Efficient Multicast Delivery for Wireless Data Center Networks

Ya-Ju Yu[1], Ching-Chih Chuang[2], Hsin-Peng Lin[2,3], and Ai-Chun Pang[1,2,4]
[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[3]Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., Taipei, Taiwan
[4]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan
E-mail: yuyaju@citi.sinica.edu.tw, d99922041@csie.ntu.edu.tw, hplin@cht.com.tw, and acpang@csie.ntu.edu.tw

***Abstract*—Recently, large-scale data centers are widely built to support various kinds of cloud services, which are mostly delivered by multicast. Even with multicast, cloud services may still generate a large amount of data traffic in some bottleneck links and, even worse, cause network congestion. Thus, how to reduce the redundancy of data transmissions to mitigate network congestion is essential. In addition to wired transmissions, modern data centers adopt wireless links to augment network capacity. Under the coexisting scenario of wired and wireless links, this paper studies multicast data delivery problem. Specifically, a multicast tree problem is defined, and the objective is to minimize the total multicast data traffic. We prove the problem is $\mathcal{NP}$-hard and propose an efficient heuristic algorithm to solve the problem. A series of experiment results shows that our proposed algorithm is very effective, compared with an optimal solution designed for traditional wired data centers.**

***Index Terms*—Data redundancy, multicast, wireless data centers**

I. Introduction

Data center networks are designed as controlled environments for housing critical computing resources. With significant economic benefits of cloud computing, large-scale data centers are widely built for supporting diverse cloud services. The cloud services are mostly provided by the coordinating applications such as web query, distributed file system [1], structured storage system [2], and distributed execution engine (e.g., MapReduce) [3]. For those coordinating applications, group communications accomplished by multicast are commonly used to deliver data traffic in data center networks. For example, a web server redirects queries to a set of indexing servers; distributed file systems replicate file chunks to a set of storage nodes, and in MapReduce, the master node distributes tasks to a group of servers for cooperative computations. In group communications, a server as a source node has to transmit one copy of the data to multiple destinations/servers. When the same data is dispersedly transmitted by different links to different destinations, the multicast traffic would occupy a large portion of network resources, which results in network congestion. Thus, it is necessary to efficiently deliver multicast traffic to reduce the data transmission redundancy for data center networks.

According to the measurement results by Microsoft, the data traffic in top-of-rack switches is heavy and may cause serious degradation in network performance [4]. To mitigate the problem in traditional wired data center networks, 60GHz wireless technologies (e.g., 802.11ad), with low cost and high data rate, are introduced to augment network capacity and provide fast connectivity. In wireless data centers, a wireless access point and a wired switch coexist on each top-of-rack. The multicast data through a top-of-rack can be transmitted by either the wireless access point or wired switch. Although the tree structure is an effective topology to deliver multicast data, how to build an efficient tree in wireless data centers is complicated and faces many challenges. The challenges mainly come from the following factors. 1) Since wireless access points are densely deployed in data centers, the interference issue among wireless access points should be carefully considered. 2) Wireless medium is broadcast in nature and more suitable for multicast, compared with wired transmissions. However, unlike a wired switch, an wireless access point could transmit data to more than one access point in its communication range and has more selections for transmission paths, especially when a directional antenna is adopted [4]. 3) The coexistence of wired and wireless links lead to an interesting issue that how to avoid wireless interference by adopting the wired links in wireless data centers such that more wireless access points can be transmitted simultaneously. We will give a simple example in Section III to describe the above mentioned challenge issues in more details.

In this paper, we address the group communication issues raised in wireless data center networks to build multicast trees, comprised of wired and wireless links. The objective is to minimize the total multicast data traffic. The contributions of this paper are as follows. First, we formulate the multicast tree building problem with the consideration of coexisting wired and wireless links in wireless data center networks. Then, we prove that our target problem is $\mathcal{NP}$-hard and propose an efficient heuristic algorithm to solve the problem. Finally, we conduct a series of simulations based on practical parameter settings, measured from real data centers, to evaluate the performance of our proposed algorithm. The simulation results demonstrate that our proposed algorithm is very effective in reducing the total data redundancy of the multicast traffic, compared with an optimal solution designed for traditional wired data centers.

The rest of the paper is organized as follows. In Section II, we review some related works on multicast tree building

978-1-4799-0537-9/13/$31.00 ©2013 IEEE

approaches. Section III describes the system model and formulates the problem. In Section IV, we prove the problem is $\mathcal{NP}$-hard and propose an efficient heuristic algorithm. Simulation results are presented in Section V. Section VI concludes the paper.

II. Related Works

To achieve group communications, multicast is used to transmit data to a group of destinations. For the Internet, the first standard of IP multicast is described in RFC 1112 [5]. Also, the Internet Group Management Protocol (IGMP) is defined to allow a host to join and leave a group, and to report its IP multicast group membership to neighboring multicast routers [6]. For multicast, the tree structure is commonly adopted to reduce redundant data transmissions and avoid unnecessary network resource wastage. The multicast trees for IP multicast can be classified into two types, i.e., source-based and share-based [7]. The source-based tree is established by the shortest-path algorithm, and each sender requires an individual tree to transmit its multicast data. This implies that the source-based multicast tree is more suitable for the applications with few senders in a multicast group. In contrast, just one shared-based tree is needed for a multicast group. Multiple senders in a common multicast group can share the tree. For both source-based and shared-based multicast trees, the tree establishment procedure is generally based on the receiver-driven manner, which would result in extra redundant links especially when there are multiple equal-cost paths between a pair of servers or switches in traditional wired data center networks [8].

On the other hand, for wireless ad-hoc networks, multicast routing has been widely studied and can be roughly classified into tree-based, mesh-based, and hybrid-based approaches. The tree-based approach, e.g., [9, 10], establishes a single path between any two nodes in a multicast group. However, with the mobility of ad-hoc nodes, the tree needs to be frequently re-established due to link failure. The significant overhead for tree maintenance and poor packet delivery ratio make tree-based multicast routing difficult to implement in wireless ad-hoc networks. Thus, some authors, e.g, [11, 12], considered the meshed-based approach with redundant paths to provide robust connectivity for group communications. However, the redundant paths definitely lead to a large amount of redundant multicast data traffic. Consequently, hybrid-based multicast routing protocols, e.g., [13, 14], were proposed. However, the above wireless multicast routing approaches cannot be applied to wireless data center networks. This is because, in wireless data center networks, the deployment density of wireless access points is extremely high and result in serious interference. Also, in wireless data centers, the connectivity under the co-existence of wired and wireless links is much more complicated than that in wireless ad-hoc networks.

Recently, some researches paid attention to multicast issues in traditional wired data centers. In [15], considering the hardware constraint in supporting multicast operations in switches, Vigfusson et al. developed a mechanism to select some of the group communication requests in multicast delivery while the remaining requests are accomplished by unicast transmissions. Then, Li et al. [8, 16] observed that the densely connected data center networks with multiple equal-cost paths given by receiver-driven multicast routing protocols, designed for the Internet, do not perform well in terms of the number of transmission links. Thus, the authors developed efficient multicast routing for wired data center networks to reduce data transmission redundancy. However, the proposed routing approach does not accommodate the wireless links adopted in modern data center networks. Furthermore, the approach only reduces the total number of used wired links, as its major performance metric, without considering different data rates requested by heterogeneous cloud services.

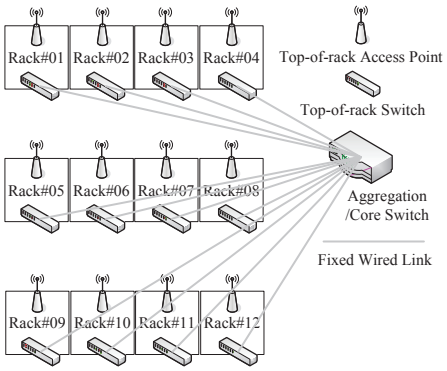


Fig. 1. A simple wireless data center architecture.

III. System Model and Problem Formulation

A. System Model

In traditional data centers, several servers are grouped in a rack and each rack is equipped with a switch. The switch is named as the top-of-rack switch which all the servers in the rack connect to. Top-of-rack switches are generally connected by aggregation switches and/or core switches based on the architectures such as hierarchical topology [17], fat-tree [17], and BCube [18]. However, the deployment cost and complexity of traditional data centers are quite high. Recently, Microsoft adopts the 60GHz wireless access technology to deploy a wireless access point on each top-of-rack switch to augment network capacity and provide fast connectivity [4]. The 60GHz access point can support high data rate with the transmission range of 10 meters. Since the deployment density of access points is extremely high in data centers, the directional antenna with narrow-beam antenna array is adopted to mitigate interference [19]. The illustration of a simple wireless data center architecture is shown in Fig. 1. There are twelve racks, each of which has one top-of-rack switch and one wireless access point. Each top-of-rack switch connects to an aggregate/core switch by the wired line, while each top-of-rack access point can transmit data to any access point within its transmission range.

Multicast data delivery is needed in wireless data center networks since cloud services are mostly provided by the coordinating applications where group communications are necessary. The multicast data delivery is accomplished by the

tree structure. The approaches for constructing multicast trees can be classified into two types [7], source-based and share-based. Since most of the group communications occurring in data centers have only one multicast sender in one multicast group, without loss of generality, this paper considers the source-based tree construction.

B. Problem Formulation

In this paper, we are interested in the source-based multicast tree construction, comprised of wired and wireless links in wireless data center networks. The objective is to minimize the total multicast data traffic (i.e., the total data redundancy). The problem formulation is described as follows. For the sake of brevity, we omit "$\forall$" when the meaning is clear from the context.

A wireless data center is modeled as a directed graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$. The $\mathbb{V} = (\mathbb{V}^F, \mathbb{V}^W)$ is a set of racks. Each rack $v \in \mathbb{V}$ includes one top-of-rack switch $s_v \in \mathbb{V}^F$ and one wireless access point $a_v \in \mathbb{V}^W$. The $\mathbb{V}^F$ is a set of top-of-rack switches and $\mathbb{V}^W$ is a set of top-of-rack access points. The link set $\mathbb{E} = (\mathbb{E}^F, \mathbb{E}^W)$ includes a set of wired (fixed) links $\mathbb{E}^F$ and a set of wireless links $\mathbb{E}^W$. Wired link $e^F_{s_i s_j} \in \mathbb{E}^F$ with capacity $C^F_{s_i s_j}$ (bps), if top-of-rack switch s_i can transmit data to top-of-rack switch s_j by the wired link. On the other hand, wireless link $e^W_{a_x a_y} \in \mathbb{E}^W$ with capacity $C^W_{a_x a_y}$ (bps) means that access point a_x can transmit data to access point a_y by the wireless link.

We consider a set of N multicast groups $\mathbb{R} = (r_1, r_2, \ldots, r_N)$, where $r_k = (\nu_k, \mathbb{D}_k, T_k)$ means that rack ν_k is the sender of multicast group k and has to transmit the multicast traffic with data rate T_k (bps) to a set of destinations (racks) $\mathbb{D}_k \subseteq \mathbb{V}$. Then, we define $l^F(k, e^F_{s_i s_j}) \in \{0, 1\}$ as indicator function, which registers 1 if the traffic of multicast group k passes through wired link $e^F_{s_i s_j}$. If wired link $e^F_{s_i s_j}$ is used and $l^F(k, e^F_{s_i s_j})$ is set at 1, top-of-rack switch s_j of rack $j \in \mathbb{V}$ can receive the multicast data. We also define $l^W(k, e^W_{a_x a_y}) \in \{0, 1\}$ to indicate whether the traffic of multicast group k uses wireless link $e^W_{a_x a_y}$ or not. If wireless link $e^W_{a_x a_y}$ is selected and $l^W(k, e^W_{a_x a_y})$ is set at 1, a set of access points of racks $\mathbb{S}_{a_x a_y} \subset \mathbb{V}$ within the coverage area of the transmission can overhear and receive the data. Our purpose is to build a multicast tree, comprised of wired and wireless links, for each multicast group. The multicast tree building is *feasible* if the following constraints are met.

Wired link capacity constraint: In order to avoid over-utilization of top-of-rack switches, Equation (1) ensures that the data rate of multicast group through each wired link cannot exceed the available capacity of each wired link.

$$\sum_{k=1}^{N} T_k \cdot \left[l^F(k, e^F_{s_i s_j}) + l^F(k, e^F_{s_j s_i}) \right] \leq C^F_{s_i s_j}, \forall e^F_{s_i s_j} \in \mathbb{E}^F \quad (1)$$

Access point capability constraint: Since wireless access points incurs interference from their neighboring access points, Equation (2) states that each access point cannot exceed its capability for data reception and transmission. $I(a_y, e^W_{a_x a_z})$ is used to indicate whether access point a_y is interfered by access point a_x, and defined based on a geometric-based protocol interference model [20]. Based on the protocol interference model, $I(a_y, e^W_{a_x a_z}) = 1$ when access point a_y is located in the transmission range of access point a_x for delivering data to access point a_z.

$$\sum_{k=1}^{N} \sum_{a_x \in \mathbb{V}^W} \sum_{a_z \in \mathbb{V}^W} \frac{I(a_y, e^W_{a_x a_z}) l^W(k, e^W_{a_x a_z}) T_k}{C^W_{a_x a_z}} + \frac{l^W(k, e^W_{a_y a_x}) T_k}{C^W_{a_y a_x}} \leq 1, \forall a_y \in \mathbb{V}^W \quad (2)$$

where

$$I(a_y, e^W_{a_x a_z}) = \begin{cases} 1, & \text{if } y \in \mathbb{S}_{a_x a_z} \\ 0, & \text{otherwise} \end{cases}$$

Delivery constraint: The destinations of each multicast group must receive their multicast data.

$$\left(\bigcup_{l^W(k, e^W_{a_x a_y}) = 1} \mathbb{S}_{a_x a_y} \right) \bigcup \left(\bigcup_{l^F(k, e^F_{s_i s_j}) = 1} j \right) \supseteq \mathbb{D}_k, \forall r_k \in \mathbb{R} \quad (3)$$

We now define the target problem formally as follows.

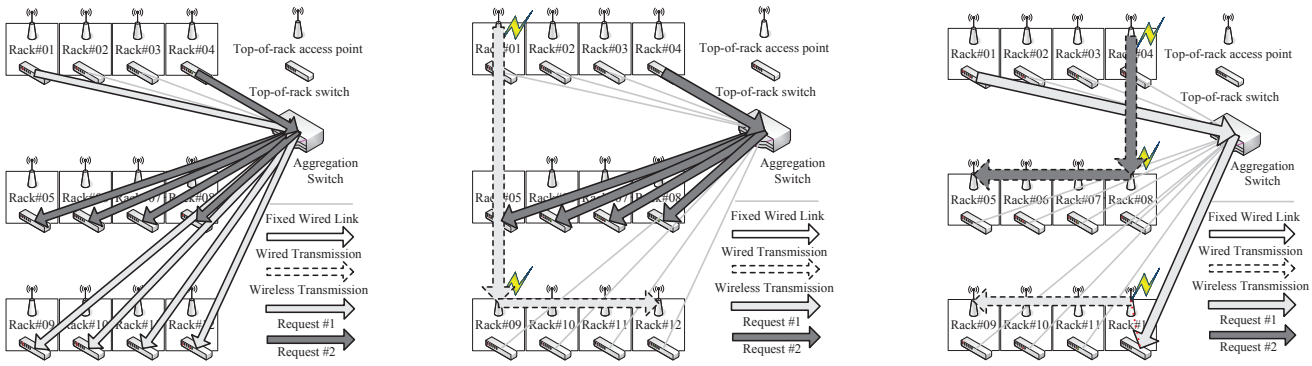
The Efficient Multicast Tree Construction Problem
Input instance: Consider a directed graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$. Each wired and wireless link has its capacity $C^F_{s_i s_j}$ and $C^W_{a_x a_y}$. There is a set of N multicast groups $\mathbb{R}$.
Objective: Our objective is to build a multicast tree, comprised of wired $l^F(k, e^F_{s_i s_j})$ and wireless links $l^W(k, e^W_{a_x a_y})$, for each multicast group such that the multicast data traffic (data redundancy) of all multicast groups is minimized. The objective function is expressed as follows.

$$\texttt{Min} \sum_{k=1}^{N} \sum_{e^F_{s_i s_j} \in \mathbb{E}^F} \sum_{e^W_{a_x a_y} \in \mathbb{E}^W} T_k \times \left[l^F(k, e^F_{s_i s_j}) + l^W(k, e^W_{a_x a_y}) \right],$$

subject to constraints (1)-(3). Table I summarizes the notations used in the problem formulation.

TABLE I
SUMMARY OF NOTATIONS

$\mathbb{V}$	The set of racks
$\mathbb{V}^F$	The set of top-of-rack switches
$\mathbb{V}^W$	The set of top-of-rack access points
$\mathbb{E}^F$	The set of wired links
$\mathbb{E}^W$	The set of wireless links
$C^F_{s_i s_j}$	The available capacity of wired link $e^F_{s_i s_j}$
$C^W_{a_x a_y}$	The available capacity of wireless link $e^W_{a_x a_y}$
$\mathbb{R}$	The set of multicast groups
$\mathbb{D}_k$	The set of destinations of multicast group k
T_k	The data rate of multicast group k
$\mathbb{S}_{a_x a_y}$	The set of access points can overhear the multicast data, when access point a_x transmits data to access point a_y
$l^F(k, e^F_{s_i s_j})$	An indicator function, which is 1 if wired link $e^F_{s_i s_j}$ is allocated to multicast group k, and 0 otherwise
$l^W(k, e^W_{a_x a_y})$	An indicator function, which is 1 if wireless link $e^W_{a_x a_y}$ is allocated to multicast group k, and 0 otherwise



(a) Multicast tree construction I (b) Multicast tree construction II (c) Multicast tree construction III

Fig. 2. An illustrative example for multicast tree construction in wireless data centers.

C. An Illustrative Example

We use a simple example, as shown in Fig. 2, to describe the multicast tree construction problem in wireless data centers. Consider a wireless data center $\mathbb{G}$ shown in Fig. 1. On each rack, there is a pair of top-of-rack switch and access point. The data sent from one top-of-rack switch to another should go through two wired links, while a top-of-rack access point can directly transmit data to another wireless access point. Moreover, since the directional antenna is adopted, the interference range of each access point is limited by its transmission direction [4]. The capacity of each link is set as 1Gbps (i.e., $C^F_{s_i s_j} = C^W_{a_x a_y} = 1\text{G}$, $\forall e^F_{s_i s_j} \in \mathbb{E}^F, e^W a_x a_y \in \mathbb{E}^W$). We consider two multicast groups in this example. For the first multicast group, the sender is placed in rack#01; the set of destinations includes rack#09, 10, 11, and 12; and the data rate of the multicast group is set as 1Gbps. For the second multicast group, the sender is set as rack#04; the set of destinations is rack#05, 06, 07, and 08; and the data rate of the multicast group is 1Gbps as well. Now, we have to build a multicast tree, comprised of wired and wireless links, for each multicast group.

As shown in Fig. 2(a), we only adopt wired links to build multicast trees as it is for traditional data centers. In this case, the senders of top-of-rack switch#01 and 04 first transmit multicast data to the aggregation switch. Then, the aggregation switch has to transmit the same multicast data through four different wired links for the four destinations. For the two multicast trees, the total number of links used is 10 and the total multicast data traffic is 1Gbps$\times$10=10Gbps. We can see that the multicast trees with purely wired links result in severe data redundancy. In Fig. 2(b), when the wireless access points are considered, the multicast data of the first multicast group can be transmitted by the access point of rack#01 to that of rack#09. Then, the wireless access point of rack#09 transmits data to the access point of rack#12. Thus, rack#10, 11, and 12 can simultaneously receive the multicast data. This multicast tree only uses two wireless links. For the second multicast group, since the access point of rack#05 is interfered by the wireless transmission of the access point on rack#01, the multicast data is selected to be transmitted by the wired links and occupies five wired links. The total multicast data traffic of the two multicast trees is 1Gbps$\times$7=7Gbps. Actually, we have a better option to build the multicast trees as shown in Fig. 2(c). Interestingly, we can utilize the wired links to avoid wireless interference such that more wireless access points can be simultaneously transmitted to further reduce the data redundancy. The data of the first multicast group can pass through the aggregation switch from rack#01 to rack#12. Then, the wireless access point of rack#12 is adopted and relays the data to the rack#09. The multicast tree for the first multicast group is comprised of two wired links and one wireless link. Then, the multicast data of the second multicast group can be transmitted by the two wireless access points on rack#04 and #08. The total data traffic for the two group communications is 1Gbps$\times(3+2)$=5Gbps.

IV. Efficient Multicast Data Delivery

In this section, we prove the $\mathcal{NP}$-hardness of the problem by a reduction from the *partition problem*, which is known to be $\mathcal{NP}$-complete [21], and propose an efficient heuristic algorithm to solve the problem.

A. Problem Hardness

Theorem 1. The efficient multicast tree building problem is $\mathcal{NP}$-hard.

Proof*:* The input instance of the partition problem is a set of M integers, $\mathbb{B} = \{b_1, b_2, \ldots, b_M\}$. The output is YES if and only if $\mathbb{B}$ can be partitioned into two subsets $\mathbb{U}$ and $\mathbb{B}\backslash\mathbb{U}$ with the same sum, i.e., $\sum_{b_m \in \mathbb{U}} b_m = \sum_{b_m \notin \mathbb{U}} b_m = \frac{1}{2}\sum_{b_m \in \mathbb{B}} b_m$.

Given an instance $\langle \mathbb{B} \rangle$ of the partition problem, we explain how to construct an instance $\langle \mathbb{G}, C^F_{s_i s_j}, C^W_{a_x a_y}, \mathbb{R}, N \rangle$ of our problem in polynomial time such that $\mathbb{B}$ can be evenly partitioned if and only if there exist M multicast trees with total data traffic $\sum_{b_m \in \mathbb{B}} b_m$. The construction is as follows: In a wireless data center, there are two racks with two top-of-rack switches $|\mathbb{V}^F| = 2$ and two top-of-rack access points $|\mathbb{V}^W| = 2$. A rack is connected to the other rack with only one wired link and one wireless link (i.e., $\mathbb{E}^{\mathbb{F}} = \{e^F_{s_1 s_2}, e^F_{s_2 s_1}\}$ and $\mathbb{E}^{\mathbb{W}} = \{e^W_{a_1 a_2}, e^W_{a_2 a_1}\}$). The capacity of each wired and wireless link is set at $\frac{1}{2}\sum_{b_m \in \mathbb{B}} b_m$ (i.e, $C^F_{s_1 s_2} = C^F_{s_2 s_1} = C^W_{a_1 a_2} = C^W_{a_2 a_1} = \frac{1}{2}\sum_{b_m \in \mathbb{B}} b_m$. There is a set of M multicast groups (i.e., $N = M$). The multicast data of M multicast groups are all transmitted from the same rack (source) to the

other rack (destination). The data rate of multicast group m is set as $T_m = b_m, \forall 1 \leq m \leq M$.

To complete the proof, we show that two partitioned subsets can be used to derive M multicast trees whose total data traffic is $\sum_{b_m \in \mathbb{B}} b_m$, and vice versa. If there are two partitioned subsets, each integer b_m corresponds to the data rate T_m required by multicast group m; a subset corresponds to the data rate of the multicast groups allocated to the wired link and the other subset corresponds to the data rate of the other multicast groups transmitted by the wireless link. Thus, the total data traffic of M multicast trees is $\sum_{b_m \in \mathbb{B}} b_m$. On the other hand, if the total data traffic of M multicast trees is $\sum_{b_m \in \mathbb{B}} b_m$, the wired link and wireless link have to respectively transmit the data rate of $\frac{1}{2}\sum_{b_m \in \mathbb{B}} b_m$. It implies that the set can be evenly partitioned by assigning the corresponding integers into the corresponding subset. The existence of a polynomial-time algorithm for the partition problem implies the same for ours, which completes the proof. ■

B. Algorithm Description

In this section, we propose an efficient algorithm for building multicast trees, comprised of wired and wireless links, for all multicast groups. The concept of this algorithm is to find some wireless access points that can cover as more destinations as possible to reduce the data redundancy of multicast traffic. Then, we find shortest paths, comprised of wired and wireless links, to connect each source with its destinations. Moreover, in order to use as few number of links as possible, for each shortest path, we will try to use wireless links first. If the wireless link cannot support the data transmission, we will utilize the wired link instead. Moreover, in order to efficiently utilize each link capacity, we will give a higher priority for the multicast group with a higher data rate to construct the multicast tree.

The pseudo-code of the proposed algorithm is shown in Algorithm 1. In Line 1, an indicator function $l^F(k, e^F_{s_i s_j})$ is used to record whether wired link $e^F_{s_i s_j}$ is allocated for transmitting the data of multicast group k, and is initialized as 0, $\forall 1 \leq k \leq N, e^F_{s_i s_j} \in \mathbb{E}^F$. In Line 2, an indicator function $l^W(k, e^W_{a_x a_y})$ is used to record whether wireless link $e^W_{a_x a_y}$ is allocated to transmit the data of multicast group k, and is initialized as 0, $\forall 1 \leq k \leq N, e^W_{a_x a_y} \in \mathbb{E}^W$. In Line 3, a variable P_k, initialized as 0, is used record the priority of multicast group k. If multicast group k has a higher value of P_k, we have a higher priority to build a multicast tree for the multicast group. In Line 4, a set $\hat{\mathbb{E}}^W_k$ is used to record which wireless links can be adopted for delivering the traffic of multicast group k. In Line 5, a set $\hat{\mathbb{S}}^W_k$ is adopted to record how many destinations of multicast group k can overhear the multicast data transmitted by the access points of the destinations (racks). In Line 6, a set $\hat{\mathbb{D}}_k$ is used to register which destinations of multicast group k can receive the data and initialized as $\emptyset$.

Then, the algorithm starts to construct a multicast tree, comprised of wireless and wired links, for each multicast group (Lines 7-29). For each multicast group k, since the directional antenna with narrow-beam is generally adopted by wireless data centers, we let each wireless access point a_x, $\forall x \in \mathbb{D}_k \bigcup \nu_k$, attempt to transmit the data of multicast group k to each wireless access point $a_y, \forall y \in \mathbb{D}_k \bigcup \nu_k$, and compute how many destinations can receive the data (Lines 7-13). In Lines 10-11, if access point a_x of rack x can transmit the data to access point a_y of rack y (i.e., $e^W_{a_x a_y} \in \mathbb{E}^W$), a set of destinations can receive the data $(i.e., \mathbb{S}^W_{a_x a_y} \bigcap \mathbb{D}_k)$; and the set $\hat{\mathbb{S}}^W_k$ is updated to $\hat{\mathbb{S}}^W_k \bigcup (\mathbb{S}^W_{a_x a_y} \bigcap \mathbb{D}_k)$. In Line 12, the wireless link $e^W_{a_x a_y}$ that can be used for transmitting the data of multicast group k is added into the set $\hat{\mathbb{E}}^W_k$. When all pairs of the access points of destinations are tried out, the priority P_k of multicast group k is set as $T_k \times |\hat{\mathbb{S}}^W_k|$ (Line 13). That is, if more destinations can overhear the data transmitted by

Algorithm 1

Input: $\mathbb{G}$, $C^F_{s_i s_j}$, $C^W_{a_x a_y}$, $\mathbb{R}$, N
Output: $l^F(k, e^F_{s_i s_j}), l^W(k, e^W_{a_x a_y})$

1: $l^F(k, e^F_{s_i s_j}) \leftarrow 0, \forall 1 \leq k \leq N, e^F_{s_i s_j} \in \mathbb{E}^F$
2: $l^W(k, e^W_{a_x a_y}) \leftarrow 0, \forall 1 \leq k \leq N, e^W_{a_x a_y} \in \mathbb{E}^W$
3: $P_k \leftarrow 0, \forall 1 \leq k \leq N$
4: $\hat{\mathbb{E}}^W_k \leftarrow \emptyset, 1 \leq k \leq N$
5: $\hat{\mathbb{S}}^W_k \leftarrow \emptyset, \forall 1 \leq k \leq N$
6: $\hat{\mathbb{D}}_k \leftarrow \emptyset, \forall 1 \leq k \leq N$
7: **for** $k = 1$ to N **do**
8: **for all** $x \in (\mathbb{D}_k \bigcup \nu_k)$ **do**
9: **for all** $y \in (\mathbb{D}_k \bigcup \nu_k)$ **do**
10: **if** $e^W_{a_x a_y} \in \mathbb{E}^W$ **then**
11: $\hat{\mathbb{S}}^W_k \leftarrow \hat{\mathbb{S}}^W_k \bigcup (\mathbb{S}^W_{a_x a_y} \bigcap \mathbb{D}_k)$
12: $\hat{\mathbb{E}}^W_k \leftarrow \hat{\mathbb{E}}^W_k \bigcup e^W_{a_x a_y}$
13: $P_k \leftarrow T_k \times |\hat{\mathbb{S}}^W_k|$
14: Re-arrange the multicast group indexes by decreasing the priority of $P_k, \forall 1 \leq k \leq N$, such that $P_1 \geq P_2 \cdots \geq P_N$
15: **for** $k = 1$ to N **do**
16: Re-arrange the wireless link indexes by decreasing the $(\mathbb{S}^W_{a_x a_y} \bigcap \mathbb{D}_k)$, $\forall e^W_{a_x a_y} \in \hat{\mathbb{E}}^W_k$
17: **for all** $e^W_{a_x a_y} \in \hat{\mathbb{E}}^W_k$ **do**
18: **if** the access point capability constraint is satisfied and $|\mathbb{D}_k \bigcap \mathbb{S}_{a_x a_y}| \geq 2$ and $\hat{\mathbb{D}}_k \bigcap \mathbb{S}_{a_x a_y} = \emptyset$ **then**
19: $\hat{\mathbb{D}}_k \leftarrow \hat{\mathbb{D}}_k \bigcup x$
20: $l^W(k, e^W_{a_x a_y}) \leftarrow 1$
21: SHORTEST-PATH(ν_k, x)
22: **for all** $v \in \mathbb{D}_k \bigcap \mathbb{S}_{a_x a_y}$ **do**
23: **if** the access point capability constraint is satisfied **then**
24: $\hat{\mathbb{D}}_k \leftarrow \hat{\mathbb{D}}_k \bigcup v$
25: **else**
26: Build a shortest path by wired links from ν_k to v and set corresponding $l^F(k, e^F_{s_i s_j})$ as 1
27: $\hat{\mathbb{D}}_k \leftarrow \hat{\mathbb{D}}_k \bigcup v$
28: **if** $\mathbb{D}_k \setminus \hat{\mathbb{D}}_k \neq \emptyset$ **then**
29: SHORTEST-PATH(ν_k, $\mathbb{D}_k \setminus \hat{\mathbb{D}}_k$)
30: **return** $l^W(k, e^W_{a_x a_y})$ and $l^F(k, e^F_{s_i s_j})$, $\forall$ $e^W_{a_x a_y} \in \mathbb{E}^W, e^F_{s_i s_j} \in \mathbb{E}^F$

the wireless access points and the traffic of multicast group k has a higher data rate, more data redundancy can be reduced. Thus, we give a higher priority for the multicast group to build multicast tree and to use wireless access points.

After the priorities of all multicast groups are set, we re-arrange the multicast group indexes by decreasing the property of P_k, $\forall 1 \leq k \leq N$, such that $P_1 \geq P_2 \cdots \geq P_N$ (Line 14). Then, we start to build a multicast tree for each multicast group and adopt the new index of multcast group, i.e., multicast group $k = 1$ has the highest property P_1 (Lines 15-29). For multicast group k, we re-arrange the wireless link indexes $e^W_{a_x a_y} \in \hat{\mathbb{E}}^W_k$ by decreasing the $(\mathbb{S}^W_{a_x a_y} \bigcap \mathbb{D}_k)$ in order to select the wireless links covering as more destinations as possible (Line 16). Then, for each wireless link $e^W_{a_x a_y} \in \hat{\mathbb{E}}^W_k$, we select access point a_x transmitting data to access point a_y if the following three conditions are met (Lines 17-18): 1) the access point can meet its capability constraint; 2) at least two destinations can simultaneously receive the multicast data (i.e., $|\mathbb{D}_k \bigcap \mathbb{S}_{a_x a_y}| \geq 2$); and 3) each destination of multicast group k cannot receive the same multicast data from more than one link in order to meet the tree properties (i.e., $\hat{\mathbb{D}}_k \bigcap \mathbb{S}_{a_x a_y} = \emptyset$). If the link is adopted, we add destination (rack) x, which can receive data, to the registered destination set $\hat{\mathbb{D}}_k$ (i.e., $\hat{\mathbb{D}}_k = \hat{\mathbb{D}}_k \bigcup x$) (Line 19) and the indicator function $l^W(k, e^W_{a_x a_y})$ is set as 1 accordingly (Line 20). Although the wireless link $e^W_{a_x a_y}$ is adopted and can transmit data to some destinations, access point a_x does not have a path to receive the multicast traffic from sender ν_k. Then, we find a shortest path, comprised of wired and wireless links, for the given pair of source ν_k and access point a_x of rack x. Whenever Procedure SHORTEST-PATH() is invoked, it attempts to find a shortest path from source ν_k of multicast group k to destination x through as few links as possible (Line 21). For the path, we try to use wireless links first. If the wireless links do not satisfy the access point capability constraint, we adopt wired links instead. Then, the corresponding indicator functions $l^W(k, e^W_{\tilde{x}\tilde{y}})$ and $l^F(k, e^F_{\tilde{i}\tilde{j}})$ are set as 1.

In Lines 22-27, although the access point a_v of the destination rack v can overhear the wireless transmission (i.e., $v \in \mathbb{D}_k \bigcap \mathbb{S}_{a_x a_y}$), it may not have enough capability to receive the data. Therefore, if the access point has capability to receive the data, we directly add the destination of rack v to the registered destination set $\hat{\mathbb{D}}_k$ (Line 24). Otherwise, we build a shortest path by wired links from sender ν_k to destination v and set corresponding $l^F(k, e^F_{s_i s_j})$ as 1 (Line 26). The destination of rack v is also added to the registered destination set $\hat{\mathbb{D}}_k$ (Line 27). Finally, if there are some remaining destinations that have no path to receive multicast data (i.e., $\mathbb{D}_k \setminus \hat{\mathbb{D}}_k \neq \emptyset$), we use Procedure SHORTEST-PATH() to find a shortest path for each remaining destination of multicast group k (Lines 28-29). Finally, we return a multicast tree, comprised of wireless and wired links, for each multicast group (Line 30).

Theorem 2. The time complexity of Algorithm 1 is $O(N\tilde{D}(\tilde{E}\omega + \tilde{D}))$. $\tilde{D} = \max_{\forall\ k} |\mathbb{D}_k|$; $\tilde{E} = \max(|\mathbb{E}^W|, |\mathbb{E}^F|)$, where ω is the running time of the shortest path algorithm.

Proof*:* The initialization process requires $O(N\tilde{E})$ time. For each multicast group k, a priority P_k is computed only once and can be done in $O(\tilde{D}^2)$. Thus, for N multicast groups, the algorithm takes $O(N\tilde{D}^2)$ time. For building a multicast tree of group k, there are at most $\tilde{D}$ destinations and $\tilde{E}$ links; and Procedure SHORTEST-PATH() is used only once for each destination. Building multicast trees for N multicast groups takes $O(N\tilde{E}\tilde{D}\omega)$. Thus, the time complexity of Algorithm 1 is $O(N\tilde{D}(\tilde{E}\omega + \tilde{D}))$. ■

V. Performance Evaluation

A. Simulation Setups

This section developed a simulation model based on a realistic wireless data center topology to evaluate our proposed algorithm, named as *Efficient Wireless Data Center Multicast Tree (EWDCMT)*, in comparison with two algorithms. The first algorithm, denoted as *steiner-tree*, was designed for wired data center networks; the algorithm can obtain an optimal multicast tree for each multicast group regardless of the link capacity constraint of each wired link. Therefore, for the performance comparison, we relax the constraint for *steiner-tree*. Note that relaxing the constraint is beneficial for the performance of *steiner-tree*. The second algorithm, represented as *shortest-path-tree*, was designed as a baseline. The algorithm built shortest path trees with the consideration of wired and wireless links in wireless data centers. For each shortest path tree, the algorithm attempts to use wireless links first. Until the available capacity of an access point is exhausted, the algorithm will adopt wired links instead. The performance metric was the total amount of transmitted data traffic for all multicast groups.

We consider a wireless data center with hierarchical topology according to the deployment of Microsoft [19]. In the wireless data center, there are 160 top-of-racks, each of which has one wired switch and one 60GHz wireless access point with a directional narrow-beam antenna. From real measurement results by Microsoft, two parallel wireless links are interfered with each other when the distance of the two links is smaller than 22 inches. Note that the width of a rack is about 24 inches [19]. Then we can calculate the transmission/interference range of wireless access points. The maximal capacity of each link is set as 1Gbps if background traffic (BT) is not considered. The impacts of heavy and light background traffic on the total multicast data traffic were investigated in experiments. Based on the real measurement results on traffic distribution in data centers, the available capacity of each link is randomly assigned from 300Mbps to 1000Mbps with uniform distribution [22] when the heavy background traffic is considered. In the case of light background traffic, we set the available capacity of each link randomly from 700Mbps to 1000Mbps.

Furthermore, we observed the impact of the number of multicast groups varying from 50 to 200 [8]. For each multicast group, the source and destinations are randomly selected from 160 top-of-racks [8]. For generating the number of destinations of each multicast group, we consider two different distributions [16]. The first one is uniform distribution between 3 to 160. The other one is power-law distribution, which

TABLE II
PARAMETER SETTINGS

Parameter	Value
Number of top-of-racks	160
Transmission range of a 60GHz wireless access point	10 meters
Available link capacity without background traffic	1 Gbps
Available link capacity with light background traffic	700-1000 Mbps
Available link capacity with heavy background traffic	300-1000 Mbps
Number of multicast groups	50-200
Data rate of each multicast group	1-100000 Kbps

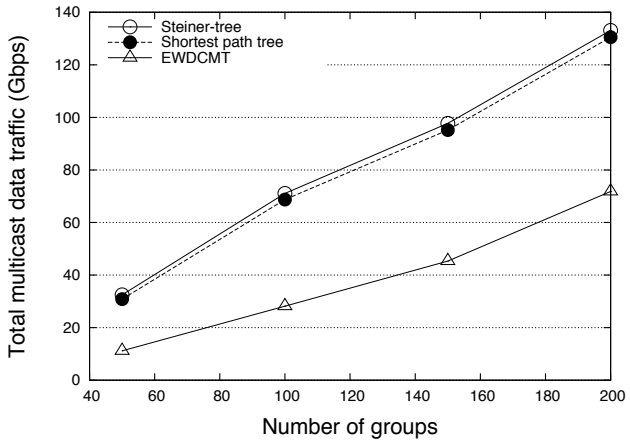
generates more small groups in the data center. Specifically, for power-law distribution, the number of destinations for each multicast group k is generated based on the following probability.

$$P(|\mathbb{D}_k|) = |\mathbb{D}_k|^{-\alpha} \quad (4)$$

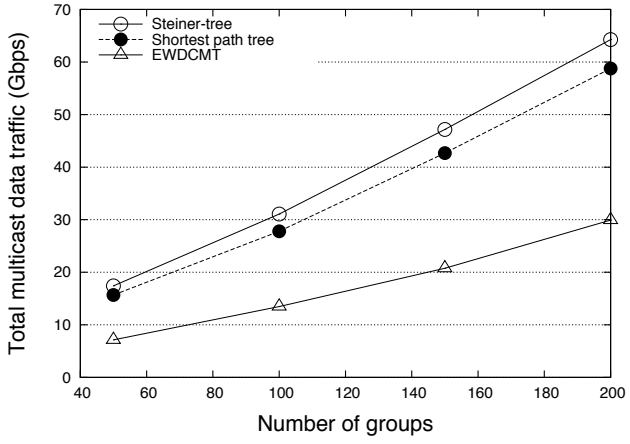
where the value of $|\mathbb{D}_k|$ is between 3 to 160, and α is set at -1 [16]. The data rate of each multicast group is set based on real data flows in data centers, measured by Microsoft [23]. That is, for the data rate of each multicast group, we roughly select one of the following six data rates, 1, 10, 100, 1000, 10000, 100000kbps, with the corresponding probabilities 0.1, 0.3, 0.2, 0.2, 0.15, and 0.05. The simulation parameters are listed in Table II. The derived simulation result is the average of the output values of 100 independent runs.

B. Simulation Results

Fig. 3 shows the impacts of the number of multicast groups under different group size distributions on the total multicast data traffic. As shown in the figure, the total multicast data traffic increases when the number of multicast groups increases under the three algorithms. The result was as expected because more multicast groups lead to more multicast data traffic and occupy more network resources. Our proposed algorithm can more efficiently reduce the total multicast data traffic, compared with *steiner tree* and *shortest path tree*. Comparing Fig. 3(a) with Fig. 3(b), the performance of *shortest path tree* is more closed to that of *steiner tree* when the uniform group size distribution is considered. The reason is that because each multicast group with the uniform group size has more members (destinations) and each member is randomly placed in the wireless data center, *shortest path tree* may rapidly exhaust the capacity of each wireless link. Thus, wired links are used instead and the performance of *shortest path tree* is similar to that of *steiner tree*. In contrast, *EWDCMT* can significantly reduce data redundancy, compared with *steiner tree* and *shortest path tree*, under the uniform group size distribution than under the power-law group size distribution. This is because our algorithm efficiently uses each wireless link and attempts to find each access point that can transmit data to as more destinations as possible. Therefore, when each multicast group has more destinations, our algorithm more efficiently utilizes the broadcast advantage of wireless medium for multicast transmissions and evidently reduce the data redundancy of multicast traffic. The simulation results show that *EWDCMT* can reduce the total data traffic, compared with *steiner-tree* and *shortest path tree*, from 45% to 72% under the uniform group size shown in Fig. 3(a) and from 49% to 55% under the power-law group size distribution.



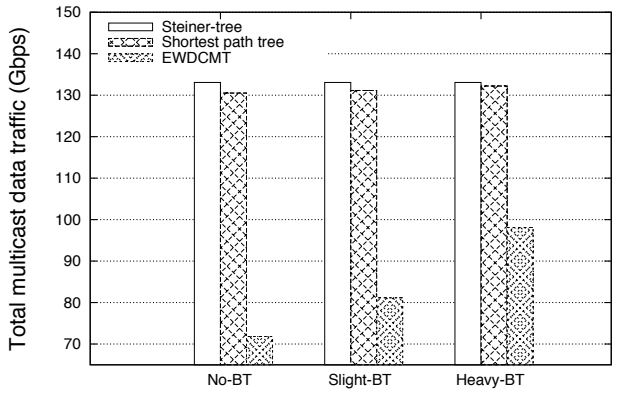
(a)



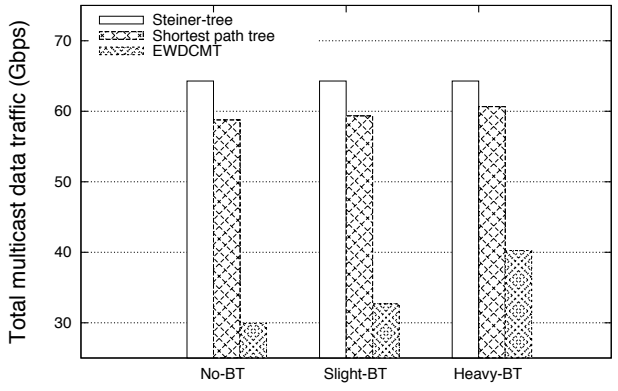
(b)

Fig. 3. Impacts of the number of multicast groups under (a) the uniform group size distribution and (b) the power-law group size distribution on the total multicast data traffic.

Fig. 4 shows the impacts of different background traffic levels on the total multicast data traffic. As we can see in this figure, the total multicast data traffic is higher, when the background traffic is heavier, under *shortest path tree* and *EWDCMT*. The reason is that when the background traffic is heavier, efficient wireless links for each multicast group may be unavailable. In order to avoid over utilization, the two algorithms must select other inefficient wireless/wired links for building multicast trees such that fewer data redundancy can be reduced. This also explains why the performance of *EWDCMT* is closed to those of *shortest path tree* and *steiner-tree* when the background traffic is heavy. On the other hand, the different levels of background traffic do not have any impact for *steiner-tree*, since *steiner-tree* does not consider the link capacity constraint of wired links. Comparing Fig. 4(a) with Fig. 4(b), the result is similar to that in Fig. 3. The performance of our proposed algorithm, compared with *steiner-tree* and *shortest path tree*, is more efficient in reducing total multicast data traffic under the uniform group size distribution, shown in 4(a),



(a)



(b)

Fig. 4. Impacts of the number of multicast groups for (a) the uniform group size distribution and (b) the power-law group size distribution on the total multicast data traffic under 200 multicast groups.

than under the power-law group size distribution, shown in 4(b). The simulation results show that *EWDCMT* outperforms *steiner-tree* and *shortest path tree*. The reduction is about 44% under the uniform group size distribution and is about 36% under the power-law group size distribution. The results also identify our motivations that constructing multicast trees in wireless data centers with densely deployed access points is not a trivial problem.

VI. Conclusion

In this paper, we have addressed the group communication issue raised in wireless data center networks. Specifically, we studied the multicast tree construction problem with the coexistence of wired and wireless links. The objective is to minimize the total multicast data traffic (i.e., total multicast data redundancy). We proved $\mathcal{NP}$-hardness of the target problem and proposed an efficient heuristic algorithm to solve the problem. For performance evaluation, we conducted simulations based on practical parameter settings measured from real data centers. The simulation results demonstrated that the proposed algorithm is very effective in reducing total multicast data traffic, compared with an optimal solution for traditional wired data centers and a baseline designed for wireless data centers.

Acknowledgement

This work was supported in part by Excellent Research Projects of National Taiwan University under Grant 102R890822, by National Science Council under Grant NSC102-2221-E-002-075-MY2 and Grant NSC101-2221-E-002-018-MY2, Chunghwa Telecom, ICL/ITRI, and Research Center for Information Technology Innovation, Academia Sinica.

References

[1] H. G. S. Ghemawat and S. Leungm, "The Google File System," *ACM SOSP*, 2003.
[2] F. Chang et. al, "Bigtable: A Distributed Storage System for Structured Data," *OSDI*, pp. 1–14, 2006.
[3] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *OSDI*, pp. 137–149, 2004.
[4] S. Kandula, J. Padhye, and P. Bahl, "Flyways To De-Congest Data Center Networks," *ACM Workshop on Hot Topics in Network*, pp. 1–6, 2009.
[5] S. Deering, "Host Extensions for IP Multicasting," *RFC 1112*, 1989.
[6] I. K. B. F. B. Cain, S. Deering and A. Thyagarajan, "Internet Group Management Protocol," *RFC 3376*, 2002.
[7] J. W. Y. Yang and M. Yang, "A Service-Centric Multicast Architecture and Routing Protocol," *IEEE Trans. on Parallel and Distributed Systems*, vol. 19, no. 1, pp. 35–51, 2008.
[8] D. Li, J. Yu, J. Yu, and J. Wu, "Exploring Efficient and Scalable Multicast Routing in Future Data Center Networks," *IEEE INFOCOM*, pp. 1368–1376, 2011.
[9] J. J. Garcia-Luna-Aceves and E. L. Madruga, "The Core-assisted Mesh Protocol," *IEEE Journal on Selected Areas in Commun.*, vol. 17, no. 8, pp. 1380–1394, 1999.
[10] J. Xie et al., "AMRoute: Ad-hoc Multicast Routing Protocol," *Mobile Networks and Applications, Multipoint Communication in Wireless Mobile Networks (WCMC)*, vol. 7, no. 6, pp. 429–439, Dec. 2002.
[11] K. Chen and K. Nahrstedt, "Effective Location-uided Tree Construction Algorithms for Small Group Multicast in MANET," *IEEE INFOCOM*, pp. 1180–1189, 2002.
[12] H. Dhillon and H. Q. Ngo, "CQMP: A Mesh-based Multicast Routing Protocol with Consolidated Query Packets," *IEEE WCNC*, pp. 2168–2174, 2005.
[13] L. Ji and M. S. Corson, "Differential Destination Multicast-A MANET Multicast Routing Protocol for Small Groups," *IEEE INFOCOM*, pp. 1192–1201, 2001.
[14] M. B. J. Biswas and S. K. Nandy, "Efficient Hybrid Multicast Routing Protocol for Ad-hoc Wireless Networks," *IEEE LCN*, pp. 180–187, 2004.
[15] H. A.-L. Y. Vigfusson and M. Balakrishnan, "Dr. Multicast: Rx for Data Center Communication Scalability," *ACM Eurosys*, pp. 349–362, 2010.
[16] Dan Li and Yuanjie Li and Jianping Wu and Sen Su and Jiangwei Yu, "ESM: Efficient and Scalable Data Center Multicast Routing," *IEEE Trans. on Networking*, vol. 20, no. 3, pp. 944–955, June 2012.
[17] A. L. M. Al-Fares and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *ACM SIGCOMM*, 2008.
[18] C. G. et al., "BCube: A High Performance, Servercentric Network Architecture for Modular Data Centers," *ACM SIGCOMM*, 2009.
[19] D. Halperin, S. Kandula, J. Padhye, P. Bahl, and D. Wetherall, "Augmenting Data Center Networks with Multi-Gigabit Wireless Links," *ACM SIGCOMM*, pp. 38–49, 2011.
[20] P. Gupta and P. R. Kumar, "The Capacity of Wireless Networks," *IEEE Trans. on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.
[21] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of $\mathcal{NP}$-Completeness*, 1st ed. New York: W. H. Freeman Co., Jan. 1979.
[22] A. A. T. Benson, A. Anand and M. Zhang, "Understanding Data Center Traffic Characteristics," *ACM SIGCOMM*, pp. 92–99, 2010.
[23] A. G. P. P. S. Kandula, S. Sengupta and R. Chaiken, "The Nature of Datacenter Traffic: Measurements & Analysis," *ACM SIGCOMM*, pp. 202–208, 2009.