Discussion Paper

AI and Energy: The Future of Data Centers in Saudi Arabia

Khaled Alshehri,[a] Marwa Mahmoud AlFattani,[b] Laila Bashmal,[c] and Ghalia Alshmmari[d]

[a]Fellow, King Abdullah Petroleum Studies and Research Center (KAPSARC); [b]Fellow, International Center for AI Research & Ethics (ICAIRE); [c]Researcher, King Saud University (KSU); [d]Fellow, ICAIRE

December 2025 | Doi: 10.30573/KS--2025-DP69

About KAPSARC

KAPSARC is an advisory think tank within global energy economics and sustainability providing advisory services to entities and authorities in the Saudi energy sector to advance Saudi Arabia's energy sector and inform global policies through evidence-based advice and applied research.

About ICAIRE

International Center for Artificial Intelligence Research and Ethics (ICAIRE) — Under the Auspices of UNESCO

ICAIRE serves as a global "lighthouse" for the advancement of ethical AI. The Center is dedicated to ensuring that AI technologies are developed and deployed in harmony with human rights and moral integrity. This is achieved through international cooperation to coordinate AI research and development, raise awareness about AI ethics, foster specialized skills, and provide expert advisory support on AI policies.

This publication is also available in Arabic.

Legal Notice

© Copyright 2026 King Abdullah Petroleum Studies and Research Center ("KAPSARC") and the International Center for Artificial Intelligence Research and Ethics ("ICAIRE"). This Document (and any information, data or materials contained therein) (the "Document") shall not be used without the proper attribution to KAPSARC and ICAIRE. The Document shall not be reproduced, in whole or in part, without the written permission of KAPSARC and ICAIRE. KAPSARC and ICAIRE make no warranty, representation or undertaking whether expressed or implied, nor do they assume any legal liability, whether direct or indirect, or responsibility for the accuracy, completeness, or usefulness of any information that is contained in the Document. Nothing in the Document constitutes or shall be implied to constitute advice, recommendation or opinion. The views and opinions expressed in this publication are those of the authors and do not necessarily reflect the official views or position of KAPSARC or ICAIRE.

KEY POINTS

1. Under a high-growth scenario, Saudi Arabia could exceed 4 GW of data center capacity by 2030, emerging as a regional AI compute hub.

2. AI data centers could consume up to 11% of national electricity by 2030 under high-growth scenarios.

3. Utilization rates and hardware efficiency have a greater impact on AI data center project costs than electricity tariffs.

4. The Kingdom's digital and energy infrastructure is central to AI readiness, data sovereignty, and economic diversification.

5. Strategic alignment of AI and energy policy positions Saudi Arabia as a competitive, climate-aware AI host nation.

Executive Summary

The rise of artificial intelligence (AI) is rapidly transforming the global economy, making AI-ready data centers essential drivers of this change. Unlike traditional data centers for general information technology (IT), AI-focused facilities use advanced chips, dense servers, and liquid cooling to support high-performance computing. This shift accelerates digital innovation but increases pressure on energy systems. Globally in 2024, total data center capacity exceeded 111,900 MW, with the United States and China accounting for more than 60%. Capacity is projected to double to 224,000 MW by 2030, with electricity use rising from 854 TWh in 2024 to nearly 1,900 TWh by 2030. AI workloads already use 5%-15% of that power and could reach 35%-50% by the end of the decade, highlighting the growing link between the energy and digital sectors.

Saudi Arabia's data center industry is expanding faster than most regions, driven by coordinated investment in AI infrastructure.

By 2024, the Kingdom had 58 operational facilities with a total IT capacity of 290.5 MW, primarily in Riyadh and Dammam, which account for nearly 80% of national data center capacity. New hubs, including NEOM, are emerging for large-scale AI projects. The sector's growth is supported by progressive digital policies like the Cloud First Policy and the Data Center Services Regulations, alongside expanding grid infrastructure and low-cost energy. This base makes Saudi Arabia the largest digital infrastructure market in the Middle East and one of the few countries globally developing AI-optimized campuses at a multi-gigawatt scale.

Electricity demand from Saudi data centers is expected to rise substantially until 2030, but actual outcomes are uncertain.

National data centers used around 2.8 TWh in 2024, or 0.85% of total electricity. By 2030, this could increase to between 10.2 TWh and 42.2 TWh, representing 2.8%-11.6% of projected national electricity demand. The increases in demand reflect a total installed capacity of roughly 2,000 MW under a moderate-growth scenario and up to 4,100 MW under a high-growth case. However, many global analysts view the pace of AI infrastructure expansion as highly uncertain – potentially resembling a "digital infrastructure bubble" in which some announced projects do not materialize or are delayed. The low-growth scenario (around 1,050 MW) remains a credible and conservative planning baseline.

The energy and environmental implications of this growth are significant and deserve careful management.

Under a fossil-fuel-dominated power mix, emissions from data centers could rise from 1.6 Mt CO_2 in 2024 to 6-24 Mt CO_2 by 2030. While this is a small share of the Kingdom's total emissions, currently estimated at around 590 Mt CO_2 per year, the growth trajectory highlights the need to include new digital loads within the broader decarbonization strategy. Achieving the national target of 50% renewable generation could reduce data center emissions by about 68%. Efficiency improvements could further mitigate the impact, reducing their electricity consumption by 13% and saving up to 5 TWh annually in the high-growth scenario. From an energy system perspective, these facilities may act as new baseload consumers, and their rapid build-out requires close coordination with utilities to ensure grid reliability and adequate capacity, and to prevent localized bottlenecks.

Saudi Arabia has notable cost advantages which could be further sustained through efficient operations and careful tariff management.

The study's cost analysis shows that data center project costs in Saudi Arabia are most sensitive to utilization and hardware efficiency, and moderately sensitive to electricity tariffs and power usage effectiveness (PUE). With low

competitive tariffs and expanding grid infrastructure, Saudi Arabia remains globally competitive even with relatively high cooling requirements. This competitiveness depends on stable tariffs, early high-utilization rates, and energy-efficient hardware. If these fundamentals are sustained, the Kingdom could become a regional hub for AI computing, serving both domestic and cross-border digital demand. Policymakers can enhance competitiveness by introducing efficiency standards and encouraging high-performance equipment. The analysis also shows that most cost gains occur as data centers move from partial to steady utilization, with diminishing returns at high load factors, a key consideration for utilities and investors optimizing grid integration and cost efficiency.

Despite its advantages, Saudi Arabia faces similar risks to other rapidly expanding markets.

The global AI data center boom has raised regulatory, environmental, and financial concerns. Projects need large tracts of land, highly skilled technical labor, and secure access to reliable electricity and water for cooling. Rising hardware costs, supply chain constraints, and geopolitical risks add further uncertainty. Financially, the surge in AI-related investment resembles a speculative "gold rush," where capital chases uncertain demand. For Saudi Arabia, this highlights the need to sequence projects, aligning expansion with realistic utilization forecasts, and integrating new loads into national energy planning to avoid overcapacity or stranded assets.

Sustainable strategies can mitigate many of these risks and strengthen competitiveness. Technologies such as modular design, AI-optimized chips, and advanced liquid or water-free cooling systems can sharply improve efficiency. Workload scheduling and AI-based energy management can reduce operational loads, while renewable energy integration through power purchase agreements and 24/7 carbon-free energy matching is emerging. Such integration also strengthens Saudi Arabia's positioning as a clean-energy AI hub, aligning digital infrastructure growth with the Kingdom's broader energy transition. Leading global companies demonstrate this shift: Microsoft's waterless cooling, Google's geothermal supply, Amazon Web Services's (AWS's) 100% renewable procurement, and Meta's heat recovery systems illustrate best practice in the industry.

The rationale for investing in AI-ready data centers goes beyond short-term returns and needs careful prioritization and coordination. These facilities create digital spillovers that strengthen the Kingdom's innovation ecosystem, data sovereignty, and economic diversification. Sustained value depends on investing in AI-ready zones with reliable power, renewable integration, and strong utilization. Coordinated planning between government entities, private developers, and global technology partners is essential to maximize benefits while ensuring energy security and climate alignment. Saudi Arabia can use its comparative advantages to become a leading, cost-efficient, and sustainable AI infrastructure hub in the region.

To achieve this balance, the analysis highlights four broad areas to guide future policy and planning. Continued investment in energy-efficient computing technologies, such as next-generation GPUs, advanced servers, and optimized cooling, can increase computing output while managing power demand. Developing AI-ready investment zones with reliable grid connections and renewable integration could attract long-term investors and strengthen Saudi Arabia's position. Expanding local research in data center efficiency, advanced cooling, and sustainable design, together with collaboration between universities, research institutions, and global technology partners, would support knowledge transfer and industrial diversification. Operationally, emphasizing efficient utilization of new facilities, flexible scheduling of AI workloads during off-peak hours, and promoting resource reuse practices such as heat recovery and water recycling could improve efficiency and system integration. From a governance and energy planning perspective, maintaining stable and transparent electricity pricing, encouraging voluntary efficiency benchmarks, and coordinating data center development with renewable energy and grid expansion initiatives would align digital growth with the Kingdom's long-term energy transition objectives.

In conclusion, the Kingdom stands at a strategic inflection point in the energy-digital nexus. Data centers are poised to become a major new source of electricity demand, but their role will depend on prudent planning, measured expansion, and operational efficiency. A cautious, efficiency-oriented approach – based on realistic demand assessment and strong coordination with the national energy transition – will allow Saudi Arabia to capture the long-term value of digital growth while safeguarding reliability, affordability, and sustainability.

Table of Contents

1 Introduction **09**

2 Fundamentals of AI Data Centers **10**

2.1 What Is an AI Data Center? 10

2.2 The Rising Demand for AI Data Centers 12

3 Overview of the Global AI Data Center Landscape **15**

3.1 Power Capacity 15

3.2 Capital Investment 17

3.3 Electricity Demand 18

3.4 Emissions 19

3.5 Regional Outlook (Up to 2030) 20

4 Saudi Arabia's AI Data Center Landscape **22**

4.1 AI Data Centers: A Pivotal Shift 23

4.2 Demand Projections 25

4.3 Emissions Projections 27

4.4 Key Enablers 29

4.5 Factors that Could Influence Projections 30

5 Cost Analysis of AI Data Centers in Saudi Arabia **32**

5.1 Calculating AI Data Center Project Costs 32

5.2 Baseline Results and Sensitivity Analysis 33

5.3 Policy Insights 38

6 Global Overview of AI Data Center Challenges and Risks **39**

6.1 Challenges 39

6.2 Risks 40

7 Towards Sustainable and More Efficient AI Data Centers **42**

7.1 Techniques for Enhancing Operational Efficiency 42

7.2 Options for Decarbonization 45

7.3 Selected Regional Policies 46

Table of Contents

8 Conclusion and Recommendations **50**

References **52**

Acknowledgments **59**

Appendices **60**

Appendix A. Scenario-Based Framework for Estimating Data Center Capacity (2025–2030) 59

Appendix B. Methodology for Estimating Data Center Energy Demand in Saudi Arabia (2025–2030) 60

Appendix C. Methodology for Estimating CO_2 Emissions from Data Centers' Electricity Consumption 62

Appendix D. Methodology for Estimating Lifetime Data Center Project Costs 64

Appendix E. Glossary of Definitions **65**

About the Authors **67**

About the Project **69**

Introduction

01

Artificial intelligence (AI) has grown explosively in recent years, becoming a driving force across industries and economies worldwide. AI adoption is accelerating with large language models (LLMs), generative AI (GenAI) applications, and automation in manufacturing and services. As models grow more capable and are deployed at massive user scale, the need for high-performance computing accelerates. Training and serving today's frontier models increasingly rely on specialized accelerators, high-bandwidth networks, and dense storage, pushing rapid evolution in the digital infrastructure that underpins the AI economy.

Data centers are at the heart of this shift. Operators are adding more facilities and upgrading capabilities: clustering tens of thousands of graphics processing units (GPUs), adopting high-speed interconnects, deploying advanced thermal management, and integrating tighter energy management and on-site power options. AI-optimized hyperscale campuses are becoming the main venues for model training and high-volume inference, while enterprises modernize colocation facilities to keep computing close to data, reducing movement and latency.

This rapid build-out has immediate energy implications. AI workloads are more power-dense and often run at high utilization, intensifying electricity demand and grid planning needs. For instance, training GPT-4 is estimated to have used approximately 42 gigawatt-hours (GWh) of electricity, and data centers are projected to account for 3% of global power sector demand and 1% of total energy sector emissions by 2030, with AI-oriented servers expected to represent about half of the sector's electricity growth (Spencer et al. 2025). At the same time, efficiency and sustainability strategies, such as efficient use, workload scheduling, renewable energy, and emerging low-carbon baseload options, are becoming central to efficient data centers. The balance between growth and sustainability will shape the AI sector's trajectory and national energy systems.

This study examines the future of AI data centers in Saudi Arabia, beginning with an overview of AI data center fundamentals, global trends shaping their expansion, investment patterns, electricity demand, and sustainability challenges. It then analyzes the current landscape in the Kingdom and outlines potential growth scenarios for AI-oriented centers. A detailed cost and efficiency analysis assesses competitiveness under different electricity pricing and infrastructure conditions. The study also discusses the main risks associated with rapid expansion, proposes mitigation measures, and reviews international approaches to developing sustainable AI infrastructure. It concludes with recommendations to support Saudi Arabia's national objectives while strengthening its position in the global data center economy.

Fundamentals of AI Data Centers

02

This chapter introduces the fundamentals of AI data centers, highlighting how they differ from traditional facilities in design, performance, and purpose. It reviews the main types of AI data centers currently in use and explains the rapid growth in demand driven by AI applications. The discussion concludes by examining the link between AI data centers and their rising energy needs.

2.1 What Is an AI Data Center?

An AI data center is a specialized facility designed to handle the intensive computational demands of AI workloads, including training, deploying, and running AI applications and services. These data centers feature advanced computing, networking, and storage infrastructures, along with robust energy and cooling capabilities, to effectively run AI algorithms at scale (IBM 2025). Unlike traditional data centers, which support general IT operations such as web hosting, enterprise applications, and data storage, AI data centers are purpose-built for high-performance computing. While traditional and AI data centers share core components such as servers, storage, and security measures, as shown in Figure 1, AI data centers differ significantly in the scale and capability of their hardware and architecture.

In other words, AI data centers are optimized environments capable of managing complex AI models with billions of parameters and delivering real-time AI inference across a wide range of applications. Table 1 summarizes the differences between traditional and AI data centers.

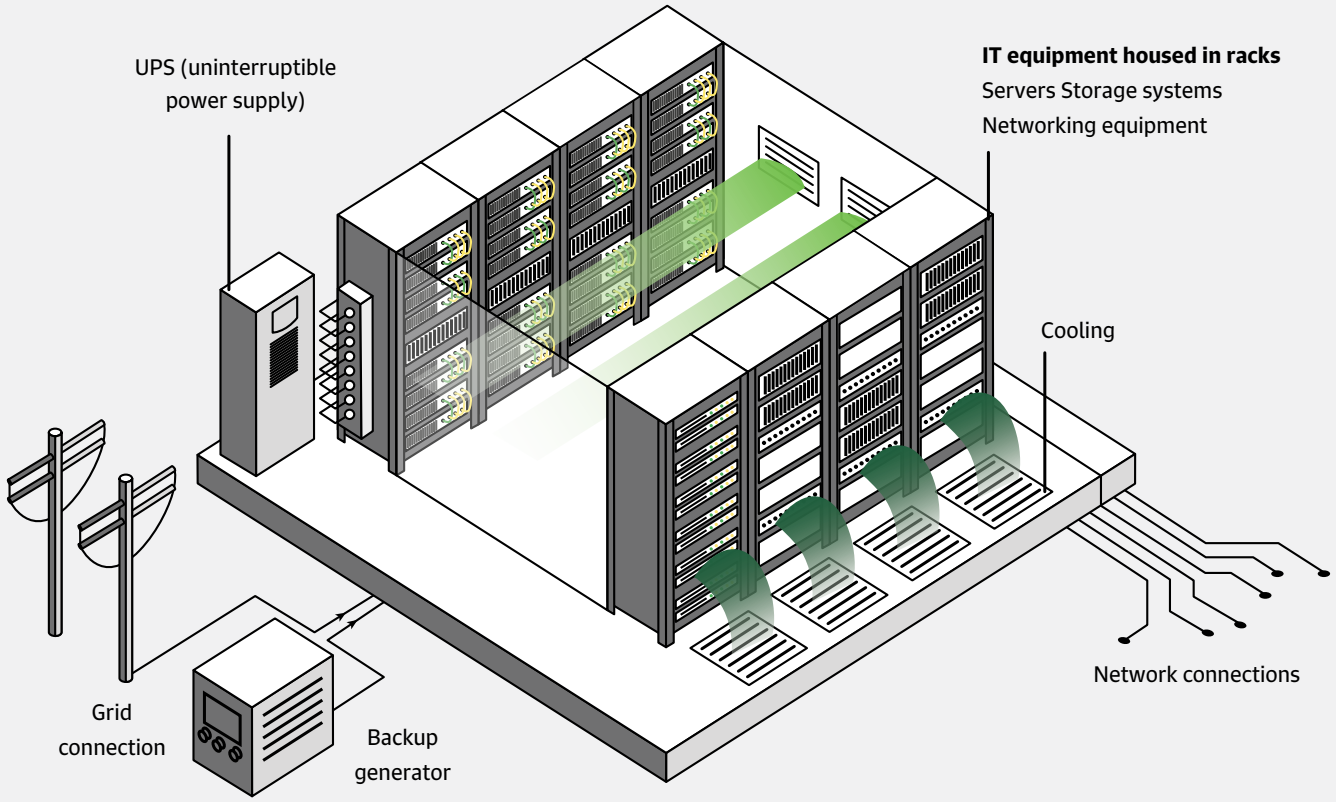
AI data centers come in several forms, reflecting different scales, ownership models, and deployment configurations. The main categories of data centers supporting AI growth today are outlined below, and Table 2 shows a comparative summary of the three types (IBM 2025; Shehabi et al. 2024; Duncan et al. 2024):

- **Hyperscale AI data centers:** Ultra-large data centers typically housing 5,000 servers and spanning at least 10,000 sq ft.[1] They are engineered for extreme scalability and designed to handle large-scale AI workloads, including training frontier AI models and providing AI services to millions of users. These facilities are often built and operated by major cloud providers, such as Amazon Web Services, Google Cloud, Microsoft Azure, and Meta.

- **Colocation AI data centers:** Large data centers owned and operated by third-party providers who rent out space, power, and network connectivity to other companies. A provider can deploy or rent ready-to-use servers inside these colocation facilities. This allows businesses of all sizes to access AI-grade infrastructure without the costs of building and maintaining their own.

[1] Sq ft = Square feet.

Figure 1. Data center's components.



Note: Schematic of a traditional data center. Network connections link the facility to external digital infrastructure, enabling continuous data exchange. The light green elements illustrate data flow between internal servers and outside networks. The dark green elements indicate electricity supply and cooling systems that maintain stable operating conditions. UPS provides short-term backup power, ensuring continuous operation during grid fluctuations or outages.
Source: IEA (2025). Reproduced by authors for improved readability.

Table 1. Differences between traditional and AI data centers.

Feature	Traditional data centers	AI data centers
Workload type	General computing, storage, web hosting, and enterprise IT	AI training and inference, and high-performance computing (HPC)
Hardware focus	Central processing units (CPUs) and modest accelerator use	GPUs, tensor processing units (TPUs), or AI accelerators (e.g., NVIDIA's H100 or GB200)
Cooling technology	Air cooling	Advanced liquid cooling, direct-to-chip or immersive cooling
Storage architecture	Optimized for structured storage: block-level block-level SAN[2] and relational database systems	Advanced, scalable, high-throughput storage systems (such as parallel file systems, NVMe-based,[3] or object storage)
Power density	5-8 kilowatts per rack	30+ kilowatts per rack
Networking	Ethernet with latency-tolerant topologies	High-bandwidth, low-latency networking
Scalability	Moderate: scale by adding standalone halls or cages	High scalability for AI clusters
Facility architecture	Single-story or low-rise buildings, standard 3-4 m ceiling, raised-floor or slab	High-bay (6-10 m) or multi-story blocks with reinforced floor loads, overhead liquid manifolds, and busways
Power capacity	10-30 MW	100-1,000+ MW

[2] SANs: Storage area networks.

[3] NVMe: Non-Volatile Memory Express.

Table 2. Differences between AI data centers.

Type	Hyperscale	Colocation	Enterprise
Ownership and use	Cloud giants	Owned by a third party and used by multiple tenants	Single organization
Floor space (sq ft)	10,000-1,000,000+	Varies	5,000-20,000
Server count	5,000+	Varies	500-2,000
Power demand (MW)	20-100+	Varies	5-10

- **Enterprise AI data centers:** Facilities owned and operated by a single organization for its own use – typically on their premises or at company-controlled sites. Enterprises maintain these facilities to retain control over sensitive data and comply with data sovereignty concerns. These centers are used by organizations with particularly sensitive data, latency-critical applications, or large-scale internal needs to deploy AI-ready infrastructure on the premises.

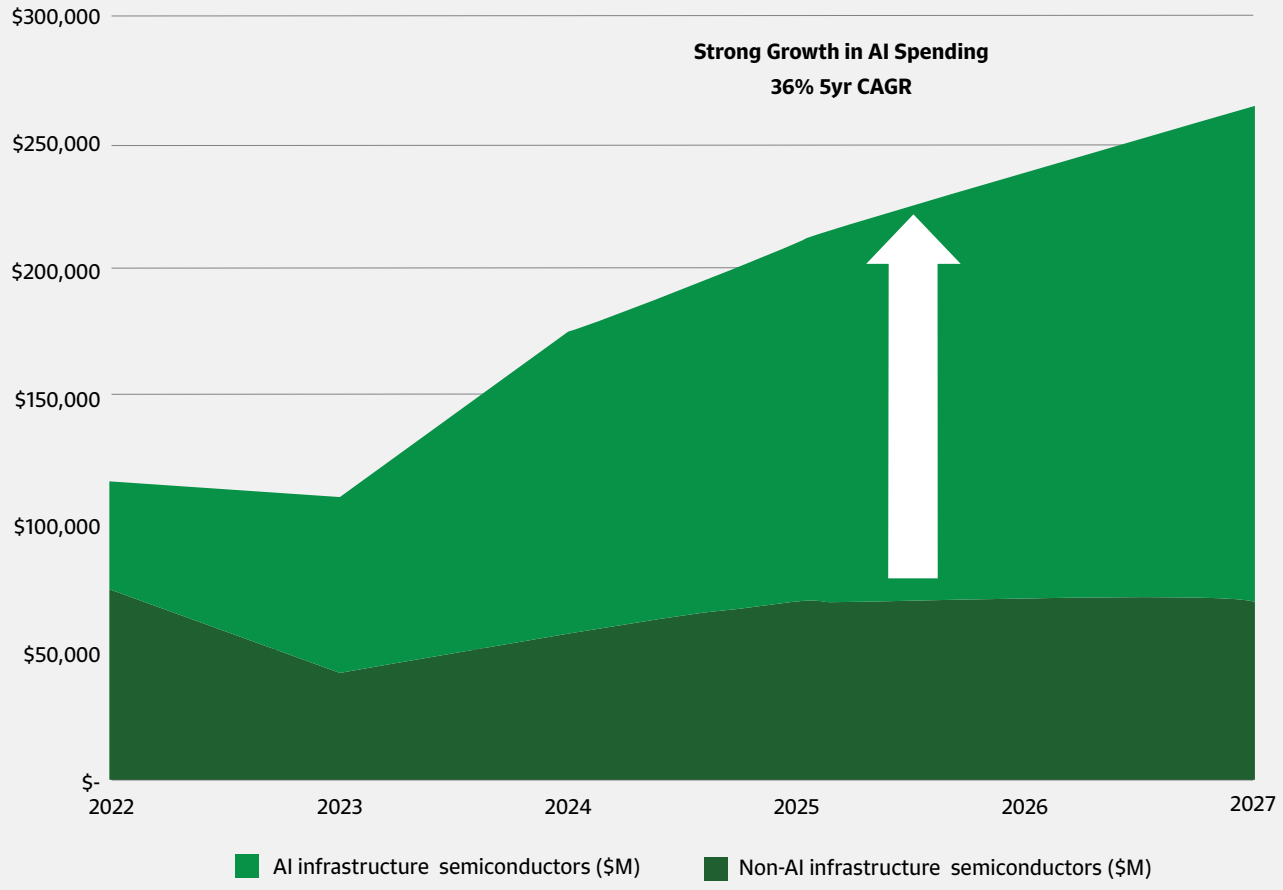
2.2 The Rising Demand for AI Data Centers

AI development and adoption have grown exponentially in recent years, driven by advances in machine learning, especially deep learning, and the explosion of big data. AI is no longer limited to research labs; it has become a key driver of economic growth, boosting productivity and strengthening both global and national competitiveness. According to the International Data Corporation (IDC), the global AI market was valued at nearly $235 billion in 2024 and is projected to reach almost $631 billion by 2028 (International Data Corporation 2024). Additionally, AI could contribute about $19.9 trillion to the global economy by 2030 and drive 3.5% of global GDP in that year (Fioretti et al. 2024).

A major force behind this rapid growth is GenAI and LLMs. These technologies have made AI more accessible, leading to record-breaking adoption rates across businesses worldwide. McKinsey reports that in 2025, nearly 78% of companies are using AI in at least one area of their operations, with 71% using GenAI specifically, a rate that has more than doubled from just 33% in 2023 (Singla et al. 2025). This clear potential for productivity and innovation has led to an explosion of AI applications, from AI copilots in software to AI-driven analytics, across sectors such as finance, healthcare, education, and manufacturing.

With this expansion, the need for robust computation resources has risen to support AI innovation. Between 2012 and 2018, the compute required for leading AI training runs doubled roughly every 3-4 months (Lohn and Musser 2022), and although advances in algorithms and architecture have improved efficiency, the emergence of LLMs has kept the demand for computing growing at a similar rate. As a result, organizations across industries, including banking, health care, and manufacturing, are investing heavily in advanced computing infrastructure to support AI development. The IDC projects that AI will continue to drive new data center investments, with spending expected to grow at a compound annual growth rate (CAGR) of 36% between 2022 and 2027 (Hoff 2024), as Figure 2 shows.

Figure 2. Growth in data center spending due to AI between 2022-2027.



Source: Hoff (2024). Reproduced by authors for improved readability.

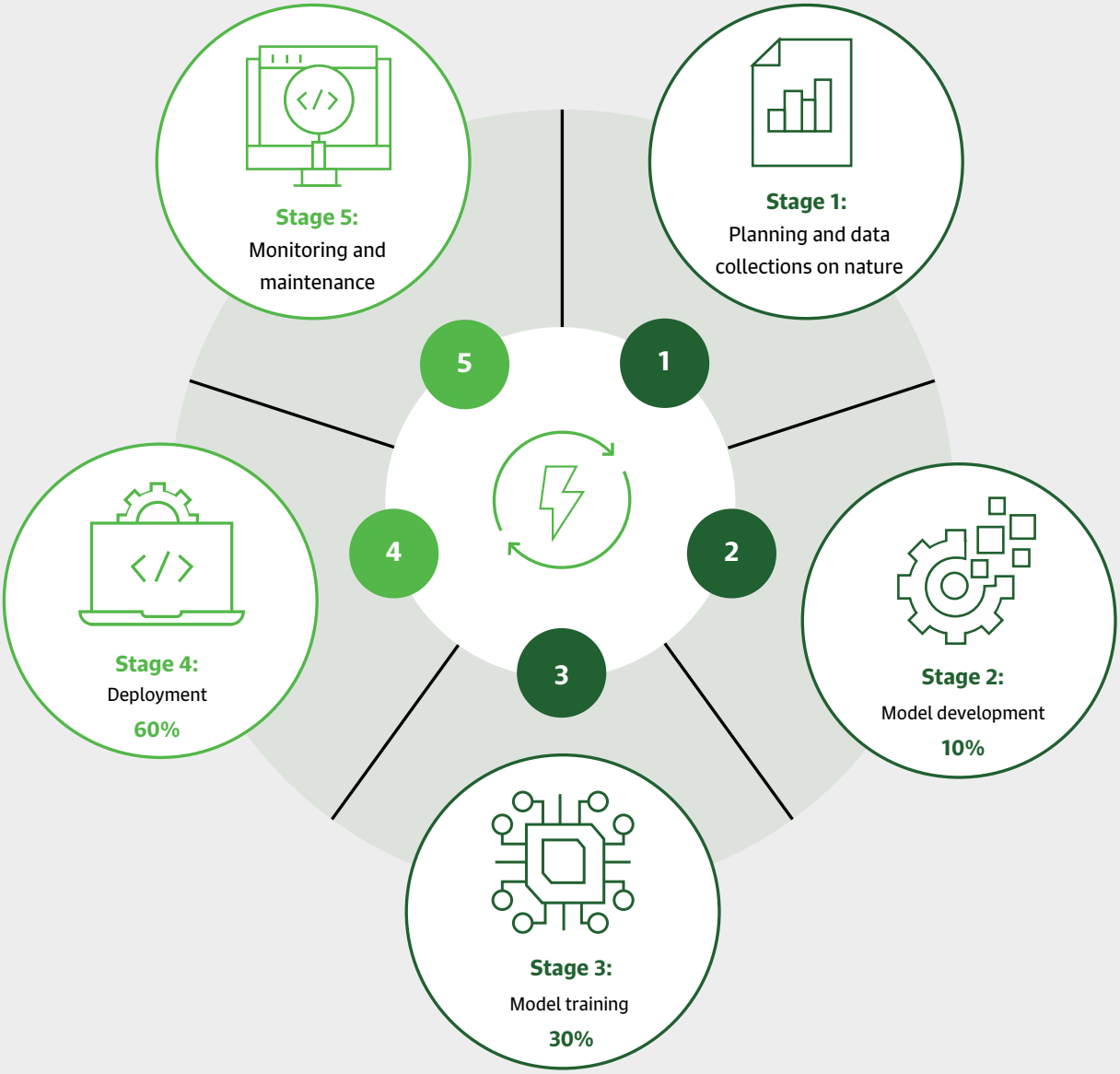
The accelerated adoption of AI and the rise of compute-intensive workloads are reshaping data center infrastructure requirements: how they are designed, how many are needed, and how large they must be. This transformation brings both opportunities and challenges for investors and policymakers. Central to these challenges are the significant power and environmental implications. High-density servers running continuous training cycles and large-scale inference workloads consume vast amounts of electricity. In some cases, a single hyperscale AI campus may use as much power as a small city.

Figure 3 shows that energy requirements vary across different stages of the AI lifecycle, and consistently exceed those of traditional digital services. Early estimates indicate that a generative AI query consumes roughly 10 times more electricity than a conventional search, rising from about 0.3 watt-hours (Wh) for a Google search to nearly 2.9 Wh for a ChatGPT request (EPRI 2024). At scale, such differences can strain local grids. Regions with clusters of data centers are already facing capacity constraints, proving the urgency of strategic planning.

Understanding the current state, trajectory, and needs of AI data centers has become an economic, environmental, and geopolitical necessity. AI data centers create opportunities for investment, innovation, and digital competitiveness, but they also pose challenges in securing reliable electricity and maintaining sustainability. Nations that can effectively align AI infrastructure growth with energy system resilience will be better positioned to reap the benefits of the AI economy.

For Saudi Arabia, understanding and leveraging these dynamics is essential to attract AI investment, expand digital infrastructure, and ensure a reliable energy supply. Effectively balancing digital expansion with energy transformation will be critical to securing a strategic role in the global AI economy and advancing national objectives for economic diversification, such as R&D priorities, telecommunications, IT, and industrial sector goals.

Figure 3. Energy consumption across AI lifecycle.



Source: Greene-Dewasmes, Higgins, and Tladi (2025). Reproduced by authors for improved readability.

Overview of the Global AI Data Center Landscape

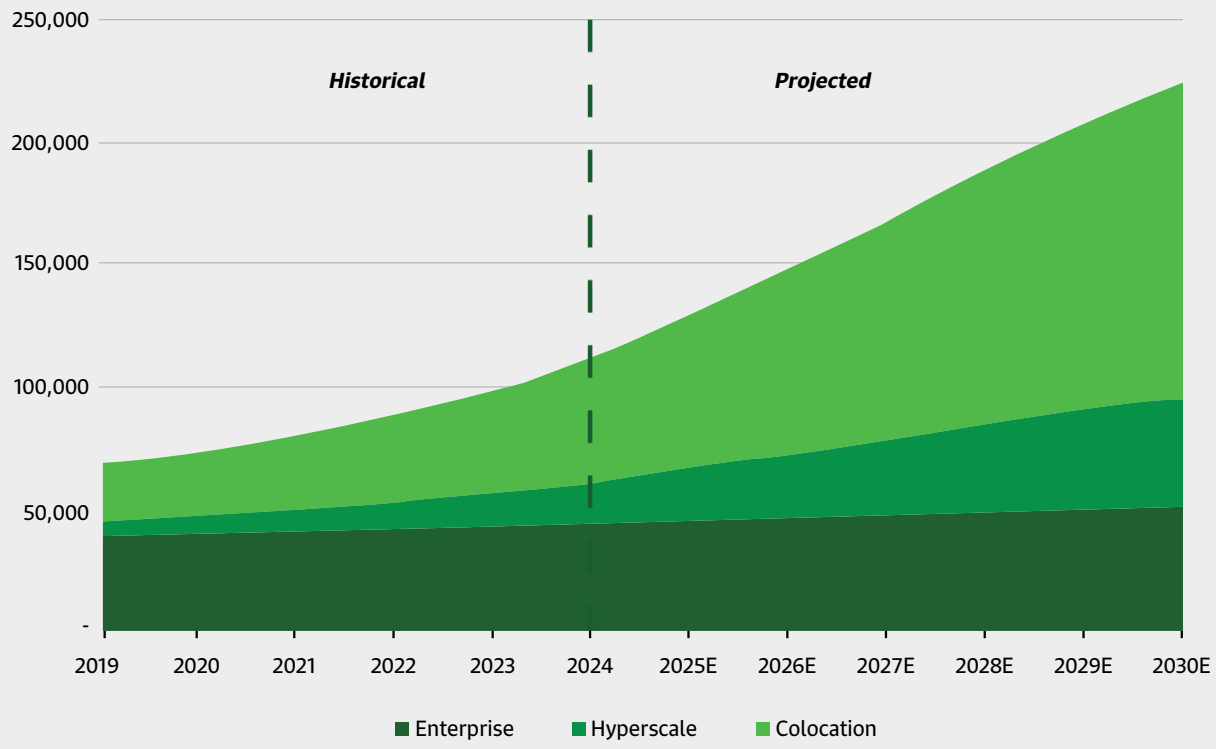
03

The rise of AI is becoming the dominant driver of demand in the global data center industry, reshaping infrastructure growth, investment flows, energy consumption, and environmental footprints. This section outlines the state of AI data centers worldwide and evaluates their implications for infrastructure, energy demand, and economics.

3.1 Power Capacity

The global data center landscape is expanding rapidly in both number and power capacity. As of March 2024, more than 11,800 facilities were operational worldwide, about twice as many as five years earlier. The U.S. leads with 5,388 data centers (around 45% of the total), followed by Germany (520), the United Kingdom (512), China (449), and Canada (336) (Fleck 2024). Within this broader build-out, hyperscale platforms that power AI workloads and the cloud had surged to about 1,189 data centers by early 2025 (Synergy Research Group 2025). According to S&P Global Data, which we primarily use for our analysis and projections throughout this paper (unless otherwise stated), the global installed capacity of data centers has been estimated to be around 111,897 MW in 2024, as shown in Figure 4. While enterprise data center capacity has remained relatively stable, significant expansion is driven by hyperscale and colocation facilities, which approximately account for 15.6% and 44.9% of the global capacity, respectively, since 2020.

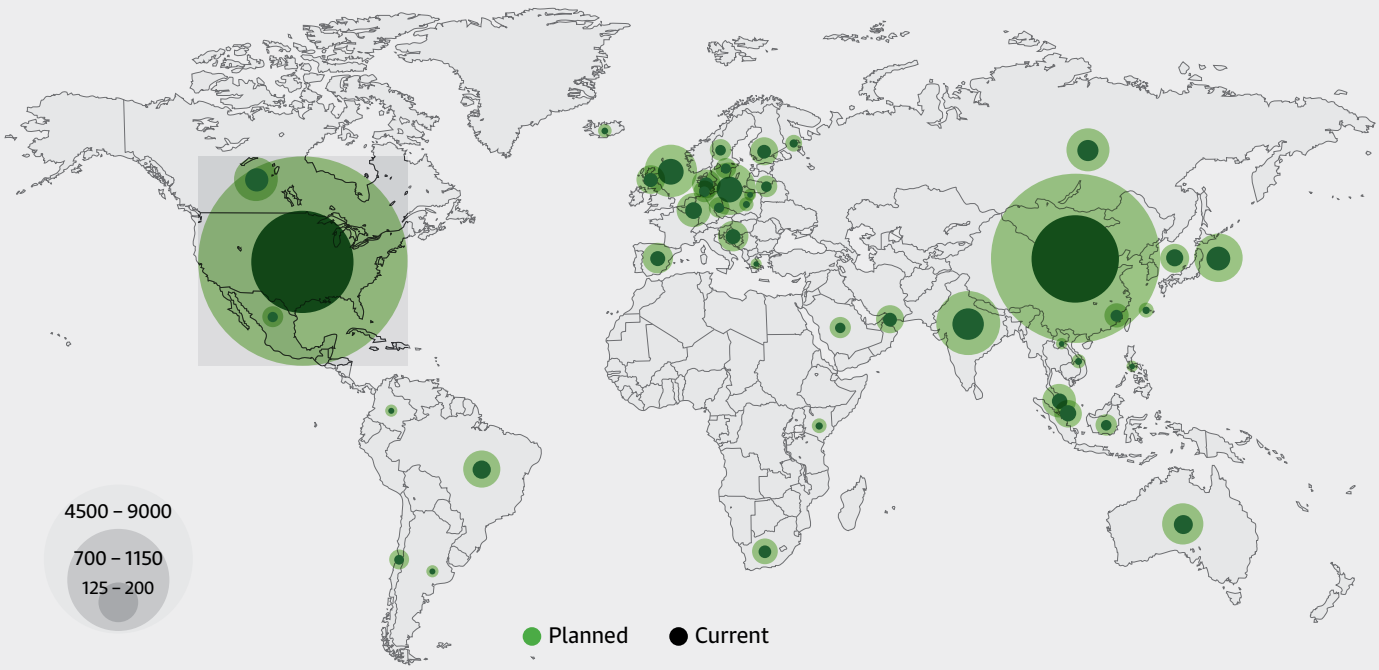
Figure 4. Historical and projected global installed data center capacity (MW).



Source: Authors.

The distribution of the overall capacity is uneven around the globe. In fact, by 2024, the U.S. and China account for roughly 60% of this capacity (38,962 MW and 28,639 MW, respectively), while the European Union holds about 20,093 MW, as Figure 5 shows.

Figure 5. Global distribution of installed data center capacity in 2025.



Source: Authors.

Looking ahead, S&P Global estimated that by 2030, the total installed capacity in data centers will be around 224,360 MW, as Figure 4 shows, with a CAGR of 12% between 2024 and 2030. Hyperscale and colocation data centers are expected to grow by a CAGR of 17%, and enterprise centers by only 5% . For AI-specific IT capacity, the IDC expected an increase to 19,600 MW by 2027, up from 3,600 MW in 2022 (Graham, Rutten, and Yashkova 2024). McKinsey, on the other hand, predicted that in a midrange scenario, the demand for AI data center capacity will rise to 33% on average between 2023 and 2030, meaning that by 2030, advanced AI-equipped facilities are expected to account for about 70% of total data center capacity, with generative AI workloads representing roughly 40% of that share (Srivathsan et al. 2024). All these factors underscore the industry's shift toward high-density, AI-optimized compute.

3.2 Capital Investment

The rapid growth of AI has triggered a global wave of capital investment in data center infrastructure. Companies and governments are allocating funds to expand data center capacity through new buildings (greenfield) and upgrades of existing facilities (brownfield) to meet the demand for AI training and inference workloads.

According to CBRE Group's[4] global survey, 95% of 92 major data center investors planned to increase spending in 2025, with 41% expecting to allocate $500 million to $2 billion (or more), up from 30% in 2024 (CBRE 2025). The IEA reported that global investment in data centers increased by nearly 70% in the last two years (Spencer et al. 2025).

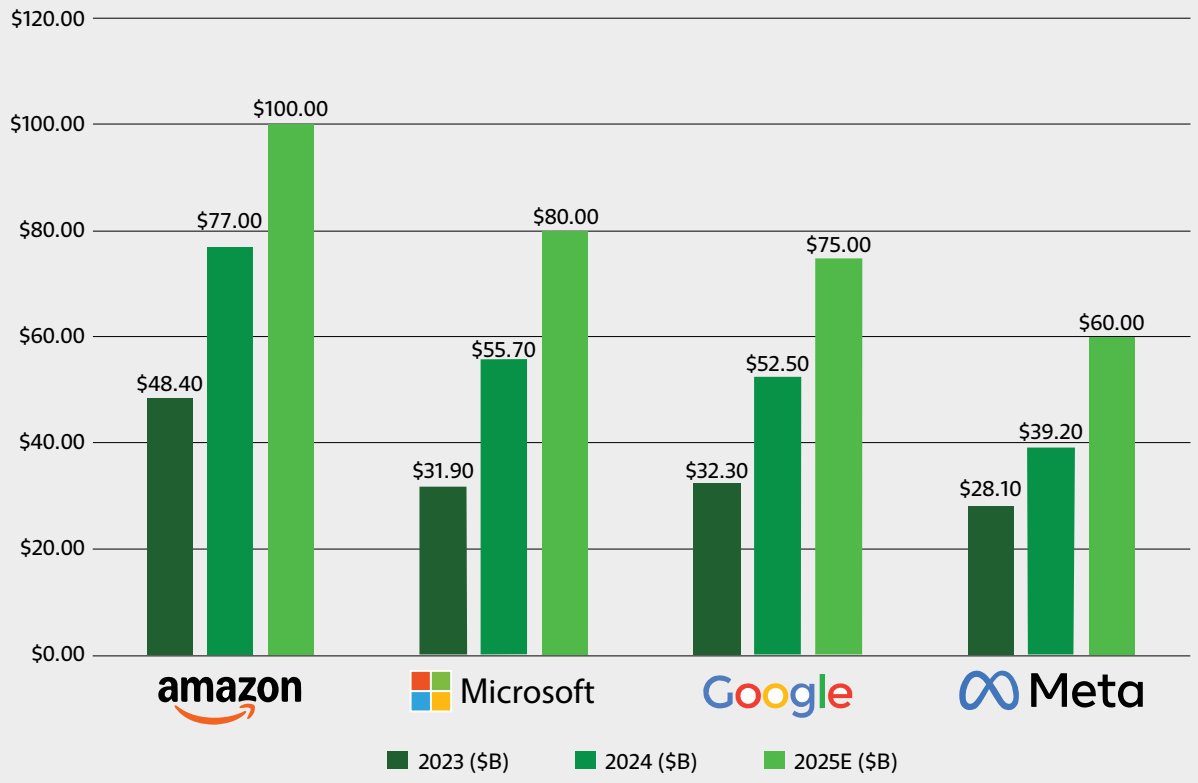
Many governments have launched initiatives to build data centers. For example, the U.S. launched the Stargate AI initiative in 2025 with a $500 billion investment to run through 2028. The European Union launched InvestAI, dedicating $207 billion to build "AI gigafactories" that provide critical computational infrastructure, including data centers (Clemmons and Graham 2025). China invested over $6.1 billion in 2024 to build eight computing hubs as part of its mega data initiative "East Data, West Computing." This led to additional data center investment from provincial governments and state-affiliated firms, totaling $27 billion by the end of the same year to further boost AI computing capacity (Stokols 2025). In parallel, France and the UAE jointly announced a plan to build a 1,000 MW AI data center in Europe, with an estimated investment of $34.8 to $58 billion (Narasimhan et al. 2025).

Hyperscale cloud providers remain the primary drivers of AI infrastructure investment. Google, Amazon, Meta, and Microsoft collectively planned to spend $320 billion in 2025, up from $230 billion in 2024, representing a 1.4-times increase and a CAGR of 40.37%. Microsoft alone committed $80 billion for AI-enabled data centers, while Meta projected $60 to $65 billion for AI supercomputing (Houlihan Lokey 2025), as shown in Figure 6.

If current trends continue, hyperscale capital expenditures (CAPEX) are expected to push annual global data center investment above $1 trillion by 2029, more than double today's level. McKinsey estimated the investment needed worldwide in data centers to meet AI demand alone to be $5.2 trillion, and it's projected to reach nearly $7 trillion by 2030 (Noffsinger et al. 2025). This increase is driven by AI-driven demand for advanced chips, cooling systems, and grid integration, as well as power and land constraints.

[4] CBRE Group: Coldwell Banker Richard Ellis.

Figure 6. Hyperscale cloud providers actual and projected spending on AI data center development in 2023-2025.



Source: Houlihan Lokey (2025). Reproduced by authors for improved readability.

3.3 Electricity Demand

Since 2017, data center electricity consumption has accelerated due to technological trends such as the growth in cloud computing, the use of social media, and the rapid expansion of AI. According to the IEA, the yearly rate of data centers' electricity consumption grew from 3% between 2005 and 2015 to 10% between 2015 and 2024 (Spencer et al. 2025). In fact, S&P Global estimated the global data center electricity consumption to be around 854 TWh.

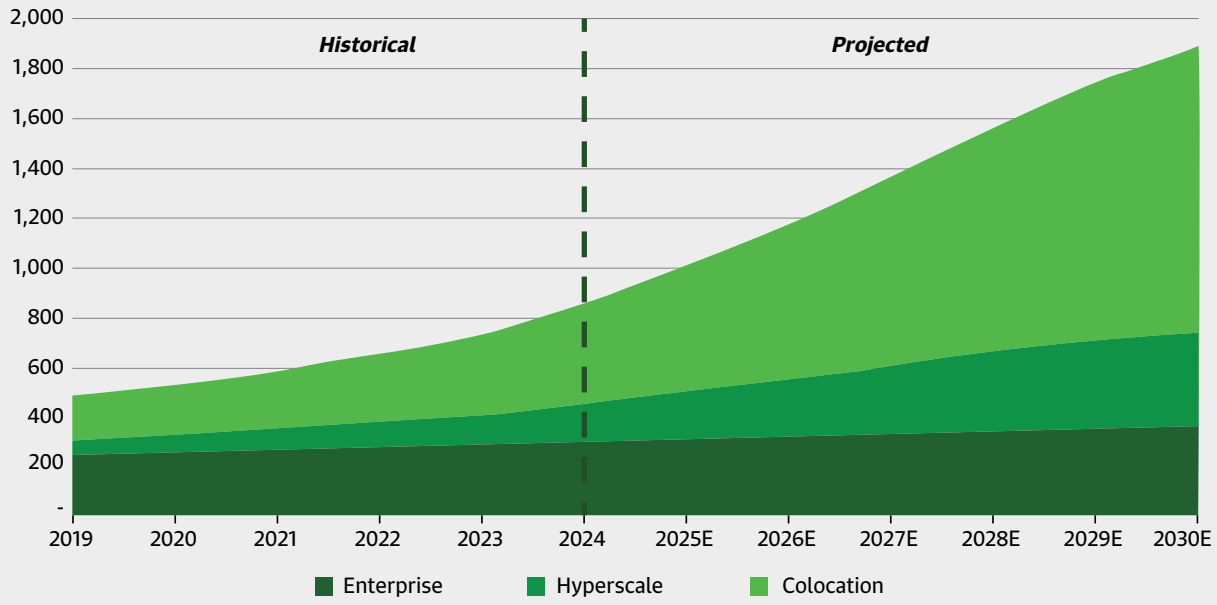
AI workloads are the most significant driver of this surge. Studies estimated that AI-specific tasks accounted for only 5%-15% of global data center energy use in 2023, but are projected to keep increasing rapidly to 35%-50% in 2030 (Kamiya and Coroamă 2025). Accordingly, the global electricity demand in data centers will double by 2030 to about 1,891 TWh with a CAGR of 14%,[5] as shown in Figure 7.

Despite these figures, the outlook is far from certain, and it is plausible that such projections may not materialize. The trajectory will depend on several evolving factors, including the pace of AI deployment, regulatory developments, and the readiness of power and grid infrastructure. In many markets, delays in securing grid connections and long lead times for high-capacity transformers have become critical bottlenecks, often extending project timelines by up to three years.

These constraints, coupled with rising costs, could slow the build-out of large-scale AI campuses and temper near-term growth expectations. For example, the IDC anticipated an AI-driven increase in data center electricity use at a CAGR of 44.7%, reaching about 146.2 TWh by 2027 (Graham, Rutten, and Yashkova 2024), while Goldman Sachs projects AI-related power use could grow to 200 TWh per year between 2023 and 2030 (Goldman Sachs 2024). Even with more conservative growth, the expanding digital infrastructure will have significant implications for the environment and energy systems, highlighting the importance of efficiency and sustainability measures in the sector's evolution.

[5] S&P Global dataset.

Figure 7. Historical and projected global data center energy consumption (TWh) for 2019-2030.



Source: Authors.

3.4 Emissions

The energy surge from AI data centers also carries a significant environmental footprint. Its carbon emissions could challenge global climate targets without proper mitigation. The IEA estimates that data centers currently account for around 180 Mt CO_2 globally, less than 1% of total energy-related emissions, yet they are among the fastest-growing sources of emissions in the power sector, projected to peak at about 320 Mt CO_2 by 2030 (Spencer et al. 2025). Accenture estimates that by 2030, emissions from AI data centers may reach 3.4% of worldwide greenhouse gases (CO_2e), a roughly 11-fold increase within a decade (Jamison et al. 2025). In this study, emissions are reported in CO_2 from fossil-fuel-based electricity generation and exclude other greenhouse gases. AI infrastructure could emit 300-350 Mt CO_2 annually by 2030, comparable to the emissions of a mid-sized country.

Most major hyperscale operators have pledged to achieve net-zero emissions, adding more complexity to the sector's energy strategy. Meeting these targets requires a growing share of renewable electricity, yet those sources are intermittent. As a result, data center operators increasingly rely on natural gas or nuclear generation for reliable power while offsetting emissions through renewable energy certificates (RECs). Developing reliable REC markets, whether domestic or international, is therefore critical, and their current immaturity could delay new projects or raise compliance costs. Alternatively, operators may look for more flexible emission targets or slower implementation timelines to balance growth with sustainability commitments.

Beyond carbon, other environmental pressures are rising, notably water use and heat emissions. Most AI data centers rely on power-hungry cooling systems, such as chilled water and air conditioning, to remove the significant heat generated by high-performance chips. By one estimate, cooling AI data centers in 2030 will require about 3.02 billion cubic meters of fresh water annually, more than countries like Norway or Sweden withdraw in a year (Jamison et al. 2025). Even a small-scale data center can have a substantial water footprint, where every 1 MW of data center capacity demands 25-30 million liters of water per year for cooling, equivalent to the drinking water needs of about 300,000 people (Spindler, Hahn-Petersen, and Hosseini 2024).

3.5 Regional Outlook (Up to 2030)

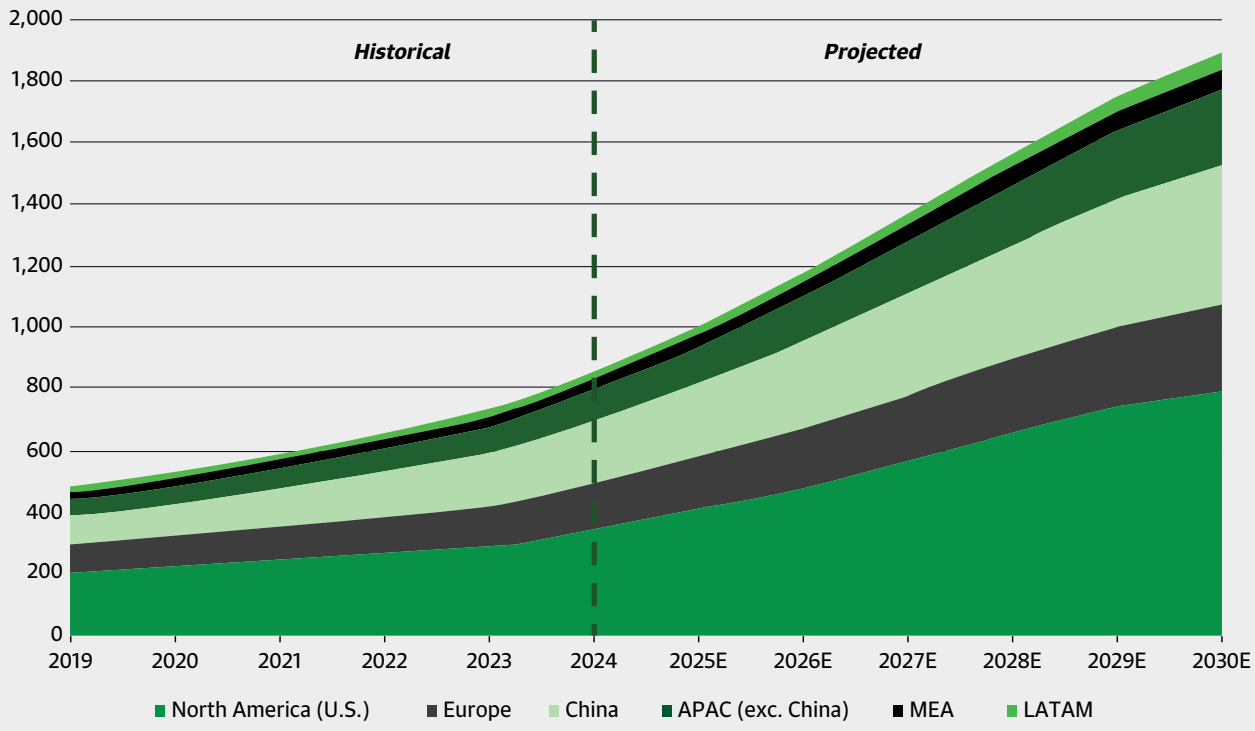
This section provides the current state and projected outlook across major geographies, highlighting demand for data center capacity, investment hotspots, and growth in electricity consumption.

North America (United States)

North America, particularly the U.S., leads the world in data center capacity and AI compute, hosting the largest concentration of hyperscale and AI facilities. U.S. cloud providers serve as the primary backend for the world's AI activity, with a significant portion of inference and training tasks performed globally being executed on servers in U.S.-operated data centers. Major hubs include the West Coast, Texas, and Virginia, with Northern Virginia, often referred to as "Data Center Alley," housing the single largest concentration of cloud and AI infrastructure in the world (Spencer et al. 2025).

According to S&P Global, the total installed data center capacity in the U.S. is projected to reach about 89,560 MW by 2030. This expansion has significant energy implications. U.S. data centers consumed roughly 322 TWh of electricity in 2024, as Figure 8 shows. This started in 2017, after a period of stability in annual energy consumption between 2014 and 2016 at about 60 TWh, driven by the growth in the installed server base and GPU-accelerated servers for AI (Shehabi et al. 2024). Looking ahead, energy consumption is expected to increase to about 769 TWh by 2030 (Green et al. 2024).

Figure 8. Historical and projected regional data center electricity consumption (TWh) for 2019-2030.



Note: APAC: Asia-Pacific, MEA: Middle East and Africa, LATAM: Latin America.
Source: Authors.

China

China is one of the world's major players in AI data center infrastructure. It hosts some of the largest data center installations and is racing to expand its AI compute capacity. Over the past decade, it has built a vast domestic network of hyperscale facilities supporting tech giants like Alibaba, Tencent, and Baidu. There are also state-led AI initiatives. In 2024, China hosted about 28,639 MW of data center capacity, which is expected to increase to more than 54,614 MW in 2030. China currently represents about 25% of global data center electricity consumption (Spencer et al. 2025), about 201 TWh of electricity. The data center sector in China started to expand in 2015, with electricity demand growing by 15% per year until 2024 (Spencer et al. 2025). This growth is expected to continue to 2030, with demand projected to reach more than double, as the S&P Global data show.

Europe

Europe is expanding its AI data center footprint. While historically behind the U.S. in hyperscale cloud, the region is now investing in its own AI compute and regional capacity. Major hubs include Dublin, Frankfurt, Amsterdam, London, and the Nordic countries. In 2024, the total installed capacity in all data centers within the European region reached around 20,000 MW, expected to increase to approximately 33,422 MW by 2030.[6] Germany led with 3,188 MW in 2024, followed by the UK with 2,764 MW and France with 1,374 MW.

This trend stems from the EU's ambition to become a global leader in AI by building AI factories and supercomputing centers across member states, with plans to triple data center capacity in the next five to seven years (European Commission 2025). Data centers consumed about 151 TWh of electricity in 2024, and this is estimated to grow by more than 281 TWh in 2030, with a CAGR of 11%. Also, a notable trend in the region is the rise of data center campuses in the Nordic countries specifically targeting AI customers with 100% green energy and high-performance computing offerings (Duncan et al. 2024).

GCC and the Middle East

The Middle East is rapidly becoming a global hotspot for data centers and AI compute, driven by economic diversification strategies and favorable energy conditions. The Gulf Cooperation Council (GCC) countries are leading this trend, capitalizing on low-cost land and energy; their strategic location bridging Asia, Africa, and Europe; and available capital. Despite the region's hot climate, which increases cooling needs, governments see data centers as key pillars of their digital and economic transformation agendas. National strategies, such as Saudi Arabia's Vision 2030, UAE Vision 2031, New Kuwait 2035, and Digital Oman 2030, have boosted the Middle East's data center market (Tohme et al. 2025).

As of 2024, the region hosted more than 290 data centers in 17 countries (Data Center Map.n.d.). Total data center capacity reached about 5,311 MW in 2024 and is expected to grow to around 8,938 MW in 2030. Industry analysts have stated that the Middle East is one of the fastest-growing regions for data centers (Shiwani, Abbasi, and Levack 2025).

Flagship projects illustrate this trend. In Saudi Arabia, major initiatives such as HUMAIN's AI factories in Dammam and NEOM's Oxagon Green AI Campus are driving the region's growth, with multi-gigawatt capacity planned by 2030. UAE has partnered with U.S. firms to develop a 5,000 MW AI data center campus through G42, aiming to build the largest AI-dedicated data center complex outside the U.S. The first 1,000 MW phase is already in progress (U.S. Department of Commerce 2025). Qatar, Oman, Bahrain, and Kuwait are also expanding. Oman, for example, has a partnership with Equinix for a large hub in Salalah, and Kuwait has announced a 1,000 MW data center park (Shiwani, Abbasi, and Levack 2025). Beyond the Gulf, secondary markets like Egypt and Morocco are investing in new data center parks and incentives to attract cloud providers. S&P Global data project that energy consumption of data centers in the Middle East will rise from 38 TWh in 2024, with a CAGR of 11%, to around 72 TWh in 2030.

The global AI data center landscape is shifting from a U.S.- and China-centered model toward a more distributed network of regional centers. As countries increasingly view AI compute as a strategic asset, regions are capitalizing on their distinct advantages to build competitive AI capabilities: the U.S. tech ecosystem, China's scale and state support, Europe's regulatory environment and green power, and the Middle East's energy resources. By 2030, a more balanced global distribution is likely to have materialized, with emerging regions contributing more substantially to AI workloads.

[6] S&P Global data.

Saudi Arabia's AI Data Center Landscape

04

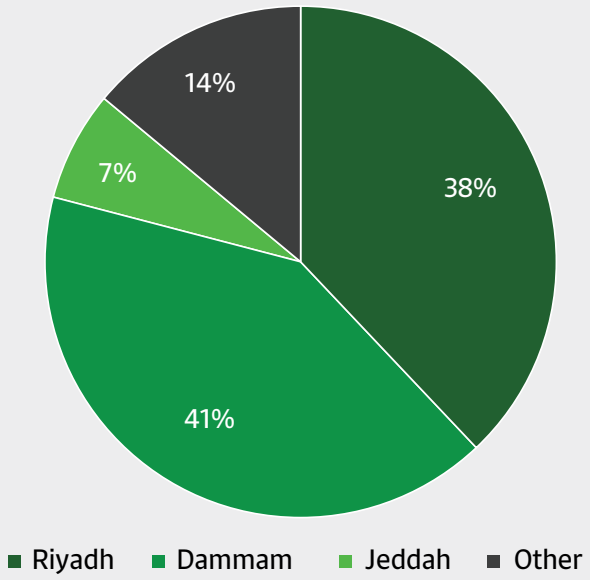
Saudi Arabia is becoming one of the most dynamic data center markets in the Middle East, with AI-driven infrastructure a key pillar of its digital and energy transformation. Rapid growth in the sector reflects the Kingdom's national digital strategies, supported by favorable regulation, a reliable power supply, and strong government and private investment. This section examines the evolution and distribution of data centers across the country, the factors enabling their growth, and the implications for electricity demand and emissions. It also develops a set of growth scenarios for the period 2025 to 2030 to assess how different operational and efficiency conditions could shape the sector's contribution to Saudi Arabia's energy and sustainability goals.

The development of data centers in Saudi Arabia has historically aligned with the Kingdom's broader digital transformation strategies. In the early 2000s, infrastructure was primarily built by national telecom operators to support government agencies, national databases, and basic internet services. As demand for digital services grew, the 2010s saw an expansion in colocation and enterprise-grade facilities driven by the financial sector, e-government initiatives, and cloud adoption. The launch of Vision 2030 in 2016 marked a turning point, positioning digital infrastructure as a cornerstone of economic diversification. Since then, the sector has evolved from general-purpose hosting to advanced, AI-capable infrastructure as part of the Kingdom's digital economy goals.

As of 2024, Saudi Arabia has around 58 data centers (CST 2024a). Between 2022 and 2024, the IT capacity of these data centers more than doubled, from 122.4 MW (MCIT 2023) to 290.5 MW (MCIT 2024), reflecting a CAGR of approximately 54% and placing the Kingdom among the fastest-growing digital infrastructure markets in the Middle East. Saudi Arabia's data centers are concentrated in a few key regions. According to S&P Global, approximately 79% of the country's installed capacity is located in two cities: Riyadh and Dammam, as shown in Figure 9.

This spatial concentration reflects the country's economic geography: the availability of critical infrastructure, such as power grids and fiber-optic networks, and the proximity to enterprise markets. Riyadh, with nearly 38% of the total capacity, has become the main hub for general-purpose cloud and enterprise data centers, as it hosts many government agencies and company headquarters. Dammam accounts for the largest share at roughly 41% of capacity. Its location in the Gulf industrial corridor, close to subsea cables and major energy companies, drives its rapid growth as a fast-emerging computing zone. Jeddah contributes about 7% of the capacity and other cities, including Madinah and Buraydah, each hold only 1%-3% of the total capacity; these smaller facilities mainly serve local government and business needs.

Figure 9. Data center capacity distribution across Saudi Arabia regions.



Source: Authors.

4.1 AI Data Centers: A Pivotal Shift

Saudi Arabia's data center capacity is growing fast, mostly from AI data centers. Until 2024, most facilities were primarily designed to support cloud services, enterprise hosting, and digital government services. Starting in 2025, there has been a pivotal shift toward AI-oriented infrastructure.

The first clear step in this direction came with the early 2025 launch of Groq Cloud's inference-first facility in Dammam, signaling the Kingdom's shift in the data center landscape as the facility is explicitly built for AI at scale (Groq 2025). At the same time, HUMAIN's first 50 MW phase is scheduled to go live late in 2025, with further expansion planned throughout the decade (Techusiness 2025). NEOM is also developing the Oxagon AI Campus, starting with 300 MW in 2028 (NEOM 2025). By the early 2030s, projects from HUMAIN, NEOM, and others are expected to push national capacity into the multi-gigawatt range.

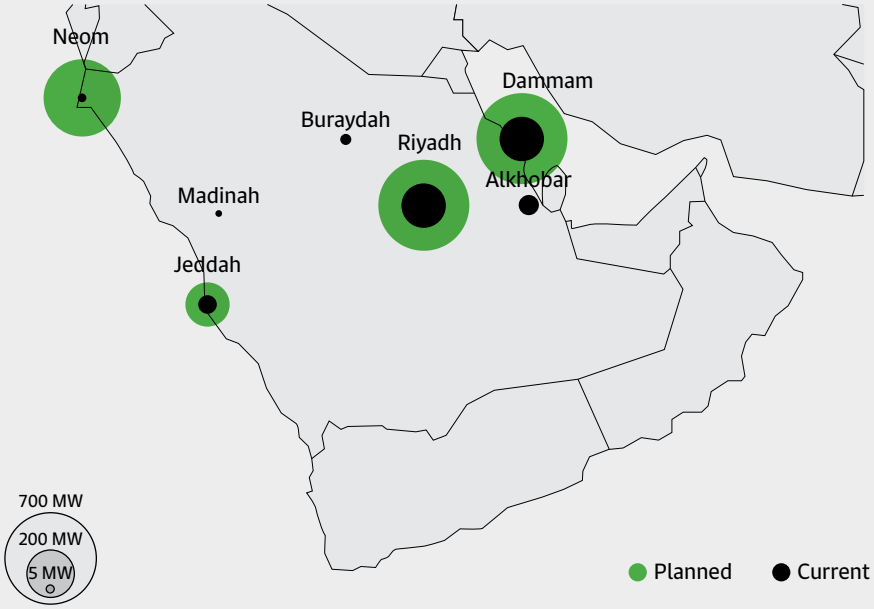
By mid-2025, announced investments in Saudi Arabia's data center sector had surpassed $25 billion, with long-term commitments projected to reach about $77 billion by 2034 (Saudi Market Research Consulting Firm 2025). HUMAIN is leading much of this growth, working with international technology partners like NVIDIA, AMD, Amazon Web Services (AWS), Qualcomm, and Cisco. NEOM's AI campus is another major contributor. If all announced projects go ahead, Saudi Arabia could account for up to 7% of global AI training and inference capacity by the early 2030s (England and Omran 2025). Notable key projects as of mid-2025 are summarized in Table 3. This transition marks a shift from earlier HPC systems, such as KAUST's Shaheen and Aramco's Dammam-7, toward large-scale, commercial AI infrastructure supporting national and global digital economic growth.

Current and announced data centers and their capacity will be distributed throughout the Kingdom. Projects will include scaling existing hubs, such as Riyadh and Dammam, and creating new ones, such as NEOM in the northwest, as shown in Figure 10. However, some announced capacities have not yet been assigned to specific locations.

Table 3. Major Saudi Arabia AI data center projects based on public announcements (as of mid-2025).

Project/ initiative	Timeline	Lead entity	Investment	Capacity target	Key objectives
HUMAIN-Groq AI Inference Center (Groq 2025)	2024-2025	HUMAIN and Groq	$1.5 B	N/A	Inference-only facility using Groq's LPUs, optimized for ultra-low-latency generative AI workloads
Oracle Infrastructure Investment (Egbert 2025)	2024-2034	Oracle	$14 B	N/A	Expansion of cloud and AI infrastructure to support national digital transformation and enterprise services
HUMAIN-AWS AI Zone (Amazon 2025)	2025-2026	HUMAIN and AWS	$5 B	N/A	Dedicated AI innovation zone integrated with AWS infrastructure
Google Cloud-PIF AI Hub (Google Cloud 2025; PIF 2024)	2025-2027	HUMAIN and Google Cloud	$10 B	N/A	Regional hub for AI development and cloud services, with a focus on enterprise and government applications
HUMAIN-AMD AI Infrastructure (Techusiness 2025; HUMAIN 2025; Grabein and Stine 2025)	2025-2030	HUMAIN, AMD, and Cisco	$10 B	500 MW (part of 1,900 MW by 2030; 6,600 MW by 2034)	Deploy AMD MI300 accelerators with Cisco networking to build open infrastructure targeting developers, startups, and national LLM projects
HUMAIN-NVIDIA AI Factories (NVIDIA 2025a)	2025-2030	HUMAIN and NVIDIA	$10 B	500 MW (part of 1,900 MW by 2030; 6,600 MW by 2034)	Initial 50 MW deployment with 18,000 NVIDIA GB300 GPUs, phased expansion to 500 MW with 180,000 GPUs
NEOM-DataVolt Oxagon Green AI Campus insert space (NEOM 2025)	2025-2031	NEOM and DataVolt	$5 B	1,500 MW	Net-zero hyperscale AI campus in Oxagon powered by solar, wind, on-site battery storage, and hydrogen backup with advanced cooling systems
HUMAIN-Qualcomm Hybrid Inference (Qualcomm 2025)	2026-2028	HUMAIN and Qualcomm	$2 B	N/A	Cloud-to-edge hybrid AI ecosystem, focusing on energy-efficient, low-latency inference for mobile and embedded systems

Figure 10. Current (2024) and planned (2030) data center hub capacities in Saudi Arabia.



Note: Planned capacity (1,716 MW) only includes projects with specified locations.

Source: Authors.

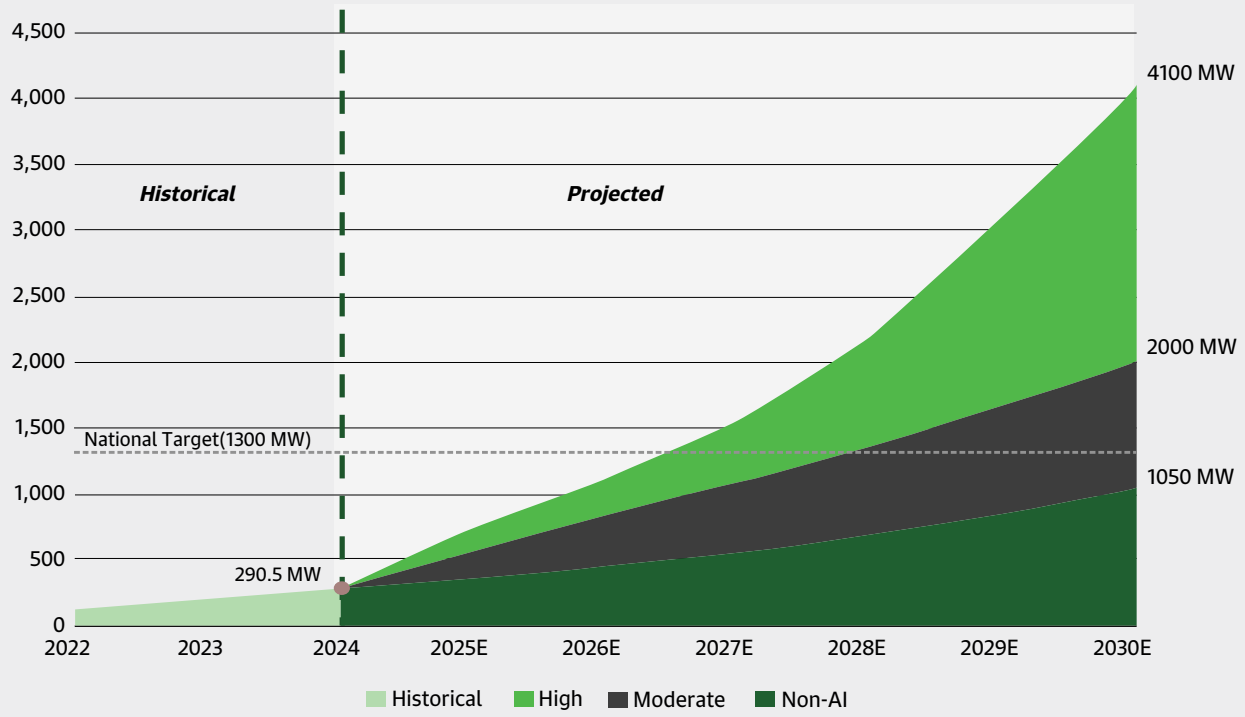
This expansion is expected to have significant implications for the Kingdom's digital infrastructure and energy landscape. Large AI campuses will place high-density demands on Saudi Arabia's electricity system in the coming years. AI introduces a new type of workload, namely continuous, GPU-intensive operations that differ from traditional data center activity. Understanding the future impacts of this shift on grid stability, energy sourcing, and sustainability will be critical for planning and coordinating the next phase of infrastructure growth.

4.2 Demand Projections

Understanding the future impact of data centers requires analyzing both the scale of potential capacity expansion and the operational conditions of facilities. While numerous projects have been announced, their ultimate realization, scale, and pace of deployment are uncertain. To account for this uncertainty, three growth scenarios for total installed capacity by 2030 are defined, as shown in Figure 11:

- **Baseline scenario:** This scenario reflects the continuation of the current trends, with growth driven primarily by general-purpose data centers, without a major expansion of AI-related centers. This serves as a reference point for capacity growth in the absence of large-scale AI transformation.
- **Moderate-growth scenario:** This looks at confirmed AI-oriented expansion, consisting of projects that are in the stages of permitting or early execution, based on an analysis of multiple data sources. This path is a realistic and achievable AI growth trajectory, with total capacity, combined with general-purpose data center expansion, potentially reaching 2,000 MW by 2030.
- **High-growth scenario:** This represents visionary growth in which all announced AI mega-projects are completed by 2030, adding roughly 3,050 MW of new AI-optimized capacity. Combined with the development of general-purpose data centers, this brings total installed capacity to around 4,100 MW by 2030. While these projects are currently under construction, they set the upper limit of potential capacity at the end of the decade.

Figure 11. Historical and projected Saudi Arabia data center capacities (MW) under three growth scenarios 2022-2030.



Note: Historical capacity is shown up to 2024 (290.5 MW). The non-AI scenario reaches about 1,050 MW by 2030, moderate growth reaches 2,000 MW, and high growth reaches 4,100 MW. The gray dashed line shows the national target of 1,300 MW.

Source: Authors.

Table 4. Modeled scenarios for data center capacity growth and operational conditions in Saudi Arabia through 2030.

		Operational condition	
		Conventional (PUE 1.5-1.7, as-is emissions)	**Sustainable** (PUE 1.3-1.5, low-carbon mix)
Growth scenario	**High growth** (4,100 MW by 2030)	**S1: High growth, conventional** • AI-first expansion • High electricity use and as-is CO_2 emissions	**S2: High growth, sustainable** • Green AI expansion • Improved efficiency from advanced cooling and chip design • Lower electricity use and CO_2 emissions
	Moderate growth (2,000 MW by 2030)	**S3: Moderate growth, conventional** • Steady AI expansion • Limited efficiency gains • Higher electricity use and CO_2 emissions	**S4: Moderate growth, sustainable** • Balanced AI growth • Enhanced efficiency from advanced cooling and hardware optimization • Lower electricity use and CO_2 emissions

While these scenarios outline how far Saudi Arabia's data-center capacity could expand by 2030, their broader impact depends on the efficiency of facilities and the sustainability of their energy supply. To explore these dynamics, two operational conditions are defined:

- **Conventional operational conditions:** Efficiency improvements are limited, advanced cooling technologies are slow to be adopted, and most electricity comes from fossil sources, typically resulting in higher power usage effectiveness (PUE)[7] values.
- **Sustainable operational conditions:** High operational efficiency achieved through advanced cooling, optimization technologies, and energy management practices. A significant proportion of electricity comes from renewable sources, resulting in lower PUE and emissions.

Combining moderate- and high-growth scenarios with these two operational conditions yields four possible combinations, shown in Table 4. These scenarios offer a unique perspective for evaluating the long-term impacts of AI-driven data center growth on the power system and sustainability goals. Importantly, these scenarios are not predictions but structured "what-if" analyses to stress-test policy and investment choices.

As data center capacity expands, translating installed IT load into electricity consumption is essential for assessing implications for the national power system. The most widely used methodology is a bottom-up approach (Masanet, Lei, and Koomey 2024), which calculates electricity use based on the power draw of IT equipment. This method provides the most accurate estimates but relies on detailed technical data, which is often unavailable for projects still in the planning stages. In such cases, electricity demand can be estimated from announced IT capacity (MW) and assumptions about PUE and usage rates. Although less detailed, this approach allows for robust projections of future demand.

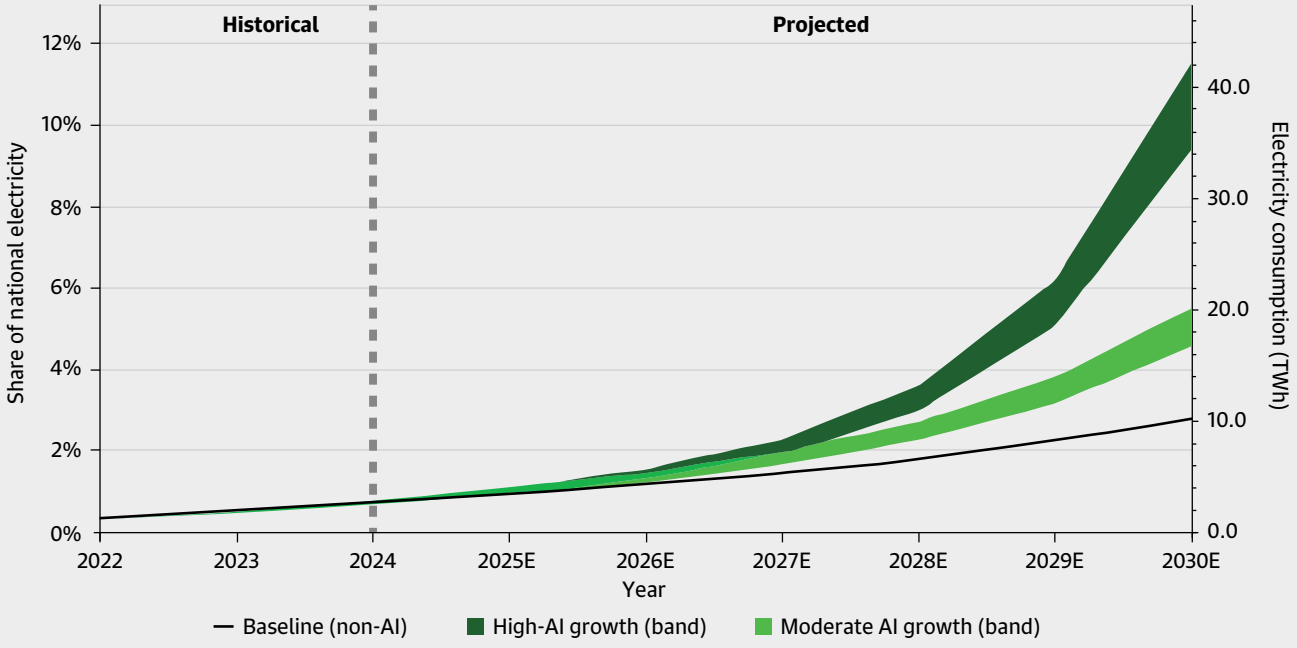
Based on this framework, Saudi Arabia's data centers consumed an estimated 2.8 TWh in 2024, accounting for around 0.85% of the country's total electricity consumption in that year.[8] By 2030, electricity consumption by data centers is projected to increase substantially, with estimates ranging from approximately 10.16 TWh to 42.23 TWh annually. The variation reflects differences in capacity expansion and the implementation of efficiency practices.

Figure 12 shows that under the baseline growth scenario, where expansion is limited to general-purpose facilities, annual consumption could reach 10.16 TWh, or 2.79% of national demand, reflecting steady but manageable growth. The introduction of AI-driven facilities significantly increases demand. Under the moderate-growth scenario, projected annual electricity demand could reach 20.15 TWh by 2030. However, this could be reduced to 17.62 TWh, a 13% reduction, with advanced cooling and improved design practices. The high-growth scenario results in the largest increase, with electricity demand rising to 42.23 TWh annually, equivalent to 11.55% of projected national electricity demand. This figure could be lowered to 36.76 TWh, a 13% reduction, by implementing efficiency standards.

[7] PUE: A data center metric that measures energy efficiency used by IT equipment.

[8] Assuming a PUE of 1.7, a use rate of 0.65, and Saudi Arabia's total electricity consumption in 2024 as 333 TWh.

Figure 12. Historical and projected electricity demand from data centers in Saudi Arabia, 2022-2030.



Note: The left axis shows the share of national electricity demand, while the right axis reports absolute consumption in TWh. The shaded bands' upper limits correspond to conventional operating practices, while the lower limits reflect sustainable operational conditions, detailed in Appendix B.

Source: Authors.

In all scenarios, electricity demand from data centers will rise over the decade. The extent of this increase will depend on the scale of AI-oriented projects and the level of efficiency in facility design and operation. Regardless of the growth scenario adopting sustainable practices can reduce energy consumption by 13% compared to conventional approaches.

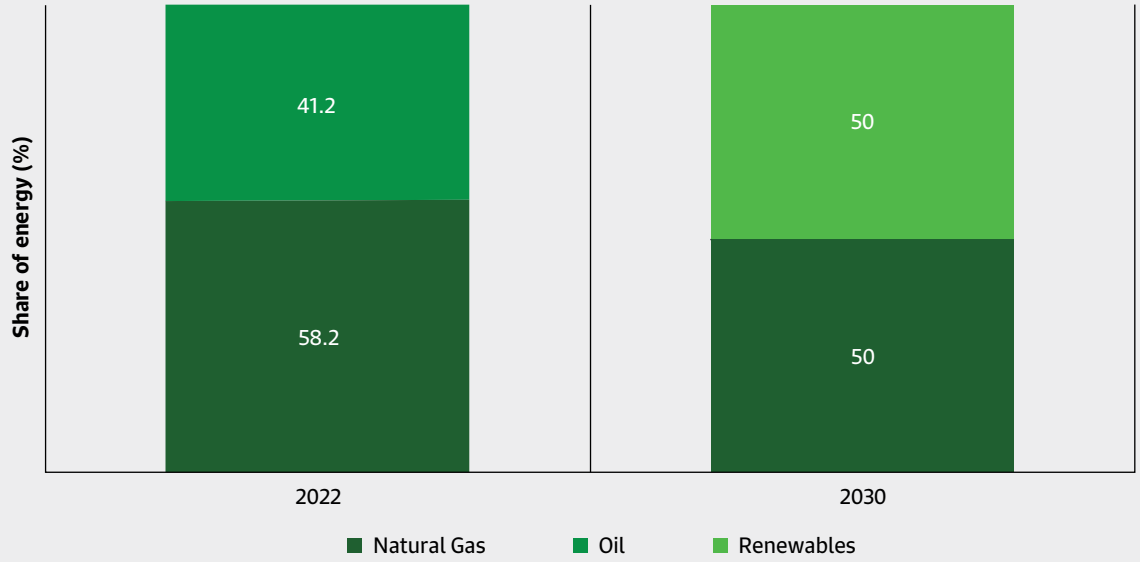
4.3 Emissions Projections

Electricity consumption is the main driver of data centers' carbon footprint, with their carbon intensity determined by the energy mix composition. In Saudi Arabia, where oil and natural gas are significant power sources, this challenge is being addressed through a commitment to achieve net-zero emissions by 2060 in line with the Paris Agreement and Saudi Vision 2030 (SGI 2025). The ambition is reinforced by the Saudi & Middle East Green Initiatives (2025) and an updated Nationally Determined Contribution that pledges to reduce annual emissions by 278 Mt CO_2 by 2030 (Saudi & Middle East Green Initiatives 2025). A central pillar of this strategy is the National Renewable Energy Program, which aims to raise the share of renewables to 50% of the energy mix by 2030 (Saudi & Middle East Green Initiatives 2025).

In this study, emissions are estimated under two scenarios: the conventional fossil fuel mix (41.2% oil and 58.2% natural gas), and the sustainable national target mix (50% renewables and 50% natural gas), as shown in Figure 13. For simplicity, we assume that the capacity mix and the electricity generation mix are the same. The 50/50 mix aligns with global projections, where nearly half of the electricity demand from data centers is expected to be met by renewable sources (Spencer et al. 2025).

With the fossil fuel mix, the electricity used by existing data centers, in 2024, produces about 1.6 Mt CO_2 each year. If growth follows a traditional and non-AI path, this could rise to 5.81 Mt CO_2 by 2030, while a moderate AI-driven scenario could increase emissions to 11.48 Mt CO_2. In the high-growth scenario, annual

Figure 13. Saudi Arabia's energy mix in 2022 compared with the 2030 national target.



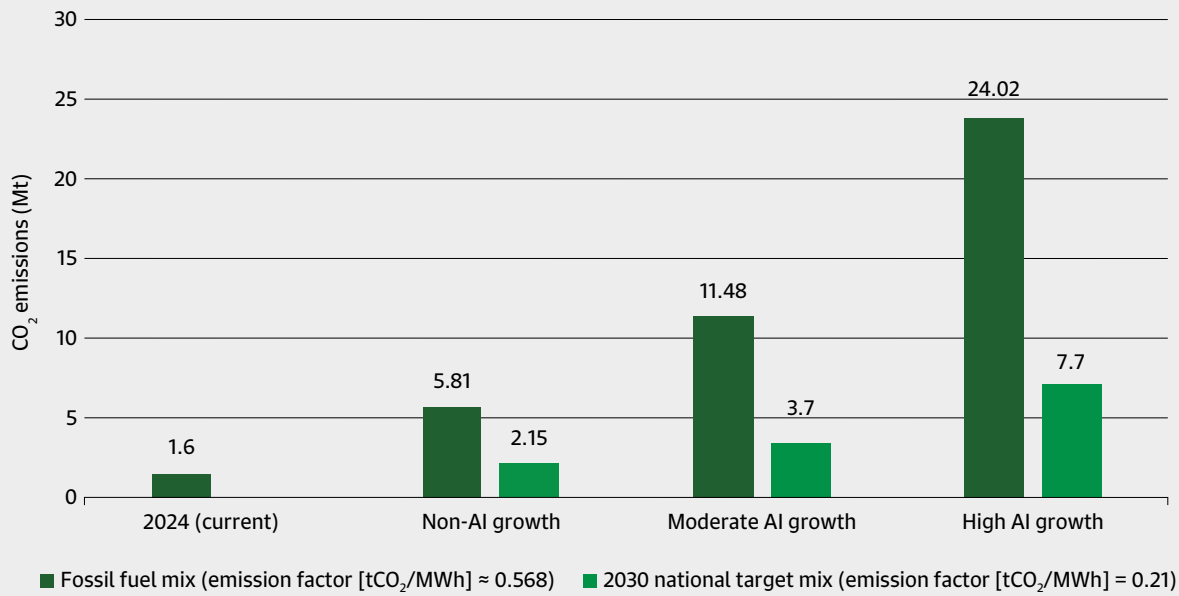
Source: Authors.

emissions could reach 24.02 Mt CO_2. However, shifting to that target energy mix would reduce emissions by roughly 68% across all scenarios, as Figure 14 shows. For example, moderate AI growth would emit 3.7 Mt CO_2 instead of 11.48 Mt CO_2, and high AI growth would drop from 24.02 Mt CO_2 to 7.7 Mt CO_2.

These reductions highlight the role of renewable integration in mitigating the climate impact of digital infrastructure expansion.

In addition to estimating total CO_2 emission from data centers, it is useful to track carbon intensity via carbon usage effectiveness

Figure 14. CO_2 emissions from data centers by scenario (Mt).



Source: Authors.

(CUE). CUE measures emissions intensity per MWh delivered to IT equipment (tCO_2/MWh-IT), where a CUE of zero indicates fully renewable or zero-carbon power. CUE is independent of load size, allowing for a clear separation between how clean operations are from how large the load is.

Conventional operations with a PUE of 1.5-1.7 on a fossil fuel mix yield a CUE of 0.85–0.97 tCO_2/MWh-IT, whereas sustainable operations with a PUE of 1.3-1.5 of the 2030 grid yield a CUE of 0.27-0.32 tCO_2/MWh-IT. This is about a threefold reduction in emissions intensity. CUE highlights that decarbonizing AI data centers depends primarily on facility efficiency (lower PUE) and the carbon intensity of the power source. High-growth scenarios raise absolute emissions unless paired with low-CUE operations, which explains the 68% drop in total CO_2 when the cleaner 2030 mix is applied.

4.4 Key Enablers

The growth of Saudi Arabia's AI data centers is supported by a strong national strategy, advanced regulation, strategic geography, and state-backed investment. The energy system is also very cost-effective, reliable, and has ambitious sustainability targets. These enablers, together with the Kingdom's digital transformation agenda, create an environment where AI infrastructure can develop competitively and sustainably. We discuss selected enablers next.

National Strategies

Vision 2030 prioritizes digital transformation and AI as key to economic diversification. Out of 96 Vision objectives, 66 are directly or indirectly tied to data and AI (SDAIA 2025b). The National Strategy for Data and Artificial Intelligence, set by the SDAIA, aims to position Saudi Arabia among the top 15 countries in AI readiness and to attract about 75 billion SAR in investments by 2030 (SDAIA 2025a). Complementing this, the National Cybersecurity Authority (NCA) published the National Cybersecurity Strategy (NCA 2018) in 2020 to create a secure and trusted cyberspace for digital infrastructure. Saudi Arabia also enjoys the world's lowest levelized costs for both wind and solar energy, along with targets for 50% renewable energy and ambitious sustainability goals. There is a clear complementarity between energy and AI-related policies that support the country's competitiveness in hosting and operating AI data centers.

Legal and Regulatory Environment

Saudi Arabia has developed a comprehensive and coordinated legal framework for information and communications technology, data governance, cloud services, sustainability, and energy efficiency, implemented through several institutions. Key regulations include the Telecommunications and Information Technology Act (Bureau of Experts of the Council of Ministers 2022), the Personal Data Protection Law (Bureau of Experts of the Council of Ministers 2021), and the Cloud First Policy (MCIT 2020), which together regulate data handling, cloud operations, and data center development. The Data Center Services Regulations (CST 2023, 2024b) enhance this framework by requiring licensed operators to prepare sustainability plans focused on energy efficiency, carbon reduction, and electronic waste management, while promoting environmentally friendly power and cooling solutions. In addition, the Global AI Hub Per website in reference Law (Istitlaa 2025) is a major step toward facilitating cross-border AI deployment through greater legal certainty and operational flexibility. It defines three categories of AI data centers: Private Hubs, Extended Hubs, and Virtual Hubs, allowing international providers to operate in Saudi Arabia under regulatory arrangements aligned with their home jurisdictions.

Strategic Location, Digital Infrastructure, and Connectivity

Saudi Arabia's geographic position at the intersection of Asia, Europe, and Africa gives it strategic advantages for regional and intercontinental data flow. The Kingdom is connected to the global internet by 16 active subsea cable systems across the Red Sea and Arabian Gulf, providing multiple landing points in five coastal cities (MCIT 2024). Domestically, fiber-optic infrastructure and 5G networks, covering approximately 65% of the population in 2024, have improved inland connectivity, supporting edge computing and other latency-sensitive applications (MCIT 2024). These infrastructure elements collectively create a technical foundation capable of supporting optimized AI data centers with reliable performance.

Energy Advantage and Grid Readiness

Saudi Arabia offers a distinctive energy advantage for AI data centers, characterized by low electricity prices, an expanding power grid, and a growing share of renewables. Electricity tariffs for cloud computing are as low as 18 halalah/kWh (about $0.048/kWh), while industrial tariffs stand at 20 halalah/kWh

(about $0.053/kWh) (SERA 2025). AI data centers typically fall within these tariff categories,[9] allowing operators to achieve a total cost of ownership up to 30% lower than in comparable international markets (Techusiness 2025). The national power grid is expanding rapidly, with transmission lines expected to increase from 96,496 km in 2024 to 160,000 km by 2030, and the number of substations rising from 1,235 to 1,650 (SEC 2025).

These developments are supported by ongoing investment in battery storage capacity (targeting 48,000 MWh) and grid modernization projects to enhance thermal efficiency and support future carbon capture systems (SPA 2024). Furthermore, the Kingdom's abundant solar irradiance and emerging wind capacity offer significant potential for green energy integration. Saudi Arabia is actively increasing its renewable capacity to 130,000 MW by 2030, with 44,000 MW already tendered and 20,000 MW added in 2023 (KAPSARC 2025). The Kingdom has also set global benchmarks for renewable energy costs, exemplified by the 2024 Al-Shuaiba solar project (Techusiness 2025), which achieved one of the world's lowest electricity generation costs at $0.0104/kWh (Bellini 2021). These developments improve the potential for long-term price stability and decarbonization for energy-intensive AI infrastructure.

Government Capital and International Collaboration

The expansion of AI data centers in the Kingdom is driven by government-backed capital, sovereign investment, and strategic international partnerships. The Public Investment Fund (PIF) plays a central role by investing in HUMAIN and other national AI infrastructure ventures. In addition, strategic collaborations with global technology leaders, including NVIDIA, AMD, and AWS, are enhancing computational capability and technology transfer. These initiatives are supported by the government, with attractive energy and sustainability offerings, land allocation, and infrastructure support, designed to attract long-term investors and speed up deployment.

4.5 Factors that Could Influence Projections

Electricity demand from AI data centers through to 2030 will depend on several interacting factors. The most important are the scale of AI adoption, advances in hardware, improvements in infrastructure and software, and, over the longer term, new computing paradigms. These factors will determine whether energy use rises steeply or can be balanced by efficiency gains. The scale of AI adoption across government, business, and society will be a major driver of future data center electricity demand. As AI is used in more critical sectors, the need for powerful computing will rise. Energy use will also depend on how complex the tasks are. Highly complex tasks require more energy due to longer context windows, deeper reasoning chains, and higher token generation. For example, a simple chatbot query may consume around 1.55 Wh, while a retrieval-augmented generation query requires roughly 2.64 Wh, and an agentic workflow requires 8.54 Wh on average (Desroches et al. 2025).

Efficiency improvements will play a central role in shaping the future energy demand of AI data centers. Advances are happening at several levels:

- **Hardware:** New generations of processors deliver far more computing power for each unit of electricity consumed. Specialized chips designed for AI, like GPUs and TPUs, are much more efficient than traditional CPU processors. Each generation also becomes more powerful as capabilities expand. For example, NVIDIA's Blackwell B200 GPU offers better energy efficiency than its predecessors.

- **Infrastructure:** Data centers are improving electricity use through better cooling systems and more innovative design. While much progress has already been made, future improvements are expected to deliver further gains.

- **Software:** Smarter algorithms help reduce the amount of computing needed for complex AI tasks, reducing energy demand while keeping performance high. For instance, DeepSeek-R1 reduced energy use with smarter design

[9] The exact tariff that data centers will be charged in the Kingdom is yet to be decided. Nonetheless, we refer to these tariffs since they are the only ones that are publicly available.

techniques, like activating only needed parts of the model, running computations in lower precision, and balancing work across hardware, resulting in notable efficiency gains.

- **New computing paradigms**: Looking further ahead, emerging technologies such as quantum and neuromorphic computing can offer potential for energy efficiency in specific AI applications. While their commercial impact is unknown, these technologies could fundamentally change the energy landscape of AI.

- **Demand response management:** Participation of data centers in demand response programs could significantly influence the electricity demand of data centers. By adjusting non-critical computing workloads during high system load or price spikes, operators can reduce peak consumption without compromising service reliability. Wider adoption of automated workload scheduling and real-time pricing mechanisms would smooth demand curves, lower grid stress, and marginally reduce total annual electricity use.

Ultimately, while efficiency gains can lower the energy required for a given amount of computation, they also tend to reduce the cost of computing. This can trigger what is known in the energy efficiency field as the rebound effect (or the Jevons Paradox), where improved efficiency lowers costs, thereby stimulating higher demand and wider adoption of AI applications. As a result, total energy use may continue to rise even as individual systems become more efficient.

Cost Analysis of AI Data Centers in Saudi Arabia

05

This section presents an in-depth cost analysis supporting the assessment of AI data center demand growth in Saudi Arabia. It examines key cost drivers, including electricity pricing, energy and computing efficiency, and other operational parameters that shape competitiveness. The analysis outlines the modeling framework applied to estimate data center costs and then applies it to a detailed case study of Saudi Arabia's market conditions.

5.1 Calculating AI Data Center Project Costs

AI workloads need substantial and ongoing computing capacity, making data center development both capital intensive and operationally complex. To evaluate whether expanding capacity adds value, investors and policymakers need a clear understanding of the full life cycle economics of building and operating AI infrastructure. This can be done through a levelized cost analysis framework, an established method for calculating costs over a project's lifespan. The framework identifies the minimum price at which computing services must operate to recover capital and operating expenses, thereby assessing overall project viability. The breakeven cost per unit of compute is evaluated over the facility's lifetime, allowing sensitivity analysis of key factors such as electricity price, PUE, and load factor.

There is a need for a quantifiable economic performance metric to evaluate data center infrastructure dedicated to AI workloads (Kristiansen Nøland, Hjelmeland, and Korpås 2024). In this analysis, we estimate the average cost of delivering each unit of compute over the facility's operational life. This involves summing all costs of construction, equipment, electricity, and maintenance over time and dividing by the total computing output produced. The results show the minimum price that must be charged per unit of computing power for the facility to break even over its lifetime. This follows the logic of the levelized cost of energy, which compares projects by measuring lifetime costs per unit of output. Applying it to data centers provides a consistent means of assessing long-term economic viability under different operational and market conditions. In this case, the analysis measures the cost per unit of compute – typically expressed in dollars per floating-point operation ($/FLOP), and often scaled to $/PFLOP or $/EFLOP due to the large computational output involved.[10]

Several factors affect the cost of data center projects: PUE, load factor, computing efficiency, CAPEX, operating expenditure (OPEX), and the weighted average cost of capital (WACC). Table 5 provides an overview of each factor.

[10] A FLOP is a single floating-point operation, which is used as a measure of computational performance. FLOPs is the total count of floating-point operations performed to complete a specific task or program. The number of FLOPs a system or processor – such as a CPU, GPU, or supercomputer – can perform in one second is used as a measure of computational performance. PFLOP (petaFLOP) = 10^{15} FLOPs. EFLOP (exaFLOP) = 10^{18} FLOPs..

Table 5. Factors that affect the average data center project costs over its lifetime.

Factor	Definition	Relevance to data centers	Ideal value	Best recorded value
PUE	The ratio of total energy consumed by the data center to the energy consumed by IT equipment only	Measures overall energy efficiency	1, where all the energy consumption goes to IT equipment	1.028, by the National Renewable Energy Laboratory (NREL) (Van Zandt 2023)
Load factor	The ratio of average IT load to peak load in a data center over a specific period	Indicates operational efficiency and utilization	100%	~90%
Computing efficiency	The number of PFLOPs/kW	AI hardware efficiency	-	~7.71 PFLOPs/kW by NVIDIA's Blackwell Ultra chip (2025)
Electricity price	Unit price of electricity	Determines ongoing operating costs	-	-
OPEX	Annual operating and maintenance expenses as a percentage of CAPEX	Represents recurring costs during operation	-	-
CAPEX	Overnight cost of construction	Required initial investment	-	-
WACC	Weighted average cost of capital	Reflects the cost of financing and investor return expectations	-	-

5.2 Baseline Results and Sensitivity Analysis

To evaluate Saudi Arabia's data center project cost, the analysis assigns default values to each factor. Table 6 summarizes these values, along with a justification for each selection. It is worth noting that some values were varied in subsequent sensitivity analyses to explore policy insights, assess sensitivity to key factors, and explore trade-offs.

Table 6. Default values used in the case study.

Factor	Default value	Justification
PUE	1.5	Average for existing data centers in Saudi Arabia (range 1.5-1.8)[11]
Load factor	80%	Illustrative mid-range assumption for AI-oriented facilities (70%-90% is typical, though actual use may vary) (Aterio 2025)
Computing efficiency	0.21 PFLOPs/kW	Median for AI hardware values (see Table 7)
Electricity price	$48/MWh	Equivalent to the official tariff for cloud-computing 18 Halalah/kWh (SERA 2025)*
CAPEX	$10,000/kW	Baseline assumption (Kristiansen Nølan, Hjelmeland, and Korpås 2024), and similar to Equinix planned hyperscale AI data center (Malik 2025)
WACC	10%	Baseline assumption (Kristiansen Nøland, Hjelmeland, and Korpås 2024). Lower financing costs are possible under domestic conditions
Annual O&M cost[12]	10% of CAPEX	
Project lifetime	15 years	

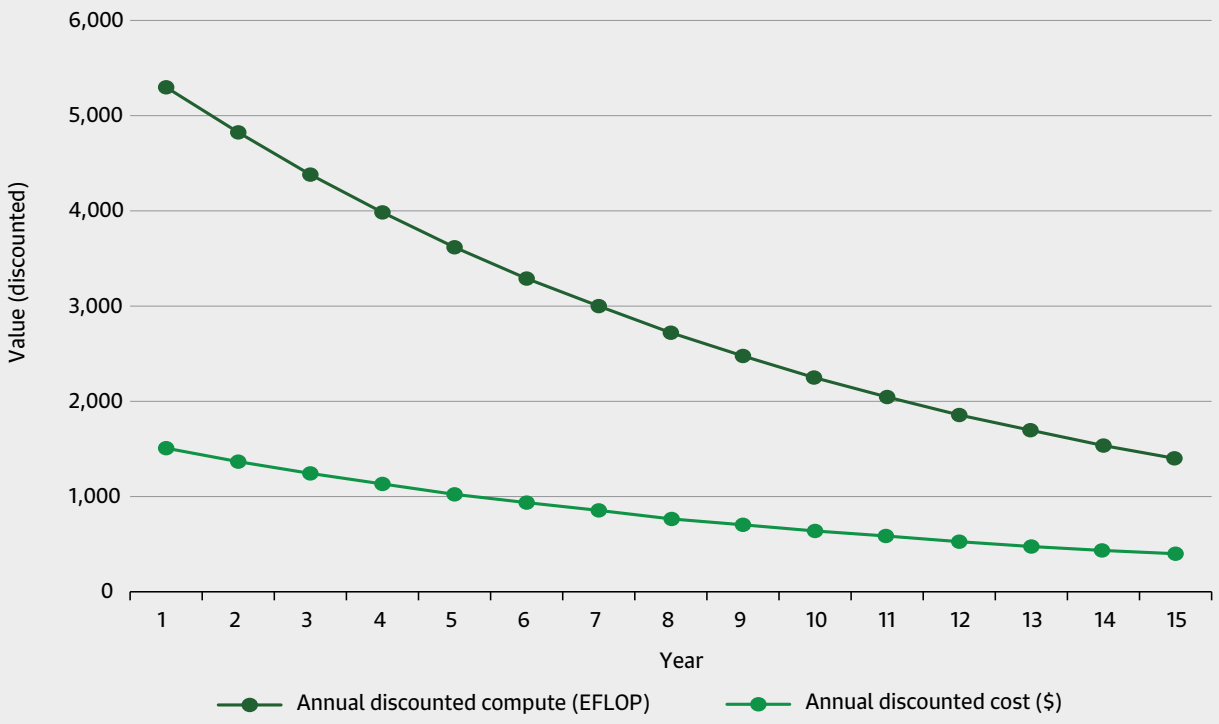
* The exact tariff that data centers will be charged in the Kingdom is yet to be decided. Nonetheless, we use this tariff as an assumption since it is publicly available.

Note: The baseline case applies representative, literature-based parameters for AI data center operations in Saudi Arabia. These values are indicative and intended to illustrate cost sensitivities rather than to forecast specific project outcomes. Actual project performance may vary depending on facility scale, technology choice, financing conditions, and operational strategy.

[11] S&P Global data.

[12] Annual O&M cost: Yearly operations and maintenance cost (OPEX) of total capital expenditures (CAPEX).

Figure 15. Annual discounted cost and compute output over time.



Source: Authors.

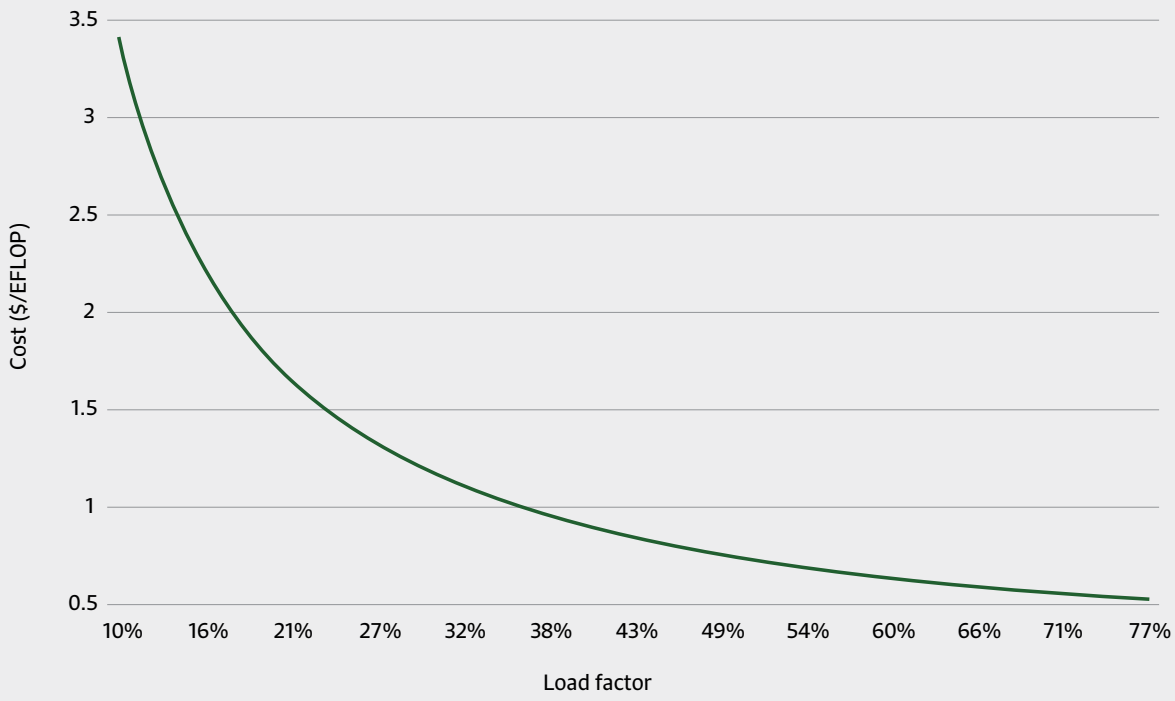
Using the default values, the baseline unit cost is estimated at $0.51/EFLOP, where 1 EFLOP = 1,000 PFLOPs. This represents the average lifetime cost of computation under typical Saudi operating conditions, assuming a PUE of 1.5, 80% load factor, and an electricity price of $48/MWh. Comparable analyses, such as Kristiansen Nøland, Hjelmeland, and Korpås (2024), report levelized computing costs of about $1.0-$1.2/EFLOP under higher electricity prices ($75-$125/MWh), a PUE of 1.12, and a computing efficiency of 0.1 PFLOPs/kW. When normalized to that same efficiency, the equivalent cost in this study would be approximately $1.07/EFLOP, placing it lower than the median of the range. This indicates that Saudi Arabia's AI data centers could be cost-competitive even with the higher PUE values typical of warm climates.

To interpret this result, Figure 15 compares the annual discounted cost with the annual discounted compute output over the project lifetime. Both measures decrease over time due to discounting, which reduces the present value of future expenditures and computing services. The decrease in discounted compute output does not reflect hardware degradation; it represents the diminishing present value of future computational capacity.

The key result is that most of the project's economic value is captured early, when effective usage is high relative to cost. This is especially true for AI data centers, where rapid technology cycles and hardware replacement rates make early operational efficiency essential for recovering capital before systems become outdated.

The unit cost of compute in any given year can be inferred by comparing the annual discounted cost with the discounted compute output. Aggregating this relationship across all years gives us the average lifetime cost of computation. While this relationship reflects standard capital recovery dynamics, it also highlights why AI data centers exhibit stronger front-loading effects than most infrastructure assets. Their high capital intensity, rapid technology turnover, and performance gains in early operating years mean that utilization achieved in the initial phase has a disproportionate effect on total cost recovery. In this sense, early and sustained high load factors are not just desirable but critical to maintaining cost competitiveness in a fast-evolving compute market.

Figure 16. Sensitivity to the load factor.



Source: Authors.

To gain further insights, the correlation between average data center project costs over its lifetime and multiple factors is examined. These factors include load factor, computing efficiency, electricity price, and PUE.

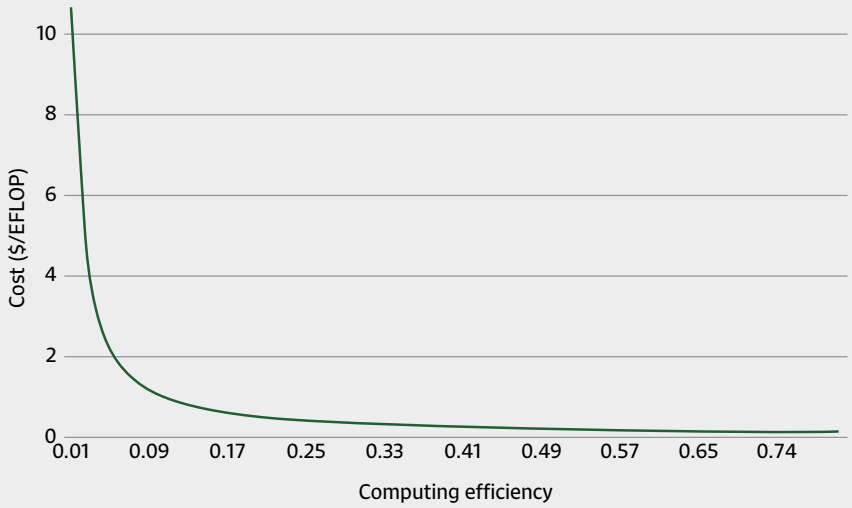
Load Factor

As the load factor increases, unit costs decline sharply, as Figure 16 shows. At low usage, costs exceed $3.40/EFLOP, but as the load factor rises toward stable operation, they quickly drop to near $0.5/EFLOP. The curve shows diminishing marginal cost savings at higher usage levels, meaning that most efficiency gains are achieved when data centers operate near steady capacity. This flattening of the cost curve highlights the importance of maintaining consistently high usage to stabilize both operating costs and power consumption. Operators often achieve this by smoothing short-term demand fluctuations through background compute tasks, such as model training, which help ensure a more constant load on the grid.

Computing Efficiency

Here, we investigate how computing efficiency – the computational output per kilowatt of IT power – affects average costs. As shown in Figure 17, improvements in computing efficiency significantly lower costs. The x-axis ranges between approximately 0.01 and 0.8 PFLOPs/kW, beyond which the curve saturates and cost reductions become minimal. At low efficiency, costs are very high, but as efficiency increases, costs fall sharply and eventually flatten. This aligns with industry trends, as hyperscale operators and AI hardware manufacturers aim for higher computing efficiency to balance throughput and cost competitiveness. This result is consistent with the growing interest of hyperscales and AI hardware manufacturers in achieving greater computing efficiency, as shown in Table 8. In essence, investing in highly efficient compute platforms not only improves throughput but also enhances long-term cost competitiveness.

Figure 17. Sensitivity to the computing efficiency.



Source: Authors.

Table 7. Computing efficiency for various AI hardware.

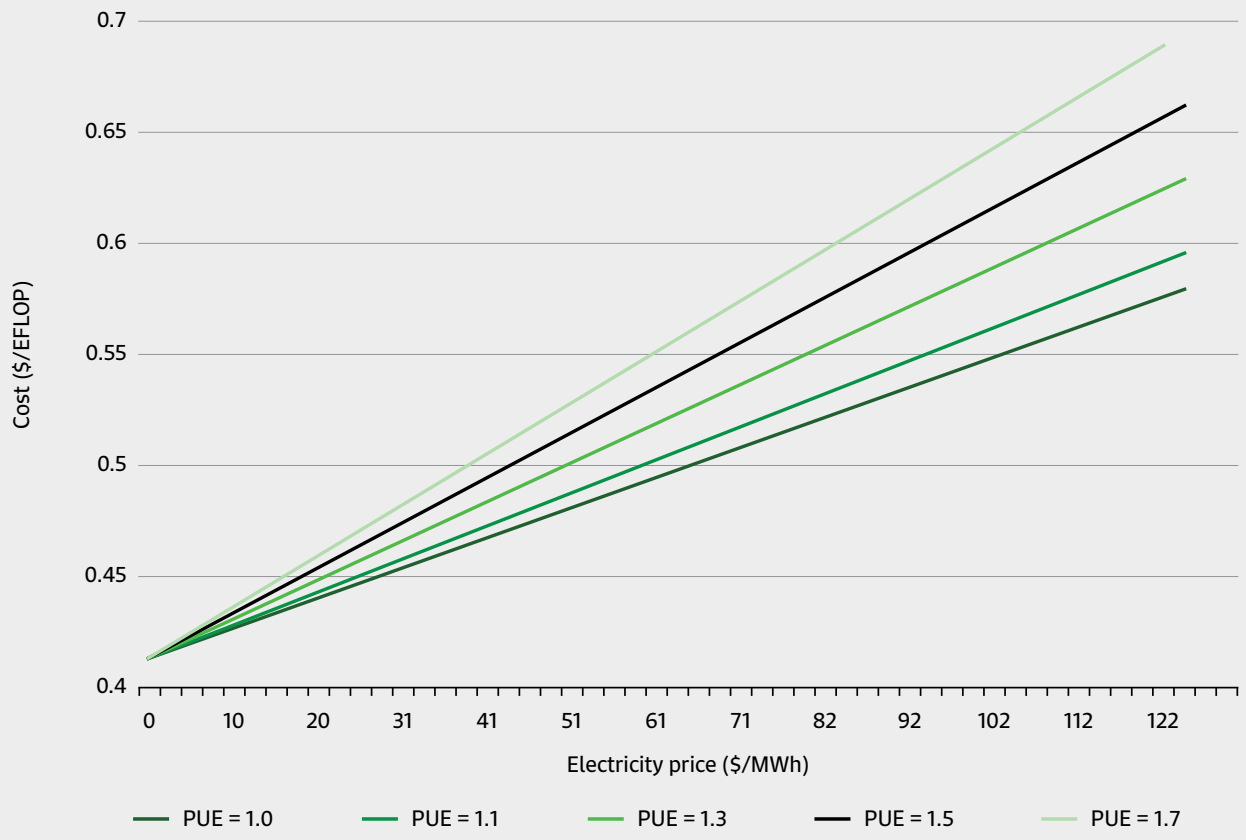
Provider	GPU model	Year released	Computing efficiency (PFLOPs/kW)
Intel	GPU Flex 140	2022	0.107
	GPU Flex 170	2022	0.107
	GPU Max 1100	2022	0.048
	GPU Max 1550	2023	0.049
AMD	Instinct MI300X	2023	0.209
Cerebras (Wang 2024)	CS-3	2024	5.43
NVIDIA	Tesla T4	2018	0.116
	Tesla V100	2017	0.052
	DGX A100	2022	0.769
	DGX H100	2022	3.137
	DGX B200	2024	5.035
	GB200 NVL72	2024	6.000
	Blackwell Ultra	2025	~7.71

Electricity Price and PUE: Trade-off Curves

A policy-sensitive lever is the trade-off between electricity price and PUE. Figure 18 presents the relationship between electricity price and data center project costs for different PUE scenarios, ranging from a PUE value of 1.0 (ideal efficiency) to 1.7 (less efficient, reflecting more cooling and overhead consumption). A PUE of 1.1 is typical for AI data centers located in cool climates, while a PUE of 1.5 represents the average for data centers in Saudi Arabia.

All scenarios originate from a common baseline at zero electricity cost, highlighting that fixed costs – CAPEX and non-electric OPEX – set a minimum threshold for computing costs. The slope of each line reflects sensitivity to electricity prices, where steeper slopes correspond to higher PUE values.

Figure 18. Trade-offs between PUE and electricity price.



Source: Authors.

This means that less efficient facilities experience a faster cost escalation as electricity prices rise. While the data center project cost is influenced by power prices, it is less sensitive compared to factors such as load factor or computing efficiency.

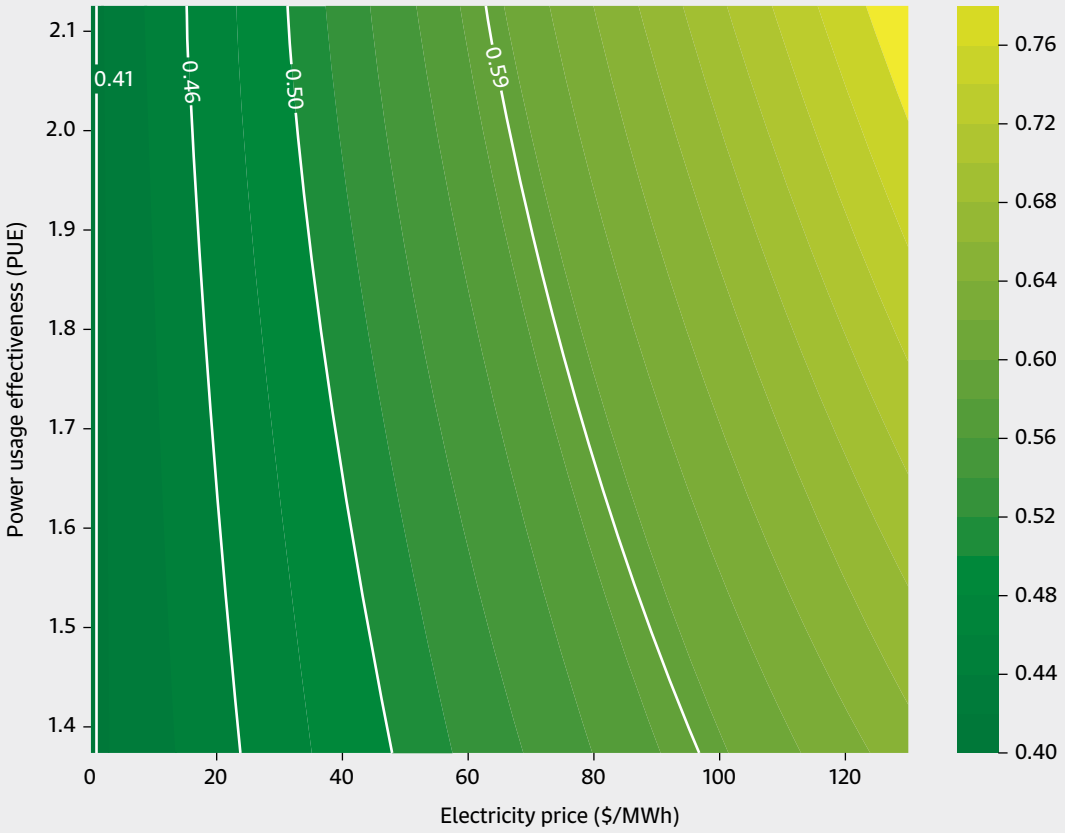
The figure also highlights an important policy and siting insight: electricity price and PUE interact multiplicatively in determining compute economics. For example, even at a relatively high PUE of 1.7, data centers in regions with low industrial tariffs, such as Saudi Arabia where electricity is priced around $0.048/kWh (18 halalas/kWh), can still maintain competitive project costs. This indicates that low-cost electricity markets allow design trade-offs, allowing developers to operate at higher PUEs within climatic or infrastructure constraints, while keeping costs at an acceptable level.

To visualize the combined effects of electricity prices and PUE on data center projects costs, Figure 19 maps the cost across different electricity prices and PUEs. At low prices of $20-$40/MWh, even inefficient data centers with a PUE above 2 can maintain a low cost. For example, at $24/MWh and a PUE above 2.0, the cost remains below $0.55/EFLOP. This shows that when electricity prices are not a limiting factor, operators can tolerate less efficient cooling infrastructure without losing their competitive edge.

At the Saudi benchmark of $48/MWh for cloud computing, a facility with a PUE of 1.7 achieves roughly $0.55/EFLOP. However, if electricity prices double to $96/MWh, the same facility's cost would rise above $0.60/EFLOP, representing at least a 10% increase in costs. This figure reinforces the idea that policymakers and developers must consider both factors together.

In Saudi Arabia, the hot climate makes achieving ultra-low PUE challenging, but electricity prices are competitive. That puts the Kingdom in a "competitive zone" on the map: even with a moderate PUE (e.g., around 1.6-1.8), the costs stay low

Figure 19. Contour map of the cost as a function of PUE and electricity price.



Notes: Each contour line represents a constant cost value, and the shading reflects the gradient of cost.

The plot includes several vertical reference lines to guide interpretation: one at $24/MWh (half of the cloud computing tariff), one at $48/MWh (the benchmark tariff for cloud computing in Saudi Arabia), and one at $96/MWh (double tariff).

at current tariffs. Data center designers can choose practical cooling solutions and still be cost-competitive. As future infrastructure is planned, maintaining favorable tariffs and enabling high-efficiency design practices will be central to sustaining this advantage.

5.3 Policy Insights

The cost analysis places Saudi Arabia favorably in the global landscape for developing AI data centers. The Kingdom's electricity tariffs are already among the most competitive in the world, and are supported by reliable grid infrastructure and cost stability. This advantage can offset the higher cooling requirements typical of warm climates and provides a strong foundation for investments in AI and cloud infrastructure.

Our analysis shows that further reductions in electricity tariffs would yield only limited benefits, while improvements in energy efficiency and use offer far greater potential to lower overall compute costs. Enhancing cooling performance, hardware efficiency, and load management is therefore the most effective path to strengthening competitiveness. Policy could focus on how energy is used rather than how cheap it is, by promoting advanced technologies and operational practices that improve efficiency and maintain high, stable usage. From a policy standpoint, keeping electricity prices stable, establishing clear performance and efficiency standards, and encouraging the use of the latest in hardware are the most effective levers for sustaining long-term competitiveness.

Global Overview of AI Data Center Challenges and Risks

06

The expansion of AI data centers has the potential to create economic and technological opportunities. However, as AI data centers proliferate and increase their capacities, various risks and challenges must be considered. Challenges reflect structural and technical difficulties to AI data center growth, while risks highlight potential adverse outcomes. This section analyzes both dimensions globally, not tied to any specific region.

6.1 Challenges

Regulatory and Policy Pressures

Governments and regulators increasingly impose new rules on data centers to meet environmental objectives and infrastructure planning needs. In mature markets, there are limits on new builds and binding sustainability requirements, such as disclosing annual energy and water use, stricter efficiency standards, and commitments to 100% renewable power. These policies reflect a broader regulatory trend in which data centers must align with national climate targets or face permitting hurdles and growth caps. For operators, this raises the bar on compliance and increases costs. Meeting new rules will require larger investments in renewable procurement, higher-efficiency cooling, heat-recovery systems, and better measurement and reporting.

Some countries' data protection and localization regulations require specific data to be stored domestically. At the same time, renewable-energy compliance mechanisms, such as RECs and green-power mandates, are becoming increasingly common, requiring data center operators to buy or generate certified renewable electricity to meet sustainability targets. Overall, regulatory pressures are rising on multiple fronts – spanning energy sourcing, water use, carbon emissions, and data governance – creating a more complex compliance landscape for developers. Failure to meet these evolving standards may raise costs or slow expansion, particularly in markets where environmental and energy regulations are tightening.

Land Use and Infrastructure Bottlenecks

As AI data centers increase, physical infrastructure and land availability become key constraints. These facilities need more than just open space; they require particular siting conditions with access to high-capacity power grids, fiber-optic networks, cooling resources, and stable geology. Globally, this has led to clustering in major metropolitan hubs, often causing local bottlenecks in grid capacity, land zoning, and fiber connectivity. While land is not inherently scarce in many countries, the challenge lies in finding sites that meet technical criteria for reliability and latency, and infrastructure requirements like

substation access, road transport, and chilled-water cooling systems. This challenge is compounded when backup power systems and cooling plants are also required.

Moreover, as data centers integrate renewable energy sources, the land required for on-site generation – such as solar PV – can become a significant constraint, particularly near urban or industrial areas. In those cases, relying on grid connections to large-scale renewable plants is often more practical and land-efficient than developing generation on site. This allows renewable power to be sourced from utility-scale plants in high-resource areas while minimizing land-use footprint and logistical complexity at the data center itself.

Talent Shortages and Workforce Constraints

The AI data center sector faces a global shortage of skilled professionals in areas such as electrical engineering, cybersecurity, network design, and AI infrastructure. The U.S. alone faces a projected shortfall of more than 300,000 workers by 2025 (Levine 2025). Similar gaps exist in Europe and Asia, where training opportunities have not kept pace with hyperscale development.

While Saudi Arabia invests in digital skilling and localization through Vision 2030, a significant talent gap remains. Specialized training for AI workloads and advanced cooling systems (e.g., liquid cooling) is not yet widespread. Relying on foreign labor may be unavoidable in the short term, and long-term growth depends on developing a skilled local workforce. Expanding university programs, establishing vocational centers focused on data center operations, and fostering industry-academic collaboration will be essential to securing operational continuity and supporting domestic employment goals.

6.2 Risks

Energy Supply Constraints

The rise in the number of AI data centers is placing significant strain on electricity grids and raising concerns about whether energy supply and grid infrastructure can keep up. Large hyperscale campuses can demand 100 MW or more – equivalent to the power use of a steel mill or hundreds of thousands of electric vehicles (Spencer and Singh 2024). The International Energy Agency (IEA) notes that delays to grid- connections are a major risk, caused by lengthy permitting processes and long transmission lead times. Connection queues in the U.S. average one to three years, and in some areas, such as Northern Virginia, can exceed seven years (Spencer et al. 2025). McKinsey likewise highlights “time to power” as the main concern for data center operators when building new sites (Green et al. 2024).

Beyond grid permitting, bottlenecks in equipment and power generation are becoming critical barriers. Transformer manufacturing lead times have more than doubled since 2020, with utilities and developers now facing wait times of more than two years for high-voltage units (IEA 2024). Similarly, gas turbine supply chains are facing unprecedented backlogs, as manufacturers prioritize delivery for grid-stability and industrial projects, further delaying the deployment of on-site backup power. These constraints exacerbate delays for connection and make it increasingly difficult to synchronize grid access, equipment delivery, and construction timelines.

Together, these challenges could delay projects, increase costs, or force reliance on carbon-intensive backup generation, undermining both growth and sustainability targets. Securing an adequate supply of grid infrastructure and critical components, along with efficient permitting processes, is now a strategic necessity for sustaining data center expansion in any region.

Environmental Footprint

The growth of data centers raises concerns about environmental sustainability. Today, data centers account for nearly 0.5% of global fuel-combustion CO_2 (about 180 Mt) and are projected by the IEA to reach about 1% by 2030, the equivalent of about 3% of global electricity (Spencer et al. 2025). Much of this environmental impact comes from the energy demand of servers and cooling systems, particularly if that energy is generated from fossil fuels.

Water consumption is also a sustainability concern due to the large water footprint of data centers. Most hyperscale facilities rely on water-based cooling to dissipate heat generated by servers. A small 1 MW data center can consume as much as 25.5 million liters of water per year (Mytton 2021), equivalent to the yearly use of more than 60 U.S. households (Sharma

2024). Larger U.S. data centers can draw as much water as 2 million households annually (Sharma 2024). Such demands are especially problematic in dry regions. With Saudi Arabia's desert climate, it is essential to balance large-scale data center growth and cooling needs with sustainable water management practices.

Investment Uncertainty

Massive capital investment is pouring into data center infrastructure to meet the growing demand for AI and cloud computing. However, the sector's long-term returns are not guaranteed. This AI-driven expansion has been likened to a "gold rush" (Graham, Rutten, and Yashkova 2024), with companies racing to build new facilities based on the assumption of sustained exponential demand growth. The risk is that projected demand is overestimated. If the economy slows down or AI fails to deliver the expected business value, hyperscale providers could reduce spending, which would slow the pace of expansion.

Another concern is the possibility of oversupply. With so many new players and investors entering the market, some regions may have more capacity than needed, which could push down occupancy rates and prices. At the same time, rising construction costs and high interest rates add financial pressure to new projects. Together, these factors mean that while growth prospects appear to be strong, the long-term profitability and usage rates of new data centers are uncertain.

Geopolitical and Supply Chain Risk

The global availability of high-performance chips, particularly GPUs and TPUs, is a critical bottleneck for scaling AI data centers. AI infrastructure relies on a few global chokepoints: semiconductor fabrication primarily in Taiwan, South Korea, and the U.S., and access to rare earth elements and batteries sourced mainly from China and Central Africa. This makes the sector highly vulnerable to geopolitical tensions, export controls, and supply chain disruptions, all of which could severely impact the availability of GPUs, memory, and power backup systems (Spencer et al. 2025).

Recent U.S. export bans on high-performance chips to China and other nations show how quickly access can be constrained. Globally, lead times for data center hardware have already increased, with critical components like power transformers facing delays of 18-24 months (Green et al. 2024). To mitigate this vulnerability, countries may develop diversified procurement strategies, localized assembly of non-sensitive hardware, and long-term contracts with suppliers. Building regional stockpiles of critical components or investing in MENA-wide programs of supply chain resilience could strengthen the Kingdom's position as a secure and reliable infrastructure host.

Toward Sustainable and More Efficient AI Data Centers

07

This section discusses different ways of reducing carbon emissions and enhancing the operational efficiency of AI data centers. We also provide an analysis of selected existing policies globally targeting more sustainable and efficient deployment and use of AI data centers.

7.1 Techniques for Enhancing Operational Efficiency

Facility Design and Modular Architecture

Modern AI data centers are moving from single mega-buildings to prefabricated "blocks" that bolt together like Lego. Each block ships with power distribution, a liquid-ready cooling loop, and sensors, so operators can add extra capacity when demand grows. Factory-set airflow paths keep hot and cold air apart and remove the "dead zones" that waste energy in custom halls. As a result, new modular server halls are already recording PUE values of approximately 1.2 (EPRI 2024), about 30% better than the global average of 1.55 in 2022 (Duncan et al. 2024).

Using data center infrastructure management tools that monitor every rack in real time and shift workloads or fan speeds can help to keep each one running at its optimal performance. Half the facility managers surveyed expect this alone to increase efficiency within five years (Donnellan 2023). When combined with using virtual servers, which has reduced physical server counts by 30%-40% in pilot projects, idle energy can fall to a single-digit share of total IT load (EPRI 2024).

Example: Meta's Data Center Heat Recycling in Denmark

Meta's aim is to run efficient, certified data centers on 100% clean and renewable energy. As of 2024, Meta's contracted portfolio is more than 11,700 MW of renewables (Meta 2024) and its hyperscale data center in Odense, Denmark, has become a showcase of circular energy innovation. The data center is linked to the city's district heating grid so that warm air from the server halls is routed to a dedicated heat-recovery system. The resulting surplus heat is delivered to the district heating network and distributed to homes and businesses through Odense's existing pipes.

According to Meta, 165,000 MWh per year of surplus heat from Odense's server halls is now supplied into the local district heating system, reaching up to 9,000 households (Meta 2025). In addition, the campus achieved LEED® Gold and was named Green Data Centre of the Year in 2021, reflecting an efficiency-first design that enables reliable heat capture.

AI Optimized Hardware: ASICs and Accelerators

The energy efficiency of hardware can be improved through specialized AI accelerators, including application-specific integrated circuits (ASICs) and TPUs. These chips are designed for AI tasks that can be carried out in parallel with each other, such as matrix multiplications, which dominate deep learning workloads. Epoch AI finds that the energy efficiency of frontier machine learning chips has been rising by approximately a factor of two every two years (Rahman 2024). Together, such accelerators form the basis of sustainable AI hardware roadmaps in hyperscale projects, cutting kWh per computation even as model sizes grow.

Advanced Cooling Techniques

Traditional air cooling is becoming insufficient for high-density AI computing, particularly with GPUs exceeding 1 kW of power per chip (Spencer et al. 2025). A conventional air cooling system can account for up to 40% of a data center's electricity use (Duncan et al. 2024). Liquid cooling technologies, such as direct-to-chip and immersion systems, offer substantial thermal and energy efficiency improvements, reducing power use by more than 50% while also enabling higher compute density and less floor space usage (Duncan et al. 2024; EPRI 2024). These improvements can lower PUE to below 1.1 (EPRI 2024). However, as liquid systems often need significant volumes of water, they bring sustainability challenges. To mitigate those challenges, rainwater harvesting and geothermal cooling have been proposed as alternative solutions that could reduce both energy and water demand for AI data center cooling (Stansbury et al. 2025).

Example: Microsoft's Waterless Cooling Innovation

As AI workloads surged, Microsoft faced a growing sustainability risk: traditional cooling consumes large volumes of potable water, which is especially problematic in arid regions with high community and regulatory expectations (Solomon 2024). In August 2024, Microsoft introduced a "zero-water" cooling architecture for AI-optimized centers, replacing a traditional evaporative system with chip-level direct liquid cooling in a closed loop. Once filled during construction, the coolant circulates continuously between servers and heat exchangers to maintain precise temperatures without using additional water. Additionally, Microsoft pairs the loop with high-efficiency chillers to minimize the energy trade-off of replacing traditional evaporation and operates at higher server inlet temperatures.

Each method is projected to save over 125 million liters of water per year compared with traditional ones, protecting local water sources and reducing dependence on municipal supplies (Solomon 2024). Microsoft has steadily improved the effectiveness of water usage in its data centers to 0.30 L/kWh, a 39% improvement since 2021 and an 80% reduction compared with its first-generation data centers in the 2000s. The method maintains reliable temperatures for dense AI loads while keeping power use minimal, enabling expansion into arid areas, de-risking permitting, and strengthening Microsoft's ability to scale AI sustainably.

AI Workload Efficiency: Model Prospective

Hardware improvements are increasingly complemented by algorithmic and architectural techniques that reduce energy per training run or inference pass. These include (EPRI 2024):

- **Model pruning:** A technique to remove unnecessary or redundant weights, neurons, or even entire attention heads in a trained neural network, leading to multi-fold reductions in computing and electricity per iteration.
- **Quantization:** A method to convert model weights to lower precision formats, such as 8-bit integers. It helps to accelerate inference and reduce memory and power consumption on embedded devices.
- **Knowledge distillation:** An approach to develop a smaller, more manageable model that reflects the functionalities of a larger one, reducing computational requirements.

These methods help cut compute time, GPU hours, and emissions without a significant loss in performance.

AI Workload Efficiency: Task Scheduling

Running some computing jobs at different times, or in different data center locations, can reduce the strain on the power grid. Deloitte's analysis shows that if AI data centers trim just 1% of their electricity use during the busiest hours, grid operators could add about 126,000 MW of new data center load without major network upgrades (Stansbury et al. 2025). According to Deloitte's survey, 68% of industry executives accept this small cut in exchange for faster grid connections (Stansbury et al. 2025). By integrating workload schedulers that respond to electricity-price or congestion signals in real time, operators can avoid costly capacity expansions, lower their carbon footprint, and even earn demand-response revenues, making task mobility a scalable option for improving sustainability.

7.2 Options for Decarbonization

Long-Term Renewable Power Agreements

Data center operators sign long-term power-purchase agreements (PPAs) with new wind and solar electricity generators, giving then a steady flow of green electricity at a fixed price for a specified period and helping new projects secure financing. Corporate PPAs support capacity of around 120,000 MW of operational renewables globally, more than 30% of which is accounted for by companies operating data centers. In fact, those PPAs cover over 20% of the estimated 415 TWh global data center electricity demand in 2024. An additional 60,000 MW of PPA-related capacity projects are being developed, much of which is already sold to the same sector (Spencer et al. 2025).

24/7 Carbon-Free Energy Matching

Traditional approaches often rely on annual clean energy credits to balance total yearly consumption. Changing to hourly carbon-free matching would need flexible solutions, like batteries that can store and shift solar or wind power, steady clean sources (e.g., nuclear or geothermal that run continuously), and, in some cases, fossil-fuel plants equipped with carbon capture and storage (EPRI 2024; Diamant 2022). A real-world pilot by Google's Belgian campus illustrates the mix, combining rooftop PV panels with a 2.75 MW/5.5 MWh lithium-ion battery, which provides a smooth onsite supply and supports grid frequency services (Spencer et al. 2025). Together, hourly tracking, batteries, and steady clean plants allow operators to match every kilowatt per hour consumed with a carbon-free supply source.

Small Modular Reactors (SMRs)

SMRs have a power capacity of up to 300 MW per unit, about one-third of a traditional nuclear power reactors' capacity (Liou 2023). Announced projects linked to data center supply would add up to 25,000 MW worldwide (Spencer et al. 2025), mostly in the U.S. Vendors such as NuScale have designs that come in 250-600 MW blocks, small enough to power a single hyperscale campus (EPRI 2024).

Fuel-Cell Backup Systems

These systems use hydrogen fuel cells, converting the chemical energy in hydrogen to electricity with few byproducts, such as water and heat (EERE 2014). For example, American Electric Power has agreed to purchase up to 1,000 MW of Bloom Energy's solid-oxide stacks for data center customers, which will replace diesel generators during power outages (Spencer et al. 2025).

Advanced Geothermal

By drilling deep into hot, dry rock, heat is transferred from underground by fluid running through sealed pipes in the drill well to a turbine on the surface. Because the well is a closed circuit, it can be sited almost anywhere, making it an attractive carbon-free power option for data center campuses (Spencer et al. 2025).

Example: Google's Geothermal Energy for Continuous Clean Energy

Google aims to power its data centers on 24/7 carbon-free energy by 2030 (Google 2025). As intermittent wind and solar alone could not deliver that round-the-clock coverage, Google partnered with Fervo Energy in 2021 on the world's first corporate agreement to develop an enhanced geothermal project. In November 2023, the plant began an enhanced geothermal pilot, with carbon-free electricity flowing to the local grid serving Google's Nevada data centers. Fervo's project test showed commercial viability, recording a 30-day flow that enabled 3.5 MW of electric production, achieved via horizontal well pairs, 191°C reservoir temperatures, and real-time fiber-optic monitoring. Building on this success, Google expanded the partnership in 2024 by contracting 115 MW through NV Energy's Clean Transition Tariff – a move which, after full commercial deployment, will enhance geothermal generation by nearly 25 times that of the pilot.

7.3 Selected Regional Policies

As AI data centers proliferate globally, governments and institutions are responding with regulations focused on the environment and sustainability to align AI data centers with climate goals using performance mandates (such as PUE targets), zoning and reporting requirements, and incentive schemes. Table 8 highlights key policies and frameworks related to AI data centers in the U.S., United Kingdom, Singapore, China, United Arab Emirates, and other countries.

United States

U.S. policy on data centers is evolving through a combination of federal initiatives, state measures, and strategic planning. In February 2024, U.S. lawmakers introduced the AI Environmental Impacts Act, which directs the Environmental Protection Agency and the National Institute of Standards and Technology to assess the environmental impact of AI in collaboration with academic, industry, and civil society stakeholders. The Act also includes a voluntary reporting framework for AI developers and operators as a first step toward common measurement standards (Crawford 2024). Although there is no binding federal efficiency mandate, the U.S. leads the Net-Zero Government Initiative to cut government-operational emissions to net-zero by 2050 (Office of the Federal Chief Sustainability Officer 2025).

Complementing this, America's AI Action Plan, published in July 2025, links AI competitiveness to infrastructure scale, energy, and regulation. It proposes faster permitting for data centers under the National Environmental Policy Act (NEPA),

Figure 20. Sustainable AI data center policy dimensions.

Energy efficiency standards

Control of energy performance through PUE thresholds and infrastructure benchmarks.

Renewable energy and carbon limits

Mandatory or voluntary transition to adopt low-carbon operations and meet emissions goals.

Infrastructure siting and grid resilience

Location-based incentives or constraints to reduce stress on grids and urban zones.

Energy reuse and heat recovery

Encouraging waste heat recovery and reuse or integration with district heating.

Transparency, disclosure, and reporting

Mandatory disclosure of energy usage, emissions, and sustainability performance.

Development controls and growth caps

Temporary bans, moratoriums, or permit limits on expansion in specific areas.

Water and environmental resource management

Response to drought and water scarcity, and site-specific sustainability.

more federal land for green data centers, modernizing the national grid for AI-scale loads, and prioritizing energy sources like nuclear and geothermal to power next-generation AI infrastructure (White House 2025).

Policies in individual states vary. Some offer tax incentives for data centers that meet high energy-efficiency or renewable-procurement standards, while others impose limits on backup-generator emissions or power usage in constrained grids. For example, since 2014 California has required data centers in the state that are larger than 1,000 sq ft to report PUE annually, and any with a PUE above 1.5 must reduce it by at least 10% per year until reaching 1.5 or lower (California Department of General Services 2014). Texas Senate Bill 6 sets strict rules for large electricity users, including data centers. It requires transparent load planning, on-site backup power, and cost-sharing for upgrades to the grid. It also authorizes the Electric Reliability Council of Texas to manage loads during grid emergencies and to use remote shutoff capabilities. The goal is to protect grid reliability for residential customers while still encouraging business growth (Paz 2025).

United Kingdom

The United Kingdom is steadily tightening sustainability requirements for data centers in line with its target of net-zero by 2050, set in the UK Climate Change Act. A key policy mechanism is the Climate Change Agreement (Environment Agency 2022) for the data center sector, a voluntary scheme that grants energy tax relief in exchange for meeting efficiency targets, with penalties for non-compliance. Large data centers also fall under the UK Emissions Trading Scheme (Department for Energy Security & Net Zero 2025), which sets prices on emissions and creates financial incentives to reduce carbon intensity.

The Streamlined Energy and Carbon Reporting framework (Department for Education and Education and Skills Funding Agency 2025) mandates operators to publicly disclose their energy usage and carbon emissions, encouraging the adoption of sustainable practices such as integrating on-site and off-site renewable energy, efficient cooling technologies, and low-carbon power procurement.

The European Union

The EU has advanced regulatory efforts to align data centers with its targets for climate-neutrality and high energy-efficiency by 2030. The 2019 EU regulation on eco-design standards for servers and data storage imposes standards for energy efficiency and sustainable materials, through improved design for durability, repairability, and recyclability (European Commission 2019).

The Corporate Sustainability Reporting Directive, published in 2022, requires large and publicly listed companies to report on greenhouse gas emissions, including those associated with IT infrastructure and third-party data providers (European Parliament and Council of the European Union 2022). The Energy Efficiency Directive, revised in 2023, requires all data centers over 500 kW to report metrics on operational sustainability, including energy use, PUE, water consumption, renewable share of energy, and waste-heat reuse (European Parliament and Council of the European Union 2023). It also requires operators to implement certified energy management systems such as ISO 50001.

A new Disclosures Delegated Act in 2024 defines a standard energy rating scheme and transparency rules for publication (European Commission 2024). Complementary EU efforts include a voluntary Code of Conduct for Data Centre Energy Efficiency (Acton, Booth, and Paci 2025) that promotes best practices across the industry and include sustainable finance rules (European Development Finance Institutions 2024) (EU Taxonomy) that direct investments toward eco-friendly projects and initiatives.

China

China has launched major initiatives to green its fast-growing digital infrastructure. The government's Eastern Data and Western Computing initiative, proposed in 2022, aims to relocate data centers to the western region, and to take advantage of natural cooling, clean energy, and cost-effective resources (Zhang et al. 2024). It includes building 10 high-density, energy-efficient, low-carbon data center clusters in eight hubs (Mengzhuo and Zhewen 2024). This is expected to reduce emissions from the data center sector by 16%-20% by 2030 (Zhang et al. 2024).

China also released the National Action Plan for New Computing Infrastructure in October 2023, prioritizing the construction of "green, low-carbon" computing systems (Chinese Ministry of Industry and Information Technology et al. 2023). The plan includes the adoption of advanced cooling technologies, such as

natural and liquid cooling, and improved computing efficiency, which, together, support China's two carbon goals: to peak carbon emissions by 2030 and to reach carbon neutrality by 2060 (Xu et al. 2023). In mid-2024, China set specific targets for data centers: reduce the average PUE to below 1.5 by 2025 and raise the renewable energy usage rate by 10% annually (Xinhua 2024). These targets are enforced through energy efficiency standards and upgrades to existing facilities.

Australia

Australia has taken significant steps toward advancing sustainable data center governance. The National Australian Built Environment Rating System (NABERS), first introduced in 1998, is the world's only mandatory energy performance labeling scheme for buildings, including data centers. Under NABERS, data centers are rated from 1 to 6 stars, with a PUE scale ranging from 2.42 (1- star) to 1.07 (6 stars) (Spencer et al. 2025). The Net-Zero Government Operations Strategy requires that by July 2025, all federal government workloads must operate in data centers rated at least 5-star, corresponding to a PUE of approximately 1.4. This is part of Australia's goal to achieve 100% renewable electricity across government operations by 2030 (Ghadially 2025).

In 2025, the Australian Energy Market Commission strengthened the country's grid resilience further by approving new rules to streamline renewable energy connections and set performance obligations for large energy users, particularly data centers and hydrogen projects. These rules would require data centers to remain stable during grid disturbances and support system security, reflecting their growing share of national electricity demand (Ghadially 2025).

Singapore

Due to land scarcity and energy constraints, Singapore has adopted a strategic, sustainability-first approach to data center growth. In 2019, it paused permits for new data center developments. Permits were restarted in 2022 through a pilot framework that favored proposals demonstrating exceptional energy efficiency and sustainability performance (Spencer et al. 2025). New data centers must meet a maximum PUE of 1.3 and obtain the BCA-IMDA Green Mark Platinum certification for data centers, which is the highest tier in Singapore's green building rating system (Yeo 2022). Singapore's national Green Data Centre Standard was updated in 2020 to align with ISO 50001 to standardize energy-management practices (Spencer et al. 2025).

In 2024, the Green Mark for Data Centres certification was enhanced to include broader sustainability metrics such as energy efficiency in IT operations, intelligent systems integration, carbon reduction strategies, and renewable energy readiness (Building and Construction Authority 2024). Additionally, and EMA launched the Green Data Centre Roadmap, which outlines Singapore's long-term ambition to lead in AI-ready, tropical climate-optimized, low-carbon data centers, including innovations like carbon-intelligent scheduling and modular retrofits (Hui Tian 2024).

United Arab Emirates

The UAE's Moro Hub is a regional benchmark for sustainable data center development. This 100 MW campus was launched in 2022 by Digital DEWA and is powered entirely by the Mohammed bin Rashid Al Maktoum Solar Park, earning a Guinness World Record as the world's largest solar-powered data center (Intel 2025). This was achieved with the help of underlying federal and local sustainability standards, such as the Dubai Green Building Regulations (Goverment of Dubai 2023), issued in 2010 and replaced in 2020 by the Al Sa'fat Green Building Rating System, which set minimum standards for green design and energy efficiency for new buildings, including digital infrastructure.

Abu Dhabi developed the Estidama Pearl Rating System in 2010, a framework to promote sustainable urban development and building practices, focusing on environmental, economic, cultural, and social aspects (Abu Dhabi Urban Planning Council 2016). It consists of five certification levels and eight criteria categories, including energy efficiency and renewable energy sources.

Together, these policies support the UAE's Net-Zero 2050 pathway for data centers.

Other Countries

Several countries are adopting targeted regulations to manage data centers' energy, environmental, and spatial impacts. Germany's Energy Efficiency Act requires all new data centers to have a PUE of 1.2 by 2026, while existing facilities must meet a PUE of 1.5 by 2027 and improve it further to 1.3 by 2030, alongside a complete transition to renewable energy by 2027. Germany also introduced the Energy Reuse Factor, requiring operators to reuse at least 10% of energy (waste heat) by 2026, rising to 15% by 2028 (Spencer et al. 2025). Table 8 summarizes different countries' PUE targets.

South Korea takes a different approach by using location incentives to steer data center development away from areas with grid congestion. The government offers a 50% reduction in electricity costs for data centers built outside the heavily populated Seoul metropolitan area, aiming to ease grid pressure and support more balanced regional developments (Spencer et al. 2025).

South Africa's National Data and Cloud Policy designates zones for future data centers, ensuring that expansion aligns with grid capacity and national infrastructure planning goals.

Beyond location-based incentives, several countries are imposing moratoriums and stricter permitting policies to manage the growth of data centers. In Ireland, as data centers reached 21% of national electricity consumption by 2020, a freeze was enacted in the Dublin area until 2028 (Duncan et al. 2024). The Netherlands enacted moratoriums in 2019 to reevaluate its data center strategies due to rapid growth in digital infrastructure. These restrictions were later lifted under revised conditions that imposed tighter standards for energy efficiency and land use.

Environmental factors such as water use also shape policy. In Chile, authorities partially reversed Google's permit to build a data center due to local concerns over water scarcity– a key issue in a region already stressed by drought. A similar debate is happening in Uruguay, where community and environmental groups have raised alarms over the water demands of planned hyperscale infrastructure.

Table 8. PUE mandates for selected countries.

Region	PUE (2023)	PUE mandate
U.S.	~1.4 (avg)	-
Australia	1.44	1.4 by 2025
China	1.56	1.5 by 2025
France	1.36	40% building energy cut by 2030
Germany	1.42	1.2 (new, 2026), 1.3 (all, 2030)
Japan	1.53	<= 1.4 since 2022
California (U.S.)	1.21	<= 1.5 since 2014

Sources: IEA (2025), Spencer et al. (2025), and Masanet et al. (2024).

Conclusion and Recommendations

08

The rapid rise of AI is reshaping the global digital landscape, with data centers at the heart of this transformation. AI-driven workloads demand powerful and robust computing systems, reliable energy, and sustainable design, creating both opportunities and challenges for countries seeking to build leadership in this new era.

This trend presents a unique chance for Saudi Arabia to position itself as a global hub for AI-ready infrastructure. The Kingdom benefits from a clear national vision, low-cost energy, and strong government support. Large-scale projects already underway, such as HUMAIN's AI factories and NEOM's Oxagon campus, show how quickly the sector is advancing. If current plans are realized, Saudi Arabia could capture a significant share of global AI computing capacity by the early 2030s, directly supporting the ambitions to drive digital transformation and diversify the economy.

However, this study also shows that the path ahead is not without risks. AI data centers place heavy demands on electricity systems and water resources. Under a high-growth scenario, data center electricity demand could increase more than tenfold by 2030, representing up to 11% of national electricity demand. Without sustainable practices, this expansion could further increase emissions and strain natural resources. By adopting advanced cooling systems and integrating renewable energy, electricity use could be cut by almost 13% and carbon emissions reduced by 68%, thus achieving growth without sacrificing environmental goals.

The economic outlook is also shaped by cost competitiveness. Thanks to affordable electricity and a favorable location, Saudi Arabia holds an advantage in the cost analysis. Yet, to sustain this advantage, the Kingdom must continue to invest in energy efficiency and workforce development, and must keep its competitive tariffs. Global risks, such as shortages of advanced chips or geopolitical tensions, also underline the need for careful planning and diversified partnerships.

Overall, we find that Saudi Arabia stands at a crossroads. By combining its energy strengths with sustainable practices and continued investment in innovation, the Kingdom can build a resilient AI data center ecosystem that both supports its economic ambitions and contributes to the global digital progress. Success will depend on balancing three priorities: ensuring a reliable and clean energy supply, building world-class digital infrastructure, and embedding sustainability at every stage of development.

To ensure global competitiveness and sustainable expansion of AI data centers in Saudi Arabia, as well as a balance between the three priorities above, several actions need to be considered across four dimensions: investment, innovation and development, operations, and governance and policy.

Investment

- **Invest in efficient technologies:** Support hardware such as next-generation GPUs and efficient servers that deliver more computing power per unit of energy.
- **Create AI-ready investment zones:** Prepare dedicated sites complete with reliable grid connections as a package with fiscal incentives for investors who commit to sustainability and renewable sourcing of energy.

Innovation and Development

- **Support local R&D:** Establish specialized national research centers dedicated to AI data centers, enabling local development of advanced cooling, efficiency solutions, and sustainable infrastructure tailored to Saudi Arabia's needs.
- **Drive innovation partnerships:** Establish partnerships between universities, research centers, and international firms to create new solutions for efficient AI systems and infrastructure.

Operations

- **Prioritize early utilization:** Maximize utilization in new data centers in the early stages, as their economic value is highest in the first years of deployment and operation.
- **Maximize efficient use:** Require operators to maintain at least high utilization by implementing workload scheduling strategies on flexible tasks such as AI training during off-peak hours to minimize idle infrastructure.
- **Adopt sustainable practices:** Encourage operators to recycle heat, reuse water, and adopt circular resource strategies where possible.

Governance and Policy

- **Preserve competitiveness with standards:** Maintain low electricity tariffs, but link incentives and permits to efficiency benchmarks and commitments to renewables.
- **Set efficiency targets:** Adopt realistic benchmarks (e.g., PUE ≤1.5 by 2030) tailored to Saudi Arabia's needs and conditions.
- **Align with the energy transition:** Coordinate data center expansion with national grid investments in solar, wind, storage, and transmission to secure a clean, reliable supply of energy.

References

Abu Dhabi Urban Planning Council. 2016. *The Pearl Rating System for Estidama: Public Realm Rating System: Design & Construction, Version 1.0*. December. Abu Dhabi Urban Planning Council. https://www.dmt.gov.ae/-/media/Project/DMT/DMT/E-Library/0001-Manuals/PRRS/PRRS-Version-10.pdf.

Acton, Mark, John Booth, and Daniele Paci. 2025. "2025 Best Practice Guidelines for the EU Code of Conduct on Data Centre Energy Efficiency." JRC Publications Repository. https://doi.org/10.2760/9449356.

Amazon. 2025. "AWS and HUMAIN Announce a More than $5B Investment to Accelerate AI Adoption in Saudi Arabia and Globally." May 13. https://www.aboutamazon.com/news/company-news/amazon-aws-humain-ai-investment-in-saudi-arabia.

Aterio. 2025. "Power Capacity Estimation ML Model." July 2. https://knowledge.aterio.io/data-products/us-data-centers/our-modeling-approach/power-capacity-estimation-ml-model.

Bellini, Emiliano. 2021. "Saudi Arabia's Second PV Tender Draws World Record Low Bid of $0.0104/kWh." *Pv Magazine International*. https://www.pv-magazine.com/2021/04/08/saudi-arabias-second-pv-tender-draws-world-record-low-bid-of-0104-kwh/.

Bizo, Daniel. 2023. "Global PUEs – Are They Going Anywhere?" *Uptime Institute Journal*. https://journal.uptimeinstitute.com/global-pues-are-they-going-anywhere/.

Building and Construction Authority (BCA) and Infocomm Media Development Authority (IMDA). 2024. *GMDC-2024: BCA-IMDA Green Mark for Data Centres (Beta Version)*. https://www1.bca.gov.sg/docs/default-source/docs-corp-buildsg/sustainability/20241008_gmdc2024_ver1.pdf?sfvrsn=407a219c_0.

Bureau of Experts at the Council of Ministers. 2021. "Personal Data Protection Law." Laws.Boe. https://laws.boe.gov.sa/boelaws/laws/lawdetails/b7cfae89-828e-4994-b167-adaa00e37188/1.

Bureau of Experts at the Council of Ministers. 2022. "Telecommunications and IT Act." Laws.Boe. https://laws.boe.gov.sa/BoeLaws/Laws/LawDetails/ae610645-e094-48ef-814e-aeb4009d244f/1.

California Department of General Services. 2014. "Requirements of Data Centers and Server Rooms - 1820.3." https://www.dgs.ca.gov/Resources/SAM/TOC/1800/1820-3.

CBRE Group Inc. 2025. 2025 Global Data Center Investor Intentions Survey. https://mktgdocs.cbre.com/2299/b889db99-6208-4f7f-97a8-55ad8ac1081f-1999143505/Global_Data_Center_Investor_In.pdf.

Chinese Ministry of Industry and Information Technology, Office of the Central Cyberspace Affairs Commission, Ministry of Education, People's Bank of China, and The State-Owned Assets Supervision and Administration Commission of the State Council. 2023. *Action Plan for the High-Quality Development of Computing Power Infrastructure*. CSET. https://cset.georgetown.edu/wp-content/uploads/t0573_compute_plan_EN.pdf.

Citi Group. 2024. "Data Center Powerplay: The Chips Have to Go Somewhere." May 29. https://www.citigroup.com/global/insights/data-center-powerplay-the-chips-have-to-go-somewhere.

Clemmons, Elisabeth, and Sean Graham. 2025. *The Macroeconomic Impacts and Implications of Datacenters: A Comprehensive Assessment*. IDC. https://my.idc.com/getdoc.jsp?containerId=US53243025.

Climatiq. 2021. "Emission Factor: Electricity Supplied from Grid – Saudi Arabia." https://www.climatiq.io/data/emission-factor/d642f6c8-32d7-4749-ac48-6463f30e0ae7.

Communications, Space and Technology Commission (CST). 2023. "Provision of Data Centers Services Regulation." https://www.cst.gov.sa/en/regulations-and-licenses/regulations/Document-1546.

Communications, Space and Technology Commission (CST). 2024a. *Annual Report 2024*. https://www.cst.gov.sa/en/knowledge-center/reports.

Communications, Space and Technology Commission (CST). 2024b. "CST Announces That the 'Data Center Services Regulations' Document Has Entered Into Force." https://www.cst.gov.sa/en/media-center/news/CST-Announces-that-the-Data-Center-Services-Regulations-Document-has-Entered-Into-Force.

Crawford, Kate. 2024. "Generative AI Is Guzzling Water and Energy." Nature. https://www.nature.com/articles/d41586-024-00478-x.pdf.

Data Center Map. n.d. "Middle East Data Centers – 298 Facilities from Operators." https://www.datacentermap.com/middle-east/.

Department for Education and Education and Skills Funding Agency. 2025. "Streamlined Energy and Carbon Reporting (SECR) for Academy Trusts." GOV.UK. https://www.gov.uk/government/publications/streamlined-energy-and-carbon-reporting-secr-for-academy-trusts.

Department for Energy Security & Net Zero. 2025. "Participating in the UK ETS." GOV.UK. https://www.gov.uk/government/publications/participating-in-the-uk-ets/participating-in-the-uk-ets.

Desroches, Clément, Martin Chauvin, Louis Ladan, Caroline Vateau, Simon Gosset, and Philippe Cordier. 2025. "Exploring the Sustainable Scaling of AI Dilemma: A Projective Study of Corporations' AI Environmental Impacts." arXiv:2501.14334. Preprint, arXiv, January 27. https://arxiv.org/abs/2501.14334.

Diamant, Adam. 2022. *24/7 Carbon Free Energy: Matching Carbon Free Energy Procurement to Hourly Electric Load.* Electric Power Research Institute (EPRI). https://restservice.epri.com/publicdownload/000000003002025290/0/Product.

Donnellan, Douglas. 2023. *Uptime Institute Data Center and IT Spending Survey 2022.* Uptime Institute. https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/DCSpendingSurvey.Report.01032023.pdf.

Duncan, Glen, Muhd Syafiq, Daniel Thorpe, and Kari Beets. 2024. *Data Centers 2024 Global Outlook.* JLL. https://leadsdell.com/wp-content/uploads/2024/09/jll-data-center-outlook-global-2024.pdf.

Egbert, Michael. 2025. "Oracle's Commitment to Saudi Arabia and President Trump's Vision for Global Prosperity." *Oracle News.* https://www.oracle.com/news/announcement/oracles-commitment-to-saudi-arabia-and-president-trumps-vision-for-global-prosperity-2025-05-13/.

Electric Power Research Institute (EPRI). 2024. *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.* https://www.epri.com/research/products/3002028905.

England, Andrew, and Ahmed Al Omran. 2025. "Saudi Arabia Seeks to Use Financial Might to Muscle into Global AI Industry." *Financial Times*, May 28. https://www.ft.com/content/176c7859-fdda-40d2-92a5-15d570f7accf.

Environment Agency. 2022. "Climate Change Agreements." GOV.UK. https://www.gov.uk/guidance/climate-change-agreements--2.

European Commission. 2019. *Commission Regulation (EU) 2019/2020. Official Journal of the European Union.* https://eur-lex.europa.eu/eli/reg/2019/2020/oj/eng.

European Commission. 2024. "Commission Notice on the Interpretation and Implementation of Certain Legal Provisions of the Disclosures Delegated Act." *Official Journal of the European Union.* https://eur-lex.europa.eu/eli/C/2024/6691/oj/eng.

European Commission. 2025. "Commission Sets Course for Europe's AI Leadership with an Ambitious AI Continent Action Plan." https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1013.

European Development Finance Institutions. 2024. *The EU Sustainable Finance Rules and Their Implications for Impact Investors.* https://edfi.eu/wp-content/uploads/2024/10/EDFI-Sustainable-Finance-Mapping-Report-FINAL.pdf.

European Parliament and Council of the European Union. 2022. *Directive (EU) 2022/2464 Of The European Parliament and of the Council.* European Union. https://eur-lex.europa.eu/eli/dir/2022/2464/oj/eng.

European Parliament and Council of the European Union. 2023. *Directive (EU) 2023/1791 on Energy Efficiency and Amending Regulation (EU) 2023/955 (Recast) (Text with EEA Relevance).* European Union. http://data.europa.eu/eli/dir/2023/1791/oj/eng.

Fioretti, Lapo, Carla La Croce, Andrea Siviero, and Elisabeth Clemmons. 2024. *The Global Impact of Artificial Intelligence on the Economy and Jobs: AI Will Steer 3.5% of GDP in 2030.* IDC. https://shorturl.at/I2BXc.

Fleck, Anna. 2024. "Which Countries Have the Most Data Centers?" Statista Daily Data, September 20. https://www.statista.com/chart/24149/data-centers-per-country.

General Authority for Statistics (GASTAT). 2023. *Electrical Energy Statistics 2023: Total Energy Delivered to Grid Increases by Approximately 5% in 2023.* Riyadh, Saudi Arabia: General Authority for Statistics. https://stats.gov.sa/documents/20117/2435281/Electrical+Energy+Statistics+2023+-+EN.pdf.

Ghadially, Farokh. 2025. "New Data Centre Rules Could Set a Global Standard." *Sustainability Matters Magazine*. https://www.sustainabilitymatters.net.au/content/sustainability/article/new-data-centre-rules-could-set-a-global-standard-545515027.

Goldman Sachs. 2024. "AI is Poised to Drive 160% Increase in Data Center Power Demand." https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand.

Google. 2025. *Google Environmental Report 2025*. https://www.gstatic.com/gumdrop/sustainability/google-2025-environmental-report.pdf.

Google Cloud. 2025. "Google Cloud and PIF Advance AI Hub in Saudi Arabia." Google Cloud Press Corner. https://www.googlecloudpresscorner.com/2025-05-13-Google-Cloud-and-PIF-Advance-AI-Hub-in-Saudi-Arabia.

Goverment of Dubai. 2023. "Al Sa'fat – Dubai Green Building System." Dubai Municipality. https://www.dm.gov.ae/municipality-business/al-safat-dubai-green-building-system/.

Grabein, Aaron, and Liz Stine. 2025. "AMD and HUMAIN Form Strategic, $10B Collaboration to Advance Global AI." AMD. https://ir.amd.com/news-events/press-releases/detail/1250/amd-and-humain-form-strategic-10b-collaboration-to-advance-global-ai.

Graham, Sean, Peter Rutten, and Olga Yashkova. 2024. *AI Datacenter Capacity, Energy Consumption, and Carbon Emission Projections*. IDC: The Premier Global Market Intelligence Company.

Green, Alastair, Humayun Tai, Jesse Noffsinger, Pankaj Sachdeva, Arjita Bhan, and Raman Sharma. 2024. *How Data Centers and the Energy Sector Can Sate AI's Hunger for Power*. McKinsey. https://www.mckinsey.com/industries/private-capital/our-insights/how-data-centers-and-the-energy-sector-can-sate-ais-hunger-for-power.

Greene-Dewasmes, Ginelle, Michael Higgins, and Thapelo Tladi. 2025. *Artificial Intelligence's Energy Paradox: Balancing Challenges and Opportunities*. World Economic Forum (WEF) in collaboation with Accenture. https://reports.weforum.org/docs/WEF_Artificial_Intelligences_Energy_Paradox_2025.pdf.

Groq. 2025. "Saudi Arabia Announces $1.5 Billion Expansion to Fuel AI-Powered Economy with Groq." https://groq.com/blog/saudi-arabia-announces-1-5-billion-expansion-to-fuel-ai-powered-economy-with-groq.

Hoff, Brandon. 2024. "Semiconductor Spend for Datacenter Infrastructure and a Breakout for AI Networking." International Data Corporation (IDC). https://my.idc.com/getdoc.jsp?containerId=US52018224.

Houlihan Lokey. 2025. *Real Estate Highlight: Navigating Market Changes in the Data Center Sector*. https://cdn.hl.com/pdf/2025/real-estate-highlight-data-centers-march-2025.pdf.

Hui Tian, Yuen. 2024. *Charting Green Growth Pathways at Scale for Data Centres in Singapore*. Infocomm Media Development Authority (IMDA). https://isomer-user-content.by.gov.sg/38/c49f426a-8053-4c7f-bbaa-8437b743a18f/files-Speeches%202024-1__Factsheet_IMDA_Green_DC_Roadmap.pdf.

HUMAIN. 2025. "Saudi Arabia's New AI Enterprise, HUMAIN, Joins Forces with AMD and Cisco to Launch Groundbreaking AI Infrastructure Collaboration." May 13. https://www.humain.ai/en/news/humain-and-amd-and-cisco/.

IBM. 2025. "What Is an AI Data Center?" https://www.ibm.com/think/topics/ai-data-center.

Intel. 2025. *Case Study: World's Largest Green Data Center*. https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2025-04/moro-hub-case-study.pdf.

International Data Corporation. 2024. "IDC's Worldwide AI and Generative AI Spending – Industry Outlook | IDC Blog." https://blogs.idc.com/2024/08/21/idcs-worldwide-ai-and-generative-ai-spending-industry-outlook/.

International Energy Agency (IEA). 2024. *Emissions Factors 2024*. https://www.iea.org/data-and-statistics/data-product/emissions-factors-2024.

International Energy Agency (IEA). 2025a. *Energy and AI*. IEA, Paris. https://www.iea.org/reports/energy-and-ai.

International Energy Agency (IEA). 2025b. *Building the Future Transmission Grid*. IEA, Paris. https://www.iea.org/reports/building-the-future-transmission-grid.

Istitlaa. 2025. "Global AI Hub Law." https://istitlaa.ncc.gov.sa/en/Transportation/citc/globalailaw/Pages/default.aspx.

Jamison, Stephanie, Sanjay Podder, Adam Burden, Bhaskar Ghosh, Senthil Ramani, Shalabh Kumar Singh, and Matthew Robinson. 2025. *Powering Sustainable AI*. Accenture. https://www.accenture.com/content/dam/accenture/final/corporate/corporate-initiatives/sustainability/document/Powering-Sustainable-AI.pdf.

Kamiya, George, and Vlad C Coroamă. 2025. *Data Centre Energy Use: Critical Review of Models and Results*. IEA-4E. https://www.iea-4e.org/wp-content/uploads/2025/05/Data-Centre-Energy-Use-Critical-Review-of-Models-and-Results.pdf.

King Abdullah Petroleum Studies and Research Center (KAPSARC). 2023. *Modeling and Projecting Regional Electricity Demand for Saudi Arabia*. June 5. https://www.kapsarc.org/our-offerings/publications/modeling-and-projecting-regional-electricity-demand-for-saudi-arabia/.

King Abdullah Petroleum Studies and Research Center (KAPSARC). 2025. "Saudi Arabia Renewables Tracker." https://apps.kapsarc.org/appboard/renewableprojects.

Kristiansen Nøland, Jonas, Martin Hjelmeland, and Magnus Korpås. 2024. "Will Energy-Hungry AI Create a Baseload Power Demand Boom?" *IEEE Access* 12: 157824–157836. https://doi.org/10.1109/ACCESS.2024.3440217.

Levine, Hannah. 2025. "There's a Labor Shortage at Data Centers." JLL. https://www.jll.com/en-us/insights/theres-a-labor-shortage-at-data-centers.html.

Liou, Joanne. 2023. "What Are Small Modular Reactors (SMRs)?" IAEA, September 13. https://www.iaea.org/newscenter/news/what-are-small-modular-reactors-smrs.

Lohn, Andrew, and Micah Musser. 2022. *AI and Compute*. Center for Security and Emerging Technology (CSET). https://cset.georgetown.edu/wp-content/uploads/AI-and-Compute-How-Much-Longer-Can-Computing-Power-Drive-Artificial-Intelligence-Progress_v2.pdf.

Malik, Saf. 2025. "What Is Driving Saudi Arabia's $21bn Data Centre Investment Surge?" Capacity Media, March 10. https://www.capacitymedia.com/article/2eir2o1uu11xqfdsdwrgg/news/what-is-driving-saudi-arabias-21bn-data-centre-investment-surge.

Masanet, Eric, Nuoa Lei, and Jonathan Koomey. 2024. "How Will the Electricity Use of AI Data Centers Evolve? To Answer This Question, Energy Analysts Need Better Data." Preprint, ResearchGate. http://dx.doi.org/10.13140/RG.2.2.11203.00801.

Mengzhuo, Liu, and Lin Zhewen. 2024. "China's East Data West Computing Initiative – Power Infrastructure as the Next Big Thing in the Global AI Race." Premia Partners. https://www.premia-partners.com/sc/insight/china-s-east-data-west-computing-initiative-power-infrastructure-as-the-next-big-thing-in-the-global-ai-race.

Meta. 2024. *For a Better Reality: 2024 Sustainability Report*. https://sustainability.atmeta.com/wp-content/uploads/2024/08/Meta-2024-Sustainability-Report.pdf.

Meta. 2025. *Meta's Odense Data Centre*. Meta. https://datacenters.atmeta.com/wp-content/uploads/2025/02/Metas-Odense-Data-Center.pdf.

Ministry of Communications and Information Technology (MCIT). 2020. *KSA Cloud First Policy*. https://www.mcit.gov.sa/sites/default/files/cloud_policy_en.pdf.

Ministry of Communications and Information Technology (MCIT). 2023. *Annual Report 2022*. https://www.mcit.gov.sa/sites/default/files/2023-07/MCIT_Annual%20Report_2022_En-Web_0.pdf.

Ministry of Communications and Information Technology (MCIT). 2024. *Annual Report 2024*. https://shorturl.at/73psX.

Mytton, David. 2021. "Data Centre Water Consumption." *NPJ Clean Water* 4 (1): 11. https://doi.org/10.1038/s41545-021-00101-w.

Narasimhan, Sridhar, Archit Johar, Jay Motani, Santiango Lavin, Abhilash Jain, and Shiven Mahajan. 2025. "AI Data Center Location Attractiveness Index." Kearney. https://www.kearney.com/industry/technology/article/ai-data-center-location-attractiveness-index.

National Cybersecurity Authority (NCA). 2018. *National Cybersecurity Strategy*. https://nca.gov.sa/national_cybersecurity_strategy-en.pdf.

NEOM. 2025. "DataVolt Signs Agreement with NEOM to Design and Develop the Region's First Truly Sustainable, Net-Zero AI Factory Campus in Oxagon ." https://www.neom.com/en-us/newsroom/datavolt-signs-agreement-with-neom.

Noffsinger, Jesse, Mark Patel, Pankaj Sachdeva, and Arjita Bhan. 2025. *The Cost of Compute: A $7 Trillion Race to Scale Data Centers*. McKinsey. https://shorturl.at/H07zW.

NVIDIA Corporation. 2025a. "HUMAIN and NVIDIA Announce Strategic Partnership to Build AI Factories of the Future in Saudi Arabia." NVIDIA Newsroom, May 13. https://nvidianews.nvidia.com/news/humain-and-nvidia-announce-strategic-partnership-to-build-ai-factories-of-the-future-in-saudi-arabia.

NVIDIA Corporation. 2025b. *Blackwell Ultra Datasheet*. NVIDIA. https://resources.nvidia.com/en-us-gpu-resources/blackwell-ultra-datasheet.

Office of the Federal Chief Sustainability Officer. 2025. "Global Initiatives." https://www.sustainability.gov/archive/biden46/federalsustainabilityplan/global-initiatives.html.

Paz, Cassia. 2025. "Texas Senate Bill 6: Data Centers Generator Fuel Plan Becomes Even More Critical." Mansfield. https://mansfield.energy/2025/07/10/texas-senate-bill-6-data-centers-generator-fuel-plan-becomes-even-more-critical.

Public Investment Fund (PIF). 2024. "PIF and Google Cloud to Create Advanced AI Hub in Saudi Arabia." https://www.pif.gov.sa/en/news-and-insights/press-releases/2024/pif-and-google-cloud-to-create-advanced-ai-hub-in-saudi-arabia/.

Qualcomm. 2025. "Qualcomm and HUMAIN to Develop State-of-the-Art AI Data Centers to Deliver Cloud-to-Edge Hybrid AI Services." May 13. https://www.qualcomm.com/news/releases/2025/05/qualcomm-and-humain-to-develop-state-of-the-art-ai-data-centers-.

Rahman, Robi. 2024. "Leading ML Hardware Becomes 40% More Energy-Efficient Each Year." Epoch AI. https://epoch.ai/data-insights/ml-hardware-energy-efficiency.

Saudi & Middle East Green Initiatives. 2025. "SGI Target: Reduce Carbon Emissions by 278 mtpa by 2030." https://www.sgi.gov.sa/about-sgi/sgi-targets/reduce-carbon-emissions.

Saudi Data and AI Authority (SDAIA). 2025a. "National Strategy for Data & AI." https://sdaia.gov.sa/en/SDAIA/SdaiaStrategies/Pages/NationalStrategyForDataAndAI.aspx.

Saudi Data and AI Authority (SDAIA). 2025b. "Saudi Data & AI Authority and Vision 2030." https://sdaia.gov.sa/en/SDAIA/SdaiaStrategies/Pages/sdaiaAnd2030Vision.aspx.

Saudi Electricity Company (SEC). 2025. *Earnings Release – Q1 2025*. May 14. https://www.se.com.sa/-/media/sec/Investors/Earning-Reports/earnings-release-q1-2025.ashx.

Saudi Electricity Regulatory Authority (SERA). 2025. "Consumption Tariff." https://sera.gov.sa/en/consumer/electric-tariff/electric-tariff-categories/consumption-tariff.

Saudi Market Research Consulting Firm. 2025. "Saudi AI Infrastructure Scaling at 30.5% CAGR by 2030." Saudi Market Research | Consulting Firm, July 10. https://saudimarketresearchconsulting.com/insights/articles/saudi-ai-infrastructure-scaling-at-30-5-cagr-by-2030.

Saudi Press Agency (SPA). 2024. "Minister of Energy Inaugurates Saudi Smart Grid Conference." https://www.spa.gov.sa/en/N2228611.

Saudi Telecom Company (STC). 2024. *Sustainability Report 2024*. Riyadh, Saudi Arabia: stc. https://www.stc.com/content/dam/groupsites/stc-annual-report-2024/assets/img/pdfs/sustainability-unsdgs.pdf.

S&P Global Inc. 2025. https://www.marketplace.spglobal.com/en/datasets.

Sharma, Lakshmee. 2024. "AI Data Centers Threaten Global Water Security." Lawfare, December 19. https://www.lawfaremedia.org/article/ai-data-centers-threaten-global-water-security.

Shehabi, Arman, Dale Sartor, Alex Hubbard, Alex Newkirk, Nuoa Lei, Md Abu Bakar Siddik, Billie Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024. *2024 United States Data Center Energy Usage Report*. Lawrence Berkeley National Laboratory. https://doi.org/10.71468/P1WC7Q.

Shiwani, Amit, Hani Abbasi, and Alistair Levack. 2025. "Unlocking the Data Centre Opportunity in the Middle East." PwC. https://www.pwc.com/m1/en/media-centre/articles/unlocking-the-data-centre-opportunity-in-the-middle-east.html.

Singla, Alex, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. 2025. *The State of AI: Global Survey*. McKinsey & Company. https://www.mckinsey.com/~/media/mckinsey/business functions/quantumblack/our insights/the state of ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf.

Solomon, Steve. 2024. "Sustainable By Design: Next-Generation Datacenters Consume Zero Water For Cooling." The Microsoft Cloud Blog, December 9. Microsoft. https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/12/09/sustainable-by-design-next-generation-datacenters-consume-zero-water-for-cooling/.

Spencer, Thomas, and Siddharth Singh. 2024. "What the Data Centre and AI Boom Could Mean for the Energy Sector." https://www.iea.org/commentaries/what-the-data-centre-and-ai-boom-could-mean-for-the-energy-sector.

Spencer, Thomas, Siddharth Singh, Laura Cozzi, Davide D'Ambrosio, Hugh Hopewell, Vincent Jacamon, Alex Martinos, Nicholas Salmon, and Brent Wanner. 2025. *Energy and AI*. International Energy Agency (IEA). https://iea.blob.core.windows.net/assets/601eaec9-ba91-4623-819b-4ded331ec9e8/EnergyandAI.pdf.

Spindler, Wesley, Luna Atamian Hahn-Petersen, and Sadaf Hosseini. 2024. "Why Circular Water Solutions Are Key to Sustainable Data Centres." World Economic Forum (WEF). https://www.weforum.org/stories/2024/11/circular-water-solutions-sustainable-data-centres/.

Springer, Cecilia, and Ali Hasanbeigi. 2025. *Data Centers in the AI Era: Energy and Emissions Impacts in the U.S. and Key States.* San Francisco, CA: Global Efficiency Intelligence, LLC. https://static1.squarespace.com/static/5877e86f9de4bb8bce72105c/t/67ab1779a68f9a35968e51f9/1739265952473/GEI+data+centers+report-+1.15.2025+clean-E.4.pdf.

Srivathsan, Bhargs, Marc Sorel, Pankaj Sachdeva, Arjita Bhan, Haripreet Batra, Raman Sharma, Rishi Gupta, and Surbhi Choudhary. 2024. *AI Power: Expanding Data Center Capacity to Meet Growing Demand*. McKinsey & Company. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand.

Stansbury, Martin, Kelly Marchese, Kate Hardin, and Carolyn Amon. 2025. "Can US Infrastructure Keep up with the AI Economy?" Deloitte. https://www.deloitte.com/us/en/insights/industry/power-and-utilities/data-center-infrastructure-artificial-intelligence.html.

Stokols, Andrew. 2025. "Energy and AI Coordination in the 'Eastern Data Western Computing' Plan." The Jamestown Foundation, March 8. https://jamestown.substack.com/p/energy-and-ai-coordination-in-the.

Synergy Research Group. 2025. "The World's Total Data Center Capacity Is Shifting Rapidly to Hyperscale Operators." https://www.srgresearch.com/articles/the-worlds-total-data-center-capacity-is-shifting-rapidly-to-hyperscale-operators.

Techusiness. 2025. *AMD Advancing AI 2025 Keynote By CEO of Humain Tareq Amin*. YouTube. https://www.youtube.com/watch?v=8UG92qu0Krg.

Tohme, Hani, Adel Belcaid, Lukas de Sonnaville, Arianna Molino, Rita Carvalho, and Hale De Vera. 2025. "A More Regenerative Digital Age: The Middle East Data Center Opportunity." Kearney. https://www.kearney.com/service/sustainability/article/a-greener-digital-age-the-middle-east-data-center-opportunity.

Tonomus. 2025. *Compute.* https://tonomus.neom.com/en-us/what-we-do/compute.html.

U.S. Department of Commerce. 2025. "UAE and US Presidents Attend the Unveiling of Phase 1 of New 5GW AI Campus in Abu Dhabi." https://www.commerce.gov/news/press-releases/2025/05/uae-and-us-presidents-attend-unveiling-phase-1-new-5gw-ai-campus-abu.

U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE). 2014. *Early Markets: Fuel Cells for Backup Power*. https://www.energy.gov/eere/fuelcells/articles/early-markets-fuel-cells-backup-power.

U.S. Environmental Protection Agency (EPA). 2024. *Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2022*. EPA 430-R-24-004. U.S. Environmental Protection Agency. https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks-1990-2022.

Van Zandt, Brooke. 2023. "A Decade of Greener Computing Blooms Inside NREL's Data Center." National Renewable Energy Laboratory (NREL) News & Feature Stories, August 2. https://www.nrel.gov/news/detail/program/2023/a-decade-of-greener-computing-blooms-inside-nrels-data-center.

Wang, James. 2024. "Cerebras CS-3 vs. Nvidia B200: 2024 AI Accelerators Compared." Cerebras, April 12. https://www.cerebras.ai/blog/cerebras-cs-3-vs-nvidia-b200-2024-ai-accelerators-compared.

The White House. 2025. *Americas AI Action Plan*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

Xinhua. 2024. "China Sets Green Targets for Data Centers." The State Council of the People's Republic of China. https://english.www.gov.cn/news/202407/24/content_WS66a0b167c6d0868f4e8e96ba.html.

Xu, Jing, Liping Li, Bin Zhang, Mingsong Hao, and Fengqiao Mei. 2023. "'Dual Carbon Goals' Enhances Policy Integration: Analysing Recent Changes in China's National Climate Policy." *Journal of Asian Public Policy* 18(3): 1–18. https://doi.org/10.1080/17516234.2024.2305320.

Yeo, Amelia. 2022. "Singapore New Data Centers." The International Trade Administration, August 11. https://www.trade.gov/market-intelligence/singapore-new-data-centers.

Zhang, Ning, Huabo Duan, Yuru Guan, Ruichang Mao, Guanghan Song, Jiakuan Yang, and Yuli Shan. 2024. "The 'Eastern Data and Western Computing' Initiative in China Contributes to Its Net-Zero Target." *Engineering* 52: 256–261. https://doi.org/10.1016/j.eng.2024.08.010.

Acknowledgments

The authors would like to thank Reema Alnuwisr, whose contributions to the cost analysis, coding, and data analytics were helpful to this study. The authors also wish to express their appreciation to the copyediting and design team at KAPSARC, especially Bree De Roche, Syed Yunus, and Christopher Bartle, for their editorial support and assistance in preparing the final publication.

Appendices

Appendix A. Scenario-Based Framework for Estimating Data Center Capacity (2025–2030)

We developed three growth scenarios to account for uncertainty in the trajectory of Saudi Arabia's data center expansion. These estimates are mainly derived from the S&P Global datasets and supplemented by information from announced projects.

Growth Scenarios

- **Baseline (non-AI growth):** Based on planned non-AI facilities, capacity expands from 290.5 MW in 2024 to about 1,050 MW by 2030.
- **Moderate AI growth:** Includes confirmed AI projects such as HUMAIN-AI Factory 1, Dammam (250 MW), HUMAIN-AI Factory 2 (250 MW), Datavolt-Riyadh (150 MW), and Datavolt-NEOM (300 MW). Adding these to the baseline scenario yields around 2,000 MW by 2030.
- **High AI growth:** Assumes that all announced mega-projects proceed, including HUMAIN's 1,900 MW target by 2030 and NEOM's AI campus (1,150 MW by 2030 of its larger 1,500 MW). With the baseline of 1,050 MW included, total capacity reaches about 4,100 MW by 2030.

The operational conditions were based on the study "Data Centers in the AI Era,"[13] which has two operational conditions: a "business-as-usual" scenario where data centers operate with conventional efficiency, and an "advanced efficiency" scenario where new technologies and best practices are widely adopted. The interaction of the capacity growth scenarios and the operational conditions gives us the following intersection:

- **High growth, conventional:** Rapid expansion to 4,100 MW by 2030, prioritizing speed over sustainability. Limited adoption of efficiency measures. This scenario produces the highest energy use and CO_2 emissions, serving as a stress test for power and resource systems.
- **High growth, sustainable:** Also achieves 4,100 MW by 2030 but integrates energy efficiency with expansion and scales up renewable energy, reducing electricity consumption and CO_2 emissions significantly due to higher operational efficiency.
- **Moderate growth, conventional:** Capacity increases steadily to around 2,000 MW with a conventional operating model. Efficiency is limited, leading to some increases in energy use and emissions, and does not align with sustainability goals.
- **Moderate growth, sustainable:** Capacity increases to 2,000 MW but with strong efficiency gains and alignment with national green targets. PUE improves through adopting best practices, and renewable integration increases, lowering both electricity consumption and CO_2 emissions compared to the conventional scenarios.

[13] Springer and Hasanbeigi (2025).

Appendix B. Methodology for Estimating Data Center Energy Demand in Saudi Arabia (2025-2030)

Estimating Data Centers' Energy Demand

This study estimates the electricity demand of data centers in Saudi Arabia for 2025-2030 using the framework proposed by Masanet et al. (2024)[14] rather than detailed bottom-up server- or rack-level measurements. Our method estimates total energy use based on installed IT capacity, adjusted by usage rate, PUE, and annual operating hours. The general formula of the estimated data center electricity use in E_{total} GWh is expressed as

$$E_{total} = p \times u \times \alpha \times t/1000 \quad (1)$$

where p is the critical IT capacity in MW, u is the average of critical IT capacity used, α is the power usage effectiveness, and t is the annual operating hours (8,760).

Because operational patterns differ between AI and general-purpose facilities, we adopted a differentiated formulation of the electricity consumption equation, in which the PUE and usage rates are assigned according to workload type. The extended equation is expressed as:

$$E_{total} = [(p_{AI} \times u_{AI} \times \alpha_{AI}) + (p_{non-AI} \times u_{non-AI} \times \alpha_{non-AI})] \times t/1000 \quad (2)$$

Where

- p_{AI}, p_{non-AI} are the installed IT capacity in MW for AI and non-AI facilities, respectively.
- u_{AI}, u_{non-AI} are the utilization rates for AI and non-AI facilities, respectively.
- α_{AI}, α_{non-AI} are the assumed PUE values for AI and non-AI facilities, respectively.
- t is the annual operating hours (8,760).

Estimating the Total Electricity Demand by 2030

To estimate the share of total electricity demand by data centers in Saudi Arabia, we used existing projections and reported statistics of national electricity consumption. Projections from the King Abdullah Petroleum Studies and Research Center (KAPSARC) indicate that demand could reach 365,400 GWh by 2030.[15] under a scenario of moderate economic growth and stable prices. For the years 2022 and 2023, we used actual electricity consumption figures published by the General Authority for Statistics (GASTAT), which were 309,524 GWh and 327,001 GWh, respectively.[16] We derived values for the years 2024 to 2029 by a linear interpolation between the official statistics and the KAPSARC projection, giving us a consistent trajectory toward the 2030 estimate.

Table B1. Projected growth in data center electricity demand and its share of total national electricity consumption.

	Capacity (MW)	Energy consumption (TWh/year)	Share of national electricity demand
Current (2024)	290.5	2.80	0.85%
Scenario 0: Non-AI growth	1,050	10.16	2.79%
Scenario 1: Moderate growth, conventional	2,000	20.15	5.52%
Scenario 2: Moderate growth, sustainable	2,000	17.62	4.8%
Scenario 3: High growth, conventional	4,100	42.23	11.55%
Scenario 4: High growth, sustainable	4,100	36.76	10.1%

[14] Masanet, Lei, and Koomey (2024).

[15] KAPSARC (2023).

[16] GASTAT (2023).

Key Operational Assumptions: PUE and Utilization Rate

PUE and utilization rate are two critical variables in data center energy modeling. In Saudi Arabia, data centers record PUE values between 1.80 and 2.10[17] due to the hot climate, though recent data reports a national average closer to 1.53.[18] New centers, such as NEOM's Oxagon campus, aim for values below 1.3 in hyperscale facilities,[19] actively pursuing efficiency improvements.

To investigate the role of efficiency on electricity demand, we looked at two distinct operational conditions: conventional and sustainable. Under the conventional conditions in our study, current general-purpose data centers using air-cooled designs are the least energy efficient, with a PUE of around 1.70, consistent with global benchmarks for infrastructure in hot climates.[20] AI-oriented centers were given a moderately lower PUE of 1.50, reflecting the adoption of liquid cooling but with limited system-wide optimization.

In the sustainable condition scenario, we assumed that improvements in efficiency will increase over time, consistent with global trends in advanced cooling and energy management. Modern general-purpose centers are expected to achieve a PUE of around 1.5 while AI-optimized centers with the latest in cooling technology could reach values near 1.30. As shown in Table B2, these assumptions recognize that efficiency profiles differ between AI-intensive and general-purpose data centers, due to variations in their workload density and cooling requirements.

The utilization rates are another key factor in modeling electricity demand. AI-focused data centers, particularly those supporting large-scale model training and continuous inference, have a higher utilization rate due to long-duration, compute-intensive workloads with minimal idle periods. Recent studies report utilization rates of around 75%-85% for AI training workloads, with lower values for inference.[21] Enterprise and colocation facilities typically have more variable workloads, resulting in a lower average utilization, commonly in the range of 50%-70%.[22] For this study, AI facilities were given a utilization rate of 0.80, and non-AI facilities a rate of 0.65.

Table B2. Assumed PUE and utilization rate values for different operational conditions and data center types in Saudi Arabia by 2030.

Scenario	Data center type	PUE (2030)	Utilization rate (2030)
Conventional	General purpose	1.70	0.65
	AI	1.50	0.80
Sustainable	General purpose	1.50	0.65
	AI	1.30	0.80

[17] STC (2024).

[18] S&P Global dataset.

[19] Tonomus (2025).

[20] Bizo (2023).

[21] Shehabi et al. (2024).

[22] Citi Group (2024).

Appendix C. Methodology for Estimating CO_2 Emissions from Data Centers' Electricity Consumption

Estimating CO_2 Emissions

To quantify the carbon footprint of Saudi Arabia's data centers, we linked projected electricity consumption to the national energy mix using standardized emissions factor methods from the IEA CO_2 Emissions Factors 2024 dataset.[23] The estimation follows the fundamental equation

$$CO_2\ Emissions\ (Mt) = E_{total} \times F \times 10^{-3} \quad (3)$$

Where

- E_{total} is the total annual energy demand of data centers (GWh).
- F is the grid emission factor, which represents the average CO_2 emitted per MWh of electricity generated, reflecting the national energy mix (tCO_2/MWh).

The grid emission factor for Saudi Arabia varies based on the energy mix:

- Fossil fuel mix: (41.2% oil, 58.2% natural gas, <1% renewables) results in a grid emission factor[24] of 0.568 tCO_2/MWh[25]:

$$F = (0.412 \times 0.783\ tCO_2/MWh) + (0.582 \times 0.423\ tCO_2/MWh) = 0.568\ tCO_2/MWh$$

- Vision 2030 target mix: (50% natural gas, 50% renewables) reduces the grid emission factor to 0.21 tCO_2/MWh:

$$F = (0.5 \times 0.42\ tCO_2/MWh) + (0.5 \times 0\ tCO_2/MWh) = 0.21\ tCO_2/MWh$$

Estimating CUE

The carbon usage effectiveness (CUE) in tCO_2/MWh-IT is calculated as the product of PUE (α) and the grid emission factor (F):

$$CUE = \alpha \times F \quad (4)$$

Using the following grid factors and PUE ranges:

- Conventional (fossil fuel mix 0.568 tCO_2/MWh; PUE 1.5-1.7): CUE = 0.852-0.966 tCO_2/MWh-IT.
- Sustainable (Vision 2030 mix 0.21 tCO_2/MWh; PUE 1.3-1.5): CUE = 0.273-0.315 tCO_2/MWh-IT.

[23] IEA (2024).

[24] Greenhouse Gas Emissions Inventory (EPA 2024).

[25] Climatiq (2021).

Table C1. Summary of data center growth scenarios in Saudi Arabia from present to 2030.

		Scenario					
Variable		Current (2024)	Scenario 0 Non-AI growth	Scenario 1 High growth, conventional	Scenario 2 High growth, sustainable	Scenario 3 Moderate growth, conventional	Scenario 4 Moderate growth, sustainable
IT load capacity (MW)		290.5 MW	1,050 MW	Non-AI: 1,050 MW AI: 3,050 MW ＝4,100 MW		Non-AI: 1,050 MW AI: 950 MW ＝2,000 MW	
Workload type		General purpose and enterprise		Predominantly AI model training and inference		Mixed: general purpose and AI	
PUE		1.7	1.7	AI: 1.5 Non-AI: 1.7	AI: 1.3 Non-AI: 1.5	AI: 1.5 Non-AI: 1.7	AI: 1.3 Non-AI: 1.5
Renewable share (%)		1%	1%	1%	50%	1%	50%
Usage rate		0.65	0.65	AI: 0.80 Non-AI: 0.65	AI: 0.80 Non-AI: 0.65	AI: 0.80 Non-AI: 0.65	AI: 0.80 Non-AI: 0.65
Derived metrics	Total electricity demand (TWh)	2.80	10.16	42.23	36.76	20.15	17.62
	Share of national electricity demand (%)	0.85%	2.79%	11.55%	10.10%	5.52%	4.80%
	CO_2 emissions ($MtCO_2$/year)	1.60	5.81	24.02	7.7	11.48	3.7
	CUE (tCO_2/MWh-IT)	0.966	0.966	0.852-0.966	0.273-0.315	0.852-0.966	0.273-0.315

Note: Scenarios vary by capacity expansion and operational conditions. The derived metrics are the total electricity demand, share of national demand, and CO_2 emissions.

• Appendix D. Methodology for Estimating Lifetime Data Center Project Costs

The cost analysis needs to consider the PUE, load factor, computing efficiency, electricity price, CAPEX, and WACC. Hence, its formula needs to be a consolidated, time-discounted view of the total costs incurred per unit of computing output over the operational lifetime of a data center. The unit is ($/PFLOP) and its formula is

$$DC\ Project\ Cost = \frac{CAPEX + OPEX}{\begin{array}{c}Present\ Value\ of\\ Total\ Output\ over\\ Project\ Lifetime\end{array}} = \frac{c + \sum_{i=1}^{n-1} \frac{\alpha pkt + dc}{(1+r)^i}}{\gamma \sum_{i=1}^{n-1} \frac{kt}{(1+r)^i}} \quad (5)$$

At a high level, the cost structure mirrors the levelized cost of energy, with the output changed from electricity generation to computational throughput. The numerator represents the total discounted costs of the data center, and the denominator represents the total discounted compute output. We discuss each term individually below:

- Numerator ($/kW):
  - c: initial overnight cost of construction (CAPEX) per kW of IT load
  - $\sum_{i=1}^{n-1} \frac{\alpha pkt + dc}{(1+r)^i}$: discounted operating expenses (OPEX) for the lifetime of the data center
    - αpkt: the energy costs, where p is the electricity price in ($/MWh), scaled by the load factor k and the power use effectiveness α, and multiplied by the average time per year t
    - *dc*: Non-energy OPEX. Here, d is a percentage of the CAPEX c.
    - $\sum_{i=1}^{n-1} \frac{1}{(1+r)^i}$: This is to discount each year for the time horizon of interest (project lifetime). The discount factor r is the weighted average cost of capital (WACC).
- Denominator (PFLOPS/kW):
  - γ: The computing efficiency of the AI hardware in (PFLOPS/kW)
  - $\sum_{i=1}^{n-1} \frac{kt}{(1+r)^i}$: The effective compute output each year, discounted over time

Formula (5) is quite intuitive. The more efficient the AI hardware is, the lower the cost, reflecting incentives in using energy-efficient hardware. This is also consistent with the trend seen in Table 2, which shows significant improvements over the past few years. The electricity price and the PUE directly affect costs via OPEX as major contributors to the energy-related costs. One can easily see that the cost is linear in the price *p* and the PUE α, but nonlinear in the computing efficiency γ as it is scaled by ($1/\gamma$). It is also nonlinear in the load factor *k*, which directly affects OPEX and compute output.

Appendix E. Glossary of Definitions

Term	Definition
Artificial intelligence (AI)	A computer science field that focuses on building systems capable of performing tasks that usually require human intelligence, such as learning, reasoning, and self-development.
AI-specialized data center	A facility designed specifically for AI workloads, typically featuring high-performance GPUs, advanced CPUs, and specialized infrastructure to support dense, compute-heavy operations.
AI workload	A category of computational tasks that includes machine learning, deep learning, generative AI, and other resource-intensive AI applications.
Carbon usage effectiveness (CUE)	A standard metric measuring the carbon emissions intensity per MWh delivered to IT equipment (tCO_2/MWh-IT).
Central processing unit (CPU)	The main processing chip in a computer that executes general-purpose tasks and controls other components.
Compound annual growth rate (CAGR)	The average annual growth rate of a value over a specified period, assuming it grows at a steady rate each year.
Compute (in AI context)	Refers to the computational power, typically measured in FLOPS, required to train or run AI models. In the context of data centers, compute includes the combined processing capacity of specialized hardware such as GPUs, TPUs, and AI accelerators, which are essential for executing complex machine learning tasks at scale.
Data center	A physical facility that houses many servers and data storage devices with high-speed connectivity to manage an organization's applications and data.
Data center infrastructure efficiency (DCIE)	A measure of a data center's energy efficiency calculated as the percentage of energy used directly by IT equipment out of the total energy consumption. Higher DCIE values signify greater efficiency in non-computational functions.
FLOPs (floating point operations)	The total count of floating-point operations performed to complete a specific task or program.
GPU (graphics processing unit)	A specialized processor designed for parallel operations, originally for rendering graphics but now widely used in AI and high-performance computing due to its ability to handle complex, large-scale computations efficiently.
Inference	A process for making predictions by applying a trained model to unlabeled examples.
IT load capacity	The maximum power demand (measured in megawatts) of servers and network equipment installed in a data center, excluding cooling and ancillary systems.
The Jevons Paradox	An economic principle stating that improvements in resource efficiency can paradoxically lead to increased overall consumption of that resource, due to reduced costs and rising demand.
Language processing unit (LPU)	A specialized chip designed to optimize natural language processing tasks such as real-time translations, speech recognition, and text analysis. LPUs are more energy-efficient for specific AI applications compared to general-purpose processors.
Model training	The process of determining or optimizing the parameters of a model based on a machine learning algorithm using training data.
Power usage effectiveness (PUE)	A metric for data center energy efficiency, calculated as the ratio of total facility power to computing equipment power. A PUE closer to 1.0 indicates higher efficiency.
Rack power density	The amount of power consumed per server rack, typically measured in kilowatts. Higher densities are common in AI and high-performance computing environments.
Utilization rate	The proportion of a data center's installed IT capacity that is actively used. Higher utilization rates indicate more efficient use of infrastructure.
Workload	A combination of tasks that run on a given computer system.

About the Authors

Khaled Alshehri

Khaled Alshehri is a Research Fellow in the Utilities and Renewables Program at the King Abdullah Petroleum Studies and Research Center (KAPSARC). His research focuses on integrating digital technologies into the power sector to optimize grid operations, enhance economic efficiency, and support the energy transition. Drawing on extensive experience across government, academia, and industry, Dr. Alshehri leverages advanced tools from game theory, control, and optimization to address complex challenges in energy systems. In 2024, Dr. Alshehri was honored with the Questrom-CEMA Best Paper Award for his co-authored work on the efficient aggregation of distributed energy resources – a recognition of significant contributions to the field. He earned his B.S. in Control and Instrumentation Engineering from King Fahd University of Petroleum and Minerals (KFUPM) and his M.S. and Ph.D. in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign (UIUC).

Marwa Khan AlFattani

Marwa Khan AlFattani is a Researcher and Technology Strategist with more than 15 years of experience specializing in artificial intelligence, digital transformation, and national technology studies. In her Strategic leadership roles, she has led research initiatives that bridge AI research, policy development, and strategic implementation, advancing the Kingdom's digital transformation agenda. Previously, AlFattani spent nearly a decade at King Abdulaziz City for Science and Technology (KACST), contributing to Saudi Arabia's first AI governance report and conducting applied research across machine learning, natural language processing, simulation systems, and smart home technologies. She is an inventor with a registered patent and has authored peer-reviewed publications in the fields of AI, digital identity, and Arabic language technologies. AlFattani holds a Master of Science in Information Technology from King Saud University and continues to focus on responsible AI and public-sector innovation.

Laila Bashmal

Laila Bashmal is an Artificial Intelligence Researcher with a Ph.D. in Computer Engineering from King Saud University. Her work focuses on leveraging unified AI systems for high-impact applications, particularly in remote sensing analysis and medical image processing. Her broader research interests include large language models, multimodal learning, and the integration of language and vision to build more general and robust AI systems, as well as examining how these capabilities reshape computational infrastructure and real-world applications.

Ghaliah Alshammari

Ghaliah Alshammari is a Senior Researcher, specializing in data science, artificial intelligence, and analytical innovation. Her work focuses on developing data-driven insights that support national digital transformation. She previously served as an Information Technology Analyst with the G20 Saudi Secretariat. Ghaliah holds an M.S. in Computer Science from The City College of New York and a B.S. in Computer Science from Princess Nourah Bint Abdulrahman University.

About the Project

This discussion paper is part of the “Quantifying the Value of Generative AI (GenAI) for the Saudi Power Sector” project. Recognizing that digital infrastructure is a prerequisite for deploying GenAI at scale, the paper provides a first comprehensive assessment of the landscape, energy demand, emissions profile, and cost dynamics of AI-oriented data centers in Saudi Arabia. It outlines key scenarios, operational considerations, and strategic enablers that will shape the Kingdom’s capacity to support emerging GenAI applications.

كابسارك
KAPSARC

unesco
Centre
Under the auspices
of UNESCO

ICAIRE
International
Center for
AI Research & Ethics

kapsarc.org
info@kapsarc.org
/kapsarc