International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-12 Issue-8, July 2023

Data Centre Efficiency Enhancement by Metrics Oriented Approach to Revamp Green Cloud Computing Concept

Saumitra Vatsal, Satya Bhushan Verma

Abstract: ***Cloud computing inherits the sharing of data from a pool of resources existing in data centres whenever demanded. The imminent requirement for this purpose is the data centre's ability to fulfil this coveted objective. A simultaneous increase in energy consumption poses a challenge to achieving energy-efficient peak performance. Energy-efficient metrics play a crucial role in achieving the desired purpose of safeguarding the environment. These metrics aim to enhance the system's proficiency. An increase in energy efficiency results in reduced consumption of energy resources, as these resources are mostly non-renewable and are the primary source of carbon and heat emissions from operational data centres. No single metric is capable of achieving enhanced energy-efficient performance in a data centre. Therefore, a collective utilisation of selected metrics about power, performance, and network traffic can improve the energy-efficient capability of data centre communication systems. The testing platform for such metrics is based on specific architectures, including D-Cell, B-Cube, Hyper-Cube, and Fat-Tree three-tier architectures.***

Keywords: Cloud Computing, Green Cloud Computing, Energy-Efficiency, Metrics

I. INTRODUCTION

Cloud-based computing is currently regarded as fundamental for IT operations globally, as it has emerged so prominently that it is successfully replacing traditional business models. It has enabled access to a plethora of software available online, along with services, in a virtual environment, thereby reducing the investment requirement for IT infrastructure. It requires only a connectivity-enabled IT infrastructure, which demands a significantly lower investment. The mode of operation is based on the "pay-as-you-go" concept, which enables direct focus on core business and the utilisation of internet-related IT services to ensure fully justified payment on demand.

Manuscript received on 15 April 2023 | Revised Manuscript received on 17 June 2023 | Manuscript Accepted on 15 July 2023 | Manuscript published on 30 July 2023.
*Correspondence Author(s)
Saumitra Vatsal*, Department of Computer Science and Engineering, Shri Ramswaroop Memorial University, Barabanki (U.P.), India. E-mail: s.vatsall@gmail.com, hod.cse@srmu.ac.in, ORCID ID: 0000-0002-5182-4507
Dr. Satya Bhushan Verma, Department of Computer Science and Engineering, Shri Ramswaroop Memorial University, Barabanki (U.P.), India. ORCID ID: 0000-0001-8256-2709

The operation of cloud computing banks on the network is related to the distribution of data centres geographically all over the world. Hence, the assessment of data centre performance becomes crucial for a comprehensive understanding of data centre-related operational facts, which serve as a foundation for designing and constructing the next generation of systems to revamp cloud computing.

The performance and efficiency of data centres can be evaluated by a correct assessment of the amount of electrical energy supplied to the system versus its actual conversion into computing power. This titration is performed using metrics. Selecting the proper metrics is crucial for achieving real performance. The performance evaluation of optimization techniques, which include task-scheduling, resource-scheduling, resource-allocation, resource provisioning and resource execution demands the right metrics to be utilised for securing optimization objectives [1]. Since the intensity of load on a virtual machine is inferred from the level of resource utilisation, it is assumed that virtual machine utilisation is proportionate to the CPU-related resource capacity being utilised for task execution. It also represents the resource-related demands to show whether the level of utilization is high or low [2]. As it is well known that the functioning of data centres is highly energy-intensive, their operation requires a massive amount of energy. The IT and cooling equipment consume 75% of this energy, while the remaining 25% is dissipated as power loss in distribution and facility operation systems. The proper performance metrics are crucial for evaluation of orchestration techniques focusing on Cloud, Fog and Edge computing related monitoring based on the MAPE-K concept (Monitoring, Analyzing, Planning, Execution – Knowledge) as introduced by IBM [3], [4], [5]. Orchestration development efforts are needed to fulfil the objectives of latency minimization, energy management and cost reduction [6], [7]. Numerous metrics have been proposed for the assessment of efficiency in energy distribution [8], [9], [10] and cooling [11], [12] about present research of energy parameters. The monitoring-related metrics can minimize fault tolerance, which is a ratio of the number of faults detected to the total number of faults existing about software or hardware associated factors [4]. They may also address the issue of the degree of heat generated by data-centre infrastructure while executing the tasks [13]. The heat generation in a data centre during tasks' execution on the underlying infrastructure poses a challenge for Cloud computing and environmental sustainability together, which can be suitably addressed by thermal-aware Cloud computing metrics [14].

Power Usage Effectiveness (PUE) is a prominent and widely used metric [15] that measures the amount of energy, in the form of electricity, shared by IT equipment. Due to the generic nature of existing metrics, it becomes difficult to differentiate the individual IT sub-systems. This can be explained by the fact that existing metrics are unable to distinguish between the efficiency of communication about data centres and that related to computing servers. The reason is that both are screened by a standard envelope of IT equipment [16], [17]. To ensure the ideal situation, the proportionality between network device power consumption and the workload must have a direct relationship. However, power consumption is practically exhibited as both fixed and variable. The fixed one pertains to line cards and the switch chassis, which maintain a constant value even when the switch is idle. The variable one pertains to transmitters operating in active mode, which adjust the transmission rate. It represents the proportionality of energy, addressing the interrelation between energy consumption and the load offered by the system or component. The current network switches exhibit a less than 8% difference in consumption between peak and idle modes of activity. If an unused port is turned off, only 1 to 2 watts are saved [18]. During the computing process, as and when a proportionality alignment is established between the workload and the power consumption of computing servers and their desired level of functioning, vis-à-vis a lower degree of utilisation, a concern arises in the form of a power consumption issue about the network. Sometimes, the consumption of power by the network accounts for 50% of the overall data centre's power consumption [19]. Sometimes, a metric-based approach is required for assessing the interrelation between energy proportionality and attached network devices. This approach serves a valuable purpose in investigating the twin aspects related to the energy proportionality of the system as a whole, as well as individual network devices. A distinction between IT equipment-related communication systems and the assessment of performance levels faces a limiting challenge due to non-computing communication processes [20]. By applying pertinent metrics for energy-aware and sustainable Cloud computing, the optimisation of energy consumption and resource utilisation can be suitably addressed through automatic resource scheduling using the energy-aware autonomic resource scheduling technique (EARTH) [21], [22]. Latency, available bandwidth, or both can serve as limiting factors. Severe constraints significantly challenge communication latency during voice conferencing, although high bandwidth availability is not the primary requirement for effective communication. The latency can be alleviated by latency-aware auto-scaling metrics, which consider multi-objective optimisation in their execution while performing real-world orchestration techniques [23]. On the contrary, functions like video streaming and cloud storage require high bandwidth for transferring large data masses but remain unaffected by network delays. In the process of cloud computing, a high traffic load is generated, but synchronisation is achieved through tight delay constraints. The evaluation of endeavour software, online transaction processing is effectively managed by metrics for security, which protect data on the cloud by using access controls and data encryption [4], [24].

The direction of the flow of information serves as the basis for categorising cloud communication, specifically into intra-cloud and cloud-to-user. The former pertains to traffic within a data centre, while the latter is related to cloud users localised in the access network domain. As evaluated by CISCO, the fastest growing data centre component is the network traffic [25]. Thus, it is inferred that to secure good performance, factors such as architectures, networking solutions, and protocols must be addressed appropriately. Specific warm-up steps are required for making the metrics available to the real world after transferring them from simulators by an approach which includes an initial relaxation of every metric, followed by further performance evaluations to secure quantitative solutions with the perspective of their real-world implementations [26].

The paper's related contribution synopsis can be unveiled as follows:

- Existing metrics analysis about energy efficiency, cooling and infrastructure effectiveness associated with data centres (Section 2).
- Assessment of communication systems related to energy efficiency and performance based on the development of a metrics-based framework (Section 3).
- An analytical comparison and evaluation of metrics based on collected traffic-related traces derived from functional data centres (Section 4) and (Section 5).

II. DATA CENTRE METRICS – BACKDROP

Metrics which address the performance, efficiency and quality of systems' cloud-related computational performance can be categorised as follows:

2.1. Energy and Power Efficiency

The metrics known as Data Centre infrastructure Efficiency (DCiE) and PUE are of paramount importance for this category. The ratio of power consumption related to the facility versus IT equipment is designated as PUE. The inverse of PUE constitutes DCiE. There exists an analogy between Energy Utilization Effectiveness (EUE) and PUE, but EUE is relatively energy-based rather than power-based [27]. The assessment of reused energy outside the data centre can be measured by two parameters, which are Energy Reuse Factor (ERF) and Energy Reuse Effectiveness (ERE) [28], while the assessment of average UPS load vis-à-vis overall UPS capacity can be assessed by the load factor of Uninterruptible Power Supply (UPS) [29]. Data Centre Energy Productivity (DCEP) and Power to Performance Effectiveness (PPE) [30] are reckoned as another two generic metrics. They respectively assess the energy consumption to evaluate the effectiveness of work and IT equipment in terms of power consumption, as well as the interrelation between performance output and energy consumption.

Retrieval Number: 100.1/ijitee.F95320512623
DOI: 10.35940/ijitee.F9532.0712823
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

2.2. AIR and Environment Related Metrics

The Return Temperature Index (RTI) serves as the most appropriate metric for environmental and air management. It facilitates the evaluation of energy performance during the isolation of heated and cooled air streams for effective air management. Evaluation of absorption of recirculated air by a rack is addressed by the Recirculation Index (RI), while the review of the air-flow fraction incoming and outgoing from a rack following a desired path is addressed by the Capture Index (CI).

2.3. Statistics About Cooling Efficiency

To address rack cooling efficiency by manufacturers' thermal guidelines, the related metric is known as the Rack Cooling Index (RCI). The assessment of power required for operating cooling equipment is addressed by Data Centre Cooling System Efficiency (DCCSE) [31]. It is represented by a ratio that exists between the average power consumption of the cooling system and the data centre load. The assessment of fans and air-circulation efficiency is addressed by Airflow Efficiency (AE) [31]. A technique designated as free-cooling is addressed by the Air Economizer Utilization Factor (AEUF) [31] which evaluates annual hourly duration for which air economizer taps external low temperature environment to secure the process of chilling the water.

Traditional communication networks, which focus mainly on network latency, bandwidth, and error rates, are addressed by metrics that prioritise these as primary indicators. Several other works address a few different aspects of data centre network analysis [32], [33]. The evaluation mainly focused on latency assessment and bandwidth for pairs of running virtual machines [32] along with analysis of capacity and related costs of the data centre network [33].

III. CLOUD COMPUTING DATA CENTRE-RELATED COMMUNICATION METRICS

These metrics relate to ascertaining the performance and energy efficiency-oriented factors associated with the cloud computing aspects of data centre communication systems. Applications in cloud computing are communication-intensive except for High Performance Computing (HPC) [20]. Hence, specific parameters can dramatically affect system performance, including error rate, bandwidth capacity, and latency. These metrics, by allowing finer granularity, can also address the undesirable aspect that exists with performance and power-related metrics, namely their inability to segregate communication systems and IT equipment classes.

Dynamic resource provisioning addresses the issue of avoiding long latencies that arise from the severe variability of growing demand during typical working hours for web applications. Demand typically rises during working hours and decreases during the nighttime and early morning hours. It also addresses the issue of diverse orchestration concerns about resource allocation, task scheduling, task placement, server consolidation, virtual machine migration and load balancing [1], [34]. The related cost minimization for gaming applications is addressed by minimizing the latency by targeting energy as an objective through a dynamic resource provisioning technique [35], [36]. The response time-related metrics evaluate the performance of productivity applications along with graphics-oriented workloads for assessing the execution of workload management for the arrival of a task at load admission to reciprocating corresponding response to the user [37]. The workload can be deciphered as processing in a given period for handling the processing of work in Cloud computing [38]. The reliability of nodes is of paramount importance, as it addresses the issue of changing adaptability under uncertain situations, such as failure in particular functions within a virtual environment. Metrics play an instrumental role in revamping the issue of adaptability by monitoring the Edge layer hosting the churn nodes, which are deciphered as those hosts which continuously can leave or join the network [39]. The prediction methods are invaluable for analysing the monitored parameters to obtain more accurate values for the planner. The relevant metrics for evaluation of accuracy of prediction include metrics like MAPE-K, Root-Mean-Square-Error (RMSE), MAE, Average Median, MSE, $R^{2,}$ and PRED [40], [41].

These metrics can be classified under the following three categories:

- Metrics about power.
- Metrics about performance.
- Metrics about network traffic

Energy efficiency in communication systems is addressed by power-related metrics that analyse how much electric power is converted into work of delivering information while executing networking and other related activities. Performance-related metrics address the analysis of capacity, communication rate, and information delivery latency. Lastly, access to the nature of transmitted information and measurement of overheads related to traffic is secured by metrics related to network traffic.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 1: Cloud Computing Related Metrics of Communication

Type	Metric	Full Form	Remarks
Power	CNEE	Communication Network Energy Efficiency	Required power for delivering a bit of information.
	NPUE	Network Power Usage Effectiveness	Power ratio between total power and consumed power in IT networking.
	EPC	Energy Proportionality Coefficient	Proportionality of system or device related energy levels.
Performance	UDCL	Uplink/Downlink Communication Latency	Data centre gateway versus servers' related time lag.
	UDHD	Uplink/Downlink Hop Distance	Data centre gateway versus servers' related hop distance.
	ISCL	Inter-Server Communication Latency	Communication time lag between servers.
	ISHD	Inter-Server Hop Distance	Distance of hop in between servers.
	DAL	Database Access Latency	Database access time.
	BOR	Bandwidth Oversubscription Ratio	Operational bandwidth with fully loaded state.
	UDER	Uplink/Downlink Error Rate	Data centre gateway and servers inter-distance path related error rate.
	ISER	Inter-Server Error Rate	Error rate between server network paths.
	ALUR	Average Link Utilization Ratio	Average traffic related load on communication links of a data centre.
	ASDC	Average Server Degree Connectivity	Per server mean number of network links.
Network Traffic	ITR	Internal Traffic Ratio	Internal data centre related exchange of traffic.
	ETR	External Traffic Ratio	Data centre related traffic efflux.
	MMTR	Management and Monitoring Traffic Ratio	Traffic generated due to monitoring and management.
	MMTE	Management and Monitoring Traffic Energy	Traffic related power consumption due to monitoring and management.

3.1. Metrics About Power

3.1.1. Communication Network Energy Efficiency

Transformation of network-related electricity is needed to fulfil the goal of delivering the information. For the measurement of its efficacy, the relevant metric is expressed as follows.

$$\text{CNEE} = \frac{\text{Network equipment power consumption}}{\text{Effective network throughput capacity}} \quad \ldots\ldots\ldots (1)$$

Data centre-related networking hardware comprises components that participate in delivering inter-server information, including server components such as routers, network switches, Network Interface Cards (NICs), and communication links. In the context of servers, the value of NICs is worth considering because servers without NICs simply discharge the function of computing and are not considered as communication equipment. The computing servers are subjected to end-to-end network-related maximum throughput, which is known as adequate network throughput capacity. The unit of CNEE is watts per bit per second. It is the energy required to deliver a unit of information. It is also equivalent to joules per bit.

3.1.2. Network Power Usage Effectiveness

It represents the portion of power consumed for data centre operational functions related to the communication system.

$$\text{NPUE} = \frac{\text{IT equipment total power consumption}}{\text{Network equipment power consumption}} \quad \ldots\ldots\ldots (2)$$

Strictly speaking, NPUE specifies the consumed power fraction due to IT equipment used for operating a data centre communication system. Similarly, the fraction of energy utilised as power by the server is also measured by PUE. The values of NPUE can range from 1 to infinity. It can be further elaborated that, if NPUE stands out as the consumption of 6 watts by IT equipment, it can be inferred that 1 watt is utilised for the operation of network equipment. An NPUE value of 1 means that network equipment is consuming all of the IT-related power, which is an undesirable state because, in such a scenario, no power seems to be available for servers' computational activities. It is not necessary that a value of 1 for NPUE necessarily indicates inefficiencies in the network; instead, it should be interpreted as an energy-efficient upgrade of computing servers.

3.1.3. Energy Proportionality Coefficient

In an ideal situation, the workload and energy consumption of network devices should be directly proportional; however, in practice, network switches or computing servers are not energy proportional. Many servers, even in an idle state, exhibit 66% power consumption at peak activity levels. Regarding switches, this ratio could be even higher, reaching up to 85%. The normalised load depicts how much variance is observed by comparing a steady workout for an ideal situation against fluctuating workloads. It is a system's energy consumption-related offered load function. It is represented as a straight line for a perfect case, as shown in Figure 1, where each increment of load, l, is accompanied by a corresponding equal increase in power, P, representing the consumption of energy. However, as revealed in practice, power consumption is not linear.

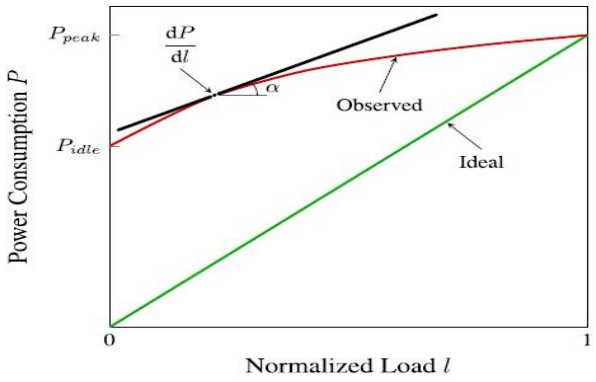

Figure 1: Representing Proportionality of Energy

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The line of inclination represents the proportionality of energy consumption to an ideal case. Analysis of this variation can be adjudged by drawing a tangent line for every point about the observed curve. Taking the observed function's first derivative into consideration, the angle α of this tangent line can be procured.

$$\tan\alpha = \frac{dP}{dl} \dots\dots\dots (3)$$

The energy proportionality measurement is defined based on tan α:

$$EPC = \int_0^1 \sin 2\alpha \; dl = \int_0^1 \frac{2tan\alpha}{1+tan^2\alpha} dl \dots\dots\dots (4)$$

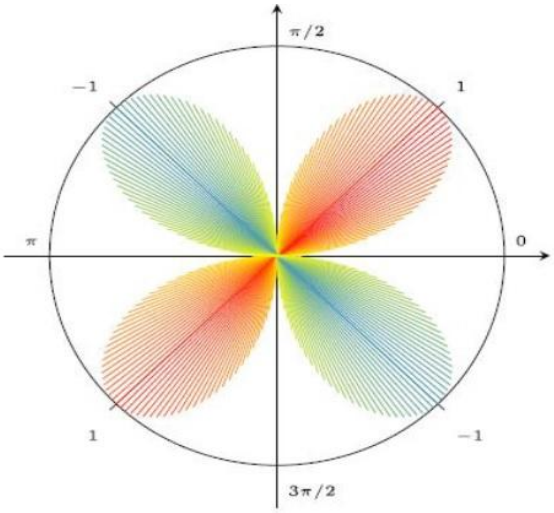

Figure 2: Energy Proportionality Coefficient (EPC)

Figure 2 depicts polar coordinates related to various values of α plotted by EPC metric values. The EPC is equal to 1 if α = π/4, thus representing that every increment of system-related load causes equivalent energy consumption. If α is equal to -π/4, it means that there is an equivalent energy consumption decrease for each system load-related increment, thus making EPC equal to -1. If energy consumption is constant and independent of load, then α is equal to 0, making the EPC also equal to 0. If α is equal to π/2, then it is represented as an asymptote of the power consumption function.

The different routing strategies, which may be energy-aware or energy-unaware, can play a role in affecting energy-related proportionality [19], [42]. The Energy Proportionality Index (EPI) pertains to the evaluation of the difference between calibrated power and ideal power. Ideal power represents power that must be consumed against a fully energy-proportional state. If the EPI value equals 0, it can be interpreted as indicating that energy consumption is in synergy with workload. A 100% EPI value indicates a fully energy-proportional device status. Thus, the expression for EPI can be calibrated against idle and peak power only.

The evaluation of the ratio existing between consumption of power during idle and peak state is measured by Linear Deviation Ratio (LDR) and Idle-to-Peak-power-Ratio (IPR) [43] concerning change in observed power consumption from fully proportional case respectively. IPR values are indicative of energy proportionality design if they tend to be zero. LDR, on the other hand, serves as a parameter for measuring maximum deviation or power consumption through a linear representation that connects the values of power consumption during peak and idle states. If the values of LDR are positive, it indicates an above-line power measurement. The values of LDR against negative values are indicative of power measurement positioned beneath the line. A consumption of power of a perfect linear representation indicates that LDR is zero.

EPC can address the energy proportionality of a device for any observed power consumption. EPI and IPR depend on the consumption of power at their idle and peak functional state. When the state is fully proportional, then the dependency of LDR remains subordinate to the absolute peak deviation value. EPC has the power to identify functions of both constant and non-constant nature.

3.2. Metrics About Performance

These metrics address delay, bandwidth, and also specific parameters such as the degree of server-specific connectivity.

3.2.1. Network Latency

Applications related to cloud show a high sensitivity for communication delays [20], [44]. Hence, for safeguarding two crucial factors, such as Service Level Agreements (SLAs) and Quality of Service (QoS), the ability to monitor and control network latency becomes an issue of paramount importance. The factors that comprise network delays include signal transmission time, as well as delays related to queuing and packet processing at every node. Hence, proportionality is established between latency-related communication and the number of hops existing between senders and receivers of information. Based on the number of hops, the Uplink/Downlink Hop Distance (UDHD) or Uplink/Downlink Communication Latency (UDCL) are considered the most prominent latency-related metrics. UDCL is the metric used to measure the time in seconds for a downlink request that reaches the computing server, or to measure the uplink request that leaves the data centre network for its destination to the end user. A faster response time is secured if UDCL is of a smaller size, which is hosted by computing servers near the data centre gateway.

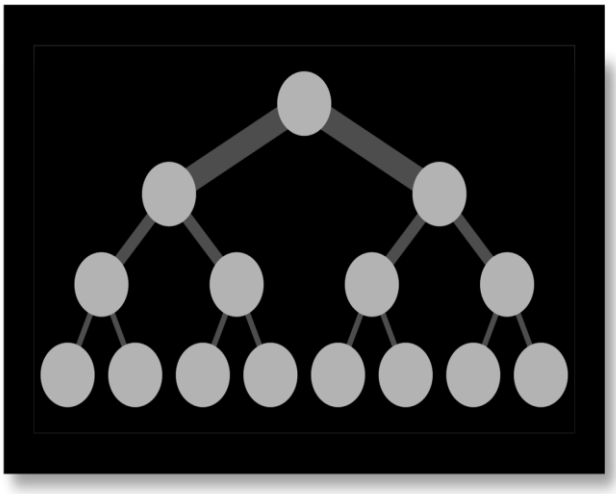

Figure 3(A): Fat Tree Three–Tier Architecture

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

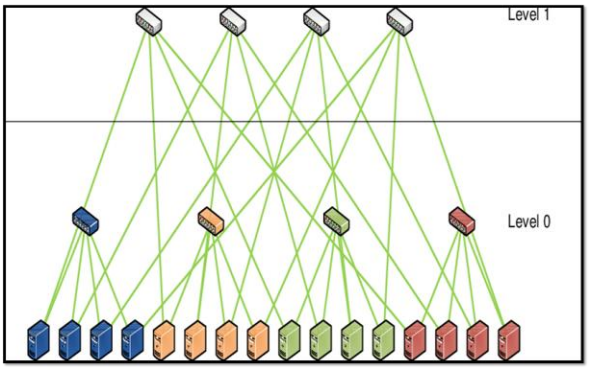

Figure 3(B): Bcube Architecture

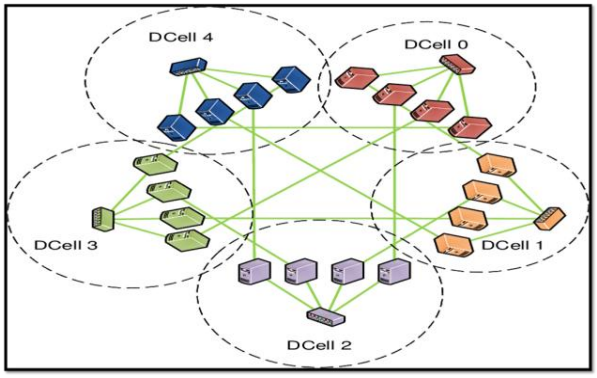

Figure 3(C): Dcell Architecture

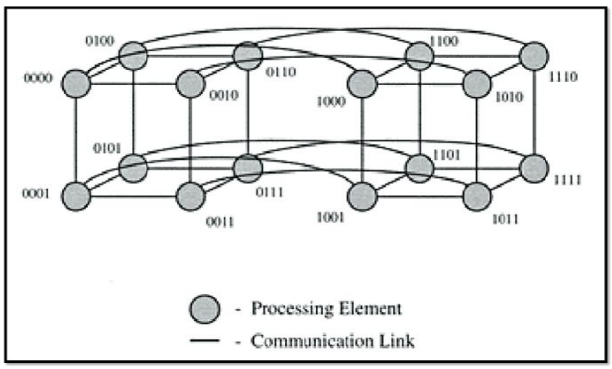

Figure 3(D): Optically Braced Hypercube

Figure 3: Various Data Centre Architectures' Related Communication Latency

The time taken in seconds by one task or the number of hops needed to communicate with another task is addressed by a metric called Inter-Server Hop Distance (ISHD) or Inter-Server Communication Latency (ISCL).

$$\text{ISHD} = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} h_{ij} \dots\dots\dots (5)$$

Where N denotes the total number of servers, hij denotes the number of hops between servers i and j.

ISCL and ISHD well address cloud applications which can exhibit parallelism in execution. Their task execution requires an exchange of data to exhibit brisk performance in network-related architectures, along with minimised inter-server hops and shorter inter-server delays. However, unrelated single-server confined applications exhibit immunity to inter-server delays. Apart from measuring average values, analysing the distribution of inter-server delays is of paramount importance. Deviation of small values signifies data centre networks related to computing servers located at small distances apart (examples: Al-Fares et al proposal, Portland [45] and VL2 [46]) and thus permitting any of the server related placement of the task, independent of its location. B Cube [47] and D Cell architectures which are server-centric, impart high inter-server delays to data centres. In this state, the consolidation of heavily communicating tasks becomes highly beneficial for reducing network delays and enhancing performance.

The Database Access Latency (DAL) represents the third delay-related metric, which is the average round-trip time (RTT) in seconds, existing between the data centre's database and servers during the process of computing. Database serves as a source to provide data for most of the cloud-related applications for storage and retrieval [20]. Performance can be enhanced by minimising the time required for transferring a query to a destination and receiving subsequent data. By applying data replication techniques, it can serve as an alternative to bring databases physically closer [48]. The delays above are illustrated in Figure 3 (a, b, c) concerning three-tier, B-Cube, and D-Cell data centre architectures.

3.2.2. Bandwidth Oversubscription Ratio

It is a network switch-related ratio that exists between the aggregate bandwidths of ingress and egress. This can be exemplified in a three-tier topology (Figure 3a), where the Top-of-Rack (ToR) switches have two 10Gb/s links, supporting nearly 48 servers each, with 1 Gb/s link connectivity.

This is equivalent to the Bandwidth Oversubscription Ratio (BOR) of (48Gb/s)/(20Gb/s), which is 2.4:1. This is equivalent to the per-server bandwidth of (1Gb/s)/(2.4), or 416 Mb/s, under full load. At the aggregation level, a bandwidth aggregation of (1.5):1 further takes place. Hence, each switch is comprised of eight 10 Gb/s links to the core network and twelve 10 Gb/s links to access the network.

The outcome is that the bandwidth available per server could be as low as (416 Mb/s)/1.5, equal to 277 Mb/s, when considered against a fully loaded topology. BOR exhibits a value of 1 because server-centric architectures avoid introducing points related to bandwidth oversubscription. The estimation of the minimum non-blocking bandwidth for each server can be achieved by computing BOR. If the available bandwidth becomes insufficient due to the generation of more traffic by computing servers, it leads to a congested state in the ToR and aggregation switches. As a result, packets are dropped from overflowing buffers, resulting in performance degradation of cloud-related applications.

3.2.3. Network Losses

Link errors may result in the loss of data packets during their transmission in a data centre network, resulting in communication delays for transport layer-based TCP protocol-related retransmissions. Hence, screening becomes imperative for ensuring the desired level of QoS and performance at the packet level and end-to-end error rates at the bit level.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The interconnecting links are dissimilar in data centres, considering the fat tree three-tier architecture, as shown in Figure 3a. It incorporates 10 Gb/s optical links, where the per-link Bit Error Rate (BER) ranges from 10^-12 to 10^-18 in both the core and access layers. The functioning of the access layer is governed by twisted-pair Gigabit-based Ethernet technology, where a BER of 10-10 is the range. Based on the link characteristics of the network and topologies, the average end-to-end error rates can be calculated by considering communication paths, such as server-to-gateway and server-to-server.

The two metrics that measure error rate estimation are Uplink/Downlink Error Rate (UDER) and Inter-Server Error Rate (ISER), the latter being the second one.

$$\text{UDER} = \frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L}\text{BER}_{nl} \dots\dots\dots (6)$$

Where N represents the number of computing servers, L denotes the hierarchical layers in the network topology, and BERln represents the link between layer l in the BER connecting server n and the data centre gateway.

Evaluation of the inter-server communication average error rate is performed by Inter-Server Error Rate (ISER):

$$\text{ISER} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\text{BER}_{ij} \dots\dots\dots (7)$$

Where N signifies the number of computing servers, and BERij represents the BER of the interconnecting paths between server i and server j. The total of BERs related to all links existing between servers i and j represents BERij.

The importance of error rate measurement is noteworthy when addressing the sensitivity of cloud-related applications to identify transmission-related errors and hardware-related faults.

3.2.4. Average Link Utilization Ratio

It represents the average load of traffic on data centre-related communication links.

$$\text{ALUR} = \frac{1}{N_i}\sum_{n=1}^{N_i}u_n \dots\dots\dots (8)$$

Where un is the utilisation ratio and Ni is the number of type i links. ALUR, being an aggregate network metric, can address traffic distribution and load levels concerning data centre networks. This can also identify network hotspots and be instrumental in avoiding network congestion-related performance degradation in cloud computing.

It is possible to measure ALUR separately for a three-tier fat tree topology, addressing aggregation, access, and network-related core segments. If any of these segments is highly congested, it will signal the need to initiate an increase in network link capacity and switches, or re-evaluate bandwidth oversubscription ratios. It is possible to measure ALUR on a server-to-switch and server-to-server segment basis for BCube and DCell topologies.

3.2.5. Average Server Degree Connectivity

Topologies of data centres are switch-centric or server-centric, depending on the data centre design strategy. The fat tree architecture is switch-centric, as it connects a single ToR switch with a single link only. The BCube and DCell architectures exemplify server-centric architecture, enhancing network capacity by providing re-adaptation capabilities in the event of node or switch total dysfunction. The enhancement of network-related capacity is achieved, revealing a high degree of connectivity that also makes the entire topology fault-tolerant, thus facilitating load balancing. However, this enhanced degree of connectivity culminates in increased network power consumption due to the use of more links and NICs for this purpose. For the analysis of high-quality connectivity in computing servers, the value of this metric needs to be estimated.

$$\text{ASDC} = \frac{1}{N}\sum_{n=1}^{N}c_n \dots\dots\dots (9)$$

Where N denotes the total number of data centre servers, and Cn denotes the connectivity of a small number of servers connected to other servers, devices, and switches.

3.3. Metrics About Network Traffic

An analysis report of network traffic properties is instrumental in evaluating the efficacy of data centre-related communication systems. Network traffic-related classification as internal or external is based on the direction of signalling.

Internal traffic constitutes 75% of the entire network-based communication within a modern data centre [25]. It comprises cloud application database interaction among independent tasks that are executed in parallel. Internal communication within a data centre remains subject to metric DAL-based database-related access delays, metric BOR-based network availability, and inter-server latency, which metric ISCL/ISHD addresses. The latency and bandwidth of a data centre network's uplink and downlink paths deliver unaffected performance for internal communication. The external traffic, which addresses the end-users, includes cloud applications-related traffic and inter-data centre traffic [25]. External traffic exhibits high sensitivity for available bandwidth, which metric BOR addresses. It is also sensitive to latency in the uplink and downlink paths, which UDCL/UDHD addresses. The inter-server bandwidth and communication latency, which the metric ISCL/ISHD addresses, remain unaffected in terms of external communications-related performance. External and internal data centre traffic exists in proportion as described below:

- The ratio between internal data centre traffic and total data centre traffic is called the Internal Traffic Ratio (ITR).

$$\text{ITR} = \frac{\text{Internal Traffic}}{\text{Total Data Centre Traffic}} \dots\dots\dots (10)$$

- External Traffic Ratio (ETR) represents the fraction of traffic that exits the data centre network.

$$\text{ETR} = 1 - \text{ITR} = \frac{\text{External Traffic}}{\text{Total Data Centre Traffic}} \dots\dots\dots (11)$$

Apart from classifying network traffic based on the target point, identifying messaging related to a user or application from the rest of the traffic becomes imperative for securing and managing aspects related to monitoring and network management. Monitoring is essential for the operation of a communication network. The transmissions that address resolutions, such as ARP and RIP/OSPF-type routing, are handled by management operations.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Management operations can also be attributed to problem detection and control messaging, such as ICMP, while SNMP can address the monitoring of operations for traffic. The Management and Monitoring Traffic Ratio (MMTR) represents network management-related traffic overhead.

$$\text{MMTR} = \frac{\text{Management and Monitoring Traffic}}{\text{Total Data Centre Traffic}} \dots\dots\dots (12)$$

The Communication Network Energy Efficiency (CNEE) metric and Management and Monitoring Traffic Energy (MMTE) metric address energy consumption during the management of network traffic, excluding traffic related to transportation applications.

$$\text{MMTE} = \text{CNEE} \cdot \text{Management and Monitoring Traffic} \dots\dots\dots (13)$$

The unit of MMTE is Joule, which represents the energy utilised by communication-related equipment for securing network-related operational status. Ideally, MMTE should exhibit near-zero values while a significant portion of energy is linked to traffic-related applications.

Processing of network is well analysed by evaluating network-related traffic at macro/microscopic levels for justifying the paramount importance of data centre traffic knowledge [49], and also for securing design traffic engineering solutions [50]. Evaluating the inter-dependencies of executed workloads and for estimation of optimised communication for several data centres which are geographically distributed [51].

IV. NUMERICAL EXAMPLES BASED ON EVALUATION

Here, category-wise metrics are proposed to address performance-related, power-related, and network traffic-related aspects for the sake of numerical comparison and evaluation.

4.1. Scenario of Evaluation

Although several data centre architectures exist [52], [53] four architectures viz. fat tree, three-tier [45], [46] BCube [47], DCell and optically cross-braced hypercube (OH) [54] are being considered for evaluation purposes. For the sake of comparison, these architectures are configured to provide backup for 4096 computing servers.

The fat tree three-tier topology comprises 128 racks with eight core switches and 16 aggregation switches, serving these servers. The interconnectivity between core and aggregation switches, as well as between aggregation access switches, is established through 10 Gb/s, 0.24 µs optical links. The computer servers and access network ToR switches are connected by 1 Gb/s, 0.01 µS twisted pair links.

The DCell and BCube architectures incorporate the arrangement of 4096 computing servers in groups of n=8. This results in the provision of a BCube architecture of level k=4, featuring commodity switches in three layers for each group of servers, and the DCell architecture of level k=2. The commodity switches are interconnected with computing servers through 1 Gb/s links. The link length for the lowest layer is 2 metres, with link lengths of 10 and 50 metres for the middle and uppermost layers, respectively. Numerous load balancers utilising 50-meter-long, 40 Gb/s optical fibres are used to establish connectivity between the gateway router and the data centre network.

In an OH architecture for supporting 4096 servers, twelve hypercube dimensions are required. For the sake of interconnection, this requirement is fulfilled by 12.212/4 = 2,228 two-by-two optical switches.

It is assumed that the support offered by optical fibres facilitates single-mode light propagation using a 1550 nm operating wavelength in all architectures.

4.2. Power Related Metrics: Evaluation

For evaluating power-related metrics, the metrics included here are NPUE, CNEE, and EPC, which cover different architectures of data centres.

4.2.1. Network Energy and Power Usage Effectiveness Evaluation

The calculation of power consumption is imperative for procuring network and computing server equipment related to NPUE and CNEE when the data centre load increases. For provoking new servers to acquire a fully operational profile from their dormant state, extra network switches are not needed, thus making the increase in load to the data centres non-linear. However, if a new rack needs to be activated, it requires power for the Top-of-Rack (ToR) switch, as well as core and aggregation switches. It is depicted as a three-tier topology, as shown in Figure 4.

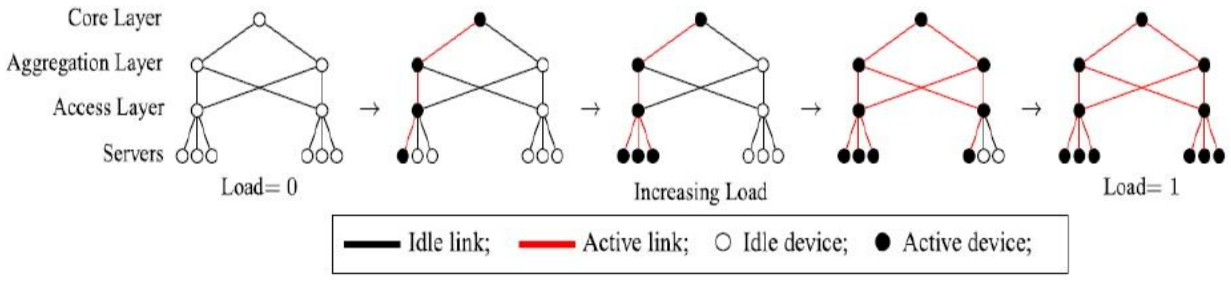

Figure 4: Powering Up Equipment as Data Centre Load Increases

For estimation of single server power consumption, most preferred models like Huawei Tecal RH228H V2, IBM System x3500 M4 and Dell PowerEdge R720 are considered for computing consumption of average power during peak and idle mode of performance [55]. The power consumption of servers can be estimated using Dynamic Voltage and Frequency Scaling (DVFS) at the server level. The power consumption P(l) is expressed as below:

$$P(l) = P_{idle} + \frac{P_{peak} - P_{idle}}{2} \cdot \left(1 + l - e^{-\frac{1}{\tau}}\right) \dots\dots\dots (14)$$

Where l represents load of the server, τ represents utilization levels for securing asymptotic power consumption in (0.5, 0.8) range.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

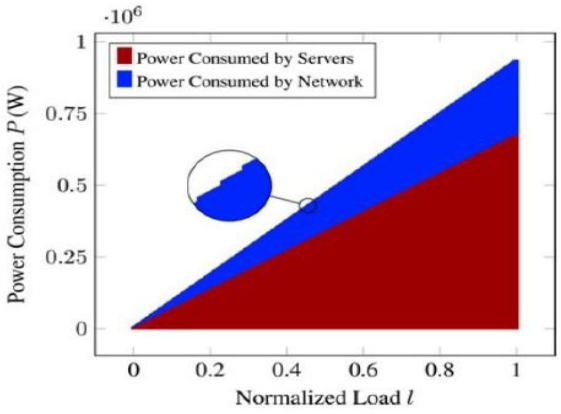

Figure 5: Fat Tree Three–Tier Data Centre Related to Power Consumption

Core layers and the aggregation issues about fat tree three-tier architecture are addressed by HP FlexFabric 11908, DCell architecture and BCube architecture related commodity and ToR switches are addressed by HP 5900 AF and for OH architectures, PRISMA-II, two optical switches are considered. The normalised power consumption concerning the fat tree three-tier architecture is depicted in Figure 5. The power consumption by the Network Interface Card is not inclusive of the power consumption by the servers, but it is additive to the power consumption by the network. A previously idle rack, when it assumes an active wake-up state in a server, is then represented as a leap, as shown in the zoomed portion of Figure 5. It culminates in power consumption-related network non-proportionality due to the activation of the core layer, aggregation, and access switches.

Table 2: Power-Related Metrics' Evaluation

	ARCHITECTURES			
METRICS	Three-Tier	BCube	DCell	OH
CNEE	0.203 mJ/bit	0.109 mJ/bit	0.027 mJ/bit	0.033 mJ/bit
NPUE	3.58	2.50	6.86	5.99

Concerning all four data centre architectures considered, the computation of CNEE is shown in the first row of Table 2. Several layers related to bandwidth oversubscription of high degree take the CNEE value to the highest level in the case of a fat tree three-tier topology. This results in energy utilisation for supporting higher bit-rates that are not fully utilised by the servers. On the contrary, the throughput achieved is 100% of the network capacity for both DCell and BCube architectures. The factors of dependence for CNEE include total network-related power consumption, while bandwidth-related oversubscription addresses the CNEE-related sensitivity issue. This fact explains why BCube-related CNEE is higher than DCell-related CNEE. BCube is comprised of (k+1).NK (2048) number of commodity switches, while in the case of DCell, it contains a single commodity switch for each group of n servers (512). OH architecture is comprised of 12,228 two-by-two optical switches, which consume significantly less power compared to commodity switches that support BCube and DCell architectures. This makes the CNEE value computed concerning the OH topology more identical to the DCell value than the value for BCube.

With the help of NPUE, an assessment of overall power effectiveness is possible while considering the energy required for transferring single-bit-related information in data centre networks. Thus, BCube appears to require the highest amount of power, as its NPUE value is the lowest among the three. This is exemplified by the fact that DCell requires more switches to be incorporated compared to a three-tier architecture. However, it includes commodity switches, which exhibit significantly less power consumption than aggregation and core-level switches. For OH architecture, although individual optical switches consume less power, the NPUE value is still lower in DCell than in OH architecture. In the case of an OH architecture, two leading causes of network power consumption are the high number of active ports and transceivers for each server.

4.2.2. Energy Proportionality Evaluation

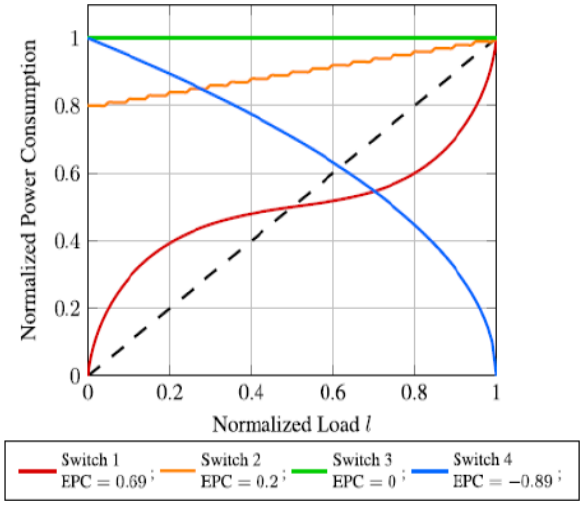

Figure 6: Different Network Switches Related Power Consumption Profiles

Figure 6 illustrates the normalised power consumption of multi-network switches for estimating computed EPC values with various profiles. For the ideal case, the EPC value equals one and is represented by a dashed line. Switch one exhibits a curvilinear behaviour when the load value falls within the intermediate range (0.2, 0.8); then, the power consumption increases at a rate lower than the workload increase rate. However, when the load values are low (less than 0.2) or high (greater than 0.8), power consumption increases more rapidly than the incoming workload. Hence, EPC becomes equal to 0.69 as a result. The energy consumption with a realistic profile is achieved in the case of switch 2, with an EPC value of 0.2, thus exhibiting a large idle part with power consumption in a step ladder pattern, attributed to its communication ports. Therefore, it very closely resembles the case profile presented by Switch 3. Switch 3 is insensitive to workload; hence, the value of EPC is exhibited as zero by switch 3. For switch 4, the EPC value is displayed to be negative (-0.89). This indicates that the device consumes less energy when the workload increases.

4.3. Performance Related Metrics: Evaluation

This section presents the evaluation results of the proposed metrics concerning connectivity (ASDC), energy losses (UDER, ISER), and network latency (UDCL,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

ISHD, UDHD, DAL, ISCL), excluding the metrics of BOR and ALUR.

The process of bandwidth multiplexing does not receive any backup points concerning server-centric architectures, and the process of oversubscription persuades BOR metrics to become equal to 1. In the computational method, per-link traffic statistics become a requirement to address the ALUR metric, and they can be procured either from detailed traces or by direct measurement from data centres during runtime.

4.3.1. Network Latency, Network Losses and Server Degree Connectivity

Test packets with values of 40 bytes and 1500 bytes were transmitted to evaluate ISCL, UDCL, UDER, DAL, and ISER, corresponding to the maximum Ethernet transmission unit and TCP acknowledgement, respectively. The UDCL and ISCL are addressed by measuring one-way transmission delay and by measuring round-trip delay in the case of DAL. To address signal losses, the copper cables and optical fibres are assigned BER values of 10^-12 and 10^-14, respectively. The queuing delays can be ignored concerning Ethernet inter-frame gap due to the absence of any other traffic in the data centre network. The configuration of a single packet network delay comprises link propagation delay (Dp) and transmission delay (Dt). The expressions for Dt and Dp represent ratios that exist between packet size s and link rate r in the case of Dt, while the ratio between link length L and signal propagation speed P defines Dp.

$$D_t = \frac{S}{R}, \qquad D_p = \frac{L}{P} \ldots\ldots\ldots (15)$$

Table 3: Precise Values Comparison Chart of Architectures

METRICS		ARCHITECTURES			
		Three-tier	BCube	DCell	OH
40 B	UDCL	1.45µs	1.38µs	1.19µs	1.16µs
	ISCL	1.98µs	3.93µs	4.73µs	1.2µs
1500 B	UDCL	15.7µs	14.47µs	15.50µs	14.42µs
	ISCL	28.34µs	73.72µs	93.92µs	24.47µs
DAL		18.11µs	17.15µs	17.15µs	15.71µs
UDHD		4	3	3	3
ISHD		5.78	7.00	8.94	3.25
UDER		$1.03 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$	$1.02 \cdot 10^{-12}$
ISER		$1.77 \cdot 10^{-12}$	$4.21 \cdot 10^{-12}$	$5.34 \cdot 10^{-12}$	$2.00 \cdot 10^{-14}$
ASDC		1	4	2.79	12

The physical characteristics of a medium are defined by P. In the case of copper, it amounts to a fraction (two-thirds) of the velocity of light, *c*. In optical fibre, the velocity of light is calibrated by considering the refractive index, which is taken to be equal to 1.468 for glass fibre. The network latency losses and connectivity metrics are presented in Table 3, along with their results. It reveals that the OH architecture supports better internal communications by considering ISCL, ISER, and ISHD, as all of these have lower values compared to other architectures. Because OH architecture has the highest ASDC value, it provides genuine assurance for providing short paths even between distant servers. A three-tier topology offers better support for internal communications compared to BCube and DCell. It appears to be a paradoxical state because the connectivity degree measures, along with ASDC for a three-tier architecture, are minimal compared to the other architectures. Although DCell and BCube both exhibit superior interconnectivity, they still require numerous hops for communication between servers that are located at a distance. BCube and DCell are primarily dependent on copper links, which poses a challenge of a very high inter-server error rate, which ISER addresses. On the contrary, the gateway and server-related rate of error remains lower in the case of BCube and DCell, as measured by UDER, because the outgoing packets from the server have to execute a lesser number of hops to reach the gateway.

4.4. Network Traffic Related Metrics Evaluation

To evaluate the metrics MMTR and MMTE, which pertain to network traffic, packet traces are obtained from real data centres UNIV1 and UNIV2. These traces and application data address OSPF, ICMP, RIP, and ARP flows.

The two-tier architecture supports both data centres, with approximately half an hour of traffic assigned to the data traces of UNIV1 and UNIV2 data centres, respectively. To evaluate the fraction of network management and traffic monitoring, the computation of MMTR is performed, which yields values of 0.79% and 0.025% for UNIV1 and UNIV2 data centres, respectively.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The results reveal that the UNIV1-related network is managed with lesser efficiency, despite being equipped with a smaller number of network devices.

Table 4: Evaluation of Management and Monitoring of Traffic Energy

	ARCHITECTURES			
MMTE	**Three-Tier**	**BCube**	**DCell**	**OH**
UNIV1	169.19 J	90.62 J	22.23 J	27.31 J
UNIV2	30.98 J	16.59 J	4.09 J	5.00 J

Table 4 addresses the energy consumption of the data centre network for processing and delivery management, as well as traffic monitoring. The metric MMTE addresses energy consumption related to traffic monitoring, considering both UNIV1 and UNIV2. It is revealed that the value is lower for all architectures related to UNIV2. The energy consumption is lowest for transferring a single bit of information in the case of DCell; therefore, DCell consistently outperforms other architectures. The fat tree three-tier architecture seems to be the most energy-hungry (vide CNEE values in Table 2).

The choice of resource allocation strategy employed certainly influences most of the metrics presented and discussed. The successful operation of a data centre mainly depends upon two parameters: the monitoring of infrastructure and the energy efficiency achieved. The process of virtual machine or workload migration increases the magnitude of monitoring and management of traffic flux concerning MMTR and MMTE metrics. The internal traffic increases in cases of metrics, such as ETR and ITR, may cause changes in the ALUR value. Thus, it focuses on the imminent conclusion that these metrics provide an essential platform for advancing the concept of resource allocation in the realm of cloud data centres, thereby paving the way for securing a novel solution for network-oriented scheduling.

V. DISCUSSION

Every Cloud provider is required to offer the service with the avoidance of SLA violations that arise due to an increase in task execution time. It can be suitably addressed by metric-oriented evaluation of workload-related performance in the domain of Cloud computing, although the SLA violations are yet to be well-defined for Edge computing and IoT applications [56]. The metrics framework offered will undoubtedly prove vital for assessing, comparing, and monitoring communication systems in a data centre. The metrics related to power enable operators of data centres to optimise investments in equipment and interconnects for networking by assessing energy efficiency with finer granularity. The delays, error rate, and throughput associated with a network are monitored and evaluated in detail using performance-related metrics. In Cloud or Edge networking, the number of hosts requested from the provider by an executor can be analysed by metric for provisioned resources, while the de-provisioned resource metric indicates the opposite action [57], [58].

These metrics have provided an energy-efficient umbrella cover for revamping relevant cloud applications, such as SaaS, which address both internal communication and user-facing communication. These metrics not only help ensure but also guarantee SLA and QoS to customers. Lastly, metrics about network traffic enable the development of infrastructure-aware resource allocation policies, facilitating the creation of effective traffic management. The metric framework for cloud-related data centre networks justifies itself in ensuring the expansion of planning capacity. It helps in capacity enhancement for designing an optimised data centre of the future.

The currently available data centre monitoring systems, such as VMware vCenter Log Insight or Cisco Prime Data Centre Network Manager, can easily merge and integrate these proposed metrics. Information needed for computing these metrics, such as link utilisation levels, error rates, or runtime power consumption, is already provided by the majority of data centre monitoring systems. For example, the data centre-related internal and outgoing data flux can be differentiated by simply examining the destination addresses. Monitoring of data related to a server is addressed by software within a data centre, such as the status of links. Thus, the computation of the ASDC metric remains subordinate to the average number of active links. The up-to-date statistical availability of traffic- and link-related information enables the design of scheduling solutions and network-aware resource allocation.

A top-level comparison of data centre architectures that have been evaluated is provided in Table 5. The measurement values and evaluation details are supplied with precision in Section 4, whereas in Table 5, these values are categorised as high (H), medium (M), and low (L) for simplicity. In the case of a three-tier architecture, the network-related total functioning capacity is limited due to high bandwidth oversubscription, resulting in the highest per-bit energy consumption per unit. DCell has the lowest per-bit energy consumption ratio. Power usage effectiveness of DCell is highest, making it the most “green” architecture amongst all the architectures. BCube is comparatively less efficient in terms of power usage effectiveness, as it incorporates the maximum number of switches. Upon scrutinising the communication latency, we find that the three-tier fat tree architectures favour server-to-server related internal communications. In contrast, the distributed data centre architectures, such as DCell and BCube, have smaller traffic-related paths directed outside of the data centre. However, OH architecture, which is server-centric, is capable of significantly reducing the number of hops between servers placed distantly. Consequently, the support they provide to internal communications is better than that of hierarchical architectures.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Table 5: Performance-Based Comparison Chart of Different Architectures

	METRICS										
ARCHITECTURES	CNEE	NPUE	UDCL	UDHD	ISCL	ISHD	DAL	UDER	ISER	ASDC	MMTE
Three-tier	H	M	M	M	L	M	H	L	H	L	H
BCube	M	L	M	L	H	H	M	L	H	M	M
DCell	L	H	M	L	H	H	M	L	H	M	L
OH	L	M	M	L	L	L	L	L	L	H	L

Values are categorized as (L) Low, (M) Medium and (H) High.

VI. CONCLUSION

Achieving network-related efficiency in communication is a prime objective to be fulfilled when dealing with cloud computing data centres. The achievement of this desired objective is facilitated by a set of metrics that address energy efficiency in the computing arena, which is discussed in this paper. These energy efficiency metrics address the perspectives related to energy, performance, and traffic. The metrics related to power measure the efficiency of the procedural task that converts electricity into information delivery. The metrics related to performance are used for analysing error rates, network latency, and available bandwidth, which are conventional characteristics of a communication system. The metrics regarding network traffic provide insight into the energy consumed in conveying traffic to different categories, as well as into the characteristics of the traffic. The framework of metrics has been assessed and validated for three-tier hierarchical and distributed (Hypercube, BCube, and DCell) data centre architectures. Several properties related to these architectures were revealed by the results obtained. These metrics will undoubtedly prove constructive for academicians and industry specialists.

ACKNOWLEDGEMENTS

I am grateful to all those involved in this endeavour.

DECLARATION

Funding/ Grants/ Financial Support	No, we did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	Saumitra Vatsal wrote this research paper. Dr. Satya Bhushan Verma supervised it.

REFERENCES

1. M. S. Aslanpour, S. S. Gill and A. N. Toosi, "Performance evaluation metrics for Cloud, Fog and Edge computing: A review, taxonomy, benchmarks and standards for future research", Internet of Things, 12, 100273, 2020. [CrossRef]
2. S. H. H. Madni, M. S. A. Latiff, and Y. Coulibaly, "Recent advancements in resource allocation techniques for cloud computing environment: a systematic review," Cluster Comput., vol. 20, no. 3, pp. 2489–2533, 2017. [CrossRef]
3. M. S. Aslanpour, M. Ghobaei-Arani, M. Heydari, and N. Mahmoudi, "LARPA: A learning automata-based resource provisioning approach for massively multiplayer online games in cloud environments," Int. J. Commun. Syst., p. e4090, 2019. [CrossRef]
4. S. S. Gill, I. Chana, M. Singh, and R. Buyya, "CHOPPER: an intelligent QoS-aware autonomic resource management approach for cloud computing," Cluster Comput., pp. 1–39, 2017. [CrossRef]
5. S. Singh, I. Chana, M. Singh, and R. Buyya, "SOCCER: self-optimization of energy-efficient cloud resources," Cluster Comput., vol. 19, no. 4, pp. 1787–1800, 2016. [CrossRef]
6. M. S. Aslanpour, M. Ghobaei-Arani, and A. Nadjaran Toosi, "Auto-scaling web applications in clouds: A cost-aware approach," J. Netw. Comput. Appl., vol. 95, 2017, doi: 10.1016/j.jnca.2017.07.012. [CrossRef]
7. S. Singh and I. Chana, "A survey on resource scheduling in cloud computing: Issues and challenges," J. Grid Comput., vol. 14, no. 2, pp. 217–264, 2016. [CrossRef]
8. M. Uddin, A. A. Rahman and A. Shah, "Criteria to select energy efficiency metrics to measure performance of data centre," Int. J. Energy Technol. Policy, vol. 8, no. 3, pp. 224-237, 2012. [CrossRef]
9. L. Wang and S. U. Khan, "Review of performance metrics for green data centers: A taxonomy study," J. Supercomput., vol. 63, no. 3, pp. 639–656, 2013. [CrossRef]
10. The Green Grid, "Harmonizing global metrics for data center energy efficiency," White Paper, 2014.
11. R. Tozer and M. Salim, "Data centre air management metrics – practical approach," Proc. of 12th IEEE Intersoc. Conf. Therm. Thermomech. Phenom. Electron. Syst., pp. 1–8, 2010. [CrossRef]
12. S. Flucker and R. Tozer, "Data centre cooling air performance metrics," Proc. of CIBSE Techn. Symp., Leicester, pp. 1–16, 2011. [CrossRef]
13. S. S. Gill, I. Chana, M. Singh, and R. Buyya, "RADAR: Self-configuring and self-healing in resource management for enhancing quality of cloud services," Concurr. Comput. Pract. Exp., p. e4834, 2018. [CrossRef]
14. S. S. Gill et al., "ThermoSim: Deep learning based framework for modelling and simulation of thermal-aware resource management for cloud computing environments," J. Syst. Softw., p. 110596, 2020. [CrossRef]

15. E. Volk, A. Tenschert, M. Gienger, A. Oleksiak, L. Siso, and J. Salom, "Improving energy efficiency in data centres and federated cloud environments: Comparison of CoolEmAll and Eco2-Clouds approaches and metrics," Proc. of 3rd Int. Conf. Cloud Green Comput., pp. 443–450, September 2013. [CrossRef]
16. D. Cole (2011), "Data centre energy efficiency-looking beyond the PUE," Available Online at: http://www.missioncriticalmagazine.com/ext/resources/MC/Home/Files/PDFs/WP_LinkedIN%20DataCenterEnergy.pdf, White Paper.
17. D. Kliazovich, P. Bouvry, F. Granelli, and N. Fonseca, "Energy consumption optimization in cloud data centers," Cloud Services, Networking, and Management, N. Fonseca and R. Boutaba, Eds., Wiley: Hoboken, NJ, USA, May 2015. [CrossRef]
18. B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee and N. McKeown, "ElasticTree: Saving energy in data centre networks" Proc. of 7th USENIX Conf. Netw. Syst. Des. Implementation, vol. 3, pp. 19–21, 2010.
19. D. Abts, M. R. Marty, P. M. Wells, P. Klausler and H. Liu, "Energy proportional datacenter networks," Proc. of ACM SIGARCH Comput. Archit. News, vol. 38, no. 3, pp. 338–347, 2010. [CrossRef]
20. D. Kliazovich, J. E. Pecero, A. Tchernykh, P. Bouvry, S. U. Khan and A. Y. Zomaya, "CA-DAG: Modelling communication-aware applications for scheduling in cloud computing," J. Grid Comput., pp. 1–17, 2015. [CrossRef]
21. S. Singh and I. Chana, "EARTH: Energy-aware autonomic resource scheduling in cloud computing," J. Intell. Fuzzy Syst., vol. 30, no. 3, pp. 1581–1600, 2016. [CrossRef]
22. S. S. Gill et al., "Holistic resource management for sustainable and reliable cloud computing: An innovative solution to global challenge," J. Syst. Softw., vol. 155, pp. 104–129, 2019. [CrossRef]
23. F. A. Salaht, F. Desprez, and A. Lebre, "An overview of service placement problem in fog and edge computing," ACM Comput. Surv., vol. 53, no. 3, pp. 1–35, 2020. [CrossRef]
24. S. S. Gill and R. Buyya, "SECURE: Self-protection approach in cloud resource management," IEEE Cloud Comput., vol. 5, no. 1, pp. 60–72, 2018. [CrossRef]
25. Cisco, "Cisco Global Cloud Index: Forecast and Methodology, 2012-2017," White paper, 2013.
26. Y. Li, Y. Chen, T. Lan, and G. Venkataramani, "Mobiqor: Pushing the envelope of mobile edge computing via quality-of-result optimization," in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 1261–1270.
27. J. Yuventi and R. Mehdizadeh (2013), "A critical analysis of power usage effectiveness and its use as data centre energy sustainability metrics," Available Online at: http://cife.stanford.edu/sites/default/files/WP131_0.pdf [CrossRef]
28. The Green Grid, "A metric for measuring the benefit of reusing energy from a data centre," White Paper, 2010.
29. (2009), "UPS load factor," Available Online at: http://hightech.lbl.gov/benchmarking-guides/data-p1.html
30. (2009), "Data centre efficiency-beyond PUE and DCiE," Available Online at: http://blogs.gartner.com/david_cappuccio/2009/02/15/data-center-efficiency-beyond-pue-and-dcie/
31. P. Mathew, "Self-benchmarking guide for data centres: Metrics, benchmarks, actions," Lawrence Berkeley National Laboratory, 2010. [CrossRef]
32. H. Khandelwal, R. R. Kompella and R. Ramasubramanian, "Cloud monitoring framework," White Paper, 2010.
33. L. Popa, S. Ratnasamy, G. Iannaccone, A. Krishnamurthy, and I. Stoica, "A cost comparison of datacenter network architectures," Proc. 6th Int. Conf., pp. 16:1 16:12, 2010. [CrossRef]
34. Y. Al-Dhuraibi, F. Paraiso, N. Djarallah, and P. Merle, "Elasticity in cloud computing: state of the art and research challenges," IEEE Trans. Serv. Comput., vol. 11, no. 2, pp. 430–447, 2018. [CrossRef]
35. L. Zhou, C.-H. Chou, L. N. Bhuyan, K. K. Ramakrishnan, and D. Wong, "Joint Server and Network Energy Saving in Data Centres for Latency-Sensitive Applications," in 2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2018, pp. 700–709. [CrossRef]
36. C.-H. Chou, L. N. Bhuyan, and D. Wong, "µDPM: Dynamic Power Management for the Microsecond Era," in 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2019, pp. 120–132.
37. S. S. Gill, P. Garraghan, and R. Buyya, "ROUTER: Fog-enabled cloud-based intelligent resource management approach for smart home IoT devices," J. Syst. Softw., vol. 154, pp. 125–138, 2019. [CrossRef]
38. M. Abdullahi and M. A. Ngadi, "Hybrid symbiotic organisms search optimization algorithm for scheduling of tasks on cloud computing environment," PLoS One, vol. 11, no. 6, p. e0158229, 2016. [CrossRef]
39. A. J. Ferrer, J. M. Marques, and J. Jorba, "Ad-Hoc Edge Cloud: A Framework for Dynamic Creation of Edge Computing Infrastructures," in 2019 28th International Conference on Computer Communication and Networks (ICCCN), 2019, pp. 1–7. [CrossRef]
40. S. S. Gill et al., "Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges," Internet of Things, vol. 8, p. 100118, 2019. [CrossRef]
41. S. S. Gill and R. Buyya, "A taxonomy and future directions for sustainable cloud computing: 360 degree view," ACM Comput. Surv., vol. 51, no. 5, pp. 1–33, 2018. [CrossRef]
42. Y. Shang, D. Li and M. Xu, "A comparison study of energy proportionality of data centre network architectures," Proc. 32nd Int. Conf. Distrib. Comput. Syst. Workshops, pp. 1–7, 2012. [CrossRef]
43. G. Varsamopoulos and S. K. S. Gupta, "Energy proportionality and the future: metrics and directions," Proc. 39th Int. Conf. Parallel Process. Workshops, pp. 461–467, 2010. [CrossRef]
44. P. Fan, J. Wang, Z. Zheng and M. Lyu, "Toward optimal deployment of communication-intensive cloud applications," Proc. IEEE Int. Conf. Cloud Comput., pp. 460–467, 2011. [CrossRef]
45. R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya and A. Vahdat, "PortLand: A scalable fault-tolerant layer 2 data centre network fabric," Proc. ACM SIGCOMM Comput. Commun. Rev.., vol. 39, no. 4, pp. 39–50, 2009. [CrossRef]
46. A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel and S. Sengupta, "VL2: A scalable and flexible data centre network," Proc. ACM SIGCOMM Comput. Commun. Rev.., vol. 39, no. 4, pp. 51–62, 2009. [CrossRef]
47. C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang and S. Lu, "BCube: A high-performance, server-centric network architecture for modular data centres," ACM SIGCOMM Comput. Communi. Rev.., vol. 39, no. 4, pp. 63–74, 2009. [CrossRef]
48. D. Boru, D. Kliazovich, F. Granelli, P. Bouvry and A. Y. Zomaya, "Energy-efficient data replication in cloud computing datacenters," Springer Cluster Comput., vol. 18, no. 1, pp. 385–402, 2015. [CrossRef]
49. T. Benson, A. Akella and D. A. Maltz, "Network traffic characteristics of data centres in the wild," Proc. 10th ACM SIGCOMM Conf. Internet Meas., pp. 267–280, 2010. [CrossRef]
50. T. Benson, A. Anand, A. Akella and M. Zhang, "Understanding data centre traffic characteristics," ACM SIGCOMM Comput. Commun. Rev.., vol. 40, no. 1, pp. 92–99, 2010. [CrossRef]
51. Y. Chen, S. Jain, V. K. Adhikari, Z.-L. Zhang and K. Xu, "A first look at inter-data centre traffic characteristics via Yahoo! datasets," Proc. IEEE INFOCOM, pp. 1620-1628, 2011. [CrossRef]
52. M. Bari, R. Boutaba, R. Esteves, L. Granville, M. Podlesny, M. Rabbani, Q. Zhang and M. Zhani, "Data centre network virtualisation: A survey," IEEE Commun. Surveys Tuts., vol. 15, no. 2, pp. 909–928, Apr.-Jun. 2013. [CrossRef]
53. A. Hammadi and L. Mhamdi (2014), "A survey on architectures and energy efficiency in data centre networks," Comput. Commun., 40, 0, pp. 1–21, Available Online at: http://www.sciencedirect.com/science/article/pii/S0140366413002727 [CrossRef]
54. H. Cui, D. Rasooly, M. R. N. Ribeiro and L. Kazovsky, "Optically cross-braced hypercube: A reconfigurable physical layer for interconnects and server-centric datacenters," Proc. Opt. Fibre Commun. Conf. Expo. Nat. Fibre Optic Eng. Conf., pp. 1–3, Mar. 2012. [CrossRef]
55. (2012), "Dell PowerEdge R720 Specification Sheet," Available Online at: http://www.dell.com/downloads/global/products/pedge/dell-poweredge-r720 -spec-sheet.pdf

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

56. S. Tuli, R. Mahmud, S. Tuli, and R. Buyya, "Fogbus: A blockchain-based lightweight framework for edge and fog computing," J. Syst. Softw., vol. 154, pp. 22–36, 2019. [CrossRef]
57. M. S. Aslanpour, S. E. Dashti, M. Ghobaei-Arani, and A. A. Rahmanian, "Resource provisioning for cloud applications: a 3-D, provident and flexible approach," J. Supercomput., 2017, doi: 10.1007/s11227-017-2156-x. [CrossRef]
58. M. S. Aslanpour and S. E. Dashti, "Proactive Auto-Scaling Algorithm (PASA) for Cloud Application," Int. J. Grid High Perform. Comput., vol. 9, no. 3, pp. 1–16, Jul. 2017, doi: 10.4018/IJGHPC.2017070101. [CrossRef]

AUTHOR PROFILES

Mr. Saumitra Vatsal is a research scholar currently pursuing doctoral research in DCSE-IoT (Department of Computer Science and Engineering, Institute of Technology) at Shri Ramswaroop Memorial University (SRMU), Uttar Pradesh, India. He is pursuing research in the topic of green Cloud computing under the supervision of Dr. Satya Bhushan Verma. Green Cloud computing is an environmentally conscious approach that curtails excessive energy consumption by Cloud data centres and reduces the carbon footprint's emission into the environment. He has attended numerous seminars, conferences and a symposium of international acclaim. He also has numerous publications of international repute. His research interests include Cloud computing in the backdrop of green computing.

Dr. Satya Bhushan Verma has completed a Ph.D. in Computer Science and Engineering from the National Institute of Technology, Durgapur, West Bengal, India. His Ph.D. thesis title is "Analysis and Modelling of Palmprint Verification System". He has published several articles in SCI and peer-reviewed journals, and he has also presented two papers at international conferences. He has filed one patent. Currently, he works as a resource person at BBA University, Lucknow (Central University). His areas of interest include biometrics, Computer Vision, Pattern Recognition, and MANET. Dr Satya Bhushan Verma is a member of IAENG (International Association of Engineers), Hong Kong. Dr. Satya Bhushan Verma is currently serving as the Head of Department (HoD) of the Department of Computer Science & Engineering (DCSE) at the Institute of Technology, Shri Ramswaroop Memorial University (SRMU), Uttar Pradesh, India.