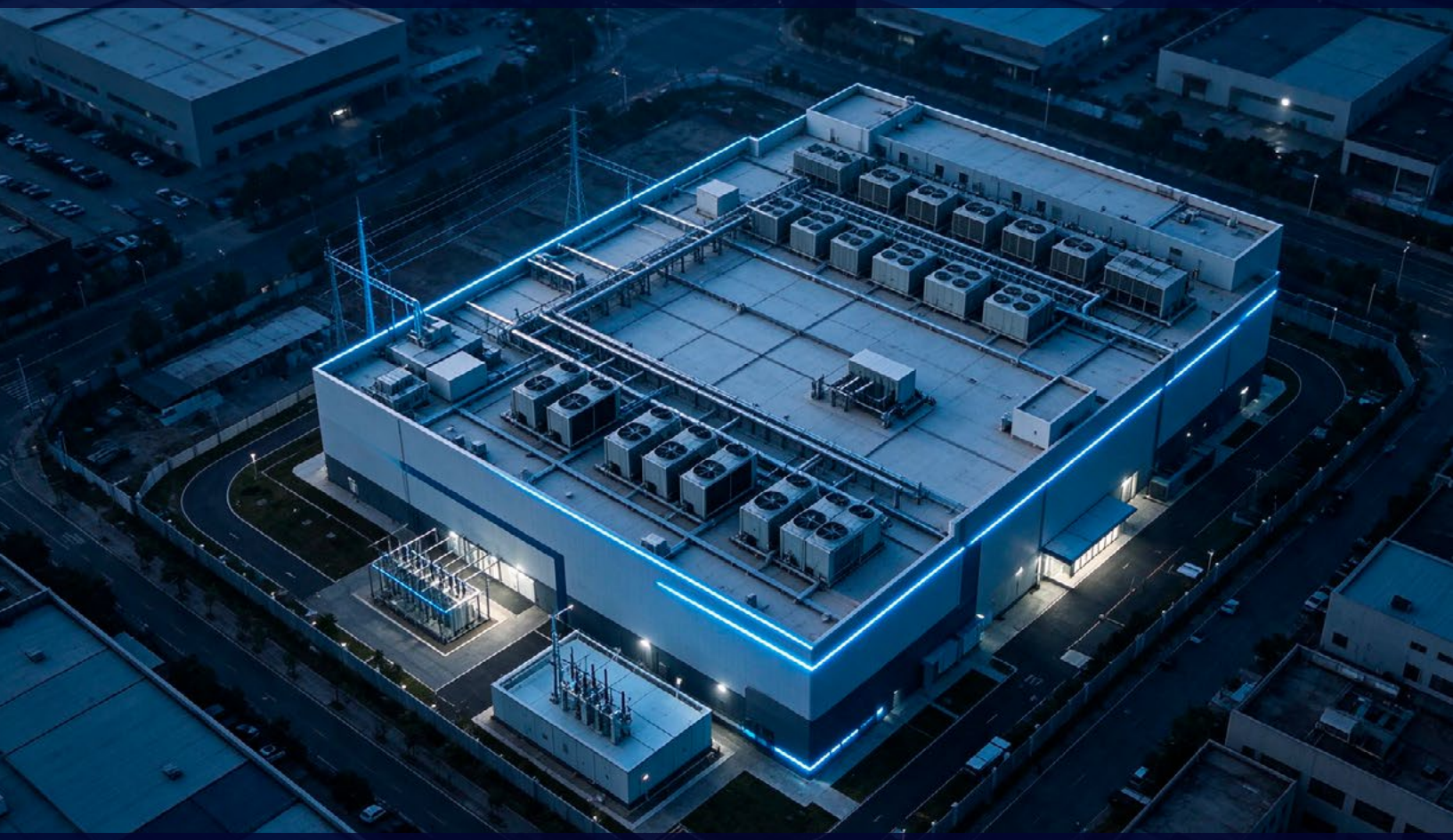
HyperFRAME RESEARCH

IN COLLABORATION WITH QumulusAI

RESEARCH BRIEF
AI INFRASTRUCTURE
DECEMBER 2026

Pockets of Power

Why the Next Megawatt Is What Really Matters



AUTHORS

Stephen Sopko
Practice Lead

Steven Dickens
CEO & Principal Analyst

CONTENTS

Inside this brief

IN BRIEF

The constraint on AI capacity is energized space. This brief tests QumulusAI's sub-50MW, energized-space model against its 2026 deployments and asks whether it can repeat in 2027.

01 Executive Summary — 03 ›
Energized space decides who adds capacity next

02 The Proof Is in the Power — 03 ›
Four signals and QumulusAI's deployment record

03 The Power Hunt: Why Five Megawatts Is a Strategy — 06 ›
Pockets of power, unit economics and exposure

04 Capacity Planning as a Partnership — 08 ›
Why the delivery clock starts with demand

05 Positioning for 2027: Repeating the Play — 08 ›
Six sites where power is ready

06 Regulatory Pushback on Large Campuses — 09 ›
Thresholds that land on the 50MW-plus campus

Looking Ahead — 11 ›

About QumulusAI & the Analysts — 12 ›

Credits — 13 ›



SECTION 01

Executive Summary

AT A GLANCE

15%

of enterprises are running mass AI deployments today.

55%

expect to reach mass deployment within six to 12 months.

HyperFRAME Lens: State of the Enterprise AI Stack, 1H 2026 (n=544) ›

While most conversations about AI capacity still center on demand and silicon, the constraint that decides who adds capacity next is energized space.

That is, buildings where power is already connected, cooling works, and racks can move in quickly. In 2027, capacity is likely to come fastest to operators that can find, finance, contract, and activate pockets of existing power.

Our March 2026 brief, The Hyperspeed Compute Era, argued that velocity was becoming the deciding infrastructure variable. This brief tests that thesis against deployments since that time: whether a sub-50MW, energized-space model holds up in production, what actually differentiates it, and whether it can repeat in 2027.

According to HyperFRAME Lens, 15% of enterprises are running mass AI deployments today, and 55% expect to reach that stage within six to 12 months (HyperFRAME Lens, State of the Enterprise AI Stack, 1H 2026, n=544). Demand moving on that clock will not wait for capacity that arrives in years.

SECTION 02

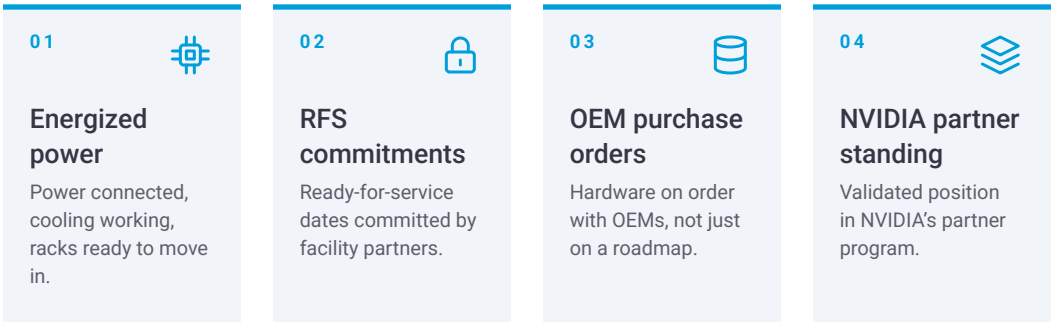
The Proof Is in the Power

Announced neoclouds crowd the field, but it is worth noting that two essential elements are often missing or delayed: energized racks and paying customers. That gap is a headwind for the operators actually delivering. Every provider describing capacity it does not control instructs buyers' predisposition to discount the next claim. A hypothetical gigawatt is worth less than a real megawatt.

Buyers still want speed first. QumulusAI management said, and our observations confirm, that customers now open conversations with when they can be running. Price comes later.

FIGURE 1

Four signals separate deployed capacity from announced capacity.



But a delivery date is only as credible as the power behind it. As demand climbs, the gap between announced and deployed capacity makes supply harder to predict. Uncommitted or delayed capacity introduces timing risks for buyers. In this environment, the buyer pays in months, but the real megawatts go to someone who asked sooner. Buyer diligence has to separate announcement from deployment.

> A hypothetical gigawatt is worth less than a real megawatt.
>
> — STEPHEN SOPKO

Four signals do that work:

- energized power
- ready-for-service (RFS) commitments from facility partners
- purchase orders with OEMs
- NVIDIA partner standing

On the OEM front, QumulusAI has deployed Lenovo and Supermicro servers on Cisco networking, and lists Dell among its OEM relationships. Providers that can point to these signals are showing compute coming online, which remains rare.

Management draws the line explicitly: QumulusAI is not primarily a powered-shell developer and does not broker powered land or GPU clusters. It deploys and operates clusters for its customers. Third-quarter results put that on the record.

Management said QumulusAI has sold every megawatt it has energized before it was turned on. On multiyear contracts, the company typically requires prepayments of 15% to 20% of contract value; contracts signed in Q3 carried $71 million in customer deposits, and deferred revenue rose to $79 million at September 30 from $30.5 million at June 30.

That validation speaks to the middle of the market. HyperFRAME Lens finds that 40% of I&O leaders are mass-market adopters who require validated business cases and proven ROI before committing (HyperFRAME Lens, 1H 2026). Those buyers are not looking to underwrite a roadmap, they need to see deployment histories before they sign.

QumulusAI's deployment history is public, and much of it is named:

ON THE FLOOR

What QumulusAI put on the floor in Q2 2026 and since.

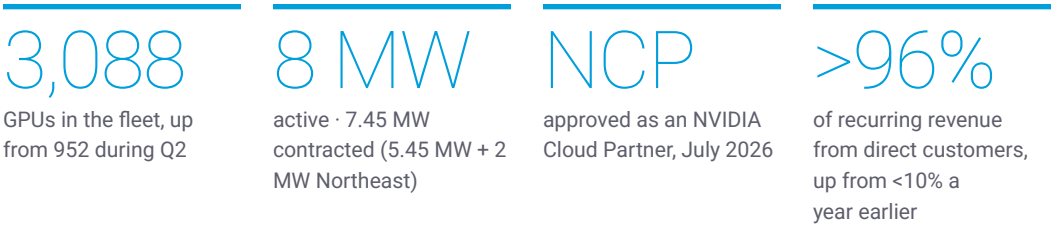


TABLE 1

QumulusAI's named deployment record, 2026. Plan against status, not announcements.

DATE	CUSTOMER · SCOPE	STATUS, NOV 2026
Early 2026	Shadeform · two NVIDIA H200 clusters	• Deployed
Aug 2026	Runpod · 616 RTX PRO 6000 Blackwell GPUs, 77 nodes	• Deployed ahead of target
Sep 1	GPU cloud marketplace · B300, Philadelphia · $18M, 2 years	• In production
Sep 8 / Oct 1	AI inference platform · B200 + B300 · $71.9M, 3 years	• In production
Q3 2026	DRW · dedicated B300 cluster · 1 year + three 1-year options	• Signed
Sep 29	Two existing customers · 2,000+ B300 · $240.9M, 3 years	• RFS Q2 2027

- QumulusAI completed deployment of 616 NVIDIA RTX PRO 6000 Blackwell GPUs across 77 nodes for Runpod, reaching full deployment in August 2026, ahead of the original September 1 target.
- At its Philadelphia colocation site, the company completed a Blackwell B300 deployment and customer handoff on September 1. That activated an $18 million, two-year take-or-pay agreement with a GPU cloud marketplace serving AI teams in more than 100 regions.
- B200 capacity under the $71.9 million, three-year agreement with an AI inference platform provider was handed off on September 8, and the B300 portion followed on October 1, putting the full agreement into production and billing roughly 10 weeks after signing.
- DRW, a global trading firm, signed for a dedicated Blackwell B300 cluster on a one-year initial term with three one-year renewal options. Management cited DRW on its Q3 call as an early example of the creditworthy enterprise customers it is adding alongside its AI-native base.
- Earlier in the year, QumulusAI announced two NVIDIA H200 cluster deployments with marketplace partner Shadeform.
- On September 29, 2026, QumulusAI announced $240.9 million in new three-year take-or-pay contracts with two existing customers, covering more than 2,000 B300 GPUs across two dedicated clusters. The company initially targeted early 2027 delivery. On its Q3 call, management said the 3.75 MW metro Atlanta block is expected to be ready for service in the second quarter of 2027, citing data center readiness, and that final equipment financing on one cluster is being aligned to facility timing. One of the two orders is fully prepaid; the other carries a 20% prepayment. At September 30, cumulative signed contract value stood at $528.7 million, with $515.0 million in remaining performance obligations.
- As of September 30, 2026, 8 MW was active (energized and billing) across five colocation sites. As of its November 10 earnings call, QumulusAI reported a further 5.45 MW of contracted capacity: 3.75 MW in metro Atlanta, expected ready for service in Q2 2027; 0.5 MW in Columbia, South Carolina, expected in December 2026; and 1.2 MW in Denver, expected in Q1 2027 and still being marketed to customers. On November 24, the company added 2 MW of contracted capacity in the U.S. Northeast, scheduled for service by December 31, 2026.

A contract backlog above $500 million signals strong customer demand, but buyers evaluating capacity should measure it against the "deployed, not announced" standard. QumulusAI's Q3 results make that test concrete: 8 MW is active, 7.45 MW is contracted,

40%

of I&O leaders are mass-market adopters who require validated business cases and proven ROI before committing.

HyperFRAME Lens: Infrastructure & Operations, 1H 2026 ›

"

Execution remains bound to final facility energization timelines, which is now the variable to watch.

— POCKETS OF POWER

and the largest contracted block has an RFS date of Q2 2027. Management said GPU and equipment allocation was arranged in advance and is not the bottleneck. Applying the four diligence questions to this pipeline shows strong OEM ties and defined GPU targets, while execution remains bound to final facility energization timelines, which is now the variable to watch.

SECTION 03

The Power Hunt: Why Five Megawatts Is a Strategy

Media attention and regulatory challenge concentrate on the 50MW-plus data center, and more so as power draw climbs into the gigawatts. Inventory that can be occupied quickly is not shaped like that.

QumulusAI defines “pockets of power” as sites of roughly 2 to 50MW where power and shell already exist or can be ready for service soon. Within that band, QumulusAI’s observed pipeline shows the fastest capacity arriving in single-digit megawatts: renovations, partial fits, facilities changing tenants, and stranded parcels of 1 to 2MW that larger buyers have passed over.

Today the fastest-moving buyers are AI-native platforms, with enterprise demand building behind them. Neither can wait out a multiyear build, and that is where small parcels win. They come to market first because they sit behind interconnections and buildings that already exist.

None of this makes gigawatt campuses a mistake. They remain essential to frontier AI. QumulusAI simply is not chasing them at the moment, a deliberate trade the company names openly: it is running a different business on a different clock.

A small parcel delivers more than its size suggests. QumulusAI’s Metro Atlanta agreement shows what a good pocket looks like. Signed for seven years, the colocation agreement covers up to 3.75MW, with a right of first offer on as much as 7MW of additional contiguous capacity. According to the announcement, that footprint is expected to support up to 2,048 B300 class GPUs. Delivery is conditioned on the provider completing the existing occupant’s transition out of the premises, and the company now targets activation in Q2 2027. Small enough to contract and fill quickly, the site also carries a path to 10.75MW if the provider builds the additional capacity and QumulusAI exercises its right of first offer.

BUYER CHECKLIST

Four Questions to Ask QumulusAI or Any Competitor

01 How many megawatts have you energized, as opposed to announced?

02 For whom, and can those customers be named?

03 How often do you deploy, and at what cadence?

04 What stands behind your next delivery date: RFS commitments, OEM purchase orders, financed hardware?

FIGURE 2

Where QumulusAI hunts. The fastest capacity arrives in single-digit megawatts inside the 2–50 MW band. Scale is not linear.



> A pocket of power that has already been sold is only as fast as the slowest party in the building.
>
> — POCKETS OF POWER

The Northeast agreement announced in November follows the same pattern: a 10-year colocation agreement for 2 MW, scheduled for service by year-end 2026, with a right of first offer on up to 9.6 MW more at the site, subject to credit criteria. QumulusAI said it is funding the first phase from cash generated by capacity already in production.

Finding and qualifying those parcels is an operating discipline. QumulusAI sources through land, power, and shell partnerships, colocation, and brownfield sites, with capital markets and procurement running in parallel. That parallel motion is the operational content behind the company's approach to developing line of sight to capacity before a customer needs it.

Unit economics hold. Management reports that recent Blackwell contracts generate $19 million to $21 million of annualized revenue per megawatt, against a blended installed base closer to $16 million. On those figures, small sites are no compromise on AI economics, which leaves sourcing discipline, more than scale, as QumulusAI's differentiator.

One objection deserves a direct answer. A portfolio of small sites multiplies the landlords, contracts, network links, and field operations that a single campus consolidates. That overhead is real. QumulusAI's model answers it three ways:

- Repeatability. One sourcing and deployment playbook runs at each site, and colocation partners carry much of the facility operation.
- Contract duration. QumulusAI's weighted average term of ~2.5 years tracks hardware generations more closely than the multiyear commitments common in large-campus deals.
- Repricing optionality. Part of the book renews each year while GPU pricing has been firming.

Speed and repricing have to outrun the added coordination cost. Gross margin as the site count grows is where that bet will show.

The Atlanta agreement also shows the model's exposure. Activation depends on the existing occupant vacating, a timeline set by the facility provider and its tenant, not by QumulusAI. A pocket of power that has already been sold is only as fast as the slowest party in the building, and it is unclear how often that party will be someone else.

Competitively, the segmentation from our Q1 2026 brief holds at category level: hyperscale providers, GPU marketplaces, and emerging GPU cloud providers. One nuance has sharpened. Some marketplaces are also QumulusAI customers, buying dedicated capacity to serve their own users. In this market, the line between competitor and channel is drawn contract by contract.

UNIT ECONOMICS

Management-reported figures. Small sites are no compromise on AI economics.

$19–21M
annualized revenue per MW on recent Blackwell contracts

$16M
per MW across the blended installed base

~2.5 yrs
weighted average contract term, closer to hardware generations

SECTION 04

Capacity Planning as a Partnership

Speed is not one number, and a fair brief should say so. A provider with an idle pool of shared, virtualized capacity can often hand over instances faster than anyone building a dedicated environment. For burst and shared workloads, that pool is a real advantage.

Even so, that speed comes with a trade. Shared-pool capacity serves many customers from the same fleet, and while reservation products can guarantee a window, capacity built for one customer's roadmap is a different product. Dedicated private clusters work differently. Where capacity is available on the floor, it can be provisioned in hours, as our March brief described. New dedicated clusters follow a demand-led path: each environment is procured, configured, and deployed for a specific customer. For that product, the delivery clock starts when demand becomes visible, not when a contract is signed.

The urgency is real. Several of the largest infrastructure providers have told investors their capacity is largely committed through 2027, with demand already building into 2028. For a buyer that needs capacity in the next few quarters, waiting for a large provider's queue is no longer a plan. The question becomes where supply can actually come from. QumulusAI's answer is to keep acquiring pockets of power and selling them before they are energized. That is a growth motion, not a backlog that has stopped growing.

> "For a dedicated cluster, the delivery clock starts when demand becomes visible, not when a contract is signed.
>
> — POCKETS OF POWER

Capacity planning belongs at the front of the relationship, because customer forecasts shared across coming months and quarters let the provider run site sourcing, financing, and hardware orders ahead of signature. QumulusAI already runs near-term deployments into available pockets alongside planning for larger future deployments. A customer's forecast decides which path its demand belongs on.

Our 2026 brief introduced capacity planning as a service within QumulusAI's F.A.C.T.S. (Flexibility, Access, Cost, Trust, and Speed) framework. Six months of deployments suggest it has become a delivery mechanism. In practice, this is where the Speed in the framework is earned. Procurement cycles were built for products that would still be on the shelf next quarter. GPU capacity is not one of them.

SECTION 05

Positioning for 2027: Repeating the Play

QumulusAI's 2027 growth case rests on executing the same sourcing motion across multiple sub-50MW projects, not one large build. While the discipline is portable, nothing in the model confines it to small sites. When a larger opportunity presents with power ready and a credible path to service, the same demand-led approach applies. Atlanta's potential path from 3.75MW toward 10.75MW, if the provider builds the additional capacity, shows a pocket growing in place without becoming a four-year construction project.

FIGURE 3

QumulusAI operating locations as of this brief. Local power availability, not corporate strategy, sets the map. Contracted capacity adds Columbia, South Carolina and a site in the U.S. Northeast.



Local power availability, not corporate strategy, sets the map. Sites are selected where power is ready, and geographic distribution follows as a byproduct. QumulusAI lists six locations as of this brief: Marietta, Georgia; Kansas City, Missouri; Denver, Colorado; Philadelphia, Pennsylvania; Watonga, Oklahoma; and Denton, Texas. The last two operated under long-term land leases.

Over time, that byproduct may become a product characteristic. In a recent earnings call, here is how QumulusAI CEO Mike Maniscalco articulated the challenge:

> "Today, distribution is a means to an end. Over time, we think it becomes the end itself. As workloads shift from training to inference, they get more particular about where they run. Budget matters. Latency starts to matter. Proximity to the end user starts to matter. So does the security and compliance posture of a specific building. When customers start choosing on those terms, a network of sites in different places stops being an operational complexity and becomes a differentiated product."

"

As workloads shift from training to inference, they get more particular about where they run.

— MIKE MANISCALCO, CEO, QUMULUSAI

QumulusAI states its near-term constraint plainly: it is not demand constrained, and land, power, and shell are the binding limit. We treat that as the operating thesis.

SECTION 06

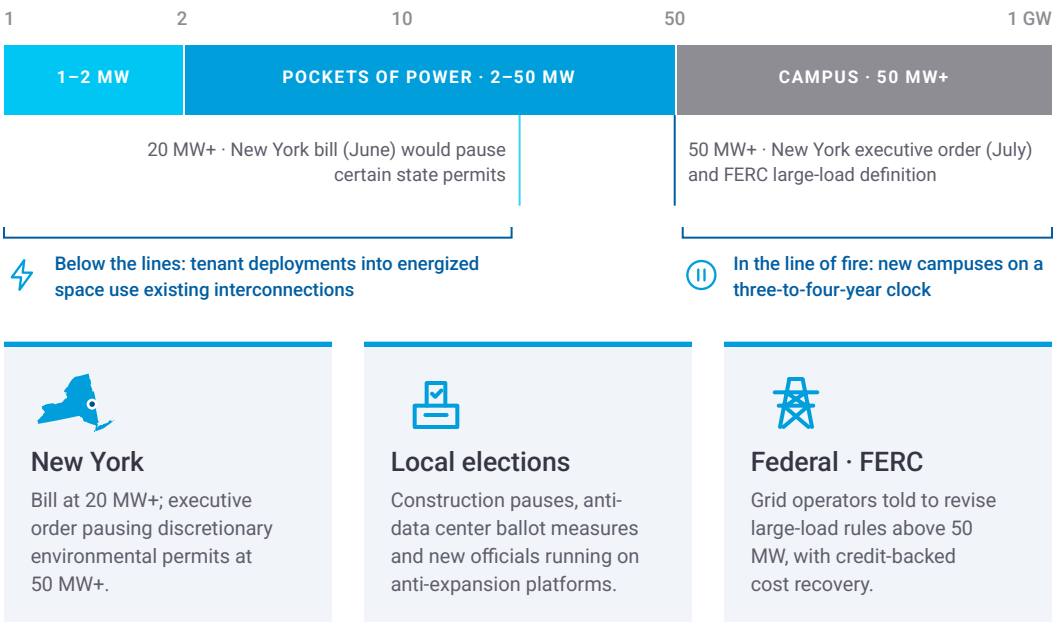
Regulatory Pushback on Large Campuses

For large data centers, the regulatory climate turned sharply in 2026, and its effects were further amplified by the outcomes of the US November 2026 midterm elections. Anti-data center sentiment emerged as a key ballot issue in several state and local races, consolidating bipartisan voter skepticism around grid reliability, rising utility rates, and land use. These results marked a turn from financial incentives toward a range of reactions from heightened oversight to outright restrictionism. Key developments include:

- New York. The legislature passed a bill in June that would pause certain state permits for data centers of 20MW or more. Governor Hochul instead issued an executive order in July pausing discretionary state environmental permits for data centers of 50MW or more while the state studies grid, water, and community impacts.

FIGURE 4

Where 2026 regulation draws its lines. Thresholds cluster on the large campus; existing energized sites sit below them.



> For a buyer whose 2027 capacity depends on a specific new campus coming online, regulatory exposure is now a deployment risk.
>
> — POCKETS OF POWER

- Local elections and municipalities. Dozens of municipalities have moved ahead with local construction pauses, with several anti-data center ballot measures passing and newly elected local officials running on explicit anti-expansion platforms.
- Federal. FERC ordered all six jurisdictional grid operators in June to justify or revise their large-load interconnection rules, defining large loads as those above 50MW at transmission voltage. Those orders aim to speed interconnection, but they also require cost recovery agreements backed by credit support, which adds financial obligations for large-campus developers.

The pattern points at scale. Midterm election results and legislative momentum indicate that political pushback lands hardest on the 50MW-plus campus, the category already running on a three-to-four-year clock. Each adds variance: a political shift in local leadership, a paused permit, a credit-support requirement, or a contested local hearing. For a buyer whose 2027 capacity depends on a specific new campus coming online, regulatory exposure is now a deployment risk. If faster federal study timelines materialize, part of the large campus's time penalty may narrow. Permitting and local opposition exposure would remain.

Smaller deployments sit largely outside the primary regulatory line of fire, though small sites are by no means exempt from scrutiny. Local ordinances, municipal moratoria, and ballot measures are frequently passed without explicit MW size thresholds, meaning even smaller builds can encounter community pushback or zoning delays. Nonetheless, a tenant deployment into an already energized facility generally relies on existing interconnections and approvals rather than entering a new large-load queue or facing lengthy environmental impact reviews. Expansion capacity, including new contiguous space, may face scrutiny akin to a new build, but existing energized space retains a lower risk profile and a shorter clock.

QumulusAI's site sizing predates the regulatory turn: our Q1 2026 brief documented QumulusAI's sub-50MW, colocation-based approach months before New York's executive order and before FERC wrote a 50MW threshold into federal policy. The company chose small, energized sites for speed. Regulators have since drawn their lines around the category it was already avoiding. A playbook built for one constraint now hedges a second, and QumulusAI has been running it since before either one drew attention.

OUTLOOK

Looking Ahead

Through 2027, capacity additions are likely to be decided by sourcing discipline more than by capital access. Capital is generally available even after market pullbacks; QumulusAI management describes more financing options now than ever. Energized space is not following the same curve; creativity, adaptability and track record are key there.

For CIOs and infrastructure leaders, three actions follow.

01

Audit current time to capacity.

Measure from first internal forecast to production workloads, not from purchase order to handoff.

02

Evaluate providers on deployed, not announced, megawatts.

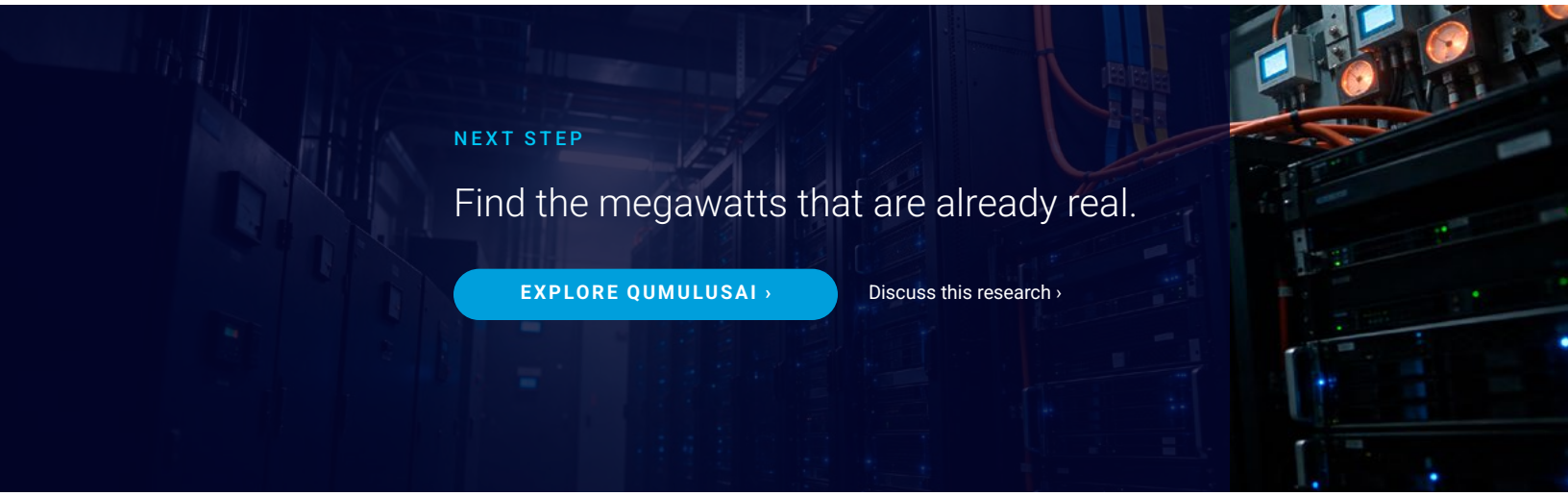
Use the four diligence questions. Ask for named customers and handoff dates.

03

Treat capacity planning as a forward partnership.

Share forecasts early, map demand across quarters, and let the provider source against them.

The gigawatt headlines will continue. The next megawatt makes revenue.



NEXT STEP

Find the megawatts that are already real.

EXPLORE QUMULUSAI ›

Discuss this research ›

ABOUT QUMULUSAI

VISIT QUMULUSAI.COM ›

QumulusAI

QumulusAI designs and deploys single-tenant GPU environments with committed capacity, customer-controlled workloads, and one infrastructure partner accountable from design through operation.

The company sources energized capacity in pockets of roughly 2 to 50 MW and deploys and operates dedicated clusters for AI-native platforms and enterprises.

8 MW energized
Billing across five colocation sites, Sept. 30, 2026.

NVIDIA Cloud Partner
Approved July 2026.

$528.7M signed
Cumulative contract value, Sept. 30, 2026.

Sponsor description based on QumulusAI's public materials; to be approved by the sponsor.



> "The gigawatt headlines will continue. The next megawatt makes revenue."
>
> — POCKETS OF POWER

ABOUT THE ANALYSTS



Stephen Sopko
Practice Lead, HyperFRAME Research

Specializes in semiconductors and the deep technologies powering today's innovation ecosystem, with decades of executive experience across Fortune 100, government and startups.



Steven Dickens
CEO & Principal Analyst, HyperFRAME Research

Top 10 analyst (AR Insights) and Forbes contributor covering hybrid cloud and mission-critical infrastructure; former senior leader at Broadcom, HPE and IBM.

CREDITS

About HyperFRAME Research

AUTHORS

Stephen Sopko
Practice Lead
Steven Dickens
CEO & Principal Analyst

CONTACT

Steven Dickens
CEO & Principal Analyst

steven.dickens@hyperframeresearch.com

+1 845 505 1678

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CONTACT US ›

FOLLOW HYPERFRAME

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between.

We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs. Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors.

DISCLOSURES

This research brief was prepared in collaboration with QumulusAI. QumulusAI reviewed it for factual accuracy only. The analysis and conclusions are HyperFRAME Research's. HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.

LICENSING

This document, including any supporting materials, is owned by HyperFRAME Research. This publication may not be reproduced, distributed, or shared in any form without the prior written permission of HyperFRAME Research.

COPYRIGHT

© 2026 HyperFRAME Research. All rights reserved. NVIDIA, Blackwell and RTX are trademarks of NVIDIA Corporation. All other trademarks are the property of their respective owners.



HYPERFRAMERESEARCH.COM



© 2026 HYPERFRAME RESEARCH. ALL RIGHTS RESERVED.