Compute 3

KEY MESSAGES

- Compute is foundational for artificial intelligence (AI), but low-income countries (LICs) and middle-income countries (MICs) have very limited resources.
  - *AI chips and servers.* In 2024, the United States dominated supply and access to AI chips and hosted 50 percent of global secure internet servers, and other high-income countries (HICs) accounted for another 41 percent, leaving just 9 percent for the rest of the world. On a per capita basis, the United States has 200 times more servers than typical MICs and 20,000 times more than LICs.
  - *High-performance computing (HPC) systems.* As of June 2025, HICs hosted 86 percent of the world's top 500 HPC systems and 97 percent of their capacity. Excluding China and India, MICs host only 3 percent of these 500 HPC systems and account for 1 percent of the capacity, despite representing 15 percent of global gross domestic product and 48 percent of the world's population.
  - *Data centers.* As of June 2025, HICs accounted for 77 percent of global co-location data center capacity (proxied by total megawatts); upper-middle-income countries (UMICs), 18 percent; lower-middle-income countries (LMICs), 5 percent; and LICs, less than 0.1 percent.
- Compute is highly tradable across borders. Governments must weigh the cost and benefits of promoting investments in domestic data centers on the basis of country context to ensure viability, sustainability, and cost-effectiveness.
  - *Growth of cloud computing exports.* The United States is estimated to account for 87 percent of global cloud computing exports, with a staggering 23 percent compound annual growth rate in 2006–23. Most countries rely heavily on US cloud services, although Chinese cloud providers have gained traction in recent years, especially in East Asia and Pacific.
  - *Cloud computing disparities.* In 2023, HICs accounted for an estimated 84 percent of US cloud computing exports, followed by 14 percent by UMICs, 2 percent by LMICs, and 0 percent by LICs. Information and communication technology services (79 percent), financial services (73 percent), and professional services (67 percent) industries led cloud integration because of their high degree of digitalization.
  - *Weighing costs and benefits.* Well-planned local data center investment can boost economic growth, jobs, and digital sovereignty, but it requires high-quality and affordable energy and digital infrastructure and large volumes of local demand, can strain power grids, can raise land costs, and can affect the environment.

- Governments should create an enabling environment to facilitate access to compute resources, catalyzing AI adoption, adaptation, and innovation:
  - *Key barriers and externalities.* High entry costs, energy and digital infrastructure constraints, limited local demand, regulatory uncertainty, lack of cloud expertise, and coordination failures often deter private investment in compute infrastructure. Compute infrastructure generates positive externalities (for example, cross-sector spillovers that boost productivity) as well as negative ones (for example, environmental costs and national security vulnerabilities).
  - *Public investment and enabling policies.* Targeted interventions such as compute subsidies for small and medium enterprises and researchers, regional data centers, and public-private partnerships (for example, for AI training infrastructure) can lower barriers, spread access, and unlock AI's full potential while managing risks through incentives and regulations on energy efficiency and data governance.

Introduction

With a simple prompt such as "a squirrel wearing a tiny business suit, aggressively negotiating for acorns with a pigeon in a park," generative AI (GenAI) tools can immediately conjure a photorealistic image with an almost unsettling speed and detail. This magical capability of AI does not materialize from nowhere. Although AI technology is intangible, it relies on a massive physical infrastructure encompassing AI chips, data storage devices, servers, and hyperscale data centers. Whether it is traditional AI or cutting-edge GenAI, these technologies require enormous data storage and computing power to analyze and predict patterns in the data.

This chapter delves into the compute infrastructure that underpins the AI revolution. It begins by analyzing the supply side of the market, highlighting its high concentration among a small handful of big tech firms and sustainability issues. It then discusses demand patterns, comparing different options for accessing storage and compute infrastructure and the factors affecting the tradability of compute across borders. The "Compute Divide" section takes stock of compute capacity across countries, analyzes the digital divide in compute, and examines factors that affect countries' performance. The final section discusses market failures and barriers for low-income countries (LICs) and middle-income countries (MICs) to access compute, the role of the government, and the pros and cons of governments' existing approaches to address the compute barrier.

Supply: Market structure, growth, and sustainability

The compute layer of the artificial intelligence (AI) ecosystem consists of many interrelated hardware, software, and data communication components. The market segments and supply chains are highly complex and constantly evolving because of rapid technological advancements and diverse application needs. New players, products, and business models regularly emerge, and the interplay between hardware advancements and software optimization adds layers of intricacy. This report defines *compute* as the ability to store, process, and transfer data at scale (Tony Blair Institute for Global Change 2023).

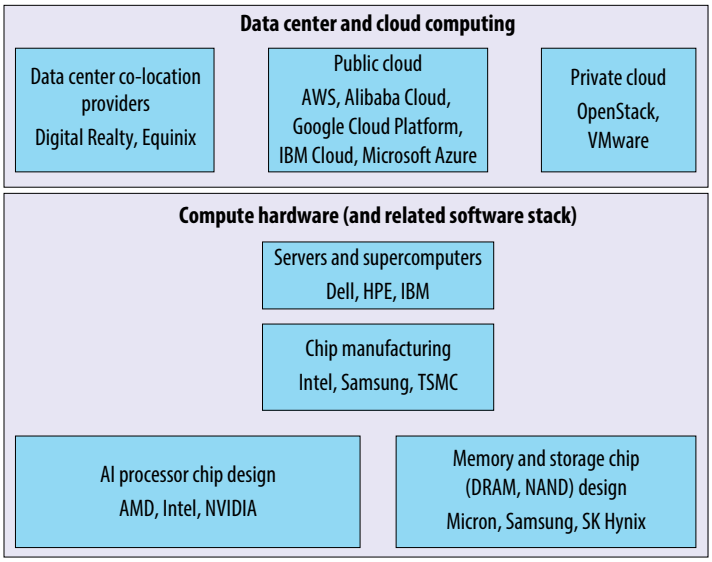
Figure 3.1 shows the market segments and key companies involved in compute. Compute is monopolized at key points in the supply chain by one or a small handful of firms. Industry concentration acts as a shaping force in how computational power is manufactured and accessed.

At the heart of AI hardware lies processor chips—the brain that carries out the computations. The term *AI chip* is broad; it includes many kinds of chips designed for the demanding compute environments required by AI tasks. Examples of popular AI chips include graphics processing units (GPUs), tensor processing units, neural processing units, field-programmable gate arrays, and application-specific integrated circuits.

The AI processor chip market has experienced rapid growth in recent years and is projected to expand 10-fold over the next decade, fueled by surging demand for AI-powered technologies across industries (Edge AI + Vision Alliance 2024). The AI chip market is highly concentrated, although competition is intensifying. NVIDIA controls 70–95 percent of the market for AI chips (Leswing 2024). It leads the data center GPU market by a long shot, with a staggering 92 percent market share as of 2023. NVIDIA's enduring leadership can be attributed to its first-mover advantage, as well as its well-established and optimized CUDA software ecosystem that provides a comprehensive development environment for GPU-accelerated computing. However, both large and small businesses are actively searching for alternatives to diversify their AI chip options.

The public cloud computing sector is dominated by "hyperscalers" such as Alibaba Cloud, Amazon Web Services (AWS), Google Cloud Platform, IBM Cloud, and Microsoft Azure. These providers offer a variety of cloud-based solutions, enabling businesses and governments to rapidly scale their operations and access cutting-edge technologies without significant upfront investments in hardware and infrastructure. For organizations preferring more control over their cloud environments, private cloud solutions are available from vendors such as VMware and the open-source OpenStack project. These solutions provide alternatives for on-premises or dedicated cloud deployments.

FIGURE 3.1 AI compute market segments and key companies



Source: Original figure for this publication.
Note: AI = artificial intelligence; DRAM = dynamic random-access memory; NAND = not and.

An increasingly common approach is the hybrid cloud, which combines public and private cloud resources to balance scalability, cost-efficiency, and control. Effective hybrid cloud strategies often involve collaboration between public and private cloud vendors, enabling interoperability, unified management tools, and shared innovation pipelines that give organizations greater flexibility and resilience.

The global cloud market is similarly experiencing rapid expansion and consolidation, bringing both benefits and challenges to the AI ecosystem. Goldman Sachs (2024) estimated that the cloud computing market will expand with a 22 percent annual growth rate from 2024 to 2030, reaching US$2 trillion by the end of the decade. AWS, Microsoft Azure, and Google Cloud collectively command two-thirds of the market share (Richter 2025). High entry barriers, network effects, and economies of scale jointly result in high market concentration, which can produce effects that are both socially beneficial and harmful.

The size and sophistication of leading cloud providers benefit their customers through improved efficiency, seamlessly integrated platforms, and advanced cybersecurity capabilities, including for incident detection and response. However, high market concentration may also reduce competition, create vendor lock-in, give unfair privileges to first-party applications over third-party equivalents, and pose systemic risks to vital sectors.

The computational resources required for AI models have been growing explosively. The amount of compute used is measured in floating point operations, and compute performance is measured in floating point operations per second. Before the deep learning era, the amount of compute used to train AI models doubled in about 24 months; when deep learning took hold around 2010, the amount of compute required started doubling every 6 months ("The Race Is On to Control the Global Supply Chain for AI Chips" 2024). Larger training data sets, more complex model architecture, and parallel processing all require a growing scale of training computation within AI, despite improvements in computing efficiency brought by new generations of chips. Once trained, applying a learned model to make predictions or decisions on new, unseen data, called *inference*, could eventually require even more compute as usage grows over time.

Provisioning the immense compute infrastructure at the scale required for leading AI research is expensive in both upfront capital expenditures and ongoing operating costs, raising questions about the sustainability of current AI development directions. New Street Research estimates that Alphabet, Amazon, Meta, and Microsoft together invested US$104 billion in building AI data centers in 2024. When spending by smaller tech firms and other industries is added in, the total AI data center investment between 2023 and 2027 could reach US$1.4 trillion ("What Could Kill" 2024).

The training process for GPT-4 incurred an estimated cost of more than US$100 million (Vipra and West 2023). The hefty training costs present a high barrier to entry in the AI model market, favoring large and well-founded companies. The cost may eventually translate to higher prices for users of generative AI (GenAI) tools and application programming interfaces. However, the Chinese start-up DeepSeek has demonstrated that training costs for advanced AI models can be slashed substantially through optimized techniques.

If not managed properly, the growing energy and cooling demand from compute could strain existing energy and water infrastructure and exacerbate environmental impacts. As the compute demand grows, the energy consumed to train models increases accordingly. It is estimated that a single ChatGPT query consumes 25 times more energy than a Google search query (Brussels Times Newsroom 2024).

The training and application of these massive AI models also generates a prodigious heat output that poses cooling challenges. A lot of water is used in cooling the servers that run AI applications. Half a liter of water evaporates for each ChatGPT conversation of about 29–50 queries (Brussels Times Newsroom 2024). These GenAI tools are used billions of times every month, so the water usage quickly adds up to huge amounts.

The private sector has undertaken various efforts to mitigate the environmental impacts of AI. These include improving the energy efficiency of hardware, optimizing training data (algorithms) to reduce computational requirements, and increasing the use of renewable energy sources. Companies are ramping up investments in renewable energy and developing energy-efficient data centers. Some companies are also exploring nuclear power as part of a broader energy mix to meet growing demand. However, views differ on whether nuclear power can scale quickly enough to address the sharp rise in energy needs anticipated this decade (Norton 2024).

Governments are also closely monitoring the development of data centers and contemplating regulations to address sustainability issues. Ireland, the Netherlands, and Singapore have temporarily paused new data center construction to reduce the strain on power grids. Singapore rolled out a new sustainability standard in 2023 as a means of lowering energy consumption (JLL n.d.). In 2024, the European Union (EU) adopted an EU-wide scheme for rating the sustainability of data centers (European Commission 2024).

Organizations operating data centers in EU nations will be required to file reports detailing water and energy consumption, as well as steps they are taking to reduce it. This scheme is intended to increase transparency and potentially to promote new designs and efficiency developments in data centers that can not only reduce energy and water consumption but also promote the use of renewable energy, increase grid efficiency, or the reuse of waste heat in nearby facilities and heat networks.

Demand: Compute access modality and trade-offs

Two ways exist to access compute infrastructure: purchasing or renting. Some organizations, particularly leading AI companies such as Alphabet, Meta, Microsoft, OpenAI, and XAI that provide foundational AI models, choose to purchase AI chips outright to train AI models. A key advantage is having exclusive access to cutting-edge hardware as soon as it becomes available. However, the upfront capital costs of procuring hardware and building and expanding data centers are massive.

Significant operational complexity also is involved in managing clustered resources, scheduling jobs, and so forth. Frequent hardware refresh cycles are required, given the rapid pace of innovation. Furthermore, owned infrastructure lacks the flexibility to easily burst to additional capacity during demand spikes. Organizations must manage their own backup and disaster recovery strategies.

For most other organizations, the primary option is to rent compute and storage capacity from cloud providers on a pay-as-you-go basis. Public cloud offerings provide multitenant shared infrastructure delivered as a service. Private clouds keep the environment dedicated, and they can be hosted internally or by a vendor. Hybrid clouds allow combining private and public cloud with data and application portability between them.

One primary benefit of cloud is low upfront costs. Scalability to instantly burst to higher capacity is another key advantage. The cloud provider manages the underlying infrastructure. Geographic flexibility is enabled through global data center footprints, reducing latency. Cloud computing also reduces management overhead and offers access to advanced services and tools that can be easily

integrated into applications. Potential downsides of cloud include high costs at hyperscale over long periods, concerns about vendor lock-in to proprietary services, limited control and customization, and data egress charges for high-volume workloads.

Different types of users have distinct demands for data storage and compute. Purchasing is often favored by organizations with strict regulatory compliance, unique performance demands, stable and predictable workloads, and the resources to manage their own infrastructure. These typically include leading foundational AI model developers and researchers with heavy AI workloads.

Cloud computing is generally recommended for businesses seeking flexibility, scalability, cost-efficiency (especially for variable workloads or experimentation), rapid deployment, access to advanced AI services, and reduced operational overhead. Large enterprises and government agencies often require substantial computing power to manage and process vast amounts of data, run AI applications, and support daily operations. They often leverage hybrid, dynamic models, mapping different workload profiles to the optimal deployment model to balance agility, cost, control, data security, and performance.

For individual workers, consumers, and small businesses, AI is primarily accessed through software platforms and cloud services such as ChatGPT and Microsoft Copilot. As end users, their experience is highly dependent on their internet speed, these platforms' hardware, and data center proximity. AI application providers such as OpenAI rely on cloud providers for hosting AI models and services on a global scale. This allows seamless access for individuals worldwide.

However, latency remains a factor—the farther a user is from a data center, the slower and more degraded their AI experience may be. For applications requiring real-time responsiveness, such as live speech recognition or robotics control, users will get a smoother experience when they are geographically close to the AI system's hosting region. Major cloud providers are continuing to build out larger global data center footprints to reduce these proximity constraints over time. Increasingly, individuals and small businesses can also access AI on their edge devices, such as smartphones, wearables, cars, and smart home devices.

Compute divide

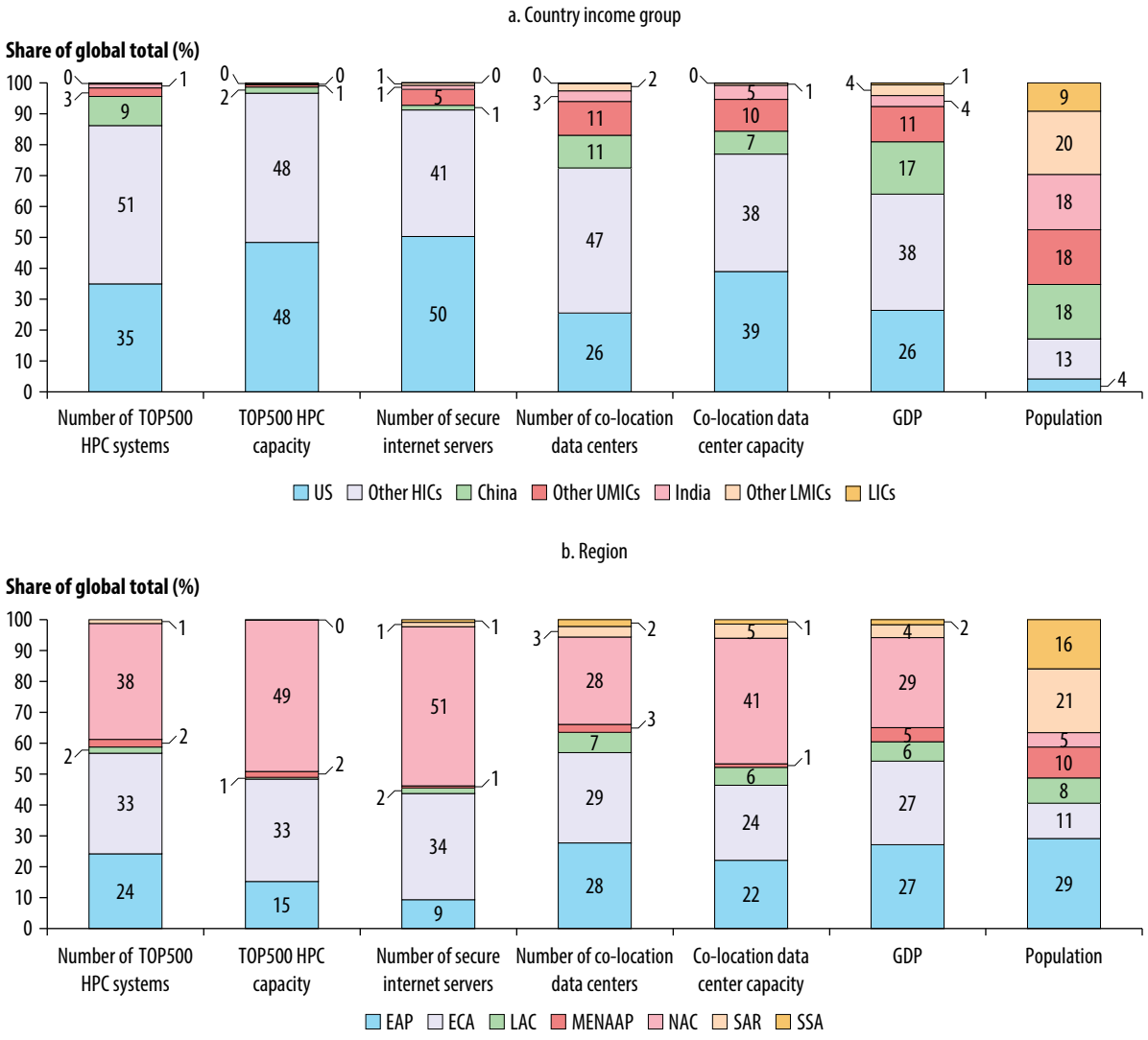
Data on countries' performance on AI chip ownership, data center capacity, and cloud adoption are scarce because of several factors. First, the rapid pace of technological advancement and competitive nature of the AI and cloud industries lead companies to closely guard detailed information about their infrastructure and capabilities as proprietary and strategic assets. Second, standardized reporting and measurement frameworks across countries and organizations are lacking, making it difficult to aggregate and compare data consistently. Third, geopolitical considerations and national security concerns may limit the transparency and availability of information, because nations view technological prowess in AI and cloud computing as critical to their strategic interests. Finally, the nascent and evolving nature of these technologies means that comprehensive data sets and tracking mechanisms are still developing, resulting in fragmented and incomplete data availability on a global scale.

This section combines data on supercomputers from TOP500, co-location data center capacity data from TeleGeography, and cloud computing adoption data to shed light on countries' performance on data storage and compute infrastructure. Given the incomplete information described previously, the findings presented here are just indicative, although the simple stocktaking still reveals a highly uneven distribution of compute access.

The United States has the highest raw compute capacity and the most-powerful supercomputers, is the global leader in AI chip production and access, and dominates the global cloud

computing market. As of June 2025, the United States hosted 175 of the top 500 high-performance computing (HPC) systems, the largest number in the world, and accounts for about 50 percent of TOP500 supercomputing capacity,[1] 50 percent of global secure internet servers, 26 percent of co-location data centers, and nearly 40 percent of co-location data center capacity (refer to figure 3.2), relative to its 27 percent of global gross domestic product (GDP) and 4 percent of population. It dominates GPU production and use, accounting for nearly half of NVIDIA's global revenue in the fiscal year ending in January 2025. US big tech companies such as Google, Meta, Microsoft, and XAI, own millions of AI chips, including NVIDIA's H100 chips, dwarfing the ownership of any other company in the world (Tony Blair Institute for Global Change 2023).

FIGURE 3.2 Compute capacity, by country income group and region



Sources: Original figures for this publication using June 2025 TOP500 data (https://top500.org). HPC compute capacity is measured with Rmax in tera floating point operations per second. Secure internet server data are from the World Development Indicators, World Bank (2024; https://databank.worldbank.org/source/world-development-indicators). Co-location data center data are from TeleGeography (June 2025; https://www2.telegeography.com/). Co-location data center capacity is measured with total megawatts.
Note: EAP = East Asia and Pacific; ECA = Europe and Central Asia; GDP = gross domestic product; HICs = high-income countries; HPC = high-performance computing; LAC = Latin America and the Caribbean; LICs = low-income countries; LMICs = lower-middle-income countries; MENAAP = Middle East, North Africa, Afghanistan, and Pakistan; NAC = North America; SAR = South Asia; SSA = Sub-Saharan Africa; UMICs = upper-middle-income countries; US = United States.

European countries, notably France, Germany, Italy, the Netherlands, and the United Kingdom, also boast substantial computing power, although they face a scarcity of large domestic cloud providers and AI chip vendors. Central Asia and Europe host one-third of global HPC systems, capacity, and internet servers, along with 28 percent of co-location data centers and 22 percent of data center capacity, relative to their 27 percent share in global GDP and 11 percent of the global population. Most of these compute resources are concentrated in EU countries, particularly the five mentioned. Germany alone is home to 41 of the top 500 HPCs, 10 percent of secure internet servers, and 5.5 percent of co-location data centers.

Despite this robust infrastructure, European countries have limited large public cloud providers and AI chip vendors. OVHcloud, a French company and the largest public cloud firm, reported an annual revenue of US$1 billion in 2024, which pales in comparison with the revenues of global giants such as Alibaba Cloud (US$16 billion), Google Cloud Platform (US$49 billion), Microsoft Azure (US$107 billion), and AWS (US$117 billion). Europe also lacks its own AI accelerator chips (Hawkins, Lehdonvirta, and Wu 2025).

China's compute capacity is strong and growing. China hosts 47 HPC systems—9 percent of the world's total—and 2 percent of supercomputing capacity, although this figure is likely an underestimate. China accounts for only 1 percent of global internet servers, 11 percent of co-location data centers, and 7 percent of co-location data center capacity, compared with its 17 percent of GDP and population share. China also accounted for 13 percent of NVIDIA's revenue in fiscal year 2024/25, although that share has been falling because of trade frictions with the United States.

Excluding China, LICs and MICs host a mere 3 percent of the 500 HPC systems, 1 percent of the supercomputing capacity, 7 percent of the secure internet servers, 17 percent of co-location data centers, and 16 percent of corresponding capacity, despite being home to 65 percent of the world's population and contributing 19 percent of global GDP. High-income countries (HICs) account for 86 percent of the top 500 HPC systems and a staggering 97 percent of the capacity. Among MICs, Brazil hosts 9 of the leading HPC systems, India hosts 6, and Thailand and Türkiye each host 2. Argentina, Morocco, and Viet Nam also recorded entries in the TOP500 list in 2025. However, the performance of supercomputers in MICs lags significantly behind frontier systems.

Beyond HPCs, Brazil, India, Indonesia, the Islamic Republic of Iran, Malaysia, South Africa, and Türkiye also have a decent number of secure internet servers and co-location data centers. Many of these nations serve as regional compute hubs, with thriving local digital economies, large market size, and business environments that attract data center investments. The compute situation is challenging for Sub-Saharan Africa (SSA) and LICs. SSA accounts for 16 percent of the global population but contributes only 2 percent of global GDP, 1 percent of secure internet servers and co-location data center capacity, and 0 percent of HPC systems. Nine percent of the world's population resides in LICs, where they account for 0.3 percent of global GDP and 0 percent of compute capacity.

Among the 4Cs (connectivity, compute, context, competency), compute is currently the most-traded element. International broadband connectivity and cloud computing offer a partial substitute for domestic data centers, creating a trade-off influenced by several factors. These include the specific technical design of the compute resources, internet latency, data transmission costs, and crucial privacy and data sovereignty considerations.

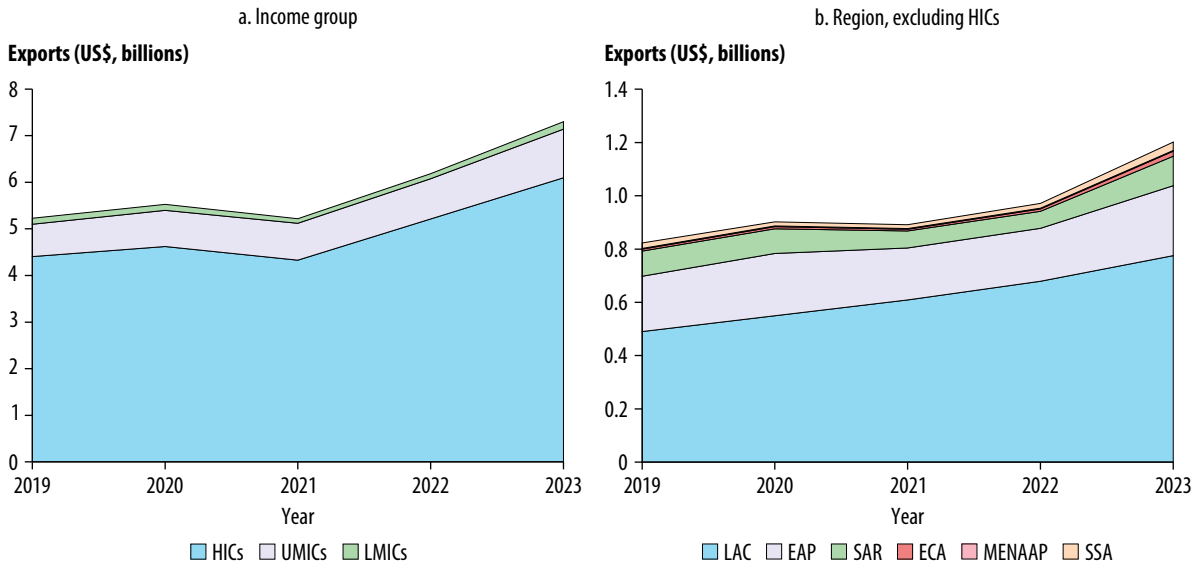
The United States contributes 87 percent of global cloud computing and data storage services exports, and during 2006–23 its exports grew annually at 23 percent. On the basis of cloud computing trade data compiled from corporate annual reports, the United States consistently accounted for 87 percent of global exports during 2016–21. China's cloud computing exports have experienced a 10-fold increase from 2016 to 2021, and its share in global exports jumped from 1 percent in 2016 to 6 percent in 2021, whereas Germany's share declined from 10 percent to 6 percent. Except for these three countries, all other economies combined account for only 1 percent of global cloud computing exports.

On the basis of World Trade Organization statistics, US cloud computing and data storage services exports grew 23 percent annually since 2006, reaching US$8.3 billion in 2023. However, this amount only represented 4 percent of the three big cloud providers' revenue.[2]

US cloud computing exports are predominantly destined for HICs (approximately 84 percent), where demand concentrates. LICs and MICs, particularly in Latin America and the Caribbean (LAC), represent a growing, although still small, market. The combined imports of France, Germany, and the United Kingdom alone rival the total share of all developing countries. Upper-middle-income countries contributed 14 percent of US cloud computing exports in 2023, whereas lower-middle-income countries contributed 2 percent and LICs, 0 percent.

In the developing market, LAC stands out as the leading destination for US cloud computing exports, exhibiting the highest import growth rate between 2019 and 2023 (refer to figure 3.3). Brazil and Mexico together account for a substantial one-third of all US cloud computing exports to developing countries. This is largely driven by LAC's relatively mature digital economy, which fuels demand for cloud computing, and its geographical proximity to the United States. Notably, China represents just 2 percent of US cloud exports, largely because of the dominance of domestic cloud providers such as Alibaba Cloud, Huawei Cloud, and Tencent Cloud.

FIGURE 3.3 US cloud computing exports, by destination country income group and region, 2019–23



Source: Original figures for this publication, based on calculations using World Trade Organization stastistics (https://stats.wto.org/).
Note: The sum of exports by income group is below US total cloud computing exports, as bilateral exports below certain value thresholds are not reported. EAP = East Asia and Pacific; ECA = Europe and Central Asia; HICs = high-income countries; LAC = Latin America and the Caribbean; LMICs = lower-middle-income countries; MENAAP = Middle East, North Africa, Afghanistan, and Pakistan; SAR = South Asia; SSA = Sub-Saharan Africa; UMICs = upper-middle-income countries.

Governments must weigh whether promoting local data center investment—through incentives, public–private partnerships (PPPs), or other measures—offers better value than relying on regional hubs or international cloud services.

Well-planned domestic data centers can drive local economic development, create jobs, catalyze structural transformation, and boost tax revenues. Northern Virginia, the densely populated area located just outside of Washington, DC, the US capital, has the highest concentration of data centers in the world, with around 200 data centers built and 117 in the development pipeline (Turner 2025). The data center industry in Virginia created nearly 15,000 direct jobs, US$1.9 billion in associated pay and benefits, and $4.5 billion in economic output in 2018 alone. Considering the economic ripple effects generated by direct investment generated, the job creation amounts to more than 45,000, with US$3.5 billion in associated pay and benefits and US$10.1 billion in economic output. Moreover, data centers generated US$600.1 million in state and local tax revenue (Northern Virginia Technology Council 2020).

In China, the mountainous province of Guizhou in the southwest has become a pioneer in China's big data industry. Guizhou, a city with one of the largest number of super-large data centers globally, had 37 big data centers in operation or under construction as of early 2023. Data center investments have powered Guizhou's digital economy development, achieving a top growth rate in China for 7 consecutive years since it became the big data pilot area in 2016.

However, these gains come with trade-offs. Data centers are highly capital and energy intensive, placing heavy demands on local power grids and land markets. Supporting data centers through tax exemptions could attract investment but might also affect electricity prices and tax revenues, potentially crowding out economic activity in labor-intensive sectors (Johansson and Kriström 2021). With the rising prices that result from demand for data center land, nearby commercial real estate and housing become less affordable for residents.

In Frankfurt, Germany, co-location data centers led to increased demand for commercial spaces, causing price hikes (Savvas 2021). To address this issue, in 2022 Frankfurt published an updated plan for commercial developments and constrained data center construction to certain areas (Judge 2022).

For developing countries, the challenge is to ensure that such investments are sustainable, wellutilized, and aligned with broader development goals. Viability depends on stable and affordable electricity, high-quality internet connectivity, competitive operating costs, and sufficient domestic demand. Other key factors include the scale and growth potential of the digital economy, the robustness of information and communication technology (ICT) skills and maintenance capacity, legal and security requirements for certain data sets, the types of data and workloads to be hosted locally, and the potential for integration with regional or global cloud networks. Careful planning, competitive neutrality, and clear sustainability strategies are essential to ensure that domestic data centers serve as enablers of the digital economy rather than costly, underused infrastructure.

How can countries achieve secure, equitable, and sustainable access to compute?

Compute has become a key bottleneck in both the training of large-scale AI models and the deployment of AI across sectors. Compute power will play a crucial role in the future economy

by supporting AI, automation, and data-driven innovation, enhancing productivity and efficiency across industries. The market alone cannot ensure equitable access or distribution, because compute supply is currently tightly controlled by a few dominant players, creating significant barriers to entry and innovation (Azoulay, Krieger, and Nagaraj 2024).

LICs and MICs often lag in access to compute, through either domestic data centers or cloud computing, limiting their ability to train and deploy AI models at scale. This section summarizes key barriers and market failures in LICs and MICs in developing local compute capacity and accessing compute and discusses different approaches to address these challenges and the pros and cons of each approach.

High market concentration, high cost, and supply constraints

High upfront costs associated with procuring AI chips and establishing data centers are prohibitive for many LICs and MICs. Cutting-edge AI chips, such as NVIDIA's Blackwell GPUs, cost between $30,000 and $40,000 per unit (Kim 2024). Large-scale AI data centers can reach costs in billions of US dollars (Haranas 2024). Those who need to access cutting-edge compute in developing countries but cannot afford to buy AI chips often turn to cloud computing, but even this option is pricey and often unavailable. For instance, AWS charges US$98 per hour to rent a virtual machine powered by an NVIDIA H100 chip, compared with US$0.004 per hour for its least-expensive non-AI virtual machines.

These exorbitant costs are compounded by limited access to investment capital in developing countries. Domestic investors may lack the resources to fund compute infrastructure, and foreign investors are often deterred by perceived risks, including political instability, economic uncertainty, and regulatory unpredictability.

The challenges of supply and affordability are deeply intertwined with the concentration of the compute supply chain. The production of AI chips is dominated by a few companies with highly specialized expertise, and the entire supply chain—from chip design to fabrication and distribution—is controlled by a few key players. Network effects, increasing returns on scale and scope, and winner-takes-most dynamics also entrench leading AI chip and cloud providers' advantage and make it challenging for new players to enter. This concentration not only limits supply but also reinforces pricing power, keeping AI chips out of reach for many developing countries.

Encouragingly, emerging models such as DeepSeek demonstrate that high AI performance can be achieved with lower compute demands, and their open-source approach offers potential to democratize access. This could reduce costs and make AI development more feasible for resource-constrained countries.

Negative externalities: Environmental costs and national security

Data centers consume huge amounts of electricity and water. This usage contributes to carbon emissions and strains local resources, and the costs are not fully reflected in service prices. E-waste from server turnover can also be substantial. Control over AI chips has become a major geopolitical tool and source of friction. Countries that rely on foreign data centers also subject the data of their citizens and organizations to foreign laws and regulations, including sensitive information, increasing the risks of unauthorized foreign access, surveillance, and national security.

Positive externalities: Research and development and cross-sector innovation

Investments in compute infrastructure can unlock broad innovation benefits—improving productivity across sectors, boosting research capacity, and supporting local start-ups. Yet these public goods characteristics mean private investors may underinvest without public support, especially in markets where returns are uncertain or long-term.

Underdeveloped energy and internet infrastructure, coordination failure, and limited demand

In many developing countries, limited or unreliable energy and internet connectivity restricts the potential utilization of cloud services, diminishing the incentive for investment. In addition, either energy infrastructure is unreliable or the cost of energy is prohibitively high, making it challenging to operate data centers efficiently. A study by the International Energy Agency (2020) found that more than 60 percent of developing countries face significant energy security challenges. This can lead to increased costs, downtime, and even project cancellations. Slow digital transformation due to coordination failures, small market size, and low purchasing power in some developing countries may not justify significant investments in data centers, especially when compared with potential returns in higher-income markets.

Lack of technical expertise and brain drain

The scarcity of local technical expertise in cloud technologies and data center management exacerbates the challenge. Talent requirements grow as compute complexity and costs grow, because specialized knowledge is needed to make the most of scarce hardware, and much of this knowledge is tacit. Yet, many developing countries lack the resources needed to cultivate a robust pool of local talent, because students often face limited access to advanced technical training, underresourced educational institutions, and insufficient digital infrastructure.

Moreover, even when local talent is developed, these countries face intense competition from advanced economies, often resulting in brain drain as trained professionals seek better opportunities abroad. This lack of technical expertise can lead to difficulties in maintaining and managing data centers and cloud infrastructure, which can result in downtime, security breaches, and other issues.

Regulatory constraints and uncertainty

Limited regulatory readiness and uncertainty can significantly deter cloud computing investment and adoption in developing countries. Clear, predictable, and modern regulatory frameworks are essential for both accessing cloud services and attracting providers to establish local infrastructure. Policies on cloud or data hosting, and data protection laws that specify legal mechanisms for cross-border data flows, directly affect where and how data can be stored and processed. Without such clarity, cloud service providers (CSPs) face legal and operational risks, and potential users—particularly in regulated sectors such as finance and health care—may be reluctant to migrate to the cloud. Outdated procurement rules can also block public agencies from leveraging cloud offerings, reducing a major source of demand that could anchor local markets.

Beyond access, measures such as intermediary liability frameworks and transparent CSP licensing regimes are important to encourage market entry, ensure a level playing field, and foster competition. In their absence, uncertainty over legal responsibilities, market access, and cost structures can delay

investment decisions, limit service availability, and ultimately slow the development of the compute market itself.

How can countries address the compute barrier?

This section provides insights into how countries can address the compute barrier and enable equitable access. It summarizes some existing approaches, discusses their pros and cons, and discusses priorities for countries at various levels of AI readiness.

Large countries and countries relatively advanced in AI

Countries at the forefront of AI development need to secure adequate computing resources to maintain their competitive advantage, foster continued innovation, safeguard national security, and drive economic growth. Governments can invest in national supercomputers and AI research facilities to provide a shared infrastructure for researchers and companies to access HPC resources. For example, the United Kingdom has announced £300 million in funding for two supercomputers to promote AI research (Reuters 2023).

However, hefty costs and lack of expertise often constrain governments' ability to build their own supercomputers in developing countries. To secure adequate compute for a thriving AI ecosystem and facilitate cutting-edge research, governments can partner with leading technology companies. For instance, India has collaborated with NVIDIA to establish the NVIDIA AI Technology Center and support the country's AI research efforts ("IIT Hyderabad, NVIDIA Establish First AI Research Center in India" 2020). The Indian government is also planning to partner with NVIDIA to offer affordable GPUs to local start-ups, researchers, and academic institutions (Vivek 2024). Such PPPs provide AI start-ups and researchers with access to compute resources, accelerating innovation.

Several countries have implemented compute voucher initiatives to provide free or subsidized cloud computing credits to AI researchers and start-ups. The Republic of Korea's AI Computing Cloud Voucher program is a notable example, granting start-ups up to US$33,000 worth of credits on local cloud platforms (Ministry of Science and ICT, Republic of Korea 2020). Multiple Chinese city governments have also pledged to provide computing vouchers worth the equivalent of US$140,000–US$280,000 to subsidize AI start-ups facing rising data center costs. These can be used for time in AI data centers to train and run the companies' large language models and to perform other tasks ("China Offers AI Computing 'Vouchers' to Its Underpowered Start-ups" 2024).

Although PPPs and targeted subsidies can help democratize access to compute to a certain degree, such programs are costly to scale and maintain over time. Although such initiatives can alleviate financial pressures on AI start-ups and researchers, they effectively transfer the financial burden to government budgets and, ultimately, taxpayers. Given that public funds are often constrained by competing priorities, these compute access programs risk being underfunded or inconsistently supported. The eligibility criteria for these programs are typically stringent, potentially skewing benefits toward well-established start-ups, prestigious universities, and communities that already have some level of compute access. Moreover, these measures do not directly tackle the underlying issue of scarcity. As such, subsidies and PPPs are only short-term strategies rather than long-term solutions to the compute scarcity problem.

No consensus exists on whether the market concentration in several segments of the compute ecosystem is too high. However, concerns are growing about how the dominant firms are limiting opportunities for newcomers. Big tech companies are aggressively pursuing investments and alliances with AI start-ups through their cloud computing arms, raising regulatory questions over their role as both suppliers and competitors in the battle to develop AI ("Big Tech Companies Use Cloud Computing Arms to Pursue Alliances with AI Groups" 2023). AI start-ups that need to train models have little choice but to rush into the arms of large companies offering essential cloud computing. Big tech companies may use their power in compute services to stifle competition in AI by giving discriminatory preference to their own AI products over new entrants and competitors.

It is critical for regulators to monitor competition dynamics in various segments of the compute market, scrutinize vertical integration, and ensure market contestability and nondiscrimination. In 2022, the US Federal Trade Commission prevented NVIDIA's proposed acquisition of Arm to stimulate innovation. The United Kingdom's Ofcom is also investigating concentration in the cloud market. France, Japan, Korea, and the Netherlands are also exploring antitrust cases in cloud computing (Vipra and West 2023).

Governments should conduct thorough cost-benefit analyses to evaluate the economic impact of data center incentives against factors such as job creation, tax revenues, and long-term digital transformation benefits. Regulatory oversight is needed to prevent data center proliferation from excessively straining electricity grids. Policy makers must strike a balance between fostering innovation and controlling energy costs. Robust climate policies are also essential to ensure sustainability targets are met as AI infrastructure grows.

Small countries with good digital infrastructure

Small nations can achieve economies of scale and risk diversification by collaborating on shared regional data center facilities, cloud platforms, and connectivity infrastructure. Concepts such as data embassies allow countries to maintain legal control over data hosted abroad. Estonia established a data embassy at a data center located in Luxembourg in 2017. The two nations struck an agreement to afford the facility the same rights as a conventional embassy. Estonia owns the hardware and uses it as a backup for its digital services and critical data sets (Meyer 2022). Monaco created a data embassy in Luxembourg in 2021 because Monaco's territory is too small to diversify the risk of natural disaster. Australia is also exploring a data embassy for Pacific nations so that these small island nations, who are unable to attract private sector investment in data centers, can use Australian data centers without compromising their sovereignty (Sharwood 2024).

However, challenges with diplomacy, data sovereignty, geopolitics, and equitable cost sharing must be navigated. Establishing more data embassies will require either a change to the Vienna Conventions that govern diplomatic and consular relations or bilateral agreements. To date, only the latter have been created. Deep trust is required among home and host governments, the private sector company, and the country in which the provider is headquartered to ensure the success of data embassies. Comprehensive risk assessments should be conducted to establish acceptable risk tolerance and associated requirements.

Accelerating government cloud adoption can stimulate local demand for AI services. This not only advances e-governance initiatives but also signals policy priorities, incentivizes private sector adoption, and creates opportunities for domestic cloud providers to emerge over time. For example,

Thailand's Ministry of Higher Education, Science, Research, and Innovation partnered with AWS in 2023 to accelerate the migration of its government's workload to the cloud. AWS will also provide training programs to upskill public servants (AWS Sector Blog Team 2023).

To fully realize the benefits of cloud computing, governments need to invest in building local cloud expertise and skills. Although hyperscalers often offer cloud credits—targeting start-ups, researchers, or government innovation programs—these resources frequently go underused in developing countries. A key reason is the widespread shortage of technical cloud skills needed to design, deploy, manage, and scale AI or other data-intensive workloads. Effective use of cloud infrastructure requires expertise in areas such as cloud architecture, cybersecurity, data governance, application migration, and cost optimization. Without this capacity, organizations may lack the confidence or ability to integrate cloud services into their operations, leaving credits unused and opportunities missed.

Governments can help close this gap by integrating cloud skills into national digital skills strategies, supporting specialized training for ICT professionals, and creating hands-on learning opportunities in partnership with cloud providers. Building a pool of cloud-literate engineers, developers, and policy makers ensures that financial incentives such as cloud credits translate into real adoption, innovation, and economic impact.

Governments can promote cloud computing investment and adoption by strengthening and modernizing their regulatory frameworks, with a focus on clarity, predictability, and alignment with international good practices. Robust data governance frameworks—covering data classification, privacy, security, and mechanisms for cross-border data flows—provide certainty for both CSPs and users, especially in sectors that handle sensitive information. Clear and transparent CSP regulations, including licensing processes, service quality standards, and security requirements, help ensure a competitive and trusted marketplace while reducing entry barriers for providers. Modernizing related areas, such as public procurement rules, can enable government agencies to leverage cloud services directly, creating anchor demand that can stimulate local market growth. In parallel, adopting intermediary liability frameworks and streamlined compliance processes can further encourage CSP investment. By combining legal clarity with pro-competition policies and safeguards for privacy and security, governments can build the trust and market conditions necessary for a thriving cloud ecosystem.

Countries lacking basic digital infrastructure

For nations lacking foundational connectivity infrastructure, prioritizing investment in high-speed broadband access for households and businesses is crucial. This enables cloud service adoption for individual users and small and medium enterprises, allowing them to access AI applications without major domestic data center investment upfront. Fiber optic networks that provide diverse paths supporting high-capacity bandwidth and low-latency and resiliency requirements are critical in enabling AI adoption in these countries. In Africa, although a growing number of underwater cables connect the continent with the wider world, a dearth of onshore lines to carry data inland leaves much of that internet capacity wasted.

Although the paths vary depending on existing capabilities, a proactive, collaborative multistakeholder approach involving governments, industry, and global institutions is critical for countries to ensure sufficient compute resources while mitigating risks and negative effects.

Notes

1. TOP500 has compiled a list of the world's 500 fastest supercomputers twice a year since 1993. The performance of a supercomputer is measured by the number of tera floating point operations it can carry out per second. The sum of the performance of all supercomputers in a country is typically used to measure the country's supercomputer capacity.
2. The cloud computing exports from balance-of-payments data could be significantly underestimated because of several factors: (1) Statistical agencies struggle to accurately capture the value of cloud computing services because of a lack of standardized definitions and reporting frameworks; (2) determining the true location of the transaction and thus export when data are stored in distributed data centers across the globe and accessed by users worldwide is challenging; (3) cloud computing is often embedded within other services or products; and (4) multinational cloud providers often have subsidiaries and data centers in multiple countries. Transactions between these entities may not be fully captured in export statistics, even though they represent cross-border service provision. Also, some cloud services are offered for free or in exchange for data.

References

AWS Sector Blog Team. 2023. "Ministry Collaborates with AWS to Boost Innovation and Upskill Public Servants." *AWS Blog*, February 1, 2023. https://aws.amazon.com/blogs/publicsector/thailands-higher-education-science-research-innovation-ministry-collaborates-aws-boost-innovation-upskill-public-servants/.

Azoulay, P., J. L. Krieger, and A. Nagaraj. 2024. "Old Moats for New Models: Openness, Control, and Competition in Generative AI." Working Paper 32474, National Bureau of Economic Research, Cambridge, MA.

"Big Tech Companies Use Cloud Computing Arms to Pursue Alliances with AI Groups." 2023. *Financial Times,* February 5, 2023. https://www.ft.com/content/5b17d011-8e0b-4ba1-bdca-4fbfdba10363.

Brussels Times Newsroom. 2024. "ChatGPT Consumes 25 Times More Energy Than Google." *Brussels Times,* May 12, 2024. https://www.brusselstimes.com/1042696/chatgpt-consumes-25-times-more-energy-than-google.

"China Offers AI Computing 'Vouchers' to Its Underpowered Start-ups." 2024. *Financial Times,* March 4, 2024. https://www.ft.com/content/9d67cda3-b157-47a0-98cb-e8e9842b2c90.

Edge AI + Vision Alliance. 2024. "AI Chip Market to Grow 10x in the Next 10 Years and Become a $300 Billion Industry." *Market Analysis,* April 30, 2024. https://www.edge-ai-vision.com/2024/04/ai-chip-market-to-grow-10x-in-the-next-ten-years-and-become-a-300-billion-industry/.

European Commission. 2024. "Commission Adopts EU-Wide Scheme for Rating Sustainability of Data Centres." News Release, March 15, 2024. https://energy.ec.europa.eu/news/commission-adopts-eu-wide-scheme-rating-sustainability-data-centres-2024-03-15_en.

Goldman Sachs. 2024. "Cloud Revenues Poised to Reach $2 Trillion by 2030 Amid AI Rollout." *Insights,* September 4, 2024. https://www.goldmansachs.com/insights/articles/cloud-revenues-poised-to-reach-2-trillion-by-2030-amid-ai-rollout.

Haranas, M. 2024. "Microsoft and OpenAI Plan $100B AI-Centric Data Center: Report." *CRN,* April 1, 2024. https://www.crn.com/news/data-center/microsoft-and-openai-plan-100b-ai-centric-data-center-report.

Hawkins, Z., V. Lehdonvirta, and B. Wu. 2025. "AI Compute Sovereignty: Infrastructure Control across Territories, Cloud Providers, and Accelerators." Preprint, posted June 24, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5312977.

Kim, T. 2024. "NVIDIA CEO Says Blackwell GPU Will Cost $30,000 to $40,000." *Barron's*. https://www.barrons.com/livecoverage/nvidia-gtc-ai-conference/card/nvidia-ceo-says-blackwell-gpu-will-cost-30-000-to-40-000-l0fnByruULe4RAdr4kPE.

"IIT Hyderabad, NVIDIA Establish First AI Research Center in India." July 9, 2020. https://www.ndtv.com/education/iit-hyderabad-nvidia-establish-first-ai-research-centre-in-india-2259850.

International Energy Agency. 2020. *World Energy Outlook 2020*. https://www.iea.org/reports/world-energy-outlook-2020/outlook-for-energy-demand.

JLL. n.d. "How Government Initiatives Are Impacting Data Centers." *Insights*. https://www.us.jll.com/en/trends-and-insights/cities/how-government-initiatives-are-impacting-data-centers.

Johansson, P.-O., and B. Kriström. 2021. "The Costs and Benefits of Supporting Data Centers: A General Equilibrium Analysis." Preprint, posted December 9, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3981376.

Judge, P. 2022. "Frankfurt Updates Its Plans for Environmental Data Center Zoning." *Data Center Dynamics*, July 27, 2022. https://www.datacenterdynamics.com/en/news/frankfurt-updates-its-plans-for-environmental-data-center-zoning/.

Leswing, K. 2024. "NVIDIA Dominates the AI Chip Market, but There's More Competition Than Ever." *CNBC,* June 2, 2024. https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html.

Meyer, T. 2022. "How Data Embassies Can Strengthen Resiliency with Sovereignty." *Google Cloud Blog,* November 11, 2022. https://cloud.google.com/blog/products/identity-security/data-embassies-strengthening-resiliency-with-sovereignty.

Ministry of Science and ICT, Republic of Korea. 2020. "Data Dam Project Begins, Being Key to Digital New Deal." Press Release, September 2, 2020. https://www.msit.go.kr/eng/bbs/view.do?sCode=eng&mId=4&mPid=2&pageIndex=&bbsSeqNo=42&nttSeqNo=453&searchOpt=&searchTxt=.

Northern Virginia Technology Council. 2020. *The Impact of Data Centers on the State and Local Economies of Virginia*. Falls Church, VA: Northern Virginia Technology Council. https://biz.loudoun.gov/wp-content/uploads/2020/02/Data_Center_Report_2020.pdf.

Norton, K. 2024. "AI Is Fueling a 'Nuclear Renaissance.' Bill Gates and Jeff Bezos Are in the Mix." *Investor's Business Daily,* July 12, 2024. https://www.investors.com/news/artificial-intelligence-ai-data-centers-demand-nuclear-energy/.

Reuters. 2023. "Britain to Invest 300 Million Pounds in AI Supercomputing." November 1, 2023; updated November 2, 2023. https://www.reuters.com/technology/britain-invest-300-million-pounds-ai-supercomputing-2023-11-01/.

Richter, F. 2025. "The Big Three Stay Ahead in Ever-Growing Cloud Market." https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/.

Savvas, A. 2021. "Growth of Data Centers in Frankfurt to Now Be Controlled." *Capacity*, May 13, 2021. https://www.capacitymedia.com/article/29ot42ikril15nn8s4bvd/news/growth-of-data-centres-in-frankfurt-to-now-be-controlled.

Sharwood, S. 2024. "'Data Embassies' Promise Bubbles of Digital Sovereignty, but India Just Cooled on the Idea." *The Register,* July 24, 2024. https://www.theregister.com/2024/07/24/data_embassies.

"The Race Is On to Control the Global Supply Chain for AI Chips." 2024. *The Economist,* July 30, 2024. https://www.economist.com/schools-brief/2024/07/30/the-race-is-on-to-control-the-global-supply-chain-for-ai-chips.

Tony Blair Institute for Global Change. 2023. *State of Compute Access: How to Bridge the New Digital Divide*. London: Tony Blair Institute for Global Change. https://www.institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide.

Turner, M. 2025. "Loudon County, Virginia: Data Center Capital of the World." https://www.loudoun.gov/ArchiveCenter/ViewFile/Item/13979.

Vipra, J., and S. M. West. 2023. *Computational Power and AI*. New York: AI Now Institute. https://ainowinstitute.org/publication/policy/compute-and-ai#b443b50d-b935-4687-a7fc-415a5cacfcb6.

Vivek, S. 2024. "India Eyes NVIDIA Partnership for Affordable AI Chips: Report." *India Today*, updated April 17, 2024. https://www.indiatoday.in/business/story/india-nvidia-partnership-affordable-ai-chips-at-subsidised-rates-report-2528346-2024-04-17.

"What Could Kill the $1trn Artificial-Intelligence Boom?" 2024. *The Economist*, July 28, 2024. https://www.economist.com/business/2024/07/28/what-are-the-threats-to-the-1trn-artificial-intelligence-boom.