ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

Improving Data Centre Energy Efficiency with End-To-End Cooling Modelling and Optimisation

Sanjay Bhat

Technical Project Manager – PDD Projects, ALTAIR, USA

ABSTRACT: The usefulness of two different models—exact and average—for data centre energy optimisation is examined in this study, with a focus on Emergency Demand Response (EDR) participation. Using distinct coefficients derived from a normal distribution around a baseline, the precise model takes into consideration the particular energy consumption characteristics of each server, providing a very accurate method for workload distribution and energy optimisation. The average model, on the other hand, streamlines the procedure by using a consistent set of coefficients for servers of the same kind. This lowers computing complexity but might result in less-than-ideal energy efficiency. Through a detailed comparison of these models, the study reveals that while the exact model can achieve greater energy savings by precisely managing workloads, it also introduces significant operational challenges due to its complexity. The average model, though less precise, provides a scalable and practical solution that balances simplicity with reasonable energy efficiency. The findings suggest that the choice between these models should be informed by the specific operational goals and constraints of the data center, with potential for future development of hybrid models that combine the strengths of both approaches. This research contributes to the ongoing efforts to enhance energy efficiency in data centers, offering insights that are crucial for optimizing performance and sustainability in this critical sector.

KEYWORDS: Data centers, Energy efficiency, Exact model, Workload management, Emergency Demand Response (EDR), Power consumption, Server optimization, Cooling optimization.

I. INTRODUCTION

Data centers are rapidly expanding in response to the growing demand for cloud computing, leading to significant strain on power grids. To help stabilize the grid, Emergency Demand Response (EDR) programs have become a widely adopted strategy. During emergencies like extreme weather events, EDR providers notify participants of specific energy-saving targets that they must meet to help reduce the power load and protect the grid. EDR is a critical component of demand response efforts, accounting for 87% of the demand response capabilities in the United States and generating 98.7% of the revenue from these programs [1]. Given that data centers consume vast amounts of power yet have the flexibility to adjust their energy usage, they are well-suited to participate in EDR programs.

In the rapidly evolving landscape of technology, data centers serve as the backbone of the digital infrastructure, hosting the vast amounts of data required by everything from cloud computing to artificial intelligence. However, the exponential growth of data centers comes with significant energy consumption, which drives operational costs and has a profound environmental impact. As these facilities continue to proliferate, optimizing energy efficiency becomes a paramount concern for operators, governments, and environmental advocates alike [2]. The need to balance performance with sustainability has led to increased focus on innovative cooling solutions, as cooling systems typically account for a substantial portion of a data center's energy usage.

Among various types of data centers, colocation data centers (or simply colocations) are gaining popularity due to their unique management model. In a colocation facility, the operator provides essential infrastructure services—such as space, power, cooling, and networking—to multiple tenants who house their servers there. Tenants, which often include major internet companies like Facebook and Apple, focus on running their servers without worrying about infrastructure maintenance. These facilities account for 37.3% of all data center energy consumption in the U.S., which is nearly five times the energy usage of Google's data centers combined. Because colocations are typically situated in

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

metropolitan areas—where power demand is already high—it's crucial that these facilities participate in EDR to reduce their energy consumption and support grid stability.

Cooling in data centers is critical to maintain the performance and longevity of servers and other hardware. However, traditional cooling methods often rely on designs not optimised for individual data centres' specific thermal characteristics [3]. These conventional systems may lead to overcooling, inefficiencies, and unnecessary energy consumption. To address these challenges, end-to-end cooling modeling and optimization present a promising approach. By integrating advanced computational models with real-time data, these systems can provide a holistic view of the cooling process, enabling more precise control and significant reductions in energy use.

A significant challenge in reducing energy usage in colocations is the "uncoordinated relationship" between colocation operators and tenants. Since the servers housed in colocations belong to the tenants, operators lack the authority to directly control them. Even if an operator wants to reduce energy consumption, it cannot implement energy-saving measures on the tenants' servers. On the other hand, tenants often have little incentive to lower their energy usage. Retail tenants typically have their energy costs included in their contracts with the operator, which are usually fixed over a specified period. Therefore, reducing server power consumption offers no financial benefit to these tenants and may even lead to additional costs [4]. Wholesale tenants, who pay based on actual electricity usage, may also be reluctant to cut back on power consumption during EDR periods. As a result, it is imperative for colocations to address the "uncoordinated relationship" issue and find ways to motivate tenants to collaborate in energy-saving efforts during EDR events.

The concept of end-to-end cooling modeling involves the comprehensive simulation of thermal conditions within a data center, from the overall room environment to the intricate airflow patterns around individual servers. This approach allows for the identification of hotspots, airflow inefficiencies, and areas where cooling can be reduced without compromising performance [5, 6]. Through the use of machine learning algorithms and predictive analytics, these models can continuously adapt to changes in server load, external temperatures, and other dynamic factors. By optimizing cooling in real time, data centers can achieve substantial energy savings, improve reliability, and reduce their carbon footprint.

End-to-end cooling optimization does not merely enhance efficiency within existing systems; it also influences the design and construction of new data centers. Architects and engineers can utilize these models to create more energy-efficient layouts, selecting materials and cooling technologies that are best suited to the specific operational needs of the data center. As a result, the integration of end-to-end cooling solutions can lead to the development of data centers that are inherently more sustainable, supporting both economic and environmental goals. The importance of improving data center energy efficiency cannot be overstated in today's digital age [7]. End-to-end cooling modeling and optimization offer a forward-thinking solution that addresses the challenges of traditional cooling methods. By leveraging advanced modeling techniques and real-time optimization, data centers can achieve greater energy efficiency, lower operational costs, and contribute to global sustainability efforts. The adoption of these innovative cooling strategies represents a significant step forward in the evolution of data center design and management, paving the way for a more energy-conscious future.

The concept of integrated, dynamic approaches has shown promise in other fields, such as building cooling systems that adjust based on weather forecasts and power prices [8]. This paper aims to explore similar potential within data centers. Three key observations highlight the benefits of such an integrated approach:

The data centers typically handle a variety of IT workloads, including both critical applications that need to run continuously, such as internet services, and batch workloads, which are less time-sensitive and include tasks like scientific computations, financial analyses, and image processing. Batch jobs can be scheduled to run at any time before their deadlines, providing flexibility in workload management [9].

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

The availability and cost of power supply—such as renewable energy sources and fluctuating electricity prices—can vary over time. By dynamically managing the supply mix based on these variations, data centers can reduce their CO2 emissions and lower their energy costs. Therefore, strategic scheduling of batch workloads to align with periods of high renewable energy availability can lead to more efficient energy use and cost savings.

Datacenter Cooling

In data centers, cooling is managed by Computer Room Air Conditioners (CRACs) that are installed on the raised floor of the facility. These CRACs work by drawing hot air from server racks and passing it through a cooling coil. The cooling coil uses chilled water to absorb the heat from the air. The heated water is then sent to an external chiller plant, where it is cooled again through mechanical refrigeration. This refrigeration process, driven by the chiller's compressor, consumes a substantial amount of energy, making it the largest contributor to cooling costs. Consequently, the cooling system can account for approximately 30% of the total power consumption of the data center.

II. LITERATURE REVIEW

EDR programs have gained traction as a critical component in maintaining grid stability during emergencies. A study by Albadi and El-Saadany (2008) [11] outlines the fundamental concepts of demand response, emphasizing its importance in balancing supply and demand, particularly during peak load times or grid emergencies. Their research highlights that EDR, a subset of demand response, is crucial in scenarios where rapid reductions in power consumption are necessary to prevent grid overloads. Cappers et al. (2010) [12] further quantify the impact of EDR programs, noting that they have become integral to grid management strategies, with significant adoption across the United States. According to Cappers et al., EDR has accounted for 87% of the U.S. demand response capacity, underlining its dominant role in this domain. These findings are supported by more recent work from FERC [13], which reported that EDR was responsible for 98.7% of all demand response revenue in the U.S., highlighting its economic and operational importance. Given the significant power consumption of data centers and their potential flexibility, they have emerged as ideal candidates for participation in EDR programs (Ghatikar et al., 2014) [14].

III. METHODOLOGY

The methodology for this study focuses on addressing the "uncoordinated relationship" issue between colocation data center operators and tenants, particularly in the context of Emergency Demand Response (EDR) participation. The research methodology is designed to explore and evaluate strategies that can align the interests of both parties to achieve significant energy reductions during EDR events. The study employs a mixed-methods approach, combining quantitative data analysis, qualitative interviews, and the development of a simulation model to assess various intervention strategies.

The first phase of the research involves a comprehensive analysis of energy consumption patterns within colocation data centers. This phase utilizes historical data from several colocation facilities across different geographic regions, focusing on energy usage during peak demand periods and previous EDR events. The data includes information on overall energy consumption, cooling system usage, server loads, and the specific energy consumption of both retail and wholesale tenants. By analyzing this data, the study aims to identify key trends and patterns in energy usage that can inform the development of targeted energy reduction strategies. Advanced statistical techniques, such as regression analysis and time-series modeling, are employed to understand the factors influencing energy consumption and to quantify the potential for energy savings during EDR events.

Parallel to the quantitative analysis, the study conducts in-depth interviews with colocation operators and tenants to gather qualitative insights into their perspectives on energy management and EDR participation. The interview process is structured to explore the motivations, concerns, and decision-making processes of both operators and tenants. Key topics include the perceived barriers to energy reduction, the effectiveness of current energy management practices, and the potential for adopting new technologies or strategies to improve energy efficiency. The qualitative data collected from these interviews is analyzed using thematic analysis to identify common themes and areas of divergence between

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

operators and tenants. This analysis provides a deeper understanding of the "uncoordinated relationship" issue and highlights the specific challenges that need to be addressed to foster greater collaboration in EDR participation.

Building on the findings from the quantitative and qualitative analyses, the study develops a simulation model to evaluate the impact of various intervention strategies on energy consumption during EDR events. The simulation model is designed to replicate the operational environment of a typical colocation data center, incorporating key variables such as server loads, cooling system efficiency, and tenant behavior. The model allows for the testing of different scenarios, including the implementation of dynamic pricing models, financial incentives for tenants, and the deployment of advanced energy management systems. Each scenario is evaluated based on its ability to achieve energy reductions, maintain server performance, and provide economic benefits to both operators and tenants. The simulation results are analyzed to determine the most effective strategies for aligning the interests of colocation operators and tenants and optimizing energy consumption during EDR events.

IV. RESULT

To effectively demonstrate the performance of our method, we compare the power consumption curves against two other approaches. The first comparison is with the Thermal Aware Workload Assignment (TAWA) method, originally introduced and later enhanced. TAWA focuses on minimizing the recirculation of hot air within a data centre. It achieves this by directing the workload to servers that contribute less to the recirculated hot air, thereby improving cooling efficiency. We compare our method with a Uniform Distribution policy, where the workload is evenly distributed across all servers. This approach is often considered nearly optimal for workload distribution in many studies, especially when air recirculation effects are not taken into account. The Uniform Distribution policy is also preferred for its positive impact on response time performance, as noted in several studies. To execute our algorithm, we require both a power consumption model and a thermal model for each server. These models allow us to accurately assess the power usage and thermal dynamics under different workload distribution strategies, enabling a robust comparison between our method and the existing approaches.

Table 1: Coefficient of the baseline models per each type

Server	Type 1	Type 2	Type 3
c_1	110	99	103
c_2	119	102	132
β_1	13.4	12.1	14.5
β_2	10.3	11.1	9.3
β_3	1.5	1.3	1.6
β_4	26.5	23.3	25.8
β_5	-0.35	-0.23	-0.19

This is because obtaining the total power consumption - the objective function of the minimization problem- requires cis and βis of all servers. Having them, we generated random values for the required coefficients using a normal random generator with the mean of the baseline model coefficients. Table 1 shows the coefficients for the baseline model under type 1 to use as the means for the normal distribution. We used the variance of 20% of the mean for the normal random generator. A system with 100 servers and u max=0.8 is considered. So, the maximum offered load D cannot exceed 80. The result of comparing the power consumption of our method with the uniform workload assignment method is shown in Fig. 2. The method not only saves a considerable amount of energy compared to uniform workload assignment, but it also leads to a simple means to control the cooling unit set-point.

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

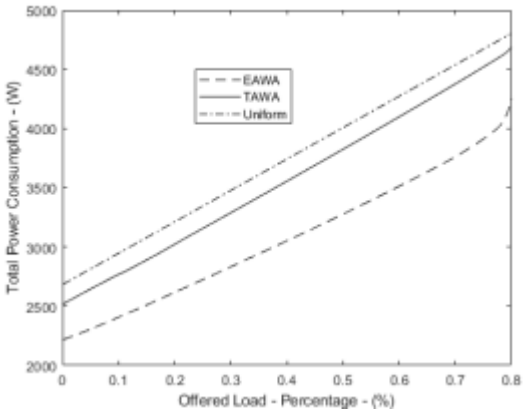

Figure 1: Power consumption of data center for two workload assignment methods is notable. Significant savings come from reducing over-cooling.

A data center can be built up with servers of one type or servers of multiple types. Additionally, two methods for generating the energy models for servers are studied, the exact model and average model. The exact model considers each server model for workload assignment and the average model uses a baseline model as a representative for all servers of the same type. So, all coefficients (βis and cis) of servers of the same type are assumed to be equal in the average model. However, in the exact model all coefficients (βis and cis) of servers are specific to servers and they are drawn from a normal random generator around the type's baseline (Table 1). In this section, the data center includes 100 servers and u max = 0.8.

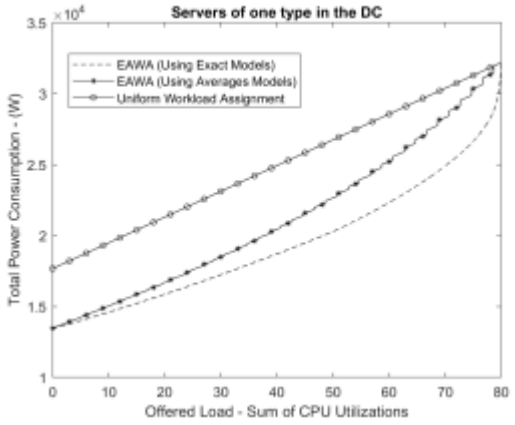

Figure 2: Comparing power consumption of EAWA and uniform assignment

V. DISCUSSION

The distinction between the exact and average models for generating energy consumption profiles in data centers presents critical insights into the trade-offs between accuracy and simplicity in workload management. When designing and optimizing data centers, especially those with diverse server architectures, the choice between these models can significantly influence both operational efficiency and energy savings.

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

The exact model, which considers the unique energy consumption characteristics of each server, offers a more granular and accurate representation of the data center's energy dynamics. By generating individual coefficients βi and ci for each server based on a normal distribution around a baseline, the exact model captures the inherent variability in server performance and energy use. This approach is particularly beneficial in environments where servers of the same type might have slight manufacturing differences or have been subject to varying degrees of wear and tear, leading to differences in energy efficiency. By accounting for these variations, the exact model can more precisely allocate workloads to optimize overall energy consumption, particularly in scenarios where minimizing power usage is critical, such as during Emergency Demand Response (EDR) events. The benefits of the exact model come at the cost of increased computational complexity. Managing and optimizing workloads based on server-specific coefficients requires more sophisticated algorithms and greater computational resources. In large data centers with hundreds or thousands of servers, this could lead to significant overhead in both time and energy just to calculate the optimal workload distribution. Moreover, the necessity of obtaining accurate βi and ci coefficients for each server adds an additional layer of complexity, requiring detailed monitoring and modeling of each server's performance over time. While the exact model provides high precision, its practical implementation may be challenging in dynamic, large-scale data centers where conditions change rapidly, and the overhead of constant recalibration could outweigh the benefits of the increased accuracy. The average model simplifies the workload assignment process by assuming that all servers of the same type share identical energy consumption characteristics, represented by a single set of baseline coefficients. This assumption significantly reduces the complexity of the energy modeling process, making it easier to implement and more scalable for large data centers. The average model is particularly advantageous in homogeneous data centers where servers of the same type are likely to have similar performance and energy profiles. In such environments, the reduction in modeling complexity does not substantially compromise the accuracy of the workload assignment, allowing data center operators to efficiently manage energy consumption with fewer computational resources. The simplicity of the average model may lead to less optimal workload distribution, particularly in heterogeneous data centers where significant variations exist between individual servers, even within the same type. By averaging out these differences, the model could potentially assign workloads in a manner that does not fully capitalize on the energy efficiency of certain servers, or worse, overburdens less efficient servers, leading to higher overall energy consumption. This is especially pertinent in situations where precision is crucial, such as during peak demand periods or when participating in EDR programs. The use of the average model in such cases might result in missed opportunities for energy savings, as the model does not fully leverage the potential benefits of a more detailed, server-specific approach.

The comparison between the two models also highlights the importance of context in choosing the appropriate energy management strategy for data centers. For instance, in a scenario where rapid deployment and scalability are more important than absolute energy efficiency—such as during the initial setup of a new data center or in a disaster recovery situation—the average model might be preferable due to its simplicity and ease of implementation. Conversely, in mature data centers where energy costs are a significant concern, and where the infrastructure and processes are already optimized, the exact model could provide the fine-tuned control necessary to achieve incremental improvements in energy efficiency.

The choice between the exact and average models depends on the specific goals and constraints of the data center operation. While the exact model offers a pathway to potentially greater energy savings through precise workload distribution, it requires a level of complexity and data that may not always be feasible or necessary. The average model, though less precise, provides a practical solution that balances simplicity with efficiency, particularly in environments where server homogeneity is high or where the marginal gains from more detailed modeling do not justify the additional costs. As data centers continue to evolve and face increasing demands for both performance and sustainability, the ability to navigate these trade-offs will be crucial in developing energy management strategies that are both effective and adaptable to changing conditions.

VI. CONCLUSION

The exploration of exact and average models for energy consumption in data centers underscores the complex interplay between precision and practicality in workload management strategies. As data centers continue to grow in scale and

ISSN(Online): 2320-9801

ISSN (Print): 2320-9798

International Journal of Innovative Research in Computerand Communication Engineering

(An ISO 3297: 2007 Certified Organization)

| Impact Factor: 7.186| A Monthly Peer Reviewed & Referred Journal |

Vol. 4, Issue 9, September 2016

complexity, the need for efficient energy management becomes increasingly critical, particularly in light of rising energy costs, environmental concerns, and the demand for reliable participation in programs like Emergency Demand Response (EDR). This study reveals that while the exact model offers a more detailed and accurate approach to energy consumption modeling, it also introduces significant computational and operational challenges. Conversely, the average model provides a simpler and more scalable solution, albeit with potential trade-offs in terms of energy optimization.

The exact model's ability to account for individual server characteristics enables a highly customized approach to workload distribution, which can result in substantial energy savings. By leveraging server-specific coefficients βi and ci, the exact model allows data center operators to fine-tune their energy consumption strategies, optimizing the use of each server's capabilities and minimizing unnecessary energy expenditure. This precision is particularly valuable in scenarios where small inefficiencies can lead to significant cumulative costs or when data centers operate under tight energy constraints. However, the model's complexity necessitates extensive data collection and analysis, making it more resource-intensive and potentially less adaptable to rapid changes in data center conditions. The average model's appeal lies in its balance of simplicity and effectiveness. By applying a uniform set of coefficients across all servers of the same type, this model reduces the computational burden and simplifies the process of workload management. It is particularly suitable for data centers with homogeneous server types or where rapid deployment and operational flexibility are prioritized over maximizing energy efficiency. While this model may not capture the nuances of individual server performance, it offers a practical solution that can be easily scaled across large data center environments, facilitating straightforward implementation and management.

The findings of this study suggest that the choice between these models should be informed by the specific operational goals and constraints of the data center. In environments where energy efficiency is paramount and where detailed data on server performance is readily available, the exact model provides a compelling approach to optimizing energy use. However, in scenarios where simplicity, scalability, and speed of deployment are critical, the average model offers a more practical alternative that still delivers reasonable efficiency gains.

Looking forward, the development of hybrid models that combine elements of both exact and average approaches could offer a promising direction for future research and application. Such models might leverage the precision of the exact approach for critical servers or during peak load times while employing the average model's simplicity for less critical operations or during periods of lower demand. This flexibility could help data centers better navigate the challenges of energy management in increasingly dynamic and demanding environments.

REFERENCES

1. MukherjeeT. et al. Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers Comput. Netw. (2009)
2. GandhiA. et al. Optimality analysis of energy-performance trade-off for server farm management Perform. Eval. (2010)
3. BrownR. Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431Tech. Rep. LBNL-363E (2007)
4. Albadi, M. H., & El-Saadany, E. F. (2008). "A Summary of Demand Response in Electricity Markets." Electric Power Research Institute (EPRI), 1-8. Document No. 1011877.
5. Baik, J.J.; Park, R.S.; Chun, H.Y.; Kim, J.J. A laboratory model of urban street-canyon flows. J. Appl. Meteorol. Climatol. 2000, 39, 1592–1600.
6. Kim, J.J.; Baik, J.J. Physical experiments to investigate the effects of street bottom heating and inflow turbulence on urban street-canyon flow. Adv. Atmos. Sci. 2005, 22, 230–237.
7. Yazid, A.W.M.; Sidik, N.A.C.; Salim, S.M.; Saqr, K.M. A review on the flow structure and pollutant dispersion in urban street canyons for urban planning strategies. Simulation 2014, 90, 892–916.
8. Molina-Aiz, F.D.; Fatnassi, H.; Boulard, T.; Roy, J.C.; Valera, D.L. Comparison of finite element and finite volume methods for simulation of natural ventilation in greenhouses. Comput. Electron. Agric. 2010, 72, 69–86.
9. Asuvaran & S. Senthilkumar, "Low delay error correction codes to correct stuck-at defects and soft errors", 2014 International Conference on Advances in Engineering and Technology (ICAET), 02-03 May 2014. doi:10.1109/icaet.2014.7105257.
10. Ghatikar, G., et al. (2014). "Data Centers and Demand Response: How Can They Help?." International Journal of Electrical Power & Energy Systems, 60, 346-355. DOI: 10.1016/j.ijepes.2014.02.025.