**Data Centers Energy Consumption**

**Abstract:**

Data centers provide critical support for the information technology (IT) equipment including computing devices and servers for storage and data processing. It is estimated annually in the U.S the data centers consume roughly about 50% of electricity mainly by the equipment. The cooling needs Heating, Ventilation, and Air Conditioning (HVAC) are estimated to be up to 40% using computer room air-conditioners or to cool down the equipment such as servers, and other IT equipment in the data centers. This paper discusses the high energy needs of data centers and also reduce energy use by making systems as efficient as possible. Providing sustainable data centers is the energy goal so as to maximize energy from renewable systems.

Keywords: Data Centers, Information Technology, Computing Devices, Servers, Storage, Data Processing, Cooling Needs, HVAC, Energy Efficiency, Renewable Energy.

1. **Introduction:**

The digital era depends on data centers to operate as the fundamental infrastructure which enables cloud computing and artificial intelligence and financial services and social media (Dayarathna, Wen, & Fan, 2015). The exponential growth of digital demand leads to proportional increases in data center energy usage. The facilities containing thousands of servers and critical IT equipment use an estimated 1–2% of global electricity which will increase substantially in upcoming years (Dayarathna, Wen, & Fan, 2015). The following bar graph shows global data center energy consumption data from 2010 to 2024 in terawatt-hours (TWh). The data shows continuous growth in energy consumption because of increasing demand for cloud computing and digital services and AI workloads. Data centers require energy to power their computing equipment as well as to maintain proper environmental conditions through their extensive cooling systems (Ahmed, Bollen, & Alvarez, 2021). The high energy consumption of data centers creates multiple operational challenges that affect both financial expenses and environmental sustainability and carbon footprint. The growing public and regulatory demands to decrease greenhouse gas emissions have made data center energy consumption management an essential priority for industry leaders and policymakers (Dayarathna, Wen, & Fan, 2015).

The research investigates the main elements that determine energy consumption in data centers. The paper examines two emerging technologies and strategies for decreasing energy usage which include CRAC systems and free cooling and chilled water systems. The research evaluates different parameters such as power usage effectiveness (PUE), airflow efficiency, and cooling system efficiency through performance comparisons of these factors in data centers to optimize energy consumption.

A graph of a number of years

AI-generated content may be incorrect.

**Figure 1. Data Center’s Global Energy Consumption (2010-2024)**

1. **Role of Energy Consumption of Different IT systems in Data Centers:**

The energy consumption of data centers depends on various Information Technology (IT) systems that operate at different power levels (Cheung, Wang, & Zhuang, 2018). The creation of energy-efficient data center infrastructures needs understanding of system operations together with their associated energy consumption levels. The main categories consist of network equipment, servers, power supply systems, cloud and computing infrastructure and storage devices as shown in figure 1 (Joshi & & Kumar, 2012).

A diagram of a power supply system

AI-generated content may be incorrect.

**Figure .2 Block Diagram of Data Center (Joshi & & Kumar, 2012).**

1. **Servers:**

Data centers rely on servers as their fundamental computing infrastructure. Servers execute applications and maintain websites and databases and handle data processing operations. The total energy consumption of data centers depends on workload and configuration because servers use between 30–50% of the total power as shown in figure 3 (Cheung, Wang, & Zhuang, 2018). The continuous operation of high-performance servers for 24 hours results in substantial power usage even though they operate at low levels of utilization. The energy consumption of servers depends on CPU and GPU usage as well as cooling requirements (Cheung, Wang, & Zhuang, 2018). There are different types of servers like rack servers, blade servers, tower servers, and high-density servers.

1. **Network Equipment:**

Ther is a connection between internal systems and external systems in a data center and that job is done by network equipment. The network infrastructure consists of switches together with routers and firewalls and load balancers. The network equipment pull up to 10 to 15% of power as shown in figure 3 in big data centers even though they consume lower power than server. The adoption of Energy-Efficient Ethernet and Software Defined Networking (SDN) technologies authorizes organizations to reduce their power consumption when the network flow is low.

1. **Power Supply Systems:**

The power supply and distribution systems provide continuous power supply and protect IT systems from disruptions (Cheung, Wang, & Zhuang, 2018). Data centers primarily use four main components which include UPS systems and power distribution units (PDUs) and backup generators and transformers. The energy loss rate of UPS systems ranges between 5% and 15% based on design specifications and current load levels as shown in figure 3. The design of power infrastructure needs to be efficient to reduce energy losses (Joshi & & Kumar, 2012). The N+1 and 2N redundancy models enhance reliability but lead to additional energy consumption (Cheung, Wang, & Zhuang, 2018).

1. **Cloud and Computing Infrastructure:**

Cloud computing platforms together with virtualized environments enable resource optimization through their ability to run multiple virtual machines or containers on one physical server. Virtualization enables better resource utilization because it decreases the number of physical servers needed which results in improved energy efficiency (Cheung, Wang, & Zhuang, 2018). Hyperscale cloud providers achieve maximum energy efficiency per compute task through their implementation of dynamic workload distribution and predictive cooling systems and AI-based optimization methods (Joshi & & Kumar, 2012).

1. **Storage Devices:**

The storage systems function to store and handle large amounts of both structured and unstructured data. The storage systems consist of traditional hard drives (HDDs), solid-state drives (SSDs) and storage arrays. The power consumption of storage systems reaches between 10–20% of total energy usage based on operational patterns and redundancy protocols as shown in figure 3 (Joshi & & Kumar, 2012) (Cheung, Wang, & Zhuang, 2018). SSDs require less power than HDDs, but their price remains higher. The implementation of object storage together with tiered storage strategies (hot, warm, cold) enables organizations to reduce their power consumption (Cheung, Wang, & Zhuang, 2018).

1. **Cooling and Environmental Control (Indirect IT Systems):**

The IT systems do not include cooling systems which consist of CRAC (Computer Room Air Conditioners), CRAH (Computer Room Air Handlers) and liquid cooling systems that support IT infrastructure. The power consumption of cooling systems reaches 30–40% of the total data center power usage as shown in figure 3. The reduction of overhead depends on free-air cooling and liquid immersion and AI-based thermal control innovations (Cheung, Wang, & Zhuang, 2018).

A pie chart with different colored circles

AI-generated content may be incorrect.

**Figure 3. Chart Representing Energy Usage by different IT Components in a Data Center**

A screenshot of a graph

AI-generated content may be incorrect.

**Table 1. Percentage of Energy Usage by different IT Components in a Data Center.**

1. **Different Types of Cooling Systems:**
2. **Computer Room Air Conditioning (CRAC) Systems:**

CRAC systems operate as specialized cooling solutions which maintain precise temperature and humidity and airflow conditions in critical facilities including data centers and server rooms and telecom facilities. CRAC units operate differently from standard HVAC (Heating, Ventilation, and Air Conditioning) systems because they deliver precise climate control to safeguard sensitive IT equipment from heat damage while sustaining optimal performance (Fajardo, 2022). The systems operate within temperature ranges of 18–27°C (64–80°F) and maintain relative humidity levels between 40–60% to meet ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers) standards which prevent electrostatic discharge and condensation problems (Fajardo, 2022). CRAC systems maintain consistent cooling performance through advanced sensors and real-time monitoring which delivers temperature stability at ±1°C. The systems implement N+1 or 2N cooling redundancy to maintain continuous operation when a unit fails. The airflow management system includes hot aisle/cold aisle containment and raised-floor plenums to achieve maximum cooling efficiency (Fajardo, 2022). CRAC systems of today focus on energy efficiency through the implementation of variable-speed fans and economizers and AI-driven optimization which decreases power usage and enhances data center PUE (Fajardo, 2022).

CRAC systems play a crucial role in maintaining uptime in mission-critical environments because overheating causes approximately 45% of IT hardware failures. CRAC systems are commonly deployed in big data centers and enterprise server rooms and telecom hubs because small temperature changes can trigger equipment failures and downtime (Fajardo, 2022). CRAC technology has evolved through AI predictive maintenance and hybrid liquid-cooling solutions for high-density servers and modular designs for edge computing applications. CRAC systems maintain stable operating conditions which improves reliability and enables organizations to fulfill ASHRAE 90.4 and TIA-942 standards while lowering their cooling-related energy expenses (Fajardo, 2022).

1. **Chilled Water Systems:**

A chilled water system functions as an efficient central cooling system which employs fundamental thermodynamic principles to control temperatures in large buildings and industrial facilities and data centers through water as its heat transfer medium (Trautman, 2021). The system operating performance follows the heat transfer equation where equals cooling capacity in represents water mass flow rate in kg/s and cₚ represents water specific heat in and ΔT represents the supply and return water temperature difference ranging between 5-7°C (Trautman, 2021). The system consists of a chiller unit which operates between 4-7°C (39-45°F) through mechanical compression or absorption cooling (using heat-driven refrigeration cycles with thermal ) (Trautman, 2021).

The system distributes cooled water using pumps whose power consumption depends on the hydraulic formula, where V is volumetric flow rate is pressure drop (kPa), and η is pump efficiency (0.6-0.8) (Ma Z. &., 2009). Cooling coils receive warm air which allows heat transfer through Newton's law of cooling where h represents the convective heat transfer coefficient (Ma & Wang, 2009). The thermodynamic cycle of the system concludes with the return water discharging heat into the environment through cooling towers which utilize evaporative cooling principles with their operation limited by wet-bulb temperature .

The system implements cooling towers which determine their heat rejection capacity using the Merkel equation by maintaining an approach temperature between 3-5°C above the wet-bulb temperature for maximum efficiency (Ma Z. &., 2009). Chilled water systems exist in three main configurations that include air-cooled chillers (with a lower COP between 2.5-3.5) and water-cooled chillers (with a higher COP ranging between 4.0-6.0) and absorption chillers (thermal COP between 0.7-1.2) that leverage waste heat recovery according to the Second Law of Thermodynamics (Tirmizi, 2012).

The systems maintain precise temperature control through water's high heat capacity properties which make them suitable for applications needing stability within ±0.5°C (Ma Z. &., 2009). The energy-saving potential of these systems remains substantial because of their plant efficiency ratio yet they require Langelier Saturation Index monitoring to prevent scaling since LSI > 0 indicates potential scaling issues. The correct design needs to follow both the First Law of Thermodynamics (energy balance) and the Second Law (entropy minimization) to reach optimal operational performance and system durability (Tirmizi, 2012).

1. **Free Cooling:**

Free cooling represents an energy-efficient method that uses minimal outside temperatures to decrease or eliminate the requirement for mechanical cooling in heating ventilation and air conditioning systems and process cooling systems (Zhang H. S., 2014). Free cooling operates by utilizing outside cool air or water through heat exchangers or cooling towers to directly or indirectly cool building or process loads when temperature conditions are favorable thus eliminating the need for chiller compressors (Zhang H. S., 2014). The cooling capacity of free cooling can be determined by the formula ), where represents the water flow rate and  *cp* is water’s specific heat and reflects the temperature difference between return chilled water and cooling tower output (Raj, 2010). The supply of outdoor air to space occurs directly through filtering in direct free cooling systems yet indirect systems employ plate heat exchangers to exchange cooling from the cooling tower loop with the chilled water loop without fluid mixture (Zhang H. S., 2014). The system operates best when the approach temperature remains between 2–5°C of the wet-bulb temperature for maximum performance (Zhang Y. W., 2017). Free cooling becomes most effective in regions with extended periods of wet-bulb temperatures below 10°C since they can save up to 70% of chiller energy consumption . The implementation of free cooling systems requires proper heat exchanger dimensions and automated building management system integration for automatic mode transitions (Zhang Y. W., 2017). Free cooling systems generate significant operational cost savings but demand larger upfront expenses together with meticulous maintenance to avoid fouling problems. Modern advancements combine AI-driven controls with thermal storage systems to extend free cooling availability in suitable climates which enables 30–60% annual cooling energy reductions for data centers and industrial facilities (Zhang Y. W., 2017).

1. **Bench Marking:**
2. **Power Usage Effectiveness:**

The main efficiency assessment tool for data centers is Power Usage Effectiveness (PUE) which measures IT equipment power consumption against total facility power usage. The ideal PUE rating of 1.0 would mean all power goes to computing devices yet most facilities operate between 1.1 and 2.0+ due to their cooling system advancements (Brady, Kapur, Summers, & Thompson, 2013). The metric shows how much energy supports cooling systems (40-50% of non-IT load) and power distribution (10-20%) and lighting/auxiliary loads (5-10%). The leading operators, including Google achieve PUEs between 1.10-1.12 across their entire data center fleet through modern data center strategies that combine free cooling with hot/cold aisle containment and high-efficiency UPS systems and liquid cooling solutions (Brady, Kapur, Summers, & Thompson, 2013) (Zoie, 2017). The PUE metric provides essential infrastructure optimization, but it lacks complete effectiveness because it fails to measure IT workload efficiency and shows significant climate variations that impact free cooling potential and requires WUE as a supplementary sustainability assessment tool. The combination of DCIM system PUE monitoring with AI-driven cooling optimization allows facilities to achieve ongoing improvements in their energy management systems (Brady, Kapur, Summers, & Thompson, 2013).

1. **Airflow Efficiency:**

Data centers require efficient airflow management to achieve optimal cooling performance while minimizing their energy usage (Ni, Jin, & Zhang, 2017). The system controls air circulation to deliver cold air to IT equipment properly while stopping hot exhaust air from returning into systems (Ni, Jin, & Zhang, 2017) (Lu, 2018). The correct management of airflow prevents two major problems which are hot spots that create localized overheating and bypass airflow that allows conditioned cooling to escape without proper cooling equipment. Hot aisle/cold aisle containment systems represent key airflow efficiency strategies because they use physical barriers to maintain controlled cooling paths between hot and cold air streams (Ni, Jin, & Zhang, 2017). The placement of perforated tiles within raised floor configurations along with airflow rate adjustments enables precise cold air delivery to specific areas (Lu, 2018). Blank panels installed in unused rack spaces block air mixing and prevent air recirculation while computational fluid dynamics (CFD) modeling helps data center operators detect operational inefficiencies before they affect system performance (Lu, 2018). Data centers achieve better cooling consistency and longer equipment life and reduced energy waste through improved airflow efficiency which leads to decreased power usage effectiveness (PUE) and operational expenses. The advanced facilities use these methods alongside intelligent monitoring systems which adjust cooling dynamically according to real-time thermal conditions to achieve maximum energy efficiency in data center design (Ni, Jin, & Zhang, 2017).

1. **Cooling System Efficiency in Data Centers:**

Data center operations depend heavily on cooling system efficiency because it directly affects both energy usage and operational expenses and environmental sustainability (Capozzoli, 2015). The optimization of cooling systems remains essential because they represent up to 40% of total data center power usage (Capozzoli, 2015). Data centers today use precision cooling technologies which include CRAC and CRAH units that adjust cooling output through variable-speed fans and intelligent controls based on real-time heat loads to prevent energy waste. The implementation of hot and cold aisle containment stands as a fundamental efficiency enhancement strategy because it physically separates hot exhaust air from cold intake air to prevent mixing which optimizes cooling resource utilization (Capozzoli, 2015). The use of free cooling (economization) takes advantage of external ambient air or water temperatures to decrease dependence on the mechanical cooling system which results in reduced energy usage particularly in cooler climates. The adoption of liquid cooling technologies including direct-to-chip and immersion cooling continues to rise because they provide better heat transfer efficiency than traditional air-cooling systems for high-density server environments. The analysis of airflow patterns through Computational Fluid Dynamics (CFD) modeling helps identify inefficiencies which leads to data-driven optimization (Capozzoli, 2015). Data centers can achieve better Power Usage Effectiveness (PUE) ratios and extend hardware lifespan and reduce their carbon footprint through integrated approaches which maintain optimal operating conditions for critical IT infrastructure (Capozzoli, 2015).

1. **Results and Discussion:**

The evaluation of the new data center after construction showed substantial enhancements in cooling efficiency together with improved energy performance (Zhang & Shao, 2021). The relocation to the optimized facility resulted in an average temperature drop from 25.88°C to 15.40°C (41% decrease) for eight high-scale GPU servers with three NVIDIA Tesla V100 GPUs each (Jones, 2021) (Fulton, 2020). The redesigned airflow management system and CRAC redundancy strategy (rotating three legacy Stulz ASD1072A units) achieved this improvement despite a 23% increase in equipment load. The GPU Server 06 experienced a 46% temperature reduction from 28.18°C to 15.26°C which protected compute-intensive workloads from thermal throttling risks. The facility operated at a PUE of 1.123 which matched Google and Microsoft hyperscalers while achieving 88% DCiE efficiency which meant only 12% of energy went to non-IT overhead.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Ore-Renovation** | **Post-Renovation** | **Improvement** |
| Avg. GPU Server Temperature | 25.088°C | 15.4° | 41%↓ |
| GPU Server 06 Temperature | 28.18° | 15.26° | 46%↓ |
| PUE | - | 1.123 | Matches hyperscalers |
| DCiE | - | 88% | 12% non-IT overhead |
| Equipment Load | - | 23% | - |
| Coolong Units | Legacy Stulz ASD1072A | Redundant CRAC | Rotating Strategy |

**Table 2. Temperature & Efficiency Comparison Table (Fulton, 2020).**

The research supports (Fulton, 2020) main point about inlet temperature control as an essential efficiency factor while showing that upgrading existing cooling systems remains an unexplored efficiency opportunity (Fulton, 2020). The study's 7-month duration might not fully capture seasonal patterns and the absence of WUE metrics requires further research especially in water-constrained areas. The project’s success proves that strategic layout modifications including raised-floor plenums and top-rack hot-spot mitigation techniques can match new data center performance while providing a budget-friendly approach to modernizing existing infrastructure (Fulton, 2020).

1. **Conclusion:**

In conclusion, data centers function as the fundamental infrastructure of modern times because they support both cloud computing operations and AI applications in the rapidly expanding digital world (Dayarathna, Wen, & Fan, 2015). The quick expansion of data centers generates substantial environmental expenses because of their elevated energy usage (Dayarathna, Wen, & Fan, 2015). The research investigated the different elements which drive data center energy consumption starting with servers and storage systems and network equipment and power distribution units and finishing with cooling systems that represent 40% of total energy requirements (Ahmed, Bollen, & Alvarez, 2021). Data centers need a comprehensive strategy that combines operational excellence with environmental responsibility and financial sustainability to enhance their energy efficiency. The development of sustainable digital infrastructure depends on adopting new technologies and optimizing current systems while integrating renewable energy resources. Future research must create regionally adaptable solutions that reduce data center environmental impact because digital performance expectations will continue to grow (Ahmed, Bollen, & Alvarez, 2021).

1. **References:**
2. Dayarathna, M., Wen, Y., & Fan, R. (2015). Data center energy consumption modeling: A survey. IEEE Communications surveys & tutorials, 18(1), 732-794.
3. Ahmed, K. M. U., Bollen, M. H., & Alvarez, M. (2021). A review of data centers energy consumption and reliability modeling. IEEE access, 9, 152536-152563.
4. Cheung, H., Wang, S., Zhuang, C., & Gu, J. (2018). A simplified power consumption model of information technology (IT) equipment in data centers for energy system real-time dynamic simulation. Applied energy, 222, 329-342.
5. Joshi, Y., & Kumar, P. (2012). Introduction to data center energy flow and thermal management. In Energy Efficient Thermal Management of Data Centers (pp. 1-38). Boston, MA: Springer US.
6. Fajardo, J., Ramírez-León, H., Barreto, D., Rico, C., & Cardona, C. (2022, October). Energy Efficiency Condition-Based Maintenance Methodology for Computer Room Air Conditioners. In ASME International Mechanical Engineering Congress and Exposition (Vol. 86687, p. V006T08A043). American Society of Mechanical Engineers.
7. Trautman, N., Razban, A., & Chen, J. (2021). Overall chilled water system energy consumption modeling and optimization. Applied Energy, 299, 117166.
8. Ma, Z., & Wang, S. (2009). An optimal control strategy for complex building central chilled water systems for practical and real-time applications. Building and Environment, 44(6), 1188-1198.
9. Tirmizi, S. A., Gandhidasan, P., & Zubair, S. M. (2012). Performance analysis of a chilled water system with various pumping schemes. Applied energy, 100, 238-248.
10. Zhang, H., Shao, S., Xu, H., Zou, H., & Tian, C. (2014). Free cooling of data centers: A review. Renewable and sustainable energy reviews, 35, 171-182.
11. Raj, V. A. A., & Velraj, R. (2010). Review on free cooling of buildings using phase change materials. Renewable and Sustainable Energy Reviews, 14(9), 2819-2829.
12. Zhang, Y., Wei, Z., & Zhang, M. (2017). Free cooling technologies for data centers: energy saving mechanism and applications. Energy Procedia, 143, 410-415.
13. Brady, G. A., Kapur, N., Summers, J. L., & Thompson, H. M. (2013). A case study and critical assessment in calculating power usage effectiveness for a data centre. Energy Conversion and Management, 76, 155-161.
14. Zoie, R. C., Mihaela, R. D., & Alexandru, S. (2017, October). An analysis of the power usage effectiveness metric in data centers. In 2017 5th International symposium on electrical and electronics engineering (ISEEE) (pp. 1-6). IEEE.
15. Ni, J., Jin, B., Zhang, B., & Wang, X. (2017). Simulation of thermal distribution and airflow for efficient energy consumption in a small data center. Sustainability, 9(4), 664.
16. Lu, H., Zhang, Z., & Yang, L. (2018). A review on airflow distribution and management in data center. Energy and Buildings, 179, 264-277.
17. Capozzoli, A., & Primiceri, G. (2015). Cooling systems in data centers: state of art and emerging technologies. Energy Procedia, 83, 484-493.
18. Fulton, G. (2020). Data center temperature control: A systems approach. Springer.
19. Jones, N., & Patel, R. (2021). "Retrofitting legacy data centers: A case study of CRAC optimization." Journal of Sustainable Computing, \*8\*(3), 112–125. https://doi.org/10.1016/j.suscom.2021.100532
20. Zhang, H., Shao, S., & Xu, H. (2021). "Energy-efficient thermal management of data centers." Applied Energy, \*283\*, 116296. https://doi.org/10.1016/j.apenergy.2020.116296