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

Saumitra Vatsal, Satya Bhushan Verma

***Abstract:** Cloud computing inherits sharing of data from pool of resources existing in data centres when ever demanded. The imminent requirement for this purpose is proficiency of the data centre for fulfilment of this coveted objective. The pursuit of energy-efficient peak performance level is challenged by a simultaneous hike of energy consumption. The energy-efficient metrics contribute a major role for attainment of desired objective of safeguarding the environment. These metrics address the enhancement of the system's proficiency. An increased energy-efficiency results into reduced consumption of energy resources since these energy resources are mostly non-renewable in nature and are the main source of carbon and heat emissions from operational data centres. As a matter of fact, any individual metric is not capable of achieving enhanced energy-efficient performance in a data centre. Therefore a collective utilization of selected metrics pertaining to 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 certain architectures which include 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 presently is reckoned 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 by virtue of an environment which is deemed to be of virtual nature, curtailing the investment requirement for raising IT infrastructure. It requires only a connectivity enabled IT infrastructure which certainly demands an investment of lesser magnitude. The mode of involved functioning is based on "pay-as-you-go" concept which empowers to focus directly on core business and to utilize internet related IT services for assuring fully justified payment, on demand basis.

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, related with distribution of data centres geographically all over the world. Hence the assessment of data centre performance becomes very crucial for fully understanding data centre related operational facts which serve as a pioneer for designing and constructing the next generation system to revamp cloud computing.

The titration of performance and efficiency of data centres can be evaluated by a correct assessment of amount of electrical energy supplied to the system vis-à-vis its actual conversion into computing power. This titration is done by means of metrics. Selection of correct metrics is critical for execution of real performance. The performance evaluation of optimization techniques which include task-scheduling, resource-scheduling, resource-allocation, resource provisioning and resource execution demands right metrics to be utilized for securing optimization objectives [1]. Since the intensity of load on a virtual machine is inferred from the level of resource utilization it is inferred that virtual machine utilization is proportionate to CPU related resource capacity which is being utilized for execution of tasks. 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 centric therefore their operation requires massive amount of energy. The IT and cooling equipment consume 75% of this energy while rest 25% is dissipated as power loss in distribution and facility operation systems. The right performance metrics are crucial for evaluation of orchestration techniques focussing on Cloud, Fog and Edge computing related monitoring on the basis of 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 assessment of efficiency pertaining to energy distribution [8], [9], [10] and cooling [11], [12] pertaining to present research of energy parameters. The monitoring related metrics can minimize fault tolerance which is a ratio of number of faults detected to the total number of faults existing pertaining with software or hardware related factors [4]. They may also address the issue of degree of heat generated by data-centre infrastructure while executing the tasks [13]. The heat generation in data centre during tasks' execution on 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) currently is reckoned prominent and popular metric used [15] as it conducts the measurement as to how much energy in the form of electricity is shared by IT equipment. Due to the fact of generic nature of existing metrics it becomes difficult in differentiating the individual IT sub-systems. It can be explained by the fact that existing metrics are unable to distinguish between efficiency of communication pertaining to data centre with efficiency related with computing servers and the reason is that both are screened by a common envelope of IT equipment [16], [17]. To assure the ideal situation the proportionality between network device power consumption to the workload must have direct relationship. But practically power consumption exhibits itself as fixed and variable. The fixed one pertains to line cards and also the switch chassis meant for maintaining a constant value even when the mode is idle for a switch while the variable one pertains to transmitters working in active mode to address the rate of transmission. It represents proportionality of energy addressing inter-relation between consumption of energy and system or a component-related offered load. The current network switches depict less than 8% difference of consumption existing during peak and idle mode of activity and if an unused port is turned off then only 1 to 2 watts are saved [18]. During computing process as and when a proportionality alignment is established between workload and consumption of power by computing servers and its level of desired functioning status vis-à-vis lower degree utilization level is attained then the concern crops up in the form of power consumption issue pertaining to network. Sometimes consumption of power by network stands out to be 50% of power consumption pertaining to the overall data centre [19]. Sometimes a metric based approach is required for the assessment of inter-relation pertaining to energy proportionality with respect to attached network devices. This approach serves a valuable purpose for investigating the twin aspects pertaining to energy proportionality of the system as a whole as well as individual network device. A distinction of IT equipment related communication systems and assessment of performance levels face a limiting challenge due to non-computing communication processes [20]. By applying pertinent metrics for energy-aware and sustainable Cloud computing the optimization of energy consumption and resource utilization can be suitably addressed by automatic resource scheduling through energy-aware autonomic resource scheduling technique (EARTH) [21], [22]. In fact the latency or available bandwidth or both together can serve as limiting factors. The communication latency is highly challenged by severe constraints during voice conferencing although the availability of high bandwidth is not the prime requirement. The latency can be alleviated by latency aware auto-scaling metrics which consider multi-objective optimization in their execution while performing the real-world orchestration techniques [23]. On contrary, the functions like video streaming & cloud storage demand high bandwidth for transference of huge data mass but remain unaffected by network delays. In the process of cloud computing high traffic load is generated but synchronization is attained by tight delay constraints. The evaluation of endeavour software, online transaction processing is effectively managed by metrics for security which protect the data on Cloud by using access controls and data encryption [4], [24].

The direction of the flow of information serves as the basis for categorizing cloud communication viz. intra-cloud and cloud-to-user. The former pertains to traffic within a data centre while latter is related with cloud users localized in access network domain. As evaluated by CISCO the fastest growing data centre component is the network traffic [25]. Thus, it is inferred that for securing good performance the factors like architectures, networking solutions and protocols have to be properly addressed. Certain 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-up for further performance evaluations to secure quantitative solutions with the perspective of their real-world implementations [26].

The paper related contribution synopsis can be unveiled as follows:

- Existing metrics analysis with regard to energy efficiency, cooling and infrastructure effectiveness associated with data centres (Section 2).
- Assessment of communication systems related energy efficiency and performance on basis of development of metrics based framework (Section 3).
- An analytical comparison and evaluation of metrics on the basis of 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 categorized as under:

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 facility versus IT equipment related power consumption is designated as PUE. Inverse of PUE constitutes DCiE. There exists an analogy between Energy Utilization Effectiveness (EUE) and PUE but EUE is rather energy based than being 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 for assessment of work and IT equipment effectiveness with regard to consumption of power and the performance output inter-relation.

2.2. Air and Environment Related Metrics

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

2.3. Metrics Pertaining to Cooling Efficiency

To address the rack cooling efficiency in accordance with manufacturers' thermal guidelines the related metric is known as Rack Cooling Index (RCI). The assessment of power required for operating cooling equipment is addressed by Data Centre Cooling System Efficiency (DCCSE) [31]. In fact it is represented by a ratio which exists between average power consumption of the cooling system and 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 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.

The traditional communication networks which focuses mainly on network latency, bandwidth and error rates is addressed by metrics which take these as main indicators. Several other works address few other aspects of data centre network analysis [32], [33]. The evaluation mainly focussed on latency assessment and bandwidth pertaining to pairs of running virtual machines [32] along with analysis of capacity and related costs of data centre network [33].

III. CLOUD COMPUTING DATA CENTRE RELATED COMMUNICATION METRICS

These metrics are related to ascertain the performance and energy efficiency oriented factors pertaining to cloud computing aspects of data centres' communication systems. Application in cloud computing are communication intensive except for High Performance Computing (HPC) [20]. Hence, certain parameters can dramatically affect the system performance and they include error rate, bandwidth capacity and latency. These metrics by virtue of allowing finer granularity can also patch-up the undesirable aspect that exist with performance and power related metrics for their inability to segregate communication systems and IT equipment class.

The dynamic resource provisioning addresses the issue of avoidance of long latencies arising due to severe variability of growing demand during the working hours pertaining with typical web applications as the demand is exhibited to rise during working hours and decrease during night to morning. It also addresses the issue of diverse orchestration concerns pertaining with 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 virtue of 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 evaluating the execution of workload management for arrival of a task at load admission to reciprocating corresponding response to user [37]. The workload can be deciphered as processing in a given time-period for handling the processing of work in Cloud computing [38]. The reliability of nodes is of paramount importance since it addresses the issue of changing adaptability under uncertain situations like failure in particular functions in virtual environment. Metrics play an instrumental role to revamp 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 for procuring more accurate values by 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 following three categories:

- Metrics pertaining to power.
- Metrics pertaining to performance.
- Metrics pertaining to network traffic

Energy efficiency pertaining to communication system is addressed by power related metrics which analyse that how much electric power actually converts into work of delivering information while executing the networking and other related activities. The analysis of capacity, communication rate and information delivery latency is addressed by performance related metrics. Lastly, an 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 Pertaining to Power

3.1.1. Communication Network Energy Efficiency

Transformation of network related electricity is needed for fulfilment of goal of delivering the information. For measurement of its efficacy the metric concerned is expressed as follows.

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

Data centre related networking hardware comprises of components which participate for discharging the function of inter-server information delivery inclusive of server parts like routers, network-switches, Network Interface Cards (NICs) and communication links. In context of servers the value of NICs is worth considering because the servers without NICs simply discharge the function of computing and they are not reckoned as communication equipment. The computing servers are subjected to end-to-end network related maximum throughput and this is known as effective network throughput capacity. The unit of CNEE is watts/bit/second. It is energy required for delivering a unit bit information. It is also equivalent to joules/bit.

3.1.2. Network Power Usage Effectiveness

It represents that part of the power which is consumed for data centre operative function pertaining to communication system.

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

Strictly speaking, NPUE specifies consumed power fraction due to IT equipment which is utilized for operating data centre communication system. Similarly, the fraction of energy utilized as power by server is also measured by PUE. The values of NPUE can range from 1 to infinity. It can be further elaborated, if NPUE stands out as consumption of 6 watts by IT equipment, it can be inferred that 1 watt is utilized for network equipment operation. 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 mean inefficiencies pertaining to network but should be deciphered as energy efficient up-gradation of computing servers.

3.1.3. Energy Proportionality Coefficient

In ideal situation the workload and network devices' energy consumption should be directly proportional but practically network switches or computing servers are not energy proportional. Many servers even in idle state exhibit 66% power consumption at the peak level activity. Pertaining to switches this ratio could be even higher reaching up to 85%. The normalized load depicts as to how much variance is observed by comparing steady workout for ideal situation against fluctuating workloads. It is a systems' energy consumption related offered load function. It is represented as a straight line for an ideal case as in Figure 1 where each increment *l* of load, is represented by a corresponding equal increased consumption P as power. But as revealed practically the consumption of power is not linear in nature.

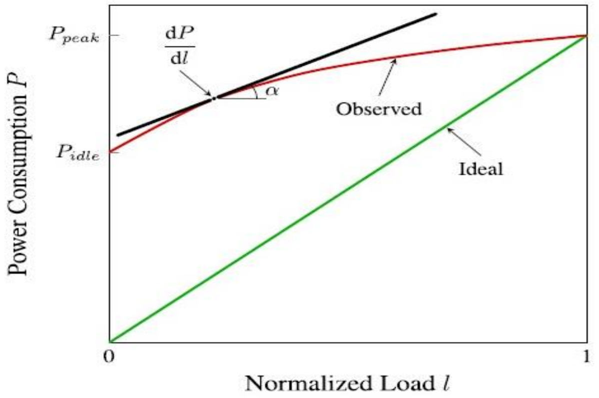


Figure 1: Representing Proportionality of Energy

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

The line of inclination with regard to proportionality of energy consumption represents change in energy consumption vis-à-vis an ideal case. Analysis of this variation can be adjudged by drawing a tangent line for every point in relation with observed curve. Taking observed function related first derivative into consideration the angle α of this tangent line can be procured.

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

The energy proportionality measurement is defined on the basis of tan α:

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

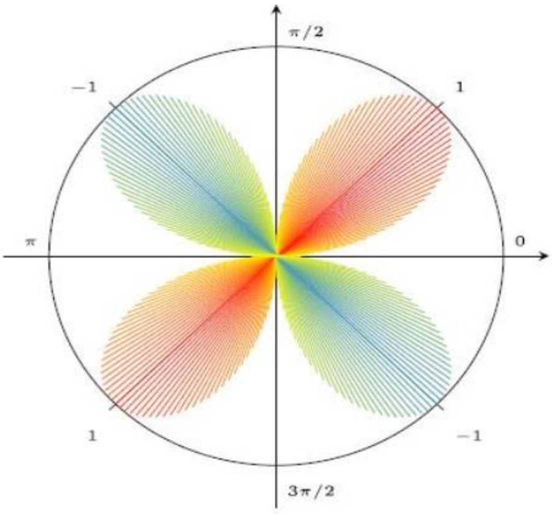


Figure 2: Energy Proportionality Coefficient (Epc)

Figure 2 depicts polar coordinates related various values of α plotted in accordance with EPC metric values. The EPC is equal to 1 if α = π/4 thus representing that every increment with system related load causes equivalent energy consumption. If α is equal to -π/4 it means that there is an equivalent energy consumption decrease negotiating 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 so as to make the EPC also equal to 0. If α is equal to π/2 then it is represented as asymptote of the power consumption function.

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

The evaluation of 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] with respect to 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 by a linear representation that connects the values of power consumption during peak and idle state. If the values of LDR are positive it is indicative of 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 perfect linear representation indicates that LDR is zero.

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

3.2. Metrics Pertaining to Performance

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

3.2.1. Network Latency

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

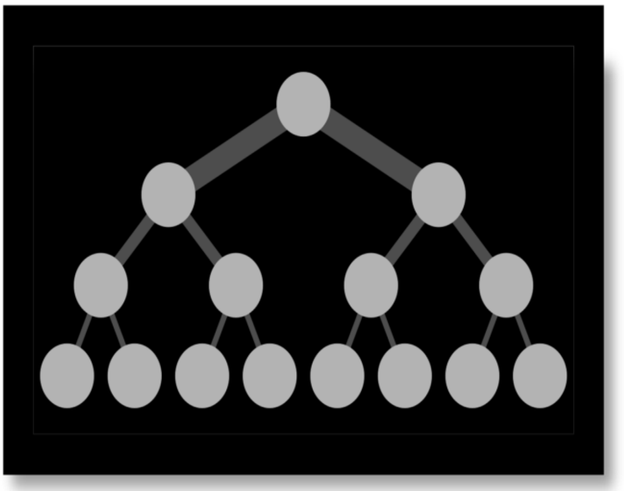


Figure 3(A): Fat Tree Three – Tier Architecture

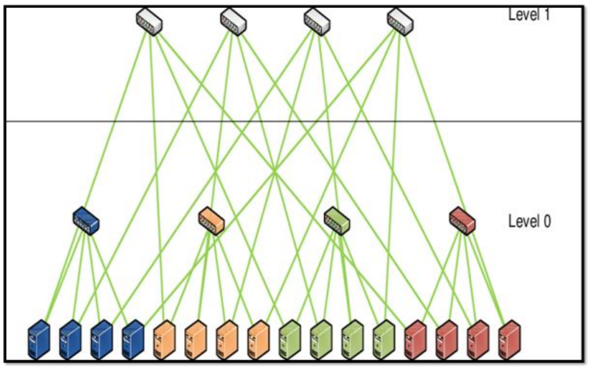


Figure 3(B): Bcube Architecture

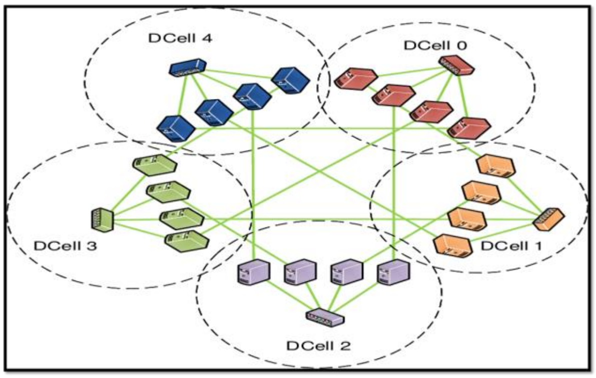


Figure 3(C): Dcell Architecture

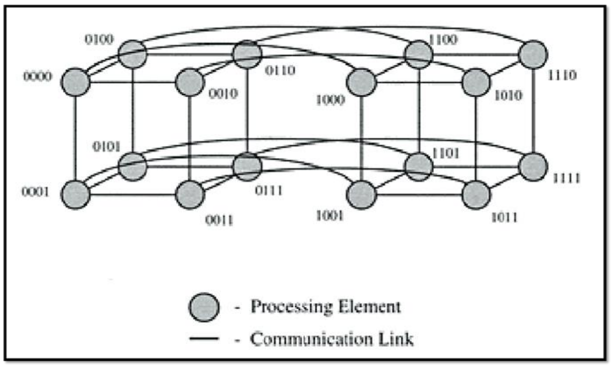


Figure 3(D): Optically Braced Hypercube

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 metric called Inter-Server Hop Distance (ISHD) or Inter-Server Communication Latency (ISCL).

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

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

Cloud applications which can exhibit parallelism in execution are well addressed by ISCL and ISHD. Their task execution requires an exchange of data so as to exhibit brisk performance in network related architectures along with minimized inter-server hops and shorter inter-server delays. However unrelated single server confined applications exhibit immunity for inter-server delays. Apart from measurement of average values, the analysis of distribution of inter-server delays is of paramount importance. Deviation of small values signifies data centre networks related with computing servers located 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 in nature, impart high inter-server delays to data centres. In this state consolidation of heavily communicating tasks for reducing network delays and enhanced performance becomes highly beneficial.

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

3.2.2. Bandwidth Oversubscription Ratio

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

This is equivalent to Bandwidth Oversubscription Ratio (BOR) of (48Gb/s)/(20Gb/s) equal to 2.4:1 which is in equivalence with per server bandwidth of (1Gb/s)/(2.4) equal to 416 Mb/s under full load. At 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 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 for the reason that server-centric architectures abstain from introducing points related to bandwidth oversubscription. The estimation of minimum non-blocking bandwidth for each and every server can be achieved by computing BOR. If the available bandwidth becomes insufficient due to generation of more traffic by computing servers it leads to a congested state pertaining to ToR and aggregation switches. As a result packets are dropped from overflowed buffers resulting in performance degradation of cloud related applications.

3.2.3. Network Losses

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

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

The inter-connecting links are dissimilar in data centres considering fat tree three-tier architecture as in Figure 3a, it incorporate 10 Gb/s optical links where per-link Bit Error Rate (BER) exists between 10^{-12} to 10^{-18} in core and access layers. Functioning of access layer is governed by twisted pair Gigabit based Ethernet technology where 10^{-10} is the range for BERs. Based upon link characteristics of network and topologies the average end-to-end error rates can be calculated by considering communication paths like server-to-gateway and server-to-server.

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

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

where N stands for number of computing servers, L stands for hierarchical layers in network topology and BER*l*n represents layer *l* link of BER connecting server n and data centre gateway.

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

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

where N signifies the computing servers' number, BERij represents server i and server j related interconnecting paths' BER. Sum total of BERs related all links existing between servers i and j represent BERij.

The importance of error rate measurement is noteworthy while addressing cloud related applications' sensitivity to identify transmission related errors and hardware related faults identification.

3.2.4. *Average Link Utilization Ratio*

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

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

where un is the utilization ratio and Ni is number of type i links. ALUR being an aggregate network metric can address traffic distribution and load levels with respect to data centre networks. This can also identify network hotspots and can be instrumental in avoidance of network congestion related performance degradation in cloud computing.

This is possible to measure ALUR separately with respect to a three-tier fat tree topology for addressing aggregation, access and network related core segments. If any of these segments is highly congested it will signal for initiating an increase in network links capacity and switches or re-evaluating bandwidth oversubscription ratios. It is possible to measure ALUR by servers-to-switches and servers-to-servers segment basis for BCube and DCell topologies.

3.2.5. *Average Server Degree Connectivity*

Topologies of data centres are switch-centric or server-centric which depend on data centre design strategy. The fat tree architecture is switch centric as it connects solitary ToR switch with single link only. The BCube and DCell architecture exemplify server-centric architecture and they enhance network capacity by providing re-adaptation capability of the node and switch total dysfunction. The enhancement of network related capacity is achieved revealing high degree of connectivity which also makes the whole topology to be fault-tolerant, thus facilitating the load balancing. But this enhanced degree of connectivity culminates in to increased network power consumption because of utilizing more links and NICs for this purpose. For analysis of computing servers' high quality connectivity along with the value of this metric needs to be estimated.

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

where N denotes total number of data centre servers, Cn denotes connectivity of small number of servers connected with other servers, devices and switches.

3.3. Metrics Pertaining to Network Traffic

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

Internal traffic constitutes 75% of entire network based communication within a modern data centre [25]. It comprises of cloud application database interaction amongst independent tasks which are in parallel execution. Internal communication within a data centre remains subjugated to metric DAL based database related access delays, metric BOR based network availability and inter-server latency addressed by metric ISCL/ISHD. The uplink and downlink path related latency or bandwidth of a data centre network delivers unaffected performance pertaining to 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 is addressed by metric BOR. It is also sensitive towards latency in uplink and downlink path which is addressed by UDCL/UDHD. The inter-server bandwidth and communication latency which is addressed by metric ISCL/ISHD remain unaffected with regard to external communications related performance. External and internal data centre traffic exists in proportion as described below:

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

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

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

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

Apart from classifying network traffic on the basis of target point the identification of messaging related with user or application from rest of the traffic it becomes imperative for securing and managing the aspect related with monitoring and network management. Monitoring is essential for communication network operation. The transmissions which address resolutions like ARP and RIP/OSPF type of routing are addressed by management operations.

Management operations furthermore can be attributed for problem detection and control messaging like ICMP while the monitoring of operation for the traffic can be addressed by SNMP. The Management and Monitoring Traffic Ratio (MMTR) represent network management related traffic overhead.

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

Communication Network Energy Efficiency (CNEE) metric and Management and Monitoring Traffic Energy (MMTE) metric address energy consumption during management of network ignoring transportation application related traffic.

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

The unit of MMTE is Joule which represents the energy utilized by communication related equipment for securing network related operational status. Ideally, MMTE should exhibit near zero values while major portion of energy is linked to traffic related with 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 executed workloads inter-dependencies and for estimation of optimized communication for several data centres which are geographically distributed [51].

IV. NUMERICAL EXAMPLES BASED 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 so configured that they provide backup for 4096 computing servers.

Fat tree three-tier topology comprises of 128 racks with 8 core and 16 aggregation switches with respect to these servers. The interconnectivity between core and aggregation switches with 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 arrangement of 4096 computing servers in groups of n=8. This results into provision of a BCube architecture of level k=4 having commodity switches in three layers pertaining to each group of servers and the DCell architecture of level k=2. The commodity switches are inter-connected with computing servers through 1Gb/s links. The link length for lowest layer is 2 metres with link length of 10 and 50 metres for middle and uppermost layers respectively. Numerous load balancers utilizing 50 metres long, 40 Gb/s optical fibres are used for establishing the connectivity between gateway router and data centre network.

In OH architecture for supporting 4096 servers twelve hypercube dimensions are required. For the sake of inter-connection this requirement is fulfilled by $12.2^{12}/4=12228$ two-by-two optical switches.

An assumption is made that support offered by optical fibres negotiate single mode light propagation using 1550 nm operating wavelength in all architectures.

4.2. Power Related Metrics: Evaluation

For the purpose of evaluation of 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 NPUE and CNEE when data centre load increases. For provoking new servers to acquire fully operational profile from their dormant profile extra network switches are not needed thus making the increase of load of the data centres non-linear. But in case a new rack has to be activated it demands power for Top-of-Rack (ToR) switch along with core and aggregation switches. It is depicted with respect to three-tier topology in Figure 4.

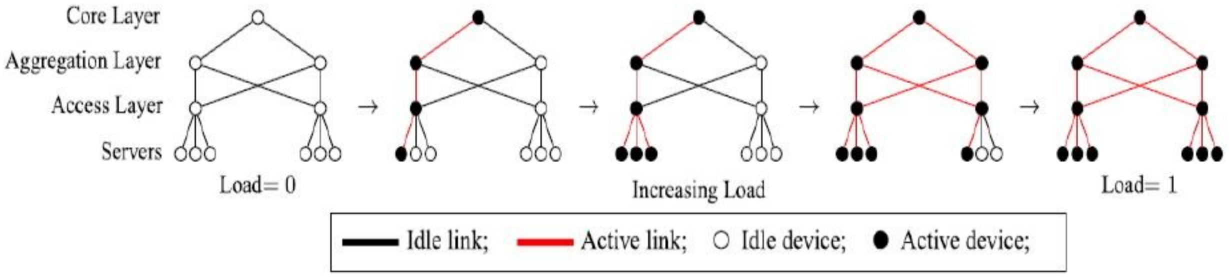


Figure 4: Powering Up Equipment as Data Centre Load Increases

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 pertaining to servers can be estimated by Dynamic Voltage and Frequency Scaling (DVFS) at server level. The power consumption P(l) is expressed as below:

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

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

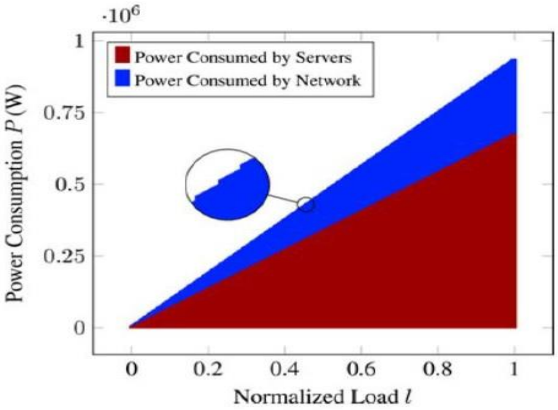


Figure 5: Fat Tree Three – Tier Data Centre Related It Power Consumption

Core layers and the aggregation issues pertaining to 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 normalized consumption of power with respect to fat tree three-tier architecture is depicted in Figure 5. The power consumption by Network Interface Card is non-inclusive for consumption of power by the servers but it is additive for consumption of power by the network. A previously idle rack when assumes an active wake up state in a server then they are represented as leaps as shown in the zoomed portion of the Figure 5. It culminates into power consumption related network non-proportionality due to activation of 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

With respect to all four data centre architectures taken into consideration the computation of CNEE is shown in first row of table 2. Several layers related bandwidth oversubscription of high degree takes the CNEE value to the highest level in case fat tree three-tier topology. It results into energy utilization for supporting higher bit-rates which are not fully utilized by the servers. On contrary the achievement of throughput is 100% of network capacity for DCell and BCube architectures. The factors of dependence for CNEE include total network related consumption of power while bandwidth related oversubscription addresses the CNEE related sensitivity issue. This fact gives an explanation as to why BCube related CNEE is higher than DCell related CNEE. BCube is comprised of (k+1).n^k(2048) number commodity switches while in case of DCell it contains a single commodity switch for each group of n servers (512). OH architecture is comprised of 12228 two-by-two optical switches which consume significantly less power as compared with commodity switches which support BCube and DCell architectures. This makes the CNEE value computed with respect to OH topology, to be more identical with DCell value than value for BCube.

With the help of NPUE an assessment of overall power effectiveness is possible while considering energy required for transferring single bit related information in data centre networks. Thus BCube seems to require highest amount of power since its NPUE value is lowest. It is exemplified by the fact that DCell exhibits more switches to be incorporated in comparison with three-tier architecture. But it includes commodity switches which exhibit a significant less power consumption than aggregation and core level switches. For OH architecture although individual optical switches consume less power but still the NPUE value is lesser in DCell than OH architecture. In case of OH architecture two main causes for network power consumption are high number of active ports and transceivers for each server.

4.2.2. Energy Proportionality Evaluation

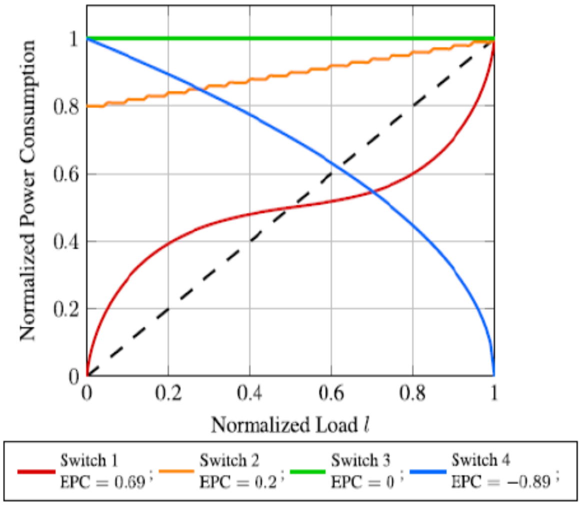


Figure 6: Different Network Switches Related Power Consumption Profiles

Figure 6 depicts normalized power consumption pertaining to multi-network switches for estimation of computed EPC values with different profiles. For ideal case the EPC value equals to 1 and is exhibited by a dashed line. Switch 1 exhibits a curvilinear behaviour when the load value falls within intermediate range (0.2, 0.8) then consumption of power increase rate is lower than workload increase rate. But when the load values are low (less than 0.2) or high (more than 0.8) then power consumption increases more rapidly than incoming workload. Hence EPC becomes equal to 0.69 as a result. The energy consumption with a realistic profile is achieved in case of switch 2 with EPC value equal to 0.2 thus exhibiting a large idle part with power consumption in a step ladder pattern for its ascription to communication ports. Thus it very closely resembles with case profile presented by switch 3. Switch 3 is totally insensitive to workload hence the value of EPC is exhibited as zero by switch 3. In case of switch 4 the EPC value is exhibited to be negative (-0.89). This shows that device is consuming less energy when there is rise in workload.

4.3. Performance Related Metrics: Evaluation

This section deals for presenting the evaluation results of proposed metrics with respect to connectivity (ASDC), energy losses (UDER, ISER) and network latency (UDCL, ISHD, UDHD, DAL, ISCL) considering metrics of BOR and ALUR as an exception.

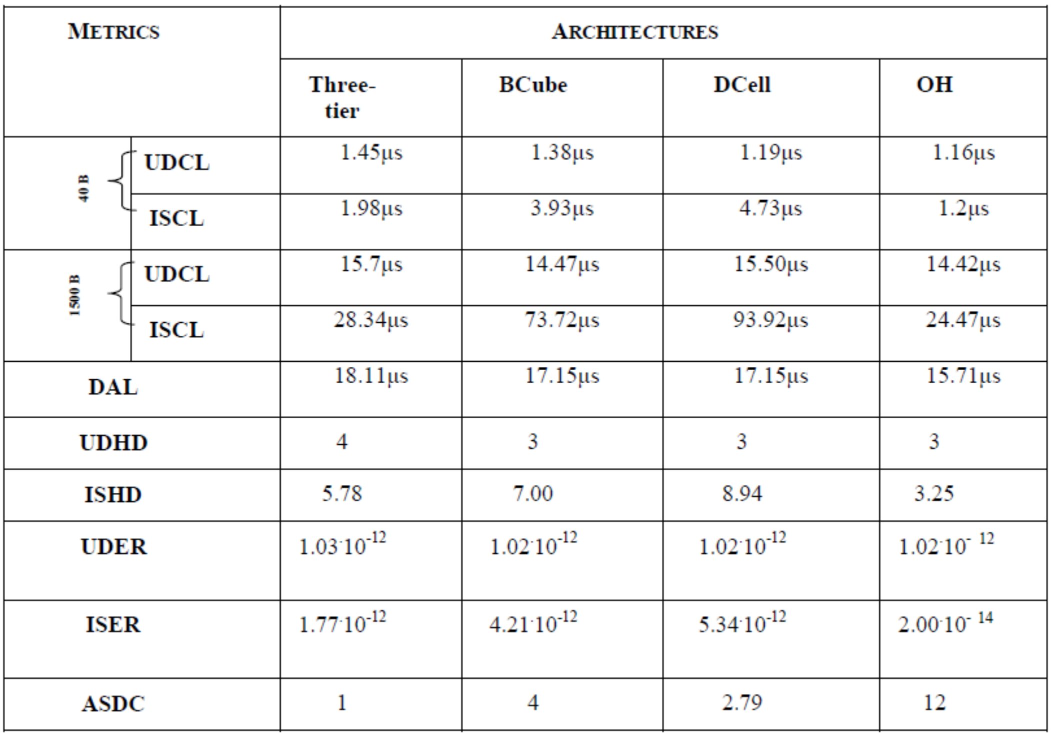
The process of bandwidth multiplexing do not receive any backup point with respect to server-centric architectures and the process of oversubscription persuade BOR metrics becoming equal to 1. In computational process per link traffic statistics becomes a requirement to address ALUR metric and it can 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 having values of 40 bytes and 1500 bytes for their transmission for evaluating ISCL, UDCL, UDER, DAL and ISER corresponding to maximum Ethernet transmission unit and TCP acknowledgement were taken into consideration. The UDCL and ISCL are addressed by measuring one-way transmission delay and by measuring round trip delay in case of DAL. To address signal losses the copper cables and optical fibres are assigned with BER values of 10^{-12} and 10^{-14} respectively. The queuing delays can be ignored with respect to Ethernet inter-frame gap due to absence of any other traffic in the data centre network. The configuration of a single packet network delay comprises of link propagation delay Dp and transmission delay Dt. The expression of Dt and Dp represents a ratio which exists between packet size s and link rate r in case of Dt while ratio between link length L with respect to signal propagation speed P defines Dp.

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

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 case of copper it amounts to be a fraction (two-thirds) of the velocity of light *c*. In optical fibre velocity of light is calibrated by considering the refractive index which is taken equal to 1.468 for glass fibre. The network latency losses and metrics pertaining to connectivity are presented in Table 3 for their results. It reveals that OH architecture supports better internal communications by considering ISCL, ISER and ISHD for the reason that all of these are of lower values in comparison to other architectures. Due to the fact that OH architecture has highest ASDC value it gives a genuine assurance for providing short paths even between distant servers. The internal communications are better supported by three-tier topology with respect to BCube and DCell. It seems to be a paradoxical state because connectivity degree measures along with ASDC for three-tier architecture is minimal as compared with rest of architectures. Although DCell and BCube both show superior inter-connectivity yet they have to deal with numerous hops for distantly placed inter-server communication. BCube and DCell are mostly dependent on copper links, so it poses a challenge of very high inter-server error rate which is addressed by ISER. On contrary gateway and server related rate of error remains lower in case of BCube and DCell as measured by UDER, for the reason that the out-coming packets from the server have to execute lesser number of hops for reaching gateway.

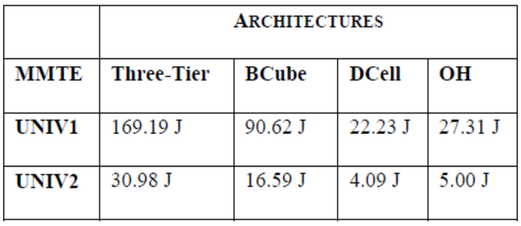
4.4. Network Traffic Related Metrics Evaluation

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

Two-tier architecture supports both data centres, and about half hour of the traffic and two and half hour of the traffic is assigned respectively for data traces of UNIV1 and UNIV2 data centres. For the purpose of evaluating the fraction of network management and for traffic monitoring the computation of MMTR is done which is found to be 0.79% and 0.025% in case of UNIV1 and UNIV2 data centres respectively.

It is revealed by the results that UNIV1 related network is managed with lesser efficiency although UNIV1 is equipped with lesser number of network devices.

Table 4: Evaluation of Management and Monitoring 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 data centre network for processing and delivery management along with traffic monitoring. The metric MMTE addresses energy consumption related with traffic monitoring while considering the UNIV1 and UNIV2. It is revealed that it is lower for all architectures pertaining to UNIV2. The energy consumption is lowest for transferring a single bit of information in case of DCell therefore DCell always outperforms other architectures and fat tree three-tier architecture seems to be most energy hungry (vide CNEE values of Table 2).

Most of the metrics which are presented and discussed are certainly influenced by choice of employed resource allocation strategy. 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 magnitude of monitoring and management of traffic flux with respect to MMTR and MMTE metrics. The internal traffic increases in case of metrics viz. ETR and ITR and it may cause changes in value of ALUR. Thus it really focuses on the imminent conclusion that these metrics offer an important platform for raising the concept of resource allocation in arena of cloud data centres thereby paving a way for securing a novel solution for network oriented scheduling.

V. DISCUSSION

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

These metrics have offered an energy efficient umbrella cover for revamping of relevant cloud applications like SaaS which address internal communication as well as fervid communication destined for users end. These metrics help to not only ensure but also guarantee SLA as well as QoS to the customers. Lastly, metrics pertaining to network traffic lead to a permit regarding resource allocation policies which are infrastructure-aware and proper traffic management development. The metric framework for cloud related data centre networks justifies itself for ensuring expansion of planning capacity. It helps in capacity enhancement for designing an optimized data centre of the future.

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

A top level comparison of data centre architectures which have been evaluated is provided in Table 5. The measurement values and the details of evaluation are provided with precision in Section 4 whereas in Table 5 these values have been mentioned as high (H), medium (M) and low (L) for simplicity. In case of three tier architecture the network related total functioning capacity is restricted for being fully operational because of high bandwidth oversubscription which ultimately causes highest per bit energy consumption. DCell has lowest per bit ratio for consumption of energy. Power usage effectiveness of DCell is highest, making it the most "green" architecture amongst all the architectures. BCube is comparatively less efficient as far as the power usage effectiveness is concerned since it incorporates maximum number of switches. On scrutinizing the communication latency we find that the three-tier fat tree architectures favours server-to-server related internal communications while on contrary the distributed data centre architectures like DCell and BCube have smaller traffic related paths to be directed outside of data centre. However, OH architecture which is server-centric is capable of significantly reducing hops' number amongst servers placed distantly. Consequently, the support they provide to internal communications is better as compared to hierarchical architectures.

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

Achievement of networks related efficiency pertaining to communication is prime objective to be fulfilled while dealing with cloud computing data centres. The achievement of this desired objective is facilitated by set of metrics which address energy efficiency in computing arena which is discussed in this paper. The perspectives related to energy, performance and traffic are addressed by these energy efficiency metrics. The metrics related to power, measure the efficiency of the procedural task which turns electricity into information delivery. The metrics related to performance are used for the analysis of error rates, network latency and available bandwidth which are conventional communication system characteristics. The metrics pertaining to network traffic provide an idea about the energy consumed in conveying traffic to different categories and also about the traffic characteristics. The framework of metrics has been assessed and vindicated for three-tier hierarchical and also for distributed (Hypercube, BCube and DCell) data centre architectures. A number of properties pertaining to these architectures were revealed by the results obtained. These metrics will undoubtedly prove constructive for academicians and industry specialists.

ACKNOWLEDGEMENTS

I am thankful to all concerned with 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 and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	Saumitra Vatsal wrote this research paper. It was supervised by Dr. Satya Bhushan Verma.

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 center 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 modeling 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 centers 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 center 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 center 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: Modeling 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 center 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 reuse energy from a data center,” White Paper, 2010.
29. (2009), “UPS load factor,” Available Online at: http://hightech.lbl.gov/benchmarking-guides/data-p1.html
30. (2009), “Data center 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 centers: 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 Centers 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 center 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 center 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 center 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 centers,” 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 centers 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 center 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 center 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 center network virtualization: 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 center 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. Fiber Commun. Conf. Expo. Nat. Fiber 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

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 & 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 environment conscious approach that curtails excessive energy consumption by Cloud data centres and emission of carbon footprint into the environment. He has attended numerous seminars, conferences and also 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 Ph.D. in Computer Science and Engineering from 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 journals in SCI and peer-reviewed journals, and he has also published 2 papers at the international conference. He has filed one patent. Currently, he is working as a resource person at BBA University Lucknow (Central University). His area of interest is Biometric, 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 officiating as the HoD (Head-of-Department) of DCSE (Department of Computer Science & Engineering), Institute of Technology at Shri Ramswaroop Memorial University (SRMU), Uttar Pradesh, India.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)/ journal and/or the editor(s). The Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP) and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.