Why Compute Is Eating the World

The Economics of Intelligence and the Infrastructure Behind It

In 2011, Marc Andreessen argued that software was eating the world. He pointed to a change already visible in the market: Amazon was taking business from traditional retailers, Netflix was replacing video stores, and software was becoming essential to communications, transportation, and finance. His argument drew much of its force from the costs and adoption figures behind those changes. At Loudcloud, for example, he estimated that hosting a basic internet application had cost about $150,000 per month in 2000, compared with $1,500 in Amazon's cloud by 2011.[1]

Fifteen years later, a new investment cycle is underway. Advanced models can write and debug software, analyze financial information, assist scientific research, and perform many tasks that used to consume substantial human time. Whether these systems should be called artificial intelligence, general intelligence, or superintelligence is a legitimate debate. Their commercial utility, however, can be evaluated without resolving it. My view is that computation is becoming a more important input to economic production, and the most direct evidence is the capital now being committed to making it available.

Several developments explain why this is happening now. Stanford's 2025 AI Index found that the price of accessing a model performing at roughly GPT-3.5's level on the MMLU benchmark fell from approximately $20 per million tokens in November 2022 to $0.07 by October 2024, a decline of more than 280-fold. Stanford also reported rapid improvements in machine-learning hardware performance and energy efficiency. A given task can therefore become cheaper even while the best available models require increasingly ambitious infrastructure.[2]

Falling costs expand the number of tasks worth automating. They also give smaller firms access to capabilities that previously required large research and engineering teams. The question for infrastructure investors is whether those new uses will expand demand for computation faster than better algorithms and hardware reduce the resources needed per task. That is an economic question, and it helps explain both the intensity of today's investment and the possibility of overbuilding.

Consider the capital budgets. Amazon has indicated approximately $220 billion of capital expenditures for 2026. Alphabet has guided to roughly $195 billion to $205 billion, and Meta expects $130 billion to $145 billion, including principal payments on finance leases. Taken together, the three companies anticipate around $545 billion to $570 billion in annual spending. These figures include assets beyond AI and should not be represented as exclusively AI expenditure. Still, the expansion of computing infrastructure is among the central reasons for the increase.[345]

Microsoft offers a more precise indication of the hardware mix. It reported approximately $41 billion in capital expenditures during the June 2026 quarter, including finance leases, and said roughly two-thirds was directed toward shorter-lived assets such as GPUs and CPUs. The remainder supported longer-lived assets, including data centers. This distinction matters: compute involves both rapidly depreciating electronics and physical facilities intended to operate for years.[6]

Demand is increasingly visible in operating results as well. Nvidia reported fiscal 2026 revenue of $215.9 billion, up 65% year over year, with data-center products central to its growth. Alphabet's Google Cloud generated approximately $24.8 billion of revenue in the June 2026 quarter, up 82% from a year earlier. Neither figure measures revenue from superintelligence. Both demonstrate the scale of commercial activity supplying accelerated computation and cloud services.[78]

A second group of companies has built businesses more directly around selling access to computing capacity. CoreWeave reported approximately $2.58 billion of revenue in the second quarter of 2026 and about $104 billion of revenue backlog as of June 30. It raised its full-year capital spending forecast to $35 billion to $39 billion. Backlog represents contracted future business subject to delivery and other conditions, not revenue already earned. Nevertheless, such commitments provide evidence that customers are willing to reserve enormous amounts of future capacity.[9]

Nebius is another revealing case. The company reported $582.3 million in second-quarter 2026 revenue, an increase of 454% year over year, and approximately $5.7 billion of capital expenditures in that quarter, primarily for GPUs and data-center expansion. Its projected 2026 capital spending of $20 billion to $25 billion

is striking relative to its current revenue. The business requires substantial financing now in expectation of cash generation over future years.[10]

IREN illustrates how companies with access to power and land can reposition themselves around this demand. Originally developed around Bitcoin mining, IREN has expanded into AI cloud infrastructure. Its fiscal 2026 AI cloud revenue reached $128.8 million, compared with $16.4 million in fiscal 2025. In its August 2026 update, IREN reported $4 billion of contracted annualized run-rate revenue associated with 2026 capacity, compared with $1 billion of operating annualized run-rate revenue at that point. These run-rate measures are different from recognized annual revenue. IREN also described $3.6 billion of GPU financing for a Microsoft contract and a further $2.8 billion of financing for other deployments.[11]

This is an important change in the financing of technology. Banks and other lenders are extending debt against GPUs, customer contracts, and expected data-center cash flows. Companies are combining equity, secured financing, and customer prepayments to deploy equipment. The resulting businesses look in part like cloud platforms and in part like infrastructure projects. Their financial success will depend not just on technology, but on borrowing costs, utilization, power prices, depreciation, and the reliability of customer payments.

Electricity is already a major constraint. Lawrence Berkeley National Laboratory estimated that American data centers used 58 terawatt-hours in 2014 and 176 terawatt-hours in 2023, or approximately 4.4% of U.S. electricity use. It projected consumption of between 325 and 580 terawatt-hours by 2028. The broad range reflects genuine uncertainty about deployment and efficiency, but even the lower estimate implies a significant increase.[12]

Globally, the International Energy Agency estimated data-center electricity consumption at 415 terawatt-hours in 2024, approximately 1.5% of worldwide demand. Its base case reaches about 945 terawatt-hours by 2030. The IEA also estimated that around $500 billion was invested in data centers globally in 2024. These are investments in a physical supply chain: power generation, transmission, transformers,

cooling, construction, networking, and servers. Grid connections and available generation can take longer to secure than a data-center building takes to construct.[13]

The implications reach far beyond California and Washington. Areas with abundant power, developable land, functioning permitting systems, and access to capital may attract computing facilities. The result could be a new geography of technology investment. Yet large projects impose costs on local communities and electricity systems, and having a development pipeline is not equivalent to completing operating capacity. It is particularly important to distinguish a company's announced megawatts from energized, revenue-producing megawatts.

The broader reason to build all this infrastructure is that computation can increase the amount of useful work produced by individuals and organizations. Google DeepMind's AlphaFold database makes more than 200 million predicted protein structures available to researchers. Those predictions do not replace laboratory verification, but they can help scientists select better questions and move through some stages of research more quickly. Similar potential exists in software engineering, finance, design, and industrial research.[14]

I expect human judgment and taste to remain crucial. Models can produce options, conduct analyses, and assist execution; people still choose objectives, decide which risks to accept, and determine what is worth building. If computational tools make technical work cheaper, entrepreneurs and researchers with good judgment may be able to accomplish far more with smaller teams. That increases the potential market for computation, while keeping its ultimate usefulness tied to human choices.

Access to computing capacity may also matter to national economic strategy. Advanced chip fabrication, specialized equipment, cloud infrastructure, and electricity are concentrated in different countries and companies. The United States has strengths in chip design, research, and capital markets; Taiwan occupies a crucial place in leading-edge semiconductor production; and equipment and memory supply chains extend across Europe and Asia. Governments seeking domestic computing capacity must assemble these resources and secure access to advanced chips.

There is, however, an obvious objection to this argument: infrastructure demand may be real while the return on today's infrastructure investments is disappointing. The railroad buildout transformed transportation,

and the internet created immense economic value, but shareholders in many early businesses lost money. A leased GPU that is fully utilized at attractive prices is an asset. The same GPU sitting idle, financed with expensive debt while new models and processors improve, can rapidly become a liability.

CoreWeave's backlog, Nebius's growth, and IREN's contracted run-rate figures are useful evidence of customer demand. They are not proof of sustainable free cash flow or strong returns on invested capital. The industry must eventually demonstrate that revenues cover energy, maintenance, depreciation, financing, and the continued replacement of advanced processors. Improvements in inference efficiency could also pressure the price suppliers can charge, even if total usage continues to grow.

The distinction between usage and profitability should shape how we evaluate the coming decade. Investors should follow recognized revenue, contracted capacity, energized megawatts, utilization, financing terms, and cash returns, rather than treating every spending announcement as value creation. Businesses with low-cost energy, strong customers, efficient hardware deployment, and disciplined financing may develop durable advantages. Others may discover that they built too early or paid too much.

Andreessen's 2011 argument was persuasive because software's improving economics could already be observed in businesses and consumer behavior. Today's evidence suggests that advanced computation has become important enough to justify a historically large industrial investment cycle. Whether it ultimately produces superintelligence remains unresolved. The consequences of building and using this infrastructure are already visible in company budgets, cloud revenues, semiconductors, electricity demand, and financing markets.

Over the next decade, I expect more industries to rely on computational capacity and more countries to regard it as economically strategic. The companies that win will not necessarily own the most hardware. They will be the ones that turn available computation into valuable products and services at attractive costs. That is why I believe compute is eating the world.

Notes

1. Marc Andreessen, "Why Software Is Eating the World," Andreessen Horowitz, August 20, 2011. https://a16z.com/why-software-is-eating-the-world/

2. Stanford HAI, "The 2025 AI Index Report," 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report

3. Amazon, 2026 capital expenditure outlook; see company Q2 2026 results and subsequent reporting. https://www.aboutamazon.com/news/company-news/amazon-q2-2026-earnings

4. Alphabet, Second Quarter 2026 Results and investor presentation. https://abc.xyz/investor/

5. Meta Platforms, "Meta Reports Second Quarter 2026 Results," July 2026. https://investor.atmeta.com/investor-news/press-release-details/2026/Meta-Reports-Second-Quarter-2026-Results/

6. Microsoft, "FY26 Q4 Earnings" investor materials. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q4

7. NVIDIA, FY2026 financial results and annual filing. https://investor.nvidia.com/financial-info/financial-reports-and-sec-filings/default.aspx

8. Alphabet, Second Quarter 2026 Results. https://abc.xyz/investor/

9. CoreWeave, "CoreWeave Reports Strong Second Quarter 2026 Results," August 2026. https://investors.coreweave.com/news/news-details/2026/CoreWeave-Reports-Strong-Second-Quarter-2026-Results/default.aspx

10. Nebius Group, "Second Quarter 2026 Shareholder Letter," August 2026. https://www.sec.gov/Archives/edgar/data/1513845/000110465926094568/tm2622968d1_ex99-2.htm

11. IREN Limited, "IREN Reports FY26 Results," August 27, 2026. https://www.sec.gov/Archives/edgar/data/1878848/000187884826000051/irenreportsfy26results.htm

12. Lawrence Berkeley National Laboratory, "2024 United States Data Center Energy Usage Report," 2024. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

13. International Energy Agency, "Energy and AI," 2025. https://www.iea.org/reports/energy-and-ai/executive-summary

14. Google DeepMind, "AlphaFold Protein Structure Database." https://deepmind.google/science/alphafold/

Works Cited

Alphabet, Second Quarter 2026 Results and investor presentation. https://abc.xyz/investor/

Alphabet, Second Quarter 2026 Results. https://abc.xyz/investor/

Amazon, 2026 capital expenditure outlook; see company Q2 2026 results and subsequent reporting. https://www.aboutamazon.com/news/company-news/amazon-q2-2026-earnings

CoreWeave, "CoreWeave Reports Strong Second Quarter 2026 Results," August 2026. https://investors.coreweave.com/news/news-details/2026/CoreWeave-Reports-Strong-Second-Quarter-2026-Results/default.aspx

Google DeepMind, "AlphaFold Protein Structure Database." https://deepmind.google/science/alphafold/

IREN Limited, "IREN Reports FY26 Results," August 27, 2026. https://www.sec.gov/Archives/edgar/data/1878848/000187884826000051/irenreportsfy26results.htm

International Energy Agency, "Energy and AI," 2025. https://www.iea.org/reports/energy-and-ai/executive-summary

Lawrence Berkeley National Laboratory, "2024 United States Data Center Energy Usage Report," 2024. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

Marc Andreessen, "Why Software Is Eating the World," Andreessen Horowitz, August 20, 2011. https://a16z.com/why-software-is-eating-the-world/

Meta Platforms, "Meta Reports Second Quarter 2026 Results," July 2026. https://investor.atmeta.com/investor-news/press-release-details/2026/Meta-Reports-Second-Quarter-2026-Results/

Microsoft, "FY26 Q4 Earnings" investor materials. https://www.microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q4

NVIDIA, FY2026 financial results and annual filing. https://investor.nvidia.com/financial-info/financial-reports-and-sec-filings/default.aspx

Nebius Group, "Second Quarter 2026 Shareholder Letter," August 2026. https://www.sec.gov/Archives/edgar/data/1513845/000110465926094568/tm2622968d1_ex99-2.htm

Stanford HAI, "The 2025 AI Index Report," 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report