CRA Insights

CRA Charles River Associates

September 2025

EFLOPS >> MWs: The economics behind data center growth

By Maxime C. Cohen and Christopher Russo

Introduction

It’s hard to go more than a few hours today without hearing about data centers and how they’re affecting the energy industry. Every conference and social media post contains a torrent of statistics on how “explosive,” “unprecedented,” and “dramatic” growth in data centers is “straining the grid,” followed by rhetorical questions wondering “how the grid will cope.”

Simultaneously, today’s market for generative artificial intelligence (AI) computing capacity is marked by extreme optimism, massive capital investment, media hype, and strategic signaling. On one side, hyperscalers, such as AWS, Microsoft Azure/OpenAI, Meta, and Google Cloud, are racing to build data centers to capture first-mover advantages and secure scarce compute resources. But on the other, the long-term demand for generative AI services (both consumer facing and enterprise) remains highly uncertain.

For the first time in many years, software has a tangible marginal cost: the cost of energy for training models, inference optimization, and large-scale deployment. But computational power is no different than any other good in the marketplace; it has a marginal cost, and so there is also a marginal value. The former is relatively straightforward to calculate, but the latter is harder to understand. It is likely that at some point, the marginal value of generative AI capacity will match the marginal cost, and the market will reach some kind of equilibrium. But what is this equilibrium outcome? Or in different terms—how many more data centers are going to be built?

In this paper, we present an alternate approach to examining this question in order to understand what’s driving the growth of generative AI, the implications for the energy industry, the market dynamics of computational capacity, and what future growth may be.

The gold rush for computing capacity and energy

The true returns on data center development are only known with certainty by hyperscalers. But we can develop an estimate by looking at the question through a slightly different lens: their publicly known return on equity (ROE).

A conservative estimate of the current ROE for most hyperscalers is in the neighborhood of 30%.[1] A reasonable assumption would be that if a hyperscaler is building a data center, it would expect to realize a return equal to at least its average return (if not, it would invest elsewhere).

There are, of course, reasons to think that the actual ROE could be higher or lower than what is currently reported. The immediate returns on generative AI investment may be different than the longer-term returns. As with many investments in the software industry, the marginal cost of providing services may drop significantly after models have already been trained, and consequently, returns may increase. There may also be additional long-term value in the form of network externalities, including technology lock-in, synergies with other products, and support and enhancements for other technologies (including ones that are not developed yet).

There is a significant first-mover advantage when it comes to generative AI capacity, both from a computational capacity perspective and from an energy perspective. Moving first can confer significant benefits to hyperscalers, such as:

- Greater control of resources, such as scarce graphics processing units (GPUs) and custom chip development.
- Locking in developers to a particular application programming interface (API), leading to positive network externalities.
- Better talent acquisition—talent often gravitates toward a market leader.
- Greater and sooner influence on policy and regulatory standards.
- First opportunities for alliances, creating strategic barriers for competitors.

It would be incomplete to omit some of the potential first-mover disadvantages for these hyperscalers, which could include high immediate capital costs and the attendant risk of overinvestment, the risk of rapid technological obsolescence as the cost of technology decreases rapidly, regulatory scrutiny and public backlash, and organizational rigidity. But at present, the marketplace and capital flows seem to suggest that the opportunities outweigh these risks.

The cost of energy may not be particularly significant, as we describe below, but these are industries (tech and utilities) that move at very different timescales. This most visibly manifests itself in the speed of construction: a midsize data center can be constructed in roughly 18 months, but construction of a power plant and its constituent steps, including regulatory approval, fuel supply, procurement of generation equipment, construction of transmission, etc., may take far longer. Much like the first-mover advantage in computing capacity, every megawatt (MW) of capacity that a hyperscaler is able to secure is a MW that one of their competitors cannot and further increases the incentives to come online quickly.

[1] As of the drafting of this paper, the *Wall Street Journal* lists Microsoft's ROE as 37% (wsj.com, retrieved July 14, 2025).

The difference in returns between hyperscalers and utilities helps explain why the industries move at such different paces. A rough ROE for a utility is around 10%, a fraction of the typical returns earned by a hyperscaler on every dollar spent. As we will see, utilities make money by building things, but they have a far lower incentive to build things as quickly as hyperscalers do.[2]

Does the cost of energy matter?

The cost of baseload electrical generating capacity[3] has increased over the past several years, and the wait for a gas-fired combustion turbine is currently reported to be several years. As energy practitioners, we tend to look at the cost of generating capacity (and the electricity it generates) as a key consideration in the economics of data centers, but capacity and energy costs may be a relatively small factor.

A reasonable estimate today for the capital cost of a hyperscale data center is approximately $12 million per MW.[4] We can further assume that a hyperscaler would need to invest approximately $1 million per year per MW-IT to keep the technology current, and we can use a typical power usage effectiveness ratio of 1.4.[5]

Our analysis shows that at a (conservative) capacity cost of $2,000 per kilowatt (kW), and under our base assumptions, it would take a cost decrease of $450/kW, or almost a 25% reduction, in energy costs to justify delaying construction by one year.

Conversely, this means that a hyperscaler would be willing to pay almost 25% more for energy to come online one year sooner. In other words, the implied financial returns on computational capacity far exceed the cost of energy to serve it, or as a pseudo-equation, EFLOPS >> MWs.

This result suggests that there will continue to be strong incentives for hyperscalers to pursue energy supplies and that they will be relatively insensitive to energy costs in the short term.

Utilities may have a strong incentive to “ride the coattails” of data center developers in order to support the perceived need for more utility infrastructure. If the demand from hyperscalers does not materialize as planned but the utility infrastructure is built to support that demand, consumers may end up bearing costs for infrastructure that is not necessary. Careful and diligent utility regulation will be important to ensure that this effect is minimized.

Scarcity signaling and the market dynamics for AI computing capacity

As noted above, the current market for generative AI computing capacity is marked by extreme optimism, massive capital investment, media hype, and strategic signaling.

[2] Of course utilities and hyperscalers also have quite different risk profiles.

[3] Baseload in this sense means constant, dispatchable output, necessary to power high-load-factor demand sources such as data centers. Today, the most common technologies suited to power generative AI data centers are natural gas-fired combined-cycle plants and, perhaps soon, new nuclear capacity.

[4] Columbia Business School’s Milstein Center, https://business.columbia.edu/milstein-center-research-lab/milstein-center/year-data-center, March 2025

[5] We further assumed in our simple analysis that the capacity in question was a combined-cycle gas turbine with a heat rate of 6,500 BTU/kWh and an average fuel cost of $3/MMBtu.

This environment has resulted in an expensive market for GPUs, particularly for high-end chips like NVIDIA's H100 and B200, which remain elevated in price due to persistent supply constraints and surging demand. Spot market prices for H100s in mid-2025 are still significantly above the manufacturer's suggested retail price and wait times for large-scale deployments remain long. This has contributed to an artificial scarcity that further incentivizes overbuilding—hyperscalers don't necessarily need computing capacity; rather, they need to signal that they have it, in order to win API share and enterprise relationships.

The market for GPUs is mirrored to some extent by the market for electrical capacity to supply data centers. Prices for gas turbines, a common type of capacity for gas turbines, have increased to over $2,000/kWh,[6] a near-tripling from several years ago, and waiting lists for gas turbines have increased to nearly five years.

This dynamic has created a kind of "computing gold rush," where the perception of scarcity fuels even more scarcity, with the dynamic manifesting itself in both the computing capacity and energy markets. Hyperscalers and cloud providers are signaling massive demand, not just to meet current AI workloads but also to shape (optimistic) market expectations and attract customers, talent, and regulators' goodwill. This in turn leads to speculative behaviors from downstream infrastructure players, including utilities and real estate developers, who may overbuild based on optimistic forecasts, supported by aligned incentives between hyperscalers and utilities.

Several recent examples illustrate the divergence between the narrative and the fundamentals:

- OpenAI (backed by Microsoft) has scaled rapidly, but revenue per token remains modest outside of a few enterprise partnerships. Many analysts argue that much of the current investment is aimed at strategic positioning and long-term dominance rather than short-term monetization.
- Meta has committed over $30 billion to AI infrastructure in 2025 alone, despite still searching for new compelling end-user AI applications beyond LLaMA .
- AWS is expanding custom chip offerings and making aggressive moves to commoditize AI infrastructure, aiming to differentiate on integration with broader AWS services.

Market equilibrium is likely to be elusive in the short term. The marginal value of additional compute resources will vary widely across use cases and industries, and the eventual stabilization point will depend on the success of monetization strategies, regulatory clarity, and the evolution of model architectures (e.g., more efficient models could reduce the marginal cost per task).

In this context, demand for generative AI capacity behaves more like a sentiment-driven commodity than a predictable infrastructure investment. The market is driven as much by narrative and signaling as it is by actual product-market fit. Much of the capacity being added is not yet fully justified by observed demand but rather by expected future demand—and the strategic advantages of being first to meet it. This feedback loop is reinforced by capital markets, where AI-linked assets (from GPU vendors to utilities to real estate developers) are rewarded for effectively signaling growth, not necessarily for observable profitability.

[6] https://gasoutlook.com/analysis/costs-to-build-gas-plants-triple-says-ceo-of-nextera-energy/, retrieved July 15, 2025

Conclusion

The implied returns on computational capacity compared to the cost of energy and capacity indicate that power is not a significant constraint on data center growth. Data center growth is more likely to be constrained by the economic value the marketplace places on AI products.

The strong incentives for generative AI developers to gain leadership in the marketplace and for energy suppliers to gain that same “pole position” are mirror images of each other. Both communities have incentives to build quickly and to create the perception of scarcity, leading to a “gold rush” for computing capacity.

These factors may lead to a “positive feedback loop” and an arms race in data center development, with multiple parties having strong incentives to create and encourage the perception of rapid growth.

About CRA’s Antitrust & Competition Practice

CRA’s competition economists provide economic analysis and testimony in competition matters around the world. Many have served in government antitrust agencies and are members of premier academic, economic, and law faculties. Their experience extends to many industries, including health care, energy, computer hardware/software, retailing, telecommunications, aerospace and defense, entertainment, transportation, natural resources, sports, chemicals, pharmaceuticals, financial services, and consumer products.

About CRA’s Energy Practice

Charles River Associates is a leading global consulting firm that offers strategic, economic, and financial expertise to major corporations and other businesses around the world. CRA’s Energy Practice provides services to a wide range of industry clients, including utilities, ISOs, RTOs, large customers, and investors. The Energy Practice has offices in Boston, London, Los Angeles, Munich, New York City, Salt Lake City, Toronto, and Washington, DC. Learn more at **www.crai.com/energy.**

Contacts

Christopher J. Russo
Vice President
Boston
+1-617-413-1180
crusso@crai.com

Maxime Cohen
Senior Consultant
Toronto
mcohen@crai.com

CRA Charles River Associates

The conclusions set forth herein are based on independent research and publicly available material. The views expressed herein do not purport to reflect or represent the views of Charles River Associates or any of the organizations with which the author is affiliated. The author and Charles River Associates accept no duty of care or liability of any kind whatsoever to any party, and no responsibility for damages, if any, suffered by any party as a result of decisions made, or not made, or actions taken, or not taken, based on this paper. If you have questions or require further information regarding this issue of *CRA Insights*, please contact the contributor or editor at Charles River Associates. This material may be considered advertising. Detailed information about Charles River Associates, a trademark of CRA International, Inc., is available at www.crai.com .