Journal of Global Research in Electronics and Communication
Volume 1, No. 12, December 2025
Available Online at: www.jgrec.info
REVIEW PAPER

A Review of Thermal and Mechanical Design Challenges in Modern Data Centers

Mr. Anoop Sunarty
Assistant Professor
Department of Computer Sciences and Applications
Mandsaur University.
Mandsaur
anoop.sunarty@meu.edu.in

***Abstract*—Thermal and mechanical design concerns in modern data centers are getting the highest priority due to increasing density and intensity of the IT work. The three-level network topology that was in existence in the past would not accommodate the modern east-west traffic, and this has led to the use of spine-leaf (Clos) topologies that offer low-latency, high-bandwidth, and scalability. The lack of uniform time-varying loads of high-density IT equipment makes thermal management even more difficult and needs advanced cooling solutions such as (but not confined to) airflow optimization, liquid cooling, and modular designs. The mechanical design must be in such a way that it provides structural integrity, use space efficiently, reduces vibration and also reduces noise as well as scalable and reliable cooling infrastructure. Efficiency, maintainability, and adaptability are the three metrics used to compare room, row, and rack-based cooling solutions. Combining the multidisciplinary thermal and mechanical aspects with the scalability of network architecture is also required to make the aspects of component reliability, operational efficiency, and energy performance sustainable in the current data centers. It is an integrated design of high-performance, durable and economical facilities.**

Keywords—Data Center Design, Thermal Management, Mechanical Design, Spine-Leaf Architecture, Cooling Infrastructure and Energy Efficiency.

I. Introduction

Data centers are the backbone of many different types of online services, including web hosting, online commerce, social media, and SaaS, PAAS, plus grid/cloud computing [1][2]. Platforms that provide generic services are proliferating; such examples are Google App Engine, Amazon Web Services (EC2), Sun Grid Engine, and Microsoft Azure [3]. Data centers are increasingly utilizing virtualization to increase server usage and enable more flexible resource allocation [4]. Virtualization is crucial for providing many of these services. To be sure, virtualization does increase the difficulty of data center administration in many ways.

Organizations' data center network architecture and operation have been radically altered by the fast development of distributed systems, cloud computing, and artificial intelligence workloads in the past several years. Traditional data center network architectures built around hierarchical, three-tier models with discrete access, aggregation, and core layers are increasingly ill-suited to the demands of modern applications and deployment paradigms. As east-west traffic patterns dominate over the north-south flows these legacy designs were optimized for, organizations require new architectural approaches. Modern enterprises face a complex set of challenges in their data center networks.

Parameters, including material qualities, part dimensions, loads, and strength, are often treated as random variables that adhere to specific statistical laws in traditional reliability design approaches [5]. It is also in accordance with this design guideline that the mathematical probability model and distribution were constructed. One of optimization's most significant subfields is reliability-based optimization. Featured in mechanical components such as gears and gear reducers [6]. When it comes to optimizing gear and planetary gear transmissions, among other types of transmissions, China has long been at the forefront. Various machine components, including radar and communication systems, have been proposed in China since the 1960s [7]. However, during the late 1970s, when the country's economy was experiencing fast growth, reform, and opening, the dependability of the propulsion system became an important issue for both civilian and military goods. Decades of work have resulted in a two-fold improvement in the dependability of military hardware.

In the contemporary digital era, businesses are continually seeking ways to leverage advanced technologies to enhance operational efficiency, innovate services, and maintain a competitive advantage [8]. Digital transformation happens when digital technology is incorporated into every part of a firm and has a profound impact on organizational operations and the delivery of consumer value [9]. The use of cloud computing, which offers inexpensive, adaptable, and scalable information technology resources over the internet, is pivotal to this change. Defense strategy advancements rely heavily on AI and ML, which have far-reaching effects on data security and quality.

A. Structure of the Paper

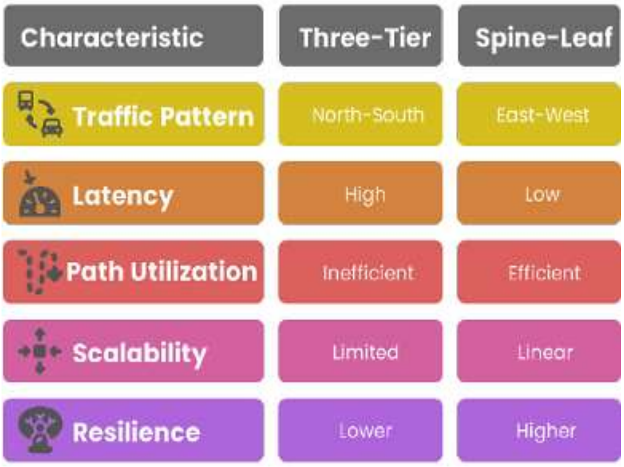
The current review is organized as follows: Section II is devoted to the architectural development of modern data centers and their role in thermal and mechanical design. Section III concerns thermal design issues, including heat sources and spatial-temporal changes in IT devices, system-level cooling designs, airflow control, modular cooling designs, and multidisciplinary design. Section IV is devoted to mechanical design issues, including structural and space constraints, integration of cooling infrastructure, etc. Lastly, Section V provides a summary of the most important results, defines the practice implications, and Section VI identifies future research directions in the thermal and mechanical design of modern data centers.

II. Architecture of Modern Data Centers

The latest data center designs have been developed to support virtualization, cloud computing, and data-intensive workloads, which generate the majority of east-west traffic. Three-tier traditional network designs, which are north-south communication-oriented, are poor in scalability, latency and bandwidth efficiency in such situations. In turn, spine-leaf (Clos) architectures have gained popularity as solutions that can provide low latency, equal bandwidth, and horizontal scalability. Use of a suitable topology varies with the scale of deployment, the nature of workloads and performance needs, between simple edge designs to multi-stage Clos designs in hyperscale data centers.

A. The Shift from Traditional Three-Tier Models to Spine-Leaf (Clos) Architectures

The three tiers of access, aggregation, and core were the conventional blueprint for data centers. The client-server applications' typical north-south traffic patterns were easily managed by this hierarchical network architecture. This layout was effective for situations when the majority of network traffic was destined for destinations other than the data center and vice versa. However, there was a substantial change toward east-west flows between servers as virtualization spread and applications grew more dispersed.



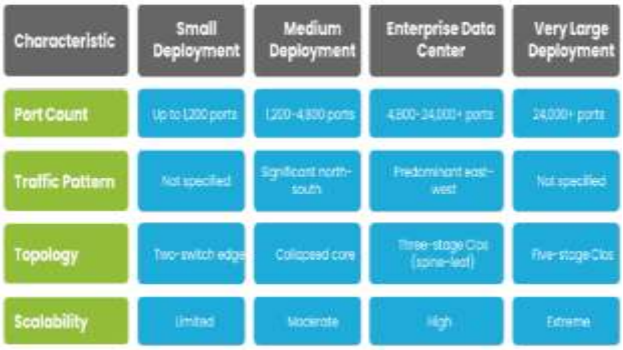
Characteristic	Three-Tier	Spine-Leaf
Traffic Pattern	North-South	East-West
Latency	High	Low
Path Utilization	Inefficient	Efficient
Scalability	Limited	Linear
Resilience	Lower	Higher

Fig. 1. Three-Tier vs. Spine-Leaf Architecture

The traditional three-tier model suffers from several limitations that make it poorly suited for modern workloads. Traffic between servers often must traverse multiple hops up and down the hierarchy, introducing latency and creating potential bottlenecks [10]. The spanning-tree protocol used to prevent loops in these Layer 2 networks blocks redundant paths, leaving significant bandwidth unused. Scalability is limited by the size of broadcast domains and the 4,096 VLAN limit (as shown in Figure 1). The current data center networks are built upon spine-leaf architecture, which offers the constant performance and horizontal scalability required by today's workloads.

B. Topology Selection Frameworks Based on Deployment Size and Use Case Requirements

Data center deployments vary significantly in size and requirements, from small rack/server rooms to massive enterprise data centers. The appropriate network topology depends largely on the scale of the deployment, as shown in Figure 2, the predominant workload types, and specific performance requirements.



Characteristic	Small Deployment	Medium Deployment	Enterprise Data Center	Very Large Deployment
Port Count	Up to 1,200 ports	1,200-4,800 ports	4,800-24,000+ ports	24,000+ ports
Traffic Pattern	Not specified	Significant north-south	Predominant east-west	Not specified
Topology	Two-switch edge	Collapsed core	Three-stage Clos (spine-leaf)	Five-stage Clos
Scalability	Limited	Moderate	High	Extreme

Fig. 2. Data Center Network Topology Comparison

For small deployments (1-25 racks, up to 1,200 ports), a two-switch edge design can provide sufficient connectivity and redundancy without unnecessary complexity. This minimalist design uses two interconnected switches to provide essential network services, making it suitable for edge locations or small data centers running cloud-native and virtualized workloads [11]. Medium-sized deployments (26-100 racks, 1,200-4,800 ports) often benefit from a collapsed core. approach that merges the traditional core and distribution layers.

This design offers a balance of simplicity and scalability, particularly when north-south traffic remains significant. For enterprise data centers (101-500+ racks, 4,800-24,000+ ports), a full three-stage Clos (spine-leaf) topology is the preferred architecture [12]. This design excels at handling east-west traffic, prevalent in virtualized environments, and provides the horizontal scalability needed as the data center grows. Very large deployments may require a five-stage Clos architecture, which adds a "super spine" tier above the spine layer. This design is especially valuable for connecting multiple data center pods or for hyperscale environments requiring extreme scalability.

III. Thermal Design Challenges in Data Centers

The primary obstacles that lead to the data center and telecoms thermal bottleneck are discussed in this section. To guarantee the electronics work efficiently and reliably, the cooling system is essential. Nevertheless, the system must deal with unknown boundary circumstances on both ends of this assignment, making it anything but trivial: (i) the ambient environment's fluctuation based on climate, and (ii) the chip's non-uniform dissipation over time.

A. Heat Generation and Thermal Characteristics of IT Equipment

Transistor Operation and Heat Load Spatial and Temporal Variations are covered in this:

1) Thermal Effects on Transistor Operation

The dissipation of power in a transistor may be broken down into two forms: the frequency-dependent dynamic power and the static power that results from leakage. At ambient temperature (below 100 °C), the overall leakage current for a typical complementary metal-oxide semiconductor (CMOS) transistor is the sum of the sub-threshold leakage and door oxide leakage [13]. The static power is approximately proportional to VDD^3, but the dynamic power is proportional to fVDD^2, where VDD is the operating voltage, and f is the frequency. As CMOS manufacturing technology has progressed towards tiny features, the leakage impact has grown. Power dissipated in technologies ranging from 45 nm to 28 nm is nearly entirely due to leakage. The main voltage-dependent component is the sub-threshold leakage current, which increases at a rate of around 10%/°C. Contrarily, the temperature sensitivity of total power has remained relatively constant, ranging from 0.5% to 2%/°C. The dependency of the operating frequency on the square root of the absolute temperature has been lowered or even reversed for contemporary manufacturing technology, when it was previously inversely proportional. The fundamental reason for this is that the threshold voltage decreases as the temperature increases. A combination of factors, including higher resistance, longer wire delays, and decreased electron mobility, contributes to this mixed temperature impact.

2) *Spatial and Temporal Variations in Heat Load*

The non-uniform and time-varying heat load makes it exceedingly difficult to keep the junction temperature within safe and efficient working limits, regardless of the actual total power dissipation. Minimizing heat loads that are not consistent throughout space is currently achieved by effective heat dispersion. Still, several advanced methods are being investigated, such as chip-integrated thermoelectric cooler arrays or micro-structured liquid-flow heat sinks with reconfigurable geometries [14]. Standard microstructures, such as single-phase liquid-cooled heat sinks with parallel microchannels, can be improved using these methods, which reduce the effect of die hotspots associated with high-power-dissipation regions and eliminate the inherent asymmetry caused by the coolant temperature increase in a streamwise manner. It is important to consider both the higher cost of such devices and the fact that certain research have demonstrated that geometrical optimization can decrease temperature gradients.

B. *System-Level Cooling Architecture Challenges in Data Centers*

Multiple heat transfer interfaces are present in these difficulties, and the following is the standardization of IT cooling equipment:

1) *Multiple Heat Transfer Interfaces*

Electronics cooling from chip to ambient provides several heat transfer cycles and interfaces due to its multiscale nature. To achieve overall increases in energy efficiency, it is important to reduce the number of intermediary connections. It may be a solution, but case-by-case investigation into direct refrigeration cycles incorporated into individual racks is necessary. No one solution can manage the entire cooling cascade, from chip to ambient, for modern datacom equipment, even if there are several cooling choices (air, water, refrigerants, etc.). The effective and dependable supply of the chips with a low DC voltage, generated from the high-voltage AC grid, is the goal of researchers now engaged in an optimization challenge pertaining to electrical power distribution. At the moment, businesses are sitting on their hands, hoping that high-voltage DC conversion standards would reduce the number of intermediate power conversions and boost energy efficiency.

2) *Standardization for IT Cooling Equipment*

Internal policies of each companies used to define the standards for cooling IT equipment. These days, other parties take care of it. The ASHRAE receives recommendations from a number of influential manufacturers in the cooling equipment business before establishing standards [15]. The following is an example of how the industry could save energy: The input temperature range for IT equipment that was operational in 2004 was 20–25 °C, but in 200 it was 18–27 °C.

3) *Acoustic Noise Emission*

Acoustic noise, especially from high fan loads, limits the functioning of conventional air-cooled racks as cooling demands rise. In the industrialized world, noise regulatory regulations are enforced by several agencies. In the United States, for instance, OSHA, which is part of the Department of Labor, is responsible for this. To save money on monitoring, training, complaint processing, and other overhead costs, datacom practices have acceptable thresholds that are lower than the levels specified in the standards [16]. There has to be a greater focus on a priori noise-mitigating design instead of posteriori problem-solving since manufacturing businesses might not always spot or address these real constraints. Since rising room temperatures often need greater fan speeds, acoustic noise emissions have become an increasingly pressing concern due to the expansion of the operational environment for datacom equipment as a result of new ASHRAE regulations.

C. *Airflow Management and Thermal Modeling*

Numerical modeling and simulation of complicated flow dynamics are commonplace while designing datacom cooling systems. Data center cooling spans such a large variety of length scales, and the tight time limitations of an industrial design process further impede the adoption of the most realistic modeling methods. It is possible to shorten the design cycle using less precise methods; nonetheless, thorough validation trials should accompany these simulations.

Mastery of modeling, validation, and result interpretation is essential for all simulation methods. Both for CFD and for getting and understanding experimental data, this is absolutely true. Verifying the accuracy of models and experiments is crucial. Model validation is often overlooked in industrial practice. Time and resource limitations cause popular validation methods like grid sensitivity analysis to be either entirely or partially neglected, even though inadequate meshing is the most common cause of problems [17]. A typical method involves contrasting the outcomes of a thorough model's simulation with those of a simpler model, which can be executed in around one hour and achieves an acceptable level of accuracy of approximately 85%. Time constraints mean that the first method is usually the best solution, since a full grid sensitivity analysis would require a lot of simulations with small grid changes. There isn't a consistent method for modeling and validation, or for representing and communicating these findings to clients, even if there are efforts to standardize equipment operating conditions.

D. *Modularity in Thermal Design*

The design of datacom cooling systems is affected to varying degrees by factors specific to the site, including as the building's architecture and the existing infrastructure. Issues may develop, for instance, when upgrading an old system and using elevated flooring to distribute cool air or when there is limited over-cabinet space because of cable racks put in for earthquake protection. Conventional CRAC systems typically fail to satisfactorily address these limitations. Rear door heat exchangers (RDHx) and other hybrid liquid cooling methods partially isolate the cooling performance of the rack from the room's HVAC system and airflow patterns, allowing for more versatility. At 60 kW per rack, an RDHx that is well-designed can absorb 80% of the total power lost. Due to its ability to manage the power dissipation of a 35 kW rack unit, less air cooling is necessary [18]. One way to reduce the initial investment costs is to use local liquid cooling, a modular 'pay as you go' cooling system technology. This allows you to create the cooling system to accommodate only the first heat load without having to worry about future load increases. As standardized thermal and/or fluidic connections are developed, modularity become even more valuable in making equipment replacement and adapting existing datacom centers easier.

IV. Mechanical Design Challenges

Contemporary data centers have become a major mechanical design problem because of the implementation of high-density IT equipment in the limited physical areas. Structural integrity, mechanical stability, scalability and increasing cooling requirements have become a critical design issue of concern. Also, there is a rise in the high-capacity cooling infrastructure that poses a problem of vibration, acoustic noise, and mechanical fatigue that may have an adverse impact on equipment reliability and operational life. Good mechanical design approaches should thus consider the load bearing, airflow and cooling consideration, vibration and noise consideration so as to be able to guarantee longer term system performance, durability and maintainability.

A. Structural and Space Constraints

The external sizes of the electronic equipment are 226 mm (length) x 210 mm (width) x 170 mm (height) as shown in Figure 3. Modular modules, screws, and side plates make up the bulk of the structure's architecture. Universalization, serialization, and modularization form the basis of the design, which enhances scalability and adaptability. The module units specially designed to meet the functional requirements are set to either single-cavity or double-cavity structures. A self-contained chamber around the constituent module units is formed by integrated milling to ensure efficient electromagnetic compatibility [19]. This method improves the equipment's seismic performance while reducing its overall weight.

The internal and external bonds of the structures with the two cavities are not that long which allows to use the space effectively and simplify the process of disassembly and maintenance. The modules in each unit are structurally set and the modules can be scaled down or up as per the requirement of the systems. This system of modules enables easy expansion and structural integrity as well as efficient utilization of limited space of installation.

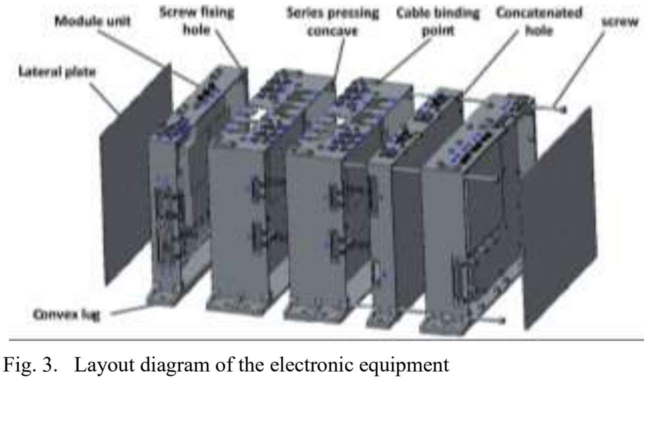


Fig. 3. Layout diagram of the electronic equipment

Figure 4 shows that four screws are used to secure one module cavity to the other. In contrast to the other module cavities, which are constructed with serial holes, the leftmost module offers screw-fixing holes. Electrical devices sandwich two neighboring module units inside of one another to protect the screw from radial shear stress. Each module structure is laid out with a series of pressing convex surfaces on the left and a series of pressing concave surfaces on the right [20]. Electronic equipment is guaranteed to be overall stiff thanks to its construction. Less vibration and impact mean less chance of damage. This design places the long side of the satellite module flush with the mounting surface, which aids in heat dissipation—an essential component of electronic equipment. Because of its short heat dissipation route, it works well for thermal design.

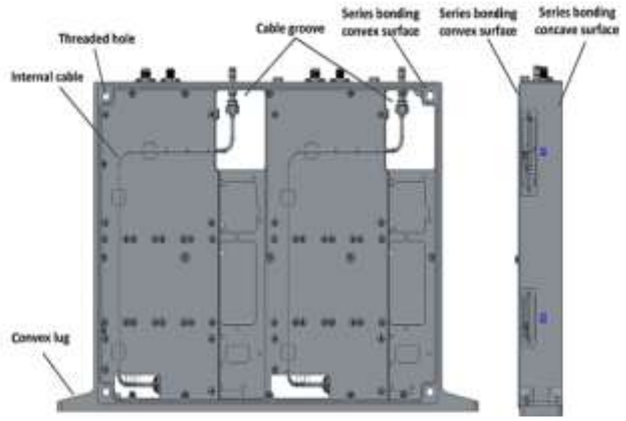

Threaded hole
Internal cable
Cable groove
Series bonding convex surface
Series bonding convex surface
Series bonding concave surface
Convex lug

Fig. 4. Schematic diagram of module unit structure

The installation of the space station and the different weather conditions must be taken into account when designing the interface for the electrical equipment structure on board. Consideration of the satellite's general layout should guide the design of the installation position and procedure for electronic equipment. On the basis of design requirements, each module's interface position and connection manner are determine. A stationary convex lug at the base of the satellite's cabin architecture holds electronic components in place. Fifteen convex lugs are the result of an exhaustive mechanical modeling. Electronic devices are fastened with M5 screws in a symmetrical, bilateral pattern. Round holes on the left-hand module unit serve as installation reference holes. Additional modules' holes are waist-shaped to guarantee installation redundancy. The module unit has cable binding points on top. Incorporating cable supports into the design allows for a binding option for cable interconnection.

B. Cooling Infrastructure Design

A data center project with a capacity of 20 MW is the subject of this research, which is based on practical learning (Figure 5). A data center located on nine stories is supplied with chilled water by a district cooling system [21]. Central chilled-water-type computer room air handling (CRAH) units make up the cooling system of the case study data center. These devices are seen in Figure 5. The IT server racks are housed in a cold aisle containment structure, which allows the comparatively hot supply air (SA) from the CRAH units to flow. The temperature and relative humidity are maintained constant with the use of 43 CRAH units per floor, plus one additional redundant unit. District cooled water is heated on one side and somewhat hotter chilled water is received on the other side of the heat exchanger.

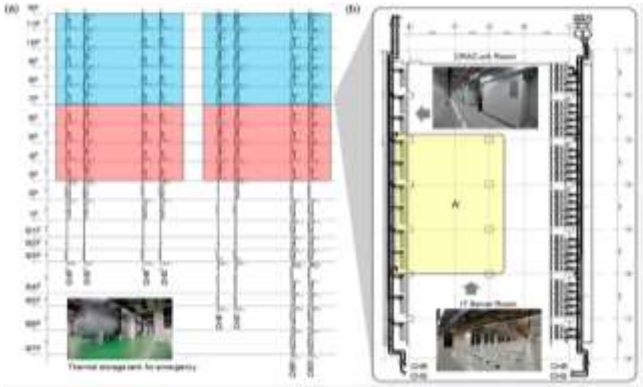


Fig. 5. Data center cooling system: (a) secondary chilled water loop; (b) IT server room on a typical floor

Additionally, there is a backup chilled water supply stored in buffer tanks in case of an emergency. The horizontal pipes of a data center's cooling system typically run in a loop on every level, with an extra riser for backup in case of an emergency. To provide a steady supply of chilled water in the case of an emergency power loss, chilled-water storage tanks were set up [22]. To find the most cost-effective storage tank size, one must calculate the window of opportunity to reuse the chilled water in the pipes without turning on the cooling system, all while keeping the IT server working environment unaffected. A preliminary determination of the pipe water level was made for this objective.

The following is a direct comparison of the pros and cons of air-cooling features depending on room, row, and rack design, as shown in Table I.

TABLE I. COMPARISON OF ROOM-, ROW-, AND RACK-BASED AIR-COOLING FEATURES AND THEIR ADVANTAGES AND DISADVANTAGES

Category	Room-Based Air Cooling	Row-Based Air Cooling	Rack-Based Air Cooling
Flexibility Advantages	Rapid changes in cooling distribution can be implemented for power densities below 3 kW; cooling capacity can be shared across the room	Convenient planning; localized cooling aligned with IT load	Convenient planning; independent of existing cooling infrastructure
Flexibility Disadvantages	Lower performance in the absence of airflow confinement	Requires a hot- and cold-aisle layout	Transferring the cooling capacity of one rack to another is not possible.
System Availability Advantages	Data redundancy can be shared across all racks.	Redundant racks in the same row or zone can be used together; cooling close to heat sources lowers temperature gradients in the vertical direction.	Avoiding hotspots and vertical temperature gradients by cooling near heat sources also reduces the likelihood of human mistakes.
System Availability Disadvantages	Restricting the passage of air is necessary.	Every row must have redundancy.	Every rack must have redundancy.
Life-Cycle Costs Advantages	Easy reconfiguration of raised/perforated floors	Reduced planning and engineering effort	Reduced planning and engineering effort through prefabrication and standardized components
Life-Cycle Costs Disadvantages	Excessive air supply volume increases energy use	Initial costs increase as data center size grows	The risk of over-planning leads to higher initial costs
Maintainability Advantages	Technicians have limited access to IT equipment due to cooling equipment being positioned outdoors or on the outside of the IT area.	Skilled technicians are not as necessary when standardized components are used.	Reduced reliance on technicians is a result of standardized components; regular maintenance may be performed by in-house personnel.
Maintainability Disadvantages	Repairs should only be performed by qualified professionals.	Technicians must work close to IT equipment	2N redundancy is necessary for parallel maintenance.
Manageability Advantages	Low number of interfaces and management points	Interface-based monitoring enables real-time analysis	Interface-based monitoring with real-time analysis
Manageability Disadvantages	No real-time analysis; advanced training is necessary.	There must be several interfaces for large-scale installations.	Numerous interfaces are necessary for large-scale deployments.

C. Vibration, Noise, and Reliability Issues

Vibration and acoustic noise have emerged to be very important issues in modern data centers, as there is more cooling required with greater rack power densities. Vibration and acoustic noise are mostly induced by high-speed rotation of axial and centrifugal fans in order to satisfy the increasing requirements of airflow. These vibrations may be transmitted over racks and enclosures, resulting in mechanical resonance, wear of moving parts, and may cause damage to delicate hardware like hard disk drives. Also, high noise levels can be beyond acceptable occupational levels, which present operational and regulatory problems. This has a large effect on hardware reliability and maintenance because consistent vibration and thermal cycling are capable of increasing component fatigue, weakening mechanical fasteners, and decreasing the life of electronic and mechanical components [23]. The potential for service outages, operational expenses, and the frequency of maintenance interventions are all impacted by failure rates. As such, to reduce vibration and noise caused by poor mechanical design, fan synchronization, damping, and other coolant methods like liquid cooling are important in mitigating vibration and noise to ensure long-term reliability.

V. LITERATURE REVIEW

The main focus of the studies summarized in Table II is on thermal and mechanical management strategies in modern data centers, addressing challenges such as hotspot mitigation, airflow optimization, high-density server design, and energy-efficient cooling. The reviewed literature employs experimental setups, CFD-based simulations, analytical indices, and system-level design frameworks to evaluate the effectiveness of advanced cooling architectures and mechanical innovations in enhancing thermal reliability, energy efficiency, and operational safety of data centers.

Guo et al. (2025) provide a foundation for area thermal control by means of an adaptive terminal device that modifies airflow in response to the power offset ratio. By redistributing cold air, this approach not only protects the privacy of IT sector data but also balances cooling supply with server heat generation. It was demonstrated that the framework could change the flow offset ratio to meet the distribution of power on servers in both static and dynamic settings through the construction of a test platform. Improvements in temperature homogeneity and a 4.44 °C drop in hotspot temperatures were seen in static testing, and a 4.48 °C drop in dynamic tests [24].

Gao et al. (2024) offer the Energy Calculus synergy index as a result of a three-stage comprehensive evaluation of the data center's whole-link heat transfer mechanism. The Temperature Cooling Index (TCI) is a new whole-link heat flow management index that has been developed with existing indices like DCP and the Thermal Matching Coefficient of the server cabinet (TMC). Ensuring the safety and high efficiency of data centers is made possible by optimizing critical parameters across the entire chain through optimal information and communication technology. These parameters include thermal resistance, airflow organization, cooling infrastructure efficiency, and chip energy utilization efficiency. regulating the environment [25].

Zhao et al. (2023) offers a synopsis of current studies examining methods for controlling local hotspots, with an emphasis on servers that are overheating, which are considered to be the most probable sites of such issues. In order to put a number on hotspots, this study defines them precisely and analyzes indicators from several angles. Poor airflow distribution or heavy server loads are common causes of overheating; therefore, this feature examines the relationship between the temperature of the air entering and leaving the system. And hot-air recirculation is the most basic way that hotspots express themselves. Methods for improving airflow distribution and reducing the length of the cold airflow route are considered in this research. Along with that, the paper offers future directions and talks about the limits of recent hotspot-elimination studies [26].

Zheng et al. (2022) This is Kuaishou's cutting-edge high-density storage server. It's unique among storage servers since it uses revolutionary mechanical and thermal methods to fit 42 individual 3.5-inch HDDs (Hard Disk Drives) in a typical 2U server footprint. The drives were relocated within the chassis instead of being placed outside in order to maximize the 2U space available for 42 HDDs. This removed the drives from convenient reach for maintenance tasks like hot-plugging or replacement, but it also raised worries about reliability, performance, and thermal issues caused by drives overheating. This high-density storage system utilizes the ideal air duct thermal design, an enhanced system design for RVI (rotation vibration isolation), and a revolutionary push-pull mechanical design to tackle these issues [27].

Yuan et al. (2022) suggests a distributed cooling system that would improve the data center's thermal environment by integrating heat pipe exchangers installed on servers with conventional air conditioning systems for computer rooms. Show the two different cooling systems compared to the original CRACs system here: one that utilizes heat exchangers mounted above the servers and another that uses them below. Since the results of the simulations match up well with the findings of the on-site measurements, and confirm that the original data center model is reliable. The findings demonstrate that data centers can significantly benefit from a decentralized cooling system when it comes to thermal conditions. examined and compared 18 scenarios to find the one with the best cooling efficiency, all of which used heat pipe exchangers set at various heights and positions [28].

Abbas et al. (2021) examine the impact on data center in-row cooling performance of utilizing aligned/staggered cooling unit layouts and top aisle confinement. Using identical data centers with varied cooling-unit arrangements (aligned or staggered) and with or without top aisle containments, four distinct configurations of cooling architecture are numerically created. Utilizing the CFD software ANSYS-IcePak, distinct models are generated for each of the four configurations. The model is tested using a physically reduced version of the data center that has already been constructed and verified by experimenting. The effectiveness of the various designs in terms of thermal environmental performance and energy efficiency are assessed and compared using air streamlines, velocity vectors, temperature distributions, two thermal performance indices (SHI and IOM), and two energy efficiency metrics (β and ηr) [29].

TABLE II. THERMAL AND MECHANICAL DESIGN IN MODERN DATA CENTERS

Authors	Focus Area	Key Findings	Approaches	Objectives	Limitations
Guo et al. (2025)	Local thermal management and airflow optimization	To develop a privacy-preserving, adaptive local thermal management framework for hotspot mitigation	Airflow management based on power offset ratio; experimental platform for static and dynamic testing; adaptive terminal device	Enhanced temperature uniformity; enhanced matching of cooling supply to server heat generation; and a 4.44 °C static and 4.48 °C dynamic reduction in hotspot temperatures.	Validation limited to test platform; scalability and long-term deployment in large-scale data centers not evaluated
Gao et al. (2024)	Whole-link thermal management and heat flow optimization	To achieve optimal ICT thermal management through holistic heat transfer analysis	Whole-link heat transfer modeling; introduction of DCP, TMC, and TCI indices; energy calculus–based synergy analysis	Proven that cooling infrastructure, airflow organization, thermal resistance, and chip performance can all be optimized to increase efficiency and safety.	Primarily analytical and index-based; lacks extensive experimental or real-world deployment validation
Zhao et al. (2023)	Local hotspot characterization and mitigation	To review and classify hotspot formation mechanisms and mitigation strategies	Literature review; hotspot definition and indicators; analysis of inlet–outlet temperature correlations and airflow behavior	Identified hot air recirculation as the primary cause of hotspots; highlighted airflow optimization and cold airflow path shortening as key solutions	Review-focused study; limited quantitative comparison of mitigation techniques and a lack of experimental validation
Zheng et al. (2022)	Designing mechanical and thermal components for servers with high-density storage	To design a compact high-density storage server while maintaining thermal reliability	Novel push–pull with rotation mechanical design; rotation vibration isolation (RVI); optimized air duct thermal design	Successfully integrated 42 HDDs in 2U space; mitigated vibration, thermal preheating, and reliability risks	Maintenance complexity remains higher than conventional designs; applicability is mainly limited to storage-centric servers
Yuan et al. (2022)	Decentralized cooling and	To improve the data center thermal environment using	CFD simulation validated with on-site measurements; server-level heat pipe	Decentralized cooling significantly improved thermal conditions; optimal placement of	Increased system complexity; economic feasibility and

	server-level heat exchange	localized cooling strategies	exchangers combined with CRACs; evaluation of 18 configurations	heat pipe exchangers enhanced cooling efficiency	maintenance implications not fully assessed

VI. Conclusion and Future Work

The high-density IT workloads, virtualization, and cloud computing are placing tremendous thermal and mechanical pressure on modern data centers. The existing network structures and standard cooling techniques are becoming inadequate, and scalable spine-leaf architectures and new strategies to cool the entire system are needed to ensure efficient operation. The thermal design should consider non-uniform and time-varying heat sources, transistor leakage, and numerous heat transfer interfaces, whereas the mechanical design should address the structural integrity, effective space use, vibration suppression, and noise. Energy efficiency, dependability, and maintainability are all enhanced by integrating cooling methods such as liquid cooling, room-based, row-based, and rack-based systems. The thermal, mechanical, and electrical engineers should collaborate to maximize performance, minimize operational costs, and extend component life. In general, the integrated approach of scalable architecture, novel thermal placement, and robust mechanical design plays a key role in the sustainable, high-performance operation of data centers.

Predictive thermal management, AI-based airflow optimization, and efficient hybrid cooling systems should be studied in the future. It can be further enhanced by more advanced modeling methods and uniform validation standards. Also, the adoption of modular, scalable mechanical design, coupled with advanced high-density computing architectures, will improve the flexibility and sustainability of operations in future data centers.

References

[1] T. R. Merlo, F. Fard, and S. Hawamdeh, "Cloud Computing's Impact on the Digital Transformation of the Enterprise: A Mixed-Methods Approach," *Sustainability*, vol. 17, no. 13, Jun. 2025, doi: 10.3390/su17135755.

[2] S. Amrale, "Proactive Resource Utilization Prediction for Scalable Cloud Systems with Machine Learning," *Int. J. Res. Anal. Rev.*, vol. 10, no. 4, 2023.

[3] B. Senapati *et al.*, "Quantum Computing and Its Potential Disruption to Data Centers and Edge Computing in Battery Cell Manufacturing Sites," in *2025 IEEE International Conference on Electro Information Technology (eIT)*, IEEE, May 2025, pp. 126–131. doi: 10.1109/eIT64391.2025.11103699.

[4] K. Kant, "Data center evolution: A tutorial on state of the art, issues, and challenges," *Comput. Networks*, vol. 53, no. 17, pp. 2939–2965, Dec. 2009, doi: 10.1016/j.comnet.2009.10.004.

[5] Y. Yuanfan, "The Study on Mechanical Reliability Design Method and Its Application," *Energy Procedia*, vol. 17, pp. 467–472, 2012, doi: 10.1016/j.egypro.2012.02.122.

[6] R. McFarlane and J. Weale, "Mechanical Design in Data Centers," in *Data Center Handbook*, Wiley, 2021, pp. 403–440. doi: 10.1002/9781119597537.ch24.

[7] U. G. Udhav, B. Ashok, D. Eshan, A. Shahjahan, G. Narayanan, and K. Pramod, "Thermal and mechanical design considerations for a switched reluctance motor," in *2016 7th India International Conference on Power Electronics (IICPE)*, IEEE, Nov. 2016, pp. 1–6. doi: 10.1109/IICPE.2016.8079409.

[8] E. O. A, "Digital Transformation: The impact of AI on Cloud Transformation," *J. Artif. Intell. Gen. Sci. ISSN3006-4023*, vol. 5, no. 1, pp. 174–183, Jun. 2024, doi: 10.60087/jaigs.v5i1.188.

[9] M. R. R. Deva, "Advancing Industry 4.0 with Cloud-Integrated Cyber-Physical Systems for Optimizing Remote Additive Manufacturing Landscape," in *2025 IEEE North-East India International Energy Conversion Conference and Exhibition (NE-IECCE)*, IEEE, Jul. 2025, pp. 1–6. doi: 10.1109/NE-IECCE64154.2025.11182940.

[10] K. Ebrahimi, G. F. Jones, and A. S. Fleischer, "A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities," *Renew. Sustain. Energy Rev.*, vol. 31, pp. 622–638, Mar. 2014, doi: 10.1016/j.rser.2013.12.007.

[11] M. Munawir *et al.*, "Penentuan Alternatif Lokasi Tempat Pembuangan Akhir (Tpa) Sampah Di Kabupaten Sidoarjo," *Energies*, 2022.

[12] A. S. George, "The Evolution of Data Center Networks: Strategies for Modern Infrastructure Design," *Partners Univers. Multidiscip. Res. J.*, vol. 02, pp. 141–159, 2025, doi: 10.5281/zenodo.15450624.

[13] R. Patel, "Survey of Digital Twin Applications in Predictive Maintenance for Industrial," *Int. J. Recent Technol. Sci. Manag.*, vol. 9, no. 4, 2024.

[14] S. V. Garimella, L.-T. Yeh, and T. Persoons, "Thermal Management Challenges in Telecommunication Systems and Data Centers," *IEEE Trans. Components, Packag. Manuf. Technol.*, vol. 2, no. 8, pp. 1307–1316, Aug. 2012, doi: 10.1109/TCPMT.2012.2185797.

[15] E. Frachtenberg, D. Lee, M. Magarelli, V. Mulay, and J. Park, "Thermal design in the open compute datacenter," in *13th InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, IEEE, May 2012, pp. 530–538. doi: 10.1109/ITHERM.2012.6231476.

[16] A. Yuksel, D. W. Demetriou, Y. Hu, and V. Mahaney, "Thermal Design of Portable Modular High Performance Computing Data Centers," in *2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, IEEE, Jun. 2021, pp. 505–509. doi: 10.1109/ITherm51669.2021.9503286.

[17] R. Patel, "Advancements in Data Center Engineering: Optimizing Thermal Management, HVAC Systems, and Structural Reliability," *Int. J. Res. Anal. Rev.*, vol. 8, no. 2, pp. 991–996, 2021.

[18] S. Garg, "AI/ML Driven Proactive Performance Monitoring, Resource Allocation and Effective Cost Management in SAAS Operations," *Int. J. Core Eng. Manag.*, vol. 6, no. 6, pp. 263–273, 2019.

[19] V. Panchal, "Thermal and Power Management Challenges in High-Performance Mobile Processors," *Int. J. Innov. Res. Sci. Eng. Technol.*, vol. 13, no. 11, pp. 18214–18229, Nov. 2024, doi: 10.15680/IJIRSET.2024.1311014.

[20] A. A. Alkrush, M. S. Salem, O. Abdelrehim, and A. A. Hegazi, "Data centers cooling: A critical review of techniques, challenges, and energy saving solutions," *Int. J. Refrig.*, vol. 160, pp. 246–262, Apr. 2024, doi: 10.1016/j.ijrefrig.2024.02.007.

[21] M. Meier and E. G. Strangas, "Cooling Systems for High-Speed Machines—Review and Design Considerations," *Energies*, vol. 18, no. 15, Jul. 2025, doi: 10.3390/en18153954.

[22] G. Barone, A. Buonomano, G. F. Giuzio, and A. Palombo, "Towards zero energy infrastructure buildings: optimal design of envelope and cooling system," *Energy*, vol. 279, Sep. 2023, doi: 10.1016/j.energy.2023.128039.

[23] A. Isazadeh, D. Ziviani, and D. E. Claridge, "Thermal management in legacy air-cooled data centers: An overview and perspectives," *Renew. Sustain. Energy Rev.*, vol. 187, Nov. 2023, doi: 10.1016/j.rser.2023.113707.

[24] H. Guo, H. Yu, S. Zhao, H. Chen, and C. Li, "Improving the local thermal environment in data centers via supply–demand matching," *Energy Build.*, vol. 347, Nov. 2025, doi: 10.1016/j.enbuild.2025.116350.

[25] P. Gao *et al.*, "Discussion on the technical path of data center information and communication thermal management," *Energy Reports*, vol. 11, pp. 2704–2714, Jun. 2024, doi: 10.1016/j.egyr.2024.02.003.

[26] R. Zhao, Y. Du, X. Yang, Z. Zhou, W. Wang, and X. Yang, "A critical review on the thermal management of data center for local hotspot elimination," *Energy Build.*, vol. 297, Oct. 2023, doi: 10.1016/j.enbuild.2023.113486.

[27] S. Zheng *et al.*, “An Advanced Mechanical and Thermal Design Optimal For High Density Storage Server Reliability,” in *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*, IEEE, May 2022, pp. 1–7. doi: 10.1109/iTherm54085.2022.9899598.

[28] X. Yuan, X. Zhou, Y. Liang, Y. Pan, R. Kosonen, and Z. Lin, “Design and Thermal Environment Analysis of a Decentralized Cooling System with Surface-Mount Heat Pipe Exchangers on Servers in Data Centers,” *Buildings*, vol. 12, no. 7, Jul. 2022, doi: 10.3390/buildings12071015.

[29] A. M. Abbas, A. S. Huzayyin, T. A. Mouneer, and S. A. Nada, “Thermal management and performance enhancement of data centers architectures using aligned/staggered in-row cooling arrangements,” *Case Stud. Therm. Eng.*, vol. 24, Apr. 2021, doi: 10.1016/j.csite.2021.100884.