Neurengy

GLOBAL DATA CENTER & AI INFRASTRUCTURE REPORT 2026

The Power–Compute Gap

Global data-center infrastructure, AI demand and the search for deployable capacity

MARKET DYNAMICS • INFRASTRUCTURE BOTTLENECKS • SOLUTIONS • ESG

How grids, connected power, modular infrastructure and distributed orchestration may reshape the next phase of AI capacity growth.

THE CENTRAL QUESTION

How can the world deploy the next wave of AI capacity when usable power, time-to-power and infrastructure integration are becoming the binding constraints?

Published by Neurengy Research • August 2026 • neurengy.com

About this report

The Power-Compute Gap examines the physical infrastructure behind the AI economy: electricity, grid access, land, cooling, connectivity, servers, capital and the operating systems that bind them together.

It combines intergovernmental and official statistics, recognised market research and clearly labelled company disclosures. Neurengy's opportunity is assessed only after the market problem, regional dynamics and solution landscape have been established.

- Data cut-off: 18 August 2026
- Regional lenses: Americas; Europe including Russia; Asia including China; Greater East covering MENA and CIS
- Publisher and editorial responsibility: Neurengy Research / Pavel Cheremisin
- ChatGPT assisted research synthesis, drafting and visual production; Pavel Cheremisin retains editorial responsibility for the published version

DISCLAIMER

This is public thought leadership, not an offer, investment recommendation, engineering design or guarantee of project economics.

EXECUTIVE SUMMARY

AI growth is becoming an infrastructure deployment problem

AI demand is expanding into an electricity system that was not planned for clusters of high-density, continuously operating digital load. IEA expects global data-center electricity consumption to rise from about 485 TWh in 2025 to around 950 TWh in 2030, while JLL sees global installed capacity approaching 200 GW by the end of the decade [1,3].

The binding constraint is increasingly access to usable power in the right place and time. A GPU order does not create a grid connection, a cooling system, a fibre route, a permitted site or a bankable service. The commercial opportunity therefore shifts toward companies able to assemble these layers faster and more flexibly.

- Hyperscale campuses remain essential for tightly coupled training and very large workloads.
- A complementary distributed layer can serve inference, batch, rendering and interruptible demand.
- Existing connected sites may shorten time-to-compute, but only after legal, temporal and engineering qualification.
- Sustainability depends on measured utilisation, energy source, water, heat reuse and avoided construction — not on the word “distributed”.

BOTTOM LINE

Window of opportunity: convert fragmented infrastructure into managed capacity without pretending that nominal power is already a data center.

The report in ten numbers

950 TWh

Global data-center electricity demand in 2030 - IEA reference [1]

~200 GW

Potential global installed capacity in 2030 - JLL [3]

3x

AI-focused data-center electricity demand, 2025-2030 [1]

~14%

Implied annual growth of global installed capacity, 2025-2030 [3]

30%

Annual growth in accelerated-server electricity use [2]

11.8%

US electricity share in LBNL 2030 reference case [5]

+170%

China data-center electricity demand, 2024-2030 [2]

+130%

US data-center electricity demand, 2024-2030 [2]

~80%

Share of global growth from China and the United States [2]

$3tn

Upper-range infrastructure and IT investment through 2030 [3]

Contents

Click any entry to open that page. The PDF outline provides the same navigation in the sidebar.

PART I The infrastructure shift 6

AI demand is accelerating faster than the... 7

The market is entering an infrastructure... 8

Deployable compute requires six systems to arrive... 9

Nominal megawatts are not deployable megawatts 10

Infrastructure architecture should follow... 11

PART II The bottlenecks 12

Time-to-power is becoming a competitive variable 13

Data centers are becoming material grid-planning... 14

Density, cooling, water and social licence now... 15

Hardware architecture, connectivity and security... 16

A $3 trillion cycle will require more than... 17

PART III Regional dynamics 18

Demand leadership meets the world's most visible... 19

Efficiency, sovereignty and heterogeneous power... 20

China anchors the fastest large-market growth in... 21

Capital-rich hubs and power-rich corridors create... 22

Comparable statistics support five major growth... 23

The strongest demand hubs are often the most... 24

PART IV The solution landscape 25

Grid expansion and on-site power solve different... 26

Interruptible compute can turn some AI demand... 27

Standard modules can convert qualified existing... 28

Software can make heterogeneous sites behave like... 29

Reuse can improve sustainability — but only with... 30

The market is moving from commitments to... 31

The climate question is net impact, not the... 32

Distributed assets need an explicit ownership and... 33

PART V Market implications 34

Four close analogues validate distinct layers of... 35

XFRA, Pacific, Perimeter and Deep Green: what... 36

Four independent models converge on... 37

A hybrid market is the most robust response to... 38

What leaders should do now - and what to watch next 39

Existing power can become a new layer of managed... 40

Discover -> Deploy -> Operate -> Orchestrate ->... 41

The opportunity remains conditional until... 42

BACK MATTER Methodology 43 | Sources 44-46

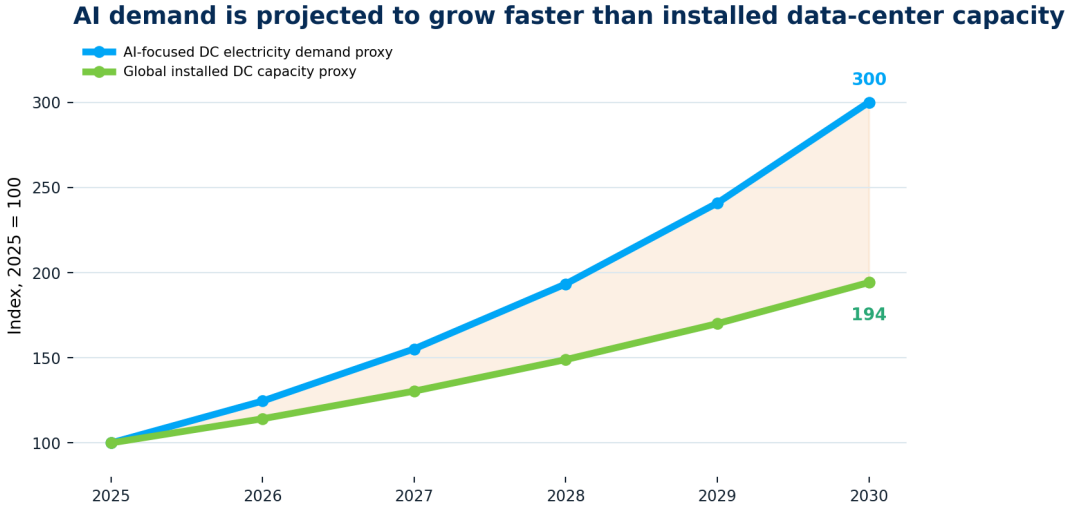
PART I

The infrastructure shift

AI demand is changing not only the quantity of capacity required, but its power density, geography and operating profile.

THE GLOBAL GAP

AI demand is accelerating faster than the capacity pipeline



Direction-of-travel comparison: IEA AI-focused demand triples; JLL capacity rises from ~103 GW to ~200 GW. Different units - not a MW shortage estimate.

The IEA expects electricity consumption in AI-focused data centers to triple from 2025 to 2030, while total data-center electricity demand rises from about 485 TWh to roughly 950 TWh. JLL's global outlook places installed data-center capacity at about 103 GW in 2025 and close to 200 GW by 2030 [1,3].

Indexed to 2025, these trajectories imply an AI-demand proxy reaching 300 versus an installed-capacity proxy near 194. This is not a direct MW shortage calculation: electricity demand and installed facility capacity are different measures. It is a directional signal that demand intensity, utilisation and deployment requirements are rising faster than the physical base.

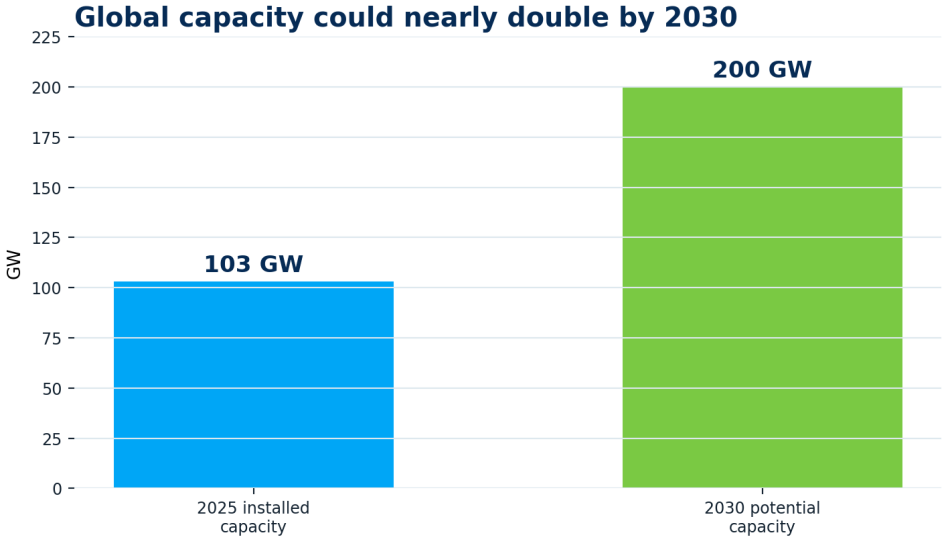
- AI-focused demand proxy: approximately 25% annual growth.
- Installed-capacity proxy: approximately 14% annual growth.
- Directional spread: roughly 11 percentage points per year.
- Real delivery will be lower than announced pipeline where power, permits or equipment slip.

BOTTOM LINE

The opportunity is created by the speed mismatch, not by a claim that every projected AI workload is already contracted.

CAPACITY

The market is entering an infrastructure investment supercycle



Source: JLL 2026 Global Data Center Outlook. Boundaries and pipeline definitions vary by market.

JLL anticipates roughly 100 GW of new capacity between 2026 and 2030 and a global installed base near 200 GW. Demand is expected to remain ahead of supply in the most constrained markets [3,24].

The opportunity is large, but announced pipelines overstate what can become operational on schedule. Power procurement, grid studies, equipment lead times and community acceptance increasingly determine delivery.

BOTTOM LINE

Capacity announcements should be discounted by power, permits, construction and customer readiness.

THE POWER–COMPUTE GAP

Deployable compute requires six systems to arrive together

AI capacity is a system, not a shipment of GPUs

Failure at any layer can strand the entire investment.

The AI economy is often described through chips and models. In practice, usable capacity emerges only when energy, connection rights, facility systems, servers, networks and workload software are integrated.

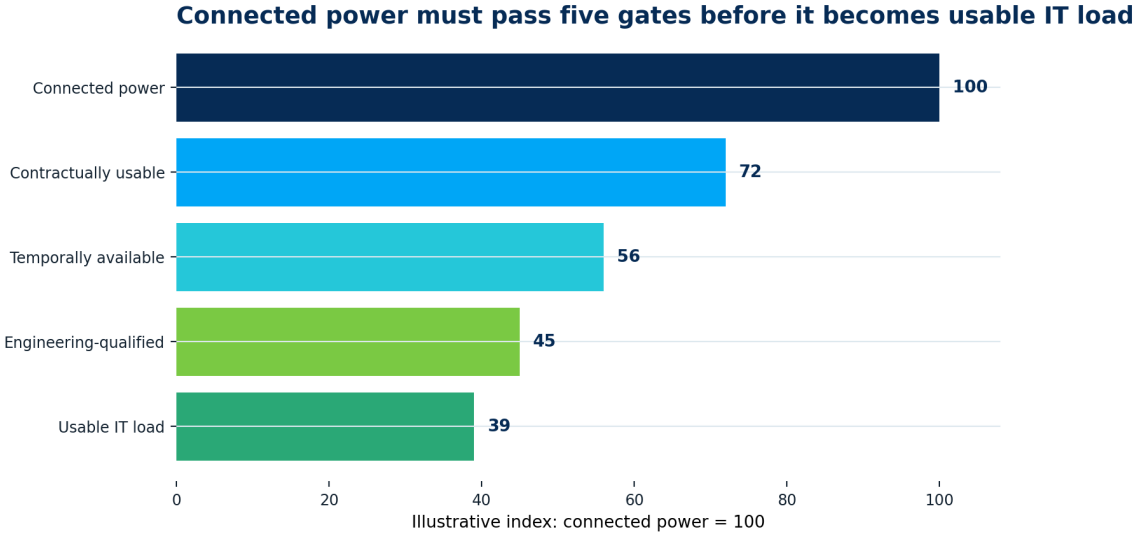
This explains why an apparently abundant supply of electricity or GPUs can coexist with scarce time-to-compute. A bottleneck at one layer can strand capital committed at every other layer.

BOTTOM LINE

The scarce product is not a GPU. It is reliable, connected and commercially usable compute.

CAPACITY DEFINITIONS

Nominal megawatts are not deployable megawatts



Method illustration only. Actual conversion depends on contracts, load profile, site engineering, connectivity and PUE.

Connected or mapped power is only the starting point. The site must permit the intended load, remain available during the required hours, support cooling and fire safety, provide connectivity and survive facility-to-IT derating.

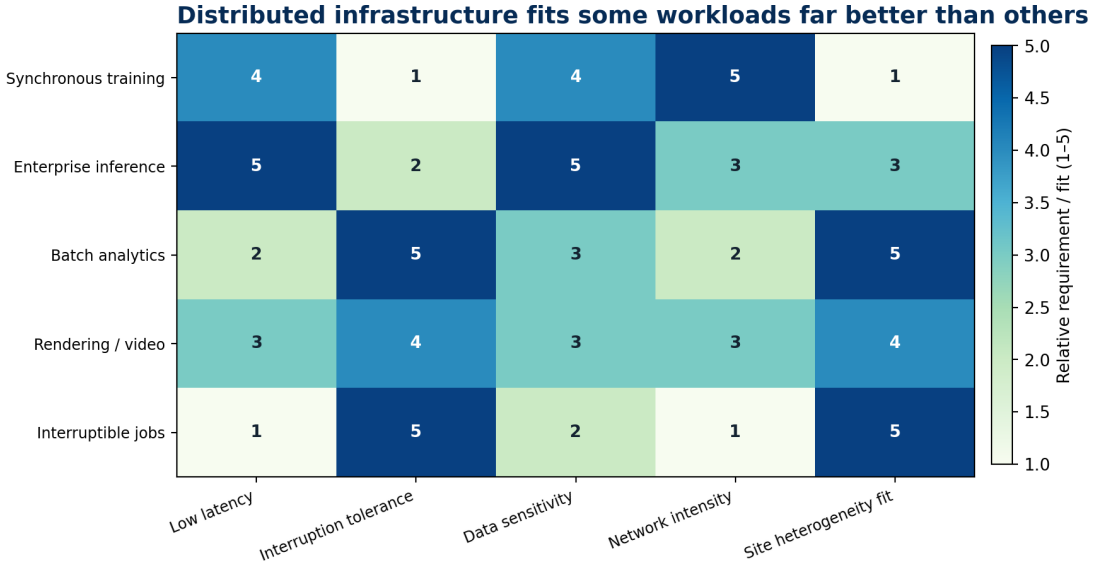
The funnel shown is an internal analytical scenario, not a market census. Its value is methodological: it prevents strategy from treating an electrical nameplate as bankable IT inventory.

BOTTOM LINE

Every site lead needs documents, telemetry, engineering review and a commercial workload.

WORKLOADS

Infrastructure architecture should follow workload physics



Synchronous training depends on homogeneous accelerators and high-bandwidth, low-latency fabrics. It is naturally concentrated. Inference, batch analytics, rendering and interruptible jobs can tolerate more diversity and geographic distribution.

A distributed provider should therefore sell workload outcomes and availability tiers, not claim universal equivalence with a hyperscale cluster.

BOTTOM LINE

The addressable market is defined by workload fit, not by total AI spending.

PART II

The bottlenecks

Power scarcity is the headline, but delivery depends on a chain of grid, engineering, regulatory and financial constraints.

GRID ACCESS

Time-to-power is becoming a competitive variable

New large loads require system studies, network reinforcement and coordination with generation and transmission plans. In the most congested hubs, connection timelines can exceed the commercial planning horizon of AI customers.

This creates two responses: move campuses toward frontier power markets, or activate smaller blocks of existing connected capacity. The second option may be faster, but it also creates a portfolio-management challenge.

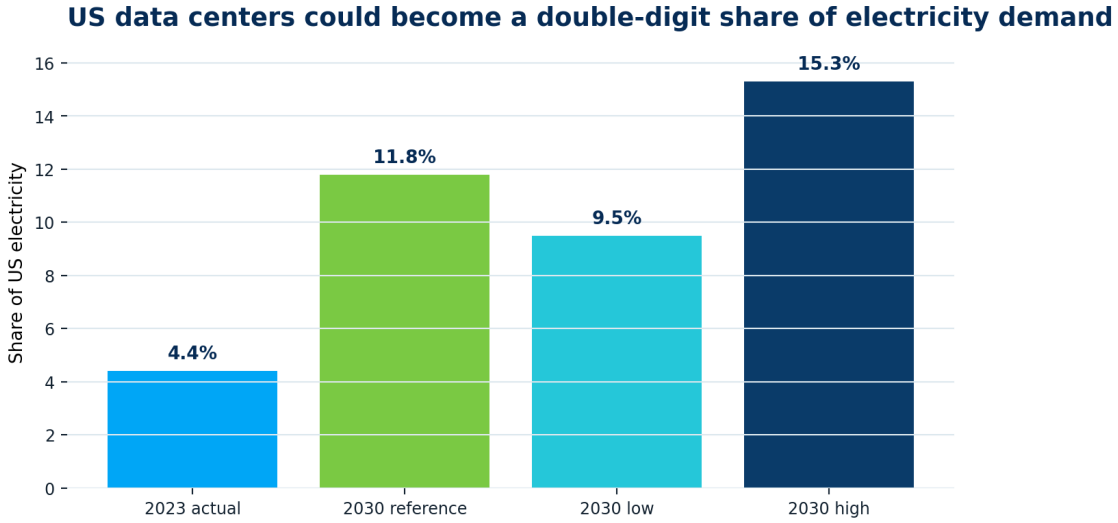
- Connection queue position is not the same as an energisation date.
- Behind-the-meter generation can accelerate supply but may increase fuel and emissions exposure.
- Flexible-load contracts can improve grid fit if workloads can genuinely be interrupted.
- Existing sites require proof that spare capacity is legally and operationally usable.

BOTTOM LINE

The winning model may optimise time-to-power, not simply energy price.

SYSTEM ADEQUACY

Data centers are becoming material grid-planning loads



Source: Lawrence Berkeley National Laboratory, 2025 Update published 2026.

LBNL's 2025 update estimates that US data centers could account for 11.8% of electricity use in 2030, with a scenario range of 9.5–15.3% [5]. Concentrated growth can affect capacity markets, network investment and local tariffs.

Large loads can also become system resources when paired with storage, generation or controllable workloads. That value requires dispatchable control and a contract with the system operator or utility — not merely the theoretical ability to shut down.

BOTTOM LINE

Compute can be a grid burden or a flexible resource; control architecture determines which.

FACILITY, WATER AND PERMISSION

Density, cooling, water and social licence now form one site constraint

Accelerator density makes liquid cooling, hydraulic design, electrical protection and fire safety part of the capacity decision. A site with spare breaker capacity is not deployable until its thermal system, structure and maintainability are proven [12–15].

Water and community impacts are local. Climate, cooling architecture, noise, backup generation and who pays for grid reinforcement increasingly affect permission to build. Reusing an existing building may improve the social-licence case, but only when impacts and local benefits are reported transparently.

- Track IT energy, facility energy, water consumption and local water stress.
- Design heat reuse around a real offtaker and seasonal demand.
- Publish noise, backup-generation and network-impact assumptions.
- Select sites on total infrastructure fit — not electricity price alone.

BOTTOM LINE

Electrical availability, cooling feasibility and permission must be qualified together.

HARDWARE, NETWORK AND TRUST

Hardware architecture, connectivity and security must be designed together

GPU economics depend on server architecture, memory, fabric, storage and utilisation. Tightly coupled training needs homogeneous accelerators and high-bandwidth networking; inference and batch create a wider design space that can be segmented by latency, data sensitivity and interruption tolerance [13–15,35].

Distribution expands the compliance perimeter. The operating layer must combine workload classification with identity, segmentation, secure boot, patching, attestation, observability, incident response and customer-specific data-location policies.

- Procure hardware against contracted workload requirements.
- Maintain an auditable site, firmware and hardware inventory.
- Separate tenant, host and platform privileges.
- Route workloads by physics, policy and recovery objectives.

BOTTOM LINE

A distributed fleet is commercially useful only when heterogeneity and trust are managed as one system.

CAPITAL

A $3 trillion cycle will require more than corporate balance sheets

The AI infrastructure cycle may require up to $3 trillion by 2030

Illustrative upper-range split based on JLL: ~1.2*tnreal − estatevalueplus* 1–2tn tenant IT investment.

JLL frames the expansion as an infrastructure supercycle: up to $1.2 trillion of real-estate value and a further $1–2 trillion of tenant IT investment by 2030 [3].

Distributed infrastructure can reduce the size of individual projects, but it does not remove capital risk. Hardware obsolescence, utilisation, host default, removal rights and residual value must be allocated among the platform, site owner, equipment SPV and lender.

BOTTOM LINE

Bankability comes from contracts and control rights, not modularity alone.

III

PART III

Regional dynamics

The global market is fragmenting into demand hubs, frontier power markets, capital hubs and sovereign-compute zones.

AMERICAS - UNITED STATES FIRST

Demand leadership meets the world's most visible power bottlenecks

The United States is the principal AI-infrastructure market by customer depth, capital, cloud ecosystems and accelerator deployment. IEA projects US data-center electricity demand to rise by around 240 TWh between 2024 and 2030 - about 130% - while LBNL's reference case places the sector at 11.8% of national electricity use in 2030 [1,5].

The constraint is local rather than national: major clusters face interconnection queues, network upgrades, land and community resistance. Canada offers lower-carbon power and cooler climates in selected provinces; Latin America adds renewable potential and growing cloud demand, but fibre, market depth and regulatory consistency vary materially.

- United States: deepest demand, most acute time-to-power pressure.
- Canada: hydro-rich options with province-specific grid constraints.
- Latin America: selective hubs rather than one homogeneous market.
- Best-fit responses: frontier campuses, flexible load, existing industrial sites and metropolitan inference.

BOTTOM LINE

The Americas support both hyperscale concentration and a complementary distributed layer close to users and spare power.

EUROPE - INCLUDING RUSSIA

Efficiency, sovereignty and heterogeneous power systems shape deployment

The EU combines mature demand with grid congestion, high energy costs, reporting obligations and growing attention to water and waste heat. Sovereign-cloud requirements and regional inference favour capacity distributed across several national markets rather than a single continental hub [10,11,41].

Russia is treated here as part of the wider European infrastructure geography, not as a separate global market. It contributes engineering capability, industrial sites and large power-system regions, while equipment access, financing, sanctions, data rules and cross-border interoperability create material constraints. Across Europe, heat reuse can create local value only where temperature, distance and year-round offtake are proven.

- Western and Northern Europe: demand depth, grid scarcity and strict disclosure.
- Nordics: lower-carbon power but finite transmission and customer-location trade-offs.
- Central and Eastern Europe: selective growth with sovereign and connectivity considerations.
- Russia: large technical base and potential sites, with elevated financing and technology-access risk.

BOTTOM LINE

Europe rewards measurable efficiency and local integration, but requires country-level legal and grid architecture.

ASIA - CHINA FIRST

China anchors the fastest large-market growth in a fragmented region

IEA projects China's data-center electricity demand to rise by around 175 TWh from 2024 to 2030 - about 170% - making China, together with the United States, responsible for nearly 80% of global growth. Domestic supply chains, sovereign platforms and large-scale renewable development support expansion, while national data rules limit assumptions about cross-border workload mobility [1].

Japan and Korea combine advanced demand with constrained land and power; India is scaling cloud and digital demand from a lower base; Singapore and southern Malaysia form a regional hub; Southeast Asian demand is expected to more than double by 2030. Cooling, water, grid carbon intensity and local operating capability vary sharply.

- China: scale, policy direction and domestic ecosystem.
- Japan and Korea: premium demand under land and grid constraints.
- India: growth potential with state-level infrastructure variation.
- Southeast Asia: fast growth, heat and humidity, and cross-border data complexity.

BOTTOM LINE

Asia requires national architectures; a regional label is not a deployable strategy.

GREATER EAST - MENA AND CIS

Capital-rich hubs and power-rich corridors create a different AI geography

MENA is emerging as a sovereign AI and capital hub. The UAE and Saudi Arabia can combine land, investment and energy development at scale, while cooling, water, supply-chain controls and the carbon profile of incremental power remain central design issues [28].

Across the CIS, Kazakhstan, Uzbekistan and other markets are using power availability, industrial zones and policy incentives to attract digital infrastructure. Announced projects are signals rather than operating capacity: bankability still depends on customer demand, fibre, equipment access, currency, data localisation and credible operators.

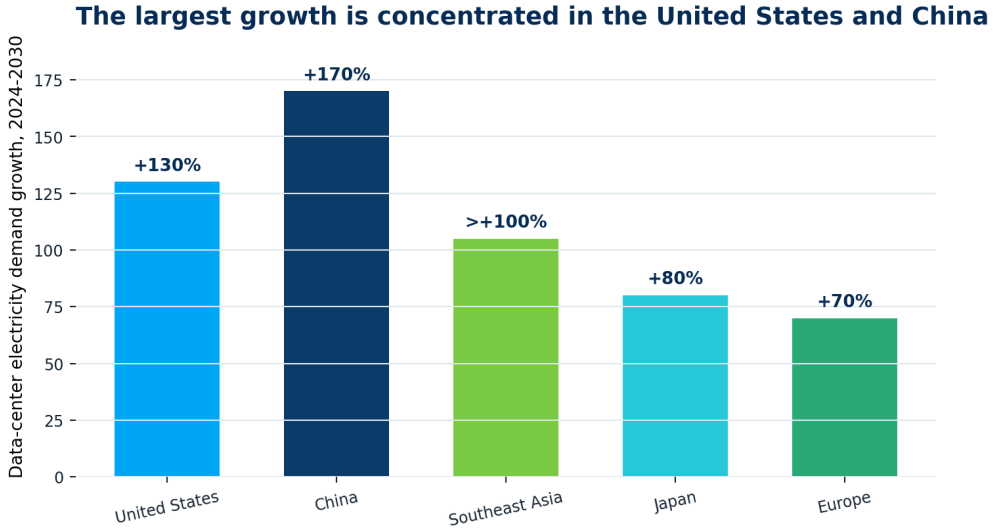
- MENA: capital and scale, with climate and technology-control exposure.
- Kazakhstan: power-led development and regional-compute potential.
- Uzbekistan: policy-led market formation and incentive zones.
- CIS portfolio: local demand and legal fit matter more than nominal energy abundance.

BOTTOM LINE

The Greater East can add capacity quickly in selected corridors, but power abundance alone does not create a compute market.

GLOBAL GROWTH CONCENTRATION

Comparable statistics support five major growth lenses - not a false top-ten ranking



Source: IEA Energy and AI. Southeast Asia is reported as more than doubling; 105% is used only to position the bar visually.

IEA's globally consistent model shows the United States and China dominating absolute growth, while Southeast Asia, Japan and Europe also expand rapidly. These are electricity-demand projections with a common methodology, making them more reliable for cross-market comparison than mixing country-level capacity announcements from incompatible datasets [1].

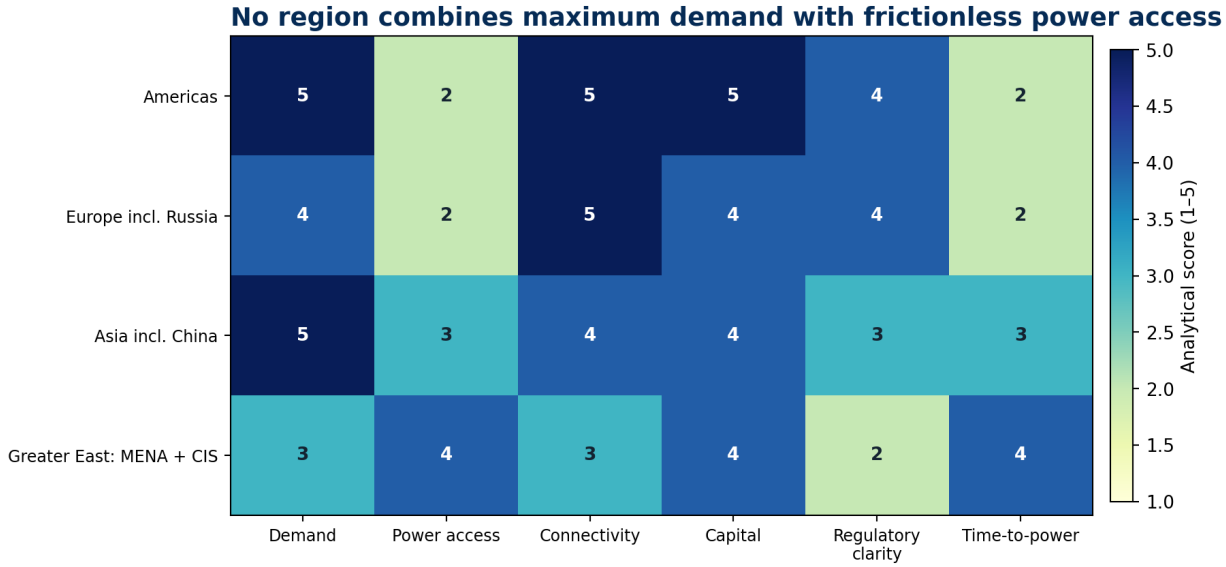
A top-ten country table is intentionally not presented. Public sources frequently mix commissioned IT load, facility power, colocation inventory, hyperscale campuses, contracted pipeline and announced projects. The report therefore uses fewer but more comparable markets and labels qualitative regional scorecards as analysis rather than statistics.

BOTTOM LINE

Comparability is more valuable than a longer ranking built from inconsistent capacity definitions.

REGIONAL SCORECARD

The strongest demand hubs are often the most power-constrained



The scorecard is an analytical framework rather than a measured index. It highlights the central trade-off: high-demand, well-connected markets typically face power and time-to-power constraints, while power-rich markets may lack customers, capital or regulatory clarity.

A portfolio model can bridge these differences only if workloads can move and customer contracts permit that movement.

BOTTOM LINE

Geographic diversification is useful when orchestration and regulation allow it.

IV

PART IV

The solution landscape

No single intervention closes the gap. The market is converging on portfolios of grid, facility, workload and financing solutions.

POWER SUPPLY

Grid expansion and on-site power solve different time horizons

Transmission, substations and new generation remain the foundation of hyperscale growth, but they often move more slowly than AI product cycles. Transparent queue milestones and coordinated planning are required to convert announced capacity into energised projects.

Gas generation, fuel cells, geothermal, renewables, storage and microgrids can shorten time-to-power. They also introduce fuel, emissions, permitting and operational risks. Distributed reuse is an interim and complementary layer — not a substitute for system investment [22,23,61].

- Align generation, network and large-load milestones.
- Use temporary generation with an explicit transition plan.
- Price fuel, carbon, maintenance and availability risk.
- Prioritise structurally advantaged energy and genuinely flexible workloads.

BOTTOM LINE

The portfolio needs both durable grid capacity and faster, risk-controlled bridging solutions.

FLEXIBLE LOAD

Interruptible compute can turn some AI demand into a grid resource

Batch analytics, rendering, checkpointable training and other delay-tolerant jobs can respond to network conditions or renewable output. The value comes from scheduling and contractual availability, not from shutting down servers at random.

Enterprise inference needs a different tier: failover, reserved capacity and clear recovery objectives. A distributed portfolio can combine firm and interruptible products, but must price them separately.

- Pre-define interruption windows and restart behaviour.
- Separate flexible queues from latency-critical service tiers.
- Measure avoided peak demand and actual dispatch performance.
- Share grid value contractually with customers and site hosts.

BOTTOM LINE

Flexibility should be designed as a product, not treated as degraded reliability.

MODULAR + RESERVE-FIRST DELIVERY

Standard modules can convert qualified existing power into capacity faster

Factory-built electrical, cooling and compute modules can compress design and commissioning after a site has passed contractual, electrical, thermal, connectivity and safety gates. Standard interfaces reduce bespoke work and make equipment easier to finance and relocate.

The strongest near-term input is not an empty plot but verified connected infrastructure at industrial plants, commercial buildings, campuses, retail estates and distressed assets. Each site still requires documents, interval telemetry, tariff analysis, access rights and engineering derating.

- Standardise the site audit as rigorously as the hardware.
- Use modularity to shorten repeat work, not to bypass qualification.
- Contract host access, operating hours and equipment-removal rights.
- Treat fragmented site diligence as a core operating capability.

BOTTOM LINE

The scalable product is a standard conversion process from verified reserve power to managed compute.

ORCHESTRATION

Software can make heterogeneous sites behave like a service

A distributed network needs workload admission, scheduling, observability, billing, security and failover. Locality-aware orchestration can place inference closer to users or data, reduce unnecessary backbone traffic and use regional capacity more effectively.

The benefit must be measured. Latency, transfer volume, model performance, job completion and energy per result should be compared with a centralised baseline.

- Route by workload physics and policy.
- Reserve homogeneous pools for demanding jobs.
- Use marketplaces only for residual capacity, not as the economic foundation.
- Expose clear service tiers to customers.

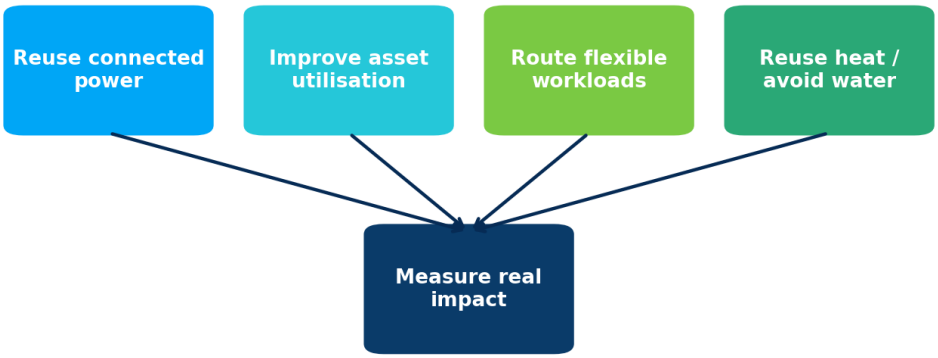
BOTTOM LINE

Orchestration converts a portfolio of assets into a commercial product.

ESG

Reuse can improve sustainability — but only with credible measurement

ESG value exists only when utilisation gains become measured outcomes



Reusing connected power and existing buildings can reduce the need for new shells and network works, improve utilisation of sunk infrastructure and create routes for heat recovery. Flexible workloads may also support grids with high renewable penetration.

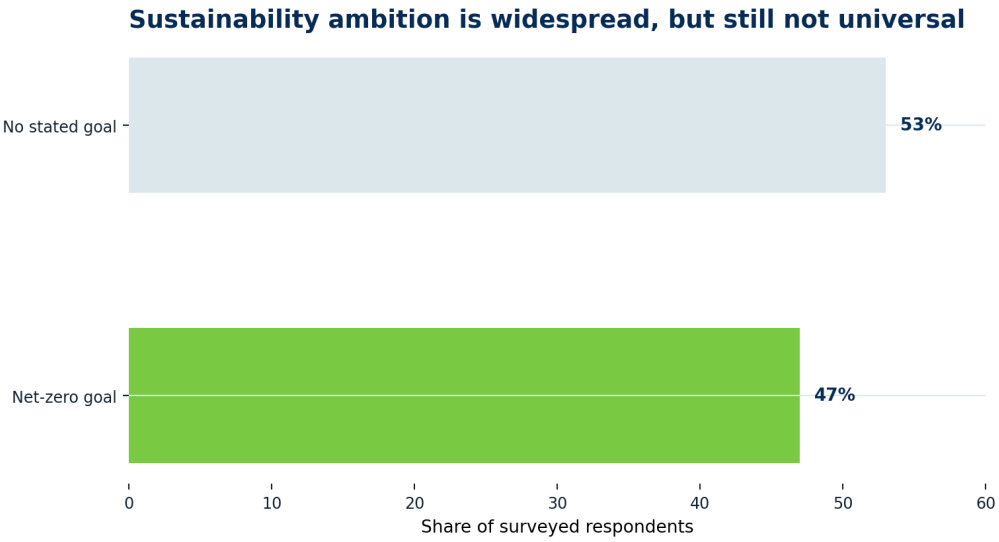
None of these benefits is automatic. Higher utilisation can increase absolute electricity use, and a reused site powered by carbon-intensive generation may perform worse than a new efficient facility. Claims require a baseline and a life-cycle boundary.

BOTTOM LINE

Neurengy’s ESG case should be an auditable performance thesis, not a green label.

ESG METRICS

The market is moving from commitments to operational evidence



Uptime Institute Sustainability Survey, 974 respondents, Dec 2024–Jan 2025.

Uptime Institute's 2024/25 sustainability survey found that 47% of respondents reported an organisational net-zero goal [39]. Regulation and customer procurement are increasingly asking for energy, water, carbon and waste-heat data.

A distributed platform can create a consistent measurement layer across small sites. That layer may become as valuable as the physical infrastructure because it turns fragmented assets into comparable, financeable and reportable units.

- PUE and useful compute per kWh
- Location- and market-based carbon intensity
- Water consumption and local water stress
- Heat captured and usefully delivered
- Embodied carbon and avoided new construction
- Hardware life, utilisation and circularity

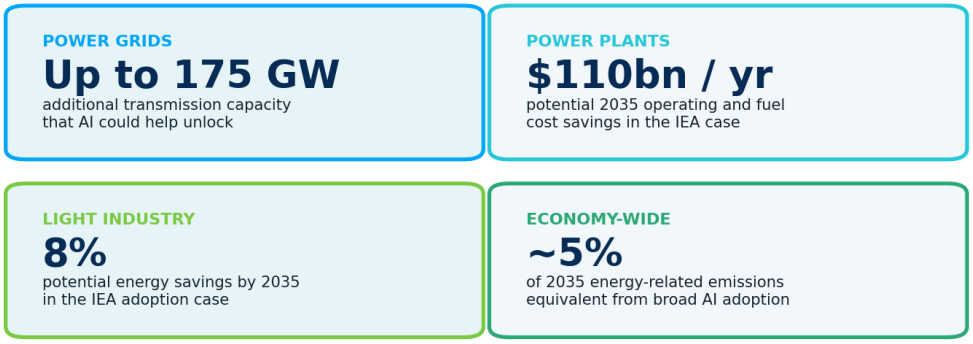
BOTTOM LINE

ESG becomes commercially relevant when it is measurable at workload and site level.

AI-ENABLED DECARBONISATION

The climate question is net impact, not the footprint of compute alone

AI can enable emissions reductions beyond its own electricity footprint



IEA Energy and AI, Widespread Adoption Case. These are conditional potentials, not guaranteed net benefits.

The IEA estimates that broad adoption of existing AI-led applications could enable emissions reductions equivalent to around 5% of energy-related emissions in 2035 — potentially much larger than data-center emissions, though far below what is required for net zero [2,68].

The enabling channels include grid forecasting and dynamic line operation, predictive maintenance, renewable integration, industrial process control, building demand response, transport optimisation, contrail avoidance and better electric-vehicle range. In the IEA adoption case, AI could help unlock up to 175 GW of transmission capacity and 8% energy savings in light industry [67].

- Count the electricity and embodied carbon of AI infrastructure.
- Measure the real-world energy or emissions outcome enabled by each workload.
- Use additionality: compare with a credible non-AI baseline.
- Account for rebound effects, induced demand and model overuse.
- Prefer workloads whose avoided emissions can be verified.

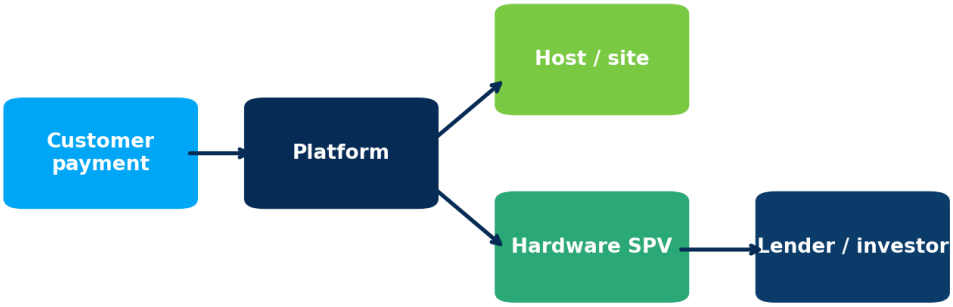
BOTTOM LINE

AI can become climate-positive only when enabled reductions are additional, measured and larger than its full life-cycle footprint.

FINANCE

Distributed assets need an explicit ownership and cash-flow architecture

Asset separation can make distributed infrastructure financeable



Separate software margin, site compensation, hardware return and debt service.

The platform, site, hardware and lender have different risk and return profiles. Separating them can keep the software company asset-light while giving equipment investors security over hardware and contracted cash flows.

The structure must include step-in rights, insurance, access, removal, replacement reserves and transparent payment priority. Otherwise a physically movable node may still be financially stranded.

BOTTOM LINE

SPVs help only when contracts make assets controllable and cash flows traceable.

V

PART V

Market implications

The next competitive frontier is the ability to assemble power, sites, workloads and capital into deployable capacity.

COMPETITIVE EVIDENCE

Four close analogues validate distinct layers of the reserve-first thesis

The closest analogues validate different layers of the Neurengy thesis

	Existing power as input	Standardised physical unit	Fleet / service orchestration	Energy or heat integration	Public stage (Aug 2026)
XFRA / SPAN	CORE	NODE	CORE	BATTERY / DR	PROOF OF CONCEPT
Pacific	CORE	POD	PLATFORM	COOLING	EARLY STAGE
Perimeter	CORE	RACK KIT	CORE	HEAT RECOVERY	OPERATING / EXPANDING
Deep Green	SITE-LED	MODULAR DC	OPERATOR	HEAT REUSE	LIVE + PIPELINE
Neurengy	CORE	NODE	CORE	OPTIONAL	VALIDATION STAGE

Company claims and announced plans; Neurengy remains pre-scale and is not presented as operationally equivalent.

XFRA/SPAN, Pacific, Perimeter Compute and Deep Green are not identical to Neurengy. Together they show four independently funded routes around conventional time-to-power: residential and small-site headroom, idle grid connections, spare commercial-building power and compact compute with useful-heat integration [23,24,72-74].

This is evidence of market direction, not proof of Neurengy's economics. Public operating metrics remain limited and company claims are treated as claims unless independently verified.

BOTTOM LINE

The category is converging around smaller, faster, power-aware and workload-specific infrastructure.

FOUR CLOSEST ANALOGUES

XFRA, Pacific, Perimeter and Deep Green: what each model proves

XFRA / SPAN

US - DISTRIBUTED INFERENCE AT THE GRID EDGE

xfra.ai

MODEL

Liquid-cooled nodes paired with SPAN controls at homes and small commercial sites.

PUBLIC EVIDENCE

Public plan: a distributed fleet using underused local electrical capacity.

WHAT IT PROVES

Proves the relevance of real-time power control and fleet orchestration. [72]

PACIFIC

US / GLOBAL AMBITION - MODULAR CAPACITY ON DELAYED SITES

pacific.space

MODEL

Relocatable AI pods for sites that already hold paid-for grid connections.

PUBLIC EVIDENCE

Public claims include rapid deployment, non-evaporative cooling and equipment finance.

WHAT IT PROVES

Proves that idle interconnection time can have monetisable value. [24]

PERIMETER COMPUTE

US - SPARE BUILDING POWER FOR METRO-EDGE INFERENCE

perimetercompute.com

MODEL

Rack kits in existing mechanical space; the operator handles retrofit, hardware and GPU cloud operations.

PUBLIC EVIDENCE

Company materials describe recurring income for owners and a growing multi-site operating team.

WHAT IT PROVES

Closest overlap with Neurengy: spare commercial power plus managed distributed inference. [73,74]

DEEP GREEN

UK - COMPACT HIGH-DENSITY COMPUTE WITH USEFUL HEAT

deepgreen.energy

MODEL

Regional AI/HPC data centers placed near heat offtakers and efficient grids.

PUBLIC EVIDENCE

Public materials show live and development projects; performance figures remain company claims.

WHAT IT PROVES

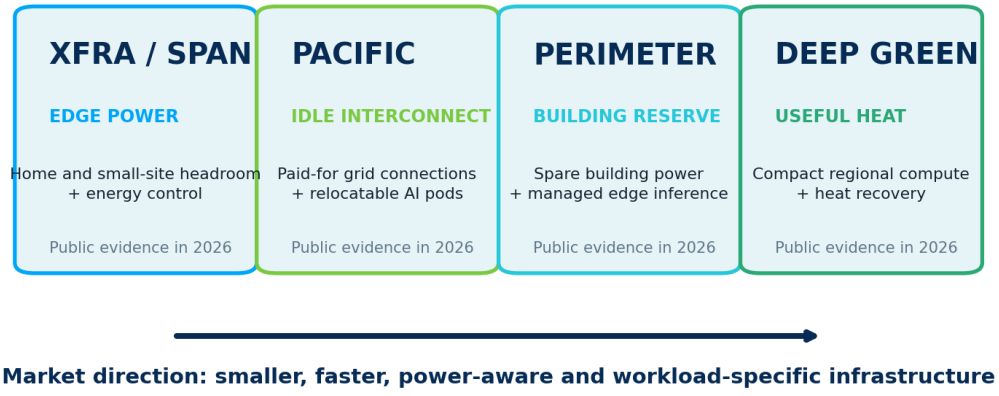
Proves that smaller compute can be embedded in a local energy system. [23,42]

Interpretation: the four companies validate direction of travel. They do not validate Neurengy's economics, site supply or SLA - those still require pilots.

MARKET VALIDATION

Four independent models converge on infrastructure reuse

Four independent operators are converging on the speed-to-power gap



Market direction: smaller, faster, power-aware and workload-specific infrastructure

XFRA uses residential and small-commercial electrical headroom; Pacific targets paid-for connections at delayed projects; Perimeter converts spare building power and mechanical space into managed metro-edge inference; Deep Green pairs compact high-density compute with heat offtake [23,24,72-74].

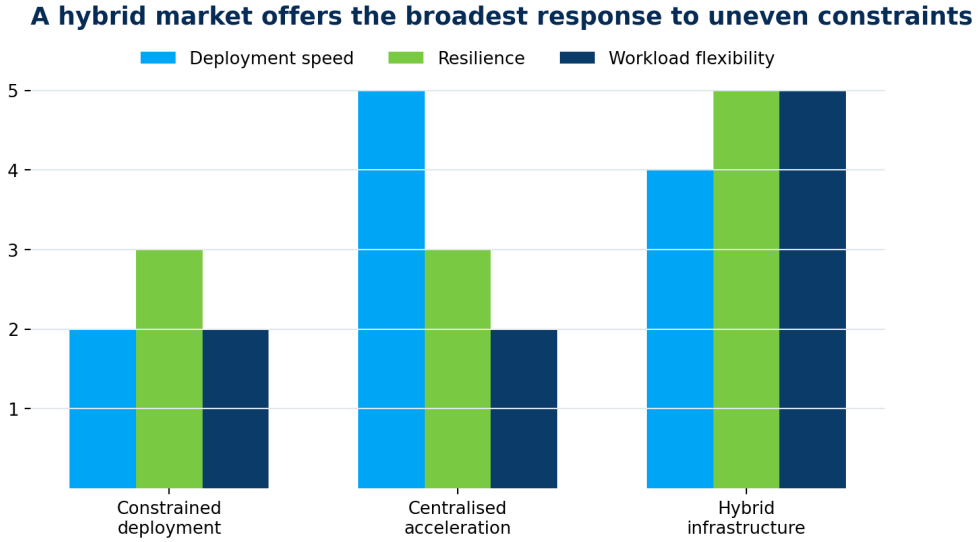
Their common pattern is commercially significant: start from an infrastructure advantage already present, add a standard physical system, and operate a workload class that does not require one hyperscale training campus. Neurengy extends this convergence into a reserve-discovery and orchestration thesis across multiple host types.

BOTTOM LINE

Neurengy's window exists where verified reserve power, workload flexibility and repeatable deployment can be combined faster than greenfield capacity.

2030 SCENARIOS

A hybrid market is the most robust response to uneven constraints



Author scenario scores; not a forecast of market share.

A constrained scenario delays projects and raises prices. Centralised acceleration delivers huge clusters but concentrates grid and capital risk. A hybrid scenario combines campuses with regional, modular, flexible and distributed capacity.

The hybrid case is not a forecast of market share. It is a resilience thesis: different workloads and regions need different infrastructure responses.

BOTTOM LINE

The market is likely to reward interoperability across architectures.

ACTIONS + SIGNALS

What leaders should do now - and what to watch next

Utilities should create transparent large-load and flexible-compute products; site owners should audit contractual and interval power; compute buyers should classify workloads before procurement; investors should separate site, hardware and platform risk.

Over the next 24-36 months, the durability of the opportunity will be determined by interconnection reform, on-site generation, liquid-cooling retrofits, accelerator availability, enterprise inference growth, data-center disclosure and the financing record of modular and distributed assets.

- Build a verified site inventory and measurable deployment funnel.
- Contract demand before hardware and publish clear service tiers.
- Track grid rules, cost allocation and flexible-load products.
- Track liquid cooling, heat reuse and water-disclosure requirements.
- Track enterprise inference relative to tightly coupled training.
- Scale only after pilots falsify or confirm the operating model.

BOTTOM LINE

Act on evidence now, but keep the strategy responsive to whether time-to-power remains structurally constrained.

NEURENGY OPPORTUNITY

Existing power can become a new layer of managed compute

Neurengy’s hypothesis is that verified underused power at existing sites can be converted into managed AI capacity faster than equivalent greenfield development. The model targets infrastructure that is too fragmented for a hyperscale campus but large enough to support useful compute.

The platform is positioned as a complement to conventional data centers. It focuses on enterprise inference, batch, rendering/video and interruptible workloads — not on claiming that small sites can replace a tightly coupled training cluster.

- Existing power. New compute.
- Reserve-first site discovery before energy-first expansion.
- One operator layer across deployment, monitoring, orchestration and billing.
- Locality-aware routing where latency, data and network economics justify it.

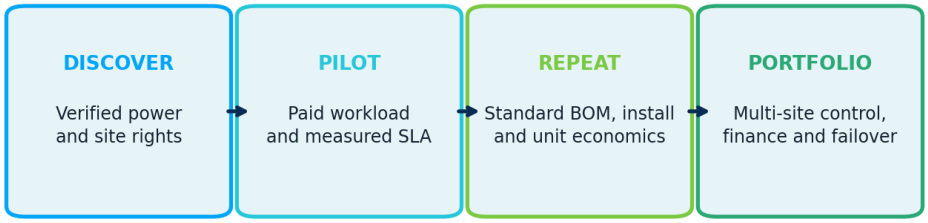
BOTTOM LINE

Neurengy has a right to exist if it can turn site fragmentation into a lower time-to-compute without sacrificing trust.

NEURENGY MODEL

Discover -> Deploy -> Operate -> Orchestrate -> Finance

Scale is earned through evidence gates, not announced capacity targets



Each gate must be supported by contracts, telemetry, engineering records and customer service data.

The operating model begins with evidence, not a capacity target. Sites are qualified through rights, interval telemetry, engineering and connectivity; standard nodes are deployed only against a workload and service design; operating data then determines whether the model can repeat across sites.

This removes public dependence on a specific gigawatt inventory or a 2030 megawatt promise. Scale becomes the result of passing discovery, paid-pilot, repeatability and portfolio-control gates.

- Discover: verified reserve power and host rights.
- Deploy: standard engineering package linked to contracted demand.
- Operate: measured uptime, utilisation, energy and security.
- Orchestrate: route by latency, data, hardware and interruption policy.
- Finance: fund repeatable assets after cash-flow and removal rights are proven.

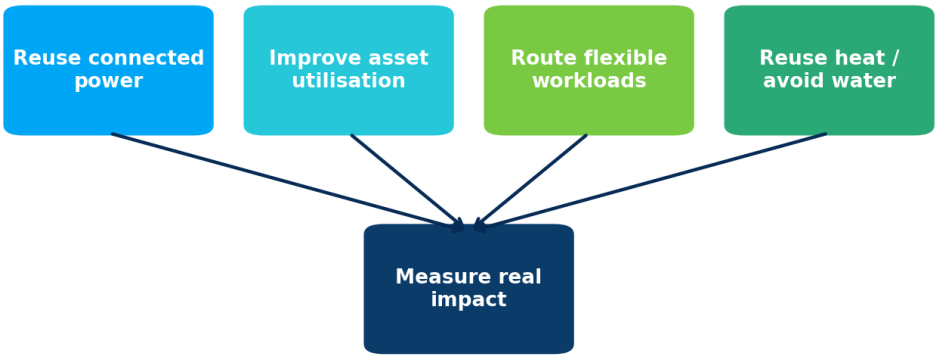
BOTTOM LINE

The defensible asset is a repeatable qualification and operating system, not an announced capacity number.

VALIDATION AND ESG

The opportunity remains conditional until portfolio evidence exists

ESG value exists only when utilisation gains become measured outcomes



Neurengy remains at validation stage. Revenue per active kW, utilisation, margin, payback, deployment speed and service quality must be demonstrated through paid demand and measured operations rather than inherited from global market forecasts.

The ESG thesis requires the same discipline: report incremental electricity, carbon, water, useful heat, hardware utilisation and avoided construction against a documented baseline. Positive outcomes and rebound effects belong in the same ledger.

- Demand: paid customer commitment for a defined workload.
- Supply: audited sites with documents, telemetry and engineering fit.
- Economics: comparable BOMs and measured contribution margin.
- Operations: uptime, failover, latency, security and job completion.
- ESG: additional, attributable and independently reviewable outcomes.

BOTTOM LINE

Conditional GO: validate a repeatable service before presenting scale as an outcome.

Methodology, definitions and limitations

Evidence hierarchy

Official and intergovernmental sources are prioritised, followed by recognised research and market consultancies. Company disclosures are used for business-model evidence and labelled as vendor claims.

Capacity boundaries

Facility power, IT load, contracted power, connected power and reliably available power are treated separately. Announced pipeline is not counted as operational capacity.

Demand-capacity gap

The indexed gap compares different measures: AI-focused electricity demand and installed facility capacity. It indicates direction and growth pressure, not an exact MW shortage.

ESG

Potential benefits are hypotheses until supported by a baseline, life-cycle boundary and measured energy, carbon, water, heat and utilisation data.

Competitive analysis

The report compares publicly visible capabilities. Qualitative scores do not represent market share, valuation or audited performance.

Editorial process

ChatGPT assisted synthesis, drafting and visual production. Pavel Cheremisin retains editorial responsibility for the published version. OpenAI does not endorse the findings.

Sources and references - 1/3

1. IEA, Key Questions on Energy and AI (2026)
   https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary
2. IEA, Energy and AI (2025)
   https://www.iea.org/reports/energy-and-ai
3. JLL, 2026 Global Data Center Outlook
   https://www.jll.com/en-us/insights/market-outlook/data-center-outlook
4. Uptime Institute, Global Data Center Survey 2025
   https://uptimeinstitute.com/resources/research-and-reports/uptime-institute-global-data-center-survey-…
5. LBNL, US Data Center Energy Usage Report: 2025 Update
   https://eta.lbl.gov/publications/united-states-data-center-energy-2025
6. US DOE, Increase in Electricity Demand from Data Centers
   https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-…
7. CBRE, Global Data Center Trends 2025
   https://www.cbre.com/insights/reports/global-data-center-trends-2025
8. System Operator of Russia, Data-center development outlook
   https://www.so-ups.ru/news/energy-news/energy-news-view/news/30412/
9. System Operator of Russia, connected DC and mining capacity
   https://www.so-ups.ru/news/press/press-release-view/news/30353/
10. EU Energy Efficiency Directive data-center context
    https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=PI_COM%3AAres%282026%293247482
11. EU Delegated Regulation 2024/1364
    https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202401364
12. LBNL, Water use of data-center workloads
    https://eta.lbl.gov/publications/water-use-data-center-workloads
13. NVIDIA H100 specifications
    https://www.nvidia.com/en-us/data-center/h100/
14. NVIDIA H200 specifications
    https://www.nvidia.com/en-us/data-center/h200/
15. NVIDIA L40S specifications
    https://www.nvidia.com/en-us/data-center/l40s/
16. Lambda GPU pricing
    https://lambda.ai/pricing
17. Selectel Cloud GPU
    https://selectel.ru/services/cloud/servers/gpu/
18. SaladCloud pricing
    https://salad.com/pricing
19. Vast.ai GPU marketplace
    https://vast.ai/
20. Akash Network
    https://akash.network/
21. io.net comparison and pricing materials
    https://io.net/blog/io-net-vs-akash-network-comparing-gpu-cloud-pricing-and-features
22. Crusoe, energy-first AI infrastructure
    https://www.crusoe.ai/
23. Soluna modular data centers
    https://www.solunacomputing.com/blog/mdc/
24. Deep Green modular heat-reuse data centers
    https://deepgreen.energy/
25. Pacific, AI factories from unused power
    https://www.pacific.space/
26. Kazakhstan Government, Data Center Valley
    https://www.gov.kz/memleket/entities/maidd/press/news/details/1240802?lang=en

Sources and references - 2/3

27. President of Uzbekistan, AI and data-center zone
https://president.uz/en/lists/view/9317

28. Reuters, Stargate UAE
https://www.reuters.com/business/media-telecom/stargate-uae-ai-datacenter-begin-...

29. iKS-Consulting, Russian commercial data-center market
https://survey.iksconsulting.ru/page82045366.html

30. NIST AI Risk Management Framework
https://www.nist.gov/itl/ai-risk-management-framework

31. AWS EC2 pricing
https://aws.amazon.com/ec2/pricing/on-demand/

32. Google Cloud GPU pricing
https://cloud.google.com/products/compute/pricing

33. CoreWeave pricing
https://www.coreweave.com/pricing

34. Runpod pricing
https://www.runpod.io/pricing

35. NVIDIA Certified Systems
https://docs.nvidia.com/certification-programs/latest/nvidia-certified-systems.html

36. IEA, Energy supply for AI
https://www.iea.org/reports/energy-and-ai/energy-supply-for-ai

37. JLL, North America Data Center Report Midyear 2026
https://www.jll.com/en-us/insights/market-dynamics/north-america-data-centers

38. JLL, The new data-center reality webinar
https://www.jll.com/en-us/webinars/2026-global-data-center-outlook

39. Uptime Institute Sustainability Survey 2024/25
https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/Dec2024.Sustainability.3pager.pdf

40. Uptime Institute, Water is local
https://journal.uptimeinstitute.com/water-is-local-generalities-do-not-apply/

41. EU Recommendation 2024/2395 on waste heat
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ%3AL_202402395

42. Deep Green Bradford heat-reuse data-center approval
https://deepgreen.energy/news/deep-green-welcomes-planning-approval-for-pioneering-bradford-heat--reuse-...

43. Government of Russia, Energy Strategy to 2050
https://minenergo.gov.ru/ministry/energy-strategy

44. National AI Strategy of the Russian Federation to 2030
https://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102608394

45. Russian low-carbon development strategy to 2050
https://government.ru/docs/43708/

46. EnergyNet, Energy Internet Architecture IDEA 2.0 (project source)
https://energynet.ru/

47. EnergyNet NTI roadmap (project source)
https://nti2035.ru/markets/energynet

48. Business Excellence, industrial metaverse article (project source)
https://ria-stk.ru/

49. Russian Energy Agency, World Energy Scenarios to 2050
https://rosenergo.gov.ru/press-center/news/stsenarii-razvitiya-mirovoy-energetiki-do-2050-goda/

50. IEA, Material Efficiency in Clean Energy Transitions
https://www.iea.org/reports/material-efficiency-in-clean-energy-transitions

51. Uptime Institute, Mapping PUE trends
https://intelligence.uptimeinstitute.com/resource/mapping-pue-trends-data-center-region-age-and-size

52. LBNL Data Center Energy and Water Modeling
https://datacenters.lbl.gov/modeling-forecasting

Sources and references - 3/3

53. IEA, Data centres and data transmission networks
https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

54. System Operator of Russia, projected power deficit context
https://www.so-ups.ru/news/smi/press-view/news/27984/

55. Russian Government, data-center registry rules
https://publication.pravo.gov.ru/document/0001202511290019

56. Russian Federal Law No. 244-FZ
https://publication.pravo.gov.ru/document/0001202507230046

57. Kazakhstan data-protection standards
https://www.gov.kz/memleket/entities/mti/press/news/details/1147770?lang=en

58. Kazakhstan Bureau of National Statistics, data exchange
https://stat.gov.kz/en/initiatives/data-exchange/

59. European Green Deal
https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

60. US DOE, clean energy resources for data-center demand
https://www.energy.gov/oe/clean-energy-resources-meet-data-center-electricity-demand

61. US DOE, geothermal and data centers
https://www.energy.gov/hgeo/geothermal/geothermal-and-data-centers

62. NVIDIA A100 specifications
https://www.nvidia.com/en-us/data-center/a100/

63. NVIDIA HGX platform
https://www.nvidia.com/en-us/data-center/hgx/

64. Selectel GPU on-prem
https://selectel.ru/services/dedicated/gpu-on-prem/

65. Deep Green Lansing proposal
https://deepgreen.energy/news/deep-green-proposes-120-million-sustainable-data-center-investment-in-...

66. LBNL, advances in data-center efficiency
https://newscenter.lbl.gov/2025/12/16/how-researchers-are-driving-advances-for-data-centers/

67. IEA, AI for energy optimisation and innovation
https://www.iea.org/reports/energy-and-ai/ai-for-energy-optimisation-and-innovation

68. IEA, AI and climate change
https://www.iea.org/reports/energy-and-ai/ai-and-climate-change

69. EU reporting requirements for data centers
https://eur-...

70. iKS Media, Russian colocation price benchmark
https://www.iksmedia.ru/news/6095999-Czeny-na-colocation-stabilizaciya.html

71. Deep Green, rapid AI colocation deployment
https://deepgreen.energy/news/fastest-ai-colocation-deployment

72. XFRA white paper: distributed data center
https://www.xfra.ai/whitepaper

73. Perimeter Compute, building-owner model
https://www.perimetercompute.com/for-building-owners/

74. Perimeter Compute, distributed inference careers and operating model
https://www.perimetercompute.com/careers/

75. OpenAI Sharing & Publication Policy
https://openai.com/policies/sharing-publication-policy/

76. EU Artificial Intelligence Act, Article 50
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:02024R1689-20260727

77. Pavel Cheremisin on LinkedIn
https://www.linkedin.com/in/pavel-cheremisin

Neurengy

Prepared by Pavel Cheremisin, CEO & Co-founder, Neurengy

LinkedIn: www.linkedin.com/in/pavel-cheremisin

neurengy.com • Published August 2026