CSIS BRIEFS

The Rise of "Watt's Law" and Why Power Could Put a Ceiling on American Innovation

Maryam Khan Cope

FEBRUARY 2026

THE ISSUE

- **A foundational advantage of the U.S. digital economy is weakening.** For decades, Moore's Law and Dennard scaling delivered steady improvements in computing performance while keeping energy costs manageable and underpinning productivity growth, industrial competitiveness, and national security capabilities.
- **AI workloads upend these assumptions.** Modern AI systems depend on massive parallelism, continuous data movement, and sustained operation across large clusters, making energy utilization and availability more important than peak chip performance.
- **This shift defines the emergence of "Watt's Law."** Under Watt's Law, AI capability scales with available power growth multiplied by system-wide innovation driven by full-stack optimization across hardware, software, networking, and operations. This can be measured by growth in tokens per joule of energy.
- **Power is now the significant binding constraint on AI progress.** National competitive advantage increasingly depends on who can most efficiently convert electricity into sustained, real-world AI output, rather than who builds the smallest transistors.
- **The implications are strategic.** Failure to adapt to a power-constrained scaling regime risks forfeiting a key source of productivity growth and long-term advantage in AI-enabled economic and national security capabilities. It also has implications for U.S. partnerships with energy-rich allies.

INTRODUCTION

For most of the modern digital economy, the United States has enjoyed a quiet structural advantage where computing kept getting better while the energy cost of using it generally fell. That was not just a happy coincidence. It was the combined effect of Moore's Law, which provided steady gains in how many transistors could be put on a chip, and Dennard scaling, which for decades ensured that smaller transistors could run at lower voltage without turning chips into miniature space heaters. Over the past decades, the United States has built numerous business models, national security capabilities, and entire industrial supply chains on the assumption that "more compute" would arrive on schedule and at tolerable power costs.

That assumption is collapsing, and countries that can mobilize power and compute at speed are on track to surpass the United States. The AI boom has pushed demand for computing beyond the comfortable world where transistors became cheaper and more efficient as a matter of course. The new constraint is increasingly mundane: elec-

tricity. Not in the abstract "energy transition" sense, but in the near-term reality of how quickly a large data center can be connected to the grid, how reliably it can be run, and how cheaply it can be powered once built. When AI systems can draw power on the scale of industrial plants, "innovation policy" starts to look like "infrastructure policy."[1]

There is a macroeconomic reason to take this seriously. Federal Reserve research suggests that the miniaturization channel—shrinking electronic components in the Moore's Law era—was not a footnote to economic growth but a measurable and significant contributor to it. A Federal Reserve Bank of New York staff report estimates that 11.74 percent to 18.63 percent of U.S. productivity growth from 1960 to 2019 can be attributed to miniaturization.[2] Put differently: A meaningful slice of U.S. GDP growth over the last half-century can be traced to the reliable march of semiconductor improvement. The same line of analysis finds that Moore's Law effects were especially strong in heavy manufacturing, where productivity gains can be tightly linked to the adoption of better capital equipment.[3] These findings do not claim that semiconductors were the only engine of growth—but they underscore how dangerous it would be to treat changes in the compute-power relationship as a niche technical issue.

The underlying physics shift is very well understood inside the chip industry. Dennard scaling faltered in the mid-2000s because voltage could no longer be reduced without unacceptable leakage, noise, and reliability problems. As voltage stopped falling, engineering changes, such as raising clock speeds each generation, ran into the "wall of heat." The industry adapted by spreading work across more cores and then, crucially, by pushing more workloads onto specialized accelerators. The result was the end of a particular kind of progress: the "automatic" kind where performance gains came with manageable power changes, and, importantly, without requiring dramatic system-level changes.

AI has made that transition away from Dennard scaling and Moore's Law impossible to ignore. Training modern models is not like running office software faster; it is like operating an always-on industrial process. AI workloads are parallel and very hungry for data movement. Their hardware is different—instead of the traditional in-and-out processing done on a simple central processing unit (CPU), they lean on thousands of chips or chiplets running together, synchronized frequently, often for weeks. In that setting, peak performance numbers matter less than the system's ability to sustain high output hour after hour without wasting power on waiting, for memory, for networking, for storage, or for poor scheduling. The question that now separates winners from also-rans is no longer "Who has the smallest transistors?" but "Who can turn megawatts into useful AI output most efficiently?"

That is why it is useful to talk about the next scaling era with a name the public can remember. This paper proposes Watt's Law: the idea that the next decade of progress will be driven less by transistor shrinkage alone and more by how effectively watts of electricity are converted into useful computation through better full-system design, including accelerators, memory, networking, software, distribution via edge computing, and operations.[4] Watt's Law is not a law of physics and not a guaranteed doubling schedule. It is a rule-of-thumb about what is increasingly determining real-world AI capability, which is system efficiency and utilization under power constraint.

To see why, consider how a modern "AI factory" actually works. The CPU, still essential, is increasingly the manager of the system rather than the main producer of AI math. CPUs run the operating system, coordinate tasks across thousands of devices, handle data preparation, and keep the workflow moving. Heavy lifting is done by AI accelerators, often GPUs, but now also neural processing units (NPUs), tensor processing units (TPUs), and a whole host of other accelerators designed for specific outcomes. They are designed to churn through the matrix operations at the heart of machine learning. But accelerators are only productive when they are fed continuously, which makes memory and data movement central to performance. High bandwidth memory (HBM), close to the chip, is expensive and limited but fast enough to keep accelerators from idling. System memory is larger but slower. Storage is larger still and slower again. If the "conveyor belt" slows, the "machines" sit idle, burning money and electricity while producing nothing.

Networking is another silent determinant. Large AI systems depend on chips talking to one another constantly. If the interconnect is congested, training slows and utilization collapses. In practice, this means AI performance is increasingly the output of a physical pipeline: CPU coordination, accelerator throughput, memory bandwidth, storage throughput, and network fabric, all bound together by software that decides how the workload is split and sched-

uled. This is why the language of "factories" fits: modern AI computing is production of AI tokens, and its productivity is measured not by peak capability but by sustained throughput of these tokens.

WATT'S LAW AND RACE FOR TOKENS PER JOULE

Aurora, the U.S. exascale system at Argonne National Laboratory, is a helpful public blueprint for this kind of co-design. Aurora is an HPE Cray EX system with 10,624 nodes, built from a pairing of 21,248 CPUs and 63,744 GPUs, and it reflects the "CPU coordinates, accelerators compute" model at enormous scale.[5] Public system descriptions emphasize that each node combines two CPUs with six GPUs, and that this architecture is designed as an integrated system rather than a loose collection of parts.[6] Aurora also illustrates the power reality: Public references widely cite a power draw around ~38.7 megawatts, which is directionally sufficient to make the point even if the exact figure shifts with operating conditions and measurement conventions.[7] Aurora's power requirements are equivalent to roughly the continuous electricity demand of 30,000–40,000 homes, all drawing power at once. At this scale, electricity supply is not just a background variable—it is a design constraint.

The most revealing evidence for Watt's Law comes from the companies most invested in the old story. ASML, the central supplier of advanced lithography tools used to manufacture smaller and smaller transistors, has made unusually direct statements in investor materials that speak to the new constraint. In its 2024 Investor Day messaging, ASML highlights that high-performance computing demand is moving "Beyond Moore's Law," that generative AI will "demand an acceleration of Moore's law," and, most importantly, that "energy use could set a ceiling on training capability" if the industry does not address cost and power challenges.[8] When the firm that sits at the center of transistor scaling says power could cap training, policymakers should hear it as a strategic warning—innovation is becoming shaped by system-wide performance, including the performance of the electrical grid as it works to supercharge the AI factory's technical infrastructure during the training of power-hungry models.

However, we should never discount the major upside of replacing Moore's Law and Dennard scaling with something more powerful. Independent trend analysis supports a measured version of the "faster-than-Moore" trajectory of AI compute innovation. The same analysis also notes that some popular "law" claims in the industry are overstated relative to observed data, but even this analysis may not stand the test of time as AI hardware systems advance rapidly.[9]

The standard way to measure how fast a computer can perform numerical calculations, such as the kind used in science, AI, graphics, and simulations, is FLOP/s (Floating Operations Per Second). Epoch AI's analysis of hundreds of GPUs finds that FLOP/s per dollar doubles about every ~2.46 years across a broad dataset, with different subsets moving faster or slower.[10] The takeaway is not that progress is slowing to a crawl. It is that technical hardware progress increasingly comes from compounding improvements across multiple layers, such as design co-optimization, pure innovation in CPU, accelerator and hardware design, memory systems, interconnects, software and compilers, and the operational innovations of keeping systems busy, rather than from transistor shrink alone.

That makes it possible to talk concretely about trajectories of how quickly AI factories can improve performance. A familiar Moore's Law benchmark is a doubling of performance every two years. By contrast, Epoch AI finds that the cost efficiency of GPUs has improved more slowly, roughly doubling every two and a half years across a broad set of products.[11] That metric misses what increasingly matters in practice—that under Watt's Law, the relevant yardstick is not transistor counts or peak performance of these GPUs, but useful AI output per unit of energy (i.e., how many AI tokens a system can produce per joule, or per megawatt-hour, in real operation).

In this Watt's Law regime, capability scales with available power multiplied by system-wide efficiency.[12] Gains come from *power availability and price*, as well as improvements across the stack, hardware design, memory and networking, software, and operational discipline that raise utilization and reduce waste. When these advances compound, well-run AI systems can deliver progress that feels faster, perhaps much faster, than Moore's Law, not because physics has changed, but because energy-constrained systems are being optimized end to end, and the power system can deliver what is needed during this optimization. This is the heart of Watt's Law: Compounding returns come from running the AI system like an industrial asset that needs increased power to fuel improvement, instead of treating it like a consumer gadget that requires software upgrades.

The policy consequences follow directly. If energy use can cap AI training capability, then the bottlenecks are not limited to chips. They include grid interconnection timelines, transmission capacity, shortages of transformers and switchgear, cooling needs, water constraints in some regions, permitting and local politics, and electricity price volatility. These are the sorts of constraints that markets can address over time, but, unfortunately, national competition rarely affords the luxury of time when rivals are investing heavily and learning quickly. And that is why the chief executives of major AI innovators have proclaimed publicly, and frequently, that energy policy is a pillar of winning the global AI race.

To keep pace with Watt's Law, technology leaders have pursued multiple strategies to rapidly add AI compute capacity, from hyperscaling training clusters and expanding large-scale inference via cloud providers to deploying edge AI and investing in national or strategic compute programs. But regardless of the pathway chosen—market-driven or state-led—the bottleneck is the same. Power supply, not silicon, now sets the upper bound on how quickly new AI compute can be deployed.

The downside scenario is not hypothetical. Export controls and chokepoints can slow an adversary, but they are not a permanent moat. CSIS analysis has argued that chip export controls have limits and can incentivize domestic substitution efforts, particularly when the targeted country has scale and strong state support.[13]

CSIS has also chronicled how rule updates and threshold adjustments are necessary because actors adapt, work around constraints, and reconfigure supply chains.[14] Another CSIS analysis highlights a more structural problem: Coordinated export controls depend on the legal authority and enforcement willingness of allies, and those authorities vary widely, complicating a unified chokepoint strategy.[15] U.S. government sources underline both the intent and the complexity: The Bureau of Industry and Security's October 7, 2022, controls demonstrate seriousness about restricting advanced computing and semiconductor manufacturing items to China for national security reasons, while Congressional Research Service and Government Accountability Office reporting highlight ongoing challenges in sustaining, enforcing, and overseeing expansive controls.[16]

MEETING THE CHINA CHALLENGE

If China or other countries develop robust homegrown AI compute, even if less efficient per chip, the strategic calculus changes. Chokepoints erode from "hard stops" into "cost taxes." Workloads can be adapted, models can be trained on larger clusters of weaker hardware, and software ecosystems can be rebuilt as resilient systems over time. Open-source reporting suggests that the direction of travel is toward domestic alternatives and large-scale aggregation strategies that may be power inefficient but still strategically useful.[17] In a world where global competitors can reliably generate their own compute, the U.S. advantage depends much less on denial and more on building durable superiority, such as cheaper power, faster deployment, higher utilization, better software, better distribution of workloads, and faster diffusion into industry and consumer use cases.

Further, CSIS's *Tech Edge* net assessment shows that China's technology ecosystem is not simply a collection of isolated capabilities, but a system designed to scale and diffuse technological capacity across sectors.[18] Beijing's ecosystem strategy combines state guidance, market incentives, supply chain mastery, and long-term capital investment in base and production technologies that underpin "stack categories" like AI compute and semiconductors. This systemic view explains why China can make strategic trade-offs that may appear inefficient at the chip level but still yield national advantage at scale, a dynamic "Watt's Law" scenario helps explain in the energy-constrained era.

THE OPPORTUNITY OF THE "WATT'S LAW" ERA

That is also where the upside case becomes compelling. Watt's Law is a once-in-a-generation opportunity. A power and stack-led era rewards the American strengths that do not require a monopoly on lithography, including leadership in systems engineering, software, cloud operations, advanced packaging innovation, and importantly, industrial integration. If the United States can pair compute strategy with power strategy, faster interconnections, more transmission capacity, modernized grid equipment supply chains, and permitting that reflects the national stakes, then it can sustain leadership not just in creating AI models, but in turning AI into economy-wide productivity.

Under conditions of geopolitical fracturing, there is an additional advantage: The manufacturing industry is the most immediate beneficiary of AI leadership. Efficient AI factories connected to physical systems can enable simulation-driven design, digital twins, predictive maintenance, and more capable robotics, shrinking manufacturing and development cycles and reducing downtime. Biotechnology stands to gain from better protein modeling, faster candidate screening, and tighter "lab-in-the-loop" systems that make discovery more iterative and less dependent on slow trial-and-error. Energy and materials science can benefit from compute-intensive exploration of catalysts, batteries, and process optimization. The much-heralded and much-needed Genesis Mission at the U.S. Department of Energy could provide a boost to AI-enabled scientific research that counters declining U.S. patents in key areas like materials science, analog circuits, and a host of other strategic industries. Further, these AI systems show incredible promise for addressing socioeconomic challenges like aging populations, safer transportation, and cost-effective, widely available healthcare.

The broader point is that Watt's Law can restore a macroeconomic tailwind—if the United States treats the energy and compute link as a first-order national priority rather than a technical afterthought.

RECOMMENDATIONS

The recommendations are therefore practical and familiar in form, even if the context is new. First and foremost, the federal government should treat large-scale AI compute more like an emergency need for strategic infrastructure and develop integrated and forward-leaning policies tied to power delivery, resilience planning, and industrial capacity.

- An urgent all-of-government effort is essential: The National Science Foundation, the Department of Energy, the National Institute of Science and Technology, and the National Institutes of Health, in conjunction with the AI infrastructure industry, should form an emergency task force to map out national needs to accelerate AI and provide recommendations to Congress, the president, the Federal Energy Regulatory Commission, the North American Electric Reliability Corporation, and state utilities and state governments across the country.
- The President's Council of Advisors on Science and Technology should form a subgroup to issue yearly reports on the nation's progress in supplying energy to AI infrastructure. A key role will be one of convening major industrial and state actors to address common challenges and develop practical policy recommendations in close consultation with industry. The recommendations should include moonshots around making kilowatt-hours of energy available to AI infrastructure on an urgent timeline.
- National test beds should be quickly established to model the energy needs of the different types of AI deployments that are needed to help industry learn how to raise utilization under real-world power and cooling constraints, similar to prior federal support for test beds for the bandwidth-constrained transitions from 4G to 5G and now 6G.
- Respect power availability as a strategy. With regard to global competitiveness, export controls should be treated for what they are—time-buying tools essential to national security that must be paired with a domestic build strategy, not a substitute for it. Foreign policy must also take into account that energy-rich U.S. allies may be ideal infrastructure partners in the race to win AI leadership.

A NEW REALITY

The United States cannot out-innovate physics, but it can out-organize competition. The next era will be defined in megawatts, uptime, utilization, and the ability to run AI factories as productive national assets. Moore's Law still matters, but it no longer tells the whole story. Watt's Law is the new reality, and now progress comes from how efficiently electricity is turned into useful computation—through better systems, better software, and better infrastructure, all underpinned by better policy. The countries that win will be the ones that can power their innovation. ■

Maryam Khan Cope *is a non-resident senior associate with Renewing American Innovation at the Center for Strategic and International Studies in Washington, D.C.*

This report is made possible by general support to CSIS. No direct sponsorship contributed to this report.

ENDNOTES

1 Christophe Fouquet, "Small Talk 2024: Global Market Trends: Industry & ASML's Technology Roadmap ESG," ASML, November 14, 2024, https://portalvhds1fxb0jchzgjph.blob.core.windows.net/press-releases-attachments/3504937/Report%20of%20foreign%20issuer%20%5BRules%2013a-16%20and%2015d-16%5D.pdf.

2 Pablo Azar, *Moore's Law and Economic Growth*, Staff Report no. 970 (New York: Federal Reserve Bank of New York, May 2021, revised October 2022), https://www.newyorkfed.org/research/staff_reports/sr970.

3 Ibid.

4 "Watt's Law" framing in this paper is a policy shorthand for stackled scaling and utilization-driven efficiency, grounded in ASML's energy-ceiling warning and observed GPU price-performance trends. Watt's Law can be expressed as "growth in token output = growth in power + growth in efficiency." At a system level, Watt's Law could also be measured as $g_\tau = g_P + g_\eta$; $g_\eta = g(arch) + g(accel) + g(mem) + g(net) + g(sw) + g(ops) + gU$. G(T)= Tokens/sec growth. G(P)= Growth in power. So if layers improve in parallel, g_η (growth in system efficiency) can exceed a single-channel Moore-style rate even when each component improves modestly.

5 "Aurora Machine Overview," Argonne Leadership Computing Facility, https://docs.alcf.anl.gov/aurora/.

6 "Aurora: Architecting Argonne's First Exascale Supercomputer for Accelerated Scientific Discovery," arXiv, https://arxiv.org/html/2509.08207v1; and "Aurora Supercomputer Blade Installation Complete," Intel Newsroom, June 22, 2023, https://newsroom.intel.com/artificial-intelligence/aurora-supercomputer-blade-installation-complete.

7 Publicly compiled Aurora power figure (~38.7 MW) used directionally for policy-scale framing. See, for example, Precious Eyabi et al., "Evaluating the Power Monitoring Capabilities of Aurora," https://sc25.supercomputing.org/proceedings/posters/poster_files/post274s2-file3.pdf.

8 Fouquet, "Small Talk 2024."

9 Marius Hobbhahn and Tamay Besiroglu, "Trends in GPU price-performance," Epoch AI, June 27, 2022, https://epoch.ai/blog/trends-in-gpu-price-performance.

10 Ibid.

11 Ibid.

12 See endnote 4 on Watt's Law.

13 Sujai Shivakumar, Charles Wessner, and Thomas Howell, "The Limits of Chip Export Controls in Meeting the China Challenge," CSIS, *Commentary*, April 14, 2025, https://www.csis.org/analysis/limits-chip-export-controls-meeting-china-challenge.

14 Emily Benson, "Updated October 7 Semiconductor Export Controls," CSIS, *Commentary*, October 18, 2023, https://www.csis.org/analysis/updated-october-7-semiconductor-export-controls.

15 Gregory C. Allen and Isaac Goldston, *Understanding U.S. Allies' Current Legal Authority to Implement AI and Semiconductor Export Controls* (Washington, DC: CSIS, March 2025), https://www.csis.org/analysis/understanding-us-allies-current-legal-authority-implement-ai-and-semiconductor-export.

16 Karen M. Sutter, *U.S. Export Controls and China: Advanced Semiconductors* (Washington, DC: Congressional Research Service, September 2025); and "Export Controls: Commerce Implemented Advanced Semiconductor Rules and Took Steps to Address Compliance Challenges," U.S. Government Accountability Office, December 2, 2024, https://www.gao.gov/assets/gao-25-107386.pdf.

17 Reporting on China's adaptation strategies illustrating the erosion of chokepoints and substitution approaches. See, for example, Hanna Dohmen, Jacob Feldgoise, and Charles Kupchan, "The Limits of the China Chip Ban," *Foreign Affairs*, July 24, 2024, https://www.foreignaffairs.com/china/limits-china-chip-ban.

18 Navin Girishankar et al., *Tech Edge: A Living Playbook for America's Technology Long Game* (Washington, DC: CSIS, January 2026), https://www.csis.org/analysis/tech-edge-living-playbook-americas-technology-long-game.