A Review of the Factors Impacting the Optimal Placement of Data Centers

Viraj Nain

Abstract

Our current digital landscape relies heavily on existing data centers to store and process information for online applications accessed by millions of users. Any failure of these operations significantly affects productivity as well as significant operations of an organization. Additionally, data centers are high energy consumers. Climate change and the scarcity of energy resources due to political or local constraints emphasize the need for us to deeply analyze the strategic placement of these data centers to understand factors contributing to risks and opportunities in data center placement.

1. Introduction

Recently, climatic conditions like extreme heat in California caused Twitter’s non-redundant data center to experience total shutdown. Cyber hackers pounce on such vulnerability and perfect their methods to exploit it. For example, a cyberattack on cooling systems of a datacenter in Atlanta seriously disrupted court, police, and airport operations in the city [1]. Further, there are political and local regulations such as the EU’s focus on regulating energy intensive technology industries. It caused Meta to abandon its plan to establish a datacenter in the small town of Zeewolde in the Netherlands [2]. All these climatic, technological, political, and local factors emphasize the challenge of determining the placement of data centers. It is now more relevant than ever before, keeping in view the increasing growth of data centers and their sheer increase in size. Additionally, the advancement in data center technology needs to be constantly upgraded to keep up with energy, storage, and capacity needs. These data centers are already forced to balance efficiency, reliability, and sustainability in their architectures. With disruptions causing temporary shutdowns of data centers, it is crucial to address these issues for the successful operation and maintenance of data centers globally.

The approach taken to tackle data center placement and its nuances was a comprehensive examination of existing works and research publications. The report delves into internet connectivity, disasters, climate conditions, and economic factors and examines their effects on data centers, to better grasp optimal placement techniques. In addition, such extensive analysis allowed me to add personal thoughts and suggestions towards optimal placement.

2. All about data centers

The following section primarily focuses on exploring data centers through their historical evolution, roles and functions, technical functionalities, network architectures, and the traffic they handle. The section categorizes data centers to understand their purposes and sales options to explain the inner workings of data centers.

2.1 The history of data centers

March 07, 2024
In the 1960s, as computers were in their infancy, the foundations of modern data centers began to take shape. These early computing systems, although basic by today’s standards, laid the groundwork for centralized data processing. It wasn’t until the 1990s that the term “data center” gained widespread recognition, encompassing hubs dedicated to managing and storing digital information.

Fast forward to today, data centers are fundamental to the digital landscape, especially with the rise of cloud computing. Platforms like Amazon Web Services (AWS) offer scalable solutions for businesses and individuals. The shift to cloud-based services has transformed how data is stored and processed, with data centers at the core of this evolution.

In the business realm, data center providers are pivotal in providing cost-effective resources. These resources streamline digital infrastructure implementation, reducing costs for businesses. Organizations leverage data center services to enhance efficiency in today’s data-driven environment. The journey from rudimentary 1960s computer operations to today’s data centers shows us that data centers are at the core of our future world that is so reliant on information and data.

2.2 Data center’s role/function

A data center, as defined, is a dedicated facility or building designed primarily to house a computer room and its associated support areas [3]. Data center providers play a pivotal role by offering physical spaces and tailored guidance to tenants, enabling them to construct and operate their data centers using their equipment. Google, for instance, defines a data center as a facility where multiple servers and communication equipment co-locate due to shared environmental and security needs. As exemplified by service models such as Data Center as a Service (DCaaS), cloud computing allows tenants to access resources on demand according to service-level agreements (SLAs). Infrastructure as a Service (IaaS) is a prevalent model wherein companies utilize computers and other resources housed within data centers.

The Data Center Network (DCN) is crucial, connecting physical components like servers and switches within data centers through specific arrangements of cables and optical fibers. The network architecture of the Datacenter is significantly influenced by the role of the data center and, subsequently, has led to emergence of technologies like server virtualization and improved architectures. The role of these data centers ranges from providing data storage, hosting web services, responding to web search requests and even performing complex computations depending on business requirements [4-5].

Recent push towards creating low-cost Data Centers has led to the creation of Modular Data Centers. These are created on-demand and sometimes even on-premises. They are generally created using shipping containers. Such low-cost data centers provide advantages beyond cost saving like quick deployment. However, these are not free from disadvantages. Apart from maintenance and compatibility issues, these might have high failure rates as well. The failure rates of these remain hidden from users most of the time on account of redundancy.

On the other hand, Green Data Centers prioritize energy efficiency, incorporating technologies for energy savings, environmentally friendly management practices, and renewable resources to create sustainable and eco-friendly data processing environments.

2.3 Technical functionality

The central element of a data center is the computer room, housing essential equipment for data processing. This includes servers, which serve as the primary components responsible for storing and processing data. Additionally, the computer room is supported by a range of critical equipment such as power supply and backup systems, cooling systems, lighting systems, fire protection systems, cabling systems, security systems, and dedicated staff for maintenance purposes.
In a data center’s infrastructure, switches play a crucial role by connecting devices and acting as controllers to facilitate the sharing of information. Servers, as mentioned earlier, take on the primary responsibility of data storage and processing. Figure 1 shows the difference in size and build between the three main types of servers. Different server types cater to varying needs: tower servers, characterized by their large shape and high performance, are suitable for small-scale businesses but are less flexible and not commonly employed in cloud environments. Rack servers, the most prevalent type, offer standard space-saving arrangements but can pose challenges in terms of heat dissipation and cabling complexity due to dense placements. Meanwhile, blade servers, noted for their low cost, high availability, and density, are garnering increasing research attention, and are poised to become the next generation of widely adopted mainstream servers.

Traditional Data Center Networks (DCNs) typically adopt a three-layer rooted tree architecture, with core switches forming the backbone that connects the data center to the internet. Figure 2 visualizes the connections between layers and switches in a graphic format. The core switches interface with aggregate switches,
which link to edge switches that are responsible for connecting to servers. The core layer facilitates high-speed packet switching and routing, acting as a backbone for data traffic. The edge layer connects end devices like servers and storage to the distribution layer, involving switches that provide network connectivity to devices within the data center. Security is generally focused on this edge layer and the aggregation layers. Meanwhile, the aggregation layer provides connectivity based on Policies set up by the organizations. It also ensures a level of queuing based on classification of connections made at the access layer. It maintains Access Control Lists and tries adding another layer of security before providing connectivity to the core layer, utilizing routers and layer 3 switches for routing between different subnets.

The main disadvantage of these traditional DCNs is scalability. This problem is aggravated when working at peak capacity. Furthermore, increasing capacity is associated with additional servers, replacement of switches, complex cabling, and associated energy management operations that adds to ever increasing cost and complexity. Organizations that have fluctuating demands experience low levels of resource utilization as well. With many devices remaining idle for the extended periods that aren’t peak times, it becomes difficult to justify the scalability and operational costs. These were more attuned to ingress (North-South) traffic and Client-Server architecture. Modern technological demands call for building redundancy at the server level. However, building such redundancy required horizontal expansion of the Data Center, thereby causing costs to increase exponentially. Moreover, adding a simple firewall to various servers becomes complex and may negatively impact the bandwidth.

In recent times, the application complexities as well as requirements have led to more East-West traffic and the traditional Data Centers do not fare well in building redundancy. This led to the development of alternative architectures. The server-centric architecture of Data Centers is comparatively more scalable and fares well in egress (East-West) traffic. It has deterministic bandwidth and latency and can do most of the local routing and networking at the server level [8].

In switch-centric design, routing information is held by switches and server generally holds a single Network Interface Card (NIC) and does not participate in forwarding. Switch-centric DCNs are faster and have lower latency. However, they are not as programmable as servers. Additionally, switches are high power consumers. Server-centric DCNs fare well on programming, however, they are not as efficient [9-10].

Every architecture has its own pros and cons. In fact, there has been research and development focused on dual-centric designs to provide additional alternatives to diverse and dynamic requirements of modern data centers.

2.4 The type of network traffic typical of a data center network

It is imperative to study and understand Data Center Network traffic patterns and characteristics to make informed decisions about the requirements as well as location of the data center. Long term strategy of an organization in terms of expansion in new markets or geographies, usage of new technologies, security and recovery directions or initiatives, all have a bearing on the Data Center Network traffic [11]. Multiple applications and tenants share the data center resources to address the needs of the users. Traffic understanding and control is necessary to effectively utilize and fairly distribute the resources.

Data center traffic can be categorized into three main types. The first type is interactive traffic. It is the traffic generated by user interactions. In such cases, the latency and response time matter. These generally cause short spikes in traffic and are responsible for smaller and shorter bandwidth usage of data centers.

Throughput oriented is another type of data center traffic. Batch computing jobs, data warehouse data transfers and large data computations come under this category. These are responsible for consistent bandwidth usage and are generally scheduled jobs.

Another important characteristic that affects data center traffic is the time-sensitivity of the data implied by the application. For example, if a specific user query does not respond within the specified time, then instead of waiting for response return zero to ensure faster response time. However, if the traffic includes
creating system backups, and distributed file systems then the application may relax the deadline and
time sensitivity and ensure that the backup is created even if the deadline has passed [12].

Growth in areas of Artificial Intelligence, gaming and streaming has fueled the data center demand and this
will continue to grow. This coupled with ease of infrastructural setup associated with Cloud operations has
led to faster adoption of Data Centers specifically Cloud Data Centers.

2.5 Tiers of Data Centers

Data centers are categorized into four tiers depending on their reliability and capacity [13]. Tier I is also
known as Basic Capacity Tier. It has an uptime of 99.671%, with a maximum downtime of 28.8 hours per
year. This is the most affordable option, especially for small businesses, wherein occasional shutdowns do
not significantly hamper the business operations. This tier lacks redundancy and is vulnerable to failures.

Tier 2 data centers are also known as Redundant Capacity. These centers have backup options to prevent
failures associated with power interruptions like generators for backup power, additional cooling equipment
to prevent server shutdowns because of heat, and prevention of power spikes and voltage fluctuations to
ensure stable electricity. These tiers have an uptime of 99.741% and a maximum downtime of 22.7 hours per
year. Data centers in these tiers try to minimize the interruptions.

Tier 3 data centers, also known as Comprehensive Redundancy Data Centers, are appropriate for medium and
small businesses, especially for business-critical operations. They provide 99.982% uptime, with a maximum
downtime of 1.6 hours per year. They are more reliable compared to lower tier data centers.

At the top is the Tier 4, Fault Tolerant Data Centers. They have an uptime of 99.95% and downtime of less
than 26 minutes per year. These are the most available Data Centers and can be crucial for businesses that
demand constant availability with continuous and uninterrupted services. The cost associated with this Tier
is highest.

The decision to choose a Data Center from one of the above-mentioned Tiers depends not only on the
business objectives but also on the cost associated with it. This classification helps in making an informed
decision to choose the most appropriate data center for specific business needs.

2.6 Real-life examples

In summer of 2022 the UK experienced a record heatwave, and this caused outage of two data centers that
supported NHS (National Health Services) hospitals. It caused significant disruption to the hospital admin-
istration causing appointment cancellation, delays and in fact a patient even missed an organ transplant.
Additionally, the staff had to rely on paper and the confusion it caused further lowered the morale of the
employees. Full restoration of services took approximately 6 weeks. A review by NHS later estimated the
associated cost of disruption of data center services to be to the tune of $1.7m [14]. The extreme heat
conditions exerted pressure on the data center infrastructure, highlighting the susceptibility of such facilities
to environmental factors and the potential impact on the digital services they support.

Similarly, Twitter’s Sacramento data center was shut down due to unprecedented heat wave in 2022, rendering
Twitter in a non-redundant state. Twitter had to take some emergency procedures to block unnecessary
updates during this time to prevent total blackout of some services in the event another data center was shut
down. by closing its Sacramento data center [15]. Fortunately, the other 2 data centers were not located
in the same region that were experiencing heat wave. These incidents show the importance of climate in
data center planning and management, as extreme weather events can pose substantial challenges to the
reliability and continuity of digital services. In fact, major technology companies are aiming to create Data
Centers in colder regions to save on some cooling costs. Additionally, they want to segregate and spread out
different and redundant data centers across different geographical locations to ensure similar environmental
factors do not affect all the centers at the same time.

3. Factors affecting data center placement
Section 3 seeks to understand the various considerations that influence the location of data centers. These factors range from costs, political factors, disasters, climate, internet connectivity, and the economy.

3.1 What is the cost to run a data center?

In the realm of data center expenses, servers account for 45% of the overall cost. Infrastructure-related costs, encompassing power distribution and cooling systems, represent approximately 25% of the total expenditure [16]. The location of the data center, its tier level, and technical support and Service Level Agreements promised by the data center are the main factors that determine the operating costs.

Location significantly influences the cost, keeping in view access to high growth markets as well as power supply and even renewable energy sources. Such markets also demand premiums on real estate, construction, and skilled labor, thereby increasing the cost. Additionally, the changing market dynamics forces data centers to continually upgrade their hardware, servers, and software to keep up with the security vulnerabilities identified.

The power consumption of the data center is generally passed on to the user organizations as is. However, electrical utility services and the network, which includes links, transit, and equipment, each contribute around 15% to the overall cost.

Data centers offering 24/7 support and providing additional redundancy and fault tolerance generally command premium pricing. Lower latency and less service interruptions are generally associated with additional costs. The ability to scale the capacity rapidly based on need also adds on to the costs. This also leads to lower utilization rates in the data centers. Risk management and fault tolerance measures have also adversely affected the utilization rates. It becomes imperative to accurately forecast the demand to proactively increase utilization rates of these data centers to bring down the costs since unpredictability of the demands is a major contributor. Rapid spikes in data center usage exceeding the conventional forecasts have mandated that we study and predict the requirements and optimize the usage.

3.2 Data Privacy and other Regulations

Some countries place data privacy regulations and thereby affect location specification of the data center. Digital Personal Data Protection Act (DPDP Act) of India allows processing of data outside of territory of India. However, it can restrict the transfer of personal data to outside of Indian territory. Similarly, General Data Protection Regulation (GDPR), Europe prescribes practices for organizations and data centers to keep sensitive data within country boundaries [17].

EN-50600 is a Europe wide transnational standard that includes comprehensive specification for planning, construction, and operation of a data center [18]. This is a regulatory standard and is more flexible and adaptable to allow for change based on business needs. This is more prescriptive in nature. Privacy laws and regulations are given utmost importance and strict adherence is expected.

3.3 Geopolitical and Country Risks

Operational stability of a data center is affected by Geopolitical and Country risks as well. Russia, Belarus, and Ukraine have not seen any investments in data centers recently, mainly, due to the risks associated with the Russian war. It might have been an appealing market keeping in view the climatic conditions; however, big data center builds ignore them on account of the geopolitical risk.

Economic instability, war and civil unrest impact the geo-political stability of a location in the country. A few years back, Ukraine would have been considered a stable and good data center location. However, the recent war and instability in the region quickly changed the profile too risky since it could disrupt operations and compromise the security of data center operations.

Environmental impact and conflict of sharing of energy resources with local community has recently caused cessation of plans of creating a new data center in Netherlands. Major corporations like Meta and Microsoft planned to make significant investments in the country keeping in view its climate, internet infrastructure,
availability of renewable resources of energy and regulations welcoming Technology industry investment. However, there was a community uprising against these corporations when locals realized that the renewable energy sources that were supposed to power their homes especially during energy crisis might be directed to these large corporations. The conflict over sharing natural resources and the subsequent community resistance caused the Government of Netherlands to ban the building of data centers in Feb 2022 [18]. Such environmental and community factors can play an important role in permission and construction of the data centers.

Scalability options that can offer easy expansion, access to skilled resources, and other logistical factors are other Geopolitical factors in the viability and success of a data center.

3.4 Internet connectivity

Proximity of the data centers to fiber networks and geographic Internet Exchanges offers advantages to the data centers in terms of low latency and optimum speed of data transmission [19]. Equally crucial is the proximity to the customer base, as users prioritize fast-loading sites and expect instantaneous access to data. Companies find that users will likely click off when faced with slow loading times, making this a necessity.

Low latency is a critical factor in optimizing the performance of data centers, especially with their proximity to Internet Exchange Points. These exchange points, where multiple networks interconnect, are often strategically clustered at major peering points. When data centers are in close geographic proximity to these Internet Exchanges, organizations utilizing their services benefit from low-latency connections and multiply redundant bandwidth. The round-trip time, influenced by the intricacies of the network’s route, is a crucial determinant of latency. As data rarely travels in a straight line between sender and recipient, passing through various networks, routers, and switches, each component introduces additional latency. Therefore, proximity of data center to its customers lowers the latency and enhances user experience and operational efficiency. This is also obvious from the inaction on relocating data centers from New York after Hurricane Sandy, even though it caused extension damage and enormous emergency efforts were required to keep the data centers running during that time in 2012. The main reason being “Manhattan is one of the most networked places in the world” [20] and it gave users quick access and low latency.

Segregating the data centers across geographically separated areas further reduces latency and enhances user experience irrespective of their size, that is, large scale or even micro. This strategic approach of directing users to nearby data centers is the mainstay of operational efficiency of the data center. Distribution of data centers across diverse locations can optimize latency, ensuring efficient and seamless interactions between their services and globally based users.

3.5 Economy-based factors

Energy usage and power consumption costs constitute a significant portion of operational expenses of data centers. In fact, there is a recent trend to focus on availability and usage of renewable energy sources as a significant factor in cost reduction as well as selecting data center location. In fact, access to output from the windmill park in Netherland was one of the prime reasons for Technology giants like Meta and Microsoft to consider it as a good location for data center creation even though the plan was later abandoned due to community conflict [21]. Renewable energy fulfills sustainability goals as well as helps in reducing energy consumption related operational costs. Strategically considering the energy landscape can enhance the overall efficiency and environmental sustainability of data center operations.

Tax incentives also play a significant role in influencing the placement of data centers [22]. Traditionally, countries and states offer organizations tax and other incentives to boost economic development and employment in their region by establishing data centers. These incentives are in the form of tax exemptions when meeting threshold investment and employment criteria hoping to contribute to economic development of the area. For example, Virginia’s Loudoun saw rapid establishment of data centers in the past fourteen years due to these incentives.

Recently, the nature and types of these incentives have shifted from focusing on growth of data centers to
sustainability of data centers and reducing carbon footprint. Now, incentives or tax breaks are offered on usage of specific energy efficient equipment, infrastructure, and usage of renewable resources. These data centers are also under scrutiny for their usage of energy and water. There are some proposed laws in the pipeline for Virginia that will focus on placing restrictions on data center incentives based on their resource consumption. Similarly, Legislators in Idaho are proposing to limit the maximum amount of financial benefits an organization can reap from the existing tax benefits.

All these tax exemptions provide a financial advantage in the establishment of data centers in those regions. Furthermore, areas with business-friendly internet regulations facilitate smooth data center operations and ensure compliance with regulatory frameworks. Energy costs and tax incentives are major factors affecting the placement of data centers even though labor laws and land costs in specific areas may impact budgeting and cost-effectiveness.

3.6 Disaster-based effects

Natural disasters like Tornadoes, earthquake, flooding, hurricanes, etc. can severely hamper data center operations. Take the example of the Peer 1 data center in New York in 2012. Although it survived the onslaught of Hurricane Sandy, it required enormous efforts on the part of the employees and even volunteers to keep the operations running during the hurricane. They had not envisioned that the basement would be flooded, and it would hamper getting fuel pumped to backup generators located on 17th floor that were burning 40 gallons of fuel an hour on a 200-gallon fuel tank. “Fuel bucket brigade” volunteers carried fuel in buckets up 17 floors throughout the ordeal and were barely saved. Another data center, Fog Creek, in the vicinity was not so fortunate and went down during that time [20].

The data center locations are, therefore, planned considering the stability and disposition of a location to natural disasters. The potential disruptions and expected downtime due to these natural disasters highlights the need to consider locations that are comparatively stable [23]. It has been established that there is a high correlation between natural disasters like earthquakes and network failure. This coupled with aftershocks and subsequent power failure further exacerbates the damage. For example, earthquake and Tsunami in Japan in 2011 led to power outages affecting 1500 telecom buildings during the main shock and an additional 700 in the aftershock. Damage associated with such disasters causes not only the data center locations to be inaccessible but also can cause significant downtime and loss of customer data. In fact, the susceptibility of the location to such disasters coupled with terrorist / cyber attacks targeting populated cities and critical infrastructure plays a significant role in the choice of a location for the data centers. It also shows the emergence of new threat landscape and disaster and the need to re-evaluate such threats / disasters periodically.

Additionally, factors such as power availability, stable power grids, access to renewable energy sources, and electricity costs play a significant role in determining the suitability of a location for hosting a data center.

In response to these risks, a disaster-aware data center and content placement approach has been developed to mitigate associated risks. This strategy involves periodically evaluating the disaster / threat probabilities and placing data centers and their contents in locations with minimal risk based on these probabilities. However, the dynamic nature of disaster probabilities and content properties introduces complexity and warrants periodic monitoring to assess if the vulnerabilities have changed over time and are there new threat actors detected, since such changes can alter the risk landscape. Considering the access latency requirements of users, the optimal placement of data centers is determined by analyzing expected content loss due to unreachability, factoring in the finite number of link-disjoint routes between the user node and a data center. This approach seeks to strike a balance, ensuring that data centers are not too far from users while also minimizing the risk of content loss during disasters or unreachability.

3.6 Climate conditions

Weather and climatic conditions also have a significant influence on the location consideration for a data center. Efficient cooling of servers and systems is important for the data center servers to function and prevent data loss and network failures. Locations with warmer temperatures require additional power and /
or water for cooling the servers. Similarly, higher altitudes, characterized by lower air density, pose challenges to air-based cooling systems’ efficiency [24]. Data center operators must adapt cooling strategies to address variations in air density at different altitudes, ensuring optimal performance and preventing overheating.

Water availability is impacted by climatic conditions and since it is a commonly used resource for cooling in data centers, it becomes a crucial factor in evaluating the feasibility of water-based cooling systems and consequently, the data center location.

Climatic humidity also affects the data center location. Extreme humidity levels can adversely affect data center equipment, leading to corrosion or static electricity issues. Therefore, maintaining optimal humidity levels is essential to preserve equipment integrity and prevent operational disruptions.

Additionally, climate significantly impacts energy requirements for cooling; warmer climates necessitate more energy to maintain lower temperatures critical for optimal equipment performance, while colder climates may require less cooling infrastructure, resulting in lower operational costs. Regions with naturally cooler temperatures can leverage free cooling methods, such as outside air or water-based systems, reducing the reliance on mechanical cooling and saving energy. Furthermore, data center operators consider climate-related regulations and incentives for using renewable energy sources, which may vary by location, when deciding on the placement of their facilities.

4. Opportunities

Data center placement does not always bring the associated risks, it also presents numerous opportunities for contribution to local economies and community. That is why several states or regions incentivize data center placement in identified regions. The significant capital investment made in building and maintaining data centers stimulates economic growth by generating employment opportunities and enhancing local infrastructure. Contributions to local tax revenues, including property taxes, sales taxes on equipment and services, and other levies, play a crucial role in supporting public services and infrastructure projects.

Additionally, the construction and operation of data centers can positively impact local property values and real estate markets, influencing land use planning and development in the surrounding areas. Moreover, direct employment opportunities, ranging from data center technicians and engineers to security personnel and administrative staff, contribute to job creation for residents. Indirectly, the presence of a data center stimulates job opportunities in related industries such as construction, maintenance, and support services, fostering a broader positive economic impact.

In Virginia, Amazon Web Services (AWS) holds data centers across three different counties which in total contribute $21.31 billion to local GDP. Oregon, Ohio, and Northern California have seen similar benefits of $6.4, $2.23, and $2.11 billion respectively [25]. AWS further estimates that their investments support jobs across the US which generate an estimated output of $65.15 billion for local businesses through thousands of jobs. The data centers in these areas play a vital role in processing and storing data for the operations of businesses that rely on AWS for their cloud computing needs. The economic impact of data center placement cannot be understated and is a boon to local communities.

5. Optimal Placement

After analyzing the various factors surrounding data center placement, it is essential for us to understand how to weigh these factors and utilize this information to make strategic decisions.

The method I propose is an algorithmic method that balances all factors discussed to make a decision. First, creating a comprehensive rating system from 1-100 to analyze internet connectivity and latency, energy costs, political stability, disaster frequency, and climate to measure the quality of various locations imputed into such an algorithm. Such information may be found available online through global indexes and indicators that use a variety of statistics and analyses to measure such factors in specific localities. This way the burden of personally researching each location can be left to existing databases which have already proven their accuracy. Finally, create a weighted sum to compare all locations. Greater weightage should be given
to more significant issues like disasters, climate, and political stability in the highest tier because these are a necessity and should have higher priority. The next most important is internet connectivity because it is needed to reduce slow speeds and ensure traffic runs smoothly. Finally, energy costs and other considerations should be prioritized, but not as much as the other factors. However, my methodology applies mainly to the more prevalent mega data centers we see today operated by large corporations. Smaller data centers or companies that have differing perspectives may change this weightage to their own satisfaction to meet their specific needs.

6. Conclusion

Overall, this paper delves into factors affecting the placement of data centers and understanding their changing landscape and role. This paper investigates various technical, statistical, economical, and historical deliberations to be considered when deciding on location of a data centers encompassing factors like costs, geopolitical stability, disaster resilience, climate conditions, and internet connectivity. Notably, it highlights the risks associated with such vulnerabilities, and emphasizes the need to adopt a comprehensive approach in deciding placement not only to prevent but also to effectively manage and mitigate such risks in unforeseen circumstances.

The challenges faced by data centers at the global stage are more pressing than ever due to rapidly changing technological and climatic environments. The paper further looks at opportunities to gain from data center placement and proposes using a comprehensive approach to weigh and balance the many factors in a practical and systematic manner to ensure the lasting success of data centers.

7. Glossary

Cloud Computing: a technology providing on-demand access to resources like servers, storage, databases, and software for the usage of businesses with various models.

Latency: the time delay between data traveling from a data center to a user’s device.

Bandwidth: the maximum rate at which data can be transmitted across a network, usually expressed in bits per second (bps).

Infrastructure as a Service (IaaS): a cloud computing model that provides servers, storage, and other infrastructure entirely over the Internet without any physical maintenance or ownership required.

Server: a computer or system that processes, stores, or manages data and services within a specified network.

Switch: a device connecting other devices within a local area, typically used to forward data between servers.

References


[7] Cisco, Data Center: Load Balancing Data Center Services SRND, 2004


17] https://gdpr.eu/tag/gdpr/


[26] CBRE Research. “Global Data Center Trends 2023”

