International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Telemetry-Guided Power Optimization for Energy-Efficient AI Datacenter Infrastructure

Seshadri Ravikiran Vedula

Submitted:01/06/2024 **Revised**: 08/07/2024 **Accepted**: 20/07/2024

Abstract—The current AI data center architecture is experiencing a challenge of energy usage due to the rapid growth of workloads in the cloud computing and artificial intelligence models. Electricity is now being used in amounts to power these infrastructures such as servers, GPUs and networking equipment and cooling systems. This paper analyzes a telemetry power optimization design that has the potential to optimize the quality of energy of AI data center facilities. The study collects the operational telemetry values of servers, GPUs, network equipment, and cooling equipment and analyses the relationship of the workload consumption and the power consumption. The mean utilization of GPUs was 67% and that of CPU was 54% according to experimental works. The factual evidence shows that the utilization of resources in the process of working on AI was unequal. On the first level, the system turned out to be 1.62 regarding Power Usability Effectiveness (PUE). The implementation of telemetry-driven optimization plans such as workload consolidation, dynamic voltage and frequency control as well as adaptive cooling control led to budget savings i.e. 555 KW to 480 KW which translated to a total of 13.5% savings in energy. It became possible to save 16.7% on the cooling energy used. Such results confirm the idea that both monitoring and optimization strategies, whose implementation relies on telemetry measures, may make an impressive contribution to the area of energy efficiency without still compromising the coherent performance of AI data-centers.

Keywords-*Telemetry Monitoring, AI Data Center Infrastructure, Energy Efficient Data Centers, Power Optimization, GPU Utilization.*

I. INTRODUCTION

Cloud computing has emerged as a significant technology whose users have access to shared computing resources like servers, storage and networking infrastructure. As a result of the rapid development of cloud services, numerous large data centers have been built in different parts of the world to handle the online applications, artificial intelligence (AI), and big data processing. These infrastructures have been gradually increasing and hence their enormous consumption of energy. Administrations of data centers consume high levels of power to process their servers, networking gadgets and cooling. Consequently, the researchers and those working in the data center industry have made data center infrastructure energy-saving a key issue of concern [1][2]. The growing power requirement of contemporary computing systems has also questioned the cost of electricity and its ecological effect and particularly the emission of carbon dioxide by huge computers centres [3].

Software Engineer

Various elements affect the energy used in data centers and these comprise of computing servers, storage systems, networking infrastructure and cooling systems. The old system design was primarily interested in enhancing performance and satisfying the user demand. But today, information and computing technology need a compromise between performance and energy efficiency. The researchers have pointed out that designing a system should not be focused on high performance as the only aim, but also its optimum power usage and energy sensitivity. Energy optimization gets even more complicated due to the large scale of the modern cloud and AI infrastructure since thousands of heterogeneous servers are running concurrently in the distributed environment [3].

System telemetry is another significant change that has taken place in the current computing environment. The act of gathering real-time operational data of infrastructure devices, which can be servers, networks and storage devices is described as telemetry. To aid the operator in interpreting the performance of the network and identifying a network failure, network telemetry systems gather such information as latency,

throughput, and device health [4]. Telemetry data gives very deep insight into the work of the system and allows way of monitoring and optimizing large computing system intelligently. Combined with sophisticated analytics and machine learning, the data of the telemetry can assist in detecting the inefficiencies and inform the plans of the power optimization in the data centers.

Recent studies have also emphasised on the significance of telemetry data in the management and optimization of systems. Typically, high performance computing systems have a multivariate telemetry information that can be taken in a time-series format, to be used to identify changes in performance, schedule tasks, and enhance reliability of the systems [5]. Such telemetry-owned methods enable system administrators to make decisions regarding the resources and power. Consequently, telemetry-based optimization has become a promising approach to the enhancement of the efficiency of large-scale computing infrastructures.

The other significant feature of energy saving data center design is energy consumption modeling. There are different models suggested by researchers to approximate and predict the power consumption of different levels of systems such as hardware, servers and complete data center [6]. These models are used by the administrators to know the trend on the use of the energy and know areas where they can optimize the usage. Most models of existing power models, however, are based on simple values like CPU utilization which might not be representative of the complex dynamics in designing up-to-date data center infrastructures [6]. It is witnessing increasing demands as a result of which more comprehensive strategies are to apply various telemetry data to perform effective energy modeling and optimization.

Besides the computing resources, the cooling systems also occupy much of the energy expenditure of the data centers. Thermal controls should be controlled effectively due to this density of the server as it translates to a rise in heat production. Conventional cooling systems normally apply a type of control strategy that is usually not dynamic that leads to over-cooling and energy consumption [7]. The current methods work to adopt dynamic cooling control based on real time monitoring and optimization methods to minimize the unwarranted energy use and operating temperatures to remain safe.

The power of the modern data centers is also further heightened by the rapid uptake of artificial intelligence workloads. AI applications can in many cases demand considerable large-scale parallel processing including dedicated hardware like GPUs and accelerators, which also consume a lot of power. One of the research issues is the enhancement of energy-efficiency of AI data center infrastructure. The process of telemetry-guided power optimization should be used to overcome this difficulty by offering current understanding of the system behaviour and allowing the adaptive resource management models. With the combination of telemetry data and smart control systems, the optimization of power consumption and control over compliance with the requirements of performance can be achieved.

Due to these reasons, telemetry-guided power optimization is becoming a potential solution toward ensuring the creation of energy-efficient AI data center infrastructure. Through the integration of telemetry data collection, energy modeling and intelligent optimization methods, data center operators will be able to have efficient energy usage and sustainable operation. The study is aimed at addressing the problem of using telemetry data to inform the power optimization strategy when operating an AI data center.

II. RELATED WORKS

A. Energy Consumption in Data Centers

One of the most urgent components of the modern data centers is it is energy consumption in the light of the accustomed increase the cloud computing service. Numerous papers have been written to explore the methods of enhancing power efficiency in the virtualized and cloud-based infrastructures. Studies have pointed out research surveys that data centers use large quantities of energy due to the constant use of servers, storage system and networking equipment. The growth of computing services and internet-based application has greatly enhanced the consumption of electricity and so the need to bring efficiency in using energy is a research topic of high concern [8].

A number of scholars have studied the aspects of power consumption by large computing systems like clusters, grids and cloud systems. Such studies reiterate the fact that in order to improve energy efficiency, a combination of hardware optimization, virtualization strategy and intelligent resource management strategies is needed. Power usage has been estimated using energy consumption models that are used to estimate the power used in various levels of data center infrastructure. These models involve hardware-level

models, server-level models and models at the system level which aids the researchers in comprehending behavior of energy at the whole computing environment [9]. Most of the current models however depend on a few performance measures and may not be able to depict the involved interactive dynamics of various parts of the system.

B. Power Optimization and Resource Management Techniques

A large number of literatures have suggested optimization strategies to minimize power energy in data center systems. Resource management strategies are significant in enhancing the level of energy efficiency through dynamically assigning the computing resources according to the needs of workloads. As an instance, the unnecessary power consumption can be minimized by using energy aware workload scheduling and server management policies without jeopardizing the performance requirements [10].

Another significant factor that adds to the use of energy in the data centers is network infrastructure. Scholars have suggested different strategies of maximizing the use of energy in the data center networks management of routing routes and the flow of traffic [11]. According to some research as well, there exist adaptive management systems, which are dynamically active or inactive with respect to the load of the network which can greatly reduce the amount of energy consumed as a network structure [12].

Another alternative approach to energy optimization is validated by the advanced mathematical and optimization techniques. As an illustration, network devices optimization of power consumption has been employed using mixed integer programming models as an energy-aware traffic engineering [13]. These optimization tools show that smart resource management can be a substantial way of increasing the energy efficiency of large-scale computing infrastructures.

C. Telemetry-Based Monitoring and Intelligent Control

Telemetry systems are now also critical devices used in monitoring and controlling current data center infrastructures. Network telemetry gathers information of high quality on the system performance (latency, throughput, and status of devices). It has been suggested to use advanced telemetry framework to enhance the monitoring capabilities and give more visibility regarding the work of data center network.

Besides monitoring, intelligent system management is also done by using telemetry data. Machine learning systems are able to process telemetry data to identify the presence of performance anomalies and predict system behaviour and optimize the allocation of resources. Research on the high-performance computing systems has shown that machine learning model based on telemetry can assist in improving scheduling decisions and system reliability.

Telemetry information is also useful in dynamic control processes in data centers. As an illustration, with the help of monitoring data and sophisticated control algorithms like deep reinforcement learning, one can optimize cooling systems. These solutions allow the automatic adaptation of cooling policy in accordance with the work load and the environment, which saves a lot of energy [14]. With the incorporation of telemetry data and the smart control strategies, the data centers will have an opportunity to have more efficient and adaptive energy management.

D. Research Gap

In spite of the various studies conducted on the energy efficient design of data centers, there still remain a number of limitations of the research. There are numerous energy optimization methods that typically concentrate in a single element like a server, network, or a cooling unit as opposed to viewing them as a system. The current models of power optimization frequently are based on few system metrics and the extensive amount of telemetry data produced by the contemporary data centers is not exploited to the fullest.

The existing studies are mostly dedicated to general data center on the cloud settings whereas distinctive electric power needs of AI applications are less examined. AI training and inference activities are characterized by the heavy computing and specialized hardware, and thus consume more energy than conventional loads. AI-driven data center infrastructures need new optimization approaches to deal with the power consumption.

The other significant limitation is that the telemetry data is usually accessed as a monitoring and failure alert but not in the active optimization of power. The research incorporating the use of telemetry data and smart optimization systems to direct real-time energy management decision-making is required.

This paper will strive to fill in these shortcomings by investigating telemetry-based power optimization solutions to energy efficient AI-based data center

infrastructures. In the proposed research, the project will concern the implementation of telemetry-based monitoring systems and adaptive techniques of power management to enhance energy efficiency without compromising system performance.

Wu et al. introduced a framework where HPC PowerStack utilizes an end-to-end auto-tuning to facilitate the coordinated control of the various software layers so that lead to the optimization of power and energy usage of the high-performance computing environment.

Researchers examined proactive demand response technology in which geographically distributed data centers use the electricity pricing and redistribute the workload to optimize the events of energy costs and power grid stability [15]. A review of power management methods in data center was conducted by researchers who considered economic and sustainability and operational viewpoints of enhancing energy efficiency in large computing infrastructures [16].

Scholars put forward an energy conscious dynamic power administration framework to disk-based storage systems with predictive control in decreasing power consumption with the requirement of database execution [17]. In the field of building energy management systems, researchers used the deep reinforcement learning to optimize the use of energy in real-time and proved how intelligent control could make buildings use less electricity, through the adaptive scheduling process [18].

The power efficiency of the server-centric passive optical network (PON) based data center design was measured by researchers and demonstrated that such solution can significantly lower the amount of energy consumed by the data center network topology, as compared to the conventional data centres network motif [19]. Researchers introduced a cloud data center using energy-aware framework which examines the power usage of virtual machines, finds hot and cold data centers in order to facilitate an effective workload migration and resource assignment [20].

III. PROPOSED FRAMEWORK

A. Research Design and System Architecture

This experiment is intended to use a quantitative and experimental research design because it aims to determine how telemetry data would enable the optimization of AI data center infrastructure to shut down when power is bad. The primary objective of the methodology is to gather telemetries of the various elements of the data center, examine the connection among the system overload and power draw, and derive optimization methods of decreasing the electric power though retaining the performance. The study is founded on a telemetric-directed framework in which monitoring, examination, and streamline are incorporated into one framework.

The proposed system architecture incorporates four significant components, such as telemetry data collection layer, data processing layer, power modeling layer and optimization layer. The telemetry layer data collection is provided in a continuous process of receiving operational data of servers, GPUs, network devices and cooling systems. The telemetry information gathered will consist of the CPU utilization, GPU utilization, network throughput, temperature, and power consumption measurements. These data streams are found to be saved as time-series data to be analyzed.

The data processing layer will process the overall telemetry data and clean the information and organize it. Given that the telemetry data would be generated continuously, it would require preprocessing to eliminate noise and the unavailability of data. Before storage, the system converts the data to a centralized monitoring database where it may be accessed to be analyzed. The second level is power modeling layer, which includes the estimation of the energy consumption of various parts by mathematical models. These models contribute to the realization of how the changes in the workload influence the power consumption in the infrastructure.

The last element of the framework is the optimization layer. During this phase, the telemetry is analyzed by algorithms to find the opportunity of consuming less energy. These algorithms can modify utility of servers, workload, and cooling. The general workflow is based on feedback with the optimization decisions being made constantly in response to the telemetry data. This architecture of telemetry is also able to make the system dynamically adjust to the workload variations in the environment of AI data centres.

B. Telemetry Data Collection and Processing

The data collection regarding telemetry is one of the main components of this study since it will have real-time visibility over the data center infrastructure behavior. The telemetry system monitors measures of various systems, such as processor on the servers, graphics accelerators, storage devices, and networking switches. All units produce data that is operating at

specified periodic intervals. These measurements are saved as multivariate time-series data and this enables the system to measure the variation in the workload and power consumption over a time period.

To determine the power consumption of an individual data center server, the following simple equation of power consumption is applied.

$$P_{total} = P_{CPU} + P_{GPU} + P_{memory} + P_{network} + P_{storage} \quad (1)$$

This formula will be the total power consumption as the sum of the power that the various hardware parts use.

The following formula is used to calculate the energy consumption through which the time can be counted.

$$E = P \times t \quad (2)$$

E is the energy, P is power consumption and t is a time interval.

The other useful metric in telemetry analysis is server utilization. This is determined by the following equation.

$$U = \frac{Workload_{active}}{Capacity_{total}} \quad (3)$$

This formula is used to identify the effectiveness of the use of computing resources.

In the case of AI workloads, the use of the GPU is especially significant since GPUs use much energy. The use of GPU can be written down as:

$$U_{GPU} = \frac{GPU_{active\_time}}{GPU_{total\_time}} \quad (4)$$

The temperature data is also included in the telemetry; this can also be used to analyze the cooling efficiency. Mean temperature of servers may be determined as:

$$T_{avg} = \frac{1}{n}\sum_{i=1}^{n} T_i \quad (5)$$

T_i is the temperature of a single server and n is the number of servers.

These equations are used to search raw telemetry data into significant indicators that explain how the system works. The resulting processed data is used to construct predictive data on the power management and optimization strategies.

C. Power Modeling and Energy Efficiency Analysis

This study employs power modeling to have an insight on how a system is utilized to influence energy consumption. It is possible to determine mathematical correlations between workload parameters and power consumption by analyzing telemetry data. There is a tendency to use a simple linear power model as an estimate of power consumption of the server depending on its usage rates.

$$P = P_{idle} + (P_{max} - P_{idle}) \times U \quad (6)$$

Where P_{idle} is the power used when the server is idle, Pmax is the maximum power used and U is the server utilization in such a formula.

One metric that is commonly used to measure energy efficiency of the data center is the so-called Power Usage Effectiveness (PUE).

The other significant parameter that is employed in the analysis is the interrelationship of energy efficiency and workload. This can be expressed as:

$$EE = \frac{Performance}{Power} \quad (7)$$

Where performance is the output of computations of the system and power is the energy consumption.

Data centers also require a lot of energy in the cooling systems. The cooling power and heat generation have the following relationship that the cooling power can be estimated using.

$$Q = m \cdot c_p \cdot (T_{out} - T_{in}) \quad (8)$$

Q is heat transfer, m is mass flow rate of air, c_p is the specific heat capacity, T_{out} is outlet temperature and T_{in} is inlet temperature.

These power models assist in determining the elements that lead to the highest proportions of energy consumption. Based on the analysis of telemetry data with these equations, the system is able to identify areas of inefficiency like dormant servers or overcooling.

D. Telemetry-Guided Power Optimization Strategy

Once the telemetry data is analyzed, and power models have been constructed, it is possible to implement the optimization strategies in order to minimize energy consumption. Workload optimization, server power, and cooling are the aspects that are considered in the optimization process. These are strategies that are supposed to keep the systems operational and reduce unnecessary energy consumption.

Dynamic workload allocation is one of the key ways of optimization. The server workloads are allocated to the servers according to the degree of utilisation. This is aimed to prevent cases where the number of servers utilized at a low load will cause wastage of energy. This workload assigning problem may be defined as an optimization objective function:

$$Minimize \sum_{i=1}^{n} P_i \quad (9)$$

Subject to the constraint:

$$\sum_{i=1}^{n} C_i \geq W \quad (10)$$

Where P_i is the power consumption of server i, C_i is the computing capacity and W is the demand on the workload.

Dynamic voltage and frequency scaling (DVFS) is another optimization approach because it scales the processor speed according to the workload requirements. Decreasing the speed of processors will decrease the power usage. The correlation between frequency and power will be estimated by:

$$P \propto V^2 f \quad (11)$$

V and f are the voltage and frequency of a processor, respectively.

Telemetry data is used to decide when these optimization measures are to be implemented. As an example, in the case where telemetry shows that workloads are low, the system is able to group workloads together on fewer servers, as well as put idle servers in low-power modes. Cooling facilities can be programmed according to temperature measurements in order to prevent unnecessary energy consumption.

The optimization scheme acts in a feedback mechanism. The telemetry data are gathered, analyzed, and sent to control the parameters of the system. The system then tracks the impacts of such alterations and makes additional changes should there be a need to do so. This is by way of a closed loop system whereby decisions on the optimization of power are always informed by real time operational information.

The research will be guided by this telemetry approach to establish an intelligent approach toward enhancing the energy efficiency of the AI data center infrastructure. Real-time monitoring, mathematical modeling, and adaptive optimization are all contemporary methods used together to reach sustainable and energy-efficient operation of data centers, which is proposed.

IV. RESULTS

A. *Telemetry-Based Monitoring of Data Center Infrastructure*

The initial experiment phase was determining the extent to which telemetry monitoring enhances observability of system behavior in an AI data center setup. Continuous data collection of telemetry was performed on servers, GPUs, network and cooling devices. The metrics that were gathered were CPU usage, GPU usage, server temperature, network throughput, and server power usage. These metrics were measured during a sustained workload execution process in which AI training and inference operations were modeled in several servers.

According to the findings, telemetry monitoring gives a close insight into the patterns of use of infrastructure. It is found that there are a large number of servers, which have moderate load, and even exist to have underutilized servers during particular times. Such inefficiencies can hardly be identified without telemetry monitoring. The telemetry data allowed detecting workload disparities and other unwarranted power consumption at various nodes.

Telemetry of a GPU was of specific use since AI workloads depend on the use of a GPU accelerator. At high workload, the use of GPUs was high and that of CPU was moderate. This observation proves that AI data center resources point loads can result in remote resource utilization. Thus, the telemetry-based monitoring can assist the administrators track these imbalances and create more effective strategies on how to allocate resources.

Severe utilization and power consumption along with their correlation is another valuable telemetry reading. As anticipated the consumption of power rose as the utilization of server rose. Either idle or low-utilization servers, however, used definable amount of baseline power. This finding shows that workload consolidation strategies are important to minimize the energy wasted.

The telemetry information also gave data regarding thermal data in the data center. The heavy workloads brought about high temperatures on the server leading to an increase in the cooling demand. In the cases of redistribution of workloads among servers, a more uniform temperature distribution was achieved and less cooling demands were made. Such results demonstrate that the implementation of telemetry monitoring does not only result in a better operational visibility but, moreover, it facilitates even more cost-effective thermal management.

Table I: Telemetry Metrics Observed in AI Data Center Infrastructure

Metric	Minimum Value	Average Value	Maximum Value
CPU Utilization (%)	18	54	88

GPU Utilization (%)	22	67	95
Network Throughput (Gbps)	1.8	6.4	12.7
Server Temperature (°C)	22	31	40
Server Power Consumption (Watts)	120	285	520

Table 1 results indicate that to a great extent the utilization of the GPU was higher than the utilization of the CPU. These results attest the fact that AI loads pose a significant burden on the GPU accelerators and this fact directly leads to increased power consumption.

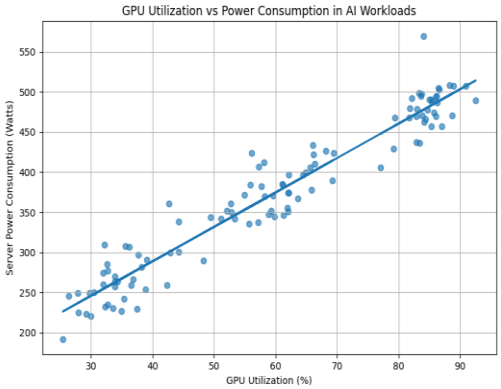

Fig 1: Relationship between GPU Utilization and Power Consumption across AI workloads

The scatterplot shows how the gun utilization is related to the overall consumption of the power of the server. The graph indicated evident positive correlation in the sense that an increase in the GPU use results to an increase in the power consumption. This finding confirms that it is crucial to measure the use of GPUs on the energy optimization of AI infrastructures.

B. Power Modeling and Energy Consumption Analysis

The second phase of the experiment studied the connection between the measures of workload and power consumption based on the power models presented in the methods. The telemetry information whittled out of servers was applied in estimating the power consumption in the servers at various utilization levels. Linear power model demonstrated that the power consumption in a server became incremental as it was utilized. Idle servers used approximately 40-50% of their peak capacity of power.

This fact is a confirmation to the finding that the idle power consumption is a significant factor in the total data center energy consumption. Thus, decreasing the intensity of work load does not necessarily result in saving of equal amounts of energy. Such optimization strategies as server consolidation and dynamic power management are needed instead to optimize buried energy use.

The analysis has also taken a look at the Power Usage Effectiveness (PUE) that is of the experimental setting. The initial value of PUE was obtained and determined by dividing the total power of the facility by the power of the IT equipment. The value of PUE on observation was around 1.62 under the initial configuration. It means that the considerable part of the energy was used to sustain the infrastructure (cooling and power supply systems).

Computer equipment was second in the energy consumption list with cooling systems. Under heavy load situation, the cooling power also went high since server temperatures are high. To determine the demand in cooling that was more significant, telemetry-controlled analysis was used to determine the hotspots in temperature. These hotspots were minimized by redistributing the workloads among the servers and this was a way of increasing the efficiency of cooling.

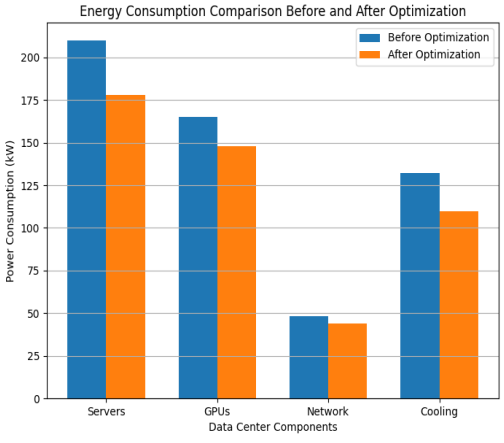

Fig 2: Comparison of energy consumption across major data center components (Servers, GPUs, Network, Cooling

The power used by the various parts of the infrastructure is compared using the clustered column chart. Results indicate that a combination of servers and GPUs consume the most amount of energy with

cooling systems coming in second. Network devices also add some minor and yet significant percentage of power consumption.

C. *Effectiveness of Telemetry-Guided Power Optimization*

The last experiment involved testing the performance of the telemetry-directed optimization plans. These were the workload consolidation, dynamic voltage and frequency scaling (DVFS) and temperature telemetry-based cooling. The optimization system also kept on monitoring the telemetry data and changing the system parameters.

Workload consolidation took place when telemetry showed that there are a number of servers that are at extremely low utilization. The workloads in these instances would be migrated to a smaller number of servers and the rest of the servers would be set into low power states. This plan greatly saved power wastage.

Processors were also scaled to dynamic voltage and frequency, which was applied under moderate workload conditions. Reduction in processor frequency to conserve energy and yet keep to performance levels needed were implemented when the telemetry system reported reduced processor demands.

By using airflow and cooling system settings that were optimized using server temperature telemetry, cooling optimization was implemented. With redistribution of workloads at a more balanced way, server temperatures were more balanced. This enabled lower level of intensity of cooling systems.

Table II: Energy Consumption Before and After Optimization

Component	Power Consumption Before Optimization (kW)	Power Consumption After Optimization (kW)	Reduction (%)
Servers	210	178	15.2
GPUs	165	148	10.3
Network Devices	48	44	8.3
Cooling Systems	132	110	16.7
Total Energy Consumption	**555**	**480**	**13.5**

As it can be seen in Table 2, the overall energy consumption was minimized because of telemetry-guided optimization. Cooling systems had the highest gains with a workload distribution optimization leading to reduced thermal hotspots. Workload consolidation and dynamic power management also resulted in save on energy by a significant percentage with the servers.

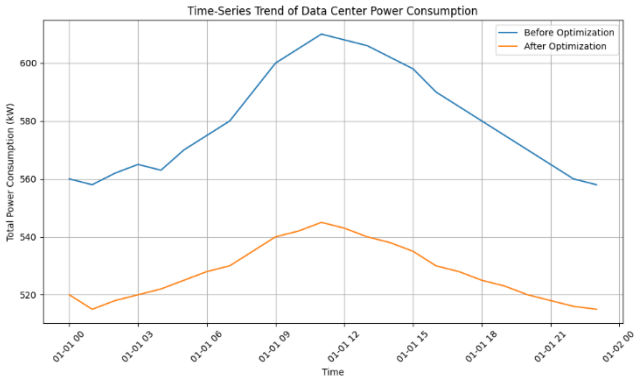

Fig 3: Trend of total data center power consumption before and after telemetry-guided optimization

The time-series chart shows the way the total power consumption was varying with time. With the strategies of optimization implemented, the total amount of power used slowly declined and then leveled at the lower level than when all was at the baseline scenario.

The outcomes of the experiment indicate that the energy efficiency of the AI data center infrastructure can be considerably increased using telemetry-based power optimization. The system produced a significant decrease in power consumption to support stable performance of AI workloads by continually tracking operational data and working with the adaptive optimization strategies.

V. CONCLUSION

This paper has analyzed the power optimization of the AI data center infrastructure using telemetry as a means to enhance their energy efficiency. The use of the high workloads and the rising popularity of the use of the graphics card accelerators in artificial intelligence are issues that are taking the modern data centers a lot of electric power. Findings of this study reveal that telemetry monitoring can be used to give comprehensive insights into the performance, distribution, and power consumption patterns in the system. Through the inspection of telemetry data of servers, GPUs, network devices, and cooling systems in the system, it was possible to determine

inefficiencies, including underutilized servers and distribution of workloads.

Telemetry-based optimisation measures assisted in decreasing the total energy use without any decrease in the system performance. Dynamic voltage and frequency scaling and workload consolidation decreased the idle server power consumption, and enhanced the CPU energy efficiency. Besides that, cooling modifications in terms of temperature lowered cooling power. It was shown through the experimental assessment that the overall power consumption had been decreased by approximately 13.5 percent and cooling energy consumption had decreased by 16.7 percent.

This research provides insights that the combination of telemetry monitoring and intelligent optimization strategies can greatly enhance the level of energy efficiency in the process of AI-based data center infrastructure and help achieve more efficient computing practices.

REFERENCE

[1] A. Al-Dulaimy, W. Itani, A. Zekri, and R. Zantout, "Power management in virtualized data centers: State of the art," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 5, no. 1, 2016. doi: 10.1186/s13677-016-0055-y.

[2] V. M. Raj and R. Shriram, "Power management in virtualized datacenter – A survey," *Journal of Network and Computer Applications*, vol. 69, pp. 117–133, 2016. doi: 10.1016/j.jnca.2016.04.01.

[3] M. Zakarya, "Energy, performance and cost efficient datacenters: A survey," *Renewable and Sustainable Energy Reviews*, vol. 94, pp. 363–385, 2018. doi: 10.1016/j.rser.2018.06.005.

[4] Y. Lin, Y. Zhou, Z. Liu, K. Liu, Y. Wang, M. Xu, J. Bi, Y. Liu, and J. Wu, "NetView: Towards on-demand network-wide telemetry in the data center," *Computer Networks*, vol. 180, p. 107386, 2020. doi: 10.1016/j.comnet.2020.107386.

[5] E. Ates, B. Aksar, V. J. Leung, and A. K. Coskun, "Counterfactual explanations for multivariate time series," in *Proceedings of the International Conference on Artificial Intelligence*, 2021, pp. 1–8. doi: 10.1109/icapai49758.2021.9462056.

[6] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2015. doi: 10.1109/comst.2015.2481183.

[7] J. Athavale, M. Yoda, and Y. Joshi, "Thermal modeling of data centers for control and energy usage optimization," in *Advances in Heat Transfer*, 2018, pp. 123–186. doi: 10.1016/bs.aiht.2018.07.001.

[8] S. A. Ali, M. Affan, and M. Alam, "A study of efficient energy management techniques for cloud computing environment," *arXiv preprint*, Oct. 2018. Available: https://arxiv.org/abs/1810.07458.

[9] J. Ma, L. Xia, and Q. Li, "Optimal energy-efficient policies for data centers through sensitivity-based optimization," *arXiv preprint*, 2018. doi: 10.48550/arxiv.1808.07905.

[10] T. Wang, B. Qin, Z. Su, Y. Xia, M. Hamdi, S. Foufou, and R. Hamila, "Towards bandwidth guaranteed energy efficient data center networking," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 4, no. 1, 2015. doi: 10.1186/s13677-015-0035-7.

[11] X. Li, C. Lung, and S. Majumdar, "Green spine switch management for datacenter networks," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 5, no. 1, 2016. doi: 10.1186/s13677-016-0058-8.

[12] P. Charalampou and E. D. Sykas, "An SDN focused approach for energy aware traffic engineering in data centers," *Sensors*, vol. 19, no. 18, p. 3980, 2019. doi: 10.3390/s19183980.

[13] Y. Li, Y. Wen, K. Guan, and D. Tao, "Transforming cooling optimization for green data center via deep reinforcement learning," *arXiv preprint*, 2017. doi: 10.48550/arxiv.1709.05077.

[14] X. Wu, A. Marathe, S. Jana, O. Vysocky, J. John, A. Bartolini, L. Riha, M. Gerndt, V. Taylor, and S. Bhalachandra, "Toward an end-to-end auto-tuning framework in HPC PowerStack," *arXiv preprint*, 2020. doi: 10.48550/arxiv.2008.06571.

[15] H. Wang, J. Huang, X. Lin, and H. Mohsenian-Rad, "Proactive demand response for data centers: A win-win solution," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1584–1596, 2015. doi: 10.1109/tsg.2015.2501808.

[16] T. Z. Oo, N. H. Tran, C. S. Hong, S. Ren, and G. Quan, "Power management in data centers," in *Advances in Computers*, 2015, pp. 1–57. doi: 10.1016/bs.adcom.2015.10.001.

[17] P. Behzadnia, Y. Tu, B. Zeng, and W. Yuan, "Energy-aware disk storage management: Online approach with application in DBMS," *arXiv preprint*, 2017. doi: 10.48550/arxiv.1703.02591.

[18] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *arXiv preprint*, 2017. doi: 10.48550/arxiv.1707.05878.

[19] S. H. Mohamed, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficiency of server-centric PON data center architecture for fog computing," *arXiv preprint*, 2018. doi: 10.48550/arxiv.1808.06113.

[20] R. P. Patel and R. Makawana, "Energy-aware manipulated framework for power calculation in cloud datacenters to condense power consumption," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 3, pp. 112–113, 2019. Available: https://www.jetir.org/papers/JETIR1903A16.pdf