Energy Access is the New Source of AI Advantage

Electricity is the big new constraint, and companies need a strategy to deal with it.

Yinuo Tang, Peking University

Eric Yanfei Zhao, University of Oxford

At the start of the gen-AI boom, the scarcest asset seemed obvious: access to the frontier model. Companies rushed to license models such as GPT-4, Claude, and Gemini, hired prompt engineers, and built proprietary copilots. Then a different constraint emerged: access to GPUs, cloud capacity, and data-center space. Now, beneath all of that, a new constraint is emerging: electricity. The new scarcity is not intelligence but the energy-intensive infrastructure required to produce and deliver it.

This shift is easy to miss, because AI still looks like software. But its economics are increasingly industrial. A model is not just code. It is chips, cooling, land, interconnection rights, and power contracts. The International Energy Agency's 2026 update projects global data-center electricity use rising from about 485 terawatt-hours in 2025 to about 950 terawatt-hours in 2030, with AI-focused data centers' power use tripling over that period. CBRE Group, Inc., a global commercial real estate services and investment firm, similarly identified a continued worldwide power shortage as a significant inhibitor of global data-center growth.

The strategic question for leaders is therefore changing. It is no longer only "Which model should we use?" or "Can we get enough GPUs?" but also "Can we get reliable, affordable, permitted power where and when the compute needs to run?"

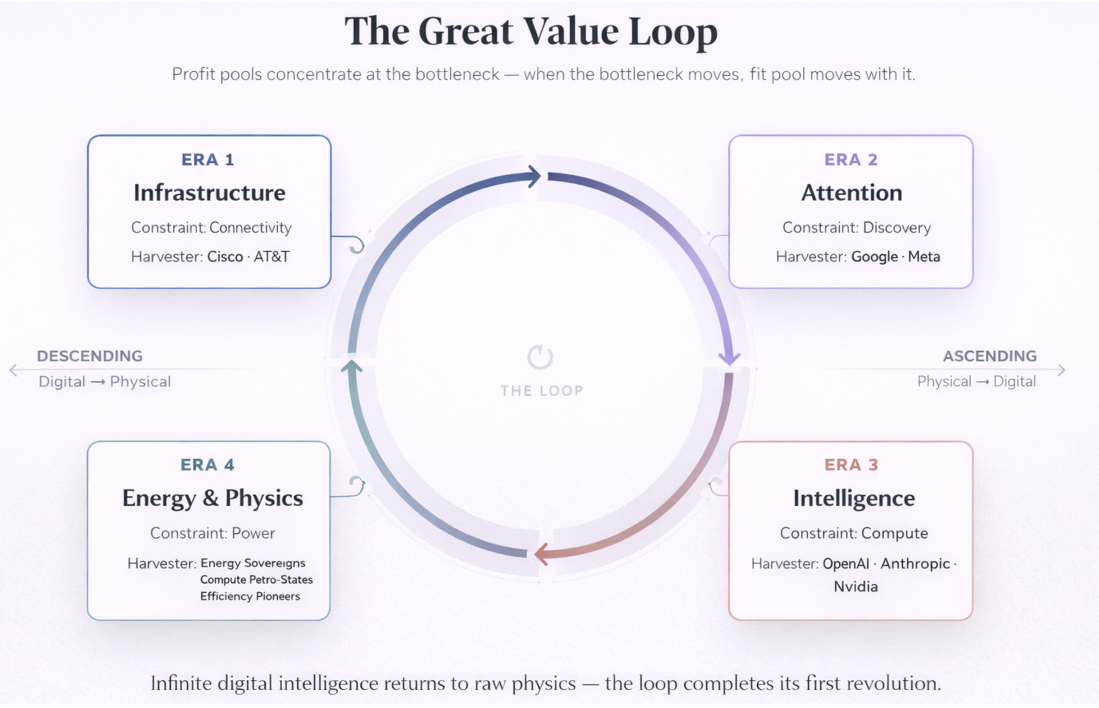
We have decades of research and operational experience working at the intersection of digital strategy and energy systems. Our research on competitive positioning and optimal distinctiveness and platform globalization and renewable-energy finance suggests that what's happening today with AI fits a recurring pattern in

how firms capture value during technological transitions. We call this recurring pattern the Great Value Loop. In the sections that follow, we explain how the Great Value Loop works, where AI is in that loop today, and how business leaders can use it to make better decisions about investment, procurement, and competitive strategy.

The Great Value Loop

The mechanism is simple. A new technology creates a scarce control point, and value then pools at that layer, because customers can't get enough of it. In the AI context, that value comes in the form of models, bandwidth, distribution, cloud capacity, chips. Then capital floods in, standards emerge, suppliers multiply, and buyers learn to procure the layer more efficiently. What was scarce becomes available; what was differentiated becomes a feature.

But adoption doesn't stop. Instead, it accelerates, pressing on the next underlying constraint. The profit pool migrates downward in the stack to whatever can't be copied, rented, or scaled fast enough. The managerial mistake is to keep investing as if yesterday's scarce layer will endure. Instead, in a disciplined way, during every cycle, you have to ask what is becoming abundant, what is becoming standardized, and what bottleneck is forming underneath.



The Great Value Loop has moved through three eras in the modern technological age, and it is now entering a fourth. Era 1 was Infrastructure: the key constraint was connectivity, and firms such as Cisco and AT&T harvested value by controlling the pipes. Era 2 was Attention: the key constraint was discovery, and firms such as Google and Meta harvested value by organizing access to information, products, and people. Era 3 was Intelligence: the key constraint was compute, and firms such as OpenAI, Anthropic, and Nvidia harvested value from frontier models, chips, and AI infrastructure. Now the loop is moving into Era 4: Energy & Physics. The key constraint is power, and the bottleneck is shifting from digital intelligence back into the physical world: electricity, cooling, land, and grid connections. In other words, the loop descends through the stack, from connectivity, to discovery, to compute, and now to power. This new era began when AI demand stopped being measured only in parameters, tokens, or cloud budgets and started being measured in megawatts. The IEA estimates that data centers consumed about 415 terawatt-hours of electricity in 2024, or roughly 1.5% of global demand. It projects that figure will more than double to about 950 terawatt-hours by 2030, with AI-focused facilities growing fastest. The issue is not only aggregate demand; it is location. AI-focused data centers are geographically concentrated, and the grids serving them face local constraints in transmission, interconnection, cooling, and permitting. The clearest signal is how aggressively hyperscalers are now moving upstream into generation itself—signing 20-year nuclear power-purchase agreements, acquiring data-center sites adjacent to reactors, issuing RFPs for gigawatts of new. Meta alone has launched an RFP targeting one to four gigawatts of new U.S. nuclear generation. These are not sustainability gestures. They are infrastructure hedges against the next AI bottleneck.

Efficiency will slow the pressure but not eliminate it. DeepSeek's 2025 release illustrated how quickly reported training costs can fall, but what we're seeing in practice is a version of the so-called Jevons paradox: Intelligence is getting cheaper but expanding the number of economically viable uses, driving total demand up rather than down. That's why what we're living through now is not a normal procurement cycle: Energy is not an input whose price can simply be renegotiated annually. It is local, permitted, slow to build, and politically contested. The firms with advantage will be

those that improve intelligence per watt, secure long-duration optionality, and place compute where reliable power exists.

The Incumbent's Energy Playbook

The arrival of this new era doesn't mean that established companies need now to transform themselves into utilities. A bank, retailer, manufacturer, insurer, or pharmaceutical company will rarely create advantage by buying a power plant. The more realistic goal is to build distinctiveness in energy access before scarcity becomes an operating emergency.

To that end, we recommend taking these five actions:

1) Make energy intensity visible.

Most firms track cloud spend, model accuracy, and AI adoption. Far fewer can answer a more basic question: How much electricity does a given AI workflow require? Leaders should ask for a quarterly AI-energy dashboard that includes energy cost per workflow, tokens or inferences per kilowatt-hour, the share of workloads that are latency-sensitive versus shiftable, cloud-region exposure to constrained grids, and cooling or water assumptions for major deployments. Tools such as Google Cloud's Carbon Footprint and Microsoft Azure Carbon Optimization now allow organizations to track workload-level energy use. Salesforce has operationalized this discipline: Its Sustainable AI framework reports carbon and energy metrics at the workload level to inform model selection, helping teams surface hidden inefficiencies such as redundant internal queries or oversized frontier models handling simple tasks, which often account for a surprising share of AI energy draw. The goal is not perfect measurement. It is to make "intelligence per watt" a management metric, not an invisible engineering variable.

2) Reduce demand before buying supply.

Many companies are overpaying for AI because they send too many tasks to models that are larger, faster, and more energy-intensive than the job requires. A customer-service summary does not always need a frontier model. A compliance search may not need to run instantly. Leaders should require AI teams to route simple tasks to

smaller models, cache repeated queries, compress prompts, quantize models where appropriate, batch nonurgent inference, and shift flexible workloads to lower-cost regions or times. These are not glamorous moves, but they are available now to firms with no control over substations, transmission lines, or generation assets. A compelling example is Pinterest, which has repeatedly optimized its core recommendation systems for scale and efficiency, including through infrastructure-budget-aware recommender design and specialized systems such as Pixie for real-time recommendations.

3) Contract for optionality, not ownership.

Microsoft has signed a 20-year power-purchase agreement with Constellation Energy that is intended to enable the restart of Three Mile Island Unit 1 as the Crane Clean Energy Center, adding roughly 835 megawatts of carbon-free electricity to the grid if the restart is completed. Non-hyperscaler incumbents are adapting versions of this playbook with less dramatic tools: long-term power-purchase agreements, virtual-power purchase agreements (VPPAs), utility green tariffs, colocation capacity reservations, and bilateral cloud or data-center contracts that specify energy exposure and compute availability. General Motors, H&M Group, AB InBev, and Target, for instance, have each signed large corporate VPPAs. Although originally framed as sustainability commitments, such contracts are increasingly being extended and repriced as AI workloads push these firms' electricity demand higher. The procurement question is no longer only "What is the cloud price?" but also "What happens to our AI cost curve if regional power prices rise or interconnection delays worsen?"

4) Redesign where compute runs.

Cloud-region selection used to be mainly about latency, compliance, and vendor architecture. It’s now also an energy decision. AWS's acquisition of Talen's data-center campus adjacent to the Susquehanna nuclear station, for example, and Google's agreement with Kairos Power for advanced nuclear capacity show that hyperscalers are already organizing AI infrastructure around energy access. Non-hyperscaler incumbents are adapting versions of this logic. JPMorgan and other large banks have moved toward selective, multi-region cloud strategies that weigh power availability alongside latency and resilience, and a growing number of European firms, from Spotify to H&M, have shifted analytics and training workloads to Nordic data-center regions

that offer cheaper, cooler, and lower-carbon power. Non-hyperscalers should ask the same questions of their cloud and colocation partners: “Where does the power come from? How constrained is the local grid? What cooling technology is available? Which workloads can move, and which truly must stay near users or headquarters?”

5) Make someone accountable.

Create a standing Compute and Energy Council, chaired jointly by the CIO, CFO, procurement leader, and sustainability or operations leader. Some companies are already moving in this direction: Salesforce has embedded sustainability metrics into its AI development process, and Microsoft provides cloud-emissions visibility and sustainability-governance tools that connect technology, finance, procurement, and sustainability decisions. This council should hold formal veto power. No major AI deployment should be approved without reviewing model efficiency, workload flexibility, cloud-region energy risk, and contract duration. Incumbents do not need to own the electrons. But they do need a distinctive way to measure, reduce, secure, and locate the compute that depends on them.

* * *

The mandate for leaders thinking AI strategy today, as electricity becomes the new constraint, is not to become utilities. It’s to stop treating energy as a background input. The companies that learn to measure, reduce, contract for, and strategically locate their compute will not need to own the electrons. But they will need a distinctive way to access them. And they’ll need to recognize that in the next phase of competition, AI strategy and energy strategy will become inseparable.