

WHITE PAPER

Sustainable AI and the rise of the energy-efficient data center

Data centers already account for up to 2% of the global consumption of electricity, and that figure is expected to double by 2030.¹ The rapid adoption of AI, particularly generative AI, is one of the main drivers of this surge in demand, pushing governments and regulators to tackle the adverse environmental implications of the technology.

The growing demand for AI is significantly straining the power infrastructure, and enterprises are starting to feel the effect in the form of soaring costs. While only a few organizations train frontier models such as ChatGPT-4, which reportedly cost up to \$100 million to develop, even midsize AI projects can consume significant power. Factor in staffing and infrastructure, and the total cost of development can escalate quickly.

As AI becomes more embedded in enterprise workflows, energy demands are rising, not just because of a few large-scale models, but also from the rapid growth of smaller, tailored models and

a surge in other AI-enabled applications spanning multiple business functions. The growing AI footprint places new demands on data center infrastructure. Enterprises now face a dual imperative to adopt more energy-efficient technology and processes while reducing costs — collectively benefiting both operational sustainability and long-term financial performance.

Fortunately, these goals need not compete. This paper discusses how IT managers must put sustainability and cost efficiency at the heart of data center optimization in the AI era.

Energy-efficient design and operations

As enterprises modernize data center architecture for AI workloads, achieving energy efficiency must be a top priority, both for the sake of sustainability obligations and the bottom line. AI workloads are inherently demanding, particularly during model training, which is immensely taxing on CPU and GPU hardware. This, in turn, generates a lot of heat.

Thermal efficiency is a major factor, with cooling accounting for about 40% of total power² consumption in a typical data center. However, that figure can vary dramatically, depending on use, data center design and even the geographic region. For instance, data centers in colder regions can significantly reduce cooling needs and, by extension, operational costs. Effective cooling is also vital for maximizing the life span of hardware components.

Case Study: Dell and AMD – Redefining Cool in the Data Center

Conventional data centers, particularly those relying on legacy infrastructure, aren't well suited to the highly parallelized nature of most AI workloads. However, data center modernization must itself be approached with care since it's easy to end up overprovisioning, resulting in avoidable wasted energy.

Addressing these challenges requires a comprehensive approach spanning the entirety of data center modernization and optimization:

1. Hardware optimizations

The processing hardware you select will directly affect workload and cooling efficiency. For instance, CPUs and GPUs optimized for AI workloads are better equipped to handle parallel processing while maintaining a lower power-to-performance ratio. Moreover, they can complete AI-related tasks faster, meaning less time spent on operations such as training and inference, resulting in reduced energy consumption overall.

2. Cooling infrastructure

Legacy cooling tends to be highly inefficient and unable to keep up with the unique demands of AI workloads. Direct-to-chip liquid cooling, which circulates coolant directly around processing units, is far more efficient and effective at keeping components cool, especially during peak use. However, if you're looking to modernize an existing air-cooled infrastructure, rather than replace it entirely, upgrading it with thermal or power sensors can enable real-time monitoring and dynamic cooling adjustments. This helps identify inefficiencies and automatically adjust airflow or workload distribution for optimal energy efficiency.

3. Virtualization and consolidation

Creating multiple virtual servers on a single physical server effectively allows you to use one machine for multiple tasks, such as data collection, model training and inference. That way, you can maximize server use while having fewer physical servers. Moreover, the smaller footprint reduces cooling requirements and physical space, further trimming capital and operational expenses.

4. Energy consumption audits

Rather than being an end goal, energy efficiency is a process of continuous optimization, driven by the availability of new technology and the ability to identify inefficiencies — particularly as AI workflows themselves change. Conducting regular energy audits will help you identify inefficiencies and areas of improvement, such as underused servers, airflow issues or the presence of specific components that have a higher-than-expected power draw.

5. AI-augmented power management

AI-powered energy management systems add a powerful software layer to any energy-efficient data center by dynamically adjusting workloads and system configurations. For instance, dynamic workload management can identify underused servers or overtaxed components and automatically adjust the allocation of resources for optimal power consumption.

Regulatory compliance and ESG standards

Data centers have become integral to global digital infrastructure to the point they now account for as much as 2% of global energy demand. That demand is only likely to grow, influenced heavily by the adoption of AI technologies across multiple industries and business use cases. Because of this, their environmental effects are facing increasing scrutiny from regulators. They're now reshaping environmental, social and governance (ESG) policies and affecting decisions among partners, investors and customers alike. It's up to data center operators to ensure they're fully aligned with those expectations.

Regulatory pressures are substantial, and highly variable between industries and jurisdictions, adding complexity to data center optimization strategies — especially among global enterprises. In California, for instance, Senate Bill 253 mandates that businesses report their greenhouse gas emissions, which also includes indirect emissions from purchased energy. Such legislation makes it all the more important that data centers prioritize renewable energy sources to reduce their carbon emissions.

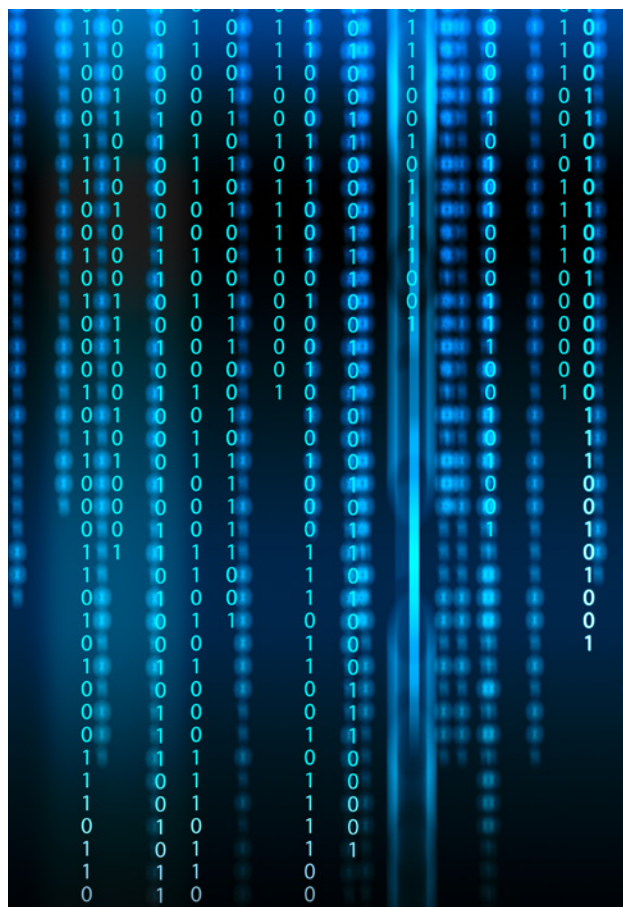
Another challenge is the lack of standardization in reporting metrics. Rather than being a universally accepted standard, ESG might be interpreted differently from one enterprise to the next, potentially leading to accusations of greenwashing. Frameworks such as the Global Reporting Initiative seek to standardize ESG reporting by highlighting accurate ways to measure key metrics, such as power use effectiveness and cooling efficiency ratio.

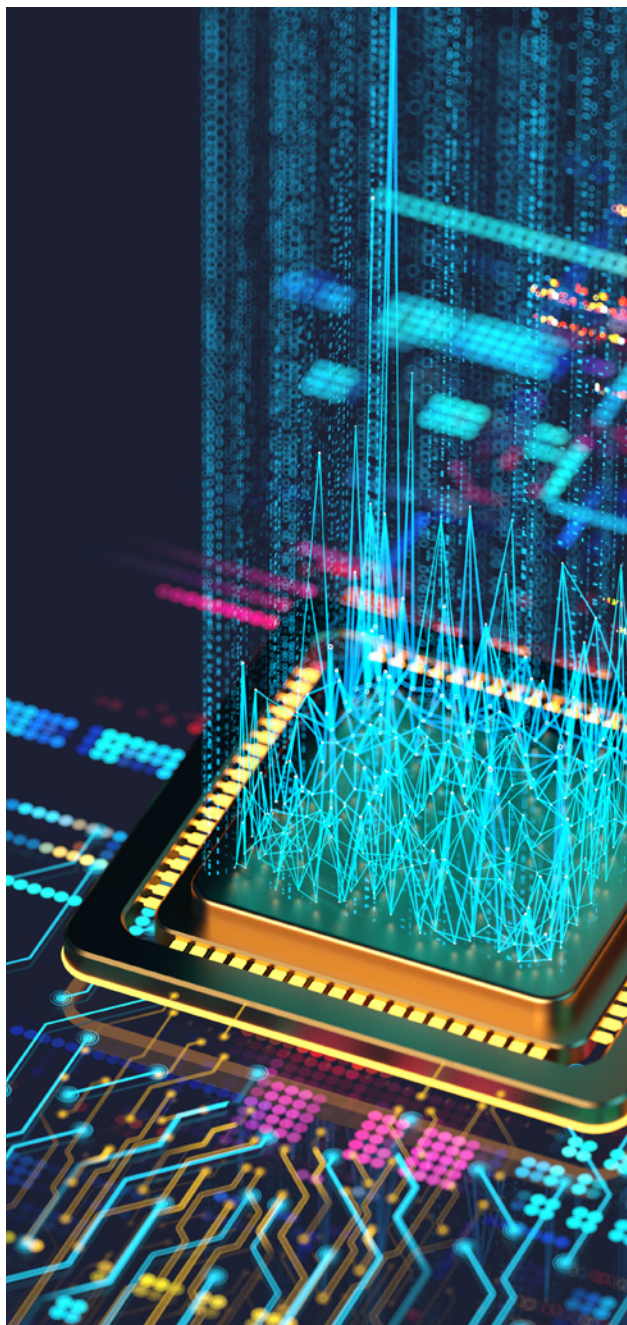
Ultimately, stakeholders — investors, customers or potential business partners — increasingly demand transparency regarding both regulatory compliance and ESG practices. Thus, building energy-efficient data centers is only the first part of the equation — you also need an established framework for evaluating and demonstrating just how energy-efficient it is.

There are several ways to address these challenges when optimizing your data center for enhanced energy efficiency:

1. ESG reporting frameworks

Any enterprise can choose how it reports its ESG efforts, but convincing stakeholders of the legitimacy of such reports is another matter. Because of this, it's better to focus on a widely adopted standard or compliance directive, even if it's not mandatory. For example, the EU AI Act establishes a framework governing the sustainable and responsible use of AI, setting the first global standard of its kind. Other standards and frameworks to consider include the GHC Protocol, EU CSRD, SEC climate disclosure rules or even recently proposed bills such as the New York State Sustainable Data Centers Act.





2. Third-party audits and certifications

When building and optimizing your AI data center, be sure to pursue recognized certifications from the outset, since doing so clearly demonstrates and validates your sustainability efforts. For example, the ISO 14001 is a globally recognized standard for environmental management systems, providing a framework for the green data center. Additionally, regular third-party audits can validate your ESG and compliance practices, while also demonstrating continuing efforts to meet certified standards

3. Internal sustainability policies

External regulation of AI, particularly with regard to its environmental effect, is still in its infancy. This will likely change over the longer term, so it's worth developing internal sustainability policies that go beyond the minimum standards currently mandated in your jurisdiction. For instance, the EU AI Act emphasizes³ the need to proactively assess and minimize the effect of AI systems on environmental sustainability and adopt energy-efficient programming and systems for its training and use. Even if the act doesn't apply to U.S. businesses (unless they're operating in the EU), it can serve as a strong starting point for building internal sustainability policies.

4. Compliance management technology

While AI-augmented energy management tools can help by continuously monitoring and optimizing energy use in the data center, they also collect valuable data that can itself help with demonstrating your compliance efforts. By layering on compliance management systems, you can streamline auditing, generate reports and monitor your compliance with your chosen frameworks and required mandates. Moreover, this enables transparent communication with stakeholders by simplifying sustainability reporting and demonstrating accountability.

Financial benefits of sustainable practices

Making sustainability an integral part of your data center optimization strategy isn't just about keeping investors happy and adhering to legal mandates and industry standards. There are significant financial benefits, too, especially with regard to operational costs.

As AI advances, it will continue to become more data-hungry, directly resulting in increased hardware and cooling requirements. For example, the cost of training frontier models increased by an average of 4,400% between the 2017 launch of Google Transformer to the 2023 launch of Google's Gemini Ultra, the latter clocking in at an estimated \$191 million.⁴

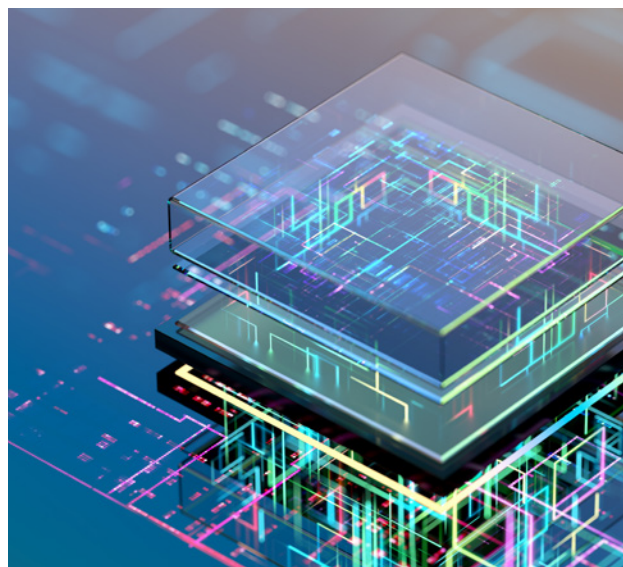
Even if you're not building a massive, general-purpose AI model, development costs can still spiral out of control. Indeed, enterprises are likely to generate only more data, and the more data they have to feed into their AI models, the more powerful those models will be. For instance, building and training a dedicated customer service chatbot might be fairly inexpensive if you're only training it on a year's worth of data spanning interactions with a few hundred customers. But, for big enterprises with tens of thousands of customers or smaller ones that need to augment their model training processes with external data sets, the technical demands — and by extension, the costs — can grow exponentially.

Since AI projects and workloads are rarely static, scale is a key consideration from the outset. Data center operators should keep this in mind by adopting modular infrastructures comprising energy-efficient components that can easily be upgraded to keep capital and operational expenses in sync with business value creation.

Using energy-efficient processors and effective cooling systems doesn't just keep operational costs under control. There are also indirect financial benefits to brand reputation, competitiveness and investor relations.

The energy-efficient AI data center can even go so far as to recover some of the costs through excess heat reuse and energy-grid contributions. Repurposing waste heat from data centers is well established in some regions, and it has an integral role in the development of smart cities. For example, in the Hyllie district of Malmö, Sweden, data centers and other facilities have partnered with local energy partner E.ON to redistribute excess energy⁵ from industrial and commercial processes into local heating supplies. Similarly, in Frankfurt, Germany, home to 60 data centers and counting, projections suggest that excess heat from data centers alone could meet the city's heating needs by 2030.⁶ In the U.S., green data centers are also becoming a thing, as exemplified by major hyperscale operators⁷ such as AWS, Google and Meta procuring renewable energy solutions to power their systems.

Examples like these highlight that renewable energy and heat reuse are not only feasible but also major potential contributors to sustainability goals. By reusing waste heat, partnering with utility providers or capitalizing on green tax incentives, data center operators can turn some of their biggest challenges into opportunities that benefit their local environments and communities, while recouping some of their operational costs.



A sustainable path forward for the AI-ready data center

AI adoption continues to accelerate as enterprises start to realize the value of developing and training their own models to address unique business challenges. However, as the technical demands grow, so, too, must the role of sustainability and energy efficiency — both of which should be central to any data center modernization strategy.

AMD provides high-performance hardware tailored specifically to the demands of AI workloads, enabling enterprises to reduce the number of servers they need and, in turn, reduce their energy costs. By selecting these AI-specialized solutions, IT leaders can modernize their data centers in line with the broader sustainability goals of the enterprise.

Learn more about AMD's energy-efficient AI solutions



References

1. AI is poised to drive 160% increase in data center power demand | Goldman Sachs
2. How Much Energy Do Data Centers Use? - Dataspan
3. Article 95: Codes of Conduct for Voluntary Application of Specific Requirements | EU Artificial Intelligence Act
4. AI Model Training Cost Have Skyrocketed by More than 4,300% Since 2020 - Edge AI and Vision Alliance
5. Learning from Scandi innovation: how Sweden's 'city of tomorrow' is leading the way in climate-smart urban development | Urban Energy Innovation | The Guardian
6. Data centres could be a source of heat for European cities, says Danfoss boss (original source is a gated article on FT)
7. Green Data Centers in the USA: The Future is Here | PRS USA