CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

JUNE 2026

The Compute Coalition: How to Build the Future of AI in the Free World

Alasdair Phillips-Robins
Teddy Tawil
Sam Winter-Levy

Technology and International Affairs Program

The Compute Coalition: How to Build the Future of AI in the Free World

Alasdair Phillips-Robins
Teddy Tawil
Sam Winter-Levy

June 2026

About the Carnegie Endowment for International Peace

In a complex, changing, and increasingly contested world, the Carnegie Endowment generates strategic ideas, supports diplomacy, and trains the next generation of international scholar-practitioners to help countries and institutions take on the most difficult global problems and advance peace. With a global network of more than 200 scholars across twenty countries, Carnegie is renowned for its independent analysis of major global problems and understanding of regional contexts.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

See online version for full sourcing

TABLE OF CONTENTS

About the Authors

Alasdair Phillips-Robins is a fellow in the Technology and International Affairs Program at the Carnegie Endowment for International Peace, where his research focuses on emerging technology and national security. From 2023 to 2025, he served as a senior policy advisor to U.S. Secretary of Commerce Gina Raimondo, where he covered emerging technology and international issues.

Teddy Tawil is a visiting researcher in the Technology and International Affairs Program at the Carnegie Endowment for International Peace, where his research focuses on the economics and geopolitics of artificial intelligence.

Sam Winter-Levy is a fellow in the Technology and International Affairs Program at the Carnegie Endowment for International Peace, where his research covers emerging technology and national security, with a focus on the geopolitics of AI.

Acknowledgments

The authors thank Noah Tan for research assistance and Jon Bateman, Michael Brown, Janet Egan, Khushboo Goel, Noah Gordon, Lennart Heim, Corey Hinderstein, Saif Khan, Andrew Leber, Anton Leicht, Milo McBride, Amelia Michael, Konstantin Pilz, David Schneider-Joseph, Ursula Solarz, Kevin Thuot, Nikita Lalwani, and Jamie Zakalik for valuable comments on the paper and the model. Thanks as well to Amanda Branom, Alana Brase, Helena Jordheim, Jessica Katz, and Jocelyn Soly for production help. All errors are our own.

Technology and International Affairs Program

The Technology and International Affairs Program develops insights to address the governance challenges and large-scale risks of new technologies. Our experts identify actionable best practices and incentives for industry and government leaders on artificial intelligence, cyber threats, cloud security, countering influence operations, reducing the risk of biotechnologies, and ensuring global digital inclusion.

Introduction

Across the world, one of the largest peacetime industrial mobilizations in history is under way. This year alone, America's biggest technology companies will spend some $670 billion, or about 2 percent of U.S. GDP, building compute clusters. Worldwide, companies and governments will pour almost $1 trillion into data centers with a single goal: building transformative AI.

The United States currently dominates the buildout: As of May 2025, almost three-quarters of the world's advanced AI computing clusters were on American soil. U.S. projects also move faster than those in most other countries. But that lead is fragile. Domestic constraints—grid capacity, permitting rules, political opposition—are tightening. Abroad, China is mobilizing to close the gap, while Gulf states are touting energy and capital to attract developers.

Many traditional U.S. allies, meanwhile, risk being left behind. Across Europe, every major publicly reported AI data center combined appears to contain less computing power than the Amazon-Anthropic mega-cluster at New Carlisle, Indiana. Some U.S. allies, including Australia, Italy, and South Korea, have no major publicly known operational AI chip concentrations at all. And future plans, such as Australia and France's gigawatt-scale projects, will at best match what industry leaders built years earlier.

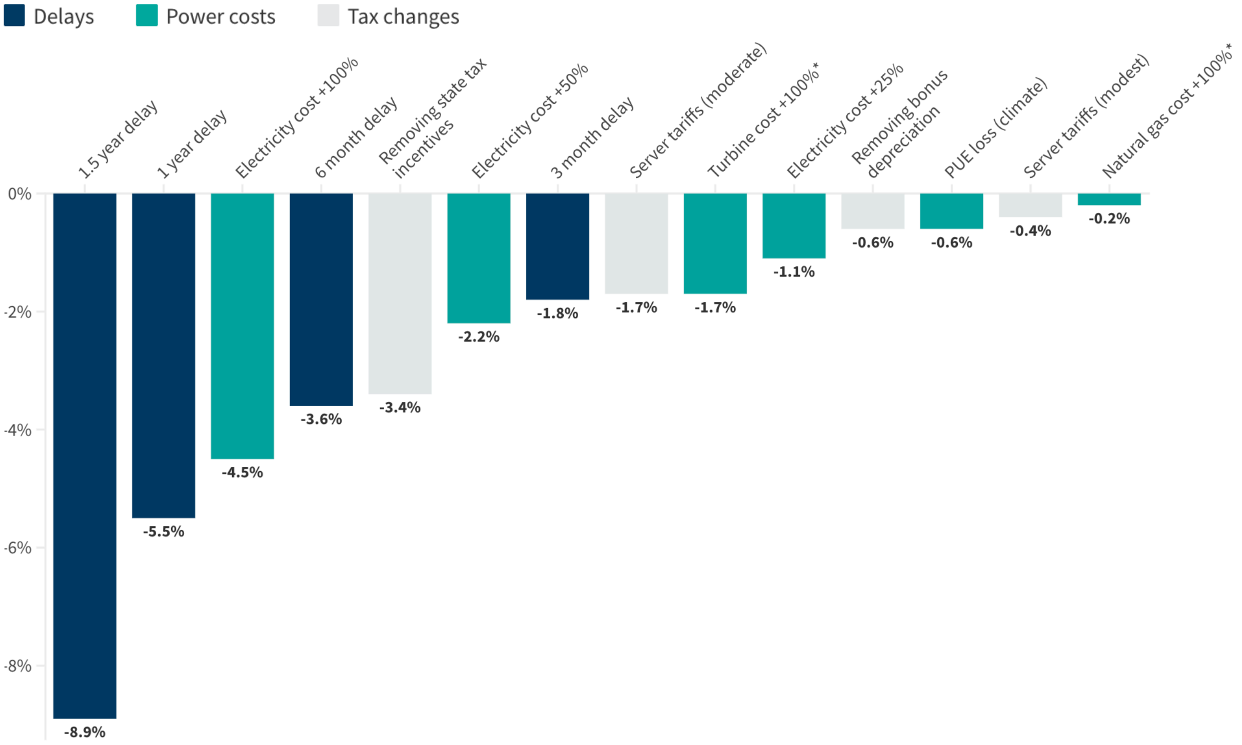
That failure is more than a question of economic returns. The countries that host the buildout will shape the future of AI: who controls it, what values it contains, how it is used. If democracies are in the lead, they will have a shot at ensuring that transformative AI is safe, secure, and reflects liberal values. If authoritarian powers pull ahead, they will gain the tools to reshape the world order in their image, through repression at home and military dominance abroad. And because authoritarian regimes are more likely to suppress dissent and hide failures, an AI race they lead raises the risk of catastrophic accidents.

The free world still has the energy, capital, talent, and industrial might to win. This paper, built on a detailed new financial model of data center competitiveness, explains what governments can do. The model considers a wide range of factors, including construction, IT, and other capital costs; operating expenses; revenues; tax systems; and operational timelines. It helps explain what is driving billion-dollar company decisions, and how governments can change them.

The core finding: time to power is what matters most. In policy circles, the conventional wisdom often emphasizes energy costs and tax incentives. Our model suggests these factors are secondary. Countries that can get data centers online quickly produce dramatically better returns than those where projects languish in permitting and grid connection queues. We estimate that a one-year delay in operation would cost an illustrative 100-megawatt U.S. data center more than $500 million over its life cycle, or more than 5 percent of its total value (see figure 1). Put differently, companies should be willing to pay at least double U.S. power prices to operate a data center just one year sooner.

Figure 1. Delays Are the Largest Driver of AI Data Center Value Erosion

Impact of changes on U.S. AI data center life cycle value, %



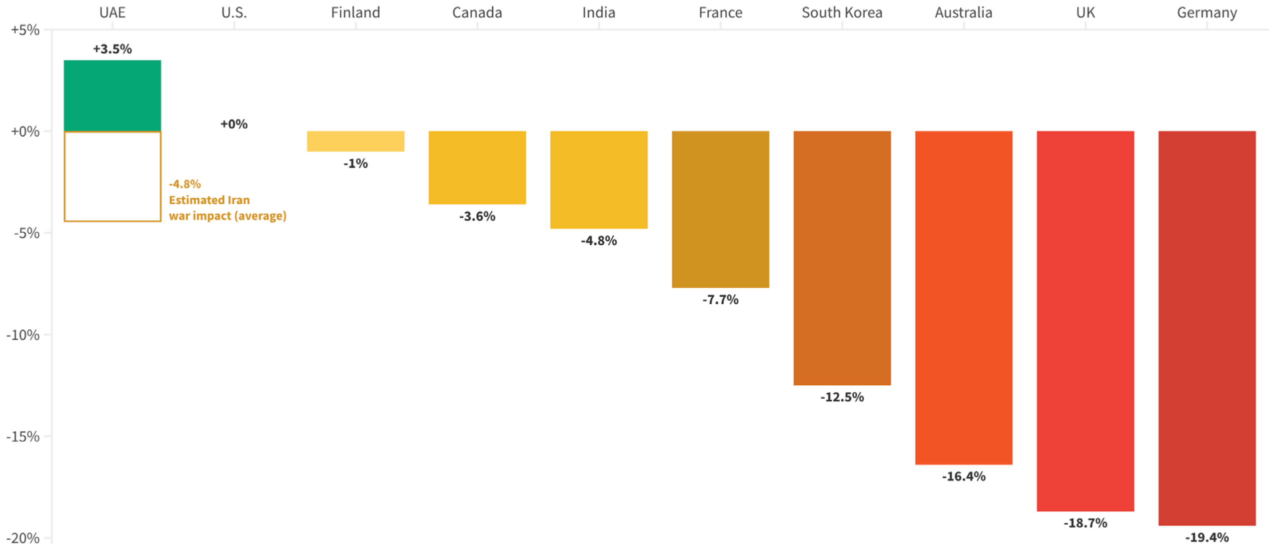
Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter “bridge” scenarios. For more information, see appendix 1.

*Behind the meter “bridge” scenario only

Source: Carnegie Endowment International AI Data Center Model

Figure 2. The U.S. Is One of the Most Economically Attractive Locations for AI Data Centers

Life cycle value of representative 100MW AI data center, % (compared to U.S.)



Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter "bridge" scenarios. For more information, see appendix 1.

Source: Carnegie Endowment International AI Data Center Model

The United States and the United Arab Emirates (UAE) top our country competitiveness rankings, in large part because major projects move fastest in those countries (see figure 2). By contrast, the country where timelines are longest, Germany, fares worst. But these rankings can shift quickly: with a one-year delay, the United States drops to fifth, behind the UAE, Finland, Canada, and India. If India sped things up by a year, it would jump ahead of Finland, Canada, and the United States, rising from fifth to second place. If war in the Middle East slows the UAE by a year and a half, it drops to fourth.

Other factors matter, but none matches the significance of speed. Even doubling electricity prices would cost a U.S. facility less than a one-year delay. Moderate tariffs on GPU servers would cost a U.S. facility about one-third as much. And removing typical state tax incentives comes in at about 60 percent of the cost of such a delay. Because IT equipment—priced similarly worldwide—accounts for the large majority of capital expenditures, national differences in non-IT factors, such as labor and land costs, have a small effect on overall project economics, and subsidies must be very large to move the needle significantly.

Our financial model focuses on what individual countries can do to compete. **But no single democracy—not even the United States—can build the world's AI infrastructure alone.** A broader coalition of democracies, one that pools its geography, power generation, capital markets, and supply chain strengths, would be stronger, more resilient, and better positioned to ensure that transformative AI is developed and governed on democratic terms.

Democracies thus need to both compete at home and cooperate abroad. This paper makes recommendations on both fronts. On the domestic level, it identifies the reforms that would most improve countries' competitiveness in attracting AI infrastructure. On the international level, it makes the case for a broader democratic coalition that can collectively outcompete authoritarian alternatives across the AI supply chain. The need is not for an overall acceleration in the AI buildout, but for a shift in relative progress to keep the center of gravity in the democratic world.

Domestically, the priority for countries that struggle to attract investment today should be removing the obstacles that slow projects down, including permitting backlogs, grid connection queues, transmission bottlenecks, and equipment shortages. Governments should not spend scarce public funds on a lucrative industry through energy subsidies, tax cuts, or other fiscal outlays. Rather, they should consider:

- Creating fast-track review processes for data center projects that meet certain criteria, such as the use of clean energy, investment in grid infrastructure, or tax commitments. As long as individual governments can actually shorten approval timelines, they will have significant flexibility to shape the bargains they strike with data center developers.

- Promoting grid flexibility and resilience by reforming interconnection queue processes, allowing operators to plan for load flexibility under the right circumstances, and encouraging the production of bottlenecked equipment.

- Supporting clean behind-the-meter power, including solar microgrids and wind farm offtake, as a bridge to grid connections.

Internationally, the democratic world needs to start building the partnerships that will let it collectively dominate the AI value chain from natural resources to last-mile deployment. Fruitful collaboration could include mutual fast tracks for allied infrastructure investment; common industrial policy in sections of the supply chain, such as critical minerals, where chokepoints lie outside the democratic world; joint safety and security research on frontier AI capabilities and safeguards; and common transparency and incident reporting standards. Together, democracies can set the pace in the development and security of powerful AI systems.

This paper proceeds as follows. Part I explains the economics of data centers: it lays out the findings from our model, shows that time to power is a key driver of companies' decisions about where to site data centers, and examines what drives differences in competitiveness between countries. Part II explains the geopolitical stakes. Those siting decisions will help determine the future of AI and the balance of global power. The paper thus makes the case for a democratic compute coalition, arguing that the United States and its partners each have strong incentives to cooperate. Part III outlines the policy agenda: domestic reforms to reduce time to power and international coordination to secure the AI supply chain.

I. The Economics of Compute

Companies are scouring the earth for land to build AI chip clusters and energy to run them. Data center economics show that the places getting chips online the fastest have a decisive advantage.

The Model

In June 2025, two major AI infrastructure projects hit milestones. In Abilene, Texas, OpenAI's Stargate data center took delivery of its first Nvidia chips. On the other side of the Atlantic, in Brussels, the European Commission received seventy-six nonbinding "expressions of interest" for its AI Gigafactories initiative—an early step in its plan to support large computing clusters across the bloc.

By December, Stargate Abilene was running an estimated 100,000 Nvidia GB200 chips, or more than 200 MW of AI computing capacity. The commission, meanwhile, was still preparing to issue its formal call for proposals—a step that has since been delayed. More than a year after the program was announced, no construction has begun.[1]

OpenAI is not the only American firm forging ahead while European leaders issue paperwork. A single American company, Google, holds roughly one-quarter of global AI computing power. In 2025, the United States hosted around 75 percent of the world's high performance AI compute. The European Union has less than 5 percent.

To understand how governments might improve their lot, we built a financial model to explain the major drivers of investment decisions. The model projects the life cycle value of a hypothetical 100 MW AI data center—large, but no longer frontier-training scale[2]—in ten countries, representing a range of players in the infrastructure race: Australia, Canada, Finland, France, Germany, India, South Korea, the United Arab Emirates, the United Kingdom, and the United States.

The model considers capital expenditures (IT equipment, construction, electrical systems, liquid cooling, on-site power generation), operating expenses (electricity, maintenance, staffing), tax treatment (corporate taxes, depreciation schedules, state and local incentives), time to operation, and projected revenues over a twelve-year life cycle, drawing on public data, equity research, and industry interviews.

The model is designed to be realistic, but its conclusions are necessarily tentative. Projecting the returns to investments of this complexity requires making many contestable assumptions, and any given real-world project will differ in multiple ways. The model is built primarily on public data, while much of the relevant information is proprietary. The model's value lies more in the trends it reveals than any specific figure. A more complete explanation of our methods is in the appendices.

Three primary findings are relevant to policymakers.

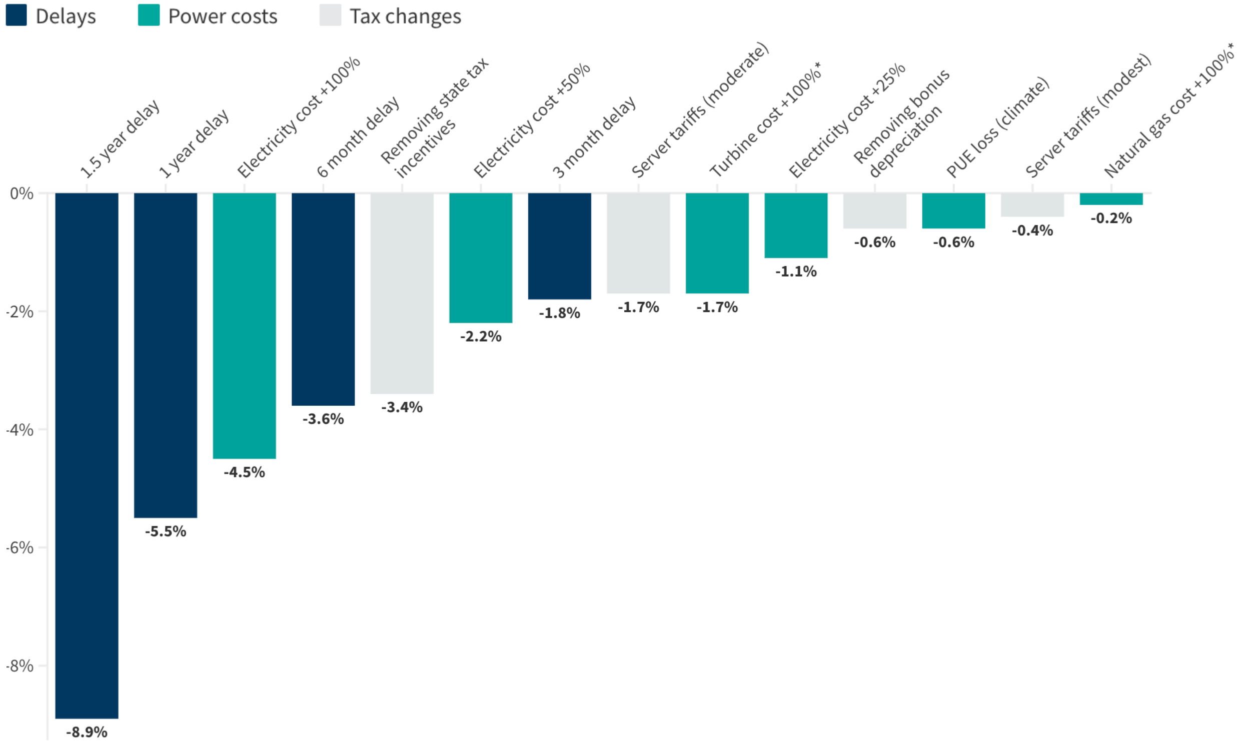
Finding 1: Time to power is the most important driver of economic returns.

Delays in getting a data center up and running are extraordinarily costly. Our model shows that for a typical 100 MW U.S. data center, each additional year of delay costs roughly $550 million in life cycle value, about 5.5 percent of the data center's roughly $10 billion life cycle value (see figure 3).[3]

Of all the external variables we considered—construction and labor costs, tariffs, local taxes, power costs, efficiency losses from the climate—delays are the number one driver of lost returns in our model, by a large margin.

Figure 3. A One-year Delay Reduces AI Data Center Value More Than Doubling Electricity Costs

Impact of changes on U.S. AI data center life cycle value, %



Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter "bridge" scenarios. For more information, see appendix 1.

*Behind the meter "bridge" scenario only

Source: Carnegie Endowment International AI Data Center Model

In the United States, our model suggests that a one-year delay costs about 25 percent more than a doubling in electricity prices from their current level ($441 million), about 60 percent more than losing typical state tax incentives ($338 million), and three times as much as moderate tariffs on servers ($172 million). Even the cost of a three-month delay—about $178 million—greatly exceeds the impact of removing upfront bonus tax depreciation ($61 million) or modest tariffs ($43 million) (see table 1).[4]

Table 1. Delays in Time-to-Power Are the Largest Driver of AI Data Center Value Erosion

Impact of changes on U.S. AI data center life cycle value, $M

Change	Value Loss ($M)
Delays	
1.5 year delay	-$877
1 year delay	-$543
6 month delay	-$353
3 month delay	-$178
Power costs	
Electricity cost +100%	-$441
Electricity cost +50%	-$221
Turbine cost +100%*	-$168
Electricity cost +25%	-$110
PUE loss (climate)	-$60
Natural gas cost +100%*	-$20
Tax changes	
Removing state tax incentives	-$338
Server tariffs (moderate)	-$172
Removing bonus depreciation	-$61
Server tariffs (modest)	-$43

Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter "bridge" scenarios. For more information, see appendix 1.

*Behind the meter "bridge" scenario only

Source: Carnegie Endowment International AI Data Center Model

Why Do Delays Matter So Much?

Three factors dominate:

Delayed revenue. Every month a data center is delayed is a month of GPU-hours that could have been sold. At current prices and utilization rates, a 100 MW Blackwell facility generates nearly $200 million in monthly revenue as it starts operating at full capacity.[5] Delays push those revenues into the future, and at any reasonable discount rate, earlier cash flows are worth substantially more than later ones.

Idle capital. Capital allocated to a planned data center earns nothing. Today, little funding may be tied up in some delayed projects, as the most expensive elements of a data center, the chips, are typically delivered at the end of the construction process. Even so, some capital will be committed in construction or labor before chips arrive. And for projects financed through debt or other mechanisms, funding often has to be secured in advance, meaning that delays, which lead to compounding interest, can cause deals to be renegotiated or to fall apart altogether.[6]

In the future, these dynamics may worsen as operators lock in longer-term contracts with chip designers or produce chips in house. Delivery schedules may become fixed in advance, in which case delays could cause chips to sit in warehouses if power isn't available to run them.

Compounding effects. The faster companies can bring data centers online, the more computational power they can devote to training their AI models and researching future breakthroughs. When the data center owner or user is a model developer, this confers a critical advantage in the AI race. Stronger models mean more customers, more revenue, and better internal tools to help produce the next generation of models. As end-users move from pilots to wider implementation of AI systems, leading in AI over the next few years could prove immensely profitable.

The result is that time swamps every other factor. For the world's latest and largest data centers, getting online months earlier is worth billions of dollars. Policymakers focused on energy prices or construction costs are missing a crucial part of the picture. The decisive variable is speed.

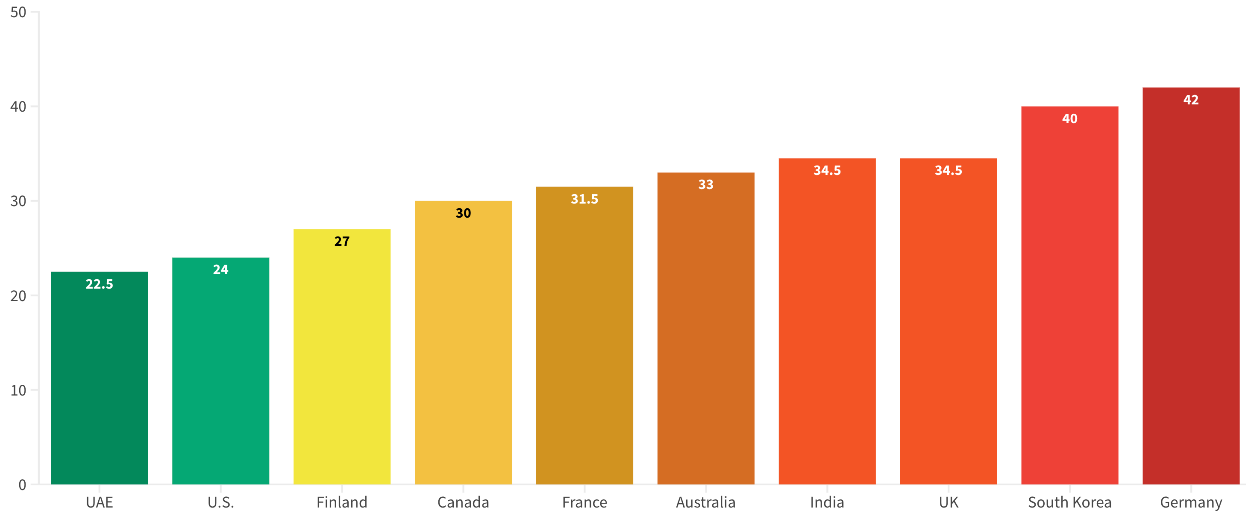
Where Delays Are Worst

Delays are widespread, but our research finds substantial variation across countries in time to first operation—how long it takes to bring a data center from project approval to running chips (see figure 4).[7]

Each estimate comes with high uncertainty, as projects vary within each country and each estimate accounts for both projects that connect to the grid and those that use behind-the-meter power. But on average—before accounting for possible delays from the war in Iran—the UAE leads at around twenty-two months.[8] The United States follows at twenty-four

Figure 4. Time-to-Power for AI Data Center Projects Varies Almost 2x Across Countries

Time from project approval to first operation, months (estimate for representative AI data center)



Note: "Project approval" refers to the point at which a data center project has secured all major regulatory permits and confirmed how it will secure power, allowing construction to begin. The estimate includes power interconnection wait times. Based on completed and announced projects, supplemented by additional research. Announced timelines were treated as indicative rather than definitive due to the possibility of delays, and actual timelines may vary widely. Estimates represent the average of two scenarios: a behind-the-meter solution using simple-cycle turbines as a temporary bridge, and grid interconnection. Assumes a reasonably favorable regulatory environment within each country (e.g., Texas rather than Massachusetts in the United States). Excludes early-stage site selection, permitting, and planning, which are difficult to measure consistently and are often expedited for large projects or handled by specialized pre-development firms. Also excludes ramp time to full capacity, estimated at roughly one year for a 100MW hyperscale data center.

Source: Carnegie Endowment International AI Data Center Model

months—competitive, but largely in favorable regulatory environments like Texas, not California.[9] Most countries fall in the two-to-three-year range, some aided by government policies to expedite approvals. France, for example, has identified a number of turnkey sites for large-scale AI data centers. Typical projects in the United Kingdom take an estimated thirty-four months, with notable differences in grid connection times between Scotland, where timelines are roughly three years, and the rest of the country, where they can stretch to five or more. (The model's estimated timelines are faster than typical grid timelines, as they also consider more rapid behind-the-meter approaches.) South Korea and Germany round out the pack at forty and forty-two months respectively. The German grid is one of Europe's most congested—some developers are reportedly turning to Belgium because of the backlog—and South Korean grid improvements are often delayed.[10]

These figures, which are derived from completed and announced projects, represent roughly median estimates for time to power for a major data center operator in each country. They are supplemented by additional research on grid and regulatory environments. They exclude initial pre-construction land permitting and planning (which are typically nonpublic, difficult to measure consistently across countries, and often contracted to third parties).[11] We discuss additional methodology details in appendix 2. Actual timelines vary widely by site, jurisdiction, and execution. But while specific projects will move faster or slower than these figures, the relative ordering is meaningful.

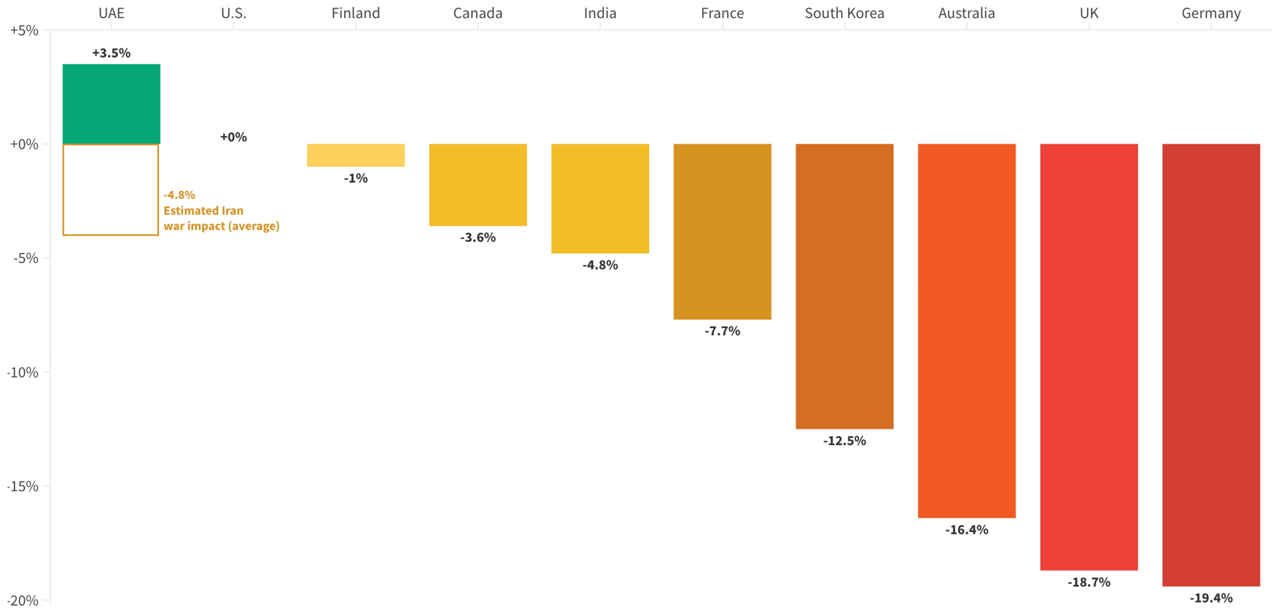
The gap between the fastest and slowest is striking: Our data suggests that the same data center could be built almost eighteen months faster in the United States than in Germany.[12] And even among democracies, the spread is substantial. OpenAI energized the first phase of its Stargate Abilene data center in Texas, with 147 MW of power, in thirteen months. In contrast, it took the American company Colt more than four times as long (six years) to begin operating a smaller data center (100 MW) in Mumbai, India, after purchasing the land.

Finding 2: The United States and the UAE are the most competitive of the countries studied, while Germany faces the biggest hurdles. Changes in timelines reshuffle the relative rankings dramatically.

Because time matters so much, small changes in project timelines can reshuffle country rankings. The model's estimates of relative returns closely track the speed at which each country can get projects up and running (see figure 5). As a result, better policy has the potential to vault countries that currently perform poorly up the table.

Figure 5. The U.S. Is One of the Most Economically Attractive Locations for AI Data Centers

Life cycle value of representative 100MW AI data center, % (compared to U.S.)



Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter "bridge" scenarios. For more information, see appendix 1.

Source: Carnegie Endowment International AI Data Center Model

Table 2. Power and Permitting Delays Reduce U.S. AI Data Center Competitiveness

Rank reflects modeled competitiveness across ten countries (U.S., UAE, India, Canada, Finland, France, South Korea, Australia, UK, and Germany) under different project delay scenarios

Scenario	Rank (of 10)
U.S. with 9 month speedup	1
Default scenario	2
U.S. with additional 1 year delay	5

Note: 100MW hyperscaler greenfield AI-optimized data center, 12 year life cycle plus terminal value, GB200 NVL72, financed by internal FCF. Average of grid interconnect and behind the meter scenarios.

Source: Carnegie Endowment International AI Data Center Model

In our default scenario, the United States ranks second among the ten countries studied in terms of data center life cycle value. But that ranking is vulnerable. If the typical project speeds up by nine months, the United States overtakes the UAE as the most competitive site for AI data centers (see table 2). Add a one-year delay and the United States falls to fifth—behind the UAE, Finland, India, and Canada.[13]

This sensitivity explains why companies are so focused on time to power. As one Wall Street analyst told us, "power rules everything right now. If you have power that you can plug a GPU into, you're going to get demand, whatever the power costs."[14] An executive choosing between building in Texas or India is generally less focused on electricity prices than on how many months sooner the Texas facility can start generating revenue. A few months' difference can be worth hundreds of millions of dollars for a 100 MW data center.

As a result, the UAE's faster general timeline translates into a substantial valuation premium. If India could build as quickly as the UAE, our model suggests it would overtake the United States, Finland, and Canada, despite the cheap power and the cool climates in the latter two countries, which increase data center efficiency.

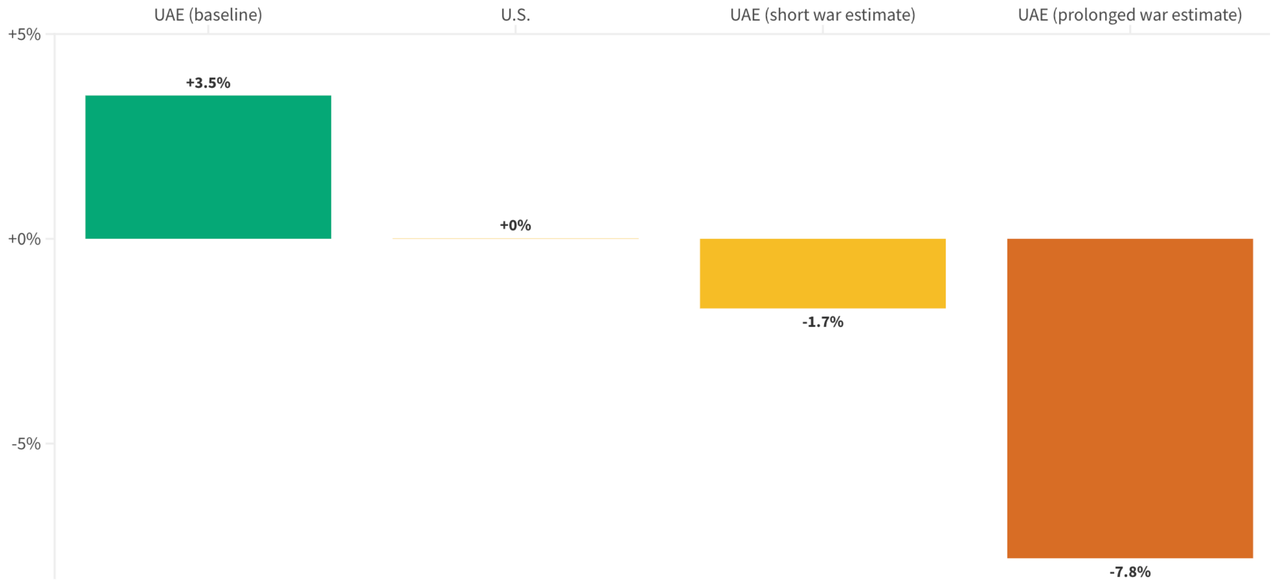
While the UAE leads in our results, the baseline scenario does not consider the impact of the recent U.S.-Iranian war and ongoing security risks in the region. The war may raise the difficulty of attracting the skilled workforce necessary for large projects—power grid engineers; construction managers; and civil, mechanical, and electrical engineers. Some data center developers in the Middle East have changed in-person work requirements due to safety risks, and one workforce analyst suggests that salaries may include additional hazard pay going forward. The war has also disrupted key Emirati ports like Jebel Ali and Fujairah, making it more difficult to receive raw materials needed for AI data center construction.

The war also poses financial risks: mounting capital expenditures for physical data center fortifications, the need for redundancy, tighter constraints on securing capital, and higher insurance costs. Many traditional commercial insurance policies do not cover risks from armed conflict, such as an AWS data center being struck by an Iranian Shahed drone. According to one estimate, the cost of Gulf "war insurance" surged nineteen-fold to $5 million for $100 million in coverage after the outbreak of hostilities, with insurers offering reduced coverage limits.[15] While some macroeconomic measures of risk have stabilized since the outbreak of hostilities, the Iranian military's public threats against AI infrastructure, including OpenAI's Stargate UAE data center, cast a long shadow over planned AI expansions in the Middle East.

Although our baseline scenario does not factor in the impact of the war, we consider here some additional scenarios to illustrate how the conflict might affect projects in the Gulf. The resulting estimates are highly uncertain, as good data on the war's effects is hard to come by, but they give a general guide to the potential impact of the conflict. The upshot is

Figure 6. The War in Iran Reduces the UAE's Competitiveness

Life cycle value of AI data center in the UAE compared to U.S. under different illustrative scenarios, %



Note: 100MW hyperscaler greenfield AI-optimized data center, 12-year life cycle plus terminal value, GB200 NVL72, financed by internal free cashflow. Average of grid interconnect and behind the meter scenarios. Short war estimate involves lower war insurance costs (annual premium of .5% of CapEx), 10% uplift to non-IT CapEx due to shipping disruptions, and a 6 month delay. Prolonged war estimate involves higher insurance costs (annual premium of 1% of CapEx), 20% uplift to non-IT CapEx due to shipping disruptions, and 1-year delay.

Source: Carnegie Endowment International AI Data Center Model

that a short war will likely do only limited damage to the UAE's competitiveness, but if the conflict drags on, things will get much worse. In the first scenario, we assume the war delays construction timelines by six months, shipping disruptions increase non-IT construction costs by 10 percent, and the annual premium for war insurance is 0.5 percent of capital expenditures. Those changes push overall economic returns in the UAE to about 1.7 percent below those in the United States. In a prolonged war scenario, where conflict delays construction by a year, non-IT construction costs rise 20 percent, and insurers charge 1 percent of capital expenditures in annual war risk premiums, the UAE would finish in sixth place, about 7.8 percent behind the United States (see figure 6).

Finding 3: Tariffs, electricity prices, tax incentives, and non-IT costs also matter.

Time, important as it is, isn't the only factor. Many projects are being planned and built in places with longer than average timelines. In projecting financial returns, several other factors are also important.

Tariffs. Moderate tariffs on GPU servers—a 20 percent tariff on Taiwan-assembled servers with 60 percent pass-through to buyers—would cost $172 million in life cycle value for a U.S. data center. Even modest tariffs (5 percent with the same pass-through) would cost $43 million. With trade tensions ongoing and semiconductor supply chains under scrutiny, tariff scenarios are plausible enough to take seriously.

Electricity prices. A 50 percent increase in U.S. electricity costs would reduce life cycle value by $221 million.[16] This matters, but less than many assume, in part because electricity is a smaller share of total costs than often believed. In the United States, electricity costs are generally less than 10 percent of annual revenue for AI data centers, and less than 5 percent under the baseline assumptions of our model (see figure 7). Even with more aggressive assumptions about power costs, the investment firm New Street Research reckons that depreciation and amortization of IT equipment (chips and servers, which are replaced every three to six years) is about seven times as costly as electricity on an annual basis.

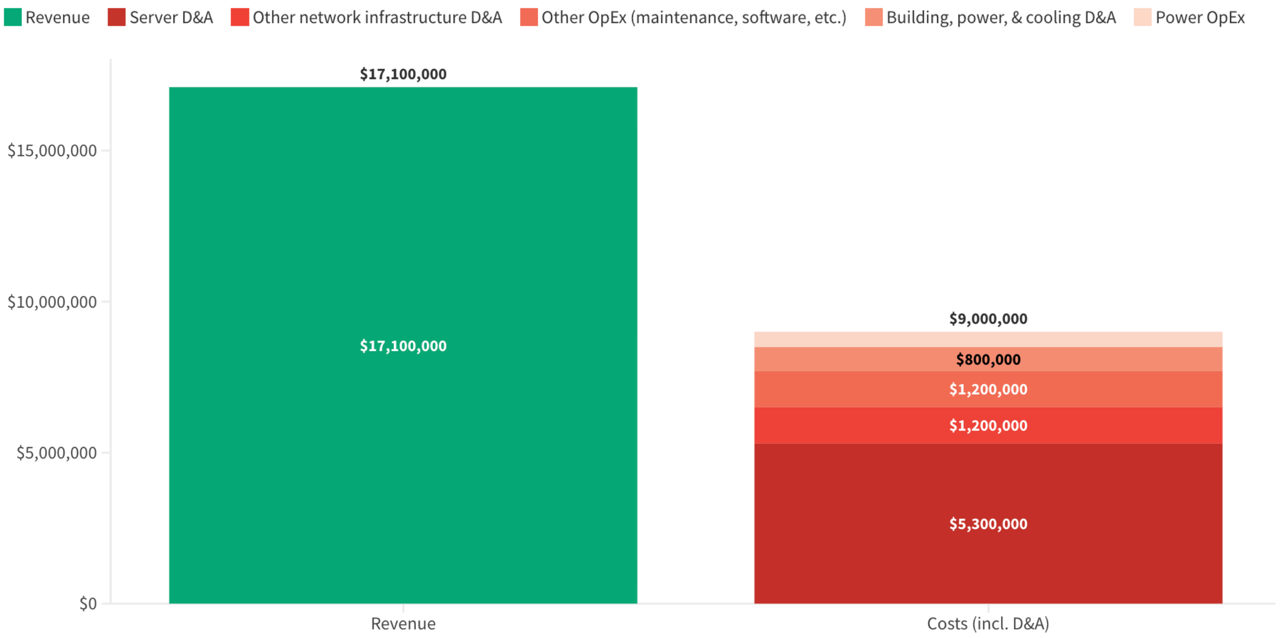
Tax treatment. Removing estimated U.S. state AI data center tax incentives costs around $338 million. Removing 100 percent bonus depreciation for equipment—the ability to deduct equipment costs upfront rather than over several years, which was made permanent under the One Big Beautiful Bill Act—costs $61 million.

> **The hierarchy is clear: delays have the largest impact, then electricity, then taxes and server tariffs, then other factors.**

Other factors. Doubling turbine capital expenditures (relevant for behind-the-meter scenarios) costs $168 million. Higher cooling costs in hot climates, captured in the model as PUE (power usage effectiveness) losses, cost $60 million compared to a cool-climate baseline like Finland.

Figure 7. For U.S. Blackwell AI Data Centers, Power Costs Are Generally <10% of Annual Revenue

Per MW annual operating economics for grid-connected U.S. Blackwell AI data center



Note: D&A, or Depreciation & Amortization, spreads the cost of assets over their useful lives. For this graph, we assume a 4-year useful life for GPUs and other IT equipment, along with a 15-year useful life for building, power, & cooling. In the U.S., we estimate all-in power costs at 7.5c/kWh. In other locations where electricity is more costly, it comprises a greater share of expenses. We rely on Epoch AI's estimate of IT costs—if they are greater, as they are in some other sources, server and network infrastructure D&A rises accordingly. 100MW hyperscaler greenfield AI-optimized data center, GB200/NVL72 spec, U.S. grid interconnect scenario (onsite/BTM power costs not considered).

Source: Carnegie Endowment International AI Data Center Model

The hierarchy is clear: delays have the largest impact, then electricity, then taxes and server tariffs, then other factors.[17] Policymakers working on data center competitiveness should allocate their attention accordingly.

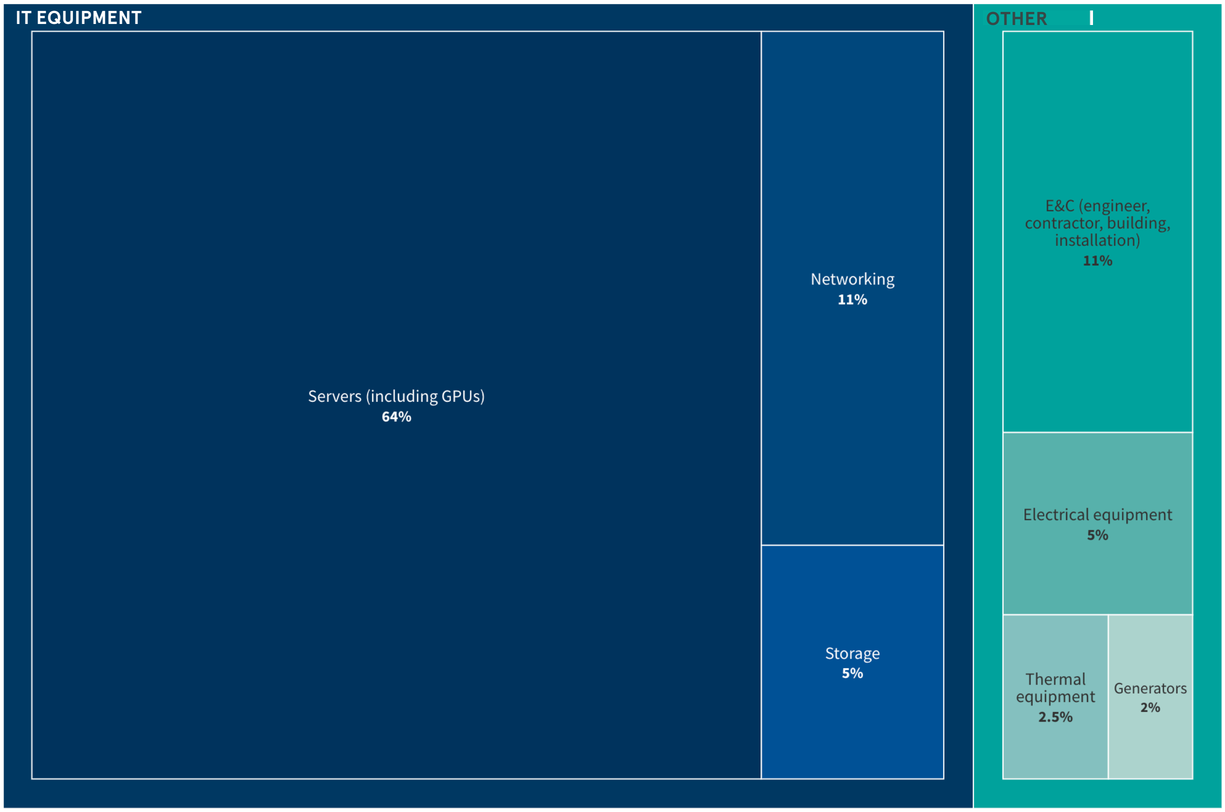
Why Non-IT Costs Matter Less Than You'd Think

One reason time dominates is that most data center capital expenditure goes to IT equipment—primarily GPUs—which is priced similarly around the world, rather than to construction, electrical systems, or cooling, where costs differ more across countries (see figure 8).

According to research by Bank of America, IT equipment accounts for about 80 percent of capital expenditures in current data centers. Other upfront costs—which include engineering, construction, electrical systems, thermal equipment, and generators—account for only about 20 percent.[18]

Figure 8. GPUs and IT Equipment Dominate AI Data Center Capital Costs

Share of AI data center capital expenditures (CapEx) by component



Note: Current architecture, CapEx only. Sums are not exact due to rounding. For future-state data centers based on Nvidia's forthcoming Rubin chip, the share of CapEx spent on IT equipment is projected to increase to 84%.

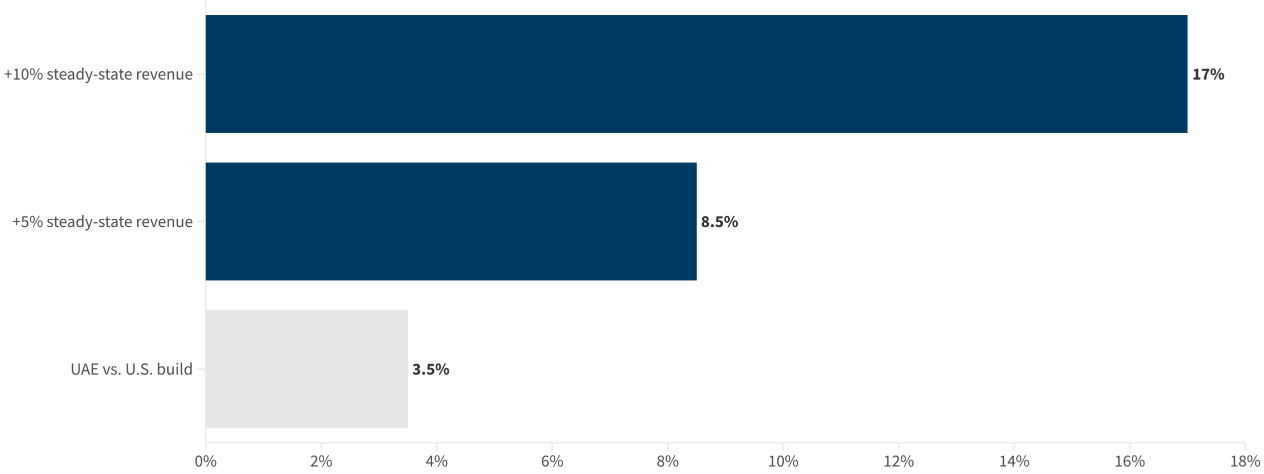
Source: Bank of America. Our model uses a similar breakdown by Epoch AI, which has IT equipment at around 70% of total capital expenditures.

This has an important implication: factors that affect non-IT costs (construction labor, permitting fees, land prices) have a relatively small impact on overall project economics. For example, despite substantially lower construction and labor costs, India's slower timelines cause it to finish behind the United States.

Countries can differentiate themselves primarily through factors that affect time (permitting, grid connections), operating costs (electricity prices, taxes), or demand (market access, data localization requirements). Of the supply side factors, time is the most important lever policymakers can pull.

Figure 9. Revenue Upside Drives More AI Data Center Value Than Location Savings

Illustrative life cycle value uplift vs. base case, 100MW Blackwell data center



Note: Base case assumes a 100MW hyperscaler greenfield AI data center using GB200/NVL72, with initial steady-state monthly revenue based on a $10 per GB200-hour price, 70% utilization, 1.15 PUE in the U.S., and a 90% accelerator server share. Valuation assumes a 12-year asset life plus terminal value (8x average EBITDA), financed through internal free cash flow. U.S. estimates reflect the average of grid-interconnect and behind-the-meter scenarios. Revenue upside also increases terminal value. To see why 1 percent revenue uplift can increase life cycle value more than 1 percent, consider the following stylized example. Suppose that a business has revenue of $100 and costs of $90, making a profit of $10. A 1 percent increase in revenue, from $100 to $101, would increase profit 10 percent, from $10 to $11.

Source: Carnegie Endowment International AI Data Center Model

Although our model focuses on supply-side factors, demand is the most important ultimate driver of returns (see figure 9). A 5 percent increase in steady-state revenue—reflecting stronger AI demand or better pricing power—boosts life cycle value by about 8.5 percent, vastly exceeding the UAE's 3.5 percent location advantage over the United States.[19] Policymakers should remember that whatever levers they can pull to enable data center development, the ultimate success of the industry will depend on customer demand, and they should pause before offering subsidies for already well-financed projects.[20]

It is also worth noting that revenue expectations in turn affect the relative importance of different policy options. When data centers are highly profitable, tax rates matter more, as there are more profits to tax. When data centers are less profitable, operating margins matter more, increasing the importance of power costs.[21] If profits are high enough, all costs become less significant, and fast access to a large power supply dominates site selection.

Understanding Time to Power

The importance of time to power makes it worth understanding what drives delays. The biggest problem for most operators is hooking up their projects to the grid. Here, much of the blame falls on regulatory delays and equipment shortages. Some companies are turning to self-help: setting up their own sources of power on site. This can speed things up but is at best a partial solution.

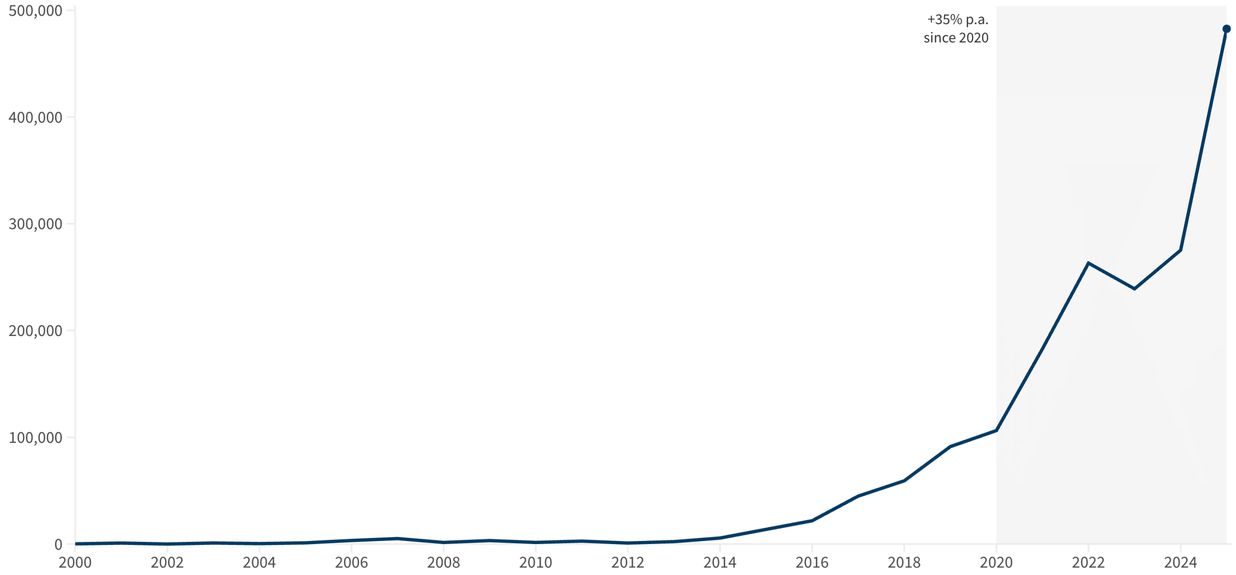
The Grid Connection Problem

Most operators want to connect to the grid because large AI data centers consume enormous amounts of power. One gigawatt is the entire output of a typical nuclear power plant; a 5 GW cluster will use more power than London on average.[22] Yet most grids simply don't have that kind of spare capacity and aren't adding it fast enough to keep up with demand.

That is because, in many countries, getting new power sources online now takes far longer than it used to. In the United States, for example, it took an average of five years to connect a new source to the grid in 2023, up from less than two years between 2000 and 2007. In 2024, new generation projects bringing more than 20 MW to the grid faced a median wait time of thirty months just to get their request approved; beginning commercial operations took another thirty-five months, up from just eleven months for projects completed in 2007.

Figure 10. Since 2020, U.S. Interconnection Demand Has Exploded

Capacity waiting in U.S. interconnection queue over time by date enqueued, MW (2000–2025)



Source: Interconnection.FYI

This is in part because the nature of electricity generation has changed. Renewable energy projects—which tend to be smaller and more numerous than traditional power plants—have overwhelmed the system for connecting new power sources to the grid, a process designed for a different era.

Equipment shortages are another problem. Electricity flows most efficiently over long distances at much higher voltages than consumers can use. To switch electrical currents between the two voltage levels, the grid uses electrical transformers. These pieces of equipment are essential to extending the grid, but in recent years they have been in short supply. Wait times for large transformers—most of which are imported—now average more than two years in North America and can stretch up to five, up from one year before the COVID-19 pandemic. Large transformers also need to be custom-made, and a shortage of skilled labor, a cyclical industry, and tariff uncertainty (many critical parts are made in China) have further increased lead times.[23]

The problem is even worse in some other countries. In India, one researcher writes that "typical interstate lines take 30 to 36 months to complete," even before accounting for right-of-way and land acquisition disputes that "routinely add years to the process." In South Korea, more than half of power upgrade projects have been delayed in recent years. The East Coast–Capital Region transmission line project, for example, has been pushed back eight years from its intended completion date in 2019. Providers in the United Kingdom have seen wait times of ten years or more to add power to the grid, although the government has introduced recent reforms in the hope of improving things.

Even when the grid has capacity to support significant new demand, the electricity still has to get to the company that wants to use it. For large new projects outside population centers—most big data centers—that means upgrading or installing new transmission lines. These are specialized pieces of equipment, and companies often face long wait times. From 2017 to 2021, the United States built an average of 700 miles of new transmission lines, down from an average of 2,000 miles between 2012 and 2016.

Such delays contrast with countries like the UAE, where officials can direct state-owned utilities to prioritize AI infrastructure. In most other countries, utilities must work through complex regulatory processes, often including lengthy grid stability studies, interconnection queues, and competing demands. The UAE's tight coordination across utilities allows fast addition of power infrastructure, in contrast to the United States, where such projects involve multiple utilities, regional grid operators, regulators, landowners, and local governments.[24]

Table 3. Data Center Grid Connection Wait Times Are up to 10 Years

Reported connection queues for new data centers in selected jurisdictions

Jurisdiction	Average time in queue
Netherlands	Up to 10 years
Germany	Up to 7 years
Northern Virginia (United States)	Up to 7 years
United Kingdom	5-7 years
Kanto (Japan)	More than 5 years
Spain	3-5 years
California (United States)	3 years
Malaysia	Under 3 years
Italy	Under 3 years
Queensland (Australia)	More than 2 years
United States	1-3 years

Note: These estimates span a different group of geographies than those in our model. Among the countries covered by both, the estimates are broadly consistent, with the IEA's tending to be slightly higher.

Source: International Energy Agency, local reporting

Behind-the-Meter Gas: A Partial Solution

Facing years-long grid connection waits, many data center operators have turned to behind-the-meter (BTM) power—generating electricity on-site rather than drawing from the grid.[25] This usually involves setting up clusters of natural gas turbines or reciprocating engines. These approaches bypass long grid queues, but they come with downsides, too: onsite power costs more, struggles to support the largest loads, and isn't a permanent solution for most projects.[26]

The most common approach today is to rely on natural gas. xAI's Colossus facility in Tennessee, for example, uses roughly 460 MW worth of gas turbines.[27] OpenAI's Stargate project in Texas relies on gas turbines for some of its initial phase. Meta's planned facilities include significant on-site generation capacity, including 400 MW in Ohio.

BTM configurations can cut time to first operation by a year or more. In the United States, it typically takes one to two years to procure BTM power. The process isn't instant, since much of the available equipment has been back-ordered, but it still shaves significant time off grid connections.[28] Although the United States has led in the deployment of BTM power, the approach is gaining traction elsewhere.

Despite their rapidity, BTM turbines and other sources of on-site power have several downsides that lead industry insiders to call them a "temporary solution at best." These include the following:

Supply constraints. As more operators turn to BTM, equipment is getting harder to find. Industry analysts told us that delivery timelines for large turbines (those with roughly 50 MW capacity or greater) have stretched to two to three years or more. GE Vernova, one of the largest turbine manufacturers, says it has sold out its inventory through 2028. Smaller turbines remain available on shorter timelines, but even those supplies are tightening.

Higher costs. On-site generation is more expensive per kilowatt-hour than grid electricity in most markets. Fuel must be delivered; equipment must be maintained; economies of scale are harder to achieve. And reliability requirements are high for AI data centers, necessitating building overcapacity and using batteries in many cases. Our model shows that BTM turbines have higher operating costs than grid-connected scenarios, partially offsetting the time advantage. The cost is worth it to shorten timelines, but operators will likely save money if they can switch projects to grid power down the line.

Environmental concerns. Gas turbines produce emissions—both carbon dioxide and local pollutants like nitrogen oxides. Google, for example, is building a gas-fired plant to power a data center in Texas that could emit more carbon dioxide by itself each year than the city of San Francisco. In addition to the direct damage it does to human health and the environment, pollution also creates political and regulatory risks, particularly in jurisdictions with aggressive climate goals. Onsite power may be preferable to keeping coal plants online, but engines and turbines still typically involve burning fossil fuels. The hyperscalers have set their own emissions reductions targets, as well, although those targets look shakier given the scale of the companies' AI ambitions.

Scale limitations. BTM works well for facilities in the hundreds of megawatts. But the largest planned data centers will draw gigawatts of power, the equivalent of a medium-sized city. Powering the largest data centers with on-site gas turbines will be difficult, thanks to reliability issues, contractors' lack of experience with BTM power at this scale, and regulatory uncertainty. At some point, grid connection will be preferable for most projects.

As a result, most operators view BTM as a bridge while they wait for grid connections or new generation capacity to come online, not as a permanent power strategy. The long-term economics favor switching to grid power in most markets.

Despite these limitations, BTM has been essential to maintaining U.S. competitiveness in AI infrastructure. With fifty states, the United States has many different state regulations on BTM power, many of them favorable toward data center developers. Without the ability to bypass grid queues, American data center timelines would look more like Germany's.

II. The Geopolitics of Compute

Compute is essential to transformative AI. The countries that host compute can tax it, control who has access to it, and regulate what they do with it. If democracies lead the development of AI, they can shape it for the better.

Technological revolutions have always elevated new resources to strategic importance. The internal combustion engine turned oil into a defining resource of the twentieth century; railroads did the same for steel and coal a generation earlier.

AI is producing a similar shift. AI systems will soon drive economic growth, create new military capabilities, and accelerate scientific and industrial research. Countries that do not adopt and deploy them at scale risk seeing their economic output, institutional capacity, and geopolitical leverage atrophy compared to those that do.

Why Domestic Compute Matters

AI is built on talent, data, and computing power, but of these, compute is the input most amenable to deliberate policy action. It is the fastest to amass: A data center begins generating value the moment it comes online, while talent pipelines and data advantages take years to build. It is a bounded physical resource, so its location and ownership can be tracked and controlled in ways that data cannot. And it sits at a chokepoint of concentrated suppliers—ASML, TSMC, Nvidia—where policy interventions can shape who has access. As researchers at Tsinghua University have put it, chips are “the one and only physical basis” for AI development. “No chip, no AI.”

Countries that host AI data centers capture direct economic value: construction spending, operational jobs, tax revenue—the last of which may be essential to public finances if AI reshapes economic activity. But equally significant is what compute enables downstream. Firms with low-latency access to abundant compute can deploy AI systems that boost productivity across sectors, from drug discovery to logistics to software development. The most compute-intensive work—training models, running large-scale simulations, accelerating scientific discovery—requires infrastructure that only a handful of countries can currently provide, and hosting it may help attract the researchers and engineers who push the field forward. Over time, this can produce a virtuous cycle: infrastructure investment attracts human capital, which generates breakthroughs, which justifies further investment.

Of course, wherever the compute is located, it still matters who owns it. A data center built by Microsoft in Kenya or by OpenAI’s cloud partners in the United Kingdom is still largely foreign-operated infrastructure. The host country does not automatically control the models trained there, and the underlying hardware, such as advanced chips and networking equipment, will remain reliant on foreign supply chains. AI systems also require frequent updates to remain competitive, meaning any version running on a local data center will fall behind without ongoing access to the provider.

Still, physical presence has advantages that remote access does not. A government that hosts compute on its soil can require cloud providers to work with local partners, enforce data residency rules, tax the data center's operators, and encourage knowledge transfer. It is easier to regulate, inspect, and negotiate with an entity that has sunk capital into your jurisdiction than one serving you remotely from another continent.

These dynamics play out across several domains.

Military and intelligence. A country that can train and run defense-relevant AI models on home soil retains control over the models, the data they process, and the operational security of the entire pipeline. Without domestic compute, a state's military and intelligence capabilities could become dependent on foreign providers. In a crisis, access might be throttled or cut when it is needed most.[29]

Governance. States with significant domestic compute are far better positioned to shape international standards on AI safety, military use, and other dimensions of governance than ones that depend entirely on others for access to the underlying infrastructure. While the countries that are home to the most advanced model developers have the first claim on setting the rules, those that provide the infrastructure on which the model is trained and run will also have a substantial say over what safety evaluations are required, what values are embedded, and what uses are permitted or prohibited.

Resilience. Dependence on foreign compute creates vulnerabilities beyond the military domain. AI capabilities cannot be stockpiled, and countries that rely on remote providers face the possibility that access could be restricted—whether because exporters want to squeeze prices or redirect limited computing capacity toward other priorities, or as a result of sanctions, export controls, corporate policy changes, or external physical or cyberattacks. As one analyst has put it, "Being a net importer of a core commodity in the modern economy is seldom an excellent place to be."

Fiscal stability. If AI-driven automation shifts a substantial share of economic output from labor income to corporate income, countries that host none of the relevant corporate activity may see their tax base erode with no corresponding new revenue stream. Locating as much frontier AI activity as possible within a country's tax jurisdiction—through data centers, local subsidiaries, or negotiated revenue-sharing arrangements—may provide a pathway to fiscal stability that pure importers of AI services will lack.

> **A country or coordinated bloc that hosts a meaningful share of global AI infrastructure could gain something analogous to what OPEC members hold in energy markets.**

A country or coordinated bloc that hosts a meaningful share of global AI infrastructure could also gain something analogous to what OPEC members hold in energy markets.

Even a single-digit share of global supply, if it can credibly be taken offline, would likely be enough to inflict real costs. Onshore compute does not guarantee sovereignty over AI. But it does make it far more likely that a country can influence how AI develops.

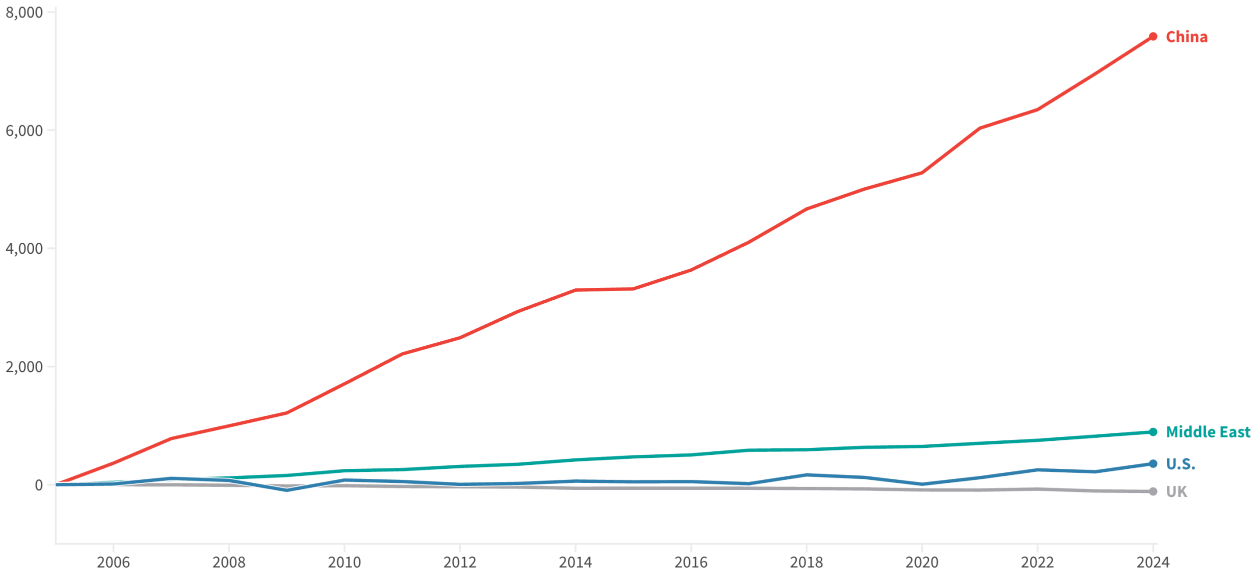
The Balance of Power

The distribution of compute will shape not only the power of individual nations, but also the contest between the United States and China and the standing of democracy on the global stage.

The U.S.-China competition. For now, the United States has a significant lead in compute: In 2025, an estimated three-quarters of the world's advanced AI computing clusters were on American soil, with just 14 percent in China. China likely reached only 1–4 percent of U.S. AI chip production in 2025 and will hit only 1–2 percent in 2026. But China has launched an all-out effort to improve its position across the AI supply chain, pumping more than $150 billion into the industry since 2014, supplemented in 2024 by an additional support package worth as much as $47.5 billion. The Chinese system can override local opposition, fast-track permitting, and direct capital with a unity of purpose that democracies rarely match. China also enjoys advantages in key inputs: it produces more STEM PhDs than the United States each year, and its electrical generation has soared from one-third of U.S. levels in 2000 to nearly two-and-a-half times U.S. levels in 2024 (see figure 11).[30]

Figure 11. Since 2005, China has Added >20x More Power to Its Grid Than the U.S.

Electricity generation growth across select countries and regions, TWh (2005–2024, cumulative)



Note: US, UK, and China data from Our World in Data, which uses Ember data supplemented with Energy Institute data where incomplete. Middle East data averaged from Energy Institute and Ember data. Middle East follows IEA categorization.

Source: Our World in Data (Ember, Energy Institute)

U.S. export controls have constrained Chinese access to the most advanced chips, but Chinese capabilities are advancing and the country has more than enough power and land to run every chip it can plausibly produce. If democratic buildouts slow while China accelerates, the balance of AI capability—and the broader balance of economic and military power that will increasingly track it—will shift.

The U.S.-China rivalry is the sharpest expression of a broader dynamic. Authoritarian governments argue with growing confidence that their model of governance delivers results—that centralized decisionmaking, state-directed investment, and the ability to override dissent produce faster growth and more effective responses to national challenges. If democratic states cannot maintain their lead in AI, they will reinforce the narrative that democratic governance is too slow and too fractious to compete in an era of rapid technological change.

> **If democratic states cannot maintain their lead in AI, they will reinforce the narrative that democratic governance is too slow and too fractious to compete in an era of rapid technological change.**

AI also poses an even deeper threat to democratic governance. Historically, authoritarian regimes have faced inherent constraints that eventually checked their power: Economic growth requires skilled workers who may demand political rights, and effective militaries risk becoming independent power centers. Advanced AI could loosen both. Automated systems could reduce dependence on workers who might organize for reform, while AI-enabled surveillance lowers the cost of monitoring and suppressing dissent. Many authoritarian regimes, including China, currently spend as much on domestic security as they do on external defense; AI could allow that control to scale more efficiently, entrenching authoritarian rule and extending its reach abroad.

There is also reason to think authoritarian AI development increases the risk of catastrophic outcomes. The conditions that produced disasters like Chernobyl and the destruction of the Aral Sea—lack of transparency, suppression of dissent, centralized target-setting, reluctance to report bad news—are precisely the conditions that AI safety researchers warn against in the development of powerful systems.[31] As Turing Award winner Yoshua Bengio has argued, authoritarian regimes are “more likely to make unsafe decisions regarding powerful AI systems, thereby increasing the likelihood of catastrophic outcomes.”

The United States and its allies cannot prevent authoritarian states from developing powerful AI indefinitely. But the tighter the race, the greater the pressure on democracies to cut corners on safety. A clear democratic lead reduces that pressure and increases the likelihood that frontier AI is developed under conditions that allow for genuine oversight. Democracies do not need a monopoly on AI. But they do need to stay ahead.

Open societies have generally outperformed closed ones because they better harness human capital and innovation. Democratic countries need to pull off that trick again by figuring out how to develop and deploy beneficial AI before their competitors do. None of this means democracies should ignore legitimate concerns about land use, energy consumption, and corporate influence. But it does underscore the stakes: The cumulative result of decisions around permitting reform and grid connections will shape the contest between political systems in the twenty-first century.

A Democratic Compute Coalition

The United States cannot build the world's AI infrastructure alone; nor should it rely solely on Gulf autocracies. Leading a broader coalition will let the United States bypass domestic constraints, ensure powerful AI is built in geopolitically dependable allies, and bind the free world into the U.S. technology ecosystem. Countries that join will generate economic returns, build geopolitical influence, and secure their seat at the AI governance table.

America Can't Build the AI Future Alone

For now, the United States is performing so well that its leaders may believe the country can build the world's AI infrastructure at home, without the need for messy foreign entanglements. It hosts the bulk of the world's AI chip clusters, and as our data suggest, large American projects can currently move faster than those in many other countries. The temptation will be to try to accelerate domestically, while largely ignoring the rest of the world.

That would be a mistake. The United States, for all its powerful advantages, cannot build the AI future alone. The physical constraints are too binding, the commercial imperatives too strong, and the political environment too risky.

> **The United States, for all its powerful advantages, cannot build the AI future alone.**

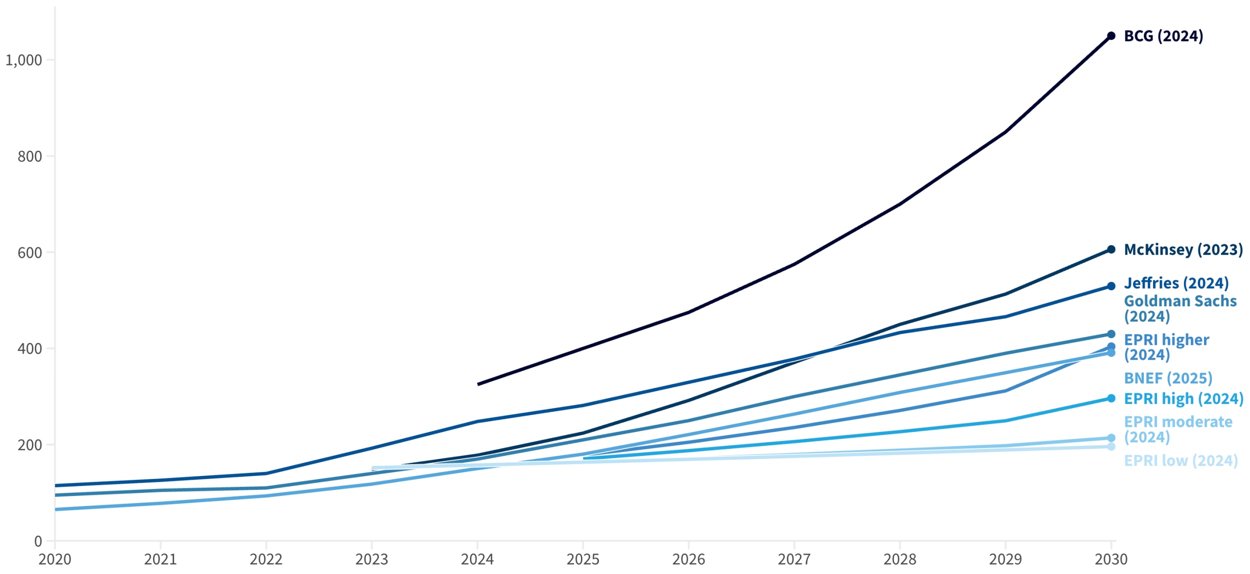
On the practical side, many customers will demand local sources of compute. Today, latency—the time it takes for a prompt and response to make the round trip between user and cloud server—is not a significant factor in most AI use, but that may change. AI services that rely on low latency, like self-driving cars, industrial robots, live translation, and real-time intelligence analysis, are set to grow in importance. In some instances, these can run on edge devices, but if the computation is sufficiently complex or memory-heavy, they will need to communicate back to a data center, which cannot be too far away or latency will degrade performance. On top of that, although copyright laws may make frontier training difficult in some jurisdictions, many countries also have privacy and data sovereignty laws that will require some workloads to be run domestically. These pressures have driven traditional cloud providers to build out global networks of cloud data centers over the past two decades. The same logic will apply this time.

Adding to the pressure, physical constraints in the United States are starting to bite. “The biggest issue we are now having is not a compute glut,” Microsoft’s chief executive Satya Nadella said in a recent interview. “It’s power.” By 2030, AI data centers will likely consume more than 100 gigawatts of power worldwide, according to estimates from the research firm Epoch AI (see figure 12). The U.S. grid today has roughly 1,250 GW in generating capacity and, according to one estimate, added only around 50 GW of baseload-equivalent capacity from 2000 to 2023. U.S. power growth has accelerated, but it’s not yet clear whether new capacity will arrive in time to meet the AI demand, even supplemented by onsite power.

Politics will also get in the way. America’s current ability to move fast is the result of jurisdictional arbitrage: Developers can dodge political and bureaucratic roadblocks by picking sites in friendly states, often far from major population centers. Yet that flexibility is coming under threat. Polls show data center construction is unpopular. Legislators in several states, including traditional data center powerhouses like Virginia and low-regulation havens like Georgia and Oklahoma, have proposed bills that would block new development. Policymakers in many cities and counties have floated similar proposals. At the national level, some lawmakers have called for a data center moratorium. The backlash is driven by a mix of concerns, from higher electricity prices to fears that AI will automate away swaths of human labor. Although the broadest efforts to block development are unlikely to succeed, political opposition is mounting rapidly.

Figure 12. AI Data Center Power Demand Is Expected to Accelerate Through 2030

Forecasts of U.S. AI data center energy use, TWh



Note: The data underlying the above forecasts include both published figures and BloombergNEF estimates, derived from visuals in third-party reports using chart extraction software and data interpolation methods. (Goldman Sachs figures corrected on Dec. 12, 2025, to display only US data.)

Source: BloombergNEF

What About the Gulf?

Even if U.S. policymakers accept those constraints, they may believe the United States can meet the world's needs by partnering with the United Arab Emirates, Saudi Arabia, and other states in the Persian Gulf. Governments in these countries can override political opposition. Regulatory processes are swift or nonexistent. And a favorable geographic location means AI providers will be able to serve customers in Europe, Africa, and the Middle East. Building American AI in the Gulf would dovetail with those states' play for global leverage as intelligence exporters—the new AI oil states.

The United States should, indeed, work with those countries; they bring important resources to the U.S. technology ecosystem. But it should take care not to over-rely on them. The Gulf does not solve latency concerns in Asia or Oceania, and it does not solve privacy requirements or data sovereignty rules anywhere. American companies and officials will need to forge a wider network.

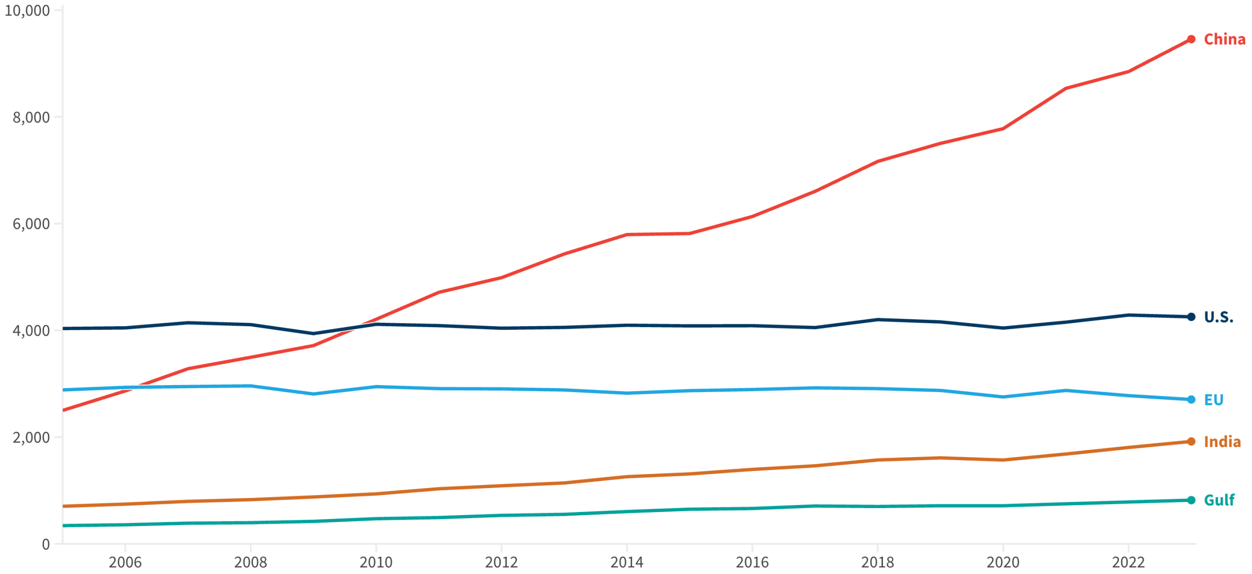
A heavy concentration in the Gulf would bring additional geopolitical risks the market may not price in. Gulf countries have uncomfortably close ties to the Chinese government, a desire to use AI for surveillance and social control, and a volatile home region. As data centers grow ever more valuable, they will become tempting military targets. In March 2026, for example, Iran struck two AWS data centers in the UAE and damaged another in Bahrain. After investing so heavily through the CHIPS and Science Act to bring back a key technology dependency from a geopolitical hotspot, the United States should not offshore the next critical technology to another such danger zone.

> **The UAE and its neighbors may struggle to support the volume of data centers the world would need.**

Nor is it clear that the Gulf countries can build at the scale required. Our data suggest that the UAE is likely able to bring some large data centers online rapidly—although even those estimates are based on pre-war projections—but the country and its neighbors may struggle to support the volume the world will need.[32] AI data centers are set to consume hundreds of terawatt hours of electricity every year by 2030. Yet taken together, Gulf countries produced a total of just over 800 terawatt hours of electricity in 2023, compared to about 1,900 TWh in India, 4,300 TWh in the United States, and 9,500 TWh in China (see figure 13). Some analysts have also argued that Gulf grids may be too unreliable to provide constant power for data centers.

Figure 13. In 2023, the U.S. Generated >5x as Much Electricity as the Gulf

Total electricity generated in each country or region, TWh (2005–2023).



Note: Gulf defined as Gulf Cooperation Council (Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and UAE)

Source: Our World in Data (Ember, Energy Institute)

Creating a Coalition

The United States cannot build the AI future alone—or with only its Gulf partners. But together, the United States and its allies and partners can dominate the AI supply chain. A coalition does not mean a formal international structure, nor does it require these countries to reach consensus on every dimension of AI policy. Instead, it means a loose network of partners that can together build, deploy, and govern transformative AI.

Leading that coalition will serve American interests. A fragmented Western world is one in which the United States bears the full burden of competing with China, faces constant pressure to share sensitive technologies through ad hoc bilateral deals, and watches as middle powers cut their own arrangements with Beijing or form countervailing blocs among themselves. A U.S.-led coalition, by contrast, will share those burdens while promoting a technology ecosystem aligned with democratic values and interests. Washington would be shortsighted not to recognize that technologically strong democracies make America stronger too.

If the United States does not lead this coalition-building effort, others will try without it. The resulting arrangements will either lack the scale to compete with China or bring democracies into uncomfortable, and often unequal, partnerships with states that do not share their strategic outlook or democratic principles. France and the UAE are already pursuing a $30+ billion AI data center joint venture, backed by the UAE's MGX investment fund. Other middle powers could try to form broader AI coalitions of their own, following the approach advocated by Canadian Prime Minister Mark Carney.[33] China, meanwhile, is working to shape the global AI order through efforts like its proposed Global AI Governance Initiative, World AI Cooperation Organization, and AI+ International Cooperation Initiative. None of these alternatives will serve U.S. interests as well as a coalition the United States helps lead. Nor will they serve most middle powers well: Worries in allied capitals about U.S. reliability are understandable, but when it comes to AI, U.S. technology, capital, and leadership remain indispensable.

Some of this coalitional logic has found its way into U.S. policy. The Trump administration's "Pax Silica" initiative seeks to build a coalition that can secure key parts of the AI supply chain; the administration has stated that it intends to create an investment consortium to support the initiative.[34] And the administration's National Security Strategy, for all its skepticism of U.S. allies on other matters, notes that "the United States must work with our treaty allies and partners—who together add another $35 trillion in economic power to our own $30 trillion national economy," totaling "more than half the world economy."

How these initiatives will work in practice remains to be seen. Pax Silica is a good first step, but the administration's rhetoric suggests that when it comes to the data center buildout, it may be largely focused on the Gulf states. And policymakers have given few indications of the tools they will use to put the broader coalition into effect, beyond the role of the AI exports program.

> **A viable strategy must be more than "buy America."**

A viable strategy must be more than "buy America." AI will reshape the labor markets that structure economic life and the military and surveillance capabilities available to governments. It will concentrate power in ways that could undermine democratic accountability. If authoritarian states dominate AI's buildout while democracies remain divided, democratic states will fall behind. The free world can shape this technology to its advantage only if it leads its development together.

Compute

Compute will be a central pillar of that coalition. As the scholars Kurt Campbell and Rush Doshi have argued in another context, the democratic world has the collective scale necessary to lead the buildout. On four major dimensions, U.S. allies can make up for American limits.

Geography. The United States faces constraints: Grid capacity and critical equipment supplies are limited, and the most promising locations often face competing demands. Allied nations can complement American buildouts: Canada's hydroelectric resources, Northern Europe's cool climates and renewable energy, Australia's solar potential. Distributing compute across allied territory also reduces concentration risk: a network of facilities is more resilient than any single point of failure.

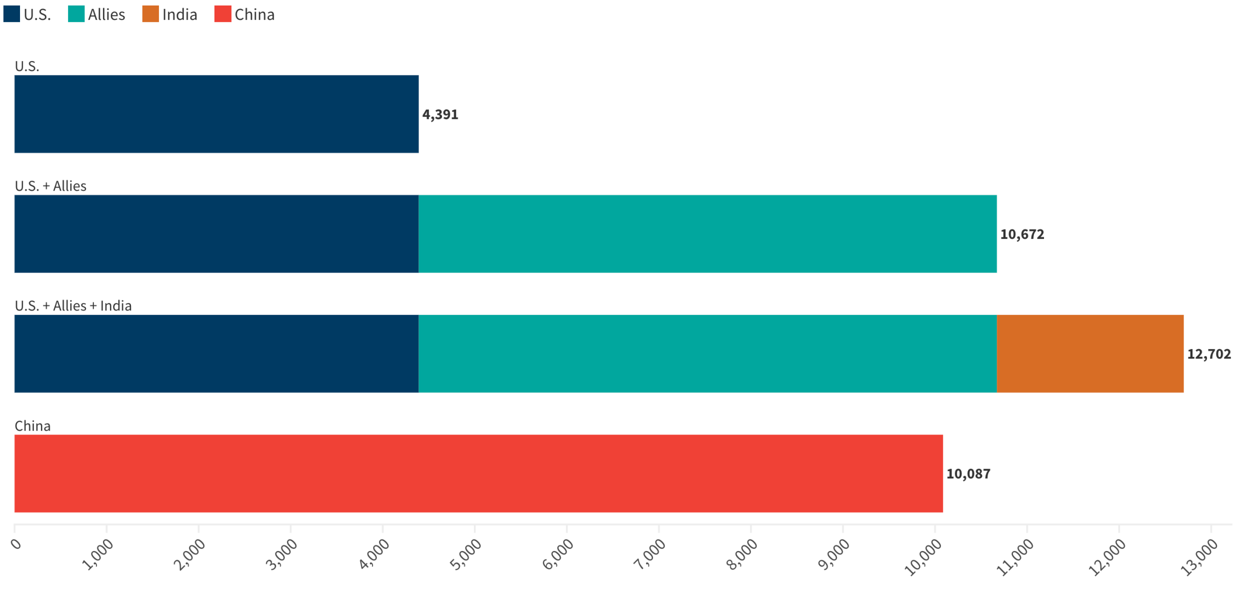
Politics. The compute buildout in the United States is running into growing opposition. Together with allies, government and industry can reduce the backlash by spreading development across multiple countries and regions. If developers can hedge their bets, and citizens of each country do not feel like they are shouldering the entire infrastructure burden, politicians are more likely to be able to convince voters that the technology's benefits are worth the costs.

Power. Today, the United States generates less electricity than China and the gap is growing rapidly. But add in major U.S. allies—Australia, Canada, the EU, Japan, South Korea, the United Kingdom—and the picture changes substantially. Include India, a democracy with its own massive energy buildout underway, and the democratic coalition pulls well ahead (see figure 14).

Capital. The investments required to fund the AI buildout run to trillions of dollars over the coming decade. The United States has the largest stock market in the world and is home to the wealthiest public tech companies serving as the primary financiers of the ongoing infrastructure buildout. Still, others could contribute: Europe's public companies have a

Figure 14. The U.S., Allies, and India Generate Nearly 3x the Electricity of the U.S. Alone

Electricity generation in 2024, TWh



Note: Calculations consider U.S. allies in the global top 50 for electricity generation.

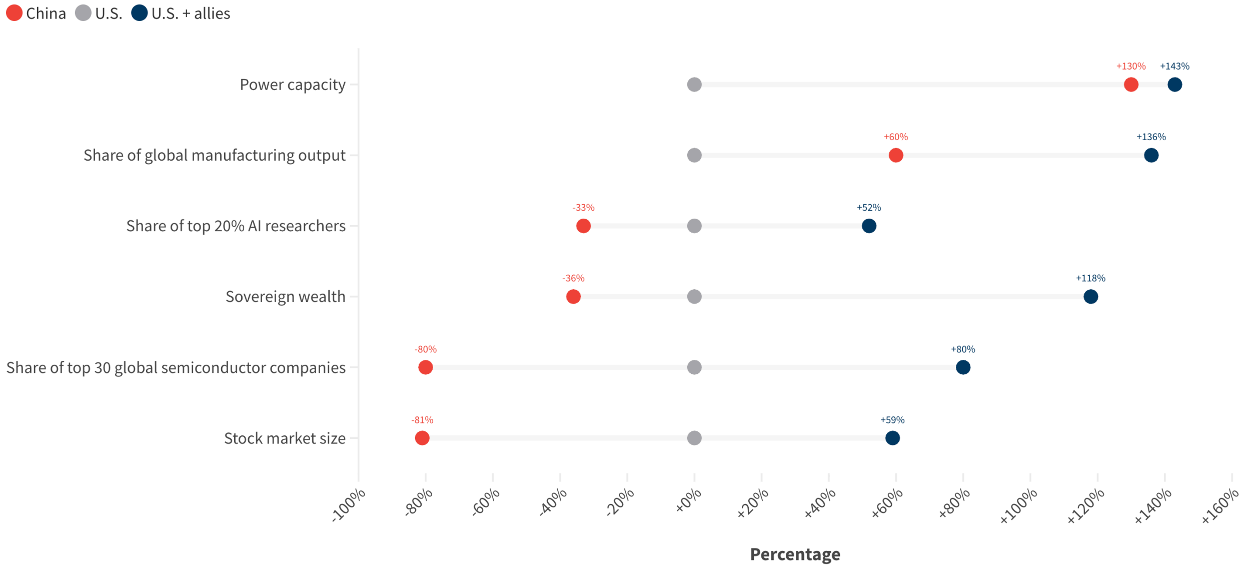
Source: Ember Energy, authors' calculations combined market capitalization of around €16 trillion (about $18 trillion) and Norway's Government Pension Fund is twice the size of Saudi Arabia's Public Investment Fund. Pooled investment vehicles could channel patient capital at the scale required without overexposing any one country.

The Wider AI Value Chain

If America's allies step up, they can win the AI infrastructure race. As the Pax Silica initiative recognizes, the United States and its partners already control most of the key steps in the AI value chain, as well as many capabilities and resources that complement it. Japan, the Netherlands, and the United States produce the most sophisticated semiconductor manufacturing equipment. Taiwan, South Korea, and the United States host advanced chip fabs. The United States, the United Kingdom, and Canada have many of the world's leading public and private AI research institutions. Australia, Canada, and the United States possess major reserves of many critical minerals. Germany, Japan, and South Korea bring extensive manufacturing expertise; European countries produce more steel, cars, ships, and civilian aircraft than the United States. U.S. capital markets are the world's deepest, and several allies, including the EU, Japan, and the United Kingdom, also have large capital bases. And from wind and hydro power in northern Europe to solar in Australia, U.S. allies possess huge energy potential (see figure 15).

Figure 15. The U.S. Benefits From the Additional Scale of Its Allies

China and U.S. allies compared to the U.S., % difference



Note: Generally defined as U.S. treaty allies. Exact countries and method depend on the dataset used; contact authors for additional details.

Source: Authors' calculations based on Ember Energy (power), World Bank (manufacturing), Paulson Institute (AI researchers), GlobalSWF (sovereign wealth), CompaniesMarketCap (semiconductor companies), SIFMA (stock market size)

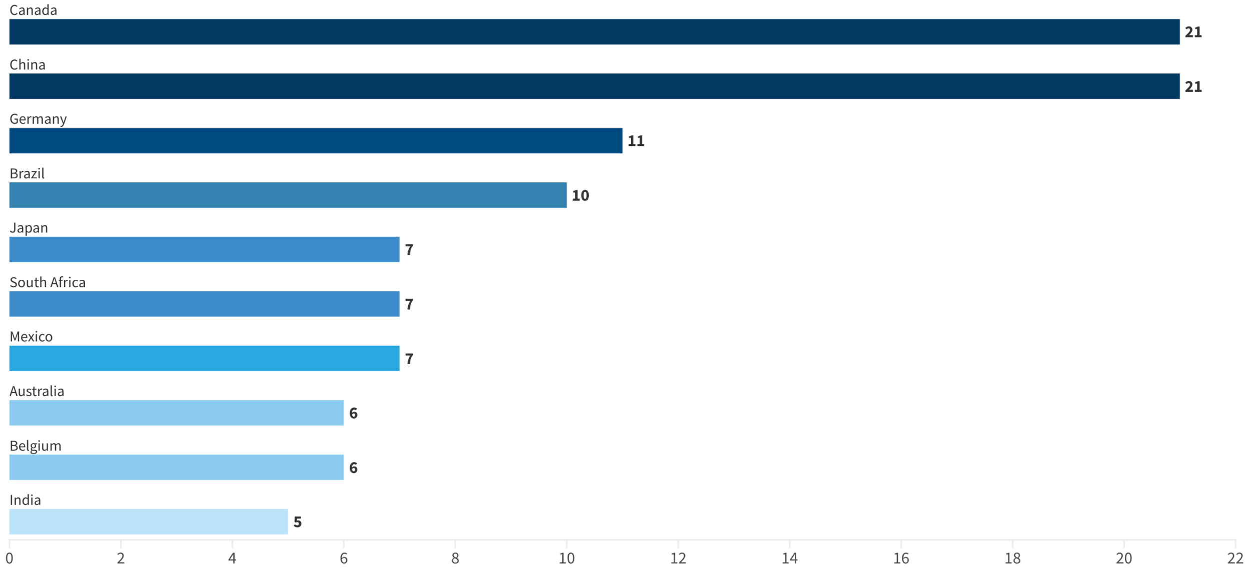
No single country has all of these. Together, the coalition can outcompete the rest of the world.

The United States depends on those countries for capabilities that no amount of domestic investment could quickly replicate. Deepening their ties to the U.S. technology ecosystem, including by facilitating data center investment, will therefore benefit both the United States, by giving its partners a greater stake in its success, and U.S. allies, who will gain a greater share of the economic returns and strategic influence over the future of AI. And for allies and partners that don't already have a niche, becoming a compute hub offers one pathway to securing a place in the AI future.

Greater coordination can also help in the areas where the United States and its partners are falling short. In critical minerals, for example, joint investment may overcome Chinese dominance of materials processing (see figure 16). Or agreements to share data on advanced manufacturing may accelerate robotics research. Those policies are not the subject of this article, but policymakers should look for ways to leverage the scale of the allied world across the AI value chain. Doing so will benefit both sides and reduce the risk that an adversary can deny the democratic world the advantages of powerful AI.

Figure 16. The U.S. Relies on Many Other Countries for Critical Minerals

Suppliers of nonfuel critical minerals for which the U.S. was >50% net import reliant, # of minerals supplied (2024)



Note: Data for 2020–2023

Source: United States Geological Survey, Mineral Commodity Summaries 2025

III. What Democracies Should Do

If democracies are to maintain their advantage in the AI contest, they must act now. At home, they need to unlock the domestic capabilities to compete. Abroad, they need to start forging the partnerships and agreements to develop, adopt, and govern the technology.

Domestic Reform

On the domestic front, democratic U.S. allies are hardly ignoring the AI infrastructure race. Over the past few years, governments have launched a range of initiatives to promote data center development. Some of the most prominent include:

- **The EU** has promised to build up to five large data centers, which it calls "Gigafactories," each containing around 100,000 AI chips.
- **India** is providing a twenty-year tax holiday for data center operators, and the central government is encouraging states to identify ready-to-operate locations for major data centers. Tax savings for an eligible 100 MW Blackwell data center could exceed half a billion dollars, rivaling the impact of time-to-power, but at a high fiscal cost.[35]
- **France** has announced investment commitments totaling 109 billion euros ($112 billion) in its AI sector and is working to identify and expedite approvals for promising data center sites.
- **Denmark, Finland, Norway, and Sweden** have offered a mix of low-tax energy, fast-track permitting, and grid investment.
- **The United Kingdom** has announced "AI Growth Zones," which will offer discounted electricity and faster planning approval and grid connections.
- **Australia** has promised regulatory priority for AI data centers that promote the country's "national interests," including by funding renewable energy and sharing benefits with local communities.

Some of these proposals are rightly focused on the time problem. The United Kingdom's AI Growth Zones, for example, hope to accelerate grid connections by up to five years. France has identified twelve "France 2030" data center sites where land permitting has already been carried out, as well as four sites near power stations "allowing for almost immediate grid connection."

Other approaches risk going in the wrong direction. The EU's Gigafactories initiative, which includes a lengthy public consultation and planning process, seems likely to direct capital toward infrastructure projects that will be built far more slowly than commercial alternatives. More broadly, Brussels's approach to AI has often seemed to focus more on regulating American technology companies than on creating the conditions for European competitors to emerge. European policymakers can no longer count on reflexive support from sympathetic politicians in Washington. If they want ongoing technology partnerships with the United States, they will need to show what they can bring to the table. That will mean figuring out how to accelerate the continent in the AI contest.

> **Money spent subsidizing electricity for data center operators is likely to be money wasted.**

Some other policy initiatives seem similarly misguided. Money spent subsidizing electricity for data center operators is likely to be money wasted: if the projects can get online, they will generate returns large enough that they do not need public funds. Even regular corporate tax incentives are less important than accelerating timelines. India, for example, has announced a twenty-year tax holiday for foreign data center operators serving global customers, a likely more expensive approach than creating fast-lane approval processes. Locking in tax shields that last decades is likely to be especially unwise given the rapid economic changes AI may bring.[36]

Rather than committing taxpayer funds to support an already lucrative industry, policymakers looking to promote AI infrastructure should focus on the timing problem.

Solving the Time-to-Power Problem

The delays that plague data center construction are not laws of nature. They are the result of policy choices—choices that can be revised.

Consider the variance within the United States. Although we estimate the average project takes two years from approval to first operation, OpenAI's Stargate Abilene project received its initial permit to install onsite gas turbines within two weeks of its application.[37] Epoch AI estimates the first phase was operational some sixteen months after construction began. In Indiana, a large Amazon-Anthropic data center similarly saw construction and air quality permitting proceed in parallel, with nearly 500 MW of capacity online within eighteen months.

These timelines are dramatically faster than the U.S. average and easily competitive with those in the UAE. The difference isn't primarily geography or resource endowments, but regulatory approach. Some U.S. states have streamlined permitting systems, accommodating utilities, and a political culture that welcomes large industrial projects. Other states and countries could take similar approaches if they chose to.

The same logic applies internationally. Finland's relatively fast timelines reflect policy choices: streamlined parallel permitting and a power system built around large-scale wind and renewables, creating excess capacity. (Though even Finland suffers from transmission bottlenecks.) Other democracies could move in this direction, but doing so requires political will to prioritize speed.

Of course, speeding up can come with costs. Shortened reviews might miss environmental risks. Accelerated grid connections might strain aging equipment. Each country will have to decide if the tradeoff is worth it. But whether from inertia or lack of attention, most democratic states appear to be underweighting the benefits of moving faster in the AI competition.

There are many potential policy fixes countries might adopt to accelerate time to power, spanning a wide range of practical and political difficulty. Some of the most significant include:

Permitting

Queue management. Companies looking to connect to the grid will sometimes submit speculative or duplicate connection requests, which can clog up review processes and suggest to grid operators that they need to build for peak demand that may never arrive. A 2024 EU study, for example, concluded that overestimating peak demand caused delays and drove up electricity prices for everyone.[38] Stronger verification rules, fees for progression through project milestones, and better tracking can give planners a more accurate sense of genuine peak demand and help real projects move to the front of the line. These policies work: Industry insiders report that when developers are asked to pay, two-thirds of data center power requests "vanish." Policymakers can also publish real-time data on current interconnect wait times and regions with spare capacity, batch requests to speed up reviews, strengthen measures to prevent speculative applications, or allow companies to trade queue positions so energy flows to the most valuable projects.

Creating fast lanes. Governments should consider creating fast-track review processes for data center projects that meet set criteria. What exactly to prioritize will be up to local policymakers, but plausible factors include the data center's ability to use clean energy, commitments by the operator to pay for new power generation and grid upgrades, the economic benefits it creates for the local and national economy, whether it brings its own backup power, its ability to accept curtailment during grid stress, and whether it will be built in an area with excess grid capacity. If data center operators can hit those marks, their projects could be moved to the front of the grid connection queue. The UK's AI Growth Zones and Australia's National AI Plan provide examples of how such a policy could be structured.[39]

Grid Reform

Load flexibility. Grid operators design systems around peak loads, but most of the time, the grid runs with excess capacity. According to the International Energy Agency, if data centers can reduce their grid demand just 1 percent of the time, the United States and the EU could accommodate some 110 GW of new data center capacity. Doing this will take new management techniques, as cloud providers cannot simply shut down critical services for hours at a time, but there are several plausible options. On-site batteries could offer backup power for short periods. Smart scheduling can shift workloads away from peak hours. And global operators can move computation to regions with lower demand. Startups are working on technical solutions, but barring a breakthrough, the key will be setting the right incentives: if flexibility can shorten grid connection times, cloud providers will find ways to deliver it.[40]

Grid flexibility. When new power generation needs to connect to an electrical grid, the grid operator will often require upgrades to transmission infrastructure to ensure it can cope with the new load, a policy sometimes known as "invest and connect." Because these upgrades must account for load peaks, they can slow connections by requiring infrastructure that won't be needed most of the time. A different approach, often called "connect and manage," allows generators to hook up to the grid before transmission upgrades are complete, and relies on the grid operator curtailing peak output to avoid overloading the system. By using this approach, the Texas grid, ERCOT, manages to connect projects in two to three years, compared with a typical five to seven years elsewhere in the United States.[41]

Incentivizing the production of key components. Where the bottleneck is the availability of grid components such as transformers, turbines, and circuit breakers, policymakers should consider targeted tax incentives, offtake agreements, loan support programs, or other mechanisms to boost production. Policymakers will have to first determine whether investment is actually the bottleneck, as opposed to shortages of talent or materials. If there is a true market failure, there is a plausible case for government support.

In some cases, policymakers may be tempted to adopt more centralized control of critical supplies. In the United States, this might involve invoking provisions of the Defense Production Act that allow the federal government to prioritize certain customers above others for parts in short supply. Yet such moves risk misfiring, especially given the fast-moving nature of the technology. Officials in Washington and overseas should generally emphasize solutions that unlock, rather than direct, supply, helping channel the market to address bottlenecks rather than undermining market forces.

Behind the Meter

Some commentators have argued that the shift to BTM power will resolve many of the industry's issues. According to an analysis by energy research firm Cleanview, after a spate of announced BTM projects in 2025, some 30 percent of planned U.S. data centers expect to use onsite power.

Policymakers can promote clean, fast BTM power in several ways. First, governments can create permitting fast lanes for BTM power sources that meet emissions criteria and are designed as a temporary bridge solution while a data center connects to power. One solar developer who works on 100 MW+ AI data center projects told us that permitting is the most important bottleneck to their business creating BTM power for data centers, as solar uses large amounts of land.[42]

Second, governments should explore ways to promote fuel cells or solar microgrids, likely through financial incentives. Fuel cells produce fewer emissions than gas turbines and can often be deployed faster than other BTM sources, but are not used widely due to unattractive unit economics. Solar panels, meanwhile, are cheap, quick to manufacture, and avoid the emissions of gas turbines. Recent research suggests that solar microgrids—small-scale, self-contained energy systems pairing solar panels with battery storage—could be cost-competitive with gas turbines for data center applications, particularly in regions with high solar potential like the American Southwest, Australia, and Spain. Meta's data center cluster in Mesa, Arizona, for example, draws on power from a co-located 300 MW group of solar farms and battery storage units. Google, meanwhile, is building a data center cluster with 1.9 GW of co-located wind, solar, and battery storage in Minnesota.

Solar is already growing rapidly in the United States and elsewhere. In 2024, the United States added nearly 40 GW of new solar capacity to its grid, while Europe added nearly 60 GW and India added some 24 GW.[43] Even if governments aren't able to speed up grid interconnect wait times, data center builders may be able to use behind-the-meter solar farms.

Solar BTM faces its own challenges: intermittency requires substantial battery storage, which adds cost and complexity, and large data centers have significant land requirements. Still, for democracies looking to speed up AI infrastructure while meeting climate commitments, solar offers an increasingly promising option as panel costs fall and battery technology improves.

Addressing Secondary Cost Drivers

Time dominates, but other factors still matter.

Tariffs. The model shows that moderate tariffs on AI servers could cost hundreds of millions of dollars per facility.[44] Policymakers should understand that tariffs on AI equipment effectively tax domestic AI infrastructure buildout. As elsewhere in the manufacturing economy, tariffs that look like protectionism for domestic workers may end up harming important domestic industries.

Tax treatment. State and local incentives, bonus depreciation, and other tax provisions affect data center economics. These provisions may be worth preserving if they serve as incentives for facilities to meet policy goals (speed, clean energy, local investment). But tax incentives are not a substitute for fixing underlying speed problems.

Electricity prices. Energy costs matter less than conventional wisdom suggests, but they still matter. Policies that keep electricity affordable—including support for new generation, efficient grid operations, and competitive wholesale markets—help at the margin. What's more, policies that speed up the development of new power sources will help bring down costs for everyone, not just data center operators.

Construction efficiency. American construction productivity has declined over decades, making everything—including data centers—more expensive to build. This is a deep structural problem beyond the scope of AI-specific policy, but a potential shortage of skilled data center engineers threatens to make things even worse. Efforts to train new workers or allow foreign engineers to immigrate would help.

The Political Challenge

Accelerating data center development involves technical questions, but it won't be a merely technocratic exercise. Policymakers may feel all they need to do is identify the right sites, solve the transformer shortage, or tweak the interconnection queue. These things matter. But AI infrastructure is also a political problem.

> **Opposition to data centers represents a genuine political groundswell that cannot be wished away.**

The Bernie Sanders–Alexandria Ocasio-Cortez moratorium proposal is unlikely to take effect—and as this paper argues, it fails to recognize that halting construction means giving up the benefits of AI progress while ceding the field to authoritarian competitors—but opposition to data centers represents a genuine political groundswell that cannot be wished away. Citizens worried about energy prices, the environment, and job displacement need answers rather than dismissal. Executives and policymakers need to articulate a positive vision for the social benefits of the technology they're building and enabling.

International Coordination

Domestic reform can get the democratic world a long way toward collective dominance of the AI supply chain, but there will be places where international coordination is needed. There will be work to enable the AI buildout itself: making cross-border investment across democracies easier, taking down trade barriers, giving AI developers and deployers common sets of operating rules.

If those initiatives ensure powerful AI systems are developed in the democratic world, governments will then be more able, if they choose, to coordinate the governance of AI itself: collaborative research on safety and security risks; shared rules for intellectual property and taxation; common standards for the use and control of AI in military and intelligence services. The exact shape policy should take in these and other areas isn't yet clear, and countries should be wary before locking themselves in. But by collectively controlling the AI supply chain, the democratic world will set itself up to control the future of AI.

Securing the Supply Chain

Much of the effort to cement control over the AI supply chain within the democratic world will involve governments getting out of their own way. Companies will naturally coordinate across borders; capital and trade flows will drive much of the buildout without central direction. But governments can easily disrupt these dynamics—through tariffs on allies, capricious diplomacy that undermines trust in domestic firms, extraterritorial regulation with sweeping compliance burdens, or prolonged legal uncertainty. A democratic compute coalition will require the opposite instincts.

This work could involve a range of policies, such as coordinated investment rules, harmonized tax systems, mutual recognition of safety tests to simplify approvals, or parallel transparency requirements. A mutual fast-track process for allied AI infrastructure investment screening would reduce the risk that national review bodies, such as the Committee on Foreign Investment in the United States, the EU's Foreign Direct Investment Screening Regulation, and similar organizations in Australia, Canada, and Japan, slow down joint investments. Common rules for transparency and third-party testing by major AI developers would prevent a patchwork of requirements from holding back deployments. A joint tax agreement might prevent a race to the bottom on the taxation of AI agents, if fiscal bases shift from income to consumption or corporate taxes.

These policies will help the democratic world improve its position across the AI supply chain. In areas like critical minerals, advanced packaging, and robotics, the United States should work with partners that have relevant capabilities to ensure that its adversaries can neither control chokepoints nor deploy AI more rapidly throughout their economies. The Trump administration has taken valuable steps in this direction, including the United States–Australia critical minerals framework. Still, more will be needed: joint research, investment, offtake agreements, prizes, and advance market commitments.

These efforts can take many different shapes, as the issues dictate: bilateral deals, small groups that act on specific issues, or broader coalitions that form to tackle the largest problems. Alleviating supply chain chokepoints will mean working directly with the countries that have the necessary resources and capabilities. In the critical mineral supply chain, for example, nations with strong domestic demand, such as India, Japan, and the United States, are well positioned to spearhead de-risking efforts like guaranteed purchase agreements, hedging mechanisms, and insurance offerings to help stabilize pricing. Resource-constrained countries like South Korea can contribute by offering affordable financing options and technological assistance to lower the risk profile of upstream ventures.

Joint AI safety research, meanwhile, can involve flexible collaboration across groups of countries with the relevant technical talent, primarily countries with well-resourced AI security institutes like the United States and United Kingdom. Preventing a race to the bottom on issues like taxation or intellectual property rights will likely need alignment across broader groups, such as the G7 or OECD. Responses to crises caused by the misuse of, or loss of control over, AI systems will fall to those with the ability and the will to act when the moment comes. The essential element is not any specific group or diplomatic channel, but the existence of a set of allied democracies with the material base, the human talent, and the political ability to develop, deploy, and govern transformative AI.

Finally, a more prescriptive role for policy in creating an AI coalition would involve governments actively steering compute toward democratic jurisdictions through export frameworks that treat coalition members preferentially. The Biden administration's AI diffusion rule offered one model: tiered access to advanced chips based in part on a country's security relationship with the United States. A more durable version of this approach, perhaps negotiated with allies rather than imposed unilaterally, could channel investment toward democracies while giving partner nations a stake in maintaining shared standards, for example on data center security provisions. The goal would be to shape commercial incentives so that building in democratic jurisdictions becomes the path of least resistance.

Shaping the AI Future

Ultimately, countries that contribute real value—faster permitting, abundant land, clean power, sovereign capital, technical capacity—will have the most leverage to shape the future of AI. If the democratic world leads, it will have far more opportunities to ensure that transformative AI systems reinforce, rather than undermine, their interests and ideals.

> **We don't yet know every governance challenge that AI will raise, but compute will serve as one of the most important levers for AI influence.**

We don't yet know every governance challenge that AI will raise, but compute will serve as one of the most important levers for AI influence. Countries that have large quantities of compute on their soil will be far better off than those without it. The United States and its partners should want the balance of that influence to lie in the democratic world.

Conclusion

The AI infrastructure race is accelerating. Over the next decade, trillions of dollars will flow into data centers around the world, building the computational foundation for systems that will reshape economies, governments, and societies. Where that infrastructure is located will influence who governs AI and whose values are embedded in the technology.

The United States has an early lead, but that lead is fragile. And many of its democratic allies are far behind. Authoritarian competitors—China through sheer mobilization, Gulf states through speed and capital—are seeking to capture an outsized share of the buildout.

Our research shows that right now, the most important factor in data center competitiveness is time. Countries that can bring facilities online quickly offer dramatically better returns than those where projects languish for years.

This finding has clear implications for policy. Democracies that want to compete need to remove the obstacles that slow them down: permitting delays, grid connection backlogs, transmission bottlenecks. These barriers are not laws of nature; they are the result of policy choices that can be revised. Countries that make speed a priority—as some U.S. states and allied nations have begun to do—can match or exceed what authoritarian competitors offer.

But speed requires a social license to operate. Enabling data center development involves technical questions, but it is fundamentally a political issue. Local opposition, environmental lawsuits, and broader public distrust can delay projects for years. Industry players need to realize that the downsides of development cannot be wished away. Free societies can mobilize extraordinary resources with extraordinary speed, but only if citizens have bought in. That means investing in host communities, addressing environmental impacts, dealing with the technology's risks, and building AI that serves the public interest. Social accommodation and rapid development need to go hand in hand.

Critics of the industry, for their part, should realize that AI really will underpin national power. Some bottlenecks to development reflect legitimate disagreements over the risks and benefits of the technology; others are the product of atrophied systems that delay all development, good or bad. The best way to shape the trajectory of AI, share the gains, and address the risks is not to attempt to block development entirely, but to strike better bargains with the developers.

Domestic reform is necessary but not sufficient. The scale of the AI buildout exceeds what any single democracy can support. What's needed is a coalition: allies working together to build infrastructure, coordinate supply chains, and ensure that frontier AI development is concentrated in democratic countries. The capabilities—capital, energy, talent, manufacturing—are distributed across the democratic world. The question is whether democracies can mobilize them at the speed the moment demands.

Democracies that rise to the challenge will help determine how AI reshapes the world. Those that do not will have little say in the matter.

Appendix 1: Model Methodology

The Carnegie Endowment for International Peace's International AI Data Center model measures costs and revenues from an illustrative 100 MW hyperscaler greenfield AI-optimized data center over a twelve-year life cycle, and then uses these cashflows to estimate the life cycle value of an AI data center across ten different countries. The model is for a Blackwell data center with GB200 chips and NVL72 racks, the most common spec for 2026 AI data center megaprojects.

We assume that the data center is financed by free cashflow from hyperscalers and includes two power scenarios.[45] In the first, the hyperscaler uses small natural gas turbines or reciprocating engines as a source of behind-the-meter power until the data center is connected to the electrical grid. In the second, the hyperscaler waits to operate the data center until it can be connected to the grid. For reporting top-line results, we average the outputs from these two scenarios. The actual mix will be somewhat similar: Bloom Energy expects 38 percent of data centers to be powered at least in part by onsite generation in 2030, but this number is likely higher for the largest projects.

Revenue estimates are subject to substantially greater uncertainty than cost estimates, but sensitivity analysis suggests that cross-country differences are primarily driven by costs and delays rather than revenue assumptions.

Our model is a life cycle model, covering costs and revenues to estimate the valuation of an AI data center. We use data from Epoch AI as a base estimate of AI data center capital expenditures and estimate how non-IT capital expenditures vary by country using Turner & Townsend's Data Centre Construction Cost Index. Data from SemiAnalysis, New Street Research, and other industry and equity research publications informed our estimates of behind-the-meter and operating costs.

To account for delays in opening the data center, we estimate time from starting construction to first operation and full completion for each country. These times vary depending on the power scenario.

Estimating these timelines was not easy, but we used a combination of reference projects, announced project timelines (which we assess with appropriate skepticism), and background research on the energy and grid climate in a country to arrive at best-guess estimates. We provide more information about our methodology in appendix 2. Epoch AI's Frontier Data Center data hub, which estimates timelines for select megaprojects based on these sources and satellite imagery, was valuable to us in this process. For each country, we assume a twelve-month ramp time during which the data center is fully brought online, aligning with the multi-phase plans generally undertaken by developers.

We discount revenues by the real Weighted Average Cost of Capital (WACC), which we estimate at around 6 percent for big tech companies. This reflects published details on AI data center deals, as well as big tech companies' balance sheets and access to low-cost capital. These rates drive the cost of delays: deferred revenue is discounted according to the WACC and is therefore less valuable, reducing profits. All cash flows are modeled in real 2025 U.S. dollars and discounted at real WACC.

We estimate monthly revenue using a bottom-up approach and rely on published on-demand chip prices. According to online marketplaces, the lowest on-demand price to rent GPUs from GB200 chips as of May 2026 is $10.50/hour, from CoreWeave. We use these prices to estimate how much a hyperscaler would receive for an hour of compute from a GB200 chip. Even if the chips are only used internally, this number can be treated as an imputed rent.

On the one hand, hyperscalers can generally sell compute for a higher price than neoclouds, as they are considered more reliable and may provide higher levels of service. On the other hand, on-demand prices generally exceed prices for enterprise buyers who receive volume discounts and commit to multi-year contracts. Prices also depend on how chips are configured and where a data center is located. Balancing these factors, we estimate the average initial price of a GB200 hour at $10.

Over time, as new chips are released, the prices of previous chips decline. We use data on historical chip price declines sourced from Augur VC and McKinsey & Company to estimate how much the price of chips drops every year after release. It is not clear whether future chip price declines will match previous ones. Chip prices are a function of the supply of AI computing power and demand, both of which are difficult to forecast given the infrastructure boom. Nvidia also announced in 2024 that it will begin releasing chips more frequently, from once every two years to once a year. This might drive further downward pressure on chip prices.

We also estimate chip utilization at 70 percent. This assumes that at the beginning of a chip refresh cycle, a data center will be used for training, driving near-full utilization, before transitioning to inference, which causes utilization to fall below 70 percent. BloombergNEF expects the average utilization rate among all data centers to be 69 percent by 2035, and 70 percent is between the values that others assume.

We assume that revenues are the same across different locations, that subsequent chip cycles have similar unit economics, and that chips are replaced every four years. Recognizing that these assumptions are controversial, we stress test the first two in addendums to the model and include a toggle to test the impact of varying chip life on results.

We test the first assumption through an analysis of on-demand Google Cloud and Amazon Web Services prices for Hopper-class chips. As of January 2026, we find that prices for most of the countries in the model are within a 10 percent band, except in South Korea, where they are 23 percent higher. This implies that South Korea could be more competitive than initially believed, though these prices are only for inference compute, not training, and could change due to the ongoing buildout.

We test the second assumption through a calculation of how the trends toward more floating point operations per second (FLOPs) per watt of power and fewer dollars per FLOP over time compare to each other. Based on data from Epoch AI, we find that they roughly cancel each other out, though with high uncertainty. We include a sensitivity table showing that if we are off even a few percentage points in our estimates of either trend, the conclusion becomes quite different.

We calculate an all-in tax rate in each country and then apply taxes according to that country's depreciation rules.

We multiply the average earnings before interest, taxes, depreciation, and amortization (EBITDA) of each data center by an exit multiple of eight and discount it beginning at the end of that data center's twelve-year operational timeline to calculate Terminal Value. Terminal Value is added to twelve-year operating cashflow to arrive at the life cycle value of each data center.

Recognizing many of the load-bearing assumptions in the model, we include a dashboard within the model allowing key parameters to be edited. The dashboard includes multiple scenarios for chip price declines and a toggle allowing initial chip prices per hour to be edited. This structure allows sensitivity analysis as well as Monte Carlo simulations based on varying multiple inputs.

It is important to note that results on the impacts of delays may be spiky because the model discounts cashflows on an annual basis. As a result, the life cycle value of the data center rises when a delay of an additional month pushes chip purchases one year back—the negative impact of the delay is outweighed by pushing billions of dollars of capital expenditures farther into the future within the model. This is an artifact of the model and does not significantly change results, but does explain why the impact of a one-year delay is not exactly twice the impact of a six-month delay.

The authors welcome additional feedback. For access to the model, please contact the authors.

Appendix 2: AI Data Center Timelines

This appendix provides a more detailed overview of how we estimated time to power for an illustrative 100 MW AI data center in the countries surveyed in the financial model.

The key points are as follows:

- **Reference projects:** We looked for the most directly comparable historical data centers as well as planned AI data centers to estimate timelines. We supplemented these estimates with background research on electrical grids and energy development for each country. Not all projects are equally relevant and the weight of this background research depends on its date of publication, source, and other factors. Therefore, the results are not a simple average; rather, they are subjective estimates informed by the authors' research.

- **Methodology:** We counted the start of the timeline as the date of the data center's announcement or the date when construction reportedly started. We counted the end of the timeline as the first point when the data center was operational or when the data center was announced as completed. When both the announcement and the construction start date were available, we preferred the announcement date. (A data center is usually announced only once the developer has a permit, but starting the clock here accounts for delays in the start date of construction due to power procurement.) When both the completion date and the date of first operation were available, we preferred the date of first operation. For the sake of simplicity, for all countries, we assume a one-year ramp time from first operation to operation of the full 100 MW data center.

- **Confidence:** These estimates should be treated as preliminary. We have moderate-to-high confidence in our estimates for the United States due to a larger sample and strong external data, and moderate-to-low confidence in our estimates for other countries in the dataset.

Below, we provide two examples that are illustrative of the process for each country and power scenario.

Example 1: U.S. Estimates

For the United States, we sourced most data from Epoch AI, a research organization that maintains a tracker database of large AI data center projects. As of April 2026, Epoch's data included twenty-three U.S. AI data centers, most far larger than 100 MW.

For each data center, we calculated the time from Epoch's listed start date to when the first phase of the data center is supposed to come online. In line with our methodology, Epoch AI counts the start of construction as when a company makes an announcement or when land is cleared. They use public sources as well as satellite imagery to approximate the completion date for each phase of an AI data center project. Our analysis of Epoch AI's data is linked here.

Epoch AI did not list start dates for three of the twenty-three projects (Meta Prometheus, Google New Albany, and Google Pryor North), so they were excluded from the analysis. From there, we used public sources to categorize projects into those that utilize behind-the-meter/onsite power as a primary source and those that rely on the grid as their primary power source (sometimes using onsite power as a backup/emergency source).

We then averaged time from the start date to the first operational phase for each category. The results are that BTM projects take an average of thirteen months, and grid projects take an average of twenty-five months.

However, additional context is important. Some of these projects were completed in 2024 and 2025; future projects will likely take longer, as soaring demand has increased the length of the interconnection queue. Epoch AI may be more likely to choose unusually large and fast projects for inclusion into their database. And some projects are not representative: xAI's Colossus 1, for example, utilized mobile natural gas generators and operated turbines before formal air permits were issued, an unusual situation unlikely to be representative of other AI data centers.

As a result, additional research served as an important supplement. For example, to arrive at our estimate, we considered quotes from industry sources outlining timelines of twelve to eighteen months for behind-the-meter projects. We considered information from sources like the IEA, news reports, and consulting firms.

This analysis—and an inclination to stay on the cautious side—led us to arrive at final estimates of eighteen months to first operation for BTM and thirty months for grid-powered projects in the United States.

Example 2: India Grid

The calculation for India's grid is more representative of many others, where there were fewer available examples and it was more difficult to find reliable data.

In these cases, we used discretion to determine which projects were the most relevant and generally weighed a few projects quite heavily. For example, in this case, we utilized two projects from Colt in Mumbai and Chennai with roughly three-and-a-half-year timelines as our primary references. They were current (with one finished in 2024, and the other currently underway), on a roughly similar scale to 100 MW (134 MW and 80 MW), and the quality of the sources was high.

We also found information on a number of older and smaller-scale projects that generally fell in the two- to three-year range, by operators including Digital Connexion, Reliance/Digital Realty/Brookfield, Yotta, Adani ConneX, and PDG.

Considering these examples and other research on the Indian grid (for example, information on the difficulty of interstate grid expansion), we arrived at a final estimate of thirty-nine months for time to power in the Indian grid.

Notes

1 Mistral AI, one of Europe's leading AI companies, has forged ahead in the meantime. In early 2026, the company invested billions into AI data centers in Sweden and France.

2 Based on its sample of AI data center projects, Epoch AI estimates that AI data centers first reached this level in mid-2024. As of March 2026, the largest data center cluster tracked by Epoch AI is an estimated 1 GW, ten times as large. Data centers as large as 2–3 GW could be built over the next five years. The financial analysis is similar for 1 GW data centers, as most AI data center costs scale with IT capacity rather than being purely fixed. The largest cost components—servers, networking, storage, and much of the engineering and construction—scale roughly with deployed compute. At large scale, costs are often modeled on a "cost per watt" basis, reflecting this proportionality. Some costs exhibit economies of scale or step-function behavior (for example, substations or interconnection), but these are typically small relative to total spend.

3 Life cycle value is the present discounted sum of all future project cash flows. It is sensitive to our valuation for calculating the Terminal Value of the data center, and should be viewed with appropriate caution. For more information, see appendix 1.

4 For the sake of simplicity, we focus on the U.S. scenario, but the results are similar for other countries. The biggest differences are driven by variance in initial costs. For example, in the United Kingdom, industrial electricity prices are about three times higher than they are in the United States. As a result, doubling electricity prices has a larger impact.

5 This depends on the rate at which chip prices decrease due to increased supply and the release of better chips. Over the life of the data center and before chips are replaced, monthly revenues could decrease into the tens of millions.

6 Most major U.S. players are hyperscalers primarily spending profits to fund AI data center builds, but some, such as Oracle, have financed them primarily with debt. Companies like OpenAI and Anthropic often serve as tenants for compute secured by investors—in their case, Microsoft and Amazon respectively.

7 Project approval refers to the stage when energy and land permits have been granted. These estimates average both grid connection and behind-the-meter strategies.

8 This estimate is driven in large part by Epoch AI's analysis of Stargate UAE. Epoch AI expects the first 1.4 GW phase of construction to be completed by January 2028 in its median scenario, or Q3 2027 in an optimistic case. This is fast given that the project was only announced and construction only commenced in mid-2025.

9 In the United States, our projections reflect typical timelines in jurisdictions where large data centers are being built, not in states where construction simply isn't happening.

10 Ironically, Germany's density of cloud data centers has in some ways harmed its competitiveness as a site for AI data centers, as the many existing data centers in hubs like Frankfurt have reduced grid capacity.

11 An industry operator explained that permitting generally proceeds in parallel at multiple sites, is conducted concurrently with planning and logistics, and is often completed by third parties that sell viable sites to hyperscaler tech companies. As a result, they believe that other parts of the construction process, like energy procurement, serve as greater bottlenecks.

12 The research organization Epoch AI projects that the first phase of Stargate UAE, a data center with 1.4 GW of power, will be completed in about 2.5 years. By contrast, a data center about half the size (720 MW) in Northumberland, England, is expected to take 9 years to become fully operational. According to permitting filings submitted to the Northumberland city council, it is expected to begin operating in 2030 and reach full completion in 2036. This example is extreme, but illustrates that the differences can be large.

13 We model taxes from India under the country's normal corporate rate rather than under the country's tax holiday for foreign-owned AI data centers. The tax holiday has not been implemented as of publication, and eligibility conditions are not yet fully clear. The relevant U.S. tax rate will also depend on how hyperscalers structure operations to take advantage. In an optimistic scenario, savings from the tax holiday could improve India's ranking above the United States.

14 Interview conducted November 7, 2025.

15 As the CEO of Constellation Energy, a major power supplier for AI data centers, told Axios, "Who's going to insure a $20 billion facility in the Middle East that can be taken out by a $5,000 drone? That's the reality we face."

16 Power use could increase in the future with the development of more dense racks, and U.S. power prices—while they vary by grid—are relatively low compared to those of other countries. A 50 percent increase in British power prices would be more significant than a 50 percent increase in U.S. power prices, because initial U.S. power prices are lower.

17 If tariffs are modest, then they may be less important.

18 We use data from Epoch AI in our model, but use this breakdown for our visualization because of the specificity of its categories. Epoch AI offers a helpful overview of other cost breakdowns, and McKinsey offers a similar breakdown by segment. In McKinsey's breakdown, hardware accounts for a lower share of spending (66 percent), perhaps because the analysis appears to include non-AI data centers. This estimate does not include land or permitting fees, both of which are small costs. Land is around 1 percent of capital expenditures, and permitting fees are a fraction of that. Even under conservative assumptions, land is a small percentage of costs. Cushman & Wakefield estimates as of October 2024 that land for data centers sells at an average of $244,000 per acre. In large purchases and nonestablished data center markets—which represent most AI data center transactions—land is substantially cheaper (see pages 10–11). A 100 MW AI data center would take up, on the high end, about 250 acres. Meta's Temple data center, which will expand to 238 MW, takes up 398 acres of land. Stream's Phoenix data center, which will expand to 280 MW, will take up 157 acres. Multiplying these numbers together yields a back of the envelope estimate of $61 million for land—about 1 percent of CapEx. Land with access to power is significantly more expensive, but that is because the price also reflects the value of the power supply. Direct permitting fees are negligible, under 1 percent of total costs; delays caused by permitting are far more consequential.

19 The finding that revenue changes are more important than location and cost changes is robust to plausible changes in the methods we use to calculate terminal value. We also calculate the unlevered internal rate of return for each data center's twelve-year operational cycle, and find that it is significantly more sensitive to changes in revenue than location or costs. Some countries, like the UAE, have attempted to subsidize demand, but it is not clear how much of a difference their efforts will make.

20 This also implies that subsidizing demand may be more useful than subsidizing supply. This helps explain why companies like Nvidia are making heavy investments in downstream AI users and the broader ecosystem.

21 The importance of power costs over time could also change as a result of evolving rack architectures. As NVIDIA releases new racks that pack more chips into a single server, the power used by each rack will rise substantially, but each rack will also contain more chips and computing power. The efficiencies that result from large, power-dense racks could actually decrease operating expenses from power, but will likely require purchasing more expensive electrical equipment upfront.

22 In 2024, Greater London's total electricity use was 34,800 GWh, implying an average hourly load of slightly under 4 GW.

23 According to the U.S. International Trade Commission, the United States imported more than $400 billion of major electrical equipment in 2025, roughly double the level from more than five years ago. The only U.S. company that makes the right kind of steel for large transformers does not produce enough for the U.S. market.

24 This speed is not without risks, however. The UAE's ambitions for a rapid AI buildout could increase outage or reliability risks if grid planners do not reinforce networks in time.

25 We use the terms "behind the meter" and "onsite" interchangeably to describe power generation that directly serves an AI data center rather than connecting to the grid first.

26 Some other data centers connect to the grid but use a smaller amount of onsite power generation as a backup source of power. Many AI data centers that use onsite generation as their primary power source now will likely transition to a hybrid approach or to using their onsite generation as backup capacity once they connect to the grid.

27 This was later decreased as a 150MW grid connection was completed.

28 Fuel cells are a notable exception. They only have a lead time of three to four months, but are significantly more expensive than other BTM power sources.

29 Such actions need not be designed to harm the country in question. If capacity is scarce and several governments are responding to a single crisis, providers may have to ration supply. Or an adversary might target data centers in third countries if it knows they serve foreign military and intelligence services. In March 2026, for example, Iran's state news agency claimed that the country's military had targeted an AWS data center in Bahrain because of its alleged role supporting the U.S. military.

30 According to the International Energy Agency, China approved nearly 100 GW of coal-fired plants in 2024, helping make 2024 the year with the most coal generation approved globally since 2015.

31 Of course, democracies aren't immune from these failures, but on average they tend to be better at airing dissent and correcting governance errors. See, generally, Amartya Sen, *Poverty and Famines: An Essay on Entitlement and Deprivation* (1983).

32 Research from GovAI agrees with our finding that the UAE does not have a substantial cost advantage over the United States in building data centers.

33 In this case, time to power and AI data center competitiveness become even more important. Computing capacity would likely be critical to relative standing within the group and help determine how much leverage middle powers have in AI discussions with the United States and China.

34 The administration has floated a goal of $1 trillion in joint investment, or more than 3 percent of U.S. GDP.

35 Authors' analysis based on Carnegie Endowment International AI Data Center model. Importantly, American companies eligible for the tax holiday would still be subject to relevant U.S. taxes on global profits.

36 For policymakers concerned that AI investment represents a bubble, accelerating time to power creates fewer risks to public finances than committing to subsidies or tax breaks.

37 Stargate Abilene's operator, Crusoe, applied for a permit from the Texas Commission on Environmental Quality on January 11, 2025, to operate ten natural gas turbines. The TCEQ granted the permit on January 22, 2025.

38 European Union Agency for the Cooperation of Energy Regulators 2024 Monitoring Report.

39 Whether these plans will provide the speed-ups developers are hoping for remains to be seen.

40 Large developers are already trying these techniques: In March 2026, Google announced that it is integrating 1 GW of "demand response" into the company's utility contracts in the South and Midwest, ramping down demand during peak hours.

41 The British grid introduced a similar reform in 2010, cutting average connection times by five years. Although the British scheme succeeded in connecting new renewable sources of power more quickly, it ran into other problems. The policy requires the grid operator to compensate power plants when it asks them to turn off generation (a practice known as curtailment) to prevent grid overloads, costing ratepayers at least a billion pounds each year, and investment in actual transmission infrastructure has lagged. ERCOT, by contrast, does not compensate for curtailment, saving money for consumers and ensuring that new projects can turn a profit on their own merits.

42 January 26–28, 2026, email correspondence on file with authors.

43 Because renewable power is intermittent, baseload equivalent power added—or the amount that can reliably run continuously—is less than nameplate solar capacity (the maximum the power source can generate). Solar panels typically require dedicated storage to provide baseload power.

44 In January 2026, the Trump administration imposed 25 percent tariffs on a narrow class of semiconductors. However, these tariffs exclude chips used in U.S. data centers among other broad exemptions.

45 Morgan Stanley predicts that of $2.9 trillion of data center spending through 2029, $1.4 trillion—about half—will be financed by hyperscaler FCF.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

1779 Massachusetts Avenue NW
Washington, DC 20036
CarnegieEndowment.org