Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

ENERGY EFFICIENCY OF DISTRIBUTED AI SERVICES IN A HYBRID CLOUD-EDGE INFRASTRUCTURE

Bekzhan Abdimanapov,
Software Engineer, USA

Abstract
This paper examines aspects of improving data center energy efficiency with the expansion of cloud and edge computing technologies. Architectural features, energy consumption benefits of decentralized data processing, and methods for reducing energy consumption are explored. Cloud and edge computing models are compared from the respective of the environmental sustainability of digital infrastructure.

Keywords: Data center, energy efficiency, cloud computing, edge-computing, energy consumption, green IT.

Introduction

The scientific novelty of this work lies in its comparative analysis of the energy efficiency of aloud and edge computing and the substantiation of the advantages of their hybrid integration for reducing the overall energy consumption of data centers.

The rapid growth of the digital economy, artificial intelligence technologies, the Internet of Things (IoT) and streaming services is leading to an unprecedented increase in the volume of data being processed. Data centers – key elements of modern IT infrastructure – are primarily responsible for storing, processing, and transmitting information. However, their operation is associated with high energy consumption and negative impact on the environment.

According to international research, the energy consumption of these centers ranges from 1 to 1.5% of global demand, with this figure expected to increase due to increased computing load [1]. Key elements determining energy consumption

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

include server hardware, cooling systems, network equipment, and uninterruptable power supply infrastructure.

The classic centralized cloud computing model is based on the concentration of computing power on a large scale within hyperscale data centers. Despite the high degree of virtualization efficiency and scalability, this structure is associated with significant energy consumption – both in processing information and in moving it through network infrastructure [2].

As energy demand and data processing latencies increase, the concept of edge computing is becoming increasingly widespread. According to this model, part of the computational processing is moved closer to the points where information is generated – to devices located at the edge of the network. This approach helps reduce the volume of data transferred, improve response times, and alleviate the burden on central server rooms [3].

Research shows that the combination of cloud and edge solutions created a hybrid computing environment that enables more efficient resource allocation and improve energy performance of digital infrastructure [4]. Key technologies in this process include adaptive load balancing, smart energy consumption control, and colling system optimization.

The relevance of this topic is driven by increasing demands for environmental sustainability and reducing the carbon footprint of IT infrastructure. Leading cloud solution providers are using renewable energy sources, heat recovery technologies, and AI-based energy management algorithms in data centers [1].

Therefore, studying the energy efficiency of data centers combining cloud and edge technologies is of great scientific and applied interest, enabling the identification of promising ways to reduce energy consumption within digital infrastructure.

Data centers play a fundamental role in the development of digital infrastructure, proving storage, processing, and transmission of information for cloud solutions, artificial intelligence technologies, the Internet of Things, and corporate IT systems. Operating such centers requires high energy consumption, making energy efficiency a critical issue for both financial feasibility and environmental sustainability.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

According to international research, the annual energy consumption of data centers is estimated at 200-250 TWh, which corresponds to approximately 1-1.5% of global electricity consumption [1]. Despite the increased computing load, the use of energy-efficient solutions has made it possible to reduce the rate of growth in energy consumption.

Data center energy consumption consists of several key components. Most of this energy is spent on computing equipment – servers, video processors, and storage systems. Significant costs are also incurred by network equipment and auxiliary systems [5].

Cooling systems play a key role in the energy consumption structure. Ensuring optimal thermal condition is crucial for the reliable operation of hardware. According to various estimates, cooling energy consumption accounts for 30 to 40% of a data center's total energy consumption [6], which stimulates the implementation of advanced heat removal solutions such as liquid and immersion cooling.

The Power Usage Effectiveness (PUE), developed by The Green Grid consortium, is widely used to analyze the energy efficiency of data centers. It is calculated as the ratio of a building's total energy consumption to the energy consumed solely by information systems. The closer the PUE value is to one, the more efficient the energy infrastructure is considered [7].

Increased energy consumption is further driven by increased computing density, driven by the use of high-performance GPU clusters and AI platforms, which leads to increased heat generation and the need for more efficient cooling and power supply solutions [1].

Modern research notes that network data delivery has a significant impact on overall energy consumption in clouds. The exchange of large volumes of information between end users and central computing nodes increases the load on telecommunications systems, which indirectly leads to increased energy costs [5].

Thus, the energy consumption of modern data centers is determined by a combination of computing, engineering, and networking parameters. Improving the overall energy efficiency of digital infrastructure requires optimizing each of these aspects.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

Cloud technologies significantly improve data center energy efficiency by consolidating and virtualizing computing infrastructure. Centralizing resources increases resource utilization and reduces the number of idle nodes.

A key element in improving energy efficiency is workload balancing and virtual machine migration, which enables the decommissioning or downtime of idle servers while maintaining stable service operation. Hyperscale clouds, equipped with smart cooling systems, renewable energy sources, and energy management algorithms based on artificial intelligence, play a significant role in improving energy consumption [1].

The cloud approach facilitates the efficient use of computing power and reduces energy consumption per volume of processed data.

Table 1 – Impact of cloud computing on energy efficiency

Factor	Mechanism of Influence	Energy Effect
Virtualization	Server Consolidation	Reducing of Number of Active Nodes
VM Migration	Load Redistribution	Shutting Down Idle Servers
Autoscaling	Dynamic Resource Allocation	Power Consumption Optimization
Hyperscale Data Centers	High Computing Density	PUE Reduction
AI Cooling Management	Thermal Load Forecasting	Reducing Cooling Energy Consumption

Edge computing involves moving data processing from central clouds to distributed nodes located closer to the data generation points. This method is widely used in IoT architectures, the Industrial Internet, and smart transportation systems.

A key advantage of edge computing is its high energy efficiency. This is achieved by minimizing the volume of data transferred, which, in turn, relieves the load on backbone networks and central data centers. Processing information directly on-site significantly reduces the energy consumption required for transmission, storage, and centralized processing.

Another important advantage is the reduced cooling requirements of large-scale data centers. Because the computing load is distributed, the need for intensive cooling of central nodes is reduced. Edge nodes, however, due to their

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

localization, have less computing power and, consequently, lower power consumption.

The centralized nature of edge computing raises challenges in power management, computing load distribution, and resource coordination. These issues continue to pose significant challenges in science and technology.

Table 2 – Energy effects of edge computing implementation

Factor	Optimization Mechanism	Energy Effect
Local Data Processing	Reduced Cloud Transfer	Reduced Network Energy Costs
Traffic Reduction	Data Filtering and Aggregation	Reduced Data Center Load
Computation Distribution	Partial Cloud Offloading	Reduced Hyperdata Center Energy Consumption
Low-power Edge Nodes	Use of Microservers	Local Energy Savings
IoT Context Processing	Processing at the Source	Optimization of Overall Energy Consumption

Energy analysis in cloud and edge computing systems plays a key role in determining the viability of future digital networks. Cloud architecture, with its concentrated computing resources, offers advantages in terms of flexible scalability and optimized utilization of computing power. However, the operation of large data centers, which require intensive cooling, and the transfer of massive amounts of data are associated with significant energy consumption. Distributed data processing at the edge, located directly at the points of information creation, offers the potential to optimize energy consumption. This model reduces network response times, reduces the volume of data transferred and partially relieves the load on central cloud servers, contributing to an overall reduction in energy consumption in the digital ecosystem.

Table 3 – Comparative Energy Efficiency of Cloud and Edge Models

Parameter	Cloud Computing	Edge Computing	Hybrid Model
Resource Centralization	High	Low	Medium
Data Transfer Energy	High	Low	Optimized
Cooling Costs	High	Low	Reduced
Processing Latency	Medium	Minimal	Minimal
Scalability	Very High	Limited	High
Overall Energy Efficiency	Medium	High Local	Highest

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

The study revealed that each of the computing models under consideration has its own unique advantages and disadvantages in terms of energy consumption energy costs by decentralization. Cloud solutions are distinguished by centralized resource management and efficient hardware utilization. Meanwhile, edge computing reduces network and infrastructure energy costs by decentralizing data processing. A hybrid architecture combining cloud and edge components appears to be the most promising for achieving high energy efficiency. It enables the flexible redistribution of computing tasks between remote data centers and local computing nodes, achieving an optimal balance between performance, scalability, and energy consumption of the digital infrastructure.

Improving data center energy efficiency is a priority for the development of modern digital infrastructure. One key solution is virtualization and server consolidation. This method helps increase physical equipment utilization and reduce the number of computing units in use.

The integration of advanced computing workload management systems that utilize artificial intelligence algorithms is essential. This enables dynamic task distribution and optimized energy costs. Furthermore, self-scaling and virtual machine migration technologies are being successfully implemented.

To achieve significant results, it is necessary to improve cooling mechanisms, such as liquid and immersion cooling, and to utilize free cooling techniques. The implementation of renewable energy sources and heat recovery technologies is a promising direction.

Thus, the combination of advanced computing, engineering and energy solutions helps to significantly reduce data center energy consumption and improve their environmental profile.

The author successfully applied distributed intelligence computing principles to the development of several projects, including web and voice assistants, such as Smart Website Assistant, and voice interface management systems. The design of these solutions involved performing partial data processing directly on the user's device (edge layer). This approach reduced the load on central servers and optimized network data exchange. This method demonstrates the potential of hybrid cloud-edge models for improving the energy efficiency of digital services and distributed processing of user requests.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

The rapid growth of digital information and increasing computing demands pose a challenge for modern IT infrastructure to improve data center energy efficiency. Research has shown that the widespread centralized cloud computing model, while highly flexible and resource-efficient, required significant energy resources. These costs are primarily associated with data transfer and the need to maintain cooling systems.

Edge computing allows for the transfer of some computing tasks closer to the data sources, reducing network traffic and energy consumption previously spent on centralized processing. A hybrid architecture combining cloud and edge technologies appears to be the most effective solution. This approach achieves an optimal balance between performance, scalability, and ongoing energy consumption.

Further improvements in energy efficiency are achieved through the implementation of virtualization technologies, intelligent load management, advanced cooling systems, and the transition to renewable energy sources. Therefore, a comprehensive approach, including a combination of cloud and edge computing, along with modern engineering developments, is key to building a sustainable and cos-effective digital infrastructure for the future.

References

1. Masanet, E., Shehabi, A., Lei, N. [and other]. Recalibrating global data center energy-use estimates // Science. - 2020. - Vol. 367, No. 6481. - P. 984–986. - DOI: 10.1126/science. aba3758.
2. Buyya, R., Yeo, C. S., Venugopal, S. Market-oriented cloud computing: Vision, hype, and reality // Future Generation Computer Systems. - 2009. - Vol. 25, No. 6. - P. 599–616. - DOI: 10.1016/j.future.2008.12.001.
3. Shi, W., Dustdar, S. The promise of edge computing // Computer. - 2016. - Vol. 49, № 5. - P. 78–81. - DOI: 10.1109/MC.2016.145.
4. Deng, R., Lu, R., Lai, C., Luan, T. H., Liang, H. Optimal workload allocation in fog–cloud computing // IEEE Transactions on Industrial Informatics. - 2016. - DOI: 10.1109/TP.2016.2569094.
5. Koomey, J. G. Growth in data center electricity use 2005–2010. - Oakland: Analytics Press, 2011.

Modern American Journal of Engineering, Technology, and Innovation
ISSN(E): 3067-7939
Volume 01, Issue 08, November, 2025
Website: usajournals.org

6. Bash, C., Forman, G. Cool job allocation: Measuring the power savings of placing jobs at cooling-efficient locations // Proceedings of the USENIX Annual Technical Conference. - 2007.
7. Belady, C. In the data center, power and cooling costs more than the IT equipment it supports // Electronics Cooling. - 2007