KPMG

The grid's client-server moment

When AI compute comes home, the distribution grid becomes the story

kpmg.com

Executive summary

For the past three years, utilities have treated load growth driven by artificial intelligence (AI) primarily as a transmission-level challenge. Planning has focused on hyperscale data centers, generation adequacy, interconnection queues, and large-load tariffs. That response was necessary, but it assumes the next wave of AI demand arrives in large, visible increments that utilities can see coming.

This paper examines a different possibility: A meaningful share of future AI demand arrives on the distribution system instead. As models become more efficient, enterprises increasingly deploy inference closer to where work occurs, AI-capable hardware proliferates through normal device refresh cycles, and privacy concerns encourage local processing of sensitive data. The result is a gradual shift from concentrated AI demand toward a more distributed pattern spanning commercial buildings, branch offices, and homes.

That possibility creates a forecast asymmetry. Utilities may be forecasting the right amount of AI demand while preparing for the wrong shape of demand. A gigawatt arriving through a handful of hyperscale campuses creates one planning problem. The same gigawatt arriving through thousands of commercial sites and millions of distributed endpoints creates another entirely. One is visible through interconnection queues and tariff negotiations. The other appears feeder by feeder, transformer by transformer, often only after load patterns have already changed.

The risk is arriving at precisely the wrong moment. The distribution grid is already absorbing electrification, distributed generation, storage, electric vehicles, and growing prosumer participation.[1] Much of the transformer fleet is aging, visibility below the meter remains limited, and utilities are increasingly being asked to manage a system characterized by millions of small decisions rather than a handful of large ones.

The first paper in this series argued that utilities must become energy orchestrators because AI demand, distributed energy, and new ecosystem participants were making coordination more valuable than ownership alone. This paper extends that argument to the distribution edge, where coordination stops being a problem of dozens of large assets and becomes a problem of millions of small ones.

The opportunity is significant. The same technologies creating this new demand can also help manage it. Utilities that strengthen distribution-level forecasting, modernize their data models, expand orchestration capabilities, and prepare for a more distributed AI future can turn a potential planning blind spot into a source of competitive advantage. The utilities best positioned to succeed will be the ones that recognize the shift early enough to shape it.

The pattern computing always follows

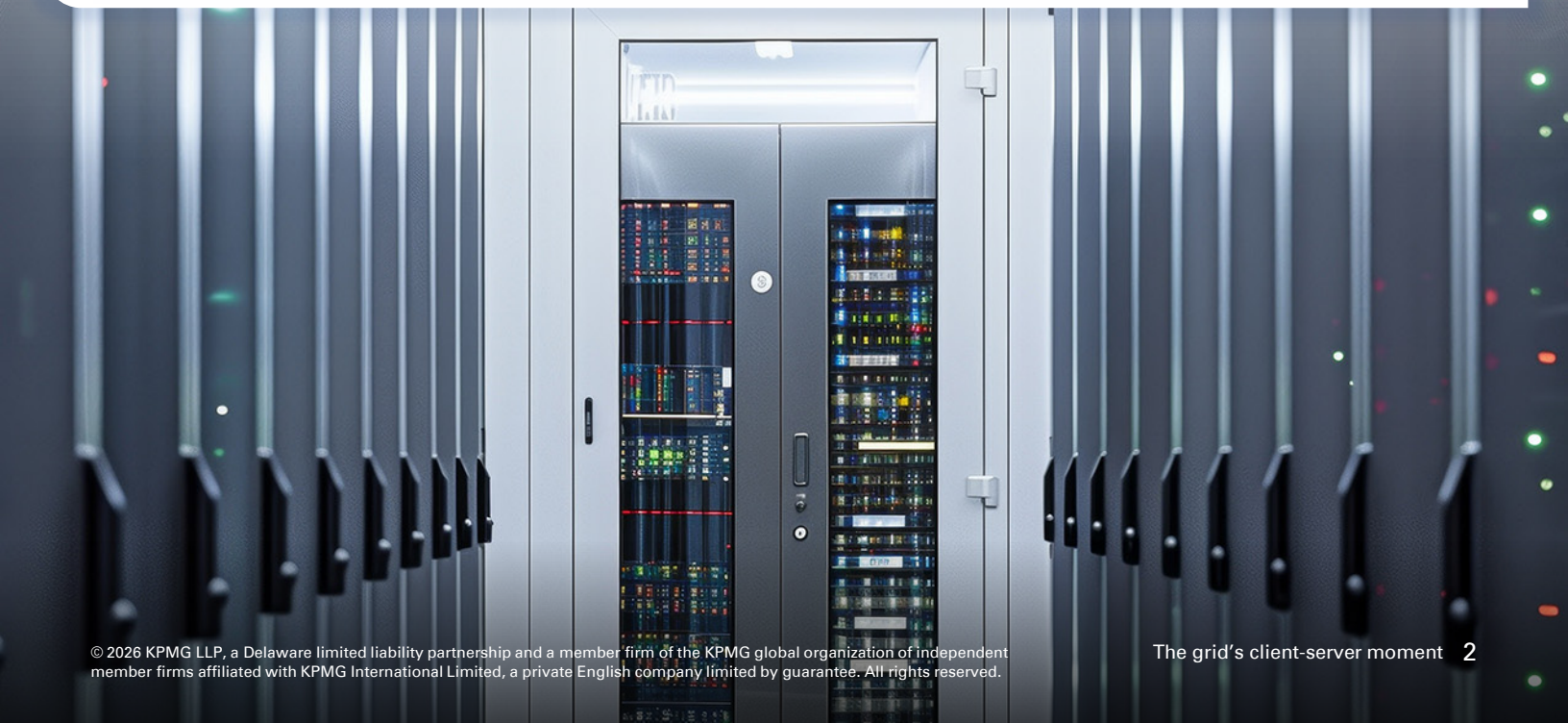
A useful way to view the history of computing is that every generation of computing has followed the same arc.

Capability is born centralized, because the hardware is scarce, expensive, and hard to operate. The mainframe era put computing in glass rooms and rationed it by the minute. Then the economics of the hardware improve faster than the economics of the network that reaches it, and the workload migrates toward the user. Minicomputers put departments in charge of their own compute. Client-server split the work between the desk and the data closet. The personal computer finished the job for an entire class of workloads, and the smartphone repeated the whole cycle in a decade.

Anyone who lived through those transitions in enterprise information technology, and most utility executives did, will remember the honest version of the story: the pendulum swings. The cloud era recentralized enormous amounts of compute, and the first phase of AI, the training phase, recentralized it further, into the largest concentrations of computing hardware ever assembled. Sceptics point at that and say decentralization is a prediction that keeps being wrong. **They are half right, but they miss the important half.**

What recentralizes is the work users never see: training, batch processing, archival storage. What migrates outward, every time, is the work that touches a human in real time. Latency, cost, and privacy pull interactive workloads toward the user with the reliability of gravity. AI inference—the act of actually using a model—is the most interactive, most latency-sensitive, most privacy-loaded workload computing has ever produced.

It will follow the pattern. The only real debate is speed.



Why AI is moving closer to the user

Stage 1 is already behind us

AI inference has begun moving from hyperscale campuses to distributed infrastructure at telecom hubs and utility substations, creating a contest over who coordinates it.[2] **What matters for this paper is what kind of story that was: a transmission-and-substation story, fought over megawatt-scale nodes big enough to negotiate with.** The migration does not stop at the substation fence, and below the fence, the loads stop negotiating.

Stage 2: The enterprise brings inference indoors

The economics that pulled inference out of the hyperscale campus are now pulling it past the substation and into the buildings where the work happens. Inference already accounts for the large majority of AI compute across major vendors,[3] and for a steady, high-utilization workload the total-cost-of-ownership math has flipped: one 2026 analysis of enterprise inference hardware found on-premises deployments breaking even against cloud in under four months, with per-token costs an order of magnitude lower over a five-year life.[4] Surveys tell the same story from the demand side. Two-thirds of data center operators report enterprises repatriating at least some workloads from the cloud, driven by cost predictability, data sovereignty, latency, and the need to keep compute close to critical data and systems.[5]

The physical numbers are what a distribution engineer cares about, so here they are.

8 GPU interference node **draws** **~10kW** under load

100 GPU enterprise pool **draws** **<200kW**

Guidance now circulating in the AI infrastructure community explicitly recommends splitting compute into pools of 100 GPU because it sits below the thresholds that trigger utility-scale infrastructure review.[6]

Read that from the utility's side of the meter. The load is not merely migrating into commercial buildings; it is being deliberately shaped to stay invisible to the interconnection process.

Nobody files paperwork for 176 kilowatts. It just shows up on the feeder.

The device in front of the user

The final stage rhymes loudest with the personal computer (PC) revolution, and it is no longer speculative. AI-capable PCs, machines with dedicated neural processing silicon, are projected to account for more than half of global PC shipments in 2026, roughly 59 percent by some estimates, up from about 39 percent in 2025.[7] Major software vendors are already building for the device rather than the cloud: Microsoft's Copilot+ PC platform, Apple's on-device Apple Intelligence, and Google's on-device Gemini Nano all run core AI features locally today, with small language models increasingly the default deployment pattern rather than the exception.

IDC projects AI PCs will go from

5% **of the installed base in 2023** ▶ **94%** **by 2028.[8]**

The capability is arriving on every desk and in every home on the ordinary refresh cycle.

What matters is not whether AI demand grows. Utilities are already planning for that. The question is whether the demand arrives in the form planners expect. The industry's response has been built around visible, centralized loads that announce themselves through interconnection requests, tariffs, and generation planning processes. But the forces described above point toward a second possibility: a meaningful share of AI demand arrives diffusely, through commercial buildings, branch offices, and homes. Utilities may be forecasting the right amount of demand while preparing for the wrong shape of demand. That is the forecast asymmetry at the center of this paper. For utilities, the significance is not the PC refresh cycle itself. It is that a growing share of AI capability may arrive through ordinary customer adoption rather than through infrastructure projects that appear in planning processes.

The efficiency turn

Something changed in the model market over the past year, and it accelerates every trend described above. The frontier stopped being the constraint. The most capable models now shipping—Anthropic's Mythos-class systems, OpenAI's latest releases, and their peers—exceed what most users and most companies can actually put to work. For many users and businesses, the latest generation of models already exceeds what is required to perform their job. Raw capability stops being where the competition happens. It moves to cost, and in AI, cost is measured in watts.

So the model makers have turned their attention to efficiency: distillation, quantization, pruning, small language models tuned for specific jobs, mixture-of-experts designs that wake only the parameters a task needs. A model that required a rack in 2024 fits comfortably on a workstation in 2026 and will fit on a laptop before your current rate case is decided. The hardware curve compounds the software one. NVIDIA's next-generation platform claims ten times lower cost per token than the generation it replaces, and each such turn of the crank moves another class of workload one step closer to the user.[9] The industry spent two years asking how big these models could get. Utilities should be asking how small they can get while still doing the job. That question matters far more because it helps determine where future AI load lands.

Power follows privacy

Efficiency makes local AI possible. Privacy is what will make it preferred. Consider the architecture of today's leading assistants and agentic tools honestly: they invite users to connect the most sensitive data they own, bank accounts, medical records, mail, payroll, customer files, and then ship that data to someone else's data center to be processed. Every consumer convenience and every enterprise pilot built this way widens the same exposure. General counsels have noticed. So have regulators, and so, increasingly, have ordinary customers.

There is a better architecture, and the efficient models described above make it practical: keep the model with the data instead of sending the data to the model. In a local-first design, personal and regulated information never leaves the device or the building. The local model does the reading, the screening, the summarizing, and the drafting. When a task genuinely needs frontier-scale reasoning or outside information, the flow inverts: The reference data comes down from the cloud rather than the private data going up, and anything that must travel goes anonymized or encrypted. The mothership gets called when it is needed, not by default. For utilities, that architectural choice matters because it changes where electricity is consumed, not just where data is processed.

Regulated industries like healthcare, financial services, legal, and government are likely to move first, because for them local-first is not a preference but the shortest path through compliance. But the pattern generalizes, and its energy consequence is the point of this paper: Every inference resolved locally is demand moved off a transmission-served campus and onto a distribution-served outlet. Privacy policy, it turns out, is load policy. Nobody in the load-forecasting business is treating it that way yet.

Why hyperscale still matters

Today's device-level NPUs (Neural Processing Unit) handle light, one-shot work: summarize this email, transcribe this call, clean up this photo. The agentic and reasoning workloads that matter most still lean on data-center silicon, and the overwhelming majority of AI compute in 2026 still runs in large facilities.[10] All true. The centralized buildout is real and will continue, but the possibility of a more distributed load future remains real and grows with each generation of hardware and models. The forecasting risk is not overestimating demand. It is assuming hyperscale demand is the only form future AI growth can take. A forecast that carries only the first half is not conservative. It is exposed.

The agent moves in

The next question is not simply where AI runs, but how often it runs.

Increasingly, AI is being embedded into software that works in the background rather than waiting for a prompt. Whether those systems ultimately take the form of personal agents, enterprise agents, or something in between matters less than the operational shift they create: AI moves from an occasional activity to a persistent one. None of this requires a breakthrough. It requires the efficiency gains and local-processing architectures described earlier in this paper, both of which are underway.

From a grid perspective, the important word in that paragraph is overnight. Chat is bursty: a question, an answer, and the silicon goes back to sleep. By contrast, AI systems embedded in day-to-day workflows operate continuously, monitoring, reading, comparing, drafting, coordinating for hours at a stretch. The duty cycle of AI at the edge moves from minutes a day toward continuous operation. The forces described earlier in this paper suggest a meaningful share of that work will run locally, because it is precisely the work that touches the most sensitive data a person or business has. Workloads tied to medical, financial, and other sensitive information are exactly the use cases local-first architectures were designed to support.

The numbers below are illustrative rather than predictive, but they highlight the planning challenge.

There are about

130 million

households in the US.

Suppose that by the early 2030s a quarter of them run capable local AI systems

~300 watts around the clock.

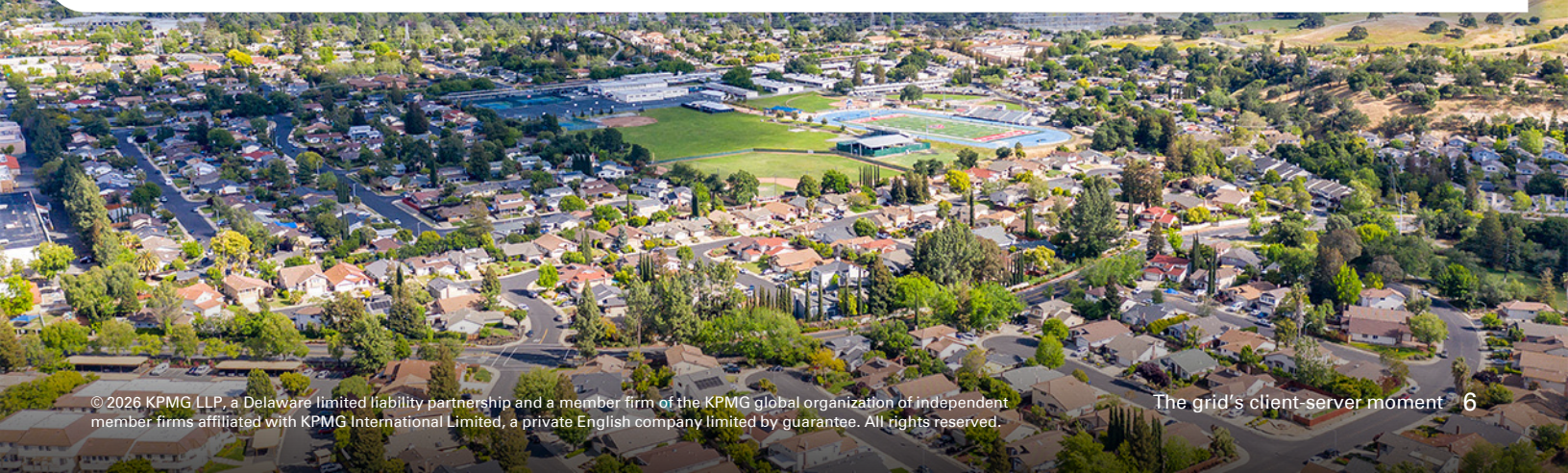
That would create

~10 gigawatts

of new, continuous, distribution-served load, the equivalent of multiple hyperscale campuses spread across every feeder in America, arriving with no interconnection request, no tariff, and no line item in anyone's forecast.

Even materially lower adoption rates still produce gigawatt-scale outcomes, and the commercial side stacks on top.

This is what residential and commercial demand growth looks like when it stops announcing itself.



The forecast asymmetry nobody is pricing

The load-growth conversation of the last three years had one shape: enormous, centralized, and visible. A hyperscale campus announces itself years in advance.[11] It files an interconnection request, signs a large-load tariff, negotiates a substation, and shows up in the rate case. Utilities have organized their response, generation procurement, transmission studies, minimum-demand contracts, around load that asks permission. At least 36 utilities now have large-load tariffs governing exactly this customer.[12] That work was necessary. It is also, quietly, a bet that AI demand keeps arriving in that shape.

There are two ways the bet loses. The first is that distributed inference nodes siphon a meaningful share of forecast hyperscale growth to sites that never enter the queue, leaving generation and transmission investments looking for the load they were built to serve. Georgia Power's agreement to financially backstop nearly 10,000 megawatts of new generation if contracted data center load fails to materialize shows that regulators are already pricing that risk.[13]

The second way is subtler, and it is this paper's core concern: The demand arrives at full size—or larger—but in a form the planning system was not built to see. Not one 500-megawatt customer at a transmission substation, but five thousand 100-kilowatt customers scattered across commercial feeders, and behind them millions of homes, adding small increments of continuous demand. The amount of load is not the surprise. The shape of the load is.

Utilities have seen a preview of this movie, and it is worth being honest about how it went. Electric vehicles (EVs) were also a diffuse, customer-driven load that arrived without interconnection applications, and the industry's consistent finding has been that EV load clusters, by neighborhood, by income band, by feeder, so that system-average adoption numbers concealed local transformer overloads for years before planners caught up.[14]

End-use compute may cluster the same way, because it is likely to concentrate first in the commercial corridors and affluent residential areas where rooftop solar, batteries, and EVs already crowd the same secondary. Unlike an EV, an agent's workload cannot always be shifted to 2 a.m.; much of it runs when people work. And unlike an EV, it never leaves the driveway.

> **A 500-megawatt data center is a negotiation. A million-kilowatt scale loads are a weather pattern. You do not negotiate with weather. You build for it.**
>
> **Todd Durocher**
> KPMG LLP

The frailest layer gets the hardest job

Now put that demand pattern on the system that has to carry it. The US distribution grid enters this decade in worse shape than any other layer of the power system, and the numbers deserve to be stated plainly. More than half of the roughly 40 million distribution transformers in the US are already beyond their expected service life.[15] Wood Mackenzie models a persistent 10 percent supply shortfall for distribution transformers and 30 percent for power transformers, with the pad-mount shortage expected to worsen through 2026 on surging demand from data centers, manufacturing, and EV charging.[16] Distribution transformer prices have risen 78 percent to 95 percent, and lead times that ran a few months in 2020 now stretch to multiple years.[17] The National Renewable Energy Laboratory estimates the country needs to expand distribution transformer supply 160 percent to 260 percent over 2021 levels by 2050 just to meet the electrification demand already in view, before anyone adds a distributed-AI scenario to the model.[18]

Against that fleet, consider what the utility can actually see. Transmission-connected load is metered, telemetered, studied, and contracted. Distribution-connected load below commercial interconnection thresholds is, for most utilities, statistically inferred.

The secondary transformer serving a strip of offices was sized for lighting, HVAC, and desktop computers whose power draw fell for two decades as laptops replaced towers. A tenant who racks 150 kilowatts of inference gear in a back room changes the duty cycle of that transformer permanently, invisibly, and with no obligation to tell anyone. Multiply by every professional services firm, clinic, design shop, and logistics office that decides its data should not leave the building, and the distribution system inherits, one anonymous increment at a time, the same load-growth story the transmission system at least got to negotiate.

The residential story used to be easy to wave off, and honest analysts did: an AI laptop adds watts, not kilowatts. The prospect of continuous local AI retires that comfort. A household running an always-on local agent, an enthusiast rig in the den, an EV in the garage, a heat pump, and an induction range is not the household the residential transformer was sized for. That transformer was the one piece of the grid designed on the assumption that you and your neighbors would never all draw hard at once. Electrification has been eroding that diversity assumption for a decade. Continuous local compute finishes the job, because it is the first major residential load with no natural off-peak.

> **The transmission system is like an aircraft carrier managing planned arrivals. The distribution system is like a beehive managing the behavior of a swarm.**
>
> **Brad Stansberry**
> KPMG LLP

When distributed compute meets distributed generation

If the story ended with a frail grid meeting a diffuse load, then this would be a paper about risk mitigation. It does not end there, because the same decade that is decentralizing compute has been decentralizing generation, and the two trends are not merely simultaneous. They feed each other.

Start with what already exists at the grid edge. More than **5 million homes** in the US have rooftop solar, and residential adoption continues to expand despite the early termination of major federal solar tax incentives.[19] The Department of Energy counts **30–60 gigawatts** of virtual power plant capacity already operating, aggregations of rooftop solar, batteries, EV chargers, thermostats, and flexible loads, and its Liftoff analysis concludes that tripling that capacity to **80–160 gigawatts** by 2030 could serve **10 percent to 20 percent** of national peak demand while avoiding more than **$10 billion** a year in conventional grid costs.[20] During the June 2025 heat emergency, one residential aggregator alone dispatched over **340 megawatts** from home batteries across five states—real capacity, delivered from living rooms, during the same event that had grid operators invoking federal emergency powers on the transmission side.[21]

Residential enrollment is the fastest-growing virtual power plant (VPP) segment, and utilities are entering 2026 making foundational investments in the distributed energy resources management system (DERMS) platforms needed to run it all.[22] In other words, utilities are already investing in many of the systems needed to manage diffuse compute, even if that is not yet how the problem is framed.

The prosumer gets a second product

The prosumer model to date has had one product, electrons, and one chronic problem: The electrons are worth the most somewhere other than where they are made. Net metering fights, export caps, and duck curves are all symptoms of the same friction. Distributed generation produces locally and must sell globally.

Local compute dissolves that friction, because compute is the first load valuable enough, flexible enough, and small enough to consume distributed generation exactly where it is produced. A home or business with solar on the roof, a battery in the garage, and an agent working behind the meter is no longer just exporting cheap midday electrons into a saturated feeder. It converts them, on site, into the highest-value commodity of the decade. Tokens, unlike electrons, do not care about feeder congestion, hosting-capacity limits, or export tariffs. They travel over fiber.

That changes the economics of every distributed resource on the system. Midday solar, the grid struggles to absorb becoming the cheapest inference fuel in the economy. Batteries earn twice, once shifting energy, once firming compute. Flexible inference can shed 30 percent to 40 percent within a minute of a grid signal, turning compute into a demand-response resource that lives on the distribution system, where relief is most valuable rather than at the end of a transmission line.[23] And the prosumer, whom the utility has spent 15 years learning to treat as a billing complication, becomes the natural building block of a grid-edge economy in which generation, storage, flexibility, and compute are aggregated and orchestrated together. The VPP was the rehearsal.

The limitations are real. Policy is currently pushing the other way: the 2025 reconciliation law cut residential solar and efficiency credits, and independent assessments of today's most mature VPPs score them roughly two on a four-point scale of delivering true plantlike performance.[24] Aggregated prosumer compute-and-power is a thesis about the direction of travel, not a description of next quarter. But betting against both the decentralization of compute and the decentralization of energy, when each improves the economics of the other, is a bet no distribution utility should make with a straight face.

AI on both sides of the meter

There is a symmetry running through this paper that deserves to be named directly:

The technology creating this load is also the only tool fast enough to manage it.

A distribution system carrying millions of small, invisible, always-on loads alongside millions of small generators cannot be planned on annual studies and operated on seasonal assumptions. It has to be observed and adjusted continuously, and no utility can hire its way to that. It has to model its way there.

The applications are concrete, and most of them run on data the utility already collects. Machine learning on interval data can detect that a load's duty cycle has structurally changed at the feeder and transformer level. It's the earliest possible warning that compute has moved in downstream.

Predictive asset-health models matter more with every year the transformer fleet ages past its design life, because in a market with multiyear replacement lead times, knowing which units will fail two summers from now is the difference between a planned swap and an outage. Forecasting improves from substation-level curves fitted to history toward feeder-level predictions that account for weather, solar back feed, EV clusters, and now compute. And on the operations side, the DERMS platforms utilities are standing up for the VPP era are, at bottom, orchestration engines whose dispatch decisions will increasingly be made by models, because it's highly unlikely that a control room staffed by humans can optimize 100,000 devices in real time.

Inside the enterprise, the same shift is arriving through the digital core. We've previously laid out how the S/4HANA foundation, with Joule as its agentic layer, gives utilities integrated process intelligence rather than siloed transactions, and the pattern is not confined to one vendor: Oracle, Microsoft, and Workday estates are growing agentic layers of their own, which means utilities running those platforms will face the same practical question: whether those agents are allowed to do real work.[25] The natural extension is a set of operational agents living on that core: an agent that watches the large-load pipeline and flags when a forecast assumption has quietly broken, an agent that assembles the regulatory data request in hours instead of weeks, an agent that walks a customer through a flexibility enrollment in plain language. Utilities that treat AI purely as a demand problem will find they adopted it internally anyway, five years late and on someone else's terms.

Follow the logic one step further and you arrive at the end state implied by these trends. The customer's orchestrating agent will manage the home's energy the way it manages the calendar: precooling ahead of a price spike, shifting the vehicle charge, deciding whether the afternoon's solar goes into the battery, the house, or the inference job, and offering the household's flexibility to whoever pays best.

On the other side of the meter, the utility's orchestration platform will negotiate with millions of those agents at machine speed. Agent to agent, at the meter, all day. That negotiation is the coordination layer emerging between utilities, distributed resources, and customer-side intelligence. **The energy orchestrator remains the entity that governs it. The open question is who earns the right to occupy that role.**

What it breaks inside the transformation program

Every paper in this series has insisted that thesis without operating consequence is just commentary, so here is where this one lands inside a utility transformation program, in the systems and data structures this readership runs:

The customer data model

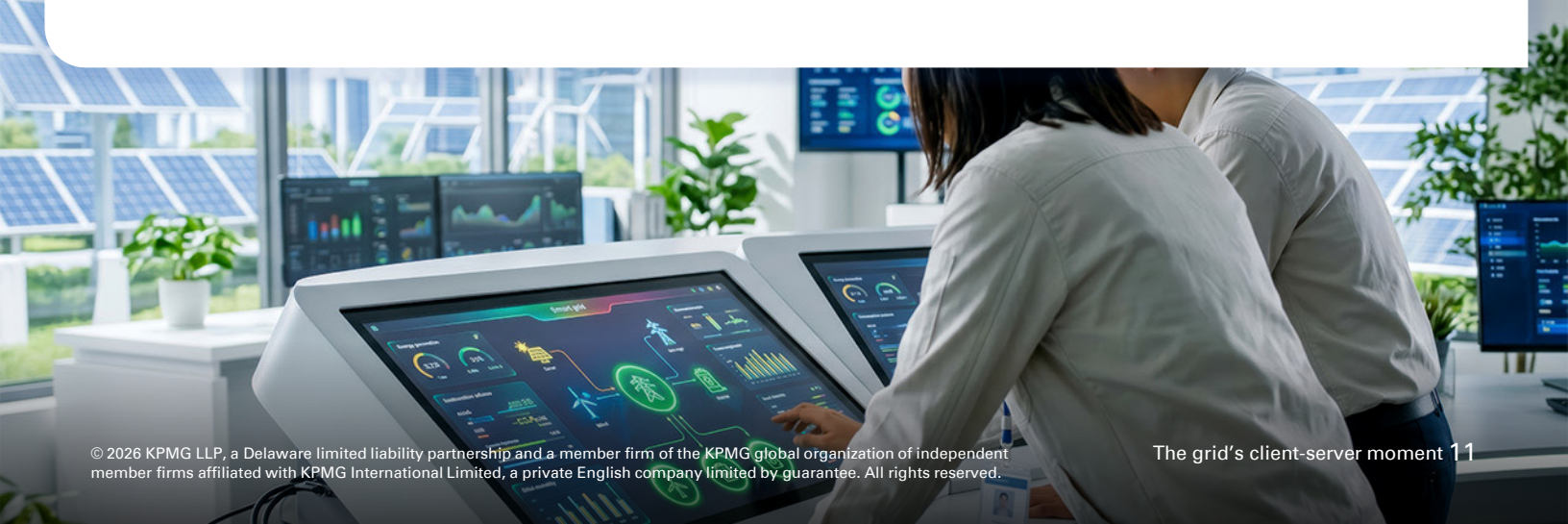
Every architecture decision in the digital core assumes a customer is a premise, a premise has a meter, and a meter has a load class that changes slowly. Diffuse compute breaks all three assumptions quietly. The commercial customer whose back office became a 150-kilowatt inference site is still coded as a small commercial account. The prosumer running generation, storage, and monetizable load behind one meter fits no load class that exists today. Master data models, whether the customer platform is SAP IS-U, Oracle's utility applications, or anything else, need a concept of dynamic load character, and they need it before the pattern is widespread, because retrofitting master data after the fact is the most expensive sentence in this industry.

Metering and analytics

Fifteen-minute interval data from smart meters, which now reach roughly three-quarters of US households, is the only instrument the utility owns that can see this load arriving.[26] Most utilities use that data to bill. The programs that win the next decade will use it to detect: load-signature analytics that flag when a feeder's duty cycle has structurally changed, transformer-level loading models refreshed from actuals rather than design assumptions, and hosting-capacity analysis that treats load as dynamically as it has learned to treat generation.

DERMS and the orchestration layer

The DERMS investments that utilities are making right now are scoped around generation-side resources: solar, storage, EVs. Scope them one category wider. Flexible compute is a distribution-connected, telemetry-capable, contract-ready demand resource, and the orchestration platform being built for the VPP era should treat a curtailable inference load the way it treats a battery: as capacity with an address on the feeder. And build the platform on the assumption that its counterparties will increasingly be software. Customer agents will interact through APIs, not call centers, and a customer platform without machine-consumable interfaces will be, to the fastest-growing class of customer, simply closed.



Rate design and regulatory strategy

Volumetric residential rates and demand charges designed for predictable commercial customers cannot see, price, or shape a load that arrives five kilowatts at a time with no application. The rate-design agenda that follows from this paper includes time-and-location-varying price signals at the distribution level, service tiers for high-duty-cycle small commercial load, and prosumer tariffs that recognize compute-plus-generation as a package rather than taxing each half separately. Price signals matter more, not less, once agents are on the other end, because software actually reads them. Utilities that codevelop these structures with their regulators early will write the template. The rest will inherit one.

Capital planning

Distribution capital planning has to absorb a scenario discipline it has never needed: not just how much load, but at what granularity. A gigawatt arriving as two hyperscale campuses and a gigawatt arriving as 200,000 distributed increments are entirely different capital programs, different transformer orders placed years apart in a market with multiyear lead times, different crew plans, different rate cases. Programs should be running at least three load-shape futures—centralized, hybrid, and diffuse—through their distribution plans this cycle because forecasts should be stress-tested in both directions. Right now most models only run in one direction.

Two load futures, two different utilities

	Centralized future	Diffuse future
Where load lands	Transmission and substations	Feeders, secondaries, behind the meter
Load shape	Large blocks, contracted and scheduled	Always-on agentic baseload plus interactive peaks
Visibility	Years of notice via interconnection queue	None; inferred from interval data after arrival
Commercial instrument	Large-load tariffs, minimum-demand contracts	Rate design, prosumer tariffs, flexibility products
Binding constraint	Generation and transmission interconnection	Distribution transformers, hosting capacity, visibility
Natural counterpart	Utility-scale generation buildout	DERMS, VPPs, prosumer generation and storage

Most utilities will live in a blend of both columns for a decade or more. The table is not asking anyone to pick a column. It is pointing out that nearly every dollar and every hour of the last three years went into the left one.

Five no-regrets moves

The uncertainty in this paper is real, and it argues for optionality, not paralysis. Utilities do not need to know exactly how much AI load will go diffuse to prepare for the possibility that some of it will. These five moves pay off across centralized, hybrid, and diffuse load futures, which is what makes them the right place to start:

1

Run the diffuse scenario now

Add a diffuse-load scenario, commercial and residential end-use compute layered on electrification, to the distribution plan alongside the centralized forecast, and let the delta between the two set the priority list for transformer procurement and feeder upgrades. In a market with multiyear transformer lead times, the scenario work is effectively a procurement decision.

2

Instrument for detection, not just billing

Stand up load-signature analytics on interval data to detect structural duty-cycle change at the feeder and transformer level, and pair it with predictive asset health on the aging transformer fleet. This is the cheapest early-warning system the utility will ever build, and it runs on data already being collected.

3

Fix the data model while the program is open

Extend the master data and integration architecture, before pressure arrives, to represent dynamic load character and combined prosumer assets, and expose customer-facing capabilities through interfaces an agent can consume. Utilities mid transformation have a once-a-decade window to build this in rather than bolt it on.

4

Widen the orchestration scope by one category

Scope the DERMS and VPP platform to treat flexible compute as a dispatchable demand-side resource from Day 1, and pilot a distribution-level flexibility product with a willing commercial compute customer in the territory.

5

Get to the commission first

Open the rate-design conversation with regulators now, framed the way this series has always framed it: the utility as the orchestrator of a grid-edge economy, seeking the pricing tools to shape a load future everyone can already see forming. The first movers will set precedent. In a regulated industry, precedent is strategy.

The cost of watching the wrong meter

One risk is letting someone else design the coordination layer above you. The quieter risk is simply watching the wrong meter. Nobody will announce the day the AI load goes diffuse. There will be no filing to object to, no queue position to negotiate, no headline groundbreaking. There will just be feeders that age faster than the model said, transformers that fail summers earlier than planned, in a market where the replacement is years away, and a slowly widening gap between the load the forecast described and the load the system actually serves. The utilities that come through this well will be the ones that noticed early, because they built the instruments to notice, and that had already positioned themselves as the orchestrator of the grid edge rather than its last uninformed party.

Computing decentralized twice before in living memory, and both times the electric industry experienced it as something that happened to its load curve rather than something it participated in. The third time arrives with a difference worth seizing. This time the decentralizing load and the decentralizing energy system increasingly reinforce one another, and the utility is the only actor positioned to see, coordinate, and shape both.

That is the energy orchestrator thesis arriving at the distribution edge:

not only coordinating dozens of visible assets, but also understanding and orchestrating millions of small decisions before they become system consequences.

How KPMG can help

The thesis of this paper cuts across planning, technology, commercial strategy, and regulation, and so does the way we help. **Five capabilities matter most, and AI runs through every one of them, because the technology driving this load story is the same technology we use to manage it.**

See the diffuse future before it arrives

The KPMG Power & Utilities practice helps utilities build the multiscenario distribution load models—centralized, hybrid, and diffuse—and quantify the capital, rate, and reliability exposure between them. Our AI and data teams then turn the forecast into a living instrument: load-signature analytics on interval data that detect duty-cycle change at the feeder and transformer level, and asset-risk models that identify which units on an aging fleet will fail before the replacement can arrive. Every model is developed under the KPMG Trusted AI framework, so the analytics that inform a rate case can withstand a commission's scrutiny.

Build a digital core that can see the grid edge

KPMG Powered Enterprise for power and gas utilities is how we help clients operationalize the energy orchestrator model in practice, across the technology platforms utilities actually run: SAP S/4HANA, Oracle, Microsoft, and Workday. We bring reference target operating models built for the sector and serve as both business integrator and system integrator, so the transformation and the technology land together rather than in sequence. On this paper's agenda, that means modernizing the customer and asset master data that will break first, unifying meter, grid, and customer data so AI can actually run on it, and integrating the digital core with the DERMS and orchestration platforms the grid-edge economy requires.

Stand up the commercial machinery

The commercial machinery needed to support a grid-edge economy—prosumer tariffs that treat compute-plus-generation as a package, distribution-level flexibility products, curtailment verification and settlement at the feeder level, and billing for load classes that change character overnight—does not sit on any utility's shelf today. Our transformation teams design those structures and stand up the operating model underneath them.

Through KPMG Powered Enterprise, we bring preconfigured, leading-practice target operating models on the client's platform of choice, an approach through which one utility realized annual benefits of roughly 1 percent of revenue.[27] For this agenda specifically, that means the billing, contract management, and settlement operations that let a utility launch a grid-edge flexibility product as a real commercial offering rather than a pilot that never scales.

Put trusted agents to work inside the utility

KPMG already operates the type of agentic operating model described in this paper. KPMG Workbench, our global multiagent AI platform, orchestrates networks of AI agents with agent-to-agent interoperability and built-in data sovereignty, drawing on our alliances with Microsoft, SAP, Oracle, Workday, and others; every agent on it carries an accreditation against our 10-pillar Trusted AI framework, and KPMG was the first organization in the world to achieve ISO 42001 certification for AI management systems. We use that platform to deliver our own work, and we help utilities apply the same pattern to theirs: operational agents on the digital core, whether that core is SAP S/4HANA with Joule, Oracle, Microsoft, or Workday, that assemble regulatory data requests, monitor the large-load pipeline for broken forecast assumptions, walk customers through flexibility enrollment in plain language, and lay the foundation for the agent-to-agent orchestration the grid edge will demand.

Design the commercial and regulatory architecture

From prosumer tariffs and distribution-level flexibility products to large-load strategy and commission engagement, our regulatory and customer transformation teams help utilities get to the commission first, with rate designs that software can read and act on, and customer platforms, built on the KPMG Connected Enterprise approach for utilities, that expose the machine-consumable interfaces the fastest-growing class of customer will expect.

Contact us

Todd Durocher
Principal, Power & Utilities Transformation Leader
KPMG US
tdurocher@kpmg.com

Todd Durocher is a leader in the Power & Utilities Advisory practice at KPMG with more than 25 years of experience in strategy, business transformation, and technology-enabled change. He advises utilities on large-scale transformation initiatives, including AI operating model modernization, cost optimization, grid modernization, acquisitions and integrations, and emerging technologies. Todd also leads KPMG's Energy SAP delivery in the US helping clients from strategizing and developing their program roadmaps, governance, business cases, and target state operating models through the implementation to post-go-live support.

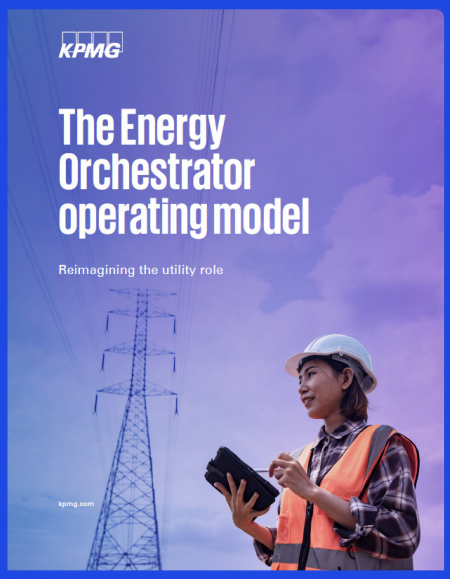
Brad Stansberry
Partner, Energy & Chemicals Advisory Leader
KPMG US
bstansberry@kpmg.com

Brad Stansberry brings more than 25 years of experience delivering results on complex projects for finance function leaders in the energy and utility industry. He also leads the Energy & Chemicals Advisory business at KPMG. His focus is on helping chief financial officers and finance executives in the energy and chemical industries to run the "business of finance" better. This includes a focus on defining finance organization strategy and objectives, improving finance processes, deploying enabling technologies, enhancing finance talent and skills, and being a more valued service provider to their business constituents.

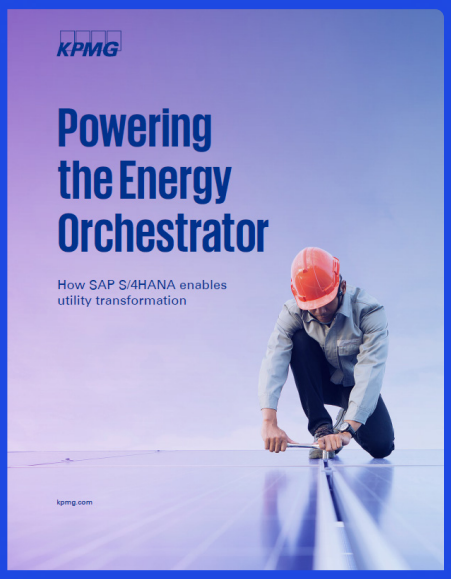
The authors would like to thank Leah Lockwood and Michael Gelfand for their contributions to this paper.

Related thought leadership

[The Energy Orchestrator: Redefining the Future of Utilities]



[The Energy Orchestrator: A New Operating Model]



[Powering the Energy Orchestrator: How SAP S/4HANA Enables Utility Transformation]

Energy Orchestrator Part IV: The new economics of power and compute

References

[1] The Pew Charitable Trusts, "Distributed Energy Can Unleash the Resilient, Affordable Grid of the Future," April 28, 2026; North American Electric Reliability Corporation, "Long-Term Reliability Assessment," January 2026.

[2] "NVIDIA, Telecom Leaders Build AI Grids to Optimize Inference on Distributed Networks," NVIDIA Blog, March 17, 2026; "EPRI, NVIDIA, Prologis, and InfraPartners Announce Collaboration to Produce Smaller Data Centers for the AI Era," EPRI press release, February 3, 2026; "The Energy Orchestrator Series Part 4: How the grid becomes the coordination layer of the AI economy," KPMG LLP, 2026.

[3] NVIDIA Corporation, Form ARS (Annual Report to Shareholders), Fiscal Year 2025, filed with the SEC.

[4] "On-Premise vs Cloud: Generative AI Total Cost of Ownership (2026 Edition)," Lenovo Press whitepaper, 2026; Nick Patience, "AI Inference: Enterprise Infrastructure and Strategic Imperatives," Futurum Group, January 27, 2026.

[5] Kirk Killian, "Preparing Enterprise Data Centers for AI Adoption," Data Center Knowledge, March 13 2026.

[6] Mitrasish, "Power-Bound, Not GPU-Bound: AI Data Center Power Constraints Are the Real 2026 Bottleneck," Spheron, June 24, 2026.

[7] "AI Advanced PCs to Surpass Half of Global Shipments in 2026 as NPU-powered Laptops Go Mainstream," Counterpoint Research, September 8, 2025.

[8] Lucas Merian, "AI PCs to surge, claiming over half the market by 2026," Computerworld, August 25, 2025.

[9] Brian Caulfield, "NVIDIA Rubin Platform, Open Models, Autonomous Driving: NVIDIA Presents Blueprint for the Future at CES," NVIDIA Company Blog, January 5, 2026.

[10] "2026 AI Outlook Report," TechInsights, 2026; Counterpoint Research and Gartner data per references 5 and 6.

[11] "Why hyperscalers are filling new data center campuses faster than expected," Datacenters.com Wholesale Colocation, February 1, 2026.

[12] Mark Jaffe, "Utilities adopt large load tariffs to cope with the costs and power demands of data centers," EUCI, December 16, 2025; "Data Center Fact Sheet," Georgia Public Service Commission, March 2026; Alander Rocha, "Georgia regulators approve massive power grid expansion to serve data centers," Georgia Recorder, December 19, 2025.

[13] Ibid.

[14] "Electric Vehicles at Scale – Phase II: Distribution System Analysis," Pacific Northwest National Laboratory, March 2022; "Influx of Electric Vehicles Accelerates Need for Grid Planning," Pacific Northwest National Laboratory news release, July 29, 2020; Jeff St. John, "How to Make EVs Get Along With the Grid," Canary Media, December 13, 2024.

[15] Sonal C. Patel, "Transformers in 2026: Shortage, Scramble, or Self-Inflicted Crisis?," POWER Magazine, January 2, 2026; Devin Thomas, Benjamin Boucher, and Michael Mendrek-Laske, "Transformer troubles: manufacturing and policy constraints hit US transformer supply," Wood Mackenzie, August 13, 2025.

[16] Ibid.

[17] "Transformer Shortages: Supply Chain Impact on Pricing," Electrical Trader, February 2, 2026.

[18] Killian McKenna, Sherin Ann Abraham, and Wenbo Wang, "Major Drivers of Long-Term Distribution Transformer Demand," National Renewable Energy Laboratory, February 2024.

[19] Solar Energy Industries Association,"6 Million Solar Installations: Powering American Communities," June 2026; Enphase Energy, "The Federal Solar Tax Credit Is Changing: What Homeowners Need to Know Before 2026," January 18, 2026.

[20] Jigar Shah, "Pathways to Commercial Liftoff for Virtual Power Plants," US Department of Energy, and DOE VPP program materials, 2023–2026.

[21] Solar Energy Industries Association, "How Virtual Power Plants Are Making the Grid More Affordable, Reliable, and Secure," July 29, 2025; U.S. Department of Energy, "Secretary Wright Issues Emergency Order to Secure Southeast Power Grid Amid Heat Wave," June 24, 2025.

[22] Herman K. Trabish, "In 2026, virtual power plants must scale or risk being left behind," Utility Dive, January 27, 2026; Kathleen Van Gorden, "What the 2025 Virtual Power Plant (VPP) Market Report Tells Us About the Storage-Led Shift: A Conversation with Ohm Analytics," Virtual Peaker, May 27, 2026.

[23] Chris Williams, et al., "Power-Flexible AI Data Centers: A New Paradigm for Grid-Responsive Compute," arXiv preprint 2606.25098, June 23, 2026; Herman K. Trabish, "Data centers are ready to negotiate flexibility for speed," Utility Dive, June 26, 2026.

[24] Herman K. Trabish, "In 2026, virtual power plants must scale or risk being left behind," Utility Dive, January 27, 2026.

[25] KPMG LLP, Powering the Energy Orchestrator: How SAP S/4HANA Enables Utility Transformation, 2025; Oracle, "The Path to Agentic AI: A Collaborative Approach," Oracle AI & Data Science Blog, June 25, 2025; Satish Thomas, "The Era of Agentic Business Applications Arrives at Convergence 2025," Microsoft Dynamics 365 Blog, December 9, 2025; Workday, "Workday Illuminate™ Expands with New AI Agents for HR, Finance, and Industry," September 16, 2025.

[26] "Distributed Energy Can Unleash the Resilient, Affordable Grid of the Future," The Pew Charitable Trusts, April 28, 2026; Amazon Web Services, Unlock Your Smart Meter Data with Meter Data Analytics 2.0 (presentation), accessed July 24, 2026.

[27] "The Energy Orchestrator Series Part 4: How the grid becomes the coordination layer of the AI economy," KPMG LLP, 2026.

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Learn about us: **kpmg.com**