

Compute, Power and the Rise of AI Infrastructure Markets

Building Price Discovery and Risk Management for the AI Economy

Udes Jha Pranav Jha

Whitepaper

October 2026

Abstract

Spending on AI infrastructure has grown much faster than the markets that might price it. Big Tech's AI capital expenditure is on course for roughly \$725 billion in 2026, most of it on GPUs and the data centers that house them, yet a GPU-hour still has no widely accepted benchmark price, no liquid forward curve, and few instruments with which a buyer or seller could hedge it. Electricity is in a similar position. The U.S. Department of Energy puts data-center consumption at 176 TWh in 2023 and as much as 580 TWh by 2028, and by June 2026 ERCOT had received about 475 GW of large-load interconnection requests, nine tenths of them from data centers. In this paper we argue that compute and power should be treated as a single market with two legs, much as natural gas and electricity are joined by the spark spread, and we describe the order in which Kalshi is building it. Prediction markets on GPU rental prices have traded on Kalshi since March 2026 and produced the first market-implied GPU forward curve in July. Perpetual futures on compute would run on the systems Kalshi already operates for its crypto and precious-metal perpetuals, and dated futures would follow for participants whose exposure falls on known dates. Alongside these we propose regional power contracts in PJM, ERCOT, CAISO and MISO, traded and cleared around the clock, since data centers do not shut down at weekends.

Keywords: Compute Markets, GPU Pricing, Electricity Markets, Perpetual Futures, Prediction Markets, Kalshi

1 Introduction

For most of the past two decades a company's compute bill was an IT expense. The firm bought servers or rented them from a cloud provider, and the price it paid rarely appeared in anyone's risk report. Large-scale AI changed that. A frontier training run occupies a very large cluster of accelerators for months, a successful product launch can multiply inference demand within days, and a lab that cannot get the newest chips may ship after its competitors. The hourly price of a GPU now drives investment decisions worth billions of dollars, but it is set mostly in private negotiation, and a company that wants to fix next year's price has very few ways of doing so.

We expect the market that forms around this exposure to have two legs. One is compute, which is already sold by hyperscale clouds, specialist GPU providers and marketplaces and through bilateral contracts, at prices that differ by chip, region, contract length and how quickly the capacity is needed. The other is electricity. Data centers have become large enough to move regional load forecasts and wholesale prices, and the price of power in turn affects where new capacity is built and what it costs to run. Commodity traders will recognise the arrangement. Gas and power are linked by the spark spread, crude and refined products by the crack spread, and corn and ethanol by the crush spread.

Kalshi is building the compute leg in three stages.

1. **Prediction markets** let traders take positions on compute prices, data-center growth and power demand before anyone has agreed on a spot price. Kalshi first listed GPU compute markets in March 2026, and published the first market-implied GPU forward curve in July 2026. Several million fully collateralized contracts have traded in the space already.
2. **Perpetual futures** give exposure to the current price of compute without owning GPUs or rolling contracts. Kalshi already runs regulated perpetuals on crypto and on precious metals.
3. **Dated futures** give exposure to particular months and years. These are the contracts AI companies, GPU operators, developers and lenders would use to hedge in the way commodity producers and consumers already do.

The power leg is being developed at the same time. We begin with four grid regions, PJM, ERCOT, CAISO and MISO, which between them capture most of where U.S. data-center load is going. Because data centers run day and night, the contracts include hourly as well as daily-average prices, and they trade and clear continuously. Figure 1 shows the sequence and what is live as of October 2026.

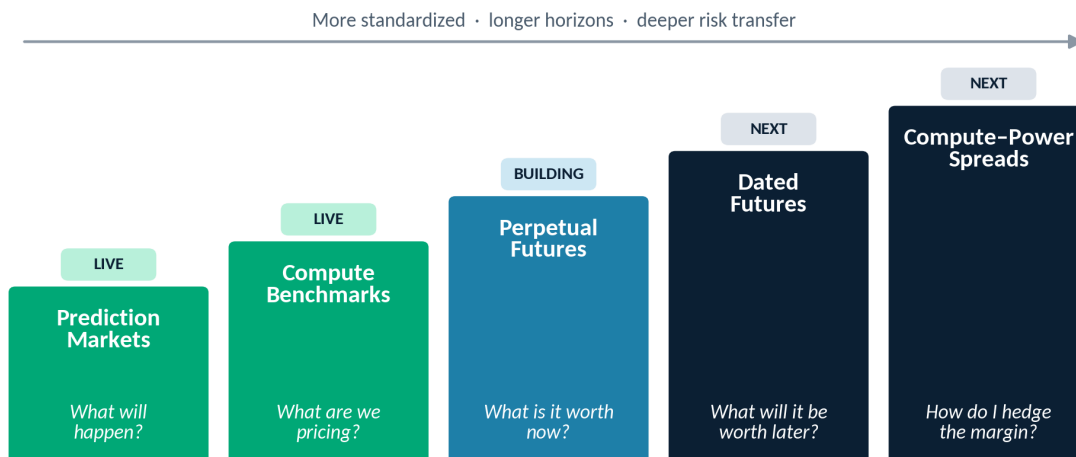


Figure 1: Stages in the development of a compute market, and the status of each at Kalshi as of October 2026. Compute perpetuals and dated futures are subject to regulatory approval. Source: Kalshi.

Section 2 describes how GPU compute is bought and sold today, and Section 3 what determines its price. Section 4 covers the growth of data-center electricity demand. Section 5 sets out the three stages in more detail, Section 6 explains the choice of regional power markets, and Section 7 the case for trading them around the clock. Section 8 links the two legs through what we call the Silicon Spark Spread. Section 9 covers clearing and risk management, Section 10 the roadmap, and Section 11 concludes.

2 The Market for GPU Compute

Compute shares several traits with physical commodities. It comes out of plants that cost billions and take years to build, so supply responds slowly to price. Location affects its value, and shortages occur. A buyer of capacity and a seller of it also lose money on opposite price moves, which is the basic condition for a hedging market to exist.

2.1 Not All GPU-Hours Are Equal

Few commodities are a single product. CME lists separate futures for WTI at Cushing, Houston and Midland as well as for Brent (CME Group, n.d.). GPU time is similarly varied. An hour on an H100 and an hour on a B200 are different things, and the price of either depends on memory, bandwidth, interconnect, cluster size, software support and where the machine is located. Table 1 lists the accelerators that account for most of the market today.

Chip	Maker / family	Memory	Max power	Market role
A100	NVIDIA Ampere (2020)	40–80 GB	400 W	Older workhorse, used for cheap inference and fine-tuning
H100	NVIDIA Hopper (2022)	80 GB	700 W	Standard training chip and the most widely quoted GPU-hour
H200	NVIDIA Hopper (2024)	141 GB	700 W	An H100 with more memory, popular for large-model inference
B200	NVIDIA Blackwell (2025)	192 GB	≈1,000 W	Current frontier chip; the scarcest and most volatile
B300 / GB300	NVIDIA Blackwell Ultra	288 GB	≈1,400 W	Sold mainly as 72-GPU racks
RTX 5090	NVIDIA consumer/pro	32 GB	575 W	Cheap inference and smaller workloads
MI325X / MI355X	AMD Instinct	256–288 GB	1,000–1,400 W	Main alternative to NVIDIA, with large memory

Table 1: The main AI accelerators. Power figures are maximum board power for the SXM and OAM versions. Sources: NVIDIA and AMD product specifications (NVIDIA, n.d.; AMD, n.d.); Kalshi.

That variety shaped our first listings. In March 2026 Kalshi opened rental-price markets on five chips, the A100, H100, H200, B200 and RTX 5090, each settling on the Ornn compute price index. Rental rates for these chips had until then been negotiated mostly in private, and the index gives traders a published figure to settle against (Kalshi, 2026a; Fortune, 2026). We did not wait for an agreed standard unit of compute before listing, and we doubt one is needed. In most commodity markets, standard grades emerged after trading had already begun.

Power draw per chip has also risen with each generation, from 400 W for the A100 to about 1,400 W for the B300 and MI355X (Figure 2). The newest hardware is therefore the most exposed to electricity prices in absolute terms.

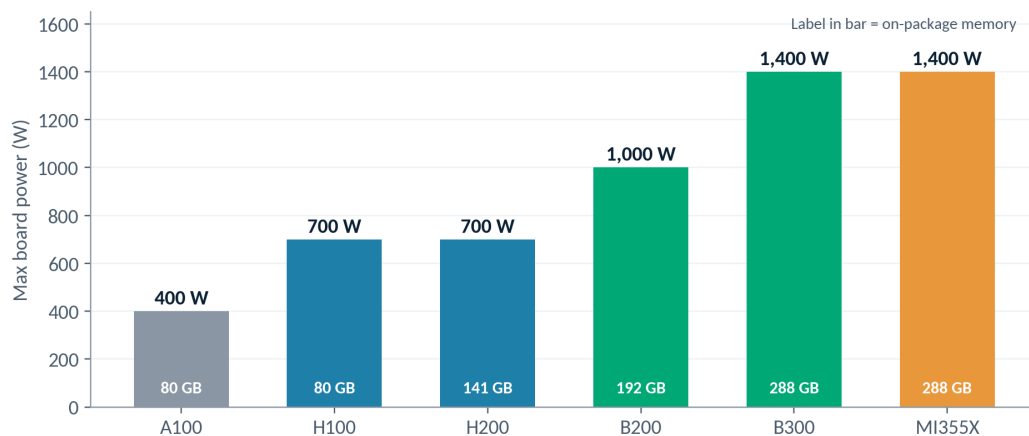


Figure 2: Maximum board power by accelerator, with on-package memory shown in each bar. Source: NVIDIA and AMD product specifications.

2.2 The Unit of Compute Is Getting Larger

Until recently capacity was discussed in terms of individual GPUs. New capacity is now designed and sold as complete systems. NVIDIA’s GB200 and GB300 NVL72 racks wire 72 GPUs together into a single machine that draws roughly 130 to 155 kW (SemiAnalysis, n.d.). The largest campuses are measured in hundreds of megawatts, and a few reach a gigawatt or more. Figure 3 follows the chain from one accelerator to the GPU-hour that is eventually sold.

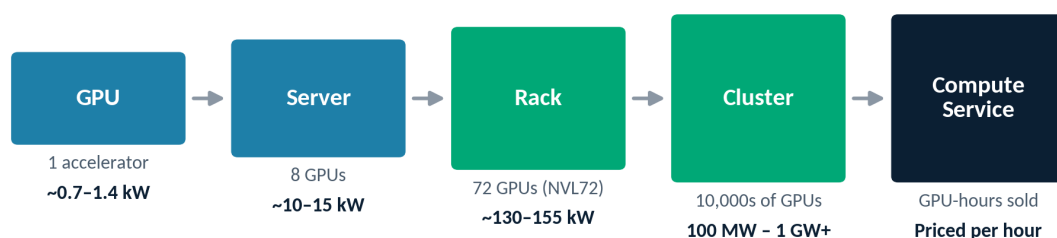


Figure 3: From a single accelerator to a priced compute service, with indicative power draw at each level. Source: NVIDIA system specifications.

2.3 Why We Begin with Individual Chips

Futures markets do not need a perfectly uniform underlying. Oil trades by grade, power by delivery location, and freight through standardized route indices. What they need is an exposure that is clearly defined and measured the same way each time. The awkward feature of compute is how fast the product changes. NVIDIA and AMD release a new generation every year or two, and any definition of “a unit of compute” fixed today could be out of date before a two-year contract expired. We therefore start with prices for named chips. Once several of those trade, the gaps between them carry information of their own. The H200/B200 spread, for instance, reflects what the market thinks about relative performance and supply, and about how quickly the older chip is losing value.

3 What Sets the Price of Compute

Figure 4 breaks the hourly price of compute into six components. The weights are illustrative and vary a good deal by chip, site and contract. Capital recovery on the hardware is the largest, at roughly 42 percent, followed by power and cooling at about 16 percent, facility and network costs at 14, financing at 12, and operations and a scarcity premium at about 8 each. The scarcity premium is small on average, but it moves faster than anything else in the stack and can jump in a shortage.

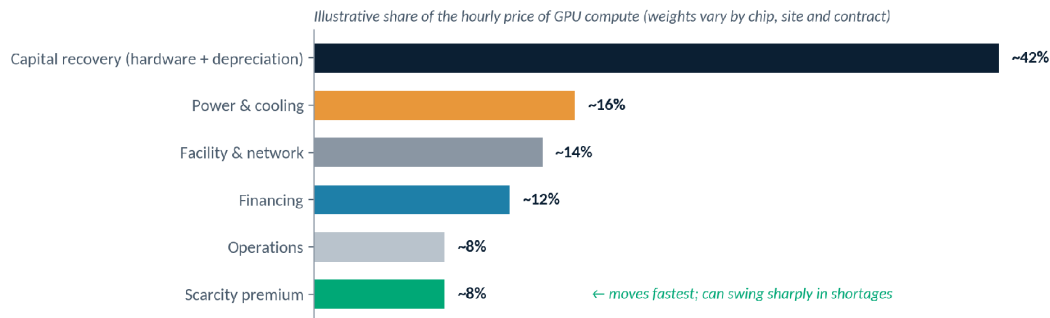


Figure 4: Illustrative composition of the hourly price of GPU compute. Weights vary by chip, site and contract. Source: Kalshi estimate.

Obsolescence. Since hardware is the largest cost, the rate at which it loses value matters most. GPUs lose economic value much faster than most industrial equipment. A chip can go on working for years after a newer generation has overtaken it on speed, memory or energy efficiency, and its rental rate falls as that happens.

Perishability. Compute cannot be stored. An hour of GPU time that goes unsold is simply lost, so utilization is central to an operator's economics, and even brief mismatches between supply and demand can move prices a long way.

Power, financing and scarcity. A single accelerator can draw more than a kilowatt and a campus several hundred megawatts, so electricity is a material cost. Operators also need to earn a return on the capital tied up in their hardware, which is why financing appears as a separate line. When a chip is scarce, buyers who need capacity immediately will pay well above what it costs to supply.

Taken together, these forces imply that compute prices should change over time in ways traders can anticipate, and the forward curves bear this out. Figure 5 reproduces the market-implied curves from our July research note (Kalshi Research, 2026). The A100, H100 and H200 curves are close to flat through 2027. The B200 curve slopes downward, from about \$7.20 to about \$5.50 per GPU-hour, which is how traders are pricing the expected increase in Blackwell supply.

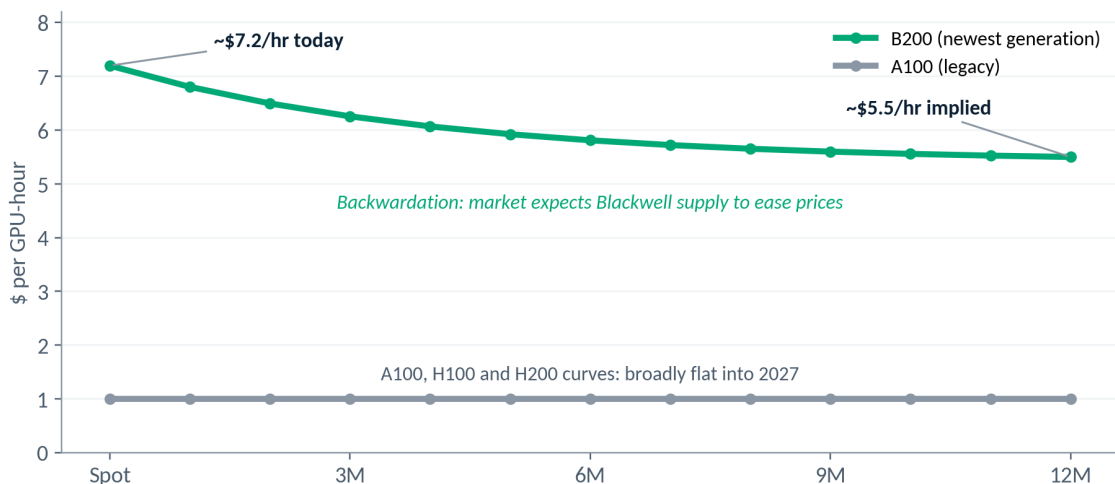


Figure 5: Market-implied GPU forward curves by chip, redrawn and simplified from Kalshi Research (2026). Levels are approximate.

With several generations trading at once, the curves also imply a rate of obsolescence. A procurement manager can then ask what an H200 hour will be worth next year relative to a B200 hour and read the answer from market prices instead of a vendor’s quote.

4 Power

Data centers turn electricity into computation, and they now use enough of it to change how grids are planned and priced. Table 2 gathers the figures used in this section.

Indicator	Value	Source
Share of U.S. electricity used by data centers, 2023	4.4%	Shehabi et al. (2024)
Projected share, 2028	6.7–12%	Shehabi et al. (2024)
Data centers’ contribution to PJM’s 5,250 MW peak-load forecast increase, 2027/28	≈5,100 MW	PJM Interconnection (2025)
PJM capacity price, 2027/28 Base Residual Auction (at cap)	\$333.44/MW-day	PJM Interconnection (2025)
ERCOT large-load interconnection requests, June 2026	≈475 GW	POWER Magazine (2026)

Table 2: Indicators of data-center load on the U.S. grid.

4.1 Data-Center Load on the Grid

The Department of Energy’s most recent estimate puts U.S. data-center consumption at about 176 TWh in 2023, or 4.4 percent of national electricity demand. Its projection for 2028 runs from 325 to 580 TWh, between 6.7 and 12 percent of the total (Shehabi et al., 2024). The width of that range is itself a sign of how uncertain the outlook is. Even the low end, though, would make data centers a significant factor in generation investment, transmission planning and power prices.

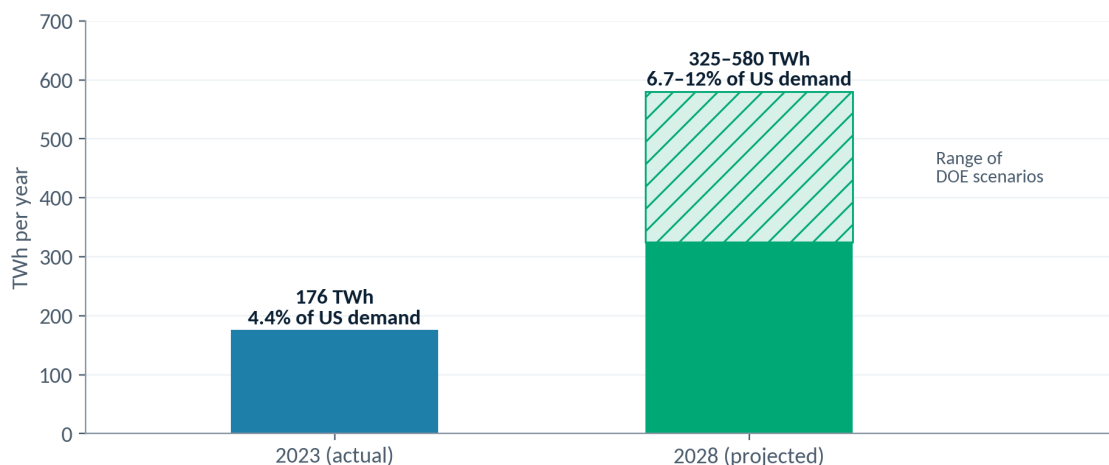


Figure 6: U.S. data-center electricity consumption in 2023 and the projected range for 2028. Source: U.S. Department of Energy and Lawrence Berkeley National Laboratory (Shehabi et al., 2024).

Grid operators are already planning around this. Of the 5,250 MW increase in PJM’s 2027/28 peak-load forecast, nearly 5,100 MW comes from data centers, and the capacity auction for that delivery year cleared at its cap of \$333.44 per MW-day (PJM Interconnection, 2025). ERCOT’s large-load interconnection queue stood at 474.7 GW in June 2026, of which 420.8 GW was identified as data centers (POWER Magazine, 2026). The highest demand the Texas grid has ever served is 91,089 MW, set on 22 July 2026. Most of the queued projects will never be built, but the queue is still more than five times that record (Figure 7).

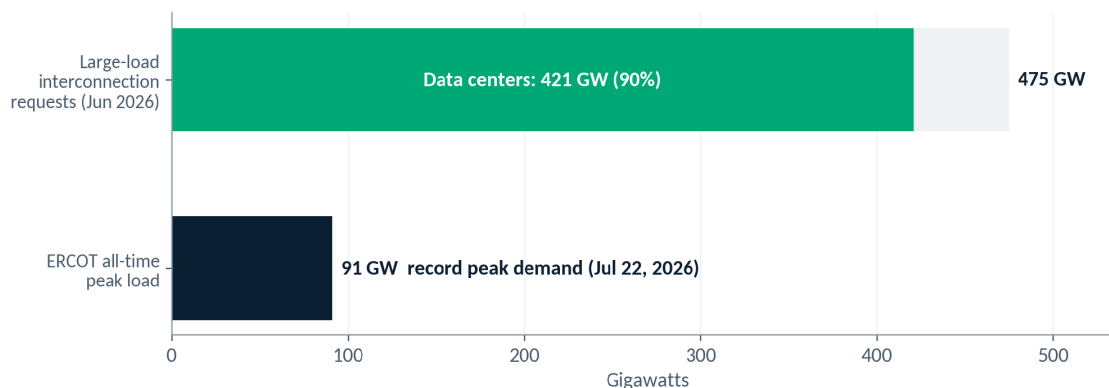


Figure 7: ERCOT large-load interconnection requests as of June 2026, compared with the grid’s record peak demand on 22 July 2026. Not all requested projects will be built. Sources: ERCOT large-load interconnection reporting; POWER Magazine (2026).

4.2 Cause and Effect in Both Directions

Power used to enter a data center’s economics as one cost among several. It now works both ways. Data-center growth is one of the main reasons power demand is rising, and the resulting prices and interconnection delays affect where new compute is built and what it costs (Figure 8).

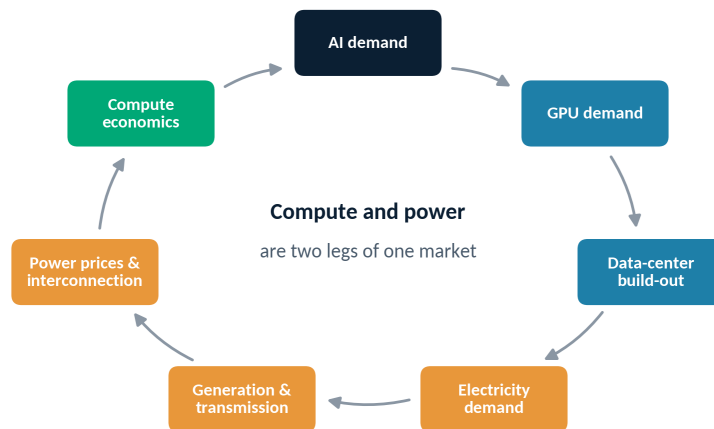


Figure 8: Feedback between AI demand, data-center construction, electricity demand and compute economics. Source: Kalshi.

An operator is in effect long compute revenue and short electricity. If it can hedge only one of the two, half of its margin is still exposed, and that is why we are building contracts on both.

5 The Three Stages in Detail

Futures markets usually grow out of an established cash market with an accepted spot price. Compute does not yet have one, and we see no reason to wait for it. Figure 9 shows the order in which we are building.

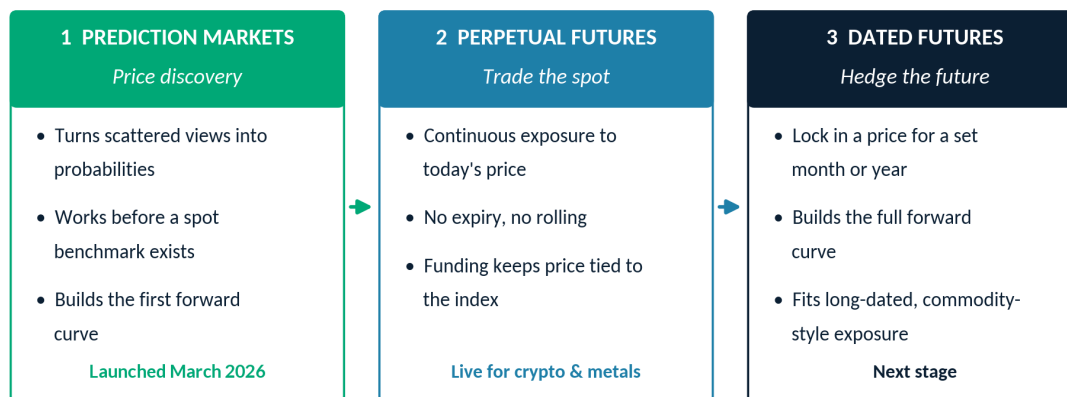


Figure 9: Prediction markets, perpetual futures and dated futures, in the order Kalshi is building them. All derivatives contracts are subject to regulatory approval. Source: Kalshi.

5.1 Stage One: Prediction Markets

Prices can be discovered before a spot market matures, provided there are well-defined questions that can be settled, and plenty already exist. Will the average H200 rental price in August be above \$5 an hour? Will the B200–H200 spread widen this quarter? How many U.S. data centers

will be active at the end of the year? Will Congress pass a bill protecting utility customers from data-center costs? None of these needs a continuous compute index to settle.

The information needed to answer them is spread across firms that rarely share it. Chipmakers see order books, cloud providers see utilization, developers see interconnection queues, power traders see load growth, and AI labs know what they intend to buy. A prediction market pools those views into a probability, and a set of strikes across several expiry dates turns the probabilities into a curve.

That is roughly what has happened on Kalshi this year. The first GPU markets opened in March 2026, settling on the Ornn index. In July we published the first market-implied GPU forward curve, built from weekly and monthly contracts, and made block trading available to hedgers (Kalshi, 2026a; Bloomberg, 2026). Related markets cover the number of active U.S. data centers, which traders expected to reach 5,120 by year-end against 4,700 in June, along with tariffs on critical minerals, chip-export rules and disruption in the Strait of Hormuz, all of which bear on the cost of compute (Kalshi Research, 2026).

A prediction-market price is a probability. It cannot serve as the settlement price for a future, and we are not suggesting it should. The two instruments do different jobs. Prediction markets show participants what others expect, and futures let them protect themselves against what actually happens.

5.2 A Benchmark Participants Can Trust

Continuous derivatives need a reference price that participants trust. At present a GPU-hour is quoted by clouds, specialist providers, marketplaces and private contracts, each on its own terms. We think a compute benchmark should do what the most widely used financial benchmarks do under the IOSCO Principles for Financial Benchmarks, which is to publish its inputs and its methodology. Over time it should also lean more on traded prices and executable quotes and less on indicative ones, the shift interest-rate markets made in moving from LIBOR to SOFR. Table 3 lists the choices a benchmark administrator has to make.

Design choice	What it determines
Eligible hardware	Which GPU generations and system configurations are included
Contributors	Whose prices are used, and the minimum number of contributors
Observation window	When prices are sampled and how they are averaged
Outlier and stale-price rules	How erroneous or out-of-date quotes are removed
Fallbacks	What happens if a major source stops publishing or data become thin
Input hierarchy	The priority given to traded prices, executable quotes and indicative quotes
Transparency	What is published: inputs, methodology, and how changes are governed

Table 3: Design choices a compute benchmark must settle. Source: Kalshi.

Several compute indices already follow many of these practices, and we expect them to tighten as trading volume grows.

5.3 Stage Two: Perpetual Futures

A perpetual future tracks an index and never expires, so the holder never has to roll into a new contract. It is a financial contract and gives no right to GPU time. Funding payments keep its price close to the index. When the perpetual trades above the index, longs pay shorts, and when it trades below, shorts pay longs.

Kalshi has run this machinery since June 2026, when it launched the first CFTC-regulated crypto perpetual futures in the United States. They traded more than \$1 billion in their first week (CNBC, 2026a). Gold and silver perpetuals followed in September and trade around the clock (CNBC, 2026b). Kalshi has also filed for perpetuals on equity indices, interest rates, foreign exchange and single stocks, and plans to add further commodities. The index, funding, margin, clearing and liquidation systems are therefore already in production, and a compute perpetual would use the same systems with a compute index in place of a crypto or metals price.

We would list single-chip perpetuals on the H100, H200 and B200 before attempting any blended index. These would give the market a visible, continuously updated price for compute and let traders take relative-value positions between generations. They would also build up the volatility and correlation history that margin models for a larger market will need.

5.4 Stage Three: Dated Futures

A perpetual tells you what compute is worth today. A dated future tells you what the market thinks it will be worth on a given date, and most commercial hedging needs the second. Kalshi has said publicly that perpetuals suit assets with continuous exposure, but that commodities with regional, seasonal or grade-specific prices also need dated contracts for participants whose physical exposure falls on known dates (Kalshi, 2026b). Compute fits that description, and we expect hedgers to use both kinds of contract. Table 4 shows who we expect them to be.

Participant	Risk	Hedge
AI company or lab	Compute prices rise before it needs capacity in 6–12 months	Buy compute futures for those months
GPU cloud operator	Rental rates fall faster than its financing assumed	Sell compute futures to lock in revenue
Data-center developer	Project returns depend on future rental prices	Use the curve to price new builds and reduce their risk
Lender or investor	Collateral values and cash flows depend on future compute prices	Use forward prices in underwriting and hedge what remains
Chipmaker or OEM	Demand and pricing shift when a new generation launches	Watch generation spreads as a market signal

Table 4: Likely users of dated compute futures and the risks they would hedge. Source: Kalshi.

We would set the first maturities to match the terms of the bilateral capacity deals already being signed, so that those exposures can be hedged from start to finish. Figure 10 lays out the resulting grid of chips and dates.

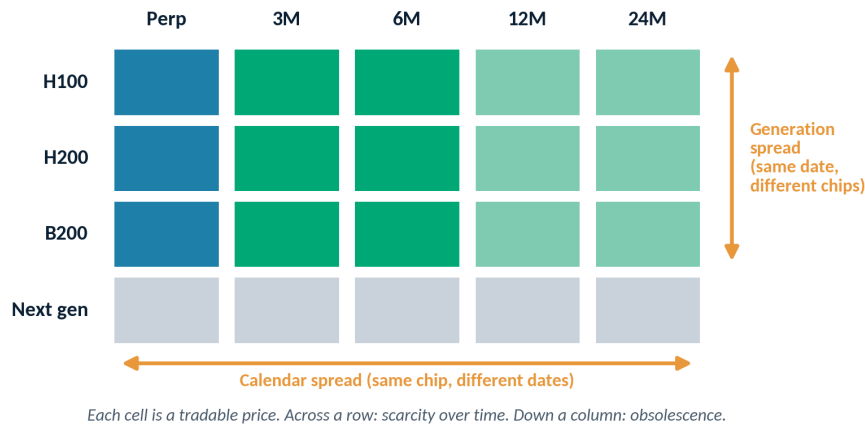


Figure 10: Grid of compute contracts by chip and maturity. Each cell is a tradable price. Reading across a row gives a calendar spread for one chip; reading down a column gives a generation spread at one date. Source: Kalshi illustration.

Reading across a row of the grid shows how the market expects one chip’s price to move over time. Reading down a column compares generations at the same date, which amounts to a market estimate of obsolescence. Infrastructure investors, equipment lenders, procurement teams and chipmakers all have use for those numbers, whether or not they trade the contracts themselves.

6 Regional Power Markets

Electricity is priced locally. A megawatt-hour delivered in Texas and one delivered in Northern Virginia are different products, because the two regions differ in fuel mix, transmission constraints, weather, market rules and load growth. Compute will increasingly be priced the same way, since a given cluster costs a different amount to run in each region. A single national power price would leave operators with a large basis risk, so we have started with four regional markets that together cover most U.S. data-center growth (Table 5). We designed them as a set, and they will form the power side of regional compute–power spreads later on.

Market	Relevance to compute	Price behavior	Typical hedger
PJM Mid-Atlantic, incl. N. Virginia	The densest data-center hub in the world. Data centers account for about 97% of the 2027/28 load-forecast increase, and capacity prices cleared at the cap.	Rising capacity costs and tight reserve margins	Virginia operators, hyperscalers, utilities
ERCOT Texas	The fastest-growing large-load queue, at about 475 GW requested, in an energy-only market that relies on scarcity pricing.	Very volatile, with large real-time spikes	Texas GPU operators; loads that can be curtailed
CAISO California	A grid with heavy solar and battery capacity, close to the West Coast technology industry.	Wide swings within the day, cheap at midday and expensive in the evening	Operators that can move workloads between hours
MISO Central U.S.	A large footprint with heavy industrial load, new data-center corridors and a generation fleet in transition.	Transmission constraints and regional basis	Midwest developers, industrial users

Table 5: Kalshi’s initial regional power markets. Sources: PJM 2027/28 Base Residual Auction results (PJM Interconnection, 2025); ERCOT large-load reporting (ERCOT, 2026); Kalshi analysis.

PJM is where AI load is reshaping a grid most visibly, mainly because of Northern Virginia, and data-center growth already shows up in its load forecasts, transmission plans and capacity prices. ERCOT combines fast data-center growth with an energy-only market that relies on scarcity pricing, and operators running flexible training or batch jobs may be able to cut consumption when prices spike. In CAISO, the large solar and battery fleet produces big price swings within a single day, so operators who can schedule work around them have something to gain. MISO covers a wide area with heavy industrial load and a generation fleet in transition, and it is becoming a destination for new data centers.

7 Trading Power Around the Clock

Data centers draw power every hour of the week, and the price risk that comes with it does not stop either. Weather forecasts change, generators trip and chip announcements are made at weekends as well as on weekdays. Physical power markets already run continuously. The gaps are in financial hedging, since most futures venues close for the weekend while a data center’s load is spread evenly across the week. Our power contracts therefore offer both hourly and daily-average prices and trade at all hours (Figure 11).

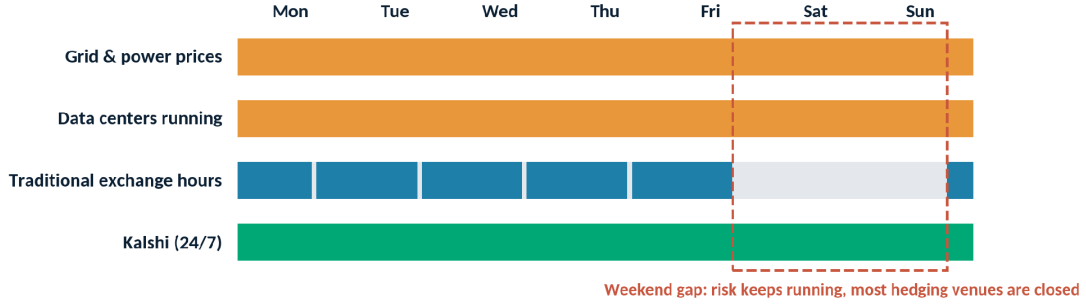


Figure 11: Power-price risk and data-center operation continue through the weekend, when most hedging venues are closed. Traditional exchange hours are illustrative of typical electronic futures sessions. Source: Kalshi.

Three examples show the problem. A heat-wave forecast revised on a Saturday morning changes the outlook for the following week's power prices. A large generator trips on a Sunday and the grid tightens. A major chip or model announcement comes out over a weekend and shifts expectations for compute. In each case, anyone without an open market to trade in carries the exposure until Monday.

Trading around the clock is only safe if risk management keeps the same hours. Margin has to be monitored continuously, collateral has to be movable at any time, and default procedures have to work on a Sunday. Most Kalshi contracts already trade 24 hours a day, seven days a week, including holidays, with initial and variation margin moving on the same schedule (Kalshi, n.d.). Kalshi Klear was the first U.S. clearinghouse to run a full default simulation on a weekend (Markets Media, n.d.). Handling a member default under stress is the most important thing a clearinghouse does, and we wanted evidence that ours could do it outside business hours. Klear also runs intraday margin monitoring, staffs an overnight margin team, accepts weekend prefunding and settles over real-time payment rails.

8 The Silicon Spark Spread

Commodity traders often care less about an input or output price on its own than about the margin between the two. Refiners watch the crack spread between crude and refined products, and gas-fired generators watch the spark spread between gas and power. A data center has the same structure, with electricity going in and compute coming out. Following the spark spread, we write the margin on a GPU-hour as

$$S_t = P_t^{\text{compute}} - h \cdot P_t^{\text{power}},$$

where P_t^{compute} is the price of a GPU-hour, P_t^{power} is the price of electricity per kilowatt-hour, and h is the energy consumed per GPU-hour including cooling, which plays the part of a power plant's heat rate. We call S_t the Silicon Spark Spread. Figure 12 shows how each leg would be hedged.

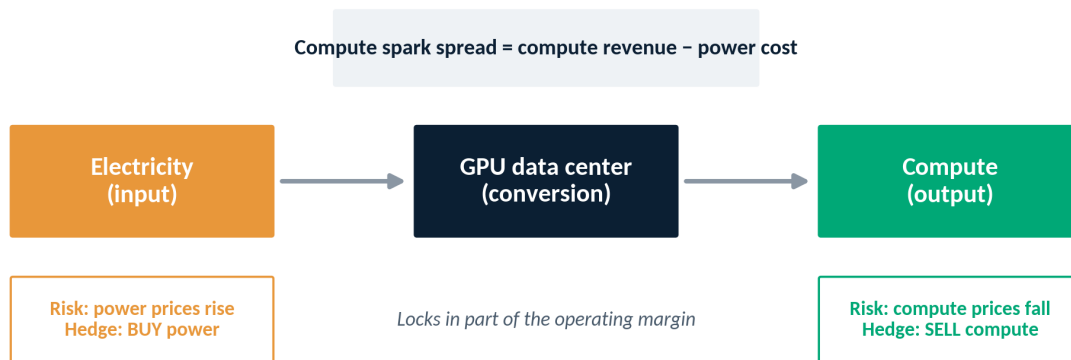


Figure 12: Hedging both legs of a data center’s operating margin: buying power against rising electricity costs and selling compute against falling rental prices. Source: Kalshi illustration.

8.1 A Worked Example

Take a GPU operator in Texas. It loses money if rental rates fall and if ERCOT power prices rise. On an exchange that lists both, it can sell compute futures and buy ERCOT power, fixing part of its margin in advance. A good deal of risk is left over: utilization, hardware depreciation, basis between the operator’s own costs and the contract prices, and fixed costs are all still there. The core of the operating margin, though, is protected.

Pairing regions adds a further layer. Virginia compute against PJM power and Texas compute against ERCOT power give two different spreads, and the gap between them is a market signal about where it is cheapest to add capacity. Table 6 lists the positions an exchange carrying both legs can support.

Position type	Example
Outright compute	Long B200 perpetual
Outright power	Long ERCOT power for next summer
Calendar spread	Long H200 3-month against short H200 12-month
Generation spread	Long B200 against short H200, same date
Regional power spread	PJM against ERCOT
Compute–power spread	Short Texas compute against long ERCOT power

Table 6: Positions supported by an exchange that lists both compute and power. Source: Kalshi.

Taken together, these positions start to resemble the spread markets that have long existed around oil, gas and power.

9 Clearing and Risk Management

Listing a contract is the easy part. Early compute derivatives will probably have thin order books, reference prices drawn from scattered sources and sharp moves around chip launches. Institutions will trade them only if margin, clearing and default management can cope with those conditions. Figure 13 shows the layers involved.

Trading	Order book · 24/7 access · price bands & velocity controls
Benchmark	Index methodology · outlier filters · fallbacks
Margin	Initial & variation margin · concentration and liquidity add-ons
Clearing (Kalshi Klear)	Central counterparty · portfolio offsets · real-time monitoring
Collateral & payments	Weekend prefunding · real-time payment rails
Default management	Default fund · weekend-tested default procedures

Figure 13: Layers of market infrastructure, from trading and benchmarks through margin, clearing, collateral and default management. Source: Kalshi.

9.1 Margin, Controls and Settlement

Initial margin has to reflect the actual risk of a position. Volatility is the starting point, but concentration and how hard the position would be to close in a stressed market also count. Variation margin is settled often enough that large unrealized gains and losses do not build up. On the trading side we use price bands, velocity limits, reference-price checks, depth-aware validation, statistical outlier detection, kill switches and trading pauses. Settlement relies on benchmark rules that still function if a data source fails, with fallbacks written down before they are needed.

9.2 Portfolio Margin

Some positions do offset one another, such as an H200 against a B200, one maturity against another, or power against compute. Recognizing these offsets can lower margin when the data justify it. We intend to be cautious here. Correlations break down, the relationship between two chips changes when a new generation launches, and regional power prices can spike on their own. Any cross-product offset should be supported by data and set conservatively. The aim is to give credit for genuine diversification without letting offsets become a source of leverage.

Kalshi Klear has been registered with the CFTC as a derivatives clearing organization since November 2024 (CFTC, 2024). Portfolio margin, concentration and liquidity add-ons, real-time monitoring, intraday variation margin and default resources are sometimes thought of as back-office matters, but institutions will not trade a new contract without them.

10 Roadmap

We do not expect a mature compute market to appear all at once. Table 7 sets out the stages, each answering a question the one before it cannot.

Stage	Question it answers	What it provides
1. Prediction markets	What does the market expect?	Probabilities on GPU prices, data-center growth, power demand and policy
2. Benchmarks	What exactly is being priced?	Published, chip-specific reference prices with defined fallbacks
3. Perpetual futures	What is compute worth now?	Continuous, tradable compute exposure with funding
4. Dated futures	What will it be worth later?	A forward curve for hedging, procurement and lending
5. Power markets	What will the main input cost?	PJM, ERCOT, CAISO and MISO contracts, traded at all hours
6. Integrated markets	How can the whole business be hedged?	Calendar, generation, regional and compute–power spreads, with portfolio margin

Table 7: Roadmap for AI infrastructure markets. Future products are subject to regulatory approval. Source: Kalshi.

Kalshi intends to cover the whole sequence, from prediction markets through perpetual and dated futures to power contracts, with Kalshi Klear clearing all of it around the clock. Having every stage on one exchange and one clearinghouse is what makes the cross-product spreads in Section 8 and the portfolio margin in Section 9 possible.

11 Conclusions

AI now depends on physical infrastructure: chips, racks, buildings, electricity, transmission lines and the capital to pay for them. Very large sums are being committed, but many of the risks that come with them are hard to price and harder to hedge. Compute prices are fragmented, forward curves are short, obsolescence is difficult to measure, and a lack of power can stall projects that would otherwise go ahead.

We have argued that markets can close much of this gap if they are built in the right order. Prediction markets come first because they can pool information before any benchmark exists. Benchmarks then give everyone a common price to refer to. Perpetual futures provide exposure to the current price, dated futures extend it forward, and regional power contracts cover the main input. The Silicon Spark Spread ties the two legs together and puts a market price on the operating margin of a data center.

Commodity derivatives have generally appeared once an exposure grew large enough that producers, consumers and lenders wanted a better way to price and transfer it. Compute has reached that stage, and Kalshi intends to build the markets for it.

References

- AMD (n.d.). AMD Instinct MI325X and MI355X Accelerators: Product Specifications.
- Bloomberg (2026). Kalshi Ramps Up Effort to Build Markets for AI Computing Power. Via *Yahoo Finance*, 14 July. <https://finance.yahoo.com/markets/options/articles/kalshi-ramps-effort-build-markets-140109761.html>
- Commodity Futures Trading Commission (2024). CFTC Grants Kalshi Klear LLC DCO Registration. Release 8957-24. Positions transferred to Kalshi Klear on 15 November 2024. <https://www.cftc.gov/PressRoom/PressReleases/8957-24>
- CME Group (n.d.). Crude Oil Futures and Options. <https://www.cmegroup.com/markets/energy/crude-oil>
- CNBC (2026a). Kalshi Perpetual Futures Trading “Perps” Crosses \$1 Billion in Volume Within a Week of Launch. 9 June. <https://www.cnbc.com/2026/06/09/kalshi-perpetual-futures-trading-perps-crosses-1-billion-in-volume-within-a-week-of-launch.html>
- CNBC (2026b). Kalshi Launches “Perps” for Gold and Silver Following CFTC Approval, Expanding Futures Offerings. 10 September. <https://www.cnbc.com/2026/09/10/kalshi-launches-perps-for-gold-and-silver-following-cftc-approval-expanding-futures-offerings.html>
- Electric Reliability Council of Texas (2026). Large Load Interconnection Update, April 2026. https://www.ercot.com/files/docs/2026/04/01/ERCOT_LargeLoad_Update_April2026_B-C_Hearing.pdf
- Fortune* (2026). Kalshi’s CEO Is Racing to Build a Futures Market for AI’s Most Precious Resource. 12 August. <https://fortune.com/2026/08/12/kalshi-compute-futures-cme-intercontinental/>
- Kalshi (2026a). Compute Has a Market-Implied Forward Curve Now. *Kalshi News*, 14 July. <https://news.kalshi.com/p/gpu-compute-forward-curve>
- Kalshi (2026b). Public Statement on Perpetual versus Dated Contracts. 2 June. <https://www.cftc.gov/filings/ptc/ptc0602265099.pdf>
- Kalshi (n.d.). What Are Trading Hours? Kalshi Help Center. <https://help.kalshi.com/trading/what-are-trading-hours>
- Kalshi Research (2026). The Economics of Compute. 22 July. Market-implied values as of publication. <https://kalshi.com/research/insights/the-economics-of-compute>
- Markets Media (n.d.). Kalshi Klear Completes Default Simulation on a Weekend. <https://www.marketsmedia.com/kalshi-klear-completes-default-simulation-on-a-weekend/>
- NVIDIA (n.d.). Data Center Product Specifications: H100, H200, B200 and GB300 NVL72. <https://www.nvidia.com/en-us/data-center/>
- PJM Interconnection (2025). PJM Auction Procures 134,479 MW of Generation Resources: 2027/2028 Base Residual Auction Results. 17 December. <https://insidelines.pjm.com/pjm-auction-procures-134479-mw-of-generation-resources/>

POWER Magazine (2026). Abbott Orders Full Audit of Texas Data Center Interconnection Queue, Threatens to Deny Grid Access. Reporting ERCOT large-load data as of June 2026 and peak demand of 22 July 2026. <https://www.powermag.com/abbott-orders-full-audit-of-texas-data-center-interconnection-queue-threatens-to-deny-grid-access/>

SemiAnalysis (n.d.). InferenceX: GB300 NVL72. 132 kW nominal rack TDP, approximately 155 kW peak. <https://inferencex.semianalysis.com/chips/gb300-nvl72>

Shehabi, A., et al. (2024). *2024 United States Data Center Energy Usage Report*. Lawrence Berkeley National Laboratory for the U.S. Department of Energy, December. <https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report>