ARTIFICIAL INTELLIGENCE

Is AI Computing Power Becoming a Commodity?

By Antti Belt and Allen Thomas

ARTICLE JULY 23, 2026 15 MIN READ

The economics of enterprise AI are shifting fast. Until recently, flat-rate subscriptions allowed large enterprise users of AI tools to focus more on driving uptake than on managing the costs of AI. But that flat-fee era for enterprise users has ended, as AI labs such as Anthropic and OpenAI have shifted to metered pricing. The next development is likely to be dynamic pricing, under which prices surge at times of peak demand. For end users, the implication is clear: AI cost exposure is rising.

Meeting that challenge will require a fundamental evolution in today's market for AI computing power (known as "AI compute"), away from heterogeneous contracts, opaque pricing, and limited tools to a system that is more liquid and transparent. How far the market moves toward complete commoditization is uncertain, owing to notable headwinds—among them the fact that AI compute is heterogeneous, varying across hardware generations, timing, and geography. But a shift is undoubtedly under way, and benchmarks and futures contracts for AI chip rentals are already emerging to enable price transparency and risk management.

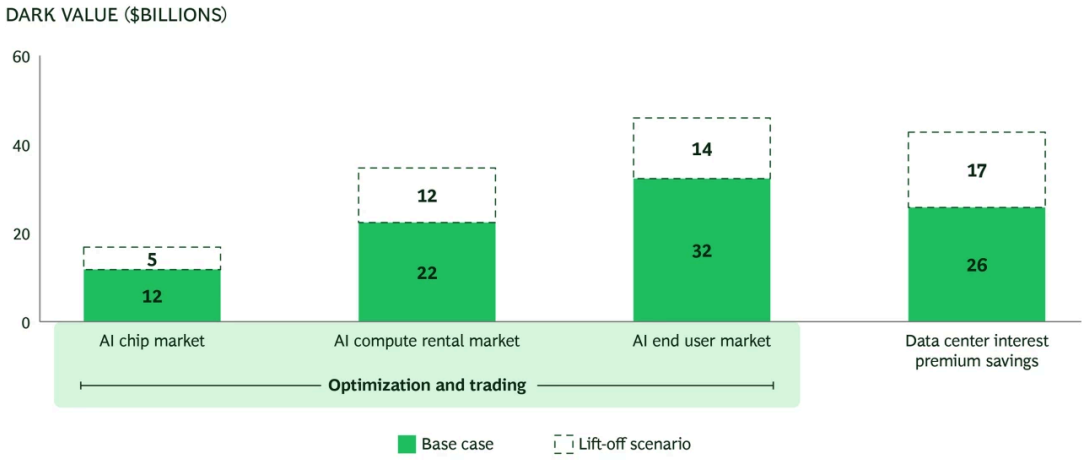
As the market grows more liquid and efficient, it creates opportunities, not only for end users but also for data center players, AI labs, and other large users of AI compute. All told, BCG Institute estimates that this change could unlock as much as $140 billion in annual "dark value"—value that exists in virtually every industry due to inefficiencies in the supply and demand of capital, goods, and services. (See Exhibit 1.) In the AI compute market, dark value is likely to concentrate in two primary areas:

- Price optimization and trading across the three primary components of the market: AI chips —largely graphical processing units (GPUs)—AI compute rental rates for those chips, and end user tokens
- Lower borrowing costs for data center players

Players that move early will be in the best position to adapt to a rapidly changing market and unlock dark value in the AI compute market.

EXHIBIT 1

AI Market Evolution Can Release up to $140 Billion of Value



Sources: LSEG; Goldman Sachs Global Institute; BCG analysis.

The Challenges of Today's Market for AI Computing Power

The market for AI compute is growing at a breakneck pace. BCG analysis projects that it will climb from $360 billion in 2025 to roughly $2.3 trillion in 2030. Yet players across the ecosystem face a number of structural challenges that a more liquid, transparent market could help address.

AI End Users: Increasing Exposure to AI Cost

As pricing approaches evolve, end users are increasingly exposed to the wholesale market for AI compute that undergirds AI models.

End-user tokens are just a way of representing AI compute use. Every time an AI model provider like Anthropic or OpenAI answers a query, it uses AI compute, which has a cost. But until recently, flat subscription pricing largely insulated end users from the true costs of AI. To manage surges in end user demand, AI providers didn't raise prices; instead, they controlled demand directly. When AI compute capacity tightened, providers responded with rate limits, queue delays, or outright throttling. Users experienced worse service but ultimately paid the same flat subscription regardless of conditions.

More recently, many providers have shifted to metered pricing. Anthropic began moving enterprise customers to a $20-per-seat base fee plus pay-per-token API rates in April 2026. The company explicitly tied the change to deepening crunch for AI compute. Similarly, OpenAI moved Codex from flat-message pricing to token metering in early April, GitHub tightened Copilot limits days later, and Windsurf replaced flat credits with hard weekly limits in March.

The next step in the market's evolution could be dynamic token pricing, where the price of inference (the computational process of running a trained AI model to generate a response) moves with the underlying supply of and demand for AI compute, much as wholesale electricity prices rise on hot afternoons and fall overnight.

With tokens, dynamic prices would reflect both demand-side workload characteristics and supply-side conditions. On the demand side, not all inference loads are equally valuable: a real-time customer-facing agent running at 9:00 AM and an overnight batch summarization job should not clear at the same price. On the supply side, available AI compute capacity fluctuates continuously. Buyers that can tolerate latency or can schedule work flexibly will pay less; those that need guaranteed throughput at a specific moment will pay a premium.

This market would work efficiently in a setting where models are substitutable. For example, a developer building a chatbot, a summarization tool, or a coding assistant can often run it on Claude, GPT, or an open-source alternative with comparable results. That keeps competition strong and ties prices closer to the provider's actual cost of AI compute in the moment.

However, not all AI activity is substitutable. Deeper ecosystem integration—such as for custom silicon, fine-tuned models, data residency arrangements, and enterprise support agreements—creates switching costs that remove the substitutability on which spot market discipline depends. In these cases, token prices will reflect negotiated relationship value as much as marginal costs for AI compute, and they will behave less like wholesale electricity and more like a long-term utility contract: stickier, less transparent, and less responsive to real-time supply conditions.

As enterprises scale their AI usage, a more developed market will help them adjust to dynamic pricing by adopting several practical measures:

- **Time flexible workloads strategically.** For substitutable, latency-tolerant jobs, enterprises should treat AI compute as they would any commodity that has predictable price cycles—scheduling batch work, model evaluations, and non-urgent processing to run when spot prices are low.
- **Use futures markets to manage price risk on nonflexible workloads.** In situations where enterprises can’t defer timing but providers remain substitutable, enterprises need visibility into real-time prices for AI compute and into financial instruments to hedge against cost volatility.
- **Scrutinize committed capacity arrangements.** For mission-critical workloads where the cost of switching providers is high, enterprises need to understand what they are actually paying for lock-in. It is crucial to negotiate committed capacity terms with full awareness of their alternatives and to avoid sleepwalking into pricing that is detached from the underlying costs of AI compute.

AI Labs and Other Large Users of AI Computing Power: Risk Management and Price Discovery

AI labs, including Anthropic and OpenAI, need vast amounts of AI compute to train new frontier models and to run end users’ inference workloads. Today’s market structure presents them with two challenges.

First, price discovery is limited. The listed prices that are readily accessible on the websites of AI compute providers rarely match the real transaction price. Often, for AI labs and other large users

of AI compute (such as autonomous vehicle players and biotech companies), providers strike prices bilaterally through negotiated discounts, reserved capacity commitments, and enterprise agreements that happen off-market and go unreported. Other markets have successfully created price discovery by establishing benchmarks. Price reporting agencies in markets such as crude oil collect real transaction data from market participants, standardize it, and publish it as a reference that the rest of the market can price against. A solid benchmark for AI compute would enable AI labs to assess whether they are paying a fair market price. At the same time, it would establish a reference price for futures markets while laying the foundation for price-indexed agreements, where fees track real supply and demand rather than trapping buyers in flat rates that are disconnected from market conditions.

Second, AI labs and other large consumers have limited tools for managing supply and price risk. That's why they are not just buying what they need today, but instead striking long-term bilateral agreements that lock in capacity at fixed or predictable rates well before they know exactly how much they will need.

This race to lock up supply is evident in data center vacancy rates of 1%, with roughly 92% of new capacity already precommitted before it is even built. But the utilization rate for some GPU clusters that enterprise users of AI compute rent or own can be as low as 5%, according to Cast.AI. Utilization is likely much higher at leading AI labs, but even there the issue does not disappear entirely. That can look inefficient from the outside, especially when some capacity is consistently underused. But from the buyer's point of view, the alternative may be worse. Undercommitting could mean not having enough AI compute to train the next model, serve the next wave of users, or keep up with competitors.

Those long-term contracts also help address price risk. According to the Silicon Data H100 rental index, the price of an H100 GPU has risen about 2.2% per month since January 2025. Over that same period, prices fell by as much as 11% at one point and rose to roughly 25% above their starting level at another.

A more developed market would enable two critical types of risk management:

- **Supply Risk Management.** In other markets, companies use long-term bilateral contracts for baseline supply planning, while a liquid short-term market lets buyers flex capacity up or down as demand changes. A secondary market enables this flexibility in both directions: companies that need more supply can buy it, and those with too much can sell the excess. The market for AI compute does not yet work that way.
- **Pricing Risk Management.** In commodities markets, futures allow firms to manage price risk by locking in future prices without physical procurement of commodities such as oil, gas, or wheat. A liquid financial market for AI compute would allow AI labs to better manage price volatility risk by hedging future costs of such power.

Data Center Players: High Borrowing Costs

Until now, the AI data center market, being young, has relied heavily on financing via hyperscaler equity and cash flow. But as the market matures, the need for debt financing is likely to grow quickly. Hyperscalers and neoclouds today borrow at relatively high interest rates to cover their data center investments. Lenders prefer predictable cash flow and minimal performance risk, but AI data center financing today breaks both requirements.

To understand why, consider the assets that secure data center loans today: GPU hardware and customer rental agreements. GPUs have uncertain residual value because they depreciate rapidly as new generations ship roughly every two years. Rental rates in customer contracts, meanwhile, have been volatile, with H100 hourly rates falling from around $8 per hour at a peak in early 2024 to $1.96 per hour by late 2025 before climbing back to $2.64 per hour in April 2026. In addition, these customer contracts often concentrate in one or two counterparties, magnifying the impact of any single renegotiation or default.

The emergence of forward markets for both GPU residual values and rental rates could change the equation by enabling two things:

- **Hedging Residual Value.** Without a market view on future residual value, lenders guess at depreciation and lend conservatively, keeping loan-to-value covenants tight and rates high. Establishing a forward curve for AI chips would permit data center players and lenders to hedge residual risk directly.
- **Hedging Rental Rates.** Rental revenue services a company's debt while the business is running. A liquid forward curve would allow companies to strike standardized contracts, giving them the same degree of revenue certainty that they now garner via large offtake agreements.

Ultimately, the development of such financial tools would translate into lower borrowing costs, enabling data center players to reallocate capital toward building more data center capacity.

The Dark Value Opportunity

As the market for AI compute evolves, data centers, AI labs, and large end users will have access to new mechanisms for managing their investment and spending. Those mechanisms will enable

them to unlock dark value in two areas: arbitrage and lower borrowing costs for data center players.

Optimization and Trading

BCG's dark value research, finds that 3% to 4% of total market size in any industry is available for capture through arbitrage—an opportunity that arises from a market's complexity. Applied to an expected $2.2 to $2.4 trillion AI compute market by 2030, this implies a potential annual value of $66 billion to $97 billion.

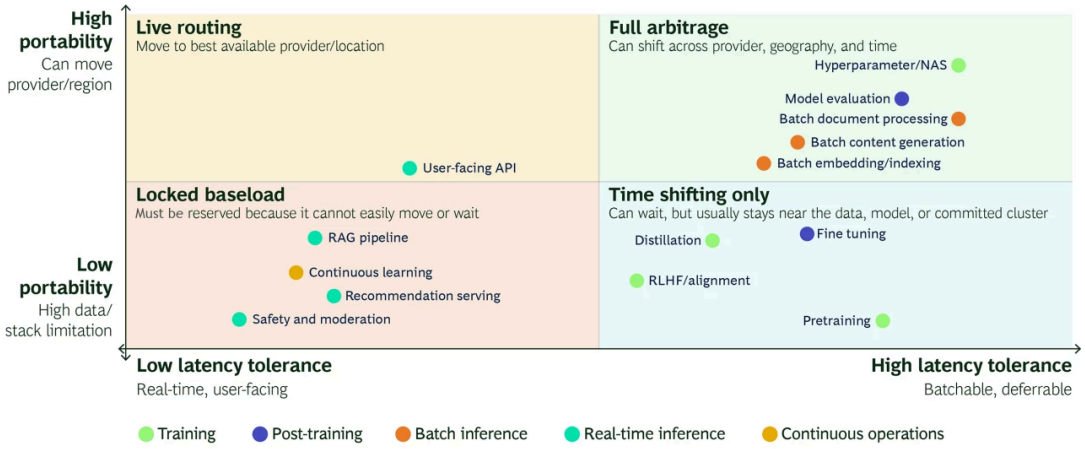
That value is distributed across the market's three primary components:

- **AI Chips.** This segment could yield $12 billion to $17 billion in annual dark value. Data center players would capture much of it.
- **AI Compute Rental.** Rental of this sort is typically measured in GPU-hours. The total annual dark value would come in at $22 billion to $34 billion. AI labs and large AI compute users would be the primary beneficiaries of this value.
- **Tokens Consumed by End Users of AI Applications.** Large enterprise users would be in a position to be well positioned to collect most of the annual dark value of $32 billion to $46 billion in this area.

Our estimate does not assume that everything will be optimized and traded. After all, most buyers of AI compute run AI workloads, and depending on the workload, arbitrage— especially geographic arbitrage—may not be feasible in practice for many of them. Some workloads cannot be transferred to a data center in another region, owing to cost constraints (data transfer fees can erase savings) or to the need for real-time turnaround. (See Exhibit 2.) This constraint may become less relevant over time, however, as companies such as Lumen Technologies advance AI infrastructure to support high-capacity, low-latency networking for moving massive AI data sets.

EXHIBIT 2

Arbitrage Opportunity Depends on Workload Portability and Latency Requirements



Sources: BCG analysis.
Note: API = application programming interface; NAS = neural architecture search; RAG = retrieval-augmented generation; RLHF = reinforcement learning from human feedback.

Reduced Borrowing Costs

To estimate the potential financing savings for data center players, we drew on loan data from LSEG/Refinitiv. First we volume-weighted investment-grade loans to calculate the average spread over the secured overnight financing rate (SOFR). Then we analyzed a subset of data center companies, including CoreWeave and xAI, to estimate spreads specific to AI infrastructure.

Assuming a $3.6 trillion investment in AI infrastructure from 2026 through 2030, we compared interest costs with no forward curve for AI compute versus an environment with a reliable forward curve. Under both scenarios, we held the financing structure constant: 70% loan-to-value, 4.5-year amortization (the median investment-grade tenor), and SOFR at 3.6%. (See Exhibit 3.)

EXHIBIT 3

Lower Interest Rates Can Save Data Center Players $116 Billion over the Next 4.5 Years

	Without forward curve	With forward curve	Difference
All-in interest rate	**6.43%**	**4.75%**	**1.68%**
Interest as a percentage of principal	**17.7%**	**13.1%**	**4.6%**
Interest across $3.6 trillion buildout	**~$446 billion**	**~$330 billion**	**$116 billion**

~$26 billion saved per year, redeployable into added capacity

Source: BCG analysis.
Note: Assumptions: $3.6 trillion buildout (2026–2030); 70% loan-to-value; 4.5-year amortization; secured overnight financing rate at 3.6%.

Companies could deploy the total savings of $116 billion, about $26 billion annually, toward adding more capacity. If AI infrastructure investment comes in at the high end of analysts' projections—$6 trillion from 2026 through 2030—the reduction in borrowing cost would exceed $40 billion annually.

Headwinds to the Development of a Liquid AI Computing Power Market

Companies have made some moves to evolve the market. Amazon Web Services, for example, launched EC2 Spot Instances in 2009 as a bidding market for unused capacity, but it later shifted to a model based on long-term supply and demand for spare capacity. More recently, independent venues such as SF Compute have pushed the idea further by building a marketplace for GPU clusters with resale built in.

However, several factors have made such efforts—and the overall shift toward a more efficient market—challenging:

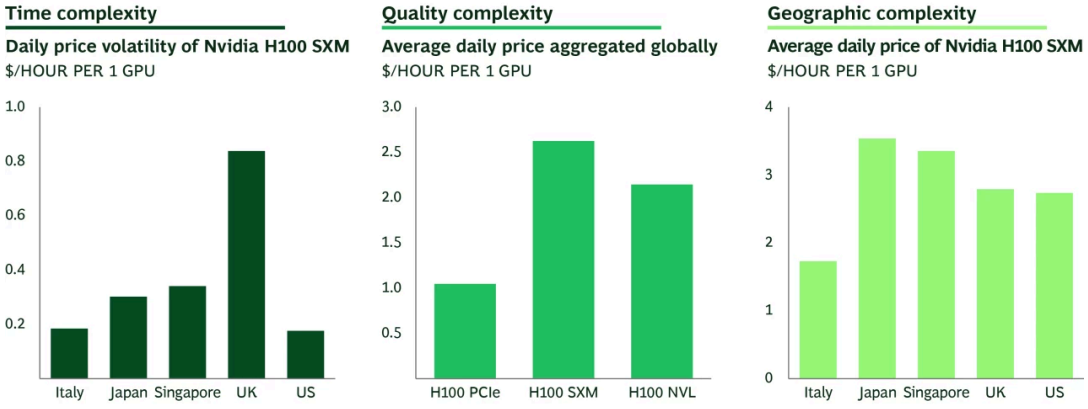
- **Technical Complexity.** AI compute differs along three dimensions. First, it can be cheaper or more available at certain times (for example off-peak hours, periods of low demand, seasonal lulls, or demand spikes immediately after the launch of a new AI model). Second, it can vary in quality. An H100 does not deliver the same performance as an H200. And non-GPU expenses such as CPU computing power, networking, storage, orchestration software, and support are often hidden, as are provider-specific factors such as downtime, setup time, debugging time, and the performance tuning required for networking and storage. Third, AI compute costs and availability differ by geographic location. (See Exhibit 4.)
- **Lack of Consensus Benchmarks.** A trusted benchmark for AI compute would enable all players across the ecosystem to compare offers, evaluate GPU-backed assets, or build the financial products to hedge price risk. In mature commodity markets such as oil, benchmarks may differ but they still give participants a common view of price. Benchmarks for AI compute are beginning to emerge, including the Ornn H100 Index and the Silicon Data H100 Rental Index, among others. But each benchmark uses a different methodology to collect data and normalize pricing on the basis of heterogeneous AI compute. As a result, the same GPU on the same day can fetch significantly different prices across indices. (See Exhibit 5.)
- **Capital Availability.** Although many AI labs and AI services companies are not yet highly profitable, equity funding is still available; and although debt is costly, they can often find

alternatives. As a result, managing price risk has only recently become a major concern.

- **Data Center Player Incentives**. Data center players have historically leaned on long-term bilateral contracts and have had little reason to standardize pricing or to support open benchmarks. Going forward, however, some data center players may see strategic value in greater market transparency if it reduces their funding costs and helps them persuade enterprise customers to switch to them from competitors that price more opaquely.

EXHIBIT 4

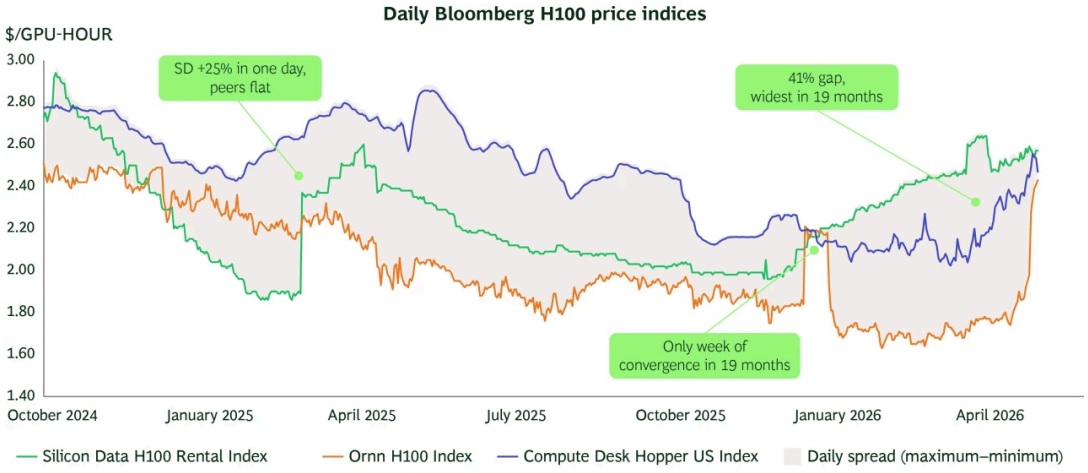
Prices for AI Computing Power Differ Significantly Across Regions, GPU Type, and Time



Sources: AWS; GCP; Vast.ai, Runpod; Lambdalabs; Nebius; Cloudrift; Verda; BCG analysis.
Note: Data used is from January 13, 2026, to April 8, 2026. Price is for spot instances. GPU = graphical processing unit.

EXHIBIT 5

Three H100 Indices Differ Markedly over Time



Sources: Bloomberg daily SDH100RT (Silicon Data), ORNN H100 (Ornn Compute), and CIHOPUS (Compute Desk Hopper US H100/H200). October 1, 2024, to May 5, 2026; n = 575 common trading days.

Lessons From Other Commodity Markets

A number of the factors that complicate the development of a more transparent and liquid AI compute market are present in other markets as well. To understand how this market could evolve, it helps to take a look at two other markets: oil and power. (See Exhibit 6.)

EXHIBIT 6

How an Evolved AI Market Could Compare to Oil and Power Markets

Dimension	Oil	Power	AI compute (what it could look like)
Balancing market How short-term imbalances are managed	**Inventories and logistics** Imbalances are absorbed through onshore storage, floating storage, and refinery adjustments	**System balancing** The day-ahead auction sets the baseline; intraday trading, ancillary services, and TSO grid balancing fine-tune the supply in real time	**Schedulability and elasticity** Load management could involve queueing and scheduling jobs, preempting and checkpointing them, and moving them across regions
Physical market Short-term trading and physical delivery	**Spot cargo market** Prompt physical barrels trade on a cargo basis under defined delivery terms and inform benchmark grade assessments	**Day-ahead market** An hourly auction clears each day at zonal prices, with power delivered physically through the grid	**PAYG, Spot, next-day blocks** Buyers could take on-demand or preemptible capacity, or reserve next-day and short-window blocks with defined hardware, region, and SLA
Long-term contracts Bilateral agreements for supply and capacity	**Term supply agreements** Producers and buyers sign offtakes at fixed or indexed prices, setting volume and quality terms	**Term supply agreements** Utilities and generators contract at fixed or indexed prices via renewable and firm PPAs	**Reserved capacity and offtakes** Players could commit through reserved instances, dedicated clusters, or private cloud
Financial market Hedging, price discovery, and investment	**Futures, swaps, options** Traders hedge and speculate, using futures alongside spreads, calendar swaps, and options	**Futures, swaps, options** Monthly, quarterly, and yearly futures trade occurs by location and product (base or peak), settling financially or physically	**Futures, swaps, options (emerging)** Cash-settled contracts on standardized AI compute indices could be split by region, accelerator class, workload, and term
Settlement mechanism and reference anchor How value is measured and contracts are settled	**Benchmark administrator** PRAs such as S&P Global Platts and Argus anchor value, with financial contracts cash-settled against them	**Exchange spot price** Exchanges with physical delivery—such as Nord Pool, EPEX, and PJM—set the reference price and settle against delivered power	**Benchmark administrator** PRAs could normalize workload performance across standardized specs, settling in cash with capacity rights over time

Source: BCG analysis.
Note: PPA = power purchase agreement; PRA = price reporting agency; SLA = service level agreement; TSO = transmission system operator.

Oil

Like AI compute, oil is heterogeneous, with crude varying in chemical composition, sulfur content, and viscosity, as well as by origin, shipping costs, and delivery logistics. The oil market solves this issue through benchmarks. Price reporting agencies (PRAs) collect real transaction data from market participants, standardize it, and publish it as a reference for the rest of the market to price against. Although Brent and WTI together account for only a small percentage of global crude output, they serve as the pricing benchmark for the rest of the market.

AI compute differs significantly from oil in terms of portability. Physical commodity markets use logistics—tankers and pipelines in oil—to link regions and to enable arbitrage that minimizes

price differences. Oil moves easily, so a barrel in Houston and a barrel in Rotterdam are priced within a few dollars of each other and effectively share a global price.

Power

Like AI compute, power is inherently regional. In many cases, power cannot easily move from where it is produced to where it is needed, so prices vary by location. Meanwhile, AI compute workloads must operate where the data is, so a GPU in Virginia is not interchangeable with one in Frankfurt.

Power illustrates how, under certain conditions, a regional commodity can still function efficiently. Three enablers keep regional price differences anchored to real costs:

- **Physical Interconnection.** When one region's price rises too far above another's, power flows from the cheaper region to the more expensive one. The added supply pushes the high price down, the reduced supply pushes the low price up, and the gap converges toward the cost of transmission.
- **Price Transparency.** Real-time prices are published at every node, so suppliers know where to sell and buyers know where to source.
- **Financial Markets.** Traders can bet on regional price differences narrowing by buying in cheaper regions and selling in more expensive ones. This activity pushes prices toward each other and helps eliminate unjustified spreads, even without moving physical power.

These mechanisms operate only within connected systems. Regions without interconnection, such as California and New Zealand, can diverge freely. But within a connected grid, price differences largely indicate real congestion and losses rather than opacity.

AI compute won't converge to a single global price, just as power doesn't. But AI compute can evolve into an efficient regional market complete with a series of supply and risk management solutions. (See the sidebar, "Building Blocks of a Liquid AI Computing Power Market.")

– Building Blocks of a Liquid AI Computing Power Market

An efficient and liquid AI compute market could include a number of valuable tools and solutions:

- **PRAs to Create Standardized Benchmarks.** PRAs would collect and normalize bilateral data into a reference price per GPU-hour that the rest of the market could then use as a reference. The benchmark should not to be

too general, as would be the case with a global H100 benchmark when AI compute is a regional asset. PRAs could also develop benchmarks for other elements of the market, such as AI chips or tokens.

- **Long-Term Indexed Contracts.** Volume-based commitments exist today, but most are set at fixed rates. In oil markets, a typical term contract is priced as "monthly average Dated Brent + $2.50/bbl," rather than as "$80/bbl fixed for three years." The buyer secures the volume, but the price moves with the market. Applying this model to AI compute, companies might agree to price a long-term contract at "OCPI H100 + provider differential," which would allow data center players to raise long-term capital against indexed revenue while protecting AI labs and other AI compute buyers from overpaying if GPU prices fall.

- **Financial Derivatives Tied to Benchmarks.** Once a trusted benchmark is in place, swaps, futures, and options can refer to it directly. This enables hedging of both rental rates and residual values, which in turn is a key step toward compressing the credit spread on AI data center debt closer to investment grade.

- **Cash Settlement Against a PRA Benchmark as the Link Between Physical and Financial Markets.** Moving GPU-hours between buyer and seller to settle contracts is impractical, so instruments must cash-settle against a reference price. The PRA's normalization, adjusted for GPU region and provider quality, enables accurate settlement across heterogeneous transactions. This mirrors oil markets, where ICE Brent and NYMEX WTI futures settle against Platts and Argus benchmarks rather than physical delivery.

- **Secondary Markets.** Today, buyers that reserve more capacity than they ultimately use or complete workloads earlier than expected have little ability to recover value from unused commitments apart from reselling the capacity to original seller. A secondary market, similar to airline seat resale, would allow reservation holders to resell firm capacity to the larger market, whether to recoup costs or to profit from favorable price movements. It would also strengthen the connection between long-term contracted capacity and near-term spot pricing, making the benchmark more robust by expanding the pool of observable transactions.

Since players are already managing supply and price risk through tools such as long-term contracts, this market evolution is likely to take place in bursts over time, rather than all at once. Whether it ultimately converges on a fully commoditized market remains uncertain. Clearly, however, each stage of that evolution, greater dark value capture becomes possible.

Recommendations for Players Across the Ecosystem

We believe that all stakeholders need to prepare for a world of increasingly dynamic pricing, greater transparency, and a more sophisticated financial layer around physical AI compute. The specific actions that each stakeholder will take, however, differ.

AI End Users

As AI labs shift toward metered and dynamic pricing, large AI end users can take three actions to manage their cost:

- **Tier demand into separate baseload and flexible categories.** AI end users should identify workload tiers—including high-value, customer-facing applications (baseload demand) that must be served regardless of price—and background tasks (batch summarization, internal research, and report generation) whose requirements are more flexible.
- **Hedge baseload demand.** Enterprises with large and growing AI cost lines should treat AI compute the same way they treat energy or foreign exchange exposure: by hedging systematically where doing so is material to earnings.
- **Throttle flexible demand.** End users can throttle, delay, or reroute flexible workloads when prices spike. In a price shock, companies should be in a position to scale down flexible consumption while preserving baseload.

AI Labs and Other Large AI Computing Power Users

AI compute is the single largest cost line for most of these players, often consuming a substantial share of early-stage capital. Managing that exposure actively, rather than treating it as a fixed input, is a strategic necessity. Several strategic steps are available:

- **Structure contracts to match exposure.** As benchmarks become trusted, buyers should push for indexed pricing (for example, “OCPI H100 + differential”) rather than fixed rates for long-term commitments. At minimum, buyers should use existing benchmarks to pressure-test and negotiate fixed-rate agreements so they aren't flying blind in their efforts to gauge fair value.
- **Manage AI compute exposure across the contract life cycle.** Long-term fixed-rate contracts reduce price volatility during the term but introduce other risks, including renewal cliffs, exposure to spot pricing for overages, and potentially being locked into paying above market prices if overall prices fall. As benchmarks and derivatives mature, financial instruments tied to indices can hedge renewal risk, cap overage costs, and offset losses on above-market contracts.
- **Arbitrage across time, quality, and geography.** As switching costs fall and AI chip substitutability increases, AI labs and other AI compute users should prepare to route workloads on the basis of price signals. Batch jobs can run overnight or in lower-cost regions; non-latency-sensitive training can move to off-peak windows.
- **Optimize model and provider selection.** Substitutability across frontier models creates real optionality. Routing each workload to the provider and model that offer the best price-performance at a given moment, rather than standardizing on a single vendor, can capture savings with minimal loss of quality.

Data Center Players

Data centers sit at the supply end of the market and stand to benefit most directly from the development of forward curves and financial instruments. They should make the following moves:

- **Use forward curves to lower financing costs.** If residual-value and rental-rate benchmarks gain liquidity, data center players should incorporate them into underwriting. They can present lenders with rental revenue and AI chip residual value projections on the basis of forward curves rather than bottom-up forecasts, compressing credit spreads toward those savings.
- **Hedge residual value and rental rate risk.** Another smart step is to transfer the two core balance-sheet risks (hardware depreciation and rental rate volatility) to financial counterparties as derivatives markets develop. By hedging in this way, data center players can offset their losses if GPU prices fall or rental rates drop.

The market for AI compute is changing quickly, with benchmarks, futures contracts, and secondary markets already beginning to emerge. Whether the market moves toward true commoditization is a question mark. But the shift toward greater liquidity and transparency is unstoppable. Early movers—whether end users managing cost exposure, data center players unlocking cheaper financing, or AI labs securing supply—stand to capture disproportionate value as this shift unfolds. The era of opaque contracts and flat-rate pricing is ending. The question now is who will seize advantage in the new landscape.

Authors

Antti Belt

Managing Director & Senior Partner; BCG Institute Fellow
Helsinki

Allen Thomas

Consultant
Chicago

ABOUT BOSTON CONSULTING GROUP

Boston Consulting Group bridges the gap between ambition and outcomes for the world's leading companies and organizations. We are built for this era of unprecedented change — bringing strategic clarity rooted in over 60 years of deep domain knowledge, combined with applied AI shaped by our practitioners. BCG works shoulder-to-shoulder with CEOs across industries and geographies to deliver transformative impact at scale: stronger returns, transferred capabilities, and change that sticks. For more information, visit bcg.com.