J Supercomput (2018) 74:334–369
https://doi.org/10.1007/s11227-017-2132-5

On energy consumption of switch-centric data center networks

Olusogo Popoola[1] · **Bernardi Pranggono**[1]

Published online: 15 September 2017

Abstract Data center network (DCN) is the core of cloud computing and accounts for 40% energy spend when compared to cooling system, power distribution and conversion of the whole data center (DC) facility. It is essential to reduce the energy consumption of DCN to ensure energy-efficient (green) data center can be achieved. An analysis of DC performance and efficiency emphasizing the effect of bandwidth provisioning and throughput on energy proportionality of two most common switch-centric DCN topologies: three-tier (3T) and fat tree (FT) based on the amount of actual energy that is turned into computing power are presented. Energy consumption of switch-centric DCNs by realistic simulations is analyzed using GreenCloud simulator. Power-related metrics were derived and adapted for the information technology equipment processes within the DCN. These metrics are acknowledged as subset of the major metrics of power usage effectiveness and data center infrastructure efficiency, known to DCs. This study suggests that although in overall FT consumes more energy, it spends less energy for transmission of a single bit of information, outperforming 3T.

Keywords Cloud computing · Data center · Data center network · Energy consumption · Energy efficient · Fat tree · Green data center · Switch-centric · Three-tier

✉ Bernardi Pranggono
b.pranggono@shu.ac.uk; bern@ieee.org

Olusogo Popoola
olusogo.j.popoola@student.shu.ac.uk; olusogo.popoola.2015@ieee.org

[1] Department of Engineering and Mathematics, Sheffield Hallam University, Sheffield, S1 1WB, UK

1 Introduction

The need for secure and efficient hosting of digital information demonstrated in converged networks (data, voice, image and video) led to the rapid evolution of data center (DC) around the world. Emergence of Web 2.0 environment with its rich enabled applications paved the way for data to become every organization most valued asset and therefore hosted with the highest degree of confidentiality, integrity and availability. The prevailing models of electronic data interchange (EDI) which demanded corporations to depend absolutely on data made DCs the live wire of the global economy [1] representing the foundation and structure upon which cloud computing was established [2]. The adoption of the cloud computing paradigm has provided the much needed avenue for data centricity of services as seen in Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) [3]. In our Internet-driven millennium, the sustainability of systems responsible for web communication deployment is vital and dependent on uninterrupted power supply, more so that the commodity price of energy is rising faster than expected.

A data center can be defined as a facility that hosts computing resources in a nexus of communication infrastructure for data storage and applications [1,4]. The capital expenditure (CAPEX) at the initial setup of a DC is equally enormous but sometimes incomparable to the operational expenditure (OPEX) [5]. The latter is needed to maintain the quality of service (QoS) in the service level agreement (SLA) and for users to have a good quality of experience (QoE). Hence, achieving a balance between appropriate service delivery, e.g., provision of more bandwidth and low latency network between communicating nodes, and reduction in energy consumption goes a long way in cutting down on OPEX.

Green DCs are designed to ensure utmost energy efficiency and minimum environmental footprint [6,7]. Therefore, recent policies of environmentalist and socialist on DC operators have sharpened the evolution of modern DC toward improving QoS and energy efficiency, coupled with breakthroughs that resulted from competition among operators to cut down on OPEX. This is visible in the proficiency employed by IT giants such as Google, Facebook, Microsoft, IBM and Amazon in becoming progenitors of cloud computing, with continuous improvement and developmental strategies to make their offerings attractive. Therefore, green DC research is seen as part of the continuous improvement needed in designing DCs that are less CAPEX when setting up the core components. This is achieved by deploying energy-efficient DCNs in a bid to further lower the 10% spent on energy as part of OPEX [8]. The approach examined introduce energy coefficient to the design of the DCN as a critical complementary consideration for qualitative performance and throughput of the network modules. Data center network topology could be switch-centric, server-centric or hybrid (dual centric) with its specific energy consumption characteristics [4]. However, studies showed that energy utilized to process workloads in switch-centric topology is more profitable as switches by default are equipped with intelligent routing algorithms and connected to servers through a single port [9], making such networks very responsive. A very responsive variant of switch-centric DCN architecture will be useful as a potential solution to the increasing demands of cloud computing DCs and help eradicated challenges faced by legacy DCN architecture.

In this article, we present an analysis of DC performance and efficiency based on the amount of actual energy that is turned into computing power. It further emphasizes the effect of bandwidth provisioning and throughput on energy proportionality of two most common switch-centric DCN topologies: three-tier (3T) and fat tree (FT). The design objective is such that will accommodate scalability, cost-effectiveness, resilience and end-to-end aggregation of bandwidth provisioned at reasonable energy spend. We have implemented a model of 3T and FT topologies based on modified network simulation (ns-2) GreenCloud [8]. We present our evaluation results and compare the performance and power-related metrics [10], bandwidth oversubscription ratio (BOR), communication network energy efficiency (CNEE) and network power usage effectiveness (NPUE) of the two DCN architectures. The energy consumption was matched with network traffic (or workload) to discover the energy awareness of the two DCN architectures.

The main contributions of the article are:

- An implementation of FT DCN architecture using GreenCloud simulator.
- Performance evaluation of 3T and FT based on power-related metrics, BOR, CNEE and NPUE. The focus of this study is on intra-DC network traffic which could generate computer-intensive workload (CIW), data-intensive workload (DIW) or balanced workload (BW) [11].
- A comparison for 3T and FT architecture based on real-world scenario (power budget).
- Introducing energy coefficient to the design and layout of DCN architecture of smaller businesses as a critical complementary consideration for qualitative performance and throughput of the network modules.

The remainder of the article is organized as follows. Section 2 provides information on the core business value of DCs with focus on topologies available for DCN implementation and the legacy techniques used in evaluating energy efficiency. Section 3 discusses the method implemented in improving energy efficiency with emphasis on the simulation of information technology equipment (ITE) to understand DCN energy consumption in line with the Greenness Initiatives. Parameters from ITE and workloads were simulated to obtain a suitable energy-aware operation scheduling and adaptations. Prior to this, the choice of data center simulator was justified, and the final part enumerates the data collection strategies and experimental methods used. In Sect. 4, the simulation results from our experiments based on modified GreenCloud simulator were discussed and evaluated in line with real-world scenarios. The analysis and performance evaluation of the components of the DCN were considered in terms of topology, server and network workloads. In Sect. 5, we offer our depth analysis on the simulation results. Finally, a conclusion is highlighted in Sect. 6.

2 Background

The design framework of green DC has focused on actualization of a scalable, efficient, extensible and cost-effective DCN architecture [12]. The legacy 3T tree-based and emerging new fabric FT which seemed to satisfy the aforementioned criteria of a green DC is exemplars of such architecture. A greater percentage of existing DCs

Table 1 Projection of power consumption for Internet services [13]

Year	Peak performance (10× per every 4 years) (PF)	Bandwidth requirement (20× per every 4 years) (PB/s)	Power consumption (MW)
2012	10	1	5
2016	100	20	10
2020	1000	400	20

implemented the traditional 3T topology at the core of their network. This has resulted in enormous energy consumption and budget increase along with the exponential growth of DCs. This is further illustrated in Table 1, where the past, present and future projections of operations, bandwidth demands and power utilization for high-performance systems are shown [13].

2.1 Data center network

A typical 3T DCN architecture is a hierarchy of three layers of switches (core, aggregation and access/edge) arranged in a tree-based topology with two of its upper layers connected with enterprise network devices (see Fig. 1). We use access and edge interchangeably in this article. The Layer 3 (L3) switches or routers at the core and aggregation layers are energy hungry in nature and therefore cannot be easily energy-managed. Due to its importance, core switches cannot be dynamically put into sleep state although it consumes a great deal of energy due to large switching capabilities, i.e., equal cost multi-path (ECMP) forwarding activities. As a result, core switches operate at the maximum transmission rates of around 85% of full load even when the DC is idle [14]. Core switches are high-capacity switches that located in the backbone network and provide access to a wide area network or the Internet. Server typically operates at 66% of full-load energy consumption when the DC is idle, making dynamic power management (DPM) and dynamic voltage frequency scaling (DVFS) approaches selective [11,15–17].

However, end-of-row (EOR) aggregation-level switches with idle module racks can be powered down. This layer is equally utilized as much as the core; hence, packet losses are more at the aggregation layer than any other layers [9]. Most DCs run

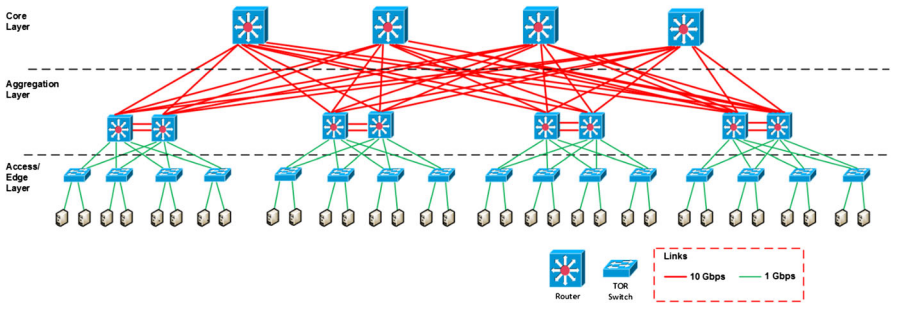

Fig. 1 Three-tier data center network topology

around 30% of their computational capacity [18]; shutting down inactive aggregation servers with prior considerations for load fluctuations that could be managed by less idle servers had always been an energy-aware decision. It was observed in [19,20] that traffic flow and packet utilization within the two upper layers are higher than the access layer, more so when the top-of-rack (TOR) switches that inhibit this lowest layer are inexpensive and low-power commodity types.

Considering the traffic characteristics in DCNs, network traffic associated with DCs could be either inter-DC or intra-DC in nature. The focus of this study is on intra-DC network traffic which could generate computer-intensive workload (CIW), data-intensive workload (DIW) or balanced workload (BW) [11]. Intra-DC network traffic is further categorized into long flows (elephant) in need of higher throughput and short (mice) control flows in demand of low latency. A further discovery was made by [19] during the analysis of existing tree-based topologies that suggest the following traffic flow procedure in organization:

- A greater number of flows in DCs are small in size with duration fewer than hundreds of milliseconds.
- 75% of cloud-based DCs have their traffic within a rack.
- Universities and private corporations DCs have 40–90% traffic prevalent through the network, i.e., from rack through the network.

Oversubscription, the ratio between the aggregate incoming and aggregate outgoing bandwidth of end hosts is introduced to reduce CAPEX during design phase. Oversubscription is considered as a drawback of 3T implementation. The typical oversubscription of 3T topology is 2.5:1 or 8:1 [21] which resulted from allocation of 10 Gbps bandwidth communication link for inter-networking between 10 Gigabit Ethernet (GE) switches in the upper layer (see Fig. 1). In addition, the multi-rooted core switches in large DCs demand multiple-path routing procedure, creating oversubscription, limiting routes or path and lookup delay due to enormous routing table.

The introduction of a new fabric with a flat network topology resolved most of 3T architecture's limitations. The FT DCN presented as folded Clos-based network fabric [5] in Fig. 2 integrates inexpensive Ethernet commodity switches to build a *k-ary* FT with links connecting each layer equally provisioned with the same bandwidth. Consequently, a bandwidth oversubscription ratio (BOR) of 1:1 is available from the core layer to the servers. FT could be implemented in a two-layered spine–leaf configuration as seen in Cisco's Massively Scalable DC (MSDC) [22] and with an additional layer provided above the spine to function in a dual capacity as a load balancer and control plane. The latter is specifically designed for enhanced routing (ECMP) between two end nodes. The control plane is provisioned with a pair of L3 switches to reduce the large switch counts in this design fabric of FT when compared with a full fledged three-layered FT DCN, thus opposing the network universal theorem that "for a given number of switches, the most optimal network exists" [22]. Moreover, the topology of spine–leaf FT architecture is scalable enough to support the explosion of east/west data traffic in web communication and the drift toward software-defined data center.

The existence of L3 lookup at leaf nodes in MSDC enhances the selection of an ideal egress port at the leaf. Intelligent routing architecture reduces the potential congestion

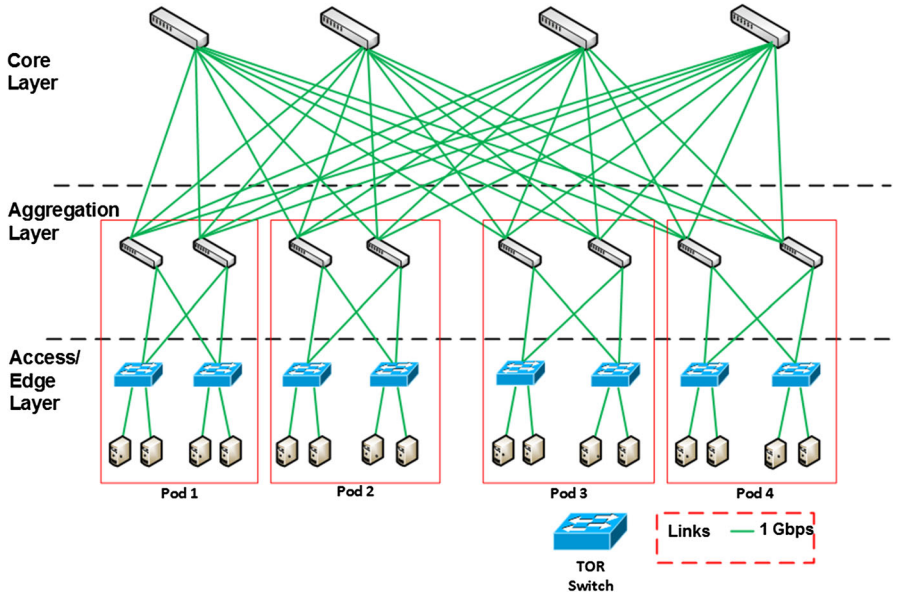

Fig. 2 Fat-tree data center network topology

in network by minimizing packet collision when several packets move toward a single egress port.

To transfer from 3T switching to FT fabric, a fiber connection is established to replace the switch with strong attention given to the channels' link loss likely to occur. Low-pass connection increases the amount of possible connections in the channels. Host-to-host (server-to-server) communication will be most efficient if "virtualization" is employed using virtual machine (VM) techniques without switch hops. Virtualization brings about more server-to-server data flow, storage to storage area network (SAN) traffic as in Storage-as-a-Service. Virtualization is considered an important technique in achieving a green DC, a concept that works with consolidation in reducing power consumption in DC with full adoption of its principle [23]. The concepts of flattening the networks of DC and emergence of visualization are essential.

2.2 Power-related metrics in data center network

In order to ensure optimum energy efficiency and minimum environmental footprint [6] as suggested by Green DC initiative [1,24], it is necessary to apply power-related metrics to evaluate the energy efficiency characteristic in DCNs. There are two main existing metrics applicable to switch-centric DCNs:

- Communication Network Energy Efficiency (CNEE): required energy to convey one bit of information.
- Network Power Usage Effectiveness (NPUE): ratio of overall IT power to power utilized by the network modules.

Although BOR is not directly power related, its computation is necessary in estimating the minimum non-blocking bandwidth available for each server. When the servers produce network traffic above the provisioned bandwidth, the edge and aggregation switches become congested, encounter overflowed bufferings and begin to drop packets [10]. The continuous loop at this point results in increased energy consumption and decreased network performance of cloud applications significantly.

Furthermore, DENS [11] recommended an energy model for switches in green DC as:

$$P_{\text{switch}} = P_{\text{chasis}} + n_{\text{linecard}} * P_{\text{linecard}} + \sum_{i=0}^{R} n_{\text{ports.r}} * P_{\text{r}} \tag{1}$$

where P_{r} = power utilized by an active port transmitting at a rate. P_{chasis} = power utilized by the switch base hardware. P_{linecard} = power utilized by an operating linecard. P_{r} operates at par with the transmitting rate of the switch, limiting the advantages of rate adaptive design because the overall utilization of switch transceivers results in 3–15% of the total energy used by the switch. On the other hand, P_{chasis} and P_{linecard} depend solely on the power status of the device and affected only when device is powered down for lack of network traffic [14].

The server energy consumption model is derived by [14,25]:

$$P = P_{\text{fixed}} + P_f * f^3 \tag{2}$$

where P_{fixed}: power consumed by memory modules, disk and I/O resources, i.e., part of the utilized power that does not scale with f the frequency of operation. P_f: power consumed by CPU, i.e., frequency dependent. f: frequency.

3 Methodology

In [2], DCN architecture was showcased as multiple-layered graph models of diverse DCNs while analyzing the characters of structurally robust DCN [18]. This is similar to the one considered in our model, where ITEs such as computational servers, network and storages devices denote the meeting point of the graph, while the interconnecting network links are the margins of the graph.

3.1 Network management scheme for simulation in GreenCloud

The scheme puts into consideration two switch-centric network architectures: 3T and FT DCN architectures. Specifically, 3T is a tree-based topology, hierarchical three-layered configuration, whereas FT is a Clos-based topology, hierarchical three-layered configuration, with the core or spine, aggregate and access/edge (TOR) layers constituting the layout. The layout also caters for redundancy to forestall points of failure in the connection. The two DCNs to be modeled are configured such that:

- It caters for network and server workload consolidation in each of the tree- and Clos-based hierarchical topologies considered.

- The same numbers of computing servers (S) are considered for task execution, computational workload and energy consumption comparison.
- The core layer switches vary for both networks with downlink speed of 10 Gbps GE medium between core—aggregation—edge switches (C_1–C_2–C_3), and 1 Gbps between edge switches—computing servers (C_3–S) in 3T, and 1 Gbps GE through all layers in FT.
- Aggregation and access/edge network layers are configured with Layer 3 (L3) and Layer 2 (L2) switches, respectively, in 3T architecture.
- Commodity switches were deployed in upper layers of the FT architecture, and the topology sometime referred to as spine–leaf network [5] with two layers.

Table 2 illustrates the configuration of the models simulated and compared in terms of energy and cost efficiencies, scalability and fault tolerance, while Table 3 is an example of a real-world configuration of these models.

3.2 Network simulation

The attributes listed in Table 4 will be considered for DC load, task scheduler and architecture. Similar task scheduling techniques defined in [11] are considered:

- Green: A unified or consolidated scheduler, designed for resolution of computational workload, allowing idle servers and network modules to be powered down.
- RoundRobin: Allocates computational and communicational jobs equally among servers and switches in a circular layout. Computational servers are not overloaded as this creates balanced network traffic. Hence, no powering down of ITE since idleness does not occur.
- BestDENS: An architecture specific technique with best-fit server selection. Attains workload consolidation for energy efficiency while averting servers and switches overload. Hence, there are more active ITEs.

The 3T simulation settings are shown in Table 5.

The 3T DCN architecture is made up of four core switches interconnected to eight aggregate and sixteen TOR switches with seventy-two 10 Gbps links (C_1–C_2, C_2–C_2 and C_2–C_3), and a total of 64 computing servers connected to the TOR switches with 1 Gbps link each (64 Gbps in total) uplink from host to edge switches. Figure 3 depicts the schematic of the modeled 3T DCN architecture.

In the FT simulation, the FT link connectivity is designed so that the three switch layers: spine/core, aggregation and leaf (TOR), all have the same number of port, which is designated as an even number n [5].

- TOR(s) connects with $\frac{n}{2}$ ports to $\frac{n}{2}$ servers.
- The remaining $\frac{n}{2}$ TOR port connects to $\frac{n}{2}$ aggregation switches.
- Aggregation switch connects with $\frac{n}{2}$ ports to the TOR switches.
- The remaining $\frac{n}{2}$ port on the aggregation switch connects to spine switches.
- FT comprises of $\frac{n}{4}^3$ servers, $\frac{n}{2}^2$ aggregation and edge(s) switches, $\frac{n}{2}^2$ core (spine) switches.

Simply put, we have $\frac{n}{4}^2$ spine switches for n^2 pod switches and n^2 servers ($\frac{n}{4}^2$ per pod) as illustrated in Fig. 4.

Table 2 Topology description of modeled 3T and FT DCN

DCN	Topology parameter	Topology setup	ITE configuration and remarks	
			Server	Switch
3T multi-routed tree-based network topology	4× NCore L3 switches	2U rack server	HP ProLiant DL385p Gen8 server, AMD Opteron^{TM} processors 6300 Proc., 768GB RAM, 32TB HDD, 2× 1GE & 1× 10GE	Cisco Nexus 7000 F series 32 port 1 and 10GE module switch for core and aggregate layer
	8× NAgg. L3 switches (EOR)	4 server per rack	Power: 750W	Power: Line card power = 10W per port; Chassis power = 385W per module
	16× NEdge L2 switches (TOR), 1-RU	1 edge switch/rack		Cisco Nexus 3064-X 48 port 1 and 10GE, four fixed QSFP i.e. 4× 10GE or 40GE for edge layer
	NCore switches (C1)	4 racks per pod		Power: Line card = 4W per port; Chassis power = 143W
	NAggregate switchs (C2)	4 edge switch/pod		
	NEdge switches (C3)	16 NRacks		
	NServers (S)	Single NIC/server		
	4C1 + 8C2 + 16C3 = 28 switches, 64S	64 NServers		
	Link(C1–C2) = 10GE			
	Link(C2–C3) = 10GE			
	Link(C2–C2) = 10GE			
	Link(C3–S) = 1GE			
FT folded Clos based network topology	8× NCore L2/L3 commodity switches (spine)	2U rack server	HP ProLiant DL385p Gen8 server, AMD Opteron^{TM} processors 6300 Proc., 768GB RAM, 32TB HDD, 2× 1GE & 1× 10GE	Cisco Nexus 3064-X 48 port 1 and 10GE, four fixed QSFP i.e. 4× 10GE or 40GE for edge layer
	16× NAgg. L2/L3 commodity switches	8 server per pod	Power: 750W	Power: Line card = 4W per port; Chassis power = 143W
	16× NEdge L2/L3 commodity switches (leaf)	2 leaf switch/pod		

Table 2 continued

DCN	Topology parameter	Topology setup	ITE configuration and remarks	
			Server	Switch
	NCore switches (C1)	2 Aggr. switch/pod		
	NAggregate switches (C2)	8 NPods		
	NEdge switches (C3)	Single NIC/server		
	NServers(S)	64 NServers		
	8 C1 + 16 C2 + 16C3 = 40 switches, 64 S			
	Link (C1–C2) = 1GE		Enhancements: Accommodate more switches in core and aggregate layer, use commodity/non-blocking switch to replace high end switches implemented *k-ary* fat tree for n pods, where $n = 8$	
	Link (C2–C3) = 1GE			
	Link (C3–S) = 1GE			

Table 3 Description of Physical Topology for 3T and FT DC Architecture

DCN	Topology parameter		Topology setup	ITE configuration and remarks	
				Server	Switch
3T multi-routed treebased network topology	4× NCore L3 switches		2U rack server	HP ProLiant DL385p Gen8 server, AMD Opteron^{TM} processors 6300 Proc., 768GB Ram, 32TB HDD, 2× 1GE & 1× 10GE	Cisco Nexus 7000F series 32 port 1 and 10GE module switch for core and aggregate layer
	8× NAgg. L3 switches (EOR)		40 server per rack	Power: 750W	Power: Line card power = 10W per port; Chassis power = 385W per module
	16× NEdge L2 switches (TOR), 1-RU		2× edge switch/rack		Cisco Nexus 3064-X 48 port 1 and 10GE, four fixed QSFP i.e. 4× 10GE or 40GE for Edge layer
	NCore switches (C1)		3 racks per pod		Power: Line card = 4W per port; Chassis power = 143W per module
	NAggregate switches (C2)		6 edge switch/pod		
	NEdge switches(C3)		24 NRacks		
	NServers (S)		2× N NIC/server		
	4C1 + 8C2 + 16C3 = 28 switches 64S		960 NServers		
	Link (C1–C2) = 10GE				
	Link (C2–C3) = 10GE				
	Link (C2–C2) = 10GE				
	Link (C3–S) = 1GE				
FT folded Clos based network topology	*3 layered:*	*2 layered:*	*3 layered:*	HP ProLiant DL385p Gen8 server, AMD Opteron^{TM} processors 6300 Proc., 768GB Ram, 32TB HDD, 2× 1GE & 1× 10GE	Cisco Nexus 3064-X 48 port 1 and 10GE, four fixed QSFP i.e. 4× 10GE or 40GE for edge layer
	8× NCore (C1)	8× NCore (C1-spine)	2U rack server	Power: 750W	Power: Line card = 4W per port; Chassis power = 143W per module

Table 3 continued

DCN	Topology parameter		Topology setup	ITE configuration and remarks	
				Server	Switch
	16× NAgg. (C2)	16× NAcc. C3-(leaf)	120 server per pod		
	16× NEdge. (C3)	64× NServers (S)	60 server per rack		
	64× NServers (S)	24 switches, 64 S	2 racks/pod		
	40 switches, 64 S		8 NPods		
	Link (C1–C2) = 1GE		Single NIC/server		
	Link (C2–C3) = 1GE		960 NServers		
	Link (C3–S) = 1GE		*2 layered:*		
			2U rack server		
			120 server/pod	Enhancements:	
			60 server/rack	Accommodate more switches in core and aggregate layers, use commodity/non-blocking switches to replace high-end switches	
			2 racks/pod	Implemented *k-ary* fat tree for n pods, where $n = 4$	
			8N pods	In the 2 layered topology an L3 protocol (OSPF and BGP) is used as a control plane and load balancer for traffic and latency reduction [22]	
			Single NIC/server		
			960 N servers		
Redundancy	N + 1			Resilience of N + 1 i.e. Tier III rated DC for both 3T and FT	

Table 4 Data center simulation attributes

DC load consideration	DC task scheduler/ID# consideration	DCN architecture consideration	DC ITE consideration
Idle (30%)	Green (G)	Three-tier (3T)	Core switch (C)
Half (50%)	RoundRobin (R)	Fat tree (FT)	Aggregation switch (Aggr.)
Full (100%)	BestDENS (D)		Edge switch (access/edge)
			Computing server (S)

Table 5 3T simulation setup

Notation	Meaning
Green (G)	
G-30%-3T	Green scheduler (G), Idle load (30%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)
G-50%-3T	Green scheduler (G), Half load (50%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10G/1 Gbps, 64Servers (64S)
G-100%-3T	Green scheduler (G), Full load (100%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)
RoundRobin (R)	
R-30%-3T	RoundRobin scheduler (R), Idle load (30%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)
R-50%-3T	RoundRobin scheduler (R), Half load (50%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10G/1 Gbps, 64Servers (64S)
R-100%-3T	RoundRobin scheduler (R), Full load (100%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)
BestDENS (D)	
D-30%-3T	BestDENS scheduler (D), Idle load (30%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)
D-50%-3T	BestDENS scheduler (D), Half load (50%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10G/1 Gbps, 64Servers (64S)
D-100%-3T	BestDENS scheduler (D), Full load (100%), 3T (3T), 4NCore (4C), 8NAggr, 4NEdge/pod: 16, 2× 10/1 Gbps, 64Servers (64S)

The desirable benefit of FT network is the ability to create large inter-connections using small-scale switches. This is mainly because its connection capacity depends on the quantity of core layer switches. Increase in number of deployed core layer switches is directly proportional to the improvement in connection capacity and likewise the cost of the network [26].

When establishing connection routes, all concurrent connections from an access switch compete for the same cluster of the core switch, thereby increasing the burden on the core switches whenever congestion occurs at the access switch. This congestion is due to simultaneous real-time request from all server-edge network interface card (NIC) at full bandwidth capacity (e.g., 1 Gbps multiplied by numbers of servers in the rack). Congestion at the TOR and non-uniformity multicast signal are responsible for the expenses associated with non-blocking multicast FT DCNs.

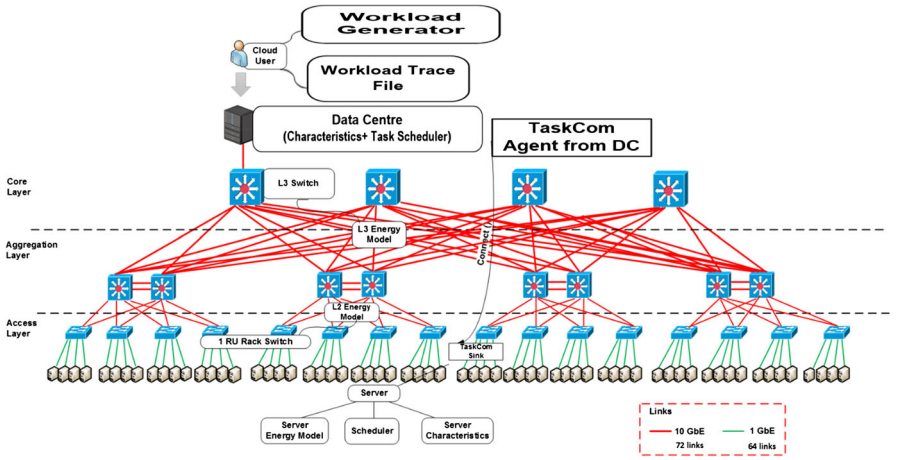

Fig. 3 Schematic of 3T DCN model

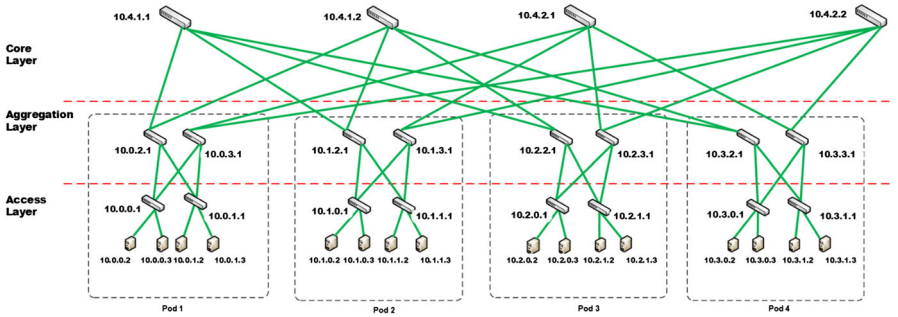

Fig. 4 Illustration of FT ($k = 4$) architecture with assigned IP. Adapted from [2]

FT topology achieves non-blocking unicast communication with a few numbers of cores, but non-blocking multicast, an imperative communication pattern utilized in most DCs still requires large numbers of core switches due to the non-uniformity of multicast traffic [26]. Instances of search queries redirection to index servers and file chunk replication in distributed servers are enhanced for high performance with non-blocking multicast communication. Thus, it is of upmost importance to decrease the cost involved in FT DCs. Otherwise, it will be a replica of the energy hungry high-end switches in the upper layers of traditional 3T architecture in terms of cost. Network module and server redundancy in high-availability (HA) DCs with six nines (99.9999%) can be used to lessen the cost of non-blocking multicast FT DCN, by adequately equipping it to handle various forms of multicast traffic by re-arrangement and re-assignment of non-blocking commodity switches to replace core and to provide high network bandwidth.

The commodity switches act as bouncing switches, implementing digit reversal bouncing (DRB), an algorithm for load balancing proposed in [27] with adequate routing condition to control traffic path within the DCN to end host, hence complementing the ECMP in splitting traffic among multiple paths easier. Packet routing

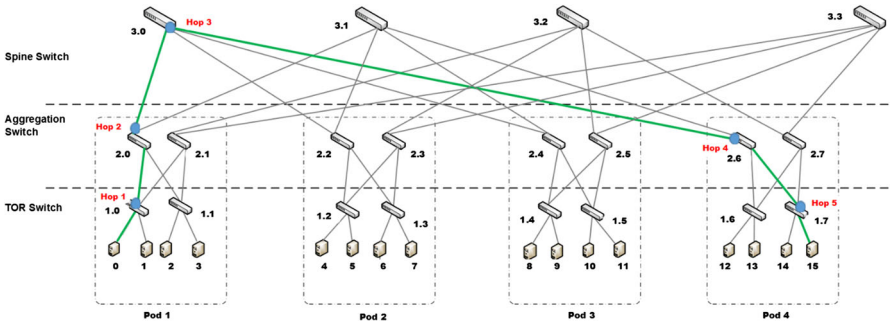

Fig. 5 FT with DRB. Adapted from [27]

Table 6 FT Simulation setup

Notation	Meaning
Green (G)	
G-30%-FT	Green scheduler (G), Idle load (30%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
G-50%-FT	Green scheduler (G), Half load (50%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
G-100%-FT	Green scheduler (G), Full load (100%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
RoundRobin (R)	
R-30%-FT	RoundRobin scheduler (R), Idle load (30%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
R-50%-FT	RoundRobin scheduler (R), Half load (50%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
R-100%-FT	RoundRobin scheduler (R), Full load (100%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
BestDENS (D)	
D-30%-FT	BestDens scheduler (D), Idle load (30%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
D-50%-FT	BestDens scheduler (D), Half load (50%), FT (FT), 8NCore (8C), 16NAggr, 16NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)
D-100%-FT	BestDens scheduler (D), Full load (100%), FT (FT), 8NCore (8C), 8NAggr, 4NEdge/pod: 16, 3× 1 Gbps, 64Servers (64S)

interaction between server 0 and 15 in Fig. 5 is an example of the spine switch bouncing the packet along a uniquely determined route, emphasizing the custom addressing and routing scheme FT architecture deployment. In essence, ECMP is used by Clos-based network to break up traffic [28]. However, hash collisions also deny ECMP from taking advantage of the full bisectional bandwidth, thereby resulting in undesirable delays with moderate network traffic [29]. On the other hand, the non-blocking switches do not cause contention in the network, enhancing the FT DCN capability of achieving full bisectional bandwidth.

The FT simulation setup is illustrated in Table 6. Therefore, the number of switches in the core/spine = the total number of commodity switches in every pod (aggrega-

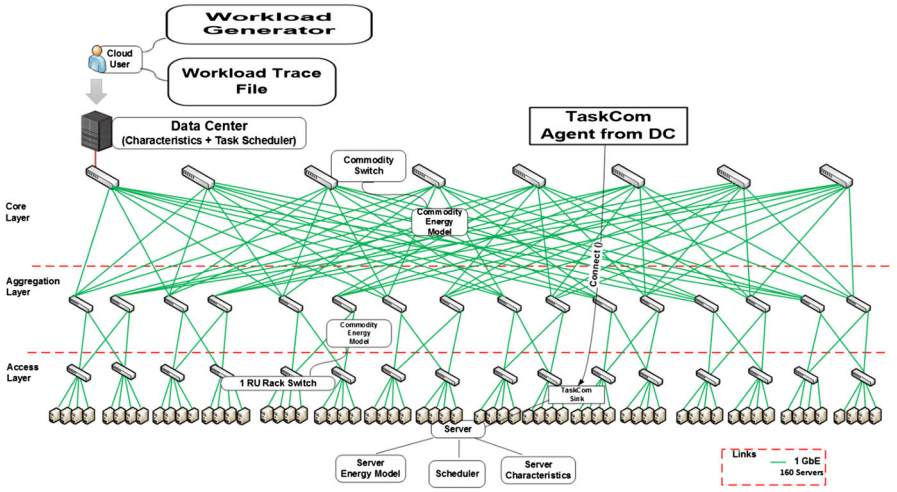

Fig. 6 Schematic of the modeled FT DCN architecture

tion + access) = the whole number of servers in each pod, all interconnected with 1 Gbps links as illustrated in Fig. 6.

4 Results

The 3T and FT DCN architecture models were simulated using modified GreenCloud simulator. Table 7 provides a summary of the results based on 3T and FT simulation setup discussed in the previous section

A total number of 64 computing servers (S) were considered for both DCN architectures, resulting DC total computing capacity of 2.256e8 MIPS for each of the eighteen simulated models. One cloud user was considered.

Overprovisioning DCN architecture for peak load and fault tolerance made DCNs to be mostly underutilized at an average load of 5–25%. Such scenario is exploitable for energy efficiency [30]. The DC load of 50% (half of the DC load capacity) as depicted in Fig. 7 is considered as the best reference point to analyze the two DCN architectures as DCs are collocated to redistribute workload. Typically, idle servers and network modules consume up to 66 and 85%, respectively, of peak load [8,31]. Furthermore, due to the inconsistency of DC workloads, overprovisioning of servers and network modules are rather emphasized to cope with such fluctuations or maximum load variations.

The 50% DC load is chosen as a more realistic workload representation of real operational DC, and it comprises of actual regular workload and workload associated with ITE overprovisioning needed to cope with expected upsurge in workload. Similarly in [8] earlier study has shown that the average load is responsible for about 30% of DC resources[32] while the remaining 70% is mostly put to sleep.

One-third of the load, i.e., idle (30%) DC load, was equally simulated (see Table 7). It creates waste of energy and inappropriate OPEX expense. For example, the I/O buses, memory and disk running account for the total tasks of 14,798. For instance the

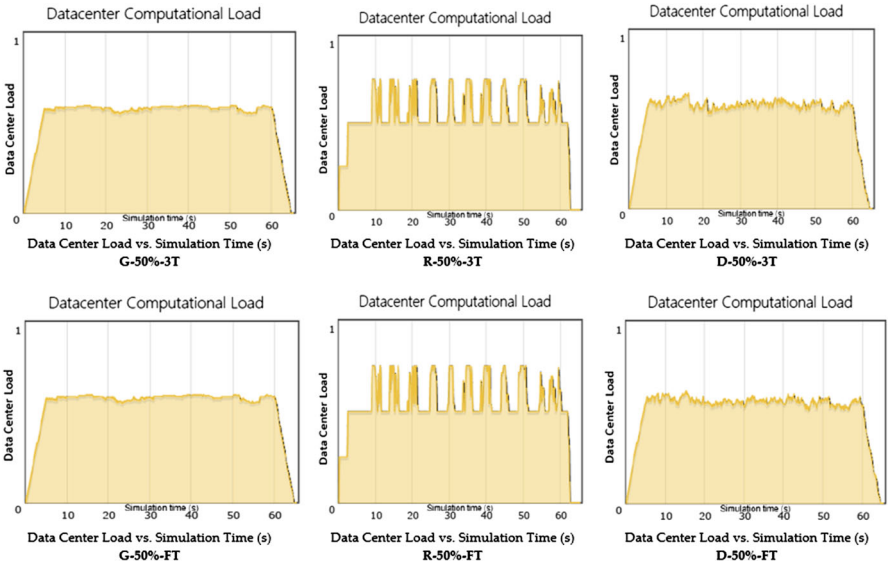

Fig. 7 DC computational load comparison among three task schedulers in 3T and FT

I/O buses, memory and disk running account for the total task of 14,798 at an average rate of 231.2 tasks per server and consuming 843.2 W*h of energy, which according to earlier study can be seen as idle servers task that consumed energy to the tune of two-third of peak lead [15].

Both DVFS and dynamic network shutdown (DNS) power management were implemented in the servers, and only DVFS was implemented in the switches as typical energy-aware scheduling solutions to:

- Consolidate workloads onto the least amount of machines; about 80% reduction of IT load is possible with virtualization and consolidation [33].
- Maximize the numbers of resources enabled for sleep state [34].

The performance of schedulers with regard to network load optimization, task execution time and effect of energy consumed for task performed suggested that the Green scheduler is the best responsive method. The choice of Green scheduler ensures incoming workloads are consolidated into minimum numbers of servers based on the jobs' processing requirements.

For 3T, we considered G-50%-3T Green task scheduler which has lowest power consumption compared to other schedulers, and higher task performed: 874W*h produced a total of 29,596 tasks at an average of 462.4 tasks per server. For FT, we considered G-50%-FT which has second lowest power consumption [35] but higher task performed. The task scheduler which has lowest power consumption is 1618.1W*h produced by BestDENS task scheduler (D-50%-FT) compared to Green's (G-50%-FT) 1620.3 W*h, but the total task performed and average task per server by D-50%-FT (i.e., 26143 and 408.5) is lower than of G-50%-FT (i.e., 29596 and 462.4). This emphasizes the fact that D-50%-FT is a best-fit scheduling algorithm for data-intensive

Table 7 Summary of DCN simulation results for 3T and FT

DC Architecture/ load (%)	Task scheduler ID	DCN energy consumption (W*h)					Power management		Total task	Average task per server
		Core switch	Aggr. switch	Access switch	Server	Total	Server	Switches		
3T										
30	G	217.3	434.6	43.1	148.2	843.2	DVFS+DNS	DVFS	14,798	231.2
50	G	217.3	434.6	43.1	179.0	874	DVFS+DNS	DVFS	29,596	462.4
100	G	217.3	434.6	43.1	220.7	915.7	DVFS+DNS	DVFS	49,327	770.7
30	R	217.3	434.6	43.1	148.1	843.1	DVFS+DNS	DVFS	14,798	231.2
50	R	217.3	434.6	43.1	179.4	874.4	DVFS+DNS	DVFS	29,596	462.4
100	R	217.3	434.6	43.1	220.9	915.9	DVFS+DNS	DVFS	49,327	770.7
30	D	217.3	434.6	43.1	149.2	844.2	DVFS+DNS	DVFS	14,798	231.2
50	D	217.3	434.6	43.1	179.1	874.1	DVFS+DNS	DVFS	27,130	423.9
100	D	217.3	434.6	43.1	223.7	918.7	DVFS+DNS	DVFS	49,327	770.7
Link (GB)		10	10	1	0.416*					
Total no. of ITE		4	8	16	64					
FT										
30	G	466.1	932.1	43.1	148.3	1589.6	DVFS+DNS	DVFS	14,798	231.2
50	G	466.1	932.1	43.1	179.3	1620.3	DVFS+DNS	DVFS	29,596	462.4
100	G	466.1	932.1	43.1	220.7	1662	DVFS+DNS	DVFS	49,327	770.7
30	R	466.1	932.1	43.1	148.1	1589.4	DVFS+DNS	DVFS	14,798	231.2
50	R	466.1	932.1	43.1	179.4	1620.7	DVFS+DNS	DVFS	29,596	462.4
100	R	466.1	932.1	43.1	220.9	1662.2	DVFS+DNS	DVFS	49,327	770.7
30	D	466.1	932.1	43.1	149.4	1590.7	DVFS+DNS	DVFS	14,798	231.2
50	D	466.1	932.1	43.1	176.8	1618.1	DVFS+DNS	DVFS	26,143	408.5
100	D	466.1	932.1	43.1	223.7	1665	DVFS+DNS	DVFS	49,327	770.7
Link (GB)		1	1	1	1					
Total no. of ITE		8	16	16	64					

Table 8 DCN link utilization in 3T and FT

DCN	Switch/network level	Type/count/speed	Link utilization	
			Downlink	Uplink
3T	Access/edge	TOR/48× 1GE/2× 10GE	4× 1GE to 4 server rack (44× 1GE Idle)	2× 10GE to Aggr. switch
	Aggregation	EOR/32× 10GE	4× 10GE to Access	4× 10GE to Core 2× 10GE to Aggr. (C_2–C_2)
	Core	Core/32× 10GE	8× 10GE	≤24× 10GE
FT	Access/edge	Commodity switch 48 port 1/10GE, 4× 10 GE or 40GE	4× 1GE to 4 server rack (44× 1GE Idle)	2× 10GE to Aggr.
	Aggregation	Commodity switch 48 port 1/10GE, 4× 10 GE or 40GE	2× 10GE to Access	8× 10GE to Core
	Core	Commodity switch 48 port 1/10GE, 4× 10 GE or 40GE	8× 10GE to access	≤40× 10GE

workloads (DIW) [11]. It is assumed that GE links are Green Ethernet (LPI enabled). The link utilization is illustrated in Table 8.

For comparison, the same number of computing nodes of 64 servers was used for both topologies while the network links to switches varied. In 3T DCN, the architecture provides bandwidth of 10 Gbps link in the core, aggregation and access layers network compared to the FT where three layers are interconnected with 1 Gbps links. Thus, the bandwidth in C_1–C_2 and C_2–C_3 links in 3T is ten times higher than the corresponding links in FT. The dissimilarity between downlink and uplink bandwidth capacity in every switch layer (BOR) of 3T is such that:

- The edge switch has two 10 Gbps links to the aggregation network and with 48 ports at 1 Gbps link downlink to support 48 servers:

$$\frac{48\,\text{Gbps}}{20\,\text{Gbps}}\text{provides a BOR of} = 2.4:1$$

and a corresponding per server bandwidth of:

$$\frac{1\,\text{Gbps}}{2.4} = 416\,\text{Mbpsatmaximumload}$$

The BOR for FTs C_1–C_2, C_2–C_3 and C_3–S is 1:1 due to the 1 Gbps links at all levels in the network. The latency experienced at all links for both topologies is 0.0033ms.

Support for ECMP routing [36] was assumed and made available in 3T with the usage of high-end switches at the core and aggregation layer [37] and the availability of 10 Gbps link between them which caters for the BOR. The extension of 10 Gbps link to the access network further provides for throughput and enhances the ECMP routing closer to the host to reduce possibility of congestion.

Similarly, it is assumed that ECMP, DRB's per packet round robin enabled routing was implemented as the adequate conditions of load balancing in FT, using BOR of 1 to an advantage and without congestion. The core switches act as bouncing switches to perform the routing [27], and commodity switches in aggregation and access layers were utilized. The same routing scheme is assigned based on the number of nodes with each pod of FT with $k = 4$.

Table 9 represents the analysis of the server—network module layout of both architectures, while Figs. 7 and 8 illustrate their energy usage.

4.1 Power utilization in information technology equipment

Equations (1) and (2) are employed in the simulator to compute the power utilization of servers and switches. It should be noted that the power factor is the same for the server, i.e., P_{fixed} as the same server specification is used. The power factor of the chassis, line card and port's transfer rate for the core and aggregate switches will be the same for 3T but different for FT as commodity switches are utilized in FT. The power factor in the access switches is the same for the two architectures.

Energy consumption ratio allocated to ITE is approximately 40% of the whole DC infrastructure [8]. The distribution varied based on the composition simulation

Table 9 ITE module layout in 3T and FT

DC architecture/load	No of DCN components								Average DCN energy (K*W) at 50%	Remarks
	Network modules				Links (Gbps)			Servers		
	Core	Aggr.	TOR	Total	10	1	Total			
Three Tier at 50%	4	8	16	28	72	64	136	64	865.5	32× 10G link for C_1–C_2 32× 10G link for C_2–C_3 8× 10G link for C_2–C_2 64× 1G link for C_3–S
Fat tree at 50%	8	16	16	40	–	160	160	64	1611.8	64× 1G link for C_1–C_2 Pods 32× 1G link interconnecting C_2–C_3 64× 1G link C_3–S

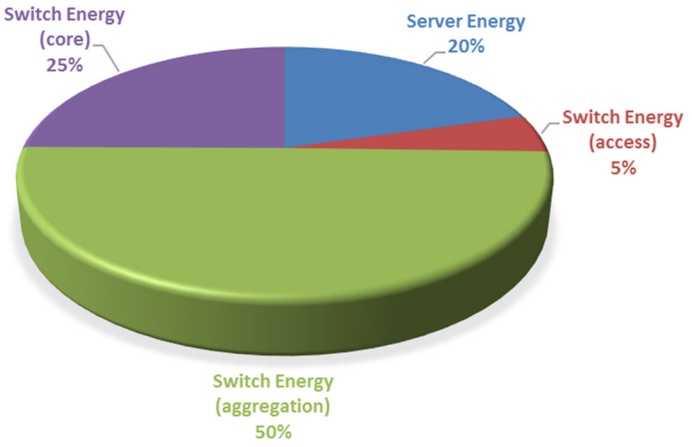

Fig. 8 Energy ratios for ITE module in 3T

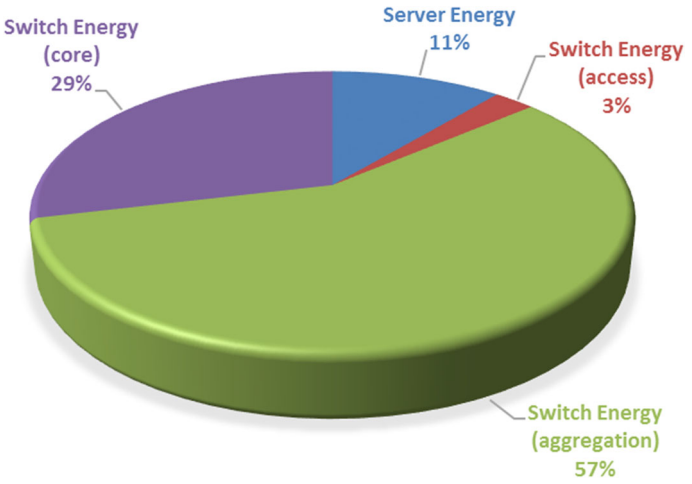

Fig. 9 Energy ratios for ITE module in FT

component. As depicted in Figs. 8 and 9 for 3T and FT, respectively, the ratio of energy consumed by network modules (all switches) to servers is approximately 4:1 in 3T and 9:1 in FT. The higher energy rate in FT is a result of the *k-ary* pods arrangement which resulted in higher numbers of commodity switches utilized to accommodate 64 servers.

The energy consumption of servers, L2 and L3 switches considered in G-50%-3T is displayed in Fig. 10. It can be observed that 93.8% of the total energy was consumed by 40 out of the 64 computing servers (62.5% of the servers) as shown in Fig. 10b. The remaining 24 servers (37.5%) consumed less than 50% of computing energy, i.e., 179.3 W. Network energy management policies of DVFS and DPM applied were responsible for varying energy use in the racks due to availability of workloads across the network [11,38]. The core and aggregation switches operated at approximately 95% of full energy in 3T [10] (see Fig. 10c, d). These layers are needed for ECMP

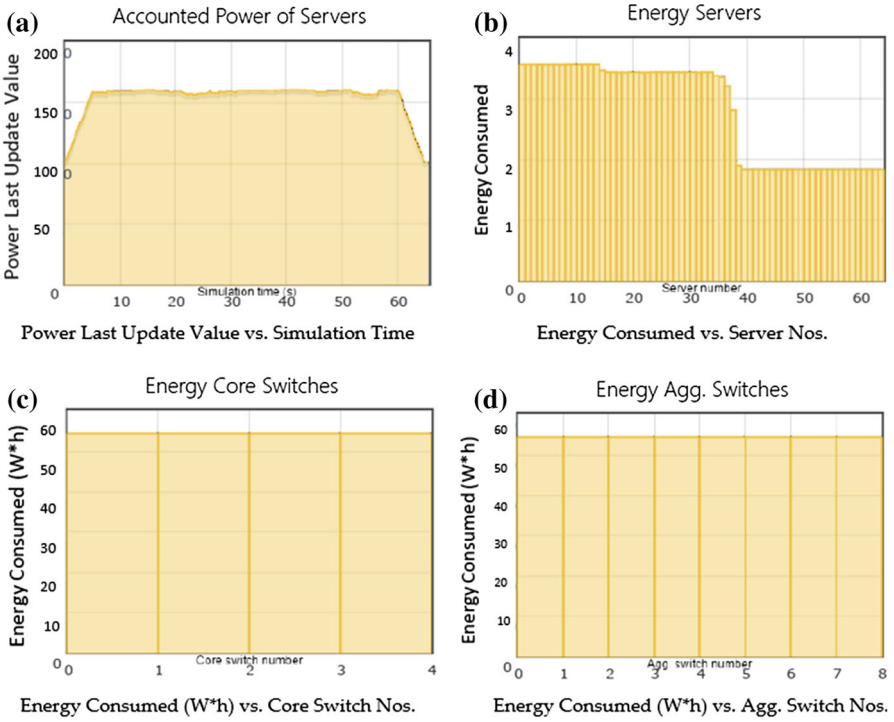

Fig. 10 Energy consumption of ITE in 3T G-50%

routing, and DNS technique is not encouraged as it may degrade network performance. The layers are also overprovisioned for this purpose. Network module overprovisioning accounted for the larger portion of consumed power by the upper layer as shown in Fig. 8.

The energy consumption of servers, L2 and L3 switches in G-50%-FT is displayed in Fig. 11. The distribution of energy usage among the 64 servers for the FT is similar to that of 3T as shown in Fig. 11b. However, the commodity switches that replaced the energy hungry enterprise switches in 3T at the upper layers are larger in quantity and are actively involved in end-to-end aggregation of bandwidth to host servers [9,10,26,27,39], resulting an increased energy consumption of the network module in FT (see Fig. 11c, d).

Both 3T and FT have the same energy utilization at the access level with 95% energy consumption and 1 Gbps bandwidth provision for each link to the computing servers.

4.2 Uplink and downlink in information technology equipment

To obtain corresponding power factor, changes were made in the setup parameter for switches to accommodate the low-power forms of the large numbers of commodity switches and port density.

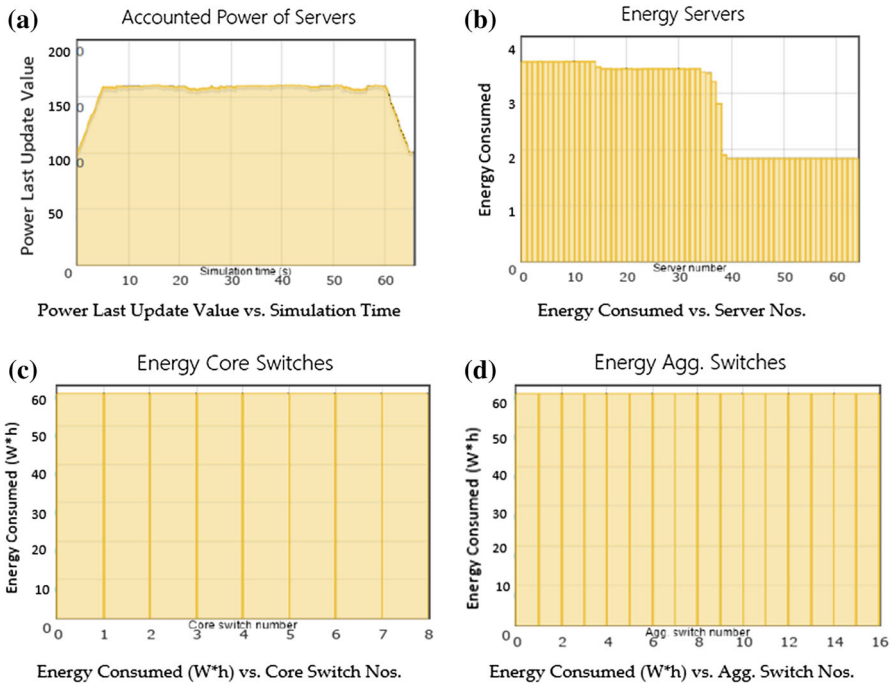

Fig. 11 Energy consumption of ITE in FT G-50%

Uplink comparison:

The uplink network traffic summary for 3T (Fig. 12) illustrates the effect of bandwidth oversubscription in the upper layers of the topology [10] with 60% of the core-access link actively utilized (see Fig. 12a), and the core usage substantially higher anticipated by smaller number of link multiplexing traffic from the layers below [9,19].

For IT load of 50%, the hosts to racks (NIC-TOR) connectivity experienced a decreasing link load from 90 to 10% out of 1 Gbps bandwidth apportioned with only 61% inter-connections active out of the 64× 1 Gbps links to the servers, i.e., BOR of $\frac{64\times1}{32\times10} = 0.2 : 1$, that is, 200 Mbps per server bandwidth (see Fig. 12a). Likewise for the EOR network, it experienced approximately 60% link load from 62.5% of the 16× 1 Gbps links supported by 10 Gbps aggregation layer links, i.e., BOR for TORs-EOR of $\frac{4\times10}{4\times10} = 1 : 1$ (see Fig. 12b).

The core layer with four core switches with a total of 40 Gbps links to the TOR, i.e., 4× 10 Gbps links to C_2/EOR experienced 93.3% link load with about 75% of the 4× links utilized (see Fig. 12c). BOR is $\frac{64\times1}{4\times10} = 0.4 : 1$. The BOR for the link favors the upper layer oversubscription, with traffic queue experienced at the 200 Mbps host to rack link (see Fig. 12d–f). DVFS and DPM factors are responsible for the nonlinear variation in the provisioned bandwidth resources across each layer [11,14].

The uplink network traffic summary shown in Fig. 13 illustrates the effect of BOR of 1:1 across all layers in the FT architecture [5,29]. For IT load of 50%, the NIC-TOR connectivity experienced a decreasing link load from 90 to 10% out of 1 Gbps

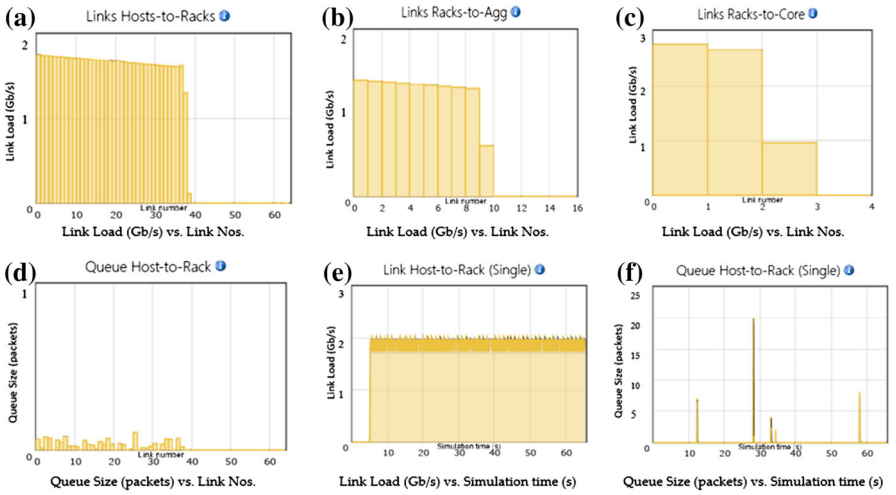

Fig. 12 3T DC network uplink

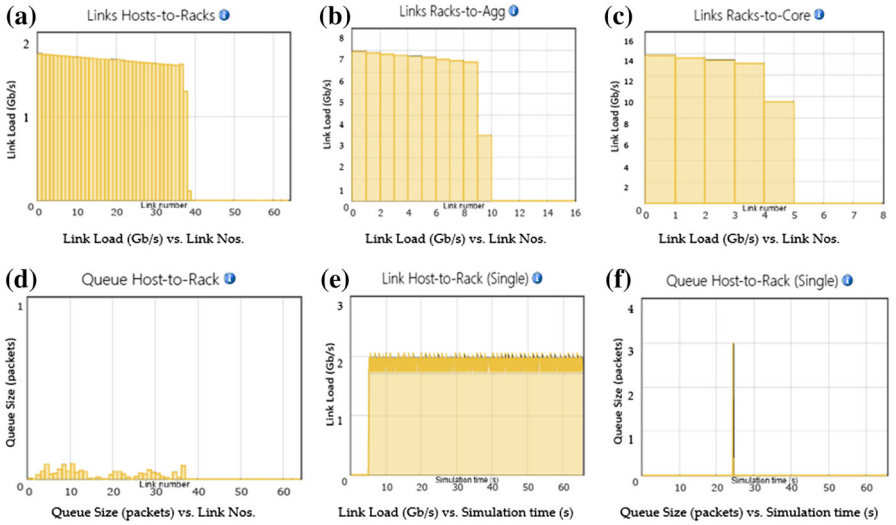

Fig. 13 FT DC network uplink

bandwidth apportioned with only 61% inter-connections active out of the 64× 1 Gbps links to the servers, i.e., BOR of $\frac{4}{4} = 1 : 1$. That is 2 Gbps per server bandwidth available but usage limited to 1 Gbps capacity of NIC (see Fig. 13a). Therefore, link capability is 10 times that of 3T, and this provides full bisection bandwidth between the hosts to rack [5], where servers within the network are able to communicate with one another arbitrarily at full bandwidth of their NIC.

Likewise for the EOR network, it experienced approximately 87.5% link load from 62.5% of the 16× 1 Gbps links at aggregation layer, i.e., BOR for TORs-EOR of $\frac{4}{2} = 2 : 1$ in line with the *k-ary* pod tree [5] (see Fig. 13b). With higher packet losses

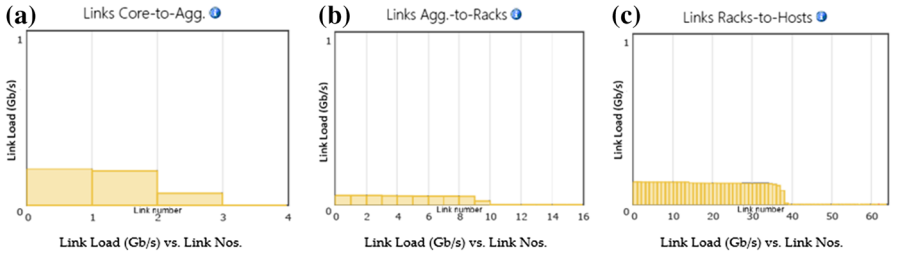

Fig. 14 3T DC network downlink

expected at the aggregation layer [9], a link BOR of 2:1 will be appropriate for the network since the layout in $k/2$ for both access and aggregation layers in Clos topology is the same. There are two access layer and two aggregation layer switches in each pod which eventually guarantees $4\times$ 1 Gbps link within this layer in a pod.

The core layer is made up of 8 core switches, with $4\times$ 1 Gbps link connected to one out of two aggregate switches in a pod. The rack to the core links ratio is such that $\left(\frac{k}{2}\right)^2$, i.e., $4\times$ 1 Gbps links available per pod to 4 computing servers as depicted in Fig. 6. Therefore, the racks to core link experienced 87.5% (14 out of 16) link load utilized by 62.5% of the links, i.e., 5 out of 8 links (see Fig. 13c). BOR is $\frac{8\times1}{8\times1} = 1:1$. The diffusion optimization of the traffic flow available with the links state is that it prohibits local congestion by assigning traffic to ports on per flow and not per host basis [5].

The flow scheduling removes global (DCNs) congestion and prevents elephant flows in need of high throughput [9] from sharing a link by assigning such flows to different links. A negligible traffic queue which lasted for less than 0.0033ms was experienced during the simulation (see Fig. 13d–f).

Downlink Comparison:

The downlink network traffic in 3T as depicted in Fig. 14 is such that a quarter of 40 Gbps total bandwidth was utilized through three out of four links to the aggregate layer (see Fig. 14a). The aggregation layer has abundant bandwidth with 40 Gbps link from $4\times$ upper level switches and same downlink bandwidth link provisioning to the TOR, i.e., BOR of $\frac{40}{40} = 1:1$, making 62.5% of the TOR switches (16) to utilize only 10% of the total link load at the aggregation layer (see Fig. 14b). The rack to host downlink is such that only 25% of the link load is utilized by 59% of the computing servers (see Fig. 14c). In case of increasing load, the 0.2:1 link BOR is insufficient as it offers only approximately 200 Mbps per server bandwidth which is lower compared to the BOR in the upper layer. In the case of CIWs where the computing server produces traffic at non-blocking bandwidth of the NIC (1 Gbps) which is more than the available bandwidth, congestion is likely to occur at the TOR and aggregation switches [10]. For BWs considered in this project, the link utilization is of equal importance as DIWs emphasize on throughput of network paths. The competition for core layer bandwidth by the TOR switches and servers associated with it is based on requested broadcast at the full bandwidth capacity of NIC [26], thereby making energy spent to support higher bit-rates enormous. However, the higher bit-rates cannot be utilized by the hosts or computing server [10]. This bottleneck of end-to-end aggregate bandwidth a.k.a.

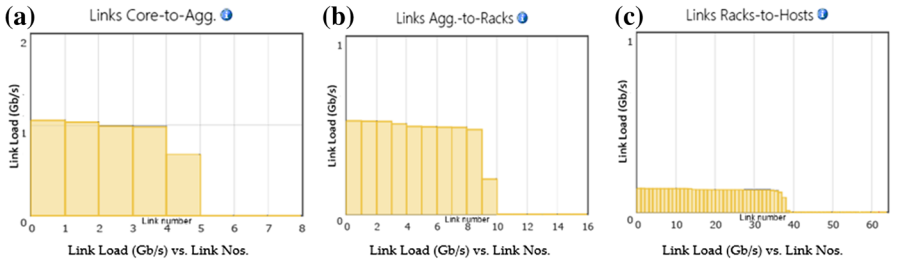

Fig. 15 FT DC network downlink

cross-section bandwidth degrades network performance in 3T [4]. Moreover, TCP incast congestion could develop at the access switch in intra-DCN for many-to-one traffic mode when multiple requests are sent toward a single server within the rack and throughput is low [40].

In FT DCN downlink illustrated in Fig. 15, approximately 50% of the link load was utilized with 62.5% ($\frac{5}{8}$) of the 8× 1 Gbps link per pod between C_1–C_2 (see Fig. 15a). Similarly, the same occurred at the C_2–C_3 links, i.e., ($\frac{10}{16}$) = 62.5% as shown in Fig. 15b. However, racks to hosts recorded a downlink link load utilization of about 20% by 59.3% ($\frac{39}{64}$) of the servers (see Fig. 15c). This indicates that the throughput between any two hosts equals 1 Gbps with the application of ECMP routing in FT, i.e., identical bandwidth path at any available bisection [41,42].

4.3 Power-related metric comparison in information technology equipment

This part focuses on the application of performance and energy-efficient metrics targeted toward communication systems in DCs. The 64 computing servers scheduled with balanced workloads have different per server bandwidths: 416 Mbps for 3T and 1 Gbps for FT. Therefore, the CNEE in Joules/bit (J/bit) and NPUE for the two DCN topologies are calculated as in [10] and derivable from Figs. 8 and 9.

$$\text{CNEE} = \frac{\text{Power consumed by network equipment (all hardware involved in information delivery)}}{\text{Effective network throughput capacity (maximum end to end)}} \tag{3}$$

$$\text{NPUE} = \frac{\text{Total power consumed by ITE}}{\text{Power consumed by network equipment}} \tag{4}$$

Assuming the GE link is energy efficient: Green Ethernet [43] and are power over Ethernet (POEs), and using the values at 50% DC load, the power-related metrics are calculated as:

$$\text{CNEE} : 3\text{T} = \frac{874.3}{416}, \text{FT} = \frac{1620.6}{1000}$$
$$\text{NPUE} : 3\text{T} = \frac{874.3}{695.0}, \text{FT} = \frac{1620.6}{1441.3}$$

5 Discussion

Having analyzed briefly the results of the simulation, the discussion will focus on the application of energy management policies setup for DCN in terms of management role, planning role and beliefs, and some of the performance and energy-efficient metrics targeted toward communication systems in DCs. Considering the theory and practice of network management policies in [44], which also encompasses the DVFS, DNS and DPM methodologies, the findings of this study suggest the following:

(1) The 3T architecture is notorious with expensive, high-end energy hungry switches at both core and aggregation layers due to the physical layout in a multiple rooted tree-based structure. To improve on the per server bandwidth of existing 3T, the aggregation to access layer link was oversubscribed with 10 Gbps link. However, the BOR limitation from upper layer still significantly affects the server uplink capability with the NIC maximum bandwidth of 1 Gbps and that of the TOR switch. This is responsible for higher CNEE. Aforementioned limitation still persists with two NICs per server.

CIWs and BWs jobs will result random congestion at the C_3–S layer at likely peak DC load due to bandwidth oversubscription at upper layer. Scalability is rather difficult as the core—aggregate layer is rigidly structured, unlikely to route task to servers outside the TOR network, hence also making it fault intolerant.

At idle DC load, the energy hungry core switches cannot be energy-managed as they are responsible for ECMP routing; thus, operators of such DC will indulge this spending as part of OPEX. Only aggregation switches with idle racks can be powered down and set with minimum wakeup time in case of load upsurge. This could still account for performance degradation in the network and decreased CNEE utilization in the topology. Consequently, unused servers in a 50% loaded DC are harder to localize as the topology is not compartmentalized, and hence, consolidation of active server into fewer rack becomes more difficult. Idle CPU still runs at 66% of peak load; thus, DVFS is not applicable. Outright shutdown with DPM is preferred but awaking server in idle rack drains a considerable amount of energy.

Lastly with reference to link utilization in Table 8, the unused links are automatically powered down with the switches and server except in cases where the port speed is step down in the aggregate layer, e.g., from 10 to 1 Gbps to save energy instead of DNS to cater for traffic fluctuation and preserve minimum level of QoS.

However, the downlink in 3T as shown in Fig. 14 confirms the problem of cross-section bandwidth attributed to end-to-end aggregation bandwidth bottleneck, alongside with those of scalability, cost and non-agility discussed earlier. The overall effect of cooling is enormous as power hungry switches have high heat dissipation, posing more power requirement challenges to heating, ventilation and air-conditioning (HVAC) system. Table 11 illustrates the energy and cost implication of the simulated models for 3T.

(2) The FT is equally switch-centric in nature with the ability to handle oversubscription and full bisectional bandwidth. As given in Table 8, symmetric end-to-end bandwidth utilization between any two nodes in the network has a BOR of 1.1 equal to 1 Gbps, making it suitable for BW jobs. The choice of Green scheduler ensures incoming workloads are consolidated into minimum numbers of server based on the

Table 10 Power-related metric evaluation

Power metrics	DC architectures		Remarks
	3T	FT	
CNEE	2.10 J/bit	1.62 J/bit	Although in overall FT consumes more energy, it spends less energy for transmission of a single bit of information, outperforming 3T
NPUE	1.25	1.12	In 3T for every 1.25 watts of spent on ITE, 1 watt is used to power the network modules equaling 44% energy spent on network module and 55.55% on servers, likewise NPUE of 1.12 for FT translates to 47% power on network module and 52.83% on servers

jobs' processing requirements. ECMP in spine and leaf network segment is similar to the assumption of adequate condition (customized addressing scheme) that has been met to enable bouncing switching in DRB routing algorithm convey a sole bit of information with a lower possible energy level, i.e., CNEE of 1.62 J/bit when compared to 3T's 2.10 J/bit.

The FT architecture is switch laden, and the larger number of switches, i.e., 40 inexpensive commodity switches when compared to 28 enterprise switches used in 3T as given in Table 9 accounted for 1441.3 W*h of energy, though as commodity switches they still consume less energy.

A considerable amount of the large number of commodity switches and the resulting port density is put to sleep using DPM scheme as DNS will degrade the network performance. Furthermore, the power factor in the commodity switches is more than 50%, less than that of 3T's core/aggregation layer formation. Table 11 illustrates the energy and economic benefit of FT architecture using real-world DCN interconnectivity.

However, the spine–leaf network organizations regarded as folded Clos which support high port count at the spine could reduce layer of the topology into two substantive layers. The uppermost layer above the spine implements L3-based routing protocol that acts as a control plane or load balancer for traffic, minimizing latency and providing congestion avoidance. The L3 routing table efficiently route packet from spine to source with egress port selection performed at the leaf, i.e., L3 lookup existing at the node. This scenario is given in Table 3. Utilization of multiple 10 Gbps links for spine–leaf connection instead of a singular 40 Gbps fiber link reduces power consumption by more than 10 times.

(3) The *k-ary* pods help consolidate traffic on fewer racks and add agility and scalability as commodity switches can be added to any layer to extend the computational capacity of the fabric, which results in more cost-effective and less energy consumption as shown in NPUE comparison in Table 10. Overall effect of using commodity switches is reduced cost in terms of CAPEX, lower energy for network modules and lower heat dissipation, reducing the OPEX on cooling also.

Cabling complexity and increased cable cost can be observed in FT as given in Table 9 with 160 links when compared to 136 interconnectivities in 3T. The Green

Table 11 Power Budget for 3T and FT DCN Architectures

DC Architecture	DCN Components Rating							Remarks
	ITE Modules					OpEx/ CapEx Spend		
	ITE/Port Density	QTY	Max. Power Rating(W)		Estimated Total Power (W)	Thermal Rating	Unit Cost(£)	
			P_{chasis}	$P_{Linecard}$	$P_{chasis} + P_{Linecard}$	(BTU/hr)		
3T	Core (32p)	4	385	10 per port	1,540 + 1,280	4 x 9,737	4 x 29,000:00	Cisco Nexus 7000 Switches:
	Aggregate (32p)	8	385	10 per port	3,080 + 2,560	8 x 9,737	8 x 29,000:00	Front to back, side to side, side to back airflow, enterprise, L3 or L2, less power efficient. Power, BTU at [46], Price at [47]
	TOR (48p)	16	143	4 per port	2,288 + 3,072	16 x 1,235	16 x 17,000:00	Cisco Nexus 3064 switches: Front and rear airflow; AC and DC power inputs; Cost effective, power efficient; Line rate Layer 2 and 3 TOR switches.
	Server (3p)	64	750	–	48,000	64 x 2,812	64 x 4,500:00	HP ProLiant DL 385p Gen8 server: Rated steady state power supply of 750W at 200 – 240 V ac and BTU at [48]. Aggregated costing at [49]
	Links-10G	72	15.4		1,108.8		10 reels x 30/(100ft)	15.4 is maximum but only about 12.95 W is assured to power device because dissipation occurs in the cable [50]. Green Ethernet [43]. Pricing at [51,52].
	Links-1G	64	15.4		985.6		9 reels x 30/(100ft)	
	Estimated TOTAL @ Full DC Load				63,914.4	359,936	908,570	It is assumed that the power consumption will be 50% of the power budget. Cooling spending on HVAC/CRAC is proportional to DCN load.
	DC Load @ 0.5				31,957.2	179,968	908,570	50 percent DC load is considered to be moderate.

Table 11 continued

DC Architecture	DCN Components Rating							Remarks
	ITE Modules					OpEx/ CapEx Spend		
	ITE/Port Density	QTY	Max. Power Rating(W)		Estimated Total Power (W)	Thermal Rating	Unit Cost(£)	
			P_{chasis}	$P_{Linecard}$	$P_{chasis} + P_{Linecard}$	(BTU/hr)		
FT	Core (48p)	8	143	4 per port	1,148 + 1,536	8 x 1,235	8 x 17,000:00	Cisco Nexus 3064 switches:
	Aggregate (48p)	16	143	4 per port	2,288 + 3,072	16 x 1,235	16 x 17,000:00	Front and rear airflow; AC and DC power inputs; Cost effective, power efficient; Line rate Layer 2 and 3 TOR switches. Power & BTU at [53], Price at [47].
	TOR (48p)	16	143	4 per port	2,288 + 3,072	16 x 1,235	16 x 17,000:00	Cisco Nexus 3064 switches: Front and rear airflow; AC and DC power inputs; Cost effective, power efficient; Line rate Layer 2 and 3 TOR switches.
	Server (3p)	64	750	–	48,000	64 x 2,812	64 x 4,500:00	HP ProLiant DL 385p Gen8 server: Rated steady state power supply of 750W at 200 – 240 V ac and BTU at [48]. Aggregated costing at [49]
	Links (1G)	160	15.4		2,464		10 reels x 30/(100ft)	15.4 is maximum but only about 12.95 W is assured to power device because dissipation occurs in the cable [50]. Green Ethernet [43]. Pricing at [51,52].
	Estimated TOTAL @ Full DC Load				63,868	229,368	968,300	50 percent DC load is considered to be moderate.
	DC Load @ 0.5				31,934	114,684	968,300	

Ethernet assumed (IEEE 802.3az standard) for the links is expected to surmount issues regarding link energy. With challenges of port count on Green Ethernet switches, turning off idle devices provides instantaneous savings. It is estimated that 80% power saving is made possible with consolidation and virtualisation; longevity is further ensured for network device in the absence of incessant heat dissipation.

Most ITEs operate at 2/3 of designed power rating, e.g., the HP server operates with a dynamic power capping tools available at the integrated light-out (iLO) user interface or set through HP insight control of the power management module. At different load variations, the network management roles, planning rules and beliefs apply. It is estimated that OPEX on energy is proportional to DC load and likewise in cooling. DVFS and DPM power management techniques were used to optimize energy efficiency while maintaining QoS/SLA. CAPEX is constant for a while, sustained by ITEs efficiency and dependent on the mean time between failures (MTBF) of ITEs. From Table 11, we observed that FT uses 23.2 watts less energy to support the same numbers of computational servers, though the initial total cost of ownership (TCO) is higher. For every 3.412 BTU/h generated 1 watt of energy is expended, and the heat dissipated is relatively proportional to workload. Reduced thermal overrun [45] through consolidation, virtualization and powering down inactive ITEs consequently bring to barest minimum the energy consumed by the computer room air-conditioning (CRAC) unit in cooling the server room. In real operation of DCs scaling to tens of thousands of servers, cooling load reduction will result in significant OPEX savings.

Reduced utilization, oversubscription, manual configuration and scale-out against scale-up, e.g., per port charge, cabling complexity and expandable cooling, are challenges faced when trying to attain DCN design goals of scalable interconnection bandwidth, throughput and load balancing at low power and cooling cost.

5.1 Related work

The energy consumption results obtained in the experiment comparing of 3T and FT DCN architectures are similar to [5,9,54]. In [54], it was concluded that network module energy consumption (approximately 4–12%) should not be ignored although the majority is consumed by servers. Energy-saving policies also influenced the outcomes and FT showed higher percentage of energy utilization. In [9], using ns-2 simulator it was demonstrated that data center TCP (DCTCP) congestion control with TCP incast/outcast in FT is better for elephant flows in need of high throughput and mice flows that needed low delays. The focus was on how the DCTCP deploys explicit congestion alertness to augment TCP congestion regulatory algorithm. This allows for optimal performance in FT, leveraging on BOR of 1 across all layers and prevents incast and outcast bottleneck, making FT a DCN of choice for large networks. The analysis by Al-Fares et al. [5] pioneered the study of FT DCN architecture as a cost-effective, cheap commodity-based topology, scalable bisection bandwidth and in terms of lower power consumption and thermal output as shown in Fig. 16 where 3T is regarded as the hierarchical design. The analysis in Fig. 17 was obtained from the

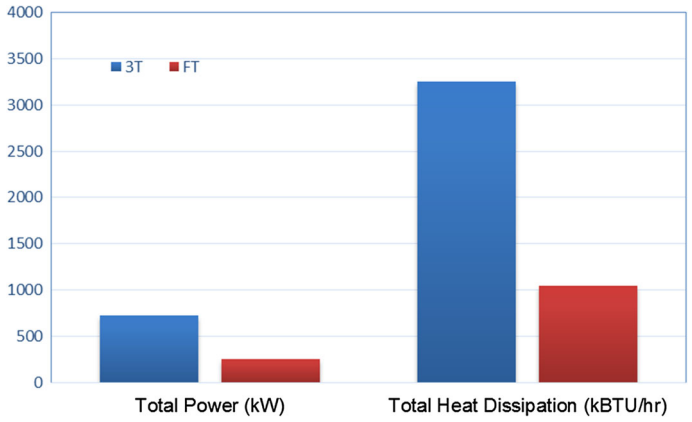

Fig. 16 Comparison of total power consumption and heat dissipation. Adapted from [5]

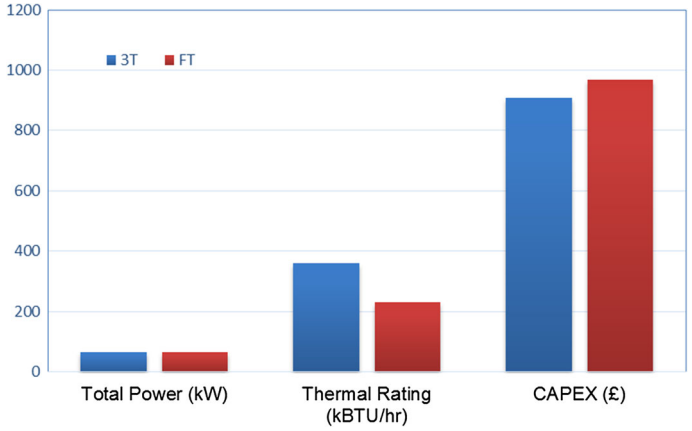

Fig. 17 Comparison of total power consumption, thermal rating and CAPEX

power budget for 3T and FT DCN architectures presented in Table 11, and the result is similar to that of [5] in Fig. 16.

It is worth mentioning that FT is DCN architecture that both a symmetric, i.e., it has organized packaging and simple physical wiring of the topology, and recursive defined in nature, i.e., numbers of levels or layers, are not fixed but increase with topology size [55]. These two factors are attributes of scalability possessed by the FT. Furthermore, the scalability and deterministic nature of FT made variants of the DCN architecture implemented by two IT giants: Google FT [5] in 2008 and Facebook FT [56] in 2014 possible. Basically, the application of these variants of FT topology was partly responsible for PUE of between 1.15 and 1.17 in 2010 cut to 1.06 in 2014 by Google [57–59] and PUE of 1.08 in 2014 recorded by Facebook [60–62], respectively.

6 Conclusion

In this article, we compared the energy-related performance of two most popular switch-centric DCN architectures: three-tier and fat tree. We also compared CAPEX, thermal and power consumption cost using real-world scenarios. The FT is equally switch-centric in nature with the ability to handle oversubscription and full bisectional bandwidth. The *k-ary* pods help consolidate traffic on fewer racks and add agility and scalability as commodity switches can be added at any layer to extend the computational capacity of the fabric which results in more cost-effective and less energy consumption DC architecture. Overall effect of using commodity switches has reduced cost in terms of CAPEX, lower energy for network modules and lower heat dissipation, decreasing the OPEX on cooling also.

References

1. Pranggono B, Tianfield H (2014) Green data center. In: Green technology applications for enterprise and academic innovation. IGI-Global, pp 179–199
2. Bilal K, Khan SU, Manzano M, Calle E, Madani SA, Hayat K et al (2015) Modeling and simulation of data center. In: Handbook on data centers, pp 945–958. Springer, New York
3. Pranggono B, Alboaneen D, Tianfield H (2015) Simulation tools for cloud computing. In: Simulation technologies in networking and communications: selecting the best tool for the test, pp 311–335. CRC Press
4. Bilal K, Malik SUR, Khalid O, Hameed A, Alvarez E, Wijaysekara V et al (2014) A taxonomy and survey on green data center networks. Future Gener Comput Syst 36:189–208
5. Al-Fares M, Loukissas A, Vahdat A (2008) A scalable, commodity data center network architecture. ACM SIGCOMM Comput Commun Rev 38:63–74
6. Rouse M (2010) Green data center. TechTarget Network
7. Song Y, Wang H, Li Y, Feng B, Sun Y (2009) Multi-tiered on-demand resource scheduling for VM-based data center. In: The 9th IEEE/ACM International Symposium on Cluster Computing and the Grid
8. Kliazovich D, Bouvry P, Khan SU (2012) GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. J Supercomput 62:1263–1283
9. Zafar S, Bashir A, Chaudhry SA (2016) On implementation of DCTCP on three-tier and fat-tree data center network topologies. SpringerPlus 5:766
10. Fiandrino C, Kliazovich D, Bouvry P, Zomaya AY (2015) Performance and energy efficiency metrics for communication systems of cloud computing data centers. In: IEEE International Conference on Cloud Computing (CLOUD), New York, USA
11. Kliazovich D, Bouvry P, Khan SU (2013) DENS: data center energy-efficient network-aware scheduling. Clust Comput 16:65–75
12. Bilal K, Khan SU, Zhang L, Li H, Hayat K, Madani SA et al (2013) Quantitative comparisons of the stateoftheart data center architectures. Concurr Comput Pract Exp 25:1771–1783
13. Pepeljugoski PK, Kash JA, Doany F, Kuchta DM, Schares L, Schow C et al (2010) Low power and high density optical interconnects for future supercomputers. In: Optical fiber communication, San Diego, p OThX2
14. Mahadevan P, Sharma P, Banerjee S, Ranganathan P (2009) A power benchmarking framework for network devices. In: Proceedings of the 8th International IFIP-TC 6 Networking Conference, Aachen

15. Chen G, He W, Liu J, Nath S, Rigas L, Xiao L et al (2008) Energy-aware server provisioning and load dispatching for connection-intensive internet services. In: 5th USENIX symposium on networked systems design and implementation, San Francisco, pp 337–350
16. Gmach D, Rolia J, Cherkasova L, Kemper A (2009) Resource pool management: reactive versus proactive or let's be friends. Comput Netw 53:2905–2922
17. Gandhi A, Harchol-Balter M, Kozuch MA (2012) Are sleep states effective in data centers? In: Green Computing Conference (IGCC), CA, San Jose, pp 1–10
18. Bilal K, Manzano M, Khan SU, Calle E, Li K, Zomaya AY (2013) On the characterization of the structural robustness of data center networks. IEEE Trans Cloud Comput 1:64–77
19. Benson T, Akella A, Maltz DA (2010) Network traffic characteristics of data centers in the wild. In: 10th ACM SIGCOMM Conference on Internet Measurement, pp 267–280
20. Alizadeh M, Greenberg A, Maltz DA, Padhye J, Patel P, Prabhakar B et al (2010) Data center tcp (dctcp). ACM SIGCOMM Comput Commun Rev 40:63–74
21. Kusic D, Kephart J, Hanson J, Kandasamy N, Jiang G (2009) Power and performance management of virtualized computing environments via lookahead control. Clust Comput 12:1–15
22. Cisco (2012) Cisco's massively scalable data center framework: network fabric for warehouse scale computer. http://www.cisco.com/c/dam/en/us/td/docs/solutions/Enterprise/Data_Center/MSDC/1-0/MSDC_Framework_1.pdf
23. Torres J, Carrera D, Beltran V, Poggi N, Hogan K, Berral JL et al (2008) Tailoring resources: the energy efficient consolidation strategy goes beyond virtualization. In: International Conference on Autonomic Computing (ICAC)
24. Modius (2016) Green data center initiatives. http://www.modius.com/green_data_center_solution
25. Kliazovich D, Bouvry P, Khan SU (2013) Simulation and performance analysis of data intensive and workload intensive cloud computing data centers. In: In optical interconnects for future data center networks, Springer, New York, pp 47–63
26. Guo Z, Yang Y (2015) On non blocking multicast fat tree data centre networks with server redundancy. IEEE Trans Comput 64:1058–1073
27. Cao J, Xia R, Yang P, Guo C, Lu G, Yuan L et al (2013) Per-packet load-balanced, low-latency routing for clos-based data center networks. In: Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies, ACM, California, pp 49–60
28. Greenberg A, Hamilton JR, Jain N, Kandula S, Kim C, Lahiri P et al (2011) VL2: a scalable and flexible data center network. Commun ACM 54:95–104
29. Al-Fares M, Radhakrishnan S, Raghavan B, Huang N, Vahdat A (2010) Hedera: dynamic flow scheduling for data center networks. NSDI 10:19–19
30. Carrega A, Singh S, Bruschi R, Bolla R (2012) Traffic merging for energy efficient datacenter network. In: Traffic merging for energy efficient datacenter network, Genoa, Italy
31. Chen G, He W, Liu J, Nath S, Rigas L, Xiao L, Zhao F (2008) Energy-aware server provisioning and load dispatching for connection-intensive internet services. In: 5th USENIX symposium on networked systems design and implementation, San Francisco, pp 337–350
32. Liu J, Zhao F, Liu X, He W (2009) Challenges towards elastic power management in internet data centers. In: Proceedings of the 2nd international workshop on cyber-physical systems (WCPS) in conjunction with ICDCS, Montreal, Quebec, Canada
33. NREL (2014) Trends in data centre design-ASHRAE leads the way to large energy savings. In: ASHRAE Conference, NREL/PR-6A40-58902, Denver, June 24, 2014
34. Ahmed A, Sabyasachi AS (2014) Cloud computing simulators: a detailed survey and future direction. In: IEEE International Advance Computing Conference (IACC), pp 866–872
35. Fan X, Weber WD, Barroso LA (2007) Power provisioning for a warehouse-sized computer. ACM SIGARCH Comput Archit News 25:13–23
36. Kephart N (2015, October 14) Visualizing traffic over ECMP and LAG paths. https://blog.thousandeyes.com/visualizing-traffic-over-ecmp-and-lag-paths/
37. Hopps C (2000) Analysis of an equal-cost multi-path algorithm. In: RFC 2992, Internet engineering task force, USA. The ACM Digital Library
38. Feng MZ, Hinton K, Ayre R, Tucker RS (2010) Reducing NGN energy consumption with IP/SDH/WDM, New York
39. Chernicoff D (2009) The shortcut guide to data center energy efficiency. Realtime Publisher, New York
40. Chen Y, Griffith R, Zats D, Joseph AD, Katz R (2012) Understanding TCP incast and its implications for big data workloads. USENIX;login: magazine, vol 37, pp 24–38

41. Solnushkin KS (2013, April 3) Fat-tree design: teach yourself fat-tree design in 60 minutes. http://clusterdesign.org/fat-trees/
42. Jain R (2013) Data centre network topologies. http://www.cse.wustl.edu/~jain/cse570-13/
43. Christensen K, Reviriego P, Nordman B, Bennett M, Mostowfi M, Maestro JA (2010) IEEE 802.3 az: the road to energy efficient ethernet. IEEE Commun Mag 48:50–56
44. Werner J, Geronimo G, Westphall C, Koch F, Freitas R (2011) Simulator improvements to validate the green cloud computing approach, vol 1–8, p 2011, Network operations and management symposium (LANOMS), 7th Latin American. IEEE
45. Chube P (2011, October) CRAC unit sizing: dos and don'ts. http://www.computerweekly.com/tip/CRAC-unit-sizing-Dos-and-donts
46. Cisco (2016, January 14) Cisco Nexus 7000 F-series modules. http://www.cisco.com/c/en/us/products/collateral/switches/nexus-7000-series-switches/data_sheet_c78-605622.html
47. itprice (2016) CISCO GPL 2016: the best CISCO global price list checking too. CISCO GPL
48. HPE. HP ProLiant DL385p Gen8 server—Specifications. http://h20564.www2.hpe.com/hpsc/doc/public/display?docId=emr_na-c03376977
49. PC-development. Computer systems. PC Development Ltd
50. GlobalCTI. Power over Ethernet (PoE) vs. power over Ethernet+(PoE+). GLOBALCTI
51. NCO. Cat6A 10G molded UTP Ethernet network patch cable. Network Cables Online
52. Cablematic. Spool UTP Cat6 cable 24AWG CCA solid green (100 m). Cablematic.com
53. Cisco (2016) Cisco Nexus 3064-X, 3064-T, and 3064-32T switches data sheet. Cisco
54. Pries R, Jarschel M, Schlosser D, Klopf M, Tran-Gia P (2011) Power consumption of data center architectures. In: International Conference on Green Communications and Networking, Springer, Berlin, pp 114–124
55. Lebiednik B, Mangal A, Tiwari N (2016) A survey and evaluation of data center network topologies. ArXiv preprint arXiv:1605.01701
56. Andreyev A (2014, November 14) Introducting data centric fabric, the next-generation Facebook data center network. http://tech.ccmbg.com/blog/introducing-data-center-fabric-the-next-generation-facebook-data-center-network-engineering-blog-facebook-code-facebook/
57. GoogleDC (2015) Efficiency: How we do it. https://www.google.co.uk/about/datacenters/efficiency/internal/
58. Chen T, Gao X, Chen G (2016) The features, hardware, and architectures of data center networks: a survey. J Parallel Distrib Comput 96:45–74
59. Singh A, Ong J, Agarwal A, Anderson G, Armistead A, Bannon R et al (2015) Juniper rising: a decade of clos topologies abd centralised control in google's datacenter network. ACM SIGCOMM Comput Commun Rev 45:183–197
60. Ray (2015, November 10) Facebook down to 1.08 PUE and counting for cold storage. http://silvertonconsulting.com/blog/2015/11/10/facebook-down-to-1-08-pue-and-counting-for-cold-storage/#sthash.W8S1gf6I.dpbs
61. Joshi Y, Kumar P (2012) Energy efficient thermal management of data centers. Springer, Atlanta
62. Farrington N, Andreyev A (2013) Facebook's data center network architecture. In: IEEE optical interconnects conference