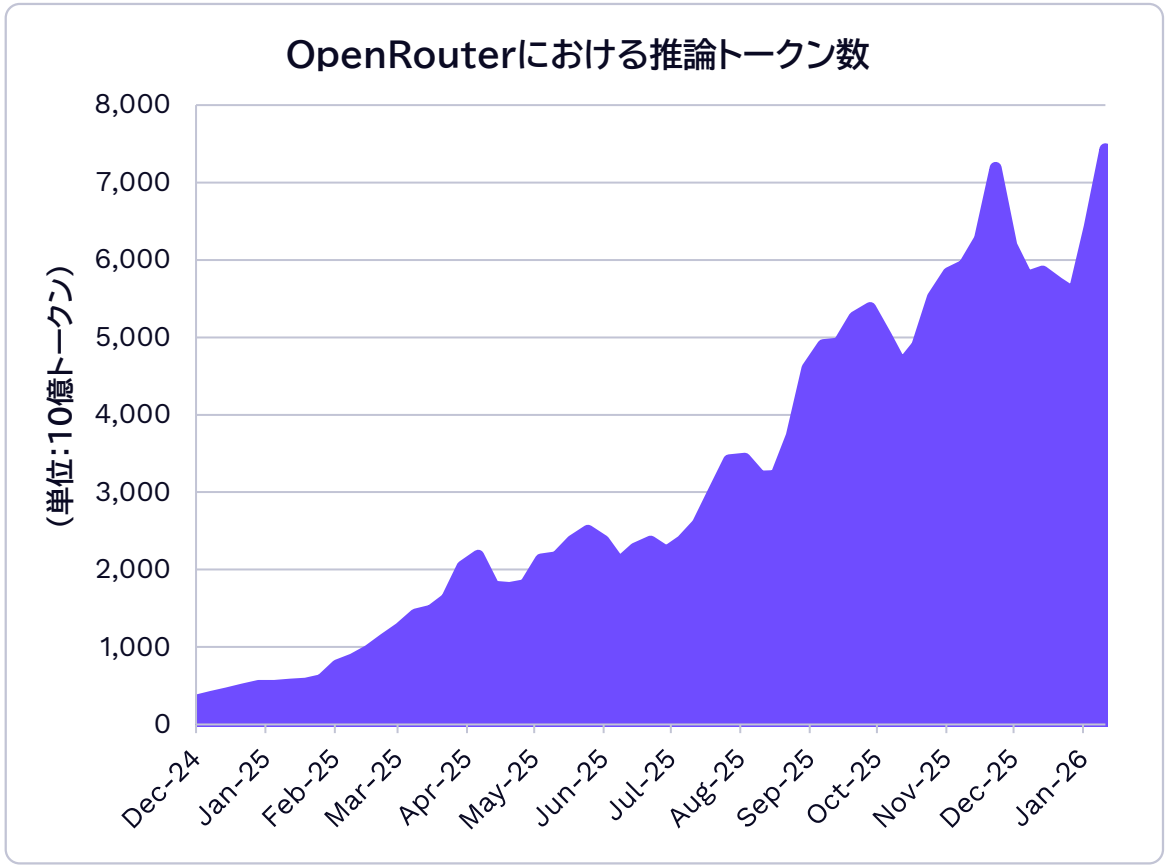
Frank Downing
(フランク・ダウニング)
リサーチ・ディレクター
AI・クラウド分野担当

Jozef Soja
(ジョゼフ・ソーヤ)
リサーチ・アナリスト
AI・クラウド分野担当

推論コストの急低下に伴うAI需要の急拡大

一部の指標によれば、推論コストは過去1年間で99%以上低下しています。AIネイティブなアプリケーションが急速に普及する中、コスト低下は開発者、企業、消費者による推論トークン数の爆発的な増加を後押ししています。大規模言語モデル(LLM)にアクセスするための統合アプリケーション・プログラミング・インターフェース(API)であるOpenRouterでは、演算需要が2024年12月以降25倍に拡大しています。

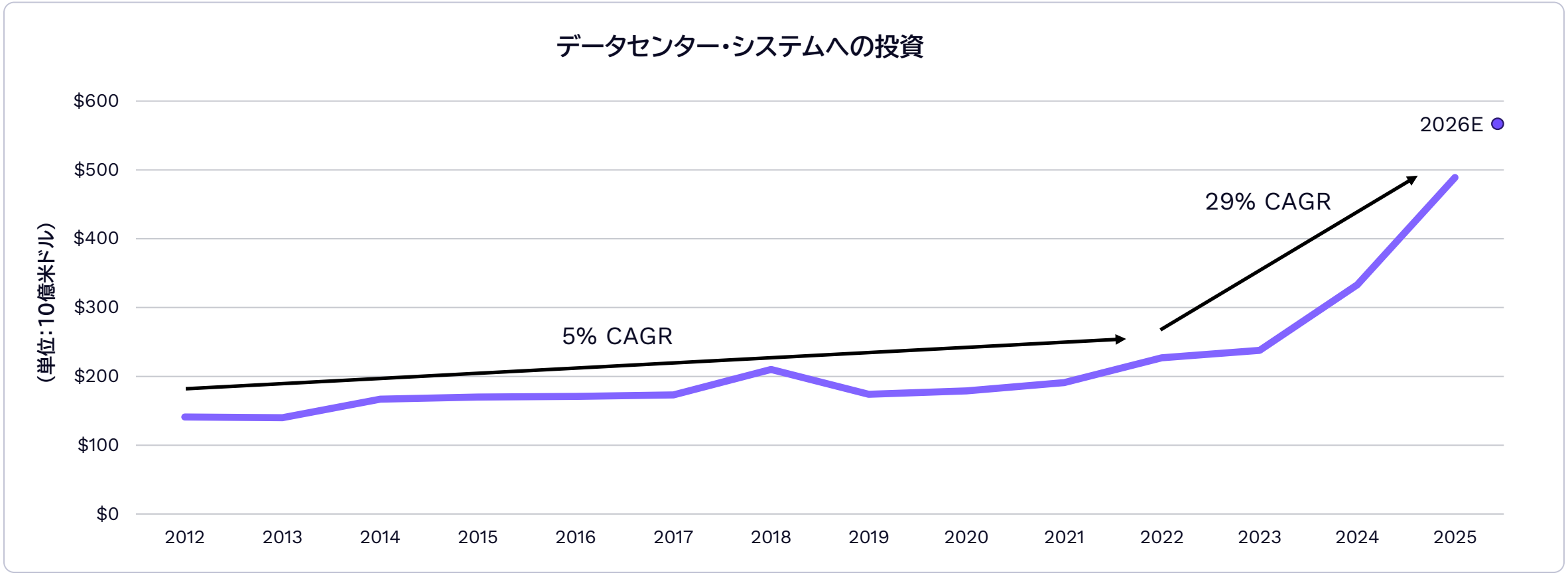




出所:ARK Investment Management LLC(2026年)。本資料は情報提供のみを目的としたものであり、特定の有価証券の売買または保有を推奨するものではありません。過去の実績は将来の成果を示唆または保証するものではありません。また、将来予測には本質的な不確実性があり、その正確性を保証するものではありません。

「ChatGPTモーメント」以降のデータセンター・システム成長率の加速（年率5%から29%へ）

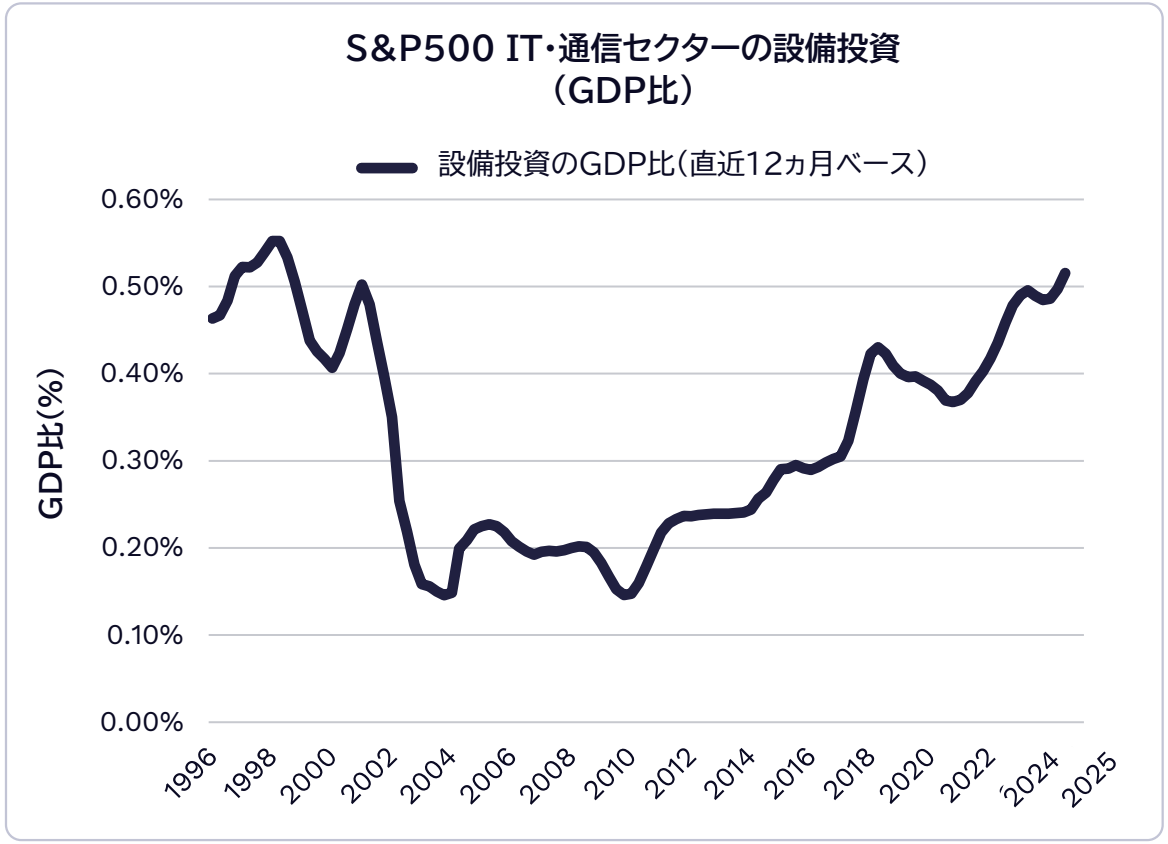
2025年には、データセンター・システムへの年間投資額は約5,000億米ドルに達し、2012年から2023年までの平均と比べて約2.5倍となりました。当社のリサーチによれば、この投資分野は今後も加速局面を迎え、2030年には約1兆4,000億米ドルへと約3倍に拡大する可能性があります。

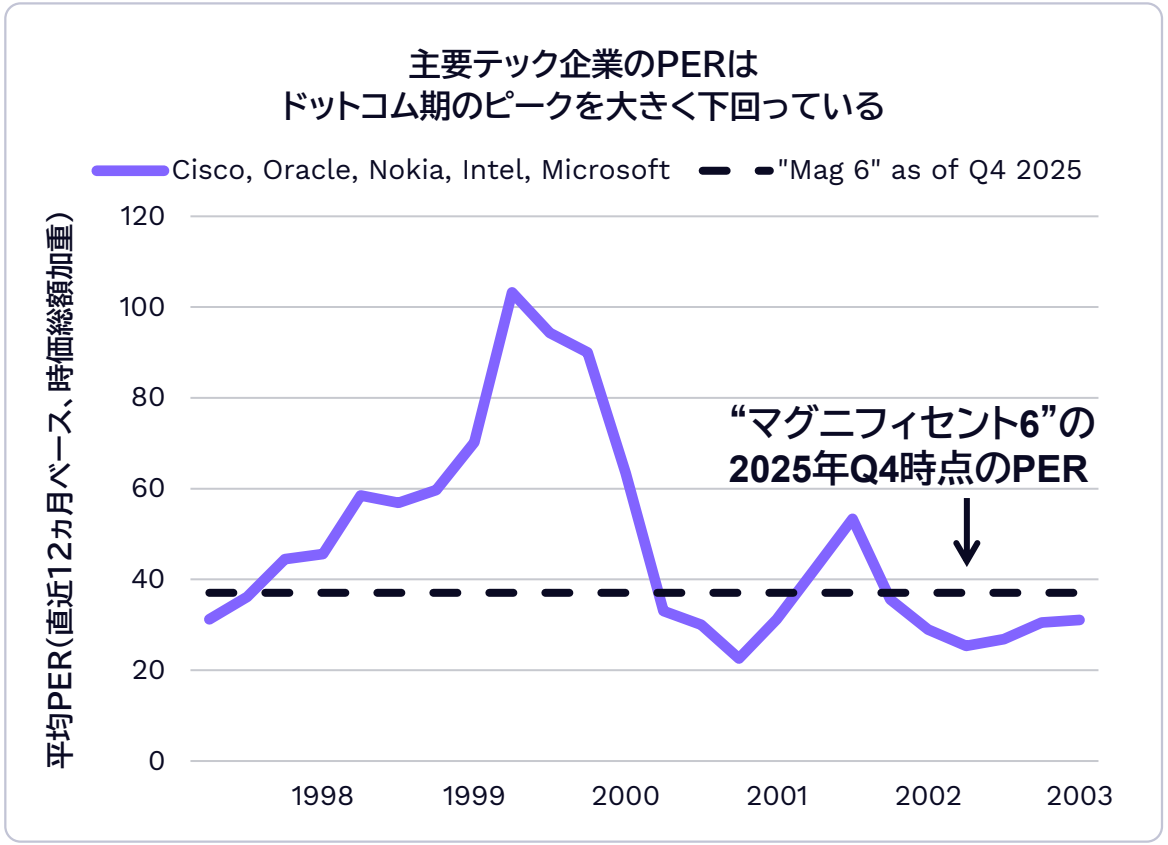


注記：「CAGR」とは、年平均成長率（Compound Annual Growth Rate）を指します。出所：ARK Investment Management LLC（2026年）。Morgan（2025年a、2025年b、2024年、2025年10月27日時点）のデータに基づいています。本資料は情報提供のみを目的としたものであり、特定の有価証券の売買または保有を推奨するものではありません。過去の実績は将来の成果を示唆または保証するものではありません。また、将来予測には本質的な不確実性があり、その正確性を保証するものではありません。

テック&テレコム・ブーム期水準にあるテック設備投資と低水準のバリュエーション

当社のリサーチによれば、大手クラウド事業者(ハイパースケーラー)各社の設備投資額(Capex)は、2026年に5,000億米ドル超に達すると見込まれます。これは、「ChatGPTモーメント」が起こる前の2021年の1,350億米ドルと比べて約3倍の水準です。情報技術および通信サービス分野における設備投資は、国内総生産(GDP)に占める割合で見ると、1998年以来の高水準に達しています。一方で、テック・セクターの株価収益率(PER)は、テック・通信バブル期のピーク時と比べると、そのごく一部の水準にとどまっています。

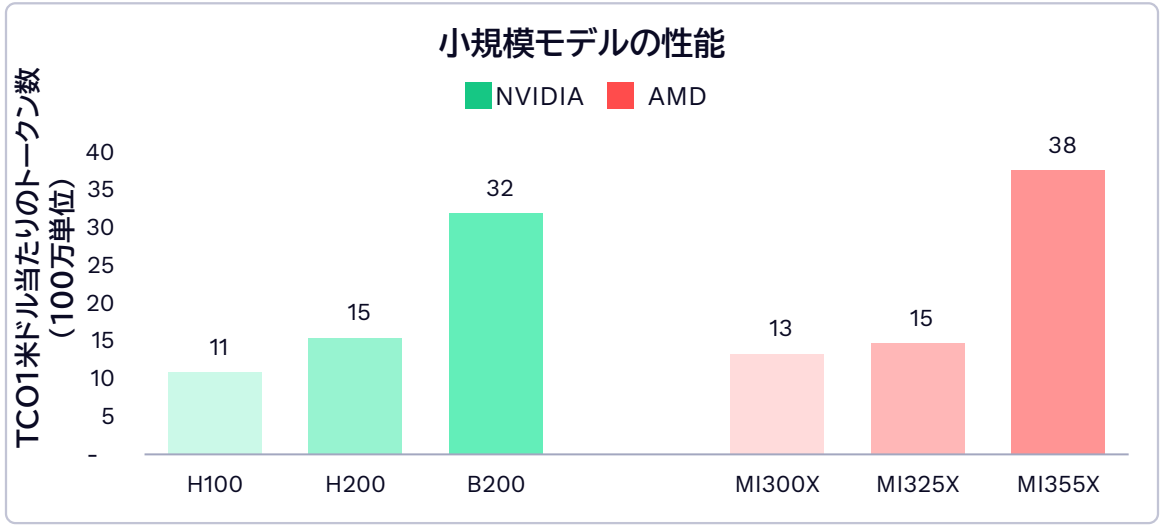


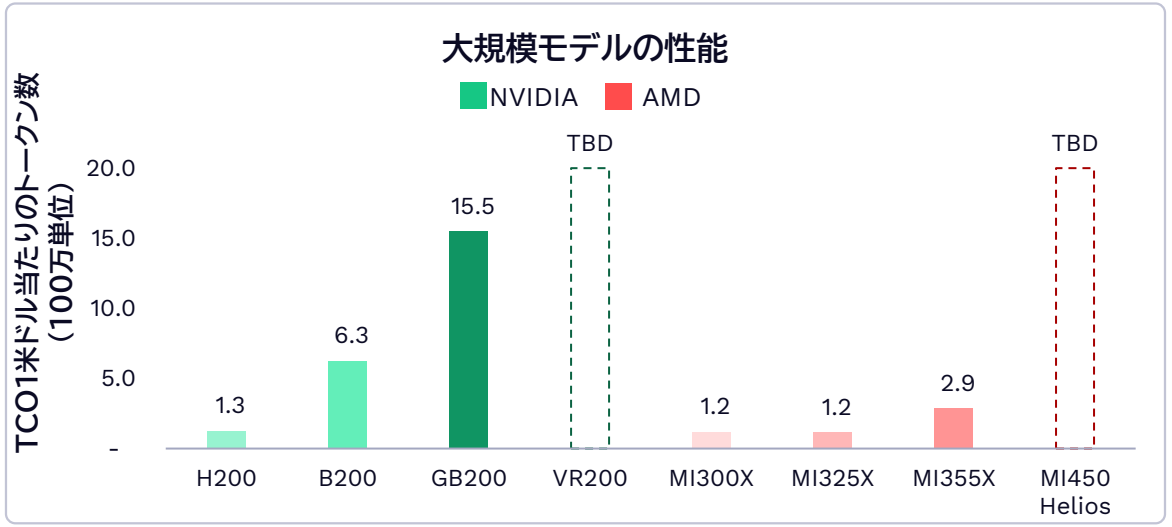


注記:「Mag 6」には、Alphabet、Apple、Amazon、Meta、Microsoft、Nvidiaが含まれます。出所:ARK Investment Management LLC(2026年)。Bloomberg(2025年a、2025年b、2026年)、FRED(2025年)、S&P(2025年、2026年1月6日時点)のデータに基づいています。これらの情報源に加え、本資料の一部には、複数の追加的な情報源を用いたARK独自の分析結果が含まれている場合があります。本資料は情報提供のみを目的としたものであり、特定の有価証券の売買または保有を推奨するものではありません。過去の実績は将来の成果を示唆または保証するものではありません。また、将来予測には本質的な不確実性があり、その正確性を保証するものではありません。

競争激化に直面するNVIDIA

NVIDIAは、AI向けチップ設計、ソフトウェア、ネットワーキングへの早期投資によって、GPU販売におけるシェアを85%まで高め、売上総利益率も75%に押し上げました。しかし現在では、AMDやGoogleといった競合企業が、小規模言語モデルの推論など特定の分野において追いつきつつあります。一方で、NVIDIAのGrace Blackwellラックスケール・システムは、大規模モデルの推論分野で優位性を保っており、最先端の基盤モデルを支えています。



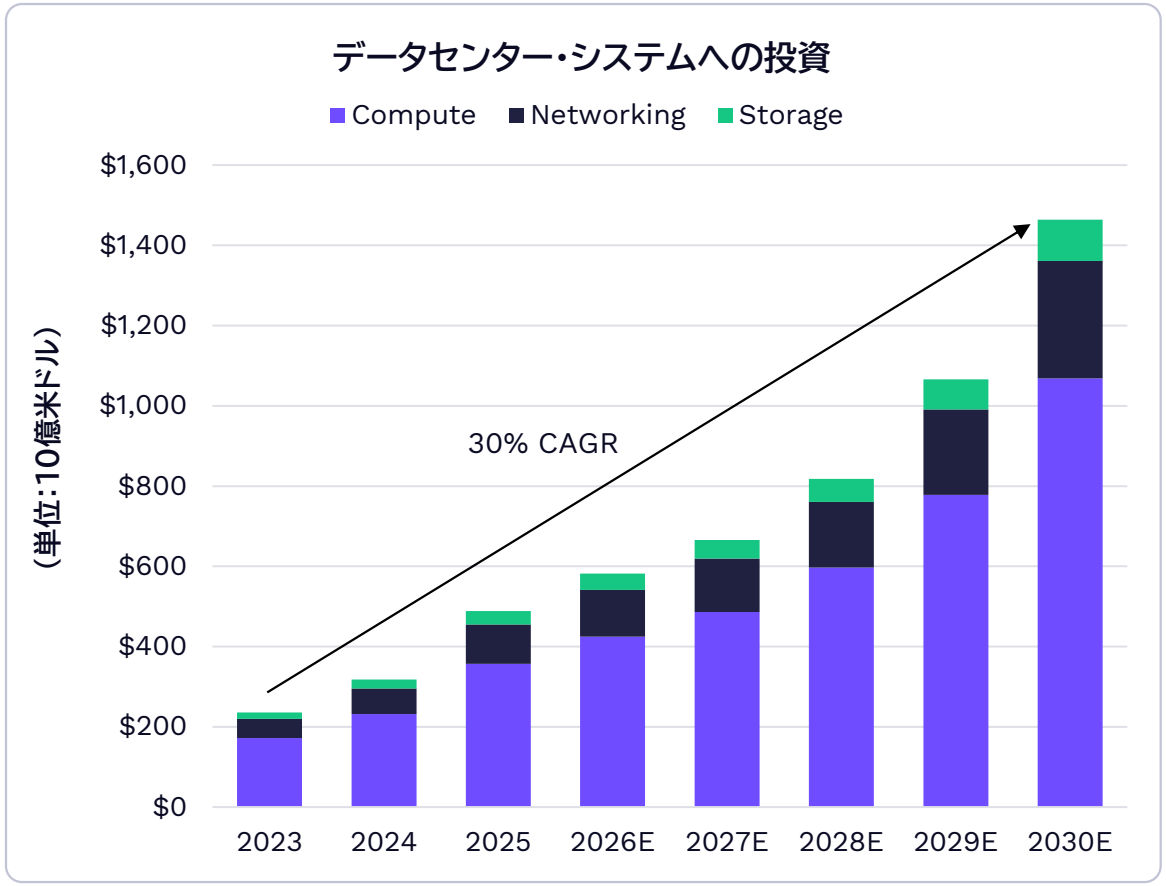


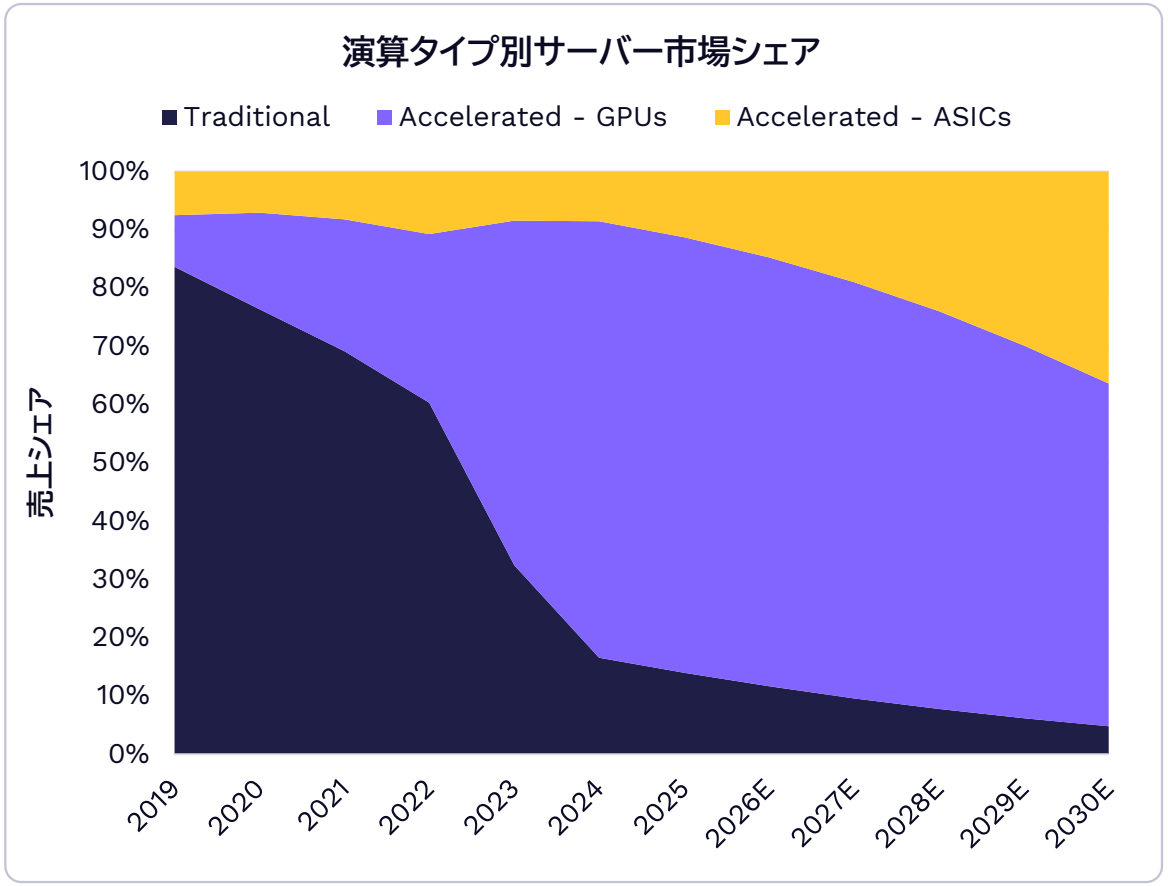
GPU	H200	B200	GB200	VR200	MI300	MI355	MI455	TPU v7
初期出荷台数	Q2 2024	Q1 2025	Q1 2025	~2H 2025	Q4 2023	Q2 2025	~2H 2026	Q4 2025
メモリ容量	141 GB	192 GB	192 GB	288 GB	192 GB	288 GB	432 GB	192 GB
消費電力（ワット）	700	1000	1200	TBD	750	1400	TBD	980
1時間当たりの総コスト（TCO）	$1.41	$1.95	$2.21	TBD	$1.13	$1.49	TBD	$1.28

注記：「TCO」とは、Total Cost of Ownership（総保有コスト）を指し、GPUの耐用期間を通じて発生する購入費用および運用費用を合算した総コストを意味します。出所：ARK Investment Management LLC（2026年）。Chenほか（2025年）、Patelほか（2025年）、InferenceMAX（2026年1月7日時点）のデータに基づいています。これらの情報源に加え、本資料の一部には、複数の追加的な情報源を用いたARK独自の分析結果が含まれている場合があります。本資料は情報提供のみを目的としたものであり、特定の有価証券の売買または保有を推奨するものではありません。過去の実績は将来の成果を示唆または保証するものではありません。また、将来予測には本質的な不確実性があり、その正確性を保証するものではありません。

AI需要が牽引する持続可能なインフラ成長

企業向けおよび消費者向けの双方の環境でAIワークロードが拡大する中、AIインフラへの投資額は2030年に1兆4,000億米ドル超に達する可能性があります。その大半は、アクセラレーテッド・サーバー向けの投資になると見込まれます。当社のリサーチによれば、AI研究機関や大手クラウド事業者（ハイパースケーラー）がコスト効率の高い演算能力を追求する中、BroadcomやAmazonのAnnapurna Labsといった企業が設計するASIC（特定用途向け集積回路）は、引き続きシェアを拡大していく可能性があります。





注記：「CAGR」とは、年平均成長率（Compound Annual Growth Rate）を指します。「ASICs」とは、Application-Specific Integrated Circuits（特定用途向け集積回路）を指します。出所：ARK Investment Management LLC（2026年）。Tegus（2025年）、The Next Platform（2025年）、IDC（2025年、2025年11月6日時点）のデータに基づいています。これらの情報源に加え、本資料の一部には、複数の追加的な情報源を用いたARK独自の分析結果が含まれている場合があります。本資料は情報提供のみを目的としたものであり、特定の有価証券の売買または保有を推奨するものではありません。過去の実績は将来の成果を示唆または保証するものではありません。また、将来予測には本質的な不確実性があり、その正確性を保証するものではありません。