8-12-2012

Deployment of a Hybrid Multicast Switch in Energy-Aware Data Center Network: A Case of Fat-Tree Topology

Tosmate Cheocherngngarn
Department of Electrical and Computer Engineering, Florida International University

Jean Andrian
Department of Electrical and Computer Engineering, Florida International University, Jean.Andrian@fiu.edu

Deng Pan
Department of Electrical and Computer Engineering, Florida International University, Deng.Pan@fiu.edu

Recommended Citation
Tosmate Cheocherngngarn, Jean Andrian, and Deng Pan, "Deployment of a Hybrid Multicast Switch in Energy-Aware Data Center Network: A Case of Fat-Tree Topology," ISRN Communications and Networking, vol. 2012, Article ID 209573, 10 pages, 2012. doi:10.5402/2012/209573

International Scholarly Research Network
ISRN Communications and Networking
Volume 2012, Article ID 209573, 10 pages
doi:10.5402/2012/209573

Research Article

Deployment of a Hybrid Multicast Switch in Energy-Aware Data Center Network: A Case of Fat-Tree Topology

Tosmate Cheocherngngarn, Jean Andrian, and Deng Pan

Department of Electrical and Computer Engineering, Florida International University, 10555 West Flagler Street, Miami, FL 33174, USA

Correspondence should be addressed to Tosmate Cheocherngngarn, tcheo001@fiu.edu

Received 18 June 2012; Accepted 12 August 2012

Academic Editors: G. Hasegawa and D. N. Serpanos

Recently, energy efficiency or green IT has become a hot issue for many IT infrastructures as they attempt to utilize energy-efficient strategies in their enterprise IT systems in order to minimize operational costs. Networking devices are shared resources connecting important IT infrastructures, especially in a data center network they are always operated 24/7 which consume a huge amount of energy, and it has been obviously shown that this energy consumption is largely independent of the traffic through the devices. As a result, power consumption in networking devices is becoming more and more a critical problem, which is of interest for both research community and general public. Multicast benefits group communications in saving link bandwidth and improving application throughput, both of which are important for green data center. In this paper, we study the deployment strategy of multicast switches in hybrid mode in energy-aware data center network: a case of famous fat-tree topology. The objective is to find the best location to deploy multicast switch not only to achieve optimal bandwidth utilization but also to minimize power consumption. We show that it is possible to easily achieve nearly 50% of energy consumption after applying our proposed algorithm.

1. Introduction

Data centers aim to provide reliable and scalable computing infrastructure for massive information and services. Accordingly, they consume huge amounts of energy and exponentially increase operational costs. According to recent literature, the annual electricity consumed by data centers in the United States is 61 billion kilowatt-hours (kWh) in 2006 (1.5 percent of total US electricity consumption) for a total electricity cost of about $4.5 billion. The energy use of the nation's servers and data centers in 2006 is estimated to be more than double the electricity that was consumed for this purpose in 2000 [1].

Energy efficiency has become nontrivial for all industries, including the information technology (IT) industry, since there is a big motivation to reduce capital and energy costs. According to Figure 1, the global information and communications technology (ICT) industry accounts for approximately 2 percent of global carbon dioxide (CO_2) emissions; the figure is equivalent to aviation in 2007. Most likely, ICT use grows faster than airline traffic in the past few years [2]. In addition, with energy management schemes, we turn to a part of the data center that consumes 10–20% of its total power: the network [3]. Thereby presenting a strong case for reducing the energy consumed by networking devices such as switches and routers, our goal is to outstandingly lower this growing recurring energy.

As a data center is to service over ten thousand servers, inflexible and insufficient bisection bandwidths have prompted researchers to explore alternatives to the traditional 2N tree topology (shown in Figure 2(a)) [4] with designs such as VL2 [5], PortLand [6], and BCube [7]. The resulting networks look more like a mesh than a tree. One such example, the famous fat-tree [4], seen in

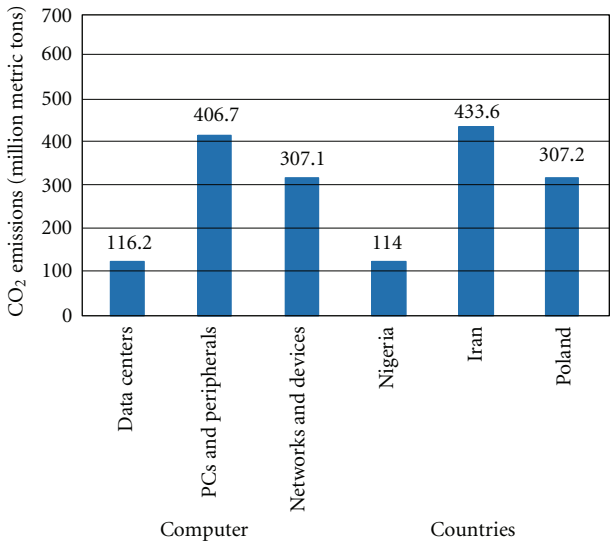


Figure 1: Carbon dioxide emissions from the energy consumed by data centers, PCs and peripherals, and networking devices are roughly equivalent to those of Nigeria, Iran, and Poland and account for 2 percent of the total world carbon footprint [25].

Figure 2(b), is built from a large number of richly connected switches/routers and can support any communication pattern (i.e., full bisection bandwidth). Traffic from clusters of servers is routed through hierarchical design of top-of-the-rack (ToR), aggregation, and core switches, respectively. The lowest layer is ToR or edge switches spreading traffic across the aggregation and core, using multipath routing, unequal load balancing, or a number of other techniques in order to deliver package to the destination server [8].

There are a number of multicast services in data center network. Servers in the data center use IP multicast to propagate information and communicate with clients or other application servers. For example, the financial services industry, particularly the market data infrastructure, depends comprehensively on IP multicast to deliver stock quotes [9]. Increased reliance on multicast in next-generation data center addresses the performance requirements for IP multicasting in the data center. Group communication widely exists in data centers hosting cloud computing [10, 11]. Multicast benefits group communications by both saving network traffic and improving application throughput. Even though multicast deployment in the Internet bears many hindrances during the past two decades for many issues such as compatibility, pricing model, and security concern, recently there is a perceptible rebirth of it, for example, the successful application of streaming videos [12], satellite radio, and so forth. The managed environment of data centers also provides a good opportunity for multicast deployment because of a single authority which is considered trustworthy.

Hybrid multicast approach is attractive to IT infrastructure for the following reasons. First, the improved bandwidth efficiency provides the incentives for network administrator to adopt the new technique as they can consolidate traffic from multiple switches onto a single switch. Secondly, in particular, wireless bandwidth is precious and mobile devices are power constrained. It makes mobile users happy for wireless hosts to move multicast packet duplication from end hosts to switches. Next, the hybrid approach allows incremental deployment of multicast switches. The hybrid approach only utilizes the packet duplication capability of multicast switches when available but does not require all switches to be multicast capable. Therefore, the network administrator can start deployment at selected areas with heavy multicast traffic as the first step. Lastly, multicast switches in the hybrid approach are transparent to end hosts. The switches can be implemented to automatically recognize and participate in P2P multicast networks, and thus no change is necessary at the end hosts. Nevertheless, it is still feasible for applications to actively detect the existence of multicast switches and utilize them as much as possible.

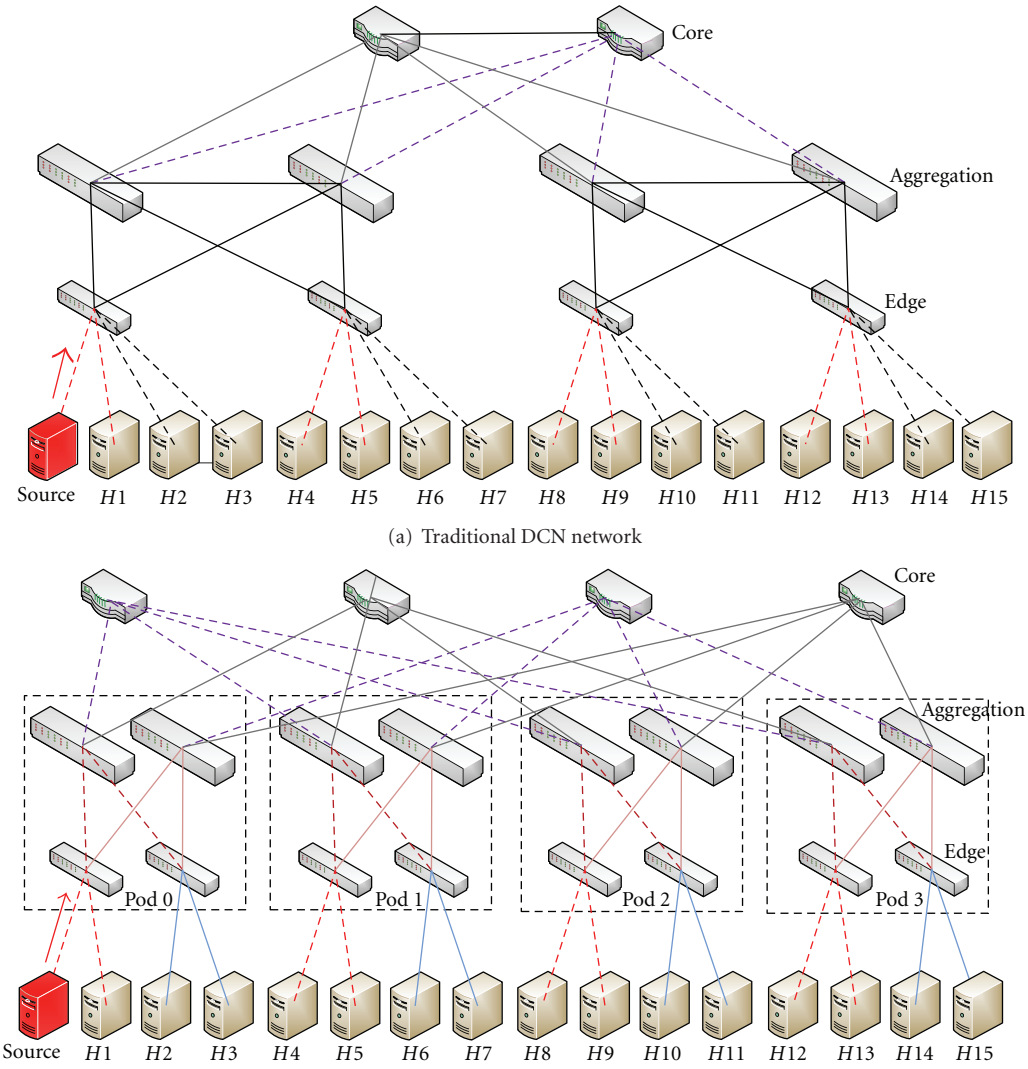
In this paper, we study the deployment strategy of multicast switches in a network to enable switch an IP multicast function. As discussed above, incremental deployment is possible and we assume that the IT infrastructure plans to deploy a fixed number of multicast switches in data center network. In addition, we assume that all servers in this data center are running many multicast traffic, such as multicast groups, broadcasting protocols to members in each individual group. Plus traffic intensity may be obtained by either measurement or estimation. The objective is therefore to find deployment locations and corresponding routing paths so as to achieve optimal bandwidth utilization and minimize power consumption.

We first formulate the selective deployment and path searching problems as linear programs. Although the linear programs obtain optimal solutions, integer linear programming is NP-complete and is not practical for large scale networks. Therefore, we propose fast polynomial algorithms to obtain quick solutions. Finally, we conduct simulations based on open-source simulator: Liu [13], and the results fully demonstrate the effectiveness of our designs.

This paper is organized as follows. In Section 2, background and related works are briefly described. In Section 3, power modeling for evaluating energy consumption in data center network is proposed apprehensively. In Section 4, we formulate the problems and present fast polynomial solutions. Simulation result of our design and discussion are offered in Section 5. Concluding remarks are presented in Section 6.

2. Background and Related Works

2.1. Data Center Multicast. Group communication is common in modern data centers running many traffic-intensity applications. Multicast is the technology to support this kind of one-to-many communication pattern, for both saving network bandwidth and decreasing sender's load. For Web search services, the incoming user query is directed to a set of indexing servers to look up the matching documents [14]. Multicast can help accelerate the directing process and reduce the response time. Moreover, distributed file system



(a) Traditional DCN network

(b) Fat-tree (with $k = 4$ pods)

Figure 2: Network topologies with a source and 15 destination receivers.

is widely used in data centers, such as GFS [15] in Google and COSMOS in Microsoft. Files are divided into many fixed size chunks, either 64 or 100 MB. Each chunk is replicated to several copies and stored in servers located in different racks to improve the reliability. Chunk replication is usually bandwidth-hungry, and multicast-based replication can save the interrack bandwidth. Multicast can also speed up the binary delivery and reduce the finishing time of any process.

Although multicast protocol is supported by most vendors' routers/switches and end hosts, it is not widely deployed in the Internet due to many technological causes, such as compatibility, pricing model, and security concern. However, we disagree that in the managed environment of data centers, multicast is a comprehensive option to support one-to-many communication in data center network. For instance, the natural pricing problem is not an issue in data centers as they are usually managed by a single authority which is considered very trusty.

Li et al. [16] are using their ESM (Efficient and Scalable Data Center Multicast Routing) technique to accommodate that challenge above. ESM, a novel multicast routing scheme in data center networks, leverages the managed environment of data centers and the topological characteristics of modern data center networks, as well as the multicast group size distribution pattern. This kind of centralized controller is widely adopted in modern data center design. For instance, in fat-tree [4], a fabric manager is responsible for managing the network fabric. In VL2 [5], a number of directory servers are used to map the AA-LA relationship. The emerging OpenFlow [17] framework also uses a controller for routing rule decision and distribution.

In this paper, we assume that ESM technique can be practically implemented in our green data center as it addresses the challenges above by exploiting the features of modern data center networks in most recent literature. It is not only flexible and scalable multicast protocol but also able

to deploy in those state-of-the-art data centers networks as proved in their breakthrough result.

2.2. Energy-Aware Data Center Network. Gupta and Singh [18] were amongst the earliest researchers to advocate conserving energy in networks. Other researchers have proposed techniques such as putting idle subcomponents (line cards, etc.) to sleep [18–20], as well as adapting the rate at which switches forward packets depending on the traffic [18, 20]. Nedevschi et al. [21] discuss the benefits and deployment models of a network proxy that would allow end hosts to sleep while the proxy keeps the network connection alive. He also proposes shaping the traffic into small bursts at edge routers to facilitate sleeping and rate adaptation. Further their work addresses edge routers in the Internet [19]. Mahadevan et al. [22] show that one of their power saving algorithms focuses on job allocation; they perform this operation from the point of view of saving power at network devices and show considerable energy savings can be achieved. Chiefly, their algorithms are for data centers and enterprise networks.

Our finding confirms that the deployment of multicast switch in energy-aware data center network including recently notable techniques, shutdown the unused links and sleep power-hungry switches/routers, can dramatically lower the total power consumption of data center. The graph of energy consumption shows 50% decrease comparing to that without power awareness.

2.3. Data Center Traffic Patterns. Figure 3 displays the plot of 7-day network traffic from the SuperJANET4 access router of service provider at Manchester recorded with MRTG [23]. The normal traffic levels for the Net North West MAN vary between 70 and 300 Mbps into the MAN (solid graph) and between 200 and 400 Mbps out of the MAN (line graph). There is a burst as visible as the sharp spikes, which occur once in a while. We can clearly see a wave pattern, with the highest instant traffic volume at about 750 Mbps and the lowest at about 50 Mbps. It is obviously seen that at night time, traffic has dropped lower than 50% of the peak regardless of incoming or outgoing direction. The key for our energy-aware DCNs to achieve power conservation during off-peak hours is to power off idle devices and shutdown unused links when possible.

Another example is in Figure 4. It might not have been included in Facebook's music launch, but Internet radio service Pandora has been adding more and more daily active users on Facebook [24]. At the end of the last year, it was near 1.4 million at the peak of the traffic wave you see above, plummeting over 30% every weekend. This famous radio streaming application is heavily based on broadcasting communication which is clearly seen that our algorithm can save vast energy on this growing application.

2.4. Data Center Topology. Recently, there is a growing interest in the community to design new data center network architectures with high bisection bandwidth to replace those old-fashioned trees [4–7]. Fat tree is the representative one

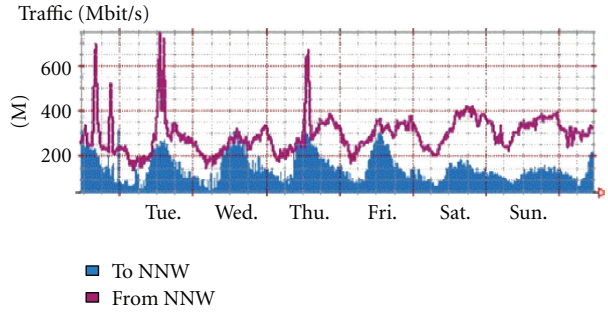


Figure 3: Weekly DCN traffic fluctuation [23].

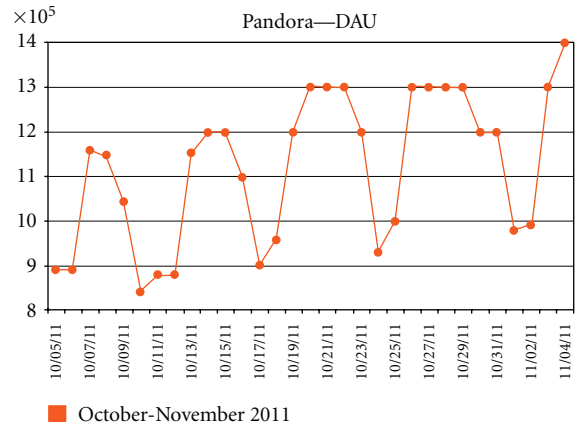


Figure 4: Fluctuating traffic pattern of Pandora satellite radio [24].

among these current three-tier architectures. Figure 2(a) illustrates the topology of fat tree, which organizes the switches in three levels. More specifically, if k is the number of ports on each single switch, then there are k pods, with each pod consisting of $k/2$ edge switches and $k/2$ aggregation switches. Each k-port switch at the edge level uses $k/2$ ports to connect the $k/2$ servers and uses the remaining $k/2$ ports to connect the $k/2$ aggregation-level switches in the same pod. At the core level, there are $(k/2)^2$ switches, and each k-port switch has one port connecting to each pod. Thus in total, there are $5k^2/4$ switches that interconnect $k^3/4$ servers. Figure 2(a) shows one such network for $k = 4$ fat tree topology.

To the best of our knowledge, we persist in ESM technique that can be practically implemented in our green data center as it can arrange those challenges above by taking the advantage of those most recent data center topologies. More importantly, combining with our hybrid multicast, ESM is proved to be operated effectively on hierarchical topology that is, fat tree, VL2, BCube, and so forth, which extensively matches our proposed framework.

3. Power Modeling

Energy consumption can be generally defined as

$$\text{Energy} = \text{AvgPower} \times \text{Time}, \tag{1}$$

where Energy and AvgPower are measured in Joule and Watt, respectively, and 1 Joule = 1 Watt × 1 Second. Energy efficiency is equivalent to the ratio of performance, measured as the rate of work done to the power used [26], and the performance can be represented by response time or throughput of the computing system:

$$\text{Energy Efficiency} = \left(\frac{\text{Work done}}{\text{Energy}}\right) = \left(\frac{\text{Performance}}{\text{Power}}\right). \tag{2}$$

To the best of our knowledge, there is no in-depth measurement study exists that quantifies the actual energy consumed by a wide range of switches under widely varying traffic conditions. However, [22] analyzed power measurements obtained from a variety of switches from well-known vendors such as Cisco, ProCurve, and Brocade. They identify various control knobs as a function of switch configurations and traffic flowing through the switch. Based on their analysis, they developed a power model to estimate the power consumed by any switch, $\text{Power}_{\text{switch}}$. Linear power model is to estimate the power consumed by any switch defined as

$$\begin{aligned}\text{Power}_{\text{switch}} &= \text{Power}_{\text{chassis}} + \text{num}_{\text{line card}} * \text{Power}_{\text{line card}} \\ &\quad + \sum_{i=0}^{\text{configs}} \text{numports}_{\text{configsi}} * \text{Power}_{\text{configsi}} \\ &\quad * \text{utilization Factor}.\end{aligned} \tag{3}$$

Table 1 lists the device categories. All power measurements including PoE already are reported in Watts (except the last column is in Mbps). 9-slot core switch is typically used as a root switch in data centers. It consumed maximum 3000 Watts when fully operated during peak hours but 555 Watts when idle. Aggregation switch is available as a modular chassis with 6 slots, with each slot capable of supporting a 24-port line card. Alternatively, 24-port 1 Gbps line card for an aggregate 24 Gbps capacity is able to be replaced by a 4-port 10 Gbps line card for an aggregate of 40 Gbps capacity operated during peak hours. Each line card consumes 35–40 Watts. For an edge switch having a line card with 48 full-duplex 1 Gbps ports, one way to fully load the switch is to attach servers to each port and ensure 1 Gbps of traffic going in and coming out of each port. Note that as the number of active ports is increased, the impact of port utilization (whether no load or fully loaded) on power consumption is under 5%.

In this paper, we follow their finding as the result is very well explained and they proved that their estimated power consumption matches the real power measured by the power meter with an error margin of under 2%. Plus, IP options set in the packet might not affect power consumption at switches performing MAC forwarding, processing packets that have IP options might impact the power consumption of a gateway router which comprehensively relate to our proposed IP multicast forwarding in those multicast switches. Moving onto the effects of traffic, packet size does not impact power consumption at all.

Also, we compute the node power saving as

$$\text{Node}_{\text{saving}}(t) = \left(\frac{\sum_i p^i_{\text{on}}(t)}{\sum_i p^i(t)}\right). \tag{4}$$

We consider a sinusoidal function reported on the node power saving as stated in [27] where the numerator is the power consumed by nodes for the energy-aware network and the denominator is the power consumed by nodes for a nongreen network. Note that $\text{Node}_{\text{saving}}(t)$ is measured during night, since the connectivity is the tightest constraint, with the offered traffic much smaller than that during peak hour. On the other hand, during the day the node power saving decreases because the traffic is very critically intense so that some unused switches are needed to be on due to path redundancy. As traffic significantly increases in peak hours, more network and link capacity are required in order to guarantee the maximum link utilization constraint. However under that scenario it would be possible to always turn off few nodes, so that a small power saving is still possible.

We run each experiment for 120 seconds thrice and report the average power over the entire duration. A similar reasoning can be applied to $\text{Link}_{\text{saving}}(t)$, which we certainly plan to do for a future work.

4. Deployment of a Multicast Switch

4.1. Problem Formulation. A network is a directed graph $G = (H \cup X, E)$, where H is the set of end hosts, X is the set of switches, and E is the set of links between hosts and/or switches. Each link $(u, v) \in E$ has a nonnegative weight $W(u, v) \geq 0$, which may be the length or average latency. A multicast group consists of a source host $s \in H$ and a set of destination hosts $D = \{d_1, \ldots, d_n\}$, for all i, $d_i \in H$. For simplicity, we assume that a host has no switching function. In the case of switching host, it can be easily represented as a nonswitching host plus a switch.

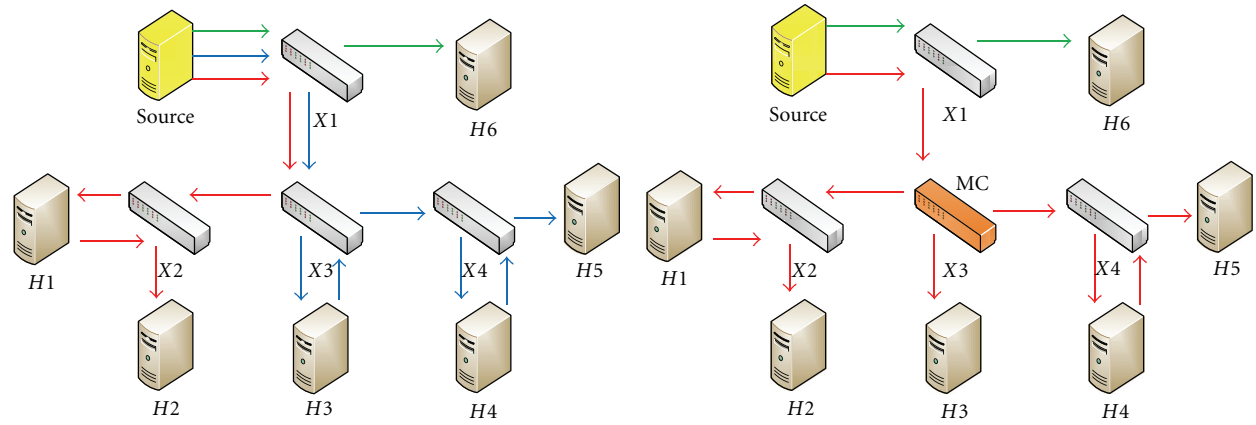
In the *P2P mode*, the switches do not perform packet duplication, and the hosts transmit the packet by unicast paths. In detail, after a destination host receives a specific packet, it forwards a copy to the next destination, as shown in Figure 5(a). Since the switches do not conduct packet duplication, the same packet may be transmitted over a link multiple times. For a link $(u, v) \in E$, define $n(u, v)$ to be the number of transmissions of the packet from u to v. Note that $n(u, v)$ may not be equal to $n(v, u)$. Define the cost of the transmission path of a packet to be the sum of the product of the weight of each link and the number of transmissions over the link, that is, $\sum_{(u,v)\in E} n(u, v)W(u, v)$. Although different packets may take different paths, we are interested in finding the optimal path with the minimum cost, which can be formulated as the following linear program:

$$\text{Minimize} \sum_{(u,v)\in E} n(u, v)W(u, v) \tag{5}$$

subject to the following constraints.

Table 1: Power consumption summary for network devices in three hierarchical layers [28].

Type	Plate power (W)	Number of ports/line card	Idle power (W)	BW (Mbps)
Core switch	3000	24	555	48000
Aggregation switch	875	24	133.5	48000
Edge switch	300	48	76.4	48000



(a) P2P multicast (b) Hybrid multicast deployment

Figure 5: Examples of P2P and hybrid multicast deployment.

Source Departure. There is at least one copy of the packet departing from the source; that is,

$$\sum_{u\in(H\cup X)} n(s,v) \geq 1. \tag{6}$$

Destination Arrival. At least one copy of the packet arrives at each destination; that is,

$$\forall i, \quad \sum_{u\in(H\cup X)} n(u,d_i) \geq 1. \tag{7}$$

Source-Destination Connectivity. Each destination must be connected with the source to avoid subtours [29]; that is,

$$\forall T, \quad d_i \in T \subseteq H \cup U \setminus \{s\}, \quad \sum_{u\in(H\cup U-T),\, v\in T} n(u,v) > 0. \tag{8}$$

Switch Conservation. A switch only transmits packets, without creating or destroying any; that is,

$$\forall u \in X, \quad \sum_{v\in(H\cup X)} n(v,u) = \sum_{v\in(H\cup X)} n(u,v). \tag{9}$$

In the *hybrid mode*, a fixed number M of switches can be upgraded with multicast support. The multicast switches can participate in the P2P multicast group and assist packet duplication when possible, as shown in Figure 5(b). If $u \in X$ is upgraded as a multicast switch, define $m(u) = 1$; otherwise, $m(u) = 0$. For $u \in X$, use Size(u) to represent the size of u, that is, the number of output ports. Note that Size(u) is not a variable but a constant for a given switch u. Our objective is still to minimize the overall cost of the transmission path of a packet by strategically deploying the multicast switches. The problem can be formulated as a linear program with the same objective function but replacing the switch conservation constraint by the following two.

Multicast Support. For a multicast switch, the difference between the number of its outgoing packet copies and that of incoming copies is less than or equal to its size minus one. In other words, after the switch receives the packet from one input, it can send a copy to each output; that is,

$$\forall u \in X, \quad \sum_{v\in H\cup X} n(u,v) - \sum_{v\in H\cup X} n(u,v) \leq m(u)(\text{Size}(u) - 1). \tag{10}$$

Fixed Number of Multicast Switches. The total number of multicast switches in the network is at most M; that is,

$$\sum_{u\in X} m(u) \leq M. \tag{11}$$

Although the above linear programs give the optimal solutions, they are NP-complete and therefore are not practical to solve the problems for large scale networks. In the following, we provide polynomial algorithms that can obtain quick solutions.

4.2. P2P Path Searching. As the basis to calculate the multicast switch deployment, we first present the algorithm to find the P2P transmission paths. The basic idea is to

separate the source and destinations into two sets. Nodes in the first set have all received a copy of the packet, and nodes in the second set have not. The algorithm then finds the minimum cost path from the first set to the second set, by which the packet reaches one more destination. The algorithm works in iterations and adds a destination host to the first set in each iteration. Use S to represent the first set and initialize it as $S = \{s\}$, and use T to represent the second set and initialize it as $T = D$. The minimum cost path from S to T can be easily found, because whenever a new host is added to S, its minimum cost path to each of the remaining hosts in T is calculated using the Dijkstra algorithm.

In summary, each iteration of the algorithm includes the following steps.

(1) Find the minimum cost path from a host $u \in S$ to a host $v \in T$. If there are multiple paths with the same minimum cost, select the one with the smallest index source (assuming each host having an ID for comparison). The reason is to consolidate traffic in certain switches so that upgrading those switches will maximize bandwidth efficiency and power off unused switches.

(2) Remove v from T and add it to S; that is, $T = T \setminus \{v\}$ and $S = S \cup \{v\}$. Calculate the minimum cost path from v to each remaining host in T.

It can be shown that the above algorithm obtains the optimal solution. Due to space limitations, the detailed proof is omitted. Since the algorithm needs $|D|$ iterations and the time complexity to calculate the shortest distance paths for the newly added host in each iteration is $O(|H \cup X|^2)$, the time complexity of the algorithm is $O(|S||H \cup X|^2)$.

4.3. Deployment of Multicast Switches with Single Multicast Group. Next we consider the multicast switch deployment problem and start with the simpler case with a single multicast group.

The main idea is to calculate the cost reduction to upgrade each switch in the P2P paths obtained above and select the one with the maximum cost reduction. Repeat the process multiple times until we have found the deployment locations of all the M multicast switches.

It can be noticed that not all switches will result in cost reduction if upgraded. We define a relaying switch to be one with an in-degree greater than one in the current transmission paths. Specifically, $u \in X$ is a relaying switch if $\sum_{v \in (H \cup X)} n(v, u) > 1$. Upgrading a relaying switch will obtain cost reduction, because packet duplication at the switch will avoid the additional incoming transmissions. In Figure 5(a), $X2$ and $X4$ are relaying switches, each with an in-degree of 2; $X1$ and $X3$ are also relaying switches each has an in-degree of 3.

After identifying the relaying switches, we need to calculate the cost reduction to upgrade such a switch, which is the total weight of the edges for the switch to receive relaying copies of the packet. In case that the relaying switch has both incoming edges from multiple neighbors, we need to determine which are the relaying edges. This can be done by a breath first search with the current path from the multicast *Source s*, and the edge from a farther node to a closer node is the relaying edge. For example, in Figure 5(a), both $(X3, X2)$ and $(H1, X2)$ are incoming edges of $X2$, and only the latter is a relaying edge. Both $(X3, X4)$ and $(H4, X4)$ are incoming edges of $X4$, and only the latter is a relaying edge. On the other hand, if a relaying switch has $n > 1$ incoming edges from the same neighbor, $n - 1$ of them are relaying edges. In Figure 5(a), $(\text{Source}, X1)$ is a relaying edge for $X1$. In case that the other node of a relaying edge is also a switch, it must be a relaying switch as well, and we need to trace back recursively until reach a host. In Figure 5(a), not only $(\text{Source}, X1)$ and $(X1, X3)$ form the relaying path for $X3$, but also $(H3, X3)$ is a relaying edge of $X3$. Calculate the total weight of all the relaying edges to obtain the cost reduction for a relaying switch, and then select the one with the maximum reduction.

To sum up, each iteration of the algorithm includes the following steps.

(1) Identify relaying switches, and calculate the cost reduction for each of them.

(2) Select the switch with the maximum cost reduction. Remove all the relaying paths and perform packet duplication at the switch instead. Stop if there are already M multicast switches.

(3) Update the cost reduction of the remaining switches after upgrading the switch selected in the above step.

In Figure 5(a), if we assume that each link has the same weight of one and $M = 1$, $X3$ has the maximum cost reduction of 3, we upgrade it to a multicast switch, and the resulting hybrid transmission network is shown in Figure 5(b). Neither $X1$, nor $X2$, $X4$ is picked as each of them has cost reduction of 2, 1, and 1, respectively.

The algorithm needs M iterations. Since there are at most $|D| - 1$ relaying paths, each with length less than $|H \cup X|$, the time complexity in each iteration is $O(|D||H \cup X|)$. Therefore, the time complexity of the algorithm is $O(M|D||H \cup X|)$.

We run each experiment for 120 seconds in order to find out the average delay three times in each scenario. A similar algorithm can be applied to multiple multicast groups, but it is not reported in this paper due to space constraints.

5. Results and Discussion

In this section, we show the simulation results to demonstrate the effectiveness of our design.

5.1. Network Delay. We set up a UDP application package with one host being the multicast source and all the remaining hosts being destinations. During the day, the traffic is very intense. *Source* host generates traffic at the rate of 100 to 1000 packets per second, and the packet size is fixed at maximum 1200 Bytes. When there is no multicast switch in the network, the packet is transmitted in the pure P2P mode and all links have identical bandwidth of 1 Gbps. However, when there is a multicast switch (M is set to 1),

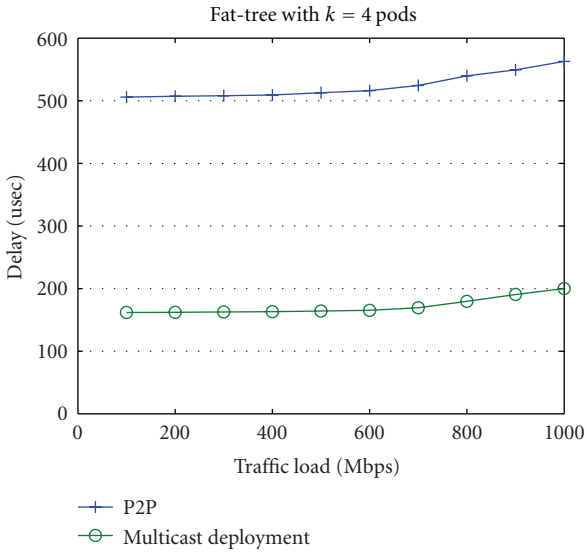


Figure 6: Average delay in green data center network.

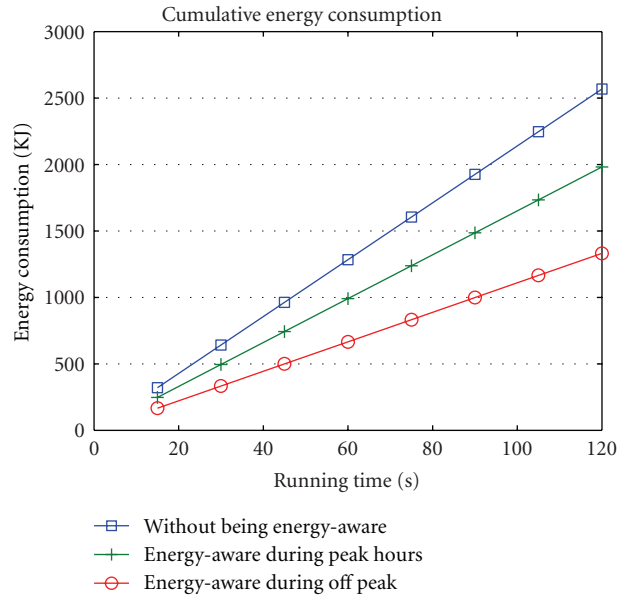


Figure 7: Energy consumption variance.

it will duplicate that packet and broadcast to remaining receivers when possible and all links between core and aggregation layer are increased to 2 Gbps in order to carry growing multicast traffic. Each simulation lasts 120 seconds.

Figure 6 shows the simulation results with the fat-tree topology which is widely adopted by modern data center network based on open-source simulator—Liu [13]. We set up a single multicast group with the *source* host being the source and the remaining hosts being the destinations as shown in Figure 2(b). We plot the average multicast delay, calculated as the average interval of all packets from the departure at the *source* host to the arrival at each destination, under two scenarios: without multicast switches and with calculated deployment. The calculated deployment curve shows the data when the left-most core switch exposed in Figure 2(b) is upgraded, which is obtained by our algorithm described in Section 4. Although those two curves grow as the packet generation rate increases, the average multicast delay with the calculated multicast switch deployment is only about one-sixth of that without multicast switches. We can see that our algorithm consistently obtains shorter average multicast delay than the P2P approach. The results fully demonstrate that our algorithm is effective in calculating good deployment locations for multicast switches to reduce the traffic amount and latency.

5.2. Energy Consumption and Power Saving. We set up a UDP application package with one host being the multicast source and all the remaining hosts being destinations. Assume that all switches are able to be configured IP multicast mode. Based on our algorithm in Section 4, a left-most core switch in Figure 2(b) is upgraded to be a multicast switch. It will duplicate that packet and broadcast to remaining receivers when possible regardless of traffic rate. During the day, the traffic is very intense. *source* host generates traffic at the rate of 100 to 1000 packets per second, and the packet size is fixed at maximum 1200 Bytes. All links have identical bandwidth of 1 Gbps, but links between core and aggregation layer are increased to 2 Gbps in order to deliver growing multicast traffic. However, as said in Section 3, the traffic during night is only 50% of the peak-hour demand. We reduce the traffic generation of *source* host at the rate of 100 to 500 packets per second according to common traffic pattern in Figures 3 and 4 having all links with identical bandwidth of 1 Gbps. Each simulation is run for 120 seconds.

Regarding power consumption summary in Table 1 and (1), after applying our proposed algorithm to enable IP multicast function in a core switch, we calculate the power consumption as exhibited in Figure 7. Energy used is reduced by half during off-peak hours. Plus, according to the data center traffic pattern in Figures 3 and 4, we can extensively deploy this scheme on weekends so that roughly 50% of the fully operated power consumption is saved. We can clearly say that the optimal energy-aware policy is also able to run during peak hours, but because of redundancy and guaranteed link utilization, we need to keep few unused switches on. Thus, energy saving is one-fourth of the maximum correspondingly. Network administrator in an enterprise or data center networks should be able to consolidate traffic from multiple switches onto a single switch so as to turn off the unused switches. As can be seen, all three curves grow as time goes by. However, energy consumption without energy-aware scheme grows much faster than the other two which demonstrates that our proposed hybrid multicast mode is effectively decreasing energy consumption.

Figure 8 reports the breakdown of the percentage of power saving after sleeping unused nodes detailing core and aggregation during both off-peak and peak hours. According to (4), where the numerator is the power consumed by nodes for the energy-aware network and the denominator is the power consumed by nodes for a non-green network,

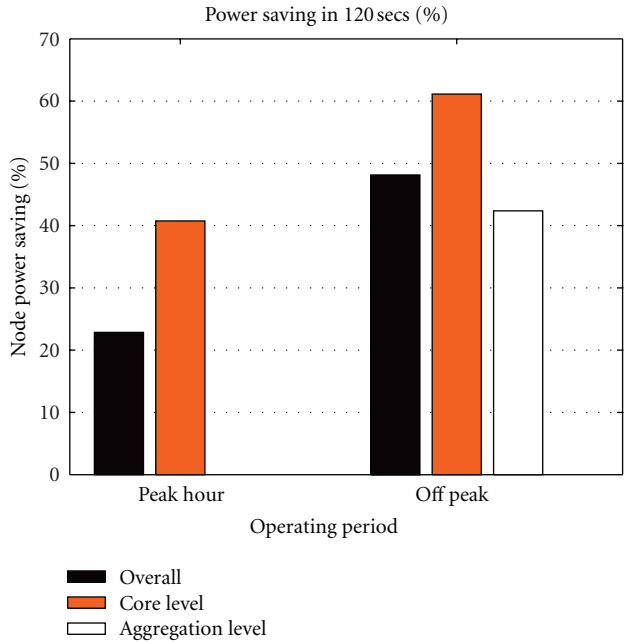


Figure 8: Node power saving comparison between peak hour and off peak.

values have been averaged over the three different runs. The plot shows that during off-peak hours, it is possible to save power of approximately 50% of nodes that are not source/destination of traffic, with the core and aggregation nodes being the largest fraction of them. This reflects the fact that the network has been designed to recover from possible faults, which requires additional resources. These additional resources are not exploited to carry traffic during off-peak time, and then they can be powered down to save energy. During peak hours, on the contrary, the saving is much lower, as only about 20% of power is not uselessly wasted, being the majority of core nodes. Note that during the day, aggregation nodes are always operatively on. These additional nodes may be required to recover from occasional faults and unexpected incidents. This obviously demonstrates that our proposed algorithm yields significant network energy saving.

6. Conclusions

Energy efficiency has become a top priority in most IT enterprise. Networking devices in data center network consist an important part of the IT infrastructure and consume massive amounts of energy. Relatively small attention has been paid to gear up the energy efficiency of data center networks thus far. In this paper, we make several contributions as follows. Firstly, we proposed the deployment of a multicast switch in a hybrid multicast network, which combines the efficiency of IP multicast and the flexibility of P2P multicast. We first formulate the problem as integer linear programming which is NP-complete and not practical for large scale networks. We further propose fast polynomial algorithms that obtain quick solutions. Accordingly, we conduct extensive simulations to evaluate the transmission cost reduction and packet delay improvement, and the simulation results fully demonstrate the effectiveness of our design; that is, the average delay of our calculated multicast switch deployment is only one-sixth of that without multicast switches. Next, we calculate power consumption after deploying a multicast switch in famous fat-tree topology. Energy used is reduced by half during off-peak hours. Besides, we can extensively deploy this scheme on weekends so that roughly 50% of the fully operated power consumption is saved. During peak hours, although we need to keep few unused switches on, energy saving is one-fourth of the maximum correspondingly. Finally, $\text{Node}_{\text{saving}}(t)$ is measured during day and night. Since the connectivity is the tightest constraint at night, the offered traffic being much smaller than during peak hour. Saving well approximately 50% of power is achievable. In contrast, during the day, it would be possible to turn off few nodes, so that a minimum 20% of power saving is promising which demonstrates that our proposed hybrid multicast mode is successful in comprehensively decreasing energy consumption.

Acknowledgment

The work described in this paper was supported by a DEA fellowship from the University Graduate School, Florida International University.

References

[1] U.S. Environmental Protection Agency, "Report to Congress on Server and Data Center Energy Efficiency," 2007, http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf.

[2] A. C. Orgerie, *Energy-Efficiency in Wired Communication Networks*, COST Training School, Brasov, Romania, 2011.

[3] A. Greenberg, J. Hamilton, D. Maltz et al., *The Cost of a Cloud: Research Problems in Data Center Networks*, ACM, SIGCOMM, CCR, 2009.

[4] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proceedings of the ACM SIGCOMM Conference on Data Communication (SIGCOMM '08)*, pp. 63–74, August 2008.

[5] A. Greenberg, N. Jain, S. Kandula et al., "VL2: a scalable and flexible data center network," in *Proceedings of the ACM Computer Communication Review (SIGCOMM '09)*, pp. 51–62, August 2009.

[6] C. Guo, G. Lu, D. Li et al., "BCube: a high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM Conference on Data Communication (SIGCOMM '09)*, pp. 63–74, August 2009.

[7] R. N. Mysore, A. Pamboris, N. Farrington et al., "PortLand: a scalable fault-tolerant layer 2 data center network fabric," in *Proceedings of the ACM Conference on Data Communication (SIGCOMM '09)*, pp. 39–50, August 2009.

[8] B. Heller, S. Seetharaman, P. Mahadevan et al., "ElasticTree: saving energy in data center network," in *Proceedings of the 7th USENIX Conference on Networked Systems Design and Implementation (NSDI '10)*, p. 17, Berkeley, Calif, USA, 2010.

[9] Cisco Systems, *Multicast with Cisco Nexus 7000*, Cisco Systems, 2009.

[10] Y. Vigfusson, H. Abu-Libdeh, M. Balakrishnan et al., "Dr. multicast: Rx for data center communication scalability," in *Proceedings of the 5th ACM EuroSys Conference on Computer Systems (EuroSys '10)*, pp. 349–362, April 2010.

[11] D. Li, H. Cui, Y. Hu et al., "Scalable data center multicast using multi-class bloom filter," in *Proceedings of the 19th IEEE International Network Protocols (ICNP '11)*, October 2011.

[12] A. Mahimkar, Z. Ge, A. Shaikh et al., "Towards automated performance diagnosis in a large IPTV network," in *Proceedings of the ACM Conference on Data Communication (SIGCOMM '09)*, pp. 231–242, August 2009.

[13] J. Liu, "Parallel Real-time Immersive network Modeling Environment (PRIME)," Modeling and Networking Systems Research Group, Florida International University, https://www.primessf.net/bin/view/Public/PRIMEProject.

[14] T. Hoff, "Google architecture," 2008, http://highscalability.com/google-architecture .

[15] S. Ghemawat, H. Gobio, and S. Leungm, "The Google file system," in *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP '03)*, pp. 29–43, 2003.

[16] D. Li, Y. Li, J. Wu et al., "ESM: efficient and scalable data center multicast routing," *IEEE Transaction on Networking*, vol. 20, no. 3, pp. 944–955, 2011.

[17] Stanford University, "OpenFlow," Stanford, CA, 2008, http://www.openflowswitch.org/.

[18] M. Gupta and S. Singh, "Greening of the internet," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '03)*, pp. 19–26, August 2003.

[19] S. Nedevschi, L. Popa, G. Iannaccone et al., "Reducing network energy consumption via rate-adaptation and sleeping," in *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI '08)*, pp. 323–336, April 2008.

[20] M. Gupta and S. Singh, "Using low-power modes for energy conservation in Ethernet LANs," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM '07)*, pp. 2451–2455, Anchorage, Alaska, USA, May 2007.

[21] S. Nedevschi, J. Chandrashenkar, B. Nordman et al., "Skilled in the art of being idle: reducing energy waste in networked systems," in *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation (NSDI '09)*, pp. 381–394, April 2009.

[22] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "Energy aware network operations," in *Proceedings of the IEEE Workshops (INFOCOM '09)*, Rio de Janeiro, Brazil, April 2009.

[23] R. Hughes-Jones, S. Parsley, and R. Spencer, "High data rate transmission in high resolution radio astronomy—VlbiGRID," *Future Generation Computer Systems*, vol. 19, no. 6, pp. 883–896, 2003.

[24] E. Eldon, "Instant message, music services are now among the most engaging Facebook apps," 2011, http://www.insidefacebook.com/author/eldon/.

[25] J. Mankoff, R. Kravets, and E. Blevis, "Some computer science issues in creating a sustainable world," *IEEE Computer Society*, vol. 41, no. 8, pp. 102–105, 2008.

[26] D. Tsirogiannis, S. Harizopoulos, and M. A. Shah, "Analyzing the energy efficiency of a database server," in *Proceedings of the International Conference on Management of Data (SIGMOD '10)*, pp. 231–242, June 2010.

[27] L. Chiaraviglio, M. Mellia, and F. Neri, "Energy-aware backbone networks: a case study," in *Proceedings of the IEEE International Conference on Communications Workshops (ICC '09)*, Dresden, Germany, June 2009.

[28] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A power benchmarking framework for network devices," in *Proceedings of the 8th International IFIP-TC 6 Networking Conference (NETWORKING '09)*, pp. 795–808, May 2009.

[29] M. Hahsler and K. Hornik, "TSP—infrastructure for the traveling salesperson problem," *Journal of Statistical Software*, vol. 23, no. 2, pp. 1–21, 2007.