Not for redistribution without written consent of Morgan Stanley

Morgan Stanley | RESEARCH

IDEA

October 26, 2026 04:01 AM GMT

Global Thematics & US Semiconductors | North America

Hot Summer, Short Fall: Increased Power Shortfall, Increased BTM Clarity, and Addressing Rubin Rack kW Misconceptions

Raising our 2026-28 net power shortfall (now including US DC spillover into APAC) from 33GW to 38GW, driven by higher DC power demand, offset by greater front-of-the-meter grid capacity (20GW from 15GW in '26-'28) & BTM. We address key pushback, incl. power requirements for NVL72 Vera Rubin racks.

AlphaWise α

Key Takeaways

- Overall power demand from US DCs increases from 97GW to 116GW. Our power shortfall before time-to-power solutions increases from 57GW to 70GW.
- Both global and US power demand ests rise, reflecting (1) increased historical demand, (2) increased chip-level demand, and (3) utilized power assumptions.
- Power shortfall net of time-to-power solutions rises from 33GW to 38GW, after accounting for greater availability of BTM solutions & US DC spillover into APAC.
- We provide more color on areas of pushback: 1) BTM, 2) Asia spillover, 3) grid optimization/enhancing technologies, and 4) Vera Rubin configurations.
- We reiterate our conviction on 234kW/rack for Vera Rubin configurations, citing spec sheets and builds rather than derived values from chip spend.

Tech Diffusion

A Morgan Stanley Research Key Theme of 2026

We maintain convicted the power shortfall is real. We have received a significant amount feedback on our power shortfall work as investors look to square estimates for $1.5T / $1.6T of '27 / '28 capex with power availability and what this means for the supply chain. We don't believe the idea of a ~33% power shortfall is a consensus view, with many investors taking the stance that companies will "creatively find ways around" these bottlenecks and citing that orders in the supply chain remain strong. That being said, we still see LPS as a key constraint to the buildout. Today, we update our Intelligence Factory model to take into account new inputs — higher

MORGAN STANLEY & CO. LLC

Stephen C Byrd
Equity Analyst
Stephen.Byrd@morganstanley.com

Joseph Moore
Equity Analyst
Joseph.Moore@morganstanley.com

David Arcaro, CFA
Equity Analyst
David.Arcaro@morganstanley.com +1 212 761-7817

Cameron McVeigh, CFA
Equity Analyst
Cameron.McVeigh@morganstanley.com

Sabrina E Hao
Equity Strategist
Sabrina.Hao@morganstanley.com +1 212 761-3315

Mason Wayne
Research Associate
Mason.Wayne@morganstanley.com

Nicole Kozhukhov
Research Associate
Nicole.Kozhukhov@morganstanley.com +1 212 761-1636

Jeremy Herring
Research Associate
Jeremy.Herring@morganstanley.com

Garo K Amerkanian
Research Associate
Garo.Amerkanian@morganstanley.com +1 212 761-1207

Amanda Huang
Research Associate
Amanda.Huang@morganstanley.com

Please see Exhibit 9— Ways to Play: Power Bottlenecks.

Morgan Stanley does and seeks to do business with companies covered in Morgan Stanley Research. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of Morgan Stanley Research. Investors should consider Morgan Stanley Research as only a single factor in making their investment decision.

For analyst certification and other important disclosures, refer to the Disclosure Section, located at the end of this report.

data center demand, partially offset by higher front-of-the-meter and behind-the-meter power availability. This leads us to a **70GW '26-28 power shortfall *before* time-to-power-solutions (up from 57GW prior), or a 38GW '26-28 power shortfall *net* of time-to-power-solutions and inclusive of US DC spillover into APAC (up from 33GW).**

We also address key areas of investor pushback to our call, which is centered around three levers of potential upside — 1) BTM, 2) Asia expansion, and 3) grid enhancing technologies / data center curtailment — as well as our Vera Rubin configuration assumptions. We'd note that while these three levers could narrow the power shortfall further, they all face constraints of their own — BTM deployment is constrained by labor and grid optimization/data center curtailment can face stakeholder buy-in issues. On Asia, we forecast 15GW of US demand spilling over into the region through 2030, and we now break this out explicitly. On Vera Rubin configurations, we feel comfortable with our assumptions, as they're in-line with NVDA and supply chain checks.

What does this mean for stocks? In our view, this shortfall will likely increase the urgency, and premia paid, to secure (A) time-to-power solutions in the US — for example, BTM players like **BE** and **SEI,** as well as the powered shell providers (**WULF, HUT, CIFR, RIOT**). For a comprehensive list of ways to play this theme, see (Exhibit 9) and (B) data center growth outside of the US — read our Asia team's note on this here. For semis, we remain most constructive on **NVDA** and **AVGO** due to conservative guidance, visibility into chip placement, geographic expansion outside the US, and **NVDA**'s higher token-per-watt profile. However, in event of a pause, we're likely to see disruption at the tail of the bullwhip and where shortage driven pricing dynamics are supporting revenue and EPS — such as downstream components like **memory, optical,** and **other components** that are likely to be over-ordered like **power management** and **analog parts** in a bear case (see more here: What Does the Power Shortfall Mean for Semis). We see a risk that **ASICs** will lose share in a power-constrained environment relative to GPUs, given the lower token production per watt for ASICs and lower fungibility across customers.

We have updated the Intelligence Factory model, our hub for Powering AI, the power shortfall, token economics, and time-to-power solutions to reflect increasing power demand. We now project 116GW of power demand from chips sold in 2026-28, increased from 97GW. This increase comes from three changes to the model:

1. Increased expectations of *utility power* necessary to power all purchased chips from increased demand of energy from storage, networking, and other rack-level components (previously excluded).
2. Increased *gross* power demand from historical ex-Gen AI data centers.
3. Proper squaring of an assumed 75% load factor within our power methodology, increasing overall power demand necessary to power chips at-capacity.

Clearer methodology in defining/converting from *gross* power (total, grid-delivered power), *critical IT* power (post-PUE power delivered to the chips), and *utilized* power (power that is "consumed" by the chips, inclusive of a load factor) is

necessary in evaluating the overall power demands of AI data centers. Our updated model clearly outlines the designation of power (gross, gross utilized, critical IT, etc.) within each line item and properly adjusts between these designations to arrive at a total power demand. Further, increased insight into potential data center demand from our data centers team (led by Cameron McVeigh; our projections of post-shortfall supply square within a 3% margin of error in 2029, with lower margins in preceding years).

Please click here to request an updated copy of the Intelligence Factory model.

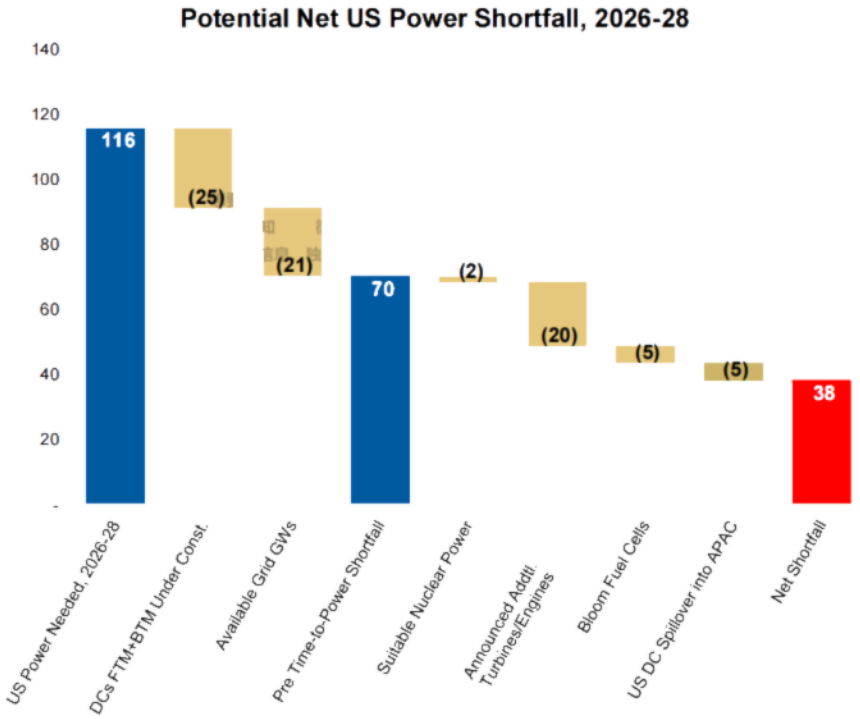
Exhibit 1: Our Power Summary Tab Outlines the Cumulative and Year-Over-Year Power Shortfall; Our '26-'28 Shortfall is Now 70GW.

Part IID.1. Annualized Cumulative Shortfall (US DCs Only)				
Year	**2026**	**2027**	**2028**	**2029**
Gross Power Needed	33	66	116	185
US DCs Under Construction (Gross Power, FTM + BTM)	(9)	(21)	(25)	(26)
Available/Contractable US Grid Capacity (Gross GW, FTM)	(4)	(8)	(21)	(39)
Shortfall Before Time to Power Solutions	**19**	**36**	**70**	**120**
Suitable Nuclear Power	(0)	(1)	(2)	(3)
Bloom Fuel Cell Pontential	(1)	(2)	(5)	(9)
Announced Addtl. BTM Turbines/Engines	(4)	(11)	(20)	(28)
Crypto Sites Available (counted in DCs Under Construction or Turbines)	(2)	(8)	(16)	(19)
Est. Midpoint US-Based Net Shortfall (Midpoint)	**14**	**22**	**43**	**80**
Additional US DCs Constructed YoY (Gross GW)	**18**	**43**	**72**	**104**
Total US DCs Constructed Incl. Shortfall EoY 2029 (Gross GW)	**82**	**107**	**136**	**168**

Part IID.2. Annualized Year-By-Year Shortfall (US DCs Only)				
Year	**2026**	**2027**	**2028**	**2029**
Power Needed	33	33	50	69
US DCs Under Construction (Gross Power, FTM + BTM)	(9)	(12)	(4)	(1)
Available/Contractable US Grid Capacity (Gross GW, FTM)	(4)	(4)	(13)	(18)
Shortfall Before Time to Power Solutions	**19**	**17**	**34**	**50**
Suitable Nuclear Power	(0)	(1)	(1)	(1)
Bloom Fuel Cell Pontential	(1)	(2)	(3)	(4)
Announced Addtl. BTM Turbines/Engines	(4)	(6)	(9)	(9)
Crypto Sites Available (counted in DCs Under Construction or Turbines)	(2)	(6)	(8)	(3)
Est. Midpoint US-Based Net Shortfall (Midpoint)	**14**	**8**	**21**	**37**
Additional US DCs Constructed YoY (Gross GW)	**18**	**25**	**29**	**32**

Source: Morgan Stanley Research

Exhibit 2: Our US 2026-28 Power Shortfall, inclusive of US DC Spillover into APAC, Rises to 38GW from 32GW.



Source: Morgan Stanley Research

Exhibit 3: Even After Time-to-Power Solutions, We Project a 43GW 2026-28 Net Shortfall, 38% of Data Center Demand. Even When Considering Global Solutions (US DC Spillover into APAC), We Still Project a 38GW Shortfall, 33% of Demanded Power.

Total US Data Center Power Demand (GW), 2026-28	**116**
Less FTM & BTM Data Centers Under Construction	(25)
Less FTM Utility Grid Access	(21)
Power Shortfall Before New "Time to Power" Solutions	**70**

	Low End	Midpoint	High End	Probability of Success
Solution #1: Addt'l BTM Turbines/Engines	6	22	38	90%
Solution #2: Bloom Energy Fuel Cells	5	6	8	90%
Solution #3: Site DC at Opt Nuclear Plant	1	3	8	70%
Solution #4: Convert PSP Sites (included in turbines or DCs under construction)	11	19	27	85%
Solution #5: US DC Spillover into APAC	5	6	8	90%
Probability-Weighted US-Based "Time to Power" Solutions through 2028	10	27	47	
Net Shortfall for US Companies Through 2028	**(60)**	**(43)**	**(23)**	
Shortfall as a % of 2026-28 US DC Deployments	*52%*	*38%*	*20%*	
Net Shortfall Through 2028 (incl. Global Solutions)	**(56)**	**(38)**	**(17)**	
Shortfall as a % of 2026-28 US DC Deployments	*48%*	*33%*	*14%*	

We believe this outcome is most likely

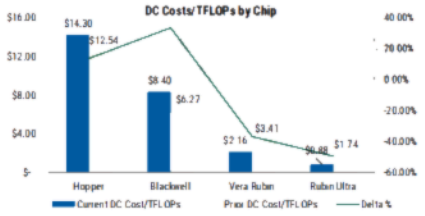
Source: Morgan Stanley Research

This increase in power comes as a result of fully factoring in the grid-level power that must flow through to data centers in order to adequately power chips through consistent inclusion of proper load factors. We reiterate that, even with the presence of a shortfall, companies are incentivized to engage with time-to-power solutions — our Intelligence Factory model projects increasing margins on tokens produced in data centers while unit costs (both per-token and production-per-watt metrics) fall with each subsequent chip generation.

We see radically reduced cost/TFLOPs production across subsequent chip

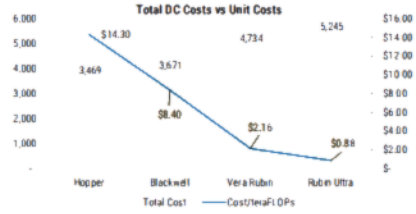
generations as total costs v. unit costs fall in kind. These production metrics are apparent in tokens produced/total wattage production, highlighting the increases in token production efficiency and lending evidence to our assertion that installing new-gen GPUs and Architectures justify the higher up-front costs.

Exhibit 4: TFLOPs Production Outpaces Relative Cost Increase...



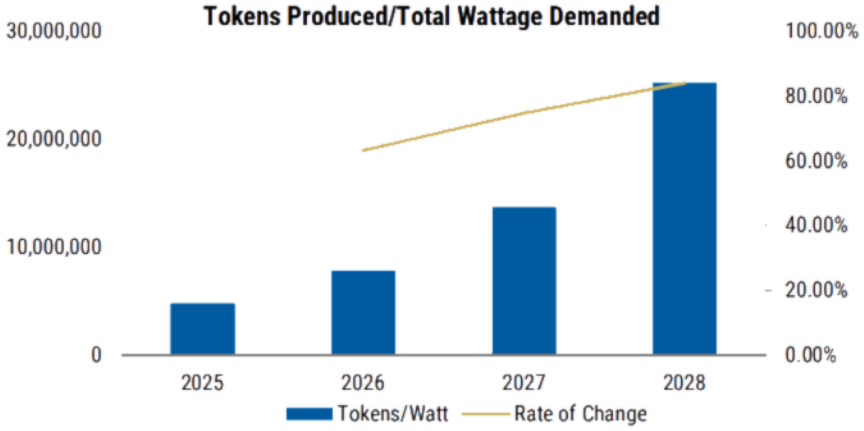
Source: NVIDIA, Morgan Stanley Research Estimates

Exhibit 5: ...Highlighting How Unit Costs Fall Faster Than Overall Cost Increases



Source: NVIDIA, Morgan Stanley Research Estimates

Exhibit 6: While Headline Power Demand Numbers Increase to a Total of 116GW of Cumulative Demand Through 2028, YoY Increases in Token Production Outpace Wattage Increase



Source: Morgan Stanley Research

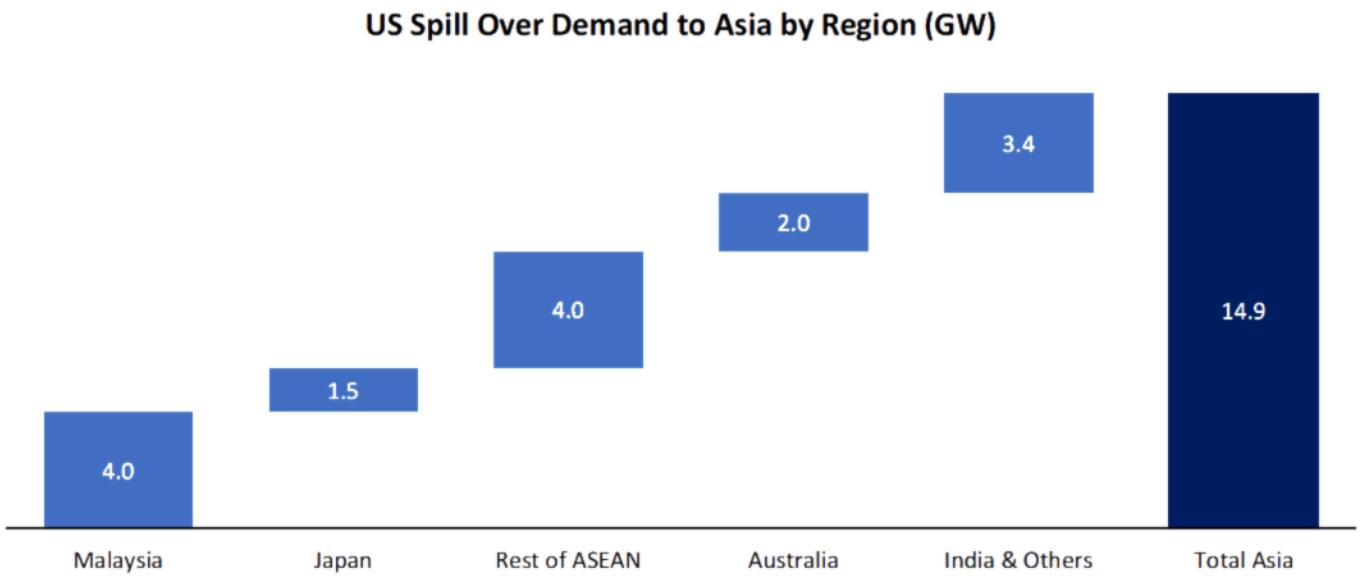
Addressing Feedback on BTM, Asia, and Grid Optimization

What has the feedback been from tech investors on the power call? Tech investors largely acknowledge that an energization shortfall exists, which is a combination of both power and people (labor shortages to put up data centers). However, we've received pushback on the magnitude of the shortfall, with many taking an optimistic view that companies will be able to creatively "figure out" ways around this constraint. We address key areas of pushback below:

- **1) Potential upside to our BTM forecasts.** We forecast 31 GW of BTM solutions coming online in 2026-28 in our base case, which reflects the average of 1) firm, contracted BTM deals we see today, and 2) total possible production *(see below for a greater explanation)*. While we agree that it's possible to further close the power shortfall when taking into account total production capacity for BTM solutions like turbines, the gating factor becomes labor. For example, our Utilities team believes the existing labor base would likely support 10-20 GW of annual gas additions (both BTM and for the grid); however, our team believes this could reach 50+ GW of annual gas additions by the end of the decade.
- **2) Greater capacity for US data center spillover into Asia: we estimate 6GW of APAC spillover in 2026-28.** Through company conversations, our Asia team has line of sight into ~15GW of US hyperscaler demand potentially spilling over into Asia through 2030 (Exhibit 7), alongside ~3GW of power demand from new semiconductor fabs in ASEAN/India/Korea. However, the total potential spillover could be higher if US companies expand further into the region. The team sees Thailand, India and Japan attracting the strongest data center commitments, with Thailand also benefiting from Chinese hyperscaler demand. Together, these markets have ~30GW of surplus power capacity as gas, coal and nuclear output ramps up, although grid constraints limit near-term serviceability. **We also now include estimates of APAC spillover in our calculation of the US shortfall; a designation is made between "Net Shortfall" (pre-APAC spillover) and "Net Shortfall Incl. Global Solutions (inclusive of APAC Spillover); Exhibit 3 .** After 6GW of APAC Spillover, the total net shortfall is 56-38GW, 48%-33% of demand for US companies. Even with the presence of international solutions, we still project a considerable shortfall in power necessary to power chips.
- **3) The potential for grid optimization or data center curtailment to help alleviate the power shortfall.** While these technologies are compelling, we view them as having a limited impact on the ability to free up US grid capacity today, and over the next several years, in reality. First, **grid enhancing technologies** — hardware and software solutions that make the grid "smarter" and enhance the capacity and flexibility of the existing infrastructure (see our work on this here) — can be a lower capex way to increase capacity of an existing line. This could be a compelling area of upside, but a key constraint is a utility's willingness to invest in this type of technology. Second, while **data center curtailment** — or where a data center promises to use less energy from the grid during "peak" times — can theoretically alleviate stress on the power grid during periods of peak demand, we believe utilities will not likely increase their data center grid connection volumes due to this dynamic. Why? Utilities manage their grids to ensure that, under the

most extreme of conditions, all customers are able to use the volume of power required. Demand response, in the form of data center curtailment, is typically of a much shorter duration than the extreme events that utility management teams plan for. In practice, a data center would not likely be able to run diesel generators for six straight days during a winter storm with rolling blackouts, in part due to limits on emissions of pollutants and particulate matter that are typically set at the state and local level. In addition, we see little actual regulatory activity across the US that would achieve the outcome being suggested.

Exhibit 7: We project 15GW of US Hyperscaler Data Center Demand Spilling Over into APAC Through 2030.



Source: Morgan Stanley Research estimates

Addressing Rubin Rack Power Budget Misconceptions

We received pushback on our updated rack-power assumptions for Vera Rubin and Rubin Ultra since publishing our September "Racking Up the Power Shortfall note," which we raised to 234kW from 149kW and 600kW from 415kW, respectively. As we construct our power shortfall from *chip purchases and rack/server-level power demand*, pushback is hyper-relevant for overall AI power demand. With industry expectations varying considerably, we take a closer look at the assumptions behind our estimates and how they compare with NVIDIA's disclosures and broader supply chain commentary.

How we arrive at our assumptions. For Vera Rubin NVL72, our 234kW estimate assumes a 2,300W TDP per GPU, or 165.6kW across 72 GPUs, with another 68.4kW for CPUs, networking and other rack components. For Rubin Ultra, we're using NVIDIA's stated 600kW rack-power figure, which the company first introduced at GTC 2025.

What has NVIDIA said? NVIDIA's own roadmap, reiterated at OCP last week, plots Vera Rubin NVL72 at ~240kW consistent with our 234kW assumption (Exhibit 7). While NVIDIA also cites 330kW for its Gen 2 Vera Rubin power architecture, the company has described this as the power capacity supported by its 110kW power shelves, rather than the expected power consumption of NVL72. NVIDIA's subsequent Gen 3 architecture supports ~570kW using liquid-cooled power shelves and busbars, broadly consistent with the ~600kW Rubin Ultra rack-power figure Jensen Huang outlined at GTC 2025. We think the ~240kW figure is a more relevant comparison for our 234kW estimate, particularly given the most common specifications coming from the supply chain. Also, at last week's OCP Global Summit, NVIDIA outlined its MGX 1.2 architecture at supporting rack deployments of up to 250kW, providing another reference point closer to our estimate.

Exhibit 8: NVIDIA Rack Power Roadmap

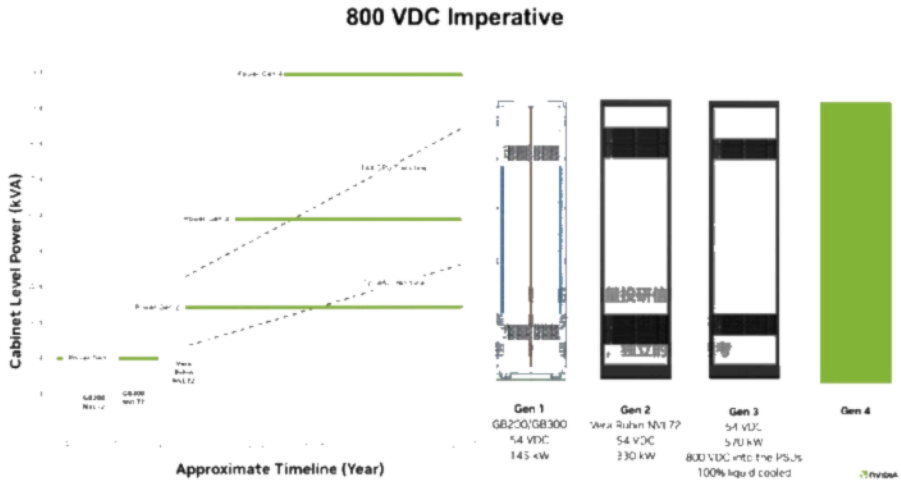


Figure 3: NVIDIA Rack Power Roadmap and the Path to Native 800 VDC Architecture

Source: NVIDIA

What is the supply chain saying? Broader industry estimates generally point to Vera Rubin rack power in the 220-250kW range. Schneider Electric has referenced Vera Rubin rack requirements of 227-246kW, while Vertiv has pointed to ~240-250kW and TrendForce estimates ~225kW, and Pegatron specifies 228kW for its NVL72 Max-P

configuration. These estimates only further reinforce our 234kW assumption, with several suppliers pointing to similar rack power requirements.

Engineering improvements could also help ease power constraints. A key takeaway from the OCP Global Summit we attended earlier this month was the industry's growing focus on making more efficient use of available power, which we think will be increasingly important as data center power constraints persist (see our OCP takeaways here). NVIDIA is developing several technologies to address this, including **1) MaxLPS, 2) Rack Intelligent Power Smoothing (IPS), and 3) 800VDC power distribution**. **1) MaxLPS** directs power to GPUs as needed and improves software and cooling efficiency, allowing more compute within the same data center power budget; **2) Rack IPS** uses energy storage and GPU-level controls to smooth rapid power fluctuations, reducing power spikes and the need to size infrastructure around peak demand; and **3) 800VDC power distribution** reduces electrical losses as power moves through the data center. While our model doesn't currently reflect the potential benefits of these technologies, we think they're worth keeping in mind, as the industry has continued to find ways to work around power constraints throughout the AI cycle.

We remain comfortable with our rack power budget assumptions. For now, we think 234kW remains a reasonable assumption for Vera Rubin given the underlying GPU TDP, NVIDIA's own roadmap and broader industry estimates. While we acknowledge the variation in reported rack power budgets, which may also reflect differences in system and facility configurations, most estimates remain consistent with our assumption. For Rubin Ultra, we continue to view ~600kW as a reasonable assumption, supported by NVIDIA's prior disclosure and broadly consistent with its Gen 3 power architecture. However, we recognize there is greater uncertainty given the cancellation of the original Kyber rack design and its replacement with a new architecture, as the final configuration and power requirements are still evolving.

Ways to Play

Exhibit 9: Ways to Play: Power Bottlenecks

Name	Ticker	Analyst	Region	Rating	Price (USD)	Price Target (USD)	Upside (Downside) to PT
Powered Shell Providers (PSPs)							
TeraWulf	WULF	Stephen Byrd	US	OW	11.32	65.00	474.2%
Hut 8	HUT	Stephen Byrd	US	OW	63.97	275.00	329.8%
Cipher Mining	CIFR	Stephen Byrd	US	OW	8.32	51.50	519.0%
Riot Platforms	RIOT	Stephen Byrd	US	OW	12.89	35.50	175.4%
Data Center REITs							
Equinix	EQIX	Cameron McVeigh	US	OW	952.82	1,250.00	31.2%
Singapore Telecom	STEL.SI	Da Wei Lee	APAC	OW	3.22 USD	4.30 USD	33.5%
Off-Grid Power Generation							
Solaris	SEI	Dave Arcaro	US	OW	58.61	102.00	74.0%
Bloom Energy	BE	Dave Arcaro	US	OW	217.77	335.00	53.8%
GE Vernova	GEV	Dave Arcaro	US	OW	827.42	1,350.00	63.2%
INNIO Group	INIO	Dave Arcaro	US	OW	15.11	40.00	164.7%
Liberty Energy	LBRT	Joe Laetsch	US	OW	15.38	28.00	82.1%
Erock Inc.	FROC	Angel Castillo	US	OW	8.28	23.00	177.8%
Siemens Energy	ENR GY	Max Yates	EMEA	OW	133.04 USD	218.60 USD	64.3%
Energy Storage							
CATL	300750 SZ	Jack Lu	APAC	OW	35.60 USD	74.91 USD	110.4%
Cummins	CMI	Angel Castillo	US	OW	519.53	761.00	46.5%
Samsung SDI	006400 KS	Young Suk Shin	APAC	OW	370.13 USD	505.58 USD	36.6%
Renewable Energy Developers / IPPs							
NextEra Energy	NEE	Dave Arcaro	US	OW	81.90	110.00	34.3%
Clearway	CWEN	Robert Kad	US	OW	31.43	52.00	65.4%
Adani Green	ADNA.NS	Girish Achhipalia	APAC	OW	12.50 USD	18.08 USD	28.7%
JSW Energy	JSWE.NS	Girish Achhipalia	APAC	OW	4.73 USD	6.90 USD	45.8%
Midstream Energy							
Williams Companies	WMB	Robert Kad	US	OW	70.63	103.00	45.8%
Data Center Electrical Equipment							
Legrand	LEGD.PA	Max Yates	EMEA	OW	139.12 USD	186.09 USD	33.8%
Rexel S.A.	RXL.PA	Max Yates	EMEA	OW	36.58 USD	47.08 USD	28.7%
Grid Equipment / Transformer							
HD Hyundai Electric	267260 KS	Ryan Kim	APAC	OW	354.66 USD	1,131.80 USD	219.1%
LS Electric	010120 KS	Ryan Kim	APAC	OW	118.40 USD	203.74 USD	72.1%
Data Center Transmission / Other							
Adani Energy	ADAI NS	Vivek Rajamani	APAC	OW	12.87 USD	16.53 USD	28.4%
Semiconductors (North America)							
NVIDIA	NVDA	Joseph Moore	US	OW	210.96	300.00	42.2%
Broadcom	AVGO	Joseph Moore	US	OW	321.68	505.00	57.0%

Source: Morgan Stanley Research, Factset. Data as of 10/27/2025

Important note regarding economic sanctions. This report references jurisdictions which may be the subject of economic sanctions. Readers are solely responsible for ensuring that their investment activities are carried out in compliance with applicable laws.