SUSTAINABLE DATA CENTERS ROADMAP

Executive Summary

October 2025

Innovation for Cool Earth Forum

EXECUTIVE SUMMARY

Chapter 1: Data Center Energy Use

The energy use of data centers can be roughly divided into two categories: information technology (IT) equipment and infrastructure equipment. IT equipment energy use refers to the electricity consumed by servers, storage arrays and/or network switches. Infrastructure energy use refers to energy consumed by non-IT equipment that provides space conditioning (such as cooling) and ensures reliable power for the data center. Ideally, the IT equipment uses most of the electricity consumed by a data center, as it performs the data center's main revenue-earning function. This concept is reflected in a key metric used to assess data center efficiency: power usage effectiveness (PUE)—the ratio of a data center's total energy usage to the energy used by its IT equipment.

Data center electricity demand has risen sharply in recent years due to increasing digitalization, greater connectivity and the artificial intelligence (AI) boom. According to recent estimates, global electricity demand from data centers doubled from around 200 TWh in 2017 to more than 400 TWh in 2024. Data centers' share of global electricity demand increased from 0.9% to 1.5% in the same period. Growth has been concentrated in the United States, China and Europe, which collectively accounted for 86% of the world's data center electricity demand in 2024.

Scenarios of future data center electricity demand vary widely based on assumptions related to market growth, technological progress, energy efficiency improvements and other factors. Recent International Energy Agency (IEA) scenarios suggest that data center electricity use could reach 670-1260 TWh globally (or 1.8-3.4% of total global electricity demand) by 2030. In some locations with high data center concentrations (e.g., Ireland and Northern Virginia), data centers' share of total power demand is much higher and is placing growing stress on regional electric grids.

Chapter 2: Data Center Energy Efficiency

Chapter 2.1: Information Technology (IT) Equipment

Computing efficiency has improved by a factor of 10 billion since 1946. Since around 2005, efficiency improvements have been doubling roughly every 2.3 years.

AI workloads are creating unprecedented energy demands that fundamentally challenge traditional data center designs. Graphics processing units (GPUs) and accelerated AI processors consume 400-1000+ watts (compared to traditional processors' 50-200 watts), while requiring sustained high-bandwidth memory and storage systems that cannot easily enter power-saving modes. Chip-level efficiency gains are increasingly offset by exponential demand growth from AI applications.

Multiple innovation pathways are emerging to address these challenges. Advanced chip packaging techniques are reducing power usage, silicon photonics are enabling power improvements in networking, and next-generation semiconductors are reaching near-perfect energy conversion efficiency. Companies are developing AI-optimized server designs, exploring power oversubscription strategies and investigating edge computing to reduce data center computational loads. Although efficiency improvement rates are slowing, these emerging technologies and operational strategies specifically tailored to AI workloads offer potential for renewed progress in reducing energy consumption.

Chapter 2.2: Software

Software efficiency depends on algorithmic design, where clever approaches can reduce computational requirements by orders of magnitude. Traditional algorithms have clear "correct" outputs that enable objective efficiency comparisons, but AI systems operate differently. Modern AI architectures are delivering substantial efficiency gains using specific techniques that reduce memory requirements and accelerate processing with minimal performance degradation.

Operational flexibility represents a powerful approach to reducing data center energy consumption and carbon emissions. Checkpoint and restart technology enables AI training workloads to pause, migrate between data centers and resume during periods of abundant renewable energy or in regions with cleaner electricity grids. However, AI inference workloads face stronger constraints due to latency requirements.

Emerging AI applications are creating new challenges. AI reasoning models allocate extra computational resources during inference but can outperform models hundreds of times larger, potentially reducing overall energy consumption despite higher per-query costs. Autonomous agents and multi-agent systems are driving dramatic increases in inference demand, consuming up to 100 times more compute than traditional chatbot interactions through continuous operation and complex reasoning chains. The rapid pace of AI development often prioritizes performance over energy optimization, while the lack of standardized benchmarks for measuring AI efficiency complicates efforts to balance computational capability with sustainable energy consumption.

Chapter 2.3: Cooling Technologies

Data centers' rising thermal loads are outpacing the capabilities of traditional air-based cooling. Air-based cooling technologies rely on circulating conditioned air to remove heat from server environments. Enhancements like hot aisle containment, evaporative cooling and air-side economization improve efficiency, but these systems struggle to manage rack power densities above ~30 kW. This has driven a transition toward more efficient and scalable cooling solutions, including direct-to-chip and immersion liquid cooling systems. These liquid-based systems offer significantly higher thermal efficiency by transferring heat through fluids with greater capacity than air. They can cool racks exceeding 100 kW and minimize water use.

Case studies from hyperscale operators including Meta, Microsoft and Amazon Web Services (AWS) illustrate diverse strategies—from precision evaporative systems and machine learning (ML)-optimized heating, ventilation and air conditioning (HVAC) to large-scale heat reuse via district heating networks. These operators are testing and deploying new approaches, including modular systems and AI-enabled controls that are enhancing resilience, adaptability and deployment speed.

Barriers to adopting advanced cooling technologies include lack of standardization, retrofit limitations, uncertain cost structures, regulatory hurdles, and undervalued water and carbon impacts. Overcoming these barriers will require a collaborative, multi-stakeholder approach.

Chapter 2.4: Heat Reuse

Most of the energy in a data center ultimately gets turned into heat. This heat can be used in beneficial ways if the data center's cooling system generates return heat with high enough temperatures. Modern liquid cooling systems offer 40-70 °C return heat, which is excellent for powering district heating systems in cold regions. Using data center waste heat for district heating systems requires collaboration between data center owners and operators and local stakeholders, such as municipal governments.

Other potential uses for data center waste heat include industrial drying, agriculture and direct air capture (DAC) of carbon dioxide (CO_2). These uses generally require heat of at least 70 °C and locations very close to data centers. They have not gone beyond pilot stages. Heat reuse by data centers has the greatest potential in cold regions.

Text Box: Data Center Energy Efficiency Metrics

By far the most widely-used metric for data center energy efficiency is “power usage effectiveness” or PUE. First introduced in 2007 by the Green Grid, PUE is the ratio of a data center’s total energy use to the energy use of its IT equipment. However, PUE does not capture the efficiency of IT equipment or the efficiency of IT workloads in a data center. Other data center energy efficiency metrics include “IT power usage effectiveness” (ITUE)—the ratio of total energy into IT equipment to the energy use of IT compute components alone; “total power usage effectiveness” (TUE)—ITUE multiplied by PUE; “data center energy productivity” (DCeP)—useful work produced divided by total energy consumed by the data center; “server energy productivity” (SEP)—the energy consumption of a server in relation to the share of compute work the server is performing; and “IT Work Capacity (ITWC)—work per unit energy.

Chapter 3: Data Center Greenhouse Gas Emissions

Chapter 3.1: On-Site Greenhouse Gas Emissions (Scope 1)

Some equipment at data centers directly emits greenhouse gases (Scope 1 emissions). Diesel generators for backup power emit CO_2. Cooling equipment and fire suppression systems can both leak hydrofluorocarbons (HFCs).

While data concerning these emissions are limited, data center operators can mitigate emissions with a variety of strategies. Natural-gas-fired generators for backup power can marginally reduce emissions if used to replace diesel generators. Drop-in diesel fuel replacements, such as hydrotreated vegetable oil (HVO), biodiesel (FAME) and synthetic paraffinic fuels, can provide deeper emissions reductions, but supplies are not universally available. Battery energy storage systems can substantially reduce emissions if charged using low-carbon power. Carbon capture is not a good candidate for mitigating emissions from on-site backup power systems because these systems have an intermittent nature and relatively small capacity.

Replacing current high global-warming-potential (GWP) refrigerants and fire suppression agents (such as R-134a and HFC-227ea, respectively) with lower-GWP alternatives can significantly reduce emissions from these refrigeration and fire suppression systems.

Chapter 3.2: Power Supply Greenhouse Gas Emissions (Scope 2)

Modern data centers consume large amounts of electricity, which can lead to high greenhouse gas emissions from power generation (Scope 2 emissions). These emissions are the largest contribution to overall lifecycle greenhouse gas emissions of the data center industry. Scope 2 emissions for data centers globally were approximately 180 megatons of CO_2 ($MtCO_2$) in 2024 (roughly 0.5% of global CO_2 emissions). This could double or triple by 2030.

Strategies for reducing Scope 2 emissions from data centers include using highly efficient cooling systems and other equipment (especially helpful on high-emissions grids) and locating on low-carbon grids with the capacity for additional wind, solar, hydroelectric, geothermal and/or nuclear power. Data centers that draw power from low-emissions grids can have Scope 2 emissions almost 100x smaller than those that draw power from high-emissions grids, so siting decisions are especially important. Implementing load flexibility by using on-site clean generation and storage and actively managing the timing of power consumption from the grid can help significantly to reduce Scope 2 emissions on grids with large daily variations in emissions intensity (such as those experiencing a solar “duck curve”).

Many low-carbon power generation technologies are being considered as part of grid capacity expansion to meet new data center load, including variable renewables (wind and solar) firmed by storage, enhanced geothermal, gas-fired generation with carbon capture and storage (CCS), solid oxide fuel cells, nuclear fission (both conventional and small modular reactors) and nuclear fusion. The key factors that determine the most appropriate technologies are dispatchability, technology readiness, location flexibility and costs. In addition to these strategies to address Scope 2 location-based emissions, many data center operators seek to further reduce their Scope 2 market-based emissions by procuring renewable energy, typically through power purchase agreements (PPAs) and the retirement of electricity energy attribute certificates (EACs). Changes to the Greenhouse Gas Protocol Scope 2 guidance (currently under development) will likely restrict, but not eliminate, the ability to pursue this strategy.

Chapter 3.3: Embodied Greenhouse Gas Emissions (Scope 3)

Data center construction can produce significant embodied greenhouse gas emissions (Scope 3 emissions). In data centers that consume mostly very low-carbon power (e.g., from renewables or nuclear), embodied emissions can exceed 40% of a data center’s total greenhouse gas emissions and may dominate the lifetime greenhouse gas emissions footprint.

The largest share of embodied greenhouse gas emissions at a data center comes from IT hardware manufacturing, both for chips and complementary systems, such as memory. This is chiefly due to the fluorinated gases (F-gases) used in manufacturing. These F-gases can be extremely strong greenhouse gases—between 100 and 24,000 times more potent than CO_2. After IT hardware, steel and cement production have the largest share of data center embodied greenhouse gas emissions.

Builders and operators of data centers could significantly reduce embodied greenhouse gas emissions with existing technologies. Innovative approaches, both in design and within supply chains, have the potential to reduce embedded emissions further. Materials substitution, carbon capture, green hydrogen and extending the life of IT equipment all have roles to play.

Chapter 4: Accelerating Low-Carbon Power with Artificial Intelligence (AI) Data Centers

Data centers powering AI workloads are among the fastest-growing sources of new power demand, often concentrated in regions where grid capacity is already constrained. Without deliberate intervention, this growth could entrench reliance on fossil fuel generation and increase costs for consumers. With the right strategies, however, AI can serve as a catalyst for clean energy deployment, driving innovation, financing and system reliability.

Aligning AI's growth with low-carbon power requires a coordinated toolkit. Advanced market commitments can bring forward clean, dependable generation. Demand flexibility enables data centers to shift workloads across time and geography, thereby turning consumption into a valuable grid asset. In parallel, strategic siting steers new facilities toward renewable-rich regions, which eases dependence on fossil capacity. At the same time, AI can optimize grid operations, improve forecasting and accelerate discovery of next-generation energy technologies. Taken together, these approaches couple rapid AI expansion with a cleaner, more reliable power system.

Realizing this opportunity will require coordinated action by technology providers, utilities, regulators and governments. Clear pathways to scale clean power, stronger partnerships to reduce the cost of reliable low-carbon supply, and thoughtful siting of new AI campuses can ensure growth supports both grid stability and community needs. Through timely intervention, AI can be harnessed not only to advance digital innovation but also to reinforce affordability and sustainability on a global scale.

Chapter 5: Data Center Water Use

Water consumption by data centers is tiny globally in relation to other sectors but can be very significant locally. For example, global data centers consume less than 0.008% of the water consumed by agriculture, and the water consumed to produce one hamburger roughly equals the water needed for 19,000 ChatGPT-3 queries.

However, in water-scarce regions, data centers can create significant pressure on water resources. Data centers are frequently clustered in dry areas due to land and energy availability. Two-thirds of data centers built or in development in the United States since 2022 are in areas with high water stress. Data centers can also degrade water quality if managed poorly.

Data center water use can be divided into three categories: on-site water use (Scope 1), water used off-site by power plants providing electricity to a data center (Scope 2) and embodied emissions in the data center's supply chain (Scope 3). Scope 1 water use is chiefly for evaporative cooling. As cooling technologies improve, data centers' Scope 1 water use will decrease. Data centers' Scope 2 water use can be significant, especially when power is drawn from grids with heavy coal or nuclear power. Data centers' Scope 3 water use comes mainly from the water used in producing concrete, steel and chips. Data on data center water use are poor—improving data quality and access should be a priority for companies and governments.

Text Box: Electronic Waste (E-Waste)

Data centers produce a small percentage of the total e-waste today, but data center e-waste will grow in the years ahead. The AI boom is projected to generate 1.2-5 million metric tons of e-waste cumulatively between 2020 and 2030. When improperly disposed of, the hazardous substances in e-waste pose environmental and health risks. Circular economy strategies, including reuse, refurbishment and material recovery, can offer substantial e-waste reductions. However, security concerns and a lack of standards, transparency and oversight (especially with international shipments) mean most data center e-waste is not recycled. Reducing data center e-waste requires public-private partnerships to establish global standards, producer responsibility laws and norms, and greater priority for circular design and responsible end-of-life management.

Chapter 6: Government Policy

Governments around the world are paying increasing attention to data centers' energy consumption and environmental impacts. The European Union and several EU countries have policies aligned with their net-zero greenhouse gas emissions goals. China, Japan and other Asian governments have energy efficiency standards and policies to encourage data centers to use renewable power. US federal policies change dramatically from administration to administration, with the current US administration emphasizing new data center construction as a high priority while deemphasizing environmental protection.

Standards for PUE are the most common policy tool. Water usage effectiveness (WUE) standards and disclosure obligations with respect to energy and/or water use are becoming more common as well. Some jurisdictions have imposed moratoria on new data center construction to address grid strain and other local impacts. Economy-wide policies not focused on data centers play an important role in data centers' energy use and environmental impacts.

Evaluating the impacts of data center policies can be challenging. Facility-level data are scarce; operators have strong incentives to improve energy efficiency whether or not policies require it; voluntary initiatives from leading firms can make it difficult to determine whether improvements in energy and environmental performance are the result of government policies; and rebound effects mean energy efficiency standards can lead to more compute rather than less energy consumption. Governments should prioritize robust data collection, adopt a broader and more sophisticated set of performance metrics beyond PUE, and facilitate the rapid buildout of clean power capacity to meet data centers' growing energy needs.

Text Box: Industry Initiatives

Many companies in the data center industry have made ambitious voluntary sustainability commitments. Hyperscalers, such as Amazon, Microsoft, Google and Meta, have pledged 100% renewable energy, net-zero carbon, water positivity and circular economy practices. Voluntary industry-driven initiatives, such as The Green Grid, the EU Code of Conduct, the Climate Neutral Data Centre Pact and the iMasons Climate Accord, have spearheaded use of standardized metrics, accelerated adoption of efficiency innovations and benchmarks, and expanded the focus to climate neutrality and embodied carbon.

Text Box: Local Opposition

Local opposition to new data centers is rising globally, driven by concerns about grid strain, electricity costs, water consumption, noise, land use and other factors. One report found that opposition has delayed or blocked more than $64 billion in data centers in the United States in the past two years. Data center projects have also been delayed or blocked in other countries, including Chile, Ireland, the United Kingdom, the Netherlands and Singapore. Local opposition often conflicts with national and regional policies that encourage data center development for economic growth and security reasons. To address community concerns, data center operators should engage local stakeholders throughout the project lifecycle, clearly communicating a project's potential benefits and risks and implementing measures to protect residential quality of life.